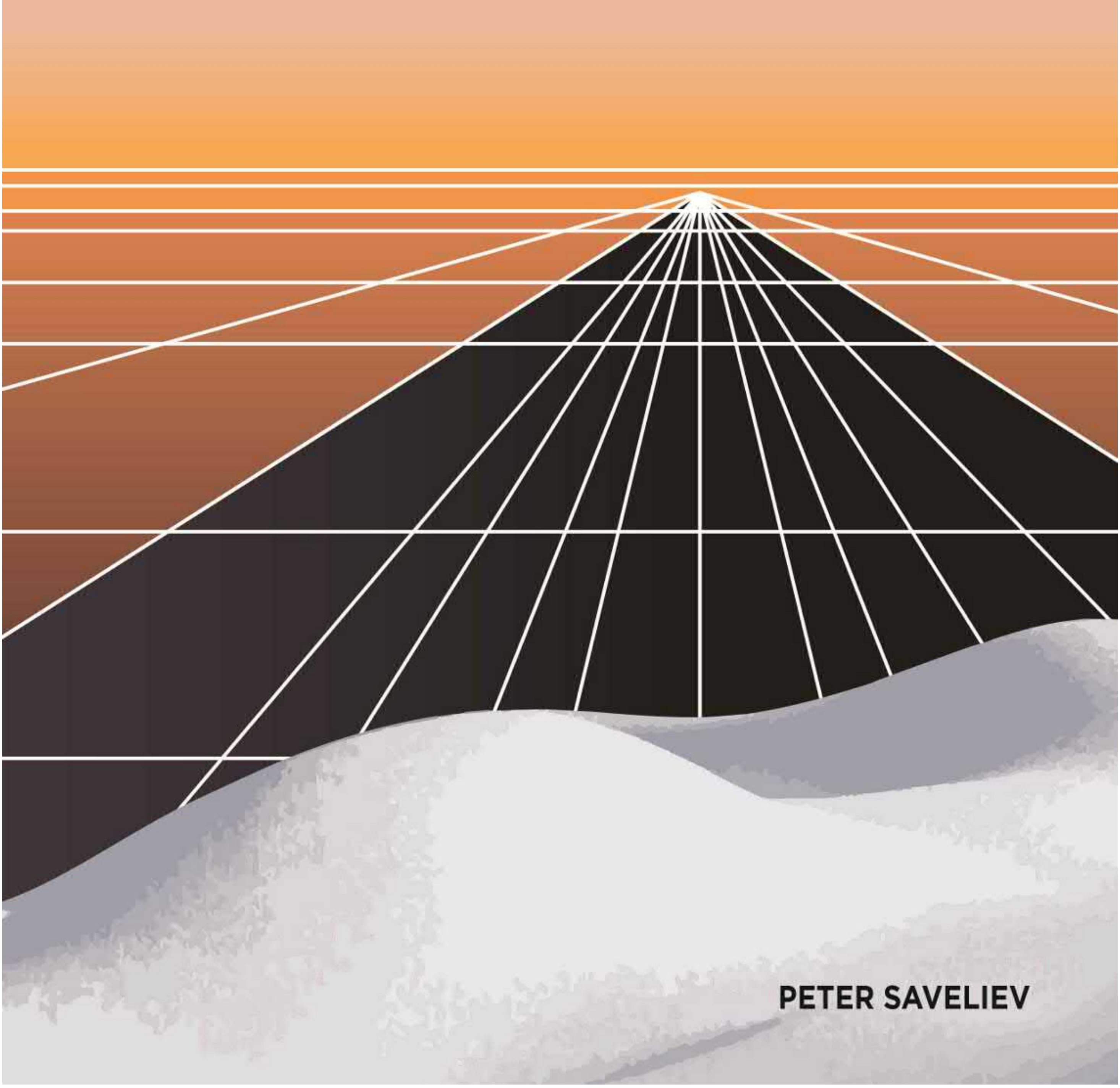
LINEAR ALGEBRA
ILLUSTRATED



PETER SAVELIEV

Contents

To the student

Mathematics is a science. Just as the rest of the scientists, mathematicians are trying to understand how the Universe operates and discover its laws. When successful, they write these laws as short statements called "theorems". In order to present these laws conclusively and precisely, a dictionary of the new concepts is also developed; its entries are called "definitions". These two make up the most important part of any mathematics book.

This is how definitions, theorems, and some other items are used as building blocks of the scientific theory we present in this text.

Every new concept is introduced with utmost specificity.

Definition 0.0.1: square root

Suppose a is a positive number. Then the *square root* of a is a positive number x, such that $x^2 = a$.

The term being introduced is given in *italics*. The definitions are then constantly referred to throughout the text.

New symbolism may also be introduced.

Square root

$$\sqrt{a}$$

Consequently, the notation is freely used throughout the text.

We may consider a specific instance of a new concept either before or after it is explicitly defined.

Example 0.0.2: length of diagonal

What is the length of the diagonal of a 1×1 square? The square is made of two right triangles and the diagonal is their shared hypotenuse. Let's call it a. Then, by the *Pythagorean Theorem*, the square of a is $1^2 + 1^2 = 2$. Consequently, we have:

$$a^2 = 2\,.$$

We immediately see the need for the square root! The length is, therefore, $a = \sqrt{2}$.

You can skip some of the examples without violating the flow of ideas, at your own risk.

All new material is followed by a few little tasks, or questions, like this.

Exercise 0.0.3

Find the height of an equilateral triangle the length of the side of which is 1.

The exercises are to be attempted (or at least considered) immediately.

Most of the in-text exercises are not elaborate. They aren't, however, entirely routine as they require understanding of, at least, the concepts that have just been introduced. Additional exercise *sets* are placed in the appendix. Do not start your study with the exercises! Keep in mind that the exercises are meant to test – indirectly and imperfectly – how well the *concepts* have been learned.

There are sometimes words of caution about common mistakes made by the students.

Warning!

In spite of the fact that $(-1)^2 = 1$, there is only one square root of 1, $\sqrt{1} = 1$.

The most important facts about the new concepts are put forward in the following manner.

Theorem 0.0.4: Product of Roots

For any two positive numbers a and b, we have the following identity:

$$\boxed{\sqrt{a} \cdot \sqrt{b} = \sqrt{a \cdot b}}$$

The theorems are constantly referred to throughout the text.

As you can see, theorems may contain formulas; a theorem supplies limitations on the applicability of the formula it contains. Furthermore, every formula is a part of a theorem, and using the former without knowing the latter is perilous.

There is no need to memorize definitions or theorems (and formulas), initially. With enough time spent with the material, the main ones will eventually become familiar as they continue to reappear in the text. Watch for words “important”, “crucial”, etc. Those new concepts that do not reappear in this text are likely to be seen in the next mathematics book that you read. You need to, however, be aware of all of the definitions and theorems and be able to find the right one when necessary.

Often, but not always, a theorem is followed by a thorough argument as a justification.

Proof.

Suppose $A = \sqrt{a}$ and $B = \sqrt{b}$. Then, according to the definition, we have the following:

$$a = A^2 \quad \text{and} \quad b = B^2 \ .$$

Therefore, we have:

$$a \cdot b = A^2 \cdot B^2 = A \cdot A \cdot B \cdot B = (A \cdot B) \cdot (A \cdot B) = (AB)^2 \ .$$

Hence, $\sqrt{ab} = A \cdot B$, again according to the definition.

Some proofs can be skipped at first reading.

Its highly detailed exposition makes the book a good choice for *self-study*. If this is your case, these are my suggestions.

While reading the book, try to make sure that you understand new concepts and ideas. Keep in mind, however, that some are more important that others; they are marked accordingly. Come back (or jump forward) as needed. Contemplate. Find other sources if necessary. You should not turn to the exercise sets until you have become comfortable with the material.

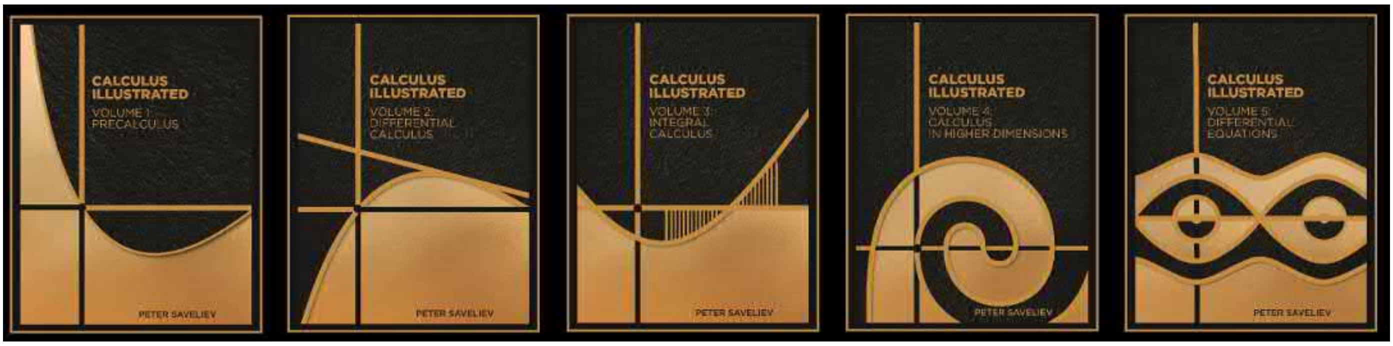
What to do about exercises when solutions aren’t provided? First, use the examples. Many of them contain a problem – with a solution. Try to solve the problem – before or after reading the solution. You can also find exercises online or make up your own problems and solve them!

I strongly suggest that your solution should be thoroughly *written*. You should write in complete sentences, including all the algebra. The standards of thoroughness are provided by the examples in the book.

Next, your solution should be thoroughly *read*. This is the time for self-criticism: Look for errors and weak spots. It should be re-read and then rewritten. Once you are convinced that the solution is correct and the presentation is solid, you may show it to a knowledgeable person for a once-over.

About the author

Peter Saveliev is a professor of mathematics at Marshall University, Huntington, West Virginia, USA. After a Ph.D. from the University of Illinois at Urbana-Champaign, he devoted the next 20 years to teaching mathematics. Peter is the author of a graduate textbook *Topology Illustrated* published in 2016 and a series *Calculus Illustrated* that started to appear in 2019. He has also been involved in research in algebraic topology and several other fields. His non-academic projects have been: digital image analysis, automated fingerprint identification, and image matching for missile navigation/guidance.



About the cover. A person standing on the top of a mountain with the sun behind him is likely to see the mountain's shadow as an almost perfect triangle.

Chapter 1: Sets and functions

Contents

1.1. Introduction to linear algebra

► PROBLEM: Suppose we have the Kenyan coffee that costs \$2 per pound and the Colombian coffee that costs \$3 per pound. How much of each do you need to have 6 pounds of blend with the total price of \$14?

We don't even try to solve the problem right away. Instead, we "translate" the setup into mathematics. The idea is to state the conditions that make a blend "acceptable".

We point out the unknowns first. Let x be the weight of the Kenyan coffee, and let y be the weight of Colombian coffee.

Since the total weight is 6, we have an equation that connects x and y:

$$\boxed{1} \quad x + y = 6\,.$$

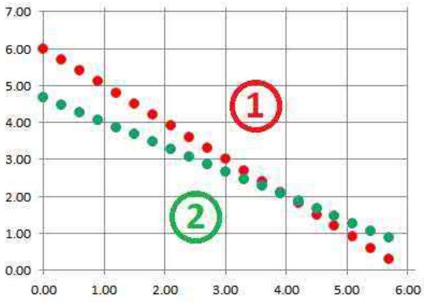
Since the total price of the blend is \$14, we have another equation for x and y:

$$\boxed{2} \quad 2x + 3y = 14\,.$$

These two are called *linear relations*.

Every x is a real number and so is each y. Both the x's and the y's come from certain *sets*. Those two sets may both be visualized as the real line $\mathbf{R}$.

Next, let's visualize possible blends all at once. We use the above sets as respectively the x- and the y-axis of the xy-plane $\mathbf{R}^2$. Then, every pair (x, y) is a point on the plane:



Furthermore, the two relations when plotted produce *straight lines*.

So, every pair (x, y) that satisfies the first equation lies on the first line and every pair that satisfied the second equation lies on the other line.

Changing gears, our conclusion is the following simple statement:

- ▶ For a combination of weights x and y to be acceptable, it must satisfy *both* of the equations.

Let's explore the logic of this statement. It can be recast as an *implication*, i.e., an "if-then" statement:

- ▶ `IF` a pair (x, y) is acceptable, `THEN` it satisfies both equations.

Exercise 1.1.1

Restate as an implication each of the following statements: (a) Every square is a rectangle. (b) Parallel lines don't intersect. (c) $(a+b)^2 = a^2 + 2ab + b^2$.

We will use the following convenient abbreviation throughout the text:

Implication

$$\Longrightarrow$$

It reads "then", "therefore", or "implies that".

Then, the above statement takes the following abbreviated form:

- ▶ A pair (x, y) is acceptable $\Longrightarrow$ It satisfies both equations.

Now, we can try to "flip" the implication of this statement, without assuming that the result will be true:

- ▶ A pair (x, y) is acceptable. $\Longleftarrow$ It satisfies both equations.

This is the abbreviation we used:

Implication

$$\Longleftarrow$$

It reads "whenever", "provided", or "only if".

In other words, we have the following implication:

- ▶ A pair (x, y) satisfies both equations. $\Longrightarrow$ It is acceptable.

The latter is called the *converse* of the original statement. It can also be stated as:

- ▶ `IF` a pair (x, y) satisfies both equations, `THEN` it is acceptable.

The converse happens to be true as well!

Exercise 1.1.2

Justify the last statement.

Exercise 1.1.3

State the converse of each of the following statements: (a) Every square is a rectangle. (b) Parallel lines don't intersect. (c) $2x = 2y$ when $x = y$.

Warning!

The converse of a true statement does not have to be true; example: $x = 1 \implies x^2 = 1$.

Exercise 1.1.4

Suggest your own example of a true statement the converse of which is false.

In our case, the implications go both ways! Combined, the statement and its converse form an *equivalence*:

- ▶ A pair (x, y) satisfies both equations `IF AND ONLY IF` it is acceptable.

We will use the following convenient abbreviation:

Equivalence

$$\iff$$

It reads "if and only if" or "is equivalent to".

Then our two statements are combined as follows:

- ▶ A pair (x, y) satisfies both equations. $\iff$ It is acceptable.

The two parts of an equivalence are interchangeable"

- ▶ A pair (x, y) is acceptable. $\iff$ It satisfies both equations.

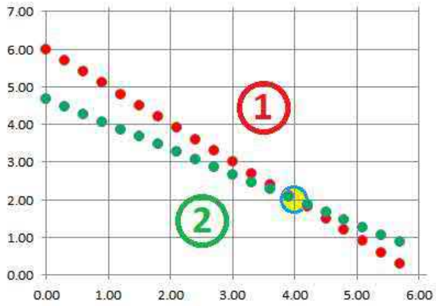
Exercise 1.1.5

Replace "both equations" with "one equation".

Let's now finish the problem. We conclude that the point (x, y) to represent an acceptable pair it has to belong to *both* of the lines!

The solution, therefore, is the point or points that the two have in common. It is their *intersection*.

Now, there seems to be just one point:



And this point can be guessed from the picture to be the following:

$$(x, y) = (4, 2).$$

The guess can be confirmed by substituting the two numbers $x = 4,\ y = 2$ into the two equations:

$$\begin{array}{llll|lllll} \boxed{1} & x & +y & =6 & \to & 4 & +2 & =6 & \texttt{TRUE} \\ \boxed{2} & 2x & +3y & =14 & \to & 2\cdot 2 & +3\cdot 2 & =14 & \texttt{TRUE} \end{array}$$

But are there any others?

Exercise 1.1.6

Make an argument that there are no others based on Euclidean geometry.

An algebraic solution may be as follows:

> From the first equation, we derive: $y = 6 - x$. Then substitute this y into the second equation: $2x + 3(6 - x) = 14$. Solve this new equation: $-x = -4$, or $x = 4$. Substitute this back into the first equation: $(4) + y = 6$, then $y = 2$.

It is routine!

We can re-write this using implications:

1. $\boxed{1} \implies y = 6 - x.$
2. #1 and $\boxed{2} \implies 2x + 3(6 - x) = 14.$
3. #2 $\implies -x = -4 \implies x = 4.$
4. #3 and $\boxed{1} \implies (4) + y = 6 \implies y = 2.$

Exercise 1.1.7

Make an argument that there are no others based on the above algebra.

Such a problem is called a *system of linear equations*:

$$\begin{cases} x & +y & =6\,, \\ 2x & +3y & =14\,. \end{cases}$$

We will develop a systematic approach to such problems.

As a preview, one of the early steps is to collect the data in tables as follows:

$$\begin{array}{|lll|}\hline 1\cdot x & +1\cdot y & =6 \\ \hline 2\cdot x & +3\cdot y & =14 \\ \hline \end{array}\,,\ \text{rewritten as:}\ \begin{array}{|c|c|c|c|c|c|c|c|c|}\hline 1 & \cdot & x & + & 1 & \cdot & y & = & 6 \\ \hline 2 & \cdot & x & + & 3 & \cdot & y & = & 14 \\ \hline \end{array}\,,\ \text{rewritten as:}\ \left[\begin{array}{cc|c} 1 & 1 & 6 \\ 2 & 3 & 14 \end{array}\right]$$

The 2-by-2 (left) part of the resulting table is made of the coefficients of x and y in the equations:

$$\begin{bmatrix} 1 & 1 \\ 2 & 3 \end{bmatrix}$$

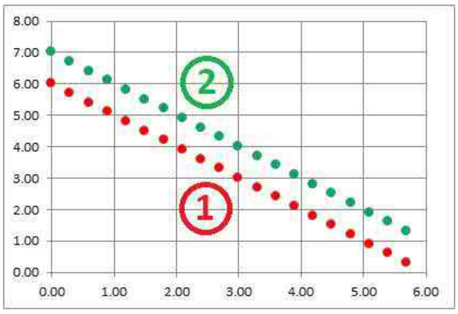
It is called a *matrix*. Meanwhile, these are the *vectors* involved:

$$\begin{bmatrix} 1 \\ 2 \end{bmatrix},\ \begin{bmatrix} 1 \\ 3 \end{bmatrix},\ \begin{bmatrix} 6 \\ 16 \end{bmatrix},\ \text{and}\ \begin{bmatrix} x \\ y \end{bmatrix}.$$

The existence of an intersection point tells us that it is *possible* to create such a blend! However, if the Colombian coffee is also priced at \$2 per pound, we have a new system of linear equations:

$$\begin{cases} x & +y & =6\,, \\ 2x & +2y & =14\,. \end{cases}$$

We discover that the lines are parallel:



There is no such mixture!

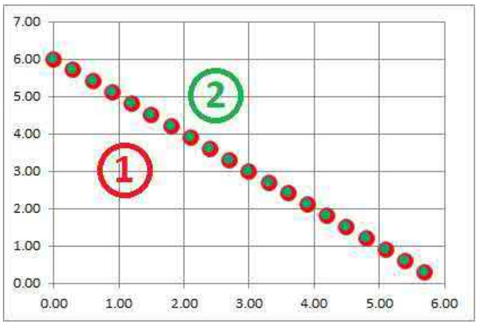
Exercise 1.1.8

Prove this fact algebraically.

The third possibility occurs when we change, in addition, the total price of the blend to $12. The system of linear equations becomes the following:

$$\begin{cases} x & +y & = 6\,, \\ 2x & +2y & = 12\,. \end{cases}$$

The lines merge:



There are infinitely many possible blends.

Exercise 1.1.9

Prove this fact algebraically.

There can be similar problems about other kinds of mixtures:

	variables	restrictions
cooking recipes	weight or count of each ingredient	total weight, total price, amounts of calories, carbs, fats, etc.
investment portfolio	amount of each stock and bond	total value, proportions of industries, tax exposure, etc.
insurance balance sheet	amounts of awards	total value, proportions of various risks, regulation requirements, etc.

Warning!

In these examples, the restrictions are likely to take the form of inequalities rather than equations.

What makes a difference, is that the number of unknowns is unlimited! There can be dozens of ingredients and thousands of securities. This number is the number of degrees of freedom of the system, while the restrictions reduce this number one equation at a time. We may, therefore, face a 10,000-dimensional space with hundreds of sets intersecting each other. Such a space cannot be visualized in the sense we saw above. The way to deal with this challenge is, of course, *algebra*.

Before we turn to this multidimensional linear algebra, we will review some basic *abstract* ideas: sets, relations, and functions. In contrast to precalculus, we will present an enhanced review:

1. We broaden the scope by dealing with sets, relations, and functions that aren't necessarily numerical.
2. We narrow the scope by dealing numerical relations and functions that are likely to be linear.

1.2. Sets

In mathematics, we refer to any loose collection of objects or entities – of any nature – as a *set.*

For example, is this a circle of marbles that we see in a bag? No, the marbles it is made of aren't connected to each other or to any location. One shake and the circle is gone:

It's the same set!

Example 1.2.1: sets as lists

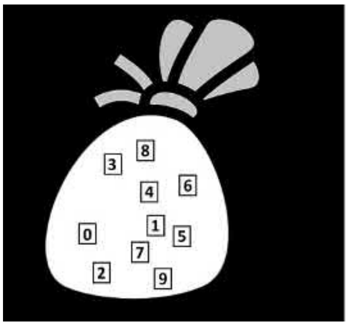
Sets given explicitly – as lists – are the simplest ones:

- A roster of students: Adams, Adkins, Arrows, ... in the alphabetical order
- A list of numbers: 1, 2, 3, 4, ... in the order of size
- A list of planets: Mercury, Venus, Earth, Mars, ... according to the distance from the Sun

Even though the items in each set appears in a special order, if we rearrange its elements, this will be the same set:

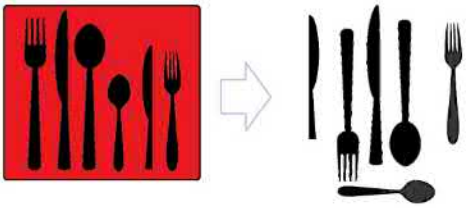
- A roster of students: Smith, Wilson, Adams, ...
- A list of numbers: 2, 1, 4, 3, ...
- A list of planets: Neptune, Venus, Earth, Mercury, ...

The order is not a part of the information we care about when we speak of sets! Here is a bag of numbers:



Example 1.2.2: "sets"

The idea of set contrasts with such expressions as "a set of silverware" when the word "set" suggests a certain structure: specific types of knives and forks with a specific place in the box:



It is the same set, mathematically, whether the items are arranged in a box or piled up on the counter. A set of encyclopedia consists of books that can be arranged alphabetically or chronologically or randomly.

Warning!

Even though we try to provide a precise definition of every new concept, the idea of set is so general that we will have to rely on examples.

What creates a set is our knowledge or ability to determine whether an object *belongs or does not belong* to it.

A list is one such method. Another is a condition to be verified.

Example 1.2.3: sets via conditions

A roster of a class produces a set of the students in this class. It's a list! On the other hand, the *female* students in the class also form a set even if there is no such list; we can just go down the roster and determine if a student belongs to this new set.

In this fashion, we can modify the example above producing more sets:

- The students *with an A*
- The *even* numbers
- The planets *the names of which start with an M*

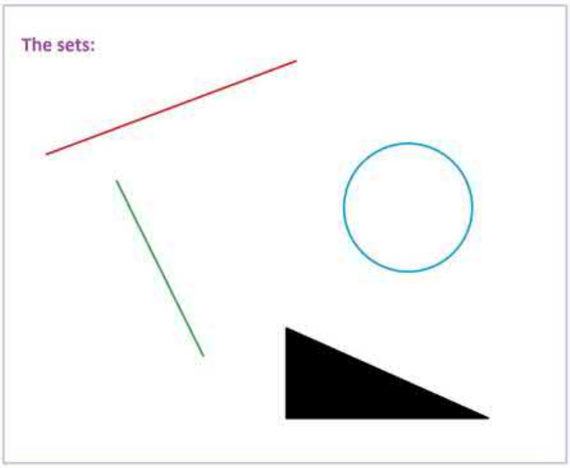
There are no lists!

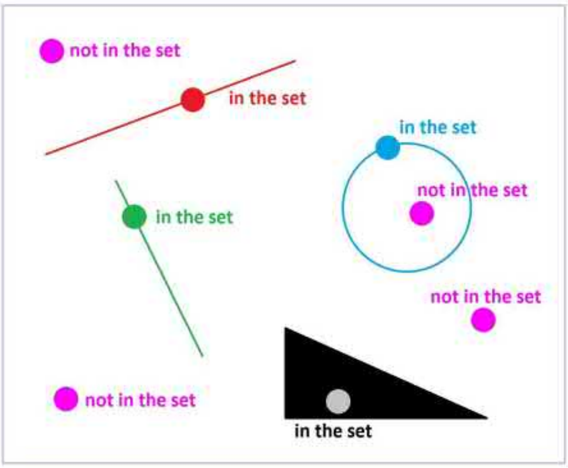
Providing a list is an explicit method of presenting a set. Providing a condition is implicit.

Example 1.2.4: sets in math

A lot of sets examined early in this book will be sets of *numbers*. For example, take the set of *even numbers*; then we know that 2 belongs to it but 3 does not. We simply check the condition: Is the number divisible by 2?

Another example from familiar parts of mathematics is sets of *points* on the plane: straight lines, triangles, circles and other curves, etc.:





We can always tell whether a point belongs to the set!

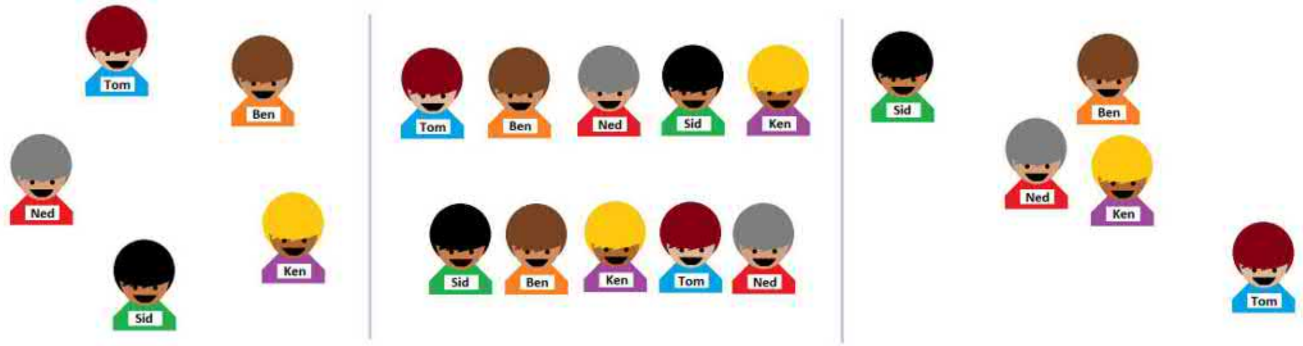
Example 1.2.5: non-sets

If the condition is vague, we don't have a set: "interesting novels", "bad paintings", etc. When the condition is nonsensical, we don't have a set either: "fast trees", "blue numbers", etc.

Exercise 1.2.6

Give your own examples of (a) sets as lists, (b) sets defined via conditions, and (c) non-sets.

In the rest of this chapter we will be using the following example. These *five boys* form a set:



On the one hand, they are individuals and can always be told from each other. On the other hand, they are unrelated to each other: We can list them in any order, we can arrange them in a circle, a square, or at random; we can change the distances between them, and so on. It's the same set! The members of a set are called its *elements*.

Our set is nothing but a *list*:

- Tom
- Ken
- Sid
- Ned
- Ben

Or: "Tom, Ken, Sid, Ned, Ben", in any order.

Warning!

As there is no order, the *elements* of a set aren't to be confused with the *terms* of a sequence as the latter are ordered.

There is a specific mathematical notation for finite sets; we put the list in *braces*:

List notation for sets

$$\{A, B, C, D\}$$

It reads "the set with elements A, B, C, D".

All of these are equally valid representations of our set:

$$\begin{aligned} & \{ \text{ Tom , Ken , Sid , Ned , Ben } \} \\ = & \{ \text{ Ned , Ken , Tom , Ben , Sid } \} \\ = & \{ \text{ Ben , Ken , Sid , Tom , Ned } \} \\ = & \ \ldots \end{aligned}$$

Exercise 1.2.7

How many such representations are there? Hint: In how many ways can you permute these five elements?

Just as the boys have names, the set also needs one. We can call this set "Team", or "Boys", etc. To keep things compact, let's give it a short name, say X:

$$X = \{ \text{ Tom , Ken , Sid , Ned , Ben } \} .$$

We say then that Tom (Ken, etc.) is an element of set X, as well as:

- Tom *belongs* to X, or
- X *contains* Tom.

Just as we want to be clear when two numbers are equal, we want the same clarity for sets. The following will be assumed to be known:

Definition 1.2.8: equal sets

Two sets X and Y are said to be *equal* to each other if the following two conditions are satisfied:

1. Every element of X is also an element of Y.
2. Every element of Y is also an element of X.

Repetitions aren't allowed! Or, at least, they are to be eliminated:

$$\{ \text{ Tom , Ken , Sid , Ned , Ben , Ben } \} \xrightarrow{\text{remove repetitions!}} \{ \text{ Tom , Ken , Sid , Ned , Ben } \}$$

It's the same set!

We can form other sets from the same elements. We can combine those five elements into any set with any number of elements as long as there is no repetition; for example, we can create these new sets:

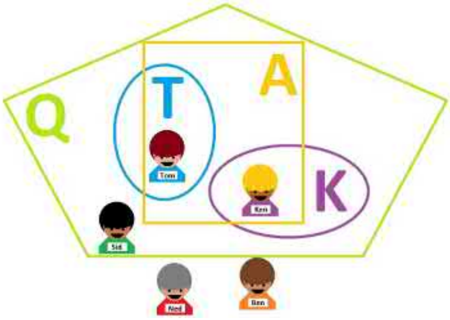
$$\begin{aligned} & T = \{ \text{ Tom } \}, \quad K = \{ \text{ Ken } \}, \quad S = \{ \text{ Sid } \}, \quad N = \{ \text{ Ned } \}, \quad \ldots \\ & A = \{ \text{ Tom , Ken } \}, \quad B = \{ \text{ Sid , Ned } \}, \quad \ldots \\ & Q = \{ \text{ Tom , Ken , Sid } \}, \quad \ldots \end{aligned}$$

The following will be routinely used.

Definition 1.2.9: subset

A set A is called a *subset* of a set X if every element of A is also an element of X.

This is how we mark subsets when the set is shown:



Exercise 1.2.10

How many subsets of 3 elements does the set have? Hint: In how many ways can you choose three elements out of five?

We will use the following notation to convey that idea:

Subset

$$A \subset X$$

The notation resembles the one for numbers: $1 < 2$, $3 < 5$, etc. Indeed, a subset is, in a sense, "smaller" than the set that contains it.

Warning!

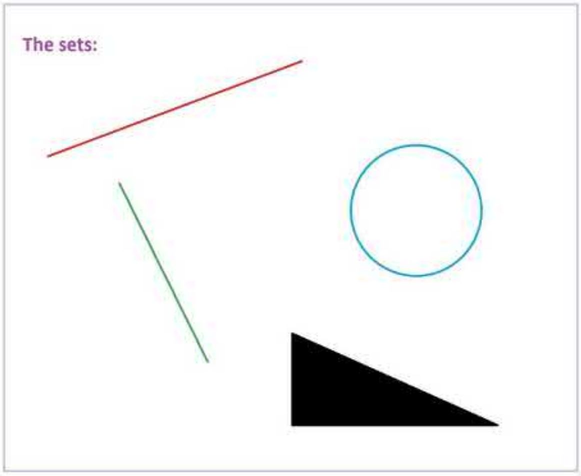
A subset doesn't *have to* be literally smaller. In fact, a set is a subset of itself. Furthermore, an infinite set might have a subset just as infinite...

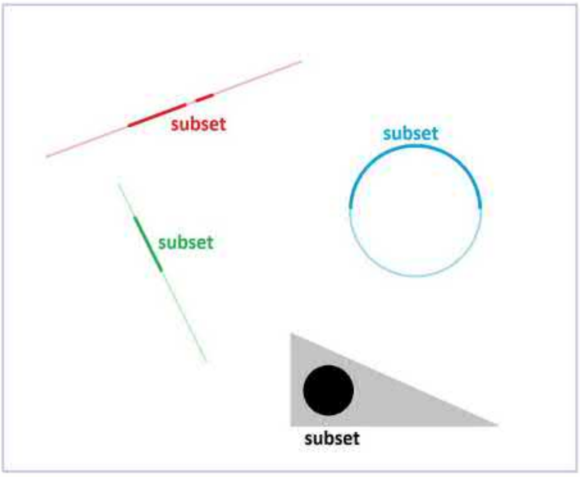
Exercise 1.2.11

Refer to the definition to determine when these are true: (a) $X \subset X$, (b) $\varnothing \subset X$, (c) $X \subset \varnothing$.

Example 1.2.12: plane shapes

We see subsets of geometric figures in the plane:





The following set is very special:

Definition 1.2.13: empty set

The set with no elements is called the *empty set*. It is denoted by:

$$\boxed{\varnothing}$$

A study of *numbers* usually starts with the following sets.

The first is the *natural numbers* used for counting:

The set of natural numbers

$\mathbf{N}$

The set is given by an infinite list:

$$\mathbf{N} = \{0, 1, 2, 3, ...\}\,.$$

The second is the *integers* used to record, addition, moving in a negative direction:

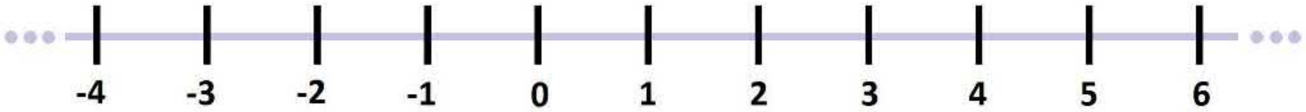
The set of integers

$\mathbf{Z}$

This is also an infinite list:

$$\mathbf{Z} = \{..., -3, -2, -1, 0, 1, 2, 3, ...\}\,.$$

Both sets can be visualized as milestones on a road:



The third is the *rational numbers* used to record fractional locations, time, etc.:

The set of rational numbers

$\mathbf{Q}$

The set is given via a description of its elements.

The fourth is the set of real numbers:

The set of real numbers

$\mathbf{R}$

It is simply visualized as the x-axis:

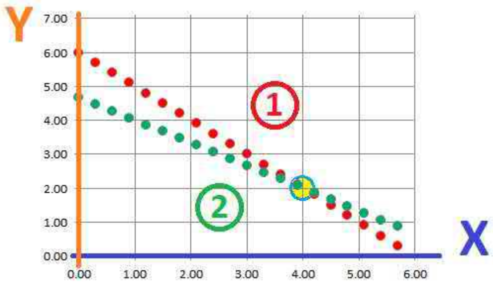
We can express the relation between these sets using the new notation:

$$\mathbf{N} \subset \mathbf{Z} \subset \mathbf{Q} \subset \mathbf{R}\,.$$

Let's now review how these concepts appeared in the problem we solved in the last section.

We made x be the weight of the Kenyan coffee and y be the weight of Colombian coffee. This means that there are two sets here: X is the set of x's and Y is the set of y's. Both are copies of the set of real numbers $\mathbf{R}$.

They are also subsets of the xy-plane $\mathbf{R}^2$:



1.3. The real number line

The starting point of studying numbers is the *natural numbers*:

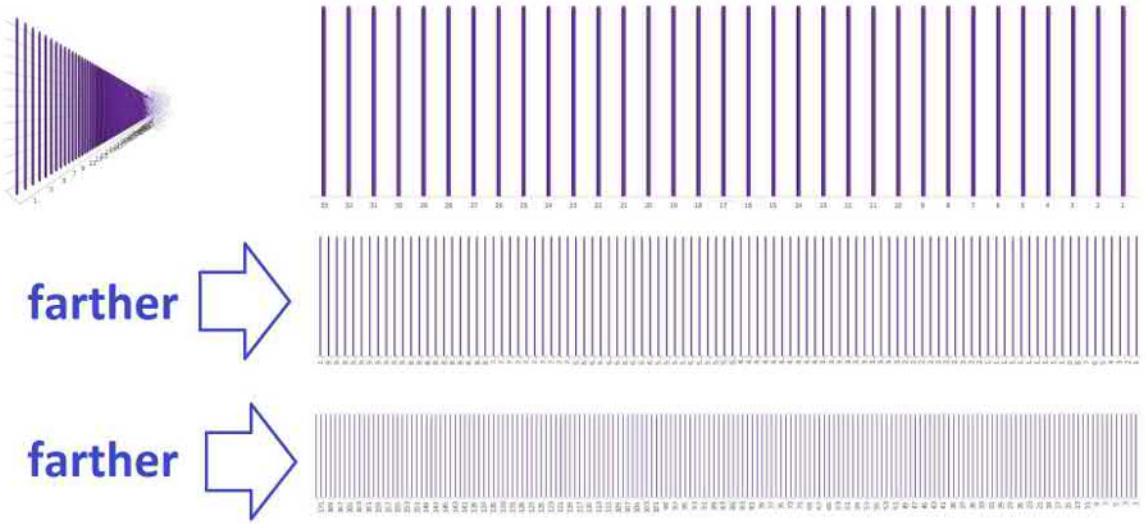
$$0, 1, 2, 3, \ldots$$

They are initially used for counting. The next step is the *integers*:

$$\ldots, -3, -2, -1, 0, 1, 2, 3, \ldots$$

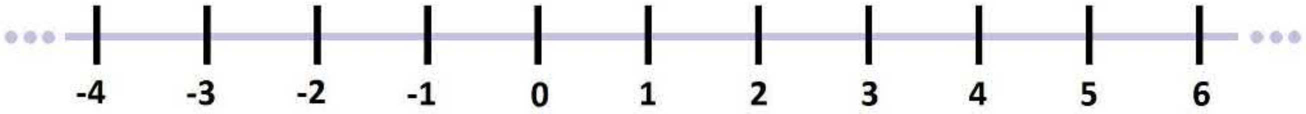
They can be used for studying the space and locations, as follows.

Imagine facing a fence so long that you can't see where it ends. We step *away* from the fence multiple times and there is still more to see:



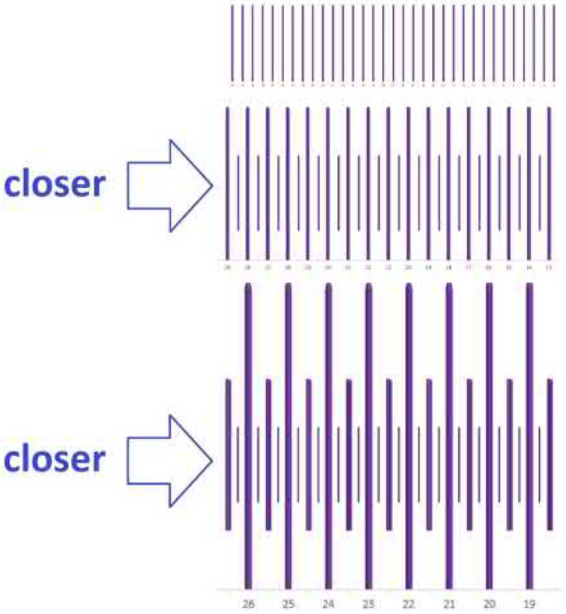
Is the number of planks *infinite*? It may be. For *convenience*, we will just assume that we can go on with this for as long as necessary.

We visualize these as markings on a straight line, according to the order of the planks:

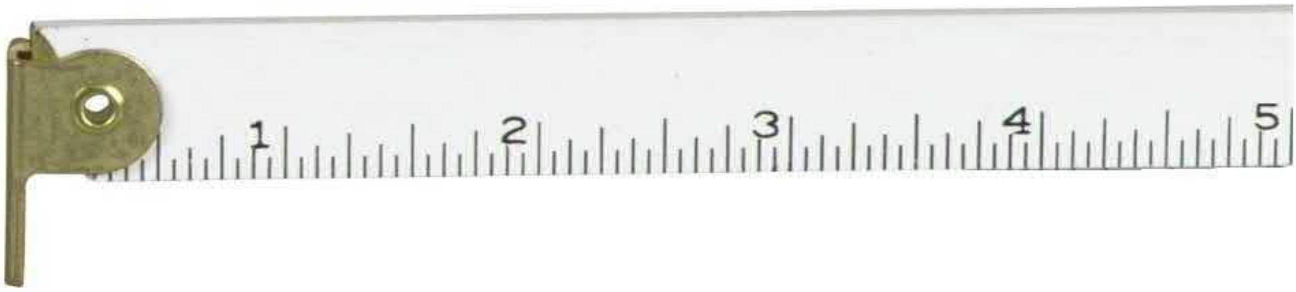


The assumption is that the line and the markings continue without stopping in both directions, which is commonly represented by "...". The same idea applies to the *milestones* on the road; they are also ordered and might continue indefinitely.

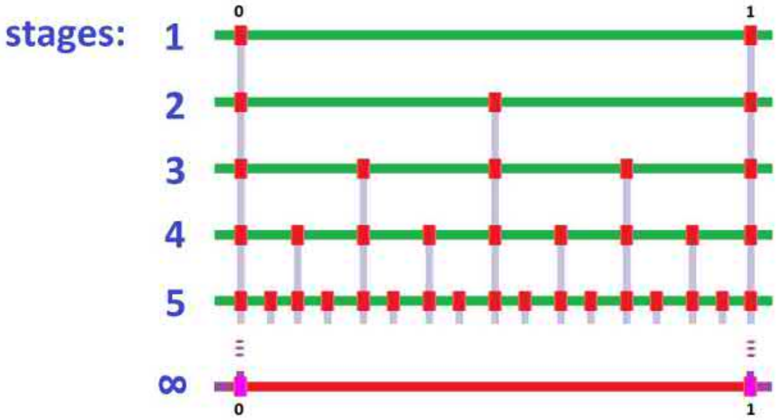
So, we zoomed *out* to see the fence. Suppose now we zoom *in* on a location on the fence. What if there is a shorter plank between every two planks? We look closer and we see more:



If we keep zooming in, the result will look similar to a *ruler*:

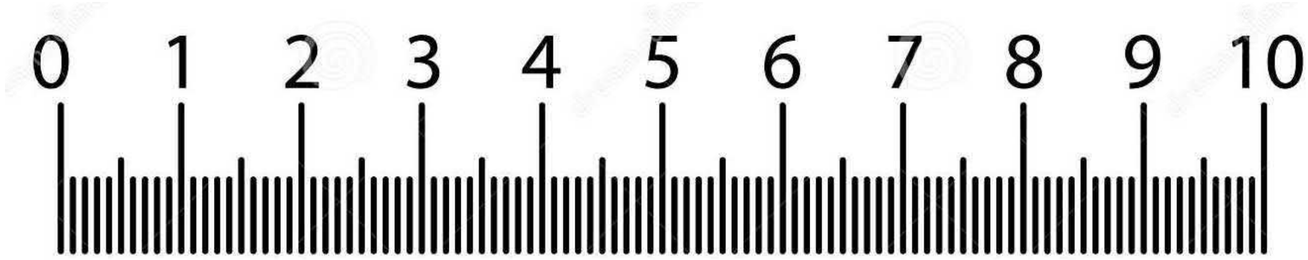


It's as if we add *one mark* between any two and then add another one between either of the two pairs we have created. We keep repeating this step. Even though this ruler goes only to 1/16 of an inch, we can imagine that the process continues indefinitely:



Is the depth *infinite*? It may be. For *convenience*, we will just assume that we can go on with this for as long as necessary.

If we add *nine marks* at a time, the result is a *metric ruler*:



Here, we go from meters to decimeters, to centimeters, to millimeters, etc.

To see it another way, we allow more and more decimals in our numbers:

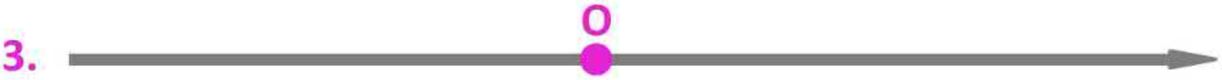
1.55 :	1.	1.5	1.55	1.550	1.5500	...
1/3 :	.3	.33	.333	.3333	.33333	...
π :	3.	3.1	3.14	3.141	3.1415	...

In order to visualize *all* numbers, we first arrange the integers in a line and then the *line of numbers* is built. It takes several steps.

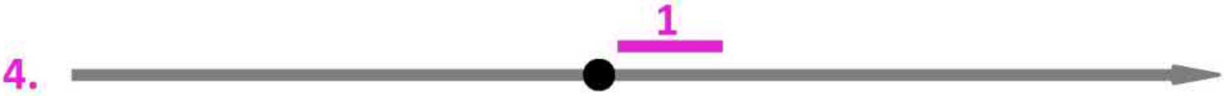
Step 1: Draw a line, called an *axis* (horizontal when convenient):

Step 2: Choose one of the two ends of the line as the *positive* direction (the one on the right when convenient), then the other is the *negative*:

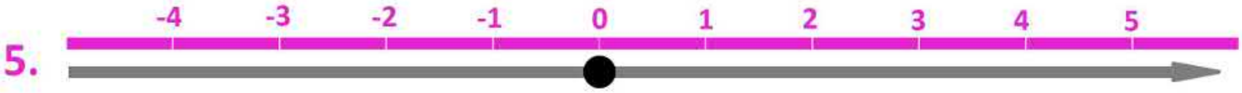
Step 3: Set a point O (a letter, not a number) as the *origin*:



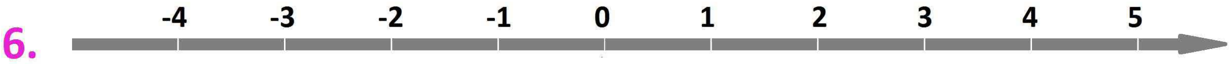
Step 4: Choose a segment of the line as the *unit* of length:



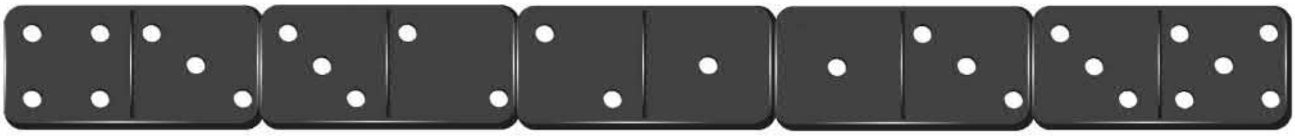
Step 5: Use the segment to measure distances to locations from the origin O (positive in the positive direction, and negative in the negative direction) and add marks, called the *coordinates*:



Step 6: Divide the segments further into fractions of the unit, etc.:

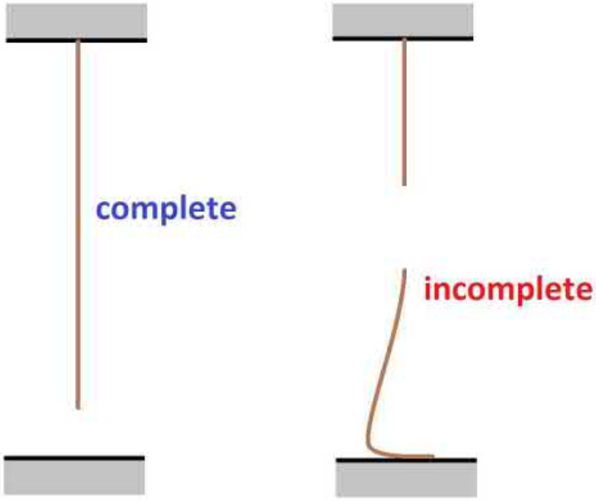


The end result depends on what the building block is. It may contain gaps and look like a ruler (or a comb) as discussed above. It may also be solid and look like a tile or a domino piece:



So, we start with integers as locations and then – by cutting these intervals further and further – also include fractions, i.e., *rational numbers*.

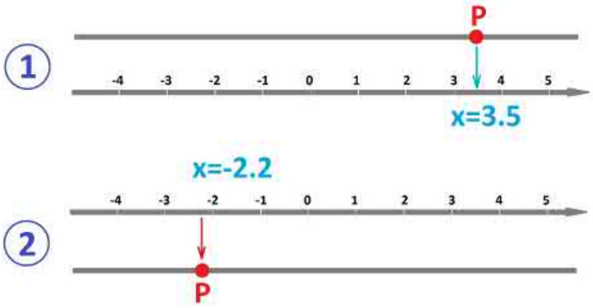
However, we then realize that some of the locations have no counterparts among these numbers. For example, $\sqrt{2}$ is the length of the diagonal of a 1×1 square (and a solution of the equation $x^2 = 2$); it's not rational. That's how the *irrational numbers* come into play. Together, they make up the *real numbers* and the *real number line*. We think of this line as *complete*; there are no missing points. As an illustration, an "incomplete" rope won't hang:



We use this setup to produce a correspondence between the locations on the line and the real numbers:

$$\boxed{\text{location } P \longleftrightarrow \text{number } x}$$

We will follow this correspondence in *both directions*, as follows:



1. First, suppose P is a *location* on the line. We then find the corresponding mark on the line. That's the "coordinate" of P: some *number* x.
2. Conversely, suppose x is a *number*. We think of it as a "coordinate" and find its mark on the line. That's the *location* of x: some point P on the line.

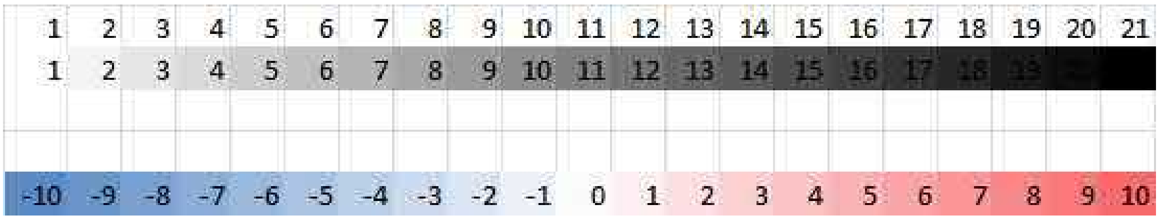
Once this system of coordinates is in place, it is acceptable to think of every location as a number, and vice versa. In fact, we often write:

$$P = x\,.$$

The result may be described as the "1-dimensional coordinate system". It is also called the *real number line* or simply *the number line*.

We have created a visual model of the real numbers. Depending on the real number or a collection of numbers that we are trying to visualize, we choose what part of the real line to exhibit; for example, the zero may or may not be in the picture. We also have to choose an appropriate length of the unit segment in order for the numbers to fit in.

In addition to the ruler, another way to visualize numbers is with *colors*. In fact, in digital imaging the levels of gray are associated with the numbers between 0 and 255. A shorter scale – $1, 2, ..., 20$ – is used in the illustration below (top):



1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21
1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	
-10	-9	-8	-7	-6	-5	-4	-3	-2	-1	0	1	2	3	4	5	6	7	8	9	10

It is also often convenient to associate blue with negative and red with positive numbers (bottom).

Exercise 1.3.1

Think of other examples when numbers are visualized with colors.

1.4. Set building

So far, all numerical sets in this chapter have been *subsets of the set of real numbers* $\mathbf{R}$. In particular, numerical sets emerge as domains and codomains of numerical functions. They may also come from *solving equations*.

Example 1.4.1: sets from equations

Unless entirely nonsensical, every statement in mathematics is true or false:

$$\begin{array}{lll}1+1 & =2 & \texttt{TRUE}\\ 1+1 & =3 & \texttt{FALSE}\end{array}$$

What about this:

$$x+1 \quad =2 \quad \texttt{TRUE OR FALSE?}$$

It depends on x, of course. We can, therefore, use equations to form sets.

Consider these:
- We face the equation $x+2=5$. After some work, we find: $x=3$.
- We face the equation $3x=15$. After some work, we find: $x=5$. Is there more?
- We face the equation $x^2-3x+2=0$. After some work, we find: $x=1$. Is that it?
- We face the equation $x^2+1=0$. After all the work, we can't find any x. Should we keep trying?

Here, x is a *label* that stands for an unspecified number that is meant to satisfy this condition. In other words, we seek such numbers that, when they replace x in the equation, we see a true statement. It could simply be trial and error.

Let's take the first equation:

$$x+2=5\,.$$

We replace x with a number above or write a number in the blank square below:

$$\square+1=5\,.$$

For example:
- Is $x=1$ a solution? Plug it in the equation: $x+2=5$ becomes $(1)+2=5$. `FALSE`. This is *not* a solution.

- Is $x = 2$ a solution? Plug it in the equation: $x + 2 = 5$ becomes $(2) + 2 = 5$. `FALSE`. This is *not* a solution.
- Is $x = 3$ a solution? Plug it in the equation: $x + 2 = 5$ becomes $(3) + 2 = 5$. `TRUE`. This *is* a solution.
- Should we stop now? Why would we? For all we know, there may be more solutions.

We never say that we have found "the" solution unless we know for sure that there is only one.

Exercise 1.4.2

Interpret each of these equations as a relation.

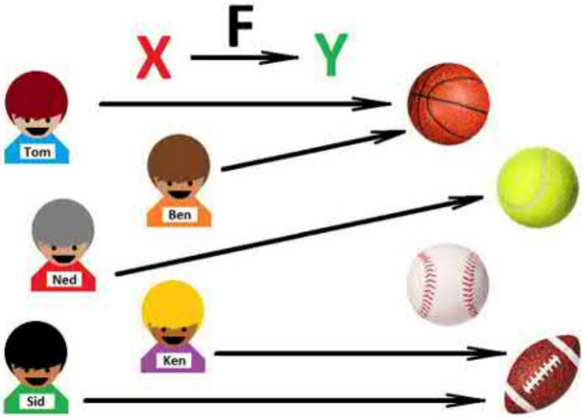
Exercise 1.4.3

Solve these equations:

$$(a)\ x^2 + 2x + 1 = 0, \quad (b)\ \frac{x}{x} = 0, \quad (c)\ \frac{x}{x} = 1\,.$$

But what does it *mean* to solve an equation? We have tried to find x that satisfies the equation... But what are we supposed to have at the end of our work?

Let's go back to our running example of boys and balls:



It tells us what game each boy prefers. What about the other way around? Which boys prefer a particular game?

- Which boys prefer basketball? The answer isn't "Tom", and it isn't "Ben"; it's "Tom and Ben and nobody else".
- Which boys prefer tennis? "just Ned".
- Which boys prefer baseball? "No one".
- Which boys prefer football? "Ken and Sid only".

But each question – one for each element of the codomain Y – is also an *equation*:

- Find x with $F(x) =$ basketball. Tom is a solution, and Ben is a solution. Combined, Tom and Ben are *the* solutions.
- Find x with $F(x) =$ tennis. Ned is the solution.
- Find x with $F(x) =$ baseball. No solutions.
- Find x with $F(x) =$ football. Ken and Sid are the solutions.

This is how we understand this idea:

> ▶ *A solution of an equation with respect to x is an element that, when put in the place of x in the equation, gives us a true statement.*

However, we must present *all* x's that satisfy the equation. In other words, the answer is a *set*:

- The solution set of the equation $F(x) = \text{basketball}$ is $\{$ Tom, Ben $\}$.
- The solution set of the equation $F(x) = \text{tennis}$ is $\{$ Ned $\}$.
- The solution set of the equation $F(x) = \text{baseball}$ has no elements.
- The solution set of the equation $F(x) = \text{football}$ is $\{$ Ken, Sid $\}$.

All of these sets are subsets of the domain X. This is the terminology we will routinely use.

Definition 1.4.4: equation and its solution

Suppose $f : X \to Y$ is a function and b is one of the elements of Y. The *solution set* of the equation $f(x) = b$ is the set of all x in X that make the equation true. To *solve an equation* means to find its solution set.

In other words, the solution set is *the* solution!

Next, the sets above can be presented in this spirit:

Set-building notation

$$\Big\{\ x : \quad \text{condition for } x\ \Big\}$$

The expression stands for the set of *all* x that satisfy the condition. What kind of condition? An equation, as above. Any condition as long as it is specific enough for us to unambiguously answer the question "does x satisfy it?". For example:

$$\{\ \text{student: 20 years old}\ \}\,.$$

The set from which we pick x's one at a time is presented or assumed to be known.

Warning!

Many sources also use:

$$\{x|\ \ \text{condition for } x\ \}\,.$$

For example, the equations above are seen as conditions. Below we list their solution sets (left):

$$\begin{array}{llll}
\{x,\ \text{boy} & :\ F(x) = \ \text{basketball}\ \} & = \{\ \text{Tom, Ben}\ \} & \\
\{x,\ \text{boy} & :\ F(x) = \ \text{tennis}\ \} & = \{\ \text{Ned}\ \} & \\
\{x,\ \text{boy} & :\ F(x) = \ \text{baseball}\ \} & = \{\quad \} & \\
\{x,\ \text{boy} & :\ F(x) = \ \text{football}\ \} & = \{\ \text{Ken, Sid}\ \} & = \varnothing
\end{array}$$

These description can sometimes be *simplified* (right). One can imagine that we simply went over the list of X and tested each of its elements.

Exercise 1.4.5

Show that the empty set is a subset of any set.

Exercise 1.4.6

Simplify the following sets:

$$\begin{array}{lll}
\{x,\ \text{boy} & : & \text{his shirt is red}\ \} \\
\{y,\ \text{ball} & : & \text{is preferred by two boys}\ \} \\
\{y,\ \text{ball} & : & \text{is round}\ \}
\end{array}$$

Example 1.4.7: inclusion vs. implication

Here is an interpretation of the definition of subset. The definition says that $A\subset B$ when the following is satisfied:

- `IF` x belongs to A, `THEN` x belongs to B;

or

- x belongs to $A \implies x$ belongs to B.

The converse of this implication is false when $A\neq B$. Furthermore, we have the following:

- x satisfies the condition for $A \implies x$ satisfies the condition for B.

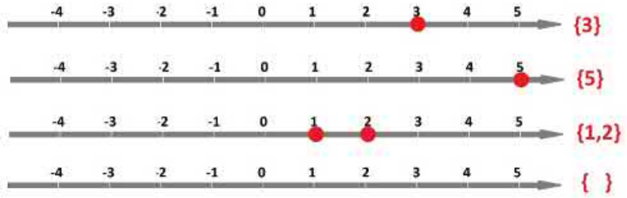
In other words, the latter condition is less restrictive.

Example 1.4.8: solution sets of equations

Let's take another look at the equations above, assuming that the "ambient" set is the set of real numbers:

equation:	answer?	solution set:
$x+2=5$	$x=3$	$\{3\}$
$3x=15$	$x=5$	$\{5\}$
$x^2-3x+2=0$	$x=1$ and...	$\{1,2\}$
$x^2+2x+1=0$	no x?	$\{\quad\}$

This is how we visualize these four sets:



Below we use the set-building notation again on the left, and then on the right, we see another, simpler, representation of the set:

$$\begin{array}{ll}\{x:\ x+2=5\} & =\{3\}\\ \{x:\ 3x=15\} & =\{5\}\\ \{x:\ x^2-3x+2=0\} & =\{1,2\}\\ \{x:\ x^2+1=0\} & =\{\quad\}=\varnothing\end{array}$$

The simplest way to represent a set is, of course, a list.

Exercise 1.4.9

Solve these equations:

$$x=x,\quad 1=1,\quad 1=0\,.$$

1.5. *Relations*

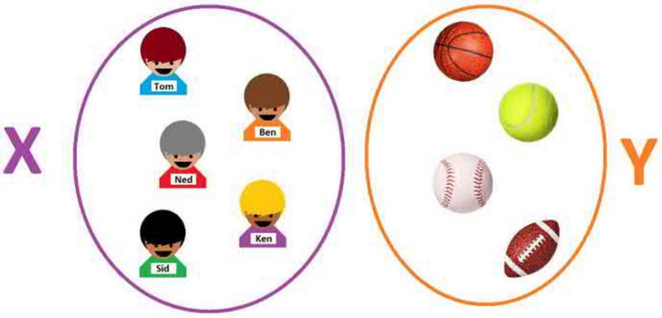
To continue with our example, suppose there is *another*, unrelated, set, say Y, the set of these four balls:

Just as X, set Y has no structure. Just as X, it's just a list:

$$\begin{aligned} Y &= \{\ \text{basketball}\ ,\ \text{tennis}\ ,\ \text{baseball}\ ,\ \text{football}\ \} \\ &= \{\ \text{football}\ ,\ \text{baseball}\ ,\ \text{tennis}\ ,\ \text{basketball}\ \} \\ &= \ldots \end{aligned}$$

We can remove balls from the set, creating subsets of Y.

Now, let's put the two sets, X and Y, next to each other and ask ourselves: Are these two sets related to each other somehow?



Yes, boys like to play sports! Let's make this idea specific. Each boy may be *interested in a particular sport* or he may not. For example, suppose this is what we know:

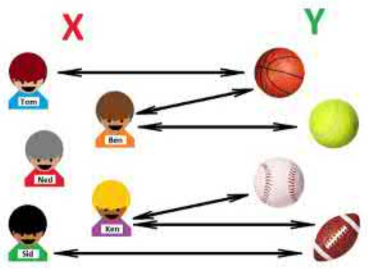
- Tom likes basketball.
- Ben likes basketball and tennis.
- Ken likes baseball and football.
- Sid likes football.

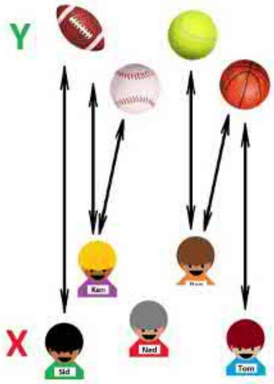
And that's all each likes.

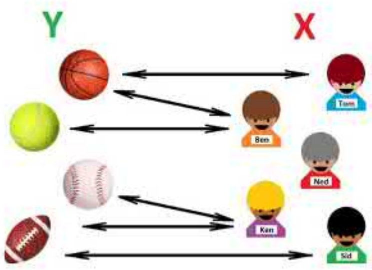
As a result, we have the following:

> ▶ An element of set X is *related* to an element of set Y.

In order to visualize these relations, let's connect each boy with the corresponding ball by a line segment with arrows at the ends, while the two sets may be placed arbitrarily against each other:







This visualization helps us discover that Ned doesn't like sports at all. As you can see, this is a two-sided correspondence: Neither of the two elements at the ends of the line comes first or second. The same applies to the sets: Neither of the two sets comes first or second. In fact, we can derive these new facts about the preferences either from the original list or from the image on the right:

- Basketball is liked by Tom and Ben.
- Tennis is liked by Ben.
- Baseball is liked by Ken.
- Football is liked by Ken and Sid.

We have, therefore, a *list of pairs*:

- Tom & basketball
- Ben & basketball
- Ben & tennis
- Ken & baseball
- Ken & football
- Ben & football

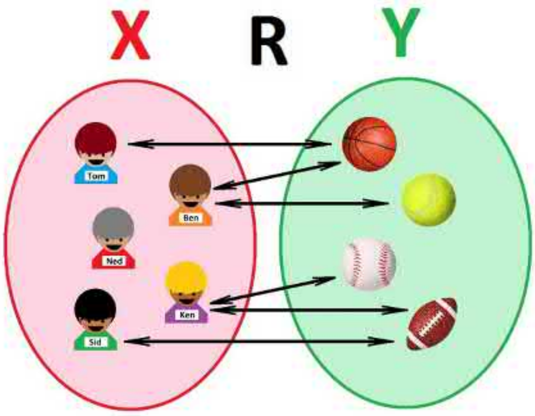
The following concept will be commonly used throughout.

Definition 1.5.1: relation between sets

Any set of pairs (x, y), with x taken from a set X and y from a set Y, is called a *relation* between sets X and Y.

In other words, this is a set of arrows.

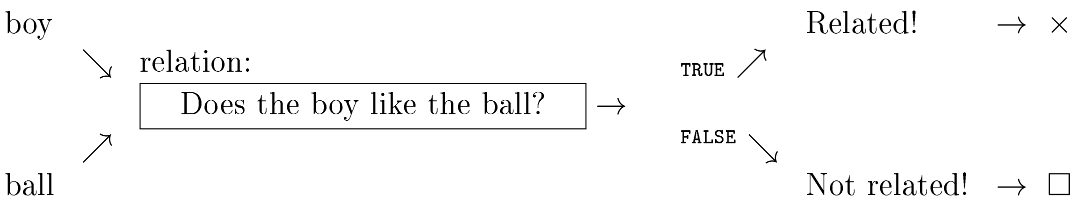
There may be many different relations between a pair sets; let's call this one R:



Warning!

We don't require *every* element to have a corresponding element in the other set.

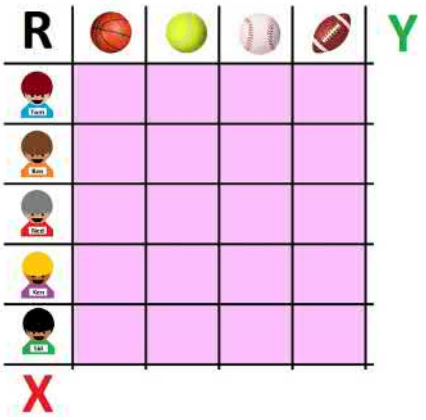
The following diagram represents the approach to relations from the point of *computing*:



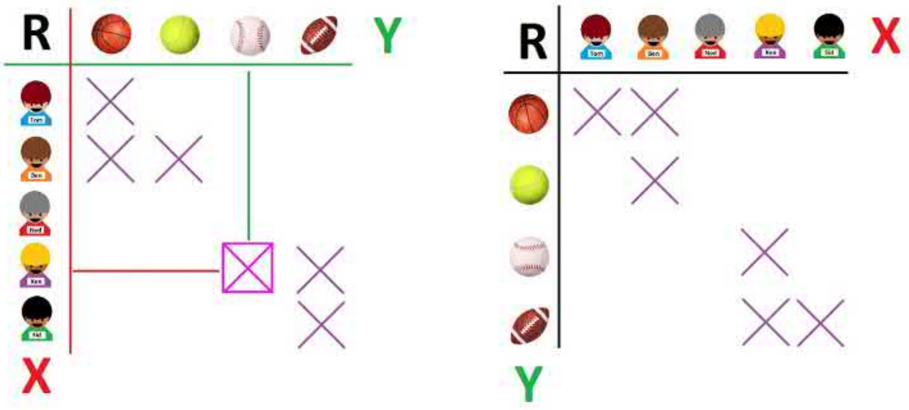
It's a simple procedure!

Next, we make a step toward *visualizing* relations.

Where do we place those crosses? When the sets are lists, we build *tables*. For the relation R above, we put the boys in the leftmost column and the balls in the top row:



There are 20 cells. Now, if the boy likes the sport, we put a cross (or another mark) in the cell that lies in the boy's row and the ball's column (left):



Or, we can put the boys in the first row and the balls in the first column (right). It's just as good! In other words, we can flip the table about its *diagonal*. These are two visualizations of the same relation.

Exercise 1.5.2

Based on the relation R presented above, create a new relation called, say, S, that relates the boys and the sports they *don't* like. Give an arrow representations and a table representations of S.

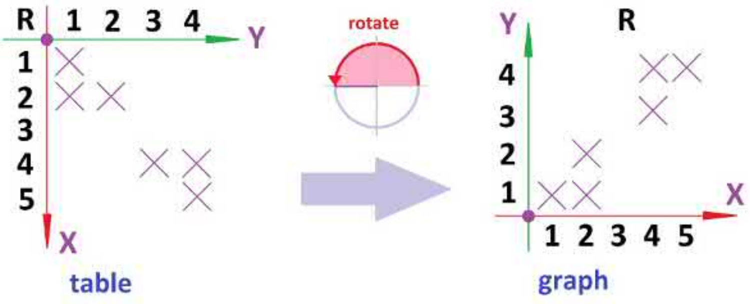
Exercise 1.5.3

Are there any relations on the subsets of the two sets?

Any collection (a set) of marks in such a table creates a relation, and conversely, a relation is nothing but a collection of marks in this table.

How do these ideas apply to *numbers*?

Let's simply *rename* the boys as numbers, $1-5$, and rename the balls as numbers, $1-4$. Even though the two sets share members, we prefer to think that the former belong to the set of real numbers, while the latter belong to another *copy* of this set. We then draw these number lines along the sides of our table (left):



These axes are also labeled to avoid confusion between the two very different sets. Furthermore, the table can be rotated 90 degrees counterclockwise (right). This table is then called the *graph* of the relation.

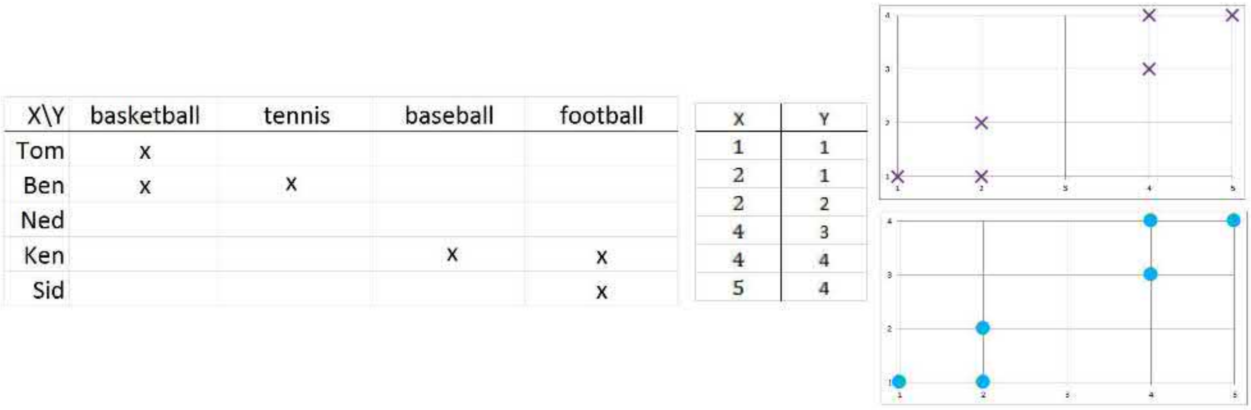
Exercise 1.5.4

When the rows and the columns are interchanged, is there anything that is preserved?

Exercise 1.5.5

Finish the sentence: "This renaming of the boys is a ___."

Suppose we have the elements of the sets renamed as numbers (left), then we capture the relation as a list of pairs of elements of X and Y (middle), and finally, the graph of the relation can be plotted automatically by the spreadsheets:



X\Y	basketball	tennis	baseball	football
Tom	x			
Ben	x	x		
Ned				
Ken			x	x
Sid				x

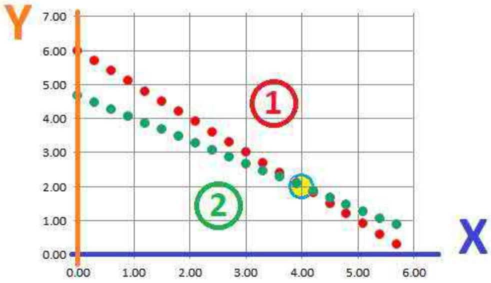
X	Y
1	1
2	1
2	2
4	3
4	4
5	4

Let's continue the review of how these concepts appeared in the problem of mixtures.

We made X to be the set of the possible weights of the Kenyan coffee and Y the set is the set of the possible weights of the Colombian coffee. Both are copies of the set of real numbers $\mathbf{R}$. We also had two relations:

1. First, since the total weight is 6, two numbers x and y are related when we have the following condition satisfied:
$$\boxed{1} \quad x + y = 6\,.$$
2. Second, since the total price of the blend is \$14, two numbers x and y are related when we have the following condition satisfied:
$$\boxed{2} \quad 2x + 3y = 14\,.$$

The graphs of the relations are these two lines:



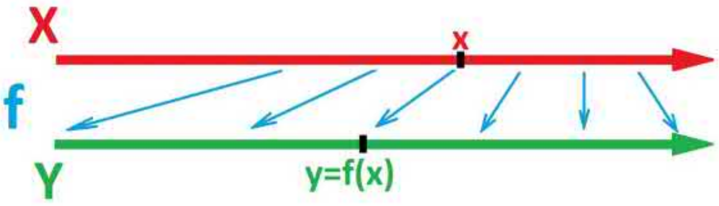
More happened though. From the first equation, we derived:

$$y = 6 - x\,.$$

This is, of course, the same relation! But its new form makes our life much easier: instead of testing *every* pair (x, y), we can just plug in as many x's as necessary producing the corresponding y's automatically. That's why we move from relations to *functions* whenever we can.

1.6. The xy-plane

A relation, or a function, deals with two sets of numbers: the domain X and the codomain Y. That's why we need two axes, one for $X = \mathbf{R}$ and one for $Y = \mathbf{R}$. How do we arrange them? We can use the method presented above: putting the axes next to each other and connecting them by arrows:



But since $X = \mathbf{R}$ is infinite, however, we would need infinitely many arrows.

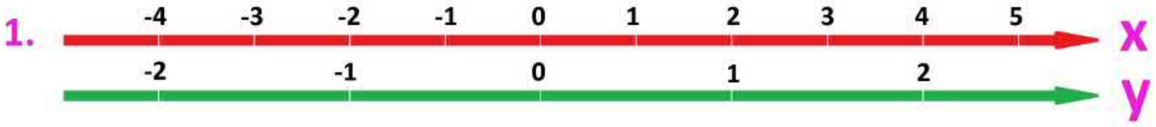
Is there a better way? We already know another approach: *tables and graphs*. Instead of side-by-side, we place X horizontally and Y vertically.

Step 1

We start with a *real line* $\mathbf{R}$, or the x-axis. That's where the real numbers live, and now X and Y are subsets of $\mathbf{R}$. So, we will need two *copies* of the real line. We give them special names:

- the *x-axis*, and
- the *y-axis*.

Just as the inputs and the outputs of a function have typically nothing to do with each other (such as time vs. space, or space vs. temperature), the two axes may have different unit segments:

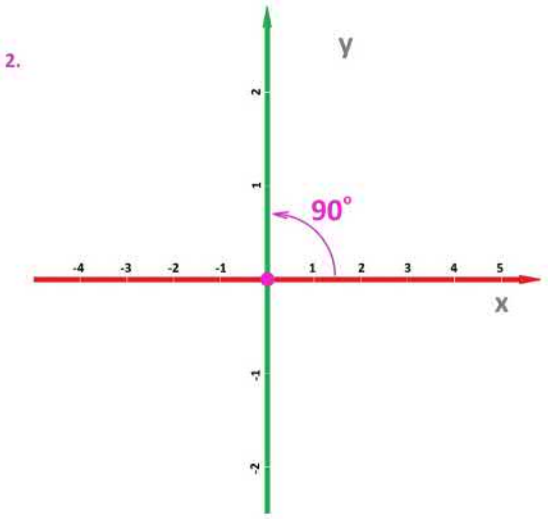


Step 2

To move toward the table we need, we arrange the two coordinate axes in a typical way as follows:

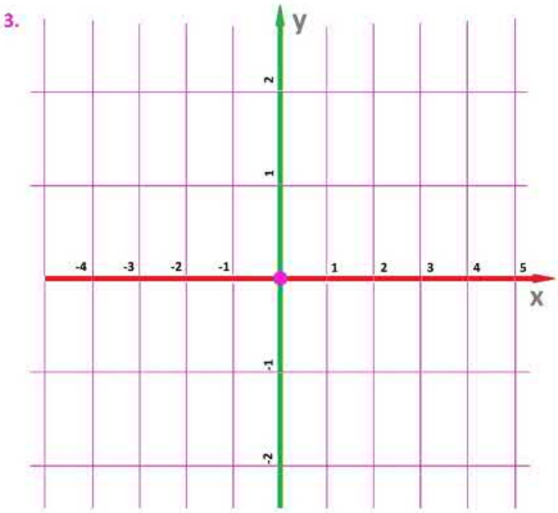
- The x-axis is horizontal, with the positive direction pointing right.
- The y-axis is vertical, with the positive direction pointing up.

Usually, the two axes are attached to each other at their origins:



Step 3

Finally, we attach “fabric” to this “frame”. We use the marks on the axes to draw a *rectangular grid*.



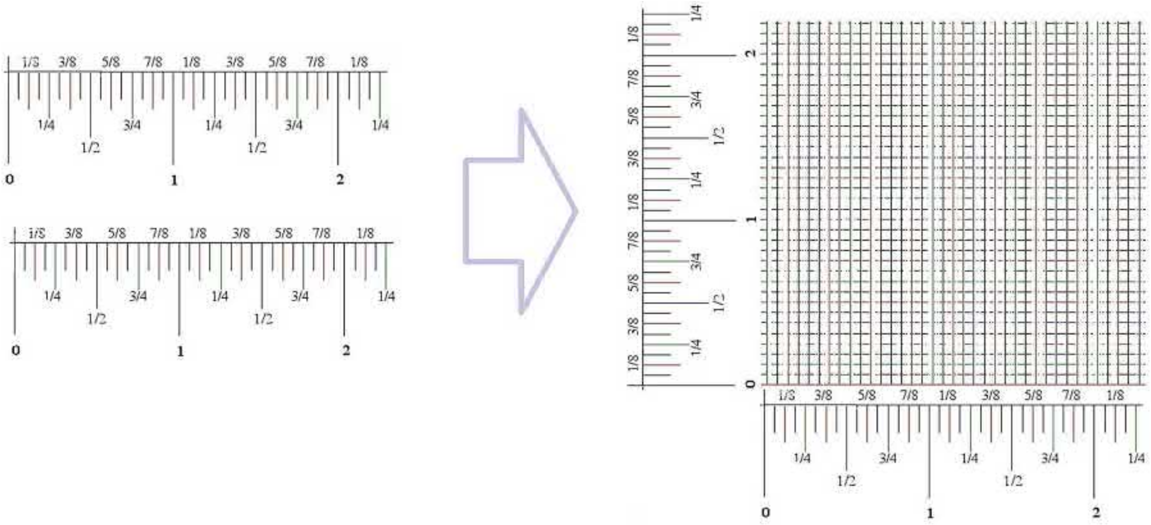
Now we have what we call the *Cartesian plane*, or simply the xy-*plane*. As it is made from a combination of two copies of $\mathbf{R}$ and is often denoted as follows:

xy-plane

$$\mathbf{R}^2$$

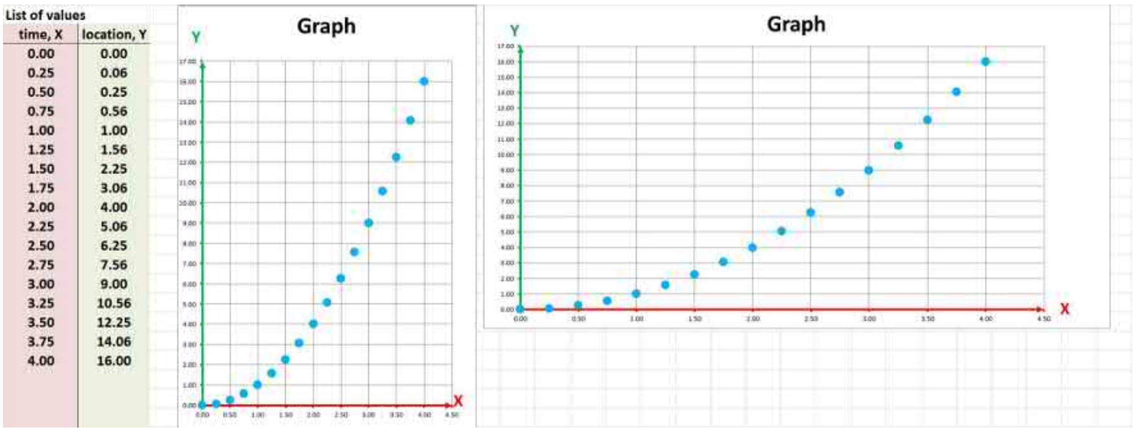
The notation is also explained by the fact that the area of an $r \times r$ square is $r \cdot r = r^2$. This is literally true, however, only when both axes measure length.

The idea that the real line is like a ruler leads to the idea that the xy-plane is like a *ruled paper*:



Example 1.6.1: resizing graphs

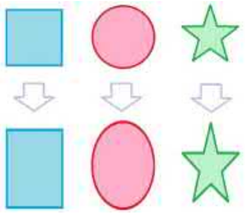
In the context of plotting graphs, it is frequently the case that the relative dimensions of x and y are unimportant, and then the xy-plane can be resized arbitrarily and *disproportionately*. The graphs change too! The chart in this spreadsheet shows how different the graph of the same function might look:



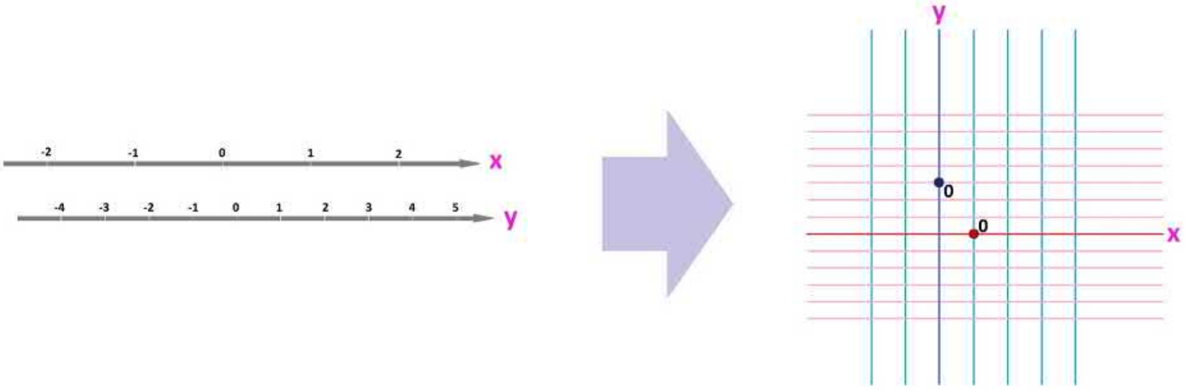
List of values

time, X	location, Y
0.00	0.00
0.25	0.06
0.50	0.25
0.75	0.56
1.00	1.00
1.25	1.56
1.50	2.25
1.75	3.06
2.00	4.00
2.25	5.06
2.50	6.25
2.75	7.56
3.00	9.00
3.25	10.56
3.50	12.25
3.75	14.06
4.00	16.00

Such resizing will turn squares into rectangles and circles into ovals:



This fact imposes an important limit on how well the graph visualizes the function. The size, of course, doesn't matter. The angles might be telling us nothing; even though the inclination – up or down – of the graph can't disappear under this re-sizing, its steepness can change. We can't, therefore, say that this line is "steep" but only that it is "steeper" than another one plotted on the same coordinate plane. In this context, it is also often acceptable to have the origins of the two axes misaligned or even absent:



The idea of the *Cartesian coordinate system* is the same as the one for the real line:

- ▶ Give a numerical representation of locations.

This time, however, this is a plane, and there are *two* axes and *two* coordinates for each point. We use the above setup to produce a correspondence:

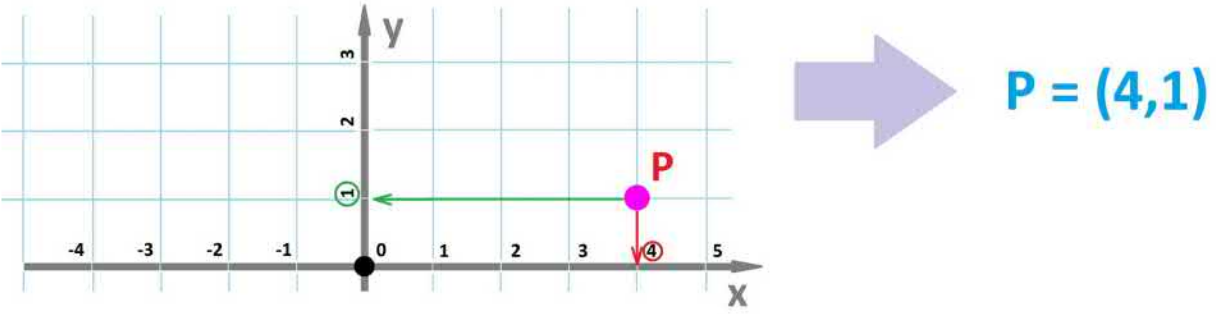
$$\boxed{\text{location } P \longleftrightarrow \text{a pair of numbers } (x,y)}$$

It works in *both directions*:

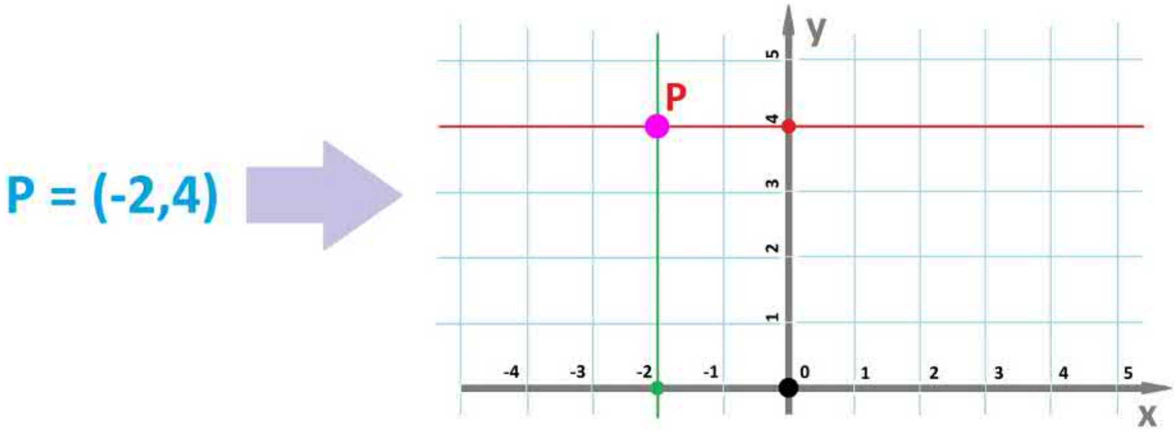
1. → First, suppose P is a *location* on the plane. We then draw a vertical line through P until it intersects the x-axis. The mark, x, of the location where they cross is the *x-coordinate* of P. We next draw a horizontal line through P until it intersects the y-axis. The mark, y, of the location where they cross is the *y-coordinate* of P.
2. ← On the flip side, suppose x and y are *numbers*. First, we find the mark x on the x-axis and draw a vertical line through this point. Second, we find the mark y on the y-axis and draw a horizontal line through this point. The intersection of these two lines is the corresponding *location* P on the plane.

Example 1.6.2: coordinates

We illustrate this idea below with a specific example. From a point to its coordinate:



From coordinates to a point:



The notation is as follows:

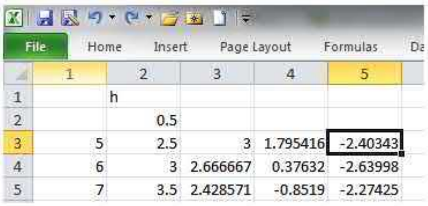
xy-coordinates

$$\Big(\ x\text{-coordinate}\,,\ \ y\text{-coordinate}\ \Big)$$

Example 1.6.3: using coordinates in computing

The 2-dimensional Cartesian system isn't as widespread as the 1-dimensional, but it is very common in certain areas of computing.

Spreadsheet applications use the Cartesian system – starting at the upper left corner – to provide a convenient way of representing locations of cells:



	1	2	3	4	5
1		h			
2		0.5			
3	5	2.5	3	1.795416	-2.40343
4	6	3	2.666667	0.37632	-2.63998
5	7	3.5	2.428571	-0.8519	-2.27425

This is the difference:

	spreadsheet	Cartesian system
1st coordinate	rows `R`, down	x, right
2nd coordinate	columns `C`, right	y, up

This idea allows us to use reference to express a value located in one cell in terms of values located in other cells. For example, the following formula computes the double of the number contained in the cell located at the intersection of row 3 and column 5:

`=2*R3C5`

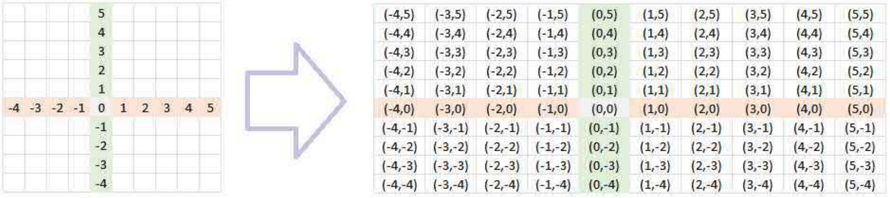
It can be placed at any other cell. The next formula computes the sum of the number contained in the cell in row 5 and column 2 and the number contained in the cell in row 1 and column 5:

`=R5C2+R1C5`

There are also relative references. For example, the following formula takes the number contained in the cell located 3 rows up and 2 columns right from the current cell:

`=R[-3]C[2]`

In contrast, this is what the proper Cartesian system for spreadsheets would look like:



(-4,5)	(-3,5)	(-2,5)	(-1,5)	(0,5)	(1,5)	(2,5)	(3,5)	(4,5)	(5,5)
(-4,4)	(-3,4)	(-2,4)	(-1,4)	(0,4)	(1,4)	(2,4)	(3,4)	(4,4)	(5,4)
(-4,3)	(-3,3)	(-2,3)	(-1,3)	(0,3)	(1,3)	(2,3)	(3,3)	(4,3)	(5,3)
(-4,2)	(-3,2)	(-2,2)	(-1,2)	(0,2)	(1,2)	(2,2)	(3,2)	(4,2)	(5,2)
(-4,1)	(-3,1)	(-2,1)	(-1,1)	(0,1)	(1,1)	(2,1)	(3,1)	(4,1)	(5,1)
(-4,0)	(-3,0)	(-2,0)	(-1,0)	(0,0)	(1,0)	(2,0)	(3,0)	(4,0)	(5,0)
(-4,-1)	(-3,-1)	(-2,-1)	(-1,-1)	(0,-1)	(1,-1)	(2,-1)	(3,-1)	(4,-1)	(5,-1)
(-4,-2)	(-3,-2)	(-2,-2)	(-1,-2)	(0,-2)	(1,-2)	(2,-2)	(3,-2)	(4,-2)	(5,-2)
(-4,-3)	(-3,-3)	(-2,-3)	(-1,-3)	(0,-3)	(1,-3)	(2,-3)	(3,-3)	(4,-3)	(5,-3)
(-4,-4)	(-3,-4)	(-2,-4)	(-1,-4)	(0,-4)	(1,-4)	(2,-4)	(3,-4)	(4,-4)	(5,-4)

Thus, every point on the xy-plane is, or can be, labeled with a pair of numbers.

Once the coordinate system is in place, it is acceptable to think of locations as pairs of numbers, and vice versa. In fact, we can write:

$$\boxed{P = (x, y)}$$

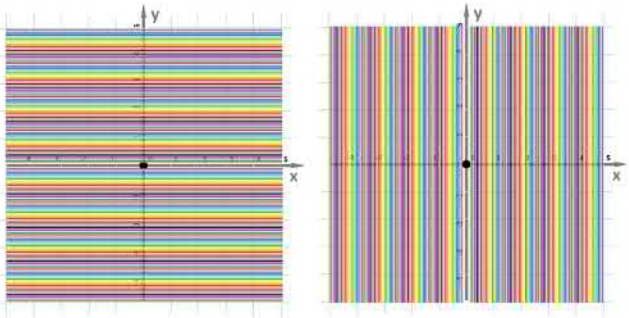
It is important to realize that what we are dealing with is a *set* too! This is the set of all pairs of real numbers presented in the set-building notation:

$$\mathbf{R}^2 = \{(x, y) : \ x \text{ real}, \ y \text{ real } \} .$$

The xy-plane is just a visualization of this set. Below we consider some of its simplest *subsets*.

Example 1.6.4: lines are fibers of plane

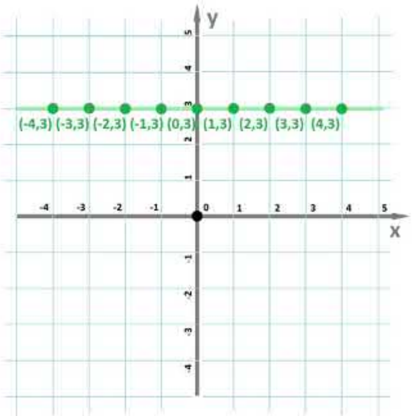
One can think of the xy-plane as a *stack* of lines, vertical or horizontal, each of which is just a shifted copy of the corresponding axis:

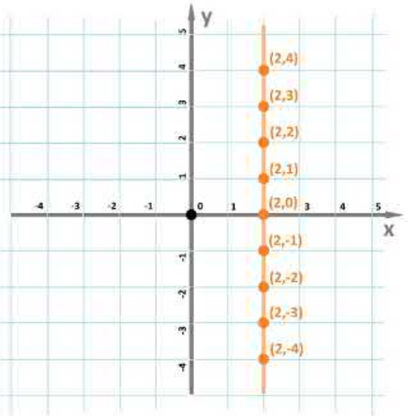


We can use this idea to reveal the internal structure of the coordinate plane:

- If L is a line parallel to the x-axis, then all points on L have the same y-coordinate. Conversely, if a set L of points on the xy-plane consists of all points with the same y-coordinate, L is a line parallel to the x-axis.
- If L is a line parallel to the y-axis, then all points on L have the same x-coordinate. Conversely, if a set L of points on the xy-plane consists of all points with the same x-coordinate, L is a line parallel to the y-axis.

Two specific examples are shown below:





Then, we have a compact way to represent these two lines:

- horizontal: $y = 3$, and
- vertical: $x = 2$.

Such an equation removes a degree of freedom! On the line, there is only one, and we are left with a single point. With two degrees of freedom on the plane, we have a line.

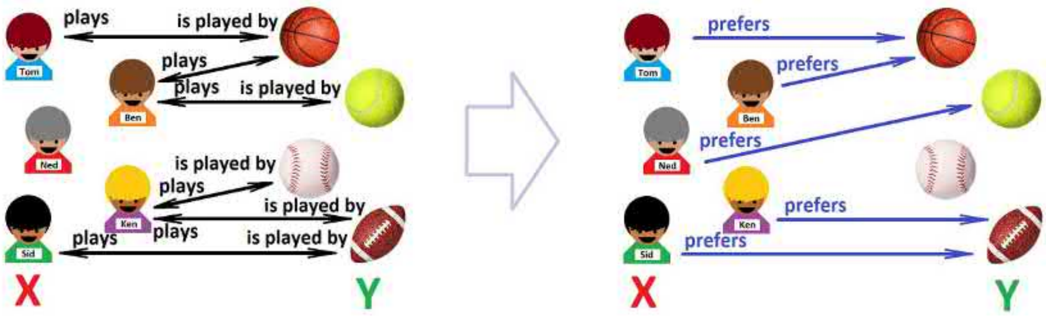
1.7. *Functions*

Let's go back to our running example and change the question from:

- "What sports has the boy played today?" to:
- "Which sport does the boy *prefer* to play?"

The idea is that everyone, even Ned, has a preference and exactly one.

Of course, the data for the first question (left) is different from that for the second (right):

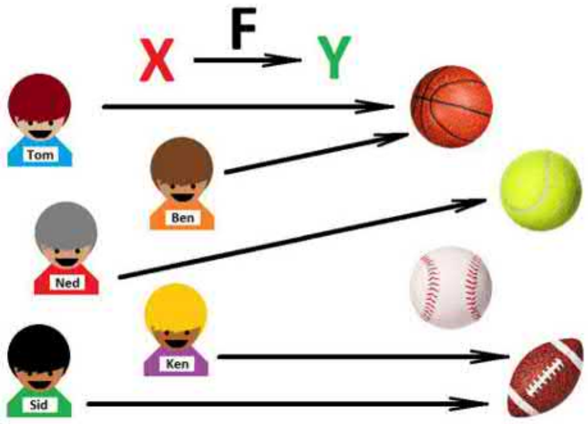


We erased one of the two arrows that start at Ben and one of the two arrows that start at Ken and we had to add an arrow for Ned.

In a relation, the two sets involved play equal roles. Instead, we now take the point of view of the boys. We will explore a new relation:

1. Tom prefers basketball.
2. Ben prefers basketball.
3. Ned prefers tennis.
4. Ken prefers football.
5. Sid prefers football.

We move from our two-ended arrows (or simply lines) to regular arrows:



We maintain the rule:

- ▶ There is exactly one y for each x.

As a result, the equality between the two sets is gone: X comes first, Y second.

This is a special kind of relation called a *function*; let's call this one F. What makes it special is that there is exactly one ball for each boy.

Below is the common notation:

Function from set to set

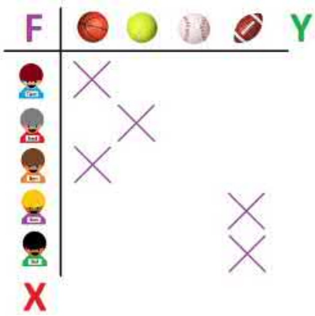
$$F : X \to Y$$

or

$$X \xrightarrow{F} Y$$

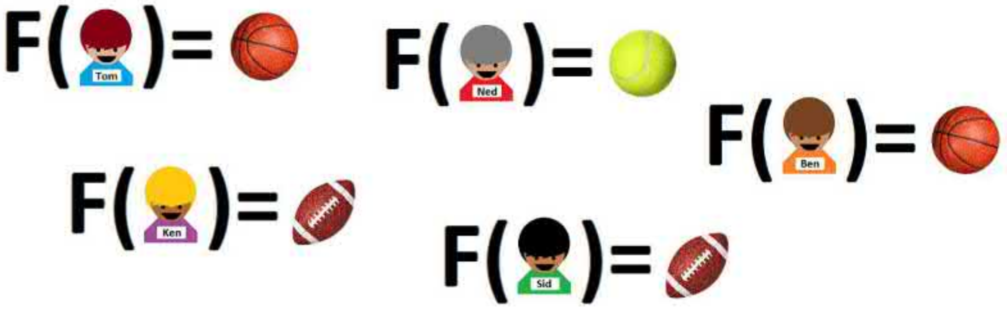
It reads "function F from X to Y".

Each element of X has only one arrow originating from it. Then, the table of this kind of relation must have exactly one mark in each row:



Our function F is a *procedure* that answers the question: "Which ball does this boy prefer to play with?" In fact, it answers *all* these questions! Conversely, a function is nothing but these answers. Each arrow clearly identifies the *input* – an element of X – of this procedure by its beginning and the *output* – an element of Y – by its end.

Each arrow in the diagram of F corresponds to a row of the table (and vice versa). The information contained in each is more commonly written in the *algebraic* manner, as follows:



The function F is then a question-answering machine: if you input the name of the boy, it will produce the name of the ball he prefers as the output.

This is the notation for the output of a function F when the input is x:

Input and output of function

$$F(x) = y$$

or

$$F : x \mapsto y$$

It reads: "F of x is y".

In other words, we have

$$F(\text{ input }) = \text{ output}$$

and

$$F : \text{ input } \mapsto \text{ output.}$$

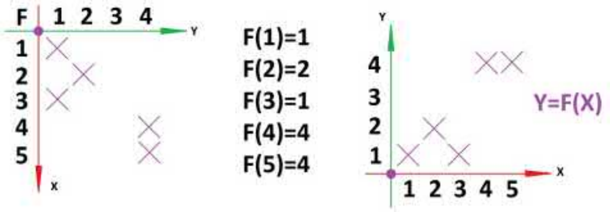
Just like any relation, a function can be represented in full by providing a list of pairs, x and y. This time, it's the *list of all inputs and their outputs*:

1. $F($ Tom $)$ = basketball
2. $F($ Ned $)$ = tennis
3. $F($ Ben $)$ = basketball
4. $F($ Ken $)$ = football
5. $F($ Sid $)$ = football

This notation will be, by far, the most common way of representing functions.

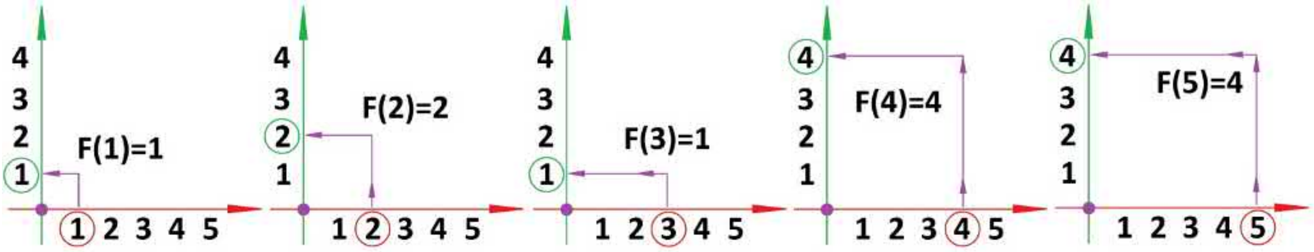
Throughout the early part of this book, we will concentrate on functions the inputs and the outputs of which are *numbers*.

To illustrate this idea, let's again *rename* the boys as numbers, $1-5$, and rename the balls as numbers, $1-4$. The table of our relation takes this form (seen on the left):



What makes the table of a *function* special is that it must have exactly one mark in each *column*. The values of F have also been rewritten (center). We then rotate the table counterclockwise (right) because it is traditional to have the inputs along a horizontal line (left to right) and the outputs along a vertical line (bottom to top).

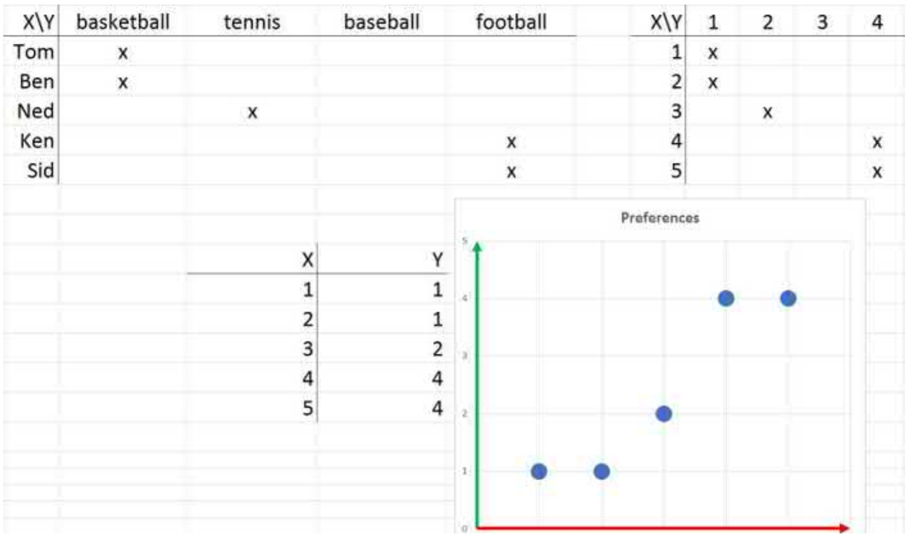
The latter table is called the *graph* of the function. The arrows can still be found:



Exercise 1.7.1

Finish the sentence: “This renaming of the boys (and the balls) is also a ___.”

We can put the data, as before, in a *spreadsheet* and then plot it automatically:



X\Y	basketball	tennis	baseball	football
Tom	x			
Ben	x			
Ned		x		
Ken				x
Sid				x

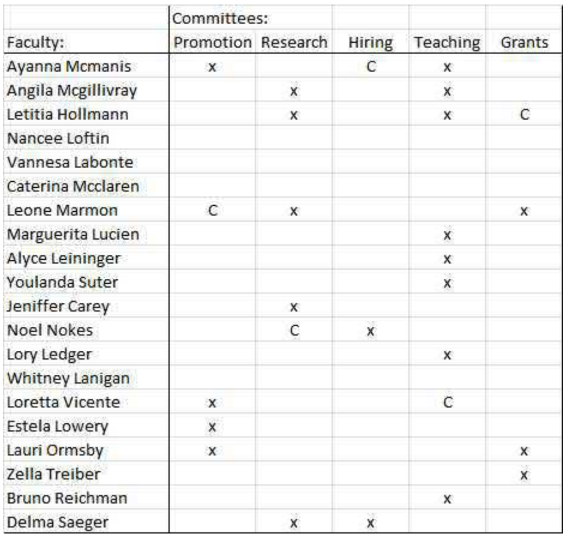
X\Y	1	2	3	4
1	x			
2	x			
3		x		
4				x
5				x

X	Y
1	1
2	1
3	2
4	4
5	4

There is only one cross in every row!

Example 1.7.2: relations and function in spreadsheets

Here is an example of how common spreadsheets are discovered to contain relations and functions. Below, we have a list of faculty members in the first column and a list of faculty committees in the first row. A cross mark indicates on which committee this faculty member sits, while “C” stands for “chair”:



	Committees:				
Faculty:	Promotion	Research	Hiring	Teaching	Grants
Ayanna Mcmanis	x		C	x	
Angila Mcgillivray		x		x	
Letitia Hollmann		x		x	C
Nancee Loftin					
Vannesa Labonte					
Caterina Mcclaren					
Leone Marmon	C	x			x
Marguerita Lucien				x	
Alyce Leininger				x	
Youlanda Suter				x	
Jeniffer Carey		x			
Noel Nokes		C	x		
Lory Ledger				x	
Whitney Lanigan					
Loretta Vicente	x			C	
Estela Lowery	x				
Lauri Ormsby	x				x
Zella Treiber					x
Bruno Reichman				x	
Delma Saeger		x	x		

The table gives a relation between these sets: $X = \{\text{ faculty }\}$ and $Y = \{\text{ committees }\}$; however, this is not a function. On the other hand, there is a function $F : Y \to X$ indicating the chair of the committee.

Exercise 1.7.3

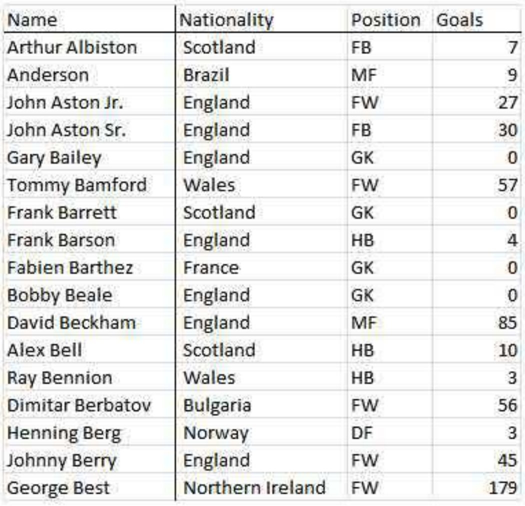
Think of other functions present in the spreadsheet.

Exercise 1.7.4

Suggest functions in the situation when an employer maintains a list of employees, with each person identified as a member of one of the projects.

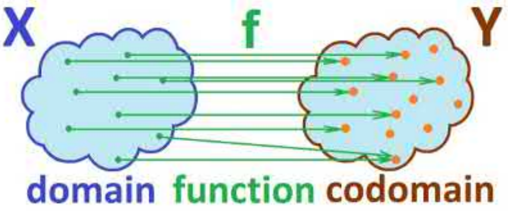
Exercise 1.7.5

What functions do you see below?



Name	Nationality	Position	Goals
Arthur Albiston	Scotland	FB	7
Anderson	Brazil	MF	9
John Aston Jr.	England	FW	27
John Aston Sr.	England	FB	30
Gary Bailey	England	GK	0
Tommy Bamford	Wales	FW	57
Frank Barrett	Scotland	GK	0
Frank Barson	England	HB	4
Fabien Barthez	France	GK	0
Bobby Beale	England	GK	0
David Beckham	England	MF	85
Alex Bell	Scotland	HB	10
Ray Bennion	Wales	HB	3
Dimitar Berbatov	Bulgaria	FW	56
Henning Berg	Norway	DF	3
Johnny Berry	England	FW	45
George Best	Northern Ireland	FW	179

A common way to visualize the concept of set – especially when the sets cannot be represented by mere lists – is to draw a shapeless blob in order to suggest the absence of any internal structure or relation between the elements. We then connect two such blobs by arrows:



This new concept is central to our study:

Definition 1.7.6: function

A *function* is a rule or procedure f that assigns to any element x in a set X, called the *input set* or the *domain* of f, exactly one element y, which is then denoted by

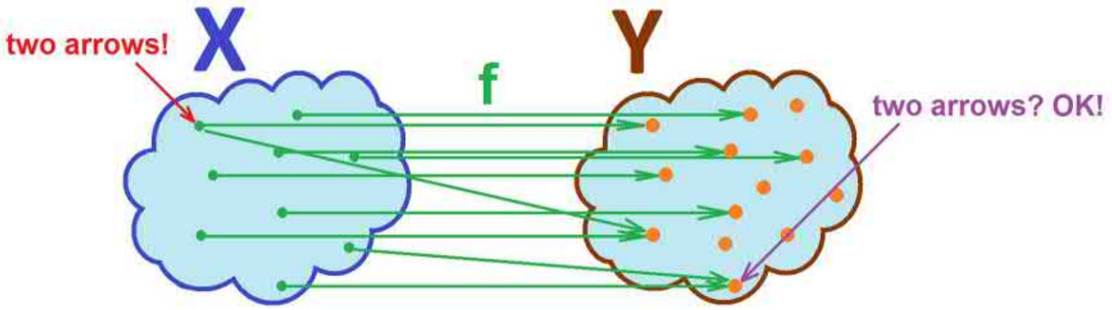
$$y = f(x)\,,$$

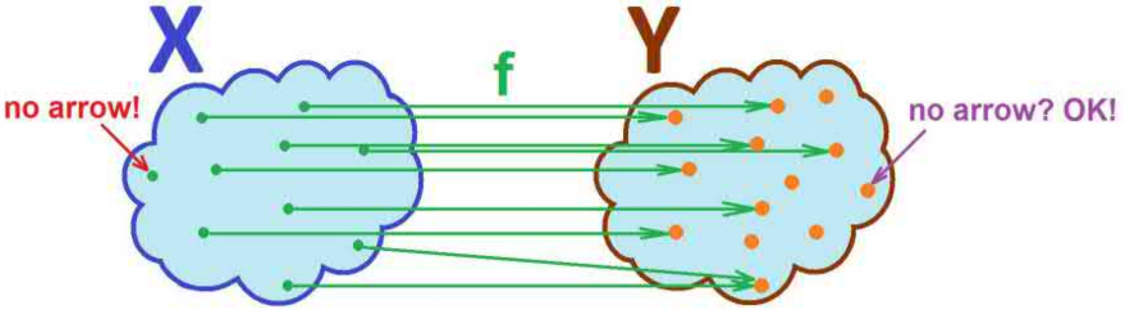
in another set Y. The latter set is called the *output set* or the *codomain* of f. The inputs are collectively called the *independent variable*; the outputs are collectively called the *dependent variable*. We also say that *the value of x under f* is y.

Warning!

In spite of the word "variable", there is no expectation of change. However, we are free to vary x, which makes y vary too.

This definition fails for a relation that has too few or too many arrows for a given x. Below, we illustrate how the requirement may be violated, in the domain (two on the left):





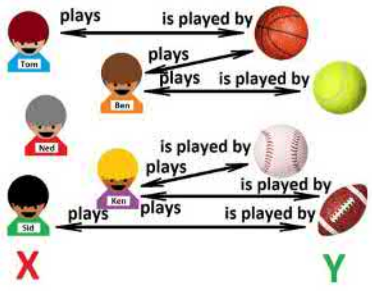
These are *not* functions. Meanwhile, we also see what shouldn't be regarded as violations, in the codomain (two on the right).

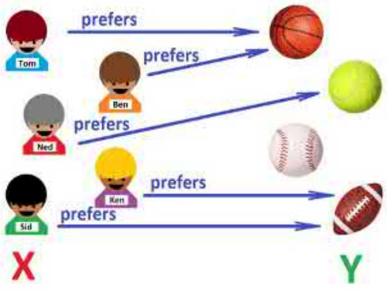
The main purpose of this section is to discuss the transition from relations to functions.

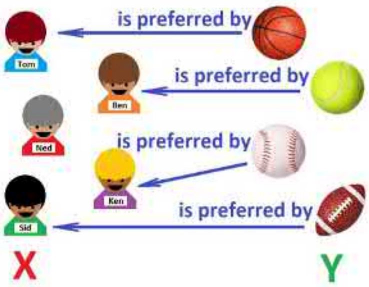
Now, this transition can happen in two ways:

1. A relation between X and Y becomes a function from X to Y.
2. A relation between X and Y becomes a function from Y to X.

We add the latter option to the example above:







In general, the following is true:

Theorem 1.7.7: When Relation Is Function

Suppose X and Y are sets and R is a relation between X and Y. Then:

- *The relation R represents some function F from X to Y, $F : X \to Y$, if and only if for each x in X there is exactly one y in Y such that x and y are related by R.*
- *The relation R represents some function G from Y to X, $G : Y \to X$, if and only if for each y in Y there is exactly one x in X such that x and y*

are related by R.

Exercise 1.7.8

Give an algebraic example of how (a) both of the two conditions can be violated, (b) both of the conditions are satisfied. Repeat the task in the boys-and-balls setting.

Exercise 1.7.9

Split either part of the theorem into a statement and its converse.

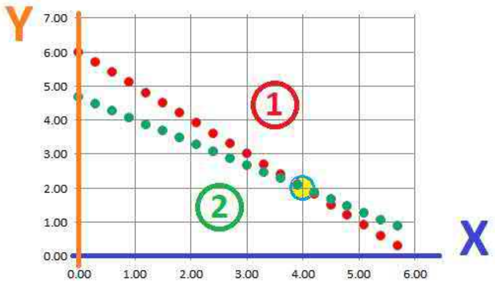
When our sets are sets of numbers, the relations are often given by *formulas*. In that case, the above issue is resolved with algebra.

Exercise 1.7.10

What function can you think of from the set X of the boys to the sets of: letters, numbers, colors, geographic locations? Think of others.

Let's again review how these concepts appeared in the problem of mixtures.

Recall that x and y are the weight of the Kenyan coffee and the weight of Colombian coffee respectively. They come from sets X and Y, both of which are copies of the set of real numbers $\mathbf{R}$. There are also two relations:



The lines are the graphs of the two relations that come from restrictions on the mixtures that we want.

These relations can be converted to functions by solving for one of the two variables. First:

$$\boxed{1} \quad x+y=6 \quad \begin{array}{l} \nearrow \quad x=6-y \\ \searrow \quad y=6-x \end{array}$$

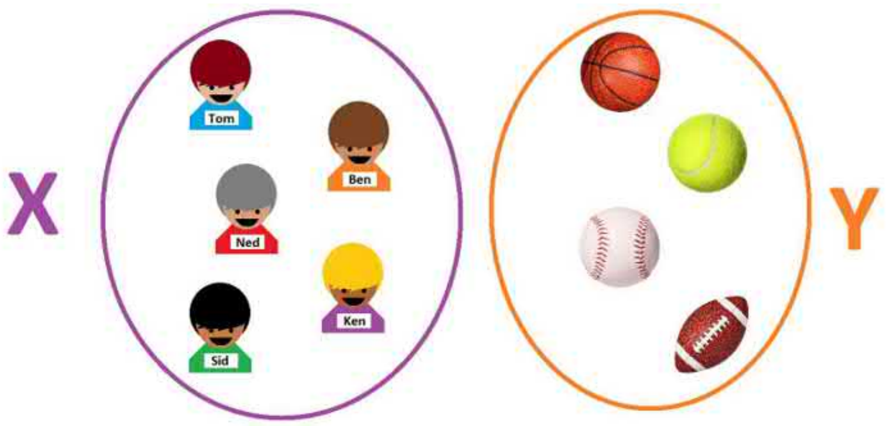
Second:

$$\boxed{2} \quad 2x+3y=14 \quad \begin{array}{l} \nearrow \quad x=\dfrac{14-3y}{2} \\ \searrow \quad y=\dfrac{14-2x}{3} \end{array}$$

As a result, we can just plug in as many x's as necessary producing the corresponding y's automatically. The same for y.

1.8. Operations on sets

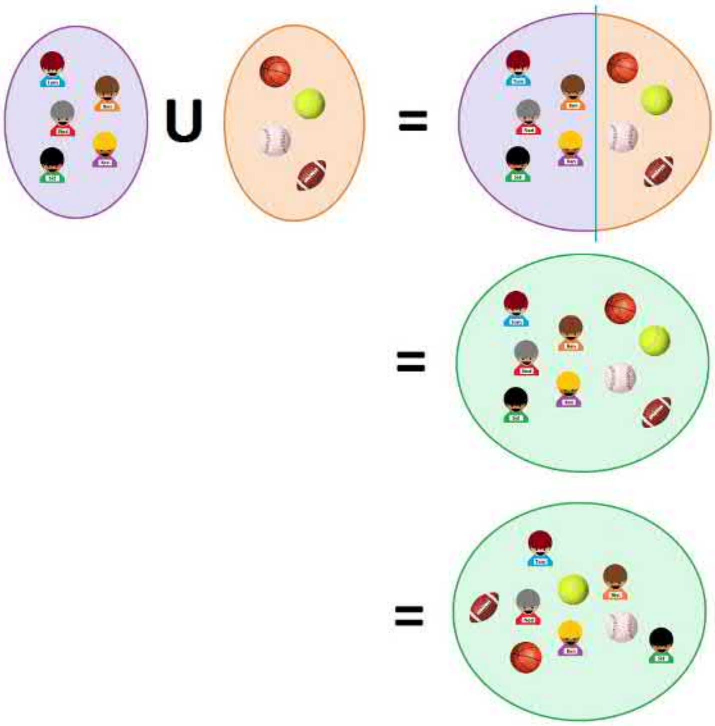
Let's go back to our example of the five boys that form a set and another set is the set of these four balls:



They are just lists without repetitions:

$$\begin{aligned} X &= \{\ \text{Tom}\ ,\ \text{Ken}\ ,\ \text{Sid}\ ,\ \text{Ned}\ ,\ \text{Ben}\ \} \\ Y &= \{\ \text{basketball}\ ,\ \text{tennis}\ ,\ \text{baseball}\ ,\ \text{football}\ \} \end{aligned}$$

We can form a new set that contains all the elements of the two sets, as follows:



We can merge the two sets together and, once the separator is gone, move the elements around. The following concept will be routinely used:

Definition 1.8.1: union of sets

The *union* of any two sets X and Y is the set that consists of the elements that belong to either X or Y. It is denoted by

$$\boxed{X \cup Y}$$

The most common case is when both sets are *subsets* of the same set. For example, all of these are subsets of $X \cup Y$:

$$\begin{array}{ll} T = \{\ \text{Tom}\ \},\quad A = \{\ \text{Tom, Ken}\ \},\quad Q = \{\ \text{Tom, Ken, Sid}\ \},\quad \ldots & \subset X \\ B = \{\ \text{basketball}\ \},\quad V = \{\ \text{basketball, tennis}\ \},\quad U = \{\ \text{basketball, tennis, baseball}\ \},\quad \ldots & \subset Y \\ \Longrightarrow & \\ \{\ \text{Tom}\ \} \cup \{\ \text{basketball}\ \} = \{\ \text{Tom, basketball}\ \} & \subset X \cup Y \\ \{\ \text{Tom, Ken}\ \} \cup \{\ \text{basketball, tennis}\ \} = \{\ \text{Tom, Ken, basketball, tennis}\ \} & \subset X \cup Y \end{array}$$

Exercise 1.8.2

Present the unions of all *pairs* of sets shown above. Now, define the unions of *triples* of sets and present them. Continue.

So, we check for each x the following:

x belongs to X `OR` x belongs to Y

Warning!

"X `OR` Y" means "X `OR` Y or both".

Exercise 1.8.3

Solve for Y:

$$X \cup Y = X\,.$$

So, finding the union of two sets given as lists is very simple:

- ▶ We merge the lists removing repetitions.

These repetitions comprise the "overlap" of the two sets:

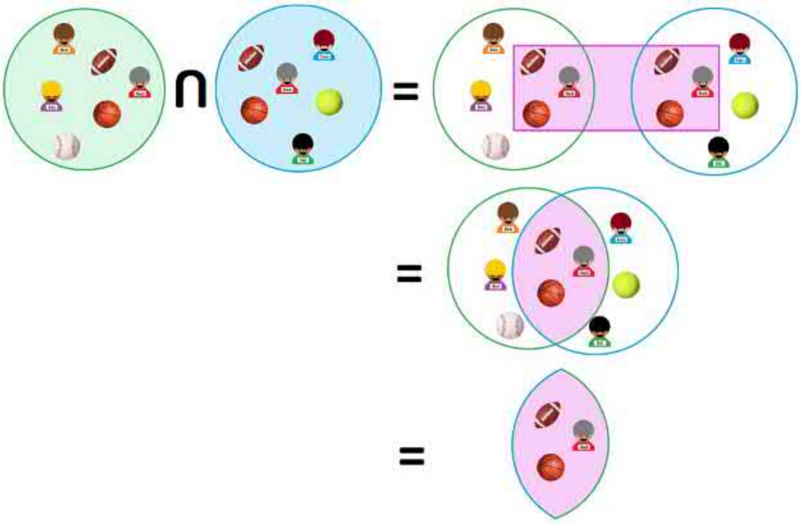
$$\begin{aligned}\{\text{ tennis, Tom, Ken }\} \cup \{\text{ basketball, tennis, Tom }\} &= \{\text{ tennis, Tom, Ken, basketball, tennis, Tom }\}\\ &= \{\text{Tom, Ken, basketball, tennis }\}.\end{aligned}$$

Example 1.8.4: unions of lists

We find the union in two steps:

$$\begin{aligned}\{1,2,3,4\} \cup \{3,4,5,6,7\} &= \{1,2,3,4,\quad 3,4,5,6,7\} && \text{Merge the lists.}\\ &= \{1,2,3,4,5,6,7\}\,. && \text{Then remove repetitions.}\end{aligned}$$

Next, the "overlap" itself. We look at what the two sets have *in common*:



We find the repeated parts and merge each into one; the rest is thrown out. The following concept will be routinely used:

Definition 1.8.5: intersection of sets

The *intersection* of any two sets X and Y is the set that consists of all the elements that belong to both X and Y. It is denoted by

$$\boxed{X \cap Y}$$

Here is an example of such a computation:

$$\{\text{ tennis, Tom, Ken }\} \cap \{\text{ basketball, tennis, Tom }\} \ = \{\text{ tennis, Tom }\}.$$

So, we check for each x the following:

$$\boxed{x \text{ belongs to } X \quad \texttt{AND} \quad x \text{ belongs to } Y}$$

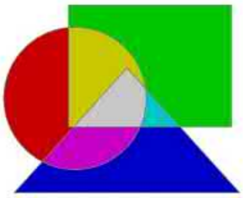
Exercise 1.8.6

Solve for Y:

$$X \cap Y = X\,.$$

Example 1.8.7: colors

The *primary colors* are red, green, and blue. The rest of the colors are seen as “combinations” of these three:



We can also interpret them as intersections.

Recall that the set-building notation is used to create a set by stating a condition these numbers are supposed to satisfy:

$$Z = \{x : \quad \text{condition for } x\ \}\,.$$

We consider what happens to sets given this way under these two operations.

Example 1.8.8: solution sets of systems of equations

Numerical sets are subsets of the real line $\mathbf{R}$ and some of them came from solving these equations, as their *solution sets*:

equation	solution set	simplified
$x^2 - 3x + 2 = 0$	$X = \{x :\ x^2 - 3x + 2 = 0\}$	$= \{1, 2\}$
$x^2 = 1$	$Y = \{x :\ x^2 = 1\}$	$= \{-1, 1\}$

What if we have both equations to be satisfied *at the same time*? We are interested in the set:

$$\begin{array}{lllll}
Z & = \{x :\ x^2 - 3x + 2 = 0 & \texttt{AND} & x^2 = 1\ \} & \\
 & & & \Longrightarrow & x \text{ belongs to both } X \quad \texttt{AND} \quad Y\ ! \\
Z & = X & \cap & Y & \\
 & = \{x :\ x^2 - 3x + 2 = 0\} & \cap & \{x :\ x^2 = 1\} & \\
 & = \{1, 2\} & \cap & \{-1, 1\} & \\
 & & = \{1\}\,. & &
\end{array}$$

Now, what if we need *just one* of the equations to be satisfied? We are interested in the set:

$$\begin{array}{lllll} Z & = \{x:\ x^2-3x+2=0 & \texttt{OR} & x^2=1\,\} & \\ & & & \Longrightarrow & x \text{ belongs to } X\ \ \texttt{OR}\ \ Y\ ! \\ Z & = X & \cup & Y & \\ & = \{x:\ x^2-3x+2=0\} & \cup & \{x:\ x^2=1\} & \\ & = \{1,2\} & \cup & \{-1,1\} & \\ & & = \{-1,1,2\}\,. & & \end{array}$$

So, sometimes the condition that defines a set splits into *two* conditions:

$$Z=\{x:\ \ \text{satisfies first condition}\ \ \texttt{AND}\ \ \text{satisfies second condition}\ \}\,.$$

The word "`AND`" is capitalized in order to emphasize that the set contains only those x's that satisfy *both* conditions simultaneously. Then we can see also that there are *two* sets:

$$X=\{x:\ \ \text{satisfies first condition}\ \}\ \text{and}\ Y=\{x:\ \ \text{satisfies second condition}\ \}\,,$$

one for either condition. We are interested in their intersection:

$$Z=X\cap Y=\{x:\ \ \text{satisfies first condition}\ \}\cap\{x:\ \ \text{satisfies second condition}\ \}\,.$$

Example 1.8.9: solution sets of inequalities

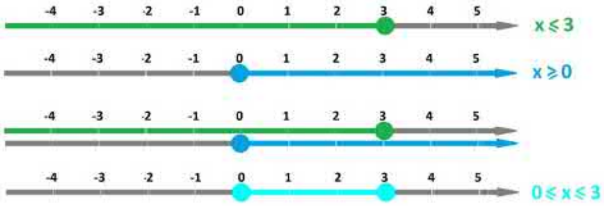
These conditions can also be inequalities. Let's consider this "double inequality":

$$0\le x\le 3\,.$$

Its solution set is all x's that satisfy the inequality. However, there are *two* inequalities in reality and two solution sets:

$$\begin{array}{llll} & \{x:\ 0\le x\le 3\} & & \\ = & \{x:\ x\ge 0 & \texttt{AND} & x\le 3\} \\ = & \{x:\ x\ge 0\} & \cap & \{x:\ x\le 3\} \\ = & [0,+\infty) & \cap & (-\infty,3] \\ = & & [0,3]\,. & \end{array}$$

The geometric interpretation is shown below:

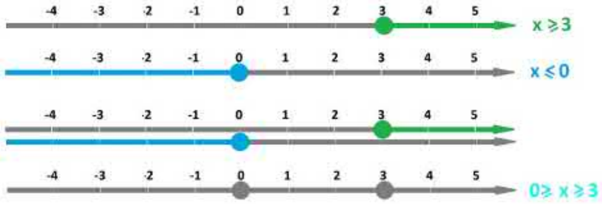


What if we "flip" both signs of this inequality:

$$0\ge x\ge 3\ ?$$

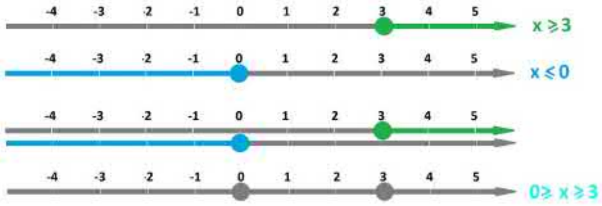
Then:

$$\begin{array}{llll} & \{x:\ 0\ge x\ge 3\} & & \\ = & \{x:\ x\le 0 & \texttt{AND} & x\ge 3\} \\ = & \{x:\ x\le 0\} & \cap & \{x:\ x\ge 3\} \\ = & (-\infty,0] & \cap & [3,+\infty) \\ = & & \varnothing\,. & \end{array}$$



What if we are interested in when the original inequality is *not* satisfied? This is our set:

$$\begin{array}{rllll} & \{x:\ 0\le x\le 3, & \texttt{FALSE}\} & & \\ = & \{x:\ x\ge 0 & & \texttt{AND} & x\le 3,\ \texttt{FALSE}\} \\ = & \{x:\ x\ge 0, & \texttt{FALSE} & \texttt{OR} & x\le 3,\ \texttt{FALSE}\} \\ = & \{x:\ x\le 0 & & \texttt{OR} & x\ge 3\} \\ = & \{x:\ x\le 0\} & & \cup & \{x:\ x\ge 3\} \\ = & (-\infty,0] & & \cup & [3,+\infty)\,. \end{array}$$



Exercise 1.8.10

What happens if you flip only one of the two inequalities above?

As a summary, let's rephrase our two definitions.

Definition 1.8.11: union and intersection of sets

$$X\cup Y=\{x:\ x \text{ belongs to } X\quad \texttt{OR}\quad x \text{ belongs to } Y\}$$

$$X\cap Y=\{x:\ x \text{ belongs to } X\quad \texttt{AND}\quad x \text{ belongs to } Y\}$$

Warning!

The word `OR` in mathematics is always meant *inclusively*: not as in "black or white" but as in "cream or sugar".

When the two sets are defined via conditions, we have the following for their intersection:

$$\begin{array}{rl} & \{x:\ \text{ first condition }\ \texttt{AND}\ \text{ second condition }\} \\ = & \{x:\ \text{ first condition }\}\cap\{x:\ \text{ second condition }\}; \end{array}$$

and for the union:

$$\begin{array}{rl} & \{x:\ \text{ first condition }\ \texttt{OR}\ \text{ second condition }\} \\ = & \{x:\ \text{ first condition }\}\cup\{x:\ \text{ second condition }\}. \end{array}$$

Example 1.8.12: truth tables

The verification of these combinations of two conditions can be sped up by the following tables:

#1	#2	#1 `AND` #2
TRUE	TRUE	TRUE
TRUE	FALSE	FALSE
FALSE	TRUE	FALSE
FALSE	FALSE	FALSE

#1	#2	#1 `OR` #2
TRUE	TRUE	TRUE
TRUE	FALSE	TRUE
FALSE	TRUE	TRUE
FALSE	FALSE	FALSE

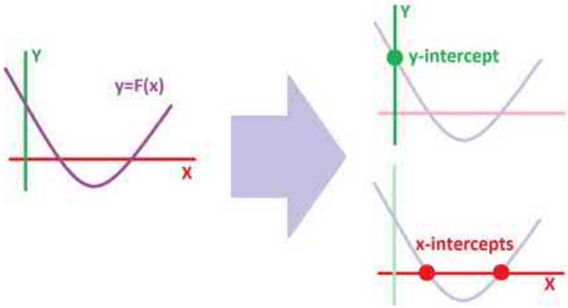
We next consider *subsets of the plane*.

Example 1.8.13: intercepts

For a numerical function F, its *graph* is the following set presented via the set-building notation:

$$\{(x,y):\ y=F(x)\}\subset\mathbf{R}^2\,.$$

Now, the two especially important subsets of the xy-plane are its two *axes*. Now, let's put these together:



The x-*intercepts* of F are the elements of the intersection of the graph of F with the x-axis:

$$\{\ x\text{-intercepts}\ \} = \text{graph of } F\ \cap\ x\text{-axis}.$$

Meanwhile, the y-*intercept* of F is the (possible) element of the intersection of the graph of F with the y-axis:

$$\{\ y\text{-intercept}\ \} = \text{graph of } F\ \cap\ y\text{-axis}.$$

Warning!

Don't confuse an element of a set and a one-element set.

Exercise 1.8.14

Give examples of functions for which these sets are empty.

Exercise 1.8.15

Explain what these sets are:

- x-axis $\cap$ y-axis
- $\{\ x\text{-intercepts}\ \}\cap\{\ y\text{-intercepts}\ \}$

Let's again consider how these concepts appeared in the problem of mixtures.

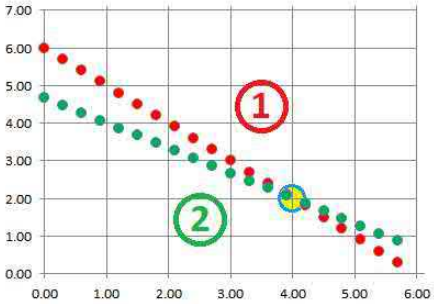
We produced two linear relations, each representing a line on the xy-plane:

$$\begin{cases} x & + & y & = & 6\,,\\ 2x & + & 3y & = & 14\,. \end{cases}$$

The graphs are respectively:

$$\{(x,y):\ x+y=6\}\ \text{ and }\ \{(x,y):\ 2x+4y=14\}\,.$$

They are these lines:



We conclude that the *intersection* of the two lines is the point (x,y) that is the solution of the *system* of equations formed by these two equations. Just as before, when there are two equations, there are two sets:

$$\begin{array}{rlcr}
 & \{(x,y):\ x+y=6 & \texttt{AND} & 2x+3y=14\}\\
= & \{(x,y):\ x+y=6\} & \cap & \{(x,y):\ 2x+3y=14\}\\
= & & \{(4,2)\}\,. &
\end{array}$$

The intersection is a single point!

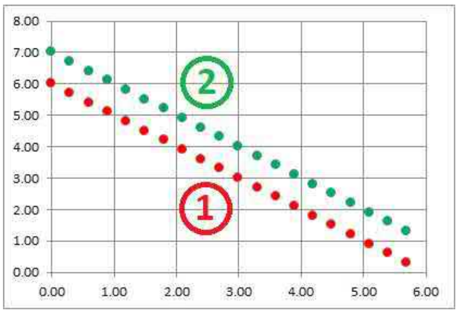
The existence of an intersection point tells us that it is *possible* to create such a blend. We say that the intersection is *non-empty*:

$$\{(x,y):\ x+y=6\}\cap\{(x,y):\ 2x+3y=14\}\neq\varnothing\,.$$

Now, what if the Colombian coffee is also priced at \$2 per pound? Then it is *impossible* to create such a blend. We say that the intersection is *empty*:

$$\{(x,y):\ x+y=6\}\cap\{(x,y):\ 2x+2y=14\}=\varnothing\,.$$

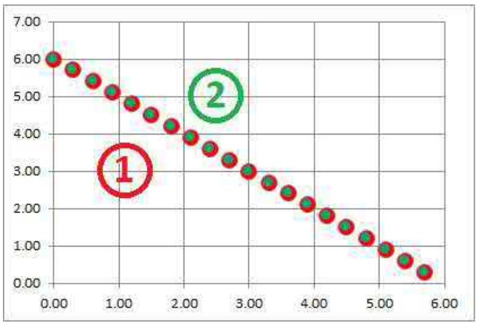
The lines are parallel:



The third possibility occurs when we change, in addition, the total price of the blend to \$12:

$$\{(x,y):\ x+y=6\}\cap\{(x,y):\ 2x+2y=12\}=\{(x,y):\ x+y=6\}\,.$$

The lines merge:



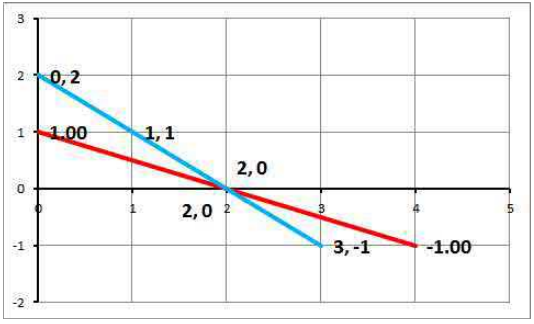
There are infinitely many possible mixtures.

Exercise 1.8.16

Carry out the computations and provide visualization for the second and third possibilities in the last example.

Exercise 1.8.17

Find algebraic representations of these two lines and repeat the analysis above:



Let's modify our problem as follows:

> ▶ PROBLEM: Suppose we have the Kenyan coffee that costs \$2 per pound and the Colombian coffee that costs \$3 per pound. How much of each do you need to have *at least* 6 pounds of blend with the total price of *at most* \$14?

The variables are the same, but the restrictions change. They are *inequalities* now!

Since the total weight is more than or equal to 6, we have an inequality that connects x and y:

$$\boxed{1} \quad x + y \geq 6\,.$$

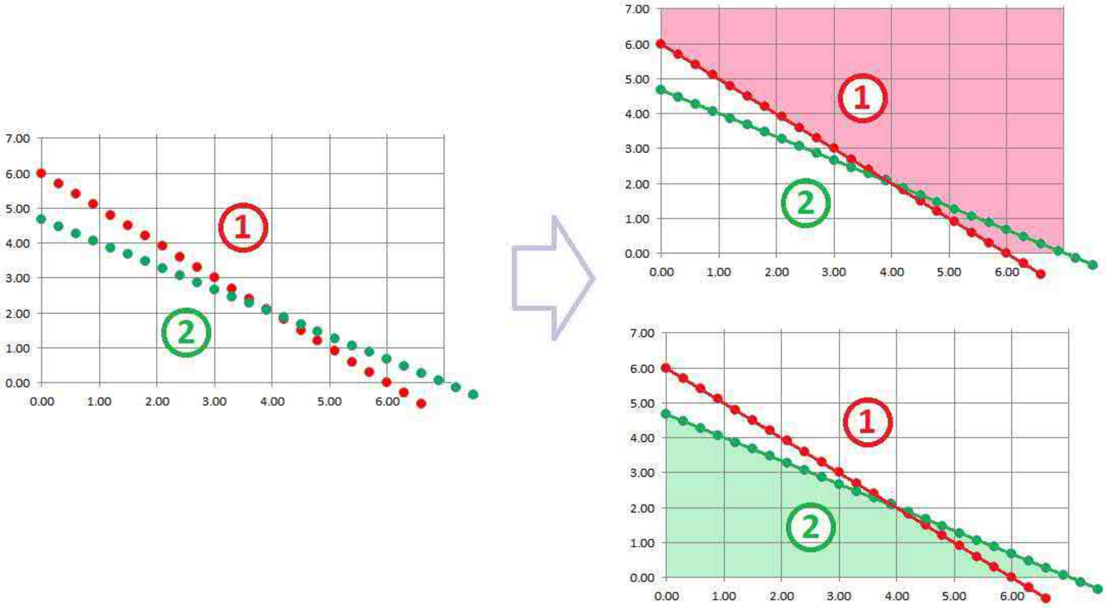
Since the total price of the blend is less than or equal to \$14, we have another inequality for x and y:

$$\boxed{2} \quad 2x + 3y \leq 14\,.$$

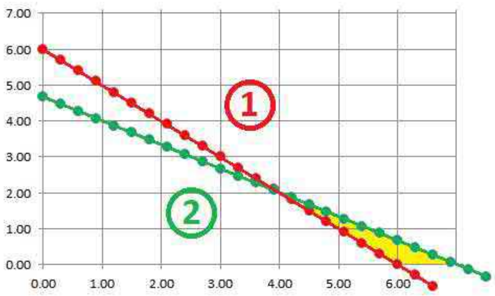
These are the sets we are interested in:

$$\{(x, y) : \; x + y \geq 6\} \quad \text{and} \quad \{(x, y) : \; 2x + 3y \leq 14\}\,.$$

We discover that these sets are the half-planes bounded by the lines that we have been talking about:



Since both of the conditions have to be satisfied, our interest is, again, the intersection:



The yellow triangle represents all acceptable mixtures.

Exercise 1.8.18

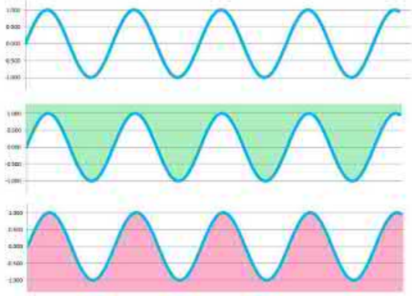
Modify the other two versions of the problem accordingly and solve them in this manner.

Exercise 1.8.19

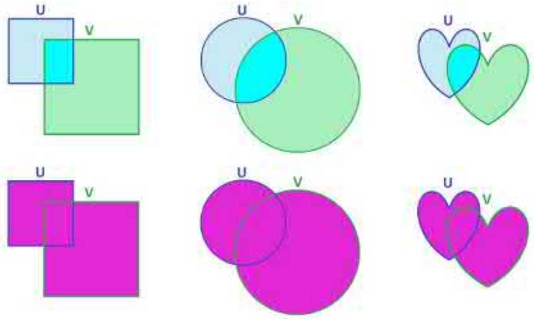
Evaluate:

$$\{(x,y):\ y \geq f(x)\ \} \cap \{(x,y):\ y \leq f(x)\ \}\,.$$

Hint:



More examples of these two operations for subsets of the plane are shown below:



These "generic" sets serve as illustrations of the idea of the union and the intersection.

1.9. *Linear relations and functions*

Recall that a relation between two sets is any pairing of their elements. This time, the sets are sets of numbers and the condition to be checked is an equation:

sets: elements:
$\mathbf{R} \to x$ Related! Add to the list.
$\searrow$ relation: `TRUE` $\nearrow$
$\boxed{x+y=2\,?} \to$
$\nearrow$ `FALSE` $\searrow$
$\mathbf{R} \to y$ Not related!

So, a numerical relation processes a pair of numbers x and y and tells only one thing: related or not related. For example:

$$\begin{array}{lllll} x=1,\ y=2 & \to & 1+2=2? & \texttt{FALSE} & \to \text{ Not related!} \\ x=1,\ y=1 & \to & 1+1=2? & \texttt{TRUE} & \to \text{ Related!} \end{array}$$

Warning!

Just because both sets are the sets of real numbers ($X = \mathbf{R}$ and $Y = \mathbf{R}$), we don't have to think of the relation as one of a set with *itself*.

Example 1.9.1: maximizing enclosure

A typical calculus problem asks to maximize a rectangular enclosure made of 100 yards of fencing. We choose width x and height y, and the two are related:

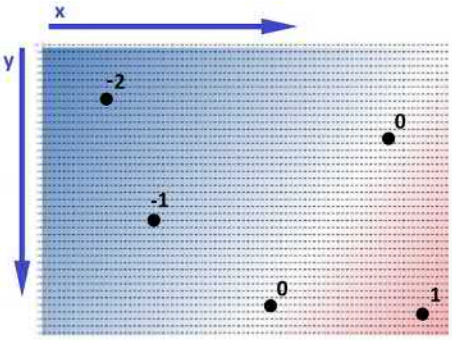
$$x + y = 50\,.$$

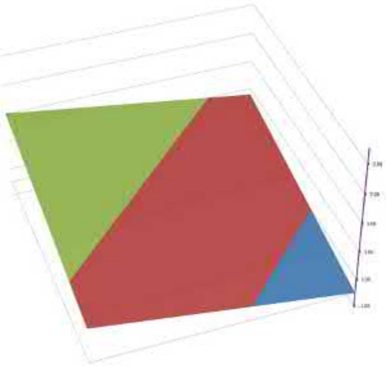
To speed up the analysis, we pre-computed *all* values of $x + y$ for every eligible pair x and y. The result is a table filled by means of the following spreadsheet formula:

`=RC2+R2C`

It is easy in a small table to color by hand the cells with value of $x + y$ equal to 50 (left); they form the yellow line:

x\y	0	10	20	30	40	50	60
0	0	10	20	30	40	50	60
10	10	20	30	40	50	60	70
20	20	30	40	50	60	70	80
30	30	40	50	60	70	80	90
40	40	50	60	70	80	90	100
50	50	60	70	80	90	100	110
60	60	70	80	90	100	110	120
70	70	80	90	100	110	120	130
80	80	90	100	110	120	130	140





In a very large array of cells (middle), we the ability of the spreadsheet to to color the cells according to the values they contain: The color of the (x, y)-cell is determined by the value of $x + y$. The linear pattern still seems conceivable. The value of $z = x + y$ can also be visualized as the elevation at that location (right).

What is the scope of possible inputs in the above diagram? Any value of x is possible and, independently, any value of y. Therefore, all *pairs* (x, y) are possible.

We discover that *plotting the graph* of a numerical relation means processing a pair of numbers (x, y), one at a time. What is the output? Related or not related, Yes or No, `TRUE` or `FALSE`, a point or no point.

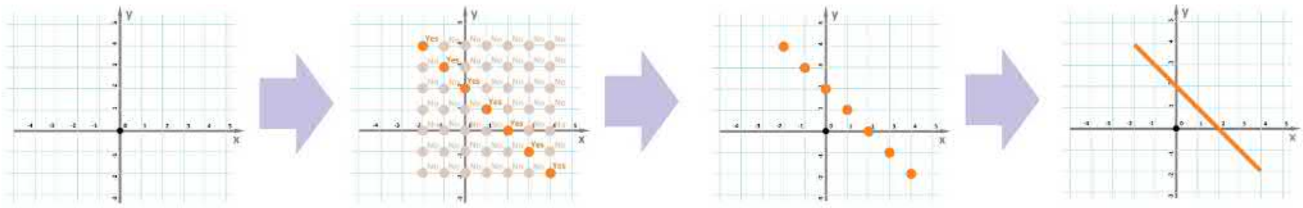
We can remake the above diagram in this spirit:

$$\begin{tikzcd} & & & \text{outcome: Plot point } (x,y). \\ \text{plane: } \mathbf{R}^2 \arrow[r] & \text{pair: } (x,y) \arrow[r] & \text{relation: } \boxed{x+y=2\,?} \arrow[ur, "\texttt{TRUE}"] \arrow[dr, "\texttt{FALSE}"'] & \\ & & & \text{Don't plot anything.} \end{tikzcd}$$

We can try to do this plotting by hand, one point at a time:

$$(0,0) \to \text{ No!} \quad (1,0) \to \text{ No!} \quad (1,1) \to \text{ Yes! ...}$$

It takes a lot of tries to produce a picture that reveals a pattern:



On the far right, we show our conjecture about the graph of the relation; it looks like a straight line. In retrospect, it's like throwing a dart at the plane trying to hit that thin line!

Exercise 1.9.2

Show that the equation $2x + 2y = 4$ represents the same relation!

In this chapter, we saw two examples of relations the graphs of which are lines:

- The relation $y = c$ produces a horizontal line because every point (x, y) is plotted as long as $y = c$ (there is no restriction on x).
- The relation $x = a$ produces a vertical line because every point (x, y) is plotted as long as $x = a$ (there is no restriction on y).

As a summary, we give a precise definition:

Definition 1.9.3: graph of relation

Suppose R is a relation between two sets X and Y of real numbers. Then, the *graph* of R is the set of all points on the xy-plane with x and y related by R:

$$\text{graph of } R \;=\; \{(x,y) : \ x \ \text{ is related to } \ y\}\,.$$

We use the set-building notation. This relation is, typically, an equation, and in this case, "most" of the points on the plane won't satisfy it. Those that do will likely form a *curve*.

We start with the simplest, and the most common, kind. *Linear relations* only allow constant multiplication and addition: the relation

$$2x + 3y = 100\,,$$

involves only multiplication of x by 2, multiplication of y by 3, and then adding them together.

The result below explains the name:

Theorem 1.9.4: Graph of Linear Relation

The graph of any linear relation, i.e.,

$$Ax + By = C\,,$$

with either A or B not equal to zero, is a straight line.

Exercise 1.9.5

What is the graph when $A = B = 0$? Hint: There are two cases.

Exercise 1.9.6

State the converse of the theorem and find out if it's true.

It is called an *implicit equation of the line*:

$$\boxed{Ax + By = C}$$

When we represent the line by a function, the equation becomes *explicit*.

The ideas of *linear algebra* have a very humble beginning.

Example 1.9.7: linear equation

Suppose we have a type of coffee that costs \$3 per pound. How much do we get for \$60?

The setup is the following. Let x be the weight of the coffee. Since the total price is 60, we have a *linear equation*:

$$3x = 60\,.$$

We solve it:

$$x = \frac{60}{3} = 20\,.$$

The algebraic operations are very simple, and the complexity comes from elsewhere: the number of variables.

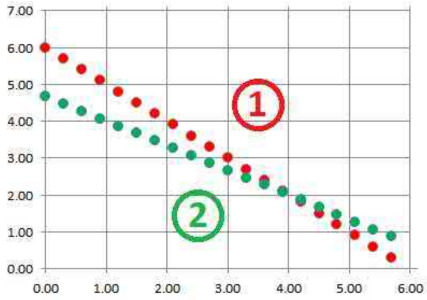
One more time: we have the Kenyan coffee that costs \$2 per pound and the Colombian coffee that costs \$3 per pound. Since the total weight is 6, we have a linear relation between x and y:

$$\boxed{1}\quad x + y = 6\,.$$

Since the total price of the blend is \$14, we have another linear relation between x and y:

$$\boxed{2}\quad 2x + 3y = 14\,.$$

According to the theorem, the graphs of the relations are lines, these lines:



Then, for a combination of weights x and y to satisfy *both* of the requirements, the point (x, y) has to belong to *both* of the lines! This point is $(x, y) = (4, 2)$, which can be confirmed by substituting the two numbers $x = 4,\ y = 2$ into the two relations:

$$\begin{array}{lllllllll}\boxed{1} & x & +y & = 6 & \to & 4 & +2 & = 6 & \texttt{TRUE}\\ \boxed{2} & 2x & +3y & = 14 & \to & 2\cdot 2 & +3\cdot 2 & = 14 & \texttt{TRUE}\end{array}$$

We also solved the problem algebraically. Solving for one of the variables creates a function from either of these relations:

$$\begin{array}{llllll}\boxed{1} & x & +y & = 6 & \Longrightarrow & y = 6 - x\\ \boxed{2} & 2x & +3y & = 14 & \Longrightarrow & y = \frac{1}{3}(14 - 2x)\end{array}$$

But this is supposed to be the same number:

$$y = 6 - x = \frac{1}{3}(14 - 2x)\,.$$

Therefore, $x = 4$. Substitute this back into the first function: $y = 6 - x = 6 - 4 = 2$.

Such a problem is called a *system of linear equations*.

Exercise 1.9.8

Solve the problem by making the relations functions of y.

Exercise 1.9.9

Set up a system of linear equations – but do not solve it – for the following problem: “An investment portfolio worth $\$1,000,000$ is to be formed from the shares of: Microsoft - \$5 per share, and Apple - \$7 per share. If you need to have twice as many shares of Microsoft than Apple, what are the numbers?”

Example 1.9.10: nutrition

Consider the following problem: “One serving of tomato soup contains 100 cal and 18 g of carbohydrates. One slice of whole bread contains 70 cal and 13 g of carbohydrates. How many servings of each should be required to obtain 230 cal and 42 g of carbohydrates?”

Let’s collect the information in a table:

	soup	bread	meal
carbs	100	70	230
cal	18	13	42

We explicitly introduce the variables:

1. Let x be the number of serving of soup.
2. Let y be the number of serving of bread.

Then the table can be enhanced:

	x servings of soup	$+\ y$ servings of bread	$=$	meal
carbs	$100x$	$+\ 70y$	$=$	230
cal	$18x$	$+\ 13y$	$=$	42

Exercise 1.9.11

Finish the problem.

Exercise 1.9.12

Solve the system of linear equations:

$$\begin{cases} x - y &= 2\,, \\ x + 2y &= 1\,. \end{cases}$$

Exercise 1.9.13

Solve the system of linear equations and geometrically represent its solution:

$$\begin{cases} x - 2y &= 1\,, \\ x + 2y &= -1\,. \end{cases}$$

Exercise 1.9.14

Geometrically represent this system of linear equations:

$$\begin{cases} x - 2y &= 1\,, \\ x + 2y &= -1\,. \end{cases}$$

Exercise 1.9.15

What if there is a third type of coffee in the example, say \$4 per pound?

There may be many relations with the same graph:

$$\begin{tikzcd}
2x+2y=4 & x+y-2=0 & -x-y+2=0 \\
x+x+y+y=4 & \boxed{x+y=2} \arrow[ul] \arrow[u] \arrow[ur] \arrow[l] \arrow[r] \arrow[dl] \arrow[d] \arrow[dr] & \textcolor{red}{y=-x+2} \\
-y=x-2 & -x=y-2 & \textcolor{green}{x=-y+2}
\end{tikzcd}$$

But two of them are special: they are *functions*. One is y in terms of x and the other x in terms of y.

Warning!

Either of those two can also have a different representation: $y=-x+2$ and $y=1-x+3$.

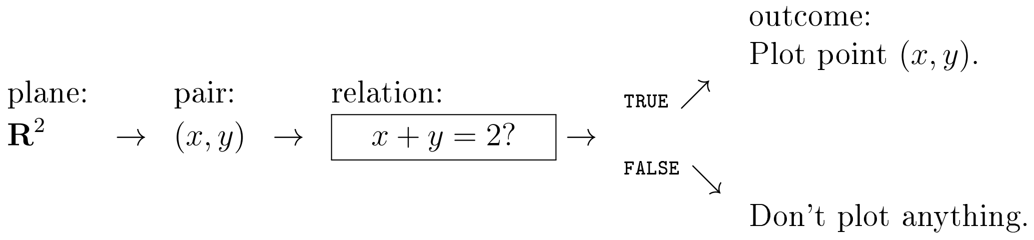
It is crucial that *graphs are sets* too; they are subsets of $\mathbf{R}^2$. In fact, we can still use the set-building notation:

$$\{(x,y):\ \text{condition on } x,\ y\}\,.$$

This condition, just as before, is often an equation; for example:

$$\{(x,y):\ x+y-2=0\}\,.$$

Because of the indirect nature of the definition of this set, plotting the graph of a numerical relation is cumbersome:



What else can we do?

We proved that the graph of any *linear* relation, i.e.,

$$Ax+By=C\,,$$

with either A or B not equal to zero, is a straight line. To create a function, all we need is to solve for x or for y.

Theorem 1.9.16: When Linear Relation Is Function

A linear relation between the sets $X=\mathbf{R}$ and $Y=\mathbf{R}$,

$$Ax+By=C\,,$$

may be represented by a function, called a linear function, as follows:

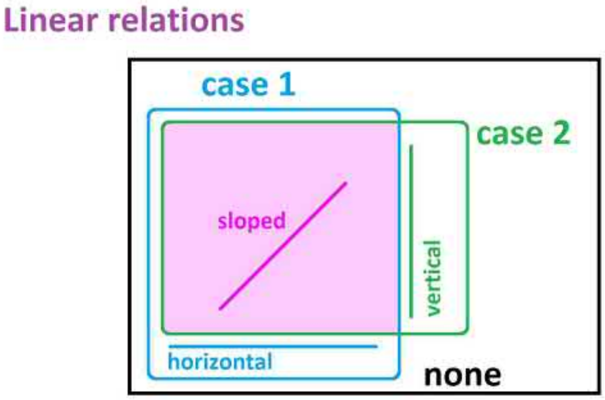
1. *When $B\neq 0$, it is a function $F:X\to Y$ given by*

$$y=F(x)=-\frac{A}{B}x+\frac{C}{B}\,.$$

2. *When $A\neq 0$, it is a function $G:Y\to X$ given by*

$$x=G(y)=-\frac{B}{A}y+\frac{C}{A}\,.$$

Indeed, every function is a relation but not every relation is a function, but when it is, there might be two. The two cases are illustrated below be referencing their graphs:



Exercise 1.9.17

Explain the intersection of the two cases and what's lies outside.

Example 1.9.18: two functions in a relation

In the relation $x+y=50$ above, we have $A=1$ and $B=1$, so we can do both:

$$y=-x+50 \quad \text{and} \quad x=-y+50\,.$$

In the relation $y=3$, we have $A=0$ and $B=1$; this is the former case and the function is constant:

$$F(x)=3\,.$$

In the relation $x=2$, we have $A=1$ and $B=0$; this is the latter case and the function is constant:

$$G(y)=2\,.$$

Exercise 1.9.19

Find all linear functions is these linear relations: (a) $3x-2y=2$, (b) $2x=3$, (c) $-y=7$.

Exercise 1.9.20

Prove the theorem.

Exercise 1.9.21

What lines are *not* included in case 1? case 2?

Exercise 1.9.22

State both cases of the theorem as implications (an "if-then" statement).

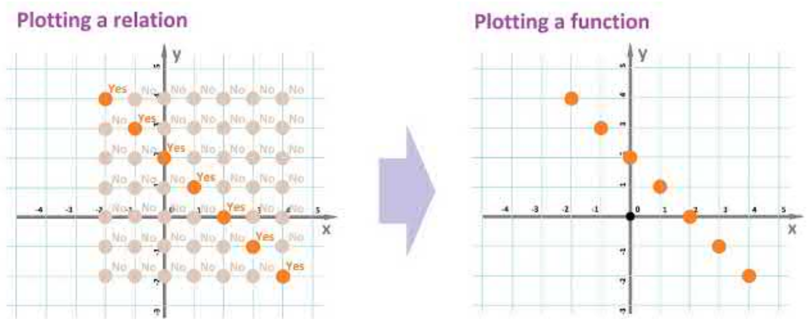
Transitioning to *functions* makes the plotting task much easier. The 49 computations are reduced to just 7:

x	-2	-1	0	1	2	3	4
-2	-4	-3	-2	-1	0	1	2
-1	-3	-2	-1	0	1	2	3
0	-2	-1	0	1	2	3	4
1	-1	0	1	2	3	4	5
2	0	1	2	3	4	5	6
3	1	2	3	4	5	6	7
4	2	3	4	5	6	7	8

$\rightarrow$

x	$y=-x+2$
-2	4
-1	3
0	2
1	1
2	0
3	-1
4	-2

Instead of testing a lot of points trying, and mostly failing, to find the ones that fit the equation, we just plug in as many values of x as necessary, producing a y and, consequently, a point on the plane every time:



The price we paid is algebra, solving for y:

$$x + y = 2 \implies y = x - 2\,.$$

In general, instead of having to run through a whole *plane* of (x, y)'s – for relations, we only need to run through a *line* of x's – for functions. We also observe that since there can be only *one* point of the graph of a function above each x, the graph of a function must be *one-point thick*; it's a curve!

Exercise 1.9.23

State the converse of "the graph of every linear function is a straight line" and find out if it's true.

1.10. Representation of functions

Functions are *explicit relations*. The two variables are still related to each other, but this relation is now unequal: The input comes first and, therefore, the output is *dependent* on the input. That is why we say that the input is the *independent variable* while the output is the *dependent variable*.

A function is a *black box*; something comes in and something comes out as a result, like this:

$$\text{input} \to \boxed{\blacksquare\blacksquare\blacksquare} \to \text{output}$$

The only law is the following:

- ▶ The same input must produce the same output.

For example, a vending machine will provide you with the item the code of which you have entered (if sufficient funds are inserted).

In the case of numerical functions, both are numbers. The black box metaphor suggests that while some computation happens inside the box, what it is exactly may be unknown:

$$\begin{array}{ccccc} \text{input} & & \text{function} & & \text{output} \\ \text{income} & \to & \boxed{\ \text{IRS}\ } & \to & \text{tax bill} \end{array}$$

How things happen might be even unimportant; what's important is the rule a function has to follow: one y for each x. For example, if you don't know how this function is computed, you can ask someone to do it for you:

$$\begin{array}{ccccc} \text{input} & & \text{function} & & \text{output} \\ x & \to & \boxed{\ \cos\ } & \to & y \end{array}$$

If we are able to peek inside, we might see something very complex or something very simple:

$$\text{input} \to \boxed{\text{multiply by } 3} \to \text{output}$$

Function is what function does! It may be simply a *sequence of instructions*.

Example 1.10.1: flowcharts represent functions

For example, for a given input x, we do the following consecutively:

- add 3,
- multiply by 2, and then
- square.

Such a procedure can be conveniently visualized with a “flowchart”:

If the input is $x = 1$, we acquire three more numbers in this order:

$$1 \ \mapsto\ 1+3=4 \ \mapsto\ 4\cdot 2=8 \ \mapsto\ 8^2=64$$

Here is the algebra of what is going on inside of each of the boxes:

$$x \ \to\ \boxed{x+3}\ \to\ y\ \to\ \boxed{y\cdot 2}\ \to\ z\ \to\ \boxed{z^2}\ \to\ u$$

We have introduced *intermediate variables* for reference. Note how the names of the variables match; we, therefore, can proceed to the next step. A sequence of algebraic steps of this process is as follows:

$$\begin{aligned} x \ \to\ x+3 &= y \\ \to\ y\cdot 2 &= z \\ \to\ z^2 &= u \end{aligned}$$

It can also be called an *algorithmic representation*.

Exercise 1.10.2

Describe the function that computes a sales tax of 5%.

Exercise 1.10.3

Describe the function that computes a discount of 10%.

Thus, we represent a function diagrammatically as a box that processes the input and produces the output:

$$\begin{array}{ccccc} \text{input} & & \text{function} & & \text{output} \\ x & \to & \boxed{\ f\ } & \to & y \end{array}$$

Here, f is *the name of the function* (in fact, “f” stands for “function”). In this example, the function is *unspecified*. We make it *specific* by describing how it works.

Functions come from many sources and can be expressed in different forms:

- a list of instructions (an algorithm)
- an algebraic formula
- a list of pairs of inputs and outputs
- a graph
- a transformation

An algorithm is commonly a list of instructions given to a computer, i.e., a *program*. It may be preferable to have a function to be handled by a person represented in the form of a *formula*. The person may appreciate a more compact form that allows to notice patterns, simplify, and further manipulate the function.

An *algebraic representation* is exemplified by $y = 3x - 1$. In order to properly introduce this as a function, we give it a name, say f, and write:

$$f(x) = 3x - 1\,.$$

Let's examine this notation:

Variables of function

$$\begin{array}{r|ccccccc} & y & = & f & (& x &) = & 3x-1 \\ & \uparrow & & \uparrow & & \uparrow & & \uparrow \\ \text{name:} & \text{dependent} & & \text{function} & & \text{independent} & & \text{independent} \\ & \text{variable} & & & & \text{variable} & & \text{variable} \end{array}$$

The letters are all just *names*! The choices for these names are mostly arbitrary. They have to vary when there is more than just one function present, for example:

$$\begin{array}{r|ccccccc} & z & = & g & (& t &) = & t+5 \\ & \uparrow & & \uparrow & & \uparrow & & \uparrow \\ \text{name:} & \text{dependent} & & \text{function} & & \text{independent} & & \text{independent} \\ & \text{variable} & & & & \text{variable} & & \text{variable} \end{array}$$

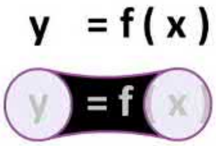
Warning!

It is often acceptable (or even preferable) to omit the name of the function and concentrate on the variables, as we did in the last example.

Thus, the independent variable is the input, and the dependent variable is the output. When the independent variable is made specific, so is the dependent variable, via the *substitution*:

$$\begin{array}{ccccc} f & (& 2 &) = & 3\cdot 2-1 \\ \uparrow & & \uparrow & & \uparrow \\ \text{function} & & \text{input} & & \text{output} \end{array}$$

We can think of this notation as a "black funnel":



Here x enters through the funnel and then (after processing) y appears from the other end. With the same effect, we can use a blank box as an entry gate instead of x:

$$\begin{array}{ccccc} f & (& \square &) = & 3\cdot\square-1 \\ & & \uparrow & & \uparrow \\ & & \text{insert input} & & \text{insert input} \end{array}$$

We can substitute functions too:

$$\begin{array}{ccccc} f & (& -t+5 &) = & 3\cdot(-t+5)-1 \\ \uparrow & & \uparrow & & \uparrow \\ \text{function} & & \text{input} & & \text{output} \end{array}$$

In summary:

- The "x" in a formula serves as a *placeholder* for: numbers, variables, and whole functions.

Exercise 1.10.4

Provide a formula for the new function $f(z^2)$ made from the function f above.

Example 1.10.5: function as sequence of steps

Let's take the function from the beginning of the section; it requires several stages:

$$y = x + 3, \quad z = y \cdot 2, \quad u = z^2 \, .$$

They can be written as follows:

$$\square \to \boxed{\square + 3} \to \boxed{\square \cdot 2} \to \boxed{\square^2} \to \square$$

For example, we compute its output for the input $x = 2$ in three consecutive steps:

$$2 \mapsto 2 + 3 = 5 \mapsto 5 \cdot 2 = 10 \mapsto 10^2 \mapsto 100$$

Example 1.10.6: decomposition of function

Consider this formula:

$$f(x) = \sqrt{x^2 - 3} + 5 \, .$$

To represent this function as a list of instructions, we just read the formula starting with x:

$$\text{input} \to \boxed{\text{square}} \to \boxed{\text{subtract } 3} \to \boxed{\text{take square root}} \to \boxed{\text{add } 5} \to \text{output}$$

We read inside out!

Exercise 1.10.7

Represent this function as a list of instructions:

$$f(x) = \left(\sqrt{x} + 2\right)^3 .$$

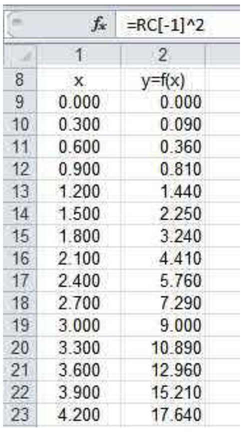
A function can also be represented by a *list of pairs of inputs and outputs.*

In the numerical case, this list is a table with two columns, for x and y:

x	$y = f(x)$
0	1
1	3
2	4
3	0
4	2
...	...

This may be called a *numerical representation* of a function as the list contains only numbers. Any list like this would do as long as there are no repetitions in the x-column!

To create larger lists, one uses a spreadsheet. Each value in the y-column is computed from the corresponding value in the x-column via some algebraic formula:



fx =RC[-1]^2

	1	2
8	x	y=f(x)
9	0.000	0.000
10	0.300	0.090
11	0.600	0.360
12	0.900	0.810
13	1.200	1.440
14	1.500	2.250
15	1.800	3.240
16	2.100	4.410
17	2.400	5.760
18	2.700	7.290
19	3.000	9.000
20	3.300	10.890
21	3.600	12.960
22	3.900	15.210
23	4.200	17.640

For example, for $y = x^2$, we have in the y-column the following spreadsheet formula:

```
=RC[-1]^2
```

It refers to the value located in: same row, previous column.

Furthermore, it is even possible that a function is pure *data* and there is no formula! One can imagine, for example, that the table has come from a measuring device (say, a thermometer) that takes readings at equal intervals of time.

Even though the data in the list represents the same function as the formula above, we can see that there are gaps in the data. We can't tell, for example, what 1.5^2 is or what 100^2 is. Thus, our algebraic representation is complete, but the numerical representation given by the list is not. However, this list is *a* function but with a smaller domain than the original.

The advantage of numerical representations is that they have been pre-computed for you so that you can see *patterns*; for example, with x increasing we see that:

- y is also increasing, and furthermore,
- y grows faster and faster.

If the last observation is hard to see in the data, we either produce more data – such as compute the difference of the sequence – or visualize the data that we do have.

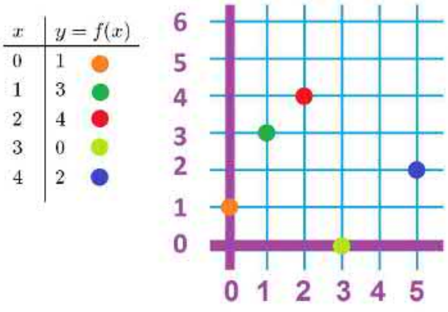
We can use the list data to plot points, which leads us to the *graphical representation* of functions. Below, the definition we have used for relations is repeated, but this one will be even more widely used:

Definition 1.10.8: graph of function

The *graph* of a numerical function $y = f(x)$ is the set of points in the xy-plane that satisfy $y = f(x)$. In other words, it is the following set:

$$\{(x, y) : \ y = f(x)\}\,.$$

For example, we can plot the above data (just the points that have been provided):



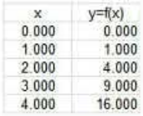
x	$y = f(x)$	
0	1	●
1	3	●
2	4	●
3	0	●
4	2	●

Warning!

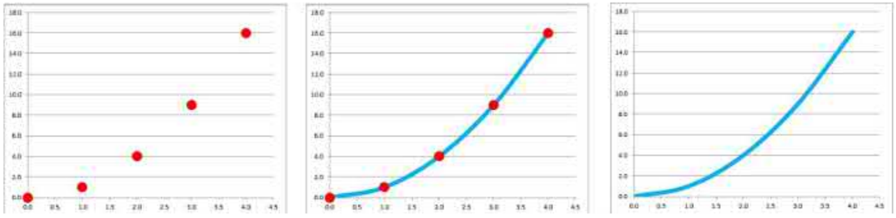
We will speak of "a graph", or "graphs", when we deal with the graph of some *function*.

Example 1.10.9: plotting points

A spreadsheet software comes with graphic capabilities. It will plot all points you have in the list:



x	y=f(x)
0.000	0.000
1.000	1.000
2.000	4.000
3.000	9.000
4.000	16.000

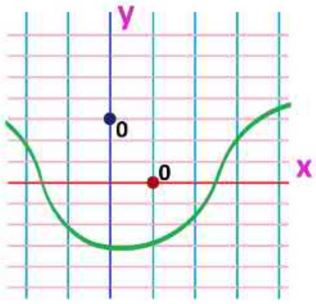


It can also automatically add a curve connecting these points.

Warning!

The first plot is the truth; the rest is a guess.

Note that when x and y represent two variables that have nothing to do with each other – such as time and location – neither do the two axes. In that case, neither the unit lengths nor the locations of the origins have to match:



A *transformation* takes the domain X, a subset of the real line, transforms it according to the function (shift, stretch, flip, etc.), and places the result on the codomain Y. It is discussed in the next chapter.

An *algorithm* is a verbal representation of a function. It may contain no explicit algebra. Instead, it tells us how to get a certain output given any input. For example,

- Question: How do we get from x to y?
- Answer: Let y be equal to the square of x.

This representation, too, conveys a *compete* information about the function.

Example 1.10.10: flowchart from formula

Describe what this function does:

$$f(x)=\frac{x^2+1}{x^2-1}\,,$$

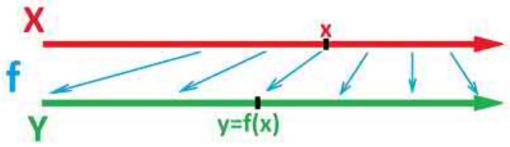
verbally:

- Step 1: Multiply x by itself, call it y.
- Step 2: Add 1 to y, call it z.
- Step 3: Subtract 1 from y, call it u.
- Step 4: Divide z by u.

There is a *fork* in the diagram:

$$\begin{tikzcd}
 & & & y+1=z \arrow[dr] & \\
x \arrow[r] & x^2=y \arrow[r] & {} \arrow[ur] \arrow[dr] & & z/u=w \\
 & & & y-1=u \arrow[ur] &
\end{tikzcd}$$

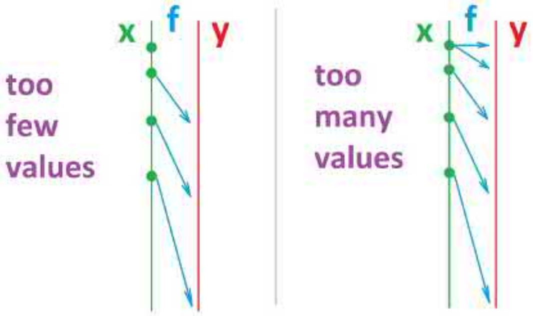
A (numerical) *function* is a rule or procedure f that assigns to any number x in a set X, called the *set of inputs* or the *domain*, one number y in another set of real numbers Y, called the *set of outputs* or the *codomain* of f.



In other words,

1. each x in X has a counterpart in Y, and
2. there is only one such counterpart.

This rule can be violated when there are too few or too many arrows for a given x:



Then this is *not a function*. It is OK, however, to have too few or too many arrows for a given y!

Next, let's revisit the rule – how to get y from x – that defines a function. It must satisfy:

- ▶ *There is only one* y *for each* x.

Let's illustrate how the rule might visibly fail for each of these four ways to represent a function.

Example 1.10.11: algebraic representation

In the following very common way to present a formula, there are two outputs for the same input (unless $x = 0$):

$$y = \pm x\,.$$

Not a function!

Example 1.10.12: numerical representation

In the following list of values, the inputs aren't ordered. It is, therefore, possible that the list might contain two rows with the same x-value and different y-values:

	x	y	
	...	...	
	0	22	
↗	...	...	↖
same!	...	...	different!
↘	...	...	↙
	0	55	
	...	...	

Not a function!

Example 1.10.13: algorithmic representation

In this list of commands, one is either ambiguous or it produces multiple outputs:

- Step 1: ...
- ...
- Step 50: Add today's date to the output of step 49.
- ...
- Step 100: ...

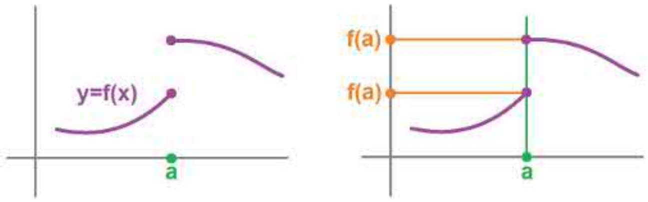
Not a function!

Exercise 1.10.14

Suggest your own examples of how formulas, lists, and algorithms can fail to give us a function.

Example 1.10.15: graphical representation

The following graph has two points – outputs – above $x = a$:



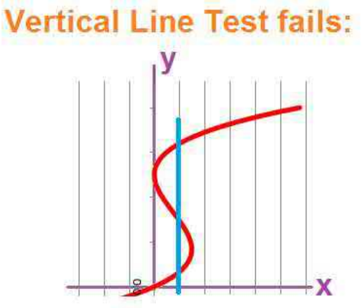
Not a function!

For the graphical representation, all it takes is a glance.

Theorem 1.10.16: Vertical Line Test For Relations

A relation is a function of x if and only if every vertical line crosses the graph at one point or none.

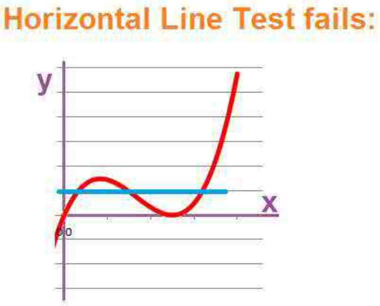
So, every vertical line is a test:



Corollary 1.10.17: Horizontal Line Test For Relations

A relation is a function of y if and only if every horizontal line crosses the graph at one point or none.

So, every horizontal line is a test:



Exercise 1.10.18

Split either theorem into a statement and its converse.

1.11. Linear functions

The dependence of x on y in a numerical function can be very simple.

However, the simplest kind of function is the one whose output does not change with the input! This is a *constant function*, i.e., it is given by a formula:

$$f(x)=k \quad \text{for each } x,$$

for some predetermined number k. Its implied domain is, of course, $X=(-\infty,\infty)$. Its computation is non-existent; for example, when $k=3$, we have the following:

$$\text{input} \quad \to \quad \boxed{\text{produce } 3} \quad \to \quad \text{output}$$

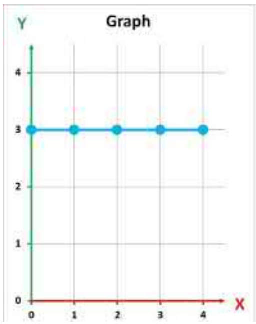
As you can see, the input is thrown away. This is the list of values of this function:

x	$y=f(x)$
0	3
1	3
2	3
3	3
4	3
...	...

Warning!

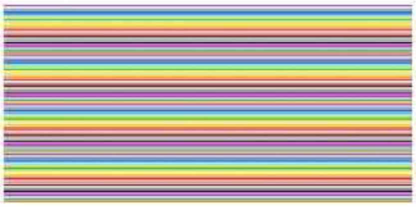
Depending on context, 3 might mean a function.

Plotting a few of these points reveals that the graph is a horizontal line:



Indeed, the corresponding relation is $y = 3$.

This is what the graphs of all constant functions combined look like:



The next simplest function is the one that *does nothing* to the input; i.e., it is given by a formula:

$$f(x) = x\,.$$

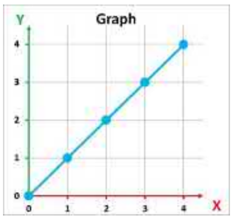
Its implied domain is, of course, $X = (-\infty, \infty)$. Its computation is trivial:

$$\text{input} \quad \to \quad \boxed{\text{pass it}} \quad \to \quad \text{output}$$

This time, the input isn't thrown away but there was still no algebra needed. This is its list of values:

x	$y = f(x)$
0	0
1	1
2	2
3	3
4	4
...	...

Plotting a few of these points reveals that the graph is a 45-degree line:



Indeed, the relation is $y = x$.

Warning!

If we say that y *is* x, then the xy-plane should have the same units for the two axes.

Exercise 1.11.1

Plot the graph of a function that represents the location as it depends on time if the speed is one foot per second.

So far, the function require no algebraic operations! Linear functions are at the next level of complexity. They may be "sloped".

Definition 1.11.2: slope

Suppose we have two points in a specified order, A then B, on the xy-plane, then *the slope of the line* from A to B is defined to be

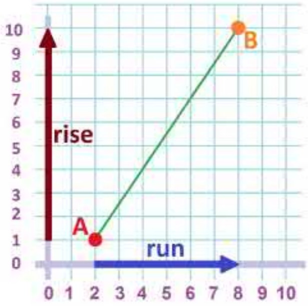
$$\boxed{\text{slope } = m = \frac{\text{rise}}{\text{run}} = \frac{\text{change of } y}{\text{change of } x}}$$

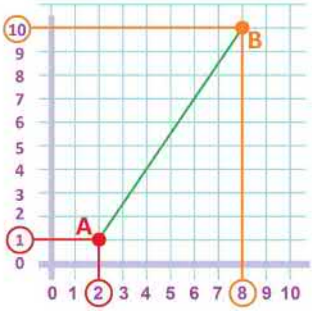
Exercise 1.11.3

Can the rise be zero? Can the run be?

Example 1.11.4: slope

The geometric meaning of the numerator and denominator is seen below:





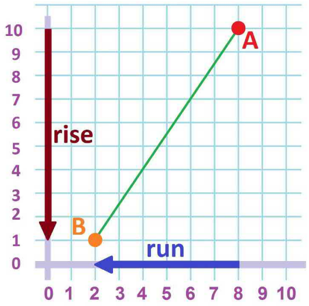
We can just count the number of steps vertically and horizontally (left):

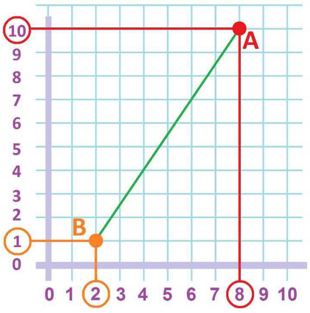
- run $= 6$, and
- rise $= 9$, therefore,
- slope $= \frac{9}{6} = \frac{3}{2} = 1.5\,.$

Or we can utilize the coordinates of the two points and subtract those of A from those of B (right):

- run $= 8 - 2 = 6$, and
- rise $= 10 - 1 = 9$, therefore,
- slope $= \frac{9}{6} = \frac{3}{2} = 1.5\,.$

"Rise" and "run" in this context aren't meant to be substitutes for "lengths of these segments" or "distances between those points". In contrast to plain geometry, one or both of them can be negative! In particular, the slope remains unchanged if we reverse the order of the two points: B first, A second:





Indeed:

- run $= 2 - 8 = -6$, and
- rise $= 1 - 10 = -9$, therefore,
- slope $= \frac{-9}{-6} = \frac{3}{2} = 1.5\,.$

Same slope! It follows that we are studying the *slope of the line* not just that of the two points.

Exercise 1.11.5

Find more pairs of points on the line with slope 1.5.

We utilize the coordinate system to find the slope. Suppose we have two distinct points on a straight line in a specified order, say,

$$A = (x_0, y_0) \text{ and } B = (x_1, y_1),$$

then the *slope* of the line they determine is given by the formula:

$$\boxed{m = \frac{\text{change from } y_0 \text{ to } y_1}{\text{change from } x_0 \text{ to } x_1} = \frac{y_1 - y_0}{x_1 - x_0}}$$

It is crucial to know the following:

Theorem 1.11.6: Slope Backwards

The slope from A to B is equal to the slope from B to A.

Proof.

If we reverse the order of the two points – B then A – both numerator and denominator simply flip their signs:

$$\text{change from } y_1 \text{ to } y_0 = -(\text{change from } y_0 \text{ to } y_1),$$

and

$$\text{change from } x_1 \text{ to } x_0 = -(\text{change from } x_0 \text{ to } x_1).$$

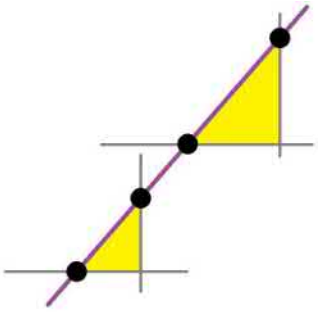
But if the numerator and denominator of a fraction flip their signs, the fraction remains intact: $(-a)/(-b) = a/b$. We have for the slope:

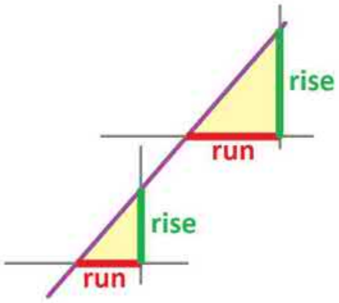
$$m = \frac{y_0 - y_1}{x_0 - x_1} = \frac{-(y_1 - y_0)}{-(x_1 - x_0)} = \frac{y_1 - y_0}{x_1 - x_0} = m.$$

Warning!

Whether it's A then B, or B then A, it must be the same for both numerator and denominator.

For the definition to make sense we need to show that any two pairs of points will produce the same slope:





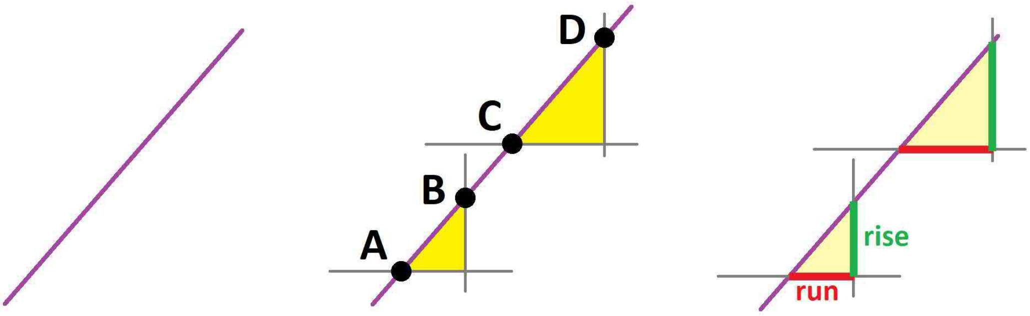
We have a stronger result below:

Theorem 1.11.7: Slope From Two Points

Any two points chosen on a straight line produce the same slope.

Proof.

Suppose we put a pair of points A and B on the line, as well as another pair C and D. We will show that these two triangles produce the same slope:



We use what we know from Euclidean geometry. The two horizontal lines (as all parallel lines) cut the same angle from our sloped line. So, the base angles are equal. Next, the two vertical lines (as all parallel lines) cut the same angle from our sloped line. So, these angles are also equal. And so are the two right angles! Therefore, these are similar triangles. This means that the lengths of their sides are proportional. In particular, the rise and run are proportional! If the second triangle is k times bigger than the first, we have:

$$\text{The slope of the second triangle } = \frac{k\cdot \text{rise}}{k\cdot \text{run}} = \frac{\text{rise}}{\text{run}} = \text{the slope of the first triangle.}$$

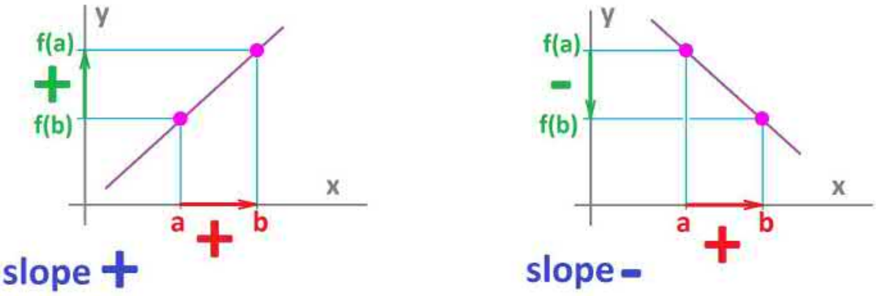
That's is what makes a straight line a straight line!

Exercise 1.11.8

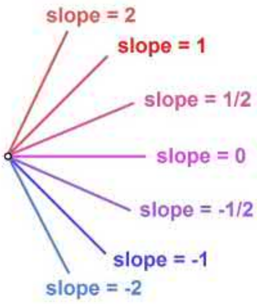
Prove the theorem using trigonometry.

What does the slope tell us about the line?

While a positive slope appears when the rise and the run have the same signs, a negative slope appears when the signs are opposite:



Below we arrange all lines according to their slopes (as if they all start at the origin):

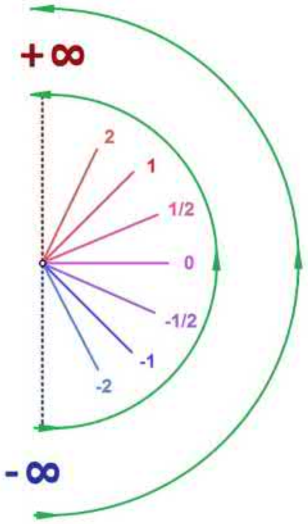


Warning!

Comparing the line with slope $m = 1$ and the line with $m = -2$ suggests that the word "steepness" as a substitute for slope should be used with caution:

$$\text{steepness } = |m|\,.$$

It's as if increasing the slope *rotates* the line counterclockwise:



We can see that the scope of possible values of slopes is $(-\infty, +\infty)$.

Warning!

It is impossible to assign a slope to a vertical line even if we are willing to use infinity: Is it $-\infty$ or $+\infty$?

Exercise 1.11.9

What happens to the slope of a line drawn on a piece of paper for different choices of the axes?

We exclude the possibility of a vertical line and an infinite slope! This is why we can concentrate on linear *functions* only.

Definition 1.11.10: linear function and polynomial

A *linear function* is a numerical function given by this formula:

$$\boxed{f(x) = m \cdot x + b}$$

for some predetermined numbers m and b. When $m \neq 0$, such a function is called a *linear polynomial*. When $b = 0$, such a function is called a *linear operator*.

Warning!

Linear polynomials exclude constant functions.

So, the simplest algebra has appeared: addition/subtraction and multiplication by a constant number. They are seen in the function's flow-chart:

$$f:\ x \to \boxed{\text{multiply by } m} \to \boxed{\text{add } b} \to y$$

The formula is commonly called the *slope-intercept form* of the linear function:

Slope-intercept form

$$f(x) = \underset{\text{slope}}{\underset{\uparrow}{m}} \cdot x + \underset{y\text{-intercept}}{\underset{\uparrow}{b}}$$

The latter is indeed the *y-intercept* of the function as defined in the last section:

$$f(0) = m \cdot 0 + b = b\,.$$

The concept of slope is central in calculus. For example, in similarity with sequences, we notice the following:

- If $m > 0$, then the outputs $y = f(x)$ are increasing as the inputs x are increasing.
- If $m < 0$, then the outputs $y = f(x)$ are decreasing as the inputs x are increasing.
- If $m = 0$, then the outputs $y = f(x)$ remain the same as the inputs x are increasing; i.e., f is a constant function.

Warning!

Even though straight lines remain straight lines if we resize the plot, the "slopes" will appear to change.

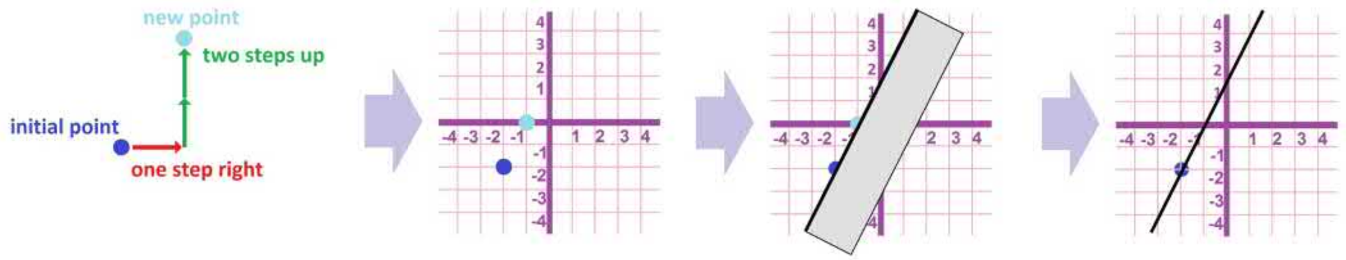
Exercise 1.11.11

Arrange all linear polynomials with the same slope according to their y-intercepts.

The slope gives us the *direction* of the line. That's how the slope-intercept formula, $y = mx + b$, works: We start at the y-intercept, $(0, b)$, and then proceed in the direction provided by the slope, m. In the same manner, we can start at *any* point. Suppose a point is given, say, $A = (x_0, y_0)$. From there, we go as described above: 1 unit right (the run) and m units up (the rise).

Example 1.11.12: plotting a line with a ruler

Let's plot the straight line with slope $m = 2$ through the point $A = (-2, -2)$. From A, we make one step right and two steps up. We have a new point, say B, with coordinates $B = (-1, 0)$ (left):



With a ruler, we draw a line through A and B (right).

Exercise 1.11.13

Plot the straight line with slope $m = -2$ through the point $A = (-1, -1)$. Make up your own parameters and plot the line. Repeat.

Exercise 1.11.14

What is the equation of the line through the points $A = (-1, 2)$ and $B = (2, 1)$?

Example 1.11.15: plotting a line without a ruler

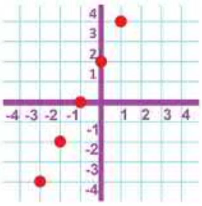
Suppose, again, a point is given, $A=(x_0,y_0)$, and the slope is known to be m. From A, we go 1 unit right and m units up, repeated as many times as necessary:

$$(x_0,y_0)\ \longrightarrow\ (x_0+1,y_0+m)\ \longrightarrow\ (x_0+2,y_0+2m)\ \longrightarrow\ (x_0+3,y_0+3m)\ \longrightarrow\ ...$$

Then we go 1 unit left and m units down, repeated as many times as necessary:

$$(x_0,y_0)\ \longrightarrow\ (x_0-1,y_0-m)\ \longrightarrow\ (x_0-2,y_0-2m)\ \longrightarrow\ (x_0-3,y_0-3m)\ \longrightarrow\ ...$$

We have a sequence of points forming a line:



How do we fill the gaps? We make half-steps: We go 1/2 unit left and $m/2$ units down.

Exercise 1.11.16

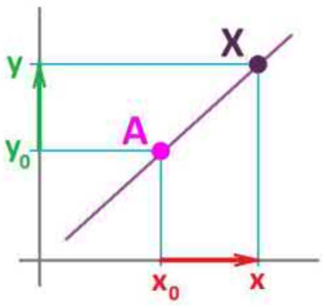
Suggest a way to plot more points.

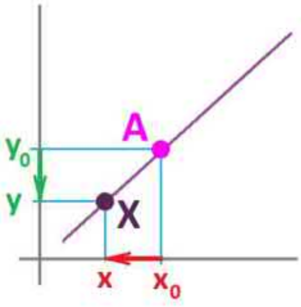
Exercise 1.11.17

Plot as many points as possible for the line from $(1,3)$ and slope -1.

Now, the *algebra*.

Suppose we have a specified point $A=(x_0,y_0)$ on a line with slope m. Let's consider an arbitrary point $X=(x,y)$ on the line:





How do we represent this point algebraically in terms of A and m?

The run is $x-x_0$ and the rise is $y-y_0$ (left or right). Therefore, the slope is

$$m=\frac{y-y_0}{x-x_0}\,.$$

So, the coordinates of X satisfy in this formula.

However, here x cannot be equal to x_0! This is an inconvenience, because the point A itself doesn't satisfy the equation. To avoid this limitation, we rewrite this formula: We multiply both sides by $x-x_0$. The result is a new and very important way to represent a line:

Theorem 1.11.18: Point-Slope Form of Line

A line with slope m passing through point (x_0,y_0) is given by the following linear *relation:*

$$\boxed{y - y_0 = m \cdot (x - x_0)}$$

Once can even consider this to be new definition of slope.

Exercise 1.11.19

What is the difference between the two relations:

$$y - y_0 = m \cdot (x - x_0) \ \text{ and } \ m = \frac{y - y_0}{x - x_0} \ ?$$

The relation can be converted to a function of y, or to a function of x provided $m \neq 0$;

$$\begin{tikzcd} & \boxed{y - y_0 = m \cdot (x - x_0)} \arrow[dl] \arrow[dr] & \\ \boxed{y = y_0 + m \cdot (x - x_0)} & & \boxed{x = x_0 + \frac{1}{m} \cdot (y - y_0)} \end{tikzcd}$$

Even though we can solve for y any time we want (and make it a function!), this form is often preferable because of the information it reveals. First, the rise and the run are clearly visible:

Point-slope form of line

$$\begin{array}{ccccc} \text{rise} & = & \text{slope} & \cdot & \text{run} \\ (y - y_0) & = & m & \cdot & (x - x_0) \end{array}$$

Second, the coordinates of the fixed point $A = (x_0, y_0)$ and a variable point $X = (x, y)$ on the graph are visible too:

Point-slope form of line

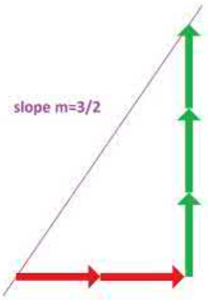
$$\overset{\text{point } X}{\overset{\downarrow}{y}} \quad - \quad \underset{\text{point } A}{\underset{\uparrow}{y_0}} \quad = \quad m \quad \cdot \quad (\overset{\text{point } X}{\overset{\downarrow}{x}} \quad - \quad \underset{\text{point } A}{\underset{\uparrow}{x_0}})$$

Exercise 1.11.20

Find the y-intercept from the point-slope form.

Example 1.11.21: plotting incremental motion

What is the slope of the line that follows this path? We make p steps right and q step up as we follow the line:



Then the equation becomes:

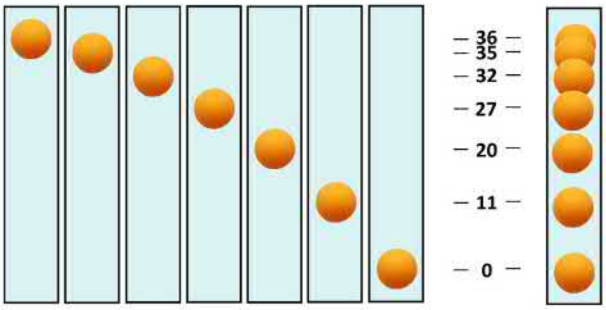
$$p(y - y_0) = q(x - x_0)\,.$$

The slope is $m = q/p$.

1.12. Sequences

Example 1.12.1: falling ball

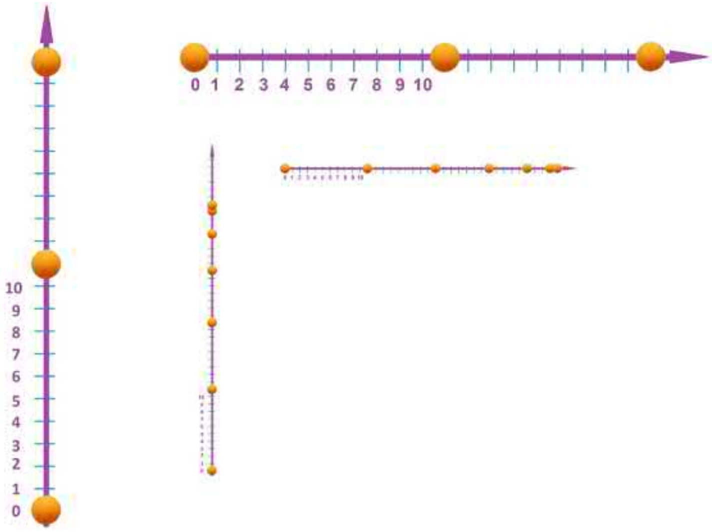
We videotape a ping-pong ball falling down and record – at equal intervals – how high it is. The result is an ever-expanding string, a sequence, of numbers. If the frames of the video are combined into one image, it will look something like this:



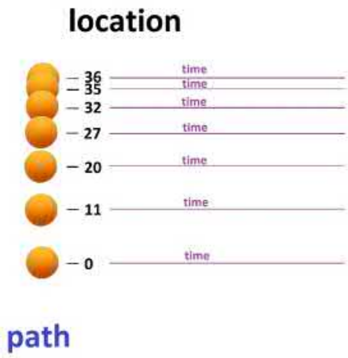
We ignore the time for now and concentrate on the locations only. We have the first few in a *list*:

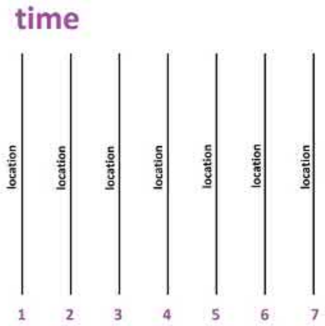
$$36,\ 35,\ 32,\ 27,\ 20,\ 11,\ 0\,.$$

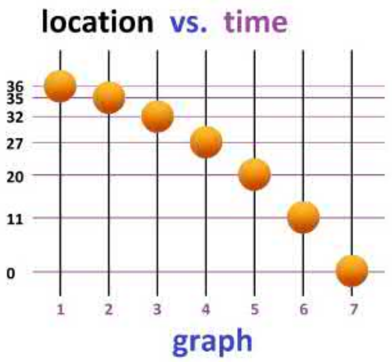
This data can be visualized by placing the ball at every coordinate location on the real line, oriented vertically or horizontally:



Though not uncommon, this method of visualization of motion, or of sequences in general, has its drawbacks: Overlapping may be inevitable and the *order* of events is lost (unless we add labels). A more popular approach is the following. The idea is to *separate time and space*, to give a separate real line, an axis, to each moment of time, and then bring them back together in one rectangular plot:







The location varies – as it does – vertically while the time progresses horizontally. The result is similar to the collection of the frames of the video as seen above. The plot is called the *graph* of the sequence. As far as the data is concerned, we have a list of *pairs*, time and location, arranged in a table:

moment	height
1	36
2	35
3	32
4	27
5	20
6	11
7	0

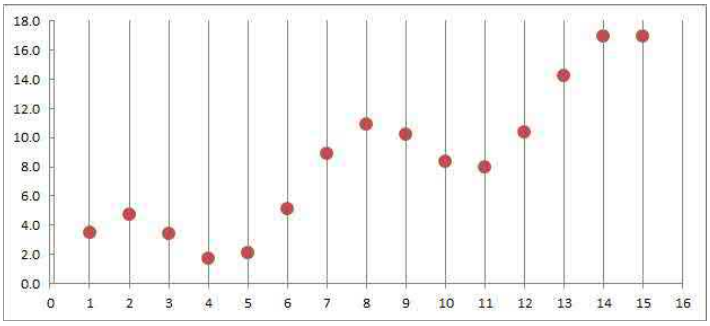
The table is just as effective representation of the data if we flip it; it's more compact:

moment:	1	2	3	4	5	6	7
height:	36	35	32	27	20	11	0

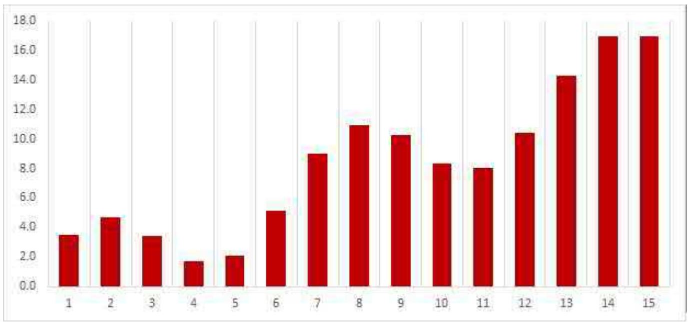
Warning!

It is entirely a matter of convenience to represent our data as a two-column table (especially in a spreadsheet) or a two-row table. In either case, it's a list of pairs of numbers.

So, the most common way to visualize a sequence of *numbers* is as a sequence of *points* on a sequence of vertical axes:



It is also common to represent the same numbers as vertical *bars*:



Warning!

The graph is just a visualization of the data.

To represent a sequence algebraically, we first give it a name, say, a, and then assign a specific variation of this name to each term of the sequence:

Indices of sequence

index:	n	1	2	3	4	5	6	7	...
term:	a_n	a_1	a_2	a_3	a_4	a_5	a_6	a_7	...

The *name* of a sequence is a letter, while the subscript called the *index* indicates the place of the term within the sequence. It reads: "a sub 1", "a sub 2", etc.

The letter "n" is often the preferred choice for the index because it might stand for "natural numbers": $1, 2, 3, 4, \ldots$. As before, "..." indicates a continuing pattern: The indices continue to grow incrementally.

Example 1.12.2: falling ball

For the last example, let's name the sequence h for "height". Then the above *table* take this form:

moment:	1	2	3	4	5	6	7	...
height:	h_1	h_2	h_3	h_4	h_5	h_6	h_7	...
	‖	‖	‖	‖	‖	‖	‖	...
height:	36	35	32	27	20	11	0	...

This is the same table aligned vertically:

moment	height	height
1	$h_1 =$	36
2	$h_2 =$	35
3	$h_3 =$	32
4	$h_4 =$	27
5	$h_5 =$	20
6	$h_6 =$	11
7	$h_7 =$	0
..	..	..

Either table is a list of identities that can be written in any order:

$$h_3 = 32 \quad h_5 = 20 \quad h_7 = 0 \qquad h_4 = 27$$

$$h_1 = 36 \quad h_2 = 35 \qquad h_4 = 27 \quad h_6 = 11$$

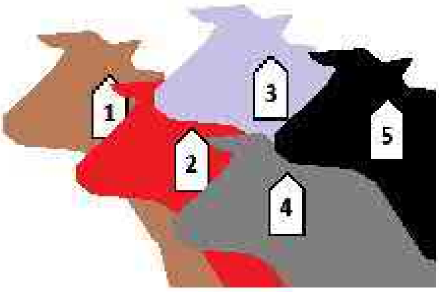
Let's deconstruct the notation:

Index of a term

$$\underset{\underset{\text{name}}{\uparrow}}{a}\ \overset{\overset{\text{index}}{\downarrow}}{n}$$

In other words, we specify a sequence first and then specify the location of the term within the sequence.

Indices serve as *tags*:



A sequence can come from a list or a table unless it's infinite. Infinite sequences often come from *formulas*.

Example 1.12.3: sequence of reciprocals

The formula:

$$a_n = 1/n\,,$$

gives rise to the sequence,

$$a_1 = 1,\ a_2 = 1/2,\ a_3 = 1/3,\ a_4 = 1/4,\ ...$$

Indeed, replacing n in the formula with 1, then 2, 3, etc. produces the numbers on the list one by one, as follows. We enter n into the formula, and a_n appears at the end of the computation. In other words, we place the current value of n inside a blank box (where n used to be) in the formula:

$$\underset{\underset{\text{insert } n}{\uparrow}}{a_{\square}} \quad = \quad \underset{\underset{\text{insert } n}{\uparrow}}{\frac{1}{\square}}\,.$$

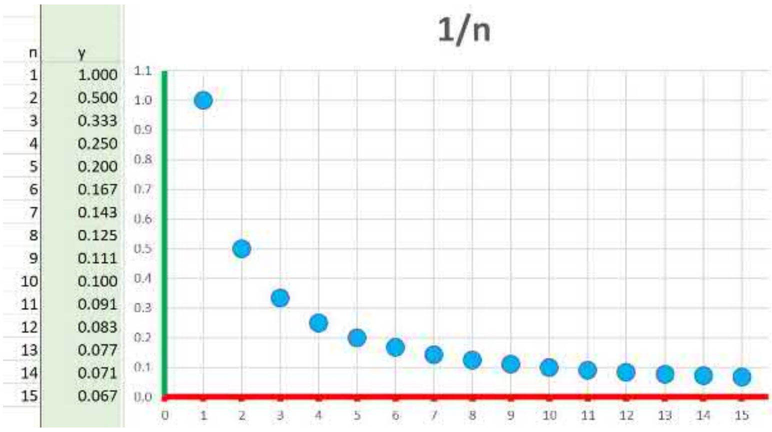
It is just substitution. We do this seven times below:

n	1	2	3	4	5	6	7	...
a_n	a_1	a_2	a_3	a_4	a_5	a_6	a_7	...
$\parallel$	$\parallel$	$\parallel$	$\parallel$	$\parallel$	$\parallel$	$\parallel$	$\parallel$	...
$\frac{1}{n}$	$\frac{1}{1}$	$\frac{1}{2}$	$\frac{1}{3}$	$\frac{1}{4}$	$\frac{1}{5}$	$\frac{1}{6}$	$\frac{1}{7}$	...

With a formula, we can use a spreadsheet (a vertical table) to produce more values with the formula:

```
=1/RC[-1]
```

We also plot these values:



n	y
1	1.000
2	0.500
3	0.333
4	0.250
5	0.200
6	0.167
7	0.143
8	0.125
9	0.111
10	0.100
11	0.091
12	0.083
13	0.077
14	0.071
15	0.067

The complete, algebraic, representation is as follows:

$$a_n = 1/n,\ \ n = 1, 2, 3, ...$$

Exercise 1.12.4

Write a few terms of the sequences given by the formulas:

1. $a_n = 3n - 1$
2. $b_n = 1 + \dfrac{1}{n}$

We will say that this is the *nth-term formula of the sequence.*

Below is the simplest kind.

Definition 1.12.5: constant sequence

A *constant sequence* has all its terms equal to each other.

In other words, we have

$$a_1 = a_2 = ... = c$$

for some number c.

Thus, *every* formula is capable of creating an infinite sequence a_n. For example, we can take these:

- $a_n = n$
- $b_n = n^2$
- $c_n = n^3$
- etc.

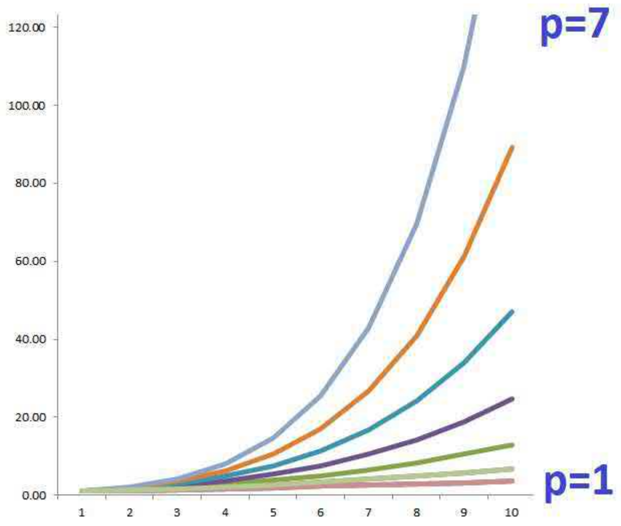
They make up a whole class of sequences.

Definition 1.12.6: power sequence

For every positive integer p, a *power sequence*, or a p-sequence, is given by the formula:

$$\boxed{a_n = n^p, \quad n = 1, 2, 3, ...}$$

The relation between the sequences becomes clear if we zoom out from their graphs ($p = 1, 2, ..., 7$):

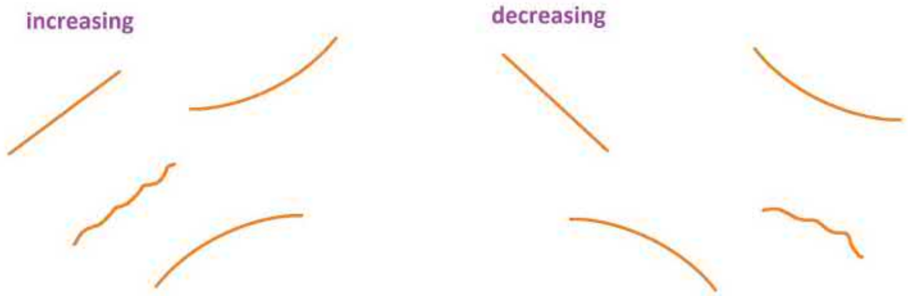


Indeed, the larger the power p, the faster the sequence grows.

The relation between the consecutive terms of each sequence is also clear: It grows! We use these words:

- "growth" or "increase" when we see the graph that *rises* left to right, and
- "decline" or "decrease" when we see the graph that *drops* left to right,

as follows:



As you can see, the behavior varies even within these two categories.

The precise definition has to rely on considering *every pair of consecutive terms* of the sequence. For example, the sequence of the falling ball,

$$36,\ 35,\ 32,\ 27,\ 20,\ 11,\ 0\,,$$

is decreasing because

$$36 > 35 > 32 > 27 > 20 > 11 > 0\,.$$

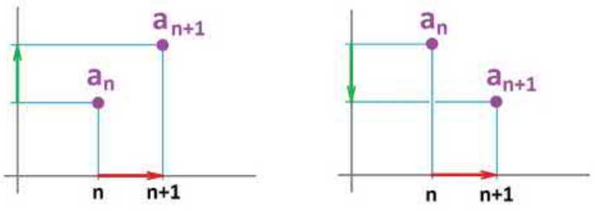
For the general case, we write:

$$a_1 > a_2 > a_3 > a_4 > a_5 > a_6\,.$$

In other words, the current term, a_n, is larger than the next, a_{n+1}:

$$\boxed{\text{last } \geq \text{ next}}$$

The meaning of this inequality is clear when we zoom in on the graph (right):



On left, the sequence is increasing:

$$\boxed{\text{next } \geq \text{ last}}$$

The following will be used throughout.

Definition 1.12.7: increasing sequence and decreasing sequence

- A sequence a_n is called *increasing* if, for all n, we have

$$a_n \leq a_{n+1}\,.$$

- A sequence is called *decreasing* if, for all n, we have

$$a_n \geq a_{n+1}\,.$$

Collectively, they are called *monotone*.

Warning!

Both increasing and decreasing sequences may have segments with no change; furthermore, a constant sequence is both increasing and decreasing.

Example 1.12.8: proving monotonicity

When the sequence is given by its formula, we use it directly. The sequence $a_n = n^2$ is proven to be increasing as follows. We need to show that for all n we have:

$$n^2 < (n+1)^2\,.$$

We simply expand the right-hand side:

$$(n+1)^2 = n^2 + 2n + 1\,.$$

As n is positive, the last part, $2n+1$, is positive too. Therefore, the expression is larger than n^2.

Similarly, we show that $\frac{1}{n}$ is decreasing via the following algebra:

$$n < n+1 \implies \frac{1}{n} > \frac{1}{n+1}\,.$$

The sequence

$$1, -1, 1, -1, 1, -1, 1, -1, 1, -1, \ldots$$

is neither increasing nor decreasing, i.e., it's not monotone.

Exercise 1.12.9

Show that $\frac{1}{n^2}$ is decreasing.

Exercise 1.12.10

Show that all power sequences are increasing.

A major reason why we study sequences is that, in addition to tables and formulas, a sequence can be defined by computing its terms in a *consecutive manner*, one at a time.

Example 1.12.11: regular deposits

A person starts to deposit \$20 every month in his bank account that already contains \$1000. Then, after the first month the account contains:

$$\$1000 + \$20 = \$1020\,,$$

after the second:

$$\$1020 + \$20 = \$1040\,,$$

and so on. In other words, we have:

$$\text{next} \;=\; \text{last} \;+20\,.$$

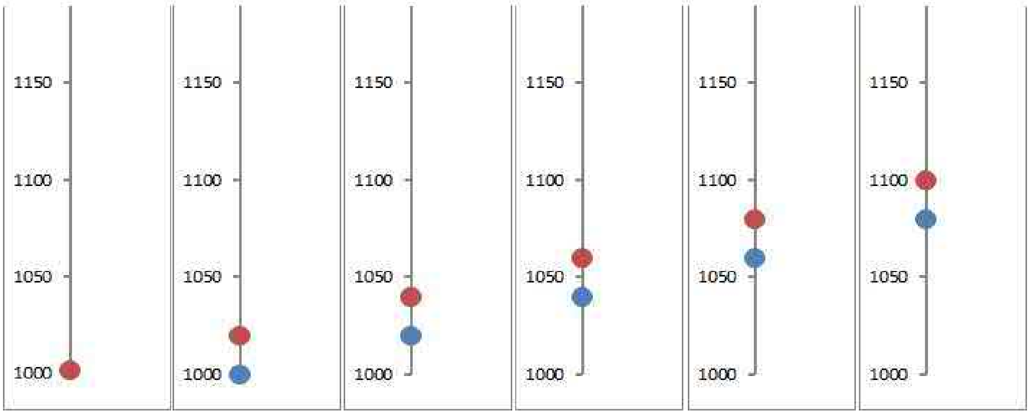
Let's make this algebraic. Suppose a_n is the amount in the bank account after n months, then we have a formula for this sequence:

$$a_{n+1} = a_n + 20\,.$$

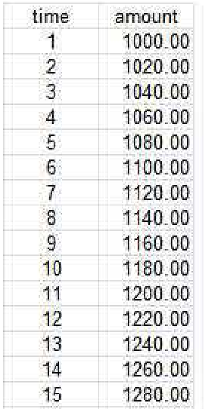
How much will he have after 50 years? We'd have to carry out $50 \cdot 12 = 600$ additions. For the spreadsheet, the formula refers to the last row and adds 20, as follows:

```
=R[-1]C+20
```

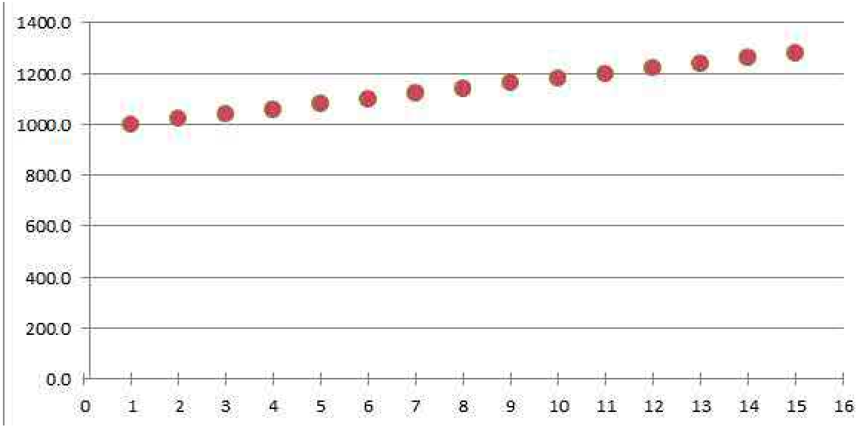
Below, the current amount is shown in blue and the next – computed from the current – is shown in red:



Plotting several terms of the sequence at once confirms that the sequence is *increasing*:



time	amount
1	1000.00
2	1020.00
3	1040.00
4	1060.00
5	1080.00
6	1100.00
7	1120.00
8	1140.00
9	1160.00
10	1180.00
11	1200.00
12	1220.00
13	1240.00
14	1260.00
15	1280.00



It also looks like a straight line.

Definition 1.12.12: recursive sequence

We say that a sequence is *recursive* when its next term is found from the current term by a specified formula, i.e., a_n determines a_{n+1}.

This is the difference between computing a sequence directly, such as $a_n = n^2$, and recursively, such as

$a_{n+1} = a_n + 20$:

$$\begin{array}{|lcl|} n & & a_n \\ \hline 1 & \to & a_1 \\ & & \\ 2 & \to & a_2 \\ & & \\ 3 & \to & a_3 \\ & & \\ .. & & .. \end{array} \qquad \begin{array}{|lcl|} n & & a_n \\ \hline 1 & & a_1 \\ & & \downarrow \\ 2 & & a_2 \\ & & \downarrow \\ 3 & & a_3 \\ & & \downarrow \\ .. & & .. \end{array}$$

The following will be routinely used.

Definition 1.12.13: arithmetic progression

A sequence defined (recursively) by the formula:

$$\boxed{a_{n+1} = a_n + b}$$

is called an *arithmetic progression* with b as its *increment.*

Exercise 1.12.14

If the increment is zero, the sequence is...

Example 1.12.15: compounded interest

An arithmetic progression describes a repetitive process. Also repetitive is the following typical situation. A person deposits $1000 in his bank account that pays 1% APR compounded annually. Then, after the first year, the interest is

$$\$1000 \cdot .01 = \$10\,,$$

and the total amount becomes $1010. After the second year, the interest is

$$\$1010 \cdot .01 = \$10.10\,,$$

and so on. In other words, the total amount is multiplied by .01 at the end of each year and then added to the total. An even simpler way to algebraically describe this is to say that the total amount is multiplied by 1.01 at the end of each year, as follows. After the first year, the total is equal to

$$\$1000 \cdot 1.01 = \$1010\,.$$

After the second year, the total is equal to

$$\$1010 \cdot 1.01 = \$1020.1\,,$$

and so on. In other words, we have:

$$\text{next} \;=\; \text{last} \;\cdot 1.01\,.$$

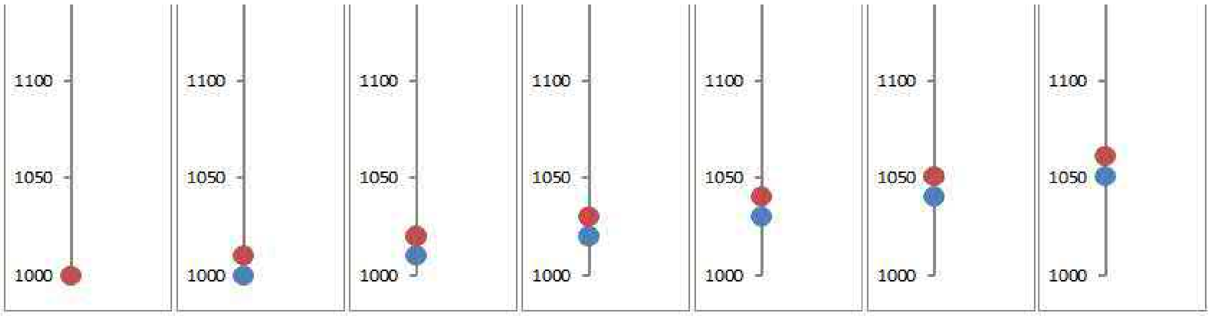
Let's make this algebraic. Suppose a_n is the amount in the bank account after n years. Then we have the following *recursive formula*:

$$a_{n+1} = a_n \cdot 1.01\,.$$

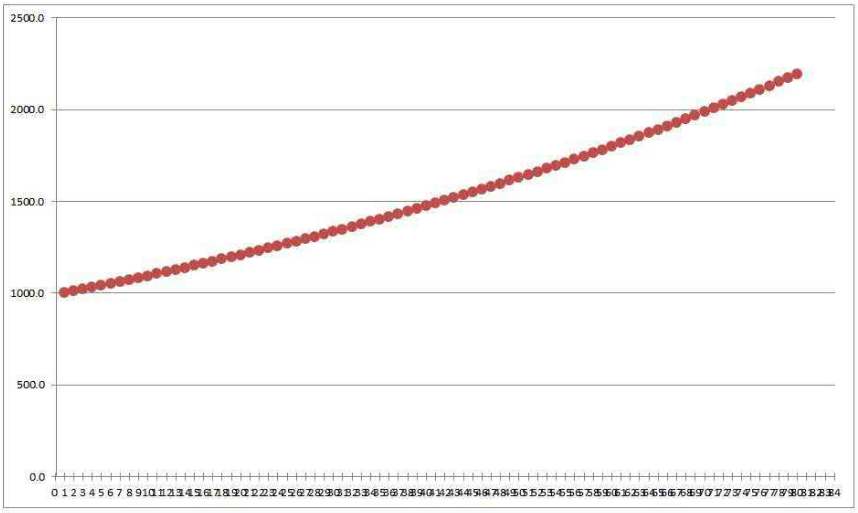
How much will he have after 50 years? We'd have to carry out 50 multiplications. For the spreadsheet, the formula refers to the last row (`R[-1]`) and multiplies by 1.01, as follows:

```
=R[-1]C*1.01
```

We plot a term and the next one:



Only after repeating the step 100 times can one see that this isn't just a straight line:



The sequence is increasing.

Exercise 1.12.16

What if, in addition to an interest of rate r, the depositor also faces inflation of rate s?

The following will be routinely used.

Definition 1.12.17: geometric progression

A sequence defined (recursively) by the formula:

$$\boxed{a_{n+1} = a_n \cdot r\,,}$$

with $r \neq 0$, is called a *geometric progression* with r as its *ratio*. We say that this is:

- a *geometric growth* when $r > 1$, and
- a *geometric decay* when $r < 1$.

Alternatively, it is called an *exponential* growth and decay, respectively.

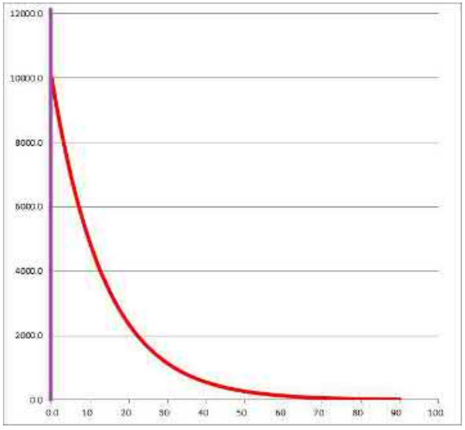
Example 1.12.18: population loss

If the population of a city declines by 3% every year, it is left with 97% of its population at the end of each year. The result is found by multiplying by .97, every time. We have, therefore:

$$\overbrace{(\underbrace{\underbrace{(1,000,000 \cdot 0.97)}_{\text{after 1 year}} \cdot 0.97)}_{\text{after 2 years}} \cdot 0.97}^{\text{after 3 years}}\,.$$

And so on. What will be the population after 50 years? We'd have to carry out 50 multiplications.

The long-term trend is clear from the graph:



This is a geometric progression with ratio $r = .97$, i.e., a geometric decay. The sequence is decreasing and eventually there is almost nobody left.

Exercise 1.12.19

Consider population dynamics with birth rate r and death rate s.

Example 1.12.20: deposits and interest, together

What if we deposit money to our bank account *and* receive interest? The recursive formula is simple, for example:

$$a_{n+1} = a_n \cdot 1.05 + 2000\,.$$

Here, the interest is 5% with a \$2000 annual deposit.

Exercise 1.12.21

What does a 2% inflation do to a dollar hidden in the mattress?

Any algebraic operation, or several operations together, can produce a recursive sequence:

$$\begin{array}{ccccccccccccc}
a_0 & \to & \boxed{\text{add } 2} & \to & a_1 & \to & \boxed{\text{add } 2} & \to & \ldots & \to & \boxed{\text{add } 2} & \to & a_n & \to & \ldots \\
a_0 & \to & \boxed{\text{divide by } 3} & \to & a_1 & \to & \boxed{\text{divide by } 3} & \to & \ldots & \to & \boxed{\text{divide by } 3} & \to & a_n & \to & \ldots \\
a_0 & \to & \boxed{\text{square it}} & \to & a_1 & \to & \boxed{\text{square it}} & \to & \ldots & \to & \boxed{\text{square it}} & \to & a_n & \to & \ldots
\end{array}$$

Exercise 1.12.22

How do these recursive sequences depend on the value of a_0?

When a sequence is defined recursively, we'd need to carry out this definition n times in order to find the nth term. This is in contrast to sequences defined directly via its *nth-term formula*, such as $a_n = n^2$, that requires a single computation to find any term.

An important recursive sequence is constructed from any sequence: the sum represents the totality of the "beginning" of the sequence, found by adding each of its terms to the next, up to that point.

Example 1.12.23: sequences given by lists

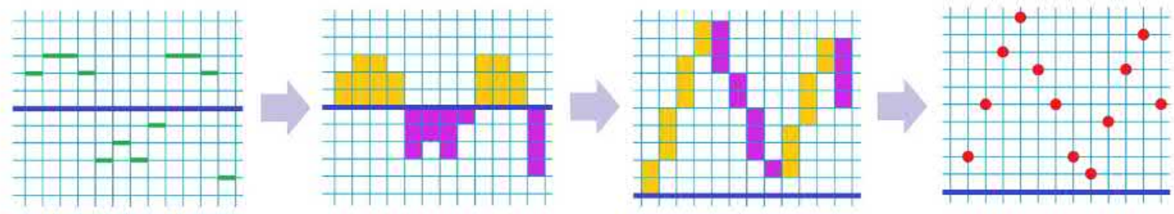
We just add the current term to what we have accumulated so far:

$$\begin{array}{rcccccccccccccccccc}
\text{sequence:} & 2 & & 4 & & & & 7 & & & & 1 & & & & -1 & & & ... \\
 & \downarrow & & \downarrow & & & & \downarrow & & & & \downarrow & & & & \downarrow & & & ... \\
\text{sums:} & 2 & & & & & & & & & & & & & & & & & \\
 & 2 & + & 4 & = & 6 & & & & & & & & & & & & & \\
 & & & & & 6 & + & 7 & = & 13 & & & & & & & & & \\
 & & & & & & & & & 13 & + & 1 & = & 14 & & & & & \\
 & & & & & & & & & & & & & 14 & + & (-1) & = & 13 & \\
 & \downarrow & & & & \downarrow & & & & \downarrow & & & & \downarrow & & & & \downarrow & ... \\
\text{new sequence:} & 2 & & & & 6 & & & & 13 & & & & 14 & & & & 13 & ...
\end{array}$$

We have a new list!

Example 1.12.24: sequences given by graphs

We treat the graph of a sequence as if made of bars and then just stack up these bars on top of each other one by one:



These stacked bars – or rather the process of stacking – make a new sequence.

Definition 1.12.25: sequence of sums

For a sequence a_n, its *sequence of sums*, or simply the sum, is a new sequence s_n defined and denoted for each $n \geq m$ within the domain of a_n by the following (recursive) formula:

$$\boxed{s_m = 0, \quad s_{n+1} = s_n + a_{n+1}}$$

In other words, we have:

$$\boxed{s_n = a_m + a_{m+1} + ... + a_n}$$

Example 1.12.26: alternating sequence

Let's do some algebra. Here a_n is the original sequence and s_n is the new one:

n	a_n	s_n	$=$	s_n
1	1	1	$=$	1
2	-1	$1-1$	$=$	0
3	1	$1-1+1$	$=$	1
$\vdots$	$\vdots$	$\vdots$		$\vdots$
n	$(-1)^n$	$1-1+1-...+(-1)^n$	$=$	1 or 0

The resulting sequence is also "alternating"!

A commonly used notation is the following:

Sigma notation for summation

$$s_n = a_m + a_{m+1} + ... + a_n = \sum_{k=m}^{n} a_k$$

Let's take a closer look at the new notation. The first choice of how to represent the sum of a segment – from m to n – of a sequence a_n is this:

$$\underbrace{a_m}_{\text{step } 1} \quad \underbrace{+a_{m+1}}_{\text{step } 2} \quad +... \quad \underbrace{+a_k}_{\text{step } k} \quad +... \quad \underbrace{+a_n}_{\text{step } n-m} \quad .$$

This notation reflects the recursive nature of the process but it can also be repetitive and cumbersome. The second choice is more compact:

$$\sum_{k=m}^{n} a_k \,.$$

Here the Greek letter Σ stands for the letter S meaning "sum".

Sigma notation

beginning and end values for k

$$\sum_{k=0}^{3} \left(k^2 + k\right) = 20 \quad \longrightarrow \quad \overset{\text{beginning and end values for } k}{\overset{\downarrow}{\begin{matrix} 3 \\ \\ k=0 \end{matrix}}} \;\; \sum \Big(\underset{\substack{\uparrow \\ \text{a specific sequence}}}{k^2 + k} \Big) \quad = \quad \underset{\substack{\uparrow \\ \text{a specific number}}}{20}$$

Warning!

It would make sense to have "$k = 0$" above the sigma:

$$\sum_{k=0}^{k=3} \left(k^2 + k\right) .$$

Example 1.12.27: expanding from sigma notation

The computation above is expanded here:

$$\sum_{k=0}^{3} \left(k^2 + k\right) =$$

k	$k^2 + k$		
0	$0^2 + 0$	$= 0$	$+$
1	$1^2 + 1$	$= 2$	$+$
2	$2^2 + 2$	$= 6$	$+$
3	$3^2 + 3$	$= 12$	
		$= 20$	

Exercise 1.12.28: contracting to sigma notation

How will the sum change if we replace $k = 0$ with $k = 1$, or $k = -1$? What if we replace 3 at the top with 4?

Example 1.12.29: contracting summation

This is how we *contract* the summation:

$$1^2+2^2+3^2+...+17^2=\sum_{k=1}^{n} k^2\,.$$

This is only possible if we find the nth-term formula for the sequence; in this case, $a_k=k^2$. And this is how we *expand* back from this compact notation, by plugging the values of $k=1,2,...,17$ into the formula:

$$\sum_{k=1}^{17} k^2=\underbrace{1^2}_{k=1}+\underbrace{2^2}_{k=2}+\underbrace{3^2}_{k=3}+...+\underbrace{17^2}_{k=17}\,.$$

Similarly, we have:

$$1+\frac{1}{2}+\frac{1}{2^2}+\frac{1}{2^3}+...+\frac{1}{2^{10}}=\sum_{k=0}^{10}\frac{1}{2^k}\,.$$

Exercise 1.12.30

Confirm that we can start at any other initial index if we just modify the formula:

$$1+\frac{1}{2}+\frac{1}{2^2}+\frac{1}{2^3}+...+\frac{1}{2^{10}}=\sum_{k=?}^{?}\frac{1}{2^{k-1}}=\sum_{k=?}^{?}\frac{1}{2^{k-2}}=...$$

Exercise 1.12.31

Contract this summation:

$$1+\frac{1}{3}+\frac{1}{9}+\frac{1}{27}=\,?$$

Exercise 1.12.32

Expand this summation:

$$\sum_{k=0}^{4}(k/2)=\,?$$

Exercise 1.12.33

Rewrite using the sigma notation:

1. $1+3+5+7+9+11+13+15$
2. $.9+.99+.999+.9999$
3. $1/2-1/4+1/8-1/16$
4. $1+1/2+1/3+1/4+...+1/n$
5. $1+1/2+1/4+1/8$
6. $2+3+5+7+11+13+17$
7. $1-4+9-16+25$

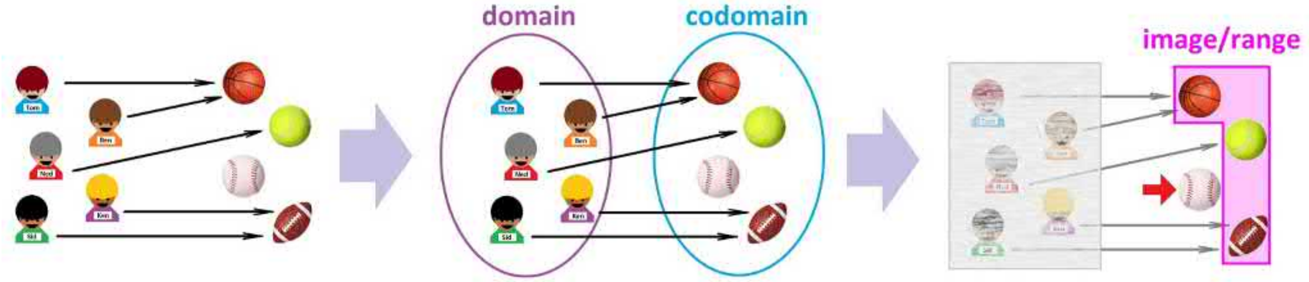
Chapter 2: Functions as transformations

Contents

2.1. The image: the range of values of a function

One of the most important characteristics of a function is the set of its values.

Let's go back to the set X of boys, the set of balls Y, and the "I prefer" function F from X to Y. A simple question we may ask about it is: What do the boys like *as a group*? It has a simple answer, a list: basketball, tennis, and football. We just have to look at Y and record each element that has an arrow drawn toward it:



This set is a subset of the codomain Y:

$$V = \{ \text{ basketball, tennis, football } \} \subset Y\,.$$

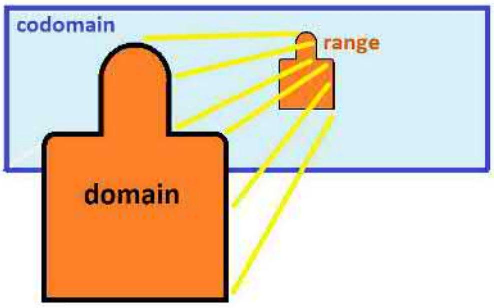
While the latter is the set of all possible or potential values of F, the former is the set of *actual* values. In other words, this is the range of values of the function. It can be, but is not in this case, the whole codomain.

Definition 2.1.1: image of function

The *image*, or the *range*, of a function $F : X \to Y$ is the set of all of its values, i.e.,

$$\{ y : \ F(x) = y\,, \quad \texttt{FOR SOME } x \}\,.$$

We see the image of the domain reflected in the codomain:



In the definition, we test each y: Is there a corresponding x? If there is, we add this y to the set.

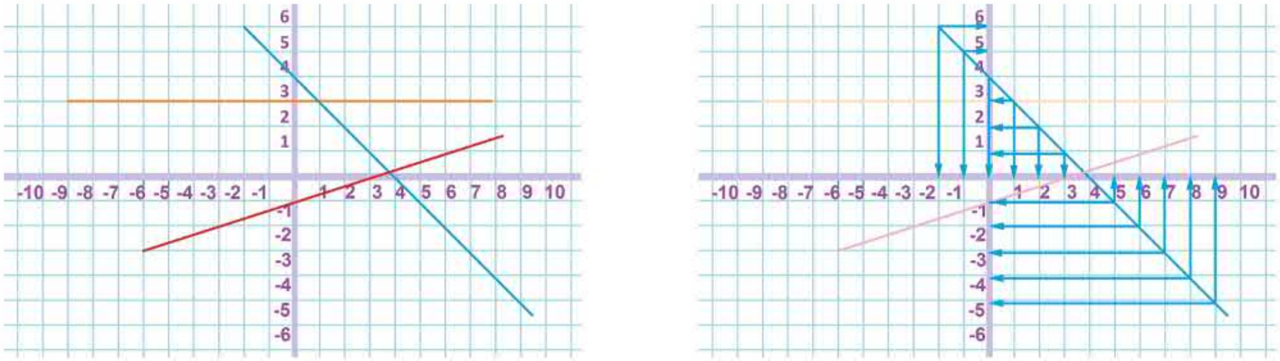
Exercise 2.1.2

Explain how one can use switch from the original codomain of F to its range V and create a new function $G : X \to V$. Are there any others?

Now *numerical functions.*

You get an idea about what the range is by simply looking at the y-column of the *table of values* of the function (just as looking at the x-column gives you an idea about the domain). However, finding the set explicitly requires some algebra.

Linear functions are simple. Consider how the arrows go from x to the graph and then to y:



All of them can be reversed, unless this is a constant function.

Now algebra. We need to try, if possible, to find an x for each y. The computations are easy too:

$$y = mx + b \implies x = \frac{y-b}{m},$$

for $m \neq 0$. So, there is an x for every y! We have proven the former part of the following:

Theorem 2.1.3: Image of Linear Function

The image of a linear function $y = mx + b$ is one of the two:

- *the set of real numbers, $V = \mathbf{R}$, when $m \neq 0$; otherwise,*
- *a single point, $V = \{b\}$.*

Exercise 2.1.4

Prove the latter part.

Exercise 2.1.5

State the theorem as an equivalence (an “if-and-only-if” statement).

Example 2.1.6: range of x^2 and x^3

Can we make the same argument for $f(x)=x^2$? Of course not: Squares can't be negative! But what about the rest of y's? We attempt to solve the equation, for each y:

$$y=x^2 \implies x=\sqrt{y} \quad \texttt{AND} \quad y\geq 0\,.$$

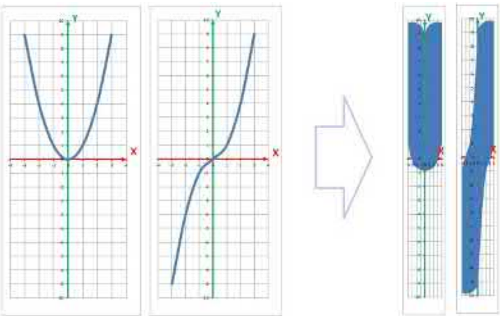
Therefore, the image of x^2 is

$$\{y:\ y\geq 0\}=[0,+\infty)\,.$$

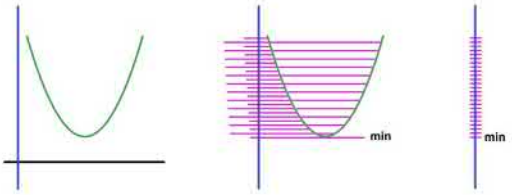
What about x^3? It works for *any* y:

$$y=x^3 \implies x=\sqrt[3]{y}\,.$$

Why such a difference? In addition to the algebra above, we will appreciate the difference between the two functions by examining their graphs. For example, we can thicken them and shrink the xy-plane:



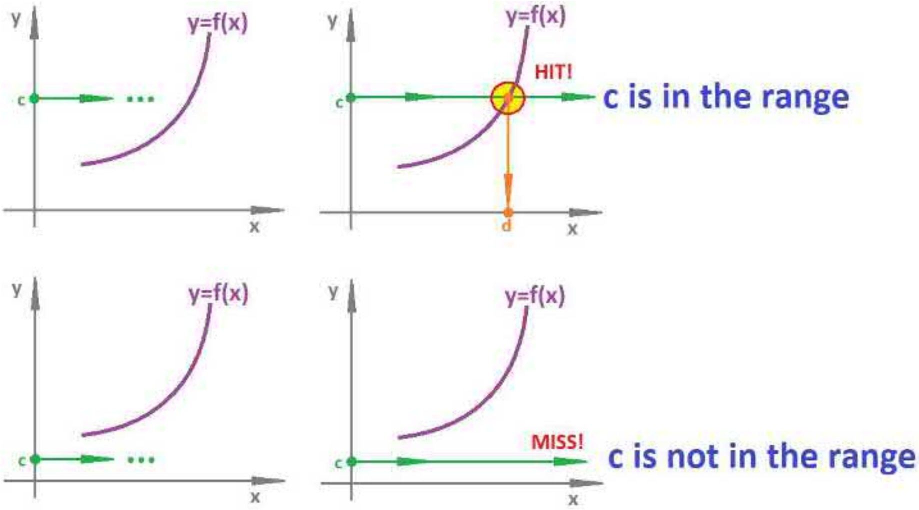
So, the vertical "spread" of the graph gives us the range (the horizontal "spread" of the graph gives us the domain). This is how the image of $y=x^2$ is seen as a ray in the y-axis:



Exercise 2.1.7

Use these two methods to find the domain of these functions.

Generally, to find the image of a numerical function, the graph of which is supplied, we test y's one at a time. From each of them, we draw a horizontal line and note whether it crosses the graph:

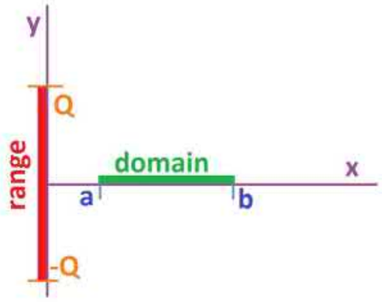


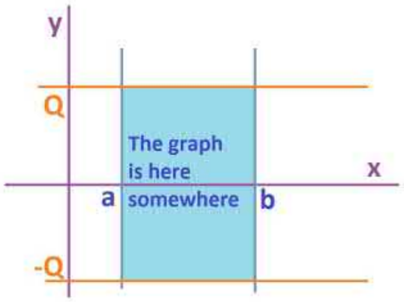
So, these graphs cannot touch or cross the x-axis, and that is the same as to say that 0 isn't in the image. Likewise, their graphs cannot touch or cross the y-axis and that is the same as to say that 0 isn't in the domain. That's the analogy – and the symmetry – of the problems of the domain and the image (not the codomain). It is the symmetry between the x-axis and the y-axis in the xy-plane.

Warning!

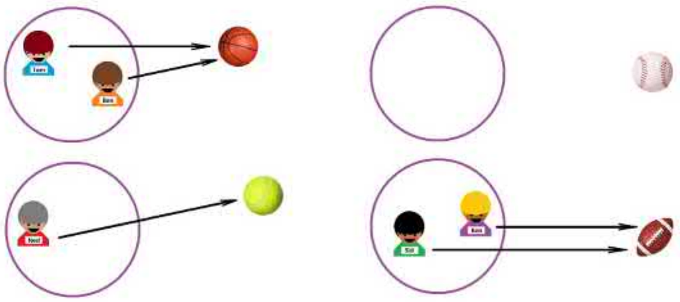
Image $\neq$ codomain.

If the domain and the range are intervals, the graph of the function is contained in the rectangle with these sides:





Another question we can ask about the boys and the balls is: Who likes basketball? or baseball, etc.? We just look at the arrow, or arrows, that is drawn towards this ball and note where it comes from. The result is a subset of X. This is what happens with the above example:



Definition 2.1.8: preimage of value

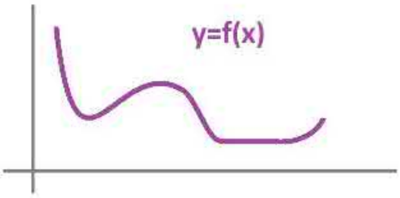
The *preimage* of an element b in a set Y under a function $F : X \to Y$ is the set of all x's the value of which under F is b, i.e.,

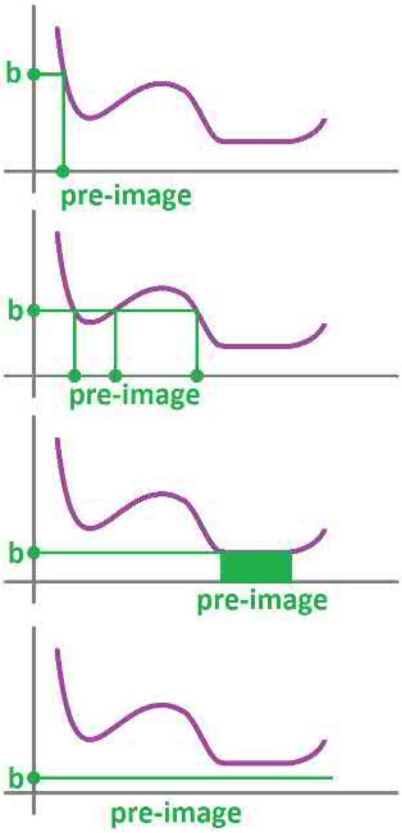
$$\{x : \ F(x) = b\}\,.$$

In other words, we are *solving equations* again.

We carry out this computation for every ball. We discover, in particular, that the preimage of baseball is the empty set. This is always the case with outputs outside the range!

The picture below illustrates how to find the preimage of a point of a numerical function:





This is what we already know:

- The preimages under a constant function are empty with an exception of a single value, the preimage of which is the whole x-axis.
- The preimages under linear (non-constant) polynomials are single points.

Exercise 2.1.9

Prove the above statements.

2.2. Numerical functions are transformations of the line

...and vice versa.

When we face a numerical function $y = f(x)$ given to us without any prior background, it is a good idea to create a *tangible representation* for it. The two main ways we are familiar with are these:

1. We think of the function as if it represents *motion*: x time, y location.
2. We plot the *graph* of the function on a piece of paper.

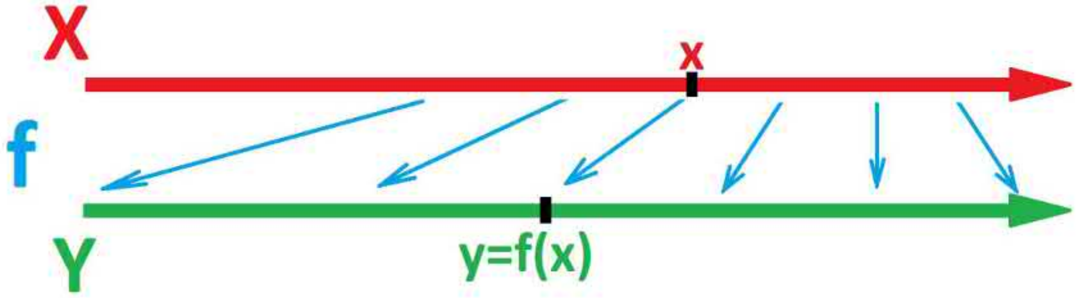
Unfortunately, both approaches fail at higher dimensions:

1. First, the time cannot be an analogue of x when we have a function $f : \mathbf{R}^2 \to \mathbf{R}$.
2. Second, the graph of $f : \mathbf{R}^2 \to \mathbf{R}^2$ simply cannot be visualized in the 3-dimensional space, let alone a piece of paper.

This why we set these two aside for now and consider a third approach:

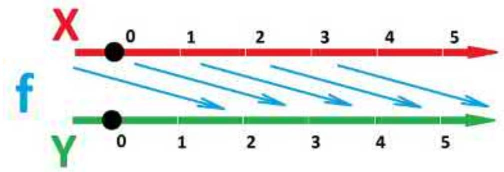
- We think of the function as a *transformation* of the real line.

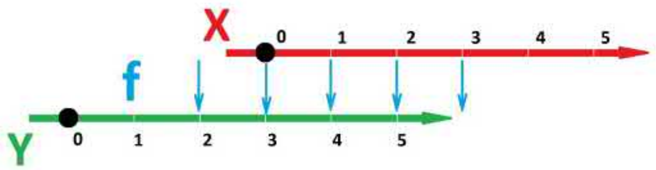
First, let's go back to representing a numerical function as a correspondence between the x-axis and the y-axis:



As you can see, these are the same axes we use to plot the graph, but they are arranged *parallel* to each other instead of perpendicular. The arrows indicate where each x lands on the y-axis. In fact, they suggest what happens to the *whole* x-axis.

Let's consider something specific:



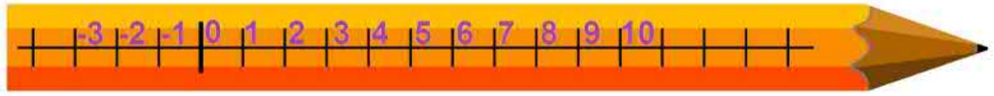


We ask: What has happened to the x-axis under this transformation?

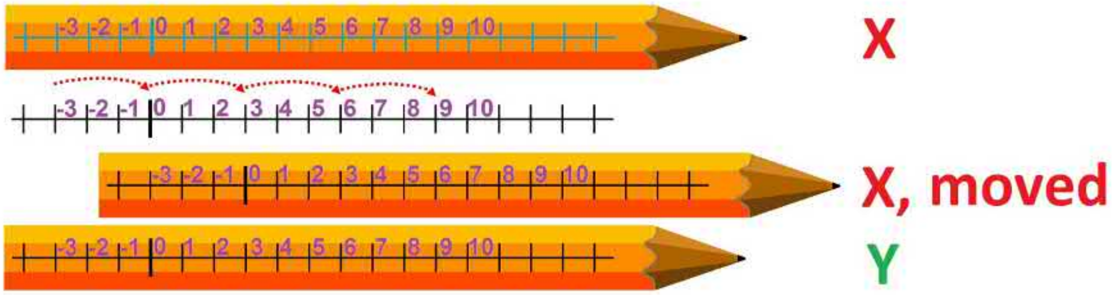
Warning!

Above, you see two ways to interpret the function: (1) arrows are between the x-axis and the intact y-axis, or (2) we move the y-axis so that $y = f(x)$ is aligned with x. The approaches give two different, even opposite, answers to the question. We will follow the former here.

To make this more tangible, we will think as if the whole x-axis, X, is drawn on a pencil:



The transformation we are starting with is a *shift*. We simply slide the pencil horizontally. Furthermore, there is another pencil, Y, to be used for reference. The markings (i.e., its coordinate system) on the second pencil show the new locations of the points on the first:

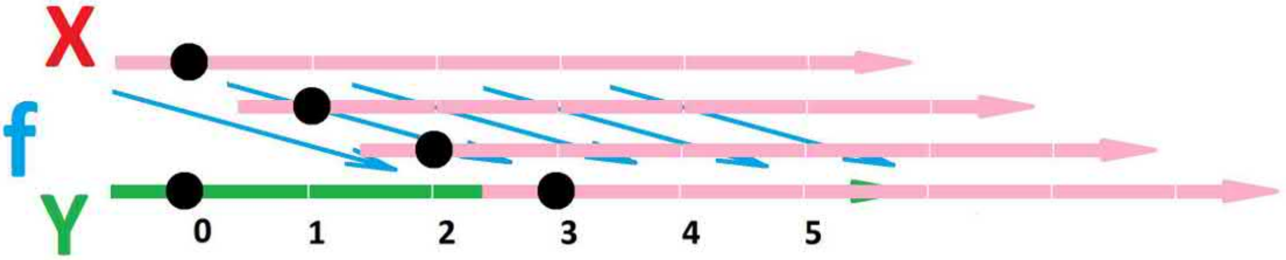


The example above is a shift of 3 units to the right.

Generally:

- ▶ When shifted $s > 0$ units right, point x moves to $x + s = y$.

This is where the algebra comes from:



> **Warning!**
>
> Replace "right" with "up" if the axes are aligned vertically; it's the positive direction that matters.

In the meantime, what about shifting left? We have the following:

- ▶ When shifted $s > 0$ units left, point x moves to $x - s$.

Of course, we can combine the two statements:

- ▶ When shifted $s > 0$ units right/left, point x moves to $x \pm s$.

Instead, we should simply allow s to be *negative*. Then we can interpret the former statement to include the latter if we understand "s units left" as "$-s$ units right". This is better:

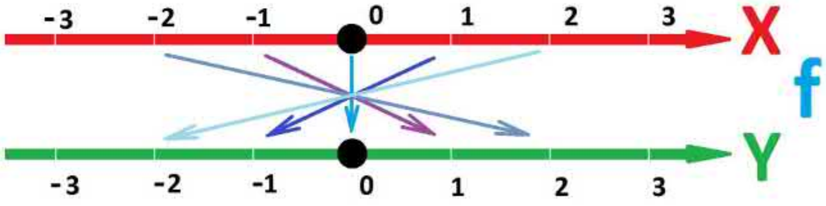
$$\boxed{x \xrightarrow{\text{right by } s} x + s}$$

In other words, we have the following function:

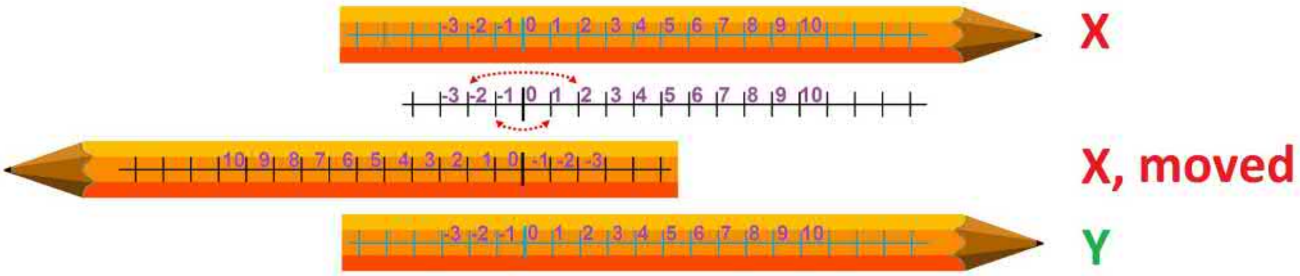
$$y = f(x) = x + s\,.$$

Of course, $s = 0$ means that there is no change.

Next, let's consider a transformation produced by the function $y = -x$. Each x then takes this trip: $x \mapsto -x$. We plot the arrows to show the origin and the destination:



To understand what is happening to the whole x-axis, we try to imagine what happens to our pencil. We lift, then *flip* the pencil with x-axis on it, and place it next to another such pencil used for reference:



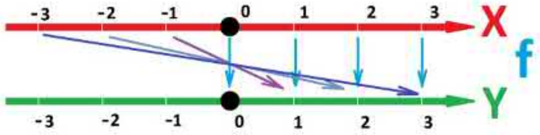
So, we have:

$$\boxed{x \xrightarrow{\text{flip}} -x}$$

The transformations that don't distort the line are called *rigid motions*. Examples of motions are the left shift and the right shift (the magnitude may vary) and the flip (the center may vary).

Example 2.2.1: fold

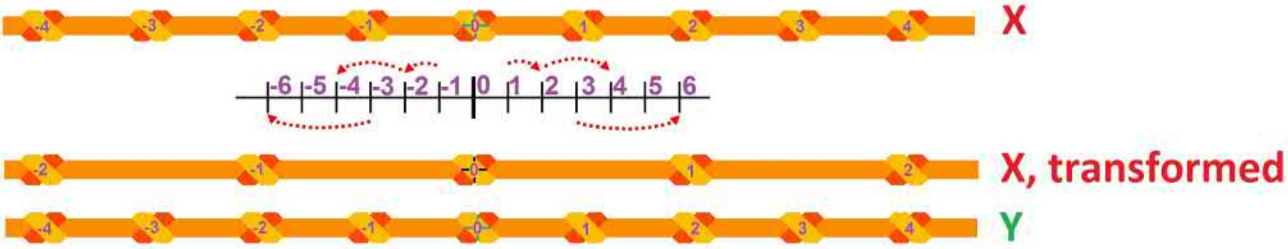
Going in the opposite direction, we might want to find a formula for a transformation we describe verbally. A different kind of transformation is a *fold*. In order to visualize it, we can't put the axis on a pencil anymore! Let's try a piece of wire:



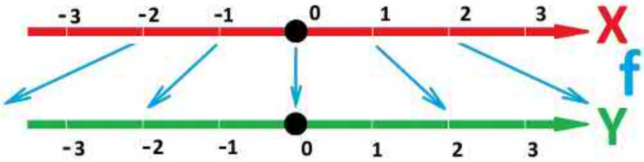
The arrow diagram on the right depicts how a half of the x-axis flips and the other stays put. What is this function? It is easy to see the range, $[0, +\infty)$, because that's where output land. This is the absolute value function, $y = |x|$! It is not linear.

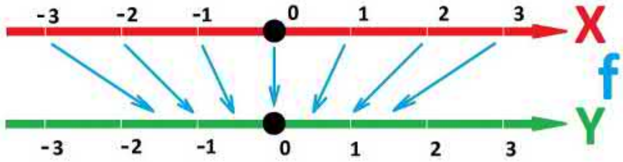
Another basic transformation is a *stretch*. The line isn't on a pencil or a wire anymore! It is a rubber string (with knots to mark locations):

We imagine that we grab it by the ends (at infinity?) and pull them apart in such a way that the center O doesn't move. For example, a stretch by a factor of 2 is shown below:



This is indeed a uniform stretch because the distance between *any* two points doubles on the left and is cut in half on the right:





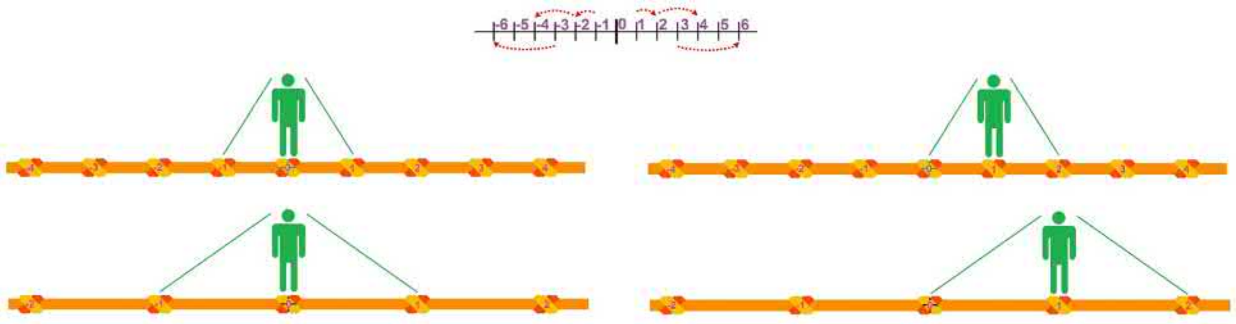
The transformations are simply $y = 2x$ and $y = x/2$.

Exercise 2.2.2

What does $f(x) = -2x$ do to the x-axis?

Example 2.2.3: expanding Universe

If you stretch something, it becomes bigger, and if you shrink something, it becomes smaller – unless it's infinite! The simplest model of the expanding Universe is the rubber band above. What we learn from it is that there is no center to this expansion or explosion; no matter where you are, you see that the distance to the nearest planet (mark) has doubled:



In contrast, a person whose world has been “shifted” or “flipped” won’t notice a thing.

Generally:

- When the line is stretched by a factor $m > 1$, point x moves to $x \cdot m$.

In the meantime, what about a *shrink*? We have the following:

- When the line is shrunk by a factor $k > 1$, point x moves to x/k .

Of course, in order to combine the two statements, we should allow m to be *less than* 1. Then we can interpret the former statement to include the latter if we understand “stretched by a factor m” as “shrunk by a factor $1/m$”. This is how we can describe it:

$$\boxed{x \quad \xrightarrow{\text{stretch by } m} \quad x \cdot m}$$

Of course, $m = 1$ means that there is no change. In other words, this is the function:

$$y = f(x) = x \cdot m\,.$$

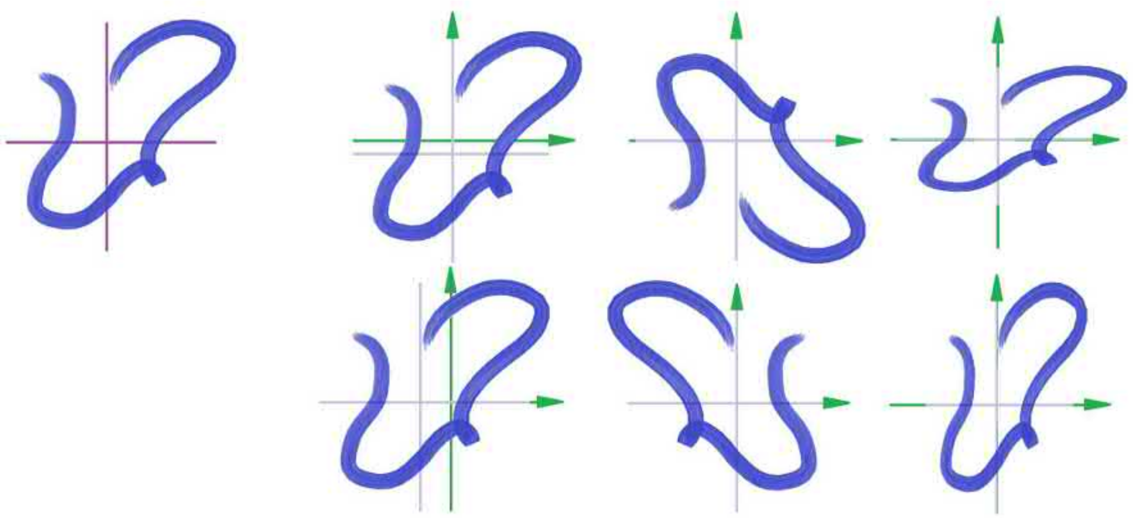
Exercise 2.2.4

What is the meaning of $|m|$ in the transformation given by $f(x) = mx + b$?

Exercise 2.2.5

What is the meaning of b in the transformation given by $f(x) = mx + b$?

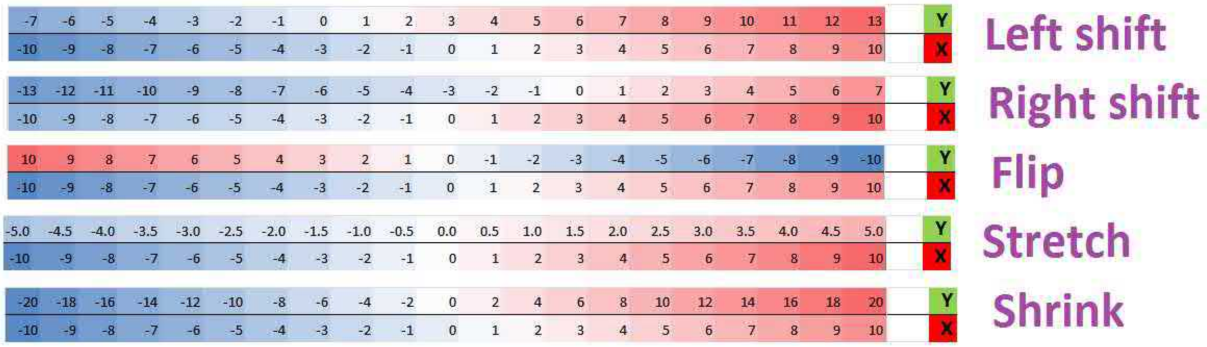
A function $f : \mathbf{R}^2 \to \mathbf{R}^2$ is a transformation of a plane to a plane:



Among examples of these are at least the ones above: shift, flip, and stretch. At least, these can be vertical and horizontal.

Example 2.2.6: coloring numbers

We will often color numbers according to their values. This is the summary of the functions we have considered:



Left shift

-7	-6	-5	-4	-3	-2	-1	0	1	2	3	4	5	6	7	8	9	10	11	12	13	Y
-10	-9	-8	-7	-6	-5	-4	-3	-2	-1	0	1	2	3	4	5	6	7	8	9	10	X

Right shift

-13	-12	-11	-10	-9	-8	-7	-6	-5	-4	-3	-2	-1	0	1	2	3	4	5	6	7	Y
-10	-9	-8	-7	-6	-5	-4	-3	-2	-1	0	1	2	3	4	5	6	7	8	9	10	X

Flip

10	9	8	7	6	5	4	3	2	1	0	-1	-2	-3	-4	-5	-6	-7	-8	-9	-10	Y
-10	-9	-8	-7	-6	-5	-4	-3	-2	-1	0	1	2	3	4	5	6	7	8	9	10	X

Stretch

-5.0	-4.5	-4.0	-3.5	-3.0	-2.5	-2.0	-1.5	-1.0	-0.5	0.0	0.5	1.0	1.5	2.0	2.5	3.0	3.5	4.0	4.5	5.0	Y
-10	-9	-8	-7	-6	-5	-4	-3	-2	-1	0	1	2	3	4	5	6	7	8	9	10	X

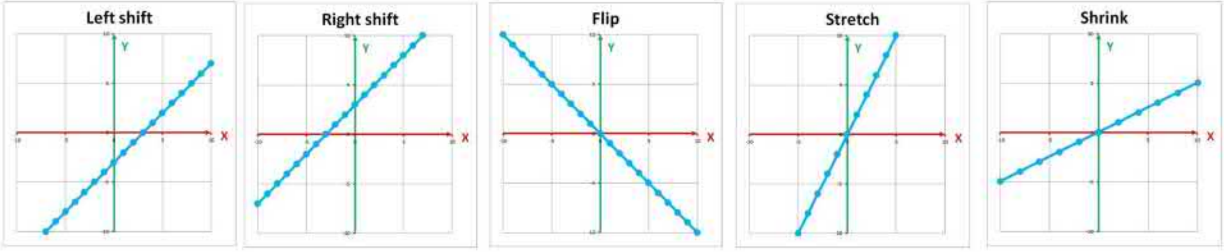
Shrink

-20	-18	-16	-14	-12	-10	-8	-6	-4	-2	0	2	4	6	8	10	12	14	16	18	20	Y
-10	-9	-8	-7	-6	-5	-4	-3	-2	-1	0	1	2	3	4	5	6	7	8	9	10	X

One can compare the colors of X to those of X as it lands on y and reach the following conclusions:

- The colors are clearly shifted in the first two images.
- In the third, the blue and the red are interchanged.
- In the fourth, the colors don't change on Y as fast as on X, while in the fifth, they are changing faster.

We also plot the graphs of these functions below:



Left shift

Right shift

Flip

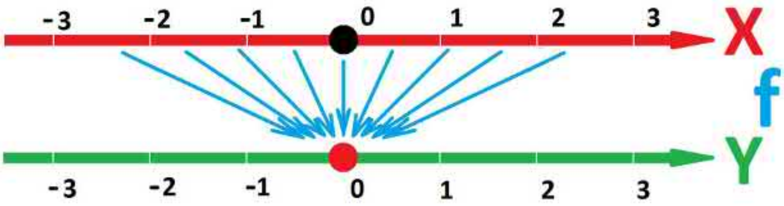
Stretch

Shrink

We can, therefore, classify the transformations based on the slopes of their graphs:

- shift (rigid motion), slope 1;
- flip (rigid motion), slope -1;
- stretch by 2, slope 2;
- shrink by 2, slope $1/2$.

A very simple, but important, class of function is the *constant functions*, $f(x) = c$. What does this function do to the x-axis? There is only one output:



The real line is shrunk to a single point; we can call this transformation *collapse*.

$$\boxed{x \xrightarrow{\text{collapse}} c}$$

Of course, one may see it as an extreme case of a stretch/shrink, with a factor $m = 0$. The slope of the graph is, of course, 0.

In summary, these are transformations created by the function $y = mx$ with various m's:

	collapse	shrink	identity	stretch
m :	0		1	

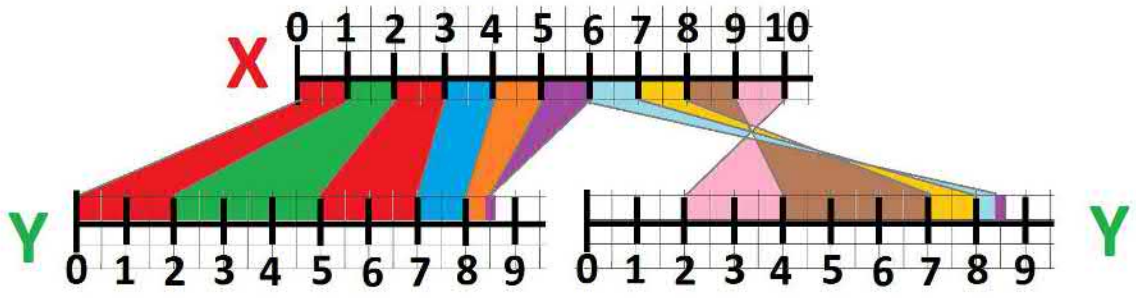
In general, it is typical to have different stretching factors at different locations.

Example 2.2.7: transformation from list of values

Let's consider this function given by its values:

x	0	1	2	3	4	5	6	7	8	9	10
y	0	2	5	7	8	8.5	8.5	8	7	4	2

We assume that the function continues between these values in a linear fashion. For example, the interval $[0, 1]$ is mapped to $[0, 2]$ linearly ($y = 2x$), the interval $[1, 2]$ to $[2, 5]$, etc. The 1-unit segments on the x-axis are stretched and shrunk at different rates, and the ones beyond $x = 6$ are also flipped over (left):



We also plot the graph of this function on the right; the stretching factors become the slopes! So, at its simplest, a non-linear function is a combination of linear patches.

Exercise 2.2.8

Represent the following function given by its values as a transformation:

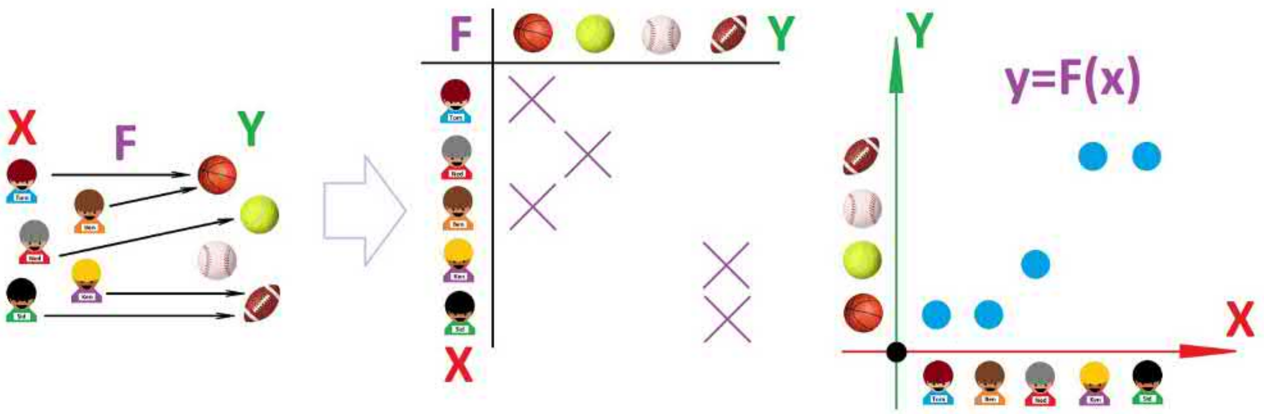
x	0	1	2	3	4	5	6	7	8	9	10
y	6	5	4	3	3	3	4	5	6	6	6

Make up your own functions and then represent them as transformations. Repeat.

2.3. *Functions with regularities: one-to-one and onto*

We go back to our example of a function that assigns to each boy a ball to play with.

In order for this to be a function from the former to the latter, the table of this relation must have exactly one mark in each *row*:

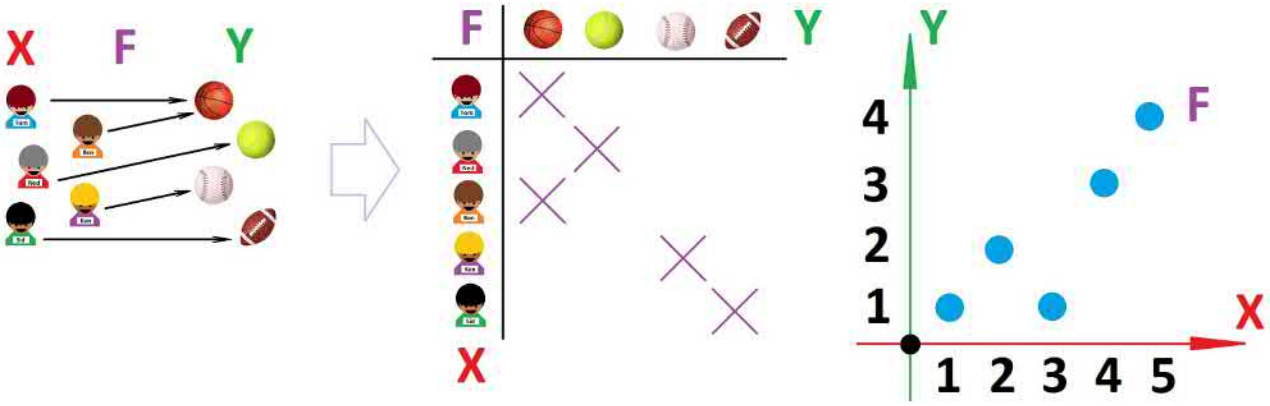


It does. But what about the *columns*?

If we further study this function, we might notice two different but related "irregularities".

First, no one seems to like baseball! There is no arrow ending at the baseball, and its column has no marks.

Let's modify the function slightly: Ken changes his preference from football to baseball. Then, there is an arrow for each ball, and all columns in the table have marks:



In the graph, there is a dot corresponding to each horizontal line.

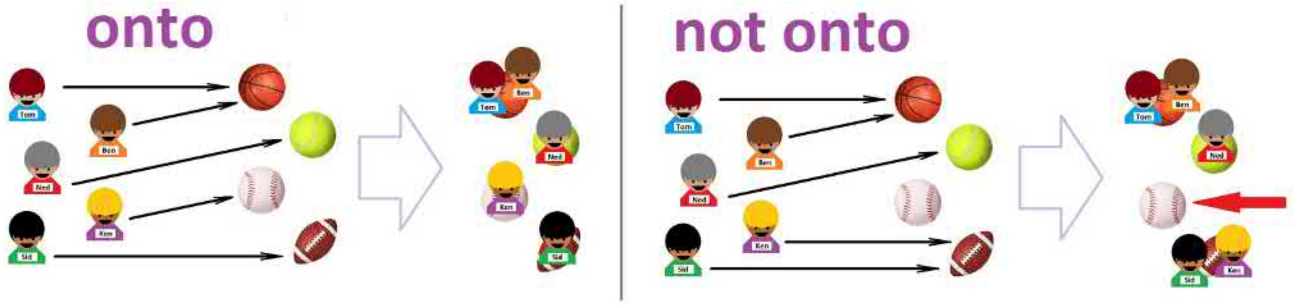
The following concept is crucial:

Definition 2.3.1: onto function

A function $F : X \to Y$ is called *onto* when there is an x for each y with $F(x) = y$.

In other words, no potential output is "wasted".

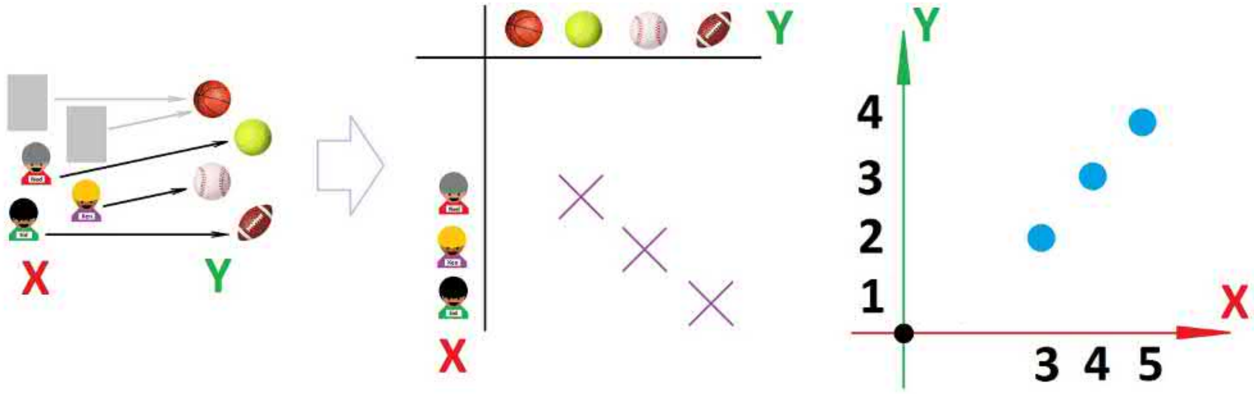
Below, the reason for this terminology is explained:



We start with an element of X and bring it – along the arrow – to the corresponding element of Y. The function is onto if all elements of Y are *covered*.

Second, both Tom and Ben prefer basketball! The two arrows converge on the basketball, and we can also see that its column has two marks. We note the same about the football.

The above function is modified: Tom and Ben have left. Then, no two arrows converge on one ball, and no column has more than one mark:



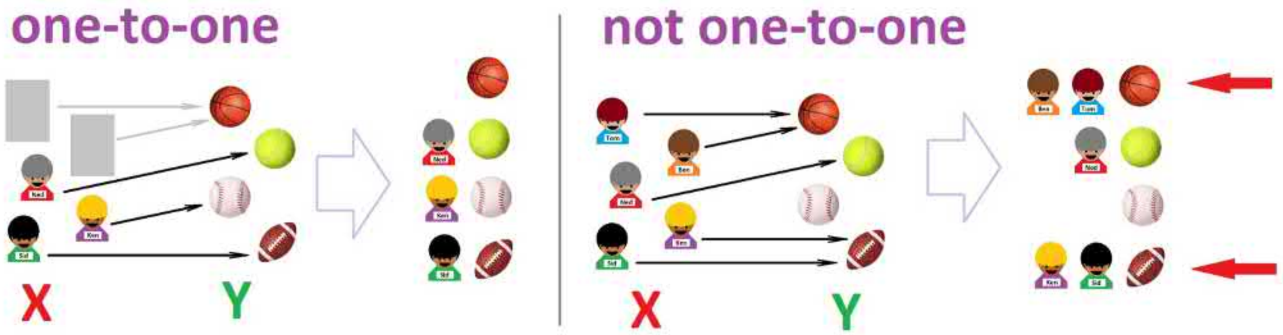
In the graph, there can only be one dot, or none, corresponding to each horizontal line.

The following concept is crucial.

Definition 2.3.2: one-to-one function

A function $F : X \to Y$ is called *one-to-one* when there is at most one x for each y with $F(x) = y$.

Below, the reason for this terminology is explained:



We start with an element of X and bring it – along the arrow – to the corresponding element of Y. The function is one-to-one if every element of Y is covered *only once*, if at all.

In summary, the two concepts are not about how many arrows *originate* from each x – it's always one – but about how many arrows *arrive* at each y. The logic of the two definitions is quite different:

- Onto: `FOR EACH` x `THERE IS` y such that $y = F(x)$.
- One-to-one: `IF` $F(x_1) = F(x_2)$ `THEN` $x_1 = x_2$.

Now *numerical functions.*

These functions are represented by their transformations and by their graphs. We will follow these two approaches to discover whether a function satisfies one of the two definitions.

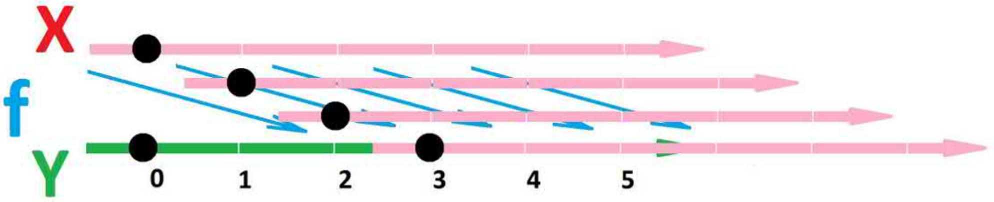
The following will be our assumption:

- ▶ The codomain is the set of all reals $\mathbf{R}$.

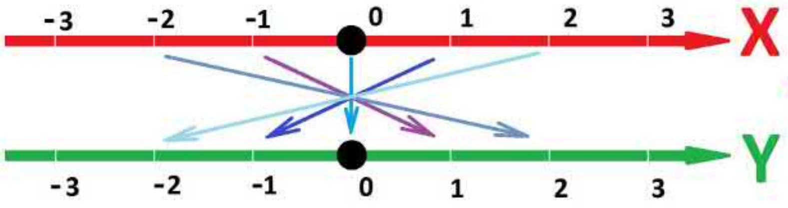
Example 2.3.3: basic transformations

Let's consider the transformations from the last section. Their descriptions – and illustrations – tell the whole story. We just need to look at how the arrows arrive at the y-axis.

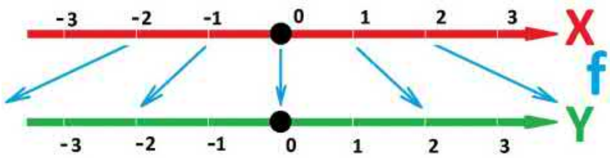
The left and right shifts:

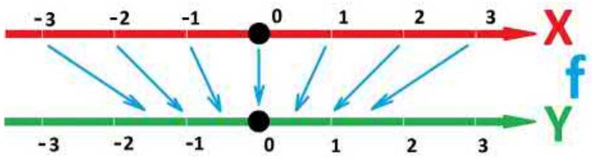


The flip:



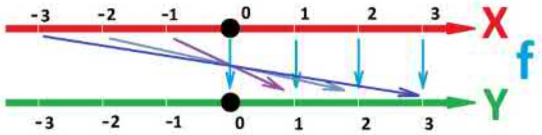
The stretch and the shrink:



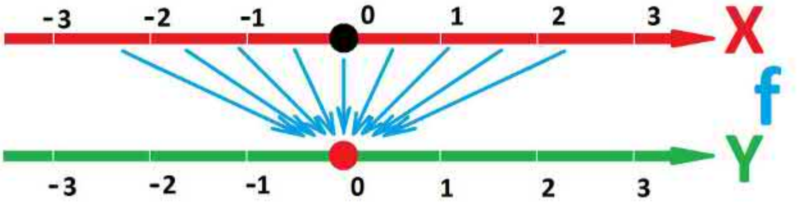


They are all both one-to-one and onto!

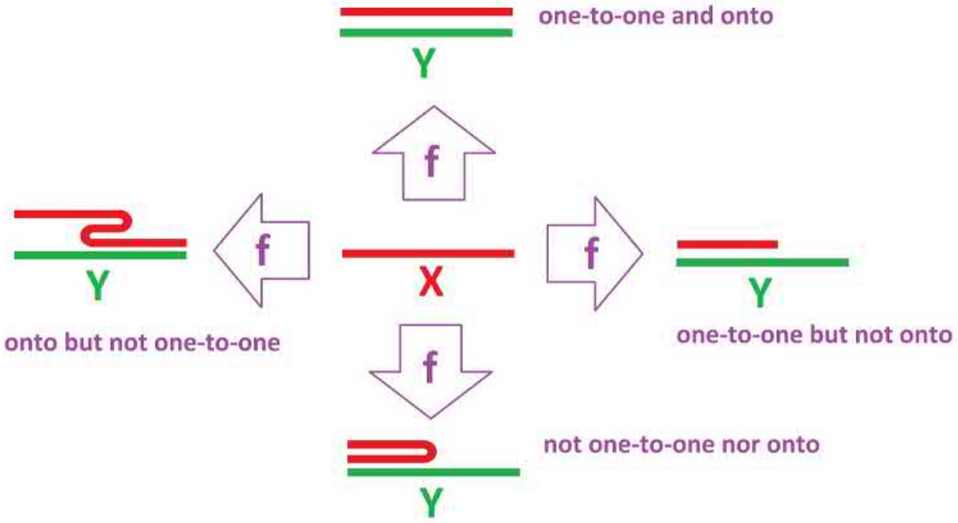
What would make a transformation to be not one-to-one? Any folding presents a visible problem:



And so does any collapsing:



So, if we imagine that the x-axis, X, is *transformed* somehow and then placed on top of the y-axis, Y, we can understand this function as a transformation. This can happen in these four basic ways:



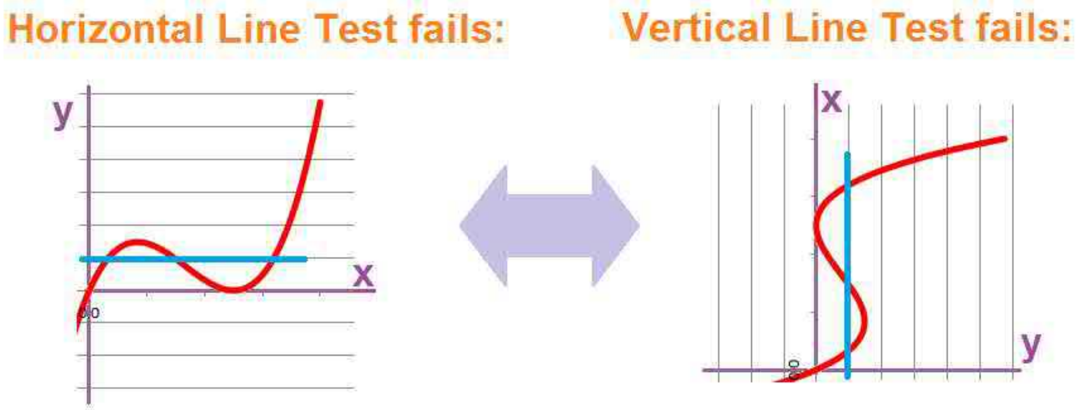
The questions we ask are these two:

- Onto: Does X cover the whole Y?
- One-to-one: Does X cover any location on Y no more than once?

What makes a difference? Two points on the graph have the same height above the x-axis! The following is a useful observation:

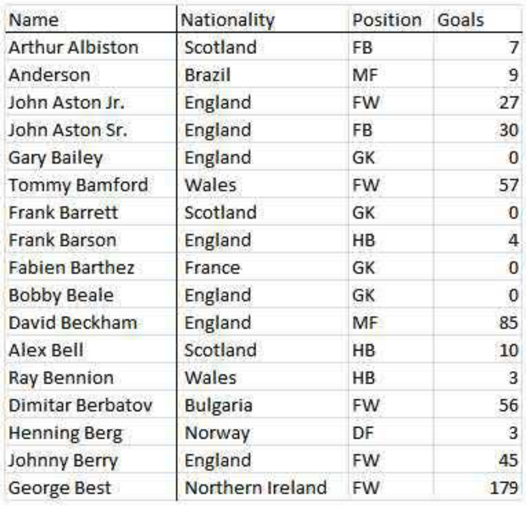
> ▶ A function is one-to-one `IF AND ONLY IF` the intersection of its graph and any *horizontal line* contains at most one point.

Notice the connection with the *Vertical Line Test for Relations* (“Is this a function?”):



Exercise 2.3.4

Identify and classify the functions below:



Name	Nationality	Position	Goals
Arthur Albiston	Scotland	FB	7
Anderson	Brazil	MF	9
John Aston Jr.	England	FW	27
John Aston Sr.	England	FB	30
Gary Bailey	England	GK	0
Tommy Bamford	Wales	FW	57
Frank Barrett	Scotland	GK	0
Frank Barson	England	HB	4
Fabien Barthez	France	GK	0
Bobby Beale	England	GK	0
David Beckham	England	MF	85
Alex Bell	Scotland	HB	10
Ray Bennion	Wales	HB	3
Dimitar Berbatov	Bulgaria	FW	56
Henning Berg	Norway	DF	3
Johnny Berry	England	FW	45
George Best	Northern Ireland	FW	179

Exercise 2.3.5

A function $y = f(x)$ is given below by a list of some of its values. Add missing values in such a way that the function is one-to-one.

x	-1	0	1	2	3	4	5
$y = f(x)$	-1		4	5		2	

Exercise 2.3.6

What codomain of the function given above should we assume to assure that it is onto?

We summarize the results in the following important theorem.

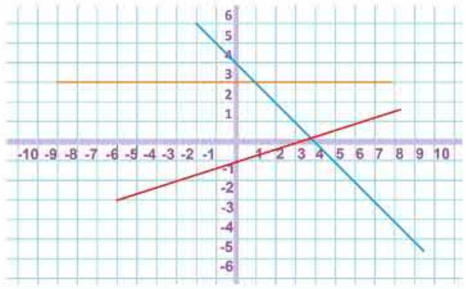
Theorem 2.3.7: Horizontal Line Test

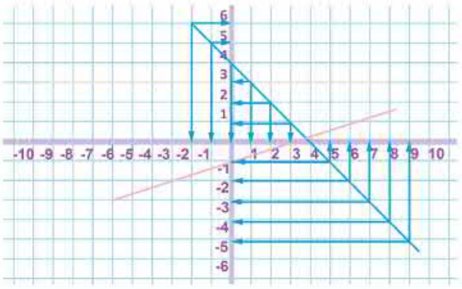
1. *A function is onto if and only if its graph and any horizontal line have at least one point in common.*
2. *A function is one-to-one if and only if its graph and any horizontal line have at most one point in common.*

Exercise 2.3.8

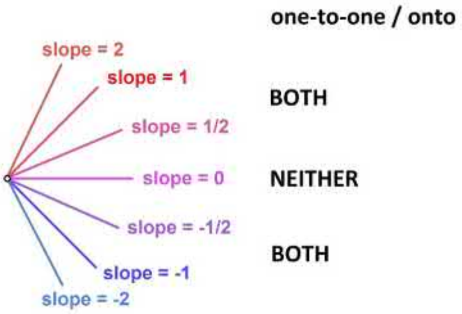
Break either part of the theorem into a statement and its converse.

The arrows for linear functions are easy to reverse except for the constant function:





We can then classify them according to their slopes:



Indeed, as we know from Euclidean geometry, two straight lines have exactly one intersection unless they are parallel; therefore the *Horizontal Line Test* is passed by all linear functions except the ones with zero slope. Those are constant functions.

However, we need to prove these facts algebraically.

Example 2.3.9: linear polynomials

Suppose we have a linear function with slope 3 and y-intercept 2:

$$F(x) = 3x + 2 \,.$$

Is it one-to-one?

Suppose we have two inputs x_1 and x_2. Can their outputs be equal under F? Let's try: Suppose $F(x_1) = F(x_2)$. We substitute and obtain the following:

$$3x_1 + 2 = 3x_2 + 2 \,.$$

Canceling 2 produces the following:

$$3x_1 = 3x_2 \,.$$

Finally, we divide by 3:

$$x_1 = x_2 \,.$$

No, the outputs are equal only when the inputs are!

Is it onto? Suppose we have an output y. Is there an x taken to y under F? We just need to solve the equation $F(x) = y$ for x, for each y. We have an equation:

$$3x + 2 = y \,,$$

with an unspecified y. No matter what y is though, we subtract 2 and then divide by 3, producing:

$$x = \frac{y-2}{3} \,.$$

Yes, there is such an x, for each y!

What made these computation possible is our ability to divide by the slope 3, which is non-zero.

The classification of linear function is simple:

Theorem 2.3.10: Linear Functions, One-to-one Onto

A linear function with slope m and y-intercept b,

$$F(x)=mx+b\,,$$

is both one-to-one and onto as long as $m\neq 0$.

Proof.

One-to-one. Suppose we have two inputs x_1 and x_2. Then we use the fact that $m\neq 0$:

$$F(x_1)=F(x_2)\implies mx_1+b=mx_2+b\implies mx_1=mx_2\implies x_1=x_2\,.$$

It's the same input.

Onto. Suppose we have an output y. We solve the equation $F(x)=y$ for x using the fact that $m\neq 0$:

$$mx+b=y\implies x=\frac{y-b}{m}\,.$$

There is such an x.

Exercise 2.3.11

Break the theorem into a statement and its converse.

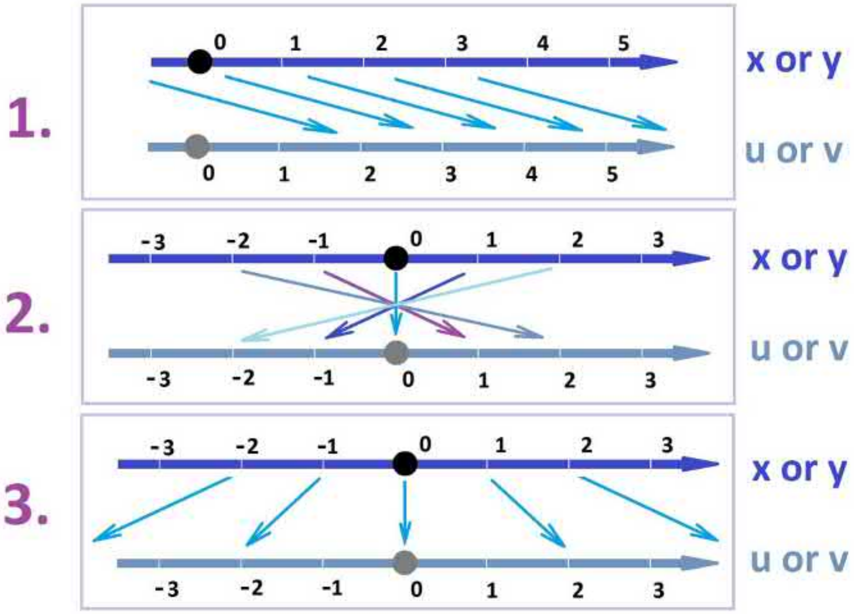
Exercise 2.3.12

Prove algebraically that $f(x)=1/x^2$ is neither one-to-one nor not onto. Show how the answer changes with your choice of the codomain.

Exercise 2.3.13

Classify the function $f(x)=x^3-x$ according to these two definitions. Prove algebraically.

The three basic transformation of the line – shift, flip, and stretch – are both one-to-one and onto:



It's no surprise to us because they are simply linear functions with the slopes respectively:

$$m = 1,\ m = -1,\ m = 2\,.$$

In the meantime, the collapse has slope $m = 0$. It's neither.

Furthermore, we see that they are one-to-one in the following. The distances between any two points remain the same in the first two cases and double in the last case. In case of a shrink, they will decrease but never to 0 as in the case of the collapse (slope $m = 0$, and all the distances become 0).

Exercise 2.3.14

Prove these statements.

The preimages of all points are also points.

We can also see that they are onto in the fact that the whole y-axis is covered by the image of the x-axis.

Exercise 2.3.15

Prove these statements.

In light of this analysis, the following is just a restatement of the two definitions:

Theorem 2.3.16: One-to-one and Onto vs. Image

1. *A function $F : X \to Y$ is one-to-one if and only if the preimage of every element of the codomain Y is a single element of the domain, X, or it's empty.*
2. *A function $F : X \to Y$ is onto if and only if its image is the whole codomain, Y.*

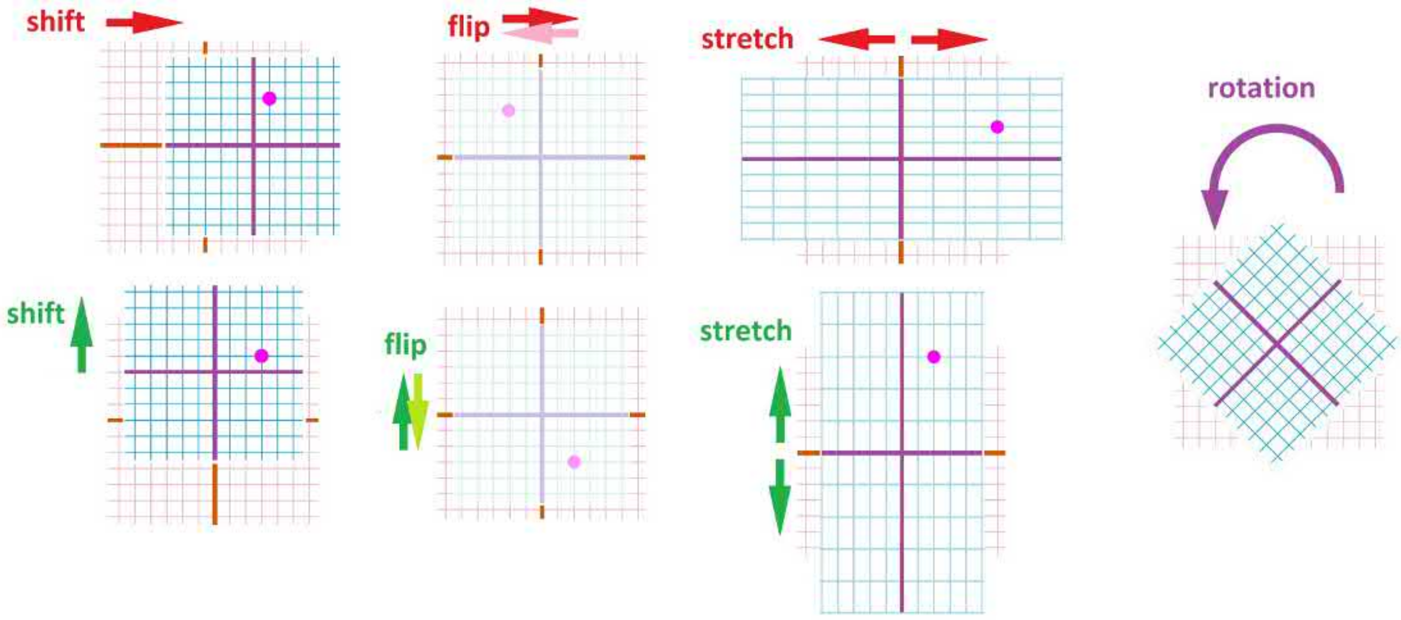
Exercise 2.3.17

Break either part of the theorem into a statement and its converse.

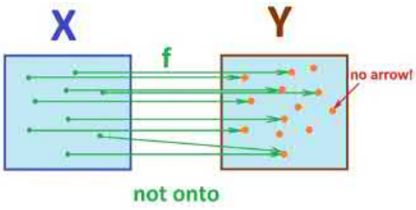
Exercise 2.3.18

Prove the theorem.

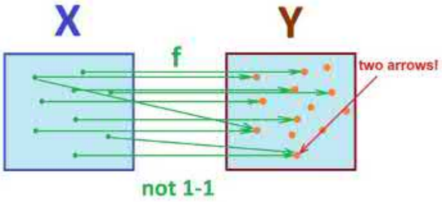
The four basic transformation of the plane – shift, flip, stretch, and rotation – are also both one-to-one and onto:



To summarize, the restrictions in these two definitions can be violated when there are too few or too many arrows arriving to a given y. These violations are seen in the codomain. This one is *not onto*:

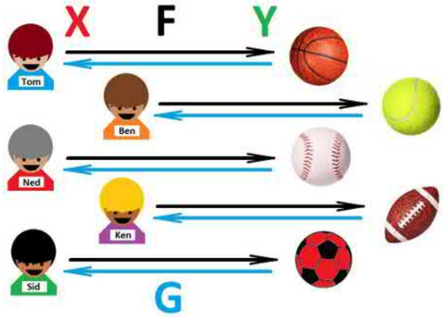


That one is *not one-to-one*:



Example 2.3.19: both one-to-one and onto

What functions are the most "regular"? The ones that are both one-to-one and onto, sometimes called *bijections*:



The function may look plain but it has an important property: the boys and the balls can be used as substitutes of each other! For example, you don't have to remember the name of every boy but just say "the one that plays basketball" to identify Tom without a chance of confusion. Or you can say "the ball that Ken plays with" without a chance of confusion.

Exercise 2.3.20

What are the smallest set X and the smallest set Y for which a function $F : X \to Y$ can be: (a) not one-to-one, (b) not onto?

Exercise 2.3.21

Sketch the graph of the function f given by its list of values below. Is it one-to-one?

x	1	2	3	4	5
$y = f(x)$	1	2	0	3	1

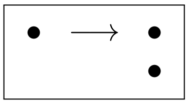
Make up your own functions by providing their lists of values and test the two definitions. Repeat.

Exercise 2.3.22

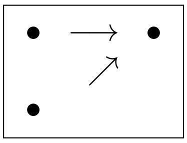
What kind of function is "many-to-one"? What about "one-to-many"?

In summary, we present the two simplest ways the two conditions can be *violated*:

Not onto:

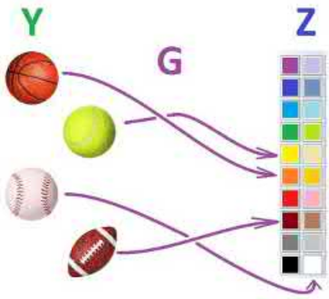


Not 1-1:



2.4. Compositions of functions

Back to our boys-and-balls example, let's note the *colors of the balls*. This has nothing to do with the boys, and it creates a new function:

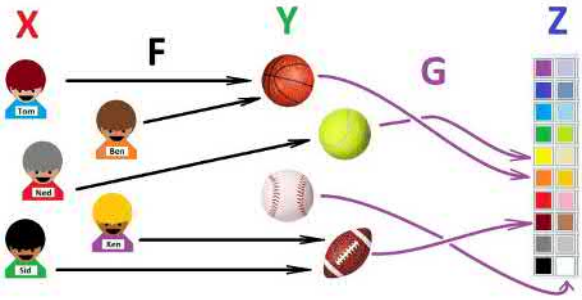


It is a function $G : Y \to Z$ from the set of all balls to the new set Z of the main colors.

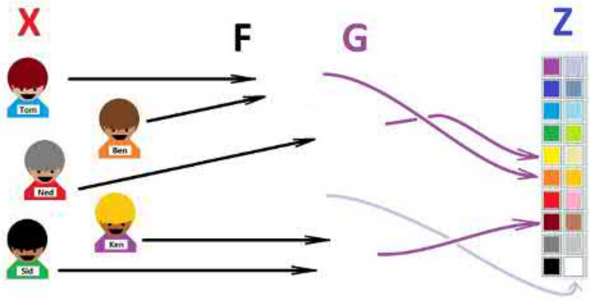
Exercise 2.4.1

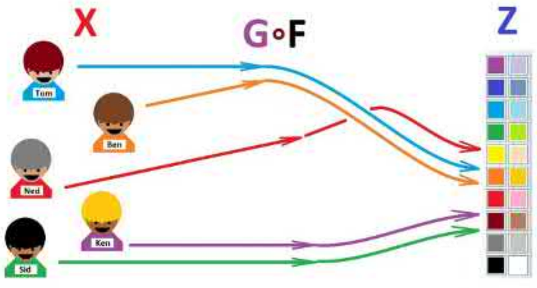
Is the new function one-to-one or onto?

We know the boys' preferences in balls, but does it entail any *preferences in colors*? In a sense. We can just combine the new function with the old:

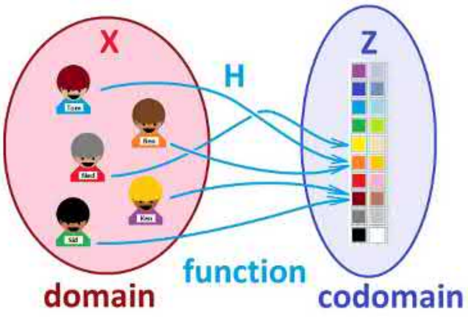


If we start with a boy on the left, we can continue with the arrows all the way to the right. This way, we will know the color of the ball the boy is playing with:





This is a new function, say $H : X \to Z$, from the set of boys X to the set Z of the main colors:



Exercise 2.4.2

Is the new function one-to-one or onto?

In general, this is the setup:

$$\begin{tikzcd} X \arrow[r, "F"] & Y \arrow[r, "G"] & Z \end{tikzcd}$$

The following concept is very important.

Definition 2.4.3: composition of functions

Suppose we have two functions (with the codomain of the former matching the domain of the latter):

$$F : X \to Y \text{ and } G : Y \to Z\,.$$

Then their *composition* is the function (from the domain of the former to the codomain of the latter)

$$H : X \to Z\,,$$

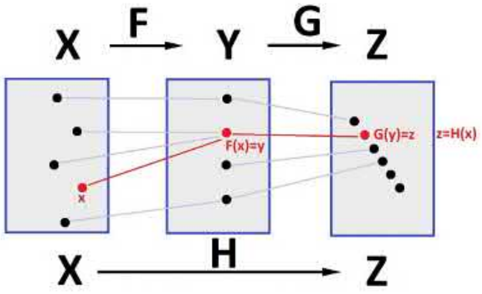
which is computed for every x in X according to the following two-step procedure:

$$x \mapsto F(x) = y \mapsto G(y) = z$$

It is denoted by

$$\boxed{G \circ F}$$

We can imagine that all three sets are copies of the plane and just follow from X along the arrows of F to Y and then along the arrows of G to Z:



Warning!

The requirement that "the codomain of the former is the domain of the latter" could be replaced with "the domain *contains* the codomain".

In other words, the new function is evaluated by the *substitution formula*:

$$\boxed{z = H(x) = G(F(x))}$$

This is the "deconstruction" of the notation:

Composition of functions

$$\underset{\substack{\uparrow\\ \text{name of the new function}}}{\left(G\circ F \right)(x)} \quad = \quad \overset{\substack{\text{names of the second and first functions}\\ \downarrow \qquad \downarrow}}{G\Big(\ F\ (x)\ \Big)}_{\substack{\uparrow\ \uparrow\ \uparrow\\ \text{substitution}}}$$

The name of the function on left reads "G composition F". The computation on right reads "G of F of x".

It is not that the operations are read from right to left, but rather *from inside out*!

If we represent the two functions as *black boxes*, we can wire them together:

$$\begin{array}{ccccccccc} \text{input} & & \text{function} & & \text{output} & & & & \\ x & \to & \boxed{\ F\ } & \to & y & & & & \\ & & & & \downarrow & & & & \\ & & & & \text{input} & & \text{function} & & \text{output} \\ & & & & y & \to & \boxed{\ G\ } & \to & z \end{array}$$

Thus, we use the output of the former as the input of the latter.

To make it clear that Y is no longer a part of the picture, we can also visualize the composition as follows:

$$\begin{tikzcd} X \arrow[r, "F"] \arrow[dr, "H"] & Y \arrow[d, "G"] \\ & Z \end{tikzcd}$$

The meaning of the diagram is as follows: Whether we follow the F-then-G route or the direct H route, the results will be the same.

If we think of functions as *lists of instructions*, we just attach the list of the latter at the bottom of the list of the former. In other words, here is the list of $G \circ F$:

- Step 1: Do F.
- Step 2: Do G.

They are executed *consecutively*; you can't start with the second until you are done with the first.

Let's test this idea on our example. We take the two *lists of values* and then cross-reference them (from left to right):

$$\boxed{\begin{array}{lcl} F(\text{ Tom }) & = & \text{basketball} \\ F(\text{ Ned }) & = & \text{tennis} \\ F(\text{ Ben }) & = & \text{basketball} \\ F(\text{ Ken }) & = & \text{football} \\ F(\text{ Sid }) & = & \text{football} \end{array}} \ , \ \boxed{\begin{array}{lcl} G(\text{ basketball }) & = & \text{orange} \\ G(\text{ tennis }) & = & \text{yellow} \\ G(\text{ football }) & = & \text{brown} \\ G(\text{ baseball }) & = & \text{white} \end{array}} \ \longrightarrow \ \boxed{\begin{array}{lcl} H(\text{ Tom }) & = & ? \\ H(\text{ Ned }) & = & ? \\ H(\text{ Ben }) & = & ? \\ H(\text{ Ken }) & = & ? \\ H(\text{ Sid }) & = & ? \end{array}}$$

We ignore any alignment between the two lists. We take the first entry in the second list,

$$G(\text{ basketball }) = \text{ orange},$$

and replace "basketball", according to the first entry of the first list, with $F(\text{ Tom })$. This is the result:

$$G\Big(\ F(\text{ Tom })\ \Big) = \text{ orange}.$$

Therefore,

$$H(\text{ Tom }) = \text{ orange}.$$

This is the first entry in the new list.

Exercise 2.4.4

Finish the list.

Once again, algebraically, composition is nothing but *substitution*!

Unfortunately, the domains of numerical functions are often infinite, and it is impossible to just follow the arrows. We deal with formulas. Fortunately, substitution is just as simple for numerical functions.

Example 2.4.5: composition of numerical functions

This time, we substitute one formula into another. For example, consider these two functions:

$$\boxed{X \xrightarrow{x^2=y} Y \xrightarrow{y^3=z} Z}$$

Then, $y = x^2$ is substituted into $z = y^3$ resulting in:

$$z = \left(x^2\right)^3 .$$

It is the same function, and it is computed by the same *two* steps! A simplification might make the extra work worthwhile:

$$z = x^6 .$$

Example 2.4.6: substitution of numerical functions

The idea of how the substitution is executed is the same as in before: Insert the input value in all of these boxes. Suppose this function on the left is understood and evaluated via the diagram on the right:

$$f\left(y\right) = \frac{2y^2 - 3y + 7}{y^3 + 2y + 1}, \quad f\left(\square\right) = \frac{2\square^2 - 3\square + 7}{\square^3 + 2\square + 1} .$$

Previously, we did the substitution $y = 3$ by inserting 3 in each of these windows:

$$f\left(\boxed{3}\right) = \frac{2\boxed{3}^2 - 3\boxed{3} + 7}{\boxed{3}^3 + 2\boxed{3} + 1} .$$

This time, let's insert $\sin x$, or, better, $(\sin x)$. This is the result of the substitution $y = \sin x$:

$$f\boxed{(\sin x)} = \frac{2\boxed{(\sin x)}^2 - 3\boxed{(\sin x)} + 7}{\boxed{(\sin x)}^3 + 2\boxed{(\sin x)} + 1} .$$

Then, we have

$$f\left(\sin x\right) = \frac{2(\sin x)^2 - 3(\sin x) + 7}{(\sin x)^3 + 2(\sin x) + 1} .$$

Note that if you don't know what the sine function (Chapter 4) does, it makes no difference! The only thing that matters is that we know that this is a *function*.

Example 2.4.7: composition from graphs

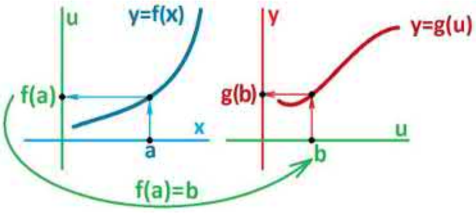
What if we have only their graphs? Suppose we have the two graphs of

$$u = f(x) \text{ and } y = g(u) ,$$

side by side, and we need to find the composition, $g \circ f$. Let's take a *single* value:

$$d = g(f(a)) .$$

Then, we use the first graph to find $c = f(a)$ on the vertical axis, travel all the way to the horizontal axis of the second, find the corresponding c on it, and finally find $d = g(c)$ on the vertical axis.



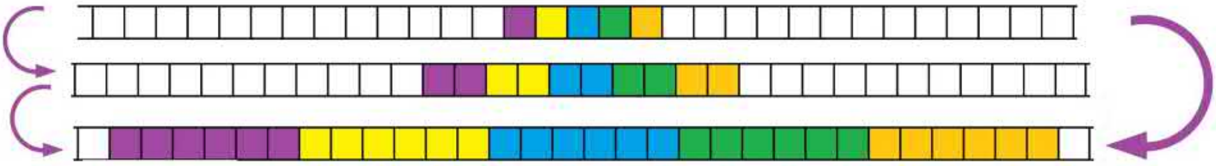
Graphs don't do a very good job of visualizing compositions...

Example 2.4.8: composition of transformations

What if we, instead, think of numerical functions as transformations? For example:

1. The first transformation is a stretch by a factor of 2.
2. The second transformation is a stretch by a factor of 3.

We visualize the first by doubling every square and the second by tripling:

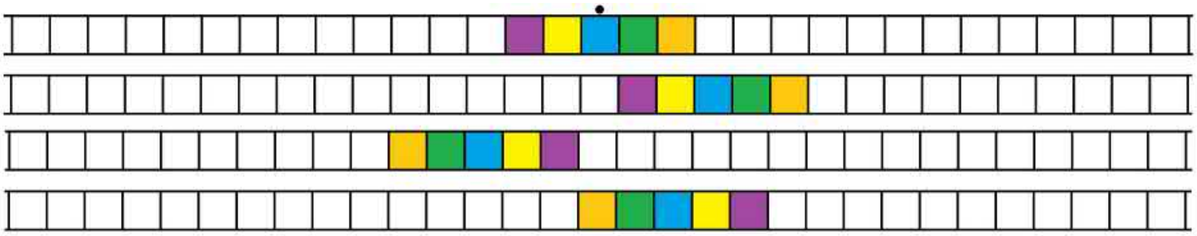


As a result there are *six* squares in the last row for each square in the first. Indeed, the composition of the two transformations is a stretch by a factor of $3 \cdot 2 = 6$.

So, we just carry out two transformations in a row. Consider this also:

1. The first transformation is a shift (right) by a 3.
2. The second transformation is a flip.
3. The third transformation is a shift (right) by a 5.

Then what their composition does is shown below:

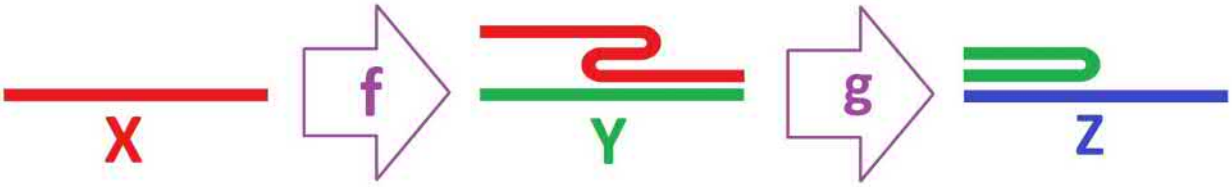


Exercise 2.4.9

Illustrate, as above, the composition of: a shift left by 5, a stretch by 2, and a flip.

Exercise 2.4.10

Illustrate the composition of the two transformations shown below:



If we think of functions as *lists of instructions* (with no forks), then each of them is already a composition! The steps on the list are the functions the composition of which creates the function. For example,

- F: Add 3. Multiply by -2. Subtract 1.

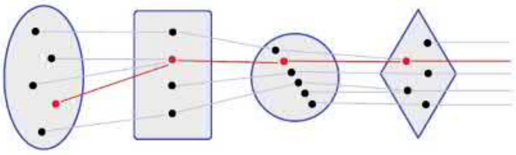
This is called a *decomposition* of F. If, furthermore, there is another function, say,

- G: Subtract 2. Apply sin.

Now, we just add the latter list to the bottom of the former:

- $G \circ F$: Add 3. Multiply by -2. Subtract 1. Subtract 2. Apply sin.

Of course, we can have compositions of many functions in a row as long the output of each function matches the input of the next:



It's as if the first function gives us direction to a destination, and at that destination, we receive the directions to our next destination where we get further directions, and so on... like a treasure hunt.

We represent functions as *black boxes* that process the input and produce the output:

$$\begin{array}{ccccccccccccc} \text{input} & & \text{function} & & \text{output} & & & & & & & & \\ x & \to & \boxed{f} & \to & y & & & & & & & & \\ & & & & \text{input} & & \text{function} & & \text{output} & & & & \\ & & & & y & \to & \boxed{g} & \to & z & & & & \\ & & & & & & & & \text{input} & & \text{function} & & \text{output} \\ & & & & & & & & z & \to & \boxed{h} & \to & u \end{array}$$

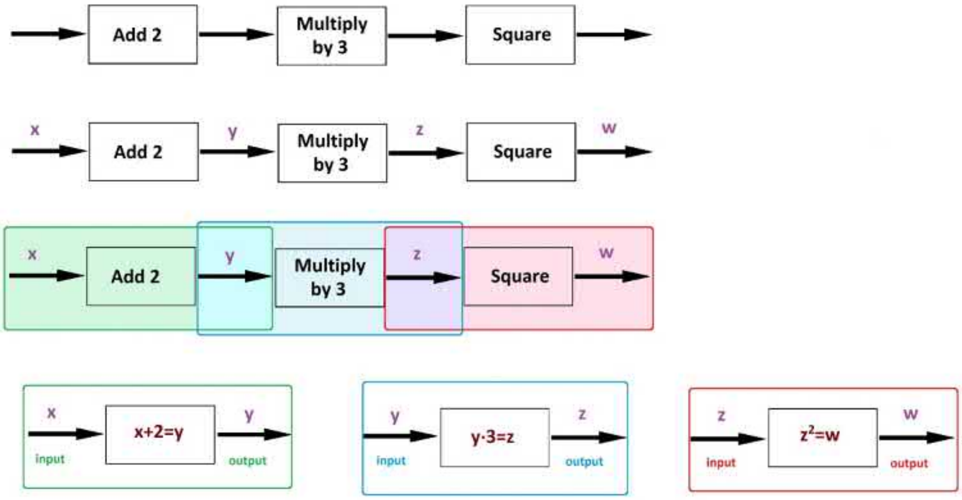
Because of the match, we can carry over the output to the next line – as the input of the next function.

This *chain of events* can be as long as we like:

$$x_1 \to \boxed{f_1} \to x_2 \to \boxed{f_2} \to x_3 \to \boxed{f_3} \to x_4 \to \boxed{f_4} \to x_5 \to \ldots$$

Example 2.4.11: decomposition of flowchart

Such a decomposition will allow us to study the function one piece at a time:



Exercise 2.4.12

Represent the following function by a single formula:

$$x \to \boxed{\text{multiply by } 2} \to y \to \boxed{\text{add } 5} \to z \to \boxed{\text{divide by } 3} \to u$$

Exercise 2.4.13

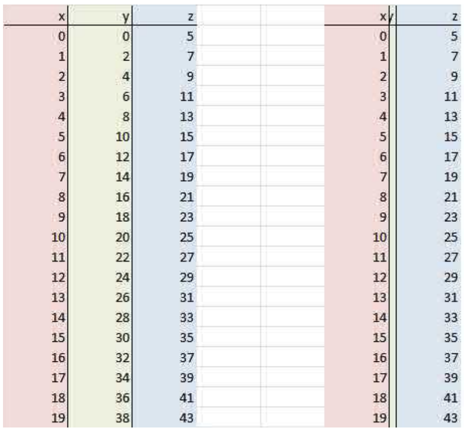
Represent the function $h(x) = (x-1)^2 + (x-1)^3$ as the composition $g \circ f$ of two functions $y = f(x)$ and $z = g(y)$.

Example 2.4.14: composition with spreadsheet, formulas

This is how the composition of several functions represented by *formulas* is computed with a spreadsheet. We start with just a list of numbers in the first column. Then we produce the values in the next column one row at a time. How? We input a formula in the next column with a reference to the last one. For example, we have in the second and third columns, respectively:

`=RC[-1]*2` `=RC[-1]+5`

referring to the cell to its left. Each of the two consecutive columns is a list of values of a function (left):



x	y	z
0	0	5
1	2	7
2	4	9
3	6	11
4	8	13
5	10	15
6	12	17
7	14	19
8	16	21
9	18	23
10	20	25
11	22	27
12	24	29
13	26	31
14	28	33
15	30	35
16	32	37
17	34	39
18	36	41
19	38	43

x	z
0	5
1	7
2	9
3	11
4	13
5	15
6	17
7	19
8	21
9	23
10	25
11	27
12	29
13	31
14	33
15	35
16	37
17	39
18	41
19	43

If we hide the middle column (right), we have the list of values of the composition. We can have as many intermediate columns as we like.

Example 2.4.15: functions given by lists, cross-referencing

What if the functions are given by nothing but their lists of values? Then we need to find a match for the output of the first function among the inputs of the second. Given the tables of values of f, g, find the table of values of $g \circ f$:

x	$y = f(x)$
0	1
1	0
2	2
3	4
4	2

followed by

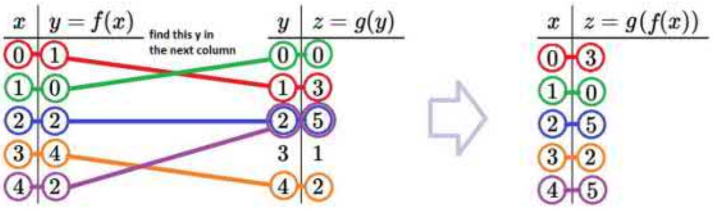
y	$z = g(y)$
0	0
1	3
2	5
3	1
4	2

is

x	$z = g(f(x))$
0	?
1	?
2	?
3	?
4	?

.

We need to fill the second column of the last table. First, we need to match the outputs of f with the inputs of g, as follows:



Alternatively, we re-arrange the rows of g according to the values of y and then remove the y-columns:

x	y	y	z
0	1	1	3
1	0	0	0
2	2	2	5
3	4	4	2
4	2	2	5

$\longrightarrow$

x	z
0	3
1	0
2	5
3	2
4	5

Exercise 2.4.16

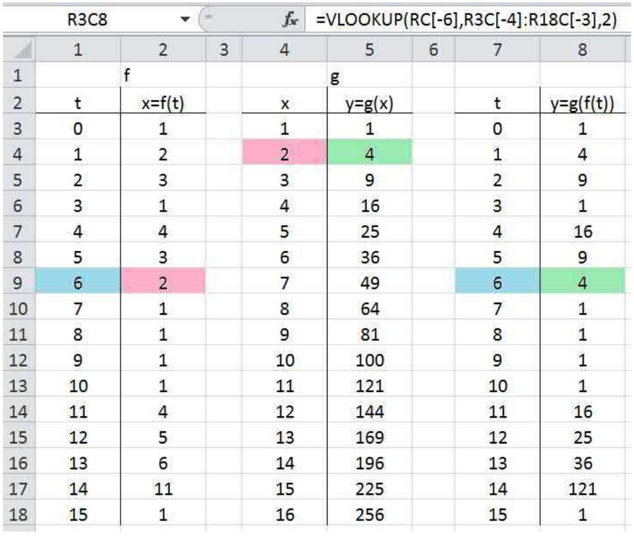
Given the tables of values of f, g below, find the table of values of $g \circ f$:

x	$y = f(x)$
0	0
1	2
2	3
3	0
4	1

y	$z = g(y)$
0	4
1	4
2	0
3	1

Example 2.4.17: composition with spreadsheet, cross-referencing

For this task, we have to use the search feature of the spreadsheet:



R3C8 =VLOOKUP(RC[-6],R3C[-4]:R18C[-3],2)

	1	2	3	4	5	6	7	8
1		f			g			
2	t	x=f(t)		x	y=g(x)		t	y=g(f(t))
3	0	1		1	1		0	1
4	1	2		2	4		1	4
5	2	3		3	9		2	9
6	3	1		4	16		3	1
7	4	4		5	25		4	16
8	5	3		6	36		5	9
9	6	2		7	49		6	4
10	7	1		8	64		7	1
11	8	1		9	81		8	1
12	9	1		10	100		9	1
13	10	1		11	121		10	1
14	11	4		12	144		11	16
15	12	5		13	169		12	25
16	13	6		14	196		13	36
17	14	11		15	225		14	121
18	15	1		16	256		15	1

For example, the search may be executed with a "look-up" function:

`=VLOOKUP(RC[-6],R3C[-4]:R18C[-3],2)`

Exercise 2.4.18

Functions $y = f(x)$ and $u = g(y)$ are given below by tables of some of their values. Present the composition $u = h(x)$ of these functions by a similar table:

x	0	1	2	3	4
$y = f(x)$	1	1	2	0	2

y	0	1	2	3	4
$u = g(y)$	3	1	2	1	0

Exercise 2.4.19

Function $y = f(x)$ is given below by a list of its values. Is the function one-to-one?

x	0	1	2	3	4
$y = f(x)$	0	1	2	1	2

Functions can be visualized as *flowcharts* and so can their compositions:

If we name the variables and use the algebraic notation, we produce a more compact version of this flowchart:

$$x \to \boxed{x+3} \to y \to \boxed{y\cdot 2} \to z \to \boxed{z^2} \to u$$

Note how the names of the variables match so that we can proceed to the next step. A purely algebraic representation of the diagram is below:

$$x+3=y, \quad y\cdot 2 = z, \quad z^2 = u\,.$$

It is also possible, but not required, to *name the functions*, say f, g, h. Then we have:

$$y = f(x) = x+3, \quad z = g(y) = y\cdot 2, \quad u = h(z) = z^2\,.$$

As we see, with the variables properly named,

composition is substitution.

In the above composition, we can carry out these two substitutions:

- We substitute $z = g(y) = y\cdot 2$ into $u = h(z) = z^2$, which results in the following:
$$u = h(z) = h(g(y)), \quad u = z^2 = (y\cdot 2)^2\,.$$
- We substitute $y = f(x) = x+3$ into $z = g(y) = y\cdot 2$, which results in the following:
$$z = g(y) = g(f(x)), \quad z = y\cdot 2 = (x+3)\cdot 2\,.$$

In general, we represent a function f diagrammatically as a *black box* that processes the input and produces the output:

$$\begin{array}{ccccc} \text{input} & & \text{function} & & \text{output} \\ x & \to & \boxed{f} & \to & y \end{array}$$

Now, suppose we have another function g:

$$\begin{array}{ccccc} \text{input} & & \text{function} & & \text{output} \\ x & \to & \boxed{g} & \to & y \end{array}$$

How do we represent their composition $g\circ f$? To represent it as a single function, we need to "wire" their diagrams together *consecutively* (instead of in parallel, as in the last section):

$$x \to \boxed{f} \to y \to ??? \to x \to \boxed{g} \to y$$

But it's only possible when the output of f matches with the input of g. We can *rename the variable* of g. For example, we can make this switch:

$$\frac{x^2-1}{x+2} \to \frac{y^2-1}{y+2}\,.$$

Warning!

If the names of the variables don't match, it might be for a good reason.

This is what we have after renaming:

$$x \to \boxed{\ f\ } \to y \to \boxed{\ g\ } \to z$$

Then we have a new diagram for a new function:

$$g \circ f:\ x \to \boxed{x \to \boxed{\ f\ } \to y \to \boxed{\ g\ } \to z} \to z$$

It's just another black box:

$$x \to \boxed{g \circ f} \to z$$

Compositions are meant to represent tasks that cannot be carried out in parallel. Imagine that you have two persons working for you, but you can't split the work in half to have them work on it at the same time because the second task cannot be started until the first is finished.

Example 2.4.20: order matters

For example, you are making a *chair*. The last two stages are polishing and painting. You can't do them at the same time:

$$\text{chair} \to \boxed{\text{polishing}} \to \boxed{\text{painting}} \to \text{finished chair}$$

You can't change the order either!

Example 2.4.21: gas mileage

Suppose a car is driven at 60 mi/h. Suppose we also know that the car uses 30 mi/gal, while the cost per gallon is \$5. Represent the expense as a function of time.

Consider the flowchart:

$$\begin{array}{ccccccc} \text{time (h)} & \xrightarrow{60\text{ mi/h}} & \text{distance (mi)} & \xrightarrow{30\text{ mi/gal}} & \text{gas used (gal)} & \xrightarrow{5\$/\text{ gal}} & \text{expense (\$)} \\ t & \xrightarrow{f} & d & \xrightarrow{k} & g & \xrightarrow{h} & e \end{array}$$

These are the participating functions:

$$60t = d \qquad \frac{d}{30} = g \qquad 5g = e$$

To answer the question, we substitute from right to left:

$$e = 5g = 5\left(\frac{d}{30}\right) = 5\left(\frac{60t}{30}\right).$$

Simplified:

$$e = 10t\,.$$

Exercise 2.4.22

Redo the example using the functions f, k, h.

2.5. *The simplest functions*

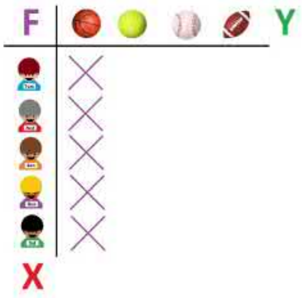
In this chapter, we will study many specific functions as well as some broad categories of functions. We start with the former.

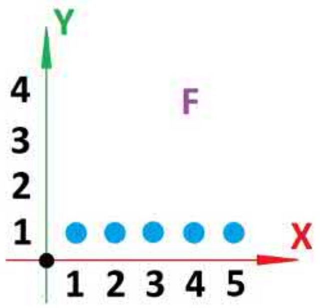
Even in the most general situation – nothing but sets – there are always two functions that are very simple.

Let's turn to the example of the two sets we considered earlier:

- X is the five boys; and
- Y is the four balls.

Now, what if *all* boys prefer basketball? Then our "preference function", F, cannot be simpler: All of its values are equal and all the arrows point to the basketball:





The table of this function F is also very simple: All crosses are in the same column. The graph is just as simple: All dots are on the same horizontal line.

The value of $y = F(x)$ doesn't vary as x varies; it is *constant*. The following concept will be routinely used.

Definition 2.5.1: constant function

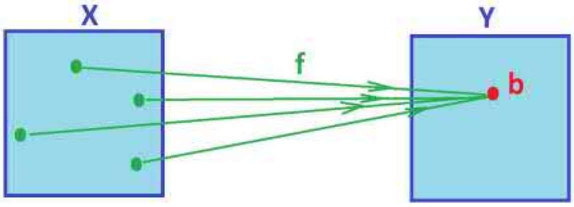
Suppose sets X and Y are given. A function $f : X \to Y$ is called a *constant function* if, for some specified element b of Y, we set:

$$\boxed{f(x) = b \quad \texttt{FOR EACH } x}$$

The process is identical for every input:

$$x \to \boxed{\text{ choose 3 }} \to y$$

In an illustration of a transformation of a plane to a plane, all arrows converge on a single output:



Exercise 2.5.2

What is the image of a constant function? What are the preimages?

What can we say about *compositions* of this special function, f, with another, g?

First, consider this diagram:

$$\begin{tikzcd}
X \arrow[r, "f"] & Y \arrow[r, "g"] & Z \\
x_1 \arrow[rd] & y_1 \arrow[r] & z_1 \\
x_2 \arrow[r] & y_2 \arrow[r] & z_2 \\
x_3 \arrow[ru] & y_3 \arrow[r] & z_3
\end{tikzcd}$$

What kind of function is $g \circ f$?

And here is another one:

$$\begin{tikzcd}
Z \arrow[r, "g"] & X \arrow[r, "f"] & Y \\
z_1 \arrow[r] & x_1 \arrow[rd] & y_1 \\
z_2 \arrow[r] & x_2 \arrow[r] & y_2 \\
z_3 \arrow[r] & x_3 \arrow[ru] & y_3
\end{tikzcd}$$

What kind of function is $f \circ g$?

By following the arrows from the left all the way to the right, we find that there is only one output. So, whether the other function comes after or before a constant function, the result is the same; this is our conclusion:

Theorem 2.5.3: Compositions with Constant Function

The composition of any function with a constant function is a constant function.

Proof.

This is the algebra for case 1:

$$x \mapsto f(x) = b \implies x \mapsto g(f(x)) = g(b)\,.$$

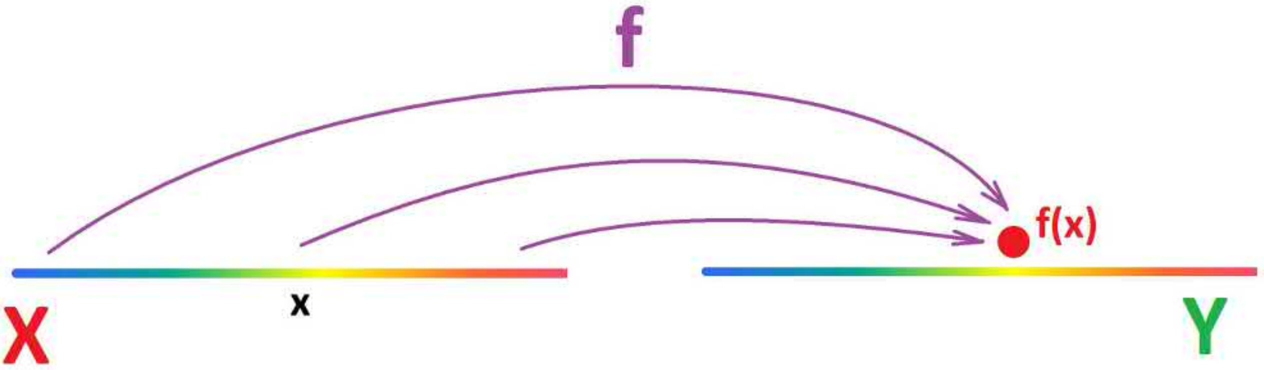
This is the algebra for case 2:

$$y \mapsto g(y) = x \implies y \mapsto f(g(y)) = b\,.$$

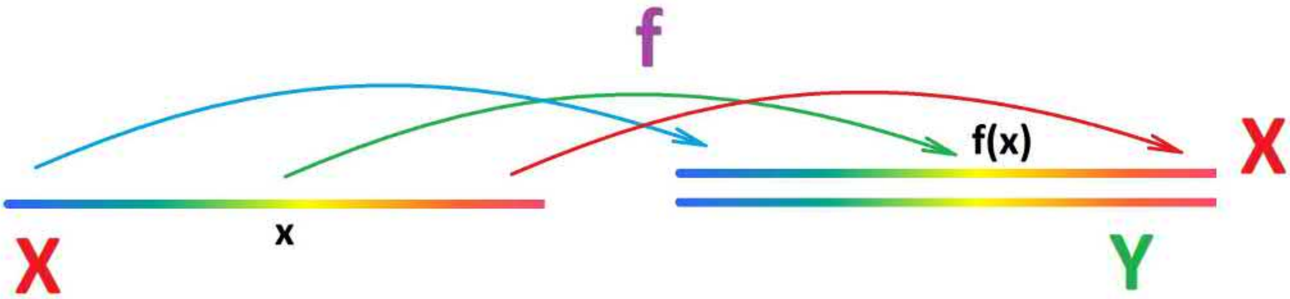
Exercise 2.5.4

Give an example of two non-constant functions the composition of which is constant.

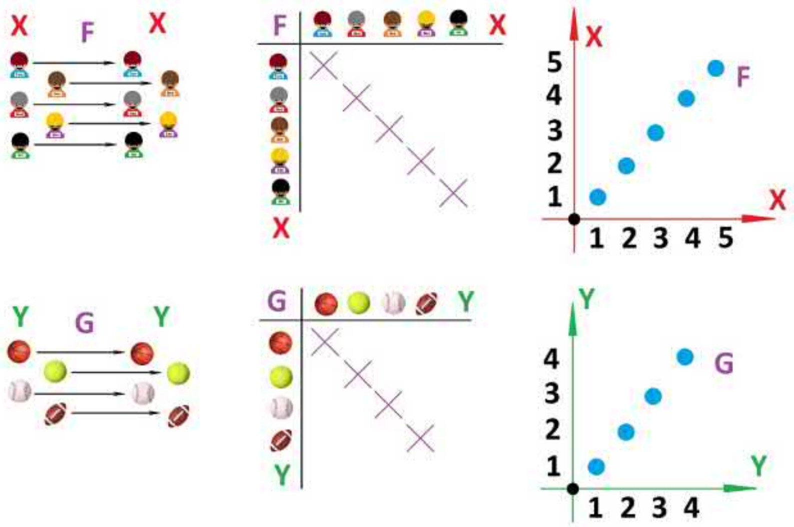
As a *transformation* of the line, the constant function is extreme; it "crushes" (or collapses) the whole line into a single point:



At the other end of the spectrum is another extreme transformation; it "does nothing" to the line:



So, for our set X of boys, we have a special function G from X to X (and another from Y to Y for the balls); each arrow comes back to the boy (or ball) it starts from:



The table of this function G is also very simple: All crosses are on the diagonal, and the graph has all dots on the diagonal. The output of G is *identical* to the input.

Definition 2.5.5: identity function

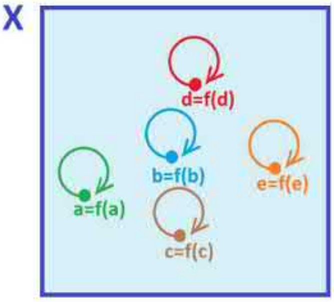
Suppose one set X is given. The *identity function* $I : X \to X$ is given by the following:

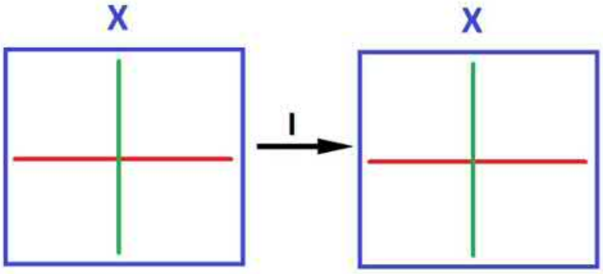
$$\boxed{I(x) = x \quad \texttt{FOR EACH} \ x}$$

The process is identical for every input:

$$x \ \to \ \boxed{\text{pass it}} \ \to \ y$$

In an illustration of a transformation of a plane to a plane, every arrow circles back to its input:





Exercise 2.5.6

What is the range of the identity function?

Exercise 2.5.7

What is the inverse of the identity function?

What can we say about the compositions of this special function, I, with another, g?

What can we say about *compositions* of this special function, f, with another, g? First, consider this diagram:

$$\begin{array}{ccccc} X & \xrightarrow{I} & X & \xrightarrow{g} & Y \\ \hline x_1 & \to & x_1 & \to & y_1 \\ x_2 & \to & x_2 & \to & y_2 \\ x_3 & \to & x_3 & \to & y_3 \end{array}$$

What kind of function is $g \circ I$?

And here is another one:

$$\begin{array}{ccccc} X & \xrightarrow{g} & Y & \xrightarrow{I} & Y \\ \hline x_1 & \to & y_1 & \to & y_1 \\ x_2 & \to & y_2 & \to & y_2 \\ x_3 & \to & y_3 & \to & y_3 \end{array}$$

What kind of function is $I \circ g$?

By following the arrows from the left all the way to the right, we find that the output always matches that of g. Whether the other function comes before or after the identity function, the result is the same; this is our conclusion:

Theorem 2.5.8: Composition with Identity Function

A composition of any function with the identity function is that function, i.e.,

$$I \circ g = g \text{ and } g \circ I = g\,.$$

Proof.

Consider this diagram:

$$x \mapsto y = I(x) = x \implies x \mapsto g(y) = g(I(x)) = g(x)\,.$$

The output is the same as the input! Here is the other one:

$$y \mapsto x = g(y) \implies y \mapsto I(x) = I(g(y)) = g(y)\,.$$

Again, the output is the same as the input!

Numbers can be represented in a number of ways, but sometimes they are *identical*:

$$1 + 1 = 2\,.$$

Similarly, functions can be represented in a number of ways, but sometimes they are *identical*:

$$x + x = 2x\,.$$

Let's make it clear what we mean when we say that *two functions are equal* (or identical):

- ▶ Two functions are equal when each possible input produces the same output for either function.

Definition 2.5.9: equal functions

Suppose f and g are two functions:

$$f, g : X \to Y\,.$$

They are *equal*,

$$\boxed{f = g}$$

if their domains are equal (as sets) and we have the following:

$$\boxed{f(x) = g(x) \quad \texttt{FOR EACH } x}$$

It is illustrated in the flowchart below:

$$\begin{array}{cccccccccc} & & & \nearrow & x & \to & \boxed{f} & \to & y & \\ x & \to & x & & & & & & & \text{same!} \\ & & & \searrow & x & \to & \boxed{g} & \to & u & \end{array}$$

Example 2.5.10: equal functions

Consider these two functions:

$$f(x) = 2x + 4 \quad \text{and} \quad g(x) = 2(x + 2)\,.$$

No matter what x is, the outputs are the same. We conclude that they are equal: $f = g$. This is just as example of how we algebraically manipulate formulas.

What does it mean when we say that *two functions are not equal*? The opposite of equal: The outputs don't fully match. In other words, the answer is: A possible input produces two different outputs for the two functions.

Definition 2.5.11: not equal functions

Suppose f and g are two functions:

$$f, g : X \to Y\,.$$

They are *not equal*,

$$\boxed{f \neq g}$$

if their domains are unequal (as sets) or we have the following:

$$\boxed{f(x) \neq g(x) \quad \texttt{FOR SOME } x}$$

In the definition, we test each x: Do the values match? It is illustrated in the flowchart below:

$$\begin{array}{cccccccccc} & & & \nearrow & 3 & \to & \boxed{f} & \to & 7 & \\ 3 & \to & 3 & & & & & & & \text{different!} \\ & & & \searrow & 3 & \to & \boxed{g} & \to & 12 & \end{array}$$

As you can see, you only need to find a single value of x for which there is a mismatch.

This case also includes the situation when the two domains are unequal:

$$\begin{array}{cccccccccc} & & & \nearrow & 3 & \to & \boxed{f} & \to & 7 & \\ 3 & \to & 3 & & & & & & & \text{different!} \\ & & & \searrow & 3 & \to & \boxed{g} & \to & \text{“not in the domain!”} & \end{array}$$

Example 2.5.12: not equal functions

Consider these two functions:

$$f(x) = \frac{x^2}{x} \quad \text{and} \quad g(x) = x\,.$$

We conclude that they aren't equal: $f \neq g$. Why? Because f is undefined at $x = 0$, which is in the implied domain of g. Replacing f with g is an example of how *not* to do symbol manipulation!

Example 2.5.13: identities

This is a familiar identity:

$$(x + 1)^2 = x^2 + 2x + 1\,.$$

A more complex two variable identity is

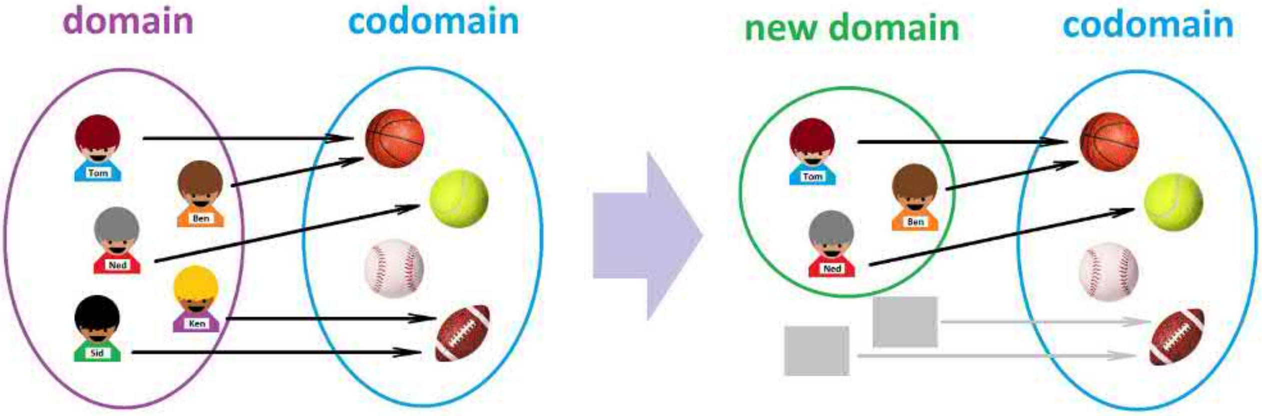
$$(x + y)^2 = x^2 + 2xy + y^2\,.$$

And so is

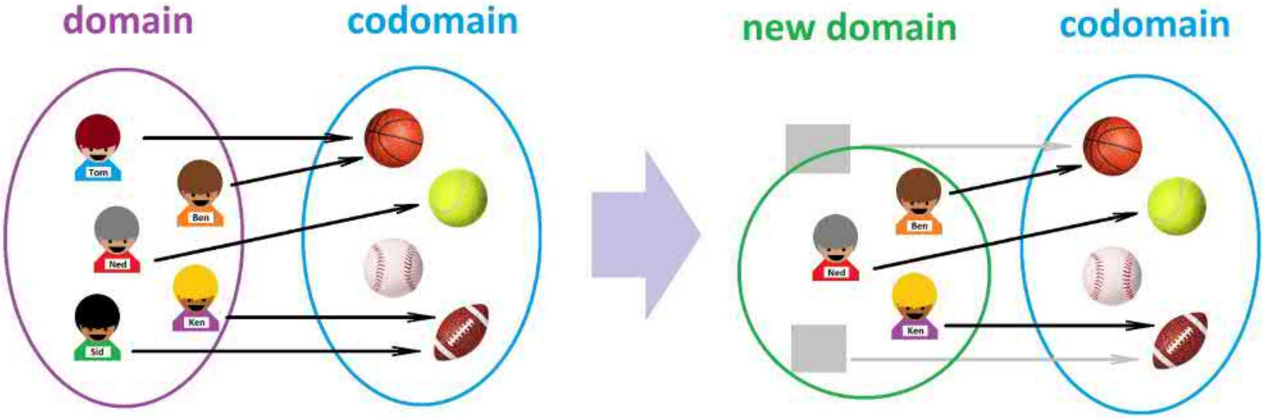
$$x^2 - y^2 = (x - y)(x + y)\,.$$

And so are all rules of exponents, logarithms, etc.

We have seen how "reducing" the domain of a function creates a new function:



Any subset of the domain can be chosen, but excluding Tom and Sid creates a function that is one-to-one:



Exercise 2.5.14

Change the codomain of the function to make it onto.

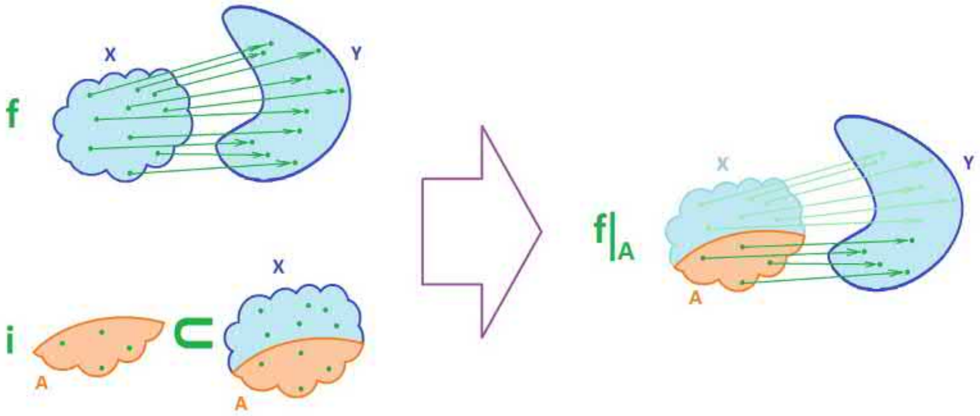
Definition 2.5.15: restriction of function

Suppose we have sets X and Y, a function $f : X \to Y$, and a subset A of X. Then the *restriction of f to A* is the function defined, and denoted, by the following:

$$\boxed{f\Big|_A(x) = f(x) \quad \texttt{FOR EACH } x}$$

in A.

The notation is reminiscent of the substitution notation. The construction is illustrated below:



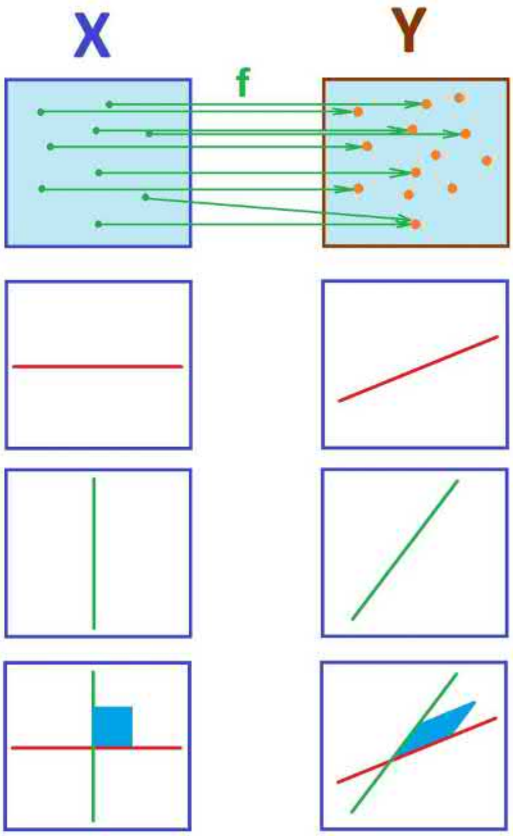
Warning!

We have $f\Big|_A \neq f$, unless $A = X$.

Here is an illustration of what restrictions can do for us. Below, we restrict the function to

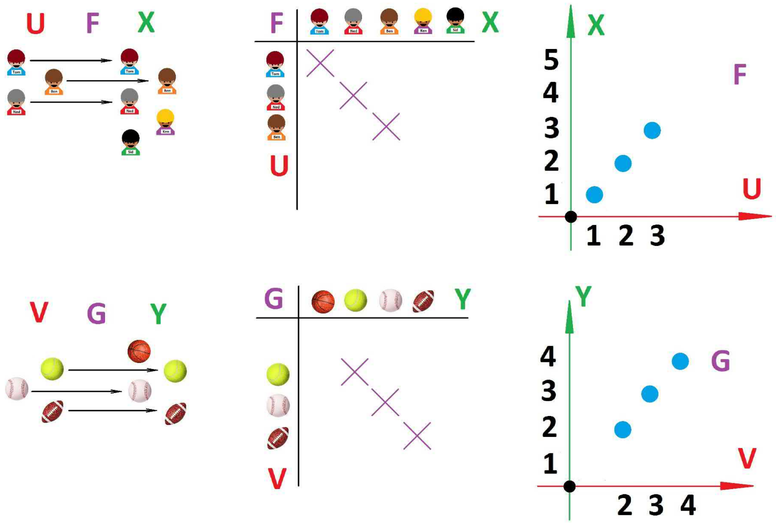
1. the x-axis,
2. the y-axis,
3. a little square in the first quadrant.

We then watch their images of these sets in Y (right column):



These images may reveal how the whole function operates.

The idea of the identity function applies even if the two sets – domain and codomain – don't match. It suffices that the former is a subset of the latter. We modify the above example below:



We see U *included* in X, and V *included* in Y. This fact creates a function.

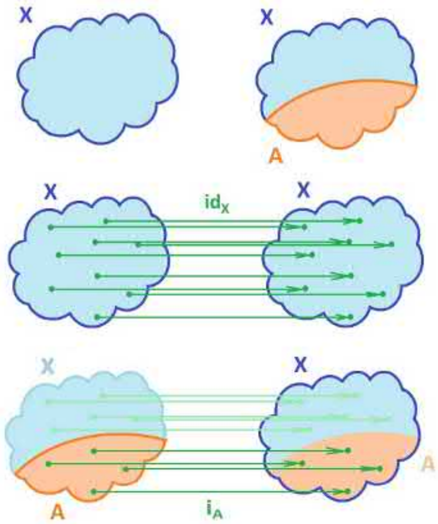
Definition 2.5.16: inclusion

Suppose we have a set X and a subset A of X. Then the *inclusion* $i : A \to X$ *of* A *into* X is the function defined by the following:

$$\boxed{i(x) = x \quad \texttt{FOR EACH } x}$$

in A.

Below, we make a copy of a set X with a subset A specified. We then construct the inclusion with identical, horizontal arrows:



Theorem 2.5.17: Restriction via Compositions

A restriction of a function is its composition with the appropriate inclusion; i.e., given a function $f : X \to Y$, a subset $A \subset X$, and its inclusion $i : A \to X$, the restriction of f to A is:

$$\boxed{f\Big|_A = f \circ i}$$

Exercise 2.5.18

Prove the theorem.

Exercise 2.5.19

Explain how inclusions are restrictions of the identity functions.

Exercise 2.5.20

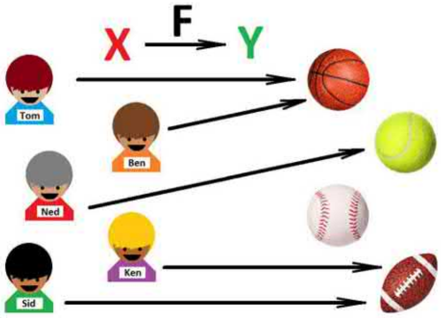
A restriction of a constant function is ___ .

2.6. *The inverse of a function*

Our running example of a function answers the question:

- *Which ball is this boy playing with?*

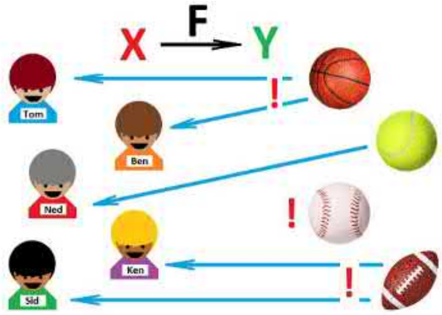
The arrow points from the boy to that ball:



Let's turn this around:

▶ *Which boy is playing with this ball?*

The solution would seem to be a simple reversal of the arrows so that the arrow should point from the ball to that boy:



We can see that, even though the latter question is asked about the same situation as the former, it cannot be answered in a positive manner! Indeed:

- There are two boys playing with the basketball – two answers.
- There is no boy playing with the baseball – no answer.

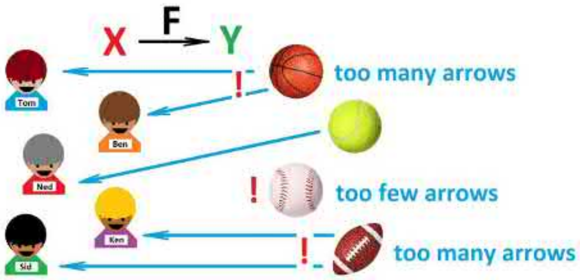
This means that there is no function this time!

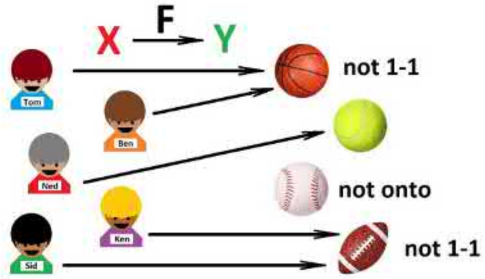
Question:

▶ *Under what circumstances would such a reversal of arrows make sense?*

All functions *from* X *to* Y are also relations *between* X and Y. However, not every relation is a function – either from X to Y or from Y to X. The reasons are the same: too many or too few arrows starting at an element of the domain set.

An especially important question is: Can we reverse the arrows of a function so that that the same relation is now seen as a new, "inverse", function? If the answer is No, can we see the reason by looking at the *original* function?

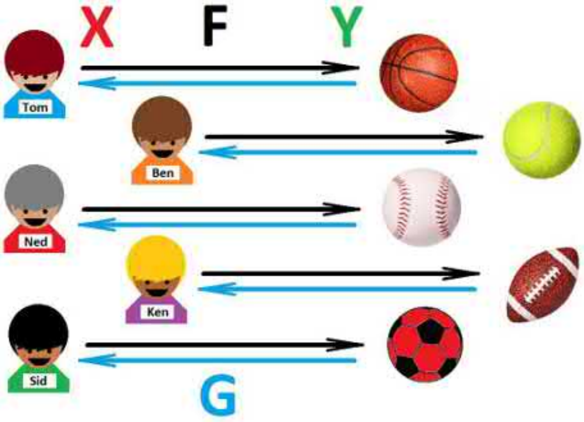




- First, some y's in Y have two or more corresponding x's in X. In other words, the function isn't one-to-one!
- Second, some y's in Y have no corresponding x's in X. In other words, the function isn't onto!

So, the original function lacks either of the two types of regularity for this to be possible.

What kind of function would make this possible? A function that is *both one-to-one and onto*:



We have added an extra ball (soccer) and have re-drawn the arrows; there is exactly one arrow for each ball. This is a *very simple*, almost uninteresting, kind of function: Each boy holds a single ball, and every ball is held by a single boy. The arrows have been safely reversed.

As a result, the elements of the two sets are all paired up:

boys	balls
Tom	basketball
Ben	tennis
Ned	baseball
Ken	football
Sid	soccer

In a sense, the only difference between the two sets is in the names. Indeed, if you don't remember Tom's name, you just say "the boy who plays with the basketball". Or, if you don't remember what that red ball is for, you just say "the game Sid plays". There is no ambiguity in this substitution.

The following concept is very important.

Definition 2.6.1: inverse of function

Suppose $F : X \to Y$ is a function. A function $G : Y \to X$ is called an *inverse function* of F when, for all x and y, we have the following match:

$$\boxed{F(x) = y \ \text{ if and only if } \ G(y) = x}$$

In other words, we have:

$$\begin{array}{ccc} x \xrightarrow{F} y & \iff \\ x \xleftarrow{G} y & \end{array}$$

Exercise 2.6.2

In the definition, interchange F and g, and x and y. What do you have?

We will rely on this important fact.

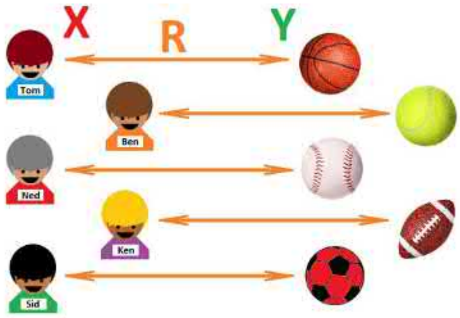
Theorem 2.6.3: One-to-one Onto vs. Inverse

A function is both one-to-one and onto if and only if it has an inverse.

Exercise 2.6.4

Prove the theorem.

Under this condition, we have a relation R that contains both of the functions:



As the function is defined indirectly, we need an assurance that there is only one.

Theorem 2.6.5: Uniqueness of Inverse

There can be only one inverse for a function.

Exercise 2.6.6

Prove the theorem.

This justifies using "the inverse" from now on. The inverse of a function F is denoted as follows:

Inverse function

$$F^{-1}$$

It reads "F inverse".

Here "F" is the *name* of the old function and "F^{-1}" is the *name* of the new function with a reference to the one it came from.

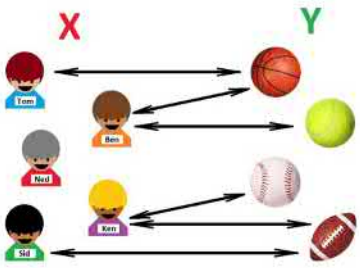
Warning!

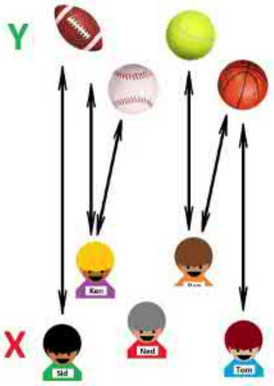
The notation is not to be confused with the power notation for the reciprocal: $2^{-1} = \frac{1}{2}$. (Warning inside a warning: the inverse of multiplication by 2 is division by 2.)

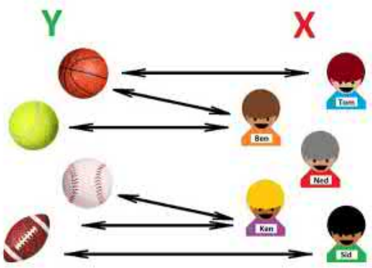
An idea to hold on to is that a function and its inverse represent the *same relation* between sets X and Y:

- x and y are related when $y = F(x)$, or
- x and y are related when $x = F^{-1}(y)$.

There is no preferred alignment:



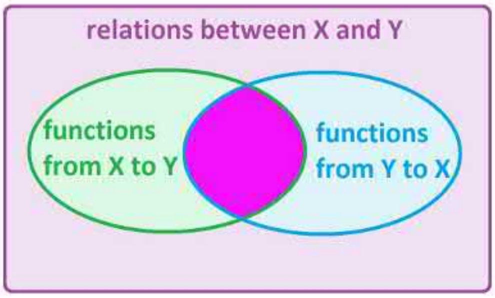




The choice between F and F^{-1} is the choice of the "roles" for X and Y, input or output, domain or codomain.

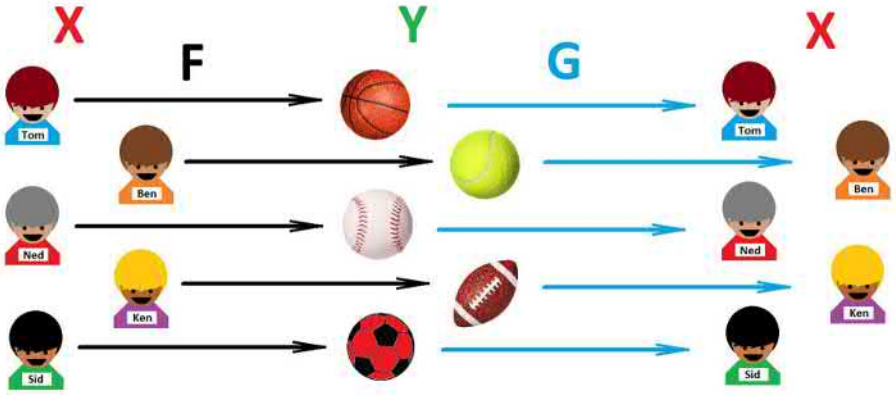
Exercise 2.6.7

Explain the picture below:



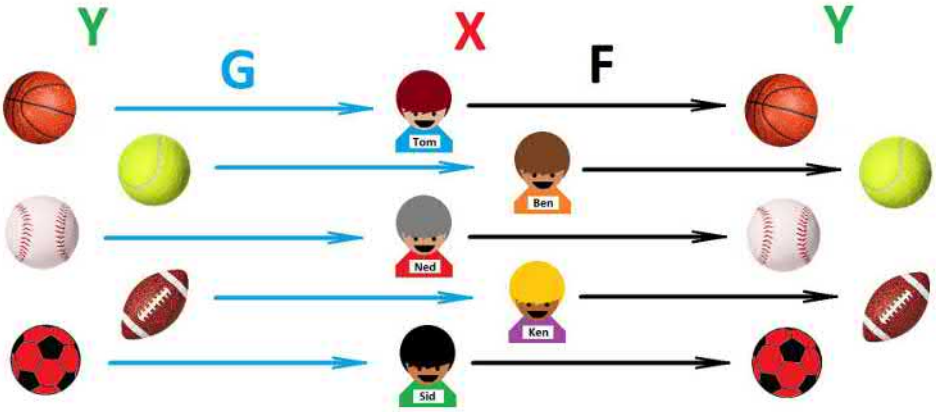
Compositions provide a different point of view on the definition of inverse.

Notice that the domain of the new function – the inverse – would have to be the codomain of the original! Their composition then makes sense:



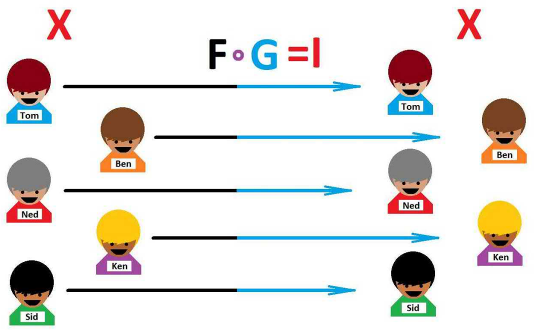
The first thing we notice about the composition is that we have made – through the two functions – a full circle from boys to balls and back to boys. Furthermore, with every two consecutive arrows, we arrive to our starting point, the exactly same boy, every time. This is the *identity function* of X.

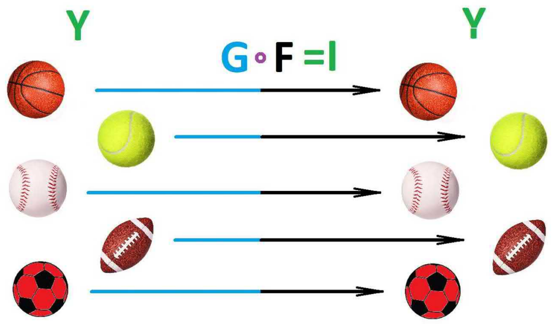
We also notice that the domain of the original function is the codomain of the inverse! Their composition – in the opposite order – then also makes sense:



Once again, we notice about the composition is that we have made – through the two functions – a full circle from balls to boys and back to balls. Furthermore, with every two consecutive arrows, we arrive to our starting point, the same ball, every time. This is the *identity function* of Y.

These are the two compositions next to each other:





Here is a flowchart representation of this idea:

$$\begin{array}{lllllllll} x & \to & \boxed{F} & \to\ y\ \to & \boxed{G} & \to & x & \text{Same } x! \\ y & \to & \boxed{G} & \to\ x\ \to & \boxed{F} & \to & y & \text{Same } y! \end{array}$$

We feed the output of F into G, and vice versa. The following is crucial.

Theorem 2.6.8: Inverse via Compositions

Suppose $F : X \to Y$ is a function that is both one-to-one and onto. Then a function $G : Y \to X$ is the inverse of F if and only if

- $G(F(x)) = x$ FOR EACH x, AND
- $F(G(y)) = y$ FOR EACH y.

Exercise 2.6.9

Prove the theorem.

The two identities in the theorem can also be written via the *identity functions*. We take two round trips:

$$\begin{tikzcd} \text{start} & x \arrow[r] \arrow[d, equal] & \boxed{f} \arrow[r] & y \arrow[d] \\ \text{finish} & x' & \boxed{f^{-1}} \arrow[l] & y \arrow[l] \end{tikzcd} \qquad \begin{tikzcd} x \arrow[r] & \boxed{f} \arrow[r] & y \arrow[d, equal] & \text{finish} \\ x \arrow[u] & \boxed{f^{-1}} \arrow[l] & y' \arrow[l] & \text{start} \end{tikzcd}$$

Both times we arrive where we started – with the same final output. We interpret this definition in terms of the identity function, as follows.

Theorem 2.6.10: Inverse via Compositions

Two functions $F : X \to Y$ and $F^{-1} : Y \to X$ are inverses of each other if and only if their compositions produce the identity functions; i.e., these two conditions are satisfied:

$$\boxed{f \circ F^{-1} = I \quad \text{AND} \quad F^{-1} \circ f = I}$$

Warning!

These are two different identity functions.

Warning!

There are *two* conditions here.

So, making circles in the diagram below won't change the value of x or y:

$$\begin{tikzcd} x \arrow[r, "F"] & y \arrow[d] \\ x \arrow[u] & y \arrow[l, "F^{-1}"] \end{tikzcd}$$

Definition 2.6.11: invertible function

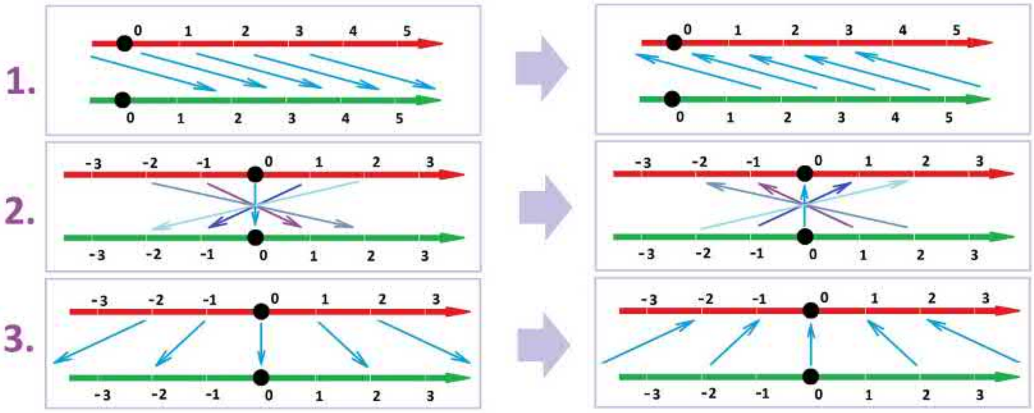
A function that is both one-to-one and onto is also called *invertible* (or a bijection).

Exercise 2.6.12

Suppose A, B, C are the sets of the one-to-one, the onto, and the invertible functions, respectively. What is the relation between these sets?

Example 2.6.13: inverse of transformation

What is the meaning of the inverse of a function when seen as a transformation of the line? It is a transformation that would *reverse* the effect of the original, as follows:



To produce the images on the right from those on the left, we simply flip all arrows.

Just by examining these simple transformations, we discover the following:

1. The inverse of the shift s units to the right is the shift of s units to the left.
2. The inverse of the flip is another flip.
3. The inverse of the stretch by $k \neq 0$ is the shrink by k (i.e., stretch by $1/k$).

In fact, we realize that they *pair up*:

1. The shift s units to the right and the shift of s units to the left are the inverses of each other.
2. The flip is the inverse with itself.
3. The stretch by $k > 0$ and the shrink by k are the inverses of each other.

When executed consecutively (in either order), the effect is nil.

Algebraically, we have these pairs of functions:

	$f = g^{-1}$	vs.	$g = f^{-1}$
1. shift	$y = x + s$		$x = y - s$
	add s		add $-s$
2. flip	$y = -x$		$x = -y$
	multiply by -1		multiply by -1
3. stretch	$y = x \cdot k$		$x = y/k$
	multiply by k		multiply by $\frac{1}{k}$

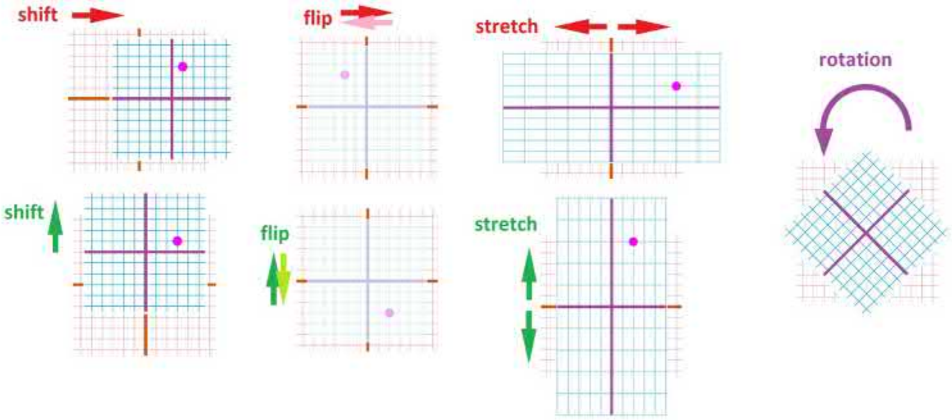
What about the fold? It can't be undone since any two points that are brought together become indistinguishable. Any further transformations will produce the same output:

$$\begin{tikzcd}
x=a \arrow[dr, "F"] & & \\
& y \arrow[r, "G?"] & x \\
x=b \arrow[ur, "F"'] & &
\end{tikzcd}$$

This function isn't one-to-one! And neither is the collapse.

Example 2.6.14: inverse of transformation

Now, the basic transformations of the plane:



We also have pairs of functions:

$f = g^{-1}$	vs.	$g = f^{-1}$
vertical shift		vertical shift back
vertical flip		vertical flip
vertical stretch		vertical shrink
horizontal shift		horizontal shift back
horizontal flip		horizontal flip
horizontal stretch		horizontal shrink
rotation		rotation back

We will examine these pairs later.

Example 2.6.15: inverse of list

How do we find the inverse of a function given by its *list of values*:

$$f = \begin{array}{|c|l|} x & y=f(x) \\ \hline 0 & 1 \\ 1 & 0 \\ 2 & 2 \\ 3 & 1 \\ 4 & 3 \\ \dots & \dots \end{array} = \begin{array}{|ccc|} x & \to & y \\ \hline 0 & \to & 1 \\ 1 & \to & 0 \\ 2 & \to & 2 \\ 3 & \to & 1 \\ 4 & \to & 3 \\ \dots & \dots & \dots \end{array}$$

To find the inverse, one can use the original table for look-up. For example, to find $f^{-1}(3)$, locate 3 in the second row and look at the entry to its left, 4. This may be a dangerous practice.

The table is understood as if there are arrows going horizontally left to right. That is why “reversing the arrows” means interchanging the columns:

$$f^{-1}=\begin{array}{|ccc|}x&\to&y\\\hline 0&\leftarrow&1\\ 1&\leftarrow&0\\ 2&\leftarrow&2\\ 3&\leftarrow&1\\ 4&\leftarrow&3\\ \ldots&\ldots&\ldots\end{array}=\begin{array}{|ccc|}y&\to&x\\\hline 1&\to&0\\ 0&\to&1\\ 2&\to&2\\ 1&\to&3\\ 3&\to&4\\ \ldots&\ldots&\ldots\end{array}=\begin{array}{|ccc|}y&\to&x\\\hline 0&\to&1\\ 1&\to&0\\ 1&\to&3\\ 2&\to&2\\ 3&\to&4\\ \ldots&\ldots&\ldots\end{array}=\begin{array}{|c|l|}y&x=f^{-1}(y)\\\hline 0&1\\ 1&0\\ 1&3\\ 2&2\\ 3&4\\ \ldots&\ldots\end{array}.$$

It may, or may not, become clear that the new function isn’t a function! To make sure, it’s a good idea, at the end, to arrange the inputs in the increasing order. Then we clearly see the conflict: $f^{-1}(1)=0$ and $f^{-1}(1)=3$. The original function, f, wasn’t one-to-one!

The general rule for *finding the inverse* of a function given by a formula follows from the definition:

- ▶ The inverse of $y=f(x)$ is found by solving this equation for x; i.e., $x=f^{-1}(y)$.

This method results in a success only when there is exactly one solution, x, for each y.

Example 2.6.16: inverse of linear polynomial

To find the inverse of a linear polynomial

$$f(x)=3x-7\,,$$

set and solve the equation (relation), as follows:

$$y=3x-7\implies y+7=3x\implies \frac{y+7}{3}=x\,.$$

Therefore, we have:

$$f^{-1}(y)=\frac{y+7}{3}\,.$$

If it is not known ahead of time whether the function is one-to-one, this fact is established, automatically, as a part of finding the inverse. For example, to find the inverse of the quadratic function

$$f(x)=x^2\,,$$

we set and solve:

$$y=x^2\implies \pm\sqrt{y}=x\,.$$

The ± sign indicates that there are two solutions ($x>0$). The original function wasn’t invertible!

A linear function,

$$f(x)=mx+b\,,$$

is one-to-one and onto whenever $m\neq 0$. Algebraically, we just solve the equation $y=mx+b$ for x. The algebraic result is below.

Theorem 2.6.17: Inverse of Linear Polynomial

The inverse of a linear polynomial

$$f(x)=mx+b,\ m\neq 0\,,$$

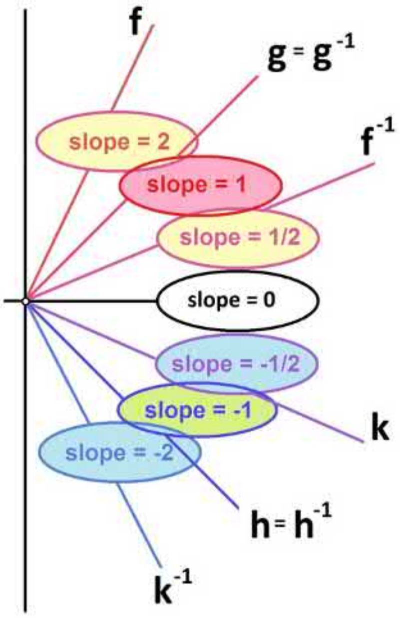
is also a linear polynomial, and its slope is the reciprocal of that of the original:

$$f^{-1}(y)=\frac{1}{m}y-\frac{b}{m}\,.$$

We know that the set of invertible functions is split into pairs of inverses. We can be more specific with the set of all linear polynomials. The pairs have reciprocal slopes, for example:

- 2 and 1/2
- -2 and $-1/2$
- 1 and 1
- -1 and -1
- etc.

We can see these pairs of a steeper line and a shallower line:



Exercise 2.6.18

What do need to do to this sheet of paper in order to make the former land on the latter?

Next, let's try to imagine how some *new algebraic operations* may have emerged.

Some emerged as the *abbreviations* for repeated familiar operations; for example, repeated addition, $2+2+2=2\cdot 3$, leads to a new operation: *multiplication*. Meanwhile, repeated multiplication, $2\cdot 2\cdot 2\cdot 2=2^4$, leads to a new operation: *exponent*. But what about subtraction and division?

Example 2.6.19: subtraction as inverse

Suppose I know how to add. Problem: With \$5 in my pocket, how much do I add to have \$12? Answer: \$7. How do I know? Solve the equation:

$$5+x=12\,.$$

This equation leads to a new operation, *subtraction*: $x=12-5$. Of course, there is also a new function. We can say that "subtraction is the inverse of addition", or more precisely, subtracting 5 is the inverse of adding 5.

Example 2.6.20: division as inverse

Suppose now I know how to multiply. Problem: If I want to make a table 20 inches wide, how many 2-by-4's do I need? Answer: 5. How do I know? Solve the equation:

$$4x=20\,.$$

This equation leads to a new operation, *division*: $x = \dfrac{20}{4}$. Division (by 4) is the inverse of multiplication (by 4).

Example 2.6.21: square root as inverse

Problem: If I want to make a square table with an area 25 square feet, what should be the width of the table? Solve the equation:

$$x \cdot x = 25 \implies x^2 = 25 \implies x = \sqrt{25}\,.$$

Thus, we have a new operation: *square root*. It is the inverse of the squaring function.

Example 2.6.22: cubic root as inverse

Problem: What is the side of a box if its volume is known to be 8 cubic feet? Solve the equation:

$$x^3 = 8 \implies x = \sqrt[3]{8}\,.$$

The cubic root is the inverse of the cubic power.

Thus, solving equations requires us to *undo* some function present in the equation:

$$\begin{array}{lrlll}
1. & x+2 = 5 & \implies (x+2)-2 = 5-2 & \implies & x = 3\\
2. & x\cdot 3 = 6 & \implies (x\cdot 3)/3 = 6/3 & \implies & x = 2\\
3. & x^2 = 4 & \implies \sqrt{x^2} = \sqrt{4} & \implies & x = 2 \ (x, y \ge 0)
\end{array}$$

We have "cancellation" on the left and simplification on the right.

We are dealing with functions! And some functions *undo the effect of others*:

1. The addition of 2 is undone by the subtraction of 2, and vice versa.
2. The multiplication by 3 is undone by the division by 3, and vice versa.
3. The second power is undone by the square root (for $x \ge 0$), and vice versa.

Each of these undoes the effect of its counterpart *under substitution*:

1. Substituting $y = x + 2$ into $x = y - 2$ gives us $x = x$.
2. Substituting $y = 3x$ into $x = \frac{1}{3}y$ gives us $x = x$.
3. Substituting $y = x^2$ into $x = \sqrt{y}$ gives us $x = x$, for $x, y \ge 0$.

And vice versa:

1. Substituting $x = y - 2$ into $y = x + 2$ gives us $y = y$.
2. Substituting $x = \frac{1}{3}y$ into $y = 3x$ gives us $y = y$.
3. Substituting $x = \sqrt{y}$ into $y = x^2$ gives us $y = y$, for $x, y \ge 0$.

Warning!

Both cancellations matter.

As we know, it is more precise to say that they undo each other *under composition*: Two numerical functions $y = f(x)$ and $x = g(y)$ are inverse of each other when for every x in the domain of f and for every y in the domain of g, we have:

$$g(f(x)) = x \quad \text{and} \quad f(g(y)) = y\,.$$

This is an alternative way of writing these compositions:

Inverses in substitution notation

$$g(y)\Bigg|_{y=f(x)}=x\quad\text{and}\quad f(x)\Bigg|_{x=g(y)}=y$$

Thus, we have three pairs of inverse functions:

$$\begin{array}{rcll}
f(x)=x+2 & \text{vs.} & f^{-1}(y)=y-2 & \\
f(x)=3x & \text{vs.} & f^{-1}(y)=\frac{1}{3}y & \\
f(x)=x^2 & \text{vs.} & f^{-1}(y)=\sqrt{y} & \text{for } x,y\geq 0
\end{array}$$

Next, it is reasonable to ask: What is the relation between the *graph* of a function and the graph of its inverse?

In other words, what do we need to do with the graph of f to get the graph f^{-1}? The answer is: Hardly anything. After all, *a function and its inverse represent the same relation*.

The graph of f illustrates how y depends on x – as well as how x depends on y. And the latter is what determines f^{-1}! So, there is no need for a new graph; the graph of f^{-1} *is* the graph of f. The only issue is that the x- and the y-axis point in the wrong directions. It's an easy fix.

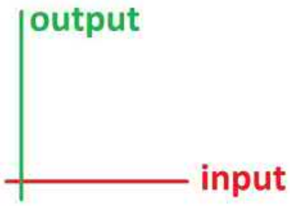
Example 2.6.23: points on the graph of inverse

Suppose we are transitioning from f to its inverse f^{-1}:

$$f=\left|\begin{array}{l|l}
x & y=f(x)\\ \hline
2 & 5\\
3 & 1\\
8 & 7\\
... &
\end{array}\right|\quad\Longrightarrow\quad f^{-1}=\left|\begin{array}{l|l}
y & x=f^{-1}(y)\\ \hline
5 & 2\\
1 & 3\\
7 & 8\\
... &
\end{array}\right|$$

These are the same pairs! Therefore, they are represented by the same points on the xy-plane.

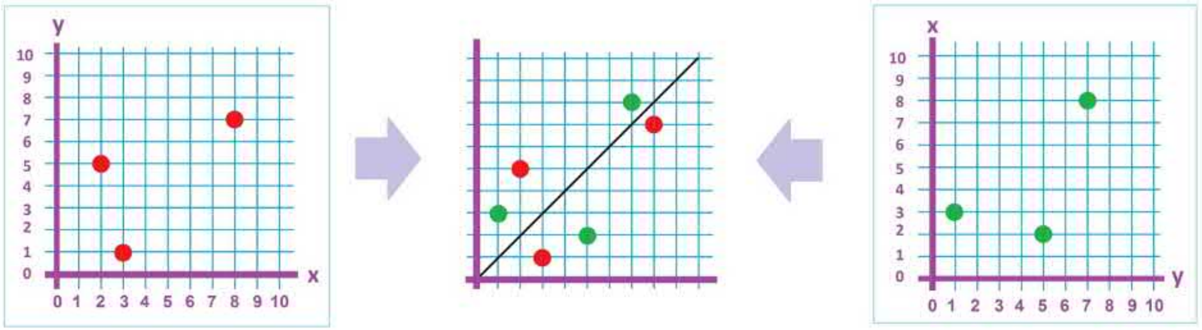
It is common, however, to put the input variable in the horizontal axis and the output in the vertical.



This makes us replace the points in the xy-plane with new points in the yx-plane:

$$\begin{array}{ccc}
(x,y) & \longrightarrow & (y,x)\\ \hline
(2,5) & \longrightarrow & (5,2)\\
(3,1) & \longrightarrow & (1,3)\\
(8,7) & \longrightarrow & (7,8)\\
... & & ...
\end{array}$$

The two coordinates are interchanged:



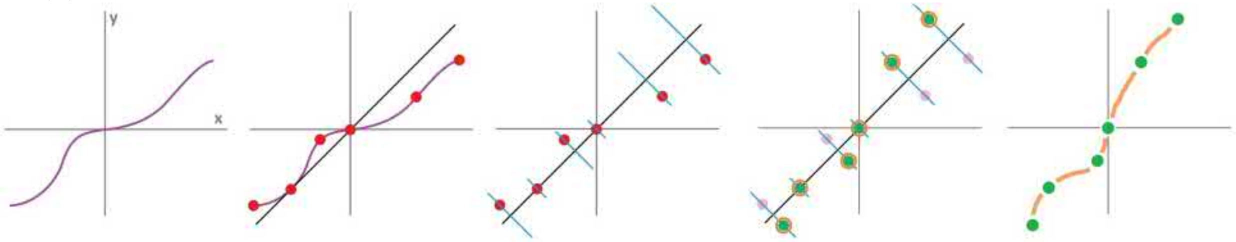
We realize that each point jumps across the diagonal line $y = x$! So, we have a match:

- ▶ Every point (x, y) in the xy-plane corresponds to the point (y, x) in the yx-plane.

Above we made a copy of the graph of f, flipped it, and then on top of the original.

Example 2.6.24: inverse graph point by point

Suppose, again, a function is given only by its graph and we need to construct the graph of the inverse $x = f^{-1}(y)$. This time we are to do this without any data:



Start with choosing a few points on the graph. Each of them will jump across the diagonal under this flip. How exactly? The general rule for plotting a counterpart of a point is the following:

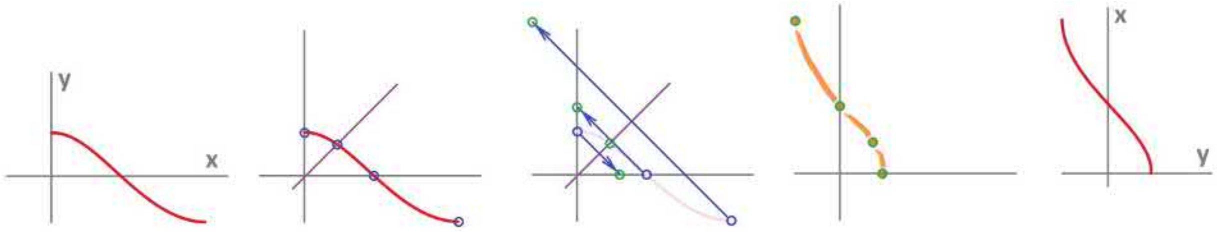
- ▶ From the point go perpendicular to the diagonal and then measure the same distance on the other side.

In other words, we plot a line through our point with slope -1.

Now, we can simplify our job by choosing the points more judiciously; we choose ones with easy-to-find counterparts. First, points on the diagonal don't move by the flip about the diagonal. Second, points on one of the axes jump to the other axis with no need for measuring. Finally, once all points are in place, finally, draw a curve that connects them.

Example 2.6.25: graph of inverse point by point

Plot the graph of the inverse of $y = f(x)$ shown below:

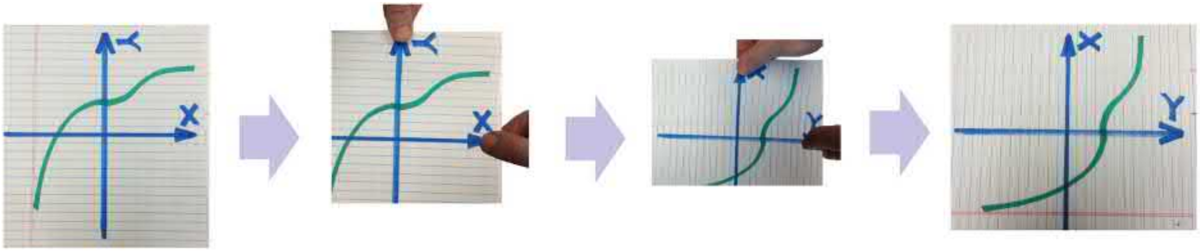


These are the steps:

- Draw the diagonal $y = x$.
- Pick a few points on the graph of f (we choose four).
- Plot a corresponding point for each of them:
  - on the line through point A that is perpendicular to the diagonal (i.e., its slope is 45 degrees down)
  - on the other side of the diagonal from A
  - at the same distance from the diagonal as A
- Draw by hand a curve from point to point.

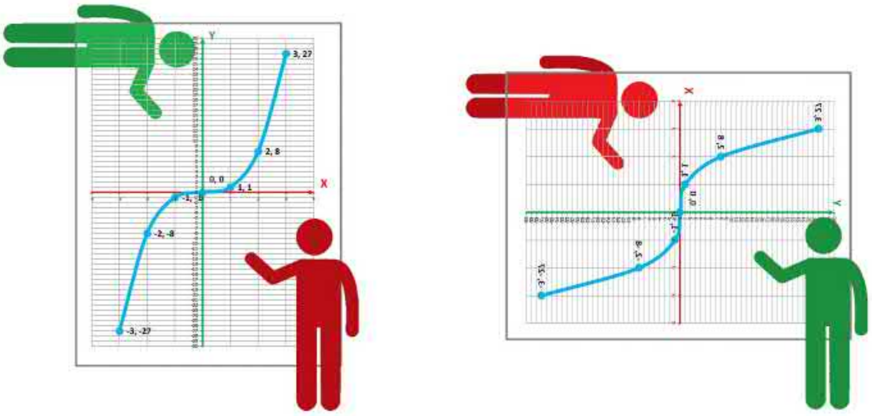
Example 2.6.26: flip graph

The following approach works *without a pen.* If we have a piece of paper with the xy-axis and the graph of $y = f(x)$ on it, we flip it by grabbing the end of the x-axis with the right hand and grabbing the end of the y-axis with the left hand then interchanging them:



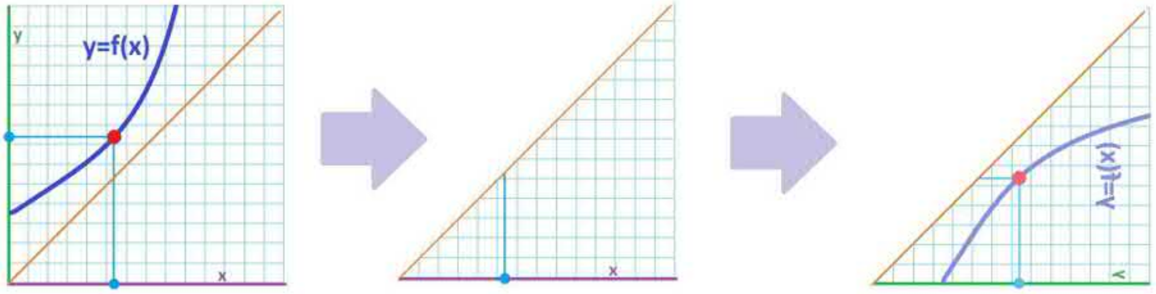
We face the *opposite side* of the sheet then, but the graph is still visible: the x-axis is now pointing up and the y-axis right, as intended. A transparent sheet of plastic would work even better.

Another way is to recognize that this is the same 3d scene seen from two different points of view:

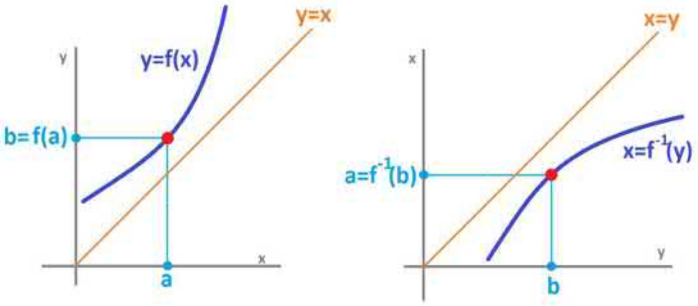


Example 2.6.27: fold graph

Alternatively, we can also fold:



The shapes of the graphs are the same but they are *mirror images* of each other:



Exercise 2.6.28

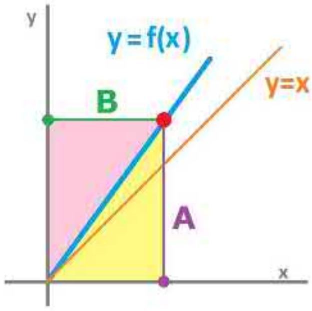
What graphs will land on themselves under this transformation?

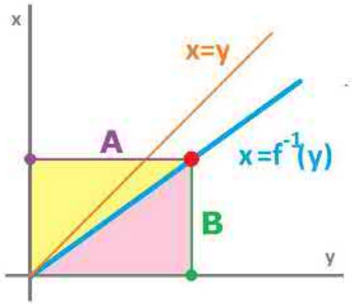
Exercise 2.6.29

What letters have this kind of symmetry?

Example 2.6.30: slope of inverse

This is what happens when we apply this flip to the graph of a linear function:





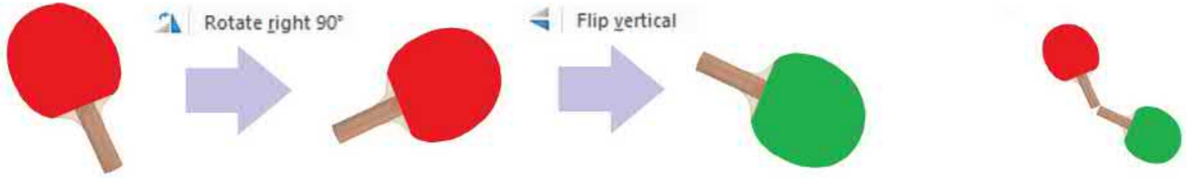
Then, we can compare:

$$\text{slope of } f \ = \frac{\text{rise}}{\text{run}} = \frac{A}{B} \ \text{ and } \ \text{slope of } f^{-1} \ = \frac{\text{rise}}{\text{run}} = \frac{B}{A}\,.$$

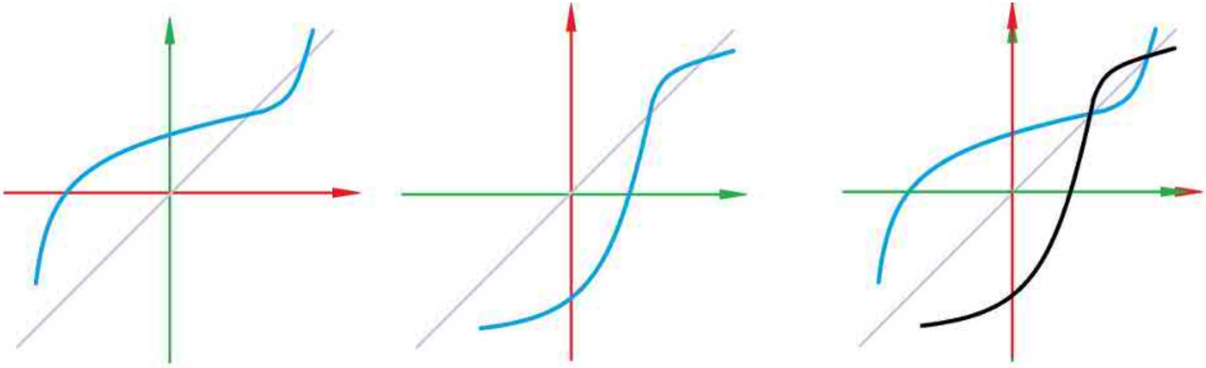
They are, as we already know, the reciprocals of each other!

Example 2.6.31: inverse graph with computer graphics

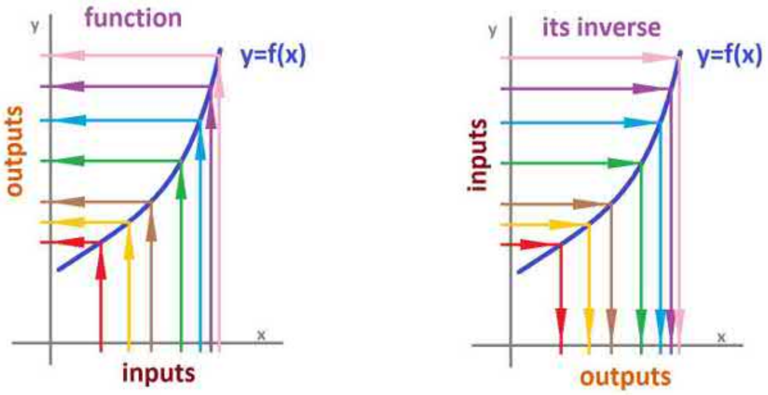
Such a transformation of the plane can be accomplished with simple image editing software by first rotating the image clockwise 90 degrees and then flipping it vertically:



This is how starting from a graph (first below), we find the graph of the inverse (second), and then bring them together for comparison (third):



Remember, we only need the graph of the original function to be able to evaluate all the values of the inverse:



Exercise 2.6.32

Function $y = f(x)$ is given below by a list its values. Find its inverse and represent it by a similar table.

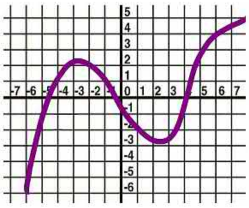
x	0	1	2	3	4
$y = f(x)$	1	2	0	4	3

Exercise 2.6.33

What kind of function is its own inverse?

Exercise 2.6.34

Plot the inverse of the function shown below, if possible:



Exercise 2.6.35

Function $y = f(x)$ is given below by a list of its values. Is the function one-to one? What about its inverse?

x	0	1	2	3	4
$y = f(x)$	7	5	3	4	6

Exercise 2.6.36

Plot the graph of the function $f(x) = \dfrac{1}{x-1}$ and the graph of its inverse. Identify its important features.

2.7. *Units conversions and changes of variables*

The variables of the functions we are considering are quantities we meet in everyday life. Frequently, there are multiple ways to measure these quantities:

- length and distance: inches, miles, meters, kilometers, ..., light years
- area: square inches, square miles, ..., acres

- volume: cubic inches, cubic miles, ..., liters, gallons
- time: minutes, seconds, hours, ..., years
- weight: pounds, grams, kilograms, karats
- temperature: degrees of Celsius, of Fahrenheit
- money: dollars, euros, pounds, yen
- etc.

Almost all conversion formulas are just multiplications, such as this one:

$$\#\ \text{ of meters } = \#\ \text{ of kilometers } \cdot 1000\,.$$

Warning!

We don't convert "pounds to kilos", we convert the *number of* pounds to the *number of* kilos.

The only exception of the temperature, because 0 degrees of Celsius doesn't correspond to 0 degrees of Fahrenheit.

This is the relation between degrees and radians:

$$\pi \text{ radians } = 180 \text{ degrees}\,.$$

In other words, the conversion between the number of degrees d and the number of radians r is the following relation:

$$\pi r = 180d\,.$$

Therefore, we convert from degrees to radians with the following function:

$$r = \frac{\pi}{180}d\,.$$

Then, we convert from radians degrees with the inverse of this function:

$$d = \frac{180}{\pi}r\,.$$

Within each of the categories, there may be complex, even circular, relations. For example, we have the following among these currencies:

$$\begin{tikzcd}
\#\ \text{of dollars} \arrow[r, "\times .9"] & \#\ \text{of euros} \arrow[d, "\times 122"] & & USD \arrow[d, "/1.3"] & EUR \arrow[l, "/.9"'] \\
\#\ \text{of pounds} \arrow[u, "\times 1.3"'] & \#\ \text{of yen} \arrow[l, "\times 0.007"'] & \text{or} & GBP \arrow[r, "/0.007"] & JPY \arrow[u, "/122"']
\end{tikzcd}$$

The arrows, of course, indicate functions, two in a row indicate compositions, and the reversed arrows are the inverses!

Exercise 2.7.1

Make your way from minutes to years.

We don't deal with these quantities one by one nor even in these pairs. We will study the *functions* that have them as variables.

We will first consider the *compositions* of these functions with the functions that represent the unit conversions.

Example 2.7.2: units of distance

Suppose t is the time and x is the location. Suppose also that a function g represents the change of units of length, such as from miles to kilometers:

$$z = g(x) = 1.6x\,.$$

Then, the change of the units will make very little difference; the coefficient, $m = 1.6$, is the only adjustment necessary. If f is the distance in miles, then h is the distance in kilometers: $h(t) = 1.6f(t)$. Thus, all the functions are replaced with their multiples. The graphs are stretched!

We call such a unit conversion a *change of variables*. Usually, it is done one at a time: either the dependent or the independent variable.

Example 2.7.3: motion and units

Suppose we study *motion* and we have a function $y = f(x)$ that relates

- x, time in minutes, to
- y, location in inches.

What if we need to switch to

- t, time in seconds, or
- z, location in feet?

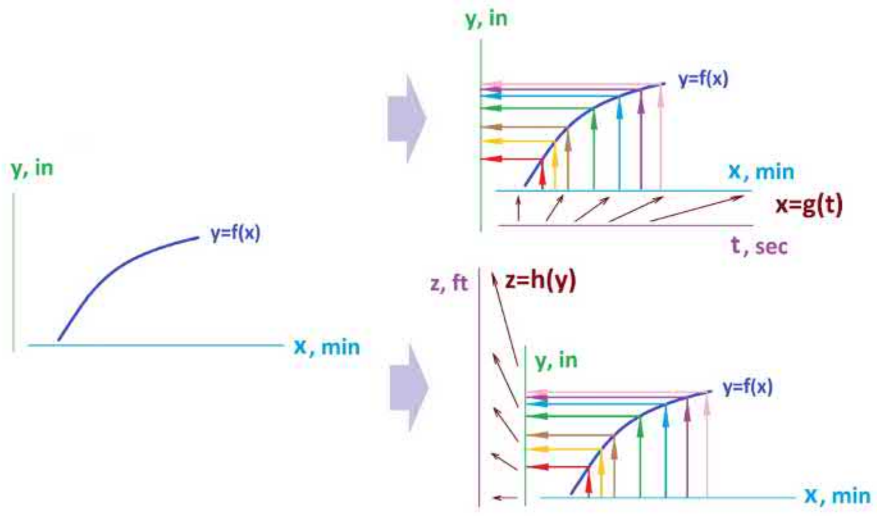
The algebra is clear:

$$x = t/60 \quad \text{and} \quad z = y/12\,.$$

Then we might have two new functions:

$$y = f(t/60) \quad \text{and} \quad z = f(x)/12\,.$$

Now, what will the new graphs look like? To answer, we combine the graph of f with the two transformations of the two axes, as follows:



The result is a vertical and a horizontal stretch/shrink. However, it's entirely up to us to choose the units on the new axes to match the old: the graph will remain the same!

Exercise 2.7.4

What is the relation between seconds and feet?

Example 2.7.5: time and temperature

Suppose we have a function f that records the temperature (in Fahrenheit) as a function of time (in minutes).

Question:

▶ What should f be replaced with if we want to record the temperature in Celsius as a function of time in seconds?

Let's name the variables:

- s is the time in seconds,
- m is the time in minutes,
- F is the temperature in Fahrenheit,
- C is the temperature in Celsius.

Suppose the original function, say,

$$F = f(m)\,,$$

is to be replaced with some new function,

$$C = g(s)\,.$$

First, we need the *conversion formulas* for these units. First, the time. This is what we know:

$$\text{1 minute } = \text{60 seconds.}$$

However, this is not the formula to be used to convert s to m because these are the *number* of seconds and the *number* of minutes, respectively. Instead, we have

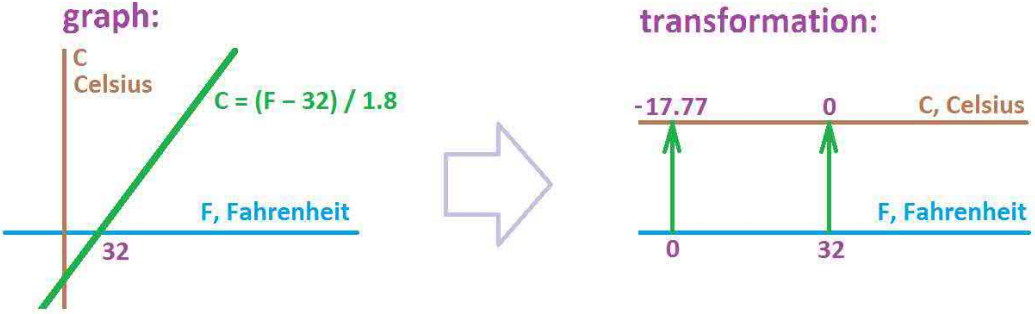
$$m = s/60\,.$$

We represent the function by its graph and as a transformation:

Second, the temperature. This is what we know:

$$C = (F - 32)/1.8\,.$$

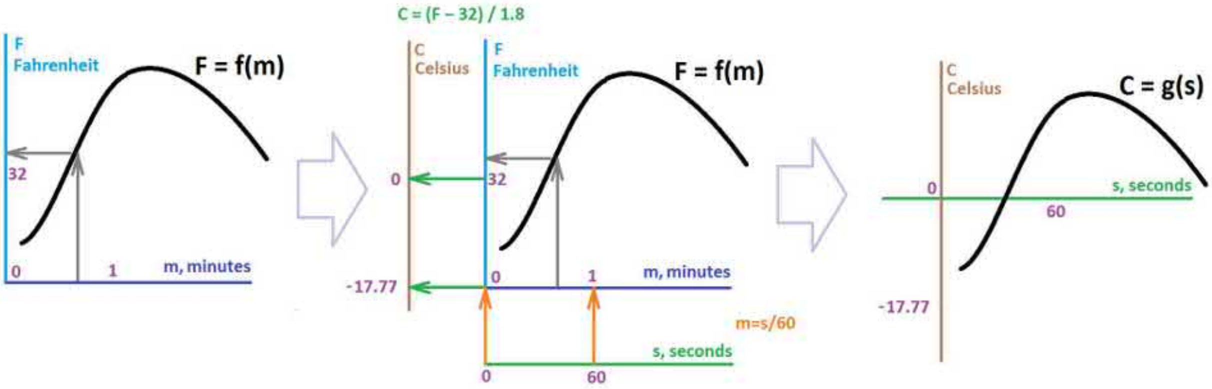
We represent the function by its graph and as a transformation:



These are the relations between the four quantities:

$$g: \quad s \xrightarrow{s/60} m \xrightarrow{f} F \xrightarrow{(F-32)/1.8} C$$

Instead of transforming the axes and, therefore, the plane, we choose to simply *relabel* them:



The answer to our question is, we replace f with g, the *composition* of the above functions:

$$F = g(s) = \big(f(s/60) - 32\big)/1.8\,.$$

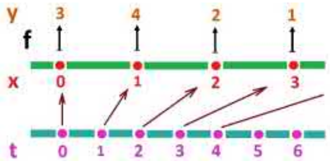
Note that both of the conversion formulas are one-to-one functions! That's what guarantees that the conversions are unambiguous and reversible. More precisely, we say that these functions are *invertible*. Indeed, these are the inverses, for the time:

$$s = 60m\,,$$

and for the temperature:

$$F = 1.8C + 32\,.$$

Note that all of the conversion formulas have been provided by *linear functions*. Then, a linear change of variables will cause the x-axis or the y-axis to shift, stretch, or flip (vertically or horizontally):



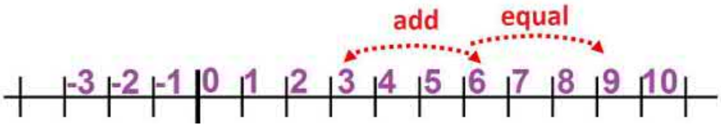
We conclude:

- ▶ A linear change of variables will cause the graph of the function to shift, stretch, or flip.

2.8. *The arithmetic operations on functions*

We would like to treat *all numerical functions as a single group.*

We find inspiration in how we have handled the *real numbers*. We put them together in the real number line, which provides us with a bird's-eye view:



But there is much more! We also recognize that these entities are interacting with each other, producing offspring via arithmetic:

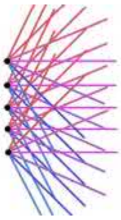
$$3 + 6 = 9,\ \ 5 \cdot 7 = 35,\ \ \text{etc.}$$

Understanding the meaning of these computations requires understanding that the beginning and the end of such a computation are just two different representations of the *same* number:

$$1+1 \;=\; 2\cdot 1 \;=\; 2\,.$$

There is a single location for each of these expressions on the real number line! Similarly, $x+x$ and $2x$ correspond to the same (albeit unspecified) location on the number line. We simply say that they are *equal.*

It is much more challenging to find such a bird's-eye view for *functions*! For example, this is what an attempt to visualize all *linear* functions would look like:



Just as with numbers, it is the *interactions* between functions that make them manageable as a whole.

For each of the four arithmetic operations on *numbers* – addition, subtraction, multiplication, and division – there is an operation on (numerical) *functions.*

But first let's make sure that we have a clear understanding of what it means for two functions to be the same. For example, these two functions are represented by two different formulas:

$$x+x \text{ and } 2x\,.$$

Are they the same function? Of course! These two functions are represented by two similar formulas:

$$x-1 \text{ and } 1-x\,.$$

Are they the same function? Of course not! How do we know?

The answer (as is the question itself) is dependent on our definition of function:

A function is a list of inputs and outputs.

To answer the question, we can simply test the formulas by plugging input values and watching the outputs:

$$\begin{array}{llll}
x+x\Big|_{x=0}=0, & 2x\Big|_{x=0}=0\,. & \text{Same!}\\
x+x\Big|_{x=1}=2, & x+x\Big|_{x=1}=2\,. & \text{Same!}\\
... & ... & ???
\end{array}$$

It seems the same. Now the second pair:

$$\begin{array}{llll}
x-1\Big|_{x=0}=-1, & 1-x\Big|_{x=0}=1\,. & \text{Different!}\\
x-1\Big|_{x=1}=0, & 1-x\Big|_{x=1}=0\,. & \text{Same!}
\end{array}$$

We stop here because a single mismatch means that they are different!

But what about these functions:

$$\frac{2x^2+2x}{2} \text{ and } x^2+x\,,$$

or those:

$$\frac{2x^2+2x}{x} \text{ and } 2x+2\,?$$

We again plug in the values:

$$\begin{array}{llll}
\left.\dfrac{2x^2+2x}{2}\right|_{x=0} = 0, & \left.x^2+x\right|_{x=0} = 0\,. & \text{Same!} \\
\left.\dfrac{2x^2+2x}{2}\right|_{x=1} = 2, & \left.x^2+x\right|_{x=1} = 2\,. & \text{Same!} \\
\ldots & \ldots & ???
\end{array}$$

The results are the same for every x! What about the latter? It breaks down:

$$\left.\frac{2x^2+2x}{x}\right|_{x=0} \text{ is undefined,} \quad \left.2x+2\right|_{x=0} = 2\,. \quad \text{Different!}$$

Plugging in $x = 0$ will produce division by 0 for the first function in the pair but not for the second. It is clear then that two functions can't be the same unless their domains are equal too (as sets).

The following is the test two function functions f and g are subjected to:

$$\begin{tikzcd}
 & & \boxed{f} \arrow[dr] & \\
x \arrow[r] & {} \arrow[ur] \arrow[dr] & & \text{same?} \\
 & & \boxed{g} \arrow[ur] &
\end{tikzcd}$$

So, f and g are called *equal*, or we say it's *the same function*, if they have the same domain and

$$f(x) = g(x) \quad \texttt{FOR EACH } x$$

in the domain.

These are our answers to the above questions. Are these two functions the same:

$$f(x) = \frac{2x^2+2x}{2} \quad \text{and} \quad g(x) = x^2 + x\ ?$$

Yes, because

$$\frac{2x^2+2x}{2} = x^2 + x \quad \text{for every } x.$$

It is crucial that the *implied domains* – for every x – of the two functions are the same.

We use the same simple notation for functions as for numbers:

Equal functions

$$f = g$$

Are these two functions the same:

$$f(x) = \frac{2x^2+2x}{x} \quad \text{and} \quad g(x) = 2x + 2\ ?$$

No, because the implied domain of the former doesn't include 0 while that of the latter does. The difference is in a single value!

We also use this simple notation:

Not equal functions

$$f \neq g$$

As you can see, once we discover that the domains don't match, we are done. However, choosing another domain will fix the problem:

$$f(x) = \frac{2x^2 + 2x}{x} \ \text{ and } \ g(x) = 2x + 2 \ \text{ are the same function on the domain } \{x: \ x \neq 0\}\,.$$

As another relevant example, these are two different functions:

- x^2 with domain $(-\infty, \infty)$;
- x^2 with domain $[0, \infty)$.

Exercise 2.8.1

Consider:

$$\frac{x}{x^2} \ \text{ vs. } \ \frac{1}{x}\,.$$

Exercise 2.8.2

Suggest your own examples of functions that differ by a single value.

The statement in the definition, such as

$$\frac{2x^2 + 2x}{2} = x^2 + x \ \text{ for every real } x,$$

is called an *identity*. The last part is often assumed and omitted from computations. The following statement is also an identity:

$$\frac{2x^2 + 2x}{x} = 2x + 2 \ \text{ for every real } x \neq 0\,.$$

However, the last part is a caveat that *cannot* be omitted! In other words, an identity is just a statement about two functions being "identically" equal, i.e., indistinguishable, within the specified domain.

This idea of transitioning from a function to its *twin* is the basis of all algebraic manipulations; they are informally called "simplifications" or "cancellations".

Now, the outputs of numerical functions are *numbers*. Therefore, any arithmetic operation on numbers – addition, subtraction, multiplication, and division – can now be applied to functions, one input at a time. Once again, functions interact and produce offspring, new functions.

The definitions of these new functions are simple:

Definition 2.8.3: sum of functions

Given two functions f and g, the *sum*, $f + g$, of f and g is the function defined by the following:

$$\boxed{(f + g)(x) = f(x) + g(x) \quad \texttt{FOR EACH } x}$$

in the intersection of the domains of f and g.

Note how the two plus signs in the formula are different: The first one is a part of the *name* of the new function while the second is the actual sign of summation of two real numbers. This is the "deconstruction" of the notation:

Sum of functions

$$\begin{array}{ccc} & & \text{names of the first and second functions}\\ & & \downarrow \qquad\quad \downarrow \\ (\, f+g\,)\ (x) & = & f(x)+g\ (x)\\ \uparrow & & \ \ \uparrow\quad\uparrow\quad\uparrow\\ \text{name of the new function} & & \text{operation on numbers}\end{array}$$

Furthermore, we now have an operation on functions: $f+g$ is a new function.

Example 2.8.4: algebra of functions

The sum of

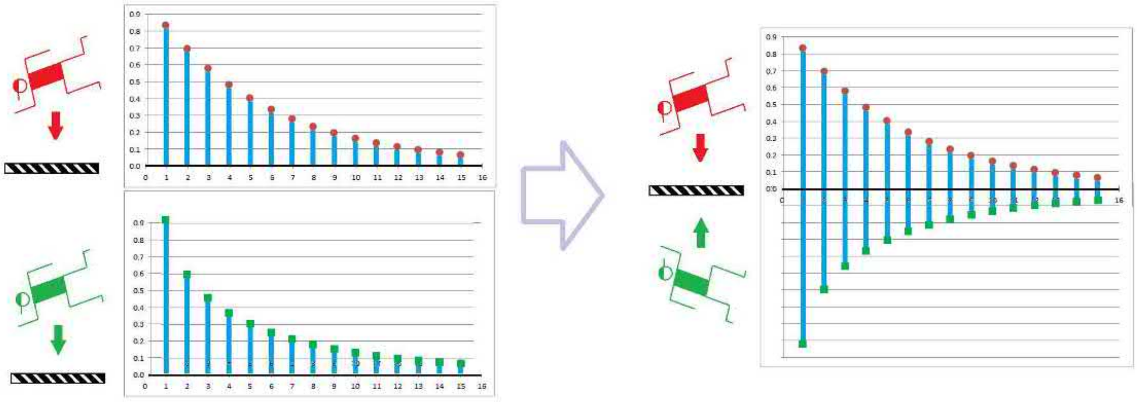
$$g(x)=x^2 \quad \text{and} \quad f(x)=x+2$$

is

$$(g+f)(x)=g(x)+f(x)=\left(x^2\right)+\left(x+2\right).$$

Whether this is to be simplified or not, a new function has been built.

This is an illustration of the meaning of the sum of two functions:



One can see how the values are added, location by location.

We represent a function f diagrammatically as a *black box* that processes the input and produces the output:

$$\begin{array}{ccccc}\text{input} & & \text{function} & & \text{output}\\ x & \to & \boxed{\ f\ } & \to & y\end{array}$$

Now, suppose we have another function g:

$$\begin{array}{ccccc}\text{input} & & \text{function} & & \text{output}\\ t & \to & \boxed{\ g\ } & \to & u\end{array}$$

How do we represent their sum $f+g$? To represent it as a single function, we need to "wire" their diagrams together side by side:

$$\begin{array}{ccccc} x & \to & \boxed{\ f\ } & \to & y\\ || & & & & \updownarrow\\ t & \to & \boxed{\ g\ } & \to & u\end{array}$$

But it's only possible when the input of f coincides with the input of g. We may have to *rename the variable* of g. We replace t with x. Then we have a new diagram for a new function:

$$\begin{array}{cccccccccccccc} & & & & \nearrow & x & \to & \boxed{\ f\ } & \to & y & \searrow & & & \\ f+g: & x & \to & x & & & & & & & & \boxed{\text{add}} \to & z & \to\ z\\ & & & & \searrow & x & \to & \boxed{\ g\ } & \to & u & \nearrow & & & \end{array}$$

We see how the input variable x is copied into the two functions, processed by them *in parallel*, and finally the two outputs are added together to produce a single output. The result can be seen as just a new black box:

$$x \ \to \ \boxed{\ f+g\ } \ \to \ y$$

Warning!

When units are involved, we must make sure that the outputs match so that we can add them.

Example 2.8.5: algebra of functions

We have combined two functions into one but we often need to go the other way and break a complex function into simpler parts that can then be studied separately. Represent

$$z = h(x) = x^2 + \sqrt[3]{x}$$

as the sum of two functions. Here is the answer:

$$x \mapsto y = x^2 \quad \text{and} \quad x \mapsto y = \sqrt[3]{x}\,.$$

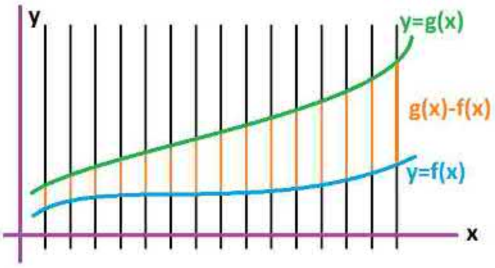
Subtraction also gives us an operation on functions.

Definition 2.8.6: difference of functions

Given two functions f and g, the *difference*, $g - f$, of f and g is the function defined by the following:

$$\boxed{(g-f)(x) = g(x) - f(x) \quad \texttt{FOR EACH } x}$$

in the intersection of the domains of f and g.



Before we get to multiplication of functions, there is a simpler but very important version of this operation.

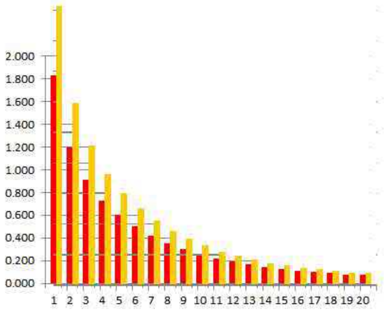
Definition 2.8.7: constant multiple of function

Given a function f, the *constant multiple* cf of f, for some real number c, is the function defined by the following:

$$(cf)(x) = cf(x) \quad \texttt{FOR EACH } x$$

in the domain of f.

In the following illustration of the meaning of a constant multiple of a function, one can see how its values are multiplied by $c = 1.3$ one location at a time:



There may be more than two functions involved in these operations or they can be combined.

Example 2.8.8: algebra of functions

Sum combined with differences:

$$h(x) = 2x^3 - \frac{5}{x} + 3x - 4\,.$$

The function is also seen as the sum of constant multiples, called a "linear combination":

$$h(x) = 2 \cdot \left(x^3\right) + (-5) \cdot \frac{1}{x} + 3 \cdot x + (-4) \cdot 1\,.$$

Example 2.8.9: algebra of functions given by tables

When two functions are represented by their lists of values, their sum (difference, etc.) can be easily computed. We simply go row by row adding the values.

Suppose we need to add these two functions, f and g, and create a new one, h, represented by a similar list:

$$\begin{array}{c|c} x & y = f(x) \\ \hline 0 & 1 \\ 1 & 2 \\ 2 & 3 \\ 3 & 0 \\ 4 & 1 \end{array} \quad + \quad \begin{array}{c|c} x & y = g(x) \\ \hline 0 & 5 \\ 1 & -1 \\ 2 & 2 \\ 3 & 3 \\ 4 & 0 \end{array} \quad = \quad ?$$

We simply add the output of the two functions for the same input. First row:

$$f : 0 \mapsto 1, \quad g : 0 \mapsto 5 \quad \Longrightarrow \quad h : 0 \mapsto 1 + 5 = 6\,.$$

Second row:

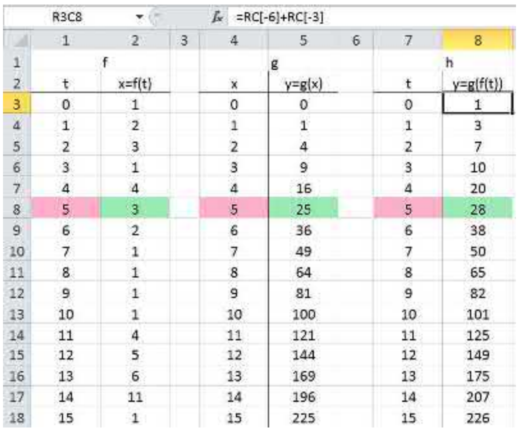
$$f : 1 \mapsto 2, \quad g : 1 \mapsto -1 \quad \Longrightarrow \quad h : 1 \mapsto 2 + (-1) = 1\,.$$

And so on. This is the whole solution:

$$\begin{array}{c|c} x & y = f(x) \\ \hline 0 & 1 \\ 1 & 2 \\ 2 & 3 \\ 3 & 0 \\ 4 & 1 \end{array} \quad + \quad \begin{array}{c|c} x & y = g(x) \\ \hline 0 & 5 \\ 1 & -1 \\ 2 & 2 \\ 3 & 3 \\ 4 & 0 \end{array} \quad = \quad \begin{array}{c|l} x & y = f(x) + g(x) \\ \hline 0 & 1 + 5 = 6 \\ 1 & 2 + (-1) = 1 \\ 2 & 3 + 2 = 5 \\ 3 & 0 + 3 = 3 \\ 4 & 1 + 0 = 1 \end{array} \quad = \quad \begin{array}{c|l} x & y = h(x) \\ \hline 0 & 6 \\ 1 & 1 \\ 2 & 5 \\ 3 & 3 \\ 4 & 1 \end{array}\,.$$

Example 2.8.10: algebra of functions with spreadsheet

This is how the sum of two functions is computed with a spreadsheet:



	1	2	3	4	5	6	7	8
1		f			g			h
2	t	x=f(t)		x	y=g(x)		t	y=g(f(t))
3	0	1		0	0		0	1
4	1	2		1	1		1	3
5	2	3		2	4		2	7
6	3	1		3	9		3	10
7	4	4		4	16		4	20
8	5	3		5	25		5	28
9	6	2		6	36		6	38
10	7	1		7	49		7	50
11	8	1		8	64		8	65
12	9	1		9	81		9	82
13	10	1		10	100		10	101
14	11	4		11	121		11	125
15	12	5		12	144		12	149
16	13	6		13	169		13	175
17	14	11		14	196		14	207
18	15	1		15	225		15	226

The formula is very simple:

`=RC[-6]+RC[-3]`

All four algebraic operations produce new functions in the same manner:

$$\begin{tikzcd}
 & & x \arrow[r] & \boxed{f} \arrow[r] & y \arrow[dr] & & \\
x \arrow[r] & x \arrow[ur] \arrow[dr] & & & & \boxed{+ \; - \; \cdot \; \div} \arrow[r] & z \\
 & & x \arrow[r] & \boxed{g} \arrow[r] & u \arrow[ur] & &
\end{tikzcd}$$

Exercise 2.8.11

Have we found the domains of these new functions?

Warning!

The algebra of functions comes from the algebra of *outputs*; the inputs don't even have to be numbers.

Composition, however, is the most important operation on functions. There is no matching operation for numbers.

2.9. Solving equations

Let's review what it means to solve an equation.

We go back to our example of boys and balls. This is our function that tells what ball each boy prefers:

$$\begin{array}{lll}
F(\text{ Tom }) & = & \text{basketball} \\
F(\text{ Ned }) & = & \text{tennis} \\
F(\text{ Ben }) & = & \text{basketball} \\
F(\text{ Ken }) & = & \text{football} \\
F(\text{ Sid }) & = & \text{football}
\end{array}$$

So, our function – in the form of this list – answers the question:

- Which ball is this boy playing with?

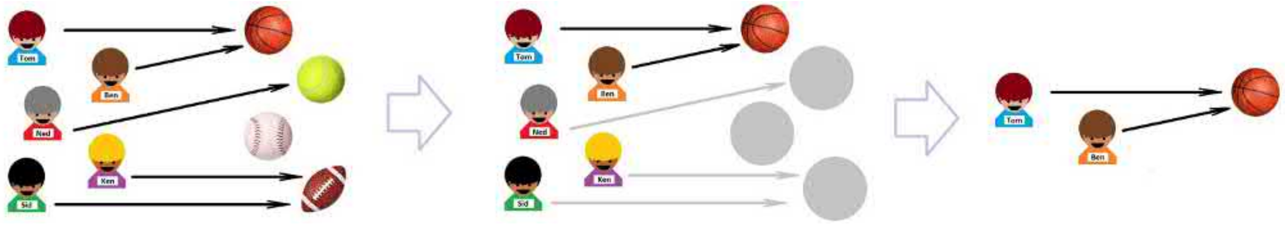
However, what if we turn this question around:

- Which boy is playing with this ball?

Let's try an example: Who is playing with the basketball? Before answering it, we can give this question a more compact form, the form of an *equation*:

$$F(\text{ boy }) = \text{ basketball } .$$

Indeed, we need to find the inputs that, under F, produce this specific output. We visualize and answer the question by erasing all irrelevant arrows:



These are a few of possible questions of this kind along with the answers:

- Who is playing with the basketball? Tom and Ben!
- Who is playing with the tennis ball? Ned!
- Who is playing with the baseball? No one!
- Who is playing with the football? Ken and Sid!

It seems that there are several answers to each of these questions. Or are there? "Tom" and "Ben" aren't *two* answers; it's one: "Tom and Ben"! Indeed, if we provide one name and not the other, we haven't fully answered the question. We know that we should write the answer as

$$\{ \text{ Tom, Ben } \} .$$

It's a set!

Let's review. The solution of an equation $f(x) = y$ with $f : X \to Y$ is always a set (a subset of X) and it may contain *any* number of elements, including none. To solve an equation with respect to x means to find *all* values of x that satisfy the equation. In other words:

1. When we substitute any of those x's into the equation, we have a true statement.
2. There are no other such x's.

For example, consider how this equation is solved:

$$x + 2 = 5 \implies x = 3 .$$

That's an abbreviated version of the following statement:

> ▶ If x satisfies the equation $x + 2 = 5$, then x satisfies the equation $x = 3$.

Plug in:

$$x + 2 = (3) + 2 = 5 .$$

It checks out! We could *try* others and they won't check out:

	$(x) + 2 =$	$\overset{?}{=} 5$	`TRUE/FALSE`	
$x = 0$	$(0) + 2 = 2$	$\neq 5$	`FALSE`	
$x = 1$	$(1) + 2 = 3$	$\neq 5$	`FALSE`	
$x = 2$	$(2) + 2 = 4$	$\neq 5$	`FALSE`	
$x = 3$	$(3) + 2 = 5$	$= 5$	`TRUE`	Add it to the list!
$x = 4$	$(4) + 2 = 6$	$\neq 5$	`FALSE`	
...	...		...	

Of course, this trial-and-error method is unfeasible because there are infinitely many possibilities.

A method of *how* we may arrive to the answer is discussed in this section.

Recall the basic methods ("rules") of *handling* equations.

Example 2.9.1: simple equations

In order to solve the equation

$$x + 2 = 5\,,$$

subtract 2 from both sides producing

$$x + 2 - 2 = 5 - 2 \implies x = 3\,.$$

In order to solve the equation

$$3x = 2\,,$$

divide by 3 both sides producing

$$3x/3 = 2/3 \implies x = 2/3\,.$$

This is the summary.

Theorem 2.9.2: Basic Algebra of Equations

- *Multiplying both sides of an equation by a number preserves it. In other words, we have:*
$$a = b \implies ka = ka \quad \text{for any } k\,.$$
- *Adding any number to both sides of an equation preserves it. In other words, we have:*
$$a = b \implies a + s = b + s \quad \text{for any } s\,.$$

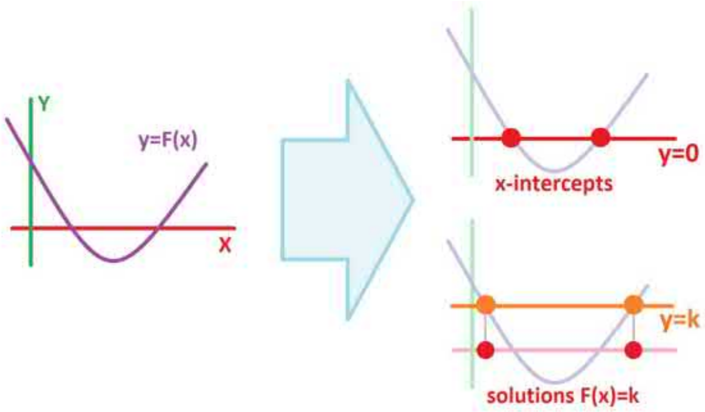
Warning!

Multiplying an equation by 0 is pointless.

Exercise 2.9.3

What about the converses?

As a reminder, in the special case when the right-hand side of the equation is zero, the solution set to this equation has a clear *geometric* meaning: An x*-intercept* of a numerical function f is any solution to the equation $f(x) = 0$. In other words, these are the x-coordinates of the *intersections* of the graph with the x-axis (top):



Furthermore, when the right-hand side is a number, say, k, the solution to this equation $f(x) = k$ gives us the intersection of the graph with the line $y = k$ (bottom).

Our interest in this section, though, is *algebra*.

We will deal with a simple kind of equation:

- ▶ x is present only once (in the left-hand side).

Like this:

$$x^2 = 17\,.$$

Starting with such an equation, our goal is – through a series of manipulations – to arrive to an even simpler kind of equation:

- ▶ x is *isolated* (in the left-hand side).

Like this:

$$x = \sqrt{17}\,.$$

Warning!

This *is* an equation.

In other words, we will try to find ways to get from left to right here:

$$\sin\left(e^{\frac{1}{\sqrt{x}}}\right) = 5 \quad\longrightarrow\quad ??? \quad\longrightarrow\quad x = \underbrace{\ldots\ldots}_{\text{no } x \text{ here}}$$

The *main idea* of how to manipulate equations is as follows:

- ▶ We apply a function to both sides of the equation, producing a new equation.

For example, if we have an equation, say,

$$x + 2 = 5\,,$$

we treat it as a number, call it y. Then we deal with this number:

$$y = x + 2 = 5,\quad \text{apply } z = y - 2 \implies (x+2) - 2 = 3 - 2 \implies x = 3 \quad \text{Solved!}$$

The idea is to produce – from an equation satisfied by x – another equation satisfied by x.

However, the challenge (and an opportunity) is that applying *any* function in this manner will produce a new equation satisfied by x! For example:

$$\begin{array}{lllllll}
y = x + 2 = 5, & \text{apply } z = y + 2 & \implies & (x+2) + 2 = 5 + 2 & \implies & x + 4 = 7 & \text{Not solved!} \\
y = x + 2 = 5, & \text{apply } z = y^2 & \implies & (x+2)^2 = 5^2 & & & \text{Not solved!} \\
y = x + 2 = 5, & \text{apply } z = \sin y & \implies & \sin(x+2) = \sin 5 & & & \text{Not solved!}
\end{array}$$

Indeed, it is the *definition* of a function that every input has exactly one output. Suppose we have a function g. Then, two equal (according to the equation) inputs of g will produce two equal outputs (another equation), always:

$$\begin{tikzcd}
\text{old equation:} & a \arrow[d, "g"] & = & b \arrow[d, "g"] \\
\text{new equation:} & g(a) & = & g(b)
\end{tikzcd}$$

In particular, we have, for any function g:

$$x + 2 = 5 \implies g(x+2) = g(5)\,.$$

That is why if the first equation is satisfied by x, then so is the second.

There are infinitely many possibilities for this function g:

$$\begin{tikzcd}
x + 4 = x + 2 + 2 = 7 & (x+2)^2 = 5^2 & \sin(x+2) = \sin 5 \\
x + 5 = x + 2 + 3 = 8 & \boxed{x + 2 = 5} \arrow[ul] \arrow[u] \arrow[ur] \arrow[l] \arrow[r] \arrow[dl] \arrow[d] \arrow[dr] & 2^{x+2} = 2^5 \\
\boxed{x = x + 2 - 2 = 3} & (x+2)^3 = 5^3 & \sqrt{x+2} = \sqrt{5}
\end{tikzcd}$$

If x satisfies the equation in the middle, it also satisfies the rest of the equations.

If we want to solve the original equation, which function – out of infinitely many – do we pick? Some of them clearly make the equation *more* complex than the original! It is the challenge for the equation solver to have enough foresight to choose a function to apply that will make the equation *simpler*.

Example 2.9.4: solving equation with function given by flowchart

Let's consider this equation:

$$3 \cdot \left(\frac{\sqrt{x}}{4} + 1 \right) = 6\,.$$

Here, several functions are consecutively applied to x. This is the flowchart of this function:

$$f:\ x \to \boxed{\text{root}} \to \boxed{\text{divide by 4}} \to \boxed{\text{add 1}} \to \boxed{\text{multiply by 3}} \to y$$

We can plug in any value on the left and get the output on the right:

$$0 \to \boxed{\text{root}} \to 0 \to \boxed{\text{divide by 4}} \to 0 \to \boxed{\text{add 1}} \to 1 \to \boxed{\text{multiply by 3}} \to 3$$

We *test* possible inputs this way. That one has failed; it's not 6!

Is there a better method?

We would like to get to x. To get to it, we will need to undo these functions one by one. In what order? Right to left, of course. We reverse the flow of the flowchart:

$$\begin{array}{rccccccccccc} f: & x & \to & \boxed{\text{root}} & \to & \boxed{\text{divide by 4}} & \to & \boxed{\text{add 1}} & \to & \boxed{\text{multiply by 3}} & \to & 6 \\ f^{-1}: & x & \leftarrow & \boxed{\text{square}} & \leftarrow & \boxed{\text{multiply by 4}} & \leftarrow & \boxed{\text{subtract 1}} & \leftarrow & \boxed{\text{divide by 3}} & \leftarrow & 6 \end{array}$$

Of course, the pairs of functions aligned vertically are the *inverses* of each other:

$$\begin{tikzcd} x \arrow[r] & \boxed{\text{root}} \arrow[r] & \boxed{\text{divide by 4}} \arrow[r] & \boxed{\text{add 1}} \arrow[r] & \boxed{\text{multiply by 3}} \arrow[r] & y \arrow[dd] \\ & \text{inverse?} & \text{inverse} & \text{inverse} & \text{inverse} & \\ x & \boxed{\text{square}} \arrow[l] & \boxed{\text{multiply by 4}} \arrow[l] & \boxed{\text{subtract 1}} \arrow[l] & \boxed{\text{divide by 3}} \arrow[l] & y \arrow[l] \end{tikzcd}$$

They, therefore, can be canceled out one pair at a time:

$$\begin{tikzcd} x \arrow[r] & \boxed{\text{root}} \arrow[r] & \boxed{\text{divide by 4}} \arrow[r] & \boxed{\text{add 1}} \arrow[d] & \boxed{\blacksquare\blacksquare\blacksquare} \\ x & \boxed{\text{square}} \arrow[l] & \boxed{\text{multiply by 4}} \arrow[l] & \boxed{\text{subtract 1}} \arrow[l] & \boxed{\blacksquare\blacksquare\blacksquare} \end{tikzcd}$$

Then, we have the following and we can cancel more:

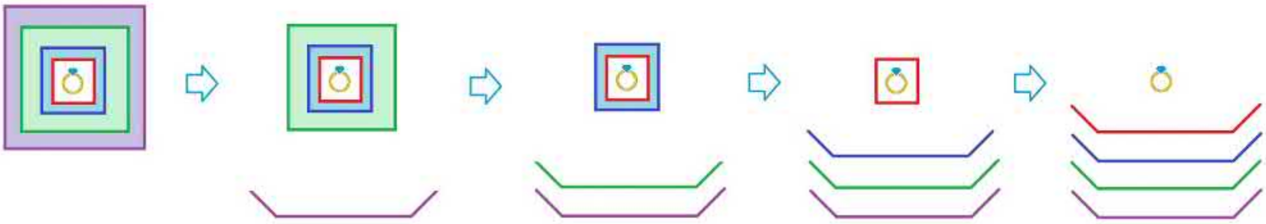
$$\begin{tikzcd} x \arrow[r] & \boxed{\text{root}} \arrow[r] & \boxed{\text{divide by 4}} \arrow[d] & \boxed{\blacksquare\blacksquare\blacksquare} & \boxed{\blacksquare\blacksquare\blacksquare} \\ x & \boxed{\text{square}} \arrow[l] & \boxed{\text{multiply by 4}} \arrow[l] & \boxed{\blacksquare\blacksquare\blacksquare} & \boxed{\blacksquare\blacksquare\blacksquare} \end{tikzcd}$$

And so on. We have demonstrated that the second row is indeed the inverse of the first.

With this fact understood, we find x by starting on the left with 6:

$$\pm 16 \leftarrow \boxed{\text{square}} \leftarrow 4 \leftarrow \boxed{\text{multiply by 4}} \leftarrow 1 \leftarrow \boxed{\text{subtract 1}} \leftarrow 2 \leftarrow \boxed{\text{divide by 3}} \leftarrow 6$$

Such equations can be visualized as follows; we see a single x "wrapped" in several layers of functions, as if a gift:



To get to the gift, the only method is to remove one wrapper at a time, from the outside in. In fact, you don't even know what kind of wrapper is the next until you've removed the last one!

As a summary, we have developed the following method of solving equations.

Theorem 2.9.5: General Algebra of Equations

Suppose g is an invertible function. Then applying g to both sides of an equation creates an equivalent equation. In other words, we have:

$$a = b \iff g(a) = g(a)\,.$$

Furthermore, the two equations have the same solution set.

Exercise 2.9.6

What is the relation between the two solution sets if instead of $\iff$, we have (a) $\Longleftarrow$, (b) $\Longrightarrow$?

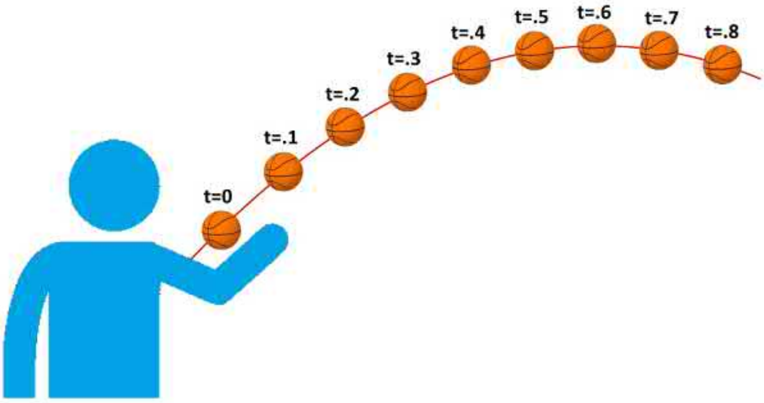
Chapter 3: The 2-dimensional space

Contents

3.1. Parametric curves on the plane

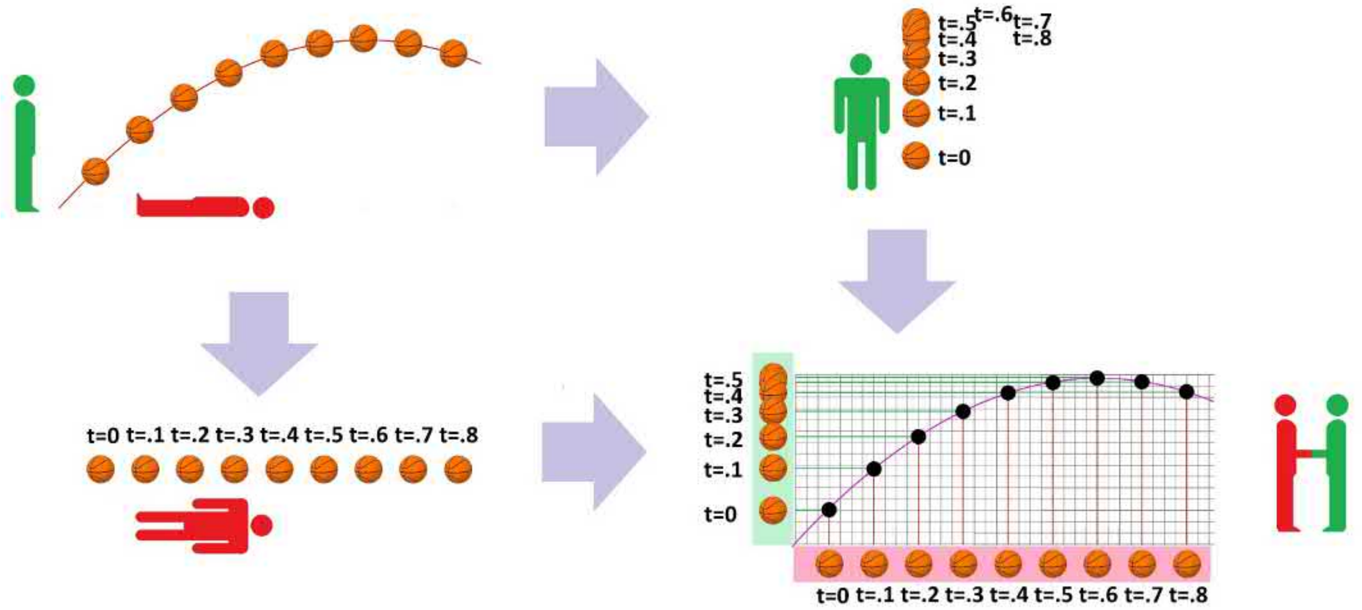
Example 3.1.1: ball

Imagine a person observing the flight of a thrown ball *from aside*:



Is there another way to capture this flight? Imagine there are two more observers:

- The first one (red) is on the ground under the path of the ball and can only see the forward progress of the ball.
- The second one (green) is behind the throw and can see only the rise and fall of the ball.



If the two make records where the ball was at what time, they can use the *time stamps* to match the two coordinates and then plot this point on the xy-plane. These points will form the ball's trajectory, what a side first observer would see. It is called a *parametric curve*

Curves aren't represented as graphs of functions. In fact, y doesn't depend on x anymore, but they are *related* to each other. The link is established by means of another variable, t. So, we have two functions that have nothing to do with each other except the inputs can be matched.

Definition 3.1.2: parametric curve

A *parametric curve* on the plane is a combination of two functions of the same variable:

$$\boxed{\begin{cases} x = f(t) \\ y = g(t) \end{cases}}$$

We can also use the Cartesian system of points to represent this curve:

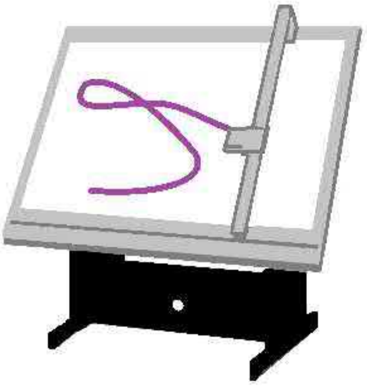
$$\boxed{(x, y) = \big(f(t), g(t)\big)}$$

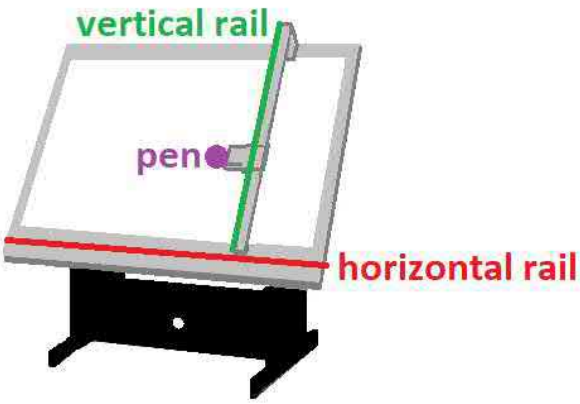
Exercise 3.1.3

Explain how a parametric curve is a relation.

Example 3.1.4: plotter

A curve may be plotted on a piece of paper by hand or by a computer by the following method. A pen is attached to a runner on a vertical bar, while that bar slides along a horizontal rail at the bottom edge of the paper:





The computer commands the next location of both as follows. At each moment of time t, we have:

1. The horizontal location of the vertical bar (and the pen) is given by $x = f(t)$.
2. The vertical location of the pen is given by $y = g(t)$.

Warning!

This view of parametric curves is most useful within the framework of multidimensional spaces and vectors.

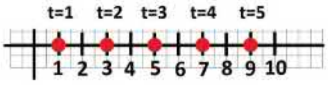
Example 3.1.5: straight lines

Let's examine motion along a straight line on the xy-plane.

First we go along the x-axis. The motion is represented by a familiar linear function of time:

$$x = 2t + 1\,.$$

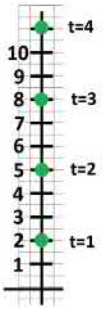
We are moving 2 feet per second to the right starting at $x = 1$. These are a few of the locations:



Second we go along the y-axis. The motion is represented by another linear function of time:

$$y = 3t + 2\,.$$

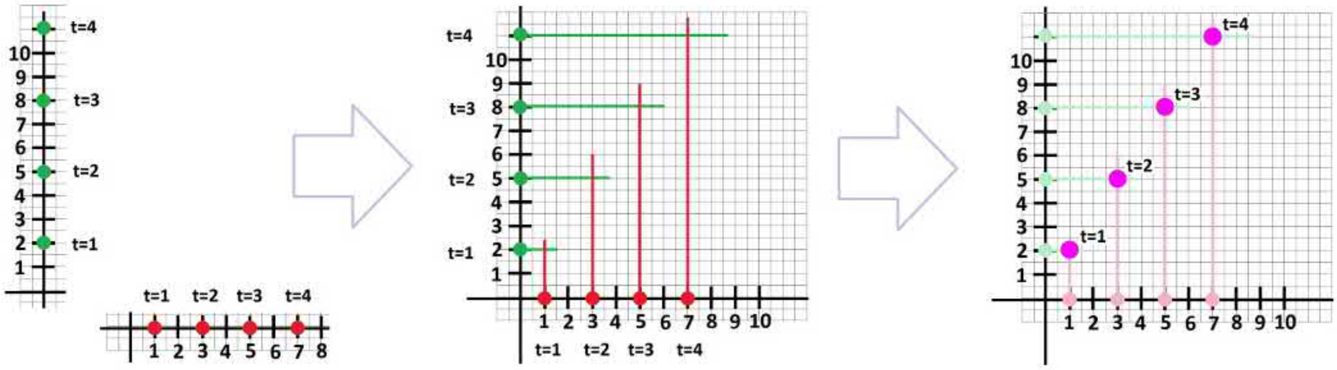
We are moving 3 feet per second up starting at $y = 2$. These are a few of the locations:



Now, what if these two are just two different views on the same motion from two different observers? Then we have:

$$\begin{cases} x = 2t + 1\,, \\ y = 3t + 2\,. \end{cases}$$

These are a few of the locations:



Exercise 3.1.6

Explain why these points lie on a straight line. Hint: triangles.

Of course, the motion metaphor – x and y are coordinates in the space – will be superseded. In contrast to this approach, we look at the two quantities and two functions that might have *nothing* to do with each other (except for t of course).

Example 3.1.7: commodities trader

Suppose a commodities trader follows the market. What he sees is the following:

- t is time.
- x is the price of wheat (say, in dollars per bushel).
- y is the price of sugar (say, in dollars per ton).

We simply have two functions and we – initially – look at them separately.

First, let's imagine that the price of *wheat* is decreasing:

$$x \searrow$$

The data comes to the observer in a pure, numerical form. To emulate this situation and to make this specific, one can choose a formula, for example:

$$x = f(t) = \frac{1}{t+1}\,.$$

To show some actual data, we evaluate x for several values of t:

t	x
0	1.00
1	.50
2	.33

With more points acquired in a spreadsheet we can plot the graph on the tx-plane:

Input, time t	Output, price of wheat x
0.00	1.00
0.20	0.83
0.40	0.71
0.60	0.63
0.80	0.56
1.00	0.50
1.20	0.45
1.40	0.42
1.60	0.38
1.80	0.36
2.00	0.33
2.20	0.31
2.40	0.29

At this point, we could, if needed, apply the available apparatus to study the symmetries, the monotonicity, the extreme points, etc. of this function.

Second, suppose that the price of *sugar* is increasing and then decreasing:

$$y \nearrow \searrow$$

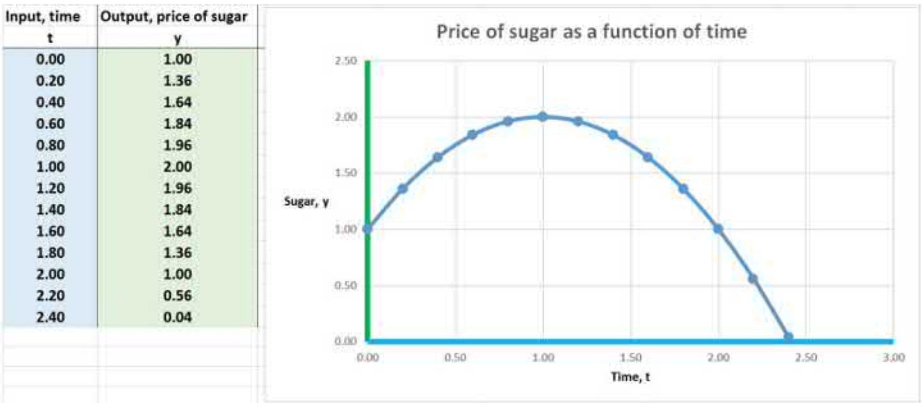
To make this specific, we can choose an upside down parabola:

$$y = g(t) = -(t-1)^2 + 2\,.$$

We then again evaluate y for several values of t:

t	y
0	1.00
1	2.00
2	1.00

With more points acquired in a spreadsheet we plot the graph on the ty-plane:



Input, time t	Output, price of sugar y
0.00	1.00
0.20	1.36
0.40	1.64
0.60	1.84
0.80	1.96
1.00	2.00
1.20	1.96
1.40	1.84
1.60	1.64
1.80	1.36
2.00	1.00
2.20	0.56
2.40	0.04

What if the trader is interested in finding hidden relations between these *two commodities*. Let's combine the data first:

t	x
0	1.00
1	.50
2	.33

and

t	y
0	1.00
1	2.00
2	1.00

$\longrightarrow$

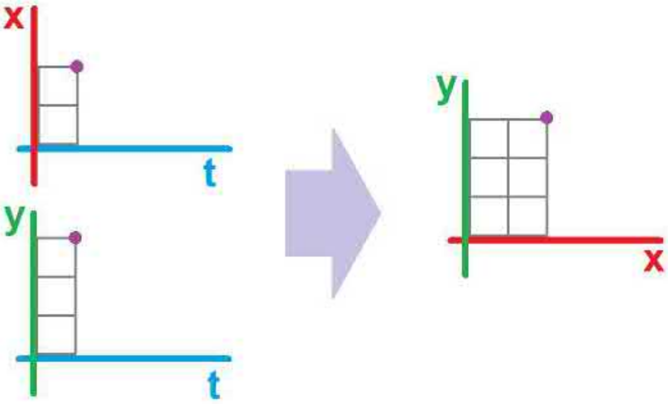
t	x	y
0	1.00	1.00
1	.50	2.00
2	.33	1.00

Since the input t is the same, we give it a single column. There seems to be two outputs. A better idea is to see *pairs* (x, y):

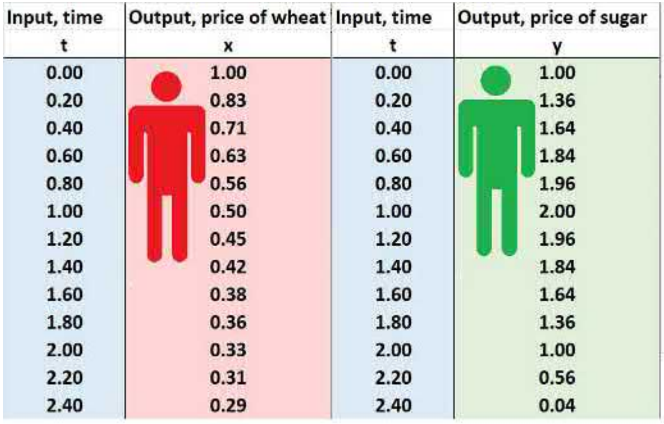
t	(	x	,	y	)
0	(	1.00	,	1.00	)
1	(	.50	,	2.00	)
2	(	.33	,	1.00	)

A value of x is paired up with a value of y when they appear along the same t in both plots.

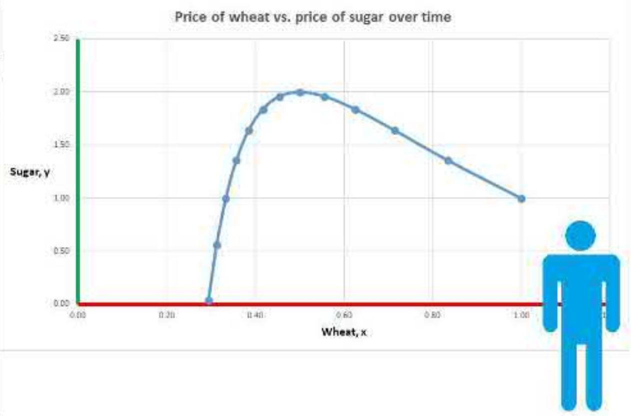
How do we combine the two *plots*? As the two plots are made of (initially) disconnected points – (t, x) and (t, y) – so is the new plot. This is what happens to each pair:



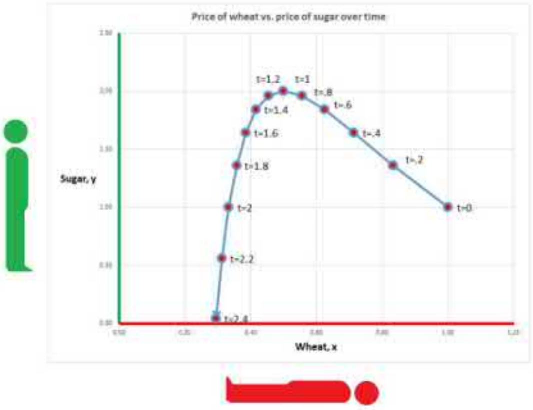
There is no t! As the independent variable is the same for both functions, only the dependent variables appear. Instead of plotting all points (t, x, y), which belong to the 3-dimensional space, we just plot (x, y) on the xy-plane – for each t. It's a "scatter plot" connected to make a curve:



Input, time	Output, price of wheat	Input, time	Output, price of sugar
t	x	t	y
0.00	1.00	0.00	1.00
0.20	0.83	0.20	1.36
0.40	0.71	0.40	1.64
0.60	0.63	0.60	1.84
0.80	0.56	0.80	1.96
1.00	0.50	1.00	2.00
1.20	0.45	1.20	1.96
1.40	0.42	1.40	1.84
1.60	0.38	1.60	1.64
1.80	0.36	1.80	1.36
2.00	0.33	2.00	1.00
2.20	0.31	2.20	0.56
2.40	0.29	2.40	0.04



The direction matters! Since t is missing, we have to make sure we know in which direction we are moving and indicate that with an arrow. Ideally, we also label the points in order to indicate not only "where" but also "when":



Thus, this is motion, just as before, but through what space? An abstract *space of prices* that we've made up. The space is comprised of all possible combinations of prices, i.e., a point (x, y) stands for a combination of two prices: x for wheat and y for sugar.

How much information about the dynamics of the two prices contained in the original functions can we recover from the new graph? A lot. We can shrink the graph vertically to de-emphasize the change of y and to reveal the *qualitative* behavior of x, and vice versa:

We see the decrease of x and then the increase followed by the decrease of y. In addition, the density of the points indicates the speed of the motion.

Example 3.1.8: abstract

We can do this in a fully abstract setting. When two functions, f, g, are represented by their respective lists of values (instead of formulas), they are easily combined into a parametric curve, F. We just need to eliminate the repeated column of inputs. Suppose we need to combine these two functions:

t	$x = f(t)$
0	1
1	2
2	3
3	0
4	1

&

t	$y = g(t)$
0	5
1	-1
2	2
3	3
4	0

$=$?

We repeat the inputs column – only once – and then repeat the outputs of either function. First row:

$$f : 0 \mapsto 1 \quad \& \quad g : 0 \mapsto 5 \quad \Longrightarrow F : 0 \mapsto (0, 5)\,.$$

Second row:

$$f : 1 \mapsto 2 \quad \& \quad g : 1 \mapsto -1 \quad \Longrightarrow F : 1 \mapsto (2, -1)\,.$$

And so on. This is the whole solution:

t	$x = f(t)$
0	1
1	2
2	3
3	0
4	1

and

t	$y = g(t)$
0	5
1	-1
2	2
3	3
4	0

$\longrightarrow$

t	$P = (f(t)\ ,\ g(t))$
0	$(1\ ,\ 5)$
1	$(2\ ,\ -1)$
2	$(3\ ,\ 2)$
3	$(0\ ,\ 3)$
4	$(1\ ,\ 0)$

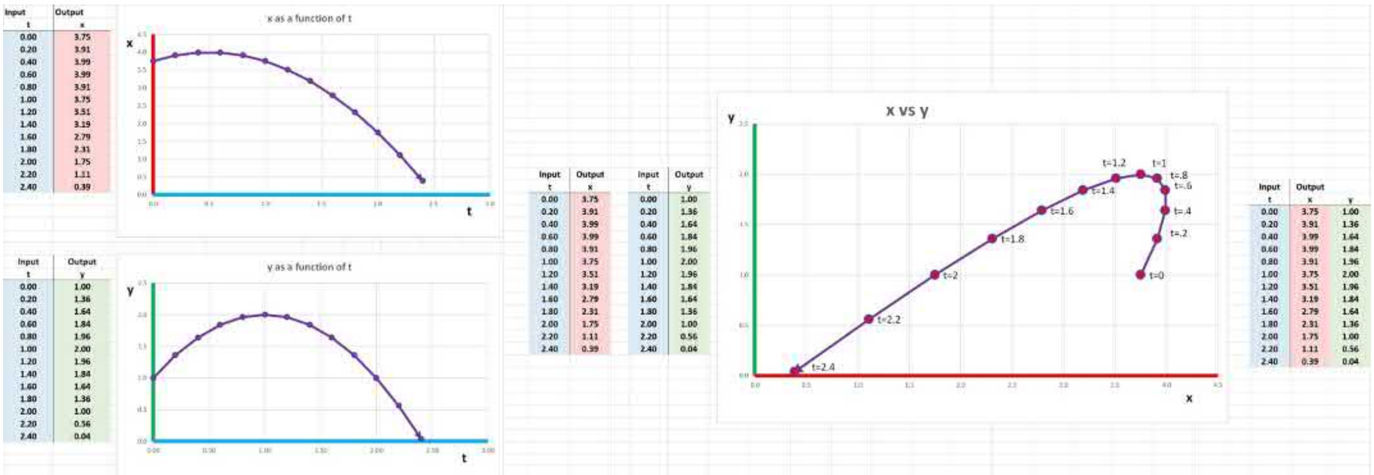
As you can see, there are no algebraic operations carried out and there is no new data, just the old data arranged in a new way. However, it is becoming clear that the list is also a function of some kind...

Warning!

The end result isn't the graph of any function.

Example 3.1.9: spreadsheet

This is a summary how the parametric curve is formed from two functions provided with a spreadsheet. The three columns – t, x, and y – are copied and then the last two are used to create a chart:



Input t	Output x
0.00	3.75
0.20	3.91
0.40	3.99
0.60	3.99
0.80	3.91
1.00	3.75
1.20	3.51
1.40	3.19
1.60	2.79
1.80	2.31
2.00	1.75
2.20	1.11
2.40	0.39

Input t	Output y
0.00	1.00
0.20	1.36
0.40	1.64
0.60	1.84
0.80	1.96
1.00	2.00
1.20	1.96
1.40	1.84
1.60	1.64
1.80	1.36
2.00	1.00
2.20	0.56
2.40	0.04

Input t	Output x	y
0.00	3.75	1.00
0.20	3.91	1.36
0.40	3.99	1.64
0.60	3.99	1.84
0.80	3.91	1.96
1.00	3.75	2.00
1.20	3.51	1.96
1.40	3.19	1.84
1.60	2.79	1.64
1.80	2.31	1.36
2.00	1.75	1.00
2.20	1.11	0.56
2.40	0.39	0.04

This chart is the *path* – not the graph – of the parametric curve. Note also that the curve isn't the graph of any function of one variable as the *Vertical Line Test* is violated.

Example 3.1.10: pattern

Plotting a parametric curve may reveal a relation between two quantities:

Input, time t	Output, price of wheat x
0.00	0.14
0.20	0.99
0.40	0.42
0.60	0.75
0.80	0.85
1.00	0.27
1.20	1.00
1.40	0.30
1.60	0.83
1.80	0.77
2.00	0.40
2.20	0.99
2.40	0.17

Input, time t	Output, price of sugar y
0.00	1.28
0.20	2.98
0.40	1.84
0.60	2.50
0.80	2.69
1.00	1.54
1.20	3.00
1.40	1.59
1.60	2.66
1.80	2.54
2.00	1.79
2.20	2.99
2.40	1.33

Parametric curves are functions!

This idea comes with certain obligations (Chapter 1). First, we have to *name* it, say F. Second, as we combine two functions, we use the following notation for this operation:

parametric curve

$$F=(f,g):\ \begin{cases}x=f(t)\\ y=g(t)\end{cases}$$

Next, what is the *independent variable*? It is t. After all, this is the input of both of the functions involved. What is the *dependent variable*? It is the "combination" of the outputs of the two functions, i.e., x and y. We know how to combine these; we form a pair, $P=(x,y)$. This P is a point on the xy-plane!

To summarize, we do what we have done many times before (addition, multiplication, etc.) – we create a new function from two old functions. We represent a function f diagrammatically as a *black box* that processes the input and produces the output:

$$\begin{array}{ccccc}\text{input} & & \text{function} & & \text{output}\\ t & \to & \boxed{f} & \to & x\end{array}$$

Now, what if we have another function g:

$$\begin{tikzcd}[row sep=small]
\text{input} & \text{function} & \text{output} \\
t \arrow[r] & \boxed{g} \arrow[r] & y
\end{tikzcd}$$

How do we represent $F=(f,g)$? To represent it as a single function, we need to "wire" their diagrams together side by side:

$$\begin{tikzcd}
t \arrow[r] \arrow[d, equal] & \boxed{f} \arrow[r] & x \arrow[d, leftrightarrow] \\
t \arrow[r] & \boxed{g} \arrow[r] & y
\end{tikzcd}$$

It is possible because the input of f is the same as the input of g. For the outputs, we can combine them even when they are of different nature. Then we have a diagram of a new function:

$$\begin{tikzcd}
 & & & t \arrow[r] & \boxed{f} \arrow[r] & x \arrow[dr] & & \\
(f,g): & t \arrow[r] & t \arrow[ur] \arrow[dr] & & & & (x,y) \arrow[r] & P \\
 & & & t \arrow[r] & \boxed{g} \arrow[r] & y \arrow[ur] & &
\end{tikzcd}$$

We see how the input variable t is copied into the two functions, processed by them *in parallel*, and finally the two outputs are combined together to produce a single output. The result can be seen as again black box:

$$\begin{tikzcd}
t \arrow[r] & \boxed{F} \arrow[r] & P\,.
\end{tikzcd}$$

The difference from all the functions we have seen until now is the nature of the output.

What about the *image* (the range of values) of $F=(f,g)$? It is supposed to be a recording of all possible outputs of F. The terminology used is often different though.

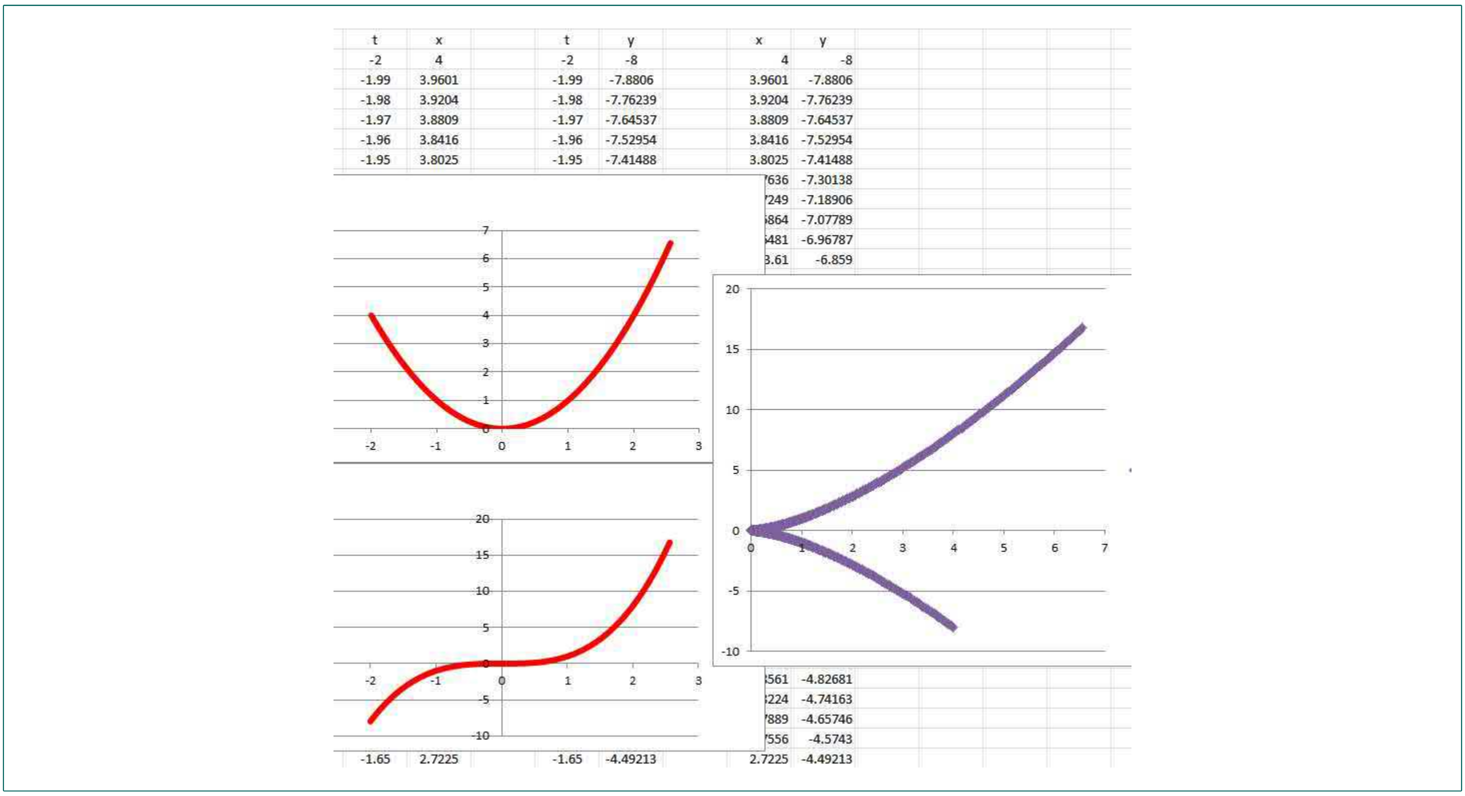
Definition 3.1.11:

The *path* of a parametric curve $x=f(t),\ y=g(t)$ is the set of all such points $P=(f(t),g(t))$ on the xy-plane.

The path is typically a curve. We plot several of them below.

Example 3.1.12: path

In general, the two processes, $x=x(t)$ and $y=y(t)$, are independent. When we combine them to see the path of the object by plotting (x,y) for each t, the result may be unexpected:



t	x	t	y	x	y
-2	4	-2	-8	4	-8
-1.99	3.9601	-1.99	-7.8806	3.9601	-7.8806
-1.98	3.9204	-1.98	-7.76239	3.9204	-7.76239
-1.97	3.8809	-1.97	-7.64537	3.8809	-7.64537
-1.96	3.8416	-1.96	-7.52954	3.8416	-7.52954
-1.95	3.8025	-1.95	-7.41488	3.8025	-7.41488
				636	-7.30138
				249	-7.18906
				864	-7.07789
				481	-6.96787
				3.61	-6.859
				561	-4.82681
				224	-4.74163
				889	-4.65746
				556	-4.5743
-1.65	2.7225	-1.65	-4.49213	2.7225	-4.49213

What about the *graph* of $F = (f, g)$? As we know from Chapter 1, the graph of a function is supposed to be a recording of all possible combinations of inputs and outputs of F. What if the outputs are 2-dimensional?

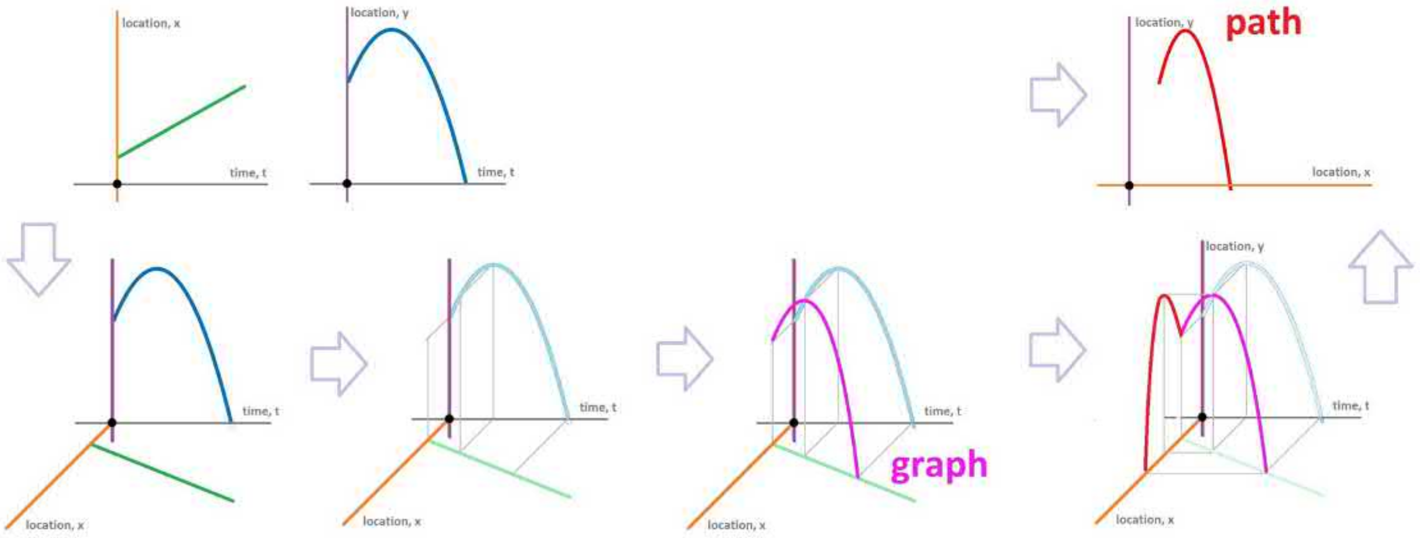
Definition 3.1.13:

The *graph* of a parametric curve $x = f(t),\ y = g(t)$ is the set of all such points $(t, x, y) = (t, f(t), g(t))$ in the txy-space.

The graph is built from:

- the graph of $x = f(t)$ on the tx-plane (the floor), and
- the graph of $y = g(t)$ on the ty-plane (the wall facing us).

It is a curve in space, akin to a piece of wire:



Then the shadow of this wire on the floor is the graph $x = f(t)$ (light from above). If the light is behind us, the shadow on the wall in front is the graph $y = g(t)$. In addition, pointing a flashlight from right to left will produce the path of the parametric curve on the xy-plane.

This is the summary of the terminology:

types of functions:	general functions	numerical functions	parametric curves	motion
the set of all outputs:	image	range	path	trajectory

3.2. *Functions of two variables*

Any formula with two independent variables and one dependent variable can be studied in this manner:

$$a = wd \quad \text{or} \quad z = x + y\,,$$

Such an expression is called a *function of two variables*. The notation is as follows:

$$g(w, d) = wd \quad \text{or} \quad f(x, y) = x + y\,.$$

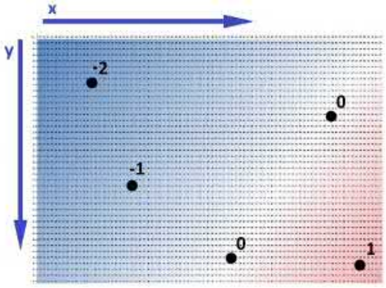
Example 3.2.1: function of two variables

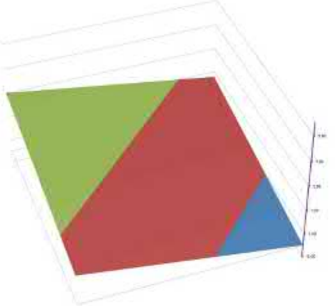
Let

$$f(x, y) = x + y\,.$$

We illustrate this new function below. First, by changing – independently — the two variables we create a table of numbers (left). We can furthermore color this array of cells (middle) so that the color of the (x, y)-cell is determined by the value of z:

x\y	0	10	20	30	40	50	60
0	0	10	20	30	40	50	60
10	10	20	30	40	50	60	70
20	20	30	40	50	60	70	80
30	30	40	50	60	70	80	90
40	40	50	60	70	80	90	100
50	50	60	70	80	90	100	110
60	60	70	80	90	100	110	120
70	70	80	90	100	110	120	130
80	80	90	100	110	120	130	140





The value of z can also be visualized as the elevation of a point at that location (right).

So, the main metaphor for a function of two variables will be *terrain*:

Each line indicates a constant elevation.

Example 3.2.2: distance

The distance formula for the Cartesian plane creates a function of two variables. This is the distance from a point (x, y) to the origin:

$$z = \sqrt{x^2 + y^2}\,.$$

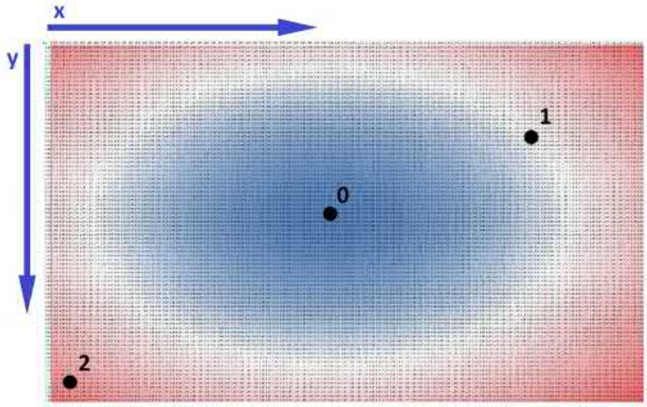
Slightly simpler is the square of the distance from a point (x, y) to the origin:

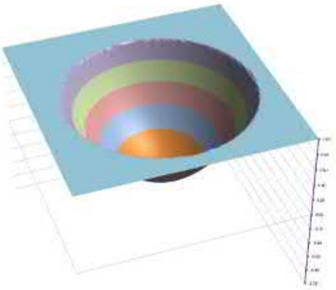
$$z = x^2 + y^2 \,.$$

We create a table of the values of the expression on the left in a spreadsheet with the formula:

`=RC1^2+R1C^2`

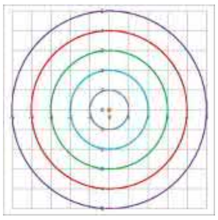
We then color the cells:





The negative values of z are in blue and the positive are in red. The circular pattern is clear.

The pattern seems to be made from concentric circles with the radius that varies with z:



For each z, we have a *relation* between x and y.

We also represent a function p diagrammatically as a *black box* that processes the inputs and produces the output:

$$\begin{tikzcd} \text{inputs} & \text{function} & \text{output} \\ x \arrow[dr] & & \\ & \boxed{p} \arrow[r] & z \\ y \arrow[ur] & & \end{tikzcd}$$

Instead, we would like to see a single input variable, (x, y), decomposed into two x and y to be processed by the function *at the same time*:

$$(x, y) \to \boxed{p} \to z$$

The difference from all the functions we have seen until now is the nature of the input.

So, even though we speak of *two* variables, the idea of function remains the same:

- ▶ There is a set (domain) and another (codomain) and the function assigns to each element of the former an element of the latter.

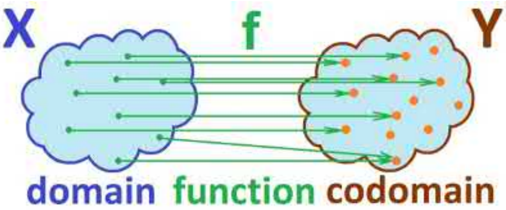
The idea is reflected in the notation we use:

$$F : X \to Z$$

or

$$X \xrightarrow{F} Z$$

A common way to visualize the concept of function – especially when the sets cannot be represented by mere lists – is to draw shapeless blobs connected by arrows:



In contrast to numerical function, however, *the domain is a subset of the* (x, y)*-plane.*

For example, we have for $f(x, y) = x + y$:

$$(0,0) \to 0,\ (0,1) \to 1,\ (1,0) \to 1,\ (1,1) \to 2,\ (1,2) \to 3,\ (2,1) \to 3 \ldots$$

Each arrow clearly identifies the *input* – an element of X – of this procedure by its beginning and the *output* – an element of Z – by its end.

This is the notation for the output of a function F when the input is x:

Input and output of function

$$F(x, y) = z$$

or

$$F : (x, y) \ \to \ z$$

It reads: "F of (x, y) is z".

We still have:

$$F(\text{ input }) = \text{ output}$$

and

$$F : \text{ input } \ \to \ \text{ output } .$$

Functions are *explicit relations.* There are *three* variables related to each other, but this relation is unequal: The two input variables come first and, therefore, the output is *dependent* on the input. That is why we say that the inputs are the *independent variables* while the output is the *dependent variable.*

Example 3.2.3: flowcharts represent functions

For example, for a given input (x, y), we first split it: x and y are the two *numerical* inputs. Then we do the following consecutively:

- add x and y,
- multiply by 2, and then
- square.

Such a procedure can be conveniently visualized with a "flowchart":

$$(x, y) \ \to \ \boxed{\quad x + y \quad} \ \to \ u \ \to \ \boxed{\quad u \cdot 2 \quad} \ \to \ z \ \to \ \boxed{\quad z^2 \quad} \ \to \ v$$

Functions of two variables come from many sources and can be expressed in different forms:

- a list of instructions (an algorithm)
- an algebraic formula
- a list of pairs of inputs and outputs
- a graph

- a transformation

An *algebraic representation* is exemplified by $z = x^2y$. In order to properly introduce this as a function, we give it a name, say f, and write:

$$f(x,y) = x^2y\,.$$

Let's examine this notation:

Function of two variables

$$\begin{array}{r|cccccccc} & z & = & f & (& x\ ,\ y &) & = & x^2y \\ & \uparrow & & \uparrow & & \uparrow\ \uparrow & & & \uparrow\uparrow \\ \text{name:} & \text{dependent} & & \text{function} & & \text{independent} & & & \text{independent} \\ & \text{variable} & & & & \text{variables} & & & \text{variables} \end{array}$$

Example 3.2.4: plug in values

Insert one input value in all of these boxes and the other in those circles. For example, this function:

$$f\,(x) = \frac{2x^2y - 3y + 7}{y^3 + 2x + 1}\ ,$$

can be understood and evaluated via this diagram:

$$f\,(\square) = \frac{2\square^2 \bigcirc -3 \bigcirc +7}{\bigcirc^3 + 2\square + 1}\,.$$

This is how $f(3,0)$ is evaluated:

$$f\left(\boxed{3}, \textcircled{0}\right) = \frac{2\boxed{3}^2\textcircled{0} - 3\textcircled{0} + 7}{\textcircled{0}^3 + 2\boxed{3} + 1}\,.$$

In summary,

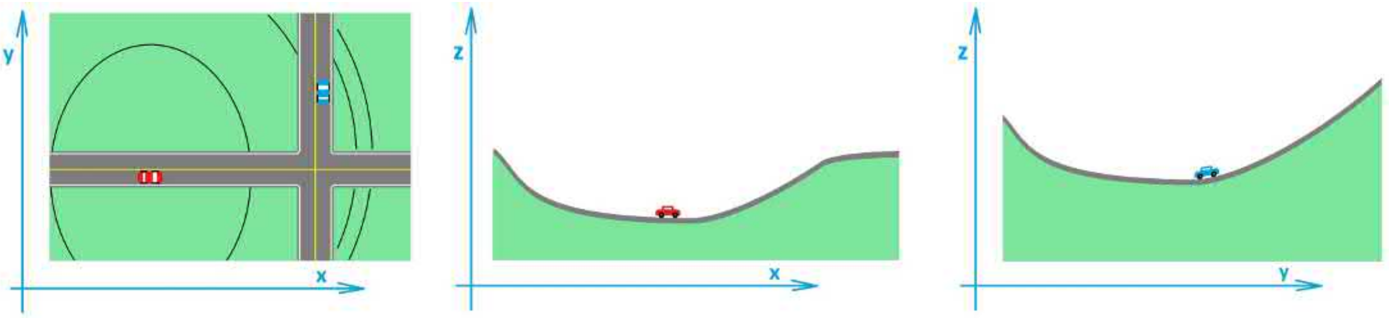
▶ "x" and "y" in a formula serve as a *placeholders* for: numbers, variables, and whole functions.

How do we study a function of two variables? We use what we know about functions of single variable.

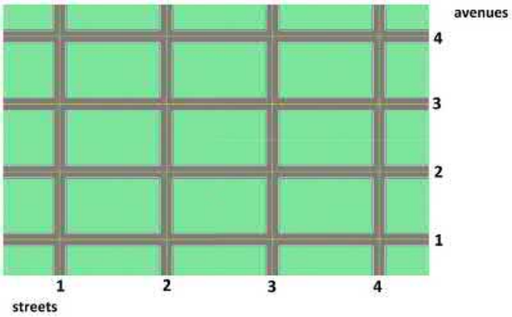
Above we looked at the curves of constant elevation of the surfaces. An alternative idea is a surveying method:

▶ In order to study a terrain we concentrate on the two main directions.

Imagine that we drive south-north and east-west separately watching how the elevation changes:

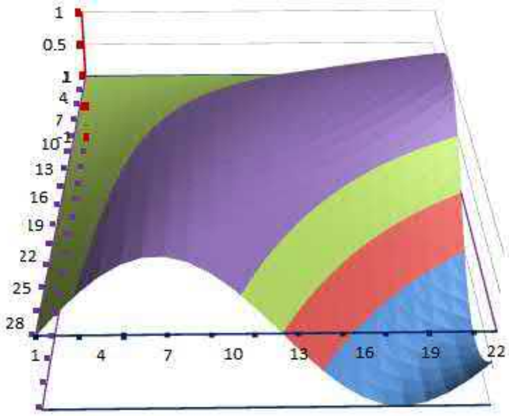


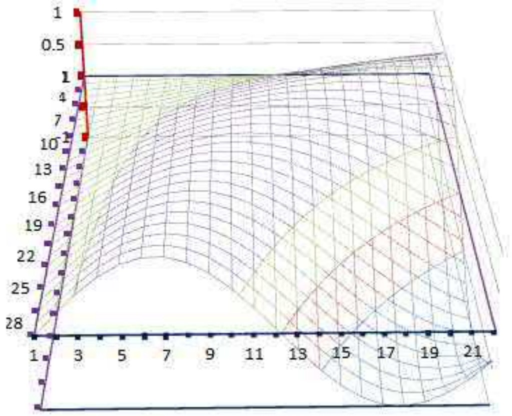
We can even imagine that we drive around a city on a hill and these trips follow the street grid:



Each of these trips creates a function of single variable, x or y.

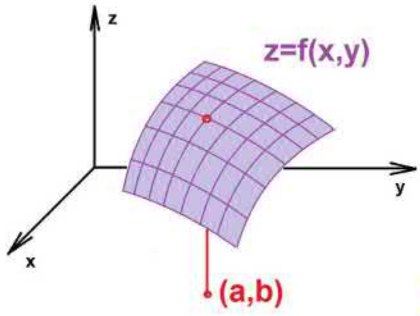
To visualize, consider the plot of $F(x,y) = \sin(xy)$ on the left:

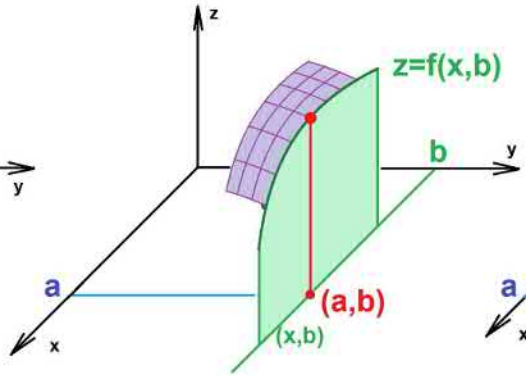


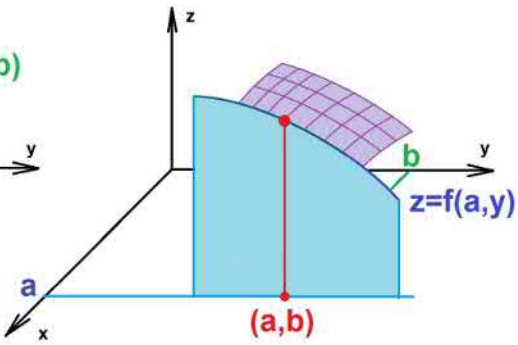


We plot the surface as a “wire-frame” on the right. Each wire is a separate trip.

The graphs of these functions are the slices cut by the vertical planes aligned with the axes from the surface that is the graph of F:



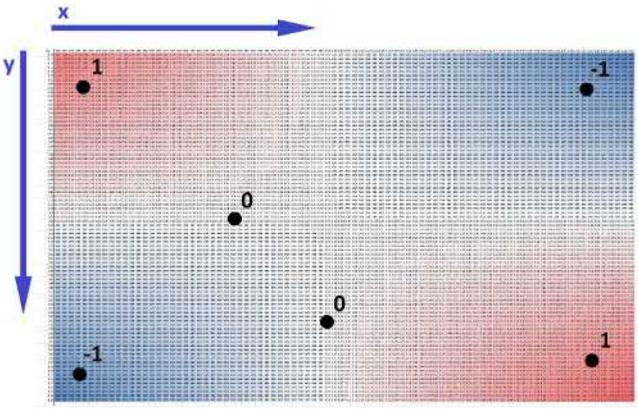


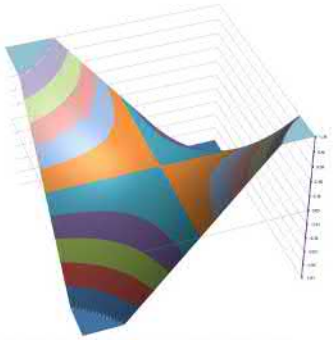


Example 3.2.5: saddle

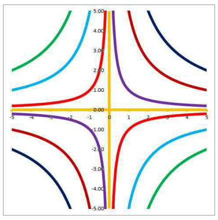
Let’s plot the graph of the function:

$$z = xy\,.$$





This is what the graphs of these relations look like plotted for various z's; they are curves called *hyperbolas*:

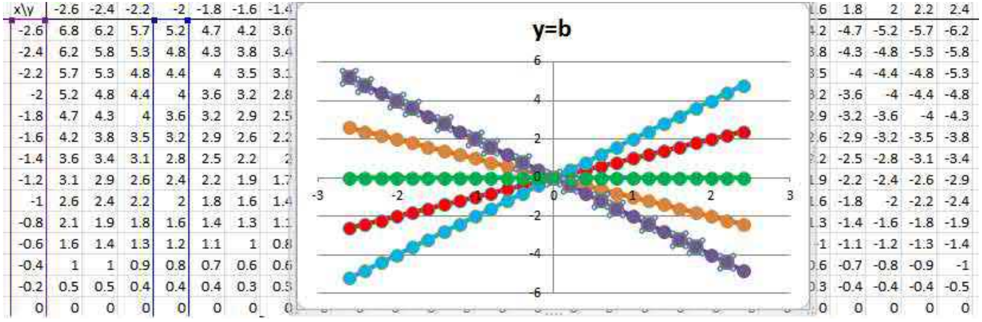


Instead, we fix one independent variable at a time.

We fix y first:

plane	equation	curve
$y=2$	$z=x\cdot 2$	line with slope 2
$y=1$	$z=x\cdot 1$	line with slope 1
$y=0$	$z=x\cdot 0=0$	line with slope 0
$y=-1$	$z=x\cdot(-1)$	line with slope 1
$y=-2$	$z=x\cdot(-2)$	line with slope -2

The view shown below is from the direction of the y-axis:



x\y	-2.6	-2.4	-2.2	-2	-1.8	-1.6	-1.4
-2.6	6.8	6.2	5.7	5.2	4.7	4.2	3.6
-2.4	6.2	5.8	5.3	4.8	4.3	3.8	3.4
-2.2	5.7	5.3	4.8	4.4	4	3.5	3.1
-2	5.2	4.8	4.4	4	3.6	3.2	2.8
-1.8	4.7	4.3	4	3.6	3.2	2.9	2.5
-1.6	4.2	3.8	3.5	3.2	2.9	2.6	2.2
-1.4	3.6	3.4	3.1	2.8	2.5	2.2	2
-1.2	3.1	2.9	2.6	2.4	2.2	1.9	1.7
-1	2.6	2.4	2.2	2	1.8	1.6	1.4
-0.8	2.1	1.9	1.8	1.6	1.4	1.3	1.1
-0.6	1.6	1.4	1.3	1.2	1.1	1	0.8
-0.4	1	1	0.9	0.8	0.7	0.6	0.6
-0.2	0.5	0.5	0.4	0.4	0.4	0.3	0.3
0	0	0	0	0	0	0	0

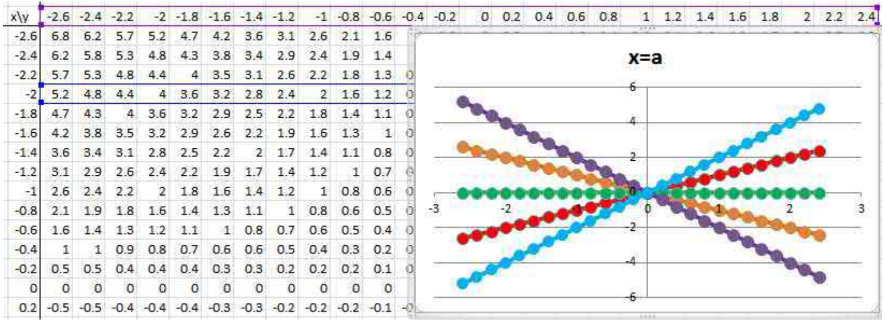
1.8	2	2.2	2.4
-4.7	-5.2	-5.7	-6.2
-4.3	-4.8	-5.3	-5.8
-4	-4.4	-4.8	-5.3
-3.6	-4	-4.4	-4.8
-3.2	-3.6	-4	-4.3
-2.9	-3.2	-3.5	-3.8
-2.5	-2.8	-3.1	-3.4
-2.2	-2.4	-2.6	-2.9
-1.8	-2	-2.2	-2.4
-1.4	-1.6	-1.8	-1.9
-1.1	-1.2	-1.3	-1.4
-0.7	-0.8	-0.9	-1
-0.4	-0.4	-0.4	-0.5
0	0	0	0

The data for each line comes from the x-column of the spreadsheet and one of the z-columns. These lines give the lines of elevation of this terrain in a particular, say, east-west direction. This is equivalent to cutting the graph by a vertical plane parallel to the xz-plane.

We fix x second:

plane	equation	curve
$x=2$	$z=2\cdot y$	line with slope 2
$x=1$	$z=1\cdot y$	line with slope 1
$x=0$	$z=0\cdot y=0$	line with slope 0
$x=-1$	$z=(-1)\cdot y$	line with slope 1
$x=-2$	$z=(-2)\cdot y$	line with slope -2

This is equivalent to cutting the graph by a vertical plane parallel to the yz-plane. The view shown below is from the direction of the x-axis:



x\y	-2.6	-2.4	-2.2	-2	-1.8	-1.6	-1.4	-1.2	-1	-0.8	-0.6
-2.6	6.8	6.2	5.7	5.2	4.7	4.2	3.6	3.1	2.6	2.1	1.6
-2.4	6.2	5.8	5.3	4.8	4.3	3.8	3.4	2.9	2.4	1.9	1.4
-2.2	5.7	5.3	4.8	4.4	4	3.5	3.1	2.6	2.2	1.8	1.3
-2	5.2	4.8	4.4	4	3.6	3.2	2.8	2.4	2	1.6	1.2
-1.8	4.7	4.3	4	3.6	3.2	2.9	2.5	2.2	1.8	1.4	1.1
-1.6	4.2	3.8	3.5	3.2	2.9	2.6	2.2	1.9	1.6	1.3	1
-1.4	3.6	3.4	3.1	2.8	2.5	2.2	2	1.7	1.4	1.1	0.8
-1.2	3.1	2.9	2.6	2.4	2.2	1.9	1.7	1.4	1.2	1	0.7
-1	2.6	2.4	2.2	2	1.8	1.6	1.4	1.2	1	0.8	0.6
-0.8	2.1	1.9	1.8	1.6	1.4	1.3	1.1	1	0.8	0.6	0.5
-0.6	1.6	1.4	1.3	1.2	1.1	1	0.8	0.7	0.6	0.5	0.4
-0.4	1	1	0.9	0.8	0.7	0.6	0.6	0.5	0.4	0.3	0.2
-0.2	0.5	0.5	0.4	0.4	0.4	0.3	0.3	0.2	0.2	0.2	0.1
0	0	0	0	0	0	0	0	0	0	0	0
0.2	-0.5	-0.5	-0.4	-0.4	-0.4	-0.3	-0.3	-0.2	-0.2	-0.2	-0.1

The data for each line comes from the y-row of the spreadsheet and one of the z-rows. These lines give the lines of elevation of this terrain in a particular, say, north-south direction.

Exercise 3.2.6

Provide a similar analysis for $f(x,y)=3x+2y$.

Example 3.2.7: baker

We will take a look at the example in the last section from a different angle. The time t is not a part of our consideration anymore but we retain the two variables representing the two *commodities*:

- x is the price of wheat.
- y is the price of sugar.

We also add a *product* to the setup:

- z is the price of a loaf of bread.

What is the relation between these three? As those two are the two major ingredients in bread, we will assume that

► z depends on x and y.

One can imagine a baker who every morning, upon receiving the updated prices of wheat and sugar, uses a *table* that he made up in advance to decide on the price of his bread for the rest of the day. Let's see how such a table might come about.

What kind of dependencies are these? Increasing prices of the ingredients in creases the cost and ultimately the price of the product:

$$\begin{array}{c} x \nearrow \Longrightarrow z \nearrow \\ y \nearrow \Longrightarrow z \nearrow \end{array}$$

At its simplest, such an increase is linear. In addition to some fixed costs,

- Each increase of x leads to a proportional increase of z.
- Each increase of y leads to a proportional increase of z.

Independently! A simple formula that captures this dependence may be this:

$$z = p(x, y) = 2x + y + 1\,.$$

In order to visualize this function, we compute a few of its values:

- $p(0, 0) = 1$
- $p(0, 1) = 2$
- $p(0, 2) = 3$
- $p(1, 0) = 3$
- $p(1, 1) = 4$
- etc.

Even though this is a list, we realize that the input variables don't fit into a list comfortably... they form a *table*!

$$\begin{array}{cccc} (0,0) & (1,0) & (2,0) & ... \\ (0,1) & (1,1) & (2,1) & ... \\ (0,2) & (1,2) & (2,2) & ... \\ ... & ... & ... & ... \end{array}$$

In fact, we can align these pairs with x in each column and y in each row:

$y\backslash x$	0	1	2	...
0	$(0,0)$	$(1,0)$	$(2,0)$	...
1	$(0,1)$	$(1,1)$	$(2,1)$	...
2	$(0,2)$	$(1,2)$	$(2,2)$	...
...	...	...	...	...

Now, the values, $z = p(x, y)$:

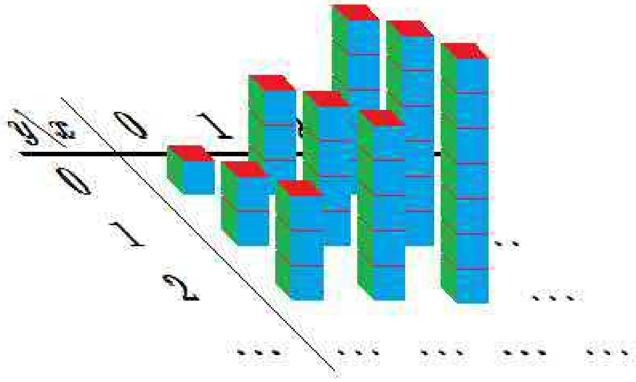
$y\backslash x$	0	1	2	...
0	1	3	5	...
1	2	4	6	...
2	3	5	7	...
...	...	...	...	...

That's what baker's table might look like...

Let's bring these two together:

$y\backslash x$	0		1		2		...
0	$(0,0)$		$(1,0)$		$(2,0)$		...
		$\searrow$		$\searrow$		$\searrow$	...
		1		3		5	...
1	$(0,1)$		$(1,1)$		$(2,1)$		...
		$\searrow$		$\searrow$		$\searrow$	...
		2		4		6	...
2	$(0,2)$		$(1,2)$		$(2,2)$		...
		$\searrow$		$\searrow$		$\searrow$	...
		3		5		7	...
...	...		...		...		...

In the past, we have visualized numerical functions by putting *bars* on top of the x-axis. Now, we visualize the values by building *columns* with appropriate heights on top of the xy-plane:



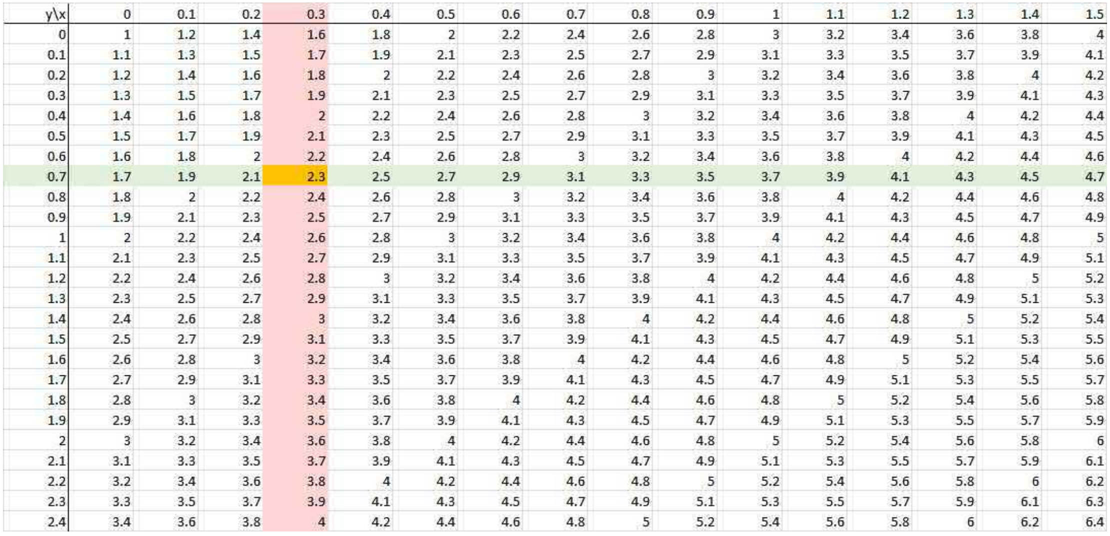
Notice that by fixing one of the variables – $x = 0, 1, 2$ or $y = 0, 1, 2$ – we create a function of *one* variable with respect to the other variable. We fix x below and extract the *columns* from the table:

$$x = 0 : \begin{array}{|c|c} y & z \\ \hline 0 & 1 \\ 1 & 2 \\ 2 & 3 \end{array}, \quad x = 1 : \begin{array}{|c|c} y & z \\ \hline 0 & 3 \\ 1 & 4 \\ 2 & 5 \end{array}, \quad x = 2 : \begin{array}{|c|c} y & z \\ \hline 0 & 5 \\ 1 & 6 \\ 2 & 7 \end{array}.$$

A *pattern* is clear: growth by 1. We next fix y and extract the *rows* from the table:

$$y = 0 : \begin{array}{c|ccc} \hline x & 0 & 1 & 2 \\ \hline z & 1 & 3 & 5 \\ \hline \end{array}, \quad y = 1 : \begin{array}{c|ccc} \hline x & 0 & 1 & 2 \\ \hline z & 2 & 4 & 6 \\ \hline \end{array}, \quad y = 2 : \begin{array}{c|ccc} \hline x & 0 & 1 & 2 \\ \hline z & 3 & 5 & 7 \\ \hline \end{array}$$

A *pattern* is clear: growth by 2. We have the total of six (linear) functions!
Let's do the same with a spreadsheet. This is the data:

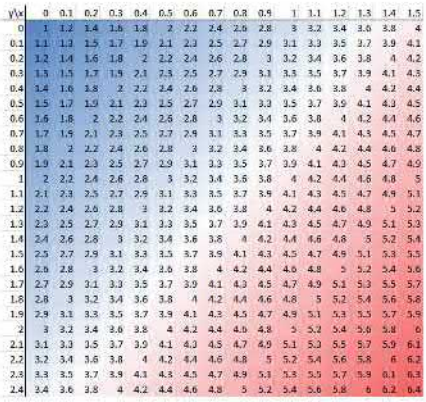


y\x	0	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1	1.1	1.2	1.3	1.4	1.5
0	1	1.2	1.4	1.6	1.8	2	2.2	2.4	2.6	2.8	3	3.2	3.4	3.6	3.8	4
0.1	1.1	1.3	1.5	1.7	1.9	2.1	2.3	2.5	2.7	2.9	3.1	3.3	3.5	3.7	3.9	4.1
0.2	1.2	1.4	1.6	1.8	2	2.2	2.4	2.6	2.8	3	3.2	3.4	3.6	3.8	4	4.2
0.3	1.3	1.5	1.7	1.9	2.1	2.3	2.5	2.7	2.9	3.1	3.3	3.5	3.7	3.9	4.1	4.3
0.4	1.4	1.6	1.8	2	2.2	2.4	2.6	2.8	3	3.2	3.4	3.6	3.8	4	4.2	4.4
0.5	1.5	1.7	1.9	2.1	2.3	2.5	2.7	2.9	3.1	3.3	3.5	3.7	3.9	4.1	4.3	4.5
0.6	1.6	1.8	2	2.2	2.4	2.6	2.8	3	3.2	3.4	3.6	3.8	4	4.2	4.4	4.6
0.7	1.7	1.9	2.1	2.3	2.5	2.7	2.9	3.1	3.3	3.5	3.7	3.9	4.1	4.3	4.5	4.7
0.8	1.8	2	2.2	2.4	2.6	2.8	3	3.2	3.4	3.6	3.8	4	4.2	4.4	4.6	4.8
0.9	1.9	2.1	2.3	2.5	2.7	2.9	3.1	3.3	3.5	3.7	3.9	4.1	4.3	4.5	4.7	4.9
1	2	2.2	2.4	2.6	2.8	3	3.2	3.4	3.6	3.8	4	4.2	4.4	4.6	4.8	5
1.1	2.1	2.3	2.5	2.7	2.9	3.1	3.3	3.5	3.7	3.9	4.1	4.3	4.5	4.7	4.9	5.1
1.2	2.2	2.4	2.6	2.8	3	3.2	3.4	3.6	3.8	4	4.2	4.4	4.6	4.8	5	5.2
1.3	2.3	2.5	2.7	2.9	3.1	3.3	3.5	3.7	3.9	4.1	4.3	4.5	4.7	4.9	5.1	5.3
1.4	2.4	2.6	2.8	3	3.2	3.4	3.6	3.8	4	4.2	4.4	4.6	4.8	5	5.2	5.4
1.5	2.5	2.7	2.9	3.1	3.3	3.5	3.7	3.9	4.1	4.3	4.5	4.7	4.9	5.1	5.3	5.5
1.6	2.6	2.8	3	3.2	3.4	3.6	3.8	4	4.2	4.4	4.6	4.8	5	5.2	5.4	5.6
1.7	2.7	2.9	3.1	3.3	3.5	3.7	3.9	4.1	4.3	4.5	4.7	4.9	5.1	5.3	5.5	5.7
1.8	2.8	3	3.2	3.4	3.6	3.8	4	4.2	4.4	4.6	4.8	5	5.2	5.4	5.6	5.8
1.9	2.9	3.1	3.3	3.5	3.7	3.9	4.1	4.3	4.5	4.7	4.9	5.1	5.3	5.5	5.7	5.9
2	3	3.2	3.4	3.6	3.8	4	4.2	4.4	4.6	4.8	5	5.2	5.4	5.6	5.8	6
2.1	3.1	3.3	3.5	3.7	3.9	4.1	4.3	4.5	4.7	4.9	5.1	5.3	5.5	5.7	5.9	6.1
2.2	3.2	3.4	3.6	3.8	4	4.2	4.4	4.6	4.8	5	5.2	5.4	5.6	5.8	6	6.2
2.3	3.3	3.5	3.7	3.9	4.1	4.3	4.5	4.7	4.9	5.1	5.3	5.5	5.7	5.9	6.1	6.3
2.4	3.4	3.6	3.8	4	4.2	4.4	4.6	4.8	5	5.2	5.4	5.6	5.8	6	6.2	6.4

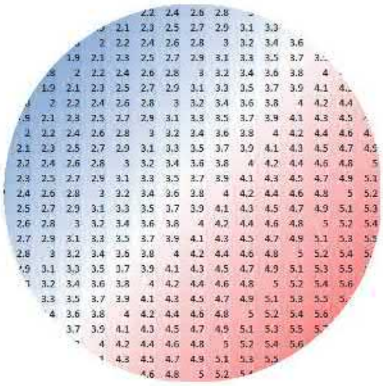
The value in each cell is computed from the corresponding value of x (all the way up) and from the corresponding value of y (all the way left). This is the formula:

`=2*R3C+RC2+1` .

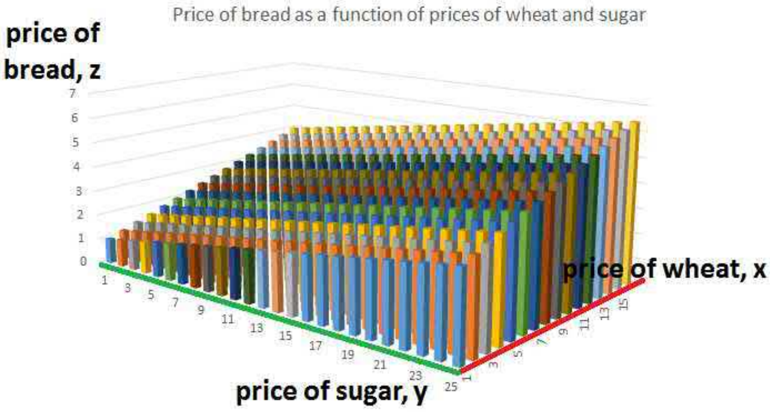
The simplest way to visualize is by coloring the cell depending on the values (common in cartography: elevation, temperature, humidity, precipitation, population density, etc.:



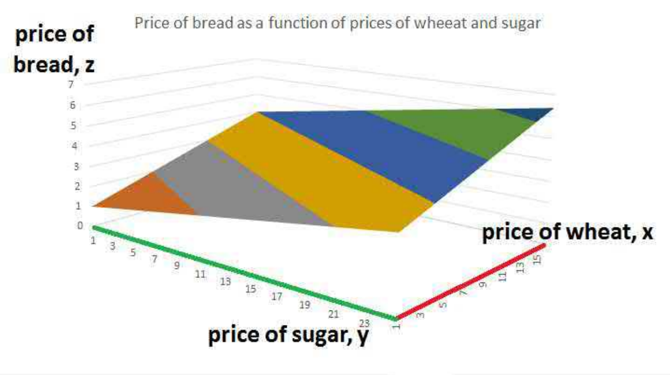
y\x	0	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1	1.1	1.2	1.3	1.4	1.5
0	1	1.2	1.4	1.6	1.8	2	2.2	2.4	2.6	2.8	3	3.2	3.4	3.6	3.8	4
0.1	1.1	1.3	1.5	1.7	1.9	2.1	2.3	2.5	2.7	2.9	3.1	3.3	3.5	3.7	3.9	4.1
0.2	1.2	1.4	1.6	1.8	2	2.2	2.4	2.6	2.8	3	3.2	3.4	3.6	3.8	4	4.2
0.3	1.3	1.5	1.7	1.9	2.1	2.3	2.5	2.7	2.9	3.1	3.3	3.5	3.7	3.9	4.1	4.3
0.4	1.4	1.6	1.8	2	2.2	2.4	2.6	2.8	3	3.2	3.4	3.6	3.8	4	4.2	4.4
0.5	1.5	1.7	1.9	2.1	2.3	2.5	2.7	2.9	3.1	3.3	3.5	3.7	3.9	4.1	4.3	4.5
0.6	1.6	1.8	2	2.2	2.4	2.6	2.8	3	3.2	3.4	3.6	3.8	4	4.2	4.4	4.6
0.7	1.7	1.9	2.1	2.3	2.5	2.7	2.9	3.1	3.3	3.5	3.7	3.9	4.1	4.3	4.5	4.7
0.8	1.8	2	2.2	2.4	2.6	2.8	3	3.2	3.4	3.6	3.8	4	4.2	4.4	4.6	4.8
0.9	1.9	2.1	2.3	2.5	2.7	2.9	3.1	3.3	3.5	3.7	3.9	4.1	4.3	4.5	4.7	4.9
1	2	2.2	2.4	2.6	2.8	3	3.2	3.4	3.6	3.8	4	4.2	4.4	4.6	4.8	5
1.1	2.1	2.3	2.5	2.7	2.9	3.1	3.3	3.5	3.7	3.9	4.1	4.3	4.5	4.7	4.9	5.1
1.2	2.2	2.4	2.6	2.8	3	3.2	3.4	3.6	3.8	4	4.2	4.4	4.6	4.8	5	5.2
1.3	2.3	2.5	2.7	2.9	3.1	3.3	3.5	3.7	3.9	4.1	4.3	4.5	4.7	4.9	5.1	5.3
1.4	2.4	2.6	2.8	3	3.2	3.4	3.6	3.8	4	4.2	4.4	4.6	4.8	5	5.2	5.4
1.5	2.5	2.7	2.9	3.1	3.3	3.5	3.7	3.9	4.1	4.3	4.5	4.7	4.9	5.1	5.3	5.5
1.6	2.6	2.8	3	3.2	3.4	3.6	3.8	4	4.2	4.4	4.6	4.8	5	5.2	5.4	5.6
1.7	2.7	2.9	3.1	3.3	3.5	3.7	3.9	4.1	4.3	4.5	4.7	4.9	5.1	5.3	5.5	5.7
1.8	2.8	3	3.2	3.4	3.6	3.8	4	4.2	4.4	4.6	4.8	5	5.2	5.4	5.6	5.8
1.9	2.9	3.1	3.3	3.5	3.7	3.9	4.1	4.3	4.5	4.7	4.9	5.1	5.3	5.5	5.7	5.9
2	3	3.2	3.4	3.6	3.8	4	4.2	4.4	4.6	4.8	5	5.2	5.4	5.6	5.8	6
2.1	3.1	3.3	3.5	3.7	3.9	4.1	4.3	4.5	4.7	4.9	5.1	5.3	5.5	5.7	5.9	6.1
2.2	3.2	3.4	3.6	3.8	4	4.2	4.4	4.6	4.8	5	5.2	5.4	5.6	5.8	6	6.2
2.3	3.3	3.5	3.7	3.9	4.1	4.3	4.5	4.7	4.9	5.1	5.3	5.5	5.7	5.9	6.1	6.3
2.4	3.4	3.6	3.8	4	4.2	4.4	4.6	4.8	5	5.2	5.4	5.6	5.8	6	6.2	6.4

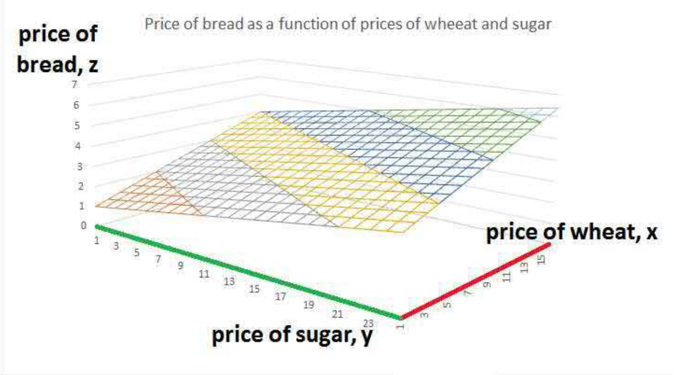


The growth is visible: it grows the most in some diagonal direction but it's *not* 45 degrees...
We can also visualize with bar chart, just as before:



If we used bars to represent the Riemann sums to compute the *area*, here we are after the *volume*... The most common way, however, to visualize a function of two variables in mathematics is with its *graph*, which, in this case, is a surface:





In this particular case, this is a *plane*. The second graph is the same surface but displayed as a wire-frame (or even a wire-fence). The wires are the graphs of those linear functions of one variable created from our function when we fix one variable at a time. Each of these wires comes from choosing either:

- the row of x's (top) and one other row in the table, or
- the column of y's (leftmost) and one other column in the table.

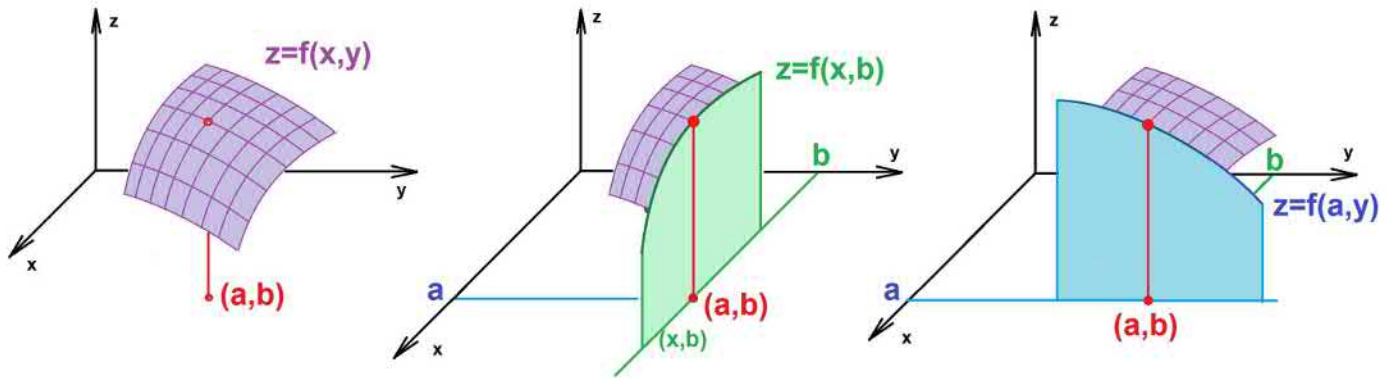
Exercise 3.2.8

Provide a similar analysis for (a) the wind-chill and (b) the heat index.

The functions of one variable created from our function $z = p(x, y)$ when we fix one variable at a time are:

$$\begin{array}{rcl} y = b & \longrightarrow & f_b(x) = p(x, b); \\ x = a & \longrightarrow & g_a(y) = p(a, y). \end{array}$$

There are infinitely many of them. Their graphs are the slices – along the axes – of the surface that is the graph of F.



Therefore, the monotonicity of these functions tells us about the monotonicity of p – in the directions of the axes!

Functions of two variables are functions...

This idea comes with certain questions to be answered. What is the input, the *independent variable*? Taking a clue from our analysis of parametric curves, we answer: it is the "combination" of the two inputs of the function, i.e., x and y that form a pair, $X = (x, y)$, which is is a point on the xy-plane. What is the output, the *dependent variable*? It is z.

We represent a function p diagrammatically as a *black box* that processes the input and produces the output:

$$\begin{tikzcd} \text{inputs} & \text{function} & \text{output} \\ x \arrow[dr] & & \\ & \boxed{p} \arrow[r, mapsto] & z \\ y \arrow[ur] & & \end{tikzcd}$$

Instead, we would like to see a single input variable, (x, y), decomposed into two x and y to be processed by the function *at the same time*:

$$(x, y) \ \to \ \boxed{\quad p \quad} \ \to \ z$$

The difference from all the functions we have seen until now is the nature of the input.

Next, what is the *domain* of p? It is supposed to be a recording of all possible inputs, i.e., all pairs (x, y) for which the output $z = p(x, y)$ of the function makes sense. This requirement create a subset of the xy-plane and, therefore, a relation between x and y.

What about the image, i.e., the range of values of p? It is a recording of all possible outputs of p.

Definition 3.2.9:

The *image* of a function of two variables $z = p(x, y)$ is the set of all such values z on the z-axis.

What about the graph of $p = (f, g)$? It is supposed to be a recording of all possible combinations of inputs and outputs of F.

Definition 3.2.10:

The *graph* of a function of two variables $z = p(x, y)$ is the set of all such points $\big(x, y, p(x, y)\big)$ in the xyz-space.

3.3. Transforming the axes transforms the plane

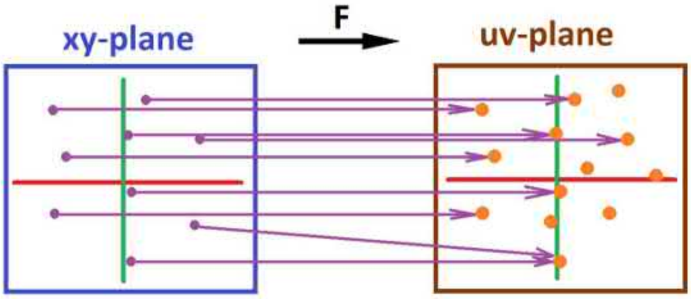
We will consider transformations of the plane as a function:

$$F : \mathbf{R}^2 \to \mathbf{R}^2 \,.$$

We know that numerical functions transform the real line:

$$f : \mathbf{R} \to \mathbf{R} \,.$$

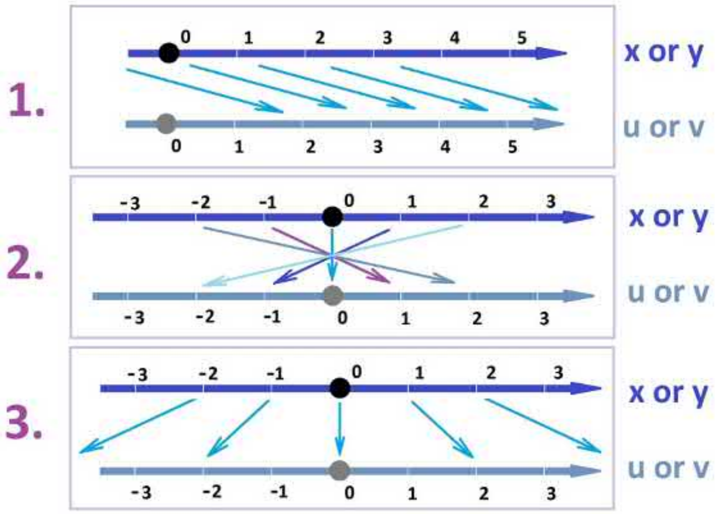
They, therefore, transform the *axes* of the xy-plane:



We narrow this down:

- ▶ How do the transformations of the axes – horizontal and vertical – affect the xy-plane?

Let's review what we know. We have these three basic transformations of *an* axis: shift, flip, and stretch. We now think of them as transformations of the x-axis to the u-axis and transformations of the y-axis to the v-axis:

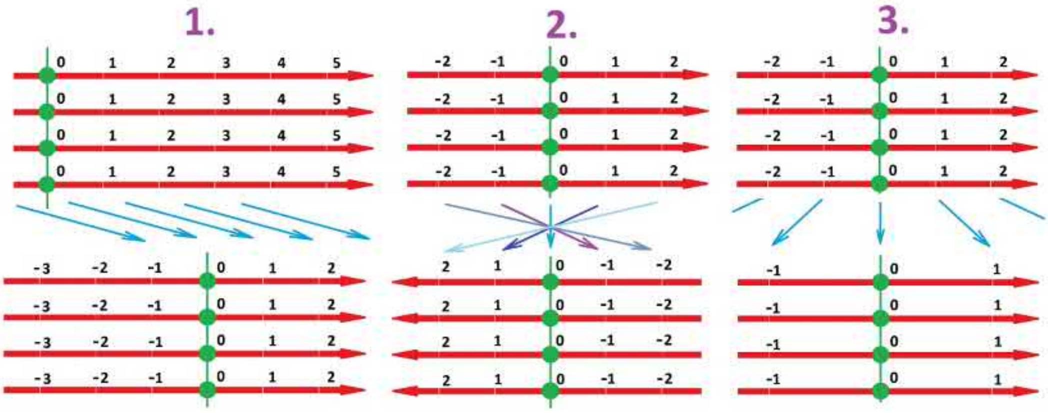


This is the algebra:

$$\begin{array}{ccc|ccc}
x & \xrightarrow{\text{shift by } s} & x+s & y & \xrightarrow{\text{shift by } s} & y+s \\
x & \xrightarrow{\text{flip}} & -x & y & \xrightarrow{\text{flip}} & -y \\
x & \xrightarrow{\text{stretch by } k} & x \cdot k & y & \xrightarrow{\text{stretch by } k} & y \cdot k
\end{array}$$

Now, let's imagine that transforming an axis transforms – in unison – all the lines on the plane parallel to it.

The x-axis and its transformations are *horizontal* and so are the transformations of the xy-plane:

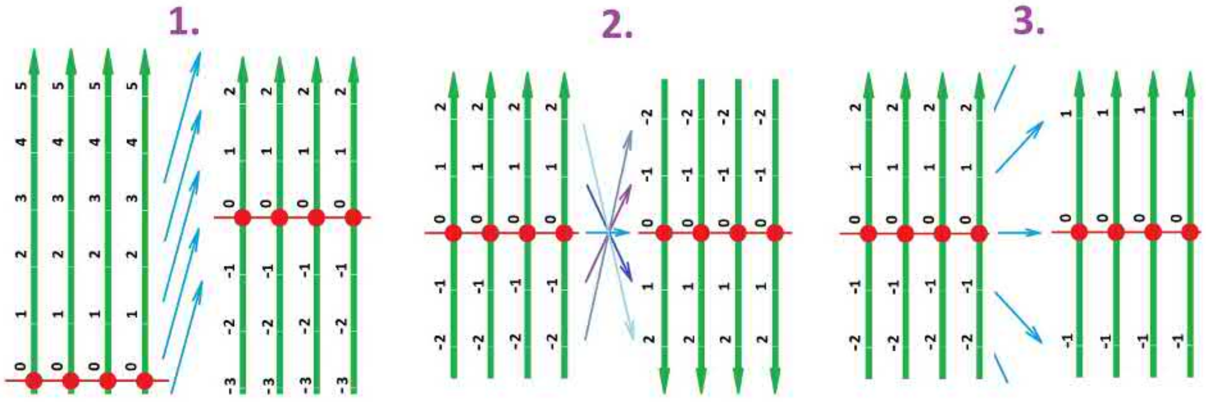


Then, the shift of the x-axis becomes a horizontal shift of the xy-plane, the flip of the x-axis becomes a horizontal flip of the xy-plane (around the vertical axis), and the stretch of the x-axis becomes a horizontal stretch of the xy-plane (away from the vertical axis).

For an algebraic representation of these transformations, we just add y, that remains unchanged, to the formula, as follows:

$$\begin{array}{ccccccc}
x\text{-axis} & & u\text{-axis} & & xy\text{-plane} & & uv\text{-plane}\\
\hline
x & \xrightarrow{\text{shift by } s} & x+s & \Longrightarrow & (x,y) & \xrightarrow{\text{horizontal shift by } s} & (x+s,y)\\
x & \xrightarrow{\text{flip}} & -x & \Longrightarrow & (x,y) & \xrightarrow{\text{horizontal flip}} & (-x,y)\\
x & \xrightarrow{\text{stretch by } k} & x\cdot k & \Longrightarrow & (x,y) & \xrightarrow{\text{horizontal stretch by } k} & (x\cdot k,y)
\end{array}$$

What about the y-axis? The y-axis and its transformations are *vertical* and so are the transformations of the xy-plane:



The shift of the y-axis produces a vertical shift of the xy-plane, the flip of the y-axis produces a vertical flip of the xy-plane (around the horizontal axis), and the stretch of the y-axis produces a vertical stretch of the xy-plane (away from the horizontal axis).

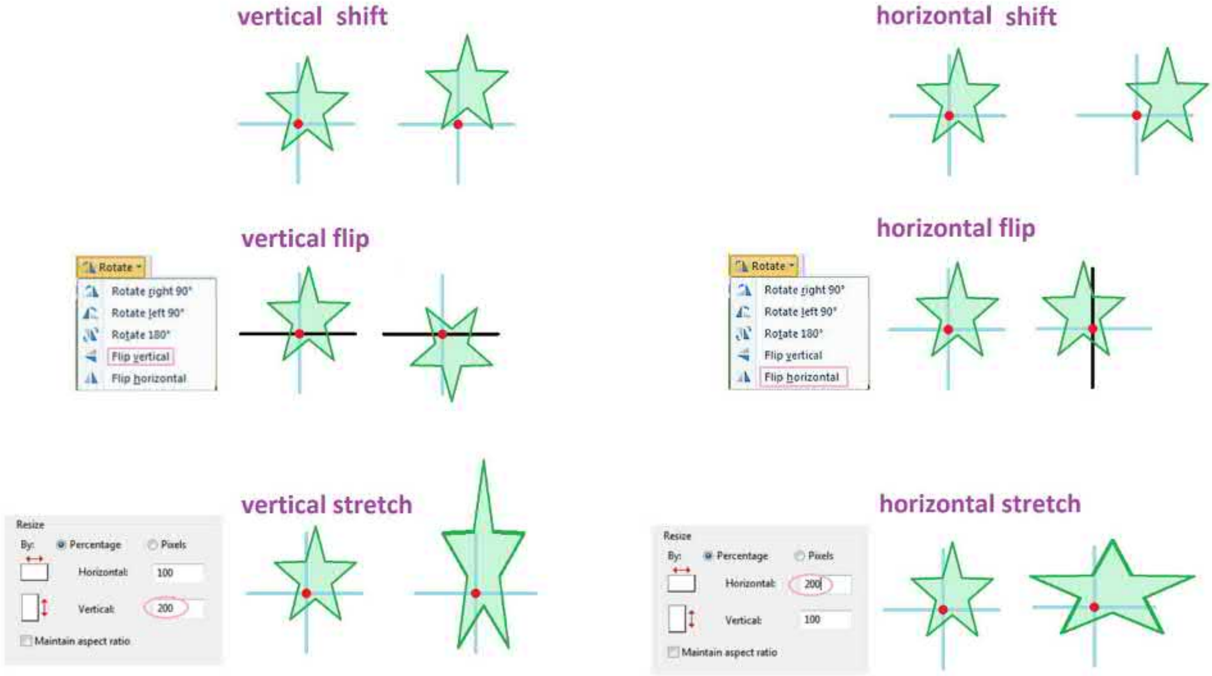
For an algebraic representation of these transformations, we just add x, that remains unchanged, to the formula:

$$\begin{array}{ccccccc}
x\text{-axis} & & u\text{-axis} & & xy\text{-plane} & & uv\text{-plane}\\
\hline
y & \xrightarrow{\text{shift by } s} & y+s & \Longrightarrow & (x,y) & \xrightarrow{\text{vertical shift by } s} & (x,y+s)\\
y & \xrightarrow{\text{flip}} & -y & \Longrightarrow & (x,y) & \xrightarrow{\text{vertical flip}} & (x,-y)\\
y & \xrightarrow{\text{stretch by } k} & y\cdot k & \Longrightarrow & (x,y) & \xrightarrow{\text{vertical stretch by } k} & (x,y\cdot k)
\end{array}$$

Horizontal transformations don't change y and vertical don't change x!

Example 3.3.1: transformations with computer graphics

We can illustrate these transformations with a graphics editor:



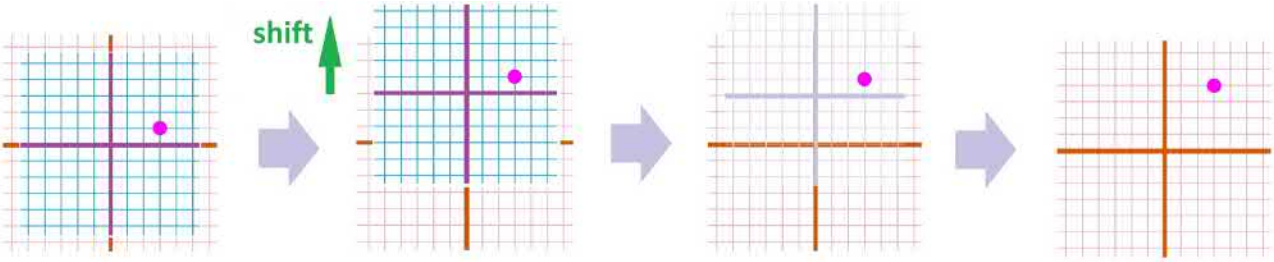
Resize — By: ● Percentage ○ Pixels; Horizontal: 100; Vertical: 200; ☐ Maintain aspect ratio

Resize — By: ● Percentage ○ Pixels; Horizontal: 200; Vertical: 100; ☐ Maintain aspect ratio

The first two rows show *rigid motions*, while the last is re-scaling.

So, the algebra of the real line creates a new algebra of the Cartesian plane. Let's revisit these six transformations one by one.

We start with a *vertical shift*. We shift the whole xy-plane as if it is printed on a sheet of paper. Furthermore, there is another sheet of paper underneath used for reference. It is to the second sheet that we transfer the resulting points. We then use its coordinate system to record the coordinates of the new point. For example, a shift of 3 units upward is shown below:



So, all vertical lines are shifted up by s. Then, the whole plane is shifted $s > 0$ units up. A generic point (x, y) makes a step up/down by s and becomes $(x, y + s)$. This is another algebraic way to present the transformation:

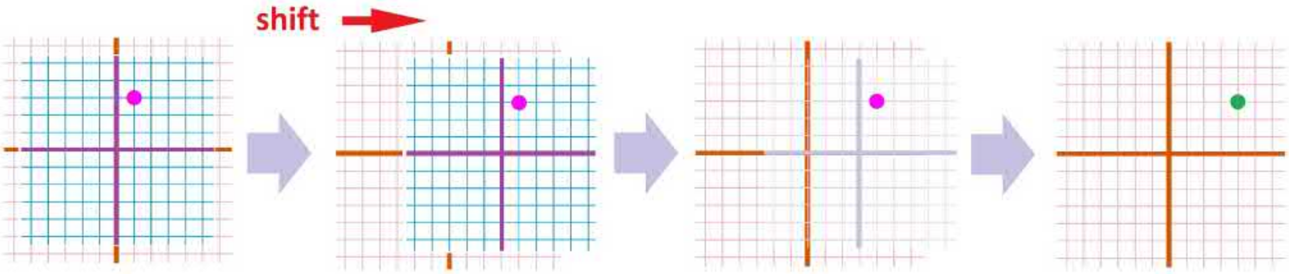
$$\boxed{\begin{array}{ccl} x & \xrightarrow{\text{nothing}} & u = x \\ y & \xrightarrow{\text{shift } s} & v = y + s \\ (x, y) & \xrightarrow{\text{up } s} & (u, v) = F(x, y) = (x, y + s) \end{array}}$$

It is as if the algebra of the flip of the y-axis given previously, $y \mapsto y + s$, is copied and then paired up with x.

Exercise 3.3.2

What is the effect of two vertical stretches executed consecutively?

What about the *horizontal shift*? For example, a shift of 2 units right is shown below:



So, all horizontal lines are shifted right by s. Then, the whole plane is shifted $s > 0$ units right. A generic point (x, y) makes a step right/left by s and becomes $(x + s, y)$. This is an algebraic way to present the transformation:

$$\begin{array}{ccc} x & \xrightarrow{\text{shift } s} & u = x + s \\ y & \xrightarrow{\text{nothing}} & v = y \\ (x, y) & \xrightarrow{\text{right } s} & (u, v) = F(x, y) = (x + s, y) \end{array}$$

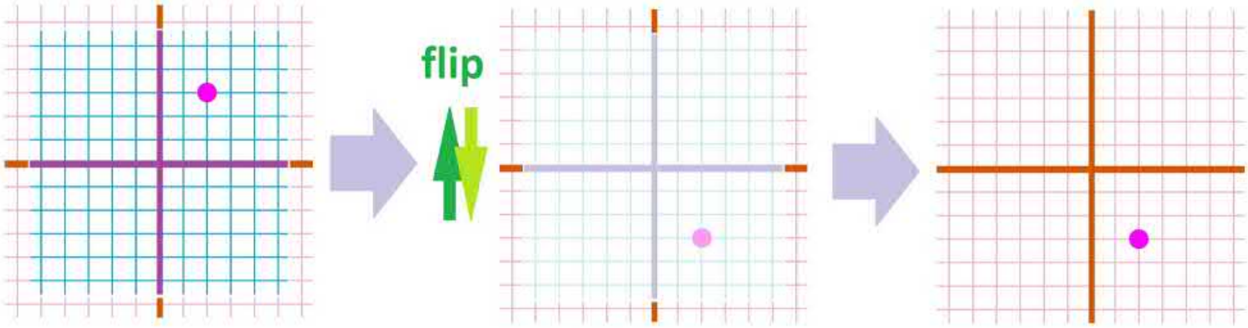
It is as if the algebra of the flip of the x-axis given previously, $x \to x + s$, is copied and then paired up with y.

Exercise 3.3.3

What is the effect of a vertical stretch and a horizontal stretch executed consecutively? What if we change the order?

These shifts can also be described as *a translation along the y-axis* and *a translation along the x-axis*, respectively.

Now a *vertical flip*. We lift, then flip the sheet of paper with the xy-plane on it, and finally place it on top of another such sheet so that the x-axes align. This flip is shown below:



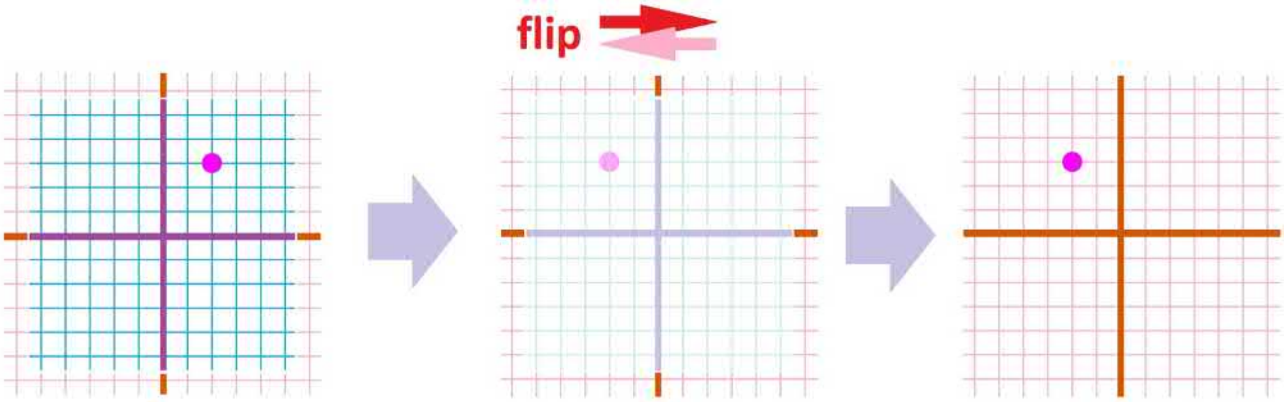
So, all vertical lines are flipped about their origins. Then, the whole plane is flipped about the x-axis. A generic point (x, y) jumps across the x-axis and becomes $(x, -y)$. This is the algebraic outcome:

$$\begin{array}{ccc} x & \xrightarrow{\text{nothing}} & u = x \\ y & \xrightarrow{\text{flip}} & v = -y \\ (x, y) & \xrightarrow{\text{vertical flip}} & (u, v) = F(x, y) = (x, -y) \end{array}$$

Exercise 3.3.4

What is the effect of two vertical flips executed consecutively?

For the *horizontal flip*, we lift, then flip the sheet of paper with the xy-plane on it, and finally place it on top of another such sheet so that the y-axes align. This flip is shown below:



So, all horizontal lines are flipped about their origins. Then, the whole plane is flipped about the y-axis. A generic point (x, y) jumps across the y-axis and becomes $(-x, y)$. This is the algebraic outcome:

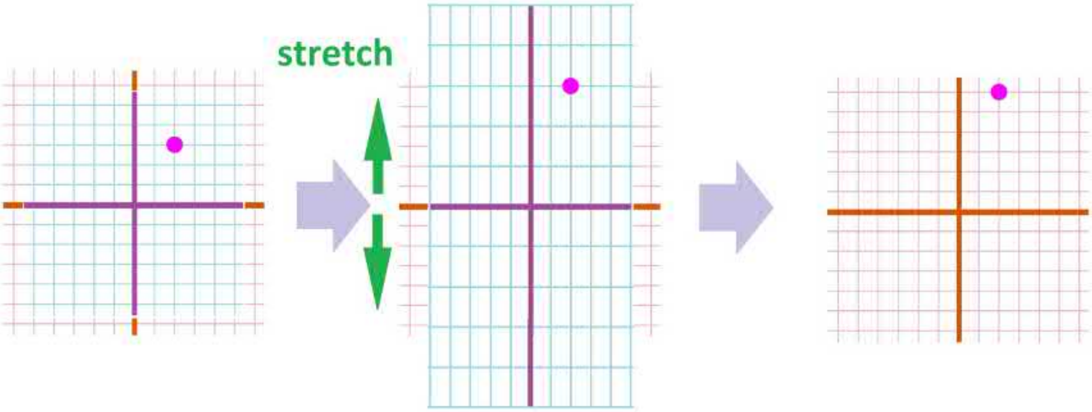
$$\begin{array}{ccc} x & \xrightarrow{\text{flip}} & u=-x \\ y & \xrightarrow{\text{nothing}} & v=y \\ (x,y) & \xrightarrow{\text{horizontal flip}} & (u,v)=F(x,y)=(-x,y) \end{array}$$

Exercise 3.3.5

What is the effect of a vertical flip and a horizontal flip executed consecutively? What if we change the order?

These flips can also be described as *a mirror reflection about the x-axis* and *a mirror reflection about the y-axis*, respectively.

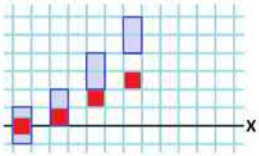
Next, a *vertical stretch*. The coordinate system isn't on a piece of paper anymore! It is on a rubber sheet. We grab it by the top and the bottom and pull them apart in such a way that the x-axis doesn't move. For example, a stretch by a factor of 2 is shown below:



So, all vertical lines are stretched by $k > 0$ away from their origins. Then, the whole plane is stretched by a factor k away from the x-axis. The distance of a generic point (x, y) from the x-axis grows proportionally to k and the point becomes $(x, y \cdot k)$. This is the algebra to describe it:

$$\begin{array}{ccc} x & \xrightarrow{\text{nothing}} & u=x \\ y & \xrightarrow{\text{stretch by } k} & v=y\cdot k \\ (x,y) & \xrightarrow{\text{vertical stretch by } k} & (u,v)=F(x,y)=(x,y\cdot k) \end{array}$$

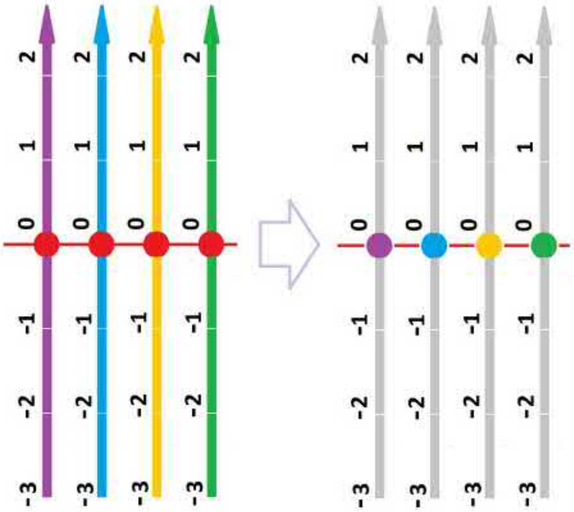
Even though the stretch is the same for all subsets of the plane, the new location will vary depending on the location of the subset relative to the x-axis:



Exercise 3.3.6

What is the effect of a vertical flip and a horizontal shift executed consecutively? What if we change the order?

The case $k = 0$ is very special. As each vertical line collapses on its x-intercept, the whole plane lands on the x-axis. It is called the *projection on the x-axis*:



This is the algebra:

$$\boxed{\begin{array}{ccc} x & \xrightarrow{\text{nothing}} & u = x \\ y & \xrightarrow{\text{collapse}} & v = 0 \\ (x,y) & \xrightarrow{\text{vertical projection}} & (u,v) = F(x,y) = (x,0) \end{array}}$$

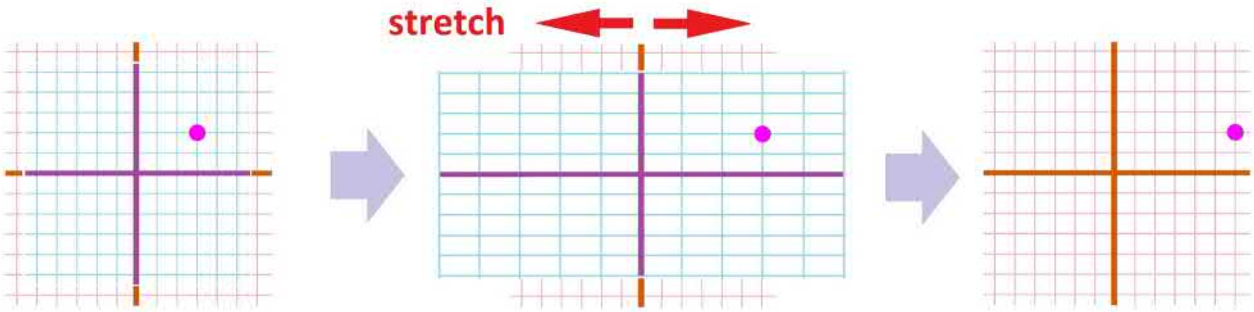
Exercise 3.3.7

What is the range of the projection?

Exercise 3.3.8

What is the effect of a vertical flip and a projection on the x-axis executed consecutively? What if we change the order?

What about *horizontal stretch*? This time, we grab it by the left and right edges of the rubber sheet and pull them apart in such a way that the y-axis doesn't move. For example, a stretch by a factor of 2 is shown below:



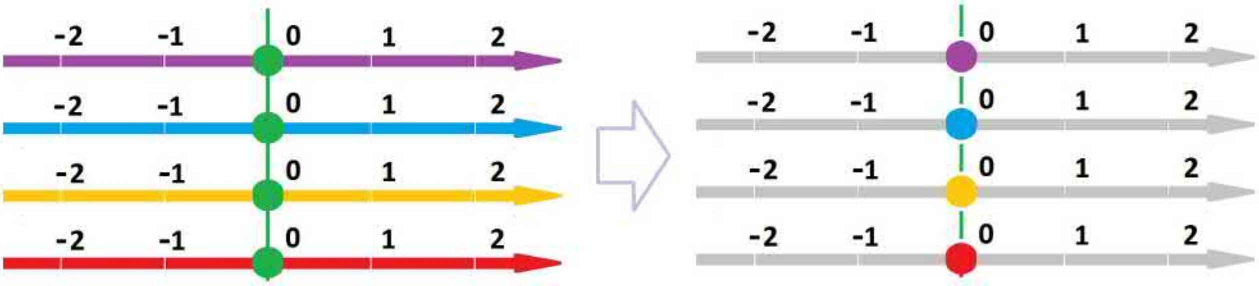
So, all horizontal lines are stretched by $k > 0$ away from their origins. Then, the whole plane is stretched by a factor k away from the y-axis. The distance of a generic point (x, y) from the y-axis grows proportionally to k and the point becomes (kx, y). This is a way describe a horizontal stretch:

$$\begin{array}{ccc} x & \xrightarrow{\text{stretch by } k} & u = x \cdot k \\ y & \xrightarrow{\text{nothing}} & v = y \\ (x, y) & \xrightarrow{\text{horizontal stretch by } k} & (u, v) = F(x, y) = (x \cdot k, y) \end{array}$$

Exercise 3.3.9

What is the effect of a vertical flip and a horizontal flip executed consecutively? What if we change the order?

The case $k = 0$ is very special. As each horizontal line collapses on its y-intercept, the whole plane lands on the y-axis. It is called the *projection on the y-axis*:



This is the algebra:

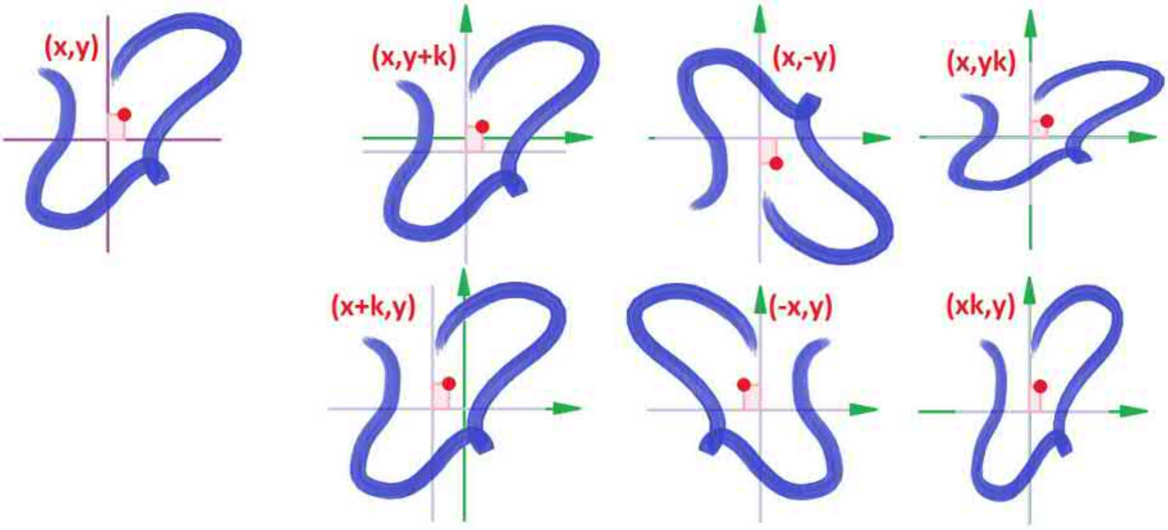
$$\begin{array}{ccc} x & \xrightarrow{\text{collapse}} & u = 0 \\ y & \xrightarrow{\text{nothing}} & v = y \\ (x, y) & \xrightarrow{\text{horizontal projection}} & (u, v) = F(x, y) = (0, y) \end{array}$$

Exercise 3.3.10

What is the effect of a vertical projection and a horizontal projection executed consecutively?

These stretches can also be described as *a uniform deformation away from the y-axis* and *a uniform deformation away from the y-axis*, respectively.

These are our six basic transformations:



The algebra below reflects the geometry above.

Theorem 3.3.11: Formulas of Transformations of Plane

The following transformations of the plane,

$$F : \mathbf{R}^2 \to \mathbf{R}^2\,,\ (u,v) = F(x,y)\,,$$

are given by their formulas:

vertical shift:	*flip:*	*stretch:*
$(\ x\ ,\ y\ )$ $\mapsto (\ x\ ,\ y+k\ )$	$(\ x\ ,\ y\ )$ $\mapsto (\ x\ ,\ y\cdot(-1)\ )$	$(\ x\ ,\ y\ )$ $\mapsto (\ x\ ,\ y\cdot k\ )$

horizontal shift:	*flip:*	*stretch:*
$(\ x\ ,\ y\ )$ $\mapsto (\ x+k\ ,\ y\ )$	$(\ x\ ,\ y\ )$ $\mapsto (\ x\cdot(-1)\ ,\ y\ )$	$(\ x\ ,\ y\ )$ $\mapsto (\ x\cdot k\ ,\ y\ )$

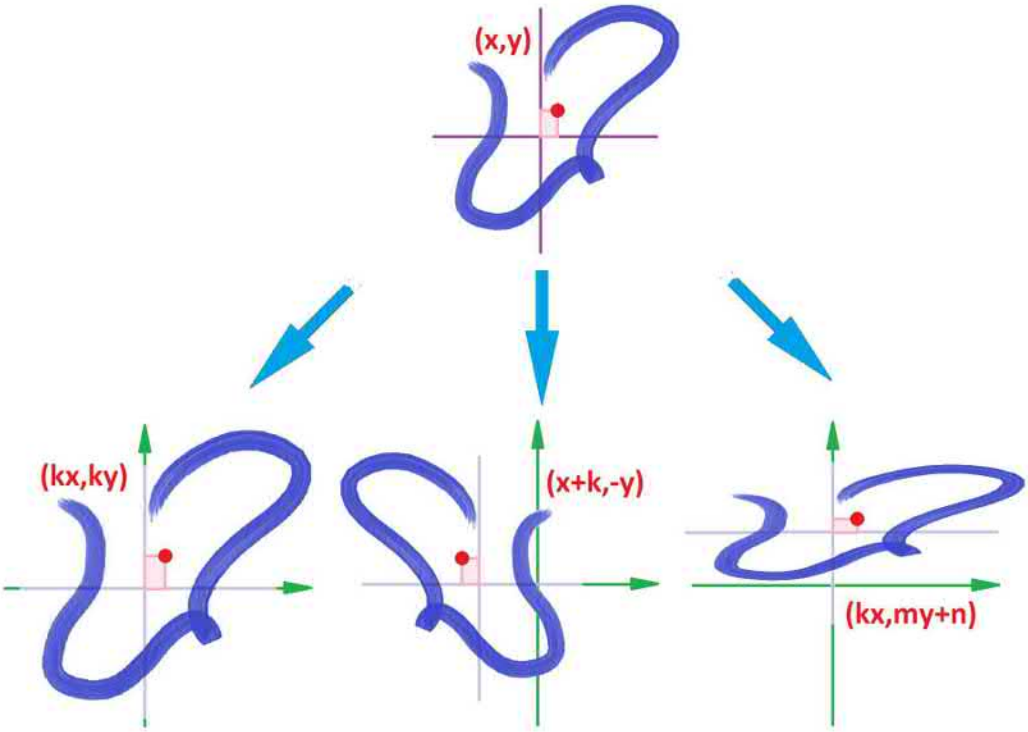
Exercise 3.3.12

What are the images of these six functions? What about the projections?

For now, each of these six operations is limited to one of the two directions: along the x-axis or along the y-axis. We *combine* them as compositions. For example,

$$\begin{array}{ccccccccc} \text{point} & \to & \boxed{\text{stretch vertically by } k} & \to & \text{point} & \to & \boxed{\text{flip horizontally}} & \to & \text{point} \\ (x,y) & \to & \boxed{\text{multiply } y \text{ by } k} & \to & (x,yk) & \to & \boxed{\text{multiply } x \text{ by } (-1)} & \to & (-x,yk) \end{array}$$

We produce a variety of results:



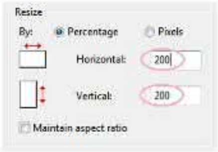
Exercise 3.3.13

Execute – both geometrically and algebraically – the following transformations:

1. Translate up by 2, then reflect about the x-axis, then translate left by 3.
2. Pull away from the y-axis by a factor of 3, then pull toward the x-axis by a factor of 2.

Example 3.3.14: transformations with computer graphics

We can illustrate these transformations with a graphics editor:



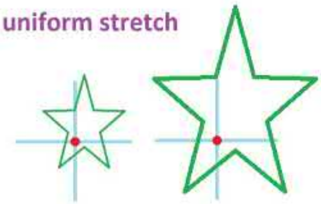
Resize

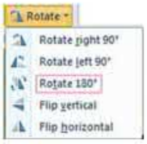
By: ● Percentage ○ Pixels

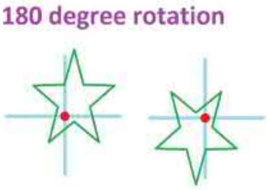
Horizontal: 200

Vertical: 200

☐ Maintain aspect ratio







Exercise 3.3.15

What sequences of basic transformations discussed above produce these results?

Exercise 3.3.16

Describe – both geometrically and algebraically – a transformation that makes a 1×1 square into a 2×1 rectangle.

Exercise 3.3.17

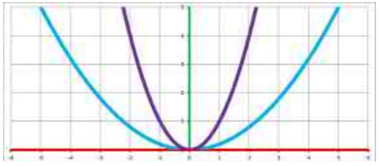
What transformations increase/decrease steepness of lines? What about their slopes?

Exercise 3.3.18

Point out the inverses of each of the six transformations of the plane on the list.

Exercise 3.3.19

Has this parabola been shrunk vertically or stretched horizontally?

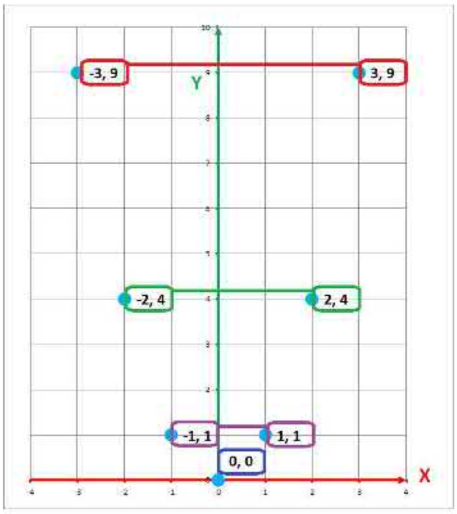


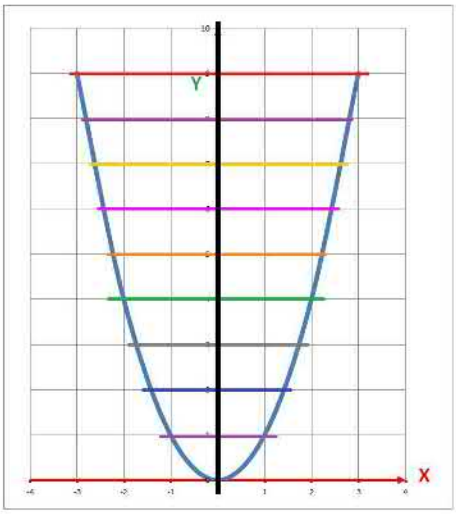
Recall some of the transformations of the plane we have seen elsewhere.

Example 3.3.20: even functions

The fact that function $y = x^2$ is even is demonstrated by observing that the parabola's left branch is a mirror image of its right branch. This operation is executed with a horizontal flip:

$$\boxed{(x, y) \mapsto (-x, y)}$$

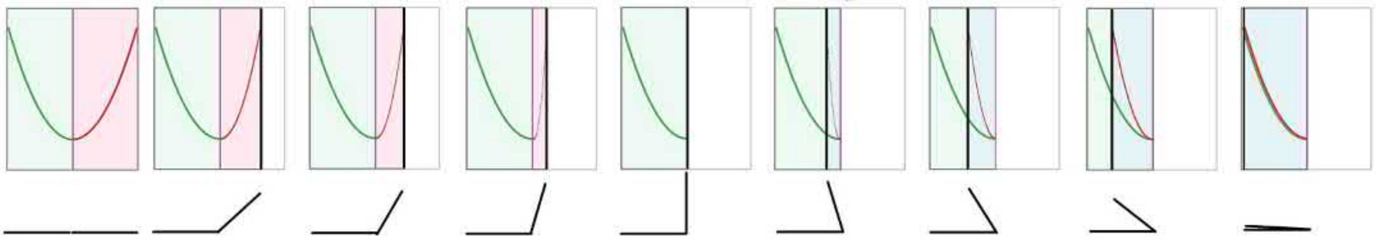




Generally, a subset A of the plane is *mirror symmetric* about the y-axis when we have:

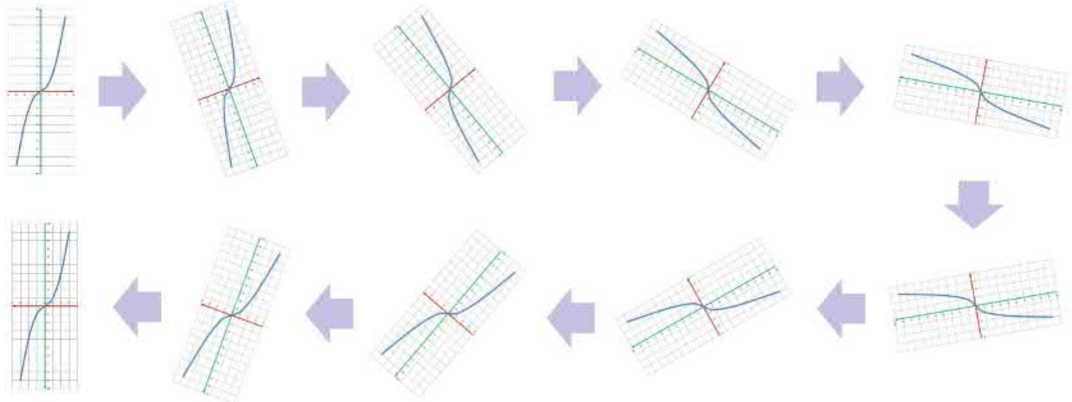
$$(x,y) \text{ belongs to } A \implies (-x,y) \text{ belongs to } A\,.$$

It is a result of a fold:

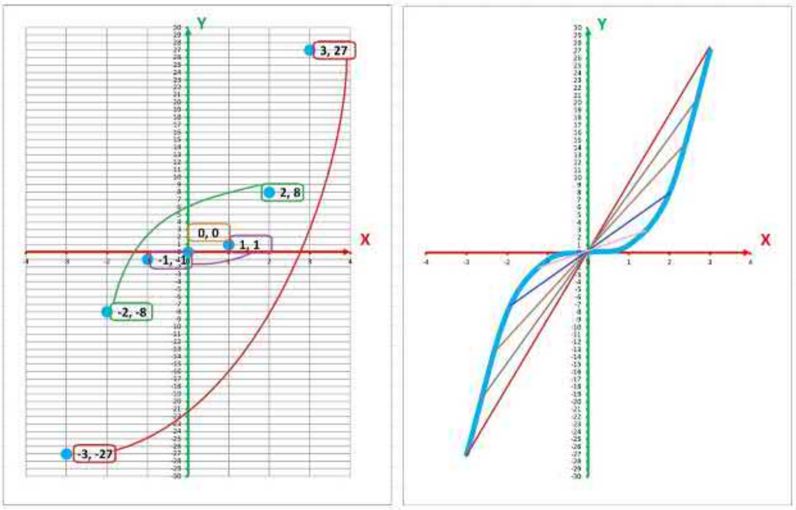


Example 3.3.21: odd functions

The fact that function $y=x^3$ is odd is demonstrated by observing that its graph's right branch is a centrally symmetric to its left branch. This operation is executed with a 180 degree rotation:



The formula can be guessed from the picture:



We achieve the same effect if we instead flip the plane about the y-axis and then about the x-axis (or vice versa):

$$\boxed{(x,y) \mapsto (-x,y) \mapsto (-x,-y)}$$

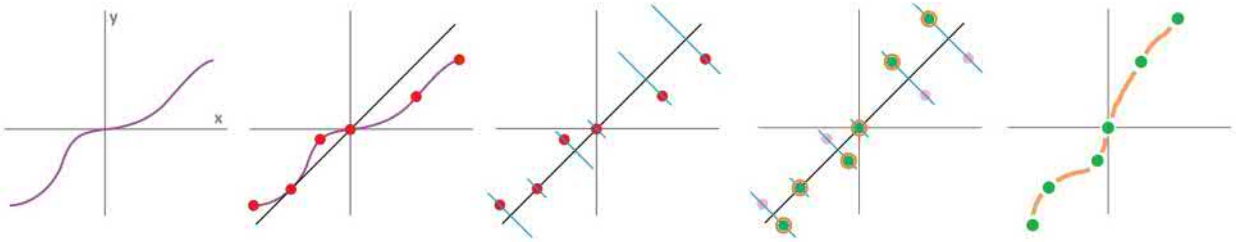
Then a subset A of the plane is *centrally symmetric* when we have:

$$(x,y) \text{ belongs to } A \implies (-x,-y) \text{ belongs to } A\,.$$

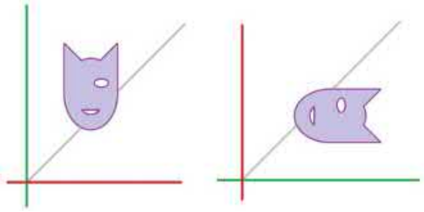
However, some transformations *cannot* be decomposed into a composition of those six basic transformations!

Example 3.3.22: inverse functions

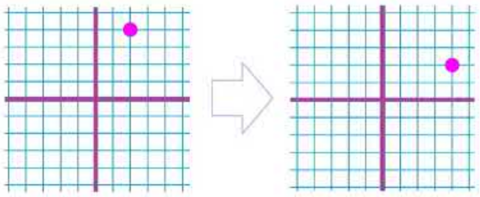
Consider the flip about the line $x = y$ that is needed to construct the graph of the inverse function:



This is the transformation:



The formula is guessed from a single point:

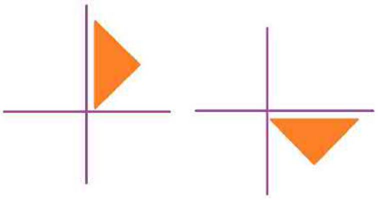


It is given by

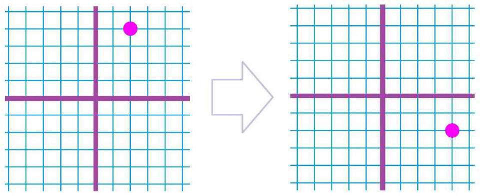
$$(x, y) \mapsto (y, x)\,.$$

Example 3.3.23: rotations

Another such example is a 90-degree rotation:



The formula is guessed from a single point:



It is:

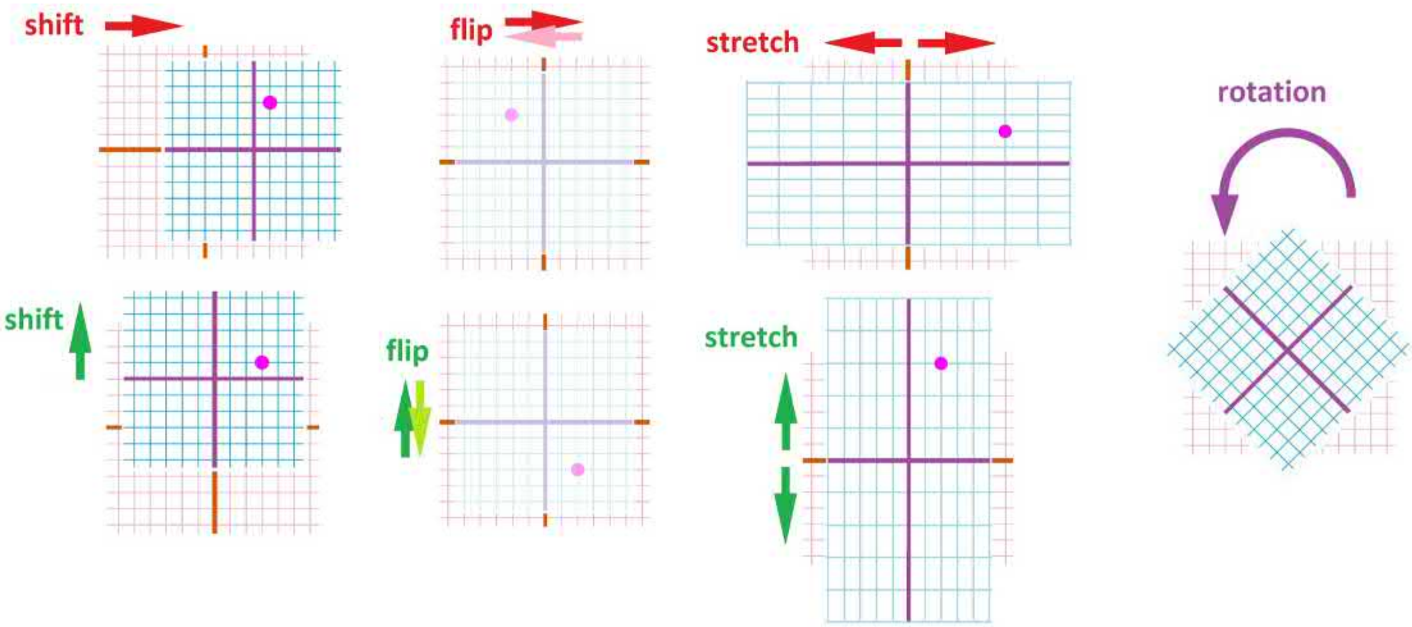
$$(x, y) \mapsto (-y, x)\,.$$

Exercise 3.3.24

What are the inverses of the transformations presented in this section?

In these transformations, x's and y's are intermixed.

It might be sufficient to consider only the compositions of these seven:



3.4. Linear transformations

We will propose a new point of view on the problems we started with.

The simpler one is: Suppose we have a type of coffee that costs \$3 per pound. How much do we get for \$60?

Let x be the weight of the coffee. Since the total price is 60, we have a *linear equation*:

$$3x = 60\,.$$

Imagine that we don't want solve it. First, the solution that we know involves division, an expensive operation. Second, we might anticipate that the total price could change every day. How do we approach the problem?

We collect data: For each possible weight x, we find the corresponding price u. Of course, this is just a *numerical function*:

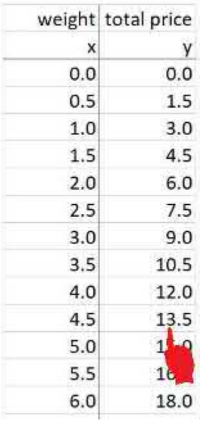
weight $x=$	0	1	2	3	...
total price $u=$	0	3	6	9	...

We can now ask someone to use this as a look-up table: To find the weight that would produce a total price of \$9, find it in the second row and then go up to read: 3 pounds. No computation necessary!

Of course, this task is just the *inverse* of the original numerical function:

total price $u=$	0	3	6	9	...
weight $x=$	0	1	2	3	...

For more data points, use a spreadsheet:



weight	total price
x	y
0.0	0.0
0.5	1.5
1.0	3.0
1.5	4.5
2.0	6.0
2.5	7.5
3.0	9.0
3.5	10.5
4.0	12.0
4.5	13.5
5.0	[illegible]
5.5	[illegible]
6.0	18.0

As a result, we can solve the same problem over and over with different data. In other words, the problem given by the equation

$$f(x) = u$$

is solved – for every u – with the inverse:

$$x = f^{-1}(u)\,.$$

Now, the next level of complexity: *two unknowns.*

We have the Kenyan coffee that costs \$2 per pound and the Colombian coffee that costs \$3 per pound. If the total weight is 6, we have a linear relation between x and y:

$$\boxed{1} \quad x + y = 6\,.$$

If the total price of the blend is \$14, we have another linear relation between x and y:

$$\boxed{2} \quad 2x + 3y = 14\,.$$

What if we solve the same problem over and over with different data?

We might anticipate that the total weight and price could change every day: What if tomorrow we expect to have to find a blend with the total weight of 8 pounds and price of \$15? How do we approach the problem?

We collect data: For each possible weights x and y, we find the corresponding total price u and total weight v. The inputs, however, won't fit in a single column as in the last example. Instead, it's an array.

We use the formulas:

$$u = x + y, \;\; v = 2x + 3y\,.$$

This is the data:

total weight u	$y\backslash x$	0	1	2	3	...
	0	0	1	2	3	...
	1	1	2	3	4	...
	2	2	3	4	5	...
	...	...	...	...	...	...

total price v	$y\backslash x$	0	1	2	3	...
	0	0	2	4	6	...
	1	3	5	7	9	...
	2	6	8	10	12	...
	...	...	...	...	...	...

We call either of these *functions of two variables*: inputs are two numbers and the output is a single number. We will utilize the general function notation though:

$$h : \mathbf{R}^2 \to \mathbf{R}\,.$$

Exercise 3.4.1

Are these functions one-to-one? Onto?

We can now ask someone to use this as look-up tables:

1. To find the weights that produce a total weight of $u = 3$ pounds, find its occurrences in the first table.

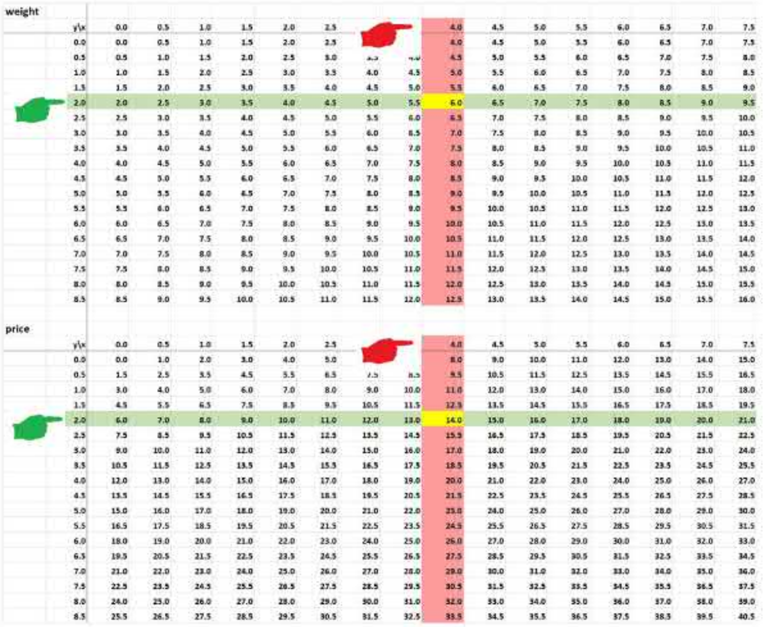
2. To find the weights that produce a total price of $v = 7$ dollars, find its occurrences in the first table.
3. Find a location within the tables that produces both: $x = 2$ and $y = 1$.

Even though no computation is necessary, there is a lot of looking around!

For more data points, use a spreadsheet with two tables.The formulas are respectively:

`=R2C+RC2` and `=2*R2C+3*RC2`

The result is as follows:



weight

y\x	0.0	0.5	1.0	1.5	2.0	2.5			4.0	4.5	5.0	5.5	6.0	6.5	7.0	7.5
0.0	0.0	0.5	1.0	1.5	2.0	2.5			4.0	4.5	5.0	5.5	6.0	6.5	7.0	7.5
0.5	0.5	1.0	1.5	2.0	2.5	3.0	3.5	4.0	4.5	5.0	5.5	6.0	6.5	7.0	7.5	8.0
1.0	1.0	1.5	2.0	2.5	3.0	3.5	4.0	4.5	5.0	5.5	6.0	6.5	7.0	7.5	8.0	8.5
1.5	1.5	2.0	2.5	3.0	3.5	4.0	4.5	5.0	5.5	6.0	6.5	7.0	7.5	8.0	8.5	9.0
2.0	2.0	2.5	3.0	3.5	4.0	4.5	5.0	5.5	6.0	6.5	7.0	7.5	8.0	8.5	9.0	9.5
2.5	2.5	3.0	3.5	4.0	4.5	5.0	5.5	6.0	6.5	7.0	7.5	8.0	8.5	9.0	9.5	10.0
3.0	3.0	3.5	4.0	4.5	5.0	5.5	6.0	6.5	7.0	7.5	8.0	8.5	9.0	9.5	10.0	10.5
3.5	3.5	4.0	4.5	5.0	5.5	6.0	6.5	7.0	7.5	8.0	8.5	9.0	9.5	10.0	10.5	11.0
4.0	4.0	4.5	5.0	5.5	6.0	6.5	7.0	7.5	8.0	8.5	9.0	9.5	10.0	10.5	11.0	11.5
4.5	4.5	5.0	5.5	6.0	6.5	7.0	7.5	8.0	8.5	9.0	9.5	10.0	10.5	11.0	11.5	12.0
5.0	5.0	5.5	6.0	6.5	7.0	7.5	8.0	8.5	9.0	9.5	10.0	10.5	11.0	11.5	12.0	12.5
5.5	5.5	6.0	6.5	7.0	7.5	8.0	8.5	9.0	9.5	10.0	10.5	11.0	11.5	12.0	12.5	13.0
6.0	6.0	6.5	7.0	7.5	8.0	8.5	9.0	9.5	10.0	10.5	11.0	11.5	12.0	12.5	13.0	13.5
6.5	6.5	7.0	7.5	8.0	8.5	9.0	9.5	10.0	10.5	11.0	11.5	12.0	12.5	13.0	13.5	14.0
7.0	7.0	7.5	8.0	8.5	9.0	9.5	10.0	10.5	11.0	11.5	12.0	12.5	13.0	13.5	14.0	14.5
7.5	7.5	8.0	8.5	9.0	9.5	10.0	10.5	11.0	11.5	12.0	12.5	13.0	13.5	14.0	14.5	15.0
8.0	8.0	8.5	9.0	9.5	10.0	10.5	11.0	11.5	12.0	12.5	13.0	13.5	14.0	14.5	15.0	15.5
8.5	8.5	9.0	9.5	10.0	10.5	11.0	11.5	12.0	12.5	13.0	13.5	14.0	14.5	15.0	15.5	16.0

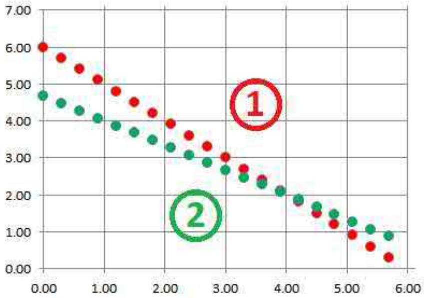
price

y\x	0.0	0.5	1.0	1.5	2.0	2.5			4.0	4.5	5.0	5.5	6.0	6.5	7.0	7.5
0.0	0.0	1.0	2.0	3.0	4.0	5.0			8.0	9.0	10.0	11.0	12.0	13.0	14.0	15.0
0.5	1.5	2.5	3.5	4.5	5.5	6.5	7.5	8.5	9.5	10.5	11.5	12.5	13.5	14.5	15.5	16.5
1.0	3.0	4.0	5.0	6.0	7.0	8.0	9.0	10.0	11.0	12.0	13.0	14.0	15.0	16.0	17.0	18.0
1.5	4.5	5.5	6.5	7.5	8.5	9.5	10.5	11.5	12.5	13.5	14.5	15.5	16.5	17.5	18.5	19.5
2.0	6.0	7.0	8.0	9.0	10.0	11.0	12.0	13.0	14.0	15.0	16.0	17.0	18.0	19.0	20.0	21.0
2.5	7.5	8.5	9.5	10.5	11.5	12.5	13.5	14.5	15.5	16.5	17.5	18.5	19.5	20.5	21.5	22.5
3.0	9.0	10.0	11.0	12.0	13.0	14.0	15.0	16.0	17.0	18.0	19.0	20.0	21.0	22.0	23.0	24.0
3.5	10.5	11.5	12.5	13.5	14.5	15.5	16.5	17.5	18.5	19.5	20.5	21.5	22.5	23.5	24.5	25.5
4.0	12.0	13.0	14.0	15.0	16.0	17.0	18.0	19.0	20.0	21.0	22.0	23.0	24.0	25.0	26.0	27.0
4.5	13.5	14.5	15.5	16.5	17.5	18.5	19.5	20.5	21.5	22.5	23.5	24.5	25.5	26.5	27.5	28.5
5.0	15.0	16.0	17.0	18.0	19.0	20.0	21.0	22.0	23.0	24.0	25.0	26.0	27.0	28.0	29.0	30.0
5.5	16.5	17.5	18.5	19.5	20.5	21.5	22.5	23.5	24.5	25.5	26.5	27.5	28.5	29.5	30.5	31.5
6.0	18.0	19.0	20.0	21.0	22.0	23.0	24.0	25.0	26.0	27.0	28.0	29.0	30.0	31.0	32.0	33.0
6.5	19.5	20.5	21.5	22.5	23.5	24.5	25.5	26.5	27.5	28.5	29.5	30.5	31.5	32.5	33.5	34.5
7.0	21.0	22.0	23.0	24.0	25.0	26.0	27.0	28.0	29.0	30.0	31.0	32.0	33.0	34.0	35.0	36.0
7.5	22.5	23.5	24.5	25.5	26.5	27.5	28.5	29.5	30.5	31.5	32.5	33.5	34.5	35.5	36.5	37.5
8.0	24.0	25.0	26.0	27.0	28.0	29.0	30.0	31.0	32.0	33.0	34.0	35.0	36.0	37.0	38.0	39.0
8.5	25.5	26.5	27.5	28.5	29.5	30.5	31.5	32.5	33.5	34.5	35.5	36.5	37.5	38.5	39.5	40.5

To find the output for a given input, we find the $x = 4$ and go down, find the $y = 2$ and go right – until intersection.

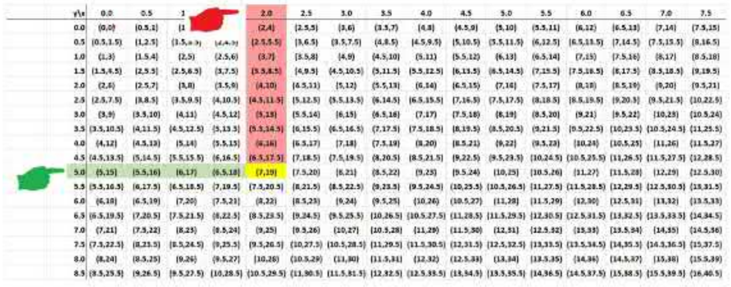
To find the input for a given output is easy but only when it is already known. In reality, there are multiple possibilities in either of the tables:

The lines are the same as those we saw in our previous analysis:



To find the answer, we'd have to move along the two lines matching the locations. Furthermore, because of the skipped values, we might never find the intersection in our table. Is there a better way?

Let's combine these two tables of the two functions of two variables into a single table. We overlap them as if they are written an plastic sheets. Then, for each x and y, we show the corresponding pair (u, v) of the total weight and total price:



We use the formula:

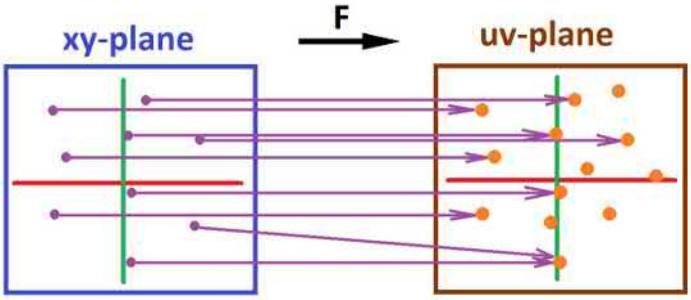
`=CONCAT("(",R2C+RC2,",",2*R2C+3*RC2,")")`

We need to search only a single table now!

Here is a very important observation:

- ▶ From a pair of weights (x, y), a pair of the total weight and total price (u, v) is found.

We, therefore, are facing a *transformation of the plane*:



So, we have combined the tables of the two functions of two variables into a single table for our new function.

The function $F : \mathbf{R}^2 \to \mathbf{R}^2$ is given by the two formulas above:

$$u = x + y, \ v = 2x + 3y \,,$$

or by a single formula presented in terms of the coordinates of the points:

$$(u, v) = F(x, y) = (x + y, 2x + 3y) \,.$$

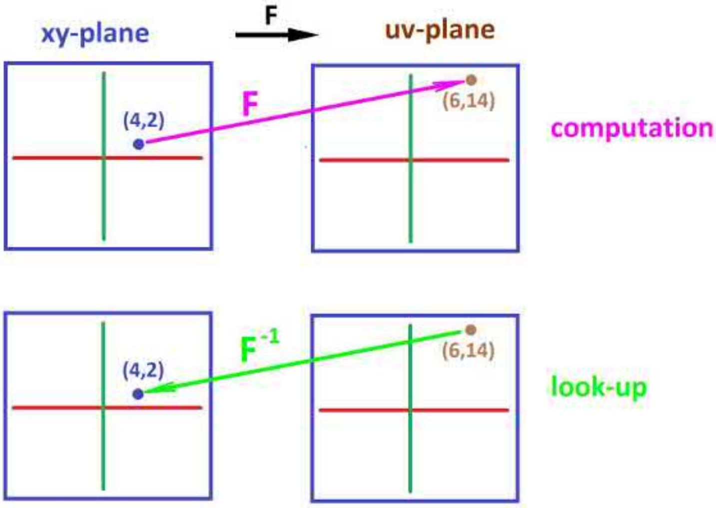
Such a function is called a *linear operator*, or transformation, or map. There can be a combination of:

1. a flip over any axis
2. a stretch in any directions and with any magnitude
3. a rotation around the origin through any angle
4. no shift!

In particular, we have

$$F(4, 2) = (6, 14) \,.$$

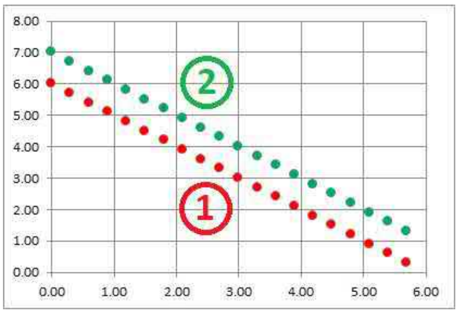
The look-up procedure is the inverse of F:



Problem solved!

However, different functions may produce different outcomes.

We saw a price combination – \$2 for either – with no solution to our problem:

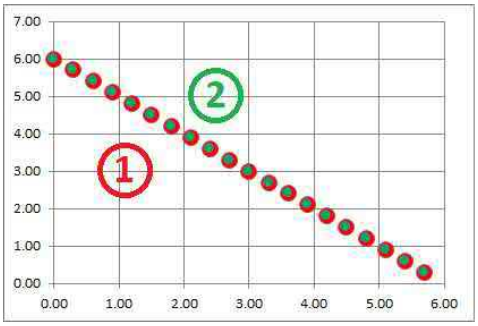


In other words, the point $(6,14)$ doesn't belong to the image of the function

$$(u,v)=G(x,y)=(x+y,2x+2y)\,.$$

We would be searching the table in vain!

For this price combination, we also saw a possibility of infinitely many solutions if the required total price and total weight – 6 and 12 – are just right:



This means that the point $(6,12)$ does belong to the image of the function G and its preimage is the whole line.

The intersection of the two lines is the preimage of the point (u,v).

So, the problem given by the *equation*

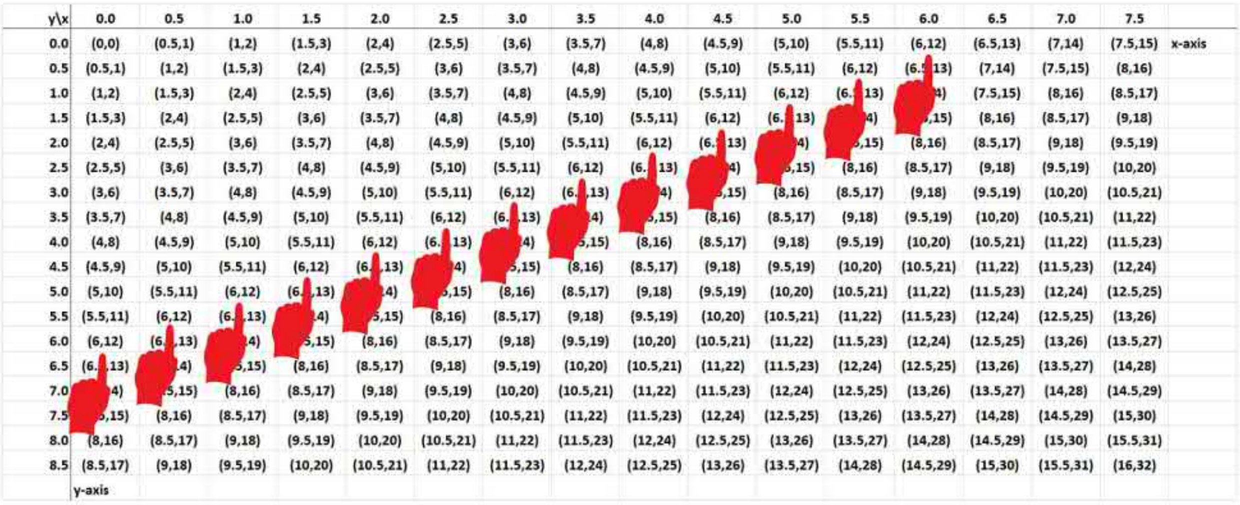
$$F(x,y)=(u,v)$$

is solved – for every (u,v) – with the inverse of the function:

$$(x,y)=F^{-1}(u,v)\,.$$

But is the function one-to-one in the first place?

The first one above, F, seems to be. But we know that the second function above, G, isn't one-to-one. The point $(6,12)$ is present (many times) but not $(6,14)$, or $(1,1)$, etc.:



y\x	0.0	0.5	1.0	1.5	2.0	2.5	3.0	3.5	4.0	4.5	5.0	5.5	6.0	6.5	7.0	7.5	
0.0	(0,0)	(0.5,1)	(1,2)	(1.5,3)	(2,4)	(2.5,5)	(3,6)	(3.5,7)	(4,8)	(4.5,9)	(5,10)	(5.5,11)	(6,12)	(6.5,13)	(7,14)	(7.5,15)	x-axis
0.5	(0.5,1)	(1,2)	(1.5,3)	(2,4)	(2.5,5)	(3,6)	(3.5,7)	(4,8)	(4.5,9)	(5,10)	(5.5,11)	(6,12)	(6. 13)	(7,14)	(7.5,15)	(8,16)	
1.0	(1,2)	(1.5,3)	(2,4)	(2.5,5)	(3,6)	(3.5,7)	(4,8)	(4.5,9)	(5,10)	(5.5,11)	(6,12)	(6. 13)	4)	(7.5,15)	(8,16)	(8.5,17)	
1.5	(1.5,3)	(2,4)	(2.5,5)	(3,6)	(3.5,7)	(4,8)	(4.5,9)	(5,10)	(5.5,11)	(6,12)	(6. 13)	4)	,15)	(8,16)	(8.5,17)	(9,18)	
2.0	(2,4)	(2.5,5)	(3,6)	(3.5,7)	(4,8)	(4.5,9)	(5,10)	(5.5,11)	(6,12)	(6. 13)	4)	5,15)	(8,16)	(8.5,17)	(9,18)	(9.5,19)	
2.5	(2.5,5)	(3,6)	(3.5,7)	(4,8)	(4.5,9)	(5,10)	(5.5,11)	(6,12)	(6. 13)	4)	5,15)	(8,16)	(8.5,17)	(9,18)	(9.5,19)	(10,20)	
3.0	(3,6)	(3.5,7)	(4,8)	(4.5,9)	(5,10)	(5.5,11)	(6,12)	(6. 13)	4)	5,15)	(8,16)	(8.5,17)	(9,18)	(9.5,19)	(10,20)	(10.5,21)	
3.5	(3.5,7)	(4,8)	(4.5,9)	(5,10)	(5.5,11)	(6,12)	(6. 13)	4)	5,15)	(8,16)	(8.5,17)	(9,18)	(9.5,19)	(10,20)	(10.5,21)	(11,22)	
4.0	(4,8)	(4.5,9)	(5,10)	(5.5,11)	(6,12)	(6. 13)	4)	5,15)	(8,16)	(8.5,17)	(9,18)	(9.5,19)	(10,20)	(10.5,21)	(11,22)	(11.5,23)	
4.5	(4.5,9)	(5,10)	(5.5,11)	(6,12)	(6. 13)	4)	5,15)	(8,16)	(8.5,17)	(9,18)	(9.5,19)	(10,20)	(10.5,21)	(11,22)	(11.5,23)	(12,24)	
5.0	(5,10)	(5.5,11)	(6,12)	(6. 13)	4)	5,15)	(8,16)	(8.5,17)	(9,18)	(9.5,19)	(10,20)	(10.5,21)	(11,22)	(11.5,23)	(12,24)	(12.5,25)	
5.5	(5.5,11)	(6,12)	(6. 13)	4)	5,15)	(8,16)	(8.5,17)	(9,18)	(9.5,19)	(10,20)	(10.5,21)	(11,22)	(11.5,23)	(12,24)	(12.5,25)	(13,26)	
6.0	(6,12)	(6. 13)	4)	5,15)	(8,16)	(8.5,17)	(9,18)	(9.5,19)	(10,20)	(10.5,21)	(11,22)	(11.5,23)	(12,24)	(12.5,25)	(13,26)	(13.5,27)	
6.5	(6. 13)	4)	5,15)	(8,16)	(8.5,17)	(9,18)	(9.5,19)	(10,20)	(10.5,21)	(11,22)	(11.5,23)	(12,24)	(12.5,25)	(13,26)	(13.5,27)	(14,28)	
7.0	4)	5,15)	(8,16)	(8.5,17)	(9,18)	(9.5,19)	(10,20)	(10.5,21)	(11,22)	(11.5,23)	(12,24)	(12.5,25)	(13,26)	(13.5,27)	(14,28)	(14.5,29)	
7.5	5,15)	(8,16)	(8.5,17)	(9,18)	(9.5,19)	(10,20)	(10.5,21)	(11,22)	(11.5,23)	(12,24)	(12.5,25)	(13,26)	(13.5,27)	(14,28)	(14.5,29)	(15,30)	
8.0	(8,16)	(8.5,17)	(9,18)	(9.5,19)	(10,20)	(10.5,21)	(11,22)	(11.5,23)	(12,24)	(12.5,25)	(13,26)	(13.5,27)	(14,28)	(14.5,29)	(15,30)	(15.5,31)	
8.5	(8.5,17)	(9,18)	(9.5,19)	(10,20)	(10.5,21)	(11,22)	(11.5,23)	(12,24)	(12.5,25)	(13,26)	(13.5,27)	(14,28)	(14.5,29)	(15,30)	(15.5,31)	(16,32)	
	y-axis																

Examining its formulas explains why:

$$\begin{aligned} u &= x + y \\ v &= 2x + 2y \end{aligned}$$

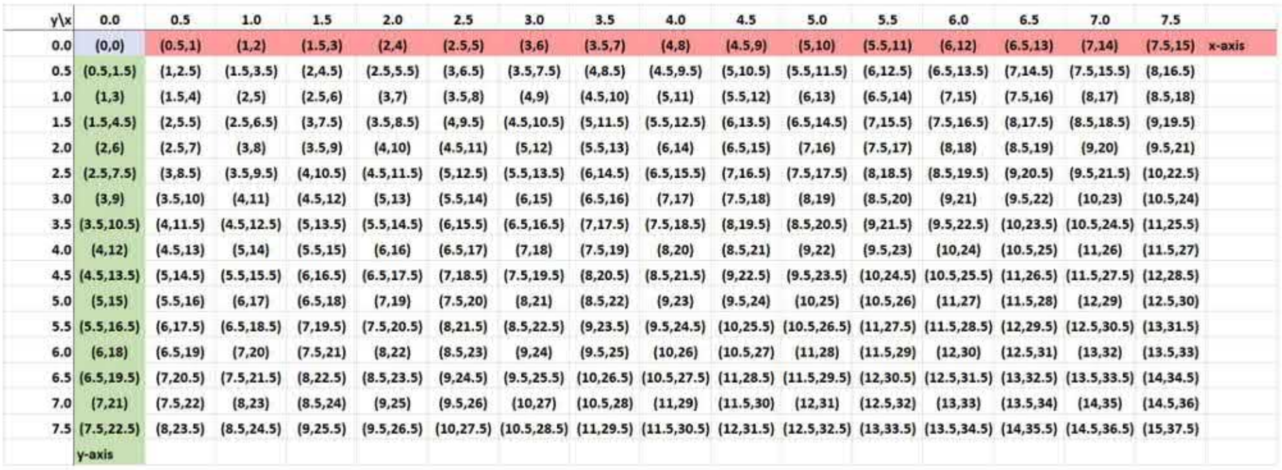
No matter what x and y are, u and v are proportional to each other:

$$v = 2u\,.$$

In other words, all points (u,v) that come from G lie on the same line. This line is the image of this function.

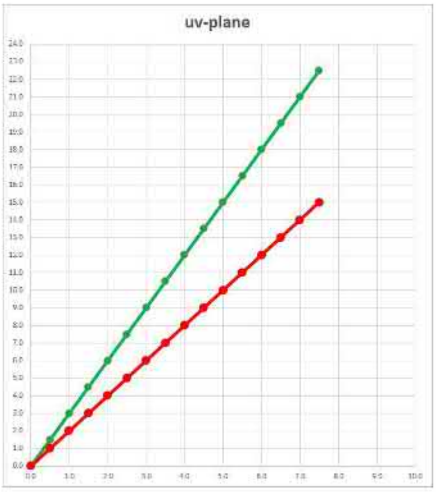
Let's try to visualize these transformations.

The initial idea is just to show where the x- and the y-axis land on the uv-plane:



y\x	0.0	0.5	1.0	1.5	2.0	2.5	3.0	3.5	4.0	4.5	5.0	5.5	6.0	6.5	7.0	7.5	
0.0	(0,0)	(0.5,1)	(1,2)	(1.5,3)	(2,4)	(2.5,5)	(3,6)	(3.5,7)	(4,8)	(4.5,9)	(5,10)	(5.5,11)	(6,12)	(6.5,13)	(7,14)	(7.5,15)	x-axis
0.5	(0.5,1.5)	(1,2.5)	(1.5,3.5)	(2,4.5)	(2.5,5.5)	(3,6.5)	(3.5,7.5)	(4,8.5)	(4.5,9.5)	(5,10.5)	(5.5,11.5)	(6,12.5)	(6.5,13.5)	(7,14.5)	(7.5,15.5)	(8,16.5)	
1.0	(1,3)	(1.5,4)	(2,5)	(2.5,6)	(3,7)	(3.5,8)	(4,9)	(4.5,10)	(5,11)	(5.5,12)	(6,13)	(6.5,14)	(7,15)	(7.5,16)	(8,17)	(8.5,18)	
1.5	(1.5,4.5)	(2,5.5)	(2.5,6.5)	(3,7.5)	(3.5,8.5)	(4,9.5)	(4.5,10.5)	(5,11.5)	(5.5,12.5)	(6,13.5)	(6.5,14.5)	(7,15.5)	(7.5,16.5)	(8,17.5)	(8.5,18.5)	(9,19.5)	
2.0	(2,6)	(2.5,7)	(3,8)	(3.5,9)	(4,10)	(4.5,11)	(5,12)	(5.5,13)	(6,14)	(6.5,15)	(7,16)	(7.5,17)	(8,18)	(8.5,19)	(9,20)	(9.5,21)	
2.5	(2.5,7.5)	(3,8.5)	(3.5,9.5)	(4,10.5)	(4.5,11.5)	(5,12.5)	(5.5,13.5)	(6,14.5)	(6.5,15.5)	(7,16.5)	(7.5,17.5)	(8,18.5)	(8.5,19.5)	(9,20.5)	(9.5,21.5)	(10,22.5)	
3.0	(3,9)	(3.5,10)	(4,11)	(4.5,12)	(5,13)	(5.5,14)	(6,15)	(6.5,16)	(7,17)	(7.5,18)	(8,19)	(8.5,20)	(9,21)	(9.5,22)	(10,23)	(10.5,24)	
3.5	(3.5,10.5)	(4,11.5)	(4.5,12.5)	(5,13.5)	(5.5,14.5)	(6,15.5)	(6.5,16.5)	(7,17.5)	(7.5,18.5)	(8,19.5)	(8.5,20.5)	(9,21.5)	(9.5,22.5)	(10,23.5)	(10.5,24.5)	(11,25.5)	
4.0	(4,12)	(4.5,13)	(5,14)	(5.5,15)	(6,16)	(6.5,17)	(7,18)	(7.5,19)	(8,20)	(8.5,21)	(9,22)	(9.5,23)	(10,24)	(10.5,25)	(11,26)	(11.5,27)	
4.5	(4.5,13.5)	(5,14.5)	(5.5,15.5)	(6,16.5)	(6.5,17.5)	(7,18.5)	(7.5,19.5)	(8,20.5)	(8.5,21.5)	(9,22.5)	(9.5,23.5)	(10,24.5)	(10.5,25.5)	(11,26.5)	(11.5,27.5)	(12,28.5)	
5.0	(5,15)	(5.5,16)	(6,17)	(6.5,18)	(7,19)	(7.5,20)	(8,21)	(8.5,22)	(9,23)	(9.5,24)	(10,25)	(10.5,26)	(11,27)	(11.5,28)	(12,29)	(12.5,30)	
5.5	(5.5,16.5)	(6,17.5)	(6.5,18.5)	(7,19.5)	(7.5,20.5)	(8,21.5)	(8.5,22.5)	(9,23.5)	(9.5,24.5)	(10,25.5)	(10.5,26.5)	(11,27.5)	(11.5,28.5)	(12,29.5)	(12.5,30.5)	(13,31.5)	
6.0	(6,18)	(6.5,19)	(7,20)	(7.5,21)	(8,22)	(8.5,23)	(9,24)	(9.5,25)	(10,26)	(10.5,27)	(11,28)	(11.5,29)	(12,30)	(12.5,31)	(13,32)	(13.5,33)	
6.5	(6.5,19.5)	(7,20.5)	(7.5,21.5)	(8,22.5)	(8.5,23.5)	(9,24.5)	(9.5,25.5)	(10,26.5)	(10.5,27.5)	(11,28.5)	(11.5,29.5)	(12,30.5)	(12.5,31.5)	(13,32.5)	(13.5,33.5)	(14,34.5)	
7.0	(7,21)	(7.5,22)	(8,23)	(8.5,24)	(9,25)	(9.5,26)	(10,27)	(10.5,28)	(11,29)	(11.5,30)	(12,31)	(12.5,32)	(13,33)	(13.5,34)	(14,35)	(14.5,36)	
7.5	(7.5,22.5)	(8,23.5)	(8.5,24.5)	(9,25.5)	(9.5,26.5)	(10,27.5)	(10.5,28.5)	(11,29.5)	(11.5,30.5)	(12,31.5)	(12.5,32.5)	(13,33.5)	(13.5,34.5)	(14,35.5)	(14.5,36.5)	(15,37.5)	
	y-axis																

Just looking at the output data on the x-axis reveals that they all satisfy $v = 2u$. The axis, therefore, ends up on the line $v = 2u$. Examining the output data on the y-axis shows that F takes it to the line $v = 3u$. We plot the two images together:



We see stretching and rotation but no flips. We will return to the topic in Chapter 5.

Exercise 3.4.2

Show what happens to the grid of the xy-plane.

Exercise 3.4.3

Repeat the above analysis for G.

Let's find the inverse of F, if possible.

The idea is the same as always:

- ▶ Solve for the independent variable.

In other words, express (x, y) in terms of (u, v). Or, express x in terms of u and v and express y in terms of u and v.

The algebra is the same as in the very beginning. We take the equations

$$u = x + y,\ v = 2x + 3y\,,$$

and solve the first one for y:

$$y = u - x\,.$$

Then substitute into the second:

$$v = 2x + 3y = 2x + 3(u - x) = -x + 3u\,.$$

Solve this for x:

$$x = 3u - v\,.$$

Substitute into the second equation:

$$y = u - x = u - (3u - v) = -2u + v\,.$$

So, the new equations are:

$$x = 3u - v,\ y = -2u + v\,.$$

The inverse is:

$$F^{-1}(u, v) = (3u - v, -2u + v)\,.$$

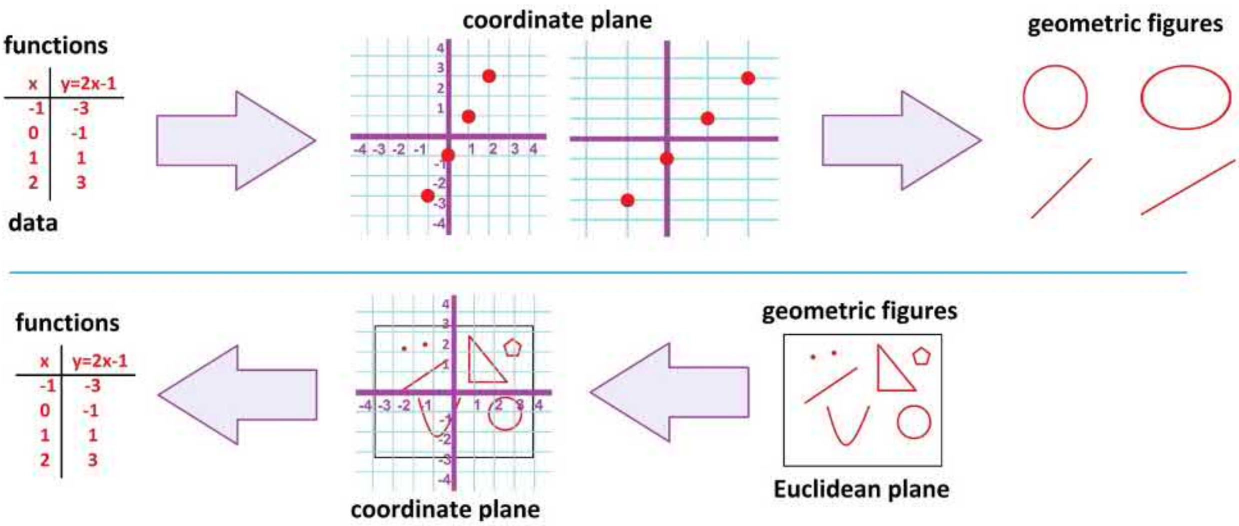
These are the lessons that we have learned and will apply to higher dimensions:

1. Combine numbers into points.
2. Combine numerical functions into functions of points.

The last plot can be re-scaled without changing its meaning. This is pure *algebra*! Now, we add *geometry*. The word means "land measuring" in Greek. So, we add the ability to indirectly measure to the xy-plane in the next section.

3.5. Analytic geometry: the Cartesian system for the Euclidean plane

We introduced the Cartesian coordinate plane as a device for visualizing *functions*; it's where the graphs live (top row):



We now approach it from another direction (bottom row): We want to study the physical space and the *Euclidean plane* as its representation. We then want to address Euclidean geometry algebraically. We start by *superimposing* – as if it is drawn on a transparent piece of plastic – the Cartesian grid over this plane. As a result all points acquire coordinates and all geometric objects acquire algebraic representations: functions and relations.

A major difference we see is that the former coordinate plane doesn't need to have a square grid as the units of x and y might be unrelated (dollars vs. hours). The latter does and the units of x and y are better be those of length (miles, feet, etc.).

The idea of "analytic geometry" is to use a coordinate system to transition between the following two:

- *geometry*: points, then lines, triangles, circles, then planes, cubes, spheres, etc.,
- *algebra*: numbers, then combinations of numbers, then relations and functions, etc.

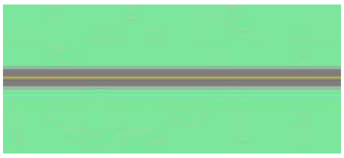
This will allow us to solve geometric problems *without measuring* – because everything is pre-measured! We will initially limit ourselves to the two simplest geometric tasks:

1. finding *distances* without a ruler: between two points, between a point and a line or curve, etc., and
2. finding *angles* without a protractor: between two lines or two curves.

We start with *dimension* 1.

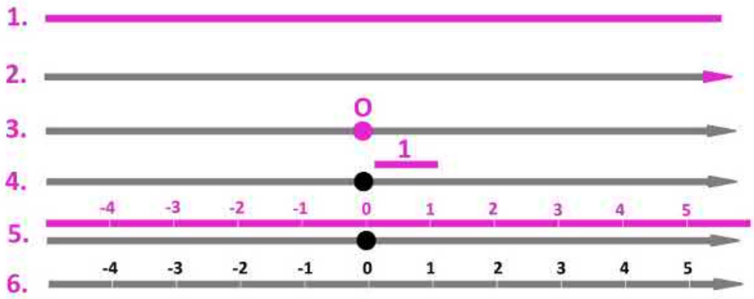
Let's first review its construction as presented in Chapter 1.

Suppose we live on a *road* surrounded by nothingness:



The coordinate system for this road is like a set of milestones. It is devised to be superimposed on the road in order to numerically capture all the locations.

It is built in several stages:



These are the steps:

1. Draw a line, the *x-axis*.
2. Choose one of the two directions on the line as *positive*, then the other is *negative*.
3. Choose a point O as the *origin*.
4. Set a segment of the line – of length 1 – as a *unit*.
5. Use the segment to measure distances to locations from the origin O – positive in the positive direction and negative in the negative direction – and add marks to the line, the *coordinates*; later the segments are further subdivided to fractions of the unit, etc.
6. We have a coordinate system on the line.

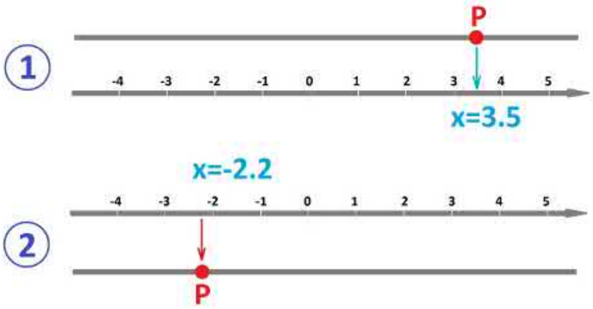
The result is a correspondence:

$$\boxed{\text{location } \mathbf{P} \longleftrightarrow \text{number } \mathbf{x}}$$

We can, therefore, refer to this line as the *real number line*.

The correspondence works in both directions.

For example, suppose P is a *location* on the line. We then find the distance from the origin – positive in the positive direction and negative in the negative direction – and the result is the coordinate of P, some *number* x. We use the nearest mark to simplify the task.



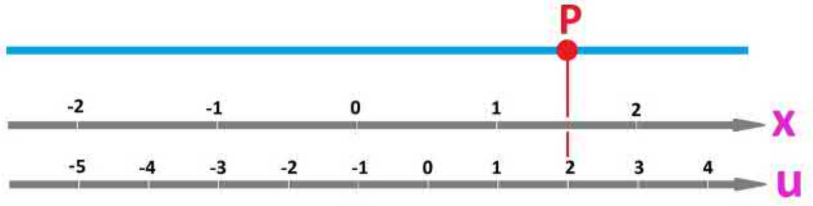
Conversely, suppose x is a *number*. We then measure x as the distance to the origin – positive in the positive direction and negative in the negative direction – and the result is a *location* P on the line. We use the nearest mark to simplify the task.

Example 3.5.1: different coordinate systems, dimension 1

Of course, we can place different coordinate systems on the same line:

1. different units (feet instead of inches, etc.)
2. different starting point O

We can place two different rulers next to our line and have two different coordinates for it:



Here, the point P shown has:

- coordinate 1.5 according to the first system and
- coordinate 2 according to the other.

We know from Chapter 3 that all the functions defined on the first axis are transformed to the ones on the second by a single function; in this particular case, it is:

$$x = (u+1)/2 .$$

Of course, all the functions defined on the second axis are transformed to the ones on the first by the *inverse* of this function:

$$u = 2x - 1 .$$

The transformation happens to be a stretch followed by a shift.

With these two functions, all the quantities – geometric or physical – defined within the two coordinate systems are transformed to each other. For example, suppose the temperature T depends on the location in terms of x as follows:

$$T = x^3 + x^2 + 2 .$$

Then we find how T depends on the location in terms of u by substitution:

$$T = \left(\frac{u+1}{2}\right)^3 + \left(\frac{u+1}{2}\right)^2 + 2 .$$

With any function, the conversion will follows this pattern:

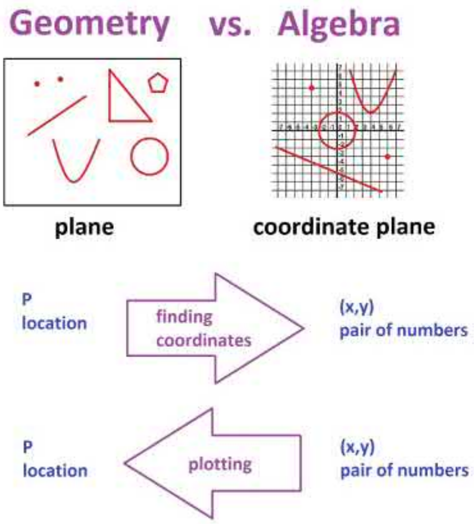
$$\begin{tikzcd}
x \arrow[r] \arrow[d] & T \\
u \arrow[ur] &
\end{tikzcd}$$

Exercise 3.5.2

What transformation isn't mentioned above? Provide an illustration and a formula for the combination of the three.

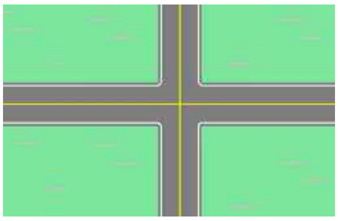
Now the coordinate system for *dimension* 2, the plane.

There is much more going on than before:

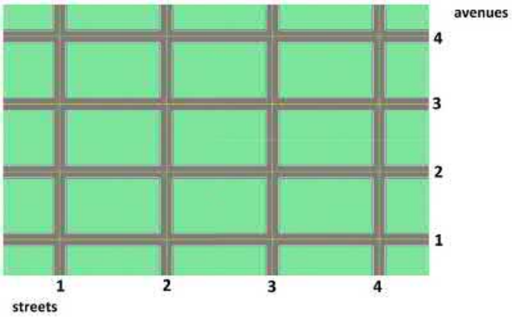


The idea is the same: solving geometric problems with algebra.

Let's repeat – with some minor changes – the construction from Chapter 2 first. Suppose we live on a *field* and we build two roads intersecting at 90 degrees:

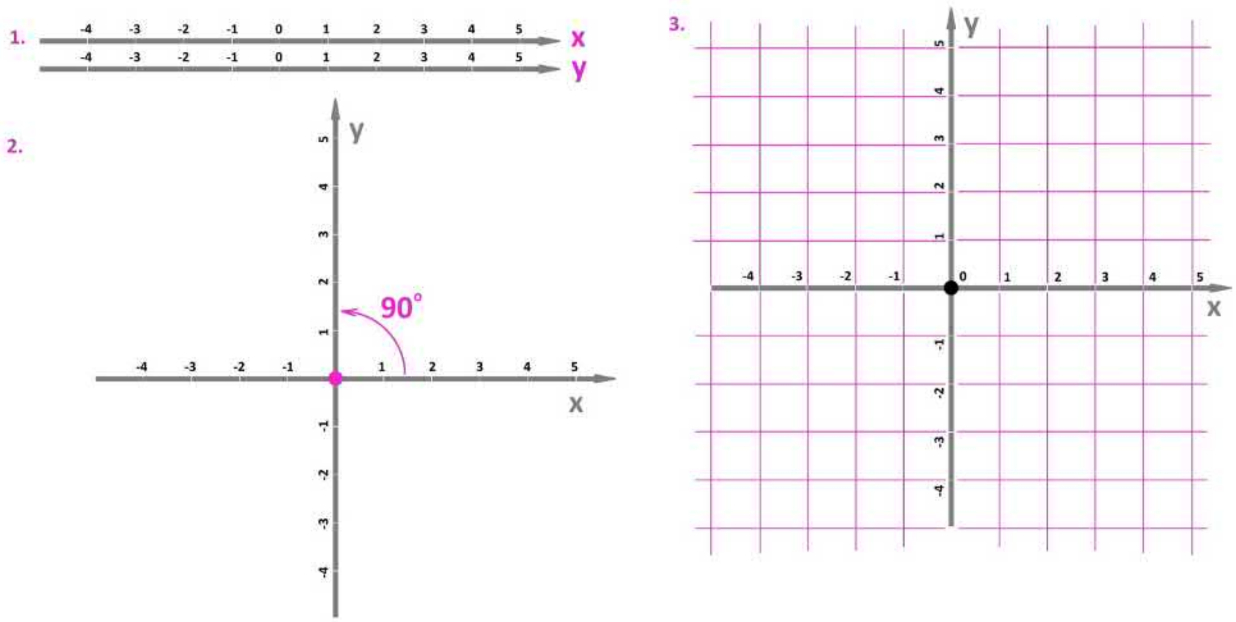


We can then treat either of the two roads as a 1-dimensional Cartesian system, as above, and use their milestones to navigate. But what about the rest of the field? How do we navigate it? We could build a city with a grid of streets:



We also number the streets. We can find locations as intersections of a numbered street and a numbered avenue.

A new coordinate system intended to capture what happens in this city or on this field is devised to be superimposed on the field. It's a grid:



It is built in several stages:

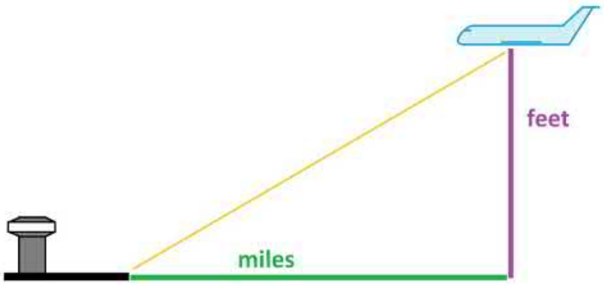
1. Choose two identical coordinate axes, the x*-axis* first and the y-axis second, with the same units.
2. Put the two axes together at their origins so that it is a 90-degree turn from the positive direction of the x-axis to the positive direction of the y-axis.
3. Use the marks on the axes to draw a grid.

Warning!

The xy-plane isn't the same as the yx-plane.

Example 3.5.3: units

It is possible though uncommon to have different units for the two axes:



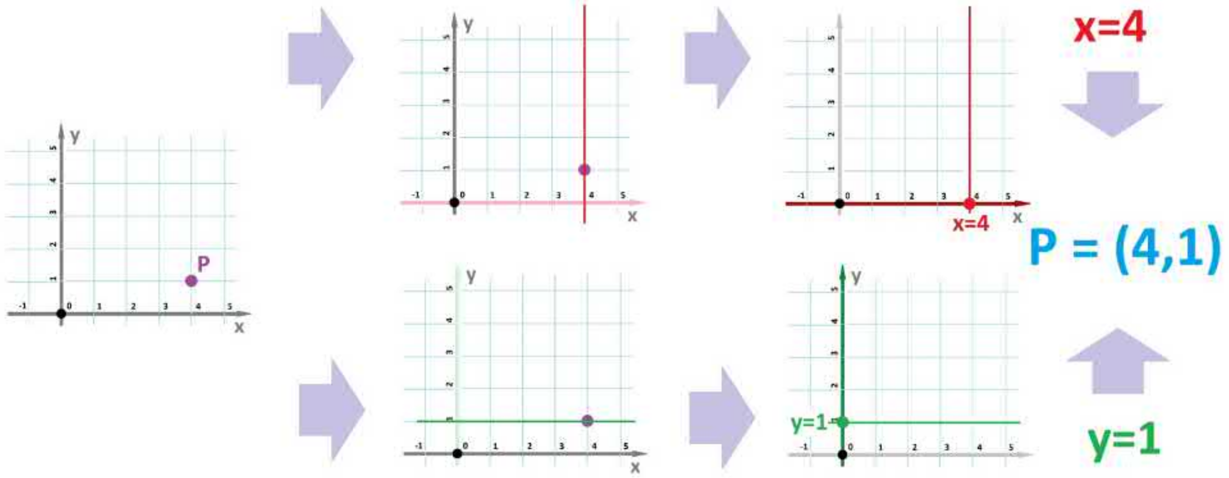
We have a correspondence that works in *both directions*:

location **P** $\longleftrightarrow$ a pair of numbers **(x,y)**

Example 3.5.4: coordinates from point and back

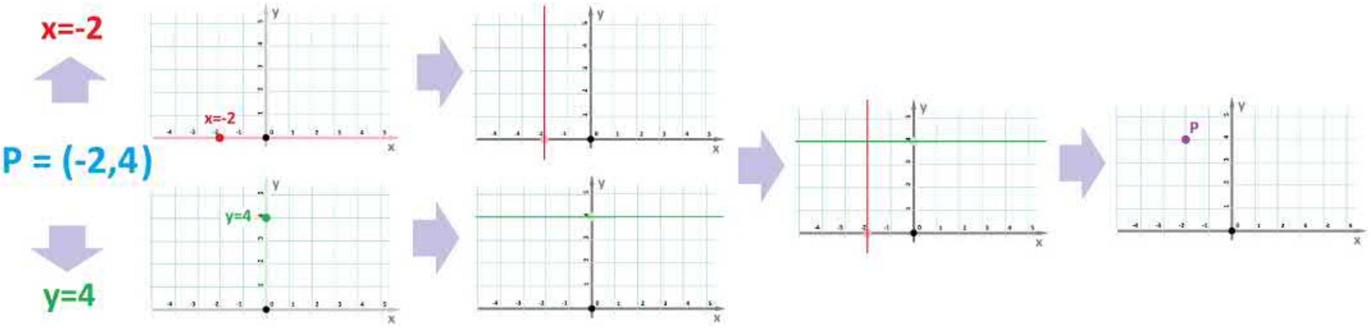
Suppose we have the Euclidean plane equipped with a Cartesian system.

Suppose P is a *point* as shown:



1. We draw a vertical line through P until it intersects the x-axis. The point of intersection then lies on this axis, which is equipped with a 1-dimensional Cartesian system. This point has a coordinate, say 4, within this system.
2. We draw a horizontal line through P until it intersects the y-axis. The point of intersection then lies on this axis, which is equipped with a 1-dimensional Cartesian system. The point has a coordinate, say 1, within this system.
3. We have discovered that our point has coordinates $P = (4, 1)$!

On the flip side, suppose we have two *numbers*, -2 and 4:



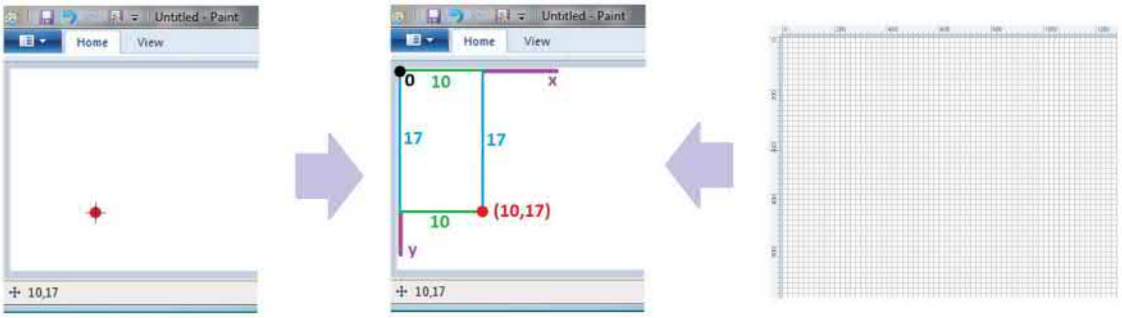
1. We find the location on the x-axis with coordinate -2. We then draw a vertical line through this point.
2. We find the location on the y-axis with coordinate 4. We then draw a horizontal line through this point.
3. The intersection of these two lines is the corresponding *point* $P = (4, 1)$ on the plane!

In summary:

- If P is a *location* on the plane, we find the distances from either of the two axes to that location – positive in the positive direction and negative in the negative direction – and the result is the two coordinates of P, some *numbers* x and y.
- If x and y are *numbers*, we measure x as the distance from the y-axis and y as the distance from the x-axis – positive in the positive direction and negative in the negative direction – and such locations together form a vertical line and a horizontal line, and the intersection of these two is the *location* P on the plane.

Example 3.5.5: coordinates used in computing

The 2-dimensional Cartesian system isn't as widespread as the one for dimension 1 (numbers). It is, however, common in certain areas of computing. For example, drawing applications allow you to make use of this system – if you understand it. The location of your mouse is shown in the status bar on the lower left, constantly updated in real time. The main difference is that the origin is in the left upper corner of the image and the y-axis is pointing *down*:

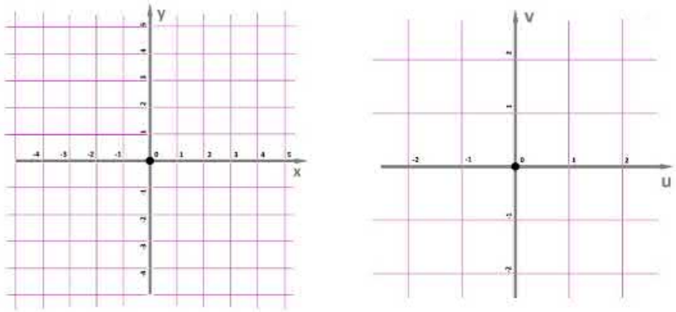


The choice is explained by the way we write: downward.

Example 3.5.6: different coordinate systems, dimension 2

Of course, we can place different coordinate systems on the same plane.

For example, we can have two systems that differ only by scale:



That's a uniform stretch! For example, a point with coordinates $(2,3)$ in the first will have coordinates $(1,1.5)$ in the second. This suggests that the conversion transformation $F:\mathbf{R}^2\to\mathbf{R}^2$ is given by:

$$u = x/2,\ v = y/2\,.$$

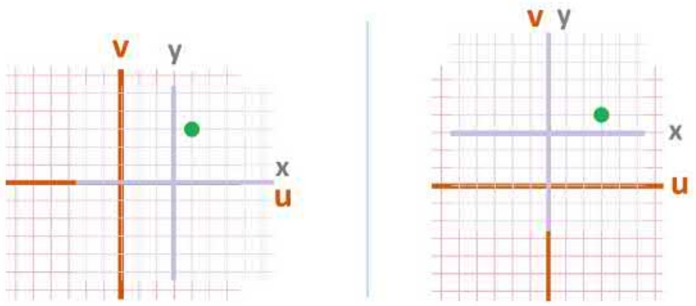
Furthermore, if the temperature of the area depends on the coordinates as, say,

$$T = x^2 + 2y\,,$$

then this dependence in the second coordinate system is

$$T = (2u)^2 + 2(2v)\,.$$

Or they can differ only by the location of the origin:



These are a horizontal and a vertical shifts respectively. The first is given by

$$u = x + 3,\ v = y\,,$$

and the second by:

$$u = x,\ v = y - 4\,.$$

We saw in Chapter 3 how we can transfer information (points, set, functions, etc.) defined on the first plane to the second. It only takes a single function. For the former example, it is

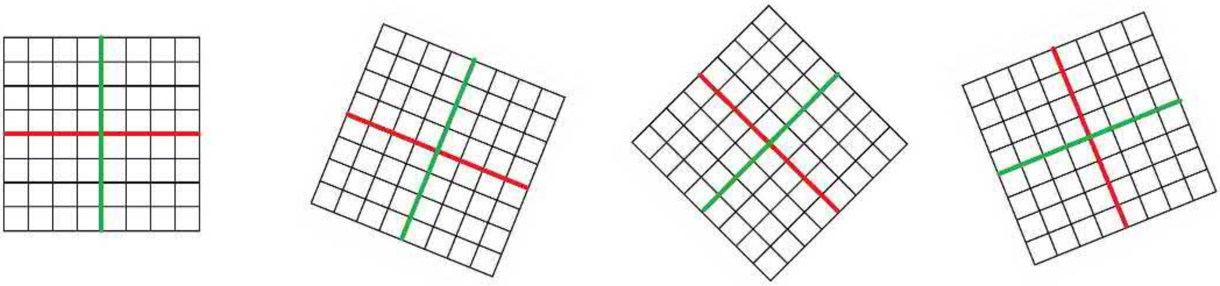
$$(x,y)\mapsto(x/2,y/2)\,.$$

For the latter example, the combination of the two shifts is:

$$(x,y) \mapsto (x-3, y-4)\,.$$

With these functions, all the quantities – geometric or physical – defined within the two coordinate systems are transformed to each other.

Moreover, the coordinate systems can also vary in terms of the *directions* of the axes:



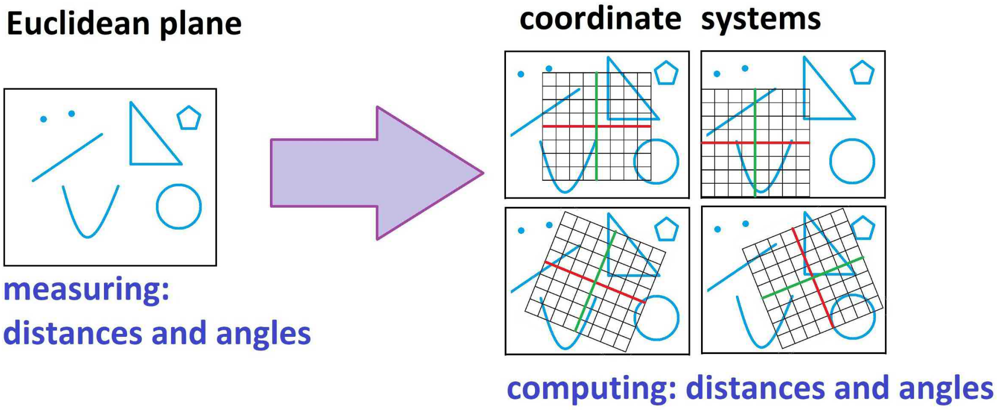
Rotations and other transformations of the plane are considered later.

Just as in the 1-dimensional case, the conversion will follows this pattern:

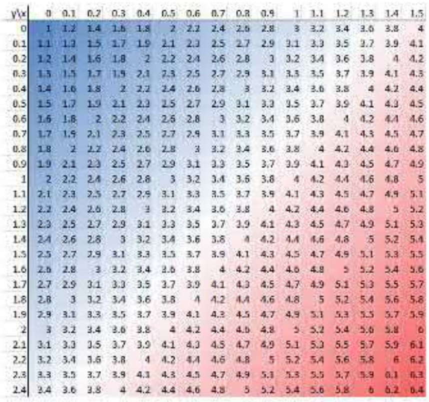
$$\begin{tikzcd}
(x,y) \arrow[r] \arrow[d] & T \\
(u,v) \arrow[ur] &
\end{tikzcd}$$

Coordinate systems...

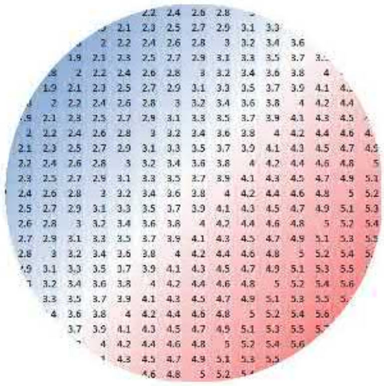
If we want to study the *Euclidean plane*, and Euclidean geometry, algebraically, we *superimpose* the Cartesian grid over this plane:



The geometry on the piece of paper then determines what is going on, not a particular choice of a coordinate system. For example, the direction of fastest growth is determined by the surface itself:

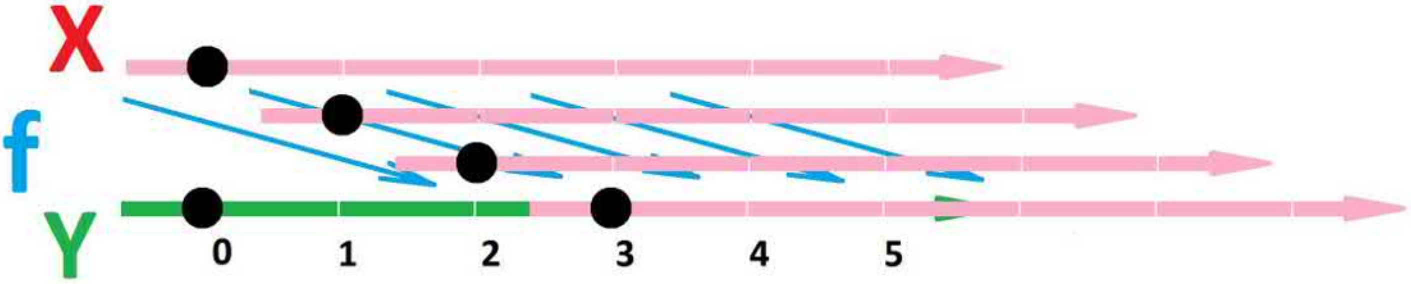


y\x	0	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1	1.1	1.2	1.3	1.4	1.5
0	1	1.2	1.4	1.6	1.8	2	2.2	2.4	2.6	2.8	3	3.2	3.4	3.6	3.8	4
0.1	1.1	1.3	1.5	1.7	1.9	2.1	2.3	2.5	2.7	2.9	3.1	3.3	3.5	3.7	3.9	4.1
0.2	1.2	1.4	1.6	1.8	2	2.2	2.4	2.6	2.8	3	3.2	3.4	3.6	3.8	4	4.2
0.3	1.3	1.5	1.7	1.9	2.1	2.3	2.5	2.7	2.9	3.1	3.3	3.5	3.7	3.9	4.1	4.3
0.4	1.4	1.6	1.8	2	2.2	2.4	2.6	2.8	3	3.2	3.4	3.6	3.8	4	4.2	4.4
0.5	1.5	1.7	1.9	2.1	2.3	2.5	2.7	2.9	3.1	3.3	3.5	3.7	3.9	4.1	4.3	4.5
0.6	1.6	1.8	2	2.2	2.4	2.6	2.8	3	3.2	3.4	3.6	3.8	4	4.2	4.4	4.6
0.7	1.7	1.9	2.1	2.3	2.5	2.7	2.9	3.1	3.3	3.5	3.7	3.9	4.1	4.3	4.5	4.7
0.8	1.8	2	2.2	2.4	2.6	2.8	3	3.2	3.4	3.6	3.8	4	4.2	4.4	4.6	4.8
0.9	1.9	2.1	2.3	2.5	2.7	2.9	3.1	3.3	3.5	3.7	3.9	4.1	4.3	4.5	4.7	4.9
1	2	2.2	2.4	2.6	2.8	3	3.2	3.4	3.6	3.8	4	4.2	4.4	4.6	4.8	5
1.1	2.1	2.3	2.5	2.7	2.9	3.1	3.3	3.5	3.7	3.9	4.1	4.3	4.5	4.7	4.9	5.1
1.2	2.2	2.4	2.6	2.8	3	3.2	3.4	3.6	3.8	4	4.2	4.4	4.6	4.8	5	5.2
1.3	2.3	2.5	2.7	2.9	3.1	3.3	3.5	3.7	3.9	4.1	4.3	4.5	4.7	4.9	5.1	5.3
1.4	2.4	2.6	2.8	3	3.2	3.4	3.6	3.8	4	4.2	4.4	4.6	4.8	5	5.2	5.4
1.5	2.5	2.7	2.9	3.1	3.3	3.5	3.7	3.9	4.1	4.3	4.5	4.7	4.9	5.1	5.3	5.5
1.6	2.6	2.8	3	3.2	3.4	3.6	3.8	4	4.2	4.4	4.6	4.8	5	5.2	5.4	5.6
1.7	2.7	2.9	3.1	3.3	3.5	3.7	3.9	4.1	4.3	4.5	4.7	4.9	5.1	5.3	5.5	5.7
1.8	2.8	3	3.2	3.4	3.6	3.8	4	4.2	4.4	4.6	4.8	5	5.2	5.4	5.6	5.8
1.9	2.9	3.1	3.3	3.5	3.7	3.9	4.1	4.3	4.5	4.7	4.9	5.1	5.3	5.5	5.7	5.9
2	3	3.2	3.4	3.6	3.8	4	4.2	4.4	4.6	4.8	5	5.2	5.4	5.6	5.8	6
2.1	3.1	3.3	3.5	3.7	3.9	4.1	4.3	4.5	4.7	4.9	5.1	5.3	5.5	5.7	5.9	6.1
2.2	3.2	3.4	3.6	3.8	4	4.2	4.4	4.6	4.8	5	5.2	5.4	5.6	5.8	6	6.2
2.3	3.3	3.5	3.7	3.9	4.1	4.3	4.5	4.7	4.9	5.1	5.3	5.5	5.7	5.9	6.1	6.3
2.4	3.4	3.6	3.8	4	4.2	4.4	4.6	4.8	5	5.2	5.4	5.6	5.8	6	6.2	6.4



We can place the coordinate system on top of our physical space in a number of ways...

We start with dimension 1. The *line* can have different coordinate systems assigned to it and those are related to each other via some transformations:

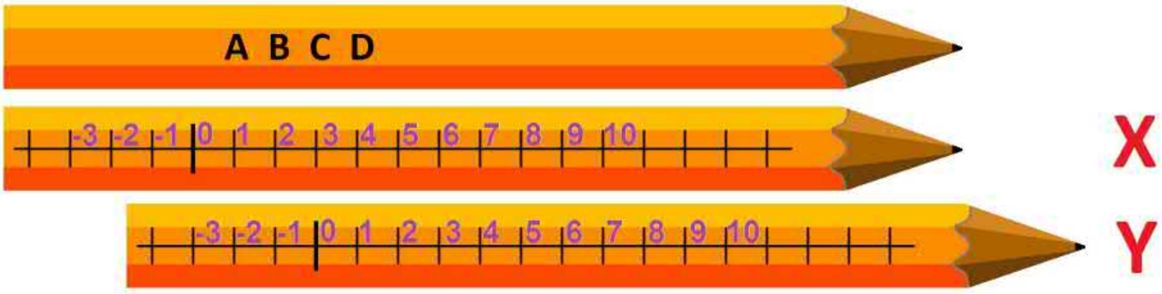


Above you see two ways to interpret the transformation:

1. The arrows are between the x-axis and the intact y-axis or
2. We move the y-axis so that $y = f(x)$ is aligned with x.

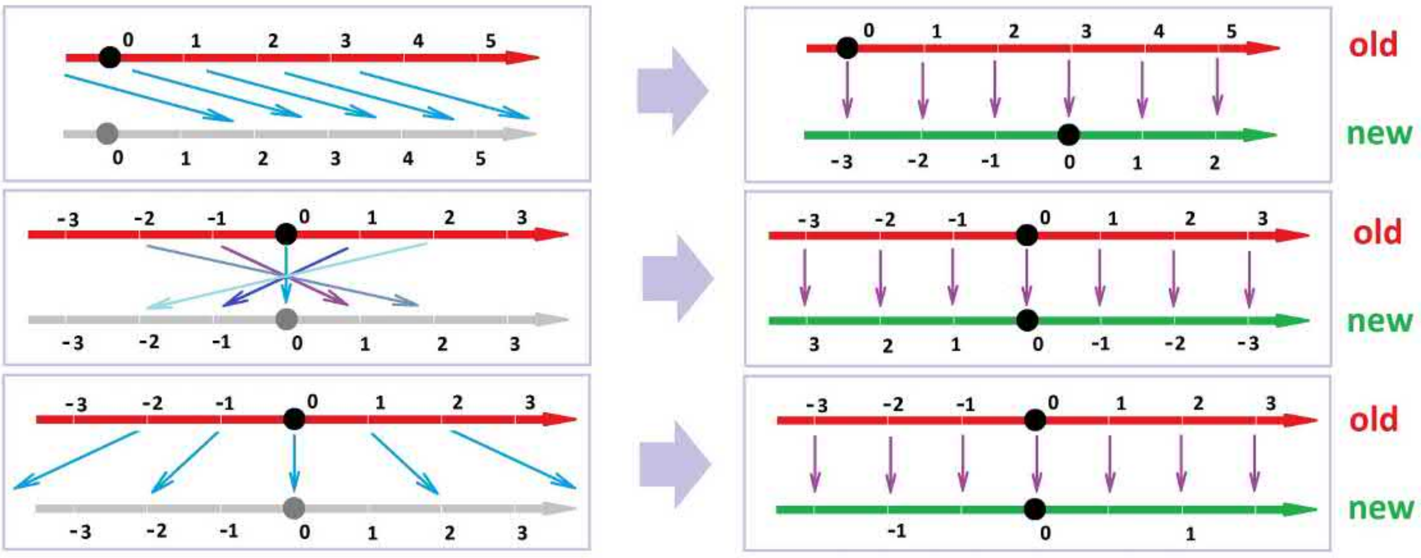
We followed the former in Chapter 1 and we will follow the latter in this section.

We can think as if the whole x-axis is drawn on a pencil:



Each letter will have a *new* coordinate in the *new* coordinate system.

These are the three main transformations of an axis: shift, flip, and stretch (left) and this is what happens to the coordinates (right):



This is the algebra for the basic transformations of the axis, the old and the new coordinates:

$$\begin{array}{ccc} t & \xrightarrow{\text{shift by } k} & x = t - k \\ t & \xrightarrow{\text{flip}} & x = -t \\ t & \xrightarrow{\text{stretch by } k} & x = t/k \end{array}$$

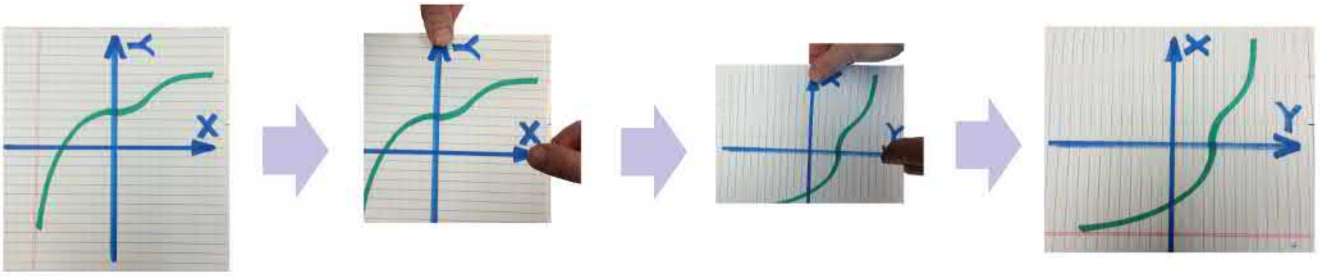
Now dimension 2, the *plane*.

Both x and y-axes can be subjected to the transformations above. The change of coordinates under the resulting six basic transformations of the xy-plane is shown below:

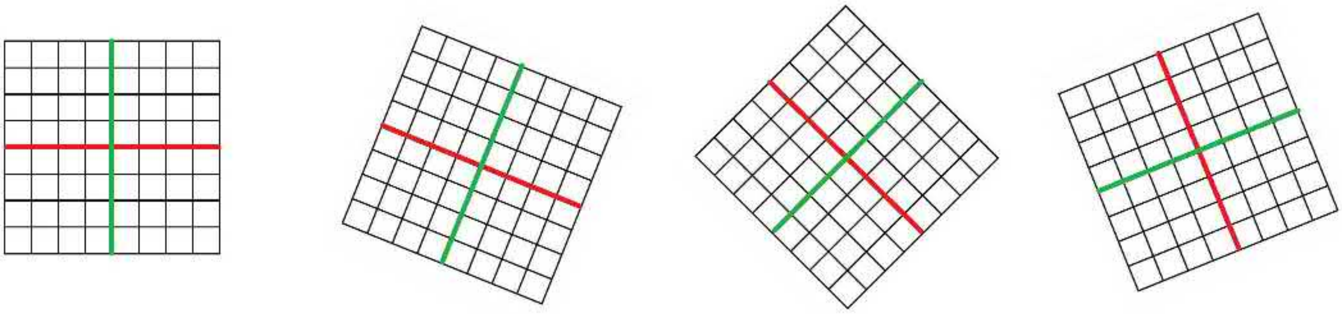
$$\begin{array}{llllll}
\text{vertical shift:} & \begin{array}{ccccc}(& x & , & y &)\\ (& x & , & y-k &)\end{array}, & \text{flip:} & \begin{array}{ccccc}(& x & , & y &)\\ (& x & , & y\cdot(-1) &)\end{array}, & \text{stretch:} & \begin{array}{ccccc}(& x & , & y &)\\ (& x & , & y/k &)\end{array},\\
\text{horizontal shift:} & \begin{array}{ccccc}(& x & , & y &)\\ (& x-k & , & y &)\end{array}, & \text{flip:} & \begin{array}{ccccc}(& x & , & y &)\\ (& x\cdot(-1) & , & y &)\end{array}, & \text{stretch:} & \begin{array}{ccccc}(& x & , & y &)\\ (& x/k & , & y &)\end{array}.
\end{array}$$

Example 3.5.7

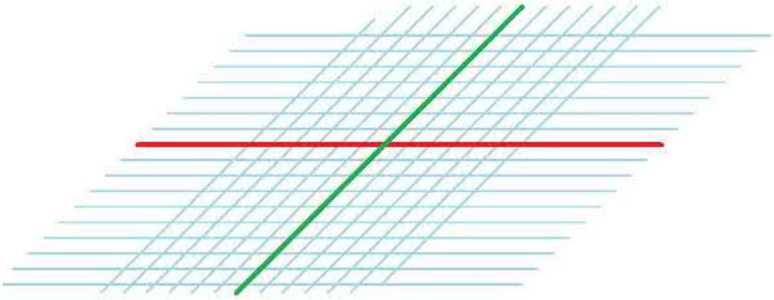
Some transformations cannot be reduced to a combination of these six. Recall from Chapter 1, that in order to find the graph of the inverse function, we execute a *flip about the diagonal* of the plane. We grab the end of the x-axis with the right hand and grab the end of the y-axis with the left hand then interchanging them:



We face the opposite side of the paper then, but the graph is still visible: the x-axis is now pointing up and the y-axis right. The axes can be *rotated*, together:



The coordinate will change but they will still unambiguously determine a location on the plane. The axes can be *skewed*:



Even then the two numbers indicating the intersection of two lines will unambiguously determine a location on the plane. And so on... Further analysis is presented in Chapter 1.

3.6. The Euclidean plane: distances

The distance – a number – between any two locations P and Q – on the line, on the plane, or in space – is assumed to be available.

The distance between two points P and Q is denoted as follows:

Distance on line

$d(P,Q)$

Warning!

The distance must never be negative.

The distance is inherited from the Euclidean line that underlies the Cartesian system. Since everything in the Cartesian system is pre-measured, we can solve some geometric problems by algebraically manipulating the coordinates of points.

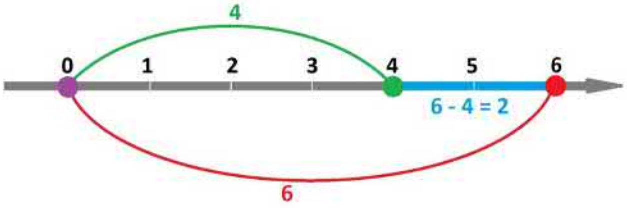
In this section, we consider the very basic geometric task of computing – as opposed to measuring – *distances*.

First, the *line*.

Now, how do we express this number in terms of their coordinates, say x and x'?

Example 3.6.1: distance dim 1

One finds the distance that has been covered on the road by *subtracting* the number on the milestone in the beginning and the number on the milestone at the end:



Here is the algebra:

$$\text{from } P = 4 \text{ to } Q = 6 \implies \text{distance } = Q - P = 6 - 4 = 2\,.$$

But what if we are moving in the opposite direction? The distance should be the same! And the computation should be the same:

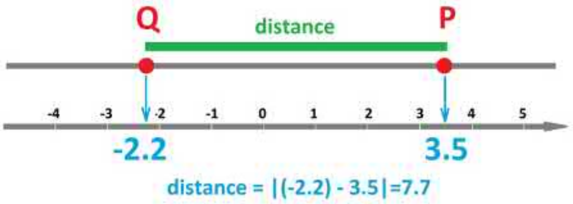
$$\text{from } Q = 6 \text{ to } P = 4 \implies \text{distance } = Q - P = 6 - 4 = 2\,.$$

In other words, one must subtract the smaller number from the larger one every time in order for the computation to make sense.

We conclude that the *distance between two locations P and Q* on the real line given by their coordinates x and x' is:

- $d(P,Q) = x' - x$ when $x < x'$,
- $d(P,Q) = x - x'$ when $x > x'$,
- $d(P,Q) = 0$ when $x = x'$.

Is there a *single formula* for this computation? The idea that the distance between two locations can never be negative suggests that this has something to do with the absolute value:



The *absolute value* function is defined to be

$$|a| = \begin{cases} -a & \text{when } a < 0\,, \\ 0 & \text{when } a = 0\,, \\ a & \text{when } a > 0\,. \end{cases}$$

We just substitute $a = x' - x$ into this formula to prove the following:

Theorem 3.6.2: Distance Formula for Dimension 1

The distance between two points on the real line with coordinates x and x' is the absolute value of their difference (in either order).

In other words, we have:

$$\boxed{d(x, x') = |x - x'| = |x' - x| = d(x', x)}$$

Exercise 3.6.3

Derive the formula from the following:

$$d(P, Q) = \begin{cases} x - x' & \text{when } x > x'\,, \\ 0 & \text{when } x = x'\,, \\ x - x' & \text{when } x < x'\,. \end{cases}$$

Exercise 3.6.4

Prove that for any two numbers x, x', we have $|x + x'| \le |x| + |x'|\,.$

Exercise 3.6.5

Prove that the point half-way between points $P = x$ and $Q = x'$ (called their "midpoint") has the coordinate $\dfrac{x + x'}{2}$.

The word "stretch" that we have used in the past now takes a precise meaning. We can rely on the idea of distance. For example, both $y = 2x$ and $y = -2x$ double the distances. The following is a general result about all linear transformations:

Theorem 3.6.6: Linear Transformations in Dimension 1

A linear function stretches the x-axis by a factor of $|m|$, where m is its slope.

Proof.

This is what happens to the distance between two points u and v, which is $|v - u|$, after a linear function $f(x) = mx + b$ is applied:

$$|f(v) - f(u)| = |(mv + b) - (mu + b)| = |mv - mu| = |m| \cdot |v - u|\,.$$

The distance has increased by a factor of $|m|$. (We say that it has decreased by a factor of m when $|m| < 1$). This stretch/shrink factor is the same everywhere.

In other words, stretching means that the distances increase or decrease proportionally.

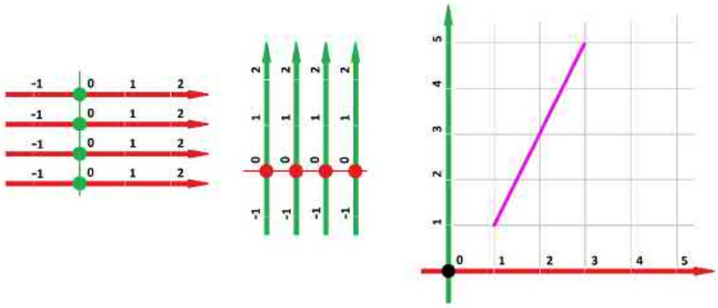
Next, the *plane*.

Once again, the value of the distance between two locations P and Q is inherited from the Euclidean plane that underlies the Cartesian system.

So, the *Distance Formula for Dimension* 1 allows us to compute distances along the axis:

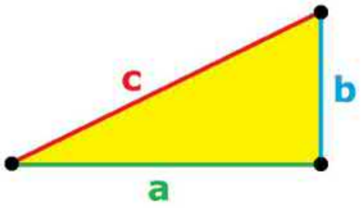
$$\boxed{d(x,x') = |x - x'|}$$

If there are two axes, the formula still allows us to compute distances along either. Moreover, we compute distances along lines *parallel* to either of the axes:



But what about the diagonal directions?

We have a tool:



We present a result that is one of the most important:

Theorem 3.6.7: Pythagorean Theorem

Suppose we have a right triangle with sides a, b, c, with c the longest one facing the right angle. Then, we have the following:

$$\boxed{a^2 + b^2 = c^2}$$

Proof.

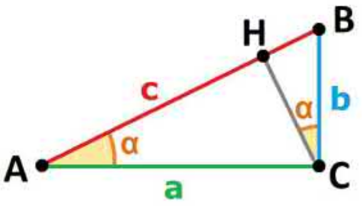
We use what we know about similar triangles (i.e., the ones with equal angles):

▶ The ratio of any two corresponding sides of similar triangles is the same.

Let ABC be our right triangle, with:

- vertex A opposite to side a,
- vertex B opposite to side b,
- vertex C opposite to side c.

We draw the height (the line perpendicular to c) from C, and call H its intersection with the side c:



The new triangle ACH is similar to our original triangle ABC, because they both have a right angle, and they share the angle at A, α. In the same way, we prove that the triangle CBH is also similar to ABC. The similarity of these two pairs of triangles leads to the equality of ratios of the corresponding sides:

$$\begin{array}{c|c} \dfrac{BC}{AB}=\dfrac{BH}{BC} & \dfrac{AC}{AB}=\dfrac{AH}{AC} \\ BC^2=AB\cdot BH & AC^2=AB\cdot AH \end{array}$$

We add the two items in the last row and factor:

$$b^2+a^2=BC^2+AC^2=AB\cdot BH+AB\cdot AH=AB\cdot(AH+BH)=AB^2=c^2\,.$$

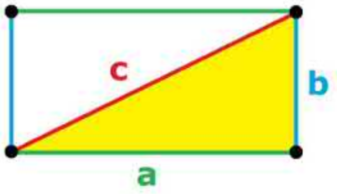
Exercise 3.6.8

State the converse of the theorem. Is it true?

Exercise 3.6.9

What do the ratios in the proof tell us about the trigonometric functions of the angle α?

We now realize that the value of the theorem for us is its treatment of rectangles:

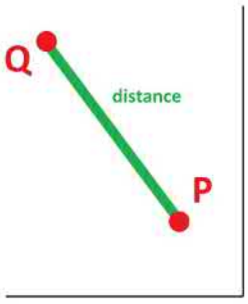


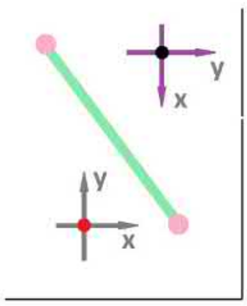
Corollary 3.6.10: Pythagorean Theorem For Rectangles

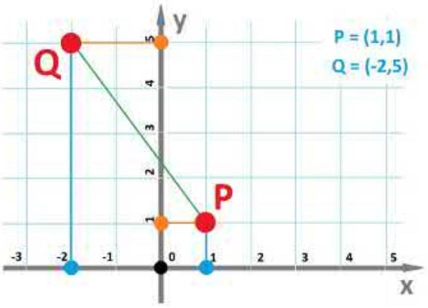
The square of the length c of the diagonal of a rectangle with sides a, b is given by the following:

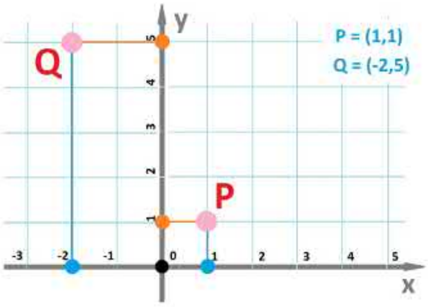
$$\boxed{c^2=a^2+b^2}$$

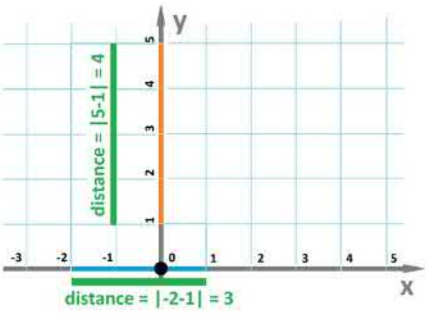
So, we have a Cartesian system placed on top of this piece of paper. How do we express the distance between P and Q in terms of their coordinates (x, y) and (x', y')? We find the distances along the axes first:

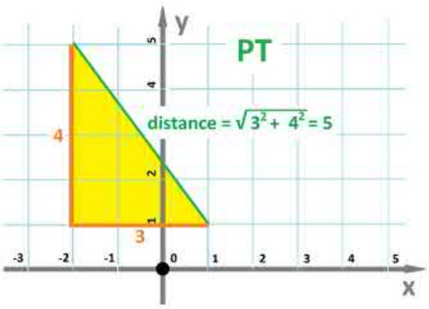












We applied the *Distance Formula for Dimension* 1 for either of the two axes and then used the *Pythagorean Theorem.*

Exercise 3.6.11

What would the affect on this computation be if we shift the grid to horizontally, or vertically, or both?

The following is one of the most useful results in geometry of the Cartesian plane.

Theorem 3.6.12: Distance Formula for Dimension 2

The distance between two points with coordinates $P = (x, y)$ *and* $Q = (x', y')$ *is*

$$\boxed{d(P, Q) = \sqrt{(x - x')^2 + (y - y')^2}}$$

Proof.

According to the formula:

- The distance between x and x' on the x-axis is $|x - x'|$.
- The distance between y and y' on the y-axis is $|y - y'|$.

Then, the segment between the points $P(x, y)$ and $Q = (x', y')$ is the hypotenuse of the right triangle with sides: $|x - x'|$ and $|y - y'|$. Then our conclusion below follows from the *Pythagorean Theorem*:

$$d(P, Q)^2 = |x - x'|^2 + |y - y'|^2 .$$

Since $|z|^2 = z^2$ for any z, we can remove the absolute value signs.

The result is so important that one can even say that the 90-degree angle between the axes was chosen so that we can produce this formula from the Pythagorean theorem.

Exercise 3.6.13

Find the distance between the points $(-5, 2)$ and $(2, -1)$.

Warning!

Combine x's with x's and y's with y's.

We now have two formulas for the two cases: dimension 1 and 2. They look quite different. However, if we square both formulas, this is how they can be matched up:

	Euclidean		Cartesian		
dimension	diagonal		x-axis		y-axis
1	$d(P, Q)^2$	$=$	$(x - x')^2$		
2	$d(P, Q)^2$	$=$	$(x - x')^2$	$+$	$(y - y')^2$

We just have an extra term for an extra axis!

This is how we can state both formulas as one, verbally:

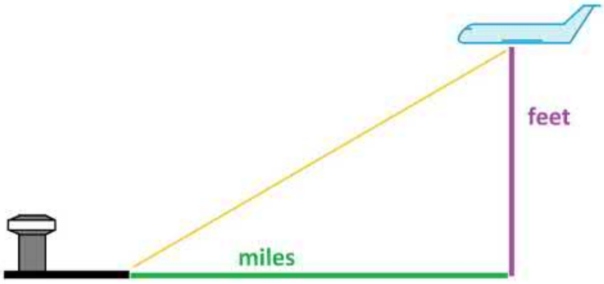
Theorem 3.6.14: Distance Formula for Dimensions 1 and 2

For dimensions 1 and 2, the square of the distance is the sum of the squares of the differences of the coordinates.

This reformulation of the two theorems allows us to guess that it might apply to higher dimensions: the sum of one, two, *three* squares, etc. We will see a continuation of this list and of this pattern in higher dimensions.

Example 3.6.15: miles and kilometers

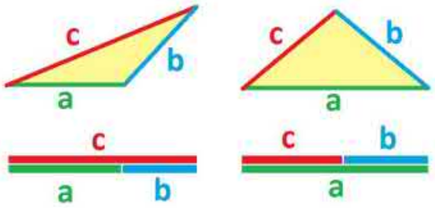
It is possible to have the x-axis measured in different units from the y-axis. For example, it is typical to measure the distance to the airport in miles but the altitude in feet:



Also, one can speak, hypothetically, of a point located "2 miles east and 5 kilometers north" from here. However, when this is the case, the Distance Formula won't be applicable anymore!

Here is a familiar property of the lengths of the sides of a triangle: The length of any side is less than the sum of the lengths of the other two. In other words, if a, b, c are these three sides, then:

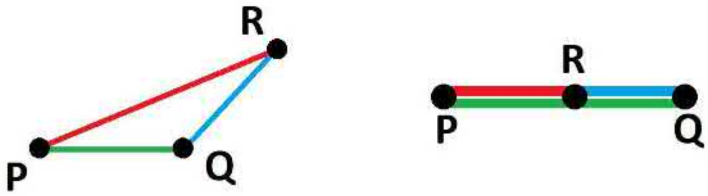
$$c < a + b\,.$$



As you can see (bottom row), even when the triangle "degenerates" to a segment, we have an equation:

$$c = a + b\,.$$

Thus, the inequality remains true though non-strict. It is, therefore, applicable to distances between points:



We restate this extended inequality as follows.

Theorem 3.6.16: Triangle Inequality

For any three points P, Q, R on the plane, we have the following:

$$\boxed{d(P,R) \le d(P,Q) + d(Q,R)}$$

In other words,

- The straight line is the shortest.

The theorem helps to confirm that the concept of distance matches our intuition.

Exercise 3.6.17

Derive the theorem from the Distance Formula.

Exercise 3.6.18

Derive the inequality for a right triangle from the Pythagorean Theorem.

When the triangle degenerates into a segment, we have the triangle inequality for dimension 1 presented above.

Exercise 3.6.19

Prove that the point M half-way between points $P=(x,y)$ and $Q=(x',y')$ – called their *midpoint* – is given by their average coordinates:

$$M=\left(\frac{x+x'}{2},\frac{y+y'}{2}\right).$$

Now, from distances to angles.

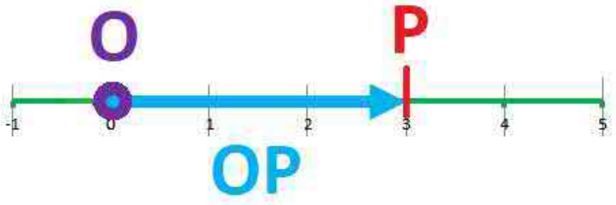
3.7. The Euclidean plane: angles

The Greek word "trigonometry" means "measuring triangles". From the study of the angles of triangles, however, it is developed into a study of the angles of rotations. We will follow this approach in this section and consider *angles between directions.*

This is the second task of analytic geometry.

What does a *direction* on the real line, or a plane, mean? We will pursue the approach via *vectors.*

First, in the 1-*dimensional* Euclidean space, i.e., nothing but the x-axis, vectors are segments of this line:



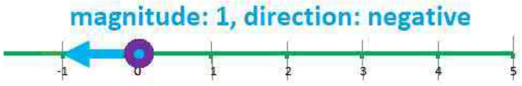
Definition 3.7.1: vector in dimension 1

If a line segment's starting point is the origin, i.e., OP for some point $P\neq O$, it is called a (1-dimensional) *vector* in $\mathbf{R}$. A vector has two attributes:

- its *magnitude*, which is the absolute value of the coordinate x of its terminal point: $|x|$; and
- its *direction*, which is either positive or negative.

The *zero vector* OO has zero magnitude and undefined direction.

For example:



Now, how do we *compare* the directions of two vectors?

Suppose we have two points: $P \neq O$, $Q \neq O$. We deal with the directions from the origin O toward locations P and Q. Of course, there can be only two outcomes:

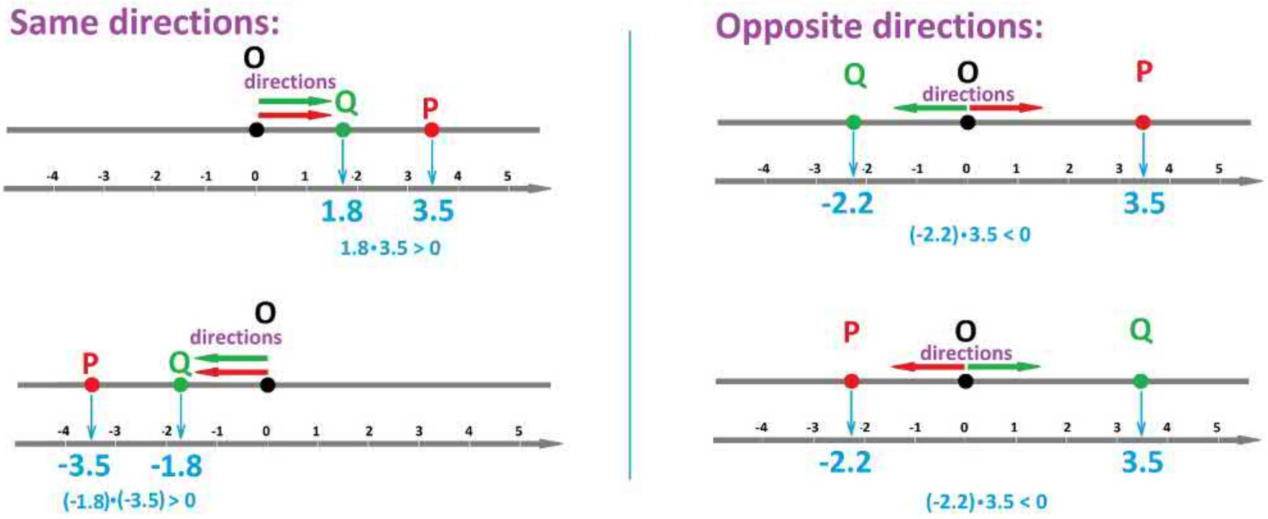
- If P and Q are on the same side of O, then the *directions are same*:

$$P\ Q\ \longleftarrow O \longrightarrow \quad \text{or} \quad \longleftarrow O \longrightarrow\ P\ Q$$

- If P and Q are on the opposite sides of O, then the *directions are opposite*:

$$P\ \longleftarrow O \longrightarrow\ Q \quad \text{or} \quad Q\ \longleftarrow O \longrightarrow\ P$$

Let's examine the coordinates of the points with those *four* possibilities:



When the two vectors are represented by their coordinates, x and x', the analysis of their directions becomes algebraic:

- If $x > 0,\ x' > 0$ or $x < 0,\ x' < 0$, then the directions are the same.
- If $x > 0,\ x' < 0$ or $x < 0,\ x' > 0$, then the directions are the opposite.

Fortunately, the *product* provides us with a single expression that makes this determination. We just multiply the coordinates:

locations:	$P\ Q\ \leftarrow O \rightarrow$	$\leftarrow O \rightarrow\ P\ Q$	same directions
signs of coordinates:	$-\ -$	$+\ +$	
sign of product:	$- \cdot - \ = \ +$	$+ \cdot + \ = \ +$	$\boxed{+}$
locations:	$P\ \leftarrow O \rightarrow Q$	$Q\ \leftarrow O \rightarrow\ P$	opposite directions
signs of coordinates:	$-\quad +$	$-\quad +$	
sign of product:	$- \cdot + \ = \ -$	$- \cdot + \ = \ -$	$\boxed{-}$

The problem is solved:

Theorem 3.7.2: Directions for Dimension 1

The directions from 0 *to* $x \neq 0$ *and* $x' \neq 0$ *are*
1. *the same when* $x \cdot x' > 0$; *and*
2. *the opposite when* $x \cdot x' < 0$.

Let's restate the theorem in terms of the sign function:

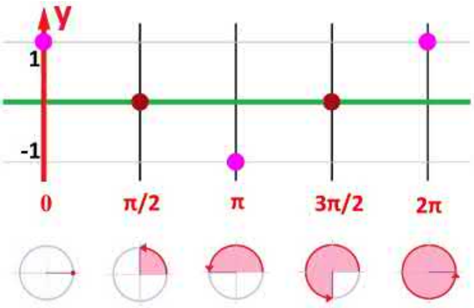
sign of the product	directions	vectors	angle
$\text{sign}(x \cdot x') = 1$	same	$\leftleftarrows \quad \rightrightarrows$	0 degrees
$\text{sign}(x \cdot x') = -1$	opposite	$\rightarrow \ \leftarrow$	180 degrees

We also measure the actual angles between the vectors (last column).

Consider the last correspondence:

$$1 \ \leftrightarrow \ 0 \ \text{ degrees and } \ -1 \ \leftrightarrow \ 180 \ \text{ degrees}$$

We've see it before:

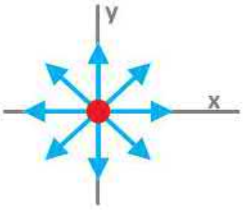


It's the *cosine*!

In summary, this is the idea we'll take to solve the problem of angles on the Cartesian plane:

- ▶ Multiplying the coordinates gives us the cosine of the angle.

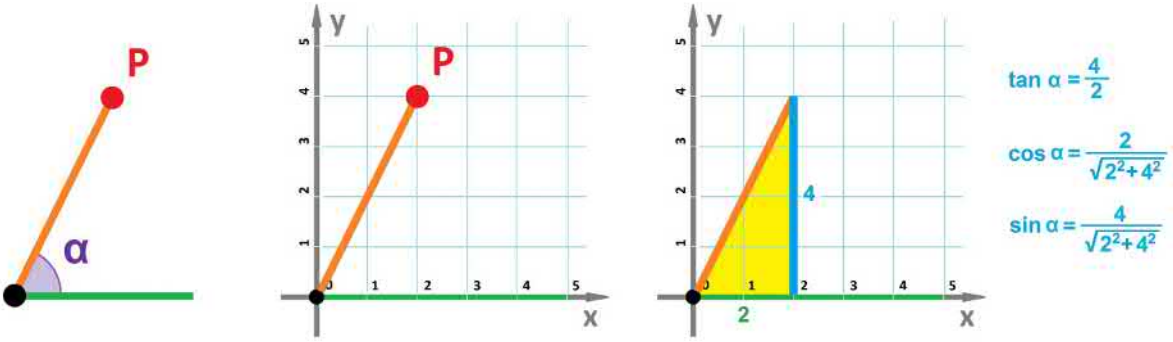
Next, the 2-*dimensional* Euclidean space, i.e., the xy-plane. There are infinitely many directions now:



The issue of the direction of a single line (or a vector) has been solved: It's the *slope*!

Example 3.7.3: trigonometry

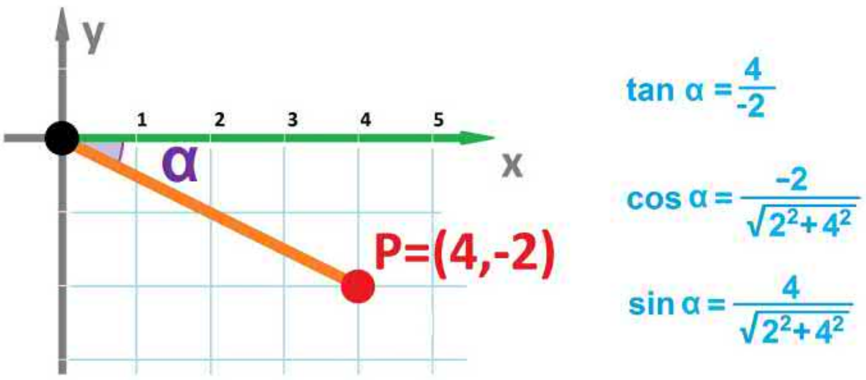
he question is the one about the angle between the line OP with the x-axis. It is determined from the coordinates of $P = (x, y)$ via this simple trigonometry:



The angle has been found:

$$\alpha = \tan^{-1} 2 \, .$$

We get much more from this: The sides of a triangle can't be negative, but the coordinates can. We don't deal with an actual triangle anymore:



Exercise 3.7.4

Explain these computations.

We put these trigonometric formulas forward as theorems. The first one has been especially important:

Theorem 3.7.5: Slope is Tangent

The tangent of the angle α between the x-axis and the line from O to a point $P=(x,y)\neq O$ is equal to the slope of this line:

$$\boxed{\tan\alpha=\frac{y}{x}}$$

The other two formulas will be used later in this section:

Theorem 3.7.6: Sine and Cosine of Direction

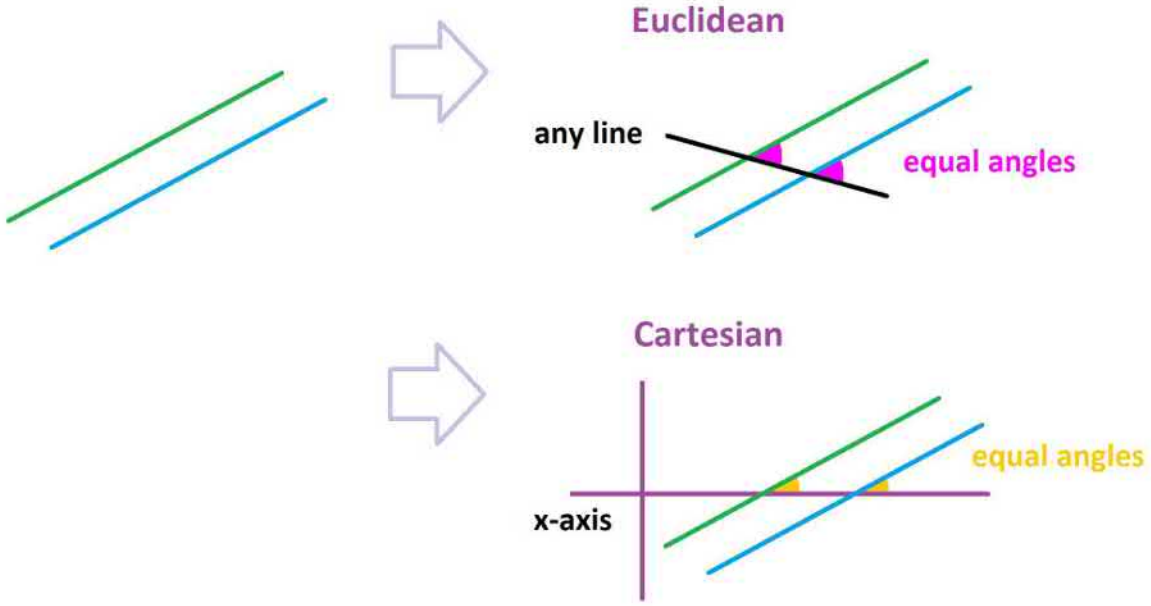
The sine and the cosine of the angle α between the x-axis and the line from O to a point $P=(x,y)\neq O$ are given by:

$$\boxed{\begin{aligned}\cos\alpha&=\frac{x}{\sqrt{x^2+y^2}}\\ \sin\alpha&=\frac{y}{\sqrt{x^2+y^2}}\end{aligned}}$$

Exercise 3.7.7

Prove the theorem.

Recall from Euclidean geometry that two lines are called *parallel* if the angle they form with another line are equal in magnitude:



If we choose this other line to be the x-axis (bottom), we can apply the above theorem to these angles:

$$\text{equal angles} \implies \text{equal tangents} \implies \text{equal slopes}$$

We conclude the following:

Theorem 3.7.8: Parallel Lines, Same Slope

Two (non-vertical) lines on the xy-plane are parallel if and only if they have equal slopes.

In other words, we have:

$$\boxed{y = mx + b \mid\mid y = m'x + b' \iff m = m'}$$

Example 3.7.9

Find the line parallel to $y = 2x$ that passes through the point $(1, 1)$. The slope of the new line will have to be 2. We then just use the point-slope formula for this line:

$$y - 1 = 2(x - 1)\,.$$

Exercise 3.7.10

Find the line parallel to $y = -3x + 2$ that passes through the point $(1, 2)$. Suggest another line and repeat.

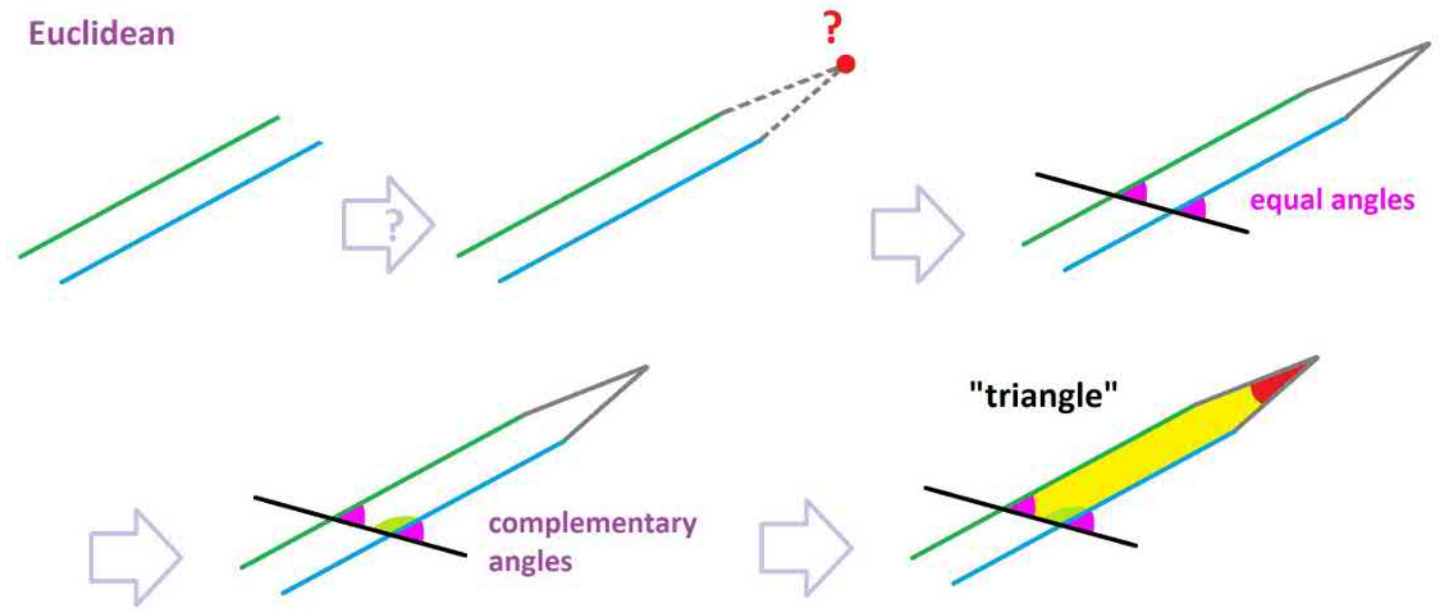
Exercise 3.7.11

Split the theorem into a statement and its converse.

Exercise 3.7.12

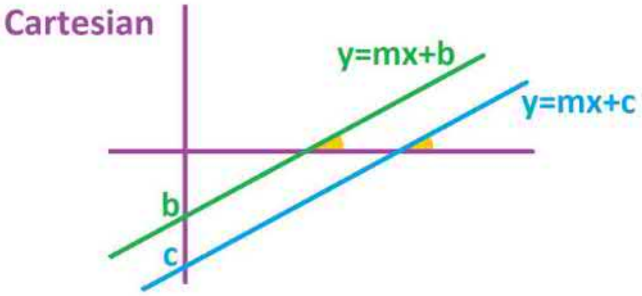
The case not covered by the theorem is a pair of two vertical lines. Show that they are all parallel to each other and not parallel to other lines.

Lines on the plane are its subsets. It is, therefore, natural to ask about their intersections. We know from Euclidean geometry that two parallel lines don't intersect:



If they did, they'd form a triangle with the sum of the angles above 180 degrees.

The Cartesian system makes it possible to prove this fact algebraically:



Corollary 3.7.13: Parallel Lines Don't Intersect

1. *Parallel lines don't intersect.*
2. *Conversely, non-parallel lines intersect.*

Proof.

1. According to the last theorem, the two lines have the same slope, m, and two y-intercepts, b and c. Suppose the y-intercepts are different, $b \neq c$. These are the equations of the lines:

$$\begin{array}{lllll} y & = & mx & +b & \texttt{AND} \\ y & = & mx & +c & \end{array}$$

For a point (x, y) to belong to the intersection, it would have to satisfy both. We have a system of linear equations! Subtracting the two equations produces: $b - c = 0$. There is no solution and, therefore, no intersection.

2. Suppose the two lines have slopes, m and n, and two y-intercepts, b and c. Suppose the slopes are different, $m \neq n$. These are the equations of the lines:

$$\begin{array}{lllll} y & = & mx & +b & \texttt{AND} \\ y & = & nx & +c & \end{array}$$

For a point (x, y) to belong to the intersection, it would have to satisfy both. We solve this system of linear equations as before.

Exercise 3.7.14

Finish the proof of the converse part of the corollary.

Exercise 3.7.15

State the corollary as an equivalence (an “if-and-only-if” statement).

Exercise 3.7.16

Prove that vertical lines don’t intersect each other and do intersect all other lines.

Example 3.7.17: mixtures, what can happen

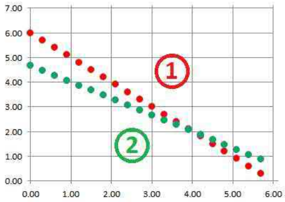
Recall an example from Chapter 2: Is it possible to create, from the Kenyan coffee (\$2 per pound) and the Colombian coffee (\$3 per pound), 6 pounds of blend with a total price of \$14? The problem is solved via a system of linear equations:

$$\left\{\begin{array}{lll} x & +y & =6 \qquad \texttt{AND} \\ 2x & +3y & =14\,. \end{array}\right.$$

Without even solving it, we follow this line of thought:

- The slopes of the two lines are different; therefore,
- the lines are not parallel; therefore,
- the lines intersect; therefore,
- the system has a solution.

Confirmed:



So, it is *possible* to create such a blend! It would be impossible if both types of coffee were priced at \$2 per pound.

Exercise 3.7.18

Justify the last statement – second possibility – using the theorem about parallel lines. What is the third possibility?

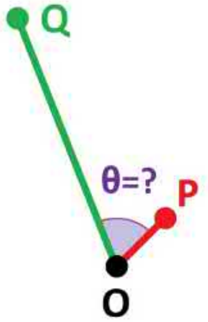
Using the slopes as stand-ins for directions of lines is very efficient but incomplete.

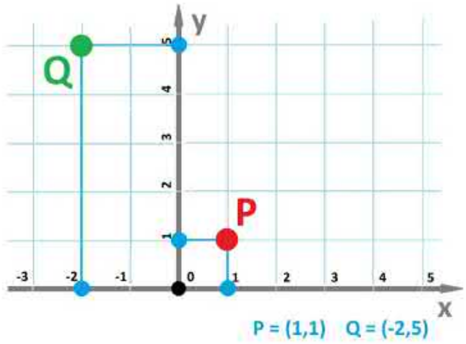
The value of the *angle between the directions* (i.e., the lines) from O to point P and from O to point Q comes from the triangle OPQ. This triangle and all of its measurements is inherited from the Euclidean plane that underlines the Cartesian system. The angle is denoted by $\widehat{QOP}$. Now, the question is:

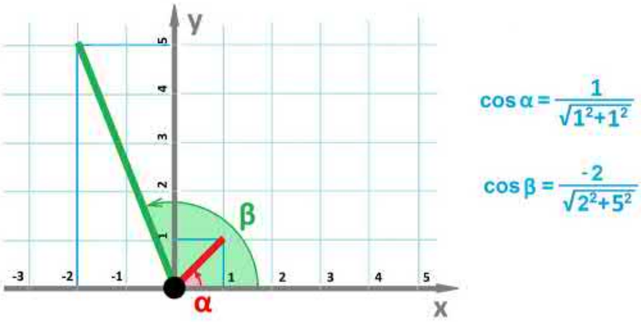
▶ How do we express $\widehat{QOP}$ in terms of the coordinates of the points $P=(x,y)$ and $Q=(x',y')$?

Above we considered a special case: $Q=(1,0)$.

The geometry is illustrated below:







Exercise 3.7.19

Find θ by finding the slopes.

We have the following from the theorem above:

	$P = (x,y)$	$Q = (x',y')$	
horizontal	$\cos\alpha = \dfrac{x}{\sqrt{x^2+y^2}}$	$\cos\beta = \dfrac{x'}{\sqrt{x'^2+y'^2}}$	$\dfrac{\text{runs}}{\text{distances}}$
vertical	$\sin\alpha = \dfrac{y}{\sqrt{x^2+y^2}}$	$\sin\beta = \dfrac{y'}{\sqrt{x'^2+y'^2}}$	$\dfrac{\text{rises}}{\text{distances}}$

The formulas look complicated, but keep in mind that the denominators are just the distances from O to P and Q, respectively. The numerators are the runs and rises.

However, our interest isn't these two angles but their difference,

$$\theta = \widehat{QOP} = \alpha - \beta\,.$$

Fortunately, there is another trigonometric formula – Sine and Cosine of Difference – that allows us to represent the cosine of this angle in terms of the four quantities above:

$$\begin{aligned}\cos\theta = \cos(\alpha-\beta) &= \cos\alpha\cos\beta + \sin\alpha\sin\beta \\ &= \frac{x}{\sqrt{x^2+y^2}}\frac{x'}{\sqrt{x'^2+y'^2}} + \frac{y}{\sqrt{x^2+y^2}}\frac{y'}{\sqrt{x'^2+y'^2}} && \text{We substitute.} \\ &= \frac{xx'+yy'}{\sqrt{x^2+y^2}\sqrt{x'^2+y'^2}}\,. && \text{And simplify.}\end{aligned}$$

The two parts of the denominator are the distances to $P = (x,y)$ and $Q = (x',y')$:

$$d(O,P) = \sqrt{x^2+y^2} \quad\text{and}\quad d(O,Q) = \sqrt{x'^2+y'^2}\,.$$

Therefore, we have:

$$\cos\theta = \frac{xx'+yy'}{d(O,P)\cdot d(O,Q)}\,.$$

But what about the numerator? We see multiplication of the coordinates, as expected.

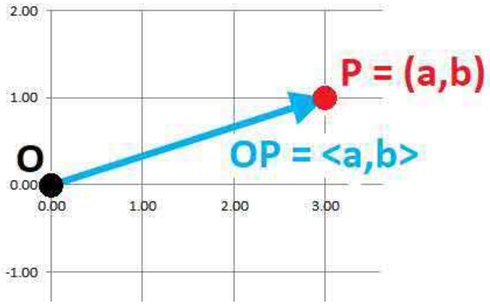
Exercise 3.7.20

Show that the angle found is dependent on the directions and not on the distances from O.

We take the *vector approach* again:

Definition 3.7.21: vector in dimension 2

If a segment's starting point is the origin, i.e., it's OP for some P, it is called a (2-dimensional) *vector* in $\mathbf{R}^2$.



Definition 3.7.22: vector in xy-plane

The *components* of vector OP are the coordinates of its terminal point P, according to the following notation:

$$\boxed{P = (a, b) \iff OP =< a, b >}$$

Warning!

It is also common to use (a, b) to denote the vector.

A vector has a direction, which is one of the two directions of the line it determines, and a magnitude, defined as follows.

Definition 3.7.23: magnitude of vector

The *magnitude* of a vector $OP =< a, b >$ is defined as the distance from O to its tip P, denoted by

$$\boxed{|| < a, b > || = \sqrt{a^2 + b^2}}$$

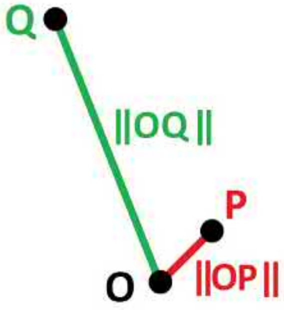
Exercise 3.7.24

Apply the definition to a vector $< x, 0 >$ and explain its relation to the one-dimensional case.

Exercise 3.7.25

Finish the sentence: "If a vector A represents the velocity, then $||A||$ represents _____ ."

The notation will help us with our formula for the angle:



It is rewritten as follows:

$$\cos\theta = \cos(\alpha-\beta) = \frac{xx' + yy'}{|| < x,y > || \cdot || < x',y' > ||}\,.$$

We now would like to make sense of the numerator of this fraction. The following will become commonly used.

Definition 3.7.26: dot product

The *dot product* of two vectors $< a,b >$ and $< c,d >$ is defined, as denoted, to be the following number:

$$\boxed{< a,b > \cdot < c,d >= ac + bd}$$

Warning!

Often, the meaning of the dot "·" has to be determined from the context.

Example 3.7.27

Let's link the dot product back to the magnitude. by the formula:

$$< a,b > \cdot < a,b >= a\cdot a + b\cdot b = a^2 + b^2 = || < a,b > ||^2\,.$$

So, the dot products contain all the information about the angles *and* about the magnitudes.

The numerator of the formula for the angle $\cos(\alpha-\beta)$ takes a simpler form now:

$$< x,y > \cdot < x',y' > \ .$$

We conclude the following:

Theorem 3.7.28: Directions for Dimension 2

The angle θ between the vectors OP and OQ, where $P = (x,y) \neq O$ and $Q = (x',y') \neq O$, is determined by the following formula:

$$\boxed{\cos\theta = \cos\widehat{QOP} = \frac{< x,y > \cdot < x',y' >}{|| < x,y > || \, || < x',y' > ||}}$$

So, the cosine of the angle can be now computed by using only addition, multiplication, and division of the four coordinates involved!

We can make the formula more compact by taking advantage of the new notation. We give either vector a single letter:

$$U = OP,\ V = OQ\,.$$

Then the angle θ between U and V satisfies:

$$\boxed{\cos\theta = \frac{U\cdot V}{||U||\,||V||}}$$

Exercise 3.7.29

(a) If the vectors represent the displacements of two objects in motion, what is the unit of this expression? (b) If the vectors represent the velocities of two objects in motion, what is the unit of this expression?

Exercise 3.7.30

What will happen to this expression if we double one of the vectors?

Exercise 3.7.31

Explain the difference between "the angle between two vectors" and "the angle between two lines".

Example 3.7.32

What is the angle between the lines $y = -2x + 3$ and $y = x - 1$? Since $+3$ and -1 are just shifts, this angle is the same as between $y = -2x$ and $y = x$. This way any point on either line will give us a vector we need. Just choose $x = 1$. Then we got $y = -2$ for the first line and $y = 1$ for the second. The two points are:

$$P = (1, -2),\ Q = (1, 1)\,.$$

We substitute these four numbers into the formula:

$$\cos\widehat{QOP} = \frac{<1,-2> \cdot <1,1>}{||<1,-2>||\,||<1,1>||} = \frac{1-2}{\sqrt{5}\sqrt{2}} = \frac{-1}{\sqrt{10}}\,.$$

Therefore,

$$\widehat{QOP} = \arccos\left(-\frac{1}{\sqrt{10}}\right)\,.$$

Exercise 3.7.33

What is the angle between the lines $x - y = 3$ and $2x + 3y = -1$? Suggest another pair of lines and repeat.

Exercise 3.7.34

Find a line that makes a 30-degree angle with the line $y = -2x + 3$? Suggest another angle and another line and repeat.

Exercise 3.7.35

Derive a formula for the sine of this angle.

Example 3.7.36: angle with itself

When two vectors are equal to each other, we have from the theorem:

$$\cos\widehat{POP} = \frac{<x,y> \cdot <x,y>}{||<x,y>||\,||<x,y>||} = \frac{xx+yy}{\sqrt{x^2+y^2}\,\sqrt{x^2+y^2}} = \frac{x^2+y^2}{x^2+y^2} = 1\,.$$

Therefore, $\widehat{POP} = 0$, as expected.

Example 3.7.37: angle within x-axis

When the terminal points of both vectors lie on the x-axis, i.e., $y = y' = 0$, the formula turns into the following:

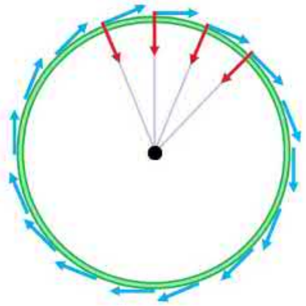
$$\cos \widehat{QOP} = \frac{xx'}{|x|\,|x'|} = \frac{x}{|x|}\frac{x'}{|x'|} = \operatorname{sign}(x)\cdot\operatorname{sign}(x')\,.$$

There are only two possibilities here, 1 or -1, and, therefore, $\widehat{QOP}$ can only be either 0 or 180 degrees.

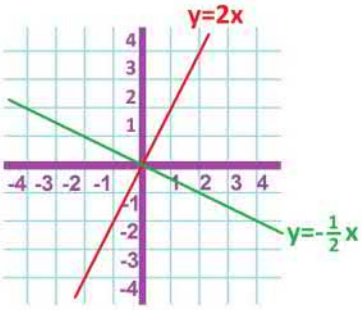
Exercise 3.7.38

Show how this fact demonstrates the theorem about the directions in dimension 1.

A case special importance is: When are two vectors or two lines *perpendicular*? For example, the velocity of the Moon orbiting the Earth is perpendicular to the force of gravity:



An example of such two lines, $y = 2x$ and $y = -\frac{1}{2}x$, suggests that the slopes will have to be *negative reciprocals* of each other:



Let's prove this fact using the theorem. For any point (x, y) on the first line and any point (x', y') on the second (other than O), we have:

$$0 = \cos\pi/2 = \frac{< x, y > \cdot < x', y' >}{|| < x, y > ||\,|| < x', y' > ||}\,.$$

Therefore,

$$< x, y > \cdot < x', y' > = 0 \iff xx' + yy' = 0 \iff xx' = -yy' \iff \frac{y}{x}\cdot\frac{y'}{x'} = -1\,.$$

But these two expressions are the slopes of the lines!

Theorem 3.7.39: Slopes of Perpendicular Lines

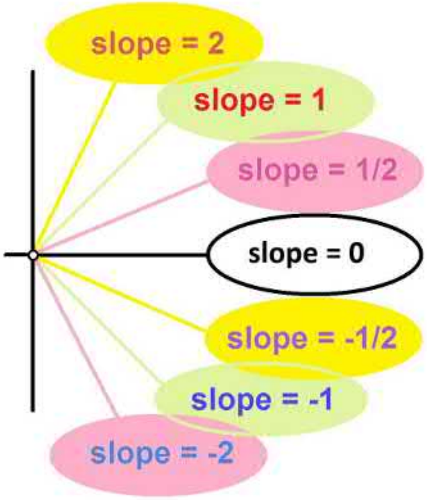
Two lines with slopes m and m' are perpendicular if and only if their slopes are negative reciprocals of each other; i.e.,

$$mm' = -1\,.$$

Exercise 3.7.40

Split the theorem into a statement and its converse.

Since any vertical line is perpendicular to any horizontal line and vice versa, we have solved the problem of perpendicularity. This is the summary:



Exercise 3.7.41

Find the line perpendicular to $y = -3x$ that passes through the point $(1, 1)$. Suggest another line and repeat.

The theory of vectors is further developed in Chapter 4.

3.8. *How complex numbers emerge*

The equation

$$x^2 + 1 = 0$$

has no solutions. Indeed, we observe the following:

$$x^2 \geq 0 \implies x^2 + 1 > 0 \implies x^2 + 1 \neq 0\,.$$

If we try to solve it the usual way, we get these:

$$x = \sqrt{-1} \quad \text{and} \quad x = -\sqrt{-1}\,.$$

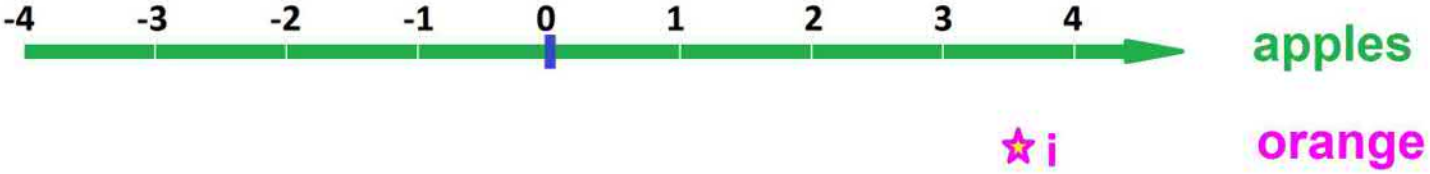
There are no such *real numbers.*

However, let's ignore this fact for a moment. Let's substitute what we have back into the equation and – blindly – follow the rules of algebra. We "confirm" that this "number" is a "solution":

$$x^2 + 1 = (\sqrt{-1})^2 + 1 = (-1) + 1 = 0\,.$$

We call this entity the *imaginary unit*, denoted by i.

We just add this "number" to the set of numbers we do algebra with:



And see what happens...

Making i a part of algebra will only require this three-part convention:

1. i is not a real number (and, in particular, $i \neq 0$), but
2. i can participate in the (four) algebraic operations with real numbers by following the same rules; also
3. $i^2 = -1$.

What algebraic rule are those? A few very basic ones:

$$x + y = y + x,\ x \cdot y = y \cdot x,\ x(y + z) = xy + xz,\ \text{etc.}$$

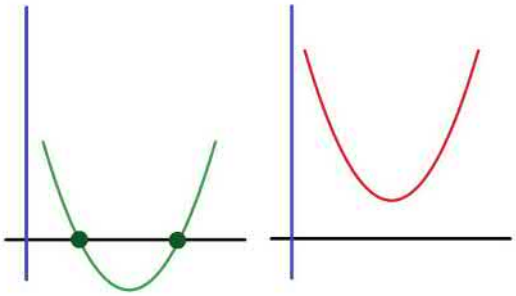
We allow one or several of these parameters to be i. For example, we have:

$$i + y = y + i,\ i \cdot y = y \cdot i,\ i(y + z) = iy + iz,\ \text{etc.}$$

What makes this extra effort worthwhile is a new look at quadratic polynomials. For example, this is how we may factor one:

$$x^2 - 1 = (x - 1)(x + 1)\,.$$

Then $x = 1$ and $x = -1$ are the x-intercepts of the polynomial:



But some polynomials, called *irreducible*, cannot be factored; there are no a, b such that:

$$x^2 + 1 = (x - a)(x - b)\,.$$

There are no *real* a, b, that is! Using our rules, we discover:

$$(x - i)(x + i) = x^2 - ix + ix - i^2 = x^2 + 1\,.$$

Of course, the number i is *not* an x-intercept of $f(x) = x^2 + 1$ as the x-axis ("the real line") consists of only (and all) real numbers.

So, multiples of i appear immediately as we start doing algebra with it.

Definition 3.8.1: imaginary numbers

The real multiples of the imaginary unit, i.e.,

$$z = ri,\ r\ \text{real},$$

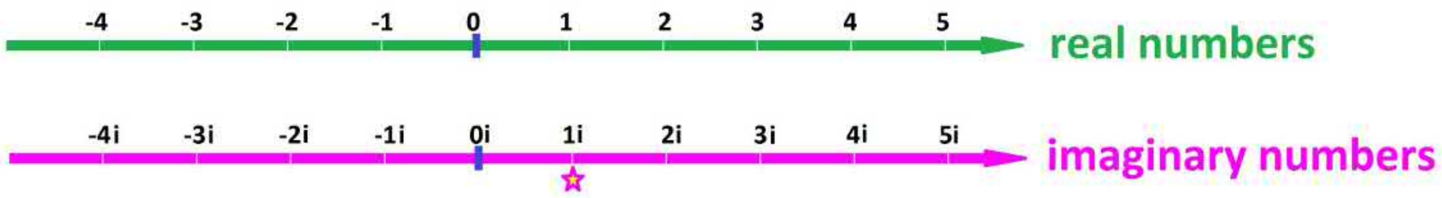
are called *imaginary numbers.*

We have created a whole class of *non-real* numbers! Of course, ri, where r is real, can't real:

$$(ri)^2 = r^2 i^2 = -r^2 < 0\,.$$

The only exception is $0i = 0$; it's real!

There are as many of them as the real numbers:



Example 3.8.2: quadratic equations

The imaginary numbers may also come from solving the simplest quadratic equations. For example, the equation

$$x^2+4=0$$

gives us via our substitution:

$$x=\pm\sqrt{-4}=\pm\sqrt{4(-1)}=\pm\sqrt{4}\sqrt{-1}=\pm 2i\,.$$

Indeed, if we substitute $x=2i$ into the equation, we have:

$$(2i)^2+4=(2)^2(i)^2+4=4(-1)+4=0\,.$$

More general quadratic equation are discussed in the next section.

Imaginary numbers obey the laws of algebra as we know them! If we need to simplify the expression, we try to manipulate it in such a way that real numbers are combines with real while i is pushed aside.

For example, we can just factor i out of all addition and subtraction:

$$5i+3i=(5+3)i=8i\,.$$

It looks exactly like middle school algebra:

$$5x+3x=(5+3)x=8x\,.$$

After all, x *could* be i. Another similarity is with the algebra of quantities that have units:

$$5\text{ in.}+3\text{ in.}=(5+3)\text{ in.}=8\text{ in.}\,.$$

So, the nature of the unit doesn't matter (if we can push it aside). Even simpler:

$$5\text{ apples }+3\text{ apples }=(5+3)\text{ apples }=8\text{ apples }\,.$$

It's "8 apples" not "8"! And so on.

This is how we multiply an imaginary number by a real number:

$$2\cdot(3i)=(2\cdot 3)i=6i\,.$$

We have a new imaginary number.

How do we multiply two imaginary numbers? It's different; after all, we don't usually multiply apples by apples! In contrast to the above, even though multiplication and division follow the same rule as always, we can, when necessary, and often have to, simplify the outcome of our algebra using our *fundamental identity*:

$$i^2=-1\,.$$

For example:

$$(5i)\cdot(3i)=(5\cdot 3)(i\cdot i)=15i^2=15(-1)=-15\,.$$

It's real!

We also simplify the outcome using the other *fundamental fact* about the imaginary unit:

$$i\neq 0\,.$$

We can divide by i! For example,

$$\frac{5i}{3i}=\frac{5}{3}\frac{i}{i}=\frac{5}{3}\cdot 1=\frac{5}{3}\,.$$

As you can see, doing algebra with imaginary numbers will often bring us back to real numbers. These two classes of numbers cannot be separated from each other!

They aren't. Let's take another look at quadratic equations. The equation

$$ax^2 + bx + c = 0,\ a \neq 0\,,$$

is solved with the familiar *Quadratic Formula*:

$$x = \frac{-b \pm \sqrt{b^2 - 4ac}}{2a}\,.$$

Let's consider

$$x^2 + 2x + 10 = 0\,.$$

Then the roots are supposed to be:

$$\begin{array}{lll} x & = \dfrac{-2 \pm \sqrt{2^2 - 4 \cdot 10}}{2} & \\ & = \dfrac{-2 \pm \sqrt{-36}}{2} & \\ & = -1 \pm \sqrt{-9} & \text{There is no real solution!} \\ & = -1 \pm \sqrt{9}\sqrt{-1} & \text{But we go on.} \\ & = -1 \pm 3i\,. & \end{array}$$

We end up adding real and imaginary numbers!

As there is no way to simplify this, we conclude the following:

- ▶ A number $a + bi$, where $a, b \neq 0$ are real, is neither real no imaginary.

Exercise 3.8.3

Explain why.

This addition is not literal. It's like "adding" apples to oranges:

$$5 \text{ apples } + 3 \text{ oranges } = \ldots$$

It's not 8 and it's not 8 fruit because we wouldn't be able to read this equality backwards. The algebra will, however, be meaningful:

$$(5a + 3o) + (2a + 4o) = (5 + 3)a + (3 + 4)o = 8a + 7o\,.$$

It is as if we collect *similar terms*, like this:

$$(5 + 3x) + (2 + 4x) = (5 + 2) + (3 + 4)x = 8 + 7x\,.$$

This idea enables us to do this:

$$(5 + 3i) + (2 + 4i) = (5 + 3) + (3 + 4)i = 8 + 7i\,.$$

The number we are facing consist of both real numbers and imaginary numbers. This fact makes them "complex"...

Definition 3.8.4: complex number

Any sum of real and imaginary numbers is called a *complex number*. The set of all complex numbers is denoted as follows:

$$\boxed{\mathbf{C}}$$

Warning!

All real numbers are complex.

Addition and subtraction are easy; we just combine *similar terms* just like in middle school. For example,

$$(1+5i)+(3-i)=1+5i+3-i=(1+3)+(5i-i)=4+4i\,.$$

To simplify *multiplication* of complex numbers, we expand and then use $i^2=-1$, as follows:

$$\begin{aligned}(1+5i)\cdot(3-i)&=1\cdot 3+5i\cdot 3+1\cdot(-i)+5i\cdot(-i)\\&=3+15i-i-5i^2\\&=(3+5)+(15i-i)\\&=8+14i\,.\end{aligned}$$

It's a bit trickier with *division*:

$$\begin{aligned}\frac{1+5i}{3-i}&=\frac{1+5i}{3-i}\frac{3+i}{3+i}\\&=\frac{(1+5i)(3+i)}{(3-i)(3+i)}\\&=\frac{-2+8i}{3^2-i^2}\\&=\frac{-2+8i}{3^2+1}\\&=\frac{1}{10}(-2+8i)\\&=-.2+.8i\,.\end{aligned}$$

The simplification of the denominator is made possible by the trick of multiplying by $3+i$. It is the same trick we used in Volume 1 to simplify fractions with roots to compute their limits:

$$\frac{1}{1-\sqrt{x}}=\frac{1}{1-\sqrt{x}}\frac{1+\sqrt{x}}{1+\sqrt{x}}=\frac{1+\sqrt{x}}{1-x}\,.$$

Definition 3.8.5: complex conjugate

The *complex conjugate* of $z=a+bi$ is defined and denoted by:

$$\bar{z}=\overline{a+bi}=a-bi\,.$$

The following is crucial.

Theorem 3.8.6: Algebra of Complex Numbers

The rules of the algebra of complex numbers are identical to those of real numbers:

- ***Commutativity of addition:*** $z+u=u+z\,.$
- ***Associativity of addition:*** $(z+u)+v=z+(u+v)\,.$
- ***Commutativity of multiplication:*** $z\cdot u=u\cdot z\,.$

- ***Associativity of multiplication:*** $(z \cdot u) \cdot v = z \cdot (u \cdot v)\,.$
- ***Distributivity:*** $z \cdot (u + v) = z \cdot u + z \cdot v\,.$

This is the *complex number system*; it follows the rules of the real number system but also contains it. This theorem will allow us to build calculus for complex functions that is almost identical to that for real functions and also contains it.

Definition 3.8.7: standard form of complex number

Every complex number x has the *standard representation*:

$$z = a + bi\,,$$

where a and b are two real numbers. The two components are named as follows:

- a is the *real part* of z, with notation:

$$a = \text{Re}(z)\,,$$

- bi is the *imaginary part* of z, with notation:

$$b = \text{Im}(z)\,.$$

Then, the purpose of the computations above were to find the standard form of a complex number that comes from algebraic operations with other complex numbers. They were literally simplifications.

The definition makes sense because of the following result:

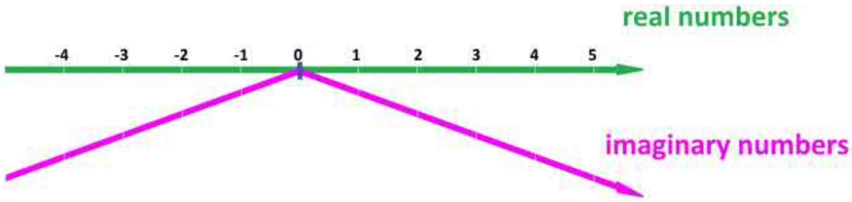
Theorem 3.8.8: Standard Form of Complex Number

Two complex numbers are equal if and only if both their real and imaginary parts are equal.

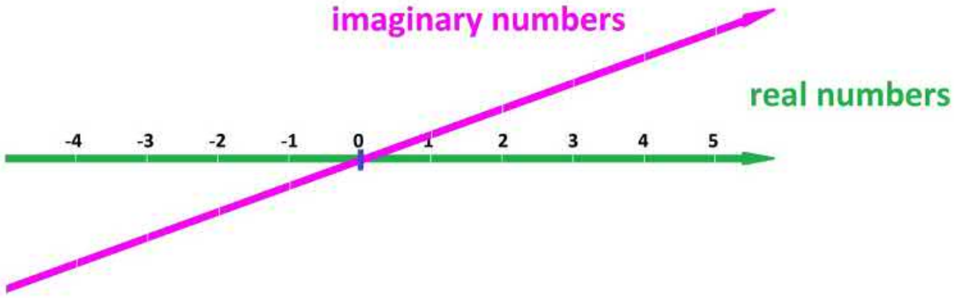
So, we have:

$$z = \text{Re}(z) + \text{Im}(z)i\,.$$

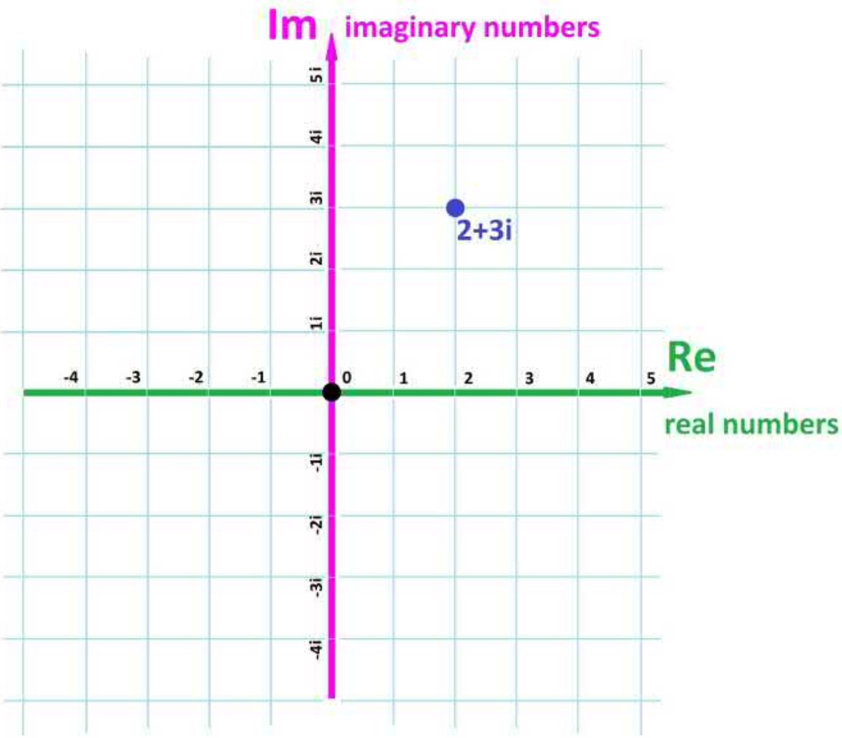
In order to see the *geometric representation of complex numbers* we need to combine the real number line and the imaginary number line. How? We realize that that have nothing in common... except $0 = 0i$ belongs to both:



We can try to combine them like that, or like this:



Or we can try to combine them in the same manner we built the xy-plane:



This representation helps us understand the main idea:

- ▶ Complex numbers are linear combinations of the real unit, 1, and the imaginary unit, i.

If $z = a + bi$, then a and b are thought of as the components of vector z in the plane. We have a one-to-one correspondence:

$$\mathbf{C} \longleftrightarrow \mathbf{R}^2\,,$$

given by

$$a + bi \ \longleftrightarrow < a, b > \ .$$

Then the x-axis of this plane consists of the real numbers and the y-axis of the imaginary numbers.

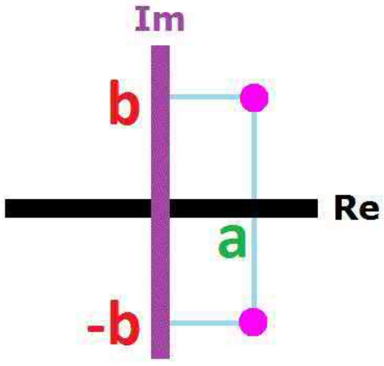
It is called the *complex plane.*

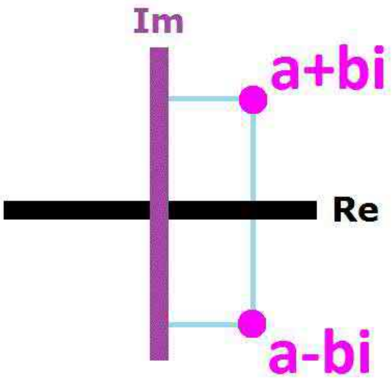
Warning!

This is just a visualization.

Then the complex conjugate of z is the complex number with the same real part as z and the imaginary part with the opposite sign:

$$\text{Re}(\bar{z}) = \text{Re}(z) \quad \text{and} \quad \text{Im}(\bar{z}) = -\,\text{Im}(z)\,.$$





Warning!

All numbers we have encountered so far are real non-complex, and so are all quantities one can encounter in day-to-day life or science: time, location, length, area, volume, mass, temperature, money, etc.

3.9. Classification of quadratic polynomials

The general quadratic equation with real coefficients,

$$ax^2 + bx + c = 0,\ \ a \neq 0\,,$$

can be simplified. Let's divide by a and study the resulting quadratic polynomial:

$$f(x) = x^2 + px + q\,,$$

where $p = b/a$ and $q = c/a$.

The *Quadratic Formula* then provides the x-intercepts of this function:

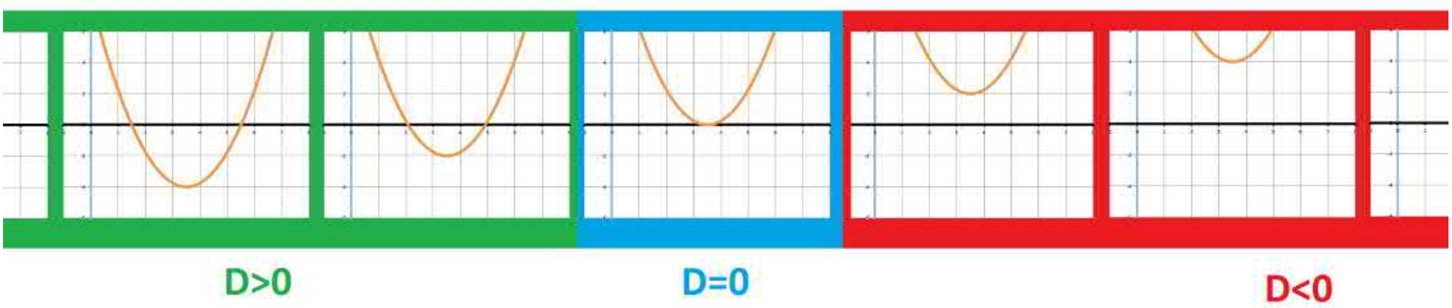
$$x = -\frac{p}{2} \pm \frac{\sqrt{p^2 - 4q}}{2}\,.$$

Of course, the x-intercepts are the real solutions of this equation and that is why the result only makes sense when the *discriminant* of the quadratic polynomial,

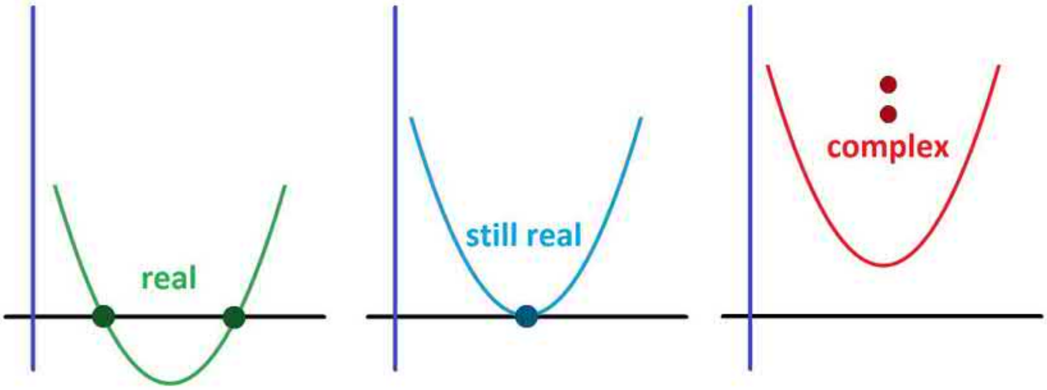
$$D = p^2 - 4q\,,$$

is non-negative.

Now, increasing the value of q (while keeping p constant) makes the graph of $y = f(x)$ shift upward and, eventually, pass the x-axis entirely. We can observe how its two x-intercepts start to get closer to each other, then merge, and finally disappear:



What is behind this event is the emergence of complex roots:



This process is explained by what is happening, with the growth of q, to the roots given by the *Quadratic Formula*:

$$x_{1,2} = -\frac{p}{2} \pm \frac{\sqrt{D}}{2}\,.$$

There are three states:

1. Starting with a positive value, D decreases, and $\frac{\sqrt{D}}{2}$ decreases.
2. Then D becomes 0 and, therefore, we have $\frac{\sqrt{D}}{2} = 0$.
3. Then D becomes negative, and $\frac{\sqrt{D}}{2}$ becomes imaginary (but $\frac{\sqrt{-D}}{2}$ is real).

Case 2 is a borderline between the other two!

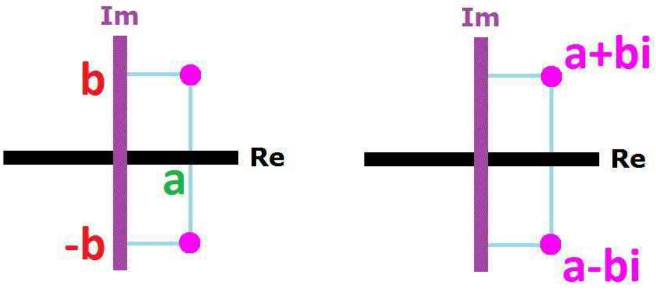
The roots are, respectively:

	discriminant	root #1	root #2
1.	$D > 0$	$x_1 = -\frac{p}{2} - \frac{\sqrt{D}}{2}$	$x_2 = -\frac{p}{2} + \frac{\sqrt{D}}{2}$
2.	$D = 0$	$x_1 = -\frac{p}{2}$	$x_2 = -\frac{p}{2}$
3.	$D < 0$	$x_1 = -\frac{p}{2} - \frac{\sqrt{-D}}{2}i$	$x_2 = -\frac{p}{2} + \frac{\sqrt{-D}}{2}i$

We make two observations:

- The two real roots ($D > 0$) are unrelated.
- The two complex roots ($D < 0$) are *conjugate* of each other.

The complex ones always come in pairs:



Exercise 3.9.1

Show that a pair of complex numbers that aren't conjugate can't be the roots of a quadratic polynomial.

Exercise 3.9.2

Show that any pair of real numbers can be the roots of a quadratic polynomial.

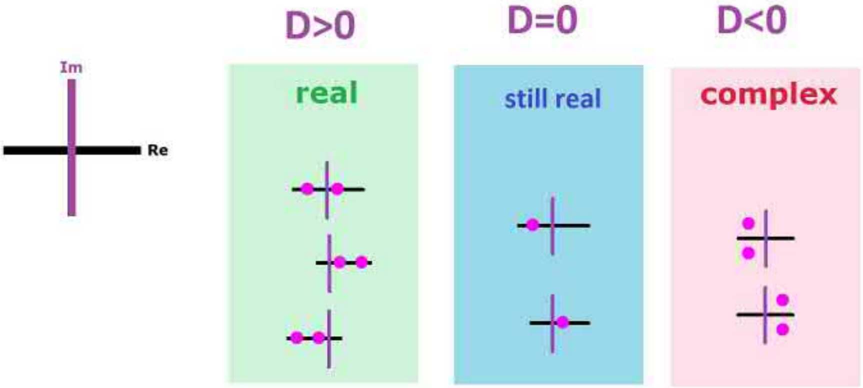
As a summary, we have the following classification the roots of quadratic polynomials in terms of *the sign of the discriminant.*

Theorem 3.9.3: Classification of Roots I

The two roots of a quadratic polynomial with real coefficients are:

1. *distinct real when its discriminant D is positive;*
2. *equal real when its discriminant D is zero;*
3. *complex conjugate of each other when its discriminant D is negative.*

In the future study of differential equations, we will need a more precise way to classify the polynomials: according to *the signs of the real parts of their roots.* The signs will determine increasing and decreasing behavior of certain solutions. Once again, these are the possibilities:



Theorem 3.9.4: Classification of Roots II

Suppose x_1 and x_2 are the two roots of a quadratic polynomial $f(x) = x^2+px+q$ with real coefficients. Then the signs of the real parts $\operatorname{Re}(x_1)$ *and* $\operatorname{Re}(x_2)$ *of x_1 and x_2 are:*

1. *same when $p^2 > 4q$ and $q \geq 0$;*
2. *opposite when $p^2 > 4q$ and $q < 0$;*
3. *same and opposite of that of p when $p^2 \leq 4q$.*

Proof.

The condition $p^2 \leq 4q$ is equivalent to $D \leq 0$. We can see in the table above that, in that case, we have $\operatorname{Re}(x_1) = \operatorname{Re}(x_2) = -\dfrac{p}{2}$. We are left with the case $D > 0$ and real roots. The case of equal signs is separated from the case of opposite signs of x_1 and x_2 by the case when both are equal to zero: $x_1 = x_2 = 0$. We solve:

$$-\frac{p}{2} - \frac{\sqrt{D}}{2} = 0 \implies p = -\sqrt{p^2 - 4q} \implies p^2 = p^2 - 4q \implies q = 0\,.$$

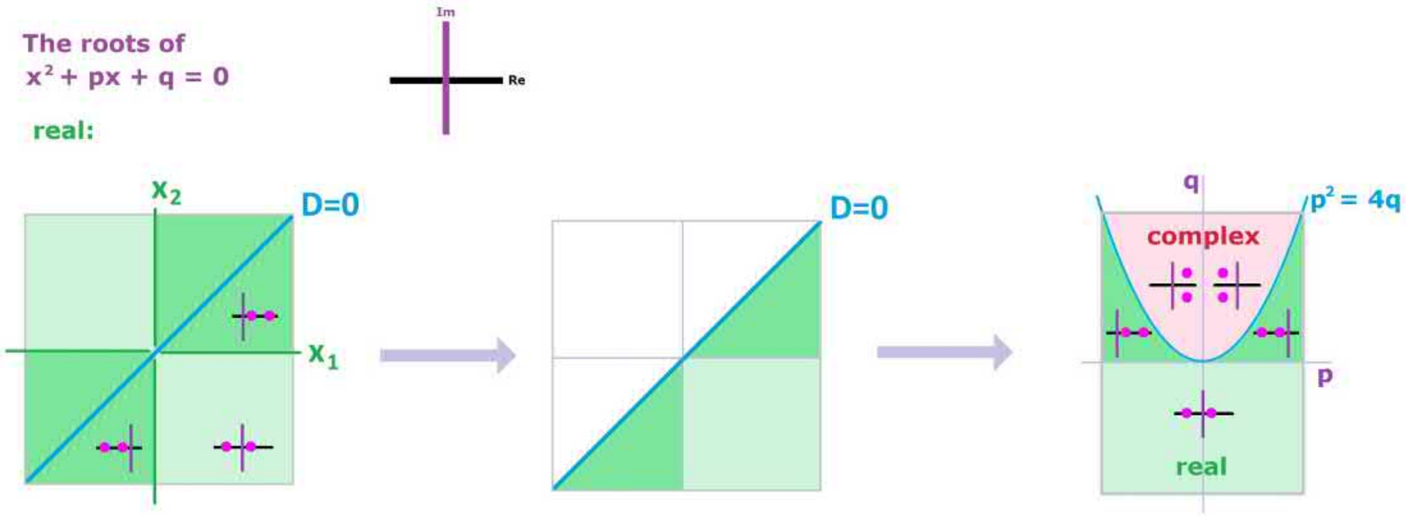
Exercise 3.9.5

Finish the proof.

Let's visualize our conclusion. We would like to show the main scenarios of what kinds of roots the polynomial might have depending on the values of its two coefficients, p and q.

First, how do we visualize *pairs of numbers*? As points on a coordinate plane of course... but only when they are *real.*

Suppose for now that they are. We start with a plane, the x_1x_2-plane to be exact, as a representation of all possible pairs of real roots (left). Then the diagonal of this plane represents the case of *equal* (and still real) roots, $x_1 = x_2$, i.e., $D = 0$. Since the order of the roots doesn't matter – (x_1, x_2) is as good as (x_2, x_1) – we need only half of the plane. We fold the plane along the diagonal (middle).



The diagonal – represented by the equation $D = 0$ – exposed this way can now serve its purpose of separating the case of real and *complex* roots. Now, let's go to the pq-plane. Here, the parabola $p^2 = 4q$ also represents the equation $D = 0$. Let's bring them together! We take our half-plane and bend its diagonal edge into the parabola $p^2 = 4q$ (right).

Note that the average of this two "solutions" will continue to provide the vertex of the parabola even though there are no x-intercepts!

Theorem 3.9.6: Vieta's Formulas

The roots x_1, x_2 *of the quadratic polynomial* $f(x) = x^2 + px + q$ *satisfy the following equations:*

$$\boxed{x_1 + x_2 = -p \quad \text{and} \quad x_1 \cdot x_2 = q}$$

Classifying polynomials this way allows one to classify matrices and understand what each of them does as a transformation of the plane, which in turn will help us understand systems of ODEs.

3.10. *The complex plane* $\mathbf{C}$ *is the Euclidean space* $\mathbf{R}^2$

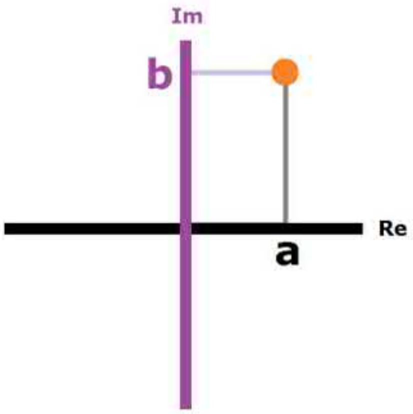
If we call complex number *numbers*, they must be subject to some *algebraic operations*.

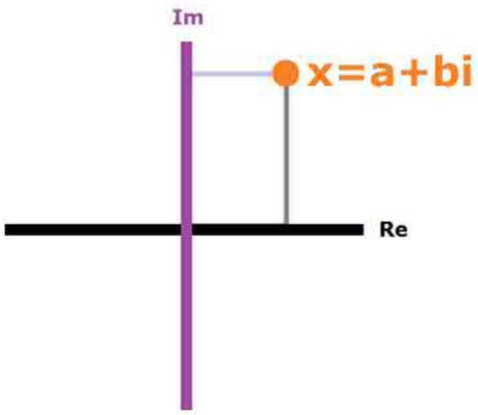
We will initially look at them through the lens of *vector algebra* of the plane $\mathbf{R}^2$.

A complex number z has the *standard representation*:

$$z = a + bi\,,$$

where a and b are two real numbers. These two can be seen in the *geometric representation* of complex numbers:





Therefore, a and b are thought of as the coordinates of z as a *point* on the plane. But any complex number is not only a point on the complex plane but also a *vector*. We have a correspondence:

$$\mathbf{C} \longleftrightarrow \mathbf{R}^2\,,$$

given by

$$\boxed{a + bi \ \longleftrightarrow\ < a, b >}$$

There is more to this than just a match; the algebra of vectors in $\mathbf{R}^2$ applies!

Warning!

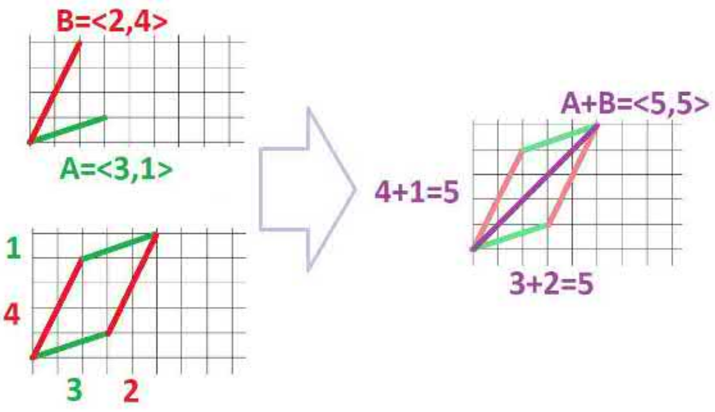
In spite of this fundamental correspondence, we will continue to think of complex numbers as *numbers* (and use the lower case letters).

Let's see how this algebra of numbers works in parallel with the algebra of 2-vectors.

First, the addition of complex numbers is done *component-wise*:

$$\begin{array}{ccccccc}(a+bi) & + & (c+di) & = & (a+c) & + & (b+d)i\,, \\ < a,b > & + & < c,d > & = & < a+c & , & b+d > \ .\end{array}$$

It corresponds to addition of vectors:



Second, we can easily multiply complex numbers by real ones:

$$\begin{array}{ccccccc}(a+bi) & c & = & (ac) & + & (bc)i\,, \\ < a,b > & c & = & < ac & , & bc > \ .\end{array}$$

It corresponds to scalar multiplication of vectors.

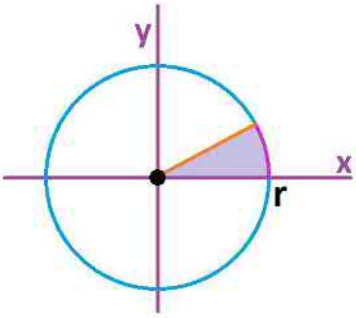
Warning!

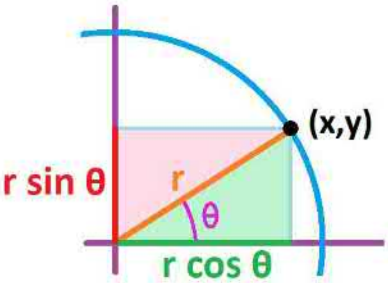
Vector algebra of $\mathbf{R}^2$ is complex algebra, but not vice versa. Complex *multiplication* is what makes it different.

Example 3.10.1: circle

We can easily represent circles on the complex plane:

$$z = r\cos\theta + r\sin\theta \cdot i\,.$$





Our study of calculus of complex numbers starts with the study of the *topology* of the complex plane. This topology is the same as that of *the Euclidean plane* $\mathbf{R}^2$!

Just as before, every function $z = f(t)$ with an appropriate domain creates a sequence:

$$z_k = f(k)\,.$$

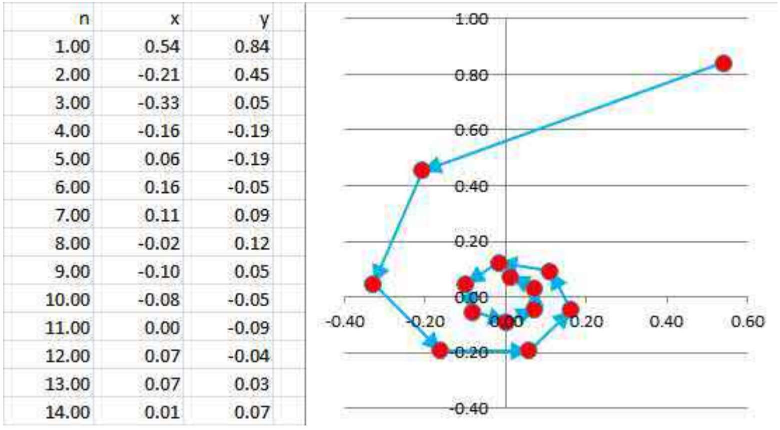
A function with complex values defined on a ray in the set of integers, $\{p, p+1, ...\}$, is called an *infinite sequence*, or simply sequence.

Example 3.10.2: spiral

A good example is that of the sequence made of the reciprocals:

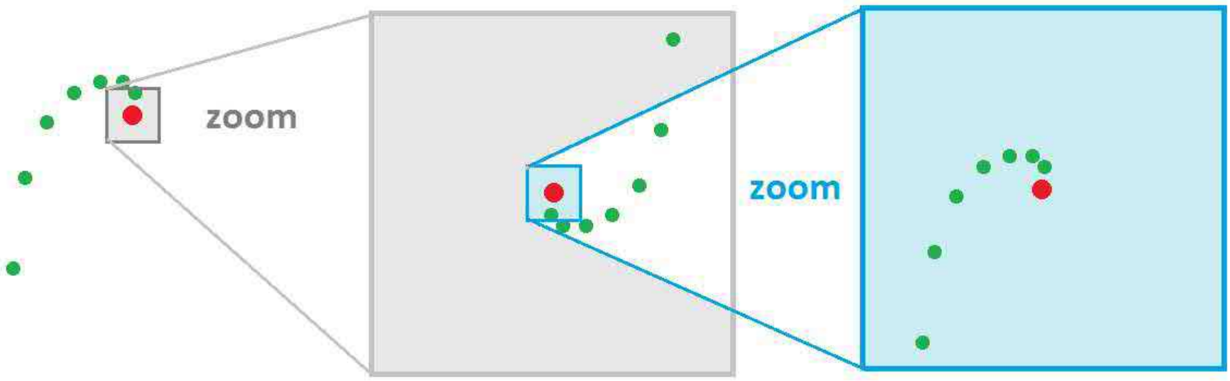
$$z_k = \frac{\cos k}{k} + \frac{\sin k}{k} i\,.$$

It *tends to* 0 while spiraling around it.

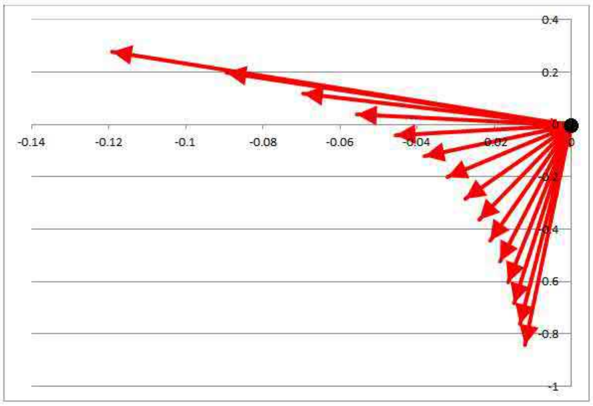


n	x	y
1.00	0.54	0.84
2.00	-0.21	0.45
3.00	-0.33	0.05
4.00	-0.16	-0.19
5.00	0.06	-0.19
6.00	0.16	-0.05
7.00	0.11	0.09
8.00	-0.02	0.12
9.00	-0.10	0.05
10.00	-0.08	-0.05
11.00	0.00	-0.09
12.00	0.07	-0.04
13.00	0.07	0.03
14.00	0.01	0.07

The starting point of calculus of complex numbers is the following. The convergence of a *sequence of complex numbers* is the convergence of its real and imaginary parts or, which is equivalent, the convergence of points (or vectors) on the complex plane seen as any plane: the distance from the kth point to the limit is getting smaller and smaller.



We use the convergence for vectors on the plane simply replacing vectors with complex numbers and "magnitude" with "modulus".



Definition 3.10.3: convergent sequence

Suppose $\{z_k :\ k = 1, 2, 3...\}$ is a sequence of complex numbers, i.e., points in **C**. We say that the sequence *converges* to another complex number z, i.e., a point in **C**, called the *limit* of the sequence, if:

$$||z_k - z|| \to 0 \text{ as } k \to \infty\,,$$

denoted by:

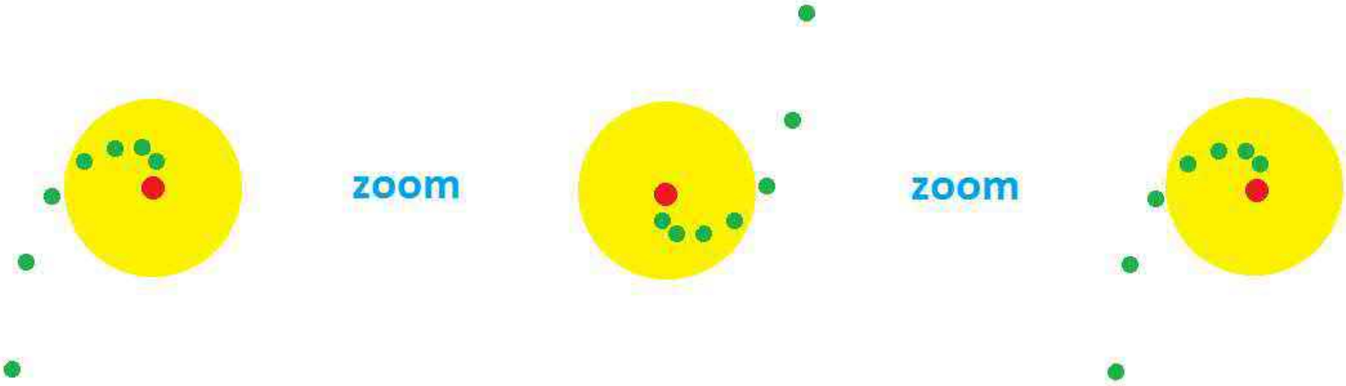
$$z_k \to z \text{ as } k \to \infty\,,$$

or

$$z = \lim_{k\to\infty} z_k\,.$$

If a sequence has a limit, we call the sequence *convergent* and say that it *converges*; otherwise it is *divergent* and we say it *diverges*.

In other words, the points start to accumulate in smaller and smaller circles around z. A way to visualize a trend in a convergent sequence is to enclose the tail of the sequence in a *disk*:



Theorem 3.10.4: Uniqueness of Limit

A sequence can have only one limit (finite or infinite); i.e., if a and b are limits of the same sequence, then $a = b$.

Definition 3.10.5: sequence tends to infinity

We say that a sequence z_k *tends to infinity* if the following condition holds: for each real number R, there exists such a natural number N that, for every natural number $k > N$, we have

$$||z_k|| > R;$$

we use the following notation:

$$z_k \to \infty \text{ as } k \to \infty\,.$$

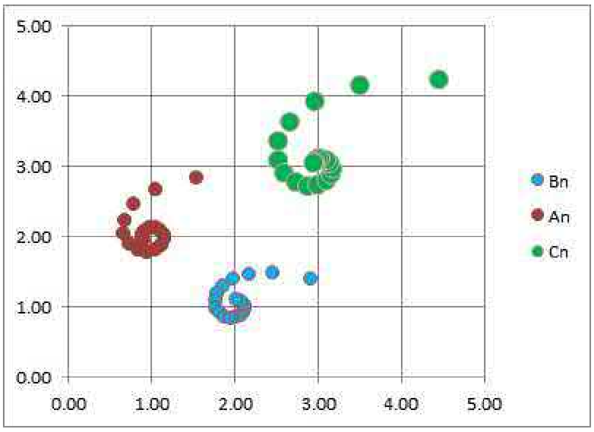
The following is another analog of a familiar theorem about the topology of the plane.

Theorem 3.10.6: Component-wise Convergence of Sequences

A sequence of complex numbers z_k in $\mathbf{C}$ converges to a complex number z if and only if both the real and the imaginary parts of z_k converge to the real and the imaginary parts of z respectively; i.e.,

$$z_k \to z \iff \operatorname{Re}(z_k) \to \operatorname{Re}(z) \quad \textit{and} \quad \operatorname{Im}(z_k) \to \operatorname{Im}(z)\,.$$

The algebraic properties of limits of sequences of complex numbers also look familiar...



Theorem 3.10.7: Sum Rule for Complex Sequences

If sequences z_k, u_k converge then so does $z_k + u_k$, and

$$\lim_{k\to\infty} (z_k + u_k) = \lim_{k\to\infty} z_k + \lim_{k\to\infty} u_k\,.$$

Theorem 3.10.8: Constant Multiple Rule for Complex Sequences

If sequence z_k *converges then so does* cz_k *for any complex* c*, and*

$$\lim_{k\to\infty} c\, z_k = c \cdot \lim_{k\to\infty} z_k\,.$$

Wouldn't calculus of complex numbers be just a copy of calculus on the plane? No, not with the possibility of *multiplication* taken into account.

Chapter 4: Multidimensional spaces

Contents

4.1. Multiple variables, multiple dimensions

Why do we need to study multidimensional spaces?

These are the main sources of spaces of *multiple dimensions*:

1. The physical space, dimension 3.
2. Multiple spaces of single dimension interconnected via functional relations: The graphs of these functions lie in higher-dimensional spaces.
3. Multiple quantities, homogeneous (such as stock and commodity prices) and non-homogeneous (such as other data): They are combined into points in abstract spaces.

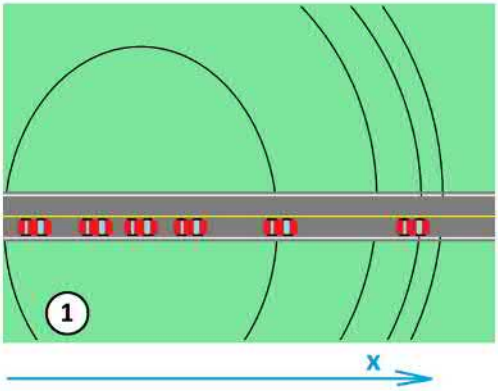
The 3-dimensional space represents a significant challenge in comparison to the plane. Furthermore, taking into account time will make it 4-dimensional.

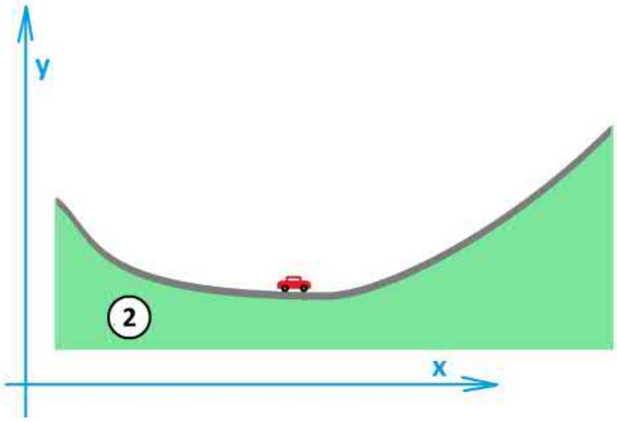
Furthermore, planning a flight of a *plane* would require 3 spatial variables, but the number increases to 6 if we are to consider the orientation of the plane in the air: the roll, the pitch, and the yaw.

Next, let's notice that even when we deal with only numerical functions, the graph of such a function lies in the xy-plane, a space of dimension 2. What if there are two such functions?

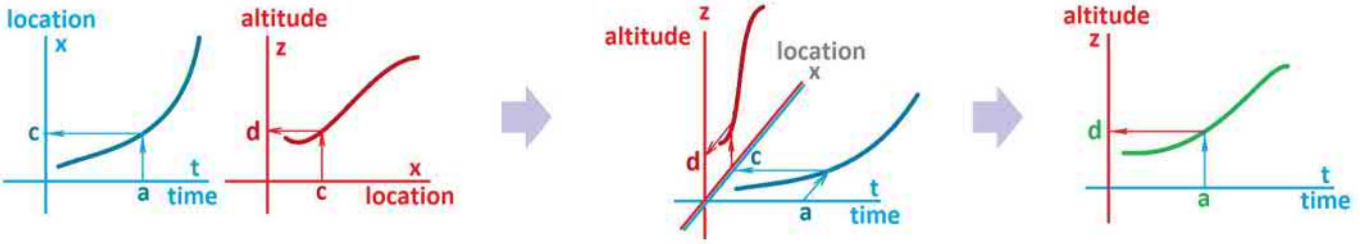
Example 4.1.1: road trip

Let's imagine a car driven through a mountain terrain. Its location and its speed, as seen on the map, are known. The grade of the road is also known. How fast is the car climbing?





The first variable is time, t. We also have two *spatial* variables: the horizontal location x and the elevation (the vertical location) z. Then z depends on x, and x depends on t. Therefore, z depends on t via the *composition*:



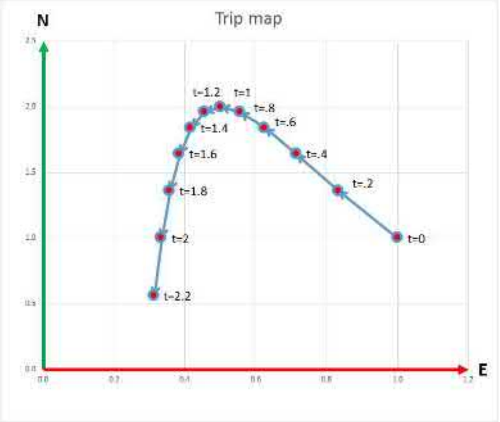
Plotting both functions together requires a 3-dimensional space.

We can take specific functions:

- The horizontal location is a *linear* function of time, $x = 2t - 1$.
- The elevation is a *linear* function of horizontal location, $z = 3x + 7$.
- Then elevation is, too, a *linear* function of time, $z = 3(2t - 1) + 7$.

Example 4.1.2: hiking

Let's now consider a more complex trip. Planning a hike, we create a *trip plan*: The times and the places are put on a simple map of the area:

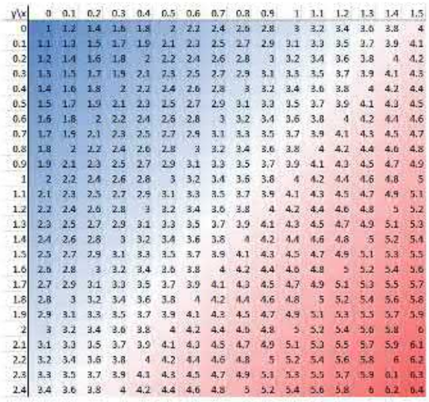


This is a *parametric curve*:

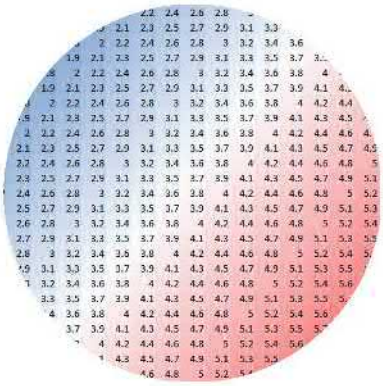
$$x = f(t),\ y = g(t)\,,$$

with x and y providing the coordinates of your location. Conversely, motion in time is a go-to metaphor for parametric curves!

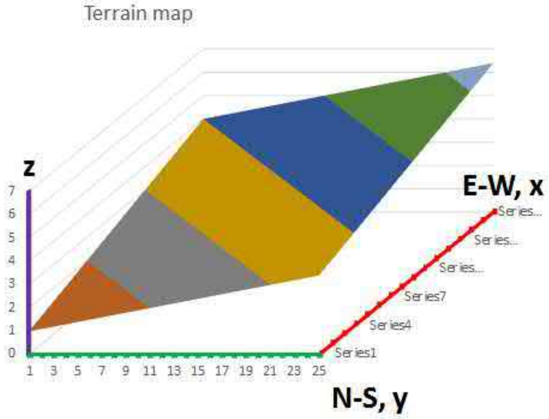
We then bring the *terrain map* of the area. The data is colored accordingly:



y\x	0	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1	1.1	1.2	1.3	1.4	1.5
0	1	1.2	1.4	1.6	1.8	2	2.2	2.4	2.6	2.8	3	3.2	3.4	3.6	3.8	4
0.1	1.1	1.3	1.5	1.7	1.9	2.1	2.3	2.5	2.7	2.9	3.1	3.3	3.5	3.7	3.9	4.1
0.2	1.2	1.4	1.6	1.8	2	2.2	2.4	2.6	2.8	3	3.2	3.4	3.6	3.8	4	4.2
0.3	1.3	1.5	1.7	1.9	2.1	2.3	2.5	2.7	2.9	3.1	3.3	3.5	3.7	3.9	4.1	4.3
0.4	1.4	1.6	1.8	2	2.2	2.4	2.6	2.8	3	3.2	3.4	3.6	3.8	4	4.2	4.4
0.5	1.5	1.7	1.9	2.1	2.3	2.5	2.7	2.9	3.1	3.3	3.5	3.7	3.9	4.1	4.3	4.5
0.6	1.6	1.8	2	2.2	2.4	2.6	2.8	3	3.2	3.4	3.6	3.8	4	4.2	4.4	4.6
0.7	1.7	1.9	2.1	2.3	2.5	2.7	2.9	3.1	3.3	3.5	3.7	3.9	4.1	4.3	4.5	4.7
0.8	1.8	2	2.2	2.4	2.6	2.8	3	3.2	3.4	3.6	3.8	4	4.2	4.4	4.6	4.8
0.9	1.9	2.1	2.3	2.5	2.7	2.9	3.1	3.3	3.5	3.7	3.9	4.1	4.3	4.5	4.7	4.9
1	2	2.2	2.4	2.6	2.8	3	3.2	3.4	3.6	3.8	4	4.2	4.4	4.6	4.8	5
1.1	2.1	2.3	2.5	2.7	2.9	3.1	3.3	3.5	3.7	3.9	4.1	4.3	4.5	4.7	4.9	5.1
1.2	2.2	2.4	2.6	2.8	3	3.2	3.4	3.6	3.8	4	4.2	4.4	4.6	4.8	5	5.2
1.3	2.3	2.5	2.7	2.9	3.1	3.3	3.5	3.7	3.9	4.1	4.3	4.5	4.7	4.9	5.1	5.3
1.4	2.4	2.6	2.8	3	3.2	3.4	3.6	3.8	4	4.2	4.4	4.6	4.8	5	5.2	5.4
1.5	2.5	2.7	2.9	3.1	3.3	3.5	3.7	3.9	4.1	4.3	4.5	4.7	4.9	5.1	5.3	5.5
1.6	2.6	2.8	3	3.2	3.4	3.6	3.8	4	4.2	4.4	4.6	4.8	5	5.2	5.4	5.6
1.7	2.7	2.9	3.1	3.3	3.5	3.7	3.9	4.1	4.3	4.5	4.7	4.9	5.1	5.3	5.5	5.7
1.8	2.8	3	3.2	3.4	3.6	3.8	4	4.2	4.4	4.6	4.8	5	5.2	5.4	5.6	5.8
1.9	2.9	3.1	3.3	3.5	3.7	3.9	4.1	4.3	4.5	4.7	4.9	5.1	5.3	5.5	5.7	5.9
2	3	3.2	3.4	3.6	3.8	4	4.2	4.4	4.6	4.8	5	5.2	5.4	5.6	5.8	6
2.1	3.1	3.3	3.5	3.7	3.9	4.1	4.3	4.5	4.7	4.9	5.1	5.3	5.5	5.7	5.9	6.1
2.2	3.2	3.4	3.6	3.8	4	4.2	4.4	4.6	4.8	5	5.2	5.4	5.6	5.8	6	6.2
2.3	3.3	3.5	3.7	3.9	4.1	4.3	4.5	4.7	4.9	5.1	5.3	5.5	5.7	5.9	6.1	6.3
2.4	3.4	3.6	3.8	4	4.2	4.4	4.6	4.8	5	5.2	5.4	5.6	5.8	6	6.2	6.4



Such a topographic map has the colors indicating the elevation of the actual terrain:

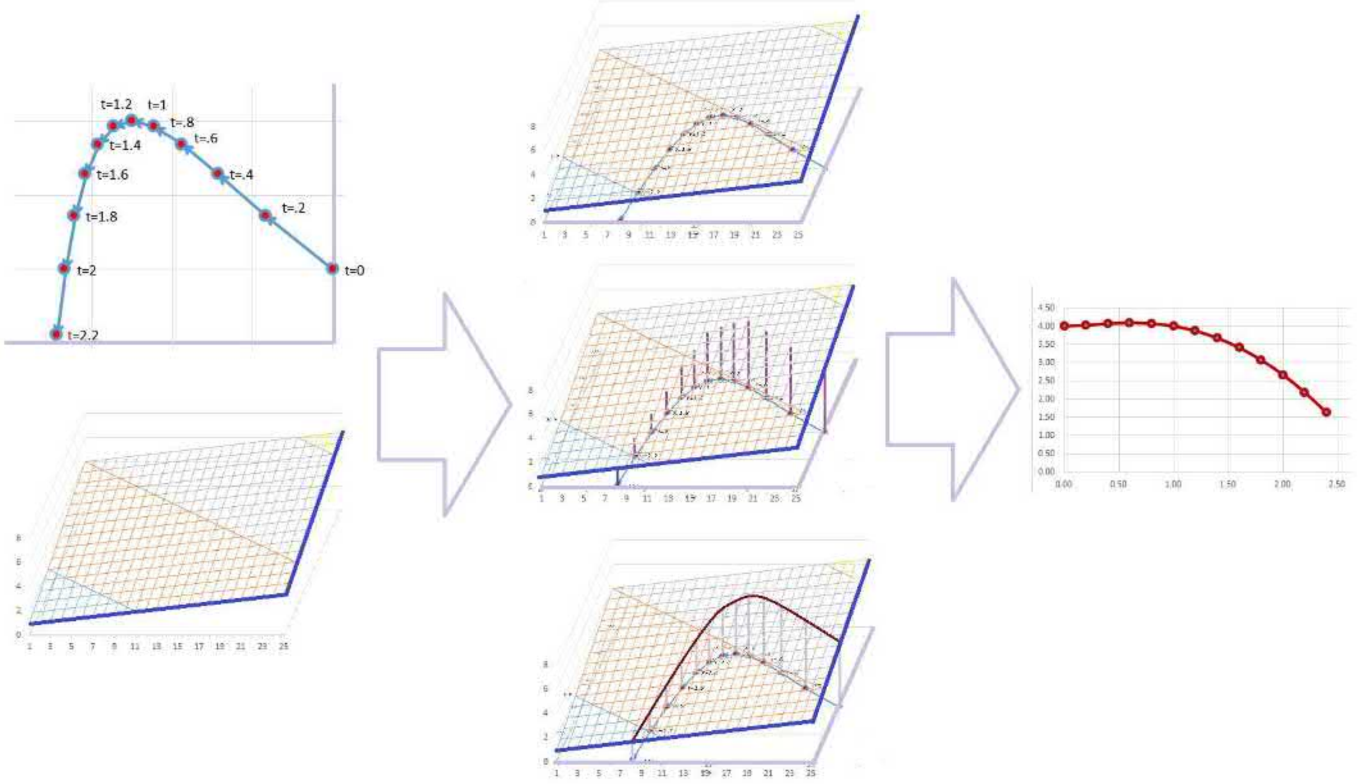


This is a *function of two variables*:

$$z = f(x, y) \,.$$

Conversely, a terrain map is a go-to metaphor for functions of two variables!

Now, back to the same question: How fast will we be climbing? The composition required is illustrated below:



We face new kinds of functions:

$$\begin{tikzcd} t \arrow[r, "\text{trip map}"] & (x,y) \arrow[r, "\text{terrain map}"'] & z \end{tikzcd}$$

Both functions deal with 3 variables at the same time, with a total of 4!

In the meantime, there are many functions of the 2 or 3 variables of location: the temperature and the pressure of the air or water, the humidity, the concentration of a particular chemical, etc.

Example 4.1.3: costs and prices

We saw an example of an abstract space: the space of prices. At its simplest, the baker does two things:

1. He watches the prices of the two ingredients of his bread: sugar and wheat.
2. He decides, based on these two numbers, what the price of the bread should be.

The space's dimension was 2, with only the two prices of the two ingredients of bread. The dependence is just as in the last example:

$$\begin{tikzcd} t \arrow[r, mapsto, "\text{costs}"] & (x,y) \arrow[r, mapsto, "\text{price}"'] & z \end{tikzcd}$$

Multiple variables lead to high-dimensional abstract spaces, such as in the case of the price of a car dependent on the prices of 1000 of its parts:

$$t \ \mapsto \ (x_1, x_2, ..., x_{1000}) \ \mapsto \ z$$

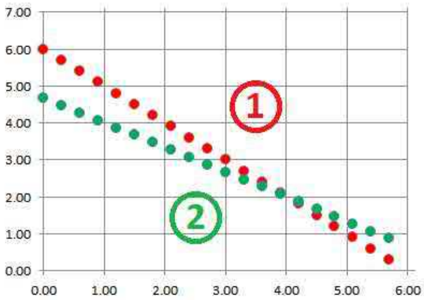
We can develop algebra, geometry, and calculus that will be applicable to a space of *any* dimension. We replace a large number of single variables with a single variable in a space of a large dimension. For example, P below is such a variable, i.e., a location in a 1000-dimensional space:

$$t \ \mapsto \ P \ \mapsto \ z$$

Example 4.1.4: mixtures

Recall the example of two types of coffee – costs \$2 per pound and \$3 per pound – used to make a blend worth \$12. The space of prices is 2-dimensional with the restriction:

$$2x + 3y = 14\,.$$



The restriction creates a 1-dimensional *subspace* – the line – within the plane. What happened is that we don't have two degrees of freedom anymore: x and y can't both be available at the same time. Once one is known, so is the other.

Now we can imagine an example of an *investment portfolio*. Suppose there are $10,000$ stocks available. The *prices per share* are known:

$$p_1,\ p_2,\ ...\ ,\ p_{10000}\,.$$

What "blend" of these would produce an investment worth \$10 million?

The question is about the *number of shares* of each stock, respectively:

$$x_1,\ x_2,\ ...\ ,\ x_{10000}\,.$$

Each investment will be recorded as a point in the $10,000$-dimensional space. And so is the price combination.

Let's combine these together:

stocks:	1	2	...	10,000	
prices:	p_1	p_2	...	p_{10000}	
# of shares:	x_1	x_2	...	x_{10000}	
subtotals:	p_1x_1	p_2x_2	...	$p_{10000}x_{10000}$	
TOTAL	p_1x_1	$+p_2x_2$	$+...$	$+p_{10000}x_{10000}$	$=10$

The restriction we have produced in the last line creates a $9,999$-dimensional *subspace* within the $10,000$-dimensional space. What happened is that we don't have $10,000$ degrees of freedom anymore: all $x_1,\ x_2,\ ...\ ,\ x_{10000}$ can't be available at the same time. Indeed, we have:

$$x_1=\frac{1}{p_1}\left(10-p_2x_2-...-p_{10000}x_{10000}\right)\,.$$

Once $9,999$ of them are known, then so is the one that is left.

Initially we will limit ourselves to dimensions that we can visualize!

Convention. We will use *upper case* letters for the entities that are (or may be) multidimensional, such as points and vectors:

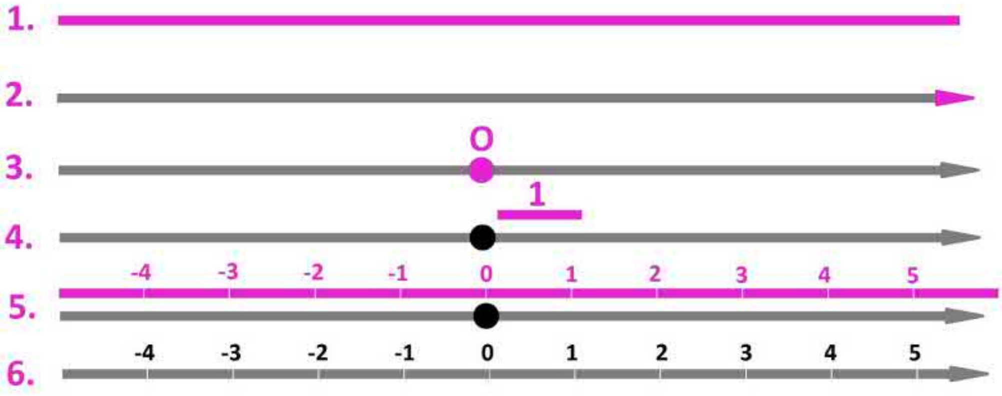
$$A,\ B,\ C,\ ...\ P,\ Q,\ ...\ ;$$

and *lower case* letters for numbers:

$$a,\ b,\ c,\ ...\ ,\ x,\ y,\ z,\ ...$$

4.2. *Euclidean spaces and Cartesian systems of dimensions* 1, 2, 3,...

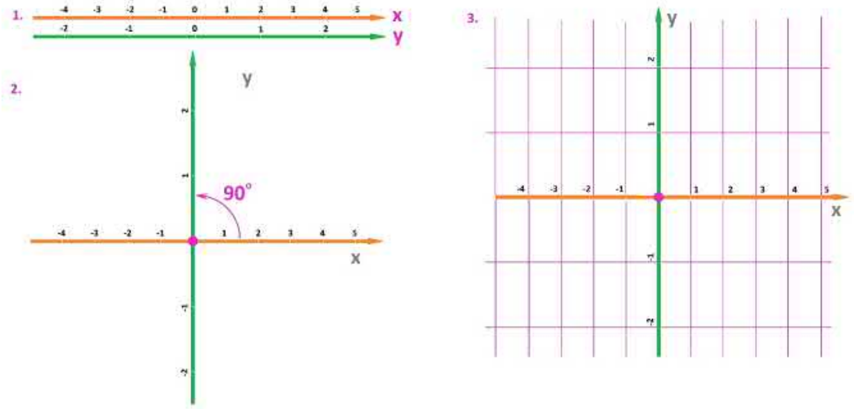
We start with the Cartesian system for dimension 1. It is a line with a certain collection of features – the origin, the positive direction, and the unit – added:



The main idea is this correspondence (i.e., a function that is one-to-one and onto):

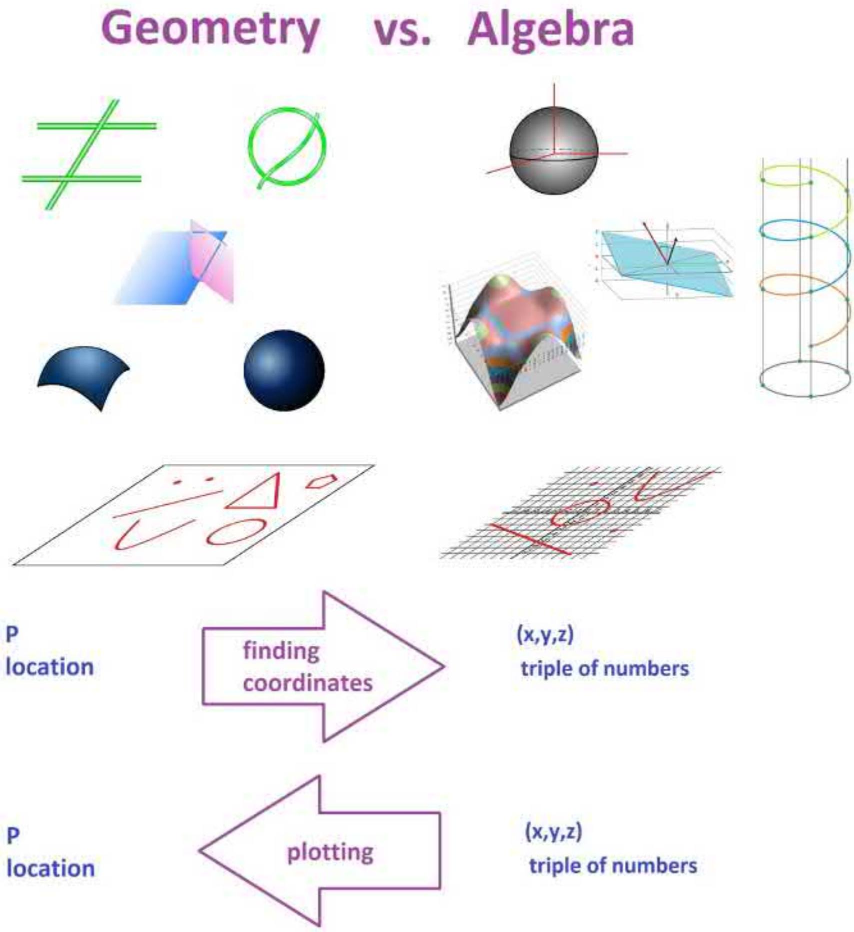
$$\boxed{\text{a location } P \longleftrightarrow \text{ a real number } x.}$$

We can have such "Cartesian lines" as many as we like and we can arrange them in any way we like. Then the Cartesian system for dimension 2 is made of *two* copies of the Cartesian system of dimension 1 aligned at 90 degrees (of rotation) from positive x to positive y:



We continue with *dimension* 3.

There is much more going on in "space" than on a plane:

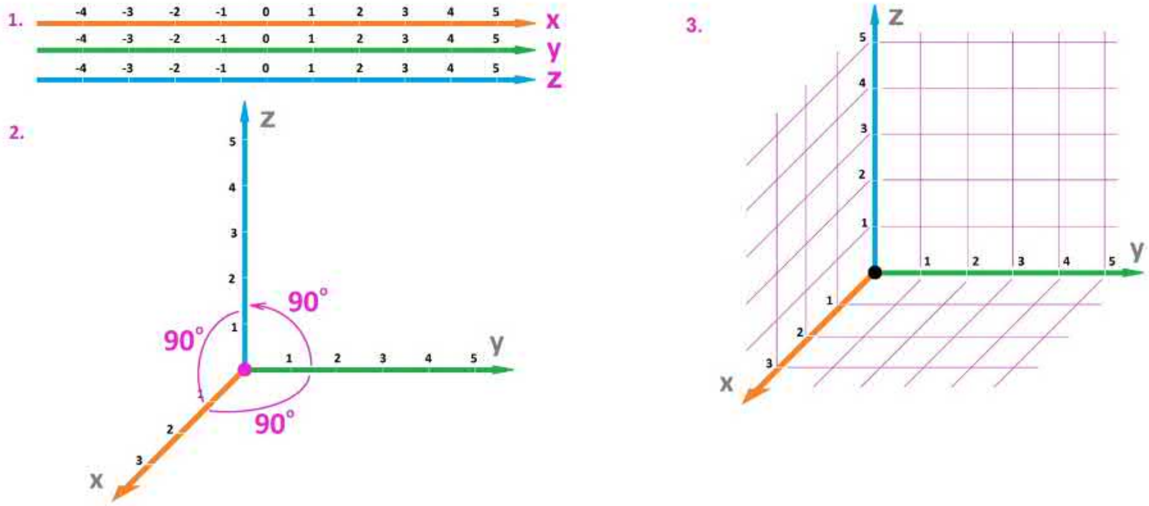


That is why we'll need *three* numbers to represent the locations.

The Cartesian system for dimension 3 is made of *three* copies of the Cartesian system of dimension 1. Just like in the case of dimension 2 above, these copies don't have to be identical; their units might be unrelated.

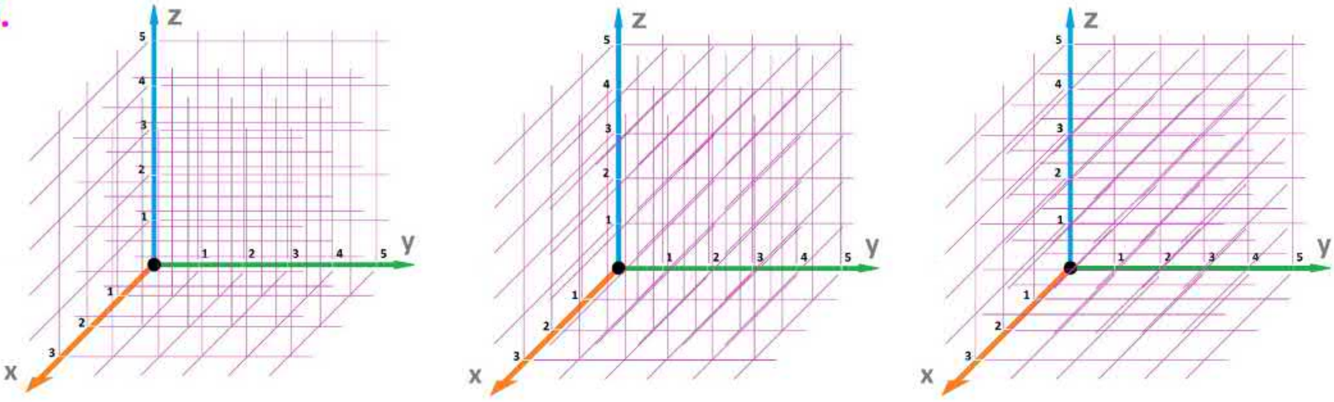
The system is built in several stages:

1. Three *coordinate axes* are chosen: the x-axis, the y-axis, and the z-axis.
2. The two axes are put together at their origins so that it is a 90-degree turn from the positive direction of one axis to the positive direction of the next – from x to y to z to x.
3. Use the marks on the axis to draw grids on the planes.
4. We repeat these three grids in parallel to create threads in space.

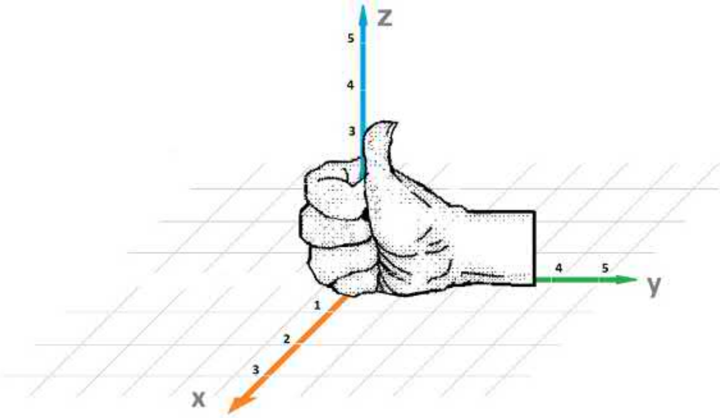


The last step is shown below:

4.



The second requirement is called the *Right Hand Rule*:

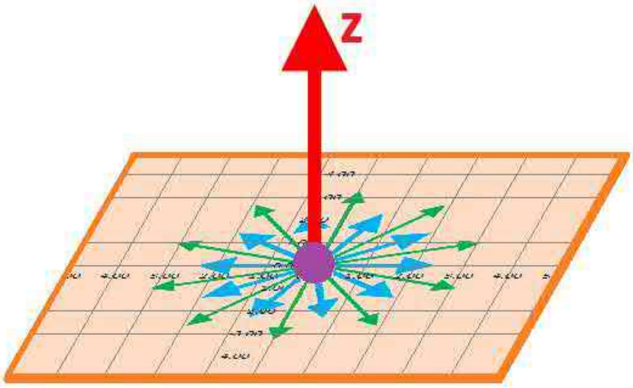


It reads:

- ▶ If we curl our fingers from the x-axis to the y-axis, our thumb will point in the direction of the z-axis.

We can also understand this idea if we imagine turning a screwdriver in this direction and seeing which way the screw goes.

The axes are perpendicular to each other, but there is more! For example, in addition to the x- and y-axis being perpendicular to the z-axis, *all* lines in the xy-plane are perpendicular to it:



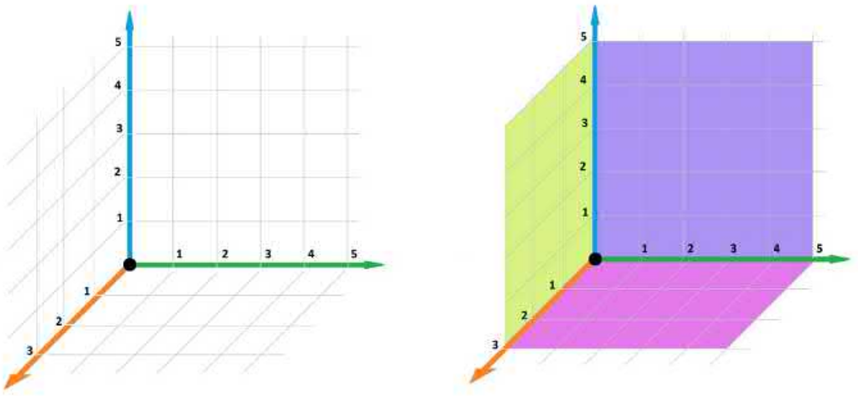
The main purpose of the Cartesian system remains the same; it is this correspondence:

$$\boxed{\text{a location } P \ \longleftrightarrow\ \textit{a triple} \text{ of real numbers } (x, y, z)}$$

Warning!

The three variables or quantities represented by the three axes may be unrelated. Then our visualization will remain valid with rectangles instead of squares, and boxes instead of cubes.

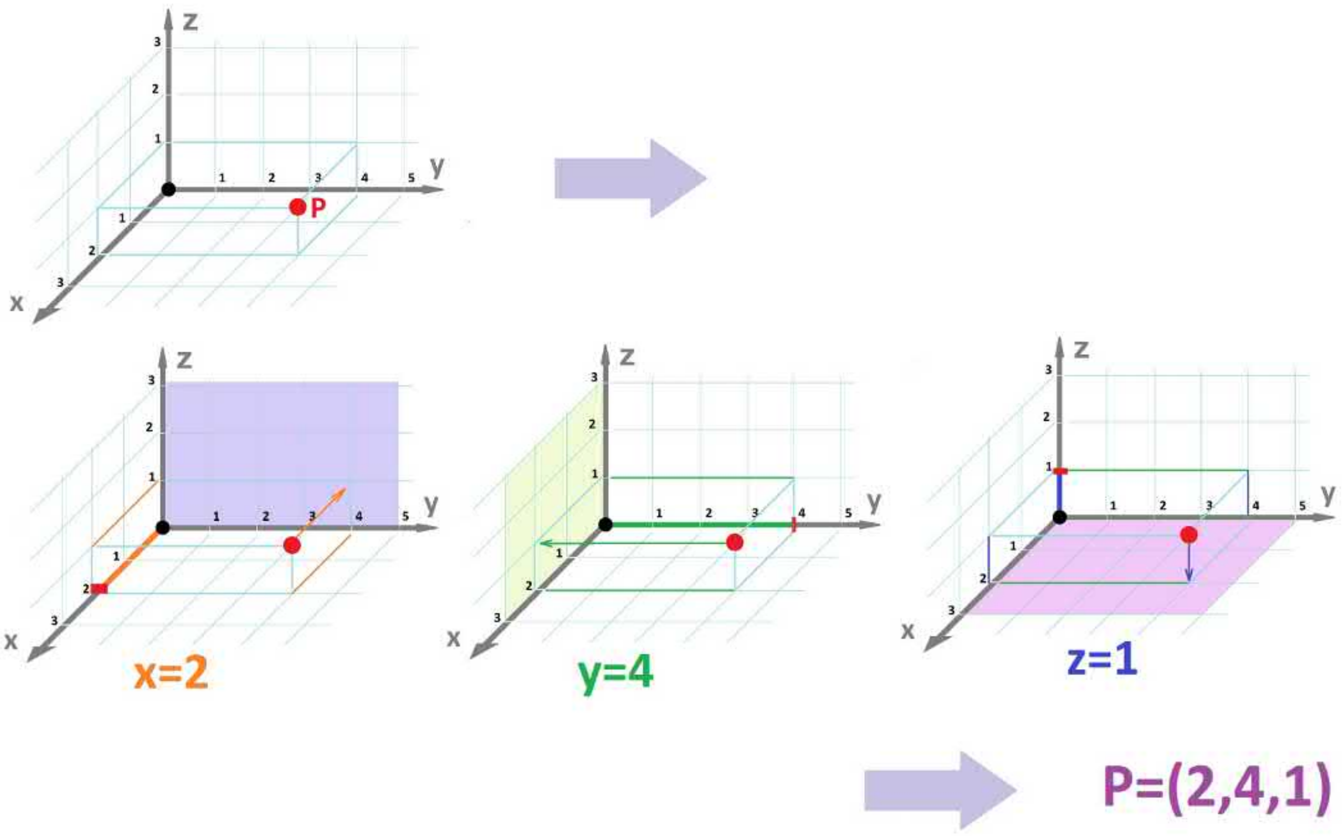
Alternatively, the system is built from three copies of the Cartesian plane: the xy-plane, the yz-plane, and the zx-plane. They are arranged at 90 degrees as walls of a room:



These planes are called the *coordinate planes*.

This is how the system works:

First, suppose P is a *location* in this space. We find the dimensions of the *box* with one corner at O and the opposite at P. We find the distances from the three planes to that location – positive in the positive direction and negative in the negative direction – and the result is the three coordinates of P, some *numbers* x, y, and z. The distance from the yz-plane is measured along the x-axis, etc. We use the nearest mark to simplify the task:

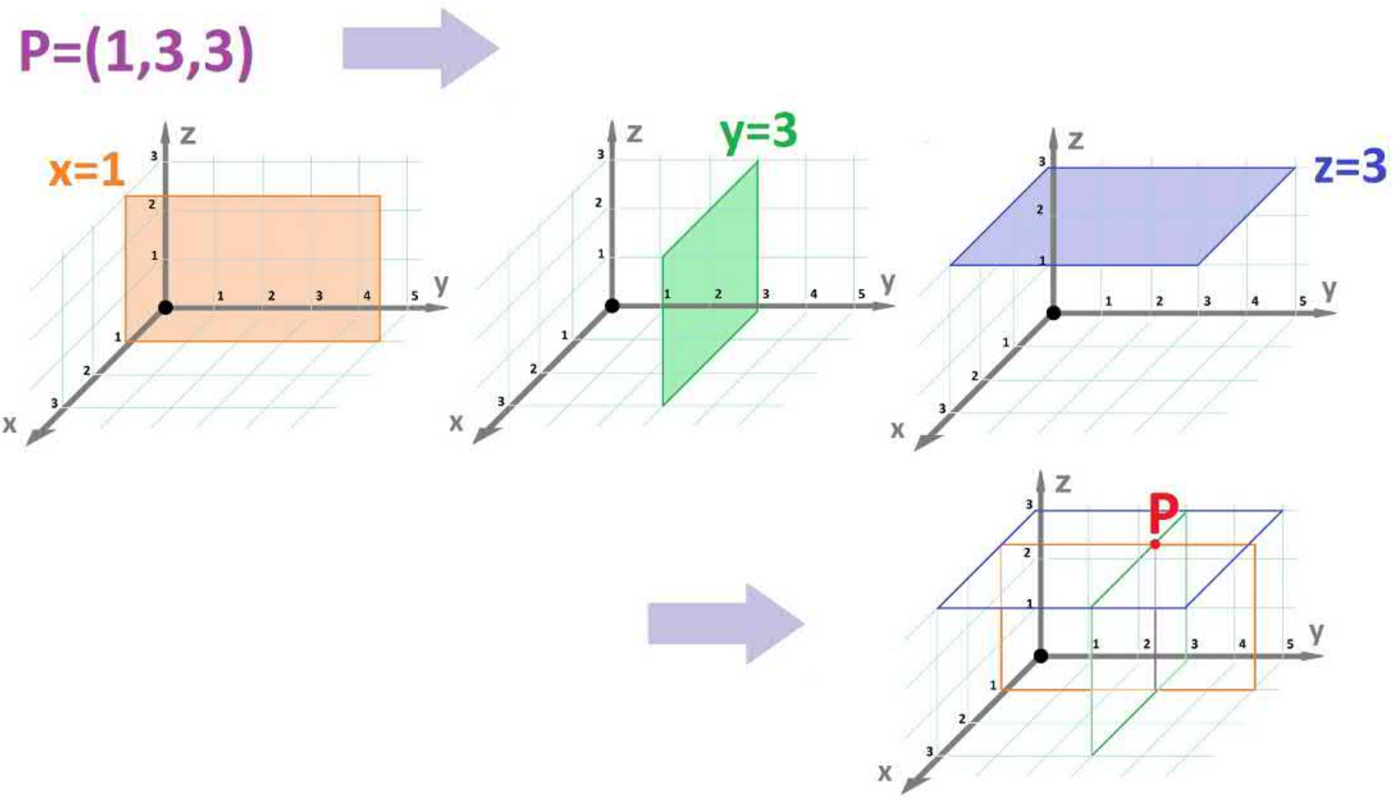


Conversely, suppose x, y, z are *numbers*. If we need to build a box with these dimensions:

- First, we measure x as the distance from the yz-plane – positive in the positive direction and negative in the negative direction – along the x-axis and create a plane parallel to the yz-plane.

- Second, we measure y as the distance from the xz-plane along the y-axis and create a plane parallel to the xz-plane.
- Third, we measure z as the distance from the xy-plane along the z-axis and create a plane parallel to the xy-plane.

The intersection of these three planes – as if these were the two walls and the floor in a room – is a *location* $P=(x,y,z)$ in the space. We use the nearest marks to simplify the task:

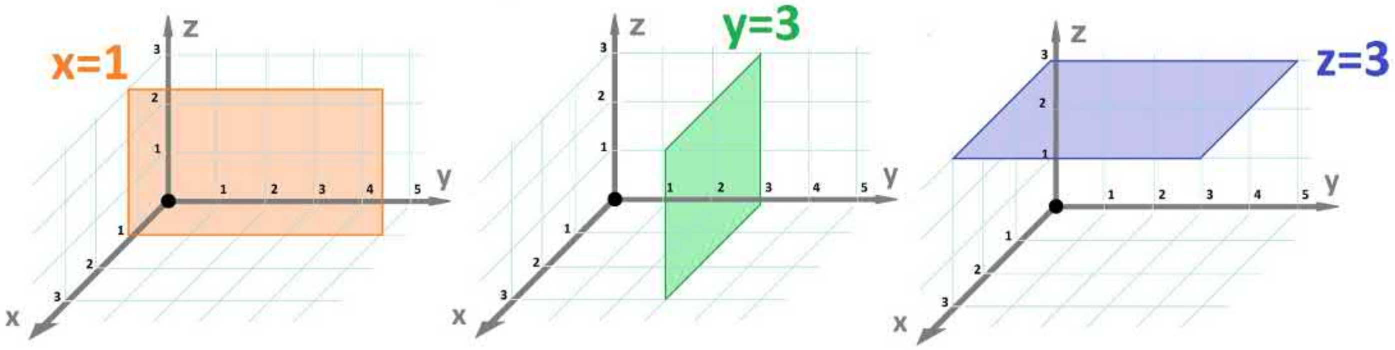


This 3-dimensional coordinate system is called *the Cartesian space* or the 3-*space*.

Once the coordinate system is in place, it is acceptable to think of location as triples of numbers and vice versa. In fact, we can write:

$$P=(x,y,z)\,.$$

Consider more of the planes parallel to the coordinate planes:



Then, we have a compact way to represent these planes:

$$x=k,\ y=k,\quad \text{or } z=k\,,$$

for some real k.

We can use this idea to reveal the internal structure of the space.

Theorem 4.2.1: Planes Parallel to Coordinate Planes

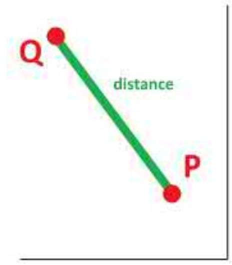
1. *If L is a plane parallel to the xy-plane, then all points on L have the same z-coordinate. Conversely, if a collection L of points consists of all points with the same z-coordinate, L is a plane parallel to the xy-plane.*
2. *If L is a plane parallel to the yz-plane, then all points on L have the same x-coordinate. Conversely, if a collection L of points consists of all points with the same x-coordinate, L is a plane parallel to the yz-plane.*
3. *If L is a plane parallel to the zx-plane, then all points on L have the same y-coordinate. Conversely, if a collection L of points consists of all points with the same y-coordinate, L is a plane parallel to the zx-plane.*

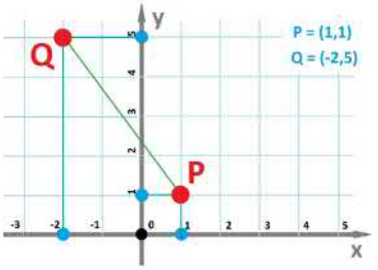
We turn to *analytic geometry* of the 3-space.

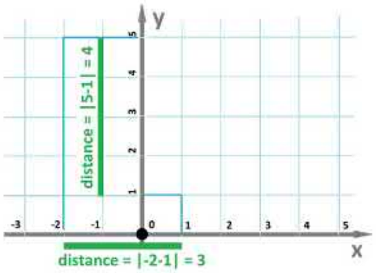
Now that everything is *pre-measured*, we can solve geometric problems by algebraically manipulating coordinates.

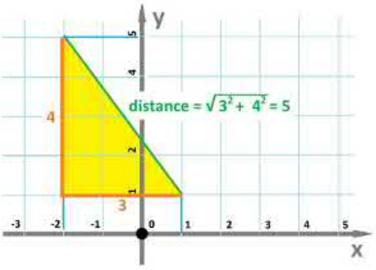
The first geometric task is finding the *distance*: What is the distance between locations P and Q in terms of their coordinates (x, y, z) and (x', y', z')?

For dimension 2, we used the distance formula from the 1-dimensional case. We found distance between two points on the plane as the length of the diagonal of the rectangle – with its sides parallel to the coordinate axes – that has these points at the opposite corners:

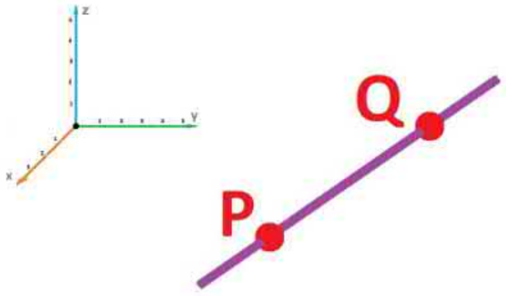






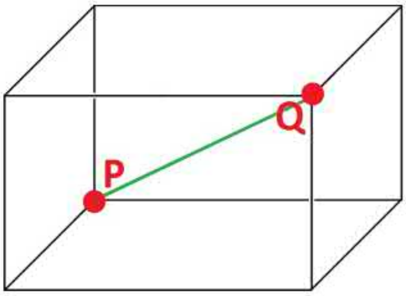


First, we need to realize that the problem itself is 1-dimensional! Indeed, any two points, in any space – 1-, 2-, 3-, or n-dimensional – can be connected by a line, and along that line – a 1-dimensional space – we measure the distance:



The coordinate system is just a means to an end.

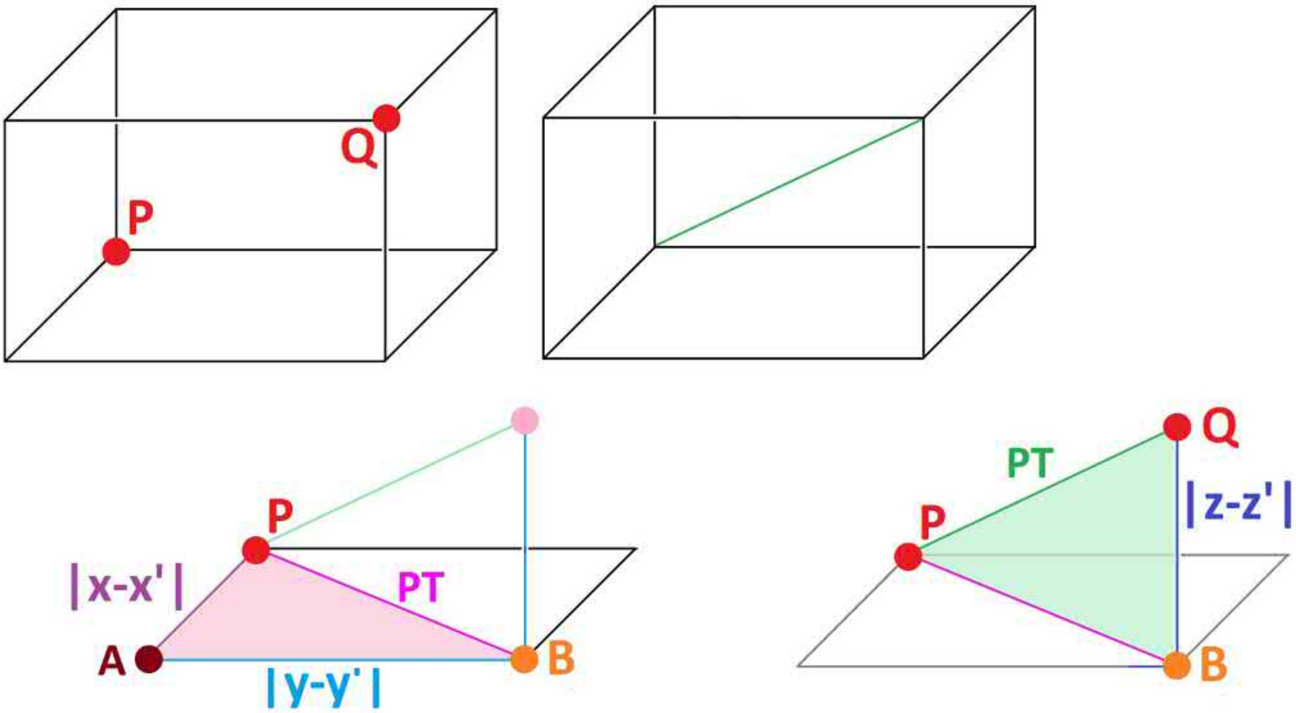
So, we need to find the distance between two points in space as the length of the diagonal of the *box* – with its edges parallel to the coordinate axes, and sides parallel to the coordinate planes – that has these points at the opposite corners:



We now utilize these two facts:

1. Every coordinate plane of the 3-space has its own, 2-dimensional, coordinate system.
2. The coordinate axes are perpendicular to the coordinate planes.

This is the outline of the construction:



The Pythagorean theorem is to be applied within the horizontal plane and then within a certain vertical plane.

The formula is, as we anticipated, symmetric with respect to the dimensions:

Theorem 4.2.2: Distance Formula for dimension 3

The distance between points with coordinates $P = (x, y, z)$ *and* $Q = (x', y', z')$ *is*

$$d(P, Q) = \sqrt{(x - x')^2 + (y - y')^2 + (z - z')^2}\,.$$

Proof.

The segment between the points $P = (x, y, z)$ and $Q = (x', y', z')$ is the diagonal of this "box". We use the distance formula from the 1-dimensional case separately for each of the three axes, as follows:

- The distance between x and x' on the x-axis is $|x - x'|$.
- The distance between y and y' on the y-axis is $|y - y'|$.
- The distance between z and z' on the z-axis is $|z - z'|$.

These are the dimensions of the box.

Next we use the *Pythagorean Theorem* twice. We first find the length of the diagonal of the bottom of the box and then the length of the main diagonal:

$$\begin{array}{lllll}
\text{PT 1:} & d(P,A)=|x-x'|, & d(A,B)=|y-y'| & & \\
 & & \Longrightarrow & d(P,B)^2 & =(x-x')^2+(y-y')^2 \\
\text{PT 2:} & d(P,B)^2=(x-x')^2+(y-y')^2, & d(B,Q)=|z-z'| & & \\
 & & \Longrightarrow & d(P,Q)^2 & =d(P,B)^2+d(B,Q)^2 \\
 & & & & =(x-x')^2+(y-y')^2+(z-z')^2
\end{array}$$

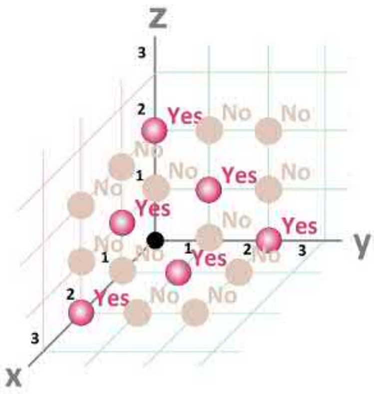
Exercise 4.2.3

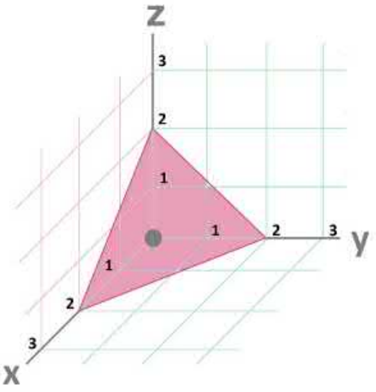
Prove that in the latter case the triangle is indeed a right triangle.

Relations are used in the same way as before but with more variables. A relation processes a triple of numbers (x,y,z) as the input and produces an output, which is: Yes or No. If we are to plot the *graph* of a relation, this output becomes: a point or no point. For example:

$$\begin{array}{ccccccl}
 & & & & & & \text{outcome:} \\
 & & & & & & \text{Plot point } (x,y,z). \\
\text{plane:} & & \text{triple:} & & \text{relation:} & \texttt{TRUE}\ \nearrow & \\
\mathbf{R}^3 & \to & (x,y,z) & \to & \boxed{x+y+z=2\,?} & \to & \\
 & & & & & \texttt{FALSE}\ \searrow & \\
 & & & & & & \text{Don't plot anything.}
\end{array}$$

We can do it by hand:





We can use, as before, the *set-building notation*:

$$\{(x,y,z): \quad \text{a condition on } x,y,z\}\,.$$

For example, the graph of the above relation is a subset of $\mathbf{R}^3$ given by:

$$\{(x,y,z):\ x+y+z=2\}\,.$$

What about dimension 4 and higher?

We cannot use our physical space as a reference anymore! We can't use it for visualization either. The space is abstract.

The idea of the n-dimensional space remains the same; it is the correspondence:

$$\boxed{\text{a location } P \longleftrightarrow \text{ a string of } n \text{ real numbers } (x_1,x_2,x_3,...,x_n)}$$

Using the same letter with subscripts is preferable even for dimension 3 as the symmetries between the axes and variables are easier to detect and utilize. However, using just P is often even better!

Because of the difficulty or even impossibility of visualization of these "locations" in dimension 4, this correspondence becomes much more than just a way to go back and forth whenever convenient. This time, we just say "It's the same thing".

Example 4.2.4: non-homogeneous variables

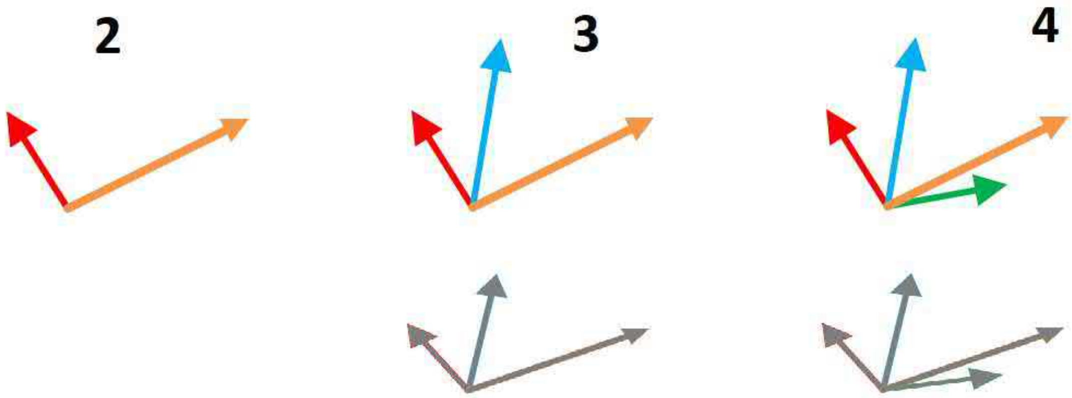
This may be the data continuously collected by a *weather center*:

1	2	3	4	5	...
temperature	pressure	precipitation	humidity	sunlight	...

They are all measured in different units and cannot be seen as an analog of our physical space.

How do we visualize this n-dimensional space?

Let's first realize that, in a sense, we have failed even with the *three*-dimensional space! We have had to squeeze these three dimensions on a *two*-dimensional piece of paper. Without the numbers telling us what to expect, we wouldn't be able to tell the dimension (top row):



At best, we are seeing the *shadows* of the lines (bottom row).

These are the spaces we will study and the notations for them:

Euclidean spaces

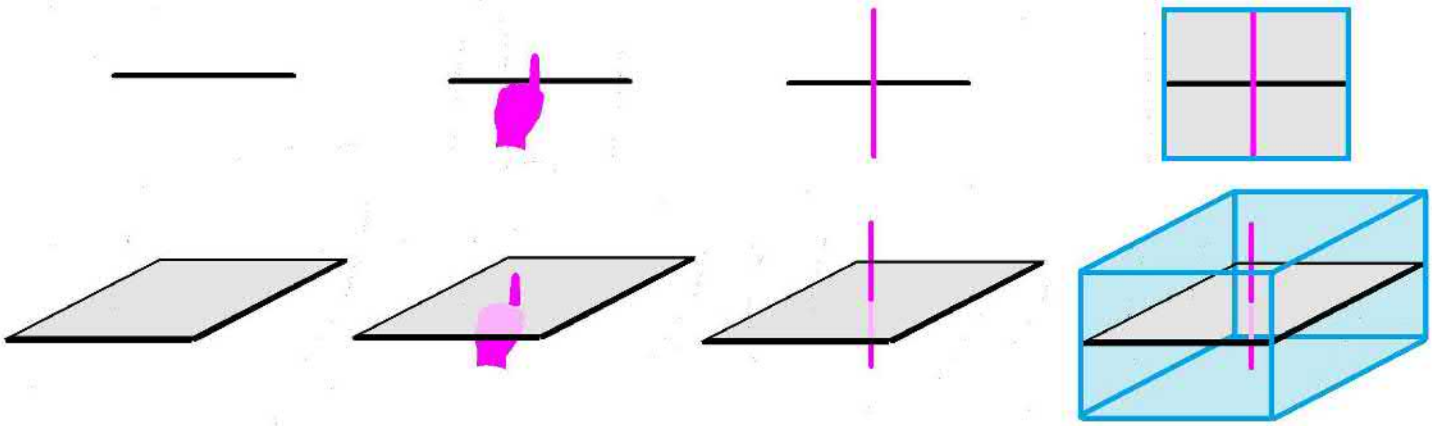
- $\mathbf{R}$, all real numbers (line)
- $\mathbf{R}^2$, all pairs of real numbers (plane)
- $\mathbf{R}^3$, all triples of real numbers (space)
- $\mathbf{R}^4$, all quadruples of real numbers
- ...
- $\mathbf{R}^n$, all strings of n real numbers
- ...

Each of them is supplied with its own algebra and geometry.

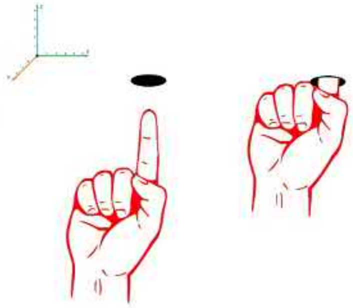
We can build these by consecutively adding one dimension at a time.

- If $\mathbf{R}$ is given, we treat it as the x-axis and then add another axis, the y-axis, perpendicular to the first.
- The result is $\mathbf{R}^2$, which we treat as the xy-plane and then add another axis, the z-axis, perpendicular to the first two.
- The result is $\mathbf{R}^3$, which we treat as the xyz-space and then add another axis perpendicular to the first three; and so on.

Here is the summary:



With our 1-dimensional finger, we puncture the space. In $\mathbf{R}^4$, the same thing happens; the finger disappears:



The formula that represents the line in the first row is:

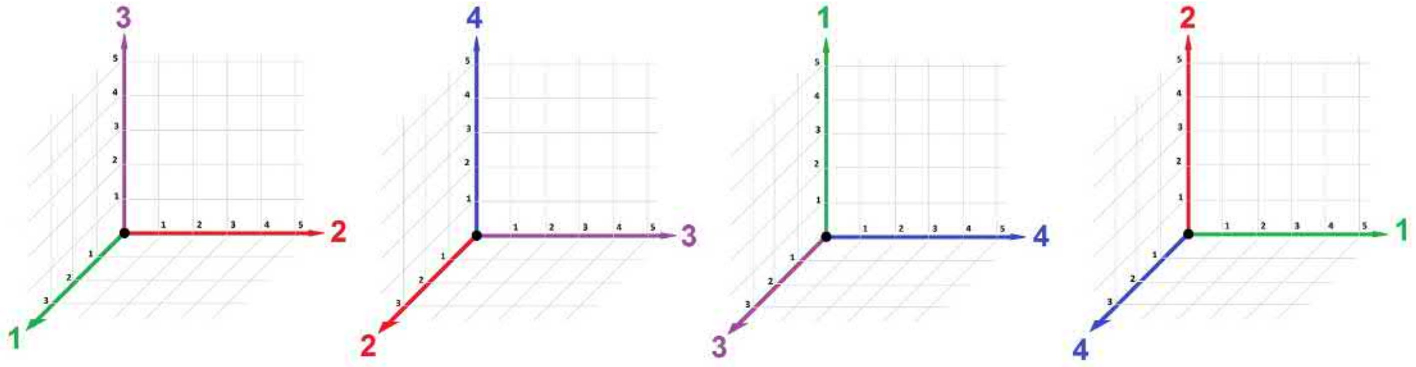
$$y = 0 \quad \text{or } x_2 = 0\,.$$

The formula that represents the plane in the second row is:

$$z = 0 \quad \text{or } x_3 = 0\,.$$

This space is abstract but is still constructed from lower-dimensional spaces:

1. Four copies of $\mathbf{R}$: the four coordinate axes.
2. Six copies of $\mathbf{R}^2$: the six coordinate planes, each spanned on a pair of those coordinate axes.
3. Four copies of $\mathbf{R}^3$: four spaces, each constructed on the frame of three of those coordinate planes.



Exercise 4.2.5

What is the formula that represents each of these spaces?

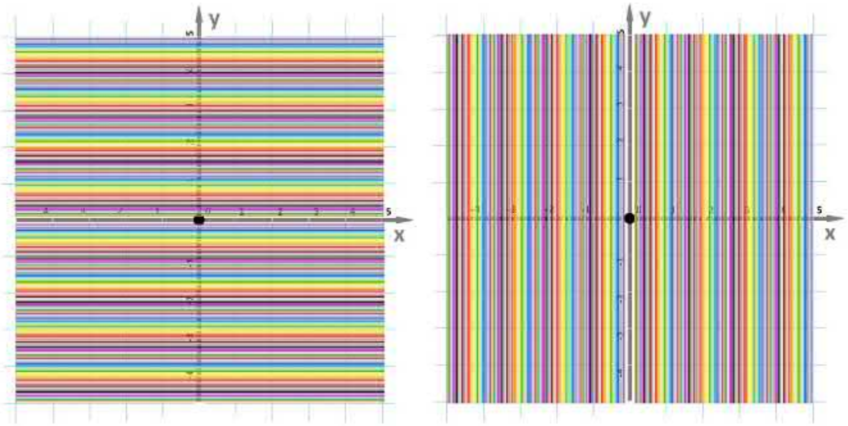
Exercise 4.2.6

How many coordinate planes are there in $\mathbf{R}^5$? $\mathbf{R}^n$? How many coordinate spaces?

So, these spaces aren't unrelated!

In order to reveal the internal structure of a spaces, we look for *lower-dimensional* spaces in it.

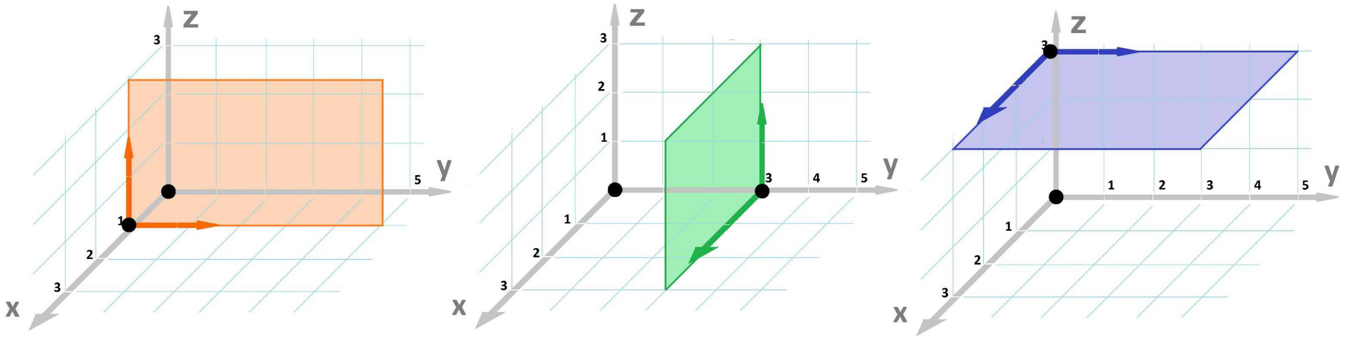
The plane is a *stack of lines*, each of which is just a copy of one of the coordinate axes:



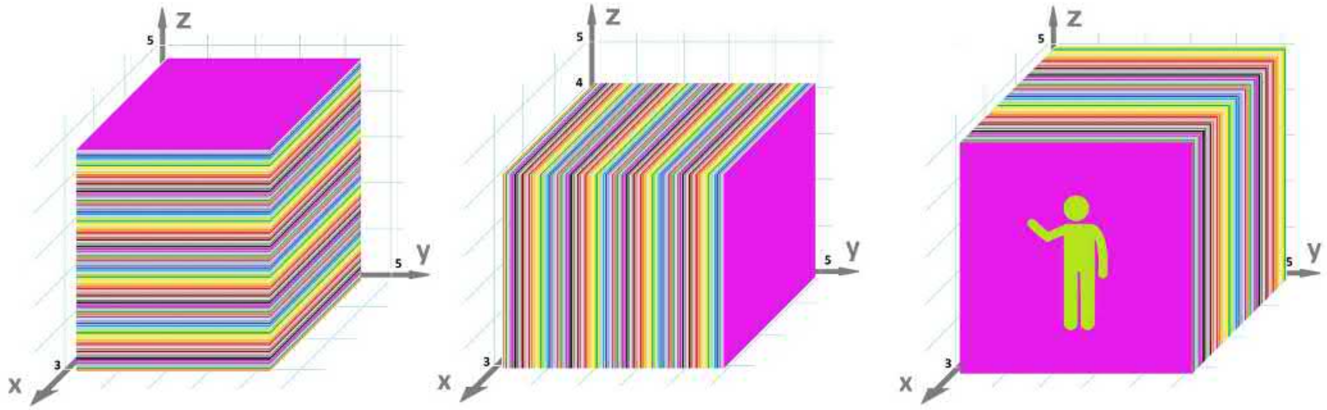
These lines are given by the equations for each real a or b:

$$\boxed{x = a,\ \ y = b}$$

They are *copies* of $\mathbf{R}$ and will have the same algebra and geometry. In fact, they can have their own coordinate systems:



Now, one can think of the 3-space as a *stack of planes*, each of which is just a copy of one of the coordinate planes:



They are given by the equations for all real a, b, c:

$$\boxed{x = a,\ \ y = b,\ \ z = c}$$

These are copies of $\mathbf{R}^2$.

If a 2-dimensional person can recognize – thinking mathematically – that the 3-space is made of layers of copies of his own space, we can see our physical 3-space as just a single "layer" in $\mathbf{R}^4$.

So, $\mathbf{R}^4$ is a "stack" of $\mathbf{R}^3$s. How they fit together is hard to visualize, but they are still copies of $\mathbf{R}^3$ given by equations:

$$\boxed{x_1 = a_1,\ x_2 = a_2,\ x_3 = a_3,\ x_4 = a_4}$$

Exercise 4.2.7

What is a line a stack of?

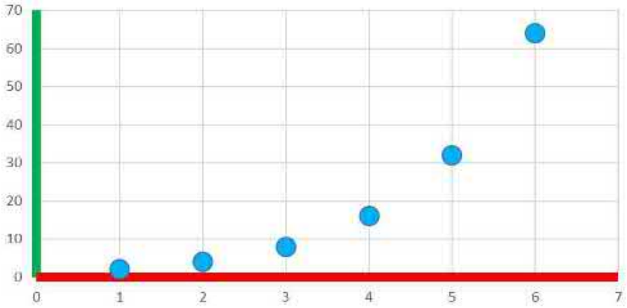
So, we can see many copies of $\mathbf{R}^m$ in $\mathbf{R}^n$, with $n > m$.

Beyond a certain point, the chance to visualize the space is gone. We, however, are still able to visualize the space one element at a time. For example, a point in the n-dimensional space is nothing but a *sequence* with n terms:

k	1	2	3	4	5	...	n
x_k	x_1	x_2	x_3	x_4	x_5	...	x_n

It's just a function, with the inputs in the first row and the outputs in the second. We visualize functions with their *graphs*. For example, this string of 6 numbers is a point in the 6-dimensional space:

n	y
1	2.0
2	4.0
3	8.0
4	16.0
5	32.0
6	64.0

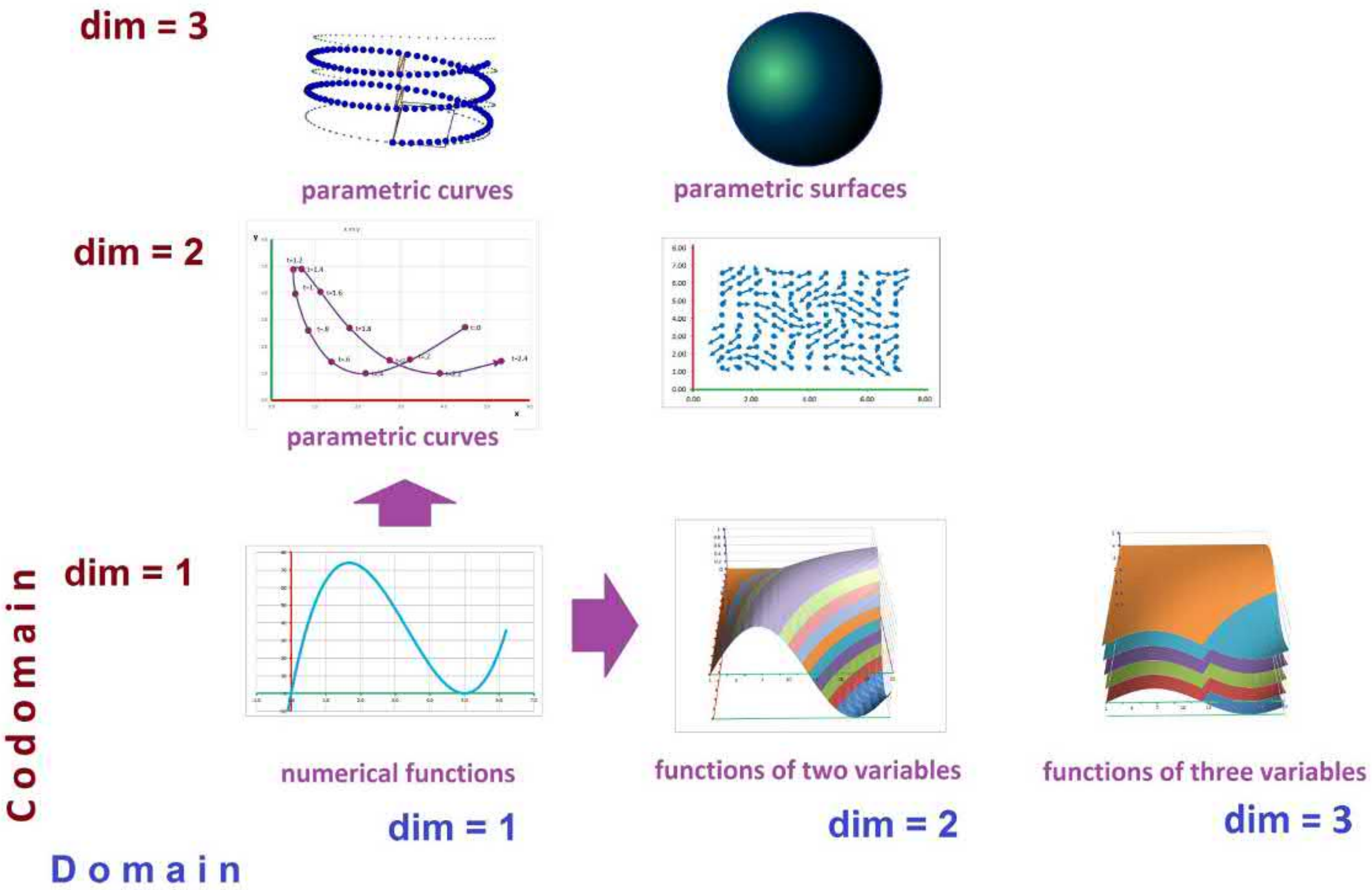


Warning!

The curve that you see is incidental because the rows of the table can be re-arranged.

Next, there is no point in studying spaces without studying *functions* between them.

Let's review multidimensional functions. We place them in a table with two axes representing the dimension of the domain and the dimension of the codomain:



We always start at the very first cell. Previously we made a step in the vertical direction and explored the first column of this table. We also moved to the right.

With all this complexity, we shouldn't overlook the general point of view on functions. We represent a function diagrammatically as a *black box* that processes the input and produces the output of whatever nature:

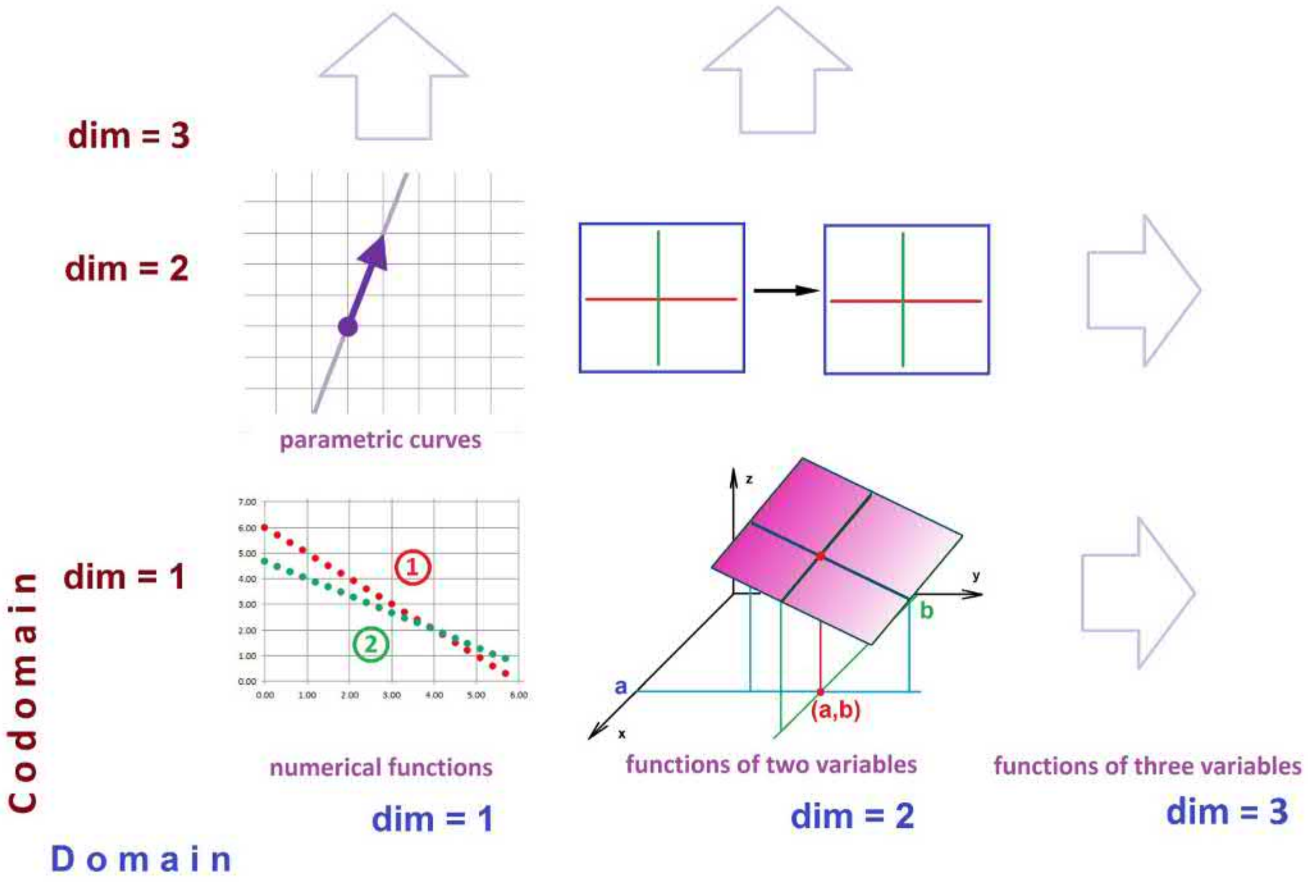
$$\boxed{\begin{array}{ccccc}\text{input} & & \text{function} & & \text{output}\\ X & \mapsto & \boxed{\ f\ } & \mapsto & Y\end{array}}$$

Let's take one from the left-most column and one from the bottom row:

$$\begin{array}{rcccl} & & \text{parametric} & & \\ \text{input} & & \text{curve} & & \text{output}\\ t & \mapsto & \boxed{\ F\ } & \mapsto & X\\ \mathbf{R} & & & & \mathbf{R}^m\\ \text{number} & & & & \text{point}\\ \hline \text{time} & & & & \text{prices of parts}\\ \text{time} & & & & \text{prices of stocks}\end{array} \qquad \begin{array}{rcccl} & & \text{function of} & & \\ \text{input} & & \text{two variables} & & \text{output}\\ X & \mapsto & \boxed{\ f\ } & \mapsto & z\\ \mathbf{R}^m & & & & \mathbf{R}\\ \text{point} & & & & \text{number}\\ \hline \text{prices of parts} & & & & \text{price of car}\\ \text{prices of stocks} & & & & \text{value of portfolio}\end{array}$$

They can be linked up and produce a composition, which is just a numerical function.

Above is a view of "generic" functions. In the linear algebra context, the functions in the table are simpler and so are their visualizations:



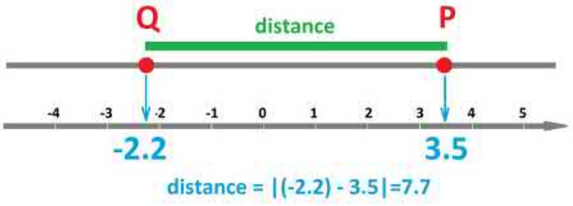
We turn to analytic geometry of n-dimensional spaces.

4.3. *Geometry of distances*

The axes of the Cartesian system $\mathbf{R}^3$ for our physical space refer to the same: distances to the coordinate planes. They are (or should be) measured in the same unit. Even though, in general, the axes of $\mathbf{R}^n$ refer to unrelated quantities, they *may* be measured in the same unit, such as the prices of n commodities being traded. When this is the case, doing geometry in $\mathbf{R}^n$ based entirely on the coordinates of points is possible.

A Cartesian system has everything in the space *pre-measured.*

In particular, we compute (rather than measure) the distances between locations because the distance can be expressed in terms of the coordinates of the locations.



Theorem 4.3.1: Distance Formula for Dimension 1

The distance from point P to point Q in $\mathbf{R}$ given by real numbers x and x' respectively is

$$\boxed{d(P,Q) = |x - x'|}$$

Here, the geometry problem of finding distances relies on the algebra of real numbers (the subtraction).

Now the coordinate system for dimension 2. The formula for the distance between locations P and Q in terms of their coordinates (x, y) and (x', y') is found by using the distance formula from the 1-dimensional case for either of the two axes in order to find

1. the distance between x and x', which is $|x - x'| = |x' - x|$, and
2. the distance between y and y', which is $|y - y'| = |y' - y|$, respectively.

Then the two numbers are put together by the *Pythagorean Theorem* taking into account this simplification:

$$|x - x'|^2 = (x - x')^2, \ |y - y'|^2 = (y - y')^2 .$$

Theorem 4.3.2: Distance Formula for Dimension 2

The distance between points P and Q in $\mathbf{R}^2$ with coordinates (x, y) and (x', y') respectively is

$$\boxed{d(P,Q) = \sqrt{(x - x')^2 + (y - y')^2}}$$

The two exceptional cases when P and Q lie on the same vertical or the same horizontal line (and the triangle "degenerates" into a segment) are treated separately.

Now the coordinate system for dimension 3. We can guess that there will be another term in the sum of the *Distance Formula.*

Theorem 4.3.3: Distance Formula for Dimension 3

The distance between points P and Q in $\mathbf{R}^3$ with coordinates (x, y, z) and (x', y', z') respectively is

$$\boxed{d(P,Q) = \sqrt{(x - x')^2 + (y - y')^2 + (z - z')^2}}$$

A pattern starts to appear:

> ▶ The square of the distance is the sum of the squares of the distances along each of the coordinates.

Thinking by analogy, we continue on to include the case of dimension 4:

dimension	points	coordinates	distance
1	P	x	
	Q	x'	$d(P,Q)^2 = (x-x')^2$
2	P	(x,y)	
	Q	(x',y')	$d(P,Q)^2 = (x-x')^2 + (y-y')^2$
3	P	(x,y,z)	
	Q	(x',y',z')	$d(P,Q)^2 = (x-x')^2 + (y-y')^2 + (z-z')^2$
4	P	(x_1,x_2,x_3,x_4)	
	Q	(x_1',x_2',x_3',x_4')	$d(P,Q)^2 = (x_1-x_1')^2 + (x_2-x_2')^2 + (x_3-x_3')^2 + (x_4-x_4')^2$
...	...	...	...

There are n terms in dimension n:

dimension	points	coordinates	distance
n	P	$(x_1,x_2,...,x_n)$	
	Q	$(x_1',x_2',...,x_n')$	$d(P,Q)^2 = (x_1-x_1')^2 + (x_2-x_2')^2 + ... + (x_n-x_n')^2$

Example 4.3.4: geometry of $\mathbf{R}^3$ in $\mathbf{R}^4$

The formula for $n = 1,2,3$ is justified by what we know about the physical space. What about $n = 4$ and above? Let's take a look at the copies of $\mathbf{R}^3$ that make up $\mathbf{R}^4$. One of them is given by $x_4 = a_4$ for some real number a_4. If we take any two points P,Q within it, the formula becomes:

$$\begin{aligned} d(P,Q) &= \sqrt{(x_1-x_1')^2 + (x_2-x_2')^2 + (x_3-x_3')^2 + (a_4-a_4)^2} \\ &= \sqrt{(x_1-x_1')^2 + (x_2-x_2')^2 + (x_3-x_3')^2}\,. \end{aligned}$$

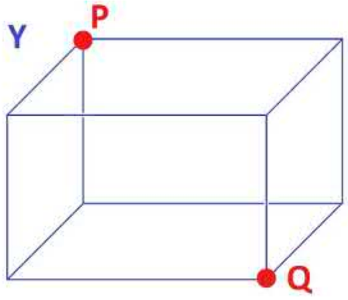
In other words, the distance is the same as the one for dimension 3. We conclude that the geometry of such a copy of $\mathbf{R}^3$ is the same as the "original"!

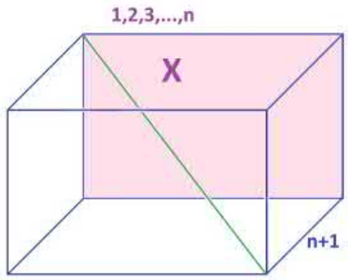
Exercise 4.3.5

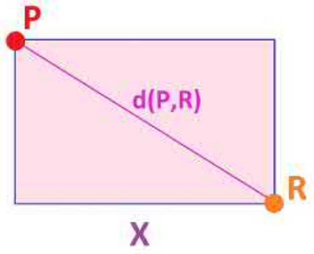
Show that the geometry of *any* plane in $\mathbf{R}^3$ and $\mathbf{R}^4$ is the same as that of $\mathbf{R}^2$.

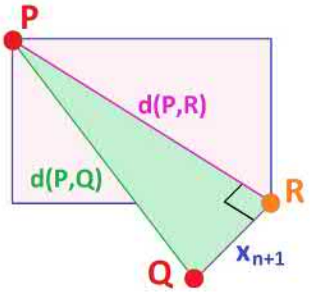
Can we justify this formula with more than just "It's a pattern"? Yes, we progress from understanding the geometry of $X = \mathbf{R}^n$ to that of $Y = \mathbf{R}^{n+1}$, every time.

Suppose the distances in $X = \mathbf{R}^n$ are computed by the above formula. Then, we add an extra axis – perpendicular to the rest – to create $Y = \mathbf{R}^{n+1}$:









Then, the *Pythagorean Theorem* is applied (green triangle). The computation is just as in the case $n = 3$ presented above:

$$d(P,Q)^2 = d(P,R)^2 + d(R,Q)^2 = (x_1-x_1')^2 + (x_2-x_2')^2 + ... + (x_n-x_n')^2 + (x_{n+1}-x_{n+1}')^2\,.$$

The formula applies to a space of any dimension n. It matches the measured distance in the physical space: $n = 1, 2, 3$. We can't say the same about the spaces of dimensions $n > 3$. They are abstract spaces. The formula, therefore, is seen as the *definition* of the distance for these spaces:

Definition 4.3.6: Euclidean metric

The *Euclidean distance between points* P and Q in $\mathbf{R}^n$ is defined to be the square root of the sum of the squares of the distances for each of the coordinates:

$$\boxed{\begin{aligned} P &= (x_1, x_2, ..., x_n) \\ Q &= (x_1', x_2', ..., x_n') \\ d(P, Q) &= \sqrt{(x_1 - x_1')^2 + (x_2 - x_2')^2 + ... + (x_n - x_n')^2} \end{aligned}}$$

We refer to the formula as the *Euclidean metric.* The space $\mathbf{R}^n$ equipped with the Euclidean metric is called the *n-dimensional Euclidean space.*

Now, the formula is somewhat complicated. Is it possible to have a few simple rules apply equally to all dimensions, without reference to the formulas?

We formulate three very simple properties of the distances. First, the distances can't be negative and, moreover, for the distance to be zero, the two points have to be the same. Second, the distance from P to Q is the same as the distance from Q to P. And so on.

Theorem 4.3.7: Axioms of Metric Space

Suppose P, Q, S are points in $\mathbf{R}^3$. Then the following properties are satisfied:

- ***Positivity:*** *$d(P, Q) \geq 0$; and $d(P, Q) = 0$ if and only if $P = Q$.*
- ***Symmetry:*** *$d(P, Q) = d(Q, P)$.*
- ***Triangle Inequality:*** *$d(P, Q) + d(Q, S) \geq d(P, S)$.*

Proof.

Suppose $d(P, Q) = 0$. Then

$$0 = d(P, Q)^2 = (x_1 - x_1')^2 + (x_2 - x_2')^2 + ... + (x_n - x_n')^2 \,.$$

Since none of the terms is negative, all have to be zero:

$$(x_1 - x_1')^2 = 0,\ (x_2 - x_2')^2 = 0,\ ...,\ (x_n - x_n')^2 = 0\,.$$

Therefore,

$$x_1 = x_1',\ x_2 = x_2',\ ...,\ x_n = x_n'\,.$$

It follows that $P = Q$.

Exercise 4.3.8

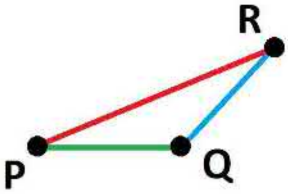
Prove the rest of the theorem.

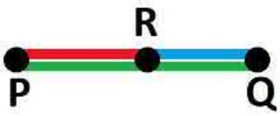
These geometric properties have been justified following the familiar geometry of the "physical space" $\mathbf{R}^3$. However, they also serve as a *starting point* for further development of linear algebra. Below, we will define the new geometry of the abstract space $\mathbf{R}^n$ and demonstrate that these "axioms" are still satisfied.

Exercise 4.3.9

The distance is a function. Explain.

We know the last property from Euclidean geometry:

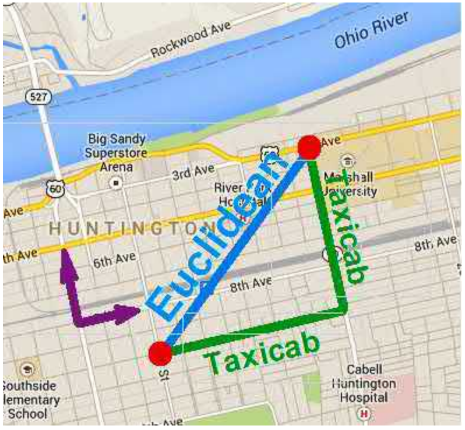




We can justify it for dimension $n \geq 4$ by referring to the following fact: Any three points lie within a single plane. This fact brings us back to Euclidean geometry... if that's what we want.

Example 4.3.10: city blocks

The Distance Formula for the plane gives us the distance measured *along a straight line* as if we are walking through a field. But what if we are walking through a city? We then cannot go diagonally as we have to follow the grid of streets. This fact dictates how we measure distances. To find the distance between two locations $P = (x, y)$ and $Q = (u, v)$, we measure *along the grid* only:



The formula is, therefore:

$$\boxed{d_T(P,Q) = |x-u| + |y-v|}$$

It is called the *taxicab metric*. It is different from the Euclidean metric as the diagonal of a 1 square is 2 units long under this geometry.

Exercise 4.3.11

Prove that the taxicab metric satisfies the three properties in the theorem.

If the physical space can be reasonably treated with non-Euclidean distances, the idea is even more applicable to higher dimensions.

If we are given space of locations or "states", $\mathbf{R}^n$, it is our choice to pick an appropriate way to compute distances from coordinates:

Definition 4.3.12: three metrics

Suppose points P and Q in $\mathbf{R}^n$ have coordinates $(x_1, x_2, ..., x_n)$ and $(x_1', x_2', ..., x_n')$ respectively.

1. The *Euclidean metric*, or the L^2-metric, is defined to be

$$\boxed{d_2(P,Q)=\sqrt{\sum_{k=1}^{n}(x_k-x_k')^2}}$$

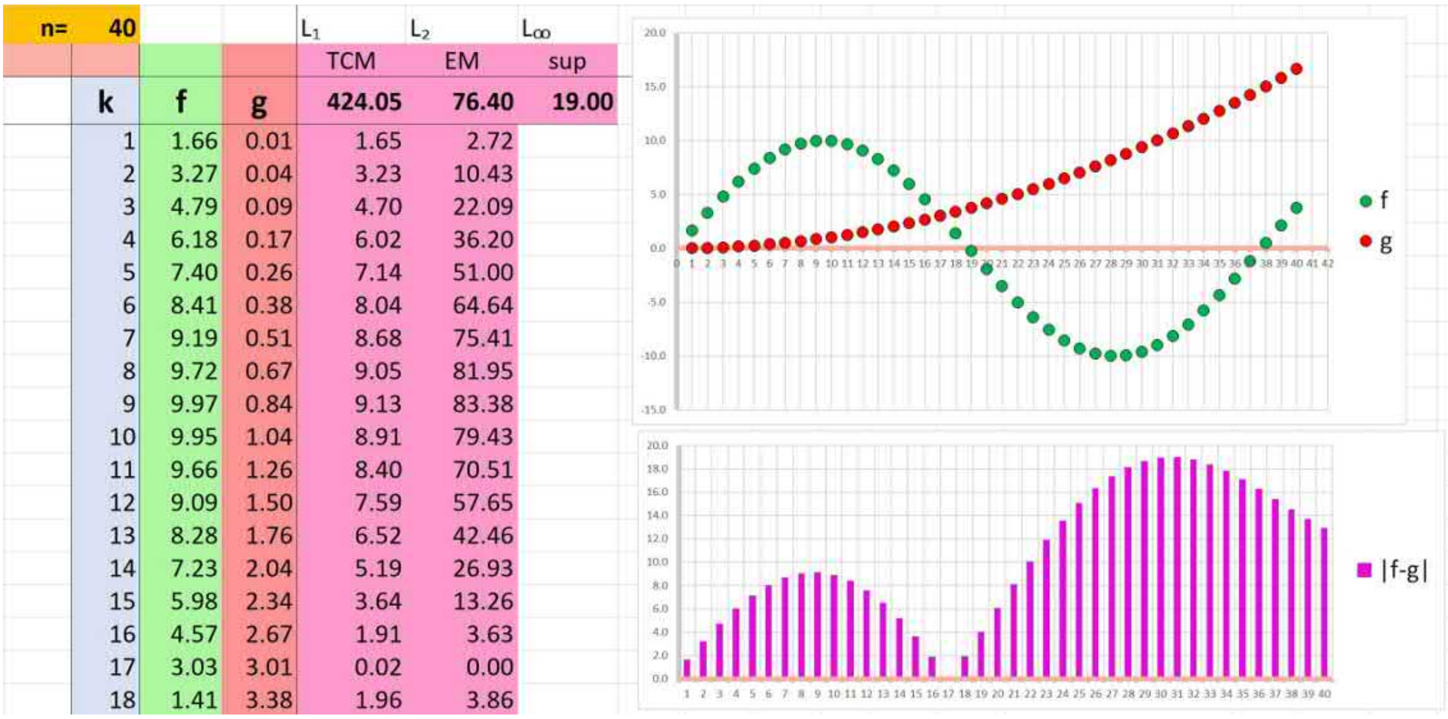
2. The *taxicab metric*, or the L^1-metric, is defined to be

$$\boxed{d_1(P,Q)=\sum_{k=1}^{n}|x_k-x_k'|}$$

3. The *max metric*, or the L^∞-metric, is defined to be

$$\boxed{d_\infty(P,Q)=\max_{k=1,\dots,n}|x_k-x_k'|}$$

They are illustrated below ($n=40$):



n=	40			L_1	L_2	L_∞
				TCM	EM	sup
	k	**f**	**g**	**424.05**	**76.40**	**19.00**
	1	1.66	0.01	1.65	2.72	
	2	3.27	0.04	3.23	10.43	
	3	4.79	0.09	4.70	22.09	
	4	6.18	0.17	6.02	36.20	
	5	7.40	0.26	7.14	51.00	
	6	8.41	0.38	8.04	64.64	
	7	9.19	0.51	8.68	75.41	
	8	9.72	0.67	9.05	81.95	
	9	9.97	0.84	9.13	83.38	
	10	9.95	1.04	8.91	79.43	
	11	9.66	1.26	8.40	70.51	
	12	9.09	1.50	7.59	57.65	
	13	8.28	1.76	6.52	42.46	
	14	7.23	2.04	5.19	26.93	
	15	5.98	2.34	3.64	13.26	
	16	4.57	2.67	1.91	3.63	
	17	3.03	3.01	0.02	0.00	
	18	1.41	3.38	1.96	3.86	

For the Euclidean metric, we compute for each row:

`=ABS(RC[-2]-RC[-1])`

We plot it, then apply this formula:

`=SUM(R[1]C:R[40]C)`

For the taxicab metric, we compute for each row:

`=RC[-1]^2`

We then apply this formula:

`=SQRT(SUM(R[1]C:R[40]C))`

The formula for max metric is simply:

`=MAX(R[1]C[-2]:R[40]C[-2])`

Exercise 4.3.13

Prove the Axioms of Metric Space for these formulas.

When we deal with the "physical space" ($n = 1, 2, 3$) as in the above theorems, the Euclidean metric is implied. For the "abstract spaces" ($n = 1, 2, ...$), the Euclidean metric is the default choice; however, there are many examples when the Euclidean geometry and, therefore, the Euclidean metric (aka the Distance Formula) don't apply.

Example 4.3.14: attributes

Let's consider the prices of wheat and sugar again. The space of prices is the same, $\mathbf{R}^2$. However, measuring the distance between two combinations of prices with the Euclidean metric leads to undesirable effects. For example, such a trivial step as changing the latter from "per ton" to "per kilogram" will change the geometry of the whole space. It is as if the space is stretched vertically. As a result, in particular, point P that used to be closer to point A than to B might now satisfy the opposite condition.

Furthermore, the two (or more) measurements or other attributes might have nothing to do with each other. In some obvious cases, they will even have different *units*. For example, we might compare two persons *built* based to the two main measurements: weight and height. Unfortunately, if we substitute such numbers into our formula, we will be adding pounds to feet!

Some of the concepts of geometry find their analogs in higher dimensions. For example, consider:

- A *circle* on the plane is defined to be the set of all points a given distance away from its center.
- A *sphere* in the space is defined to be the set of all points a given distance away from its center.

What about higher dimensions? The pattern is clear:

- A *hypersphere* in $\mathbf{R}^n$ is defined to be the set of all points a given distance away from its center.

In other words, each point P on the hypersphere satisfies:

$$d(P, Q) = R\,,$$

where Q is its center and R is its radius.

Example 4.3.15: Newton's Law of Gravity

According to the law, the force of gravity between two objects is

- proportional to either of their masses,
- inversely proportional to the square of the distance between their centers.

In other words, the force is given by the formula:

$$F = G\frac{mM}{r^2}\,,$$

where:

- F is the force between the objects;
- G is the gravitational constant;
- m is the mass of the first object;
- M is the mass of the second object;
- r is the distance between the centers of the mass of the two.

The dependence of F on m and M is very simple and, furthermore, we can assume that the masses of planets are remain the same. We are left with a function of one variable:

$$F(r) = G\frac{mM}{r^2}\,.$$

More precisely, r depends on the *location* P of the second object in the 3-dimensional space:

$$r = d(O, P)\,,$$

if, for simplicity, we assume that the first object is located at the origin. Note that this force is constant along any of the spheres centered at O.

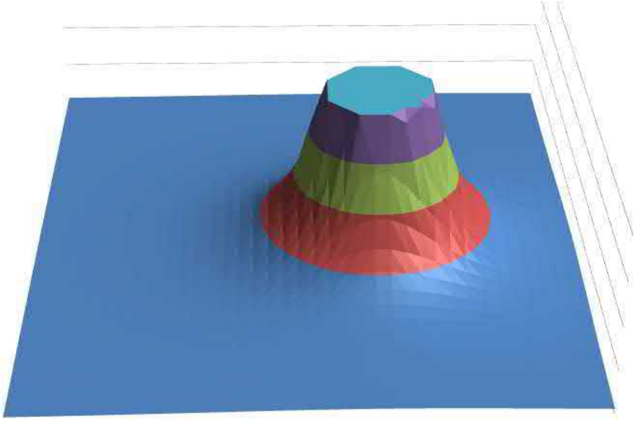
Now, we can re-write the law as a function of P:

$$\boxed{F(P)=\frac{GmM}{d(O,P)^2}}$$

Furthermore, if we suppose that the three spatial variables x, y, z are the coordinates of P, we can re-write the law as a function of three variables:

$$F(x,y,z)=\frac{GmM}{d(O,P)^2}=\frac{GmM}{\left(\sqrt{x^2+y^2+z^2}\right)^2}=\frac{GmM}{x^2+y^2+z^2}\,.$$

If we ignore the third variable ($z=0$), we can plot the graph of the resulting function of two variables:



But what about the *direction* of this force? This question is addressed in the next section.

Exercise 4.3.16

Visualize the function for the case of 3 dimensions.

4.4. *Where vectors come from*

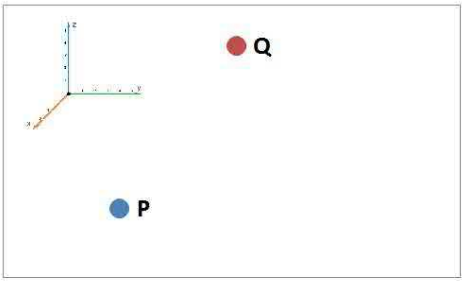
We introduced vectors in previously to properly handle the geometric issue of *directions* and angles between directions. However, vectors also appear frequently in our study of the natural world.

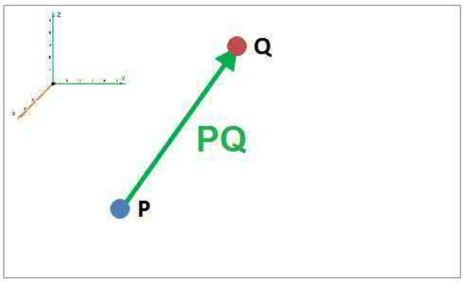
Definition 4.4.1: displacement

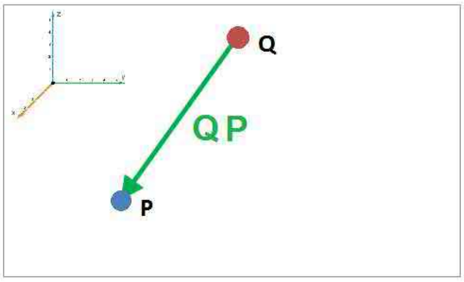
When the points in $\mathbf{R}^n$ are called locations or positions, the vectors are called *displacements*. In particular, if P and Q are two locations, then the vector PQ is the *displacement* from P to Q.

The idea applies to any space $\mathbf{R}^n$ but we will start with the physical space devoid of a Cartesian system.

From this point of view, a *vector* is a pair, PQ, of locations P and Q.







Warning!

"Vector" is not synonymous with "segment"; it's not even a set. The segment that you see is just a visualization.

We saw vectors in action previously, but the goal was limited to using vectors to understand directions and angles between them. Our interest here is the *algebraic operations* on vectors.

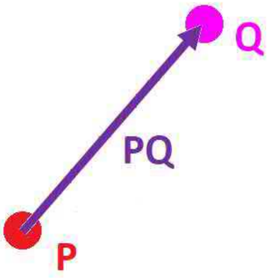
If a vector is an *ordered* pair, this means that $PQ \neq QP$. But is there a relation? The displacement from P to Q is the opposite to the displacement from Q to P:

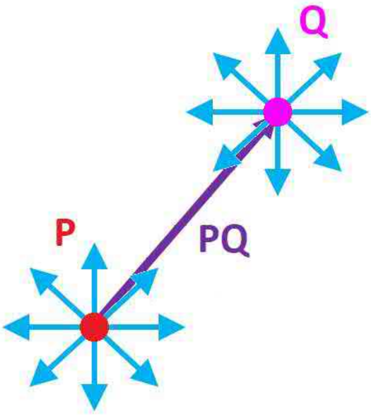
$$\boxed{QP = -PQ}$$

The locations and displacements and, therefore, points and vectors are subject to algebraic operations that connect them:

$$\boxed{P + PQ = Q}$$

As you can see, we add a vector to a point that is its initial point and the result is its terminal point.





It follows:

$$PQ = Q - P .$$

As you can see, the vector is the difference of its terminal and its initial points. It follows that

$$QP = P - Q = -(Q - P) = -PQ .$$

We are back to the above formula.

Definition 4.4.2: affine space

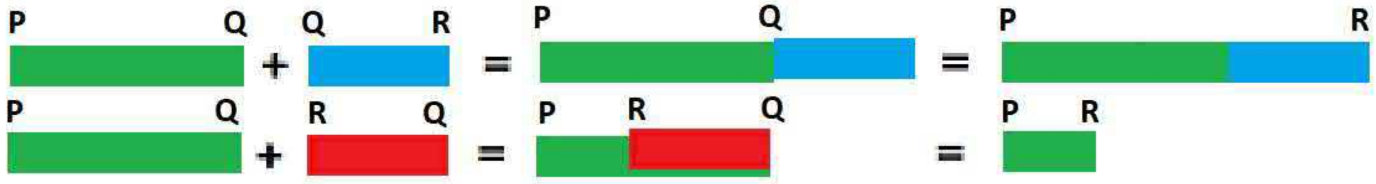
The *affine space* of the Euclidean space $\mathbf{R}^n$ is the set of ordered pairs PQ of points P and Q in this space. These pairs are called *vectors*.

We now review the algebra.

First, dimension 1.

Even though the algebra of vectors is the algebra of real numbers, we can still, even without a Cartesian system, think of the algebra of *directed segments*.

The *addition* of two vectors is executed by attaching the head of the second vector to the tail of the first, as illustrated below:

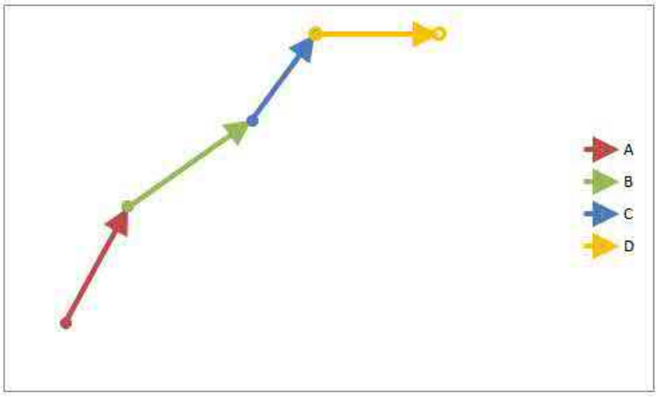


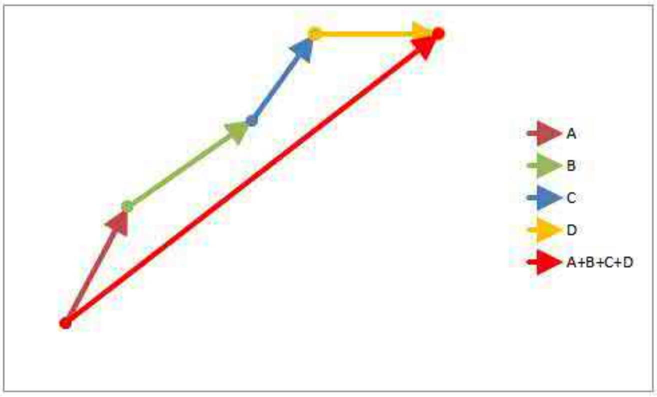
The negative number (red) is a segment directed backwards so that its tail is on its left.

Now dimension 2.

Example 4.4.3: consecutive displacements

We move point to point through the space:





This is how we can understand addition of vectors as displacements:

initial location	displacement	terminal location		
P	PQ	$P+PQ$	$=Q$	
Q	QR	$Q+QR$	$=R$	$=P+(PQ+QR)$
R	RS	$R+RS$	$=S$	$=P+(PQ+QR+RS)$
S	ST	$S+ST$	$=T$	$=P+(PQ+QR+RS+ST)$

The right column shows how adding vector to point can be seen as an alternative approach: We add vector to vector first.

For the general case of m steps, we have these two representations:

$$\begin{aligned}\sum_{k=0}^{m} X_kX_{k+1} &= X_0X_1+\ldots && +X_{m-1}X_m && = X_0X_m\\ \sum_{k=0}^{m}(X_{k+1}-X_k) &= (X_1-X_0)+\ldots && +(X_m-X_{m-1}) && = X_m-X_0\end{aligned}$$

Since moving from P to Q and then from Q to R amounts to moving from P to R, the construction is, again, a "head-to-tail" alignment of vectors:

$$PR=PQ+QR\,.$$

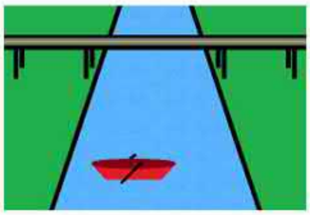
Warning!

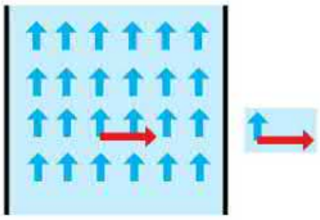
We use the same symbol "+" as for addition of numbers.

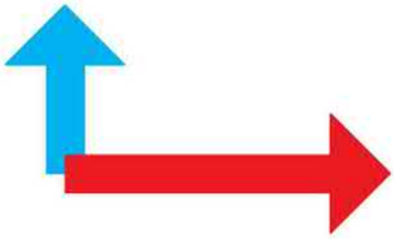
However, in the physical world, there are other "metaphors" for vectors besides the displacements.

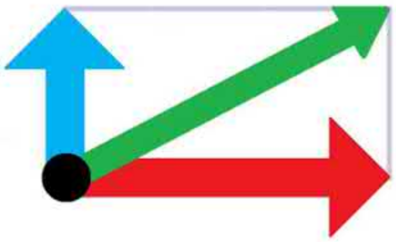
Example 4.4.4: velocity of stream

We look at the velocities of particles in a stream at each location. Then they may be combined with the speed of rowing of the boat:









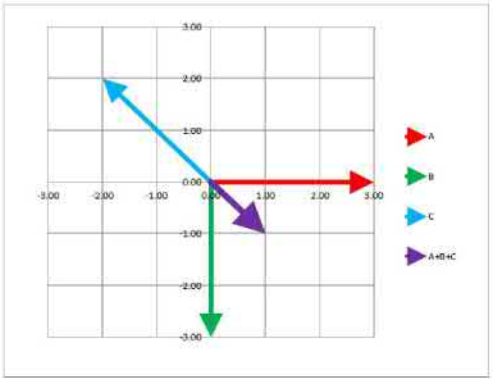
We are to add these two vectors at each location, but they, in contrast to the displacements, start at the *same* point!

Exercise 4.4.5

With the velocities as shown, what is the best strategy to cross the canal?

Example 4.4.6: forces

Let's also look at *forces as vectors*. For example, springs attached to an object will pull it in their respective directions:



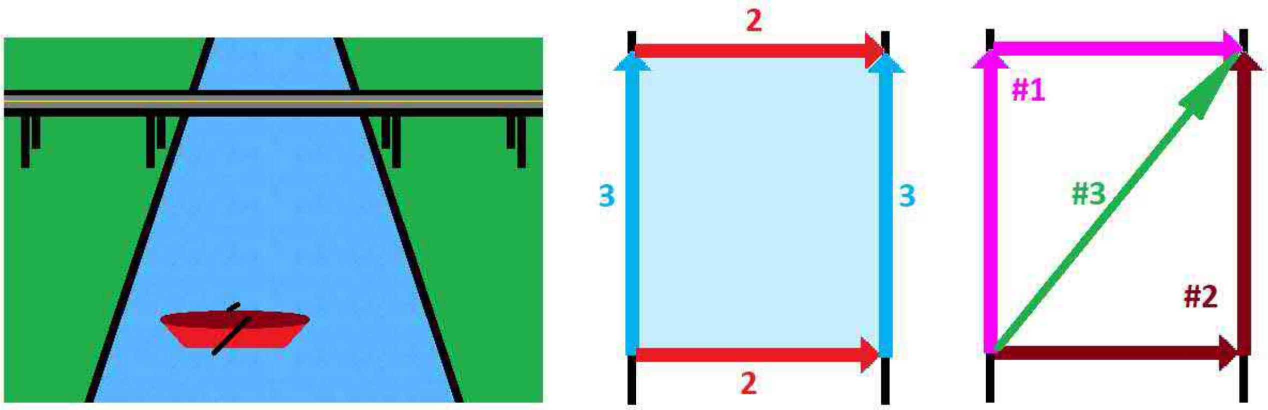
We add these vectors to find the combined force as if produced by a single spring. The forces are vectors that start at the same location.

Example 4.4.7: displacements

We can interpret the displacements, too, as vectors aligned to their starting points. Imagine we are crossing a river 3 miles wide and we know that the current (with no rowing) takes us 2 miles downstream. Three different ways this can happen:

1. a free-flow trip 3 miles north followed by a walk 2 miles east over a bridge; or
2. a walk 2 miles east over another bridge followed by a free-flow trip 3 miles north; but also
3. a rowing trip along the diagonal of a rectangle with one side going 3 miles north and another 2 miles east.

The three outcomes are the same:



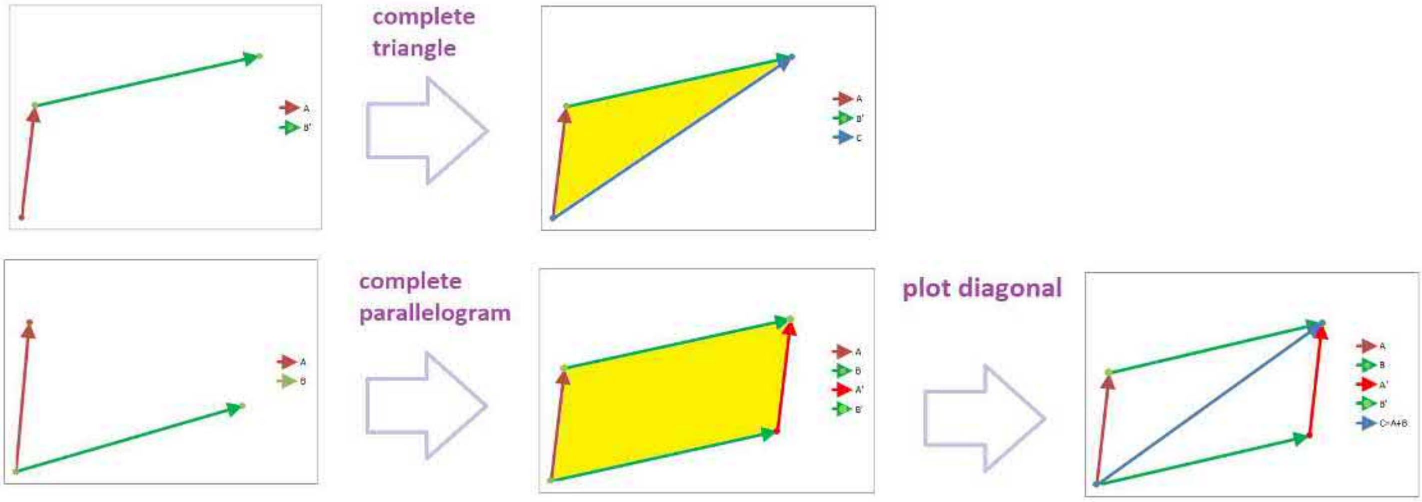
The sum of these two vectors is the same in any order:

3 miles north and 2 miles east

So, to add two vectors, we follow either

1. The head-to-tail: the triangle construction.
2. The tail-with-tail: the parallelogram construction.

They have to produce the same result! They do, as illustrated below:



1. For the former, we make a *copy* B' of B, attach it to the end of A, and then create a new vector with the initial point that of A and terminal point that of B'.
2. For the latter, we make a copy B' of B, attach it to the end of A, also make a copy A' of A, attach it to the end of B. Then the *sum* $A+B$ *of two vectors* A *and* B with the same initial point is the vector with the same initial point that is the diagonal of the parallelogram with sides A and B.

Exercise 4.4.8

Prove that the result is the same according to what we know from Euclidean geometry.

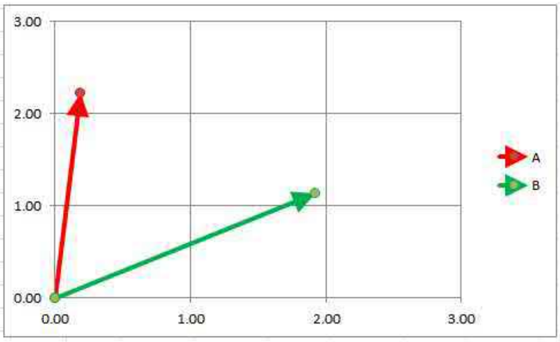
It is the same construction.

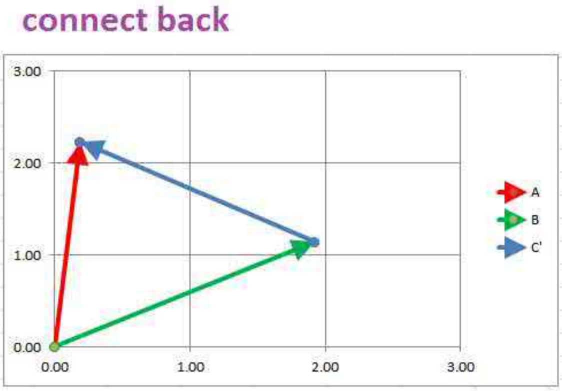
We think about vectors as line segments in a Euclidean space. As such, it has a direction and the length. It is possible to have the length to be 0; that's the *zero vector*. Its direction is undefined.

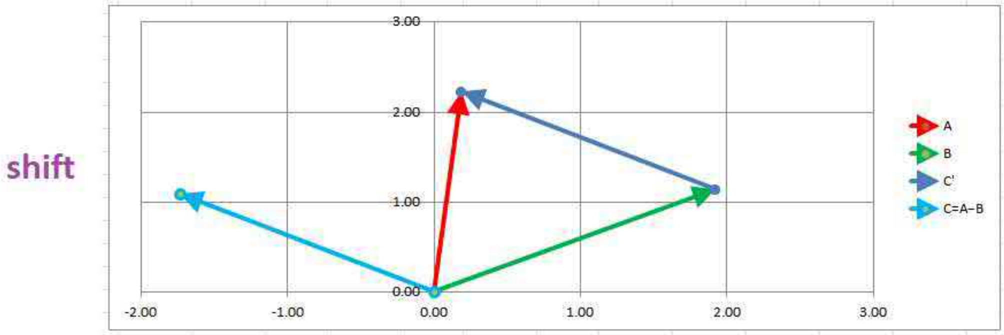
Once we know addition, *subtraction* is its inverse operation. Indeed, given vectors A and B, finding the vector C such that $B+C=A$ amounts to solving an equation, just as with numbers:

$$A-B=C \implies B+C=A\,.$$

In other words, what do I add to B to get A? An examination reveals the answer:



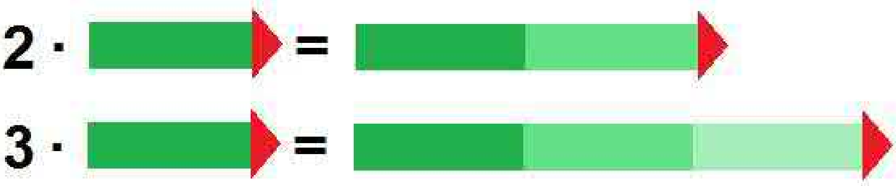




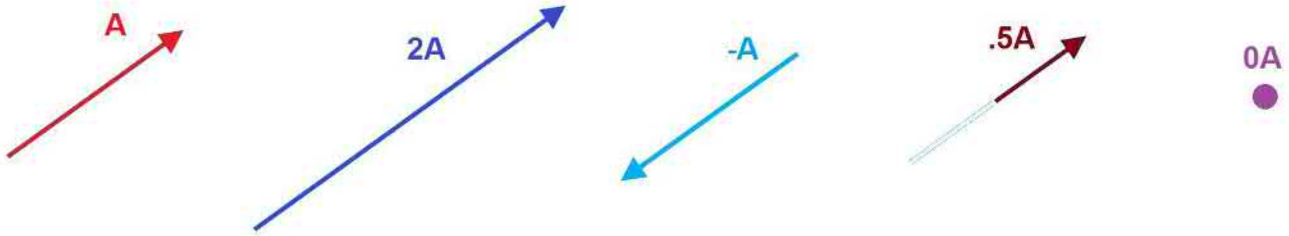
So, we construct the vector from the end of B to the end of A. One more step: make a copy of C with the same starting point as A.

If we want to go faster, we row twice as hard; the vector has to be stretched! Or, one can attach two springs in a consecutive manner to double the force, or cut any portion of the spring to reduce the force proportionally. A force might keep its direction but change its magnitude! It might also change the direction to the opposite.

There is then another algebraic operation on vectors. This is dimension 1:



This is dimension 2:



As you can see, every point has a special vector attached to it. For every point P, the *zero vector* is:

$$0 = PP\,.$$

We say that 0 serve as the *identity*.

Thus, the scalar product $c \cdot A$ of a vector A and a real number c is the vector with the same initial point as A, with the direction which is

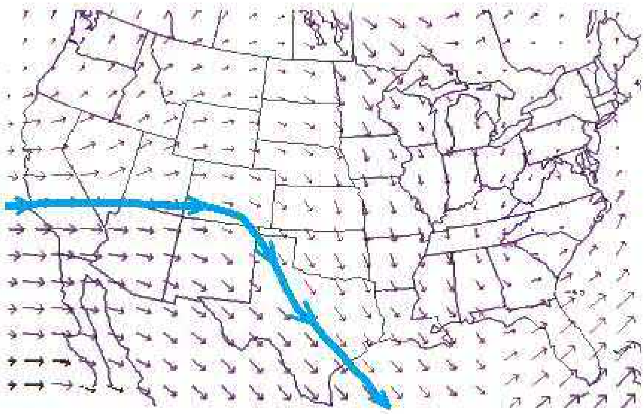
- same as that of A when $c > 0$,
- opposite to that of A when $c < 0$, and
- zero when $c = 0$.

Warning!

We use the same symbol "·" as for multiplication of numbers.

Example 4.4.9: velocity of wind

Velocities appear as the wind speed at different locations:



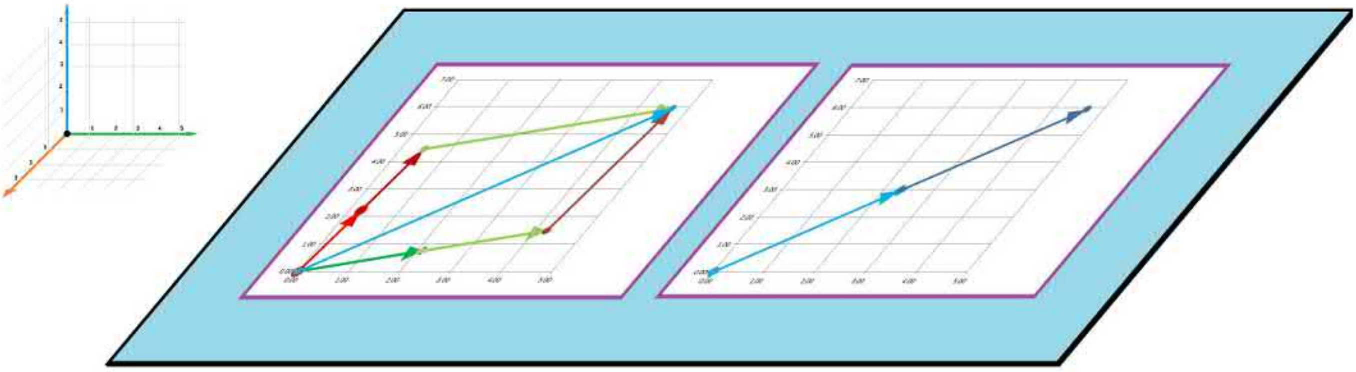
If the velocities are combined with the time increments, we can find the displacements of the particles of the air. We can also plot a whole trip of one such particle, just as in the beginning of the section. With the velocities denoted by $V_k,\ k = 1, 2, ..., m$, respectively, takes the form:

$$\sum_{k=0}^{m} V_k \cdot \Delta t = X_m - X_0 \,.$$

Exercise 4.4.10

Plot a few more paths.

What is the dimension of the space? As we know from Euclidean geometry, two lines and, therefore, two vectors, determine a *plane*. This is why we imagine that the operations, as we have defined them, are limited to a certain plane within a possibly higher-dimensional space:



Example 4.4.11: units

Out of caution, we should look at the units of the scalar. Yes, the force is a multiple of the acceleration:

$$F = ma \,.$$

However, these two have different units and, therefore, cannot be added together!

Also, the displacement is the time multiplied by the velocity:

$$\Delta X = \Delta t \cdot V \,.$$

But these two have different units and, therefore, cannot be added! They live in two different spaces.

We will continue, throughout the chapter, to use *capitalization* to help to tell vectors from numbers.

Warning!

To indicate vectors, many sources use:

- an arrow above the letter, $\vec{v}$, or
- the bold face, $\mathbf{v}$.

In this section, we introduced to $\mathbf{R}^n$, a space of points, a new entity – a vector. It is an ordered pair PQ of points, P and Q, linked back to points by this algebra:

$$\boxed{PQ = Q - P \quad \text{or} \quad Q = P + PQ}$$

The vectors can be added together and multiplied by a number according to the procedures described above.

The operations satisfy the following properties:

Theorem 4.4.12: Axioms of Affine Space

The points and the vectors in the affine space of $\mathbf{R}^n$ *satisfy the following properties:*

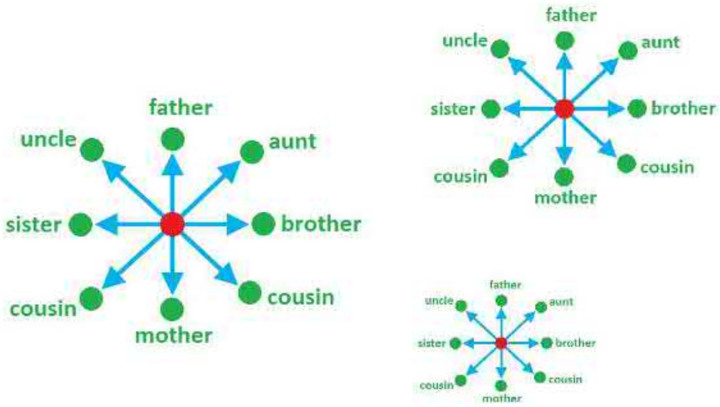
1. ***Identity***: *For every point* P, *we have for some vector denoted by* 0 *the following:*
$$P + 0 = P\,.$$
2. ***Associativity***: *For every point* P *and any vectors* V *and* W *starting at* P, *we have:*
$$(P + V) + W = P + (V + W)\,.$$
3. ***Free and transitive action***: *For every point* P *and every point* Q, *there is a vector* V *such that*
$$P + V = Q\,.$$

Proof.

1. Choose:
$$0 = PP\,.$$
2. Compare:
$$(P + PQ) + QR = R \quad \text{and} \quad P + (PQ + QR) = P + PR = R\,.$$
3. Choose:
$$V = PQ\,.$$

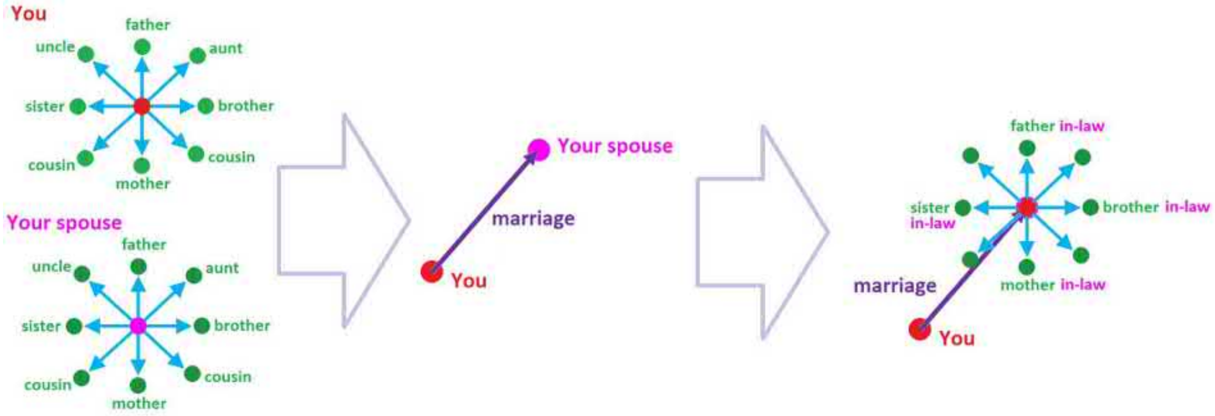
Example 4.4.13: family relations

Let's imagine that every point in the space stands for a person. Now, each person is linked by a vector to one's family:



Every person is the center of such a system of links. Also, potentially, each person is linked to any other person.

Now, a *marriage* will link one such center to another, and renaming commences:



The word "affine" means "in-law".

Let's simplify.

This is what we have in particular:

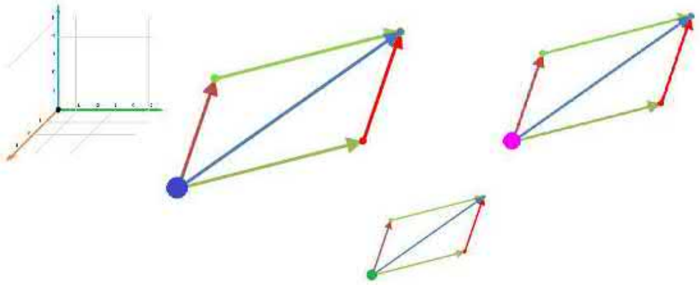
1. Two vectors with the same origin can be added together, producing another one with the same origin.
2. A vector can be multiplied by a number, producing another one with the same origin.

These operations are carried out according to the procedures described above. Given points P and Q and a real number k, we have for some points R and S:

$$\boxed{PQ + PR = PR \quad \text{and} \quad k \cdot PQ = PS}$$

Indeed, the result is another vector with the same initial point as the original(s)!

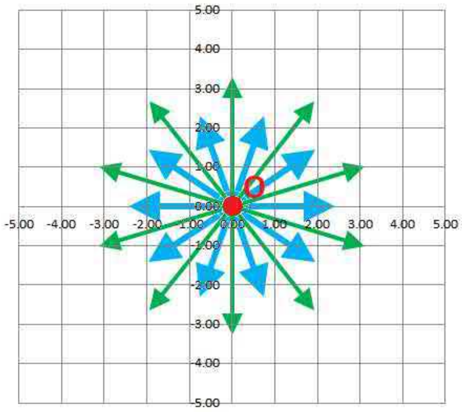
This algebra of vectors (including the scalar multiplication) can, therefore, be carried out separately at every location:



And this algebra is identical for every location! Therefore, a single initial point will be sufficient for our study of vector algebra.

But the choice of the initial point is crucial. This point will be assumed to be O, the origin, unless otherwise indicated.

This is what our collection of vectors looks like, for dimension 2:



Such a "space of vectors" is called a *vector space*. It is equipped with two operations, the vector addition:

$$\boxed{\text{vector} \ + \ \text{vector} \ = \ \text{vector}}$$

and the scalar multiplication:

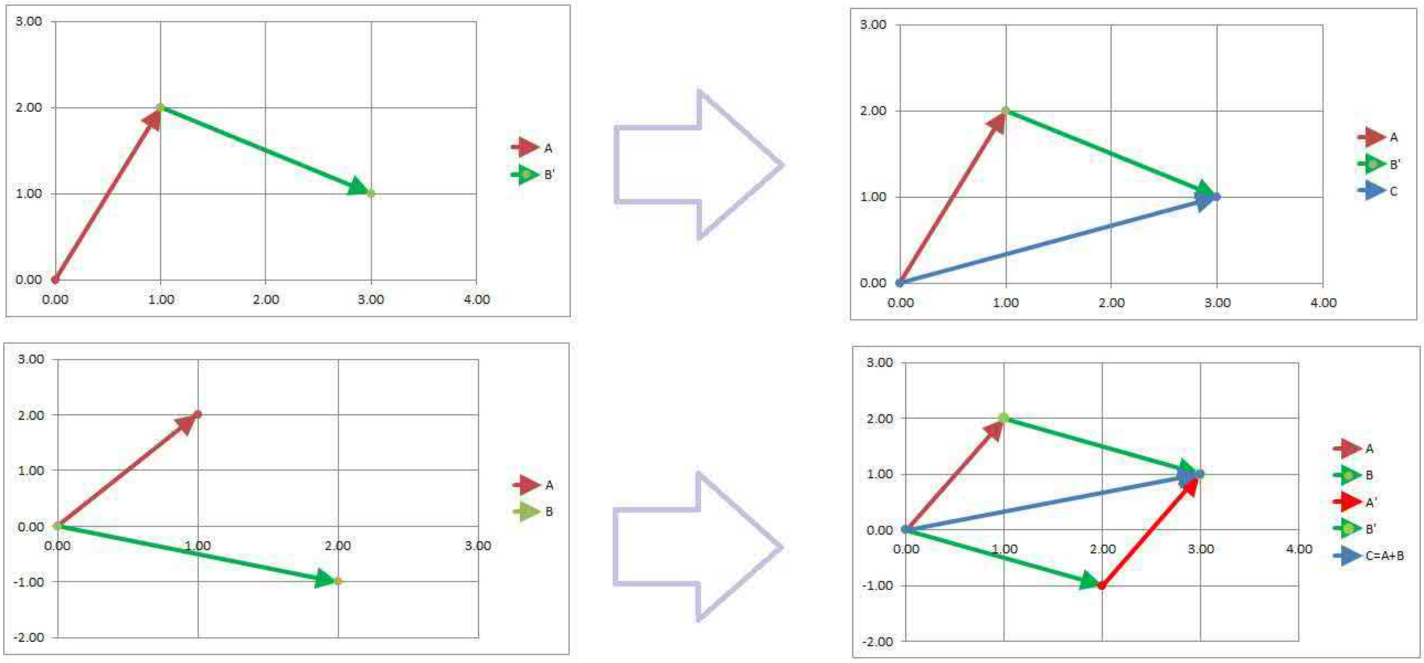
$$\boxed{\text{number} \ \cdot \ \text{vector} \ = \ \text{vector}}$$

Warning!

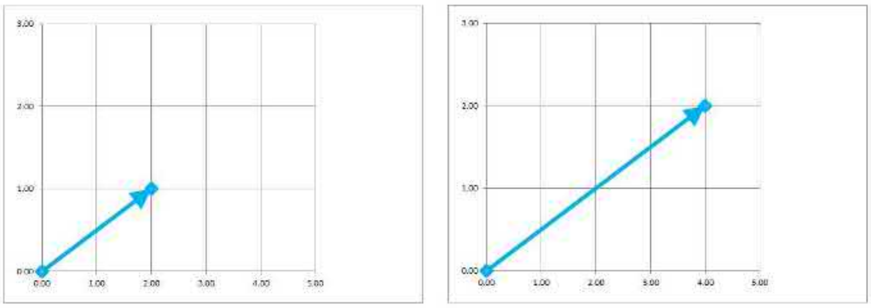
There are no points in a vector space.

Let's add a Cartesian system to our plane (with the origin already chosen).

We can watch what happens to the coordinate of the end-points of the vectors as we carry out our algebra. Sum implies coordinate-wise addition:



Scalar product implies multiplication of the coordinates by the same number:



We are in $\mathbf{R}^2$ now.

4.5. Vectors in $\mathbf{R}^n$

We now understand vectors. Or at least we understand them in the lower-dimensional setting.

Now, we move to the next stage:

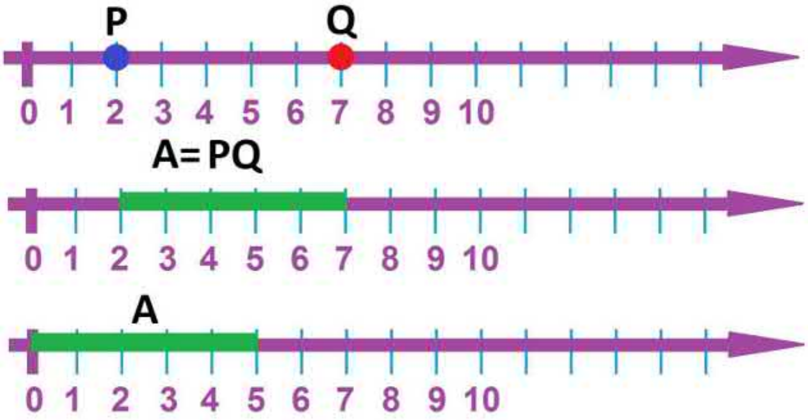
1. We add a Cartesian system to these spaces: line, plane, and space.
2. We also consider the abstract spaces of arbitrary dimensions, $\mathbf{R}^n$.

We will use the former to make sure that the approach to the latter makes sense.

A vector is still a pair PQ of points P and Q in $\mathbf{R}^n$. Now, either of these two points corresponds to a string of n numbers called its coordinates.

On the line $\mathbf{R}^1$, points are numbers and the vectors are simply differences of these numbers:

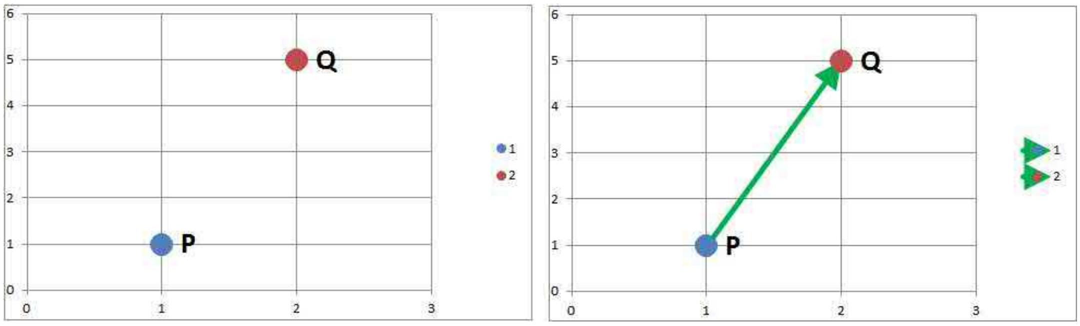
$$PQ = Q - P\,.$$



On the plane $\mathbf{R}^2$, we might have:

$$P = (1,2) \text{ and } Q = (2,5)\,.$$

How can we express vector PQ in terms of these four numbers?



We look at the *change* from P to Q:

- The change with respect to x, which is $2-1=1$.
- The change with respect to y, which is $5-2=3$.

We combine these into a new *pair of numbers* (with triangular brackets to distinguish these from points):

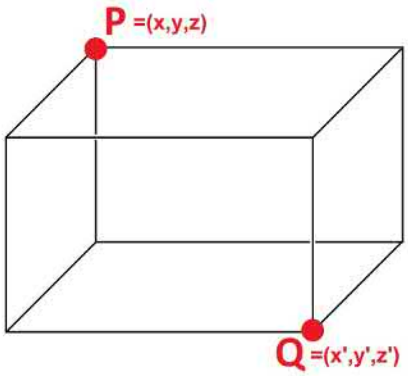
$$PQ=<1,3>\ .$$

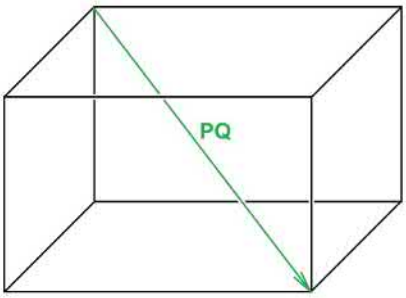
Technically, however, we have to mention the initial point $P=(1,2)$ of the vector.

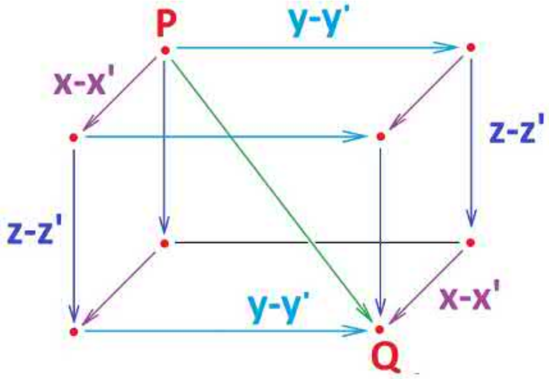
Now in $\mathbf{R}^3$, we might have:

$$P=(x,y,z) \text{ and } Q=(x',y',z')\,.$$

How can we express vector PQ in terms of these six numbers?







There are three changes (differences) along the three axes, i.e., a *triple*:

$$PQ=<x'-x,y'-y,z'-z>\ .$$

Definition 4.5.1: vector and its components

A *vector* PQ in $\mathbf{R}^n$ with its initial point

$$P=(x_1,x_3,...,x_n)$$

and its terminal point

$$Q=(x_1',x_2',...,x_n')$$

is given by the string of n numbers called the *components* of the vector:

$$x_1'-x_1,\ x_2'-x_3,\ ...,\ x_n'-x_n\,.$$

The definition matches the one that relies on directed segments.

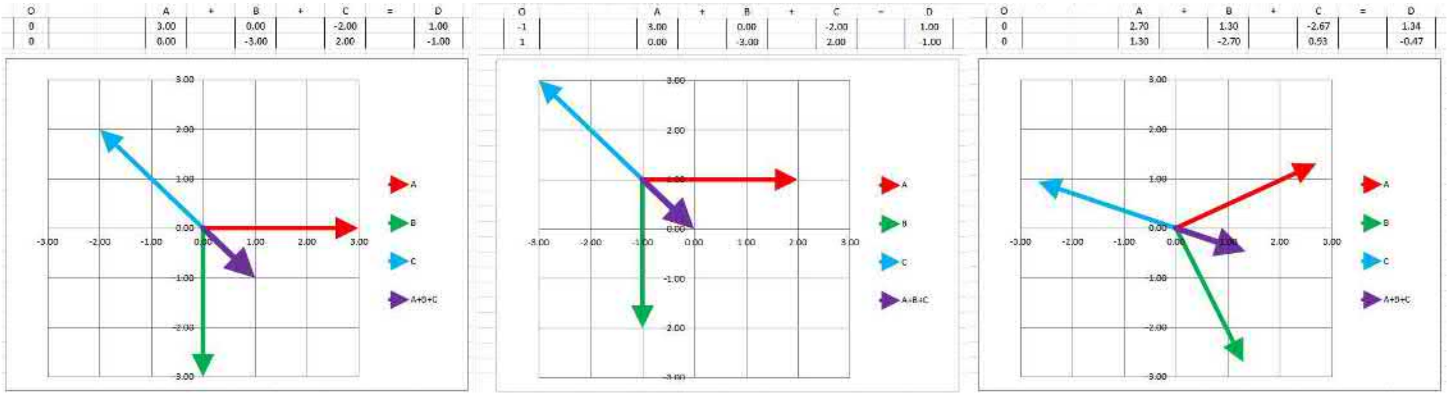
A vector may emerge from its initial and terminal points or independently. In either case, we assemble the components according to the following notation.

Row and column vectors

$$\left[\begin{array}{c} a_1\\ a_2\\ ...\\ a_n \end{array}\right] =<a_1,a_2,...,a_n>$$

The former is preferred; the latter is its abbreviation.

Once again we can only carry out vector addition on vectors with the same initial point. What happens if we change the initial point while leaving the components of the vectors intact? Not only is each vector "copied" but so are the results of all algebraic operations. They are the same, just shifted to a new location:



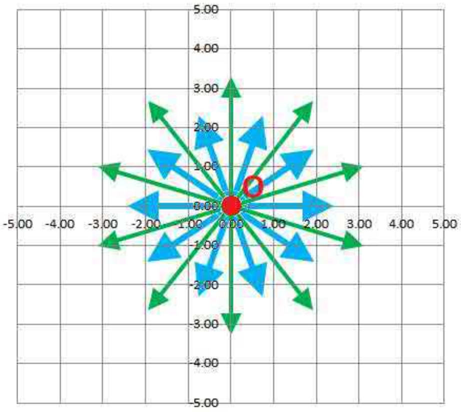
O	A	+	B	+	C	=	D
0	3.00		0.00		-2.00		1.00
0	0.00		-3.00		2.00		-1.00

O	A	+	B	+	C	=	D
-1	3.00		0.00		-2.00		1.00
1	0.00		-3.00		2.00		-1.00

O	A	+	B	+	C	=	D
0	2.70		1.30		-2.67		1.34
0	1.30		-2.70		0.53		-0.47

We can even rotate the coordinate system (right).

It is then sufficient to provide results for the vectors that start at the *origin* O only! Only these are allowed:



In that case, *the components of a vectors are simply the coordinates of its end*:

$$\boxed{\begin{array}{rlllll} P & = (x_1, & x_3, & \dots, & x_n) & \Longrightarrow \\ OP & =< x_1, & x_2, & \dots, & x_n > & \end{array}}$$

Warning!

The difference between points and vectors lies in the algebraic operations to which they are subject.

Next we consider the familiar algebraic operations but this time the vectors are represented by their components.

We carry out operations *componentwise*.

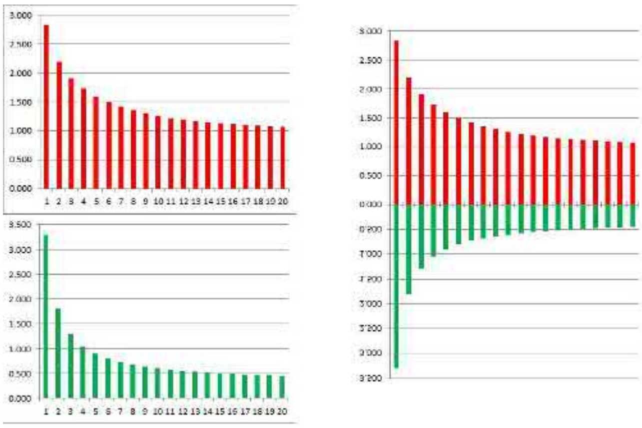
We demonstrate these operations for dimension $n = 3$ and for both the row and the column styles of notation. The vector addition is defined by:

$$\begin{array}{rlllll} A & =< & x, & y, & z & > \\ + & & & & & \\ B & =< & u, & v, & w & > \\ \hline A+B & =< & x+u, & y+v, & z+w & > \end{array} \qquad \begin{bmatrix} x \\ y \\ z \end{bmatrix} + \begin{bmatrix} u \\ v \\ w \end{bmatrix} = \begin{bmatrix} x+u \\ y+v \\ z+w \end{bmatrix}$$

This is how we progress to the definition that doesn't involve points:

$$\boxed{\begin{array}{ll} P=(x_1,...,x_n) & OP\\ Q=(y_1,...,y_n) & OQ\\ \hline R=(x_1+y_1,...,x_n+y_n) & OP+OQ=OR \end{array}} \to \boxed{\begin{array}{ll} U & =<x_1,...,x_n>\\ V & =<y_1,...,y_n>\\ \hline U+V & =<x_1+y_1,...,x_n+y_n> \end{array}}$$

A visualization of this operation in $\mathbf{R}^{20}$ is below:



The scalar multiplication is defined by:

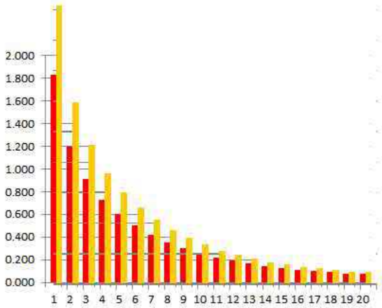
$$\begin{array}{cccccc} A & =< & x, & y, & z & > \\ \times & & & & & \\ k & & & & & \\ \hline kA & =< & kx, & ky, & kz & > \end{array} \qquad k\cdot\begin{bmatrix} x\\ y\\ z\end{bmatrix}=\begin{bmatrix} kx\\ ky\\ kz\end{bmatrix}$$

In either case, the components are aligned. Even though both seem equally convenient, the former will be seen as an abbreviation of the latter.

This is how we progress to the definition that doesn't involve points:

$$\boxed{\begin{array}{ll} k \text{ real} & \\ P=(x_1,...,x_n) & OP\\ \hline R=(kx_1,...,kx_n) & k\cdot OP=OR \end{array}} \to \boxed{\begin{array}{ll} k \text{ real} & \\ U & =<x_1,...,x_n>\\ \hline k\cdot U & =<kx_1,...,kx_n> \end{array}}$$

A visualization of this operation in $\mathbf{R}^{20}$ is below ($k=1.3$):

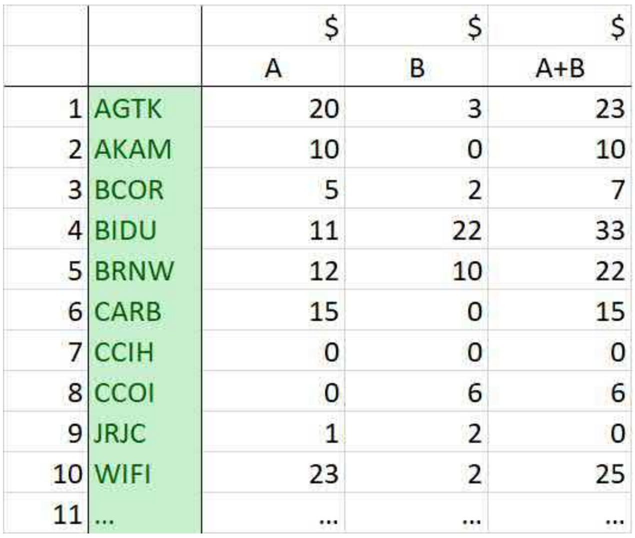


The scalar k is also known as the *constant multiple*.

Example 4.5.2: investment portfolios

If there are $10,000$ stocks on the stock market, every investment portfolio can be seen as a $10,000$-dimensional vector.

Then, *merging* two or more portfolios will *add* their vectors:

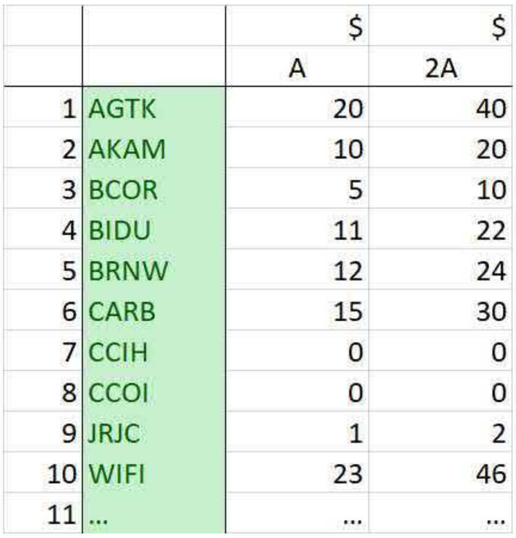


		$	$	$
		A	B	A+B
1	AGTK	20	3	23
2	AKAM	10	0	10
3	BCOR	5	2	7
4	BIDU	11	22	33
5	BRNW	12	10	22
6	CARB	15	0	15
7	CCIH	0	0	0
8	CCOI	0	6	6
9	JRJC	1	2	0
10	WIFI	23	2	25
11	...	...	...	...

We use the formula:

```
=RC[-2]+RC[-1]
```

Second, *doubling* or tripling a portfolio while preserving the proportion (or weight) of each stock will *scalar multiply* its vector:



		$	$
		A	2A
1	AGTK	20	40
2	AKAM	10	20
3	BCOR	5	10
4	BIDU	11	22
5	BRNW	12	24
6	CARB	15	30
7	CCIH	0	0
8	CCOI	0	0
9	JRJC	1	2
10	WIFI	23	46
11	...	...	...

We use the formula:

```
=2*RC[-1]
```

Even *non-homogeneous* holdings are subject to these operations:

$$< \text{10000 tons of wheat }, \text{20000 barrels of oil }, ... > ,$$

or

$$< \$100000,\ \yen 1000000,\ ... > \ .$$

Definition 4.5.3: sum of vectors

For two vectors in $\mathbf{R}^n$, their *sum* is defined to be the vector acquired by their componentwise addition:

$$\boxed{\begin{array}{lll} & A & =< a_1, ..., a_n > \\ & B & =< b_1, ..., b_n > \\ \Longrightarrow & A+B & =< a_1+b_1, ..., a_n+b_n > \end{array}}$$

Definition 4.5.4: scalar product

For a number k and a vector in $\mathbf{R}^n$, their *scalar product* is defined to be the vector acquired by their componentwise multiplication:

$$\boxed{\begin{array}{lll} & k & \text{real} \\ & A & =< a_1, ..., a_n > \\ \Longrightarrow & kA & =< ka_1, ..., ka_n > \end{array}}$$

We thus have the algebra of vectors for a space of any dimension!

These operations can be proven to satisfy the same properties as the vectors in $\mathbf{R}^3$ (next section). The proof is straight-forward and relies on the corresponding property of real numbers. For example, to prove the commutativity of vector addition, we use the commutativity of addition of numbers as follows:

$$A + B = \begin{bmatrix} x \\ y \\ z \end{bmatrix} + \begin{bmatrix} u \\ v \\ w \end{bmatrix} = \begin{bmatrix} x+u \\ y+v \\ z+w \end{bmatrix} = \begin{bmatrix} u+x \\ v+y \\ w+z \end{bmatrix} = \begin{bmatrix} u \\ v \\ w \end{bmatrix} + \begin{bmatrix} x \\ y \\ z \end{bmatrix} = B + A\,.$$

As a result, we are able to treat vectors as if they were numbers.

Furthermore, the *special* vectors deserve special attention:

Definition 4.5.5: zero vector

The *zero vector* in $\mathbf{R}^n$ has only zero components:

$$\boxed{0 =< 0, ..., 0 >}$$

Definition 4.5.6: the negative of vector

The *negative vector* of a vector A in $\mathbf{R}^n$ has its components the negatives of those of A:

$$\boxed{-A =< -a_1, ..., -a_n >}$$

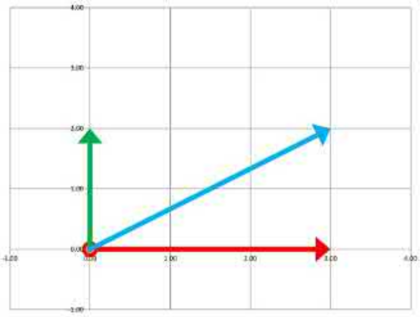
Exercise 4.5.7

Prove the eight axioms of vector spaces for the algebraic operations defined this way for (a) $\mathbf{R}^3$, (b) $\mathbf{R}^n$.

Let's explain the reason for the word "component".

A vector A is *decomposed* into the sum of other vectors. We chose those vectors to be special: Each is aligned with one of the axes. For example, we decompose:

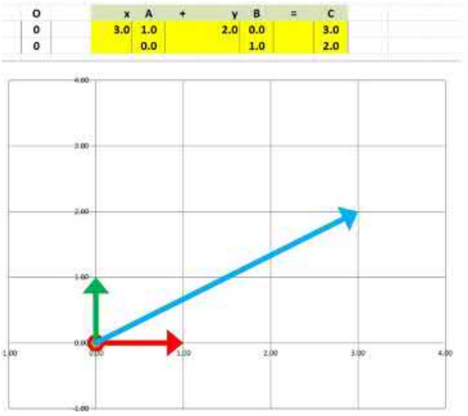
$$A =< 3, 2 > = < 3, 0 > + < 0, 2 > .$$



Then the two vectors are called the *component vectors* of A. We take this analysis one step further with scalar multiplication:

$$A =< 3, 2 > =< 3, 0 > + < 0, 2 >= 3 < 1, 0 > + 2 < 0, 1 > .$$

This is a *decomposition* of A:



O	x	A	+	y	B	=	C
0	3.0	1.0		2.0	0.0		3.0
0		0.0			1.0		2.0

Similarly, *any* vector can be represented in such a way:

$$< a, b >= a < 1, 0 > +b < 0, 1 > \ .$$

We use the following notation for these special vectors in $\mathbf{R}^2$:

Basis vectors in R^2

$$i =< 1, 0 >, \ j =< 0, 1 >$$

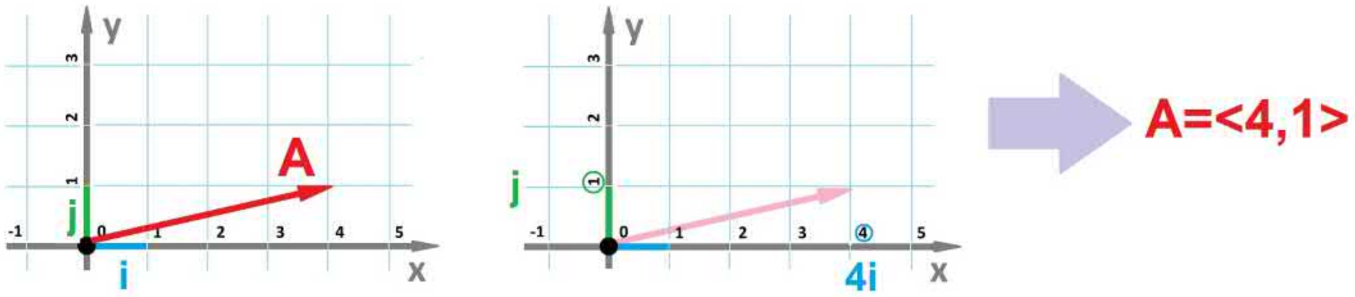
Then, any vector is as a *linear combination* of these two:

$$\boxed{< a, b >= ai + bj}$$

Example 4.5.8: decomposition

Below we present a new point of view. Before, we'd consider a point and its coordinates:

$$P = (4, 1) \, .$$



Now we write a vector and its components:

$$4i + j =< 4, 1 > \ .$$

Thus, representing a vector in terms of its components is just a way (a single way, in fact) to represent it in terms of a pair of specified unit vectors aligned with the axis.

We, furthermore, use the following notation for such vectors in $\mathbf{R}^3$:

Basis vectors in R^3

$$i =< 1,0,0 >,\ j =< 0,1,0 >,\ k =< 0,0,1 >$$

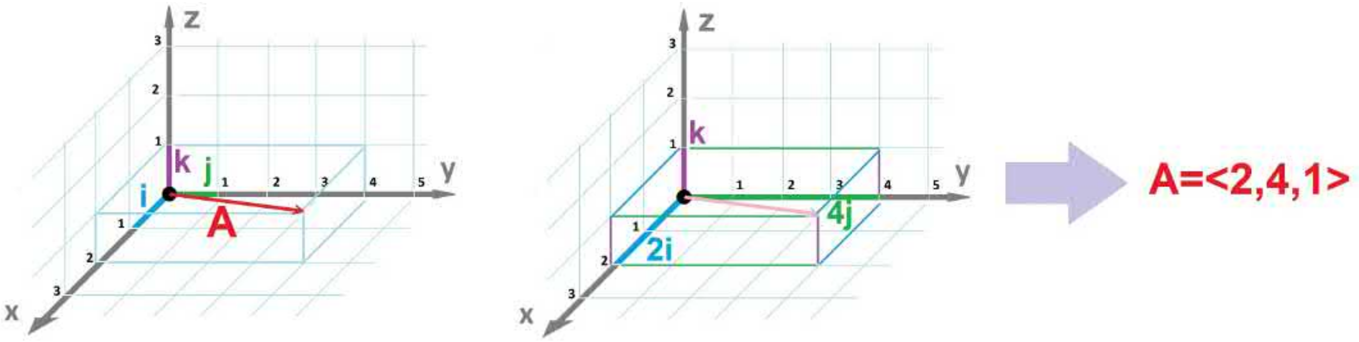
For every vector, we have the following representation:

$$\boxed{< a,b,c >= ai + bj + ck}$$

Example 4.5.9: decomposition

Now we write a vector and its components:

$$2i + 4j + k =< 2,4,1 > \ .$$



Definition 4.5.10: basis vectors

The *basis vectors* in $\mathbf{R}^n$ are defined and denoted by

$$\begin{array}{l} e_1 =< 1,0,0,...,0 > \\ e_2 =< 0,1,0,...,0 > \\ ... \\ e_n =< 0,0,0,...,1 > \end{array}$$

Together they form a *basis* of $\mathbf{R}^n$.

Then, any vector is as a *linear combination* of these n vectors:

$$\boxed{< x_1, x_2, ..., x_n >= x_1e_1 + x_2e_2 + ... + x_ne_n}$$

We have come to a new understanding:

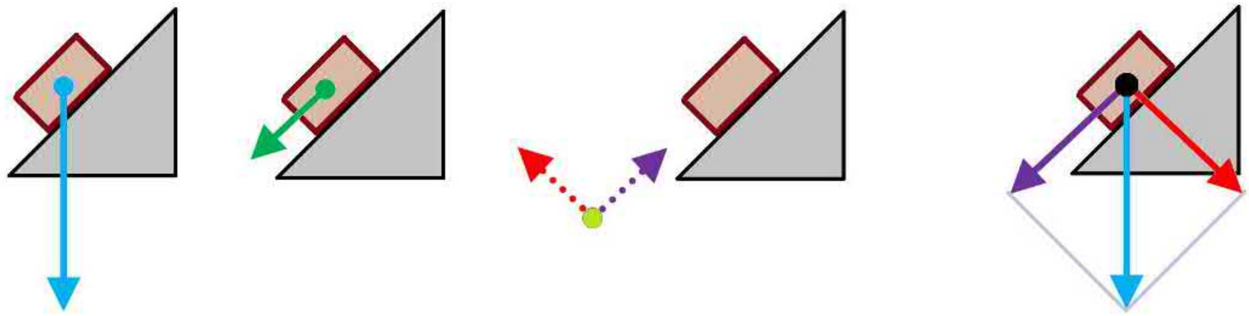
- old: Cartesian system = the axes
- new: Cartesian system = the origin and the basis vectors

Of course, we can choose a different Cartesian system by choosing a new set of basis vectors. The choice of the basis vectors is dictated by the problem to be solved.

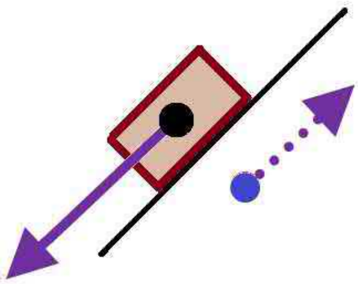
Example 4.5.11: compound motion

Suppose we are to study the motion of an object sliding down a slope.

Even though gravity is pulling it vertically down, the motion is restricted to the surface of the slope. It is then beneficial to choose the first basis vector i to be parallel to the surface and the second j perpendicular:



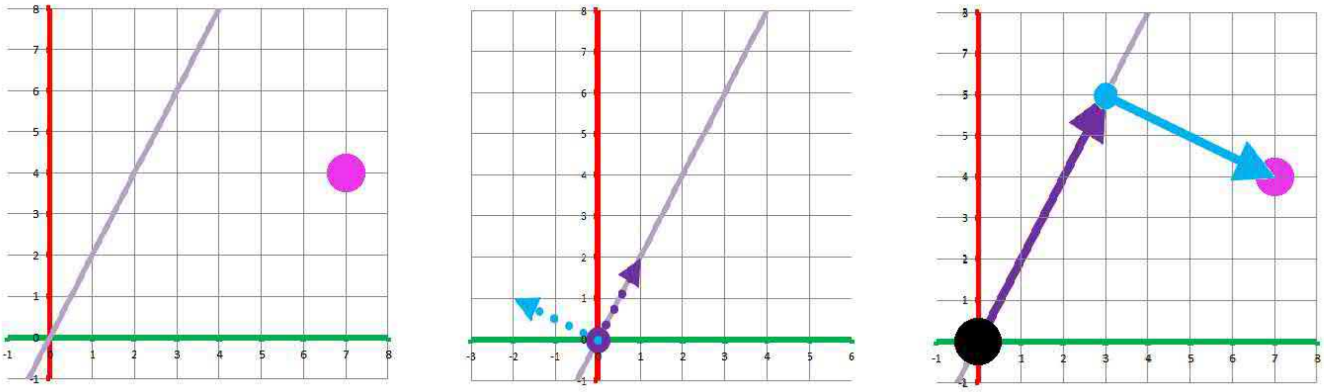
The gravity force is then decomposed into the sum of two vectors. It is the first one that affects the object and is to be analyzed in order to find the acceleration (Newton's Second Law). The second is perpendicular and is canceled by the resistance of the surface (Newton's Third Law). The problem becomes 1-dimensional:



Example 4.5.12: investing

Even when we deal with the abstract spaces $\mathbf{R}^n$, such decompositions may be useful.

For example, an investment advice might be to hold the proportion of stocks and bonds 1-to-2. We plot each possible portfolio as a point on the xy-plane, where x is the amount of stocks and y is the amount of bonds in it. Then the "ideal" portfolios lie on the line $y = 2x$. Furthermore, we would like to evaluate how well portfolios follow this advice. We choose the first basis vector to be $i =< 2, 1 >$ and the second perpendicular to it, $j =< -1, 2 >$:



Then the first coordinate – with respect to this new coordinate system – of your portfolio reflects how far you have followed the advice, and the second how much you've deviated from it. Now we just need to learn how to compute distances and angles in such a space.

4.6. Algebra of vectors

We will look for similarities with the algebra of numbers: the *laws of algebra*.

For example, we freely use the following shortcut when we deal with numbers:

$$x \cdot (y + z) = (x \cdot y) + (x \cdot z)\,.$$

Is there a similar rule for vectors? Yes, in a sense.

The complexity and the number of these laws will be higher because the participants are of two different types: numbers and vectors. They are also intermixed. For example, we may write an analog of the above formulas as follows:

$$x \cdot (Y + Z) = (x \cdot Y) + (x \cdot Z)\,.$$

Here x is still a number, but Y and Z are vectors. The formula can, however, be easily verified in specific situations. The left-hand side:

$$2 \cdot (< 3, 4 > + < 5, 6 >) = 2 \cdot (< 3 + 5, 4 + 6 >= 2\cdot < 8, 10 >=< 16, 20 > \ .$$

The right-hand side:

$$2\cdot < 3, 4 > +2\cdot < 5, 6 >=< 2\cdot 3, 2\cdot 4 > + < 2\cdot 5, 2\cdot 6 >=< 6, 8 > + < 10, 12 >=< 6{+}10, 8{+}12 >=< 16, 20 > \ .$$

We will first explore these rules and short-cuts as they appear independently from componentwise representations of vectors.

In dimension 1, we deal with the algebra of directed segments. As these segments now all start at 0, this *is* the algebra of real numbers. Nonetheless, we keep the two types apart, with an eye on the higher dimensions.

The following simple idea connects vector addition to scalar multiplication:

$$A + A = 2A\,.$$

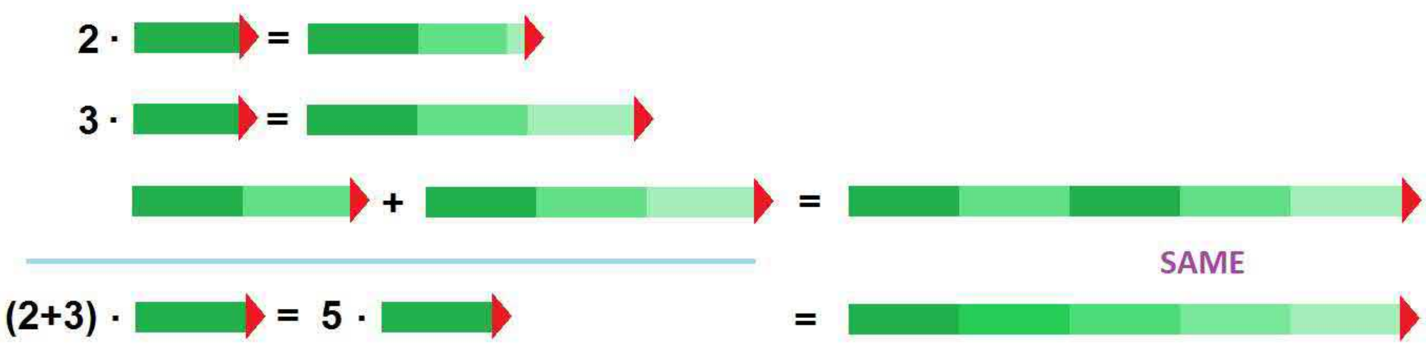
Its generalization is the *First Distributivity Property of Vector Algebra*:

$$\boxed{aA + bA = (a + b)A}$$

It's just factoring:

- ▶ We factor a *vector* out.

For example, let's add a double to a triple:



The result is the same if we quintuple the original vector.

In other words, we distribute scalar multiplication over addition of real numbers.

If we are to be precise, the symbol "+" stands for two different things above:

- $aA + bA$, addition of vectors
- $a + b$, addition of numbers

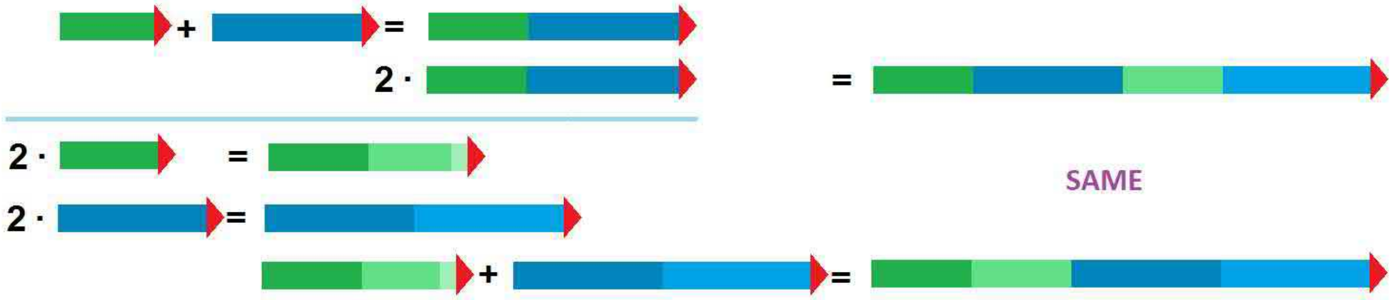
Next, we can also distribute multiplication of real numbers over addition of vectors. The *Second Distributivity Property of Vector Algebra* is:

$$\boxed{aA + aB = a(A + B)}$$

It's just factoring again:

- ▶ We factor a *number* out.

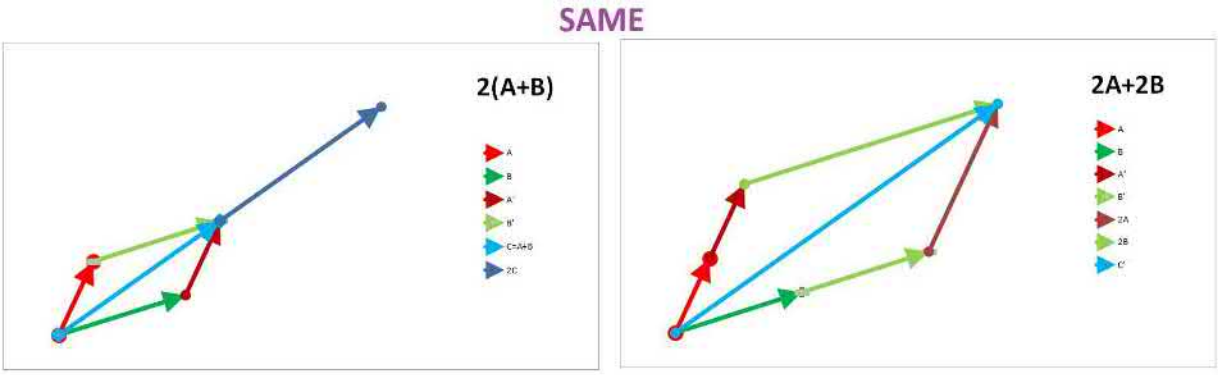
For example, let's double the sum:

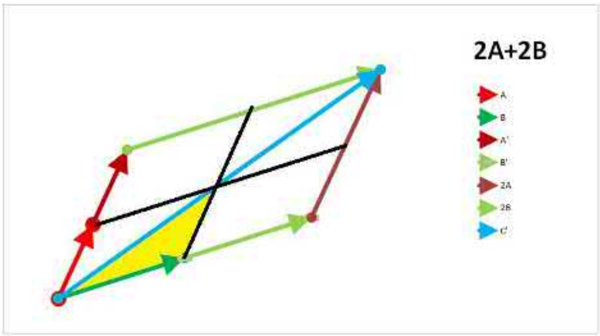


The result is the same if we add the doubles.

Dimension 2.

Below, we add two vectors and then stretch the result (left) and we stretch two vectors and then add them (right):





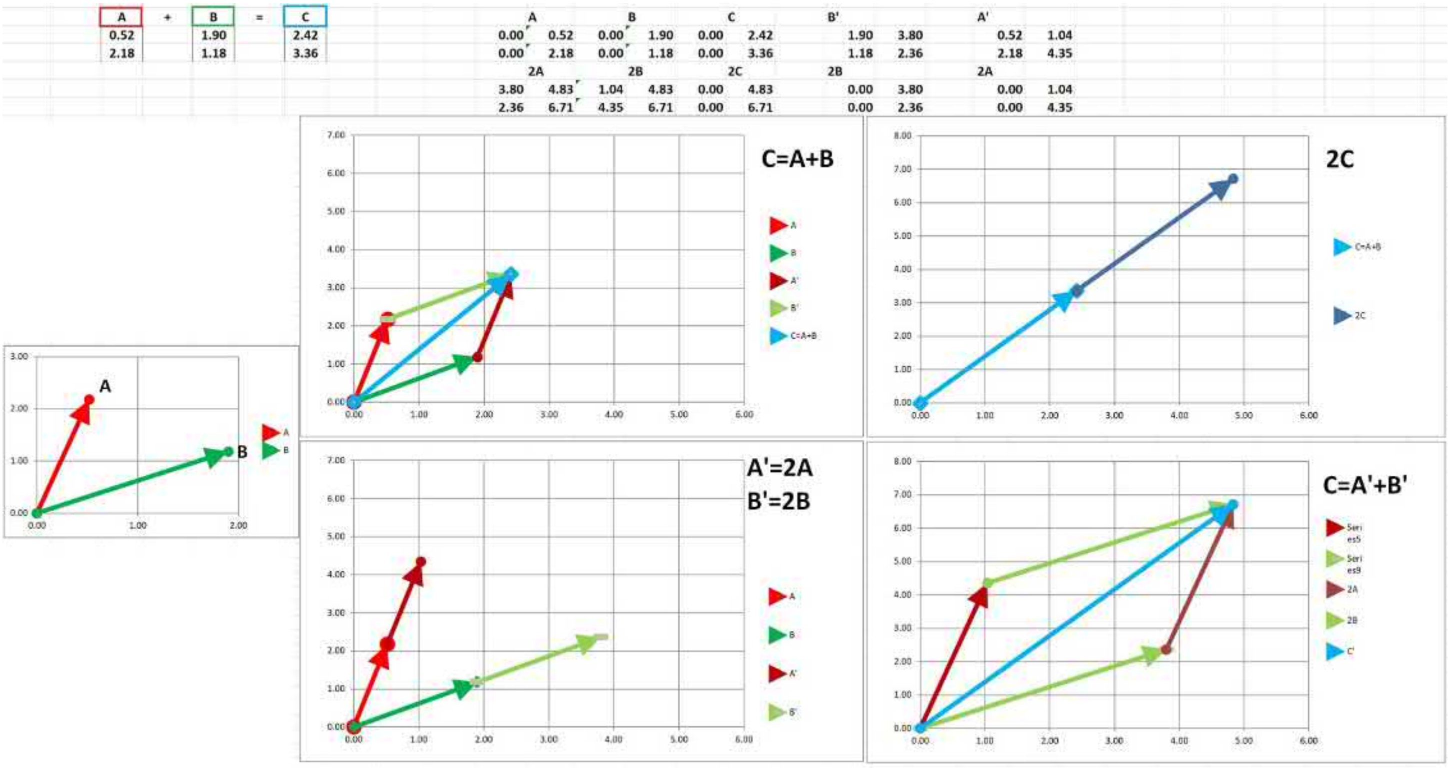
The result is the same.

Exercise 4.6.1

Explain why the results are the same. Hint: Similar triangles.

This algebraic rule, and others still to come, has been justified following the familiar algebra and geometry of the "physical space" $\mathbf{R}^2$ and $\mathbf{R}^3$. However, they also serve as a *starting point* for further development of linear algebra. In the last section, we defined the algebra of the abstract space $\mathbf{R}^n$ and now demonstrate that these "axioms" are still satisfied.

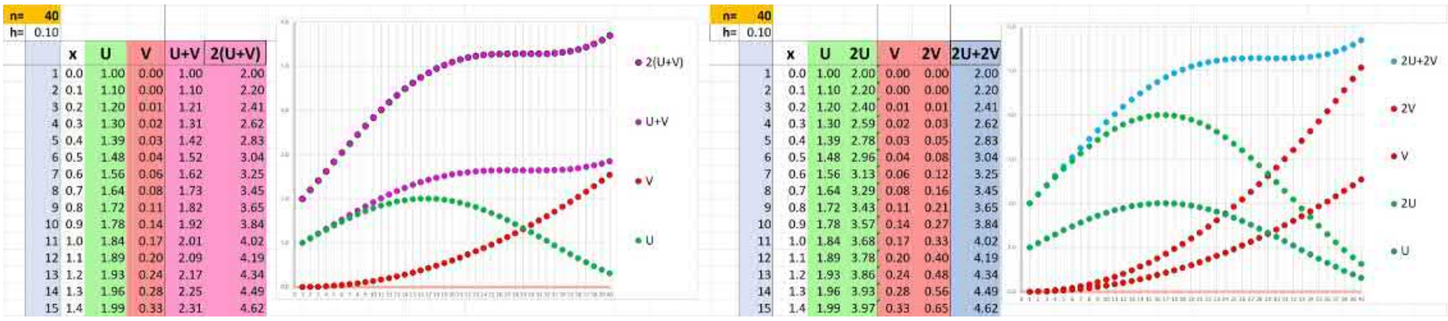
Below is the same formula for dimension 2, presented along with the components of the vectors:



A	+	B	=	C
0.52		1.90		2.42
2.18		1.18		3.36

A		B		C		B'		A'	
0.00	0.52	0.00	1.90	0.00	2.42	1.90	3.80	0.52	1.04
0.00	2.18	0.00	1.18	0.00	3.36	1.18	2.36	2.18	4.35
2A		2B		2C		2B		2A	
3.80	4.83	1.04	4.83	0.00	4.83	0.00	3.80	0.00	1.04
2.36	6.71	4.35	6.71	0.00	6.71	0.00	2.36	0.00	4.35

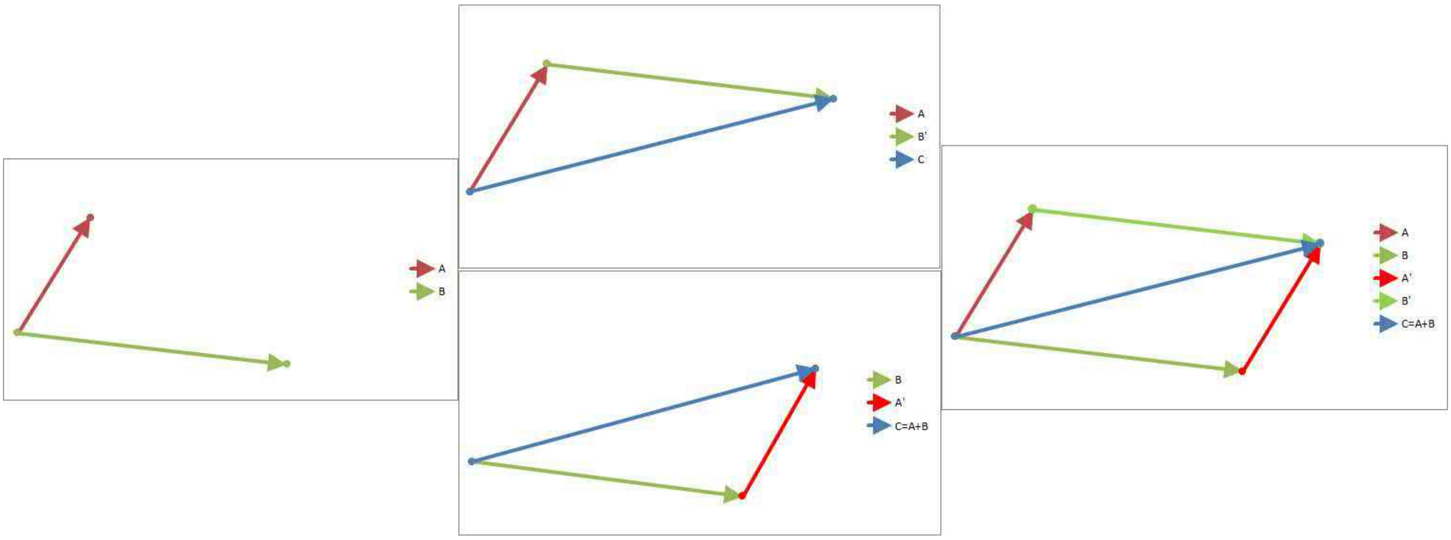
The formula is illustrated for dimension 40:



n=	40					
h=	0.10					
		x	U	V	U+V	2(U+V)
	1	0.0	1.00	0.00	1.00	2.00
	2	0.1	1.10	0.00	1.10	2.20
	3	0.2	1.20	0.01	1.21	2.41
	4	0.3	1.30	0.02	1.31	2.62
	5	0.4	1.39	0.03	1.42	2.83
	6	0.5	1.48	0.04	1.52	3.04
	7	0.6	1.56	0.06	1.62	3.25
	8	0.7	1.64	0.08	1.73	3.45
	9	0.8	1.72	0.11	1.82	3.65
	10	0.9	1.78	0.14	1.92	3.84
	11	1.0	1.84	0.17	2.01	4.02
	12	1.1	1.89	0.20	2.09	4.19
	13	1.2	1.93	0.24	2.17	4.34
	14	1.3	1.96	0.28	2.25	4.49
	15	1.4	1.99	0.33	2.31	4.62

n=	40						
h=	0.10						
	x	U	2U	V	2V	2U+2V	
1	0.0	1.00	2.00	0.00	0.00	2.00	
2	0.1	1.10	2.20	0.00	0.00	2.20	
3	0.2	1.20	2.40	0.01	0.01	2.41	
4	0.3	1.30	2.59	0.02	0.03	2.62	
5	0.4	1.39	2.78	0.03	0.05	2.83	
6	0.5	1.48	2.96	0.04	0.08	3.04	
7	0.6	1.56	3.13	0.06	0.12	3.25	
8	0.7	1.64	3.29	0.08	0.16	3.45	
9	0.8	1.72	3.43	0.11	0.21	3.65	
10	0.9	1.78	3.57	0.14	0.27	3.84	
11	1.0	1.84	3.68	0.17	0.33	4.02	
12	1.1	1.89	3.78	0.20	0.40	4.19	
13	1.2	1.93	3.86	0.24	0.48	4.34	
14	1.3	1.96	3.93	0.28	0.56	4.49	
15	1.4	1.99	3.97	0.33	0.65	4.62	

Recall that to find $A+B$, we make a *copy* B' of B, attach it to the end of A, and then create a new vector with the initial point that of A and terminal point that of B'. Now, to find $B+A$, we make a *copy* A' of A, attach it to the end of B, and then create a new vector with the initial point that of B and terminal point that of A':



The results are the same.

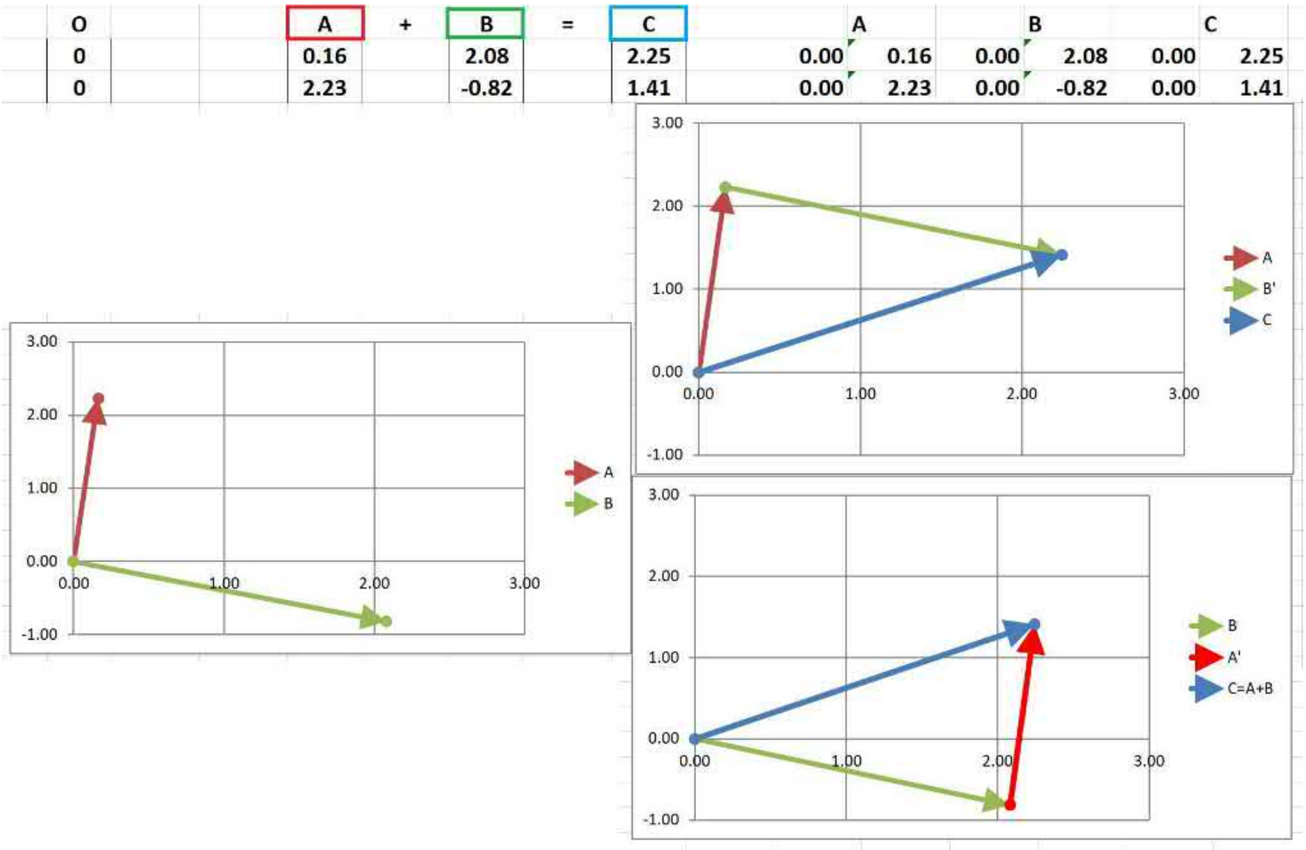
Exercise 4.6.2

Explain why the results are the same. Hint: Similar triangles.

Therefore, we have the *Commutativity Property of Vector Addition*:

$$\boxed{A + B = B + A}$$

Now with the coordinate system present, the components of the vectors are combined as follows:



O	A	+	B	=	C	A		B		C	
0	0.16		2.08		2.25	0.00	0.16	0.00	2.08	0.00	2.25
0	2.23		-0.82		1.41	0.00	2.23	0.00	-0.82	0.00	1.41

Next, we know that we can ignore the parentheses when we are adding numbers:

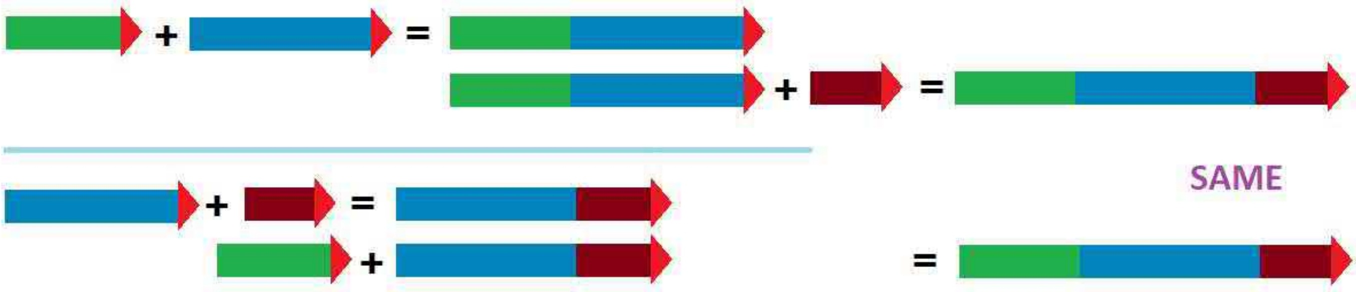
$$(1 + 2) + 3 = 1 + (2 + 3) = 1 + 2 + 3\,.$$

Identical is the *Associativity Property of Vector Addition*:

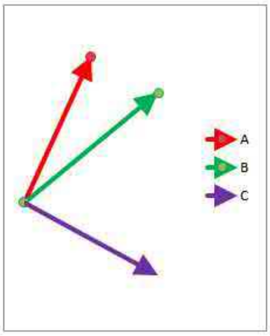
$$\boxed{A + (B + C) = (A + B) + C}$$

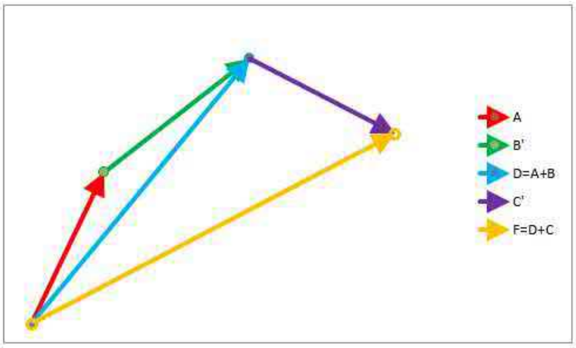
The order of addition doesn't matter!

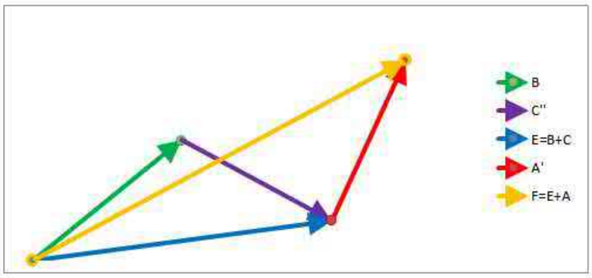
This is the property of dimension 1:



This is dimension 2:



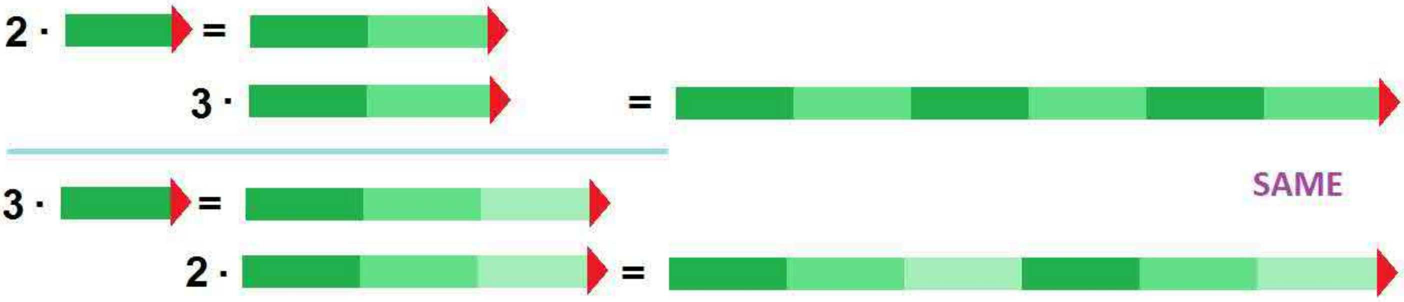




Let's consider next how we can apply two (or more) scalar multiplications in a row. Given a vector A and real numbers a and b, we can create several new vectors:

- $B = aA$ from A and then $C = bB$ from B
- $D = bA$ from A and then $C = aD$ from D
- $C = (ab)A$ directly from A

The results are the same. For example, below we double, then triple:



The result is the same if we sextuple.

If we are to be precise, the missing symbol "$\cdot$" stands for two different things above:

- $B = a \cdot A$, scalar multiplication of a vector
- $D = b \cdot A$, scalar multiplication of a vector
- $C = (a \cdot b)A$, multiplication of numbers

Our conclusion is the *Associativity Property of Scalar Multiplication*:

$$\boxed{b(aA) = (ba)A}$$

Exercise 4.6.3

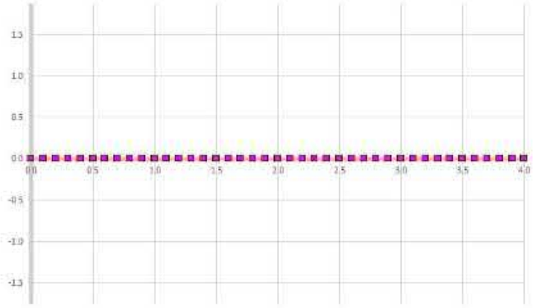
Provide an illustration for dimension 2.

There are some *special numbers*: 0 and 1. These formulas is what makes them special:

$$0 + x = x,\ 0 \cdot x = 0,\ 1 \cdot x = x\,.$$

Are there *special vectors*?

Consider vector addition. The *zero vector* is special. It has no magnitude nor direction and would have to be visualized as the dot O itself. This is what the zero vector is in $\mathbf{R}^{40}$:



There is 0, the real number, and then there is 0, the vector. The latter can mean no displacement, no motion (zero velocity), forces that cancel each other, etc. The two are related:

$$\boxed{0 \cdot A = 0}$$

This is a simple expression with a tricky algebraic meaning:

$$\text{number} \quad \cdot \quad \text{vector} \quad = \quad \text{vector}$$

Of course, the following holds for all vectors:

$$\boxed{A + 0 = A}$$

Now, scalar multiplication. Consider:

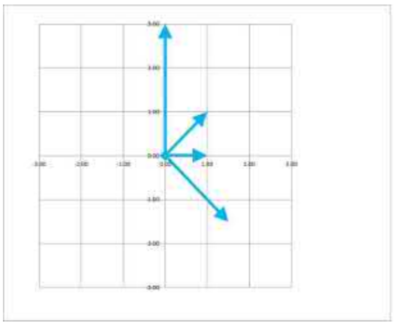
$$\boxed{1 \cdot A = A}$$

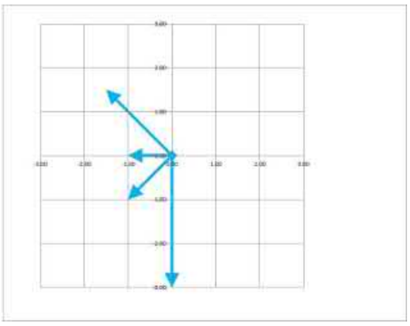
We have the same participants here as above:

$$\text{number} \quad \cdot \quad \text{vector} \quad = \quad \text{vector}$$

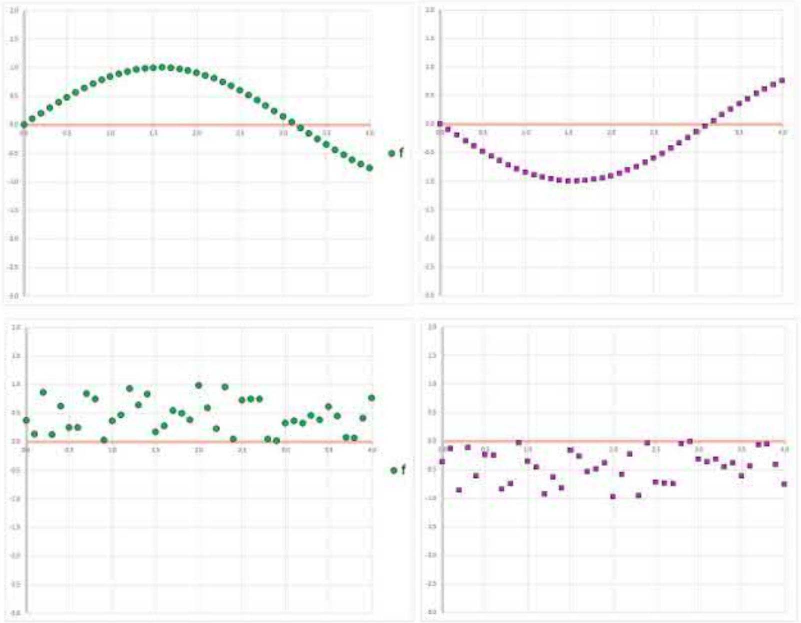
So, 1 remains special in vector algebra.

Next, since $PQ = -QP$, we have the *negative* $-A$ of a vector A, as the vector that goes in reverse of A. They are acquired by the central symmetry of the plane:





This is what happens in $\mathbf{R}^{40}$:



From the algebra, we also discover that

$$\boxed{-A = (-1) \cdot A}$$

As a summary, this is the complete list of rules one needs to carry out algebra with vectors:

Theorem 4.6.4: Axioms of Vector Space

The two operations – addition of two vectors and multiplication of a vector by a scalar – in $\mathbf{R}^n$ satisfy the following properties:

1. $X + Y = Y + X$ *for all* X *and* Y.
2. $X + (Y + Z) = (X + Y) + Z$ *for all* X, Y, *and* Z.
3. $X + 0 = X = 0 + X$ *for some vector* 0 *and all* X.
4. $X + (-X) = 0$ *for any* X *and some vector* $-X$.
5. $a(bX) = (ab)X$ *for all* X *and all scalars* a, b.
6. $1X = X$ *for all* X.
7. $a(X + Y) = aX + aY$ *for all* X *and* Y.
8. $(a + b)X = aX + bX$ *for all* X *and all scalars* a, b.

Taken together, these properties of vectors match the properties of numbers perfectly!

We put forward the following idea:

> ▶ *All manipulations of algebraic expressions that we have done with numbers are now allowed with vectors – as long as the expression itself makes sense.*

In other words, we just need to avoid operations that haven't been defined: no multiplication of vectors, no division of vectors, no adding numbers to vectors (of course!), etc.

All spaces of vectors, vector spaces if you like, we have seen so far have been only $\mathbf{R}^n$. Are there others?

Example 4.6.5: subsets and subspaces

Let's fix the last coordinate in $\mathbf{R}^n$ and look at the algebra:

$$\begin{array}{llllll} < & a_1 & a_2 & ... & a_{n-1} & a_n > \\ + & & & & & \\ < & b_1 & b_2 & ... & b_{n-1} & b_n > \\ \hline < & a_1+b_1 & a_2+b_2 & ... & a_{n-1}+b_{n-1} & a_n+b_n > \end{array} \to \begin{array}{llllll} < & a_1 & a_2 & ... & a_{n-1} & 0 > \\ + & & & & & \\ < & b_1 & b_2 & ... & b_{n-1} & 0 > \\ \hline < & a_1+b_1 & a_2+b_2 & ... & a_{n-1}+b_{n-1} & 0 > \end{array}$$

It works exactly the same!

How scalar multiplication works is also matched. Furthermore, we would anticipate that the eight properties in the theorem are satisfied.

Let's denote this set as follows:

$$\mathbf{R}^n_0 = \{< x_1, x_2, ..., x_{n-1}, 0 >\} \subset \mathbf{R}^n\,.$$

If we drop the redundant 0's, we realize that this is just a copy of $\mathbf{R}^{n-1}$:

$$\begin{array}{cccccc} & < & a_1 & a_2 & ... & a_{n-1} > \\ & + & & & & \\ \to & < & b_1 & b_2 & ... & b_{n-1} > \\ \cline{2-6} & < & a_1+b_1 & a_2+b_2 & ... & a_{n-1}+b_{n-1} > \end{array}$$

Definition 4.6.6: vector space

Any set with two operations that satisfy the conclusions of the theorem is called a *a vector space*.

So, $\mathbf{R}^{n-1}_0$ is a vector space, a *subspace* of $\mathbf{R}^n$.

Every subset of $\mathbf{R}^n$ is subject to the algebraic operations of the ambient space. How do we determine when this is a vector space? We just need to make sure that the algebra makes sense:

Theorem 4.6.7: Subspaces

Suppose U is a subset of a vector space that satisfies:

1. *If X and Y belong to U, then so does $X+Y$.*
2. *If X belongs to U, then so does kX for any number k.*

Then U is a vector space.

Exercise 4.6.8

For the last example, show that setting the last coordinate to a non-zero number won't create a vector space.

Exercise 4.6.9

For the last example, show that setting the several coordinates to zero will create a vector space.

Exercise 4.6.10

Prove that a line through 0 on the plane is a vector space.

Exercise 4.6.11

Prove the theorem.

4.7. *Convex, affine, and linear combinations of vectors*

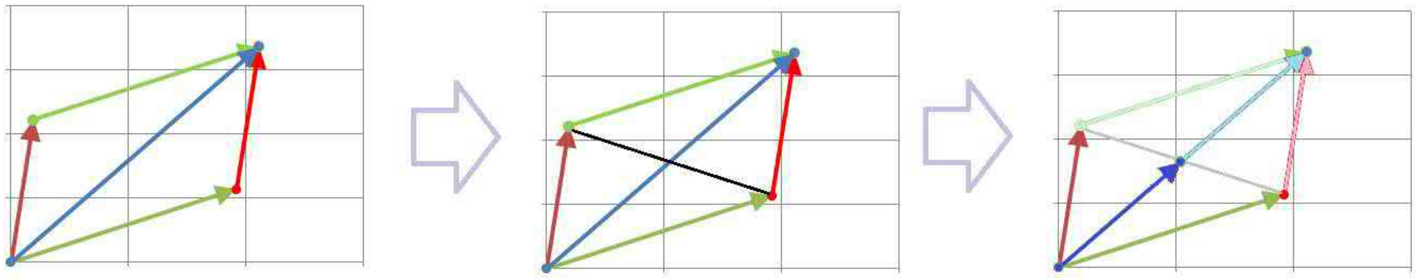
The average of two numbers is defined to be their half sum:

$$\frac{x+y}{2} = \frac{1}{2}x + \frac{1}{2}y\,.$$

Since the algebra of vectors mimics that of numbers, nothing stops us from defining the *average of vectors* U and V in the same manner:

$$\frac{1}{2}U + \frac{1}{2}V = \frac{1}{2}(U+V)\,.$$

It is a convenient concept illustrated below for dimension 2:

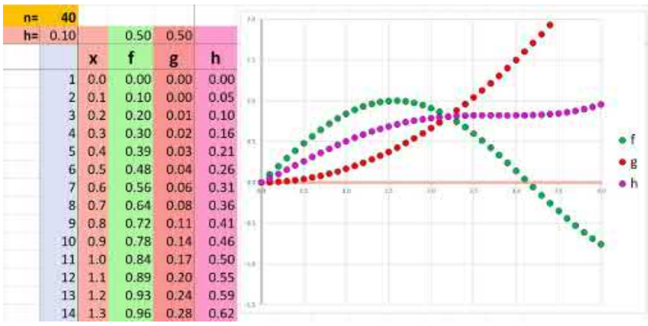


The average of two vectors is the vector that goes to the center of the parallelogram formed by the two.

Exercise 4.7.1

Prove the last statement.

In dimension 40, the graph of the average lies half-way (vertically) between the two:



n=	40				
h=	0.10		0.50	0.50	
		x	f	g	h
	1	0.0	0.00	0.00	0.00
	2	0.1	0.10	0.00	0.05
	3	0.2	0.20	0.01	0.10
	4	0.3	0.30	0.02	0.16
	5	0.4	0.39	0.03	0.21
	6	0.5	0.48	0.04	0.26
	7	0.6	0.56	0.06	0.31
	8	0.7	0.64	0.08	0.36
	9	0.8	0.72	0.11	0.41
	10	0.9	0.78	0.14	0.46
	11	1.0	0.84	0.17	0.50
	12	1.1	0.89	0.20	0.55
	13	1.2	0.93	0.24	0.59
	14	1.3	0.96	0.28	0.62

Let's take this idea one step further.

The weighted average of two numbers is defined to be a combination like this:

$$\alpha x + \beta y\,,$$

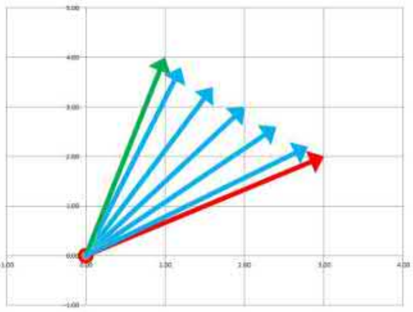
where $\alpha \geq 0$ and $\beta \geq 0$ add up to 1:

$$\alpha + \beta = 1\,.$$

Similarly, the *weighted average of vectors* U and V is defined to be:

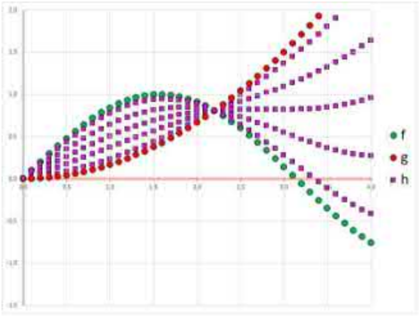
$$\alpha U + \beta V\,.$$

We can see in dimension 2 how one vector is gradually transformed into the other as *alpha* runs from 1 to 0 (and β from 0 to 1):



These intermediate stages are also called *convex combinations* of the two vectors. Together their ends form a line segment; it runs from the end of U to the end of V.

In dimension 40, we can also see a gradual transition from one vector to the other:

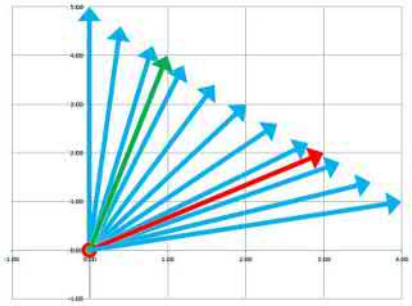


Now, we saw in dimension 2 that the average is also a straight segment between the two vectors. Of course, it is a straight segment between the two for dimension 3 or any dimension that we can visualize.

If we remove the restriction $\alpha \geq 0$ and $\beta \geq 0$, our combinations

$$\alpha U + \beta V$$

are called the *affine combinations* of U and V. Together, their ends form a whole line; it passes through the end of U and the end of V:



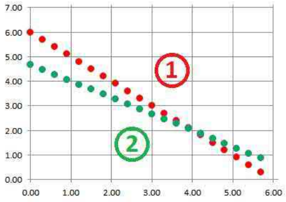
Let's recall the problem we have been using to illustrate many new ideas.

Problem: We are given the Kenyan coffee at \$2 per pound and the Colombian coffee at \$3 per pound. How much of each do you need to have 6 pounds of blend with the total price of \$14?

We let x be the weight of the Kenyan coffee and let y be the weight of the Colombian coffee. Then the total price of the blend is \$14. Therefore, we have a system:

$$\begin{array}{rcrl} x & + & y & = 6 \\ 2x & + & 3y & = 14 \end{array}$$

The solution to the system as presented initially had a clear *geometric* meaning. We thought of the two equations as equations about the coordinates of *points*, (x,y), in the plane. In fact, either equation is a representation of a line on the plane. Then the solution $(x,y) = (4,2)$ is the point of their intersection:



The second interpretation was in terms of a *function* defined on the plane. A function $F : \mathbf{R}^2 \to \mathbf{R}^2$ is given by:

$$F(x,y) = (x+y, 2x+3y)\,.$$

Then our solution is:

$$(x,y) = F^{-1}(6,14).$$

We now have a new interpretation – in terms of *vectors* in the plane.

We re-write the system as a vector equation:

$$\begin{array}{rcrl} x & + & y & = 6 \\ 2x & + & 3y & = 14 \end{array} \quad \Longrightarrow \quad \begin{bmatrix} x & + & y \\ 2x & + & 3y \end{bmatrix} = \begin{bmatrix} 6 \\ 14 \end{bmatrix}.$$

The first vector's components are computed via some algebra. We will try to interpret this algebra of numbers in terms of our algebra of vectors.

We split the vector up:

$$\begin{bmatrix} x & + & y \\ 2x & + & 3y \end{bmatrix} = \begin{bmatrix} x \\ 2x \end{bmatrix} + \begin{bmatrix} y \\ 3x \end{bmatrix}$$

We factor the repeated coefficients out:

$$\begin{bmatrix} x & + & y \\ 2x & + & 3y \end{bmatrix} = x \begin{bmatrix} 1 \\ 2 \end{bmatrix} + y \begin{bmatrix} 1 \\ 3 \end{bmatrix}.$$

Our system has been reduced to a single *vector equation*:

$$x \begin{bmatrix} 1 \\ 2 \end{bmatrix} + y \begin{bmatrix} 1 \\ 3 \end{bmatrix} = \begin{bmatrix} 6 \\ 14 \end{bmatrix}.$$

Let's analyze this equation and the problem it presents.

Given two vectors:

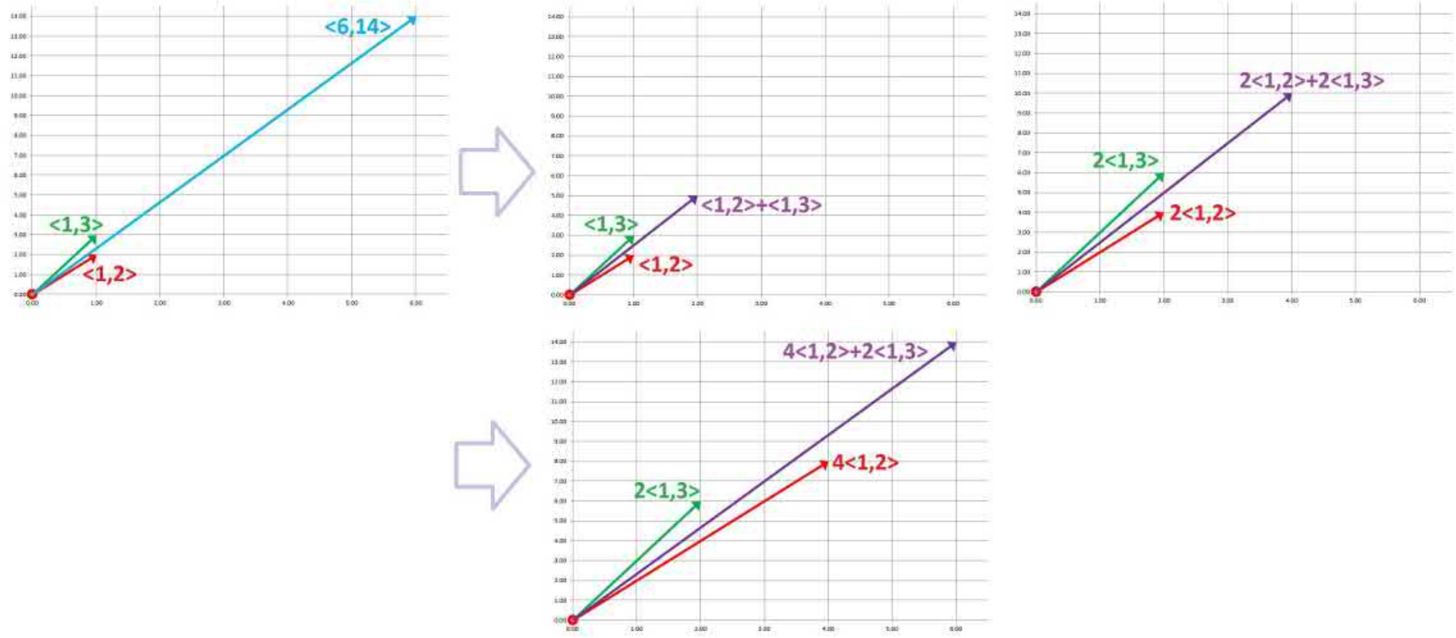
$$U = \begin{bmatrix} 1 \\ 2 \end{bmatrix} \quad \text{and} \quad V = \begin{bmatrix} 1 \\ 3 \end{bmatrix},$$

find two numbers x and y so that we have:

$$xU + yV = \begin{bmatrix} 6 \\ 14 \end{bmatrix}.$$

It may appear that we just need to represent the vector $< 6, 14 >$ as an *affine combination* of the vectors U and V. However, there is no restriction that x and y must add up to 1. We speak of *linear combinations.*

So, we need to find a way to *stretch* either of these two vectors so that their sum is the third vector. The setup is on the left followed by a trial-and-error:



Just adding the two vectors or adding their proportional multiples fails; it is clear that the angle can't match. Hypothetically, we go through all linear combinations of these two vectors to find one that is just right.

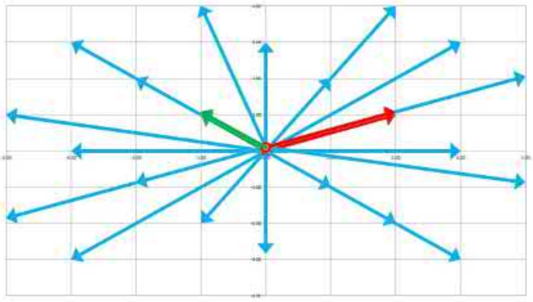
So, the new point of view on the problem of mixtures is different:

▶ Instead of the *locations*, we are after the *directions*.

In general, these are all linear combinations $\alpha U + \beta V$, in addition to the affine combinations:

$\alpha\backslash\beta$	...	-2	-1	0	1	2	...
...	...	...	...	...	...	...	...
-2	...	$-2U-2V$	$-2U-V$	$-2U$	$-2U+V$	$-2U+2V$	...
-1	...	$-U-2V$	$-U-V$	$-U$	$-U+V$	$-U+2V$	...
0	...	$-2V$	$-V$	0	V	$2V$	...
1	...	$U-2V$	$U-V$	U	$U+V$	$U+2V$	...
2	...	$2U-2V$	$2U-V$	$2U$	$2U+V$	$2U+2V$	...
...	...	...	...	...	...	...	...

These are the linear combinations with integer coefficients of $U =< 2,1 >$ and $V =< -1,1 >$:



They seem to cover the whole plane.

Exercise 4.7.2

Is it true?

In summary, we have these three combinations:

Definition 4.7.3: linear, affine, and convex combinations

1. A *linear combination* of two vectors U and V is defined to be the following expression with any real numbers α and β called its coefficients:

$$\alpha U + \beta V\,.$$

2. An *affine combination* of two vectors U and V is defined to be their linear combination with coefficients α and β that add up to 1:

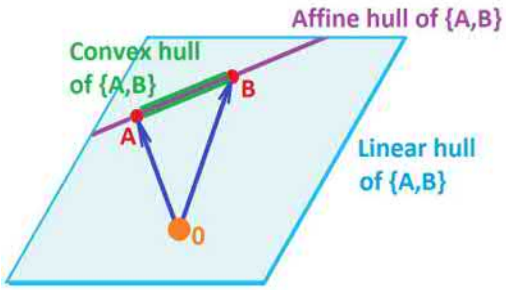
$$\alpha + \beta = 1\,.$$

3. A *convex combination* of two vectors U and V is defined to be their affine combination with coefficients α and β that are non-negative:

$$\alpha \geq 0,\ \beta \geq 0\,.$$

Example 4.7.4: hulls

These concepts have analogs for points. We define three *hulls*:



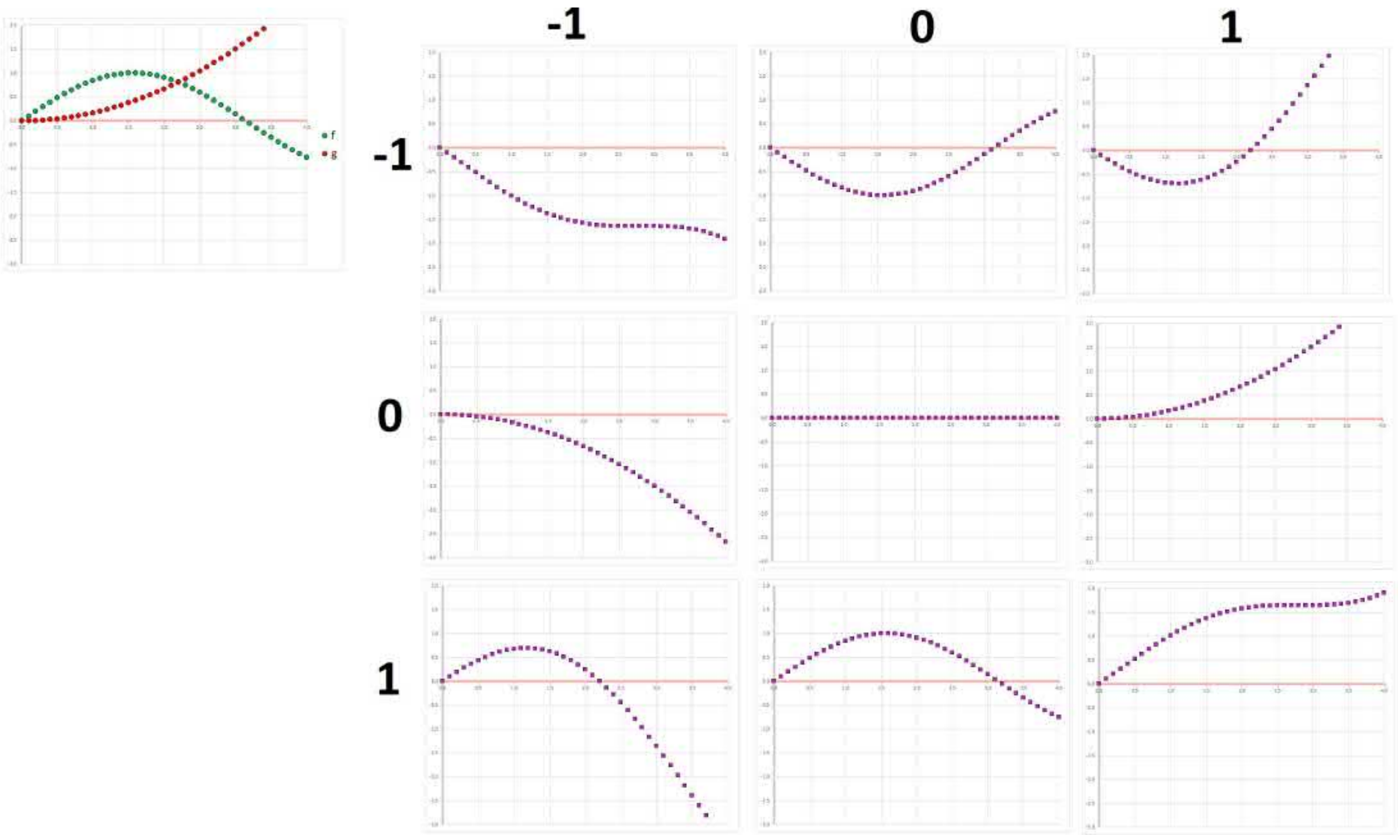
As you can see, the convex and the affine hulls are independent of the location of the origin.

Exercise 4.7.5

What about the linear hull?

In the general Euclidean space, we consider the results of all possible computations with the two algebraic operations that we have: vector addition and scalar multiplication.

This is what a few of them $(\alpha f + \beta g)$ look like in $\mathbf{R}^{40}$:



Exercise 4.7.6

Label these.

The next definition is an extension of this idea to an unlimited number of vectors:

Definition 4.7.7: linear combination

Suppose $V_1, ..., V_m$ are vectors in $\mathbf{R}^n$. Then, the *linear combination of the vectors* $V_1, ..., V_m$ *with coefficients* $r_1, ..., r_m$ is the following vector:

$$r_1 V_1 + ... + r_m V_m \,.$$

Then, the set of all linear combinations of a single vector is simply the set of its multiples (a line).

The set of all linear combinations of two vectors in the plane $\mathbf{R}^2$ is the whole plane, unless the two are multiples of each other.

Exercise 4.7.8

Prove the last statement.

Exercise 4.7.9

Finish the sentence: "The set of all linear combinations of three vectors in the 3-space is the whole 3-space, unless _ _ _ _ _ _ _ _ _".

An important fact is the following:

Theorem 4.7.10: Linear Combination of Basis Vectors

Every vector in $\mathbf{R}^n$ is a linear combination of the basis vectors:

$$< a_1, ..., a_n >= a_1 e_1 + ... + a_n e_n \,.$$

Exercise 4.7.11

Prove the theorem.

Example 4.7.12: polynomials

A polynomial is a linear combination of the power functions:

$$a_0 + a_1 x^1 + ... + a_n x^n \,.$$

In this sense, the space of all polynomials of degree up to n is indistinguishable from $\mathbf{R}^{n+1}$.

Exercise 4.7.13

Show that the multiples of a given vector in a vector space form a vector space.

Exercise 4.7.14

Show that the linear combinations of a given pair of vector in a vector space form a vector space.

Exercise 4.7.15

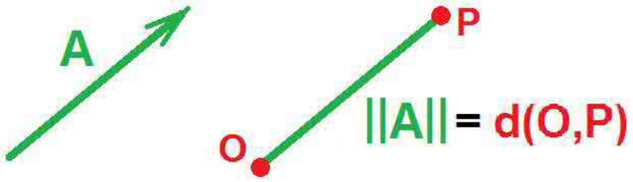
What is the next statement in this sequence?

4.8. *The magnitude of a vector*

A vector is a directed segment. Its attributes are, therefore, the direction and the magnitude. It may be hard to explain what *direction* means without referring, circularly, to vectors. That is why we look at the *magnitude* first:

- ▶ The magnitude of a vector is what's left of it when it's stripped off its direction.

When we interpret the vector as a displacement, we'd rather talk about its *length*. The meaning of this number is clear when the vector is given by two points, PQ. The length is the distance $d(P,Q)$ between them:



We intentionally make no reference to a Cartesian system:

Definition 4.8.1: magnitude of a vector

The *magnitude* or the *length of a vector* in $\mathbf{R}^n$ is defined to be the distance between its initial and terminal points:

$$\boxed{||PQ|| = d(P, Q)}$$

This number is also called the *norm* of the vector.

In particular, we have:

1. If $d(P, Q)$ stands for the Euclidean metric, $||PQ||$ is called the *Euclidean norm*.
2. If $d(P, Q)$ stands for the taxicab metric, $||PQ||$ is called the *taxicab norm*.

The notation resembles the *absolute value* and not by accident; they are the same in the 1-dimensional case, $\mathbf{R}$.

We also intentionally make no reference to a specific distance formula. The approach is as follows:

> ▶ We now look at each of the three *Axioms of Metric Space* and – using the above formula – translate it into a property of vectors.

First, the *Positivity*:

$$\boxed{d(P, Q) \geq 0; \quad \text{and} \quad d(P, Q) = 0 \ \text{ if and only if } P = Q}$$

We rewrite according to the definition above:

$$\boxed{||PQ|| \geq 0; \quad \text{and} \quad ||PQ|| = 0 \ \text{ if and only if } P = Q}$$

But the vector PP is just the zero vector! Therefore, we have this new form of the property:

$$\boxed{||A|| \geq 0; \quad \text{and} \quad ||A|| = 0 \ \text{ if and only if } \ A = 0}$$

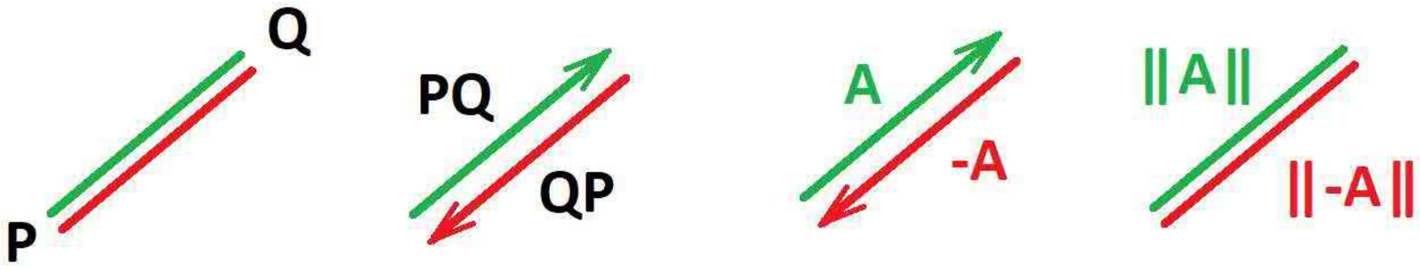
Second, the *Symmetry*:

$$\boxed{d(P, Q) = d(Q, P)}$$

We rewrite according to the definition above:

$$\boxed{||PQ|| = ||QP||}$$

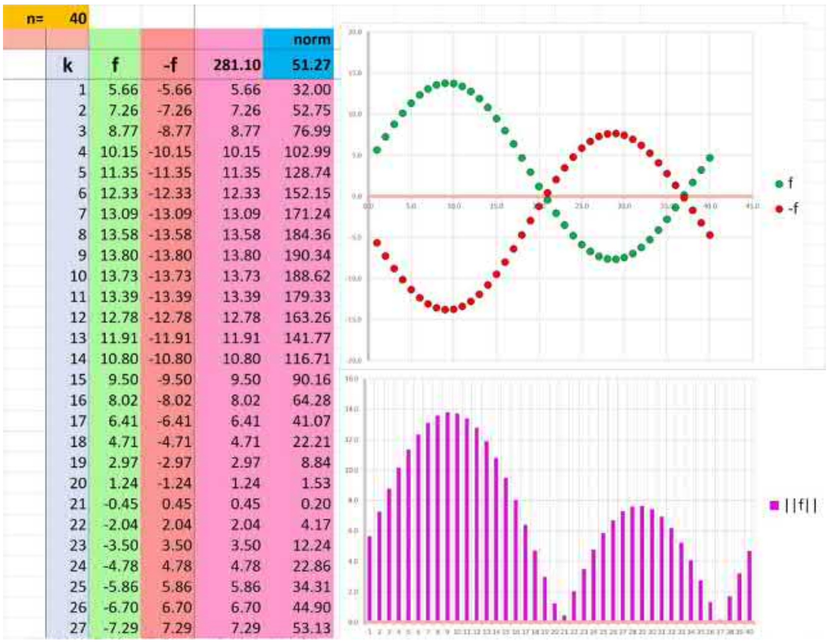
But the vector PQ is the negative of QP:



Therefore, we have this new form of the property:

$$\boxed{||A|| = || - A||}$$

Its meaning is visualized for $\mathbf{R}^{40}$ below:



n=	40				
					norm
	k	f	-f	281.10	51.27
	1	5.66	-5.66	5.66	32.00
	2	7.26	-7.26	7.26	52.75
	3	8.77	-8.77	8.77	76.99
	4	10.15	-10.15	10.15	102.99
	5	11.35	-11.35	11.35	128.74
	6	12.33	-12.33	12.33	152.15
	7	13.09	-13.09	13.09	171.24
	8	13.58	-13.58	13.58	184.36
	9	13.80	-13.80	13.80	190.34
	10	13.73	-13.73	13.73	188.62
	11	13.39	-13.39	13.39	179.33
	12	12.78	-12.78	12.78	163.26
	13	11.91	-11.91	11.91	141.77
	14	10.80	-10.80	10.80	116.71
	15	9.50	-9.50	9.50	90.16
	16	8.02	-8.02	8.02	64.28
	17	6.41	-6.41	6.41	41.07
	18	4.71	-4.71	4.71	22.21
	19	2.97	-2.97	2.97	8.84
	20	1.24	-1.24	1.24	1.53
	21	-0.45	0.45	0.45	0.20
	22	-2.04	2.04	2.04	4.17
	23	-3.50	3.50	3.50	12.24
	24	-4.78	4.78	4.78	22.86
	25	-5.86	5.86	5.86	34.31
	26	-6.70	6.70	6.70	44.90
	27	-7.29	7.29	7.29	53.13

The norm is the purple area at the bottom.

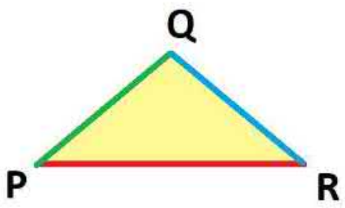
Third, the *Triangle Inequality*:

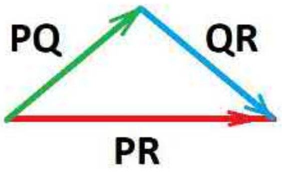
$$\boxed{d(P,Q) + d(Q,R) \geq d(P,R)}$$

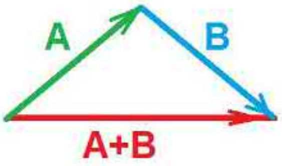
We rewrite according to the definition above:

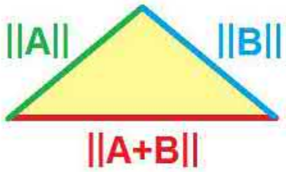
$$\boxed{||PQ|| + ||QR|| \geq ||PR||}$$

But $PQ + QR = PR$:









Therefore, we have this new form of the property:

$$\boxed{||A|| + ||B|| \geq ||A + B||}$$

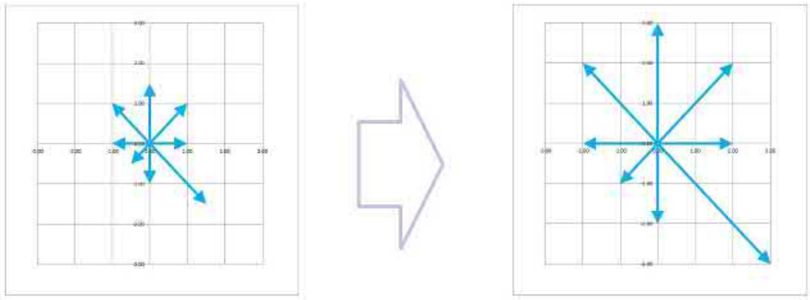
So, we have moved from the *geometry of points* to the *algebra of vectors.*

In summary:

1. The magnitude cannot be negative, and only the zero vector has a zero magnitude.
2. The magnitude of the negative of a vector is equal to that of the vector.
3. The magnitude of the sum of two vectors is larger than or equal to that of their sum.

That is how the magnitude interacts with *vector addition*. What about *scalar multiplication*? There is no corresponding property of distances.

All vectors double in length whether we multiply by 2 or -2, while their directions are preserved or flipped. When the direction doesn't matter, we just multiply the components by this stretching factor, $2 = |2| = |-2|$:



The result is another convenient property:

Theorem 4.8.2: Homogeneity of Norm

Both the Euclidean and the taxicab norms satisfy the following for any vector A *and any scalar* k*:*

$$\boxed{||k \cdot A|| = |k| \cdot ||A||}$$

Proof.

For the Euclidean norm in $\mathbf{R}^2$:

$$||k < a, b > || = || < ka, kb > || = \sqrt{(ka)^2 + (kb)^2} = |c| \cdot \sqrt{a^2 + b^2} = |k| \cdot || < a, b > || \,.$$

For the taxicab norm in $\mathbf{R}^2$:

$$||k < a, b > || = || < ka, kb > || = |ka| + |kb| = |k| \cdot |a| + |k| \cdot |b| = |k| \cdot (|a| + |b|) = k \cdot || < a, b > || \,.$$

Warning!

The *Symmetry* above is now redundant as it is incorporated into this new property:

$$-A = (-1)A \,.$$

These properties are applicable to all dimensions and are used to manipulate vector expressions.

We now turn around and ask:

- What's left of a vector when its magnitude is stripped off?

If we "remove" the magnitude from consideration, we are left with nothing but the direction. We can only say this:

- Vectors with the same direction are (positive) multiples of each other.

But what is the simplest vector among those?

To study directions, we limit our attention to some special vectors:

Definition 4.8.3: unit vector

Every vector with magnitude equal to 1 is called a *unit vector*:

$$||X|| = 1\,.$$

We can make such a vector from any vector – "normalize" it – except 0, by dividing by its magnitude:

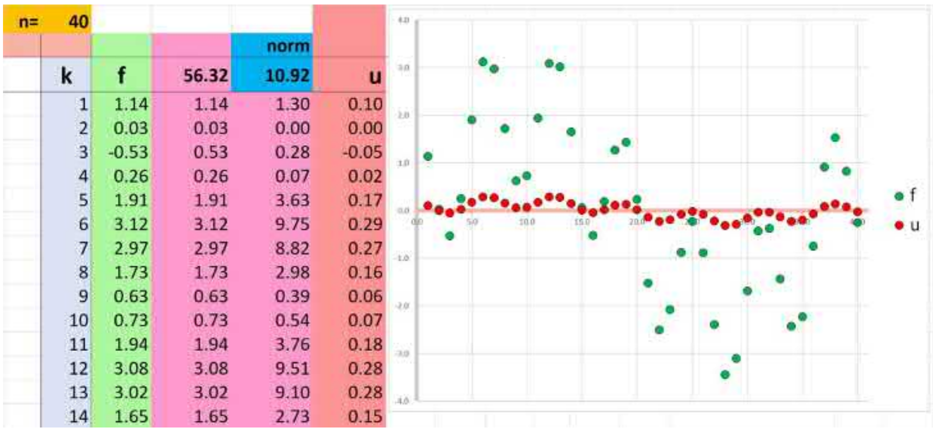
Theorem 4.8.4: Normalization of Vectors

For any vector $X \neq 0$*, the vector*

$$\frac{X}{||X||}$$

is a unit vector.

It's simply a re-scaled version of the original. This is what it looks like in $\mathbf{R}^{40}$:

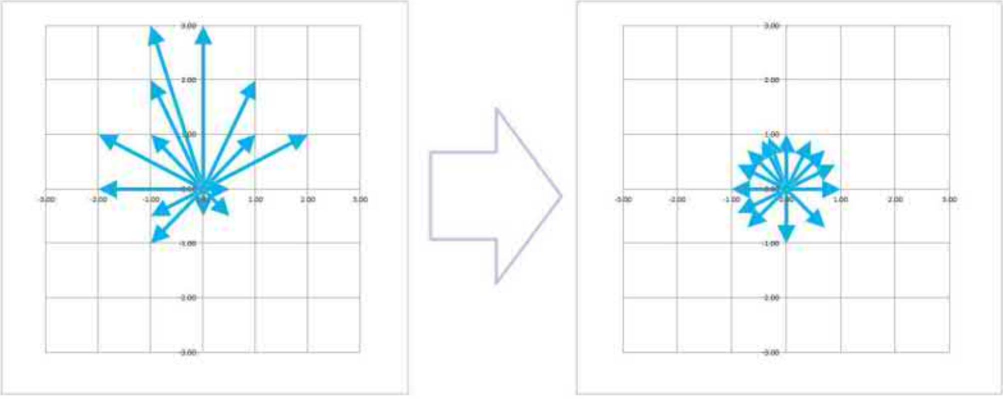


n= 40				
			norm	
k	f	56.32	10.92	u
1	1.14	1.14	1.30	0.10
2	0.03	0.03	0.00	0.00
3	-0.53	0.53	0.28	-0.05
4	0.26	0.26	0.07	0.02
5	1.91	1.91	3.63	0.17
6	3.12	3.12	9.75	0.29
7	2.97	2.97	8.82	0.27
8	1.73	1.73	2.98	0.16
9	0.63	0.63	0.39	0.06
10	0.73	0.73	0.54	0.07
11	1.94	1.94	3.76	0.18
12	3.08	3.08	9.51	0.28
13	3.02	3.02	9.10	0.28
14	1.65	1.65	2.73	0.15

Exercise 4.8.5

Prove the theorem.

The effect of normalization is that the vectors that are too long are shrunk and the ones that are too short are stretched – radially – toward the unit circle:



Unit vectors capture nothing but the direction:

Theorem 4.8.6: Multiples of Vectors

Suppose two vectors have equal or opposite unit vectors:

$$\frac{V}{||V||} = \pm\frac{W}{||W||}\,.$$

Then they are multiples of each other:

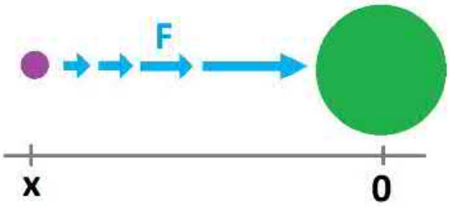
$$V = kW\,.$$

Exercise 4.8.7

Prove the theorem.

Example 4.8.8: Newton's Law of Gravity

Recall the law of gravity. Its force pulls two objects the harder the closer to each other they are:



The gravity is a function of two variables, x, y, that give the location, or better, it is a function of points, P, in $\mathbf{R}^2$ with real values:

$$f : \mathbf{R}^2 \to \mathbf{R}\,.$$

Algebraically, the law says that the force of gravity between two objects of masses M and m located at points O and P is given by:

$$\boxed{f(P) = G\frac{mM}{d(O,P)^2}}$$

Next, the law, in addition to the formula, includes the statement that the *force is directed from P to O*. This is implicitly the language of *vectors*: The direction of the force depends on the direction of the location vector. Let's sort this out.

Let's take care of the magnitudes first. We have two vectors:

- Gravity is a force and, therefore, a vector.
- The location of the second object is its displacement from the first and, therefore, a vector.

The function will have both vector inputs and vector outputs:

$$F : \mathbf{R}^2 \to \mathbf{R}^2\,.$$

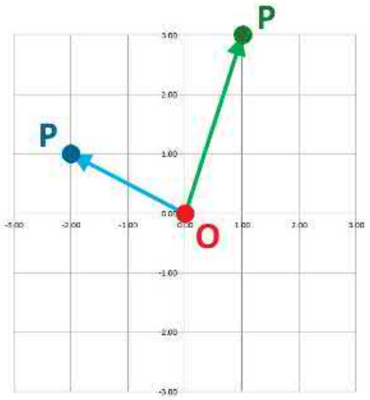
We place the origin O at the location of the first object (maybe the Sun). Then $OP = X$, and we can re-write the formula as follows:

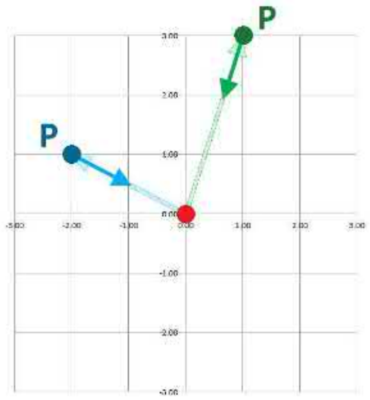
$$\boxed{||F(X)|| = G\frac{mM}{||X||^2}}$$

Next, what about the directions? Let's derive the *vector form* of the law. The law states:

> ▶ The force of gravity affecting either of the two objects is directed towards the other object.

We see this below:





In other words, F points in the opposite direction to X, i.e., it's direction is that of $-X$. Therefore, the unit vectors of $F(X)$ and $-X$ are equal:

$$\frac{F}{||F||} = -\frac{X}{||X||} .$$

Therefore, the vectors themselves are multiples of each other:

$$\boxed{F(X) = c(-X)}$$

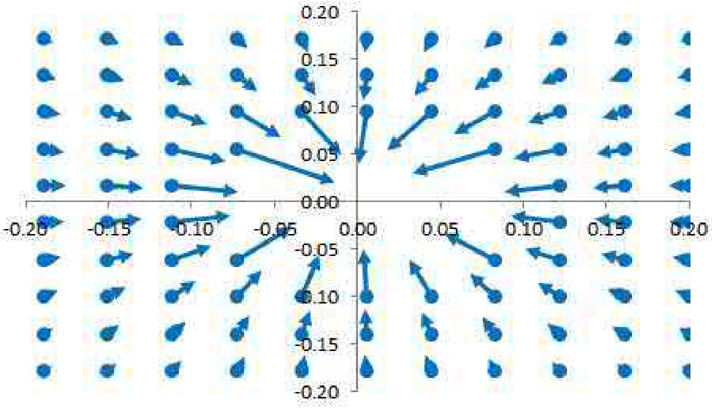
That's all we need except for the coefficient. We now use *Homogeneity* to find it:

$$G\frac{mM}{||X||^2} = ||F|| = |c| \cdot ||(-X)|| = |c| \cdot ||X|| .$$

The final form is the following:

$$\boxed{F(X) = -G\frac{mM}{||X||^3}X}$$

Now that both the input and the output are 2-dimensional vectors, how do we visualize this kind of function? Even though this is just a (non-linear) transformation of $\mathbf{R}^2$ (or $\mathbf{R}^3$), there is a better way. First, we think of the input as a *point* and the output as a *vector* and then we attach the latter to the former. Below, we plot vector $F(X)$ starting at location X on the plane:



It is called a *vector field*.

Exercise 4.8.9

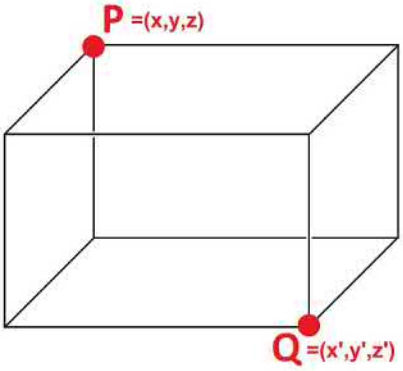
Suggest other examples of location-dependent forces.

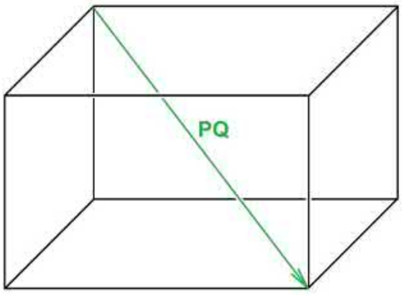
These geometric properties have been justified following the familiar geometry of the "physical space" $\mathbf{R}^3$. However, they also serve as a *starting point* for a further development of linear algebra.

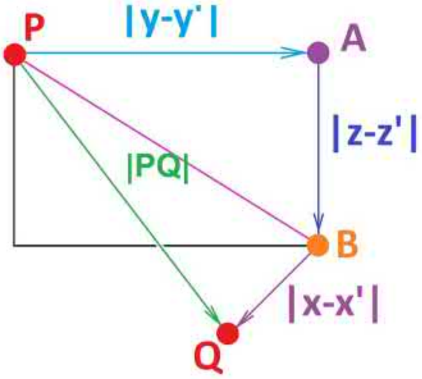
When a Cartesian system is provided, we have the *Euclidean metrics*, i.e., the distance between points P and Q in $\mathbf{R}^3$ with coordinates (x, y, z) and (x', y', z') respectively is

$$d(P, Q) = \sqrt{(x - x')^2 + (y - y')^2 + (z - z')^2}\,.$$

The three terms are recognized as the three components of the vector PQ:







If we apply this computation to a vector OP with $P = (a, b, c)$, we conclude:

$$|| < a, b, c > || = \sqrt{a^2 + b^2 + c^2}\,.$$

Meanwhile, the taxicab norm of this vector is

$$|| < a, b, c > || = |a| + |b| + |c|\,.$$

In general we have the following:

Theorem 4.8.10: Magnitude of Vector

Suppose we have a vector

$$A =< a_1, ..., a_n >$$

in $\mathbf{R}^n$. *Then we have:*

1. *The Euclidean norm of* A *is equal to*

$$\boxed{||A|| = \sqrt{a_1^2 + ... + a_n^2}}$$

2. *The taxicab norm of a vector* A *is equal to*

$$\boxed{||A|| = |a_1| + ... + |a_n|}$$

In the sigma notation, we have, respectively:

$$||A|| = \sqrt{\sum_{k=1}^{n} a_k^2}\,,$$

and

$$||A|| = \sum_{k=1}^{n} |a_k|\,.$$

As a summary, these are the properties of the magnitudes of vectors.

Theorem 4.8.11: Axioms of Normed Space

For any vectors A, B in $\mathbf{R}^n$ and any real k, the following properties are satisfied by both the Euclidean norm and the taxicab norm:

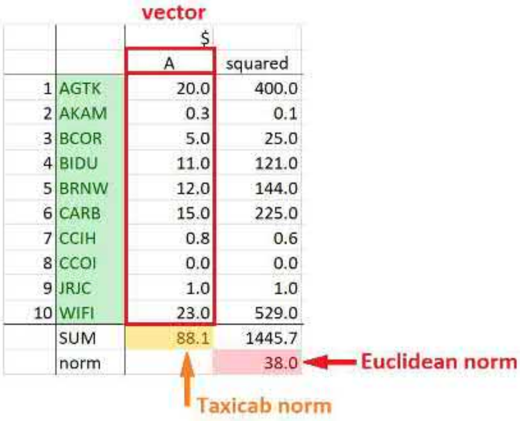
1. ***Positivity:*** $||A|| \geq 0$; *and* $||A|| = 0$ *if and only if* $A = 0$.
2. ***Triangle Inequality:*** $||A|| + ||B|| \geq ||A + B||$.
3. ***Homogeneity:*** $||k \cdot A|| = |k| \cdot ||A||$.

Exercise 4.8.12

Demonstrate that the formulas for the norms satisfy those three properties.

Example 4.8.13: investment portfolios

A portfolio of stocks can be subject to these operations. Assuming that there are only these 10 stocks available, all portfolios are vectors (or points) in $\mathbf{R}^{10}$:

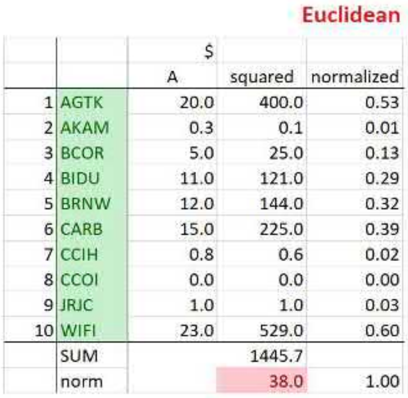


		$	
		A	squared
1	AGTK	20.0	400.0
2	AKAM	0.3	0.1
3	BCOR	5.0	25.0
4	BIDU	11.0	121.0
5	BRNW	12.0	144.0
6	CARB	15.0	225.0
7	CCIH	0.8	0.6
8	CCOI	0.0	0.0
9	JRJC	1.0	1.0
10	WIFI	23.0	529.0
	SUM	88.1	1445.7
	norm		38.0

The taxicab norm (yellow) is just the total value of the portfolio. The Euclidean norm is in pink.

Let's consider the "direction" of this portfolio. We normalize this vector by dividing by 88.1 for the taxicab norm and by 38.0 for the Euclidean norm:

		$	%
		A	weights
1	AGTK	20.0	0.23
2	AKAM	0.3	0.00
3	BCOR	5.0	0.06
4	BIDU	11.0	0.12
5	BRNW	12.0	0.14
6	CARB	15.0	0.17
7	CCIH	0.8	0.01
8	CCOI	0.0	0.00
9	JRJC	1.0	0.01
10	WIFI	23.0	0.26
	SUM	88.1	1.00



		$		
		A	squared	normalized
1	AGTK	20.0	400.0	0.53
2	AKAM	0.3	0.1	0.01
3	BCOR	5.0	25.0	0.13
4	BIDU	11.0	121.0	0.29
5	BRNW	12.0	144.0	0.32
6	CARB	15.0	225.0	0.39
7	CCIH	0.8	0.6	0.02
8	CCOI	0.0	0.0	0.00
9	JRJC	1.0	1.0	0.03
10	WIFI	23.0	529.0	0.60
	SUM		1445.7	
	norm		38.0	1.00

The former simply consists of the percentages of the stocks within the portfolio.

Warning!

It wouldn't make sense to have the norm of a portfolio of *non-homogeneous* items, such as commodities:

$$< 10000 \text{ tons of wheat}, 20000 \text{ barrels of oil}, ... > ,$$

or currencies:

$$< \$100000,\ ¥1000000,\ ... > .$$

Exercise 4.8.14

When is the norm equal to the sum of the components?

4.9. Lines as parametric curves

Functions may process an input of any nature and produce an output of any nature.

We represent a function diagrammatically as a *black box* that processes the input and produces the output:

$$\begin{array}{ccccc} \text{input} & & \text{function} & & \text{output}\\ x & \mapsto & \boxed{\ f\ } & \mapsto & y \end{array}$$

Convention. We will use the *upper case* letters for the functions the outputs of which are (or may be) multidimensional, such as points and vectors:

$$F,\ G,\ P,\ Q,\ ...$$

We will use the *lower case* letters for the functions with numerical outputs:

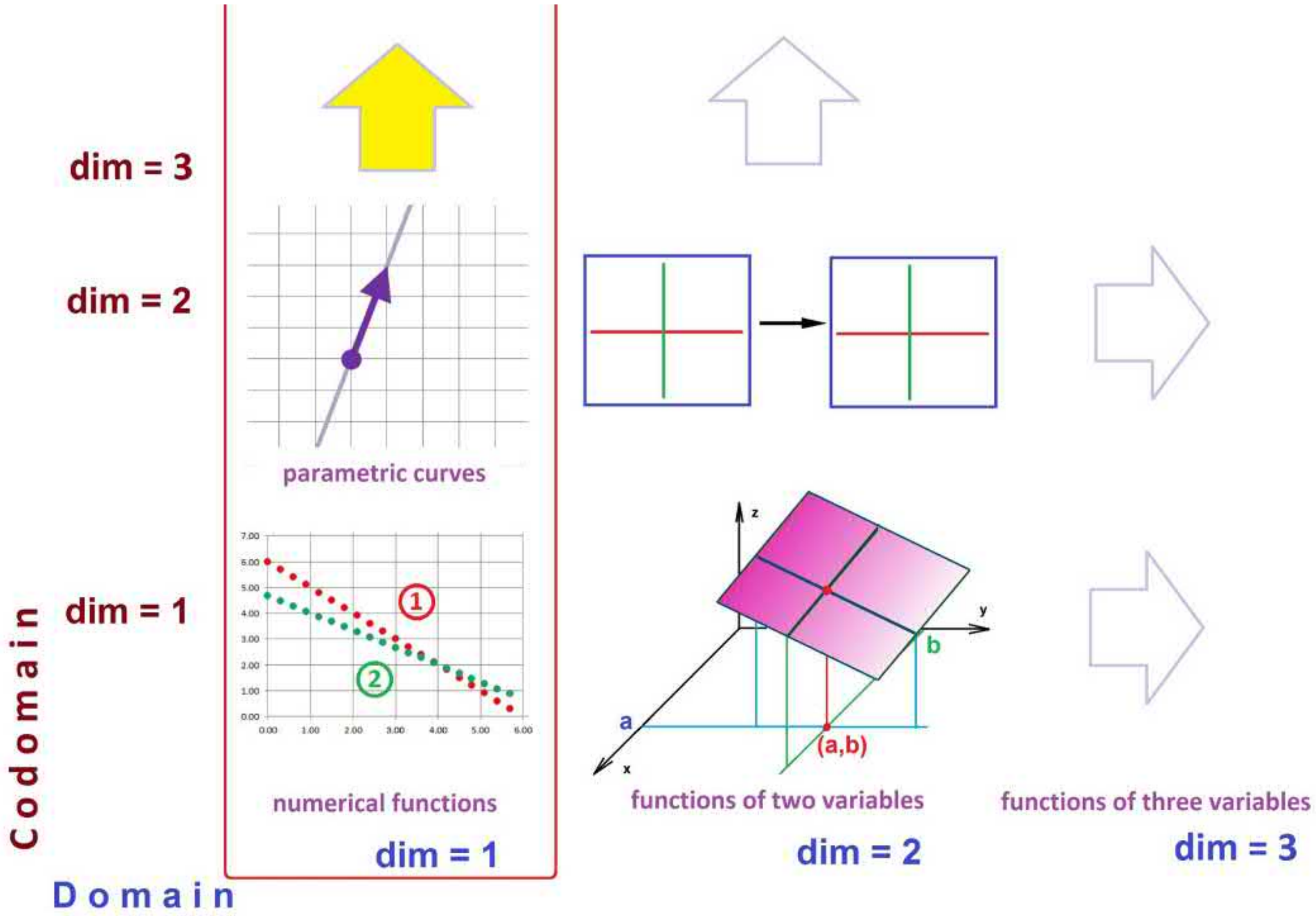
$$f,\ g,\ h,\ ...$$

Functions in multidimensional spaces take points or vectors as the input and produce points or vectors of various dimensions as the output. We can say that the input X is in $\mathbf{R}^n$ and the output $U=F(X)$ of X is in $\mathbf{R}^m$:

$$\begin{array}{llll} F: & P & \mapsto & U\\ & \text{in } \mathbf{R}^n & & \text{in } \mathbf{R}^m \end{array}$$

Then, the domain of such a function is in $\mathbf{R}^n$ and the range (image) is in $\mathbf{R}^m$. The domain can be less than the whole space.

Below we illustrate the four (linear) possibilities for $n=1,2$ and $m=1,2$:

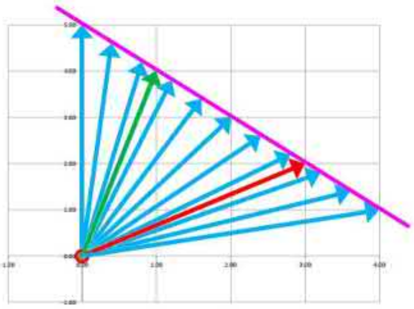


We will concentrate in this section on the first (infinite) column: parametric curves.

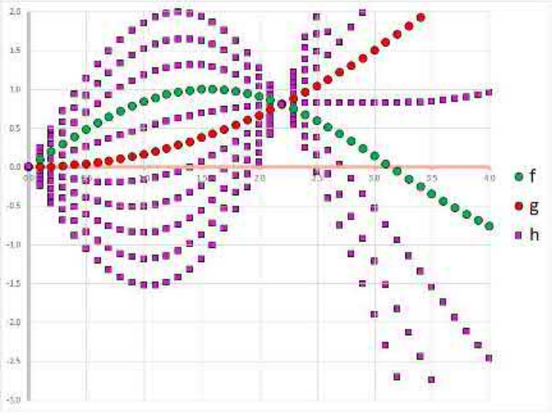
We will refer to as a *parametric curve* to

- any function of the real variable, i.e., the domain lies inside $\mathbf{R}$, and
- with its values in $\mathbf{R}^m$ for some $m = 1, 2, 3....$

Recall from earlier in this chapter how straight lines appear as affine combinations of the two vectors on the plane:



And this is the line in $\mathbf{R}^{40}$ that passes through the two points shown in red and green:



In this section, we will limit ourselves to the interpretation of these functions via *motion*. The independent variable is the *time*, and the value is the *location*.

A *point* is the simplest curve. Such a curve with no motion is provided by a *constant function*.

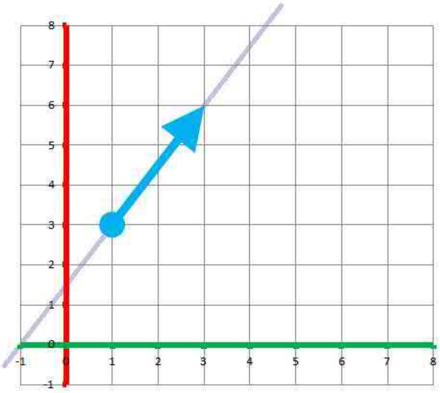
A *straight line* is the second simplest curve.

We start with lines in $\mathbf{R}^2$. We already know how to represent straight lines on the plane:

1. The first method is the *slope-intercept form*:
$$y = mx + b\,.$$
This method excludes the vertical lines! This is too limiting because in our study of curves, there are no preferred directions.
2. The second method is *implicit*:
$$px + qy = r\,.$$
The case of $p \neq 0,\ q = 0$ gives us a vertical line.
3. The third method is *parametric*. It has a dynamic interpretation (below).

Example 4.9.1: straight motion

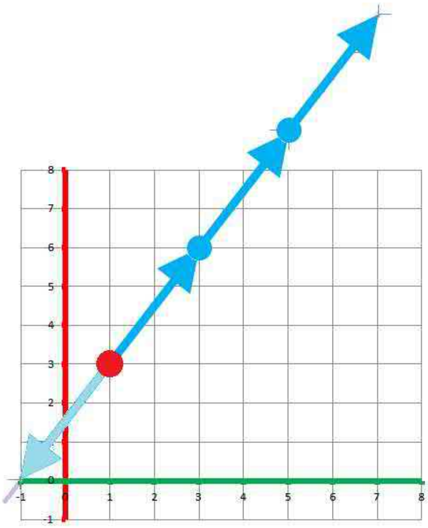
Suppose we would like to trace the line that starts at the point $(1,3)$ and proceeds in the direction of the vector $< 2,3 >$.



We use *motion* as a starting point and as well as a metaphor for parametric curves, as follows. We start moving:

- from the point $P_0 = (1,3)$,
- under a constant velocity of $V =< 2,3 >$.

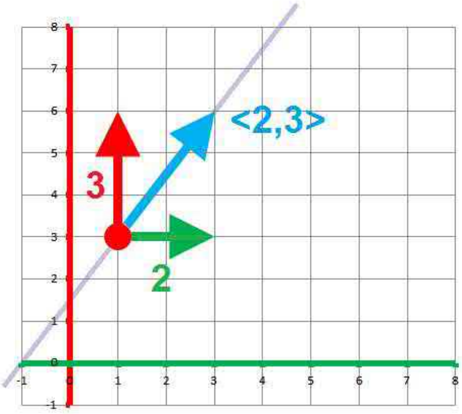
To get the rest of the path, we introduce another variable, time t. When $t = 1,2,3,...$ is increasing incrementally, we have a sequence of locations on the plane:



For the negative t's, we go in the opposite direction.

Let's initially treat x and y separately:

- Horizontal: We move from 1 at 2 feet per second.
- Vertical: We move from 3 at 3 feet per second.



Let's find the formulas for the two. These are two consecutive locations:

$$x(0) = 1,\ x(1) = 3 \ \text{ and } \ y(0) = 3,\ y(1) = 6\,.$$

The functions x and y must be linear:

$$x(t) = 1 + 2t \ \text{ and } \ y(t) = 3 + 3t\,.$$

Combined, this is a parametric curve. Now, let's translate these formulas into the language of vectors.

In terms of vectors, if we are at point P now, we will be at point $P+V$ after one second. For example, we are at $P_1 = P_0 + V = (1,3)+ < 2,3 >= (3,6)$ at time $t = 1$. We define this function:

$$P : \mathbf{R} \to \mathbf{R}^2\,.$$

It is made of two numerical functions:

$$P(t) = (x(t), y(t))\,.$$

We already have two points on our parametric curve P:

$$P(0) = P_0 = (1,3) \ \text{ and } \ P(1) = P_1 = (3,6)\,.$$

What is its formula?

We need to convert this to vectors:

$$x(t) = 1 + 2t,\ y(t) = 3 + 3t\,.$$

Let's assemble the two coordinate functions into one parametric curve:

$$P(t) = (x(t), y(t)) = (1 + 2t,\ 3 + 3t)\,.$$

This is still not good enough; we'd rather see the P_0 and V in the formula. We continue by using vector algebra:

$$\begin{array}{lll} P(t) & = (1+2t,\ 3+3t) & \text{We undo vector addition.} \\ & = (1,3)+ < 2t,\ 3t > & \text{Then we undo scalar multiplication.} \\ & = (1,3) + t < 2,\ 3 > & \text{And finally we have the answer.} \\ & = P_0 + tV\,. & \end{array}$$

So, the four coefficients, of course, come from the specific numbers that give us P_0 and V.

Exercise 4.9.2

The line is not the graph of the function P but its _____ .

We have discovered a vector representation of a straight uniform motion. The *location* P is given by:

$$P(t) = P_0 + tV\,,$$

where P_0 is the initial location and V is the (constant) *velocity*. Then tV is the *displacement*.

Warning!

One can, of course, move along a straight line at a *variable* velocity.

So, we have:

$$\boxed{\text{position at time } t \ = \ \text{initial position} \ + \ t \cdot \ \text{velocity}}$$

We used this approach for dimension 1; only the context has changed.

The pattern becomes clear. The line starting at the point (a, b) in the direction of the vector $< u, v >$ is represented parametrically as follows:

$$P(t) = (a, b) + t < u, v > \ .$$

Similarly for dimension 3, the line starting at the point (a, b, c) in the direction of the vector $< u, v, w >$ is represented as:

$$P(t) = (a, b, c) + t < u, v, w > \ .$$

And so on.

At the next level, we'd rather have no references to neither the dimension of the space nor the specific coordinates:

Definition 4.9.3: parametric curve of the uniform motion

Suppose P_0 is a point in $\mathbf{R}^m$ and V is a vector. Then the *parametric curve of* *the uniform motion through P_0 with the initial velocity of V* is the following:

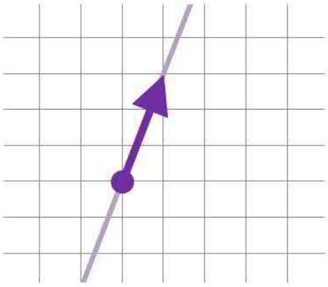
$$\boxed{P(t) = P_0 + tV}$$

Then, the *line through P_0 in the direction of V* is the path (image) of this parametric curve.

Stated for dimension $m = 1$, the definition produces the familiar point-slope form:

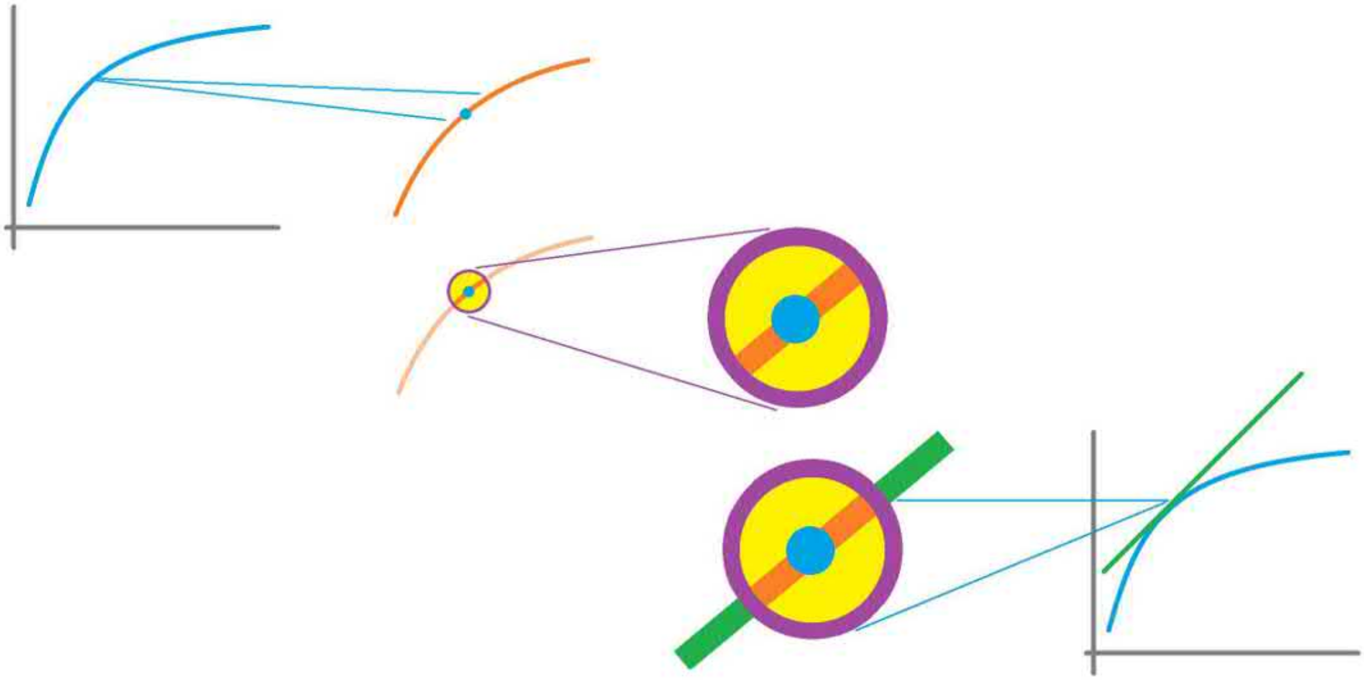
$$P(t) = P_0 + tV\,.$$

Indeed, P_0 is the y-intercept and V is the slope. The rate of change is a single number because the change is entirely within the y-axis. What has changed is the context as there are infinitely many directions in $\mathbf{R}^2$ for change:



That is why the change and the rate of change is a vector.

The importance of straight lines stems from the fact that, under common restrictions, *every* curve is likely to look like a straight line in the short term:

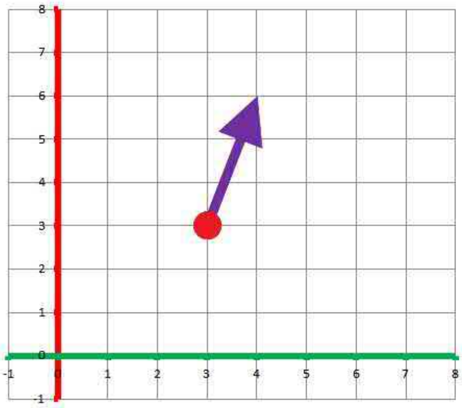


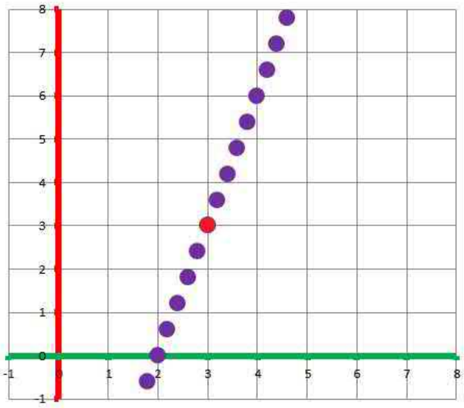
Example 4.9.4: recursive formulas

These are the recursive formulas that give the location as a function of time when the velocity is constant ($k = 0, 1, ...$):

$$\begin{array}{llll} x: & p_{k+1} & = p_k & +v\Delta t \\ y: & q_{k+1} & = q_k & +u\Delta t \end{array}$$

These points are plotted on the right for $p_0 = 3$, $q_0 = 3$, $v = 1$, $u = 3$, $\Delta t = 1/5$:





These quantities are now combined into points and vectors on the plane:

$$P_k = (p_k, q_k),\ V =< v, u > \ .$$

The equations take a vector form too:

$$P_{k+1} = P_k + V\Delta t\,.$$

Exercise 4.9.5

Consider the case when the velocity isn't constant.

Example 4.9.6: price dynamics

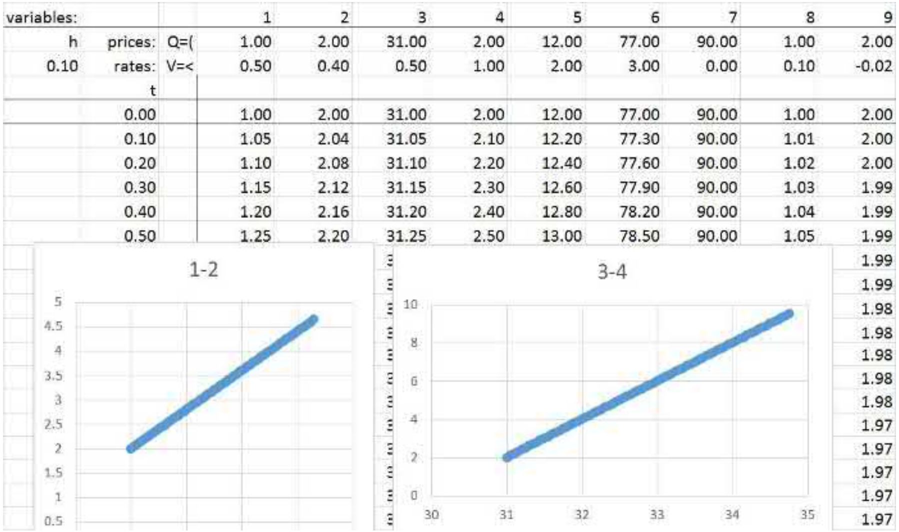
The definition applies to spaces of data. Suppose $\mathbf{R}^m$ is the space of prices (of stocks or commodities); we might have $m = 10,000$.

The prices recorded continuously or incrementally will produce a parametric curve and this curve might be a straight line. This happens when the prices are growing (or declining) *proportionally* but, possibly, at different rates.

Recursive formulas are especially easy to implement with a spreadsheet. In each column, we use the same formula for the kth price:

$$x_k(t + \Delta t) = x_k(t) + v_k\Delta t\,,$$

where v_k is the kth rate of change shown at the top:



variables:			1	2	3	4	5	6	7	8	9
h	prices:	Q=(	1.00	2.00	31.00	2.00	12.00	77.00	90.00	1.00	2.00
0.10	rates:	V=<	0.50	0.40	0.50	1.00	2.00	3.00	0.00	0.10	-0.02
	t										
	0.00		1.00	2.00	31.00	2.00	12.00	77.00	90.00	1.00	2.00
	0.10		1.05	2.04	31.05	2.10	12.20	77.30	90.00	1.01	2.00
	0.20		1.10	2.08	31.10	2.20	12.40	77.60	90.00	1.02	2.00
	0.30		1.15	2.12	31.15	2.30	12.60	77.90	90.00	1.03	1.99
	0.40		1.20	2.16	31.20	2.40	12.80	78.20	90.00	1.04	1.99
	0.50		1.25	2.20	31.25	2.50	13.00	78.50	90.00	1.05	1.99
											1.99
											1.99
											1.98
											1.98
											1.98
											1.98
											1.98
											1.97
											1.97
											1.97
											1.97
											1.97

The table gives us our curve. It lies in the 10,000-dimensional space. Can we visualize such a curve in any way? Very imperfectly. We pick two columns at a time and plot that curve on the plane. Since these columns correspond to the axes, we are plotting a "shadow" (a projection) of our curve cast on the corresponding coordinate plane. They are all straight lines. A similar (short-term) dynamics may be exhibited by other data such as, for example, the vitals of a person:

1. body temperature
2. blood pressure
3. pulse (heart rate)
4. breathing rate

Exercise 4.9.7

Find a parametric representation of the line through two distinct points P and Q.

In the physical space, a straight line is followed by an object when there are no forces at play. Even a constant force leads to acceleration which may change the direction of the motion.

The advantage of the vector approach is that the choice of the coordinate system is no longer a concern!

Example 4.9.8: from relation to parametric

Suppose we have a line given by its relation:

$$y-3=2(x-1)\,.$$

What is its parametric representation?

Let's examine the equation. From its the slope-intercept form we derive:

1. 2 is the slope.
2. $(1,3)$ is the point.

So, let's just move

1. from the point $(1,3)$,
2. along the vector $<1,2>$ every second.

We have:

$$(x,y)=(1,3)+t<1,2>\ .$$

Exercise 4.9.9

What if we move faster?

The example suggests a shortcut for $\mathbf{R}^2$:

$$\boxed{\text{slope }=\frac{\text{rise}}{\text{run}}\ \Longrightarrow\ \text{direction }=<\text{run},\text{rise}>}$$

4.10. The angles between vectors; the dot product

Recall that a Cartesian system pre-measures the space $\mathbf{R}^n$ so that we can do *analytic geometry*:

- ▶ Using the coordinates of points and the components of vectors, we compute distances and angles.

In this chapter, we applied this idea to the distances between points and, therefore, to the magnitudes of vectors. What about the *angles*? Let's first review what we did for dimensions 1 and 2.

Dimension 1 first.

What is the difference between the vectors OP and OQ (P, Q are not equal to O) represented in terms of their components x and x'? There can be only two possibilities:

- If P and Q are on the same side of O then the *directions are the same*,
- If P and Q are on the opposite sides of O then the *directions are the opposite.*

Then the theorem about the *directions for dimension* 1 is stated as follows:The angle between the vectors OP and OQ with components $x \neq 0$ and $x' \neq 0$ is

- 0 when $x \cdot x' > 0$; and
- π when $x \cdot x' < 0$.

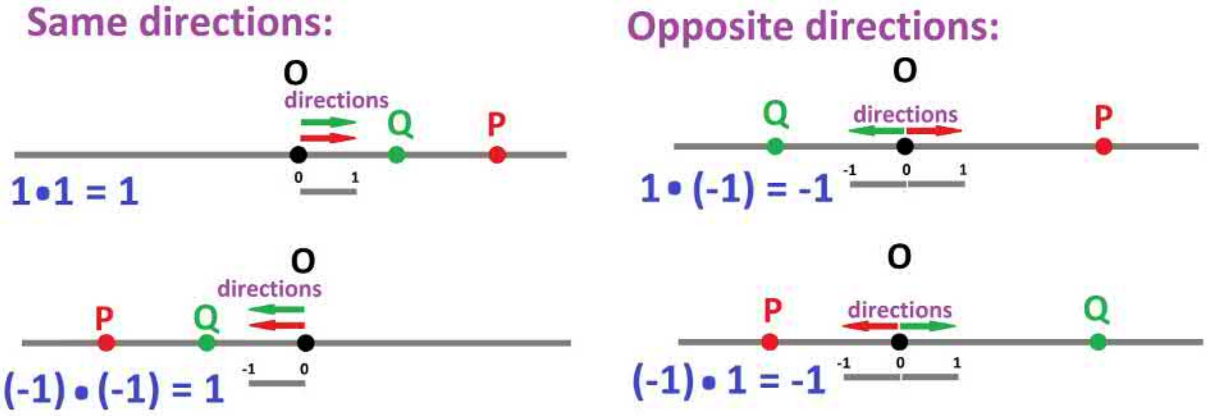
However, we have made some progress since we faced this task. Mainly, it is this realization:

▶ The direction of the vector *is* its normalization, a unit vector.

Indeed, only the directions of the vectors matter and not the sizes! We can then make the same statement but about the unit vectors:

$$\frac{x}{|x|} \text{ and } \frac{x'}{|x'|} .$$

The advantage is that they can only take two possible values, 1 and -1, the positive direction and the negative direction. And so does their product:



We can then restate the result: The angle between vectors OP and OQ with components $x \neq 0$ and $x' \neq 0$ is the following:

normalization of x	$\cdot$	normalization of y		product	angle
$\frac{x}{\lvert x\rvert}$	$\cdot$	$\frac{x'}{\lvert x'\rvert}$	$=$	1	0
$\frac{x}{\lvert x\rvert}$	$\cdot$	$\frac{x'}{\lvert x'\rvert}$	$=$	-1	π

Matching these four numbers,

$$0 \mapsto 1 \text{ and } \pi \mapsto -1 ,$$

we realize that this is the cosine:

$$\cos 0 = 1 \text{ and } \cos \pi = -1 .$$

We then have a new version of our theorem:

Theorem 4.10.1: Angles for Dimension 1

If θ is the angle between vectors OP and OQ in $\mathbf{R}$ *with components $x \neq 0$ and $x' \neq 0$ then*

$$\boxed{\cos \theta = \frac{x}{|x|} \cdot \frac{x'}{|x'|}}$$

Now the coordinate system for dimension 2.

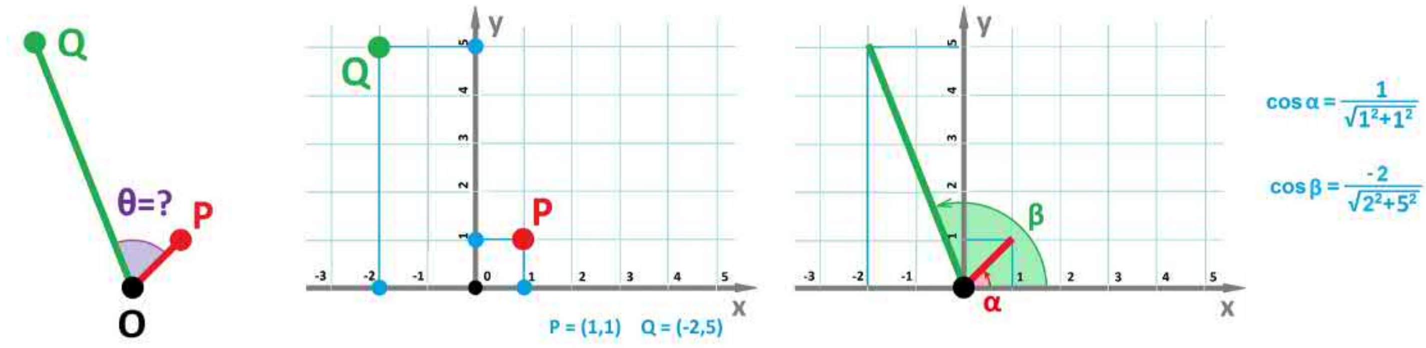
What is the difference between the directions from the origin O toward points P and Q (other than O) represented in terms of their coordinates (x, y) and (x', y')? We are talking about *angle* between the two non-zero vectors:

$$OP =< x, y > \quad \text{and} \quad OQ =< x', y' > \, .$$

Warning!

If one of the vectors is zero, there is no angle because there is no direction.

We know how to find the angles with the x-axis, α and β:



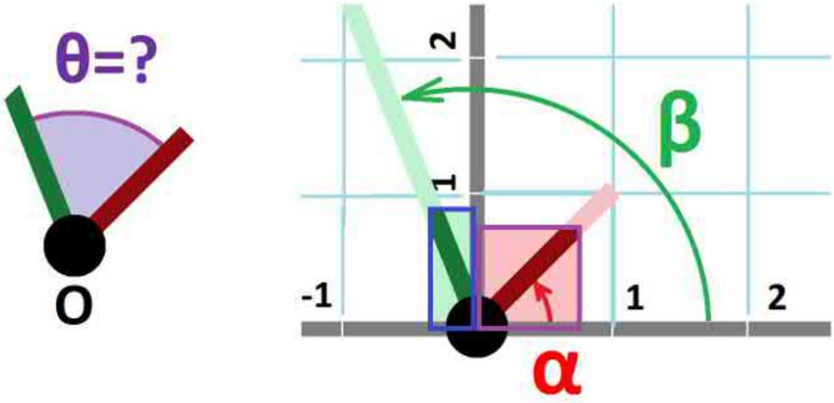
The angle we are looking for is:

$$\theta = \widehat{QOP} = \beta - \alpha \, .$$

The cosine of this angle can be found from the trigonometric functions of these two angles according to the following formula:

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

Let's exclude the magnitudes of the vectors from consideration:



Instead of the original vectors OP and OQ, we look their *normalizations*, U and V, respectively:

$$\begin{aligned} OP =< x, y > \quad &\Longrightarrow \quad U = \frac{< x, y >}{|| < x, y > ||} &= \left\langle \frac{x}{|| < x, y > ||}, \frac{y}{|| < x, y > ||} \right\rangle \\ OQ =< x', y' > \quad &\Longrightarrow \quad V = \frac{< x', y' >}{|| < x', y' > ||} &= \left\langle \frac{x}{|| < x', y' > ||}, \frac{y}{|| < x', y' > ||} \right\rangle \end{aligned}$$

The sines and cosines of these angles are found in terms of the four components of these two vectors U and V. These sines and cosines are exactly these components because the magnitude of the vector and, therefore, the hypotenuse is 1 in either case:

$$\begin{aligned} \cos\alpha &= \frac{x}{|| < x, y > ||} & \sin\alpha &= \frac{y}{|| < x, y > ||} \\ \cos\beta &= \frac{x'}{|| < x', y' > ||} & \sin\beta &= \frac{y'}{|| < x', y' > ||} \end{aligned}$$

Therefore, according to the formula, we have:

$$\begin{aligned}\cos\theta &= \frac{x}{||<x,y>||}\cdot\frac{x'}{||<x',y'>||}+\frac{y}{||<x,y>||}\cdot\frac{y'}{||<x',y'>||}\\ &= \frac{xx'+yy'}{||<x,y>||\cdot||<x',y'>||}\end{aligned}$$

We will have a special name for the numerator of this fraction:

Definition 4.10.2: dot product

The *dot product* of vectors $<x,y>$ and $<x',y'>$ in $\mathbf{R}^2$ is defined by:

$$\boxed{<x,y>\cdot<x',y'>=xx'+yy'}$$

Thus, the dot product is computed, as other vector operations, componentwise.

We now re-state our theorem about the directions:

Theorem 4.10.3: Angles for Dimension 2

If θ is the angle between vectors $A\neq 0$ and $B\neq 0$ in $\mathbf{R}^2$, then:

$$\boxed{\cos\theta=\frac{A\cdot B}{||A||\ ||B||}}$$

Warning!

It makes sense not to put "$\cdot$" in the denominator to avoid confusion.

The presence of the magnitudes in the denominator suggests (to be proven later) that the result is, as expected, depends only on the directions.

Example 4.10.4: simple vectors

Let's test the theorem on simple vectors.

First the two basis vectors:

$$i=<1,0>,\ j=<0,1>\implies i\cdot j=1\cdot 0+0\cdot 1=0\,.$$

Indeed, they are perpendicular and $\cos\pi/2=0$. Similarly,

$$<1,1>\cdot<-1,1>=1\cdot(-1)+1\cdot 1=0\,.$$

However,

$$<1,0>\cdot<1,1>=1\cdot 1+0\cdot 1=1\,.$$

To see the correct angle of 45 degrees, we apply the formula from the theorem:

$$\cos\theta=\frac{<1,0>\cdot<1,1>}{||<1,0>||\ ||<1,1>||}=\frac{1}{1\ \sqrt{2}}=\frac{\sqrt{2}}{2}\,.$$

The following is a very convenient result:

Corollary 4.10.5: Right Angle, Zero Dot Product

Two non-zero vectors are perpendicular if and only if their dot product is zero; i.e.,

$$\boxed{A \perp B \iff A \cdot B = 0}$$

Example 4.10.6: lines

Suppose we have two lines given by their relations:

$$y - 3 = 2(x - 1) \ \text{ and } \ y + 1 = -3(x - 3)\,.$$

What is the angle θ between them?

Do we need their parametric representations? No, just the direction vectors. The slope of the first is 2, so we can choose the direction vector to be $V =< 1, 2 >$. The slope of the second is -3, so we can choose the direction vector to be $U =< 1, -3 >$. Therefore,

$$\cos\theta = \frac{< 1, 2 > \cdot < 1, -3 >}{|| < 1, 2 > ||\ || < 1, -3 > ||} = \frac{1 - 6}{\sqrt{5}\,\sqrt{10}} = -\frac{5}{\sqrt{50}}\,.$$

We start to climb the dimensions.

Definition 4.10.7: dot product

The *dot product* of vectors $< x, y, z >$ and $< x', y', z' >$ in $\mathbf{R}^3$ is defined by:

$$\boxed{< x, y, z > \cdot < x', y', z' >= xx' + yy' + zz'}$$

The dot product is componentwise operation:

$$\begin{array}{rccccc} A & =< & x, & y, & z & > \\ \cdot & & & & & \\ B & =< & u, & v, & w & > \\ \hline A \cdot B & = & x \cdot u+ & y \cdot v+ & z \cdot w & \end{array} \qquad A \cdot B = \begin{bmatrix} x \\ y \\ z \end{bmatrix} \cdot \begin{bmatrix} u \\ v \\ w \end{bmatrix} = \begin{array}{l} x \cdot u+ \\ y \cdot v+ \\ z \cdot w \end{array}$$

Our theorem about the directions remains valid:

Theorem 4.10.8: Angles for Dimension 3

If θ is the angle between vectors $A \neq 0$ and $B \neq 0$ in $\mathbf{R}^3$, then:

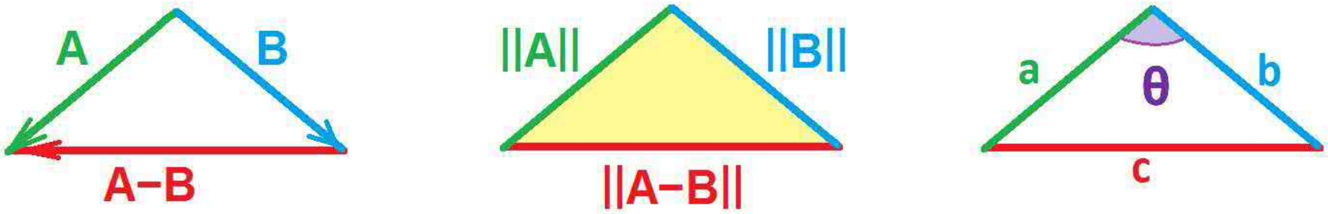
$$\boxed{\cos\theta = \frac{A \cdot B}{||A||\ ||B||}}$$

Proof.

Instead of trigonometric formulas we used for case $n = 2$, we will rely on the *algebraic properties of the dot product*. We start with the *Law of Cosines* (cosine is what we are looking for anyway) which states:

$$c^2 = a^2 + b^2 - 2ab\cos\theta\,,$$

for any triangle with sides a, b, c and angle θ between a and b.



We interpret the lengths of the sides of the triangle in terms of the lengths of vectors:

$$a = ||A||,\ b = ||B||,\ c = ||A - B||\,.$$

Then we translate the law into the language of vectors:

$$||A - B||^2 = ||A||^2 + ||B||^2 - 2||A||\ ||B||\cos\theta\,.$$

Instead of solving for $\cos\gamma$, we expand the left-hand side:

$$\begin{aligned} ||A - B||^2 &= (A - B)\cdot(A - B) && \text{Normalization.}\\ &= A\cdot A + A\cdot(-B) + (-B)\cdot A + (-B)\cdot(-B) && \text{Distributivity.}\\ &= ||A||^2 - 2A\cdot B + ||B||^2 && \text{Associativity and Normalization.} \end{aligned}$$

The Law of Cosines then takes the following form:

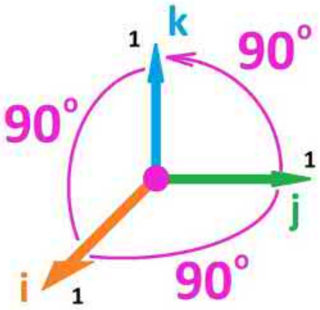
$$||A||^2 - 2A\cdot B + ||B||^2 = ||A||^2 + ||B||^2 - 2||A||\ ||B||\cos\theta\,.$$

Now we cancel the repeated terms in the two sides of the equation and obtain the following:

$$-2A\cdot B = -2||A||\ ||B||\cos\theta\,.$$

Example 4.10.9: basis vectors

It is once again easy to confirm that the basis vectors are perpendicular to each other:

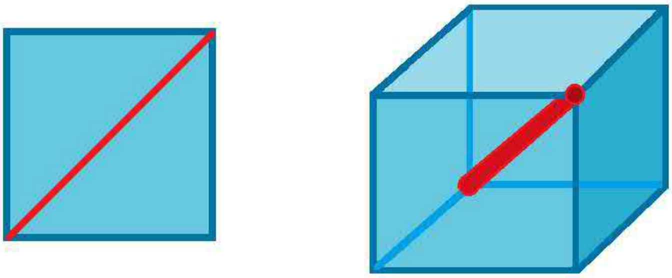


The 1 and 0 are mismatched:

$$\begin{aligned} i\cdot j &=< 1,0,0 > \cdot < 0,1,0 > &&= 1\cdot 0 + 0\cdot 1 + 0\cdot 0 &&= 0\\ j\cdot k &=< 0,1,0 > \cdot < 0,0,1 > &&= 0\cdot 0 + 1\cdot 0 + 0\cdot 1 &&= 0\\ k\cdot i &=< 0,0,1 > \cdot < 1,0,0 > &&= 0\cdot 1 + 0\cdot 0 + 1\cdot 0 &&= 0 \end{aligned}$$

Example 4.10.10: diagonals

The angle between the sides and the diagonal in a square is 45 degrees. Now, what is the angle between the diagonal of a *cube* and any of its edges? Try to guess from the picture if the angle is 45 degrees:



A hard trigonometric problem is solved easily with the dot-product.

First we choose vectors to represent the edges: the three basis vectors for the outside edges and $A =< 1,1,1 >$ for the diagonal. We have for the angle:

$$\cos\theta = \frac{< 1,1,1 > \cdot < 1,0,0 >}{|| < 1,1,1 > || \, || < 1,0,0 > ||} = \frac{1}{\sqrt{3}}\,.$$

What is this angle? We only know that

$$\frac{1}{\sqrt{3}} < \frac{1}{\sqrt{2}}\,.$$

Because cosine is a decreasing function, it follows that this angle is larger than $\pi/4$. There is more room for maneuver than on the plane!

Exercise 4.10.11

Find all the angles from the center of a cube to its corners.

Can we make sense of directions and angles in $\mathbf{R}^n$?

We previously "extrapolated" the definition of the magnitude (and the distance before that) to produce the Euclidean norm:

dimension	vector	components	norm
1	A	a	$\|A\| = \|a\|$
2	A	$< a,b >$	$\|\|A\|\|^2 = a^2 + b^2$
3	A	$< a,b,c >$	$\|\|A\|\|^2 = a^2 + b^2 + c^2$
...	...	...	...
n	A	$< a_1, a_2, ..., a_n >$	$\|\|A\|\|^2 = a_1^2 + a_2^2 + ... + a_n^2$

We do the same for the definition of the dot product:

dimension	vectors	components	dot product
1	A	a	
	B	u	$A \cdot B = a \cdot u$
2	A	$< a, b >$	
	B	$< u, v >$	$A \cdot B = a \cdot u + b \cdot v$
3	A	$< a, b, c >$	
	B	$< u, v, w >$	$A \cdot B = a \cdot u + b \cdot v + c \cdot w$
...	...	...	...
n	A	$< a_1, a_2, ..., a_n >$	
	B	$< b_1, b_2, ..., b_n >$	$A \cdot B = a_1 \cdot b_1 + a_2 \cdot b_2 + ... + a_n \cdot b_n$

Definition 4.10.12: dot product in n-space

The *dot product* of vectors A and B is defined to be the sum of the products of their components:

$$\boxed{\begin{array}{rlllll} A & =< a_1, & a_2, & ..., & a_n & > \\ B & =< b_1, & b_2, & ..., & b_n & > \\ A \cdot B & = \quad a_1 b_1 & +a_2 b_2 & +... & +a_n b_n & \end{array}}$$

In the sigma notation:

$$\boxed{A \cdot B = \sum_{k=1}^{n} a_k b_k}$$

Exercise 4.10.13

What is the angle between the diagonal of a 4-dimensional cube and any of its edges?

Exercise 4.10.14

What is the angle between the diagonal of a n-dimensional cube and any of its edges? What value does the angle approach when n approaches infinity?

Below we see how this new operation (last row) compares with the other vector operations:

vector addition	A	$+$	B	$=$	C
	vector		vector		vector
scalar multiplication	c	$\cdot$	A	$=$	C
	number		vector		vector
dot product	A	$\cdot$	B	$=$	s
	vector		vector		number

Warning!

The last two might be confusing without a *context*; for example, consider the three possible meanings of the following:

$$0 \cdot A = 0\,.$$

Let's consider the *properties of the dot product.*

If we just set $Y = X$, we have the so-called *Normalization*:

$$\boxed{||X||^2 = X \cdot X}$$

One can, therefore, recover the magnitude from the dot product just as before. So, once we have the dot product, we don't need to introduce the magnitude independently.

The *Positivity* of the norm then requires a similar property for the dot product:

$$\boxed{V \cdot V \geq 0; \quad \text{and} \quad V \cdot V = 0 \iff V = 0}$$

Next, *Commutativity* or *Symmetry*:

$$\boxed{A \cdot B = B \cdot A}$$

This means that the angle is *between* A and B; i.e., the same from A to B as from B to A.

Next, *Associativity*:

$$\boxed{(kA) \cdot B = k(A \cdot B) = A \cdot (kB)}$$

So, the effect of stretching on the dot product is a multiple and the angle doesn't change for $k > 0$ or is replaced with the opposite when $k < 0$.

We can see now that only the normalizations matter for the angle between two vectors. We just choose these values for k in the last formula:

$$\frac{1}{||A||\;||B||},\ \frac{1}{||A||},\ \text{ and }\ \frac{1}{||B||}$$

to re-write our formula from the last section:

$$\cos\theta = \frac{A \cdot B}{||A||\;||B||} = \frac{1}{||A||\;||B||} A \cdot B = \frac{1}{||A||} A \cdot \frac{1}{||B||} B = \frac{A}{||A||} \cdot \frac{B}{||B||}\,.$$

The result suggests that *the dot product is independent from the coordinate system.* Certainly, this system is just a tool that we introduce into the space the geometry of which we study, and we don't expect that changing the components of vectors will also change the distances and the angles. But it's also true in $\mathbf{R}^n$!

Next, *Distributivity* or *Linearity*:

$$\boxed{A \cdot (B + C) = A \cdot B + A \cdot C}$$

Treated componentwise, the Commutativity, Associativity, Distributivity properties for the dot product of vectors follow from the Commutativity, Associativity, Distributivity for numbers. For example, this is the whole proof of the Commutativity for $n = 2$:

$$\begin{bmatrix} a \\ b \end{bmatrix} \cdot \begin{bmatrix} u \\ v \end{bmatrix} = au + bv = ua + vb = \begin{bmatrix} u \\ v \end{bmatrix} \cdot \begin{bmatrix} a \\ b \end{bmatrix}$$

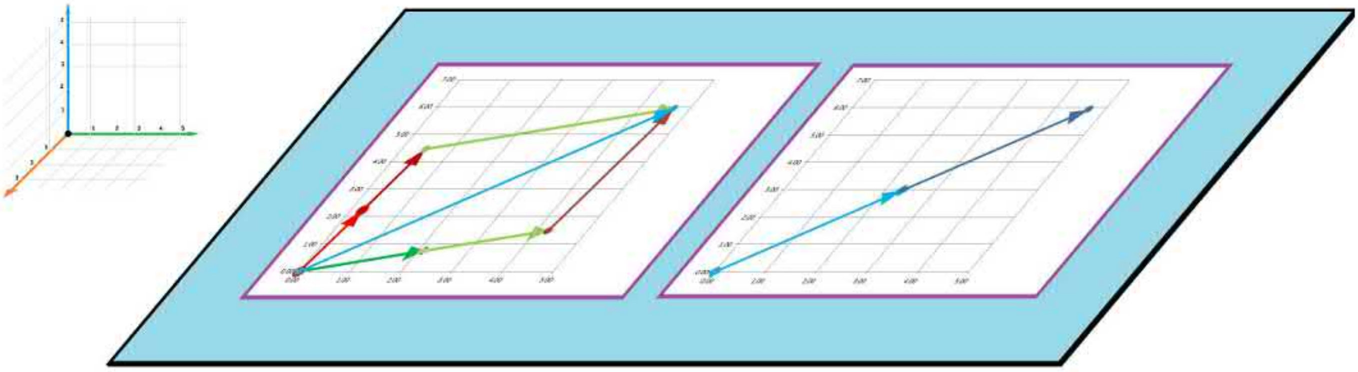
Once again, these properties allow us to use the usual algebraic manipulation steps for numbers as long as the expressions make sense to begin with.

How do we understand the geometry of this made-up space? After all, when $n > 3$, there is no reality test for this concept and we can't verify the formulas we are to use!

We have come to understand the distances in $\mathbf{R}^n$ and now ask:

▶ What is the meaning of the angle between two vectors A and B in $\mathbf{R}^n$?

The answer is to reduce the multidimensional case to the case of $n = 2$. Indeed, every two vectors define a plane and this plane has the same vector algebra operations – including the dot product – as the ambient space $\mathbf{R}^n$:



The Distributivity will require 3 dimensions. In the meantime, the plane has the well-understood Euclidean geometry: The lengths of vectors and the angles between vectors can be measured.

The definition is abstract but it matches the lower dimensions $n = 1, 2, 3$:

Definition 4.10.15: angles for dimension n

The angle θ between vectors $A \neq 0$ and $B \neq 0$ in $\mathbf{R}^n$ is defined to satisfy:

$$\boxed{\cos\theta = \frac{A \cdot B}{||A||\ ||B||}}$$

For the record, we summarize the rules of the dot product:

Theorem 4.10.16: Axioms of Inner Product Space

The dot product in $\mathbf{R}^n$ satisfies the following three properties for all vectors U, V, W and all scalars a, b:

- ***Symmetry:*** $U \cdot V = V \cdot U$.
- ***Linearity:*** $U \cdot (aV + bW) = a(U \cdot V) + b(U \cdot W)$.
- ***Positive-definiteness:*** $V \cdot V \geq 0$; *and* $V \cdot V = 0 \iff V = 0$.

We have added a third vector operation to the toolkit but vector algebra still looks like that of numbers!

From the inequality

$$|\cos\theta| \leq 1\,,$$

we derive the following.

Corollary 4.10.17: Cauchy Inequality

For any pair of vectors $A \neq 0$ *and* $B \neq 0$ *in* $\mathbf{R}^n$*, we have:*

$$\boxed{|A \cdot B| \leq ||A||\ ||B||}$$

In other words, if we rotate one of the vectors, the dot product reaches its maximum when they are parallel to each other.

So, what *is* the dot product of two vectors? Two partial answers:

1. The dot product is the cosine of the angle when the vectors are unit vectors.
2. The dot product is the square of the magnitude when the vectors are equal.

A complete, geometric answer may simply be our formula:

$$\boxed{A \cdot B = ||A||\ ||B|| \cos\theta}$$

Example 4.10.18: inner product for taxicab?

Once can see how the dot product emerged from the Euclidean metric and norm:

dimension	vectors	components	Euclidean norm and dot product
n	A	$< a_1, a_2, ..., a_n >$	$\|\|A\|\|^2 = a_1 \cdot a_1 + a_2 \cdot a_2 + ... + a_n \cdot a_n$
n	A	$< a_1, a_2, ..., a_n >$	
	B	$< b_1, b_2, ..., b_n >$	$A \cdot B = a_1 \cdot b_1 + a_2 \cdot b_2 + ... + a_n \cdot b_n$

But what about the taxicab metric? We can suggest the following candidate for the inner product:

dimension	vectors	components	taxicab norm and inner product?
n	A	$< a_1, a_2, ..., a_n >$	$\|\|A\|\| = \|a_1\| + \|a_2\| + ... + \|a_n\|$
n	A	$< a_1, a_2, ..., a_n >$	
	B	$< b_1, b_2, ..., b_n >$	$A \cdot B = \sqrt{\|a_1\| \cdot \|b_1\|} + \sqrt{\|a_2\| \cdot \|b_2\|} + ... + \sqrt{\|a_n\| \cdot \|b_n\|}$

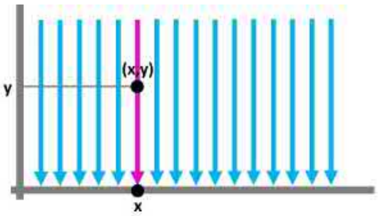
Unfortunately, the Linearity fails! We, therefore, will be unable to deal with angles in this space.

Exercise 4.10.19

Prove the last statement for $n = 2$.

4.11. Projections and decompositions of vectors

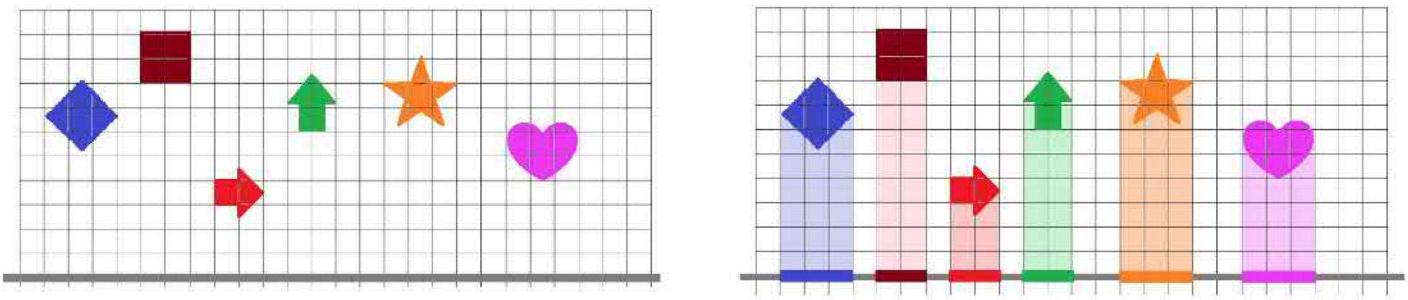
To find the x-coordinate of a point on the xy-plane, we go vertically from that points until we reach the x-axis:



It's a familiar function:

$$F : \mathbf{R}^2 \to \mathbf{R},\ < x, y > \mapsto x\,.$$

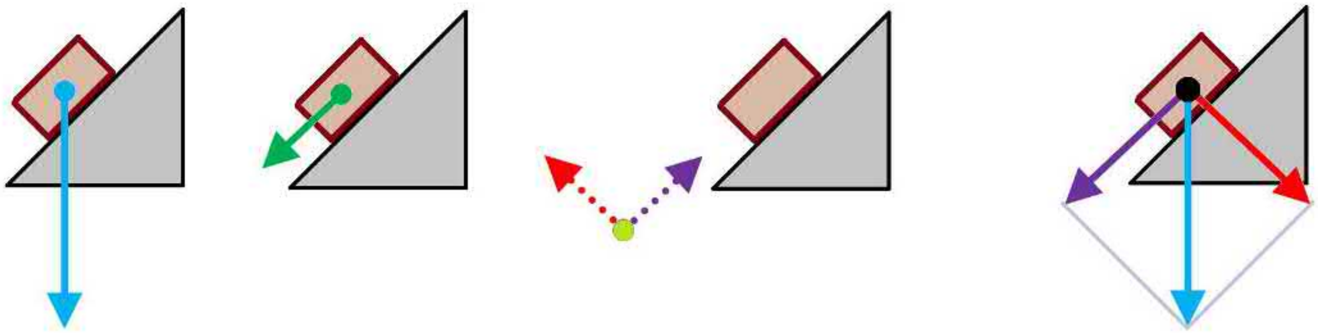
The result resembles *shadows* cast by points, vectors, or subsets onto the x-axis with the light cast from above (or from the right for the y-axis):



It is called the *projection* of the point on the x-axis. Same for the y-axis.

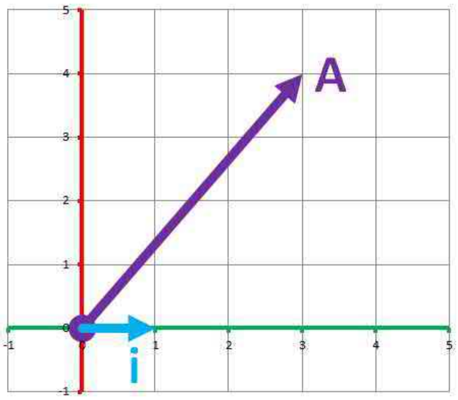
Similarly, the projection of a *vector* on the x-axis gives its x-component. If several coordinate systems co-exit, a transition from one to another will require expressing the new coordinates of a point or the new components of a vector in terms of the old. We do that one axis or basis vector at a time.

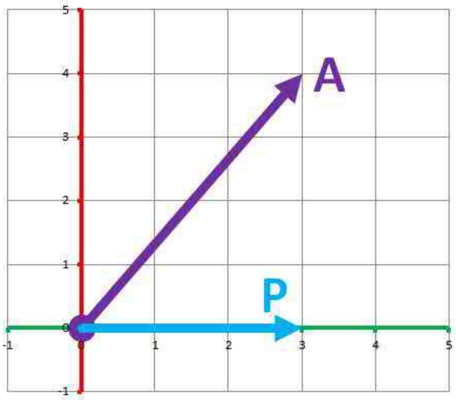
As a reason for this study here is the example of compound motion from earlier in this chapter, an object sliding down a slope:



In order to concentrate on the relevant part of the motion, one would choose the first basis vector to be parallel to the surface and the second perpendicular.

For vectors, the projection on the x-axis will require choosing a representative vector on it; it's i. Any vector A is expressed in terms of i by finding A's component P on the x-axis. For example, below the component is 3:



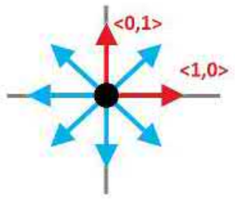


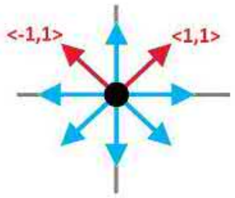
Specifically, we have a linear combination of i and j:

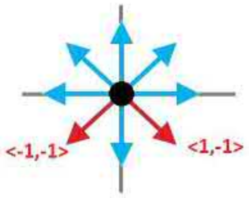
$$A = < 3, 4 > = 3i + 4j \,.$$

What if instead of i we have an arbitrary vector V?

But first, we consider a shortcut for finding a *perpendicular vector*:







The problem of rotating a given vector V in the plane through $\pi/2$ has an easy solution:

Theorem 4.11.1: Orthogonality in 2-space

For any vector $V = < u, v >$ on the plane, the following vector, called a normal vector of V, denoted by $V^\perp$, has the angle of $\pi/2$ with V:

$$\boxed{V^\perp = < u, v >^\perp = < -v, u >}$$

Proof.

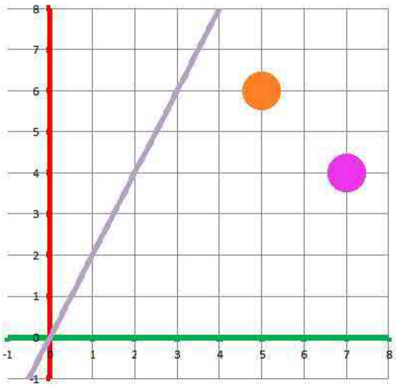
This is easy to confirm:

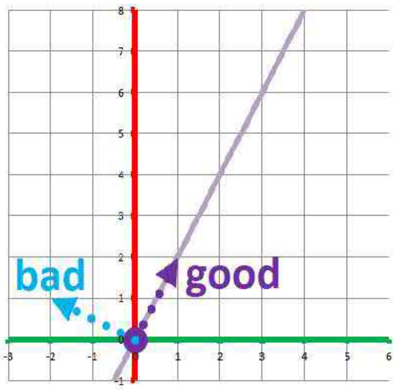
$$V \cdot V^\perp = < u, v > \cdot < -v, u > = u(-v) + vu = 0 \,.$$

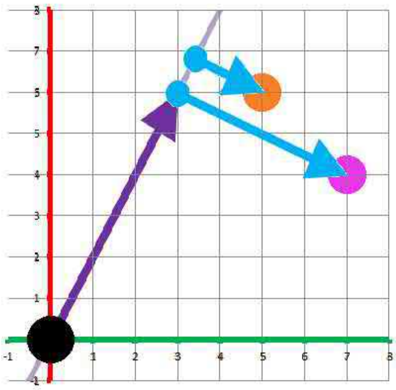
We have then a special operation on vectors in dimension 2. Every vector has exactly two normal *unit* vectors.

Example 4.11.2: investing advice

An initially investment advice might be very simple, for example: hold the proportion of stocks and bonds 1-to-2. If x is the amount of stocks and y is the amount of bonds in it, the "ideal" portfolios lie on the line $y = 2x$, i.e., they are the ends of vectors that are multiples of $V = < 2, 1 >$:







How do we determine how close is each portfolio to the ideal? We can use V as a measuring stick. What about the bad? We can find a vector perpendicular to V via the last theorem. Though not unit vectors, these two vectors will give the good and the bad components of any portfolio:

$$g =< 1, 2 > \quad \text{and} \quad b =< -2, 1 > \ .$$

So, we need to represent every portfolio as a linear combination of these two. For example:

$$< 5, 6 >= p < 1, 2 > +q < -2, 1 > \quad \text{and} \quad < 7, 4 >= u < 1, 2 > +v < -2, 1 > \ .$$

In other words, we need to solve two systems of linear equations:

$$\begin{cases} p - 2q = 5 \\ 2p + q = 6 \end{cases} \quad \text{and} \quad \begin{cases} u - 2v = 7 \\ 2u + v = 4 \end{cases}$$

These are the solutions:

$$p = 17/5, \ q = -4/5 \ \text{ and } \ u = 4, \ v = -2 \, .$$

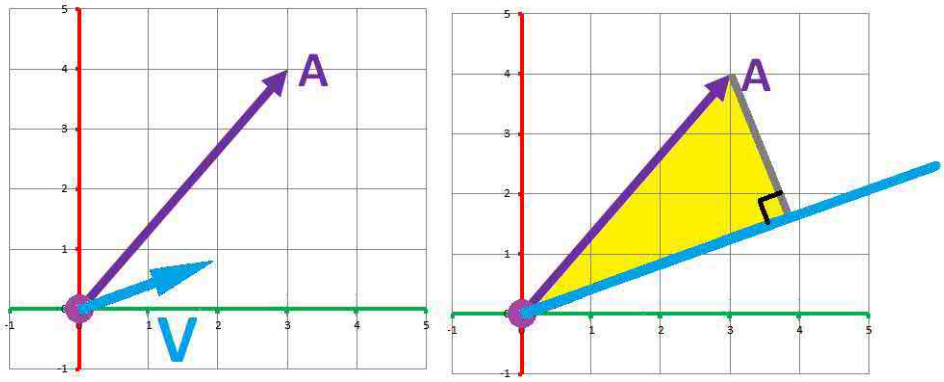
The second numbers in these pairs indicate how far it is from the ideal. The second portfolio is farther.

What if we have many investment vehicles in each portfolio? For example, we may assume that these portfolios live in $\mathbf{R}^7$:

					$	$	$
		%	weights	unit	25	Portfolios	
		V	V/100	V/||V||	line	A1	A2
	magnitude	41.23	0.41	1.00	10.31	10.39	5.20
1	growth stocks	20	0.20	0.49	5	4	0
2	value stocks	10	0.10	0.24	3	2	3
3	corporate bonds	5	0.05	0.12	1	4	2
4	minicipal bonds	15	0.15	0.36	4	2	1
5	federal bonds	25	0.25	0.61	6	8	0
6	real estate	15	0.15	0.36	4	0	3
7	cash	10	0.10	0.24	3	2	2
	SUM	100	1.00	2.43	25	22	11

Unfortunately, the shortcut of the last theorem is not available anymore. We will need a further analysis.

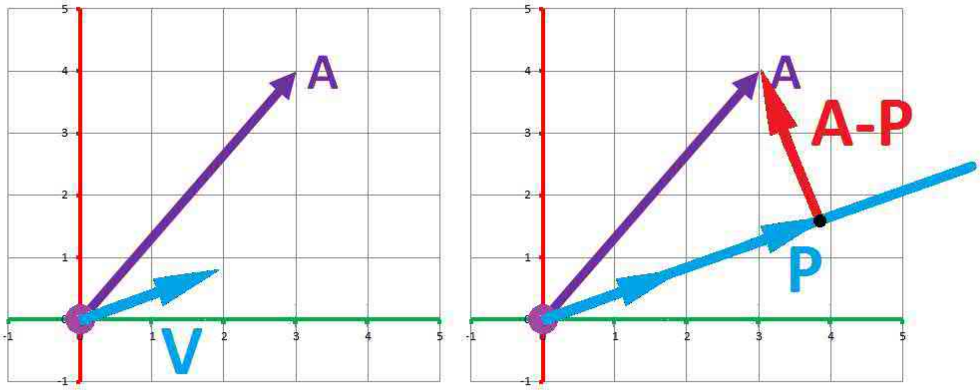
This is what a projection on a line defined by a vector looks like:



The question becomes: How much does A "protrude" in the direction of V? More precisely:

- ▶ Vector A is expressed in terms of V by finding A's projection P on the line created by V.

So, every vector A needs to be expressed as a linear combination of V and some other vector perpendicular to V:



If P is the projection, what's the other vector? We simply subtract:

$$A = P + (A - P)\,.$$

The vector we are after is described indirectly:

Definition 4.11.3: orthogonal projection

Suppose A and $V \neq 0$ are two vectors in $\mathbf{R}^n$. Then the *orthogonal projection of* A *onto* V is a vector P that satisfies the following:

1. P is parallel to V.
2. $A - P$ is perpendicular to V.

Let's find an explicit formula.

First, "parallel" simply means a multiple! Therefore, the first property means that there is a number k – this is the one we are looking for – such that:

$$P = kV\,.$$

The second property is expressed in terms of the dot product:

$$V \cdot (A - P) = 0\,.$$

We substitute:

$$V \cdot (A - kV) = 0\,,$$

and use *Distributivity* and *Associativity*:

$$V \cdot A - kV \cdot V = 0\,.$$

Next we use *Normalization*:

$$V \cdot A = k||V||^2 \,.$$

Then,

$$k = \frac{V \cdot A}{||V||^2} \,.$$

This is the multiple of V that gives us P. Thus, we have proven the following result:

Theorem 4.11.4: Projection Via Dot Product

The orthogonal projection of a vector A onto a vector $V \neq 0$ is the vector P given by:

$$\boxed{P = \frac{V \cdot A}{||V||^2} V}$$

Notice that the formula – as expected – depends only on the *direction* of V; only unit vectors are involved:

$$P = \left(\frac{V}{||V||} \cdot A \right) \frac{V}{||V||} \,.$$

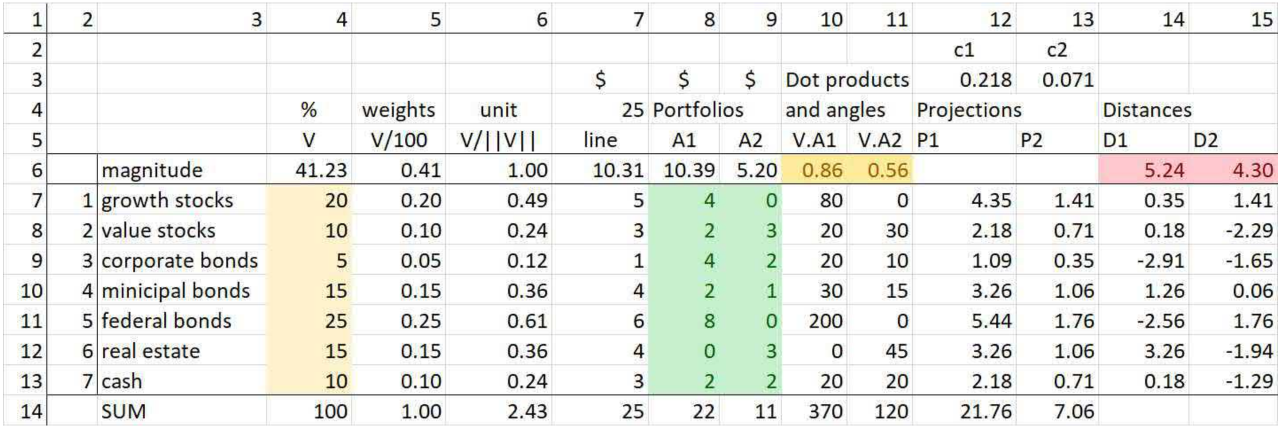
Exercise 4.11.5

Prove the last formula.

Example 4.11.6: investing advice continued

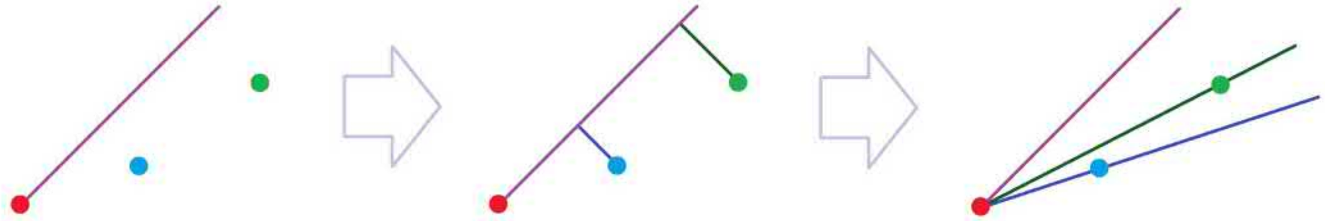
We have 7 instruments in each portfolio. The portfolio *advice* V is shown in column 4 and the two competing portfolios A_1 and A_2 in columns 8 and 9. We carry out the following computations:

1. We find the coordinatewise products of A_1 and A_2 with V (columns 10 and 11) and then by adding those obtain the two dot products (bottoms of the columns).
2. From those two, we find the multiples c_1 and c_2 for the projections according to the last theorem (tops of columns 12 and 13).
3. We use those two to multiply V componentwise to obtain the projections P_1 and P_2 of A_1 and A_2 (columns 12 and 13).
4. The differences of A_1 and A_2 from V are found (columns 14 and 15).



1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
2											c1	c2		
3						$	$	$	Dot products		0.218	0.071		
4			%	weights	unit	25	Portfolios		and angles		Projections		Distances	
5			V	V/100	V/||V||	line	A1	A2	V.A1	V.A2	P1	P2	D1	D2
6		magnitude	41.23	0.41	1.00	10.31	10.39	5.20	0.86	0.56			5.24	4.30
7	1	growth stocks	20	0.20	0.49	5	4	0	80	0	4.35	1.41	0.35	1.41
8	2	value stocks	10	0.10	0.24	3	2	3	20	30	2.18	0.71	0.18	-2.29
9	3	corporate bonds	5	0.05	0.12	1	4	2	20	10	1.09	0.35	-2.91	-1.65
10	4	minicipal bonds	15	0.15	0.36	4	2	1	30	15	3.26	1.06	1.26	0.06
11	5	federal bonds	25	0.25	0.61	6	8	0	200	0	5.44	1.76	-2.56	1.76
12	6	real estate	15	0.15	0.36	4	0	3	0	45	3.26	1.06	3.26	-1.94
13	7	cash	10	0.10	0.24	3	2	2	20	20	2.18	0.71	0.18	-1.29
14		SUM	100	1.00	2.43	25	22	11	370	120	21.76	7.06		

Finally, we compute the magnitudes of those two (tops of columns 14 and 15). We conclude that the *second* portfolio is closer to the line that represents the ideal:



Alternatively, we realize that these distances are proportional to the size of the investment, which skews the conclusions. We then turn to the *angles* instead. Their cosines are computed (tops of columns 10 and 11) with the formula from the last section. We conclude that the *first* portfolio is closer to the line that represents the ideal.

Exercise 4.11.7

What is the set of all portfolios with the total investment of 1?

Chapter 5: Linear operators

Contents

5.1. Where matrices come from

In this chapter, we set aside the geometry of the Euclidean spaces – measuring distances and angles – and concentrate on *pure algebra*.

Let's recall these problems about *mixtures*.

Problem, dimension 1*:* Suppose we have a type of coffee that costs \$3 per pound. How much do we get for \$60?

Let x be the weight of the coffee. Then we have:

$$\text{Setup: } 3x = 60\,. \quad \text{Solution: } x = \frac{60}{3}\,.$$

Problem, dimension 2*:* Suppose the Kenyan coffee costs \$2 per pound and the Colombian coffee costs \$3 per pound. How much of each do you need to have 6 pounds of blend with a total price of \$14?

Let x be the weight of the Kenyan coffee and let y be the weight of Colombian coffee. Then we have:

$$\text{Setup: } \begin{array}{lllll} x & + & y & = 6 \\ 2x & + & 3y & = 14 \end{array}$$

Solution: From the first equation, we derive: $y = 6 - x$. Then substitute it into the second equation: $2x + 3(6 - x) = 14$. Solve the new equation: $-x = -4$, or $x = 4$. Substitute this back into the first equation: $(4) + y = 6$, then $y = 2$.

But it was so much simpler for the former problem! Is it possible to mimic the setup, i.e., the equation, and the solution of the 1-dimensional case for the 2-dimensional case? The existence of vector algebra suggests that it might be possible.

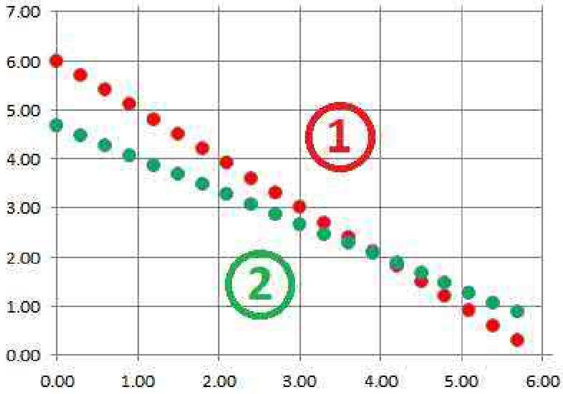
Let's recall the ways we have interpreted the problem.

First: points and lines.

We think of the two equations as equations about the coordinates of points, (x, y), in the plane:

$$\begin{cases} x + y = 6, \\ 2x + 3y = 14. \end{cases}$$

Either equation is a line on the plane. The solution $(x, y) = (4, 2)$ is the point of their intersection:



Second: vectors and their linear combinations.

Let's put the equations in these tables:

$1 \cdot x \ + \ 1 \cdot y \ = \ 6$
$2 \cdot x \ + \ 3 \cdot y \ = \ 14$

The table is split horizontally to reveal the equations. Next, we start to split vertically and realize that we see a componentwise *addition of vectors*:

$1 \cdot x$	$+$	$1 \cdot y$	$=$	6
$2 \cdot x$	$+$	$3 \cdot y$	$=$	14

We have:

$1 \cdot x$	$+$	$1 \cdot y$	$=$	6
$2 \cdot x$		$3 \cdot y$		14

But x's and y's are repeated! We realize that we see a componentwise *scalar multiplication of vectors*:

1 2	$\cdot$	x	$+$	1 3	$\cdot$	y	$=$	6 14

Vectors start to appear. Indeed, our system has been reduced to a single *vector* equation:

$$\begin{bmatrix} 1 \\ 2 \end{bmatrix} x + \begin{bmatrix} 1 \\ 3 \end{bmatrix} y = \begin{bmatrix} 6 \\ 14 \end{bmatrix}$$

We see three vectors of the same dimension 2. This isn't a coincidence. They are of the same nature:

$$\begin{bmatrix}1\\2\end{bmatrix},\ \begin{bmatrix}1\\3\end{bmatrix},\ \begin{bmatrix}6\\14\end{bmatrix}\quad\text{are}\quad\begin{bmatrix}\text{weight (in pounds)}\\ \text{cost (in dollars)}\end{bmatrix}.$$

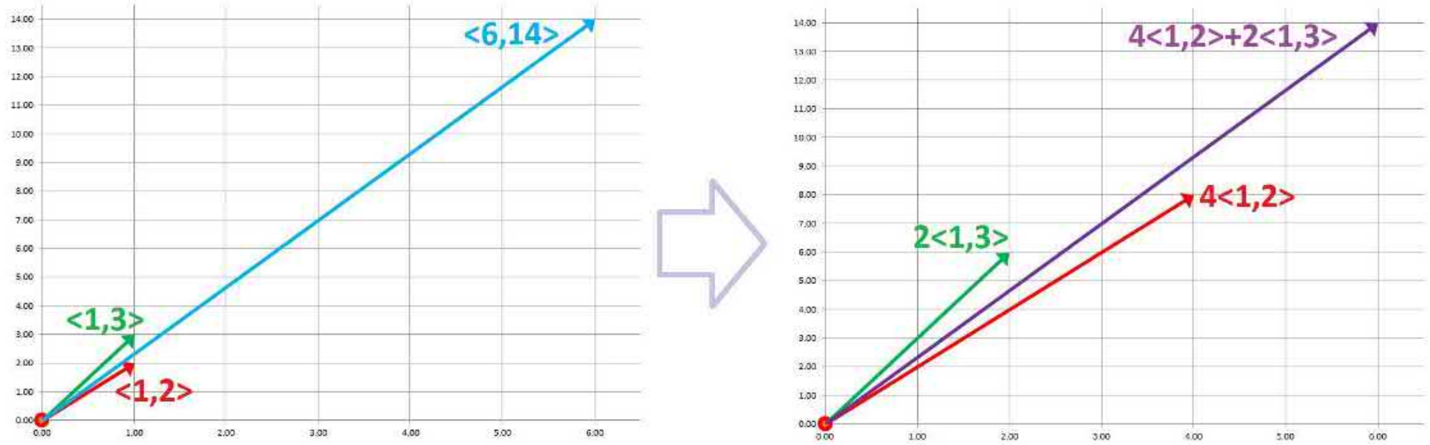
They live in the same space $\mathbf{R}^2$ and, therefore, subject to the operations of vector algebra.

The solution to the system in this interpretation has the following algebraic meaning. We can think of the two equations as a single equation about the coefficients, x and y, of these vectors in the plane:

$$x\begin{bmatrix}1\\2\end{bmatrix}+y\begin{bmatrix}1\\3\end{bmatrix}=\begin{bmatrix}6\\14\end{bmatrix}.$$

Geometrically, we need to find a way to stretch these two vectors so that after adding them the result is the vector on the right. We speak of *linear combinations.*

The setup is on the left followed by a trial-and-error on the right:



So, the new point of view has changed: Instead of the *locations*, we are after the *directions*.

Exercise 5.1.1

Are there other vectors here?

Third: transformations.

Initially, we use points.

Dimension 1 problem:

- A transformation $f:\mathbf{R}\to\mathbf{R}$ is given by

$$f(x)=30x\,.$$

- Solve the equation:

$$f(x)=60\,.$$

Dimension 2 problem:

- A transformation $f:\mathbf{R}^2\to\mathbf{R}^2$ is given by

$$F(x,y)=(x+y,2x+3y)\,.$$

- Solve the equation:

$$F(x,y)=(6,14)\,.$$

Now, we prefer to use vectors. Let's find all 2-dimensional vectors in the equations:

$$\boxed{\begin{array}{rcrcr} 1\cdot x & + & 1\cdot y & = & 6\\ 2\cdot x & + & 3\cdot y & = & 14 \end{array}}$$

The first is on the right; it consists of the two "free" terms (free of x's and y's!) on the right-hand side:

$$B=\begin{bmatrix}6\\14\end{bmatrix}.$$

Another one is less visible; it is made up of the two unknowns:

$$X=\begin{bmatrix}x\\y\end{bmatrix}.$$

Even though its dimension is also 2, it's not of the same nature as the others:

$$\begin{bmatrix}x\\y\end{bmatrix}\quad\text{is}\quad\begin{bmatrix}\#\text{ of pounds}\\\#\text{ of pounds}\end{bmatrix}.$$

It lives in a different $\mathbf{R}^2$.

Then, we have a function between these two spaces:

$$F:\mathbf{R}^2\to\mathbf{R}^2,\quad Y=F(X)\,.$$

Its formula can be written in terms of vectors:

$$F\left(\begin{bmatrix}x\\y\end{bmatrix}\right)=\begin{bmatrix}x+y\\2x+3y\end{bmatrix}.$$

Our problem becomes a problem of solving an equation for X:

$$F(X)=B\,.$$

Warning!

Since we aren't doing any geometry but we are doing vector algebra, the vector approach will be preferred throughout the chapter.

The algebraic operations needed to compute F are so simple that they will be easy to abbreviate.

Let's review the setup. The problem for dimension n has n ingredients:

	dim 1	dim 2
the unknown	x	$X=<x,y>$
multiplied by	3	?
is equal to	60	$B=<6,14>$

We have transitioned from numbers to vectors. But what is the operation that makes B from X? None of the familiar ones.

The four coefficients of x, y form a table:

$$F = \begin{bmatrix} 1 & 1 \\ 2 & 3 \end{bmatrix} .$$

It has two rows and two columns. In other words, this is a 2×2 *matrix*.

Both X and B are *column-vectors* in dimension 2, and matrix F turns X into B. This is very similar to multiplication of numbers; after all, they are vectors of dimension 1. Let's match the setups of the two problems:

$$\begin{aligned} \dim 1: &\quad m \cdot x = b \\ \dim 2: &\quad F \cdot X = B \end{aligned}$$

If we can just make sense of the new algebra!

Here $FX = B$ is a *matrix equation*, and it's supposed to capture the system of equations. Let's compare the original system of equations to $FX = B$:

$$\begin{array}{rrr} x & +y & = 6 \\ 2x & +3y & = 14 \end{array} \text{, rewritten as } \begin{bmatrix} 1 & 1 \\ 2 & 3 \end{bmatrix} \cdot \begin{bmatrix} x \\ y \end{bmatrix} = \begin{bmatrix} 6 \\ 14 \end{bmatrix} .$$

We can see these equations on the right-hand side as these two *dot products*. First:

$$1 \cdot x + 1 \cdot y = 6 \text{ , rewritten as } \begin{bmatrix} 1 & 1 \end{bmatrix} \cdot \begin{bmatrix} x \\ y \end{bmatrix} = 6 .$$

Second:

$$2x + 3y = 14 \text{ , rewritten as } \begin{bmatrix} 2 & 3 \end{bmatrix} \cdot \begin{bmatrix} x \\ y \end{bmatrix} = 14 .$$

This suggests what the meaning of FX should be. We "multiply" either row in A, as a vector, by the vector X – via the dot product:

Definition 5.1.2: product of matrix and vector

The *product* FX *of a* 2×2 *matrix* F *and a* 2*-vector* X,

$$F = \begin{bmatrix} a & b \\ c & d \end{bmatrix} , \; X = \begin{bmatrix} x \\ y \end{bmatrix} ,$$

is defined to be the following 2-vector:

$$\boxed{FX = \begin{bmatrix} a & b \\ c & d \end{bmatrix} \cdot \begin{bmatrix} x \\ y \end{bmatrix} = \begin{bmatrix} ax + by \\ cx + dy \end{bmatrix}}$$

We can still see these dot products in the result:

$$\begin{bmatrix} a & b \\ c & d \end{bmatrix} \cdot \begin{bmatrix} x \\ y \end{bmatrix} = \begin{bmatrix} ax + by \\ cx + dy \end{bmatrix} \quad \text{and} \quad \begin{bmatrix} a & b \\ c & d \end{bmatrix} \cdot \begin{bmatrix} x \\ y \end{bmatrix} = \begin{bmatrix} ax + by \\ cx + dy \end{bmatrix} .$$

Warning!

A matrix is nothing but an abbreviation of a transformation of the plane:

$$FX = F(X)\,.$$

However, not all transformations can be represented by matrices.

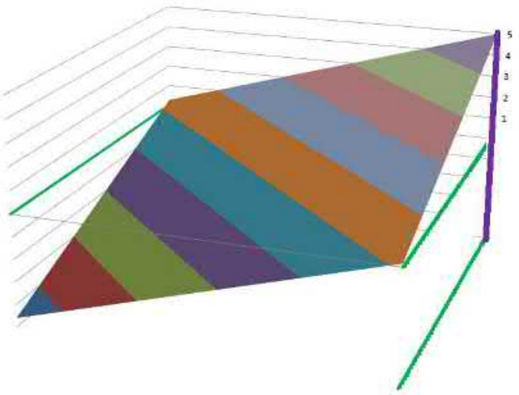
Example 5.1.3: 3 variables

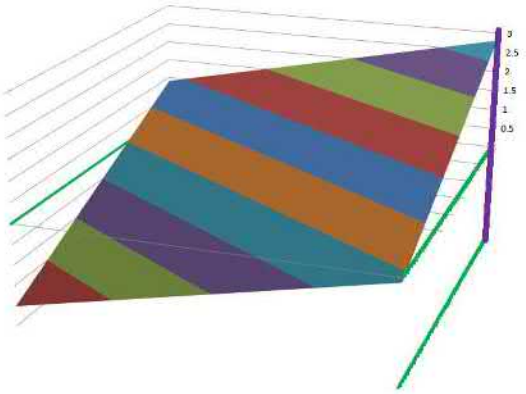
What if the blend is to contain another, third, type of coffee? Given three prices per pound, $2, 3, 5$, how much of each do you need to have 6 pounds of blend with a total price of 14?

Let x, y, and z be the weights of the three types of coffee, respectively. Then the total price of the blend is 14. Therefore, we have a system:

$$\begin{cases} x & + & y & + & z & = 6 \\ 2x & + & 3y & + & 5z & = 14 \end{cases}$$

Either of these equations represents a plane in $\mathbf{R}^3$. The solution set then comes from their intersection:





There are, of course, infinitely many solutions. An additional restriction in the form of another linear equation may reduce the number to one... or none. The variety of possible outcomes is, by far, higher than in the 2-dimensional case; they are not discussed in this chapter.

The vector algebra, however, is the same! The three weights can be written in a vector, $< 1, 1, 1 >$, and the first equation becomes the dot product:

$$< 1, 1, 1 > \cdot < x, y, z >= 6\,.$$

The three prices per pound can be written in a vector, $< 2, 3, 5 >$, and the first equation becomes the dot product:

$$< 2, 3, 5 > \cdot < x, y, z >= 14\,.$$

Finally, we have a *matrix equation*:

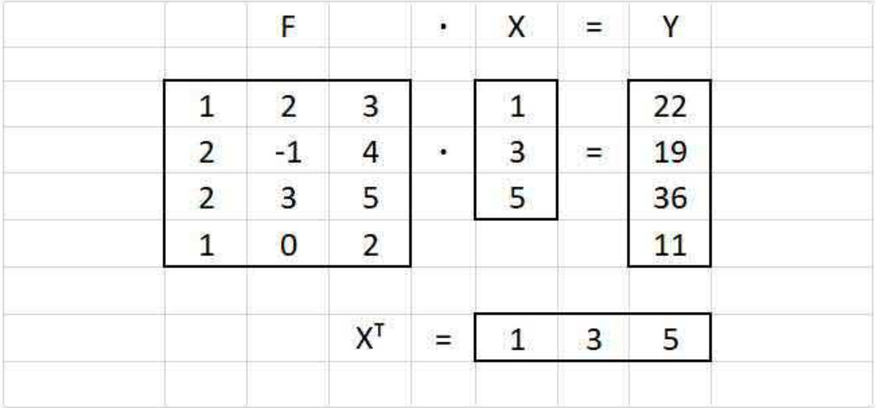
$$\begin{bmatrix} 1 & 1 & 1 \\ 2 & 3 & 5 \end{bmatrix} \cdot \begin{bmatrix} x \\ y \\ z \end{bmatrix} = \begin{bmatrix} 6 \\ 14 \end{bmatrix}\,.$$

Without harm, we can make the matrix square:

$$\begin{bmatrix} 1 & 1 & 1 \\ 2 & 3 & 5 \\ 9 & 0 & 0 \end{bmatrix} \cdot \begin{bmatrix} x \\ y \\ z \end{bmatrix} = \begin{bmatrix} 6 \\ 14 \\ 0 \end{bmatrix}\,.$$

Example 5.1.4: spreadsheet

One can utilize a spreadsheet and other software to this multiplication for matrices of any dimensions. In order to make this work, the vector X has to be "transposed" (bottom):



		F		·	X	=	Y	
	1	2	3		1		22	
	2	-1	4	·	3	=	19	
	2	3	5		5		36	
	1	0	2				11	
			X^T	=	1	3	5	

This is the code for the transpose of X:

```
=TRANSPOSE(R[-5]C:R[-3]C)
```

This is the code for Y:

```
=SUMPRODUCT(RC2:RC4,R8C[-2]:R8C)
```

The whole system can be written in the form of exactly the same matrix equation:

$$FX = B\,.$$

The multiplication is executed in the same way too:

$$\begin{bmatrix} \mathrm{r} & \mathrm{r} & \mathrm{o} & \mathrm{o} & \mathrm{w} & \mathrm{w} \end{bmatrix} \cdot \begin{bmatrix} \mathrm{c} \\ \mathrm{o} \\ \mathrm{l} \\ \mathrm{u} \\ \mathrm{m} \\ \mathrm{n} \end{bmatrix} = rc + ro + ol + ou + wm + wn$$

Generally, we face a system with:

1. the number of variables m, and
2. the number of equations n.

We will have an $n \times m$ matrix:

$$\begin{array}{c|ccccc|} & 1 & 2 & 3 & \dots & m \\ 1 & 2 & 0 & 3 & \dots & 2 \\ 2 & 0 & 6 & 2 & \dots & 0 \\ \vdots & \vdots & \vdots & \vdots & \dots & \vdots \\ n & 3 & 1 & 0 & \dots & 12 \end{array}$$

Here, the number at the ij-position is the coefficient of the jth variable in the ith equation.

Recall how an index points at a location within a sequence. Similarly, we use *double* index to point at the correct location within a table.

Definition 5.1.5: entries of matrix

The *ij-entry in an $n\times m$ matrix A* is the number at the ith row and jth column, denoted by

$$\boxed{A_{ij}}$$

for each $i=1,2,...,m$ and each $j=1,2,...,n$.

For example, we have for the above matrix:

$$\begin{array}{c|ccccc|}
i\backslash j & 1 & 2 & 3 & ... & n \\
1 & A_{1,1}=2 & A_{1,2}=0 & A_{1,3}=3 & ... & A_{1,n}=2 \\
2 & A_{2,1}=0 & A_{2,2}=6 & A_{2,3}=2 & ... & A_{2,n}=0 \\
\vdots & \vdots & \vdots & \vdots & ... & \vdots \\
m & A_{m,1}=3 & A_{m,2}=1 & A_{m,3}=0 & ... & A_{m,n}=12
\end{array}$$

Warning!

What is the difference between tables of numbers and matrices? The algebraic operations discussed here.

5.2. *Transformations of the plane*

Transformations of the plane are made up of two real-valued functions of two variables.

Definition 5.2.1: transformation of the plane

A *transformation of the plane* is a function

$$F:\mathbf{R}^2\to\mathbf{R}^2\,,$$

given by any pair of functions f,g of two variables:

$$F(x,y)=(u,v)=\Big(f(x,y),\ g(x,y)\Big)\,.$$

When appropriate, we can also look at the inputs and outputs as *vectors*:

$$<x,y>=\begin{bmatrix}x\\y\end{bmatrix}\mapsto<u,v>=\begin{bmatrix}u\\v\end{bmatrix}\,,$$

instead of points.

Let's review a few examples of such transformations.

Example 5.2.2: two functions of two variables

We need two real-valued functions of two variables. Consider $u = f(x,y) = 2x - 3y$:

$$\begin{aligned} f : \mathbf{R}^2 &\to \mathbf{R}, && \text{meaning} \\ f : (x,y) &\to u && = 2x - 3y \end{aligned}$$

Consider also $v = g(x,y) = x + 5y$:

$$\begin{aligned} g : \mathbf{R}^2 &\to \mathbf{R}, && \text{meaning} \\ g : (x,y) &\to v && = x + 5y \end{aligned}$$

Let's build a new function from these two. We take the input to be the same – a point in the plane – and we *combine* the two outputs into a single point (u,v) – in another plane. Then what we have is a single function:

$$\begin{aligned} F : \mathbf{R}^2 &\to \mathbf{R}^2, && \text{meaning} \\ F : (x,y) &\to (u,v) && = (2x - 3y,\ x + 5y) \end{aligned}$$

In short, this is the formula for this function:

$$F(x,y) = (2x - 3y,\ x + 5y)\,.$$

In terms of vectors:

$$F : \begin{bmatrix} x \\ y \end{bmatrix} \mapsto \begin{bmatrix} 2x - 3y \\ x + 5y \end{bmatrix}$$

The coefficients of the matrix of F are read from that representation:

$$F = \begin{bmatrix} 2 & -3 \\ 1 & 5 \end{bmatrix}$$

What this function does to the plane remains to be determined.

Example 5.2.3: vertical shift

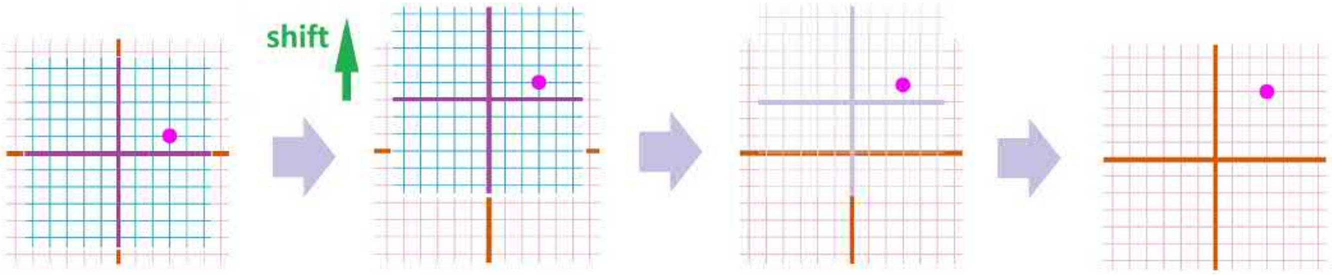
The function defined by

$$F(x,y) = (x, y + 3)$$

is a *vertical shift*:

$$(x,y) \xrightarrow{\text{up } k} (x, y + k).$$

We visualize these transformations by drawing something on the original plane (the domain) and then see what that looks like in the new plane (the co-domain):



Predictably, the formula:

$$F(x,y) = (x + a, y + b) = (x,y) + < a, b > ,$$

gives the *shift by vector* $< a, b >$.

Exercise 5.2.4

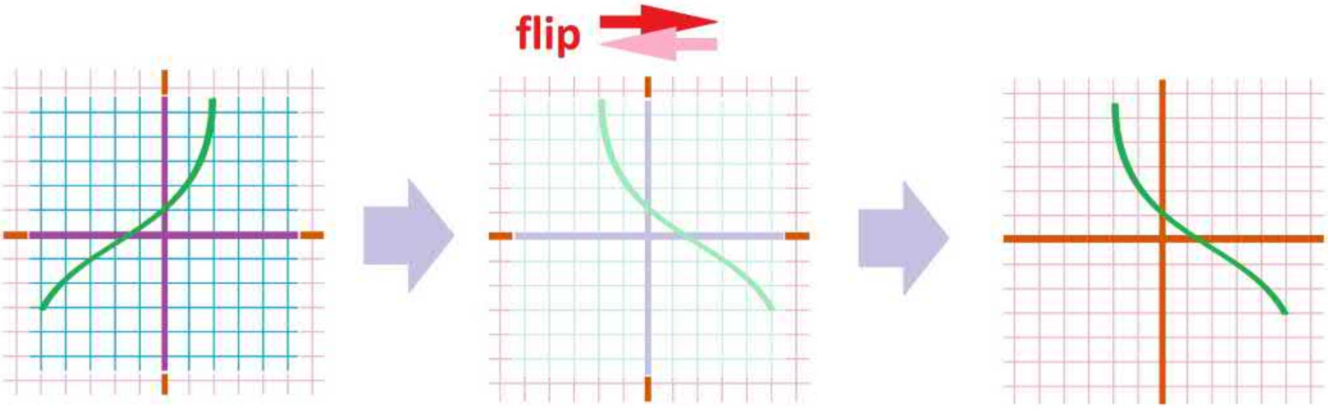
Explain why there is no matrix.

Example 5.2.5: horizontal and vertical flip

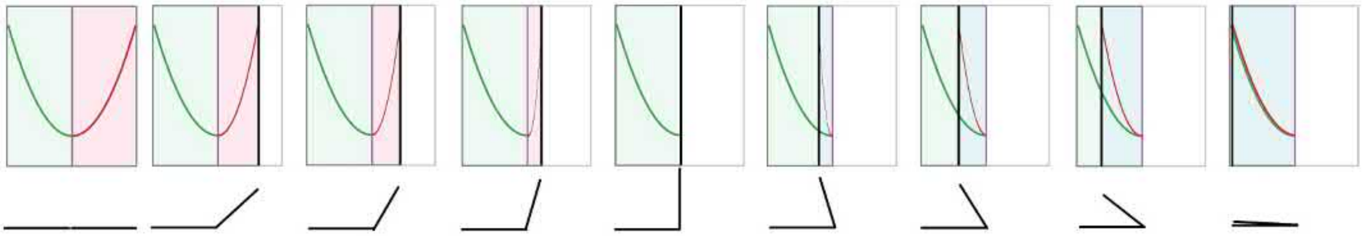
Now the *horizontal flip*. We lift, then flip the sheet of paper with xy-plane on it, and finally place it on top of another such sheet so that the y-axes align. If the function is given by

$$F(x,y)=(-x,y)\,,$$

then we have the following:



Below we illustrate the fact that the parabola's left branch is a mirror image of its right branch:



We can also represent this transformation via vectors:

$$F:\begin{bmatrix}x\\y\end{bmatrix}\mapsto\begin{bmatrix}-x\\y\end{bmatrix}=\begin{bmatrix}(-1)x+0y\\0x+1y\end{bmatrix}$$

Then, we have its matrix:

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

Indeed,

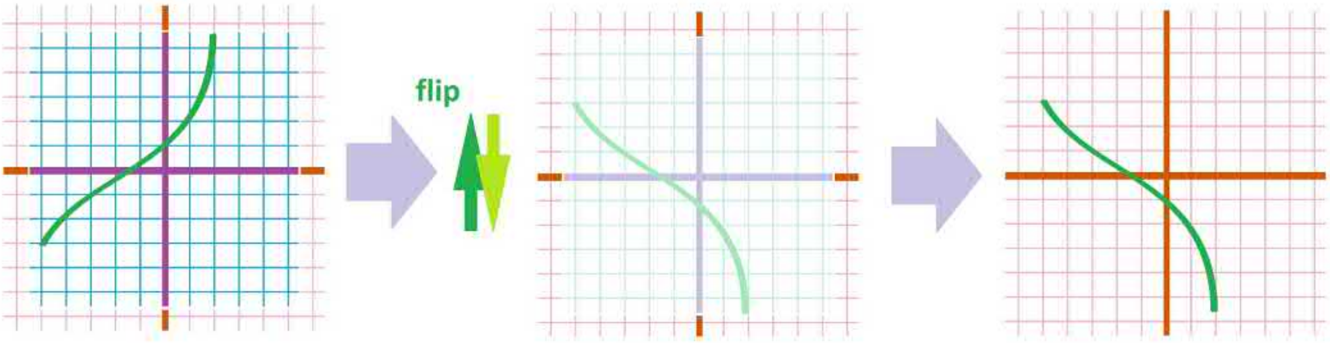
$$F\left(\begin{bmatrix}x\\y\end{bmatrix}\right)=F\begin{bmatrix}x\\y\end{bmatrix}=\begin{bmatrix}-1&0\\0&1\end{bmatrix}\begin{bmatrix}x\\y\end{bmatrix}=\begin{bmatrix}-x\\y\end{bmatrix}.$$

Next, consider *vertical flip*. We lift, then flip the sheet of paper with xy-plane on it, and finally place it on top of another such sheet so that the x-axes align. If

$$G(x,y)=(x,-y)\,,$$

then we have:

$$(x,y)\xrightarrow{\text{vertical flip}}(x,-y).$$



We can also represent this transformation via a matrix:

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

Indeed,

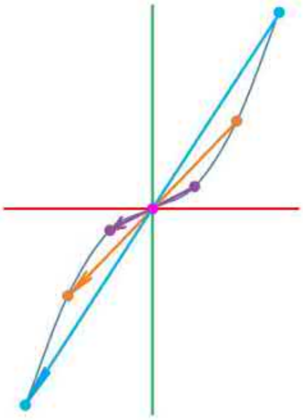
$$G\left(\begin{bmatrix} x \\ y \end{bmatrix}\right) = \begin{bmatrix} -1 & 0 \\ 0 & 1 \end{bmatrix} \begin{bmatrix} x \\ y \end{bmatrix} = \begin{bmatrix} x \\ -y \end{bmatrix} .$$

Example 5.2.6: central symmetry

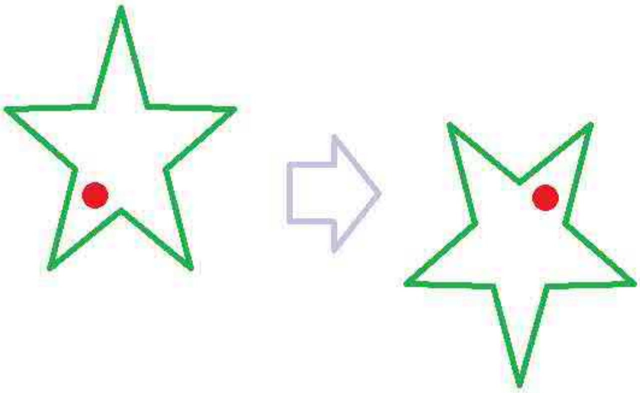
How about the *flip about the origin*? This is the formula,

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

of what is also known as the central symmetry:



This is what the transformation does to a star:



We can also represent this transformation via vectors:

$$F : \begin{bmatrix} x \\ y \end{bmatrix} \mapsto \begin{bmatrix} -x \\ y \end{bmatrix} = \begin{bmatrix} (-1)x + 0y \\ 0x + (-1)y \end{bmatrix}$$

Then, we have its matrix:

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

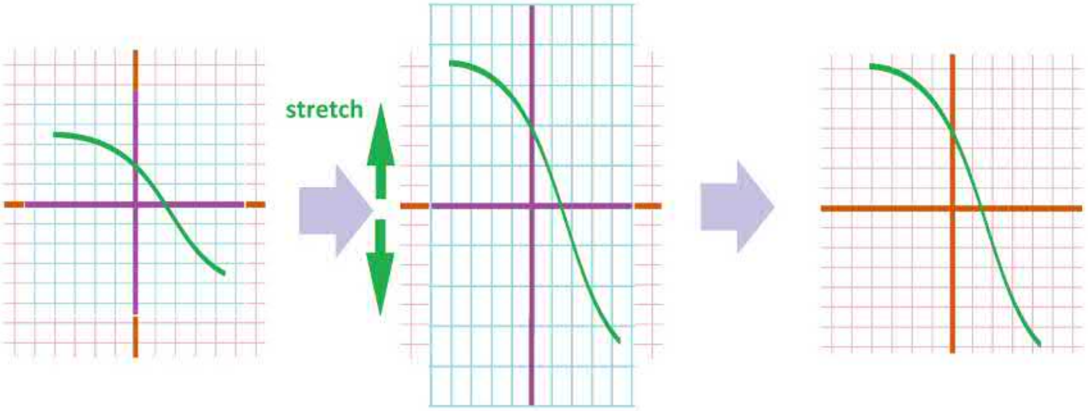
Indeed,

$$F\left(\begin{bmatrix} x \\ y \end{bmatrix}\right) = F\begin{bmatrix} x \\ y \end{bmatrix} = \begin{bmatrix} -1 & 0 \\ 0 & -1 \end{bmatrix}\begin{bmatrix} x \\ y \end{bmatrix} = \begin{bmatrix} -x \\ -y \end{bmatrix} .$$

Example 5.2.7: horizontal and vertical stretch

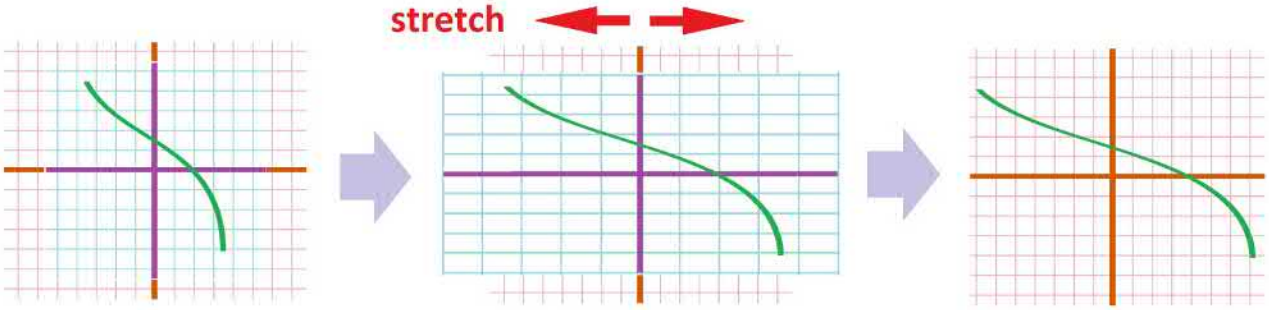
Now the *horizontal stretch.* We grab a rubber sheet by the top and the bottom and pull them apart in such a way that the y-axis doesn't move. Here,

$$F(x,y) = (kx,y) .$$



Similarly, the *horizontal stretch* is given by:

$$G(x,y) = (x,ky) .$$



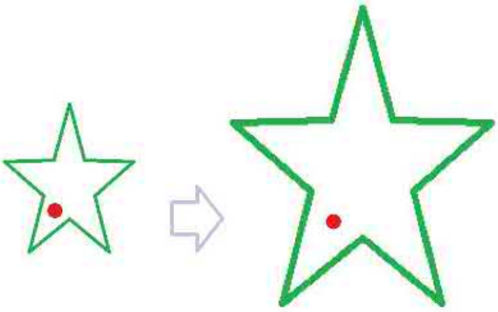
We can also represent these transformations via matrices:

$$F = \begin{bmatrix} k & 0 \\ 0 & 1 \end{bmatrix} \quad \text{and} \quad G = \begin{bmatrix} 1 & 0 \\ 0 & k \end{bmatrix} .$$

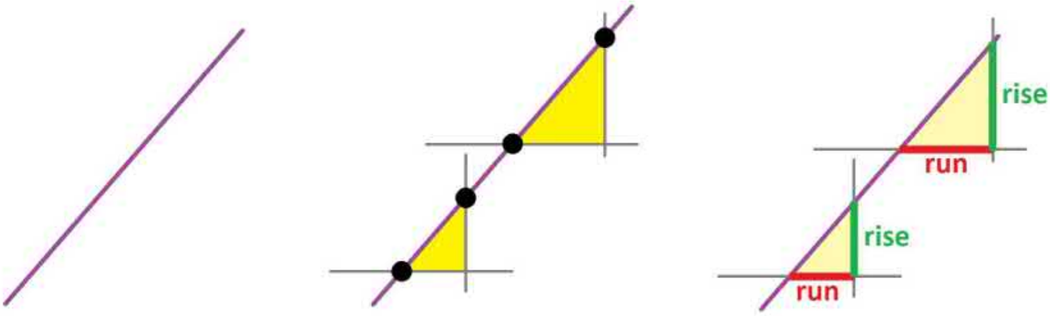
Example 5.2.8: re-scaling

How about the *uniform stretch* (same in all directions)? This is its formula:

$$F(x,y) = (kx,ky) .$$



The result is *re-scaling*. The reason comes from what we know from geometry: Similar triangles have the same angles. Here is an illustration why:



We can also represent this transformations via a matrix:

$$F = \begin{bmatrix} k & 0 \\ 0 & k \end{bmatrix} .$$

Exercise 5.2.9

Find the matrix for a disproportional stretch.

Just as before, we put all newly introduced functions in broad categories.

Some of these categories – such as monotonicity – have become irrelevant.

Others – such as symmetry – have become by far more complex.

Two that will be pursued are one-to-one and onto.

Recall:

- We call a function *one-to-one* if there is no more that one input for each output.
- We call a function *onto* if there is at least one input for each output.

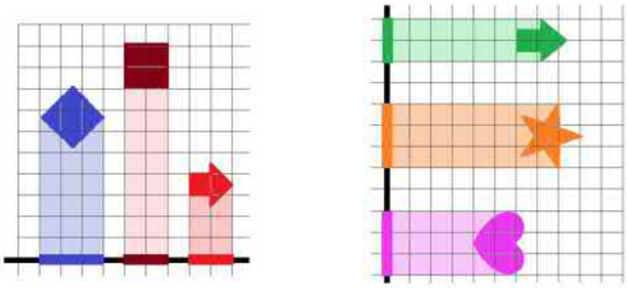
The functions above are all both one-to-one and onto.

Exercise 5.2.10

Prove the last statement.

Example 5.2.11: projections

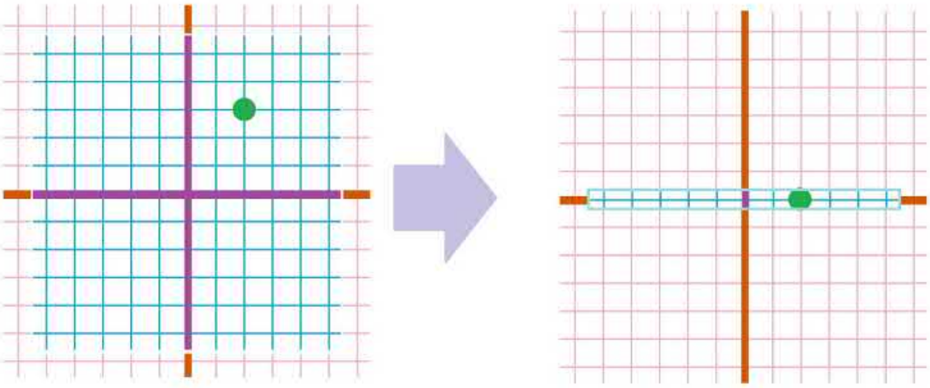
The functions that are not one-to-one or onto are the *projections*. There are at least two types:



This is the vertical one:

$$F(x,y)=(x,0)\,.$$

It is the projection on the x-axis. It's as if the sheet of the xy-plane is rolled into a thin scroll:



We can also represent this transformations via a matrix:

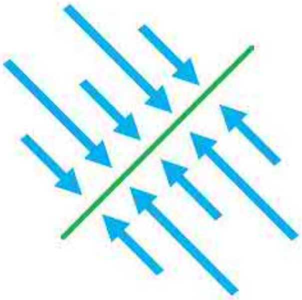
$$F=\begin{bmatrix}1&0\\0&0\end{bmatrix}.$$

Exercise 5.2.12

Find the formula and the matrix for the projection on the y-axis.

Exercise 5.2.13

Find the formula and the matrix for the projection on the line $y=x$.



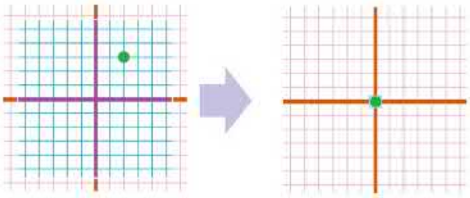
Example 5.2.14: collapse

Finally, we have the *collapse*:

It is a constant function:

$$F(x,y)=(x_0,y_0)\,.$$

It's as if the sheet is crushed into a tiny ball:



There is a matrix representation when this point is the origin:

$$F=\left[\begin{array}{cc}0&0\\0&0\end{array}\right].$$

Exercise 5.2.15

Show that there is no matrix unless $x_0=0,\ y_0=0$.

The meaning of each number in the matrix depends on its location:

$$\begin{array}{c|cc} & x & y\\ \hline x & a & b\\ y & c & d\end{array}$$

This is a special case that we have learned about so far:

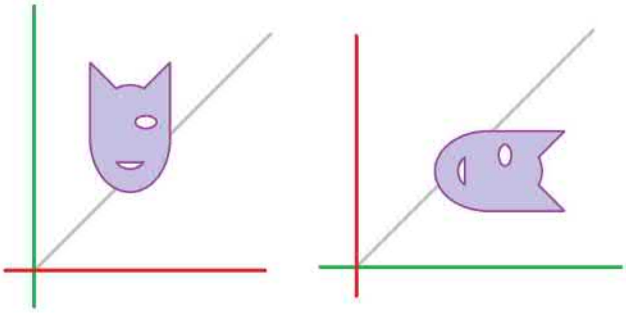
Matrix Deconstruction

$$\begin{array}{rcccl}\text{stretched or flipped } x & \to & a, & 0 & \leftarrow \text{ there is no interaction between } x,y\\ \text{there is no interaction between } x,y & \to & 0, & d & \leftarrow \text{ stretched or flipped } y\end{array}$$

There are many more transformations, however, that aren't covered so far because they cannot be represented in terms of the vertical and horizontal transformations.

Example 5.2.16: flip about diagonal

A *flip about the line* $x=y$ that appeared in the context of finding the graph of the inverse function:

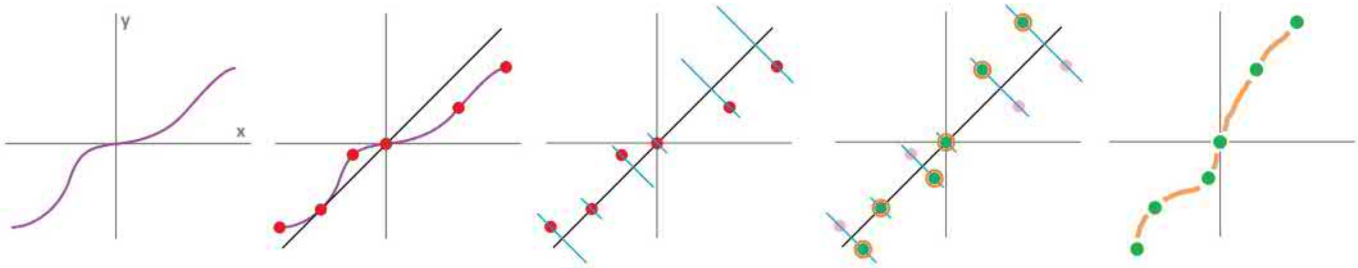


As we acquire the inverse by interchanging x and y, we have the same here:

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

The matrix is:

$$F=\left[\begin{array}{cc}0&1\\1&0\end{array}\right].$$



We see here some new features:

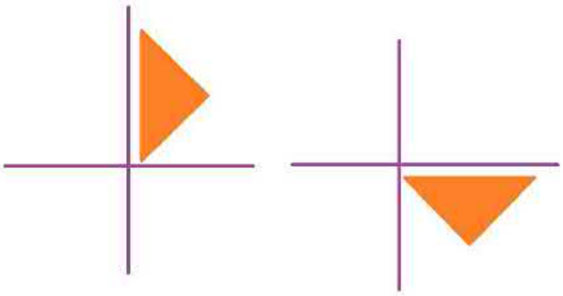
Matrix Deconstruction

x doesn't depends on x $\to$ 0, 1 $\leftarrow$ x depends on y

y depends on x $\to$ 1, 0 $\leftarrow$ y doesn't depends on y

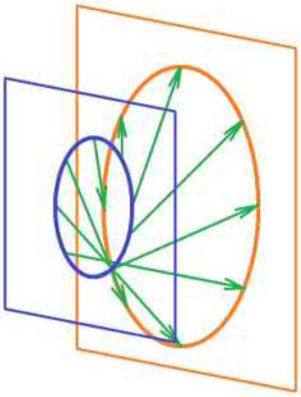
Exercise 5.2.17

Find the matrix for this rotation:



Example 5.2.18: compositions

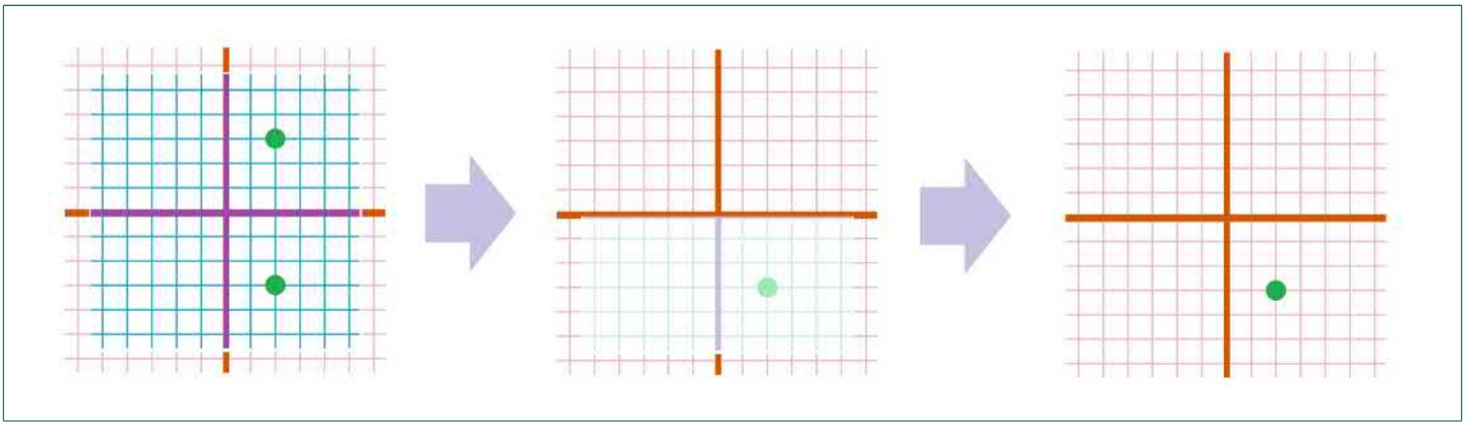
More complex transformations, however, will require further study. Below, we visualize the 90-degree rotation with a stretch with a factor of 2:



Example 5.2.19: folding

Among others, we may consider *folding* the plane:

$$F(x,y) = (|x|, y)\,.$$



Exercise 5.2.20

Confirm that this function cannot be represented by a matrix.

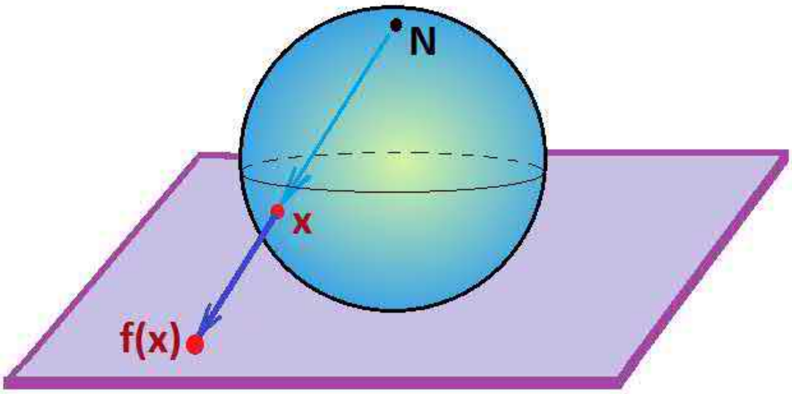
Of course, any Euclidean space $\mathbf{R}^n$ can be – in a similar manner – rotated (around various axes), stretched (in various directions), projected (onto various lines or planes), or collapsed.

For example, this is how the projection on the xy-plane

$$F(x,y,z)=(x,y,0)$$

works:

This is how maps are made (stereographic projection):



5.3. *Linear operators*

We consider transformations of the plane as before but we think of the points on the plane as the ends of 2-*vectors*. It makes no difference when they are expressed in terms of their components:

$$<x,y> \mapsto <u,v>$$

Then, what happens to the functions? It used to be the case that the coordinates of the output depend on the coordinates of the input:

$$F(x,y)=(u,v)=2x-3y,\ x+5y)\,.$$

Now our function shows how the components of the output depend on the components of the input:

$$F\big(< x, y > \big) =< u, v >=< 2x - 3y,\ x + 5y > \ .$$

Here $< x, y >$ is the input and $< u, v >$ is the output:

$$\begin{array}{ccccc} \text{input} & & \text{function} & & \text{output} \\ < x, y > & \to & \boxed{\quad F \quad} & \to & < u, v > \end{array}$$

Both are vectors; it's a *vector function*:

$$F : \mathbf{R}^2 \to \mathbf{R}^2 \,.$$

Warning!

We could interpret F as a vector field, but we won't.

Example 5.3.1: re-scaling

Some of the benefits of this point of view are immediate. For example, even if we didn't know that the transformation given by

$$F(x, y) = (2x, 2y)$$

is a uniform stretch, we can discover that fact with our knowledge of vector algebra. We write this function in terms of vectors and discover scalar multiplication:

$$F(< x, y >) =< 2x, 2y >= 2 < x, y > \ .$$

In fact, this idea will work in any dimension. This is how you stretch the space by a factor of 2, using the component-free approach:

$$F : \mathbf{R}^n \to \mathbf{R}^n \ \text{ given by } \ F(X) = 2X \,.$$

The approach gives us a better representation when the functions that make up the transformation happen to be *linear*. Matrices reply on vector algebra:

$$F(X) = \big(f(x, y), g(x, y)\big) = \begin{bmatrix} a & b \\ c & d \end{bmatrix} \cdot \begin{bmatrix} x \\ y \end{bmatrix} = F \cdot X \,.$$

Exercise 5.3.2

Does it work when the function isn't linear? Try $F(x, y) = \left(e^x, \dfrac{1}{y}\right)$.

Conversely, if we do have a matrix, we can always understand it as a function, as follows:

$$\begin{bmatrix} a & b \\ c & d \end{bmatrix} \begin{bmatrix} x \\ y \end{bmatrix} = \begin{bmatrix} ax + by \\ cx + dy \end{bmatrix} =< ax + by,\ cx + dy > ,$$

for some a, b, c, d fixed. So, matrix F contains all the information about function F. One can think of F (a table) as an abbreviation of FX (a formula).

Warning!

We will continue to use the same letter for the function and the matrix.

Clearly, a function given by a matrix is a *special* one. What is so special about it?

Now, the domain $\mathbf{R}^2$ of this function is a *vector space*, and so is its codomain. How does such a function interact with the algebra of these two spaces? What happens to the *vector operations* under F?

Suppose we have addition and scalar multiplication carried out in the domain space of F. Once F has transformed the plane, what do these operations look like now?

Example 5.3.3: dimension 1

The multiplication by 2, $f(x) = 2x$ "preserves addition":

$$f(x + x') = 2(x + x') = 2x + 2x' = f(x) + f(x')\,.$$

After all, this is just a stretch by a factor of 2.

The computation is just an abbreviation of the following diagram:

$$\begin{tikzcd} x, x' \arrow[r, "+"] \arrow[d, "\cdot 2"] & x + x' \arrow[d, "\cdot 2"] \\ 2x, 2x' \arrow[r, "+"] & 2x + 2x' = 2(x + x') \end{tikzcd}$$

In the diagram, we start with a pair of numbers at the top left and then we proceed in two ways:

- Right: Add them. Then down: Apply the function to the result.
- Down: Apply the function to them. Then right: Add the results.

A shift by 1, $f(x) = x + 1$, doesn't preserve addition:

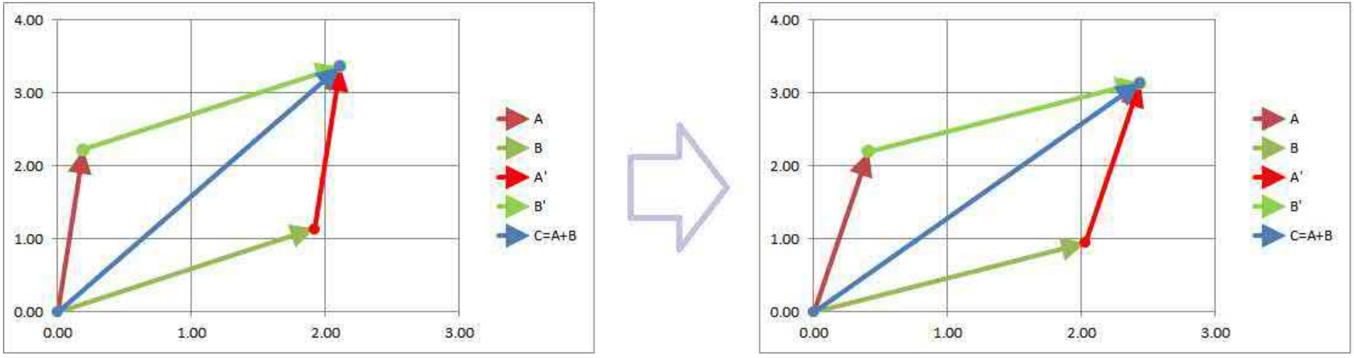
$$f(x + x') = (x + x') + 1 = x + x' + 1 \neq x + x' + 2 = (x + 1) + (x' + 1) = f(x) + f(x')\,.$$

Exercise 5.3.4

What effect does the function $f : \mathbf{R} \to \mathbf{R}$ given by $f(x) = 2x$ have on multiplication in its domain?

Example 5.3.5: addition under rotation

What happens to an addition diagram (the parallelogram construction) when the *plane* is transformed? If such a diagram is *rotated*, as if on a piece of paper, it will remain an addition diagram:



We see the parallelogram rule of addition on the left and on the right. This is the algebra:

$$\begin{tikzcd} A, B \arrow[r, "+"] \arrow[d, "\text{rotated}"] & A+B \arrow[d, "\text{rotated}"] \\ F(A), F(B) \arrow[r, "+"] & F(A)+F(B) = F(A+B) \end{tikzcd}$$

What happens to an addition diagram when the vector space is transformed? When it is still an addition diagram, this is the language we will use:

Definition 5.3.6: addition is preserved

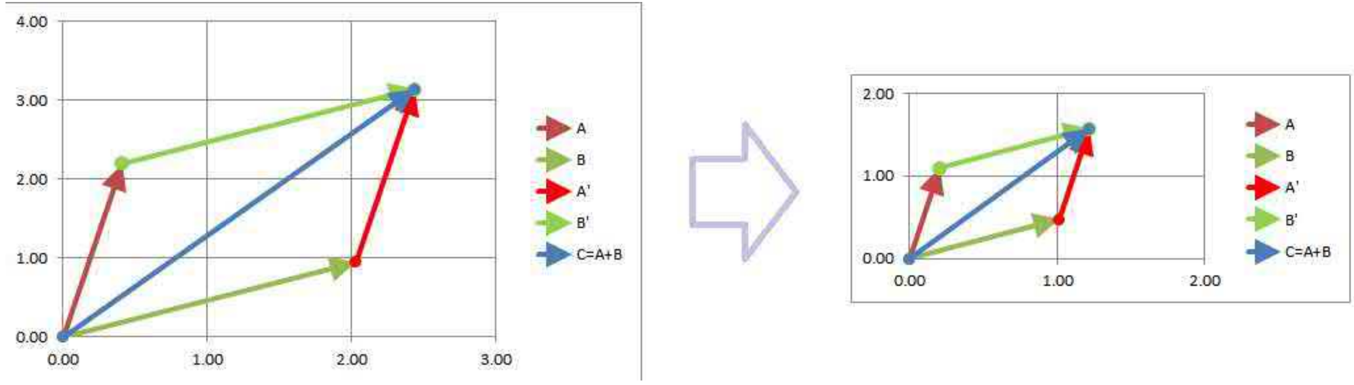
We say that *addition is preserved* under a function $F : \mathbf{R}^n \to \mathbf{R}^m$ if

$$\boxed{F(A+B) = F(A) + F(B)}$$

for any vectors A and B.

Example 5.3.7: stretch dimension 2

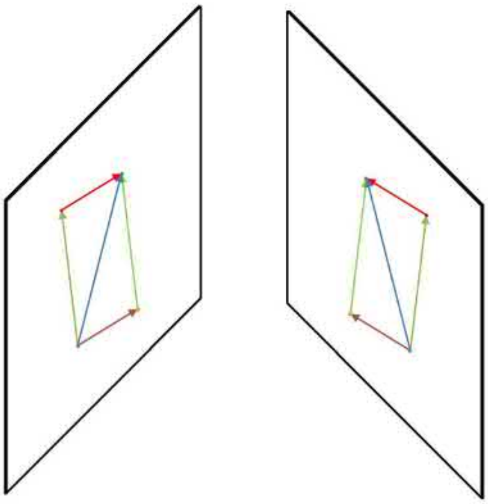
Furthermore, the example of a stretch below shows that the triangles of the diagram, if not identical, are *similar* to the original:



It's as if we just stepped away from the piece of paper that has the addition diagram on it or put the diagram under a magnifying glass.

Example 5.3.8: reflection dimension 2

We can also see the addition diagram in the mirror, and it's still an addition diagram:



Exercise 5.3.9

Show that a fold doesn't preserve vector addition. Suggest other examples.

What about the general case?

Theorem 5.3.10: Preserving Addition

If a function $F : \mathbf{R}^n \to \mathbf{R}^m$ *is given by a matrix,* $F(X) = FX$*, it preserves addition.*

Proof.

Consider $F : \mathbf{R}^2 \to \mathbf{R}^2$ and two input vectors:

$$A = \begin{bmatrix} x \\ y \end{bmatrix} \quad \text{and} \quad B = \begin{bmatrix} x' \\ y' \end{bmatrix}.$$

Let's confirm the formula:

$$F(A+B) = F(A) + F(B)\,.$$

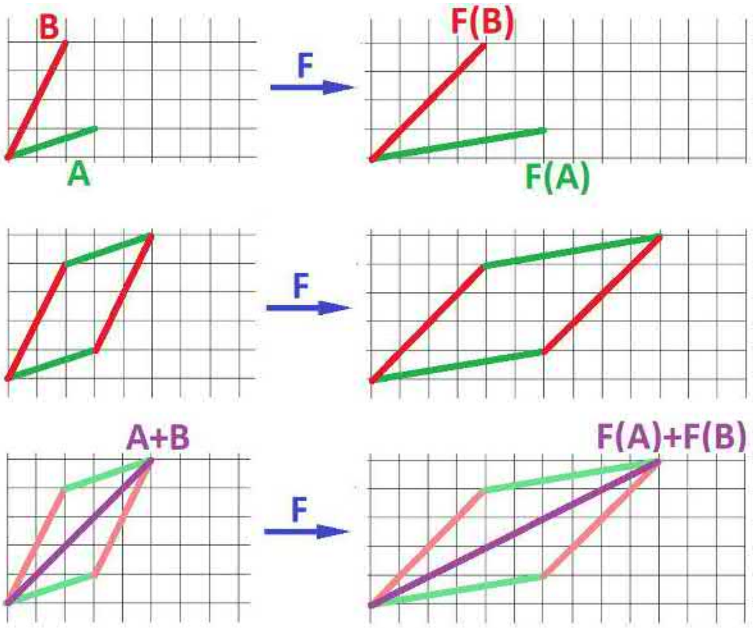
Let's compare:

Left-hand side:	Right-hand side:
$F\left(\begin{bmatrix} x \\ y \end{bmatrix} + \begin{bmatrix} x' \\ y' \end{bmatrix}\right)$	$F\begin{bmatrix} x \\ y \end{bmatrix} + F\begin{bmatrix} x' \\ y' \end{bmatrix}$
$= \begin{bmatrix} a & b \\ c & d \end{bmatrix}\left(\begin{bmatrix} x \\ y \end{bmatrix} + \begin{bmatrix} x' \\ y' \end{bmatrix}\right)$	$= \begin{bmatrix} a & b \\ c & d \end{bmatrix}\begin{bmatrix} x \\ y \end{bmatrix} + \begin{bmatrix} a & b \\ c & d \end{bmatrix}\begin{bmatrix} x' \\ y' \end{bmatrix}$
$= \begin{bmatrix} a & b \\ c & d \end{bmatrix}\begin{bmatrix} x+x' \\ y+y' \end{bmatrix}$	$= \begin{bmatrix} ax+by \\ cx+dy \end{bmatrix} + \begin{bmatrix} ax'+by' \\ cx'+dy' \end{bmatrix}$
$= \begin{bmatrix} a(x+x')+b(y+y') \\ c(x+x')+d(y+y') \end{bmatrix}$	$= \begin{bmatrix} ax+by+ax'+by' \\ cx+dy+cx'+dy' \end{bmatrix}$

These are the same, after factoring.

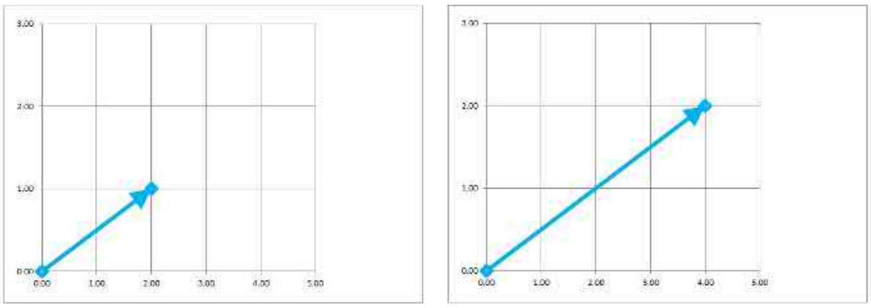
Example 5.3.11: stretch dimension 2

Here is a confirmation of the result for a horizontal stretch:



We simply compare what happens in the domain with its “reflection” in the codomain.

The diagram of scalar multiplication is much simpler:



It is a stretch of the vector; rotated or stretched, it remains a stretch. This is the language we will use:

Definition 5.3.12: scalar multiplication is preserved

We say that *scalar multiplication is preserved* under a function $F : \mathbf{R}^n \to \mathbf{R}^m$ if

$$\boxed{F(kA) = kF(A)}$$

for any vector X and real k.

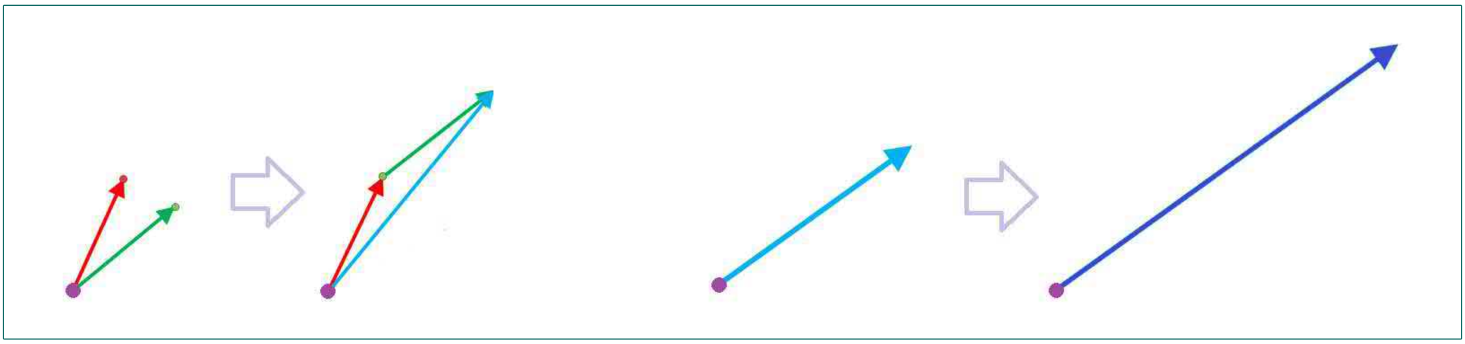
The formulas in the two definitions are just abbreviations of these diagrams:

$$\begin{tikzcd}
A, B \arrow[r, "+"] \arrow[d, "F"] & A+B \arrow[d, "F"] & A \arrow[r, "\cdot k"] \arrow[d, "F"] & kA \arrow[d, "F"] \\
F(A), F(B) \arrow[r, "+"] & F(A)+F(B) = F(A+B) & F(A) \arrow[r, "\cdot k"] & kF(A) = F(kA)
\end{tikzcd}$$

In other words, the order of these operations makes no difference.

Example 5.3.13: motions

The same conclusion is quickly reached for the flip and other motions: The triangles of the new diagram are identical to the original. We can just imagine that the addition diagram is drawn on a piece of paper with no grid, which then has been rotated:



Theorem 5.3.14: Preserving Scalar Multiplication

If a function $F : \mathbf{R}^n \to \mathbf{R}^n$ is given by a matrix, $F(A) = FA$, it preserves scalar multiplication.

Proof.

Consider an input vector $X =< x, y >$ and a scalar k. Then,

$$\begin{bmatrix} a & b \\ c & d \end{bmatrix} \left(k \begin{bmatrix} x \\ y \end{bmatrix} \right) = \begin{bmatrix} a & b \\ c & d \end{bmatrix} \left(\begin{bmatrix} kx \\ ky \end{bmatrix} \right) = \begin{bmatrix} akx + bky \\ ckx + dky \end{bmatrix} = k \begin{bmatrix} ax + by \\ cx + dy \end{bmatrix} = k \begin{bmatrix} a & b \\ c & d \end{bmatrix} \begin{bmatrix} x \\ y \end{bmatrix} .$$

Now, this is the general case:

Definition 5.3.15: linear operator

A function $F : \mathbf{R}^n \to \mathbf{R}^m$ that preserves both addition and scalar multiplication is called a *linear operator* (or a linear map); i.e.,:

$$\boxed{\begin{aligned} F(U+V) &= F(U) + F(V) \\ F(kV) &= kF(V) \end{aligned}}$$

Warning!

Previously, $y = ax + b$ has been called a “linear function”. Now, $y = ax$ is called a “linear operator”.

We combine the two operations together:

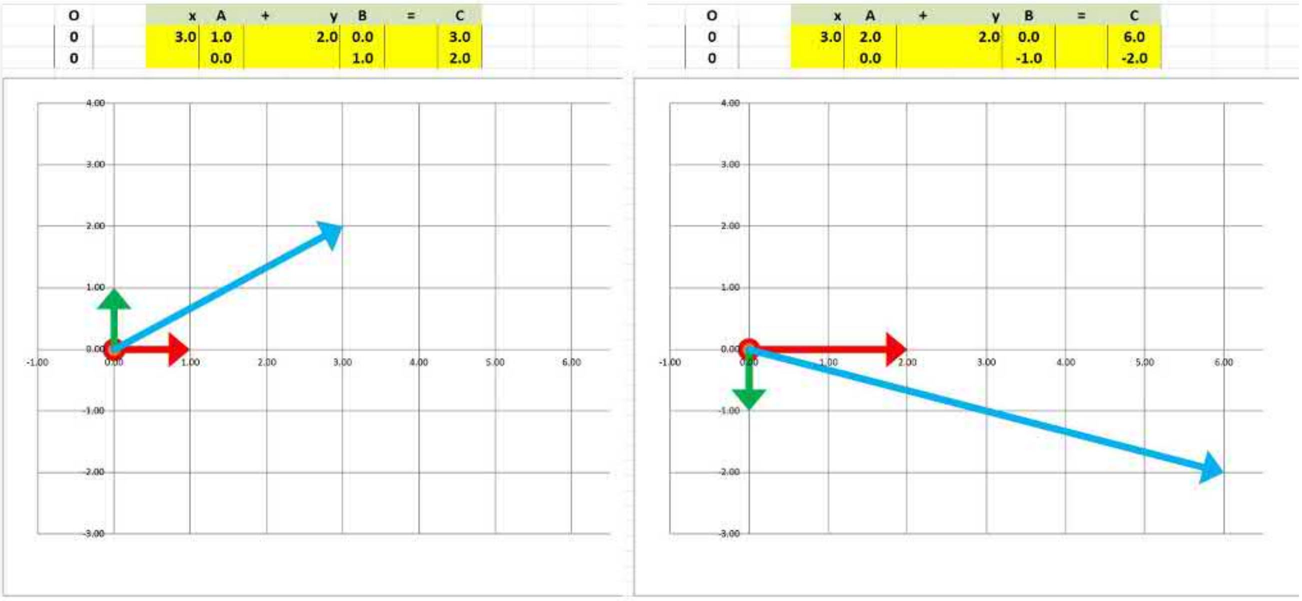
Theorem 5.3.16: Linear Operators and Linear Combinations

A linear operator $F : \mathbf{R}^n \to \mathbf{R}^m$ preserves linear combinations; i.e.,

$$X = xA + yB \implies F(X) = xF(A) + yF(B) ,$$

for any vectors A and B and any real coefficients x and y.

In other words, the diagram of a linear combination will remain such under a linear operator:



o	x	A	+	y	B	=	C
0	3.0	1.0		2.0	0.0		3.0
0		0.0			1.0		2.0

o	x	A	+	y	B	=	C
0	3.0	2.0		2.0	0.0		6.0
0		0.0			-1.0		-2.0

Exercise 5.3.17

Describe what this linear operator does and find its matrix.

This is the summary of our analysis.

Theorem 5.3.18: Linear Operators vs. Matrices

- *The function* $F : \mathbf{R}^2 \to \mathbf{R}^2$ *defined via multiplication by a* 2×2 *matrix* F,
$$F(X) = FX\,,$$
is a linear operator.
- *Conversely, every linear operator* $F : \mathbf{R}^2 \to \mathbf{R}^2$ *is defined via multiplication by some* 2×2 *matrix.*

Proof.

The first part of the theorem follows from the two theorems above. The converse is proven in the next section.

Warning!

Linear operators and matrices aren't interchangeable, because matrices emerge only when a Cartesian system has been specified.

Corollary 5.3.19: Linear Operator at 0

A linear operator takes the zero vector to the zero vector:

$$F(0) = 0\,.$$

Exercise 5.3.20

Prove the corollary (a) from the definition of a linear operator, (b) by examining the matrix multiplication.

The conclusion will be visible in the examples in the next section.

The latter part of the theorem is materialized when a matrix is found for a linear operator described by what it does. We start with a couple of simple examples.

Let's not forget:

▶ Linear operators are functions.

The two simplest functions – no matter what the domain and codomain are – are the constant function and the identity function.

However, according to the last corollary, there can be only one constant, 0, and, therefore, only one constant linear operator. This is the simplest linear operator:

Definition 5.3.21: zero operator

The function $F : \mathbf{R}^n \to \mathbf{R}^m$ that takes every vector to the zero vector,

$$F(0) = 0\,,$$

is called the *zero operator*. The notation is as follows:

$$0 : \mathbf{R}^n \to \mathbf{R}^m, \quad 0(X) = 0\,.$$

The matrix of the zero operator consists, of course, of all zeros:

$$0 = \begin{bmatrix} 0 & 0 & 0 & ... & 0 \\ 0 & 0 & 0 & ... & 0 \\ ... & ... & ... & ... & ... \\ 0 & 0 & 0 & ... & 0 \end{bmatrix}$$

We can write:

$$0_{ij} = 0\,.$$

What about the identity function? The dimensions of the domain and the codomain must be the same:

Definition 5.3.22: identity operator

The function $F : \mathbf{R}^n \to \mathbf{R}^n$ that takes every vector to itself,

$$F(X) = X\,,$$

is called the *identity operator*. The notation is as follows:

$$I : \mathbf{R}^n \to \mathbf{R}^n, \quad I(X) = X\,.$$

The matrix of the identity operator is the following:

$$I = \begin{bmatrix} 1 & 0 & 0 & ... & 0 \\ 0 & 1 & 0 & ... & 0 \\ ... & ... & ... & ... & ... \\ 0 & 0 & 0 & ... & 1 \end{bmatrix}$$

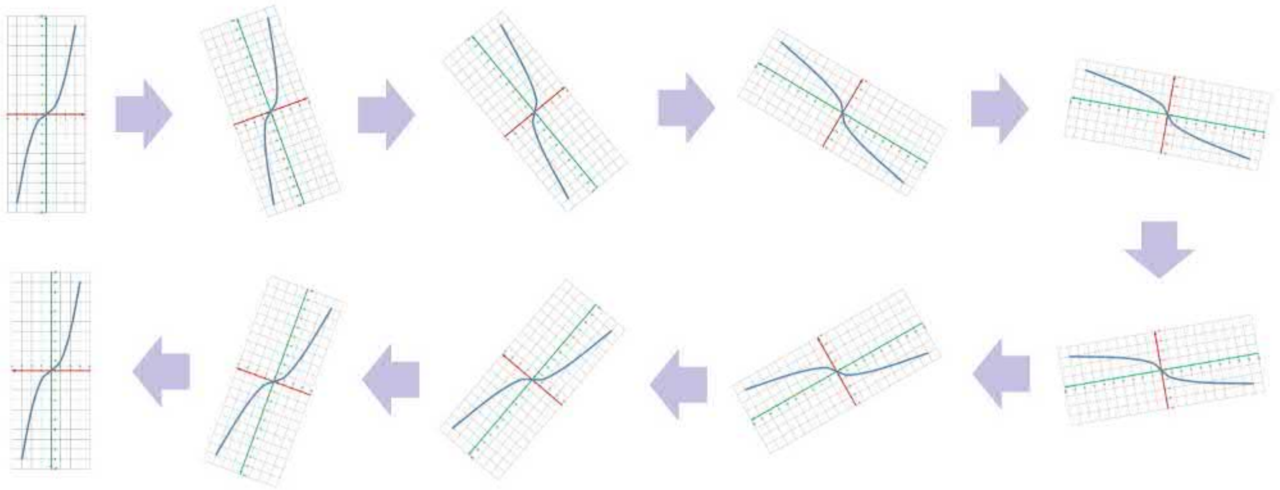
It has 1's on the main diagonal and 0's elsewhere. We can write:

$$I_{ij} = \begin{cases} 1 & \text{if } i = j\,, \\ 0 & \text{if } i \neq j\,. \end{cases}$$

5.4. Examining and building linear operators

The transformations of the plane are illustrated previously with curves plotted on the original plane and then seen transformed. Beyond just saying that the curve was drawn on a piece of paper or a sheet of rubber, what exactly happens to those curves?

Using the graphs of functions to represent these curves in the domain of the transformation fails. For example, rotating such a curve, $y = f(x)$, is likely to produce a curve that isn't the graph of any function, $u = g(v)$, in the codomain:



Our choice is, then, *parametric curves*:

$$P : \mathbf{R} \to \mathbf{R}^2 ,$$

given by:

$$X = P(t) \ \text{ or } \ x = x(t),\ y = y(t) .$$

Example 5.4.1: non-uniform re-scale

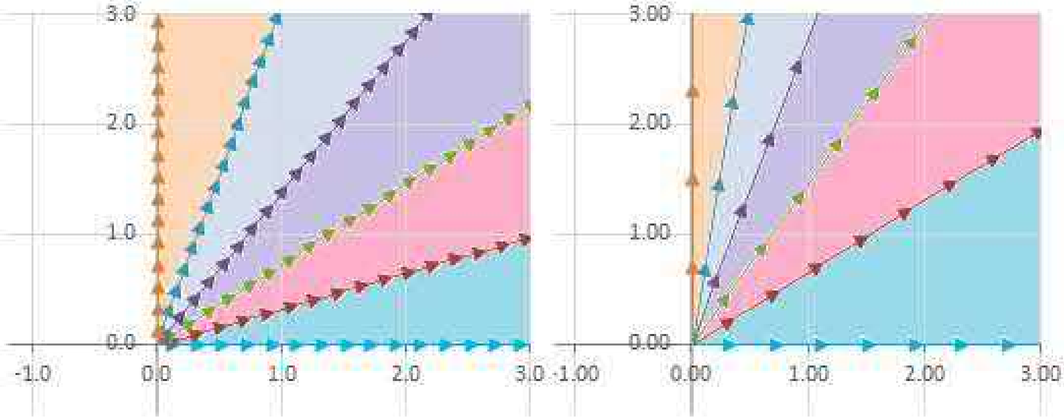
Let's consider this transformation:

$$\begin{cases} u &= 2x \\ v &= 4y \end{cases}$$

Here, this function is given by the matrix:

$$F = \begin{bmatrix} 2 & 0 \\ 0 & 4 \end{bmatrix}$$

We can see what happens to the lines through 0:

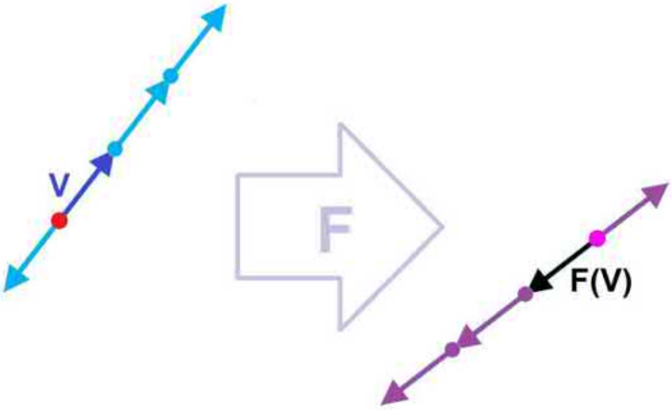


They all remain straight even though some of them rotate.

Let's take a closer look at the straight lines that pass through the origin. The equation of such a line is very simple:

$$P(t) = tV\,,$$

where $V \neq 0$ is some fixed vector. As V is transformed by F, then so are all of its multiples:



It's another straight line. Linear operators don't bend!

We confirm the observation:

Theorem 5.4.2: Images of Lines

The image of a straight line through the origin under a linear operator will produce another straight line thorough the origin.

Proof.

Suppose F is such an operator and V is the direction vector of the line. Then:

$$P(t) = tV \quad \xrightarrow{F} \quad (F \circ P)(t) = F(tV) = tF(V)$$

This is a line with $F(V)$ as the direction vector.

In general, we witness the following:

$$\text{parametric curve} \quad \xrightarrow{F} \quad \text{parametric curve}$$

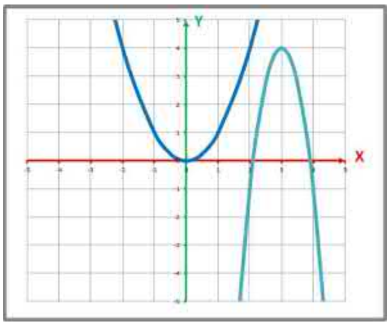
Let's consider two more basic types of curves.

The graph of a quadratic polynomial is called a "parabola". This is what we care about:

- ▶ Any parabola can be acquired from *the* parabola of $f(x) = x^2$ via a vertical stretch or shrink.

It follows from the fact that every quadratic polynomial has its *vertex form*:

$$f(x) = a(x-h)^2 + k\,.$$

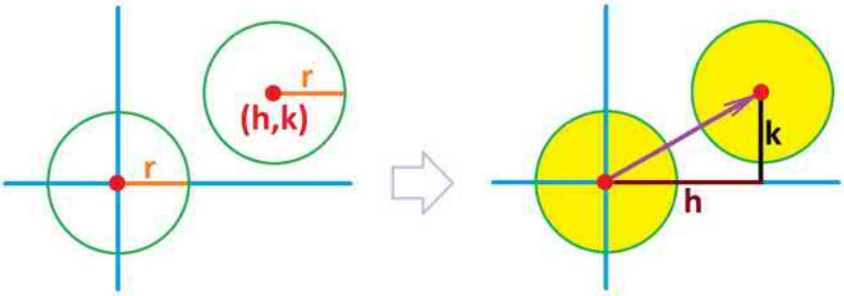


There are many circles on the plane. This is what we care about:

- ▶ Any circle can be acquired from *the* circle of $x^2 + y^2 = 1$ via a uniform stretch or shrink.

It follows from the fact that every circle has its *centered form*:

$$(x-h)^2 + (y-k)^2 = r^2 .$$



Exercise 5.4.3

Explain how the graph of any exponential function a^x can be acquired from the natural one e^x via linear transformations. Same for the logarithms.

We take advantage of our knowledge of vector algebra to summarize these three cases:

	"template"	relation	parametric
/	line: $y = x$	$y - 3 = 2(x-1)$ vertical stretch by 2, shift up by $<1,3>$	$(x,y) = (1,3) + t < 1,2 >$
⌣	parabola: $y = x^2$	$y - 3 = 2(x-1)^2$ vertical stretch by 2, shift up by $<1,3>$	$(x,y) = (1,3) + < t, 2t^2 >$
○	circle: $x^2 + y^2 = 1$	$(y-3)^2 + (x-1)^2 = 2^2$ uniform stretch by 2, shift up by $<1,3>$	$(x,y) = (1,3) + 2 < \cos t, \sin t >$

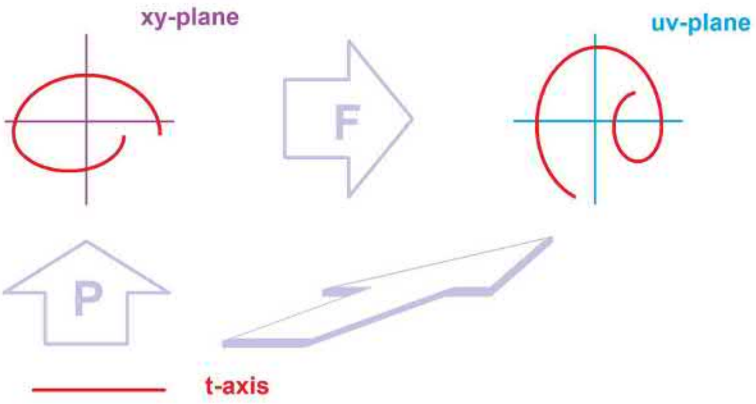
We will need to visualize examples of linear operators on the plane:

$$F : \mathbf{R}^2 \to \mathbf{R}^2 .$$

We illustrate them by creating marks on the original plane and then see what happens to them as they appear on the new plane. Each of these marks will be a parametric curve in the domain:

$$P : \mathbf{R} \to \mathbf{R}^2 ,\ X = P(t) .$$

We then plot its image in the codomain under the transformation $Y = F(X)$:



To be precise, we plot the image of the *composition* of the two functions:

$$F \circ P : \mathbf{R} \to \mathbf{R}^2,\ Y = F(P(t))\,.$$

It is also a parametric curve.

In other words, we have:

$$\begin{tikzcd} \mathbf{R}^2 \arrow[r, "F"] & \mathbf{R}^2 \\ \mathbf{R} \arrow[u, "P"'] \arrow[ur, "F\circ P"'] & \end{tikzcd}$$

We will plot many pairs of curves each time:

$$\begin{tikzcd} & \mathbf{R}^2 & & & & \mathbf{R}^2 \\ \text{old curve:} & & \longrightarrow & \text{new curve:} & & \\ & \mathbf{R} \arrow[uu, "P"'] & & & \mathbf{R} \arrow[uur, "F\circ P"'] & \end{tikzcd}$$

We can use this setup in two ways:

1. We can study a curve by applying various transformations to the plane.
2. We can study a transformation by applying it to various curves.

We did the former in the beginning of the section. Now the latter.

Example 5.4.4: stretch

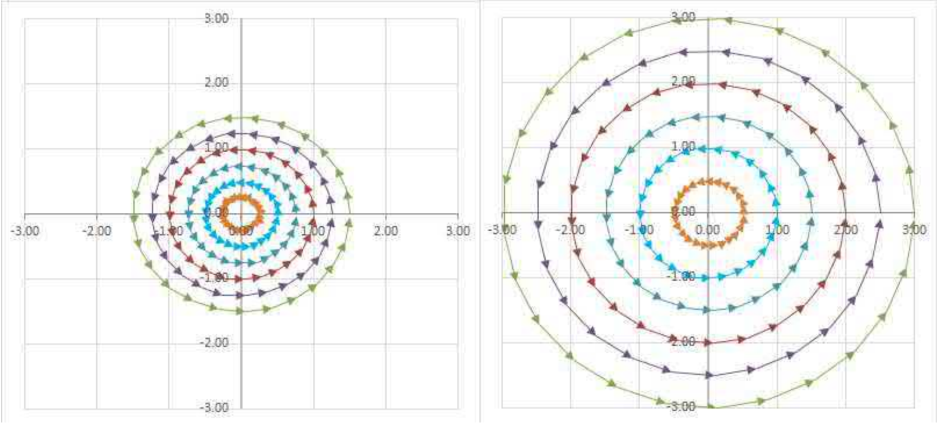
So, applying transformations to curves will give us new curves. For example, we start with the circles:

$$P(t) = r < \cos t, \sin t >,\ r > 0\,.$$

Then, using scalar multiplication by 2 on all vectors means *stretching radially* the whole space. We then discover that the image of the curve is given by:

$$Q(t) = 2P(t) = 2r < \cos t, \sin t > \ .$$

It is a parametric curve of the circle of radius $2r$:



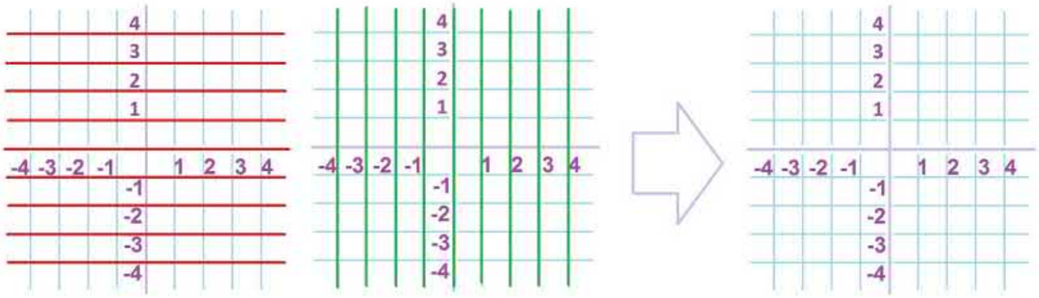
What curves do we choose? Familiar ones: straight lines and circles. How many? The whole *grid*.

We have two possibilities:

Cartesian grid: a rectangular grid of lines

Polar grid: a grid of circles and radii

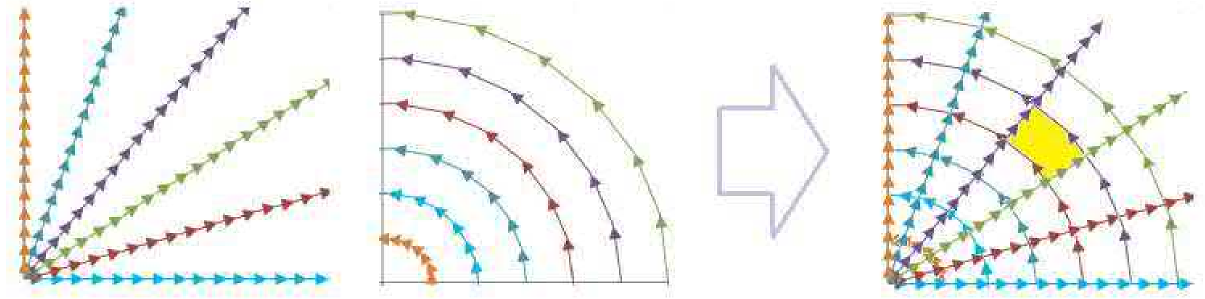
The Cartesian grid is created with these two types of lines:



These lines are defined parametrically:

1. horizontal: $x = t,\ y = k,\ k = ... - 3, -2, -1, 0, 1, 2, 3, ...$
2. vertical: $x = k,\ y = t,\ k = ... - 3, -2, -1, 0, 1, 2, 3, ...$

The polar grid is created with these two types of lines:



Here they are defined parametrically:

1. rays: $x = at,\ y = bt,\ a, b$ real; and
2. circles: $x = r\cos t,\ y = r\sin t,\ r$ real.

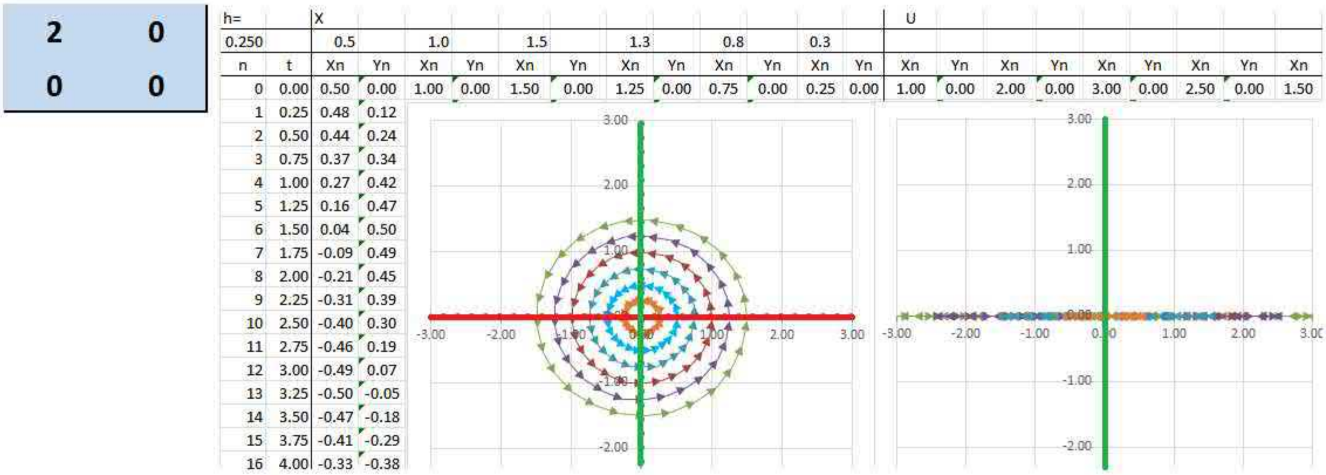
We will apply the Cartesian or the polar as needed.

Example 5.4.5: collapse on axis

Let's consider this very simple function:

$$\begin{cases} u &= 2x \\ v &= 0 \end{cases}, \text{ re-written: } \begin{bmatrix} u \\ v \end{bmatrix} = \begin{bmatrix} 2 & 0 \\ 0 & 0 \end{bmatrix} \cdot \begin{bmatrix} x \\ y \end{bmatrix}.$$

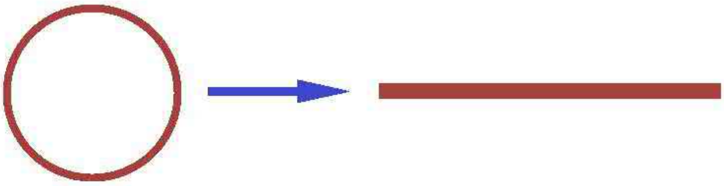
Below, one can see how this function collapses the whole plane to the x-axis:



2	0
0	0

h=		X												U								
0.250		0.5		1.0		1.5		1.3		0.8		0.3										
n	t	Xn	Yn	Xn	Yn	Xn	Yn	Xn	Yn	Xn	Yn	Xn	Yn	Xn	Yn	Xn	Yn	Xn	Yn	Xn	Yn	Xn
0	0.00	0.50	0.00	1.00	0.00	1.50	0.00	1.25	0.00	0.75	0.00	0.25	0.00	1.00	0.00	2.00	0.00	3.00	0.00	2.50	0.00	1.50
1	0.25	0.48	0.12																			
2	0.50	0.44	0.24																			
3	0.75	0.37	0.34																			
4	1.00	0.27	0.42																			
5	1.25	0.16	0.47																			
6	1.50	0.04	0.50																			
7	1.75	-0.09	0.49																			
8	2.00	-0.21	0.45																			
9	2.25	-0.31	0.39																			
10	2.50	-0.40	0.30																			
11	2.75	-0.46	0.19																			
12	3.00	-0.49	0.07																			
13	3.25	-0.50	-0.05																			
14	3.50	-0.47	-0.18																			
15	3.75	-0.41	-0.29																			
16	4.00	-0.33	-0.38																			

This is what happens to every circle:



In the meantime, the x-axis is stretched by a factor of 2. We can see both in the matrix:

$$\left.\begin{array}{rcccc}\text{stretch of } x & \to & 2, & 0 & \leftarrow \\ y \text{ doesn't depends on } x & \to & 0, & 0 & \leftarrow\end{array}\right\}\text{collapse}$$

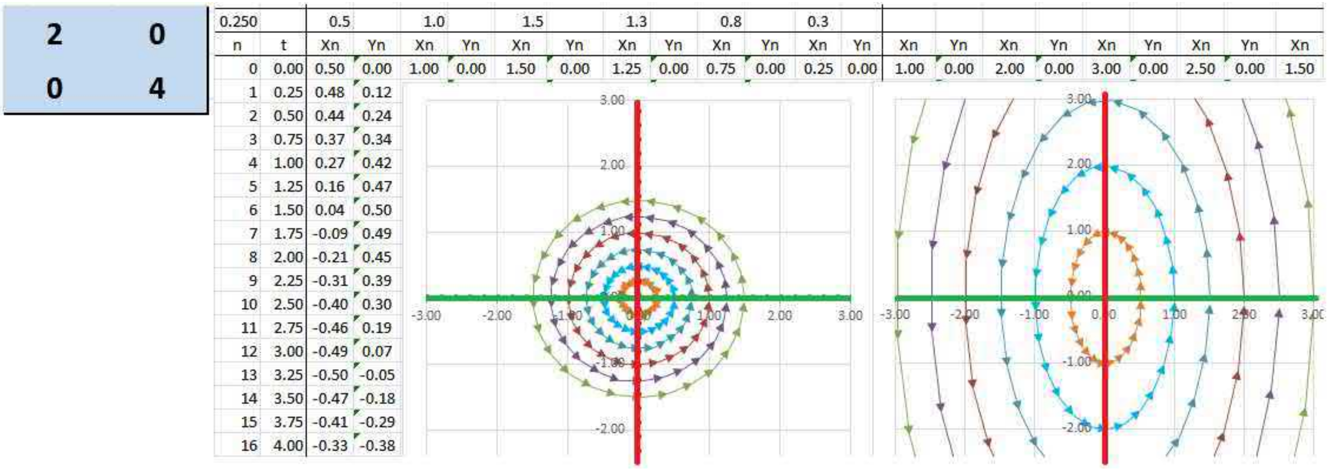
Because of the collapse of the y-axis, the function is neither one-to-one nor onto.

Example 5.4.6: stretch-shrink along axes

Let's revisit this linear operator:

$$F = \begin{bmatrix} 2 & 0 \\ 0 & 4 \end{bmatrix}.$$

We look at the circles first:



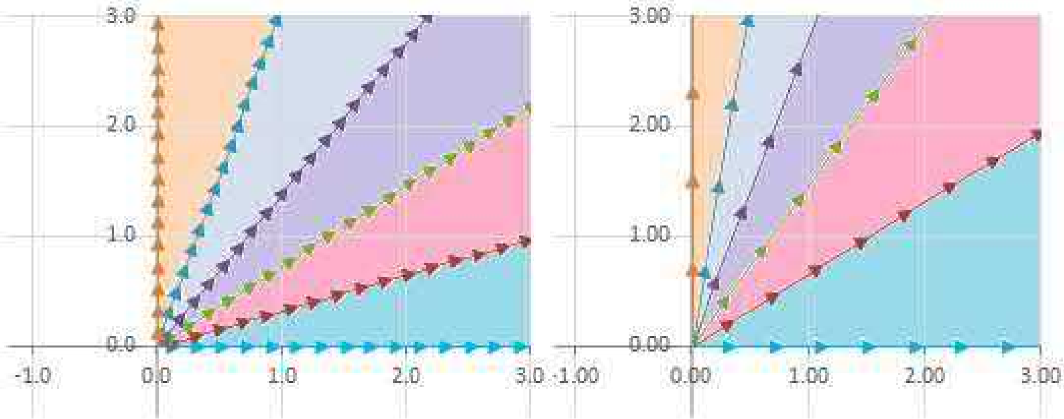
2	0
0	4

0.250		0.5		1.0		1.5		1.3		0.8		0.3										
n	t	Xn	Yn	Xn	Yn	Xn	Yn	Xn	Yn	Xn	Yn	Xn	Yn	Xn	Yn	Xn	Yn	Xn	Yn	Xn	Yn	Xn
0	0.00	0.50	0.00	1.00	0.00	1.50	0.00	1.25	0.00	0.75	0.00	0.25	0.00	1.00	0.00	2.00	0.00	3.00	0.00	2.50	0.00	1.50
1	0.25	0.48	0.12																			
2	0.50	0.44	0.24																			
3	0.75	0.37	0.34																			
4	1.00	0.27	0.42																			
5	1.25	0.16	0.47																			
6	1.50	0.04	0.50																			
7	1.75	-0.09	0.49																			
8	2.00	-0.21	0.45																			
9	2.25	-0.31	0.39																			
10	2.50	-0.40	0.30																			
11	2.75	-0.46	0.19																			
12	3.00	-0.49	0.07																			
13	3.25	-0.50	-0.05																			
14	3.50	-0.47	-0.18																			
15	3.75	-0.41	-0.29																			
16	4.00	-0.33	-0.38																			

The circles have become ellipses! We can see what happens in the matrix:

$$\begin{array}{rccccl}\text{stretch of } x & \to & 2, & 0 & \leftarrow & x \text{ doesn't depends on } y \\ y \text{ doesn't depends on } x & \to & 0, & 4 & \leftarrow & \text{stretch of } y\end{array}$$

The axes stay put. What happens to the rest of the plane? Let's look at the lines now:



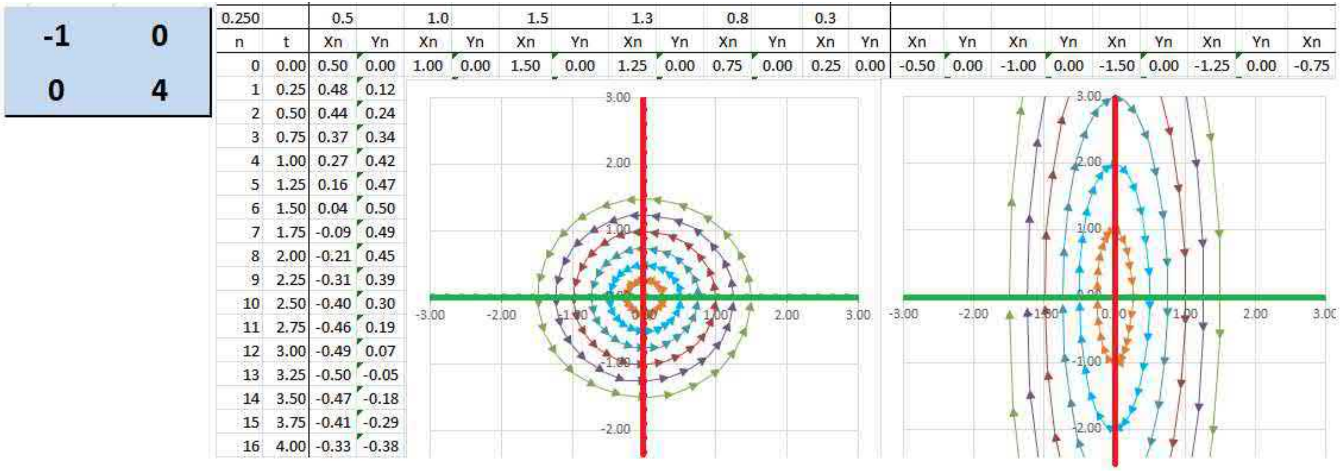
Since the stretching is non-uniform, the vectors turn. However, since the basis vectors e_1 and e_2 don't turn, this is not a rotation but rather a "fanning out" of the vectors. Their slopes have increased. We also discover that the function is both one-to-one and onto.

Example 5.4.7: stretch-shrink along axes

A slightly different function is:

$$\begin{cases} u &= -x \\ v &= 4y \end{cases}$$

It is simple because the two variables are fully separated. Just the circles:



-1	0
0	4

0.250		0.5		1.0		1.5		1.3		0.8		0.3										
n	t	Xn	Yn	Xn	Yn	Xn	Yn	Xn	Yn	Xn	Yn	Xn	Yn	Xn	Yn	Xn	Yn	Xn	Yn	Xn	Yn	Xn
0	0.00	0.50	0.00	1.00	0.00	1.50	0.00	1.25	0.00	0.75	0.00	0.25	0.00	-0.50	0.00	-1.00	0.00	-1.50	0.00	-1.25	0.00	-0.75
1	0.25	0.48	0.12																			
2	0.50	0.44	0.24																			
3	0.75	0.37	0.34																			
4	1.00	0.27	0.42																			
5	1.25	0.16	0.47																			
6	1.50	0.04	0.50																			
7	1.75	-0.09	0.49																			
8	2.00	-0.21	0.45																			
9	2.25	-0.31	0.39																			
10	2.50	-0.40	0.30																			
11	2.75	-0.46	0.19																			
12	3.00	-0.49	0.07																			
13	3.25	-0.50	-0.05																			
14	3.50	-0.47	-0.18																			
15	3.75	-0.41	-0.29																			
16	4.00	-0.33	-0.38																			

The slight change to the function produces a similar but different pattern: We see the reversal of the direction of the ellipse around the origin. We way that *the orientation has changed*. The matrix of F is still diagonal:

$$F = \begin{bmatrix} -1 & 0 \\ 0 & 4 \end{bmatrix}.$$

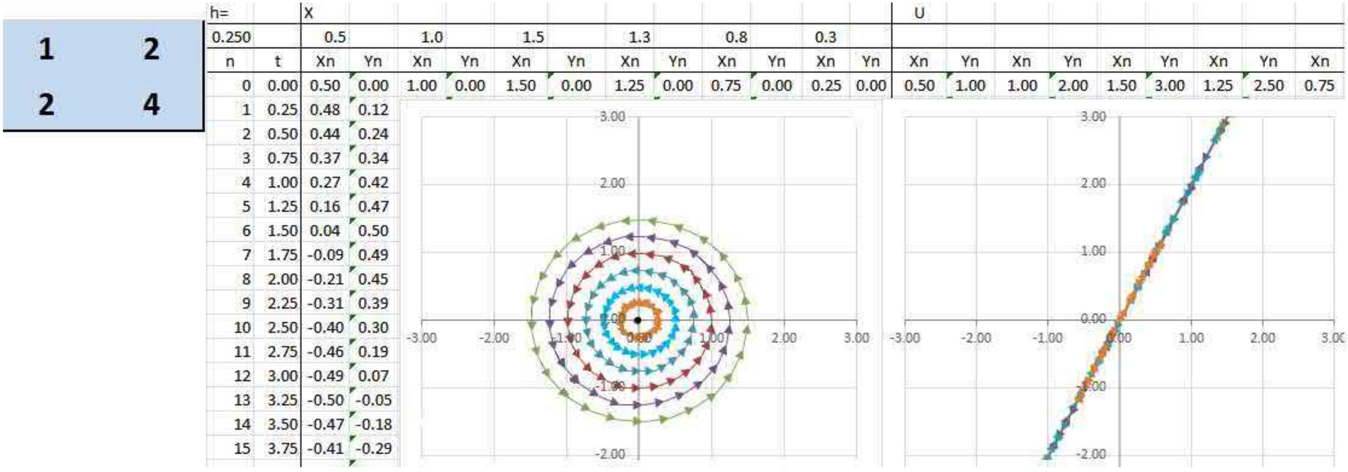
The function is both one-to-one and onto.

Example 5.4.8: experiment

Let's consider a more general function:

$$\begin{cases} u &= x & +2y \\ v &= 2x & +4y \end{cases} \implies F = \begin{bmatrix} 1 & 2 \\ 2 & 4 \end{bmatrix}.$$

It is hard to tell what it does, judging by its matrix. We *experiment*:



1	2
2	4

h=		X												U								
0.250		0.5		1.0		1.5		1.3		0.8		0.3										
n	t	Xn	Yn	Xn	Yn	Xn	Yn	Xn	Yn	Xn	Yn	Xn	Yn	Xn	Yn	Xn	Yn	Xn	Yn	Xn	Yn	Xn
0	0.00	0.50	0.00	1.00	0.00	1.50	0.00	1.25	0.00	0.75	0.00	0.25	0.00	0.50	1.00	1.00	2.00	1.50	3.00	1.25	2.50	0.75
1	0.25	0.48	0.12																			
2	0.50	0.44	0.24																			
3	0.75	0.37	0.34																			
4	1.00	0.27	0.42																			
5	1.25	0.16	0.47																			
6	1.50	0.04	0.50																			
7	1.75	-0.09	0.49																			
8	2.00	-0.21	0.45																			
9	2.25	-0.31	0.39																			
10	2.50	-0.40	0.30																			
11	2.75	-0.46	0.19																			
12	3.00	-0.49	0.07																			
13	3.25	-0.50	-0.05																			
14	3.50	-0.47	-0.18																			
15	3.75	-0.41	-0.29																			

It appears that the function is stretching the plane in one direction and collapsing in another. That's why there is a whole line of points X with $FX = 0$. To find it, we solve this equation:

$$\begin{cases} x & +2y & = 0 \\ 2x & +4y & = 0 \end{cases} \implies x = -2y.$$

The vector $< 1, 2 >$ is, in fact, visible in the matrix. Because of the collapse of the green line to the origin, the function is neither one-to-one nor onto.

Exercise 5.4.9

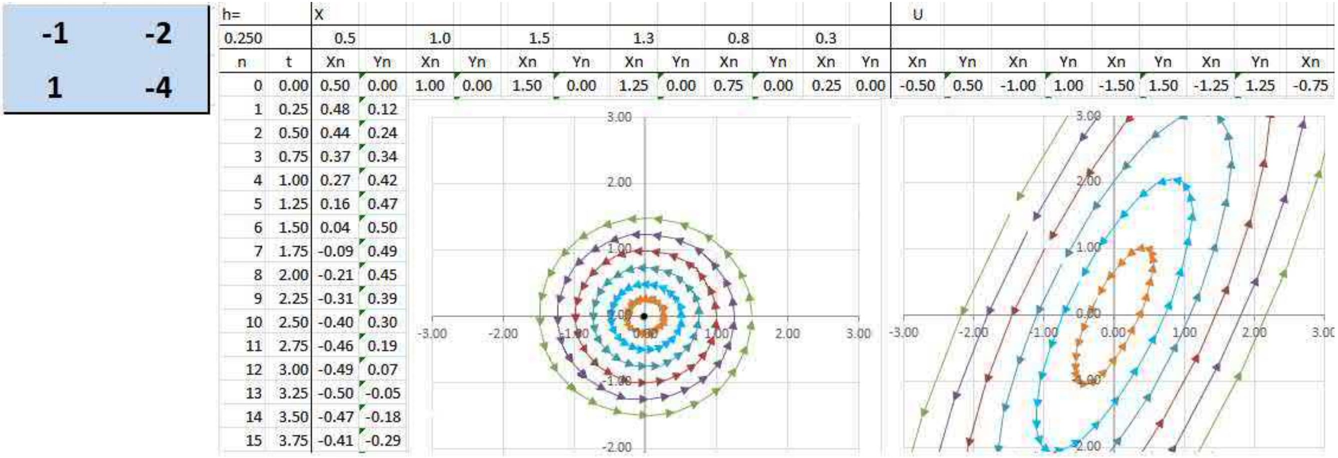
Show where each point goes.

Example 5.4.10: experiment

Consider the following matrix F:

$$F = \begin{bmatrix} -1 & -2 \\ 1 & -4 \end{bmatrix} .$$

Again, it's too complex to reveal what it does, and we have to experiment:



-1	-2
1	-4

h=		X												U								
0.250		0.5		1.0		1.5		1.3		0.8		0.3										
n	t	Xn	Yn	Xn	Yn	Xn	Yn	Xn	Yn	Xn	Yn	Xn	Yn	Xn	Yn	Xn	Yn	Xn	Yn	Xn	Yn	Xn
0	0.00	0.50	0.00	1.00	0.00	1.50	0.00	1.25	0.00	0.75	0.00	0.25	0.00	-0.50	0.50	-1.00	1.00	-1.50	1.50	-1.25	1.25	-0.75
1	0.25	0.48	0.12																			
2	0.50	0.44	0.24																			
3	0.75	0.37	0.34																			
4	1.00	0.27	0.42																			
5	1.25	0.16	0.47																			
6	1.50	0.04	0.50																			
7	1.75	-0.09	0.49																			
8	2.00	-0.21	0.45																			
9	2.25	-0.31	0.39																			
10	2.50	-0.40	0.30																			
11	2.75	-0.46	0.19																			
12	3.00	-0.49	0.07																			
13	3.25	-0.50	-0.05																			
14	3.50	-0.47	-0.18																			
15	3.75	-0.41	-0.29																			

It looks like a non-uniform stretch along diagonal directions. The function is both one-to-one and onto.

Exercise 5.4.11

Describe what is happening under this operator F:

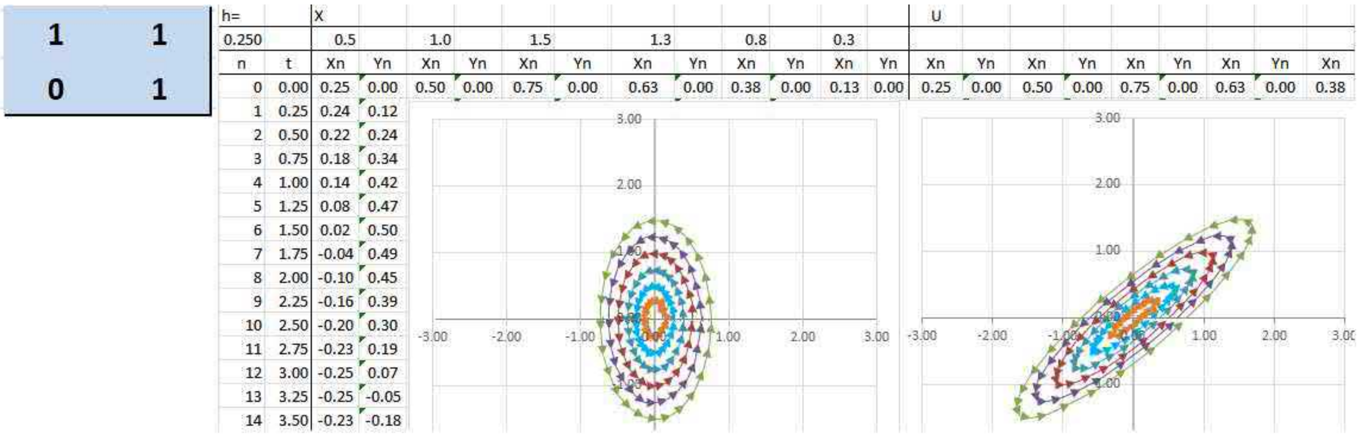
$$F = \begin{bmatrix} 1 & 2 \\ 3 & 2 \end{bmatrix} .$$

Example 5.4.12: skewing-shearing

Consider this matrix:

$$F = \begin{bmatrix} 1 & 1 \\ 0 & 1 \end{bmatrix} .$$

Below, we replace circles with *ellipses* and see what happens to them under such a function:

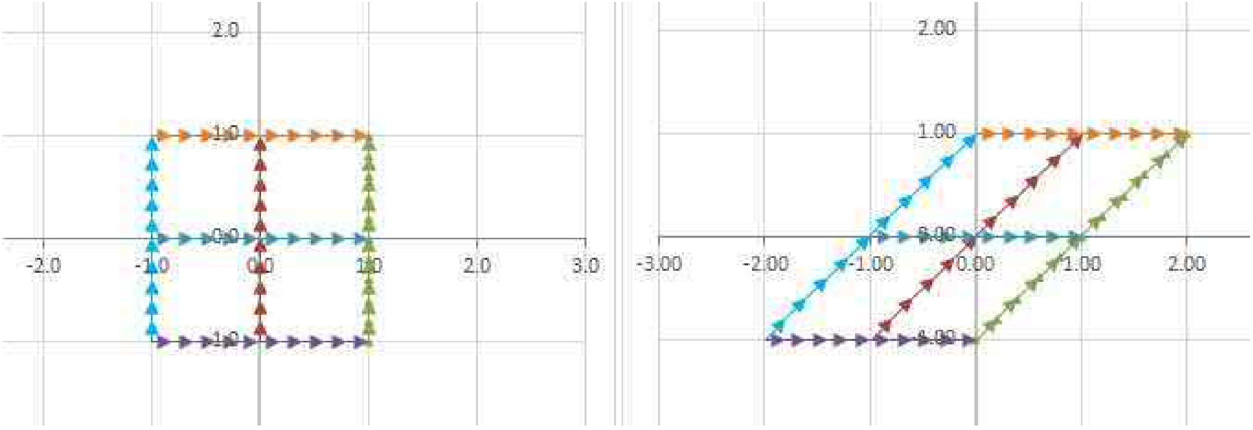


1	1
0	1

h=		X												U								
0.250		0.5		1.0		1.5		1.3		0.8		0.3										
n	t	Xn	Yn	Xn	Yn	Xn	Yn	Xn	Yn	Xn	Yn	Xn	Yn	Xn	Yn	Xn	Yn	Xn	Yn	Xn	Yn	Xn
0	0.00	0.25	0.00	0.50	0.00	0.75	0.00	0.63	0.00	0.38	0.00	0.13	0.00	0.25	0.00	0.50	0.00	0.75	0.00	0.63	0.00	0.38
1	0.25	0.24	0.12																			
2	0.50	0.22	0.24																			
3	0.75	0.18	0.34																			
4	1.00	0.14	0.42																			
5	1.25	0.08	0.47																			
6	1.50	0.02	0.50																			
7	1.75	-0.04	0.49																			
8	2.00	-0.10	0.45																			
9	2.25	-0.16	0.39																			
10	2.50	-0.20	0.30																			
11	2.75	-0.23	0.19																			
12	3.00	-0.25	0.07																			
13	3.25	-0.25	-0.05																			
14	3.50	-0.23	-0.18																			

There seems to be no stretch along the x-axis. There is still angular stretch-shrink but this time it is between the two ends of the same line.

To see more clearly, consider the Cartesian grid. This is what happens to a square:



The plane is skewed, like a deck of cards:

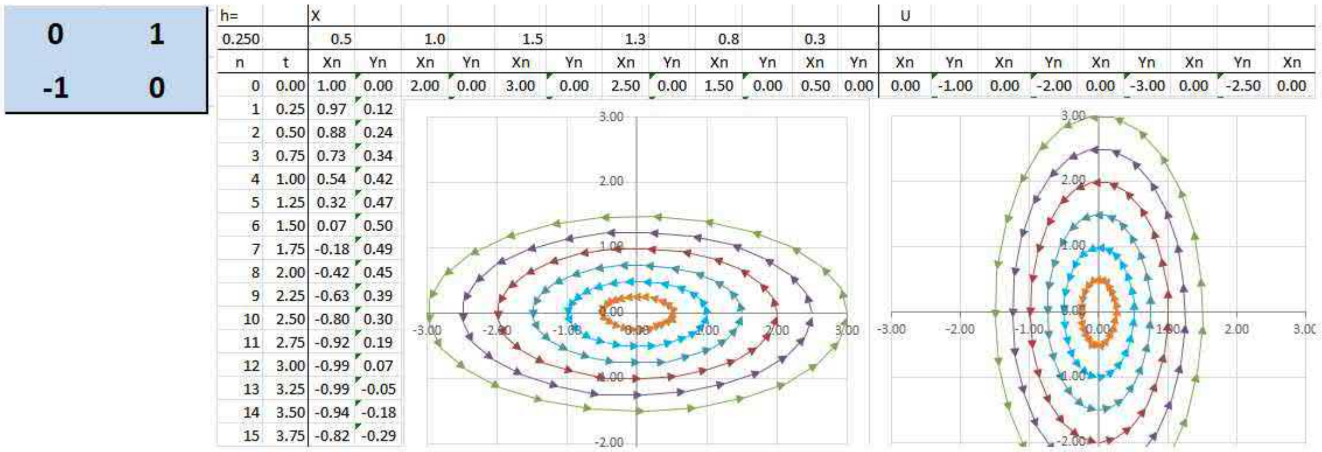
Such a skewing can be carried out with any image-editing software. The function is both one-to-one and onto.

Example 5.4.13: rotation $\pi/2$

Consider a rotation through 90 degrees: (x,y) becomes $(-y,x)$. We have:

$$\begin{cases} u &= \quad -y \\ v &= x \end{cases} \quad \text{, re-written:} \quad \begin{bmatrix} u \\ v \end{bmatrix} = \begin{bmatrix} 0 & 1 \\ -1 & 0 \end{bmatrix} \cdot \begin{bmatrix} x \\ y \end{bmatrix} .$$

The experiment confirms what we know:



0	1
-1	0

h=		X												U								
0.250		0.5		1.0		1.5		1.3		0.8		0.3										
n	t	Xn	Yn	Xn	Yn	Xn	Yn	Xn	Yn	Xn	Yn	Xn	Yn	Xn	Yn	Xn	Yn	Xn	Yn	Xn	Yn	Xn
0	0.00	1.00	0.00	2.00	0.00	3.00	0.00	2.50	0.00	1.50	0.00	0.50	0.00	0.00	-1.00	0.00	-2.00	0.00	-3.00	0.00	-2.50	0.00
1	0.25	0.97	0.12																			
2	0.50	0.88	0.24																			
3	0.75	0.73	0.34																			
4	1.00	0.54	0.42																			
5	1.25	0.32	0.47																			
6	1.50	0.07	0.50																			
7	1.75	-0.18	0.49																			
8	2.00	-0.42	0.45																			
9	2.25	-0.63	0.39																			
10	2.50	-0.80	0.30																			
11	2.75	-0.92	0.19																			
12	3.00	-0.99	0.07																			
13	3.25	-0.99	-0.05																			
14	3.50	-0.94	-0.18																			
15	3.75	-0.82	-0.29																			

We've had many examples, but how do we *build* a linear operator from a description?

The solution relies on the following simple observation:

Theorem 5.4.14: Columns are Values of Basis Vectors

The two columns of the matrix of a linear operator are the values of the two basis vectors under this operator:

$$F=\left[\begin{array}{cc} a & b \\ c & d \end{array}\right] \Longrightarrow F\left[\begin{array}{c} 1 \\ 0 \end{array}\right]=\left[\begin{array}{c} a \\ c \end{array}\right] \quad \text{and} \quad F\left[\begin{array}{c} 0 \\ 1 \end{array}\right]=\left[\begin{array}{c} b \\ d \end{array}\right].$$

Exercise 5.4.15

Prove the theorem.

The converse is just as important:

Theorem 5.4.16: Values of Basis Vectors Are Columns

The matrix of a linear operator is fully determined by the values of the two basis vectors under this operator.

In other words, we *merge* the two column-vectors into a matrix:

$$F:\left[\begin{array}{c} 1 \\ 0 \end{array}\right] \mapsto \left[\begin{array}{c} a \\ c \end{array}\right], \quad F:\left[\begin{array}{c} 0 \\ 1 \end{array}\right] \mapsto \left[\begin{array}{c} b \\ d \end{array}\right] \quad \text{Merge:} \quad F=\left[\begin{array}{cc} a & b \\ c & d \end{array}\right].$$

Example 5.4.17: matrices from values

This is the zero operator:

$$\left[\begin{array}{c} 1 \\ 0 \end{array}\right] \mapsto \left[\begin{array}{c} 0 \\ 0 \end{array}\right], \quad \left[\begin{array}{c} 0 \\ 1 \end{array}\right] \mapsto \left[\begin{array}{c} 0 \\ 0 \end{array}\right] \quad \text{Merge:} \quad 0=\left[\begin{array}{cc} 0 & 0 \\ 0 & 0 \end{array}\right].$$

This is the identity:

$$\left[\begin{array}{c} 1 \\ 0 \end{array}\right] \mapsto \left[\begin{array}{c} 1 \\ 0 \end{array}\right], \quad \left[\begin{array}{c} 0 \\ 1 \end{array}\right] \mapsto \left[\begin{array}{c} 0 \\ 1 \end{array}\right] \quad \text{Merge:} \quad I=\left[\begin{array}{cc} 1 & 0 \\ 0 & 1 \end{array}\right].$$

This is the horizontal stretch by 2:

$$\begin{bmatrix}1\\0\end{bmatrix} \mapsto \begin{bmatrix}2\\0\end{bmatrix}, \quad \begin{bmatrix}0\\1\end{bmatrix} \mapsto \begin{bmatrix}0\\1\end{bmatrix} \qquad \text{Merge:} \quad S_x = \begin{bmatrix}2&0\\0&1\end{bmatrix}.$$

And this is the vertical stretch by 3:

$$\begin{bmatrix}1\\0\end{bmatrix} \mapsto \begin{bmatrix}1\\0\end{bmatrix}, \quad \begin{bmatrix}0\\1\end{bmatrix} \mapsto \begin{bmatrix}0\\3\end{bmatrix} \qquad \text{Merge:} \quad S_y = \begin{bmatrix}1&0\\0&3\end{bmatrix}.$$

This is the horizontal flip:

$$\begin{bmatrix}1\\0\end{bmatrix} \mapsto \begin{bmatrix}-1\\0\end{bmatrix}, \quad \begin{bmatrix}0\\1\end{bmatrix} \mapsto \begin{bmatrix}0\\1\end{bmatrix} \qquad \text{Merge:} \quad F_x = \begin{bmatrix}-1&0\\0&1\end{bmatrix}.$$

And this is the vertical flip:

$$\begin{bmatrix}1\\0\end{bmatrix} \mapsto \begin{bmatrix}1\\0\end{bmatrix}, \quad \begin{bmatrix}0\\1\end{bmatrix} \mapsto \begin{bmatrix}0\\-1\end{bmatrix} \qquad \text{Merge:} \quad F_y = \begin{bmatrix}1&0\\0&-1\end{bmatrix}.$$

This is the flip about the diagonal:

$$\begin{bmatrix}1\\0\end{bmatrix} \mapsto \begin{bmatrix}0\\1\end{bmatrix}, \quad \begin{bmatrix}0\\1\end{bmatrix} \mapsto \begin{bmatrix}1\\0\end{bmatrix} \qquad \text{Merge:} \quad F_d = \begin{bmatrix}0&1\\1&0\end{bmatrix}.$$

Warning!

Its matrix is just an abbreviated representation of a linear operator.

Exercise 5.4.18

Suppose a linear operator A:
- leaves the x-axis intact, and
- stretches the y-axis by a factor of 2.

Find the matrix of A.

Exercise 5.4.19

Suppose a linear operator A:
- rotates the x-axis 45 degrees clockwise, and
- flips the y-axis.

Find the matrix of A.

Exercise 5.4.20

Suppose a linear operator A:
- leaves the x-axis intact, and
- stretches the diagonal $y = x$ by a factor of 2.

Find the matrix of A.

Exercise 5.4.21

Make up your own linear operator and find its matrix. Repeat.

Let's apply this result to some transformations we have been interested in.

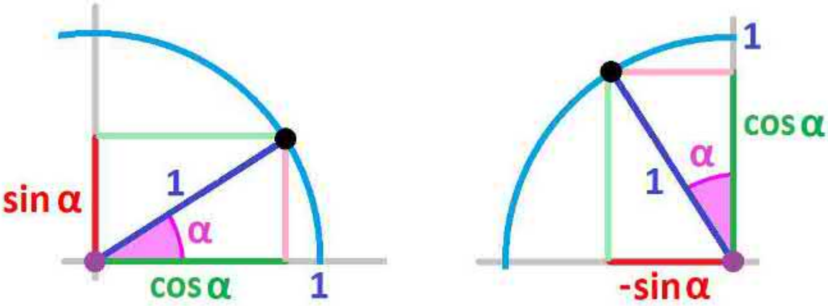
Theorem 5.4.22: Matrix of Rotation

The linear operator of rotation through an angle α is given by the following matrix:

$$\boxed{R=\left[\begin{array}{cc}\cos\alpha & -\sin\alpha\\ \sin\alpha & \cos\alpha\end{array}\right]}$$

Proof.

We only need to see where the basis vectors $<1,0>$ and $<0,1>$ go.



The first one is simple:

$$R\left[\begin{array}{c}1\\ 0\end{array}\right]=\left[\begin{array}{c}\cos\alpha\\ \sin\alpha\end{array}\right].$$

The second one flips the sign of the x-component:

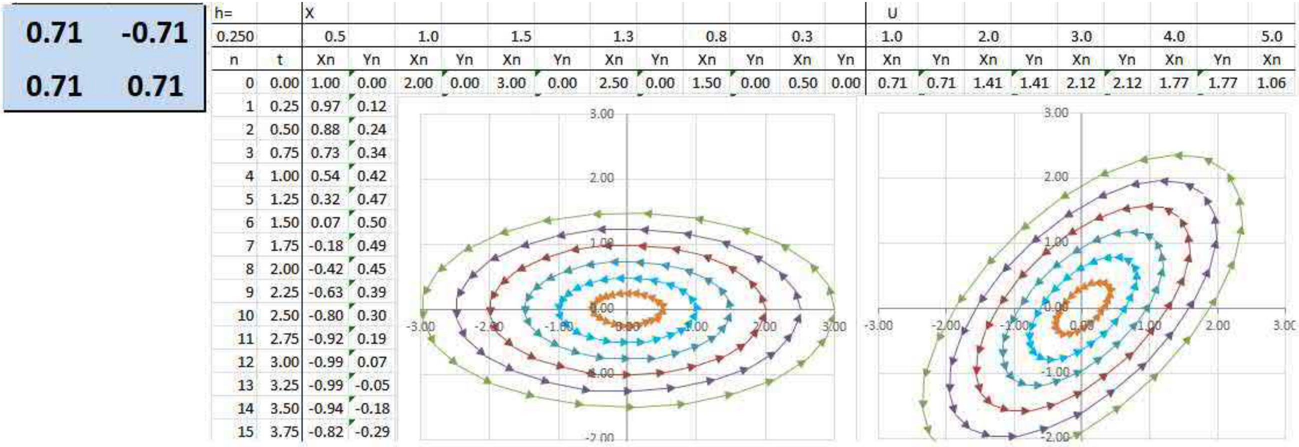
$$R\left[\begin{array}{c}0\\ 1\end{array}\right]=\left[\begin{array}{c}-\sin\alpha\\ \cos\alpha\end{array}\right].$$

Example 5.4.23: rotation $\pi/4$

Consider a rotation through 45 degrees:

$$\begin{cases} u &= \cos\frac{\pi}{4}x \quad -\sin\frac{\pi}{4}y\\ v &= \sin\frac{\pi}{4}x \quad +\cos\frac{\pi}{4}y\end{cases}, \text{ re-written: } \left[\begin{array}{c}u\\ v\end{array}\right]=\left[\begin{array}{cc}\cos\frac{\pi}{4} & -\sin\frac{\pi}{4}\\ \sin\frac{\pi}{4} & \cos\frac{\pi}{4}\end{array}\right]\cdot\left[\begin{array}{c}x\\ y\end{array}\right].$$

We plot ellipses instead of circles to make it easier to see what is happening:



0.71	-0.71
0.71	0.71

h=		X												U								
0.250		0.5		1.0		1.5		1.3		0.8		0.3		1.0		2.0		3.0		4.0		5.0
n	t	Xn	Yn	Xn	Yn	Xn	Yn	Xn	Yn	Xn	Yn	Xn	Yn	Xn	Yn	Xn	Yn	Xn	Yn	Xn	Yn	Xn
0	0.00	1.00	0.00	2.00	0.00	3.00	0.00	2.50	0.00	1.50	0.00	0.50	0.00	0.71	0.71	1.41	1.41	2.12	2.12	1.77	1.77	1.06
1	0.25	0.97	0.12																			
2	0.50	0.88	0.24																			
3	0.75	0.73	0.34																			
4	1.00	0.54	0.42																			
5	1.25	0.32	0.47																			
6	1.50	0.07	0.50																			
7	1.75	-0.18	0.49																			
8	2.00	-0.42	0.45																			
9	2.25	-0.63	0.39																			
10	2.50	-0.80	0.30																			
11	2.75	-0.92	0.19																			
12	3.00	-0.99	0.07																			
13	3.25	-0.99	-0.05																			
14	3.50	-0.94	-0.18																			
15	3.75	-0.82	-0.29																			

The function is both one-to-one and onto.

Example 5.4.24: rotation with stretch-shrink

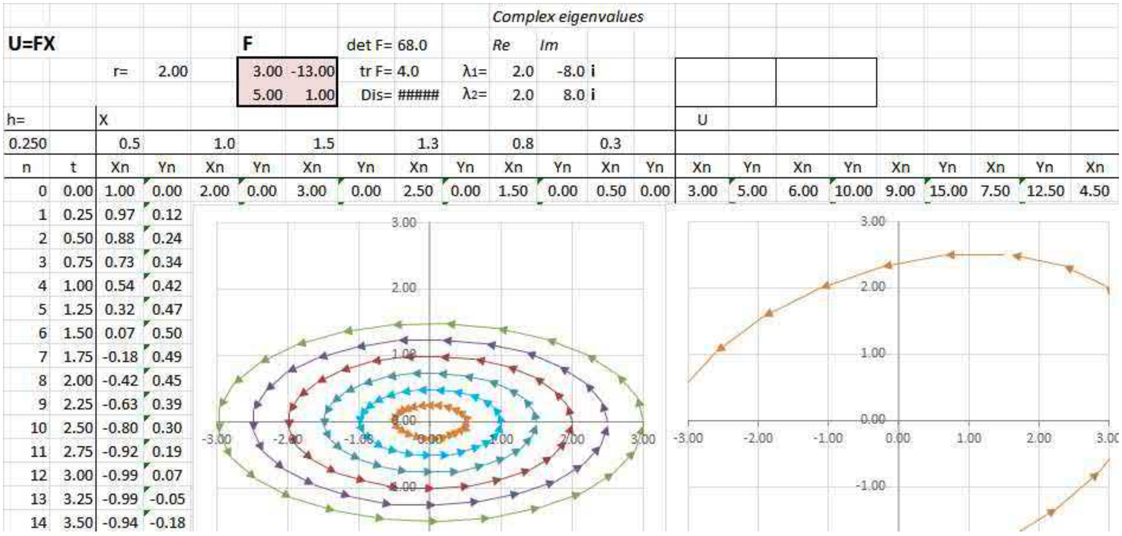
Let's consider a more complex function:

$$\begin{cases} u & = 3x \quad -13y \\ v & = 5x \quad +y \end{cases}.$$

Here, the matrix of F is not diagonal:

$$F = \left[\begin{array}{cc} 3 & -13 \\ 5 & 1 \end{array}\right].$$

This is what happens to our ellipses:



Complex eigenvalues

U=FX **F** det F= 68.0 *Re* *Im*

r= 2.00 3.00 -13.00 tr F= 4.0 λ1= 2.0 -8.0 i

5.00 1.00 Dis= ##### λ2= 2.0 8.0 i

h=		X												U								
0.250		0.5		1.0		1.5		1.3		0.8		0.3										
n	t	Xn	Yn	Xn	Yn	Xn	Yn	Xn	Yn	Xn	Yn	Xn	Yn	Xn	Yn	Xn	Yn	Xn	Yn	Xn	Yn	Xn
0	0.00	1.00	0.00	2.00	0.00	3.00	0.00	2.50	0.00	1.50	0.00	0.50	0.00	3.00	5.00	6.00	10.00	9.00	15.00	7.50	12.50	4.50
1	0.25	0.97	0.12																			
2	0.50	0.88	0.24																			
3	0.75	0.73	0.34																			
4	1.00	0.54	0.42																			
5	1.25	0.32	0.47																			
6	1.50	0.07	0.50																			
7	1.75	-0.18	0.49																			
8	2.00	-0.42	0.45																			
9	2.25	-0.63	0.39																			
10	2.50	-0.80	0.30																			
11	2.75	-0.92	0.19																			
12	3.00	-0.99	0.07																			
13	3.25	-0.99	-0.05																			
14	3.50	-0.94	-0.18																			

We seem to have a combination of stretching, flipping, and rotating... The function is both one-to-one and onto.

For linear operators, there is an easy way to answer the question we have been asking: When is it one-to-one?

Theorem 5.4.25: One-to-one Linear Operator

A linear operator F is one-to-one if and only if the equation $F(X) = 0$ has only the zero solution.

Proof.

Suppose there are two distinct solutions $X \neq Y$. Then, we conclude:

$$F(X) = F(Y) \implies F(X) - F(Y) = 0 \implies F(X - Y) = 0\,.$$

In other words, we have found such a $Z = X - Y \neq 0$ that $F(Z) = 0 = F(0)$.

Exercise 5.4.26

Prove the rest of the theorem.

In other words, F is one-to-one when

$$F(X) = 0 \implies X = 0\,.$$

So, to determine whether a mixture problem has a single solution, we choose, in a twist, to replace it with a mixture problem that requires to produce *zeros* in all equations. Then we ask if this problem has a non-zero solution.

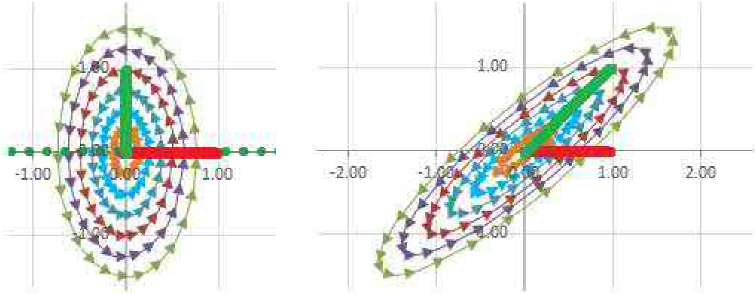
5.5. The determinant of a matrix

Example 5.5.1: one-to-one and not

Consider this matrix:

$$A = \begin{bmatrix} 1 & 1 \\ 0 & 1 \end{bmatrix} .$$

Is the linear operator one-to-one? Every one of our ellipses in the domain has been stretched and maybe rotated but they still cover the the whole plane in the codomain:

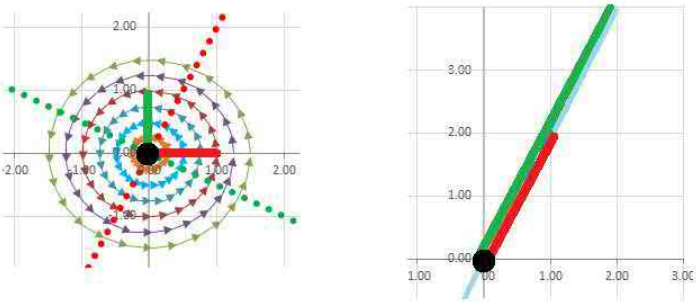


It is one-to-one.

This one is different:

$$A = \begin{bmatrix} 1 & 2 \\ 2 & 4 \end{bmatrix} .$$

We just watch where the two basis vectors go:



What is the difference from the former case? They go to multiples of each other: Their values are proportional to the vector $< 1, 2 >$. We can see that in the matrix that the second column is twice the first:

$$2 \begin{bmatrix} 1 \\ 2 \end{bmatrix} = \begin{bmatrix} 2 \\ 4 \end{bmatrix} .$$

In fact, a whole line of vectors goes to 0; it's not one-to-one!

Definition 5.5.2: singular matrix

A 2×2 matrix A is called *singular* when its two columns are multiples of each other.

So, for a singular matrix

$$A = \begin{bmatrix} a & b \\ c & d \end{bmatrix},$$

there is such an x that:

$$\begin{bmatrix} a \\ c \end{bmatrix} = x \begin{bmatrix} b \\ d \end{bmatrix}.$$

Let's examine this idea: Under what circumstances is a matrix singular?

We break the vector equation above into two scalar equations:

$$a = xb, \;\; c = xd\,.$$

Instead of solving them for x, we assume that there is such an x. We multiply the first by d, and the second by b, and then subtract to eliminate x:

$$\begin{array}{ll} ad & = xbd \\ cb & = xbd \\ \hline ad - bc & = 0 \end{array}$$

We conclude that if such an x exists, then

$$ad - bc = 0\,.$$

In this expression, the terms of the matrix are cross-multiplied and subtracted:

$$\begin{bmatrix} a & b \\ c & d \end{bmatrix} \to ad - bc\,.$$

This number is an important characteristic of the matrix:

Definition 5.5.3: determinant

The *determinant* of a 2×2 matrix A is defined and denoted as follows:

$$\boxed{\det A = \det \begin{bmatrix} a & b \\ c & d \end{bmatrix} = ad - bc}$$

What does the determinant determine?

Theorem 5.5.4: Singular Matrix and Determinant

A 2×2 *matrix* A *is singular if and only if* $\det A = 0$.

Proof.

($\Rightarrow$) Suppose A is singular, then

$$\begin{bmatrix} a \\ c \end{bmatrix} = x \begin{bmatrix} b \\ d \end{bmatrix} \implies \begin{cases} a = xb \\ c = xd \end{cases} \implies \det A = ad - bc = (xb)d - b(xd) = 0\,.$$

($\Leftarrow$) Suppose $ad - bc = 0$, then let's find x, the multiple.

- Case 1: Assume $b \neq 0$, then choose $x = \frac{a}{b}$. Then

$$\begin{aligned} xb &= \frac{a}{b}b & &= a, \\ xd &= \frac{a}{b}d &= \frac{ad}{b} = \frac{bc}{b} &= c. \end{aligned}$$

So

$$x \begin{bmatrix} b \\ d \end{bmatrix} = \begin{bmatrix} a \\ c \end{bmatrix} .$$

- Case 2: Assume $a \neq 0$ ...

Exercise 5.5.5

Finish the proof.

We make the following observation about the determinant, which will also reappear in the case of $n \times n$ matrices:

▶ The determinant is an alternating sum of terms, each of which is the product of n of the matrix's entries, exactly one from each row and exactly one from each column.

Let's consider a special matrix equation, with the zero right-hand side:

$$F(X) = 0\,.$$

It is called a *homogeneous* equation.

We know that there is always at least one solution, the zero vector!

The question is then becomes

▶ Are there any non-zero solutions?

We know that the answer may be provided by a more basic question:

▶ Is the function one-to-one?

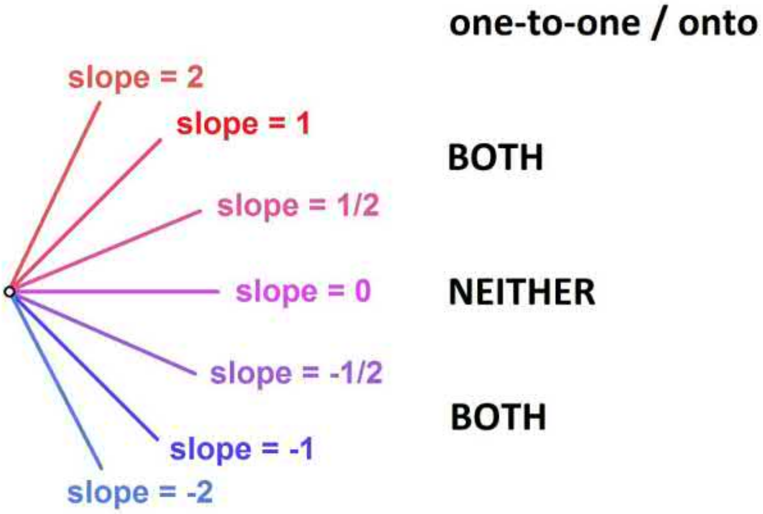
Let's start with dimension 1:

$$f(x) = mx, \quad \text{solve} \quad f(x) = 0\,.$$

This is simple:

$$mx = 0 \implies x = 0 \text{ ... unless } m = 0.$$

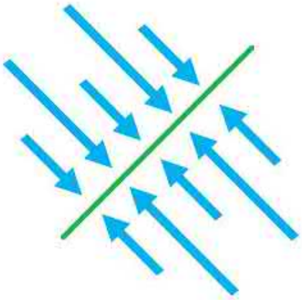
All functions except the constant zero function are one-to-one and, therefore, can only produce a single solution for our equation:



Is there a similar decisive condition in the two-dimensional case? We do have the zero operator (matrix):

$$F(X) = 0 \quad \text{for all} \quad X\,.$$

It collapses the whole plane to the origin. However, there are other, less extreme collapses, projections onto lines:



Then a whole line is taken to 0. They are also not one-to-one!

We deploy simple algebra in order to resolve this issue.

Theorem 5.5.6: Non-Zero Solutions

Suppose A is a 2×2 matrix. Then, $\det A \neq 0$ if and only if the solution set of the matrix equation $AX = 0$ consists of only 0.

Proof.

We will use the Zero Factor Property: The product of two numbers is zero if and only if either one of them (or both) is zero; i.e.,

$$a = 0 \quad \texttt{OR} \quad b = 0 \iff ab = 0\,.$$

Let's solve the system of linear equations:

$$\begin{cases} ax & + & by & = 0, & (1)\\ cx & + & dy & = 0. & (2)\end{cases}$$

From (1), we derive:

$$y = -ax/b, \text{ provided } b \neq 0. \quad (3)$$

Substitute this into (2):

$$cx + d(-ax/b) = 0\,.$$

Then

$$x(c - da/b) = 0\,,$$

or, alternatively,

$$x(cb-da)=0,\ \text{when } b\neq 0\,.$$

One possibility is $x=0$; it follows from (3) that $y=0$ too. Then, we have two cases for $b\neq 0$:

- Case 1: $x=0,\ y=0$, or
- Case 2: $ad-bc=0$.

Case 1 doesn't interest us. In case 2, x is arbitrary and there may be non-zero solutions.

Now, we apply this analysis to y in (1) instead of x; we have for $a\neq 0$:

- Case 1: $x=0,\ y=0$, or
- Case 2: $ad-bc=0$.

The result is the same! Furthermore, if we apply this analysis for x and y in (2) instead of (1), we have the same two cases. Thus, whenever one of the four coefficients, a,b,c,d, is non-zero, we have these cases:

- Case 1: $x=0,\ y=0$, or
- Case 2: $ad-bc=0$.

But when $a=b=c=d=0$, Case 2 is satisfied... and we can have any values for x and y!

According to the analysis above:

$$\det A\neq 0 \implies x=y=0\,.$$

The converse is also true. Indeed, let's consider our system of linear equations again:

$$\begin{cases} ax & + & by & =0, & (1)\\ cx & + & dy & =0. & (2)\end{cases}$$

We multiply (1) by c and (2) by a. Then we have:

$$\left\{\begin{array}{llllll} cax & + & cby & =0 & c\cdot(1)\\ acx & + & ady & =0 & a\cdot(2)\\ \hline (ca-ac)x & + & (cb-ad)y & =0 & \\ 0\cdot x & - & \det A\cdot y & =0 & \end{array}\right.$$

The third equation is the result of subtraction of the first two. The whole equation is zero when $\det A=0$! This means that equations (1) and (2) represent two identical lines on the plane. It follows that the original system has infinitely many solutions.

Exercise 5.5.7

What if $a=0$ or $c=0$?

Example 5.5.8: computing determinants

The flip over the y-axis:

$$\det\begin{bmatrix}-1 & 0\\ 0 & 1\end{bmatrix}=(-1)\cdot 1-0\cdot 0=-1\,.$$

The stretch:

$$\det\begin{bmatrix}\lambda & 0\\ 0 & \mu\end{bmatrix}=\lambda\cdot\mu-0\cdot 0=\lambda\cdot\mu\,.$$

The rotation:

$$\det \begin{bmatrix} \cos\alpha & -\sin\alpha \\ \sin\alpha & \cos\alpha \end{bmatrix} = \cos\alpha\cdot\cos\alpha - (-\sin\alpha)\cdot\sin\alpha = \cos^2\alpha + \sin^2\alpha = 1\,,$$

by the Pythagorean Theorem.

These have non-zero determinants. Meanwhile, the projection on the x-axis has a zero determinant:

$$\det \begin{bmatrix} \lambda & 0 \\ 0 & 0 \end{bmatrix} = \lambda\cdot 0 - 0\cdot 0 = 0\,.$$

As you can see, we can derive some more information from the value of the determinant than just that it's one-to-one.

The following is the *contra-positive* form of the theorem:

Corollary 5.5.9: Non-Zero Solutions

Suppose A *is a* 2×2 *matrix. Then, there is such an* $X\neq 0$ *that* $AX=0$ *if and only if* $\det A = 0$.

Since $A(0)=0$, this indicates that A isn't one-to-one. There is more:

Corollary 5.5.10: Bijections and Determinants

Suppose A *is a* 2×2 *matrix. It is a bijection if and only if* $\det A \neq 0$.

Exercise 5.5.11

Prove the rest of the theorem.

So, a zero determinant indicates that some non-zero vector X is taken to 0 by A. It follows that all the multiples, kX, of X are also taken to 0:

$$A(kX) = kA(X) = k0 = 0\,.$$

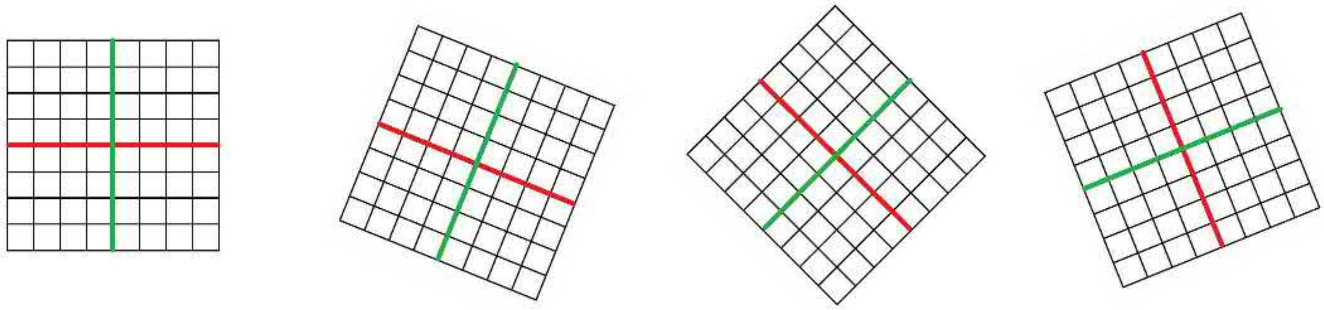
In other words, the whole line is collapsed to 0.

Theorem 5.5.12: Line Collapses

If a vector is taken to zero by a linear operator, then the whole line in the direction of this vector from the origin is taken to zero:

$$A(X)=0 \implies A\Big(\{Y:\ Y=kX,\ k \text{ real}\,\}\Big)=0\,.$$

We can place different coordinate systems on the same plane. The origin is the same, but the units and the angles of the axes may be different:



We know that the vector algebra remains the same. However, a component representation of a vector does depend on our choice of the Cartesian system. Therefore, a matrix representation of a linear operator depends on our choice of the Cartesian system too. Remarkably, this isn't true for the determinant! The following important fact is accepted without proof:

Corollary 5.5.13: Determinant Is Intrinsic

The determinant of a linear operator remains the same in any Cartesian coordinate system.

The determinant will tell us a lot about the linear operator:

- $\det A < 0$ indicates the presence of a flip.
- $|\det A| = 1$ indicates that this is a motion.
- $\det A = 0$ indicates the collapse or the presence of a projection.

But how do we detect stretches or rotations?

5.6. It's a stretch: eigenvalues and eigenvectors

An easy observation about rotation is that if just one vector isn't rotated, there is no rotation!

If a vector isn't rotated, what can possibly happen to it under a linear operator? A stretch (with a possible flip). In other words, it's scalar multiplication:

$$V \ \mapsto\ \lambda V\,,$$

for some real λ.

Example 5.6.1: re-scaling

Consider a linear operator given by the matrix:

$$A = \begin{bmatrix} 2 & 0 \\ 0 & 3 \end{bmatrix}.$$

What exactly does this transformation of the plane do to it? To answer, just consider where A takes the standard basis vectors:

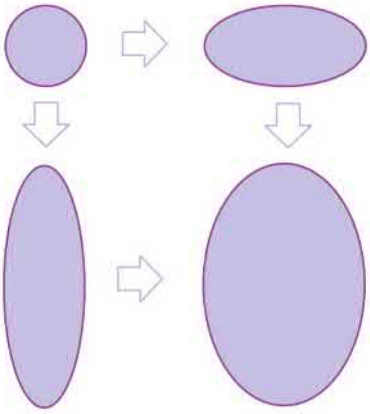
$$A: \quad e_1 = \begin{bmatrix} 1 \\ 0 \end{bmatrix} \quad \mapsto \quad \begin{bmatrix} 2 \\ 0 \end{bmatrix} \quad = 2e_1$$

$$A: \quad e_2 = \begin{bmatrix} 0 \\ 1 \end{bmatrix} \quad \mapsto \quad \begin{bmatrix} 0 \\ 3 \end{bmatrix} \quad = 3e_2$$

In other words, what happens to either is a (different) scalar multiplication:

$$A(e_1) = 2e_1 \text{ and } A(e_2) = 3e_2$$

Furthermore, the entirety of each of the axes is stretched this way. So, we can say that A stretches the plane horizontally by a factor of 2 and vertically by 3, in either order:



Even though we speak of stretching the plane, this is not to say that all *vectors* are stretched. Indeed, other vectors may be rotated; for example, the values of $< 1, 1 >$ isn't its multiple:

$$A \begin{bmatrix} 1 \\ 1 \end{bmatrix} = \begin{bmatrix} 2 \\ 3 \end{bmatrix} \neq \lambda \begin{bmatrix} 1 \\ 1 \end{bmatrix},$$

for any real λ.

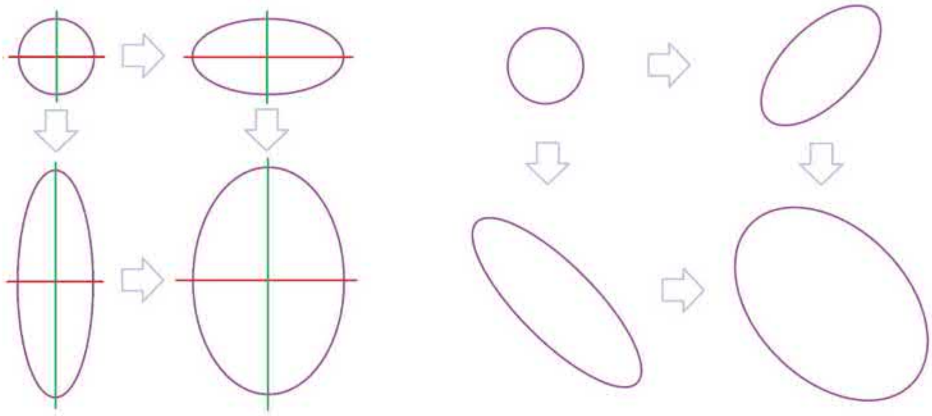
Exercise 5.6.2

Analyze a linear operator with a diagonal matrix:

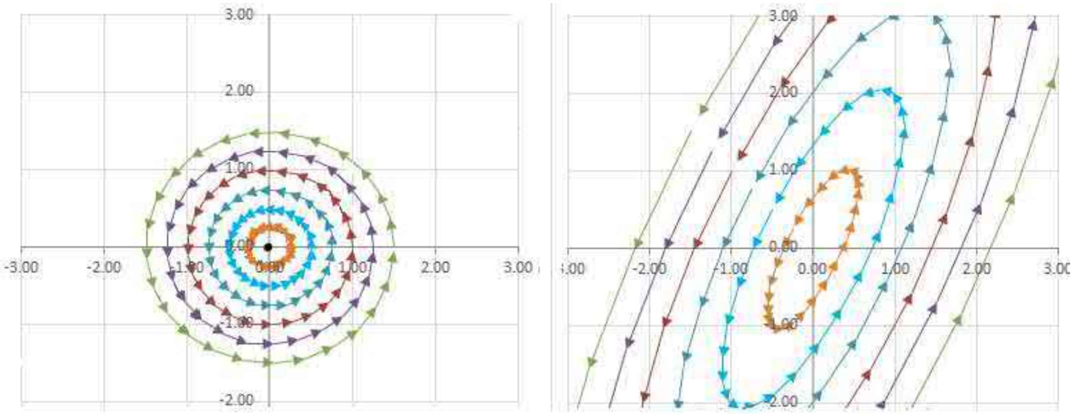
$$A = \begin{bmatrix} h & 0 \\ 0 & v \end{bmatrix}.$$

Example 5.6.3: stretch along other axes

What if the operator stretches along *other* lines? Here we simply rotate the picture in the last example through 45 degrees to make this point:



The circle is stretched, but in what direction or directions? Is there a rotation too? It is hard to tell without prior knowledge. We also have seen this:

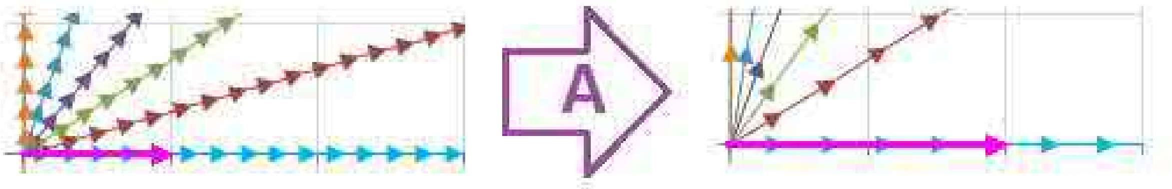


The plane is visibly stretched, but in what direction or directions? It is hard to tell because the result simply looks skewed.

It might be typical then that a linear operator A rotates some vectors, but A also stretches other vectors. On such a vector V, A acts as a scalar multiplication:

$$A(V) = \lambda V\,,$$

for some number λ. For example, we see *disproportional* horizontal and vertical stretching:



This idea bring us to the following important concept:

Definition 5.6.4: eigenvalue

Given a linear operator $A : \mathbf{R}^2 \to \mathbf{R}^2$, a (real) number λ is called an *eigenvalue* of A if it satisfies:

$$\boxed{A(V) = \lambda V}$$

for some non-zero vector V in $\mathbf{R}^2$. Then, V is called an *eigenvector* of A corresponding to λ.

Warning!

Vector $V = 0$ is excluded because we always have $A(0) = 0$.

Note that "eigen" means "characteristic" in German.

Now, how do we find these?

Example 5.6.5: identity operator

If this is the identity matrix, $A = I$, the equation is easy to solve:

$$\lambda V = AV = IV = V.$$

So, $\lambda = 1$. This is the only eigenvalue. What are its eigenvectors? All vectors but 0. Indeed, no vector is rotated!

Exercise 5.6.6

What about a stretch by a factor of k?

Example 5.6.7: diagonal matrix

Let's revisit this diagonal matrix:

$$A = \begin{bmatrix} 2 & 0 \\ 0 & 3 \end{bmatrix} .$$

Then our vector equation $AV = \lambda V$ becomes:

$$\begin{bmatrix} 2 & 0 \\ 0 & 3 \end{bmatrix} \begin{bmatrix} x \\ y \end{bmatrix} = \lambda \begin{bmatrix} x \\ y \end{bmatrix} .$$

Let's rewrite:

$$\begin{bmatrix} 2x \\ 3y \end{bmatrix} = \begin{bmatrix} \lambda x \\ \lambda y \end{bmatrix} \implies \begin{cases} 2x & = \lambda x \ \texttt{AND} \\ 3y & = \lambda y \end{cases} \implies \begin{cases} x(2-\lambda) & = 0, & (1) \\ y(3-\lambda) & = 0. & (2) \end{cases}$$

The two equations must be satisfied simultaneously.

We will use the *Zero Factor Rule* again. Now, we have $V =< x, y >\neq 0$, so either $x \neq 0$ or $y \neq 0$. Let's use the above equations to consider these two cases:

- Case 1: $x \neq 0$, then from (1), we have: $2 - \lambda = 0 \implies \lambda = 2$.
- Case 2: $y \neq 0$, then from (2), we have: $3 - \lambda = 0 \implies \lambda = 3$.

These are the only two possibilities. We have found the eigenvalues!

The second part is to find the eigenvectors. If $\lambda = 2$, then $y = 0$ from (2). Therefore, the corresponding eigenvectors are:

$$\begin{bmatrix} x \\ 0 \end{bmatrix}, \ x \neq 0 , \quad A \begin{bmatrix} x \\ 0 \end{bmatrix} = 2 \begin{bmatrix} x \\ 0 \end{bmatrix} .$$

If $\lambda = 3$, then $x = 0$ from (1). Therefore, the corresponding eigenvectors are:

$$\begin{bmatrix} 0 \\ y \end{bmatrix}, \ y \neq 0 , \quad A \begin{bmatrix} 0 \\ y \end{bmatrix} = 3 \begin{bmatrix} 0 \\ y \end{bmatrix} .$$

These two sets are *almost* equal to the two axes! If we append 0 to these sets of eigenvectors, we have the following. For $\lambda = 2$, the set is the x-axis:

$$\left\{ \begin{bmatrix} x \\ 0 \end{bmatrix} : \ x \text{ real} \right\}$$

And for $\lambda = 3$, the set is the y-axis:

$$\left\{ \begin{bmatrix} 0 \\ y \end{bmatrix} : \ y \text{ real} \right\}$$

The solution is time-consuming, but there will be a short-cut later. We have confirmed the fact that, because of the non-uniform re-scaling, all vectors are rotated except for the vertical and horizontal ones.

Exercise 5.6.8

Analyze a linear operator with a diagonal matrix:

$$A = \begin{bmatrix} h & 0 \\ 0 & v \end{bmatrix} .$$

From the example, we can guess a pattern.

Theorem 5.6.9: Multiples of Eigenvectors

Any non-zero multiple of an eigenvector is also an eigenvector – with respect to the same eigenvalue.

Proof.

Suppose V is an eigenvector of a linear operator A corresponding to the eigenvalue λ:

$$AV = \lambda V .$$

If $W = kV$, then

$$\begin{aligned} AW &= A(kV) && \text{Substitute.} \\ &= kAV && \text{Use the fact that it preserves scalar multiplication.} \\ &= k\lambda V && \text{Use the fact that this is an eigenvector of } \lambda. \\ &= \lambda(kV) && \text{Rearrange.} \\ &= \lambda W && \text{Substitute back.} \end{aligned}$$

The whole line is made up of eigenvectors. It's a copy of **R**! More general is the following:

Definition 5.6.10: eigenspace

For an eigenvalue λ of a linear operator A, the *eigenspace* of A corresponding to λ is defined and denoted by the following:

$$E(\lambda) = \{V : \ A(V) = \lambda V\} .$$

It's all eigenvectors of λ plus 0. We include it in order to make this set into a *space*, a vector space.

From the examples above, we derive the following.

Example 5.6.11: identity matrix

For the identity matrix, we have:

$$E(1) = \mathbf{R}^2 .$$

Example 5.6.12: diagonal matrix

For

$$A = \begin{bmatrix} 2 & 0 \\ 0 & 3 \end{bmatrix} ,$$

we have:

- $E(2)$ is the x-axis.
- $E(3)$ is the y-axis.

Two copies of $\mathbf{R}$!

Example 5.6.13: rotation

A rotation doesn't stretch any vectors. Therefore, there are no (real) eigenvalues. Therefore, there are no eigenvectors and no eigenspaces.

Example 5.6.14: zero matrix

For the zero matrix, $A = 0$, we have:

$$AV = \lambda V, \text{ or } 0 = \lambda V\,.$$

Therefore, $\lambda = 0$ since $V \neq 0$. Furthermore:

$$E(0) = \mathbf{R}^2\,.$$

Example 5.6.15: projection

Consider now the projection on the x-axis,

$$P = \begin{bmatrix} 1 & 0 \\ 0 & 0 \end{bmatrix}.$$

Then our matrix equation is solved as follows:

$$\begin{bmatrix} 1 & 0 \\ 0 & 0 \end{bmatrix} \begin{bmatrix} x \\ y \end{bmatrix} = \lambda \begin{bmatrix} x \\ y \end{bmatrix} \Longrightarrow \begin{bmatrix} x \\ 0 \end{bmatrix} = \begin{bmatrix} \lambda x \\ \lambda y \end{bmatrix} \Longrightarrow \begin{cases} x = \lambda x & \texttt{AND} \\ 0 = \lambda y \end{cases}$$

So, the only possible cases are:

$$\lambda = 0 \quad \text{and} \quad \lambda = 1\,.$$

It appears that the operator is projecting in one direction and doing nothing in another.

Next, in order to find the corresponding eigenvectors, we now go back to the system of linear equations for x and y. We consider these two cases. First:

$$\text{Case 1: } \lambda = 0 \Longrightarrow \begin{cases} x = 0 \cdot x & \texttt{AND} \\ 0 = 0 \cdot y \end{cases} \Longrightarrow \begin{cases} x = 0 & \texttt{AND} \\ y & \text{any} \end{cases} \Longrightarrow E(0) = \left\{ \begin{bmatrix} 0 \\ y \end{bmatrix} :\ y \text{ real} \right\}$$

This is the y-axis. Second:

$$\text{Case 2: } \lambda = 1 \Longrightarrow \begin{cases} x = 1 \cdot x & \texttt{AND} \\ 0 = 1 \cdot y \end{cases} \Longrightarrow \begin{cases} x & \text{any} \ \texttt{AND} \\ y = 0 \end{cases} \Longrightarrow E(1) = \left\{ \begin{bmatrix} x \\ 0 \end{bmatrix} :\ x \text{ real} \right\}$$

This is the x-axis.

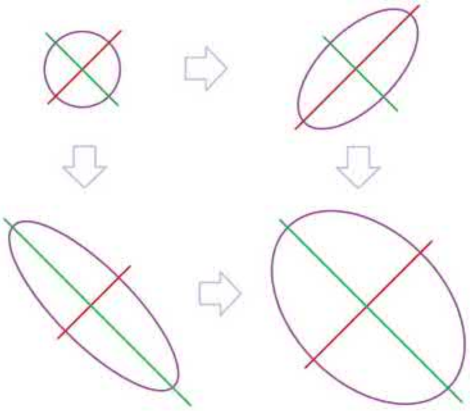
Typically, we have two eigenvectors that aren't multiples of each other:

- $A(V_1) = \lambda_1 V_1$ and
- $A(V_2) = \lambda_2 V_2$,

for some numbers $\lambda_1 \neq \lambda_2$.

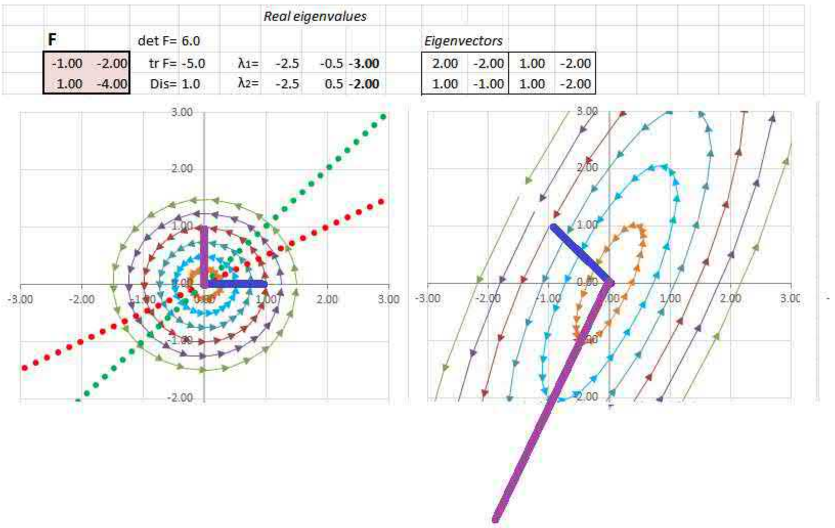
Example 5.6.16: stretch along other axes

Let's revisit the stretch along special lines. It might look like this:



However, how would we even find these special directions?

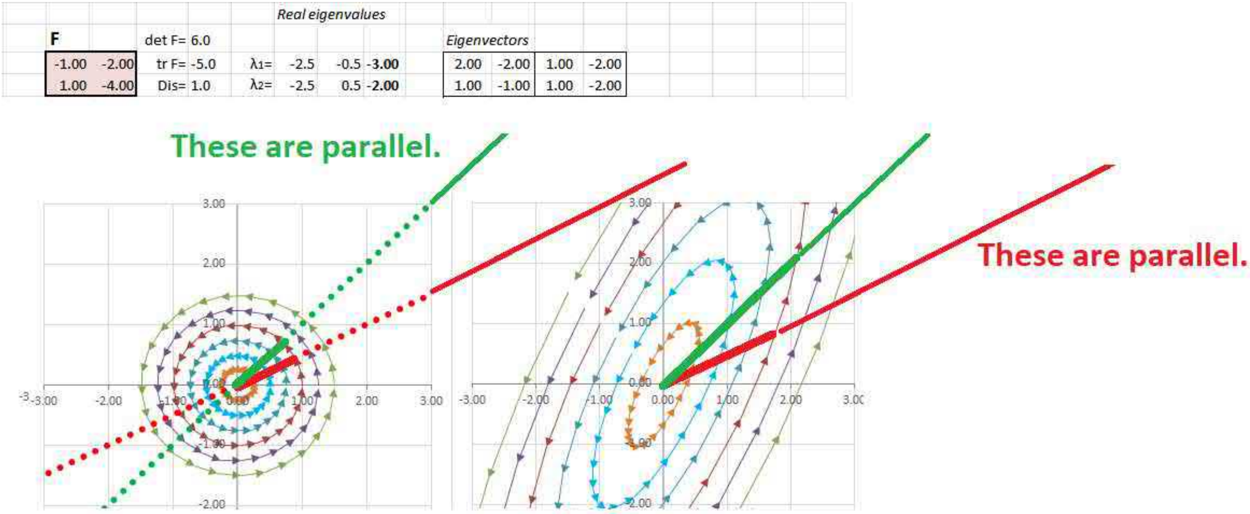
Below, we try the basis vectors, look at where they go (from the matrix itself), and see that they have rotated:



Real eigenvalues

F		det F= 6.0					*Eigenvectors*			
-1.00	-2.00	tr F= -5.0	λ_1=	-2.5	-0.5	**-3.00**	2.00	-2.00	1.00	-2.00
1.00	-4.00	Dis= 1.0	λ_2=	-2.5	0.5	**-2.00**	1.00	-1.00	1.00	-2.00

They are not eigenvectors! The eigenvectors below may be found by trial and error or by the method presented below; they, indeed, don't rotate:



Real eigenvalues

F		det F= 6.0					*Eigenvectors*			
-1.00	-2.00	tr F= -5.0	λ_1=	-2.5	-0.5	**-3.00**	2.00	-2.00	1.00	-2.00
1.00	-4.00	Dis= 1.0	λ_2=	-2.5	0.5	**-2.00**	1.00	-1.00	1.00	-2.00

Example 5.6.17: no eigenvectors

Can we derive what a linear operator does from its matrix only, without visualization? Suppose A is given:

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

Then, to find the eigenvalues, we consider this system of linear equations:

$$AV = \lambda V \implies \begin{bmatrix} 0 & -1 \\ 1 & 0 \end{bmatrix} \begin{bmatrix} x \\ y \end{bmatrix} = \lambda \begin{bmatrix} x \\ y \end{bmatrix} .$$

We solve it as follows:

$$\implies \begin{cases} -y & = \lambda x \ \texttt{AND} \\ x & = \lambda y \end{cases} \implies \begin{cases} -xy & = \lambda x^2 \ \texttt{AND} \\ xy & = \lambda y^2 \end{cases} \implies \lambda x^2 = -\lambda y^2 \implies x^2 = -y^2 \ \texttt{OR} \ \lambda = 0 \, .$$

A direct examination reveals that $\lambda = 0$ is *not* an eigenvalue:

$$AV = 0 \cdot V \implies \begin{bmatrix} 0 & -1 \\ 1 & 0 \end{bmatrix} \begin{bmatrix} x \\ y \end{bmatrix} = 0 \implies x = 0, \ y = 0 \, .$$

In the meantime, the equation $x^2 = -y^2$ is impossible unless both x and y are zeros, which is not allowed.

There seems to be no eigenvalues, certainly not *real* ones... This means that every vector is rotated. Maybe this *is* a rotation? Yes, we recognize the matrix of the 90-degree rotation.

Exercise 5.6.18

How does the determinant of A tell you whether 0 is an eigenvalue?

Exercise 5.6.19

Show that a zero eigenvalue implies a collapse.

Example 5.6.20: need for homogeneous system

Let's revisit this linear operator:

$$A = \begin{bmatrix} -1 & -2 \\ 1 & -4 \end{bmatrix} .$$

Our vector equation becomes:

$$\begin{bmatrix} -1 & -2 \\ 1 & -4 \end{bmatrix} \begin{bmatrix} x \\ y \end{bmatrix} = \lambda \begin{bmatrix} x \\ y \end{bmatrix} .$$

We rewrite, again, as a system of linear equations:

$$\begin{bmatrix} -x - 2y \\ x - 4y \end{bmatrix} = \begin{bmatrix} \lambda x \\ \lambda y \end{bmatrix} \implies \begin{cases} (-1-\lambda)x & - & 2y & = 0 \ \texttt{AND} \\ x & + & (-4-\lambda)y & = 0 \end{cases}$$

This is another system of linear equations to be solved, again. It is more complex than the ones we saw above and none of the shortcuts are available... The system corresponds to a *homogeneous* vector equation:

$$\begin{bmatrix} -1-\lambda & -2 \\ 1 & -4-\lambda \end{bmatrix} \begin{bmatrix} x \\ y \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \end{bmatrix} .$$

What do we know about those?

Let's review.

Suppose we have a linear operator A and we need to find its eigenvalues and eigenvectors. Let's, for now, concentrate on the former. Suppose λ is an eigenvalue of A. This means that λ is a real number and there is some non-zero vector V that satisfies:

$$\boxed{AV = \lambda V}$$

Let's do some vector algebra:

$$AV = \lambda V \implies AV - \lambda V = 0 .$$

We want to turn this equation of vectors and matrices into one entirely of matrices. We can take this equation one step further by observing that

$$\lambda V = \lambda I V ,$$

where I is the identity matrix. The linearity of these operators allows to factor V out of our equation. It takes a new form:

$$\boxed{(A - \lambda I)V = 0}$$

The equation characterizes an eigenvector and its eigenvalues in a space of any dimension.

Example 5.6.21: dimension 2

In the $\mathbf{R}^2$ case, we make this specific when our linear operator A is specific. Suppose it is given by a matrix:

$$A = \begin{bmatrix} a & b \\ c & d \end{bmatrix} .$$

We carry out these computations:

$$AV = \begin{bmatrix} a & b \\ c & d \end{bmatrix} \begin{bmatrix} x \\ y \end{bmatrix} = \begin{bmatrix} ax+by \\ cx+dy \end{bmatrix} \quad \text{and} \quad \lambda \begin{bmatrix} x \\ y \end{bmatrix} = \begin{bmatrix} \lambda x \\ \lambda y \end{bmatrix} .$$

These two vectors are supposed to be equal, so we have a system of linear equations, which is then transformed into a *homogeneous* form:

$$\begin{cases} ax & + & by & = \lambda x \quad \texttt{AND} \\ cx & + & dy & = \lambda y. \end{cases} \iff \begin{cases} (a-\lambda)x & + & by & & = 0 \quad \texttt{AND} \\ cx & & + & (d-\lambda)y & = 0. \end{cases}$$

The matrix of this system is:

$$G = \begin{bmatrix} a-\lambda & b \\ c & d-\lambda \end{bmatrix} .$$

Following these computations, we recognize some matrix algebra:

$$G = \begin{bmatrix} a & b \\ c & d \end{bmatrix} - \lambda \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} .$$

We go back to our compact representation:

Theorem 5.6.22: Eigenvalues and Eigenvectors

Suppose A is a linear operator. Then every pair of an eigenvalue λ and its eigenvector V of A satisfy the following matrix equation:

$$\boxed{GV = 0, \quad \text{where } G = A - \lambda I}$$

Now, the question about the eigenvalues of the matrix A becomes one about the matrix G:

- Under what circumstances does the system $GV = 0$ have a non-zero solution?

We know the answer from the last section:

- The system $GV = 0$ has a non-zero solution if and only if $\det G = 0$.

We have proven the following result:

Theorem 5.6.23: Eigenvalues as Roots

Suppose A is a linear operator $\mathbf{R}^2$. Then every eigenvalue λ of A is a solution to the following equation:

$$\boxed{\det(A - \lambda I) = 0}$$

In contrast to the *matrix* equation in the last theorem, this simple *algebraic* equation allows to discover eigenvalues and then, possibly, their eigenvectors.

We codify this idea below:

Definition 5.6.24: characteristic polynomial

The *characteristic polynomial* of a 2×2 matrix A is defined to be:

$$\boxed{\chi_A(\lambda) = \det(A - \lambda I)}$$

Meanwhile, the equation $\chi_A(\lambda) = 0$ is called the *characteristic equation*.

This is the convenient form of the equation we are to solve for dimension $n = 2$:

$$\chi_A(\lambda) = \det \begin{bmatrix} a - \lambda & b \\ c & d - \lambda \end{bmatrix} = (a - \lambda)(d - \lambda) - bc = 0\,.$$

It's *quadratic*!

We don't know what a linear operator does but – even without the eigenvectors – we can tell a lot from its eigenvalues. We rediscover some of the information about familiar operators below.

Example 5.6.25: re-scaling?

Consider again:

$$A = \begin{bmatrix} 2 & 0 \\ 0 & 3 \end{bmatrix}\,.$$

Then we solve:

$$\chi_A(\lambda) = \det \begin{bmatrix} 2-\lambda & 0 \\ 0 & 3-\lambda \end{bmatrix} = (2-\lambda)(3-\lambda) = 0\,.$$

Therefore, we have:

$$\lambda = 2,\ 3\,.$$

We conclude that the linear operator *stretches the plane by these factors in two different directions.* What are those directions? We can't tell without finding the eigenvectors.

Example 5.6.26: projection?

If

$$A = \begin{bmatrix} 1 & 0 \\ 0 & 0 \end{bmatrix},$$

the characteristic equation is:

$$\chi_A(\lambda) = \det \begin{bmatrix} 1-\lambda & 0 \\ 0 & -\lambda \end{bmatrix} = (1-\lambda)(-\lambda) = 0\,.$$

Then,

$$\lambda = 1,\ 0\,.$$

We conclude that the linear operator *does nothing in one direction and collapses in another.* That's a projection! What are those directions? We don't know without the eigenvectors.

Example 5.6.27: rotation?

Consider

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

Then,

$$\chi_A(\lambda) = \det \begin{bmatrix} -\lambda & -1 \\ 1 & -\lambda \end{bmatrix} = \lambda^2 + 1 = 0\,.$$

No real solutions! So, no non-zero vector is taken by A to its own multiple. Maybe this is a rotation...

These three examples suggest a classification of linear operators of the plane. But first a quick review of quadratic polynomials.

Consider one:

$$f(x) = x^2 + px + q\,.$$

The *Quadratic Formula* then provides the x-intercepts of this function:

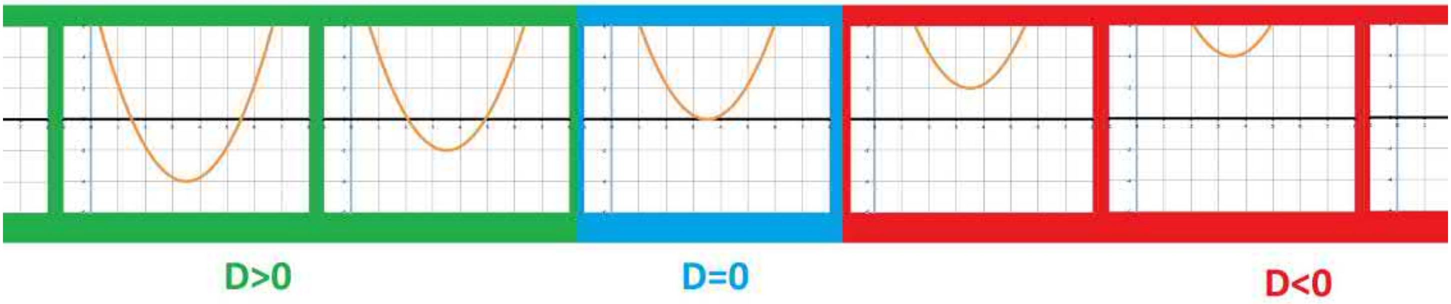
$$x = -\frac{p}{2} \pm \frac{\sqrt{p^2 - 4q}}{2}\,.$$

Of course, the x-intercepts are the real solutions of this equation and that is why the result only makes sense when the *discriminant* of the quadratic polynomial,

$$D = p^2 - 4q\,,$$

is non-negative.

Increasing the value of the free term q makes the graph of $y = f(x)$ shift upward and, eventually, pass the x-axis entirely. We can observe how its two x-intercepts start to get closer to each other, then merge, and finally disappear:



This process is explained by what is happening, with the growth of q, to the roots given by the *Quadratic Formula*:

$$x_{1,2} = -\frac{p}{2} \pm \frac{\sqrt{D}}{2}\,.$$

There are three states:

1. Starting with a positive value, D decreases, and $\dfrac{\sqrt{D}}{2}$ decreases.
2. Then D becomes 0 and, therefore, we have $\dfrac{\sqrt{D}}{2} = 0$.
3. Then D becomes negative, and there are no real roots (complex roots are discussed in the next chapter).

So, we have:

- The eigenvalues are the real roots of the (quadratic) characteristic polynomial χ_A.
- Therefore, the number of eigenvalues is less than or equal to 2, counting their multiplicities.

Let's try to expand the characterstic polynomial and see if patterns emerge:

$$\begin{aligned}
\chi(\lambda) &= \det \begin{bmatrix} a-\lambda & b \\ c & d-\lambda \end{bmatrix} \\
&= (a-\lambda)(d-\lambda) - bc \\
&= ad - a\lambda - \lambda d + \lambda^2 - bc \\
&= \lambda^2 - (a+d)\lambda + (ad-bc) \\
&= \lambda^2 - \operatorname{tr} A\,\lambda + \det A\,.
\end{aligned}$$

The term in the middle is defined as follows:

Definition 5.6.28: trace of matrix

The *trace* of a matrix A is the sum of its diagonal elements. It is denoted by:

$$\boxed{\operatorname{tr} A = \operatorname{tr} \begin{bmatrix} a & b \\ c & d \end{bmatrix} = a + d}$$

So, the trace appears – along with the determinant – in the characteristic polynomial:

Theorem 5.6.29: Characteristic Polynomial

The characteristic polynomial of matrix A takes this form:

$$\boxed{\chi_A(\lambda) = \lambda^2 - \operatorname{tr} A \cdot \lambda + \det A}$$

It is known that not only the determinant but also the trace are independent of our choice of a Cartesian system. Therefore, so is the characteristic polynomial.

The discriminant of the characteristic polynomial can now be used to tell what the linear operators does:

$$D = (\operatorname{tr} A)^2 - 4 \det A \,.$$

Theorem 5.6.30: Classification of Linear Operators

Suppose A is a linear operator given by a 2×2 matrix and D is the discriminant of its characteristic polynomial. Then we have three cases:

1. *$D > 0$. The eigenvalues are distinct: Operator A non-uniformly re-scales the plane in the distinct directions of the corresponding eigenvectors.*
2. *$D = 0$. The eigenvalues are equal: Operator A uniformly re-scales the plane in all directions unless the eigenvectors are all multiples of each other.*
3. *$D < 0$. There are no eigenvalues: Operator A rotates the plane (with a possible re-scaling).*

Proof.

For Part 1, we already know that the directions are distinct because if two eigenvectors are multiples of each other, then they have the same eigenvalue. Indeed:

$$A(V) = \lambda V \implies A(kV) = kA(V) = k\lambda V = \lambda(kV)\,.$$

Parts 2 and 3 are addressed later in the chapter.

5.7. The significance of eigenvectors

We have shown how one can *visualize* the way a linear operator transforms the plane: by examining what happens to various curves in the domain. By mapping these curves, one can discover stretching, shrinking in various directions, rotations, etc.

In the last section, we also saw how one can *understand* the way a linear operator transforms the plane: by examining its eigenvalues. The method is entirely algebraic rather than experimental. We simply find the directions of pure stretch for F:

$$\boxed{FV = \lambda V}$$

The visualizations are produced by a spreadsheet. The spreadsheet also computes the eigenvector and its eigenvalues; they are shown above the graphs.The spreadsheet also shows eigenspaces as two (or one, or none) straight lines; they remain in place under the transformation.

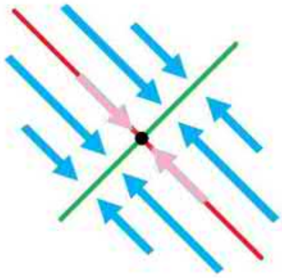
We would like to learn how to predict the outcome by examining only its matrix.

Below is a familiar fact that will take us down that road:

Theorem 5.7.1: Preimages of Zero

If the image of $V \neq 0$ *under a linear operator* F *is zero, then so is that of any of its multiples* kV.

In other words, the whole line with V as its direction vector is collapsed to 0 by F:



Example 5.7.2: collapse on axis

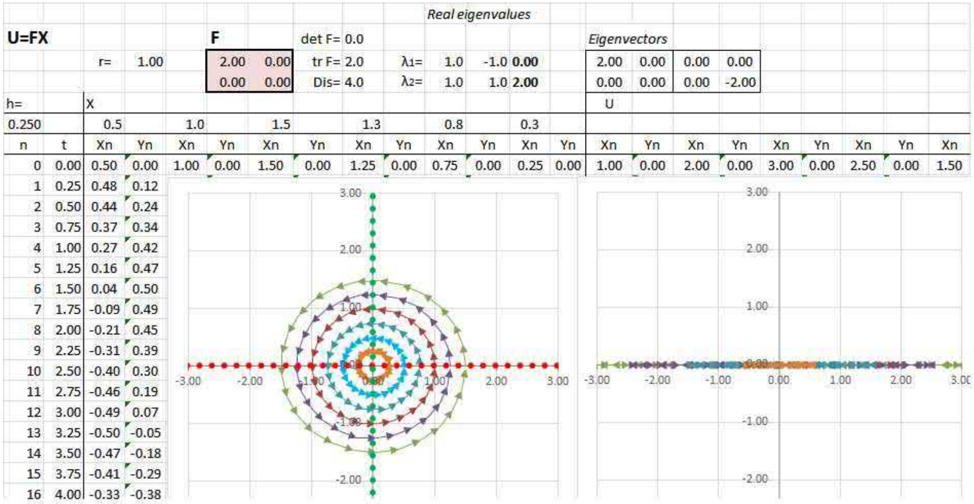
We start with a familiar example:

$$\begin{cases} u &= 2x \\ v &= 0 \end{cases} \quad \text{is re-written as} \quad \begin{bmatrix} u \\ v \end{bmatrix} = \begin{bmatrix} 2 & 0 \\ 0 & 0 \end{bmatrix} \begin{bmatrix} x \\ y \end{bmatrix}.$$

Even without the characteristic equation, we can guess the eigenvalue-eigenvector pairs:

$$\begin{bmatrix} 2 & 0 \\ 0 & 0 \end{bmatrix} \begin{bmatrix} 1 \\ 0 \end{bmatrix} = 2 \begin{bmatrix} 1 \\ 0 \end{bmatrix}, \quad \begin{bmatrix} 2 & 0 \\ 0 & 0 \end{bmatrix} \begin{bmatrix} 0 \\ 1 \end{bmatrix} = 0 \begin{bmatrix} 0 \\ 1 \end{bmatrix}.$$

Below, one can see how this operator projects the whole plane to the x-axis:



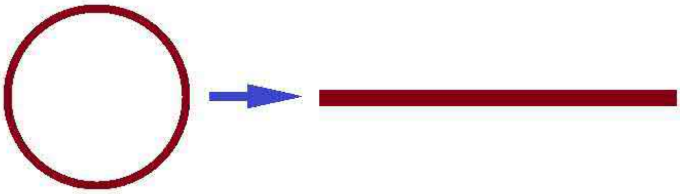
Real eigenvalues

U=FX **F** det F= 0.0 *Eigenvectors*

r= 1.00 2.00 0.00 tr F= 2.0 λ1= 1.0 -1.0 **0.00** 2.00 0.00 0.00 0.00

0.00 0.00 Dis= 4.0 λ2= 1.0 1.0 **2.00** 0.00 0.00 0.00 -2.00

h=		X												U								
0.250		0.5		1.0		1.5		1.3		0.8		0.3										
n	t	Xn	Yn	Xn	Yn	Xn	Yn	Xn	Yn	Xn	Yn	Xn	Yn	Xn	Yn	Xn	Yn	Xn	Yn	Xn	Yn	Xn
0	0.00	0.50	0.00	1.00	0.00	1.50	0.00	1.25	0.00	0.75	0.00	0.25	0.00	1.00	0.00	2.00	0.00	3.00	0.00	2.50	0.00	1.50
1	0.25	0.48	0.12																			
2	0.50	0.44	0.24																			
3	0.75	0.37	0.34																			
4	1.00	0.27	0.42																			
5	1.25	0.16	0.47																			
6	1.50	0.04	0.50																			
7	1.75	-0.09	0.49																			
8	2.00	-0.21	0.45																			
9	2.25	-0.31	0.39																			
10	2.50	-0.40	0.30																			
11	2.75	-0.46	0.19																			
12	3.00	-0.49	0.07																			
13	3.25	-0.50	-0.05																			
14	3.50	-0.47	-0.18																			
15	3.75	-0.41	-0.29																			
16	4.00	-0.33	-0.38																			



The operator collapses the y-axis to 0, while the x-axis is stretched by a factor of 2. The standard basis vectors happen to be eigenvectors! That's the reason why the matrix is so simple.

Example 5.7.3: stretch-shrink along axes

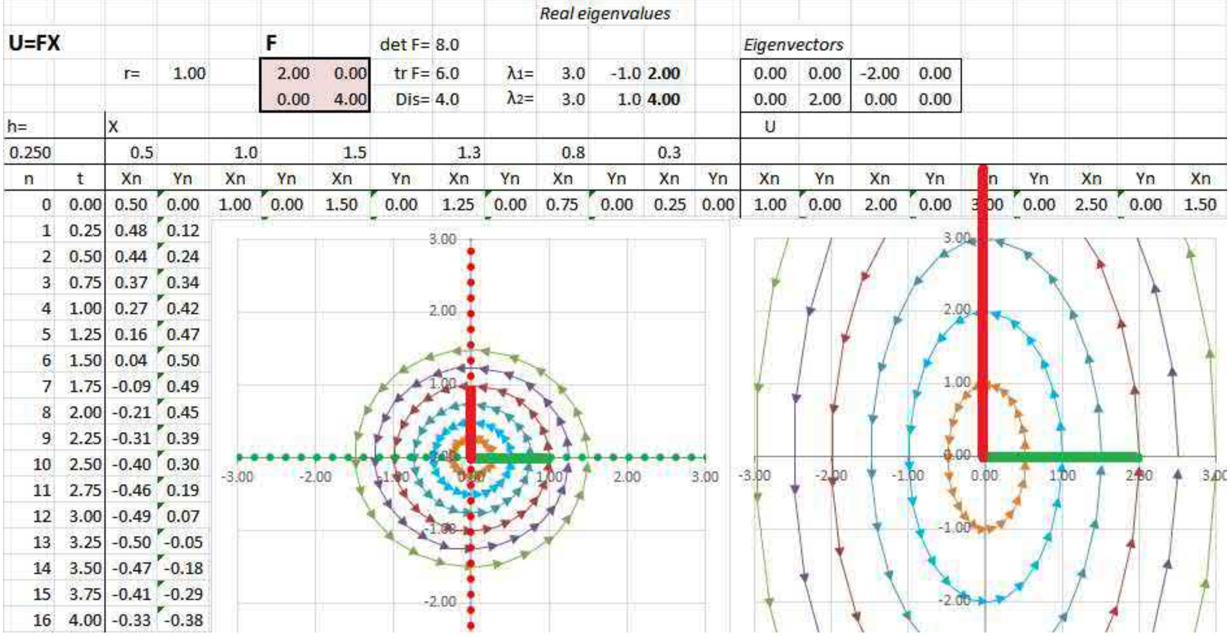
Let's consider this linear operator and its matrix

$$\begin{cases} u &= 2x \\ v &= 4y \end{cases} \quad \text{and} \quad F = \begin{bmatrix} 2 & 0 \\ 0 & 4 \end{bmatrix}.$$

Once again, we don't need the characteristic equation to suggest the eigenvalues (and then find the eigenvectors):

$$\begin{bmatrix} 2 & 0 \\ 0 & 0 \end{bmatrix} \begin{bmatrix} 1 \\ 0 \end{bmatrix} = 2 \begin{bmatrix} 1 \\ 0 \end{bmatrix}, \quad \begin{bmatrix} 2 & 0 \\ 0 & 0 \end{bmatrix} \begin{bmatrix} 0 \\ 1 \end{bmatrix} = 4 \begin{bmatrix} 0 \\ 1 \end{bmatrix}.$$

As it turns out, we only need to track the values of the basis vectors, and the rest of the values are seen as a *linear combination* of these:



Real eigenvalues

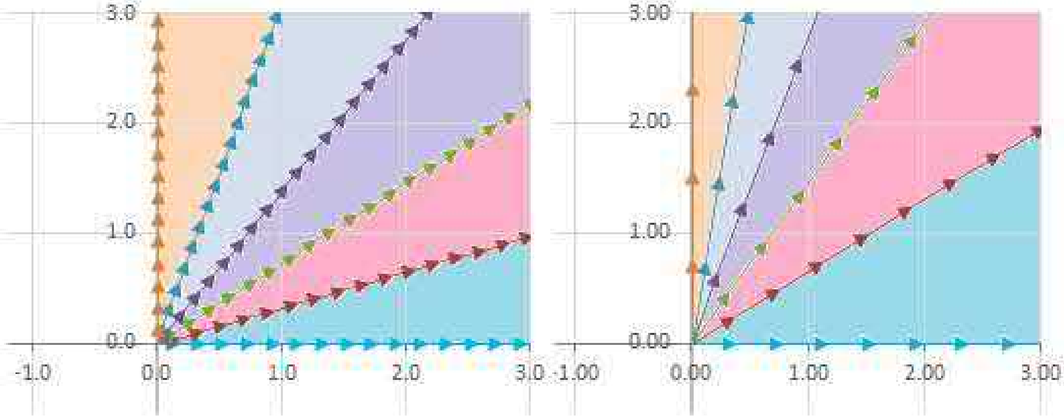
U=FX | F | det F= 8.0 | Eigenvectors

r= 1.00 | 2.00 0.00 | tr F= 6.0 | λ1= 3.0 -1.0 **2.00** | 0.00 0.00 -2.00 0.00

0.00 4.00 | Dis= 4.0 | λ2= 3.0 1.0 **4.00** | 0.00 2.00 0.00 0.00

h=		X												U								
0.250		0.5		1.0		1.5		1.3		0.8		0.3										
n	t	Xn	Yn	Xn	Yn	Xn	Yn	Xn	Yn	Xn	Yn	Xn	Yn	Xn	Yn	Xn	Yn	n	Yn	Xn	Yn	Xn
0	0.00	0.50	0.00	1.00	0.00	1.50	0.00	1.25	0.00	0.75	0.00	0.25	0.00	1.00	0.00	2.00	0.00	3.00	0.00	2.50	0.00	1.50
1	0.25	0.48	0.12																			
2	0.50	0.44	0.24																			
3	0.75	0.37	0.34																			
4	1.00	0.27	0.42																			
5	1.25	0.16	0.47																			
6	1.50	0.04	0.50																			
7	1.75	-0.09	0.49																			
8	2.00	-0.21	0.45																			
9	2.25	-0.31	0.39																			
10	2.50	-0.40	0.30																			
11	2.75	-0.46	0.19																			
12	3.00	-0.49	0.07																			
13	3.25	-0.50	-0.05																			
14	3.50	-0.47	-0.18																			
15	3.75	-0.41	-0.29																			
16	4.00	-0.33	-0.38																			

The rest of the vectors turn non-uniformly; i.e., they "fan out":



This is what happens to an arbitrary vector $X = < x, y >$:

$$FX = \begin{bmatrix} 2 & 0 \\ 0 & 4 \end{bmatrix} \begin{bmatrix} x \\ y \end{bmatrix} = x \begin{bmatrix} 2 \\ 0 \end{bmatrix} + y \begin{bmatrix} 0 \\ 4 \end{bmatrix} = 2x \begin{bmatrix} 1 \\ 0 \end{bmatrix} + 4y \begin{bmatrix} 0 \\ 1 \end{bmatrix} = 2xe_1 + 4ye_2 .$$

The last expression is a linear combination of the values of the two *standard* basis vectors. The middle, however, is also a linear combination but with respect to these two vectors:

$$V_1 = \begin{bmatrix} 2 \\ 0 \end{bmatrix} \quad \text{and} \quad V_2 = \begin{bmatrix} 0 \\ 4 \end{bmatrix}.$$

They can be thought of as forming a "non-standard" basis. Though not unit vectors as the standard ones, they are still aligned with the axes. Now, what is the point? Every vector can be expressed as a linear combination of the two:

$$< x, y >= \frac{x}{2} V_1 + \frac{y}{4} V_2 \,.$$

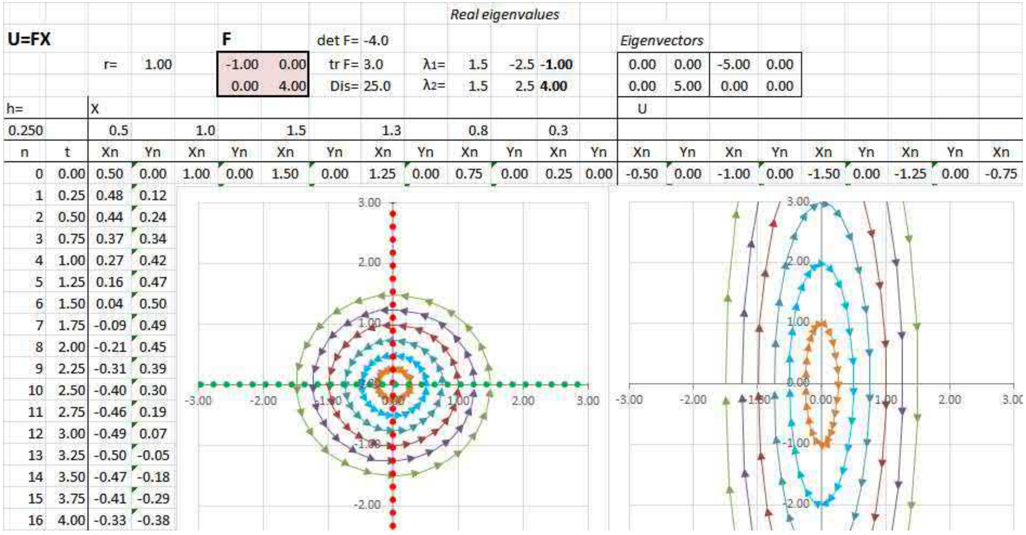
Furthermore, to know where any X goes under F, we need only to know where these two go: It's pure scalar multiplication! We will see that any pair of eigenvectors – when not multiples of each other – would do.

Example 5.7.4: stretch-shrink along axes

A slightly different operator is the following:

$$\begin{cases} u &= -x \\ v &= 4y \end{cases} \quad \text{and} \quad F = \begin{bmatrix} -1 & 0 \\ 0 & 4 \end{bmatrix}.$$

It is still simple because the two variables remain fully separated. As a result, the two transformations of the axes can be thought of as transformations of the whole plane:



Real eigenvalues

U=FX

F det F= -4.0

r= 1.00

-1.00	0.00
0.00	4.00

tr F= 3.0 λ1= 1.5 -2.5 **-1.00**

Dis= 25.0 λ2= 1.5 2.5 **4.00**

Eigenvectors

0.00	0.00	-5.00	0.00
0.00	5.00	0.00	0.00

h=		X												U								
0.250		0.5		1.0		1.5		1.3		0.8		0.3										
n	t	Xn	Yn	Xn	Yn	Xn	Yn	Xn	Yn	Xn	Yn	Xn	Yn	Xn	Yn	Xn	Yn	Xn	Yn	Xn	Yn	Xn
0	0.00	0.50	0.00	1.00	0.00	1.50	0.00	1.25	0.00	0.75	0.00	0.25	0.00	-0.50	0.00	-1.00	0.00	-1.50	0.00	-1.25	0.00	-0.75
1	0.25	0.48	0.12																			
2	0.50	0.44	0.24																			
3	0.75	0.37	0.34																			
4	1.00	0.27	0.42																			
5	1.25	0.16	0.47																			
6	1.50	0.04	0.50																			
7	1.75	-0.09	0.49																			
8	2.00	-0.21	0.45																			
9	2.25	-0.31	0.39																			
10	2.50	-0.40	0.30																			
11	2.75	-0.46	0.19																			
12	3.00	-0.49	0.07																			
13	3.25	-0.50	-0.05																			
14	3.50	-0.47	-0.18																			
15	3.75	-0.41	-0.29																			
16	4.00	-0.33	-0.38																			

The negative sign produces a different pattern: We see the reversal of the direction of the ellipse around the origin. Algebraically, we have as before:

$$FX = \begin{bmatrix} -1 & 0 \\ 0 & 4 \end{bmatrix} \begin{bmatrix} x \\ y \end{bmatrix} = x \begin{bmatrix} -1 \\ 0 \end{bmatrix} + y \begin{bmatrix} 0 \\ 4 \end{bmatrix} = -x \begin{bmatrix} 1 \\ 0 \end{bmatrix} + 4y \begin{bmatrix} 0 \\ 1 \end{bmatrix} = -xe_1 + 4ye_2 \,.$$

Once again, the last expression is a linear combination of the values of the two *standard* basis vectors, while the middle is a linear combination but with respect to *another basis* made up of the eigenvectors:

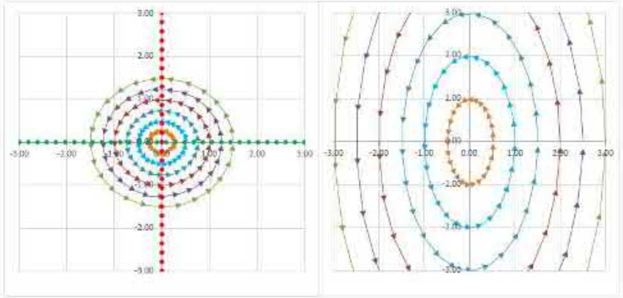
$$V_1 = \begin{bmatrix} -1 \\ 0 \end{bmatrix} \quad \text{and} \quad V_2 = \begin{bmatrix} 0 \\ 4 \end{bmatrix}.$$

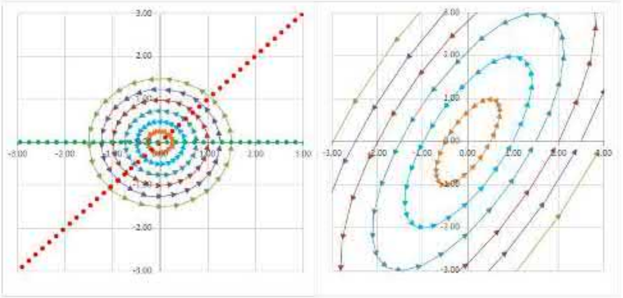
What if the matrix isn't diagonal?

Exercise 5.7.5

Show that the standard basis vectors e_1, e_2 are eigenvectors of diagonal matrices.

Instead of the standard basis vectors, we will concentrate on the eigenvectors of the operator in order to understand what the operator does. To illustrate, just imagine that the picture on the left has been skewed, resulting in the image of the right:



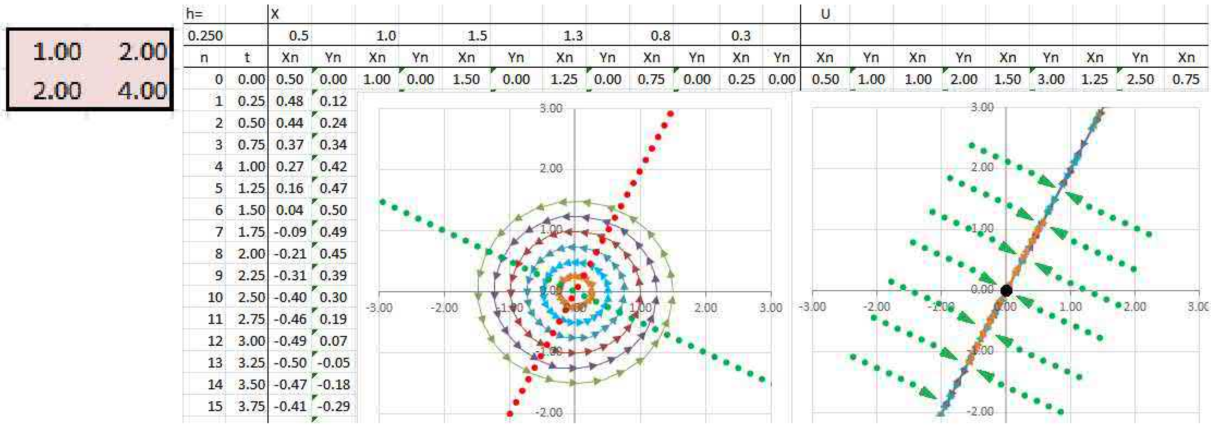


The eigenvectors of the matrix will serve as an alternative basis.

Example 5.7.6: collapse

Let's consider a more general linear operator:

$$\begin{cases} u &= x &+ &2y \\ v &= 2x &+ &4y \end{cases} \quad \Longrightarrow \quad F = \left[\begin{array}{cc} 1 & 2 \\ 2 & 4 \end{array}\right].$$



1.00	2.00
2.00	4.00

h=		X												U								
0.250		0.5		1.0		1.5		1.3		0.8		0.3										
n	t	Xn	Yn	Xn	Yn	Xn	Yn	Xn	Yn	Xn	Yn	Xn	Yn	Xn	Yn	Xn	Yn	Xn	Yn	Xn	Yn	Xn
0	0.00	0.50	0.00	1.00	0.00	1.50	0.00	1.25	0.00	0.75	0.00	0.25	0.00	0.50	1.00	1.00	2.00	1.50	3.00	1.25	2.50	0.75
1	0.25	0.48	0.12																			
2	0.50	0.44	0.24																			
3	0.75	0.37	0.34																			
4	1.00	0.27	0.42																			
5	1.25	0.16	0.47																			
6	1.50	0.04	0.50																			
7	1.75	-0.09	0.49																			
8	2.00	-0.21	0.45																			
9	2.25	-0.31	0.39																			
10	2.50	-0.40	0.30																			
11	2.75	-0.46	0.19																			
12	3.00	-0.49	0.07																			
13	3.25	-0.50	-0.05																			
14	3.50	-0.47	-0.18																			
15	3.75	-0.41	-0.29																			

It appears that the function has a stretching in one direction and a collapse in another. What are those directions? Linear algebra gives the answer.

Even without looking for eigenvectors, we know that we can use the fact that *the determinant is zero*:

$$\det F = \det \left[\begin{array}{cc} 1 & 2 \\ 2 & 4 \end{array}\right] = 1 \cdot 4 - 2 \cdot 2 = 0\,.$$

It's not one-to-one and, in fact, there is a whole line of points X with $FX = 0$. To find it, we solve this equation by solving this system of equations:

$$\begin{cases} x &+ &2y &= 0 \\ 2x &+ &4y &= 0 \end{cases} \quad \Longrightarrow \quad x = -2y\,.$$

The two equations are equivalent and represent the same line. We have, indirectly, found the eigenspace and, of course, the eigenvectors of the zero eigenvalue $\lambda_1 = 0$. We can take this eigenvector for further use:

$$V_1 = \left[\begin{array}{c} 2 \\ -1 \end{array}\right] \quad \Longrightarrow \quad FV_1 = 0\,.$$

Let's instead turn to the characteristic polynomial:

$$\det(F - \lambda I) = \det \left[\begin{array}{cc} 1-\lambda & 2 \\ 2 & 4-\lambda \end{array}\right] = \lambda^2 - 5\lambda = \lambda(\lambda - 5) \quad \Longrightarrow \quad \lambda_1 = 0,\ \lambda_2 = 5\,.$$

Let's find the eigenvectors for $\lambda_2 = 5$. We need to solve the vector equation:

$$FV = 5V\,,$$

i.e.,

$$FV = \begin{bmatrix} 1 & 2 \\ 2 & 4 \end{bmatrix} \begin{bmatrix} x \\ y \end{bmatrix} = 5 \begin{bmatrix} x \\ y \end{bmatrix}\,.$$

This gives as the following system of two linear equations (it's the same equation):

$$\begin{cases} x & + & 2y & = 5x \texttt{ AND} \\ 2x & + & 4y & = 5y \end{cases} \implies \begin{cases} -4x & + & 2y & = 0 \texttt{ AND} \\ 2x & - & y & = 0 \end{cases} \implies y = 2x\,.$$

This line is the eigenspace. We choose a vector along this line to as the eigenvector for further use:

$$V_2 = \begin{bmatrix} 1 \\ 2 \end{bmatrix}\,.$$

We summarize what F does:

- A projection along the vector $< 2, -1 >$: The line $x = -2y$ is collapsed to 0.
- A stretch by a factor of 5 along the vector $< 1, 2 >$: The line $y = 2x$ is stretched without any rotation.

We have confirmed the illustration above!

Furthermore, the two eigenvectors aren't multiples of each other. That is why every vector is a linear combination of the two eigenvectors and, therefore, its value under F is a linear combination of the eigenvectors too:

$$X = xV_1 + yV_2 \implies FX = x \cdot 0 \cdot V_1 + y \cdot 5 \cdot V_2\,.$$

We derive where X goes from the above summary!

Exercise 5.7.7

Find the line of the projection.

We are able to summarize what the operator does from the algebra only. The idea is uncomplicated:

> ▶ The linear operator *within the eigenspace* is "1-dimensional"; it can then be represented by a single number.

This number, the stretch-shrink factor, is of course the eigenvalue.

If we know these two numbers, how do we find the rest of the values of the linear operator? In two steps.

First, every X that lies within the eigenspace, which is a line, is a multiple of the eigenvector, and its value under F can be easily computed:

$$X = rV \implies U = F(rV) = rFV = r\lambda V\,.$$

Second, the rest of the values are found by following this idea: Try to express the value as a linear combination of two values found so far.

Let's provide a foundation for this idea.

5.8. Bases

The matrix representation of a linear operator is determined by our choice of the Cartesian system. On the other hand, what it does may be described with such words as “stretch”, “rotation”, “flip”, etc. These descriptions have nothing to do with the coordinate system. And neither do such algebraic characteristics of the operator as the trace, the determinant, the eigenvalues and the eigenvectors. We are on the right track!

Once again, dealing with *vectors* instead of points requires a different approach.

The *standard basis* of $\mathbf{R}^2$, as before, consists of these two:

$$e_1 =< 1,0 >,\ e_2 =< 0,1 > \ .$$

The components of a vector $X =< a,b >$ with respect to this basis are, as before, a,b:

$$< a,b >= a < 1,0 > +b < 0,1 >= ae_1 + be_2 \, .$$

In other words, every vector can be represented as a *linear combination* of these two vectors.

However, they aren’t the only ones with this property! For example, let’s rewrite the above representation:

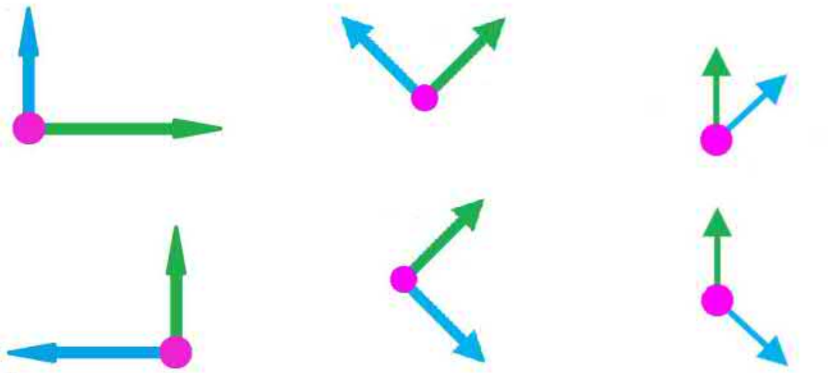
$$< a,b >= a < 1,0 > +b < 0,1 >= ae_1 + (-b)(-e_2) \, .$$

So, this vector is a linear combination of the two vectors $V_1 = e_1$ and $V_2 = -e_2$.

Exercise 5.8.1

Represent vector $< a,b >$ in terms of e_1 and $e_1 + e_2$.

All of these pairs of vectors may serve in such representations:



Exercise 5.8.2

Show that vectors that are multiples of each other can’t be used for this purpose.

A very important concept below captures this idea:

Definition 5.8.3: basis

A *basis* of $\mathbf{R}^2$ is any such pair of vectors V_1, V_2 that every vector can be represented as a linear combination of these vectors:

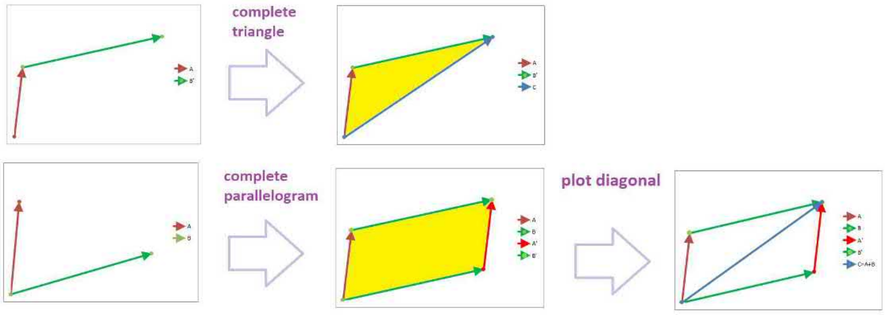
$$X = r_1V_1 + r_2V_2 \, .$$

Then the coefficients r_1, r_2 are called the *components* of X with respect to the basis.

Exercise 5.8.4

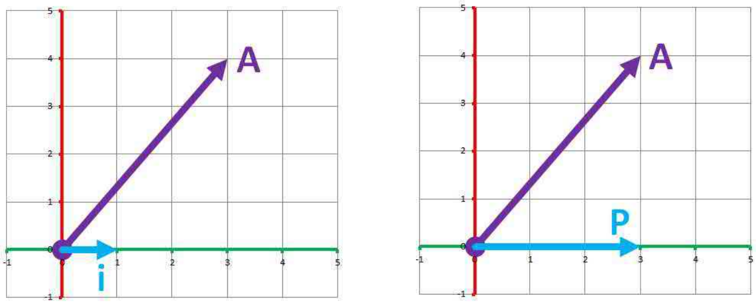
Prove that there is only one such representation.

The reasoning is that the algebra of vectors was established *before* the Cartesian system was added to the vector space and before the two operations were expressed in terms of the components of vectors. For example, this is how we add vectors:

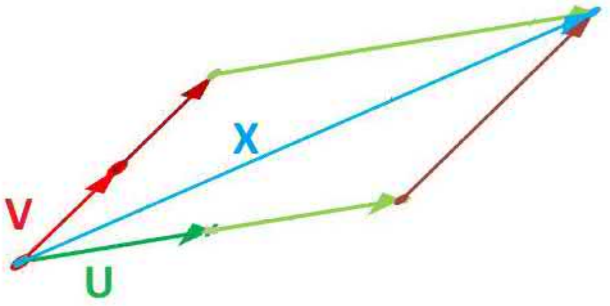


Example 5.8.5

The components are found for the standard basis are found via the orthogonal projections:



A very different choice of basis U, V is shown below:



Here, vector X has components 2 and 2 with respect to this basis:

$$X = 2U + 2V\,.$$

Example 5.8.6: component algebra

Linear combinations are behind the components algebra. Collecting common terms after addition is what happens to the components:

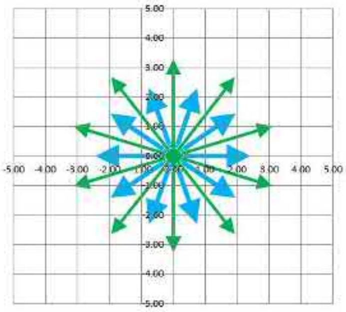
	components	linear combination
U	$=< 2, 3 >$	$= 2e_1 + 3e_2$
W	$=< -1, 2 >$	$= -1e_1 + 2e_2$
$U + V$	$=< 2, 3 > + < -1, 2 >$	$= (2e_1 + 3e_2) + (-1e_1 + 2e_2)$
	$=< 2 - 1, 3 + 2 >$	$= (2e_1 - 1e_1) + (3e_2 + 2e_2)$
	$=< 1, 5 >$	$= 1e_1 + 5e_2$

Replacing e_1, e_2 with, say, V_1, V_2 won't change anything in this computation. Same for scalar multiplication:

	components	linear combination
U	$=< 2, 3 >$	$= 2e_1 + 3e_2$
r	$= -3$	
rU	$= (-3) < 2, 3 >$	$= (-3)(2e_1 + 3e_2)$
	$=< (-3)2, (-3)3 >$	$= (-3)2e_1 + (-3)(1e_1)$
	$=< -6, -9 >$	$= -6e_1 - 9e_2$

Example 5.8.7: non-basis

We need to cover all the vectors:



Let's try $V_1 =< 1, 0 >, V_2 =< 2, 0 >$. What are all possible linear combinations? For all pairs r_1, r_2, we have

$$X = r_1V_1 + r_2V_2 = r_1 < 1, 0 > +r_2 < 2, 0 >=< r_1 + 2r_2, 0 > \ .$$

The second component will remain 0 no matter what the coefficients are! This is not a basis because we can't represent some of the vectors.

The general result is as follows:

Theorem 5.8.8: Basis on the Plane

Any two vectors that aren't multiples of each other and only they form a basis of $\mathbf{R}^2$*; i.e.,*

$$V_1 = rV_2 \iff \{V_1, V_2\} \ \text{ is not a basis.}$$

Exercise 5.8.9

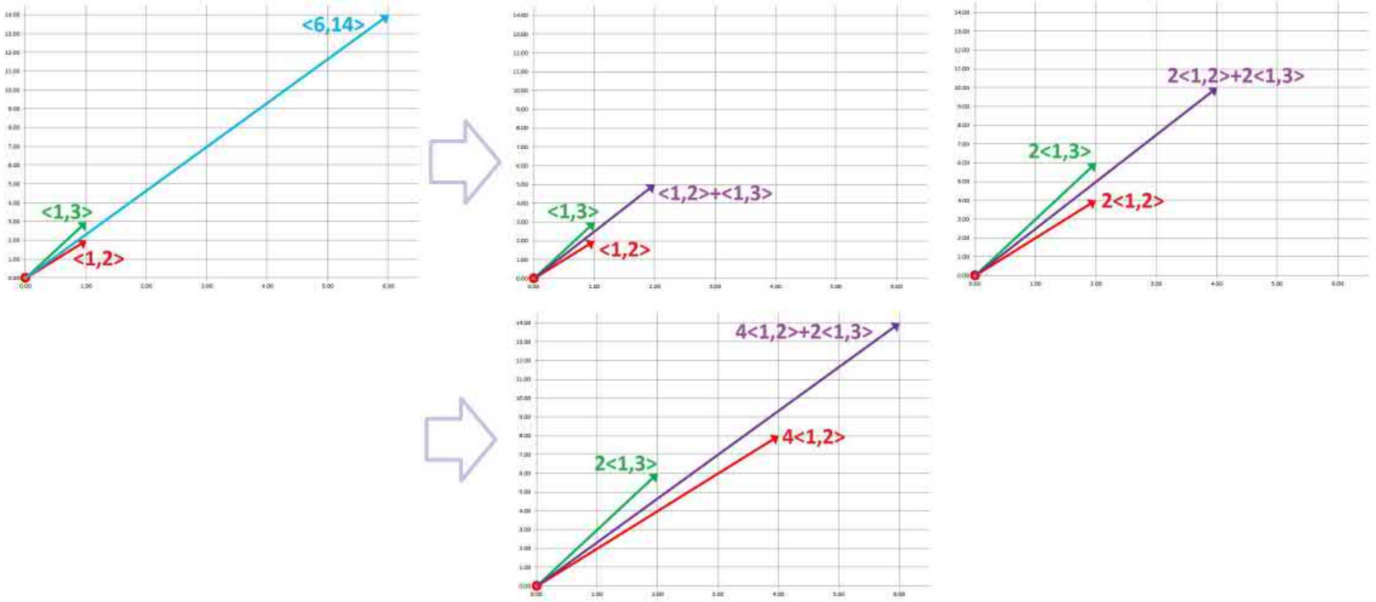
Prove the theorem.

Example 5.8.10: linear system

Recall from the beginning of the chapter how the solution to a system of linear equations was seen as if the two equations were equations about the coefficients, x and y, of *vectors* in the plane:

$$x \begin{bmatrix} 1 \\ 2 \end{bmatrix} + y \begin{bmatrix} 1 \\ 3 \end{bmatrix} = \begin{bmatrix} 6 \\ 14 \end{bmatrix} .$$

To solve the system is to find a way to stretch these two vectors so that after adding them the result is the vector on the right:

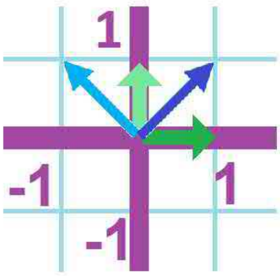


Now we conclude that we can guarantee that there is a solution only when the two vectors form a basis! For example, this mixture problem doesn't have a solution:

$$x \begin{bmatrix} 1 \\ 2 \end{bmatrix} + y \begin{bmatrix} 2 \\ 6 \end{bmatrix} = \begin{bmatrix} 6 \\ 14 \end{bmatrix} .$$

Example 5.8.11: coordinates

Consider an alternative basis along with the standard one:



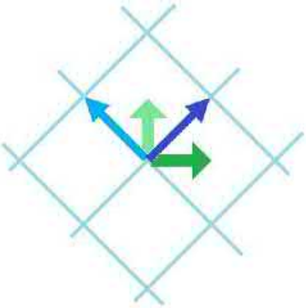
It's easy to express the new vectors in terms of old:

$$V_1 =< 1, 1 >, \ V_2 =< -1, 1 > .$$

But what about vice versa? It's harder; we need to find a and b so that

$$e_1 = aV_1 + bV_2 ,$$

and the same for e_2. We draw a grid for the new system to help:



The answer is:

$$e_1 = \frac{1}{2}V_1 - \frac{1}{2}V_2,\ e_2 = \frac{1}{2}V_1 + \frac{1}{2}V_2\,.$$

We can then re-write these vectors in the language of components *with respect to the new basis*:

$$e_1 = \left\langle \frac{1}{2}, -\frac{1}{2} \right\rangle = \frac{1}{2} < 1, -1 >,\ e_2 = \left\langle \frac{1}{2}, \frac{1}{2} \right\rangle = \frac{1}{2} < 1, 1 >\ .$$

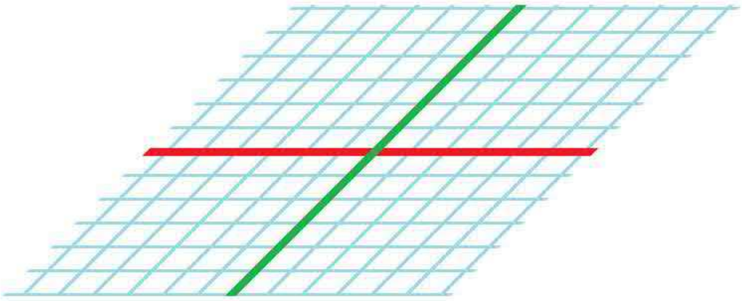
In summary, we have different coefficients and, therefore, different components of the same vectors with respect to different bases:

bases:	$\{e_1, e_2\}$	$\{V_1, V_2\}$
$V_1 =$	$< 1, 1 >$	$=< 1, 0 >$
$V_2 =$	$< -1, 1 >$	$=< 0, 1 >$

and

bases:	$\{e_1, e_2\}$	$\{V_1, V_2\}$
$e_1 =$	$< 1, 0 >$	$= \left\langle \frac{1}{2}, -\frac{1}{2} \right\rangle$
$e_2 =$	$< 0, 1 >$	$= \left\langle \frac{1}{2}, -\frac{1}{2} \right\rangle$

In general, the vectors of the alternative basis might have any angle between them (as long as it's not zero). Then, we have a skewed grid:



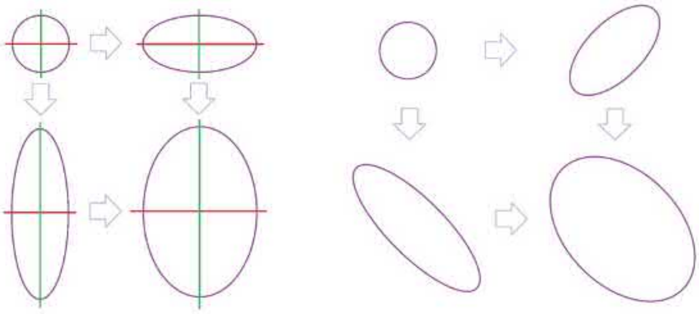
Thus, the component-wise algebra is fully operational whatever basis we choose:

$$< a, b > + < c, d > = < a + c, b + d > \quad \text{and} \quad r < a, b > = < ra, rb >\ .$$

Warning!

Unlike algebra, the geometry of the Cartesian system relies on the Pythagorean Theorem. As a result, the formulas for magnitudes and the dot products fail in the current form if used for nonperpendicular bases.

Now, the linear operators. They can be described *apart* from the Cartesian system (that was added to the vector space), i.e., rotation, stretching, etc.:



However, just as vector algebra works componentwise, so do linear operators. Furthermore, this approach works with respect to any basis:

Theorem 5.8.12: Linear Operator in Terms of Basis

Suppose $\{V_1, V_2\}$ *is a basis. Then, all values of a linear operator* $Y = F(X)$ *are expressed as linear combinations of its values on these vectors; i.e., for any pair of real coefficients* r_1 *and* r_2*, we have:*

$$X = r_1V_1 + r_2V_2 \implies F(X) = r_1F(V_1) + r_2F(V_2)\,.$$

Exercise 5.8.13

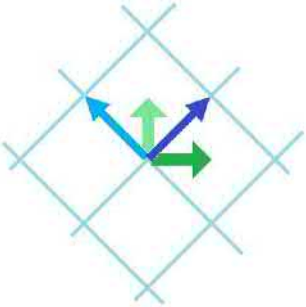
Prove the theorem.

In other words, the operator is fully determined by its values on the basis vectors – just as with the standard basis. But, just as with vectors, we have different matrices for the same linear operator with respect to different bases. The columns of the matrix of A are the values of the basis vectors under the operator:

$$A(V_1) \text{ and } A(V_2)\,.$$

Example 5.8.14: matrices

We, again, consider this alternative basis along with the standard one:



The mutual representations are:

$$e_1 = \frac{1}{2}V_1 - \frac{1}{2}V_2,\ e_2 = \frac{1}{2}V_1 + \frac{1}{2}V_2\,,$$

and

$$V_1 = e_1 + e_2,\ V_2 = -e_1 + e_2\,.$$

Suppose an operator A stretches the plane along the x-axis (in other words, along e_1) by a factor of

2. Then the matrix of A with respect to $\{e_1, e_2\}$ is:

$$A = \begin{bmatrix} 2 & 0 \\ 0 & 1 \end{bmatrix}.$$

What about the other basis, $\{V_1, V_2\}$? It is hard to guess this time because – unlike for the standard basis – the change is not aligned with the basis vectors. Let's use the "convenient" basis and then switch to the one we are interested in. We write the formulas in terms of e_1, e_2 first:

$$\begin{aligned} A(V_1) &= A(e_1 + e_2) &&= A(e_1) + A(e_2) &&= 2e_1 + e_2. \\ A(V_2) &= A(-e_1 + e_2) &&= -A(e_1) + A(e_2) &&= -2e_1 + e_2. \end{aligned}$$

Then we substitute the values of e_1, e_2 in terms of V_1, V_2 as written above:

$$\begin{aligned} A(V_1) &= 2e_1 + e_2 &&= 2\left(\frac{1}{2}V_1 - \frac{1}{2}V_2\right) + \left(\frac{1}{2}V_1 + \frac{1}{2}V_2\right) &&= \frac{3}{2}V_1 - \frac{1}{2}V_2. \\ A(V_2) &= -2e_1 + e_2 &&= -2\left(\frac{1}{2}V_1 - \frac{1}{2}V_2\right) + \left(\frac{1}{2}V_1 + \frac{1}{2}V_2\right) &&= -\frac{1}{2}V_1 + \frac{3}{2}V_2. \end{aligned}$$

Therefore, the matrix of the linear operator with respect to $\{V_1, V_2\}$ is:

$$A = \begin{bmatrix} \frac{3}{2} & -\frac{1}{2} \\ -\frac{1}{2} & \frac{3}{2} \end{bmatrix}.$$

What if the stretch was along V_1? This operator's matrix with respect to $\{V_1, V_2\}$ is simple:

$$B = \begin{bmatrix} 2 & 0 \\ 0 & 1 \end{bmatrix}.$$

Exercise 5.8.15

Rewrite the above computation in terms of components.

5.9. Classification of linear operators according to their eigenvalues

We apply the last theorem to eigenvectors.

Corollary 5.9.1: Representation in Terms of Eigenvectors

Suppose V_1 and V_2 are two eigenvectors of a linear operator F that correspond to two (possibly equal) eigenvalues λ_1 and λ_2. Suppose also that V_1 and V_2 aren't multiples of each other. Then, all values of the linear operator $Y = F(X)$ are represented as linear combinations of its values on the eigenvectors:

$$X = r_1V_1 + r_2V_2 \implies F(X) = r_1\lambda_1V_1 + r_2\lambda_2V_2\,,$$

with some real coefficients r_1 and r_2.

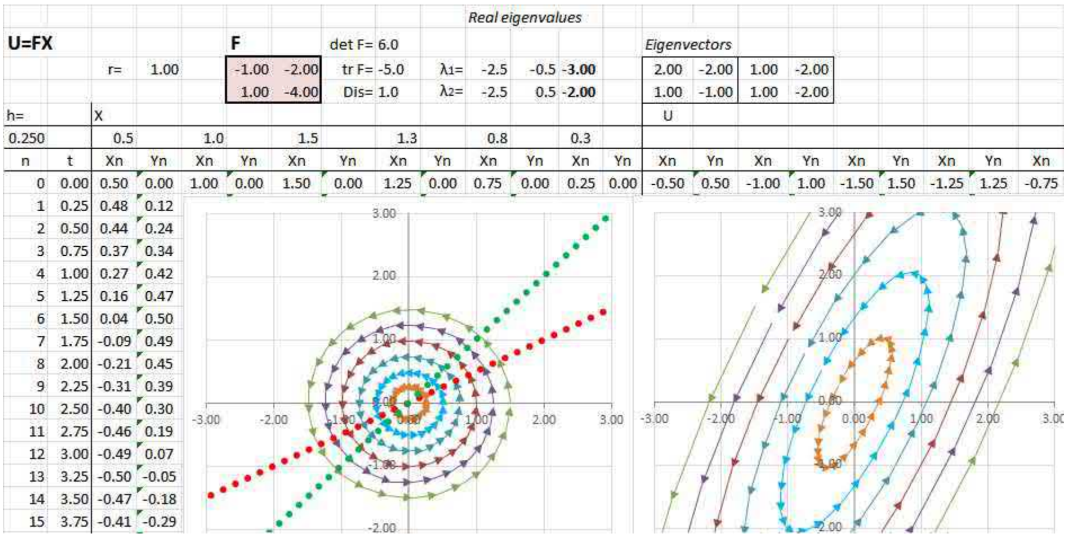
In other words, the matrix of F with respect to the basis $\{V_1, V_2\}$ of eigenvectors is diagonal with the eigenvalues on the diagonal:

$$F = \begin{bmatrix} \lambda_1 & 0 \\ 0 & \lambda_2 \end{bmatrix}.$$

Example 5.9.2: stretch-shrink

Let's consider this function:

$$\begin{cases} u &= -x &- &2y \\ v &= x &- &4y \end{cases} \Longrightarrow F = \begin{bmatrix} -1 & -2 \\ 1 & -4 \end{bmatrix}.$$



Real eigenvalues

U=FX **F** det F= 6.0 *Eigenvectors*

r= 1.00 -1.00 -2.00 tr F= -5.0 λ1= -2.5 -0.5 **-3.00** 2.00 -2.00 1.00 -2.00

1.00 -4.00 Dis= 1.0 λ2= -2.5 0.5 **-2.00** 1.00 -1.00 1.00 -2.00

h=		X												U								
0.250		0.5		1.0		1.5		1.3		0.8		0.3										
n	t	Xn	Yn	Xn	Yn	Xn	Yn	Xn	Yn	Xn	Yn	Xn	Yn	Xn	Yn	Xn	Yn	Xn	Yn	Xn	Yn	Xn
0	0.00	0.50	0.00	1.00	0.00	1.50	0.00	1.25	0.00	0.75	0.00	0.25	0.00	-0.50	0.50	-1.00	1.00	-1.50	1.50	-1.25	1.25	-0.75
1	0.25	0.48	0.12																			
2	0.50	0.44	0.24																			
3	0.75	0.37	0.34																			
4	1.00	0.27	0.42																			
5	1.25	0.16	0.47																			
6	1.50	0.04	0.50																			
7	1.75	-0.09	0.49																			
8	2.00	-0.21	0.45																			
9	2.25	-0.31	0.39																			
10	2.50	-0.40	0.30																			
11	2.75	-0.46	0.19																			
12	3.00	-0.49	0.07																			
13	3.25	-0.50	-0.05																			
14	3.50	-0.47	-0.18																			
15	3.75	-0.41	-0.29																			

Let's confirm what is shown above. The analysis starts with the characteristic polynomial:

$$\det(F - \lambda I) = \det \begin{bmatrix} -1-\lambda & -2 \\ 1 & -4-\lambda \end{bmatrix} = \lambda^2 - 5\lambda + 6\,.$$

Therefore, the *eigenvalues* are:

$$\lambda_1 = -3,\ \lambda_2 = -2\,.$$

To find the *eigenvectors*, we solve the two vector equations:

$$FV_i = \lambda_i V_i,\ i = 1, 2\,.$$

The first, $\lambda_1 = -3$:

$$FV_1 = \begin{bmatrix} -1 & -2 \\ 1 & -4 \end{bmatrix} \begin{bmatrix} x \\ y \end{bmatrix} = -3 \begin{bmatrix} x \\ y \end{bmatrix}.$$

This gives us the following system of linear equations:

$$\begin{cases} -x &- &2y &= -3x \\ x &- &4y &= -3y \end{cases} \Longrightarrow \begin{cases} 2x &- &2y &= 0 \\ x &- &y &= 0 \end{cases} \Longrightarrow x = y\,.$$

We have discovered, again, that this is the same equation; this line gives us the eigenspace. We choose one eigenvector:

$$V_1 = \begin{bmatrix} 1 \\ 1 \end{bmatrix}.$$

The second eigenvalue:

$$FV_2 = \begin{bmatrix} -1 & -2 \\ 1 & -4 \end{bmatrix} \begin{bmatrix} x \\ y \end{bmatrix} = -2 \begin{bmatrix} x \\ y \end{bmatrix} .$$

We have the following system (same equation):

$$\begin{cases} -x & - & 2y & = -2x \\ x & - & 4y & = -2y \end{cases} \Longrightarrow \begin{cases} x & - & 2y & = 0 \\ x & - & 2y & = 0 \end{cases} \Longrightarrow x = 2y\,.$$

This line is the eigenspace of $\lambda_2 = -2$. We choose one eigenvector:

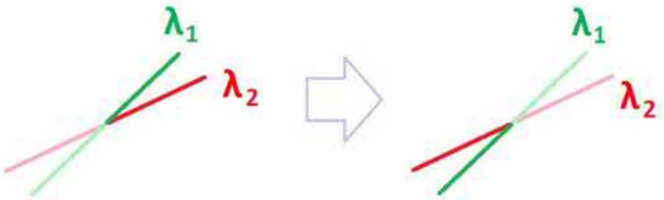
$$V_2 = \begin{bmatrix} 2 \\ 1 \end{bmatrix} .$$

The pair $\{V_1, V_2\}$ is a basis!

We summarize what F does:

1. A flip and stretch along the vector $< 1, 1 >$: The line $y = x$ remains intact.
2. A flip and stretch along the vector $< 2, 1 >$: The line $x = 2y$ remains intact.

We also conclude that there is no change of orientation. Stretching aside, this looks like central symmetry:



We observe fanning between these two lines. For the rest of the vectors, we have:

$$X = xV_1 + yV_2 \Longrightarrow FX = -3x \begin{bmatrix} 2 \\ -1 \end{bmatrix} - 2y \begin{bmatrix} 1 \\ 2 \end{bmatrix} .$$

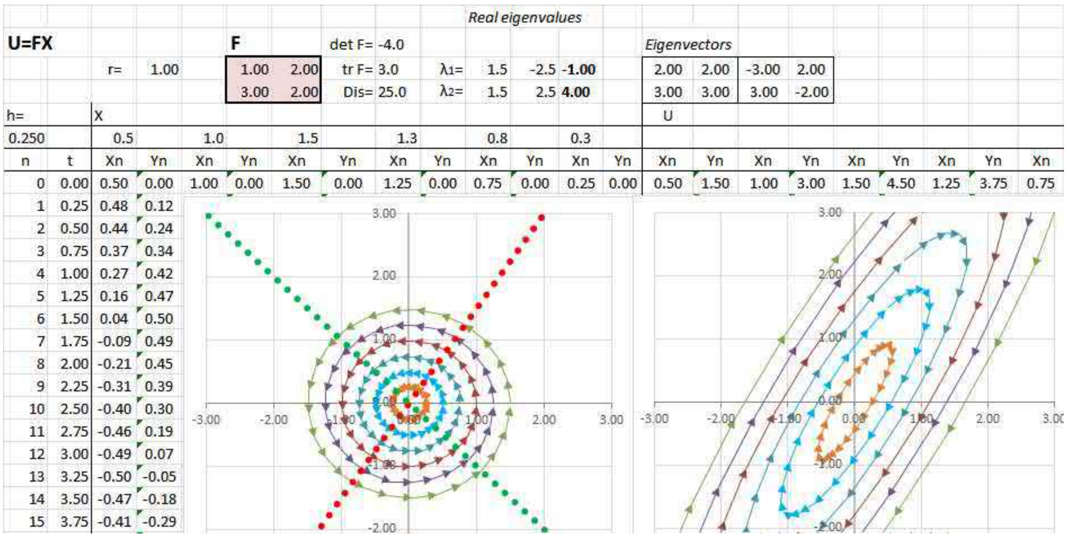
Therefore, matrix of F with respect to the basis $\{V_1, V_2\}$ is diagonal:

$$F = \begin{bmatrix} -3 & 0 \\ 0 & -2 \end{bmatrix} .$$

Example 5.9.3: stretch-shrink

Let's consider this linear operator:

$$\begin{cases} u & = x & + & 2y \\ v & = 3x & + & 2y \end{cases} \Longrightarrow F = \begin{bmatrix} 1 & 2 \\ 3 & 2 \end{bmatrix} .$$



Real eigenvalues

U=FX — **F** — det F= -4.0 — *Eigenvectors*

r=	1.00	1.00	2.00	tr F= 3.0	λ1=	1.5	-2.5	**-1.00**	2.00	2.00	-3.00	2.00
		3.00	2.00	Dis= 25.0	λ2=	1.5	2.5	**4.00**	3.00	3.00	3.00	-2.00

h=		X												U								
0.250		0.5		1.0		1.5		1.3		0.8		0.3										
n	t	Xn	Yn	Xn	Yn	Xn	Yn	Xn	Yn	Xn	Yn	Xn	Yn	Xn	Yn	Xn	Yn	Xn	Yn	Xn	Yn	Xn
0	0.00	0.50	0.00	1.00	0.00	1.50	0.00	1.25	0.00	0.75	0.00	0.25	0.00	0.50	1.50	1.00	3.00	1.50	4.50	1.25	3.75	0.75
1	0.25	0.48	0.12																			
2	0.50	0.44	0.24																			
3	0.75	0.37	0.34																			
4	1.00	0.27	0.42																			
5	1.25	0.16	0.47																			
6	1.50	0.04	0.50																			
7	1.75	-0.09	0.49																			
8	2.00	-0.21	0.45																			
9	2.25	-0.31	0.39																			
10	2.50	-0.40	0.30																			
11	2.75	-0.46	0.19																			
12	3.00	-0.49	0.07																			
13	3.25	-0.50	-0.05																			
14	3.50	-0.47	-0.18																			
15	3.75	-0.41	-0.29																			

Let's find the eigenvectors:

$$\det(F - \lambda I) = \det \begin{bmatrix} 1-\lambda & 2 \\ 3 & 2-\lambda \end{bmatrix} = \lambda^2 - 3\lambda - 4\,.$$

Therefore, the eigenvalues are:

$$\lambda_1 = -1,\ \lambda_2 = 4\,.$$

Now we find the eigenvectors. We solve the two equations:

$$FV_i = \lambda_i V_i,\ i = 1, 2\,.$$

The first:

$$FV_1 = \begin{bmatrix} 1 & 2 \\ 3 & 2 \end{bmatrix} \begin{bmatrix} x \\ y \end{bmatrix} = -1 \begin{bmatrix} x \\ y \end{bmatrix}\,.$$

This gives as the following system of linear equations:

$$\begin{cases} x & + & 2y & = -x \\ 3x & + & 2y & = -y \end{cases} \Longrightarrow \begin{cases} 2x & + & 2y & = 0 \\ 3x & + & 3y & = 0 \end{cases} \Longrightarrow x = -y\,.$$

We choose:

$$V_1 = \begin{bmatrix} 1 \\ -1 \end{bmatrix}\,.$$

Every value within this eigenspace (the line $y = -x$) is a multiple of this eigenvector:

$$X = \lambda_1 V_1 = - \begin{bmatrix} 1 \\ -1 \end{bmatrix}\,.$$

The second eigenvalue:

$$FV_2 = \begin{bmatrix} 1 & 2 \\ 3 & 2 \end{bmatrix} \begin{bmatrix} x \\ y \end{bmatrix} = 4 \begin{bmatrix} x \\ y \end{bmatrix}\,.$$

We have the following system:

$$\begin{cases} x & + & 2y & = 4x \\ 3x & + & 2y & = 4y \end{cases} \Longrightarrow \begin{cases} -3x & + & 2y & = 0 \\ 3x & - & 2y & = 0 \end{cases} \Longrightarrow x = 2y/3\,.$$

We choose:

$$V_2 = \begin{bmatrix} 1 \\ 3/2 \end{bmatrix}.$$

Every value within this eigenspace (the line $y = 3x/2$) is a multiple of this eigenvector:

$$X = \lambda_2 V_2 = 4 \begin{bmatrix} 1 \\ 3/2 \end{bmatrix}.$$

The pair $\{V_1, V_2\}$ is a basis! Then,

$$X = xV_1 + yV_2 \implies U = F(X) = -xV_1 + 4yV_2 .$$

The matrix of F with respect to the basis $\{V_1, V_2\}$ is:

$$F = \begin{bmatrix} -1 & 0 \\ 0 & 4 \end{bmatrix}.$$

We observe fanning between these two lines.

Let's summarize the results.

Theorem 5.9.4: Classification of Linear Operators – Real Eigenvalues

Suppose matrix F has two real non-zero eigenvalues λ_1 and λ_2. Then, the function $U = F(X)$ stretches/shrinks the two eigenspace by factors $|\lambda_1|$ and $|\lambda_2|$ respectively and, furthermore:

- *If λ_1 and λ_2 have the same sign, it preserves the orientation of a closed curve around the origin.*
- *If λ_1 and λ_2 have the opposite signs, it reverses the orientation of a closed curve around the origin.*

Exercise 5.9.5

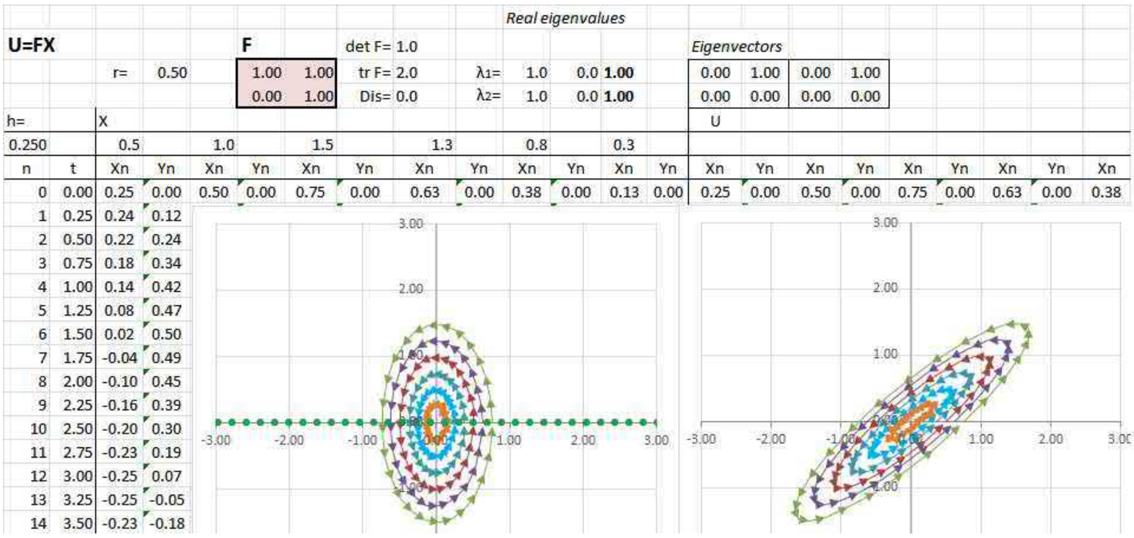
Apply the theorem to the last example.

Example 5.9.6: skewing-shearing

Consider this matrix:

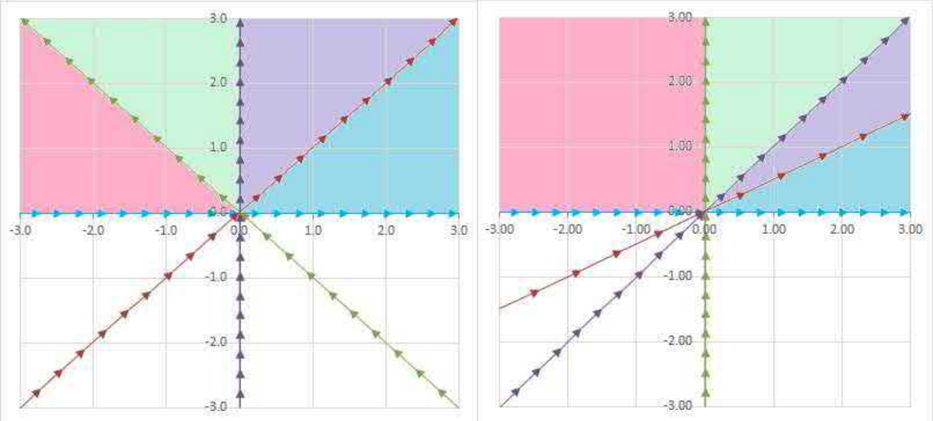
$$F = \begin{bmatrix} 1 & 1 \\ 0 & 1 \end{bmatrix}.$$

Below, we replace a circle with an ellipse to see what happens to it under such a function:

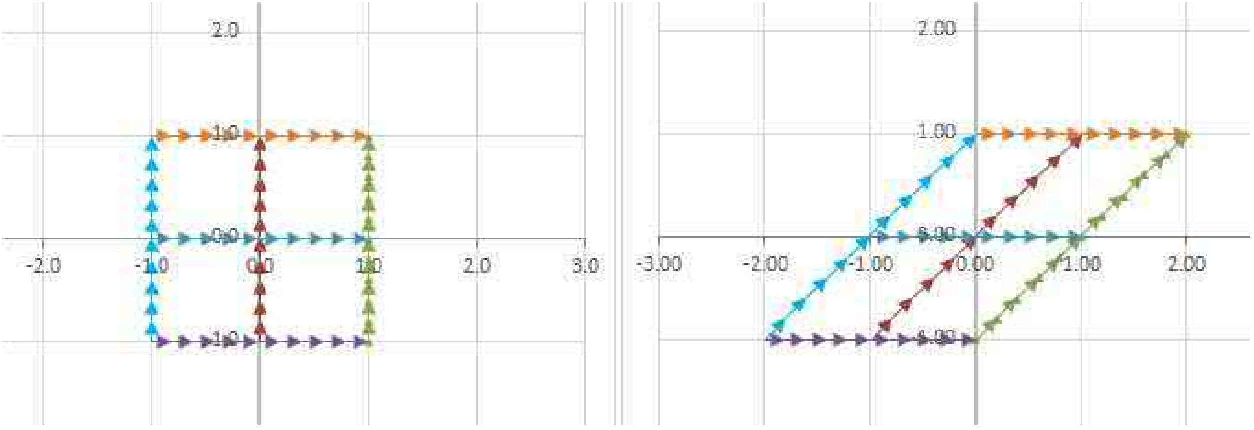


										Real eigenvalues												
U=FX					**F**		det F=	1.0						*Eigenvectors*								
		r=	0.50		1.00	1.00	tr F=	2.0	λ1=	1.0	0.0	**1.00**		0.00	1.00	0.00	1.00					
					0.00	1.00	Dis=	0.0	λ2=	1.0	0.0	**1.00**		0.00	0.00	0.00	0.00					
h=		X												U								
0.250		0.5		1.0		1.5		1.3		0.8		0.3										
n	t	Xn	Yn	Xn	Yn	Xn	Yn	Xn	Yn	Xn	Yn	Xn	Yn	Xn	Yn	Xn	Yn	Xn	Yn	Xn	Yn	Xn
0	0.00	0.25	0.00	0.50	0.00	0.75	0.00	0.63	0.00	0.38	0.00	0.13	0.00	0.25	0.00	0.50	0.00	0.75	0.00	0.63	0.00	0.38
1	0.25	0.24	0.12																			
2	0.50	0.22	0.24																			
3	0.75	0.18	0.34																			
4	1.00	0.14	0.42																			
5	1.25	0.08	0.47																			
6	1.50	0.02	0.50																			
7	1.75	-0.04	0.49																			
8	2.00	-0.10	0.45																			
9	2.25	-0.16	0.39																			
10	2.50	-0.20	0.30																			
11	2.75	-0.23	0.19																			
12	3.00	-0.25	0.07																			
13	3.25	-0.25	-0.05																			
14	3.50	-0.23	-0.18																			

There is still angular stretch-shrink but this time it is between the two ends of the same line. We see "fanning out" again:



This time, however, the fan is fully open! It makes a difference that the fanning happens to a whole half-plane. To see more clearly, consider what happens to a square:



This is the characteristic polynomial:

$$\det(F - \lambda I) = \det \begin{bmatrix} 1-\lambda & 1 \\ 0 & 1-\lambda \end{bmatrix} = (1-\lambda)^2 \, .$$

Therefore, the eigenvalues are

$$\lambda_1 = \lambda_2 = 1 \, .$$

What are the eigenvectors?

$$FV = \begin{bmatrix} 1 & 1 \\ 0 & 1 \end{bmatrix} \begin{bmatrix} x \\ y \end{bmatrix} = 1 \begin{bmatrix} x \\ y \end{bmatrix} \, .$$

This gives as the following system of linear equations:

$$\begin{cases} x & + & y & = x \ \texttt{AND} \\ & & y & = y \end{cases} \implies x \text{ any}, \ y = 0\,.$$

The only eigenvectors are horizontal! Therefore, our classification theorem doesn't apply. There is no diagonal matrix for this operator.

Example 5.9.7: rotations

There are other outcomes that the theorem doesn't cover. Recall the characteristic polynomial of the matrix A of the 90-degree rotation:

$$\chi_A(\lambda) = \det \begin{bmatrix} -\lambda & -1 \\ 1 & -\lambda \end{bmatrix} = \lambda^2 + 1\,.$$

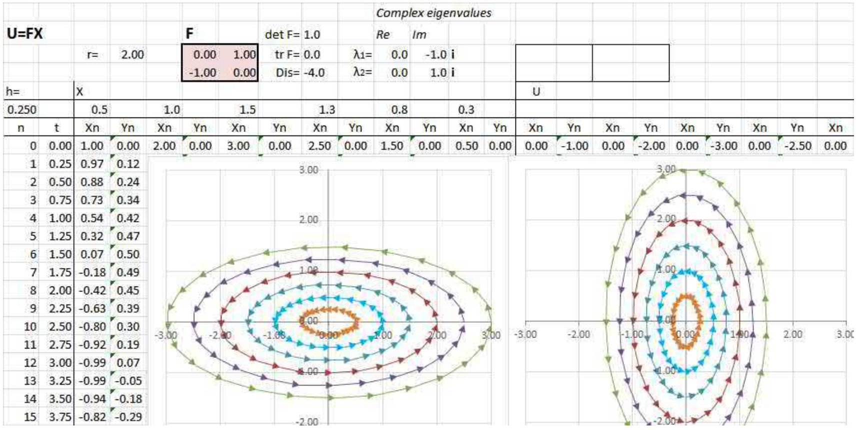
But the characteristic equation,

$$x^2 + 1 = 0\,,$$

has no solutions! Are we done then? Not if we are willing to use *complex numbers* (next chapter):

$$\lambda_{1,2} = i \ \text{ and } \ -i\,.$$

This is the effect of rotation:



Complex eigenvalues

U=FX

F

det F= 1.0 | *Re* | *Im*

r= 2.00

0.00	1.00
-1.00	0.00

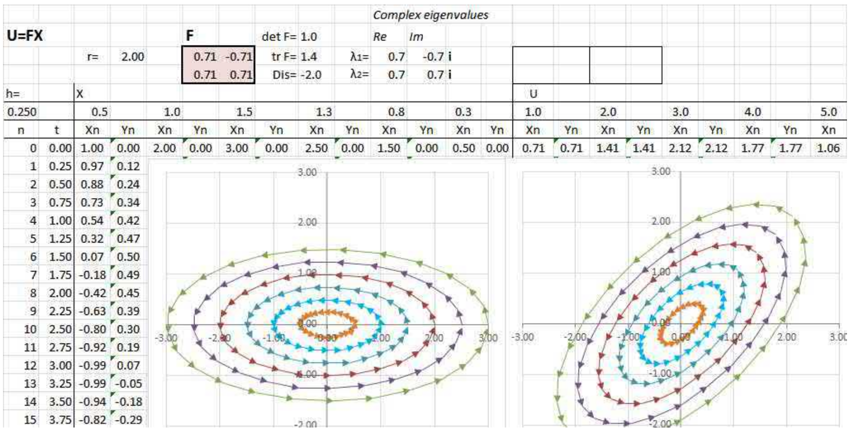
tr F= 0.0 λ1= 0.0 -1.0 i

Dis= -4.0 λ2= 0.0 1.0 i

h=		X												U								
0.250		0.5		1.0		1.5		1.3		0.8		0.3										
n	t	Xn	Yn	Xn	Yn	Xn	Yn	Xn	Yn	Xn	Yn	Xn	Yn	Xn	Yn	Xn	Yn	Xn	Yn	Xn	Yn	Xn
0	0.00	1.00	0.00	2.00	0.00	3.00	0.00	2.50	0.00	1.50	0.00	0.50	0.00	0.00	-1.00	0.00	-2.00	0.00	-3.00	0.00	-2.50	0.00
1	0.25	0.97	0.12																			
2	0.50	0.88	0.24																			
3	0.75	0.73	0.34																			
4	1.00	0.54	0.42																			
5	1.25	0.32	0.47																			
6	1.50	0.07	0.50																			
7	1.75	-0.18	0.49																			
8	2.00	-0.42	0.45																			
9	2.25	-0.63	0.39																			
10	2.50	-0.80	0.30																			
11	2.75	-0.92	0.19																			
12	3.00	-0.99	0.07																			
13	3.25	-0.99	-0.05																			
14	3.50	-0.94	-0.18																			
15	3.75	-0.82	-0.29																			

Let's consider a *rotation through an arbitrary angle* θ:

$$\begin{bmatrix} u \\ v \end{bmatrix} = \begin{bmatrix} \cos\theta & -\sin\theta \\ \sin\theta & \cos\theta \end{bmatrix} \begin{bmatrix} x \\ y \end{bmatrix}.$$



Complex eigenvalues

U=FX

F

det F= 1.0 | *Re* | *Im*

r= 2.00

0.71	-0.71
0.71	0.71

tr F= 1.4 λ1= 0.7 -0.7 i

Dis= -2.0 λ2= 0.7 0.7 i

h=		X												U								
0.250		0.5		1.0		1.5		1.3		0.8		0.3		1.0		2.0		3.0		4.0		5.0
n	t	Xn	Yn	Xn	Yn	Xn	Yn	Xn	Yn	Xn	Yn	Xn	Yn	Xn	Yn	Xn	Yn	Xn	Yn	Xn	Yn	Xn
0	0.00	1.00	0.00	2.00	0.00	3.00	0.00	2.50	0.00	1.50	0.00	0.50	0.00	0.71	0.71	1.41	1.41	2.12	2.12	1.77	1.77	1.06
1	0.25	0.97	0.12																			
2	0.50	0.88	0.24																			
3	0.75	0.73	0.34																			
4	1.00	0.54	0.42																			
5	1.25	0.32	0.47																			
6	1.50	0.07	0.50																			
7	1.75	-0.18	0.49																			
8	2.00	-0.42	0.45																			
9	2.25	-0.63	0.39																			
10	2.50	-0.80	0.30																			
11	2.75	-0.92	0.19																			
12	3.00	-0.99	0.07																			
13	3.25	-0.99	-0.05																			
14	3.50	-0.94	-0.18																			
15	3.75	-0.82	-0.29																			

The angle can be seen in the characteristic polynomial:

$$\chi_A(\lambda) = (\cos\theta - \lambda)^2 + \sin^2\theta = \cos^2\theta - 2\cos\theta\,\lambda + \lambda^2 + \sin^2\theta = \lambda^2 - 2\cos\theta\,\lambda + 1\,.$$

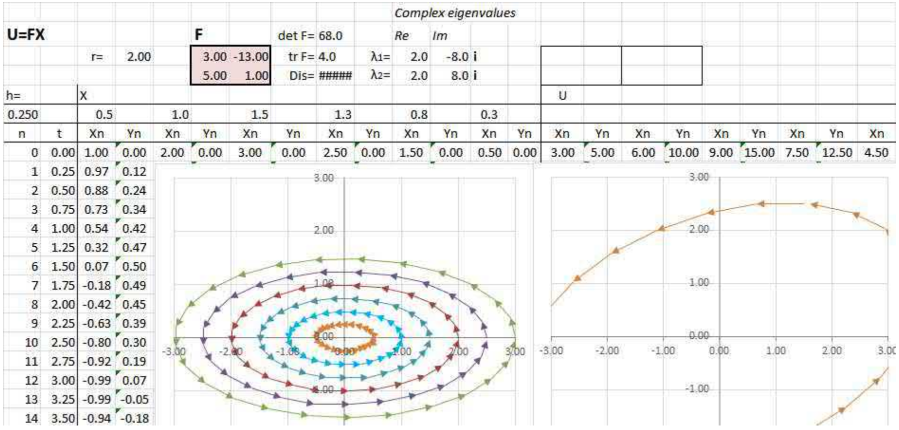
The discriminant of this polynomial is negative. Therefore, it has no real roots. The result makes sense: A rotation cannot possibly have eigenvectors because all vectors are rotated!

Example 5.9.8: rotation with stretch-shrink

Let's consider this linear operator:

$$\begin{cases} u &= 3x & -13y\,, \\ v &= 5x & +y\,, \end{cases} \quad \text{and} \quad F = \begin{bmatrix} 3 & -13 \\ 5 & 1 \end{bmatrix}.$$

Below we can recognize both rotation and re-scaling:



Complex eigenvalues

U=FX

F

det F= 68.0

Re Im

r= 2.00

3.00 -13.00

5.00 1.00

tr F= 4.0

Dis= ######

λ1= 2.0 -8.0 i

λ2= 2.0 8.0 i

h=		X												U								
0.250		0.5		1.0		1.5		1.3		0.8		0.3										
n	t	Xn	Yn	Xn	Yn	Xn	Yn	Xn	Yn	Xn	Yn	Xn	Yn	Xn	Yn	Xn	Yn	Xn	Yn	Xn	Yn	Xn
0	0.00	1.00	0.00	2.00	0.00	3.00	0.00	2.50	0.00	1.50	0.00	0.50	0.00	3.00	5.00	6.00	10.00	9.00	15.00	7.50	12.50	4.50
1	0.25	0.97	0.12																			
2	0.50	0.88	0.24																			
3	0.75	0.73	0.34																			
4	1.00	0.54	0.42																			
5	1.25	0.32	0.47																			
6	1.50	0.07	0.50																			
7	1.75	-0.18	0.49																			
8	2.00	-0.42	0.45																			
9	2.25	-0.63	0.39																			
10	2.50	-0.80	0.30																			
11	2.75	-0.92	0.19																			
12	3.00	-0.99	0.07																			
13	3.25	-0.99	-0.05																			
14	3.50	-0.94	-0.18																			

This is our characteristic polynomial:

$$\chi(\lambda) = \det(F - \lambda I) = \det \begin{bmatrix} 3-\lambda & -13 \\ 5 & 1-\lambda \end{bmatrix} = \lambda^2 - 4\lambda + 68\,.$$

Exercise 5.9.9

What does it tell us?

Our interpretation of the characteristic polynomial in terms of the trace of the matrix:

$$\chi(\lambda) = \lambda^2 - \operatorname{tr} F\,\lambda + \det F\,.$$

allows us to prove in the next chapter the following result for the case of no real eigenvalues:

Corollary 5.9.10: Trace and Discriminant

Suppose the discriminant of the characteristic polynomial of a matrix F satisfies:

$$D = (\operatorname{tr} F)^2 - 4\det F \leq 0\,.$$

Then, the operator $U = FX$ does the following:

1. *It rotates the real plane through the following angle:*

$$\theta = \sin^{-1}\left(\frac{1}{2}\sqrt{\frac{4 - (\operatorname{tr} F)^2}{\det F}}\right).$$

2. *It re-scales the plane uniformly by the following factor:*

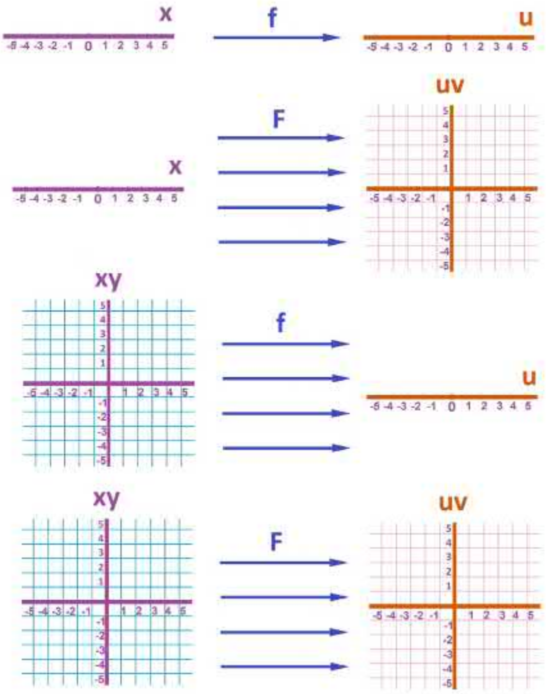
$$s = \sqrt{\det F}\,.$$

Exercise 5.9.11

Apply the corollary to the last example.

5.10. Algebra of linear operators and matrices

We will take a broader view at linear operator and include the lower dimensions. These are the four possibilities:



We want to understand how these operators are represented by matrices and how these matrices are combined to produce *compositions*.

The rule remains:

> ▶ The value under F of each basis vector in the domain of F becomes a column in the matrix of F.

Let's apply the rule to these four situations using the standard bases.

Example 5.10.1: dimensions 1 and 1

Suppose we have a linear operator, which is just a numerical function:

$$f : \mathbf{R} \to \mathbf{R} \ \text{ defined by } \ f(x) = 3x\,.$$

It is a stretch by a factor of 3. What is its matrix? The basis of the x-axis is $< 1 >$ and the basis of the u-axis is $< 1 >$. The operator works as follows:

$$f(< 1 >) = 3 < 1 > \ .$$

Therefore, its matrix is

$$f = [3]\,.$$

Example 5.10.2: dimensions 1 and 2

Suppose we have a linear operator, which is just a parametric curve:

$$F : \mathbf{R} \to \mathbf{R}^2 \ \text{ defined by } \ F(x) =< 3x, 2x > \ .$$

It stretches the x-axis on the uv-plane along the vector $< 3, 2 >$. What is its matrix? The basis of the x-axis is $< 1 >$ and the basis of the uv-plane is $< 1, 0 >$ and $< 0, 1 >$. The operator works as follows:

$$F(< 1 >) =< 3, 2 > \ .$$

Therefore, its matrix is

$$F = \begin{bmatrix} 3 \\ 2 \end{bmatrix} \ .$$

Example 5.10.3: dimensions 2 and 1

Suppose we have a linear operator, which is just a function of two variables:

$$f : \mathbf{R}^2 \to \mathbf{R} \ \text{ defined by } \ f(x, y) = 3x + 2y \, .$$

It rolls the xy-plane on the u-axis. What is its matrix? The basis of the xy-plane is $< 1, 0 >$ and $< 0, 1 >$ and the basis of the u-axis is $< 1 >$. The operator works as follows:

$$f(< 1, 0 >) =< 3 > \quad \text{and} \quad f(< 0, 1 >) =< 2 > \ .$$

Therefore, its matrix is

$$f = [3 \ 2] \, .$$

Example 5.10.4: dimensions 2 and 2

Suppose we have a linear operator, which is just a transformation of the plane:

$$F : \mathbf{R}^2 \to \mathbf{R}^2 \ \text{ defined by } \ f(x, y) =< 3x + 2y, 5x - y > \ .$$

We'd need the eigenvector analysis in order to determine what it does... What is its matrix? The basis of the xy-plane is $< 1, 0 >$ and $< 0, 1 >$ and the basis of the u-axis is $< 1, 0 >$ and $< 0, 1 >$. The operator works as follows:

$$F(< 1, 0 >) =< 3, 5 > \quad \text{and} \quad F(< 0, 1 >) =< 2, -1 > \ .$$

Therefore, its matrix is

$$F = \begin{bmatrix} 3 & 2 \\ 5 & -1 \end{bmatrix} \ .$$

As you can see, we can jump ahead of the rule described above and write the coefficients present in the formula of the linear operator straight into the matrix.

Warning!

Its matrix is just an abbreviated representation of a linear operator.

Exercise 5.10.5

Include the dimension 0 for domains and codomains in the above analysis.

Whenever there is algebra in the output space, we can use it to do algebra of the functions. If the codomain of functions is a vector space, we can add these functions and multiply them by a constant. We just narrow down this idea to linear operators:

Definition 5.10.6: addition of linear operators

Given two linear operators:

$$F, G : \mathbf{R}^n \to \mathbf{R}^m ,$$

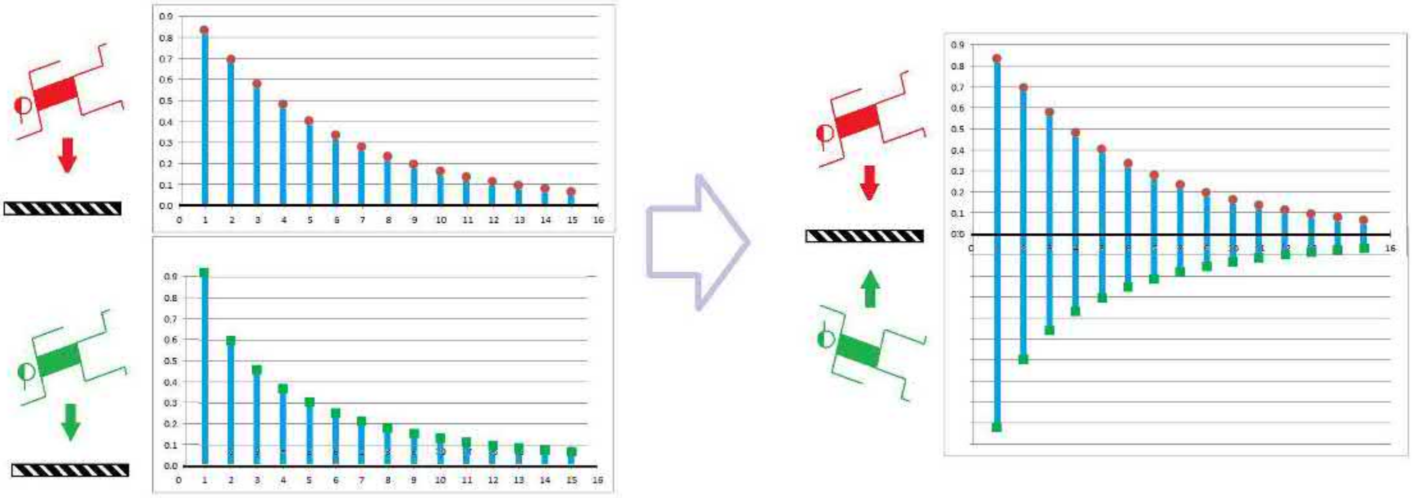
their *sum* is linear operator:

$$F + G : \mathbf{R}^n \to \mathbf{R}^m ,$$

defined by:

$$(F + G)(x) = F(x) + G(x) .$$

We illustrate this operation just as before:



But if these are linear operators, what happens to their matrices? We go though the same four cases below.

Example 5.10.7: dimensions 1 and 1

Given linear operators (numerical functions):

$$f, g : \mathbf{R} \to \mathbf{R} \ \text{ defined by } \ f(x) = 3x \quad \text{and} \quad g(x) = 2x .$$

Their matrices are:

$$f = [3] \quad \text{and} \quad g = [2] .$$

What about their sum? It is an operator with the same domain and codomain and it is computed as follows:

$$f + g : \mathbf{R} \to \mathbf{R} \ \text{ defined by } \ (f + g)(x) = f(x) + g(x) = 3x + 2x = 5x .$$

Its matrix is

$$f + g = [5] .$$

Of course, this new number is just the sum of the two original numbers.

Example 5.10.8: dimensions 1 and 2

Given linear operators (parametric curves):

$$F, G : \mathbf{R} \to \mathbf{R}^2 \ \text{ defined by } \ F(x) =< 3x, 2x > \quad \text{and} \quad G(x) =< 5x, -x > .$$

Their matrices are:

$$F=\begin{bmatrix}3\\2\end{bmatrix}\quad\text{and}\quad G=\begin{bmatrix}5\\-1\end{bmatrix}.$$

What about their sum? It is an operator with the same domain and codomain and it is computed as follows:

$$F+G:\mathbf{R}\to\mathbf{R}^2\ \text{ defined by }\ (F+G)(x)=F(x)+G(x)=<3x,2x>+<5x,-x>=<8x,x>\,.$$

Its matrix is

$$F+G=\begin{bmatrix}8\\1\end{bmatrix}.$$

Of course, this is just the sum of the two as if they were vectors.

Example 5.10.9: dimensions 2 and 1

Given linear operators (functions of two variables):

$$f,g:\mathbf{R}^2\to\mathbf{R}\ \text{ defined by }\ f(x,y)=3x+2y\quad\text{and}\quad g(x,y)=5x-2y\,.$$

Their matrices are:

$$f=[3,\ 2]\quad\text{and}\quad g=[5,\ -2]\,.$$

Their sum is an operator with the same domain and codomain and it is computed as follows (this is vector addition):

$$f+g:\mathbf{R}^2\to\mathbf{R}\ \text{ defined by }\ (f+g)(x,y)=f(x,y)+g(x,y)=3x+2y+5x-2y=8x\,.$$

Its matrix is

$$f=[8,\ 0]\,,$$

the sum – componentwise – of the two.

Example 5.10.10: dimensions 2 and 2

Given linear operators (transformations of the plane):

$$F,G:\mathbf{R}^2\to\mathbf{R}^2\ \text{ defined by }\ F(x,y)=<3x+2y,5x-y>\quad\text{and}\quad G(x,y)=<5x+y,x+y>\,.$$

Their matrices are:

$$F=\begin{bmatrix}3&2\\5&-1\end{bmatrix}\quad\text{and}\quad G=\begin{bmatrix}5&1\\1&1\end{bmatrix}.$$

Their sum is an operator with the same domain and codomain and it is computed as follows (this is vector addition):

$$F+G:\mathbf{R}^2\to\mathbf{R}^2\ \text{ defined by }\ (F+G)(x,y)=<3x+2y,5x-y>+<5x+y,x+y>=<8x+3y,6x>\,.$$

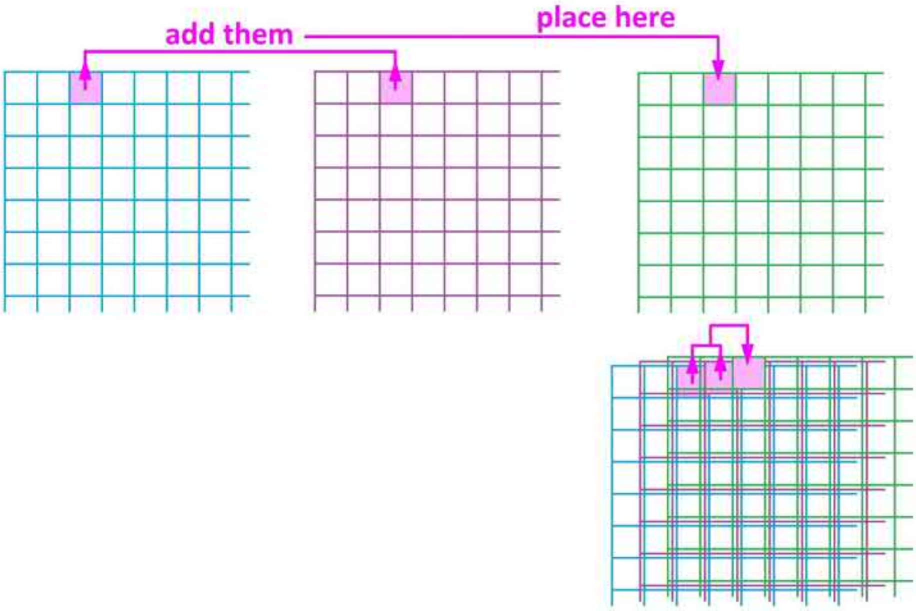
Its matrix is

$$F+G=\begin{bmatrix}8&3\\6&0\end{bmatrix},$$

the sum – componentwise – of the two:

$$F+G=\begin{bmatrix}3&2\\5&-1\end{bmatrix}+\begin{bmatrix}5&1\\1&1\end{bmatrix}=\begin{bmatrix}3+5&2+1\\5+1&(-1)+1\end{bmatrix}=\begin{bmatrix}8&3\\6&0\end{bmatrix}.$$

This operation of matrix addition is componentwise: The two operators have the same domain and codomain, so that the dimensions of the matrices are equal too. They can then be overlapped so that the entries are aligned and added accordingly:



Definition 5.10.11: addition of matrices

Suppose A and B are two $m \times n$ matrices. Then their *sum* is the $m \times n$ matrix, denoted by:

$$\boxed{A + B}$$

the ij-entry of which is the sum of the ij-entries of A and B.

In other words, if $A = a_{ij}$, $B = b_{ij}$, and $C = A + B = c_{ij}$, then

$$c_{ij} = a_{ij} + b_{ij}\,,$$

for each $i = 1, 2, ..., m$ and each $j = 1, 2, ..., n$.

It is simpler for scalar multiplication:

Definition 5.10.12: scalar multiplication of linear operator

Given a linear operator:

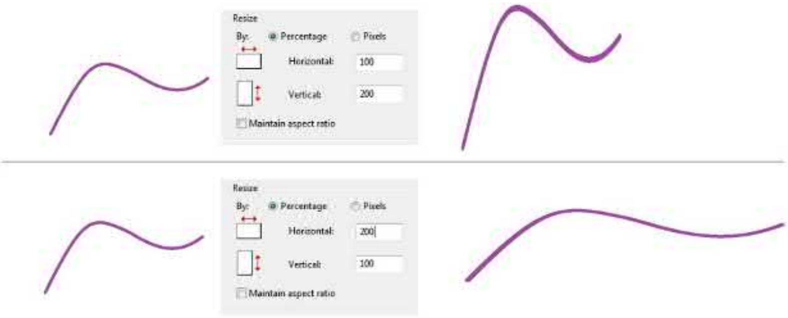
$$F : \mathbf{R}^n \to \mathbf{R}^m\,,$$

its *scalar product* with a real number r is linear operator:

$$rF : \mathbf{R}^n \to \mathbf{R}^m\,,$$

defined by:

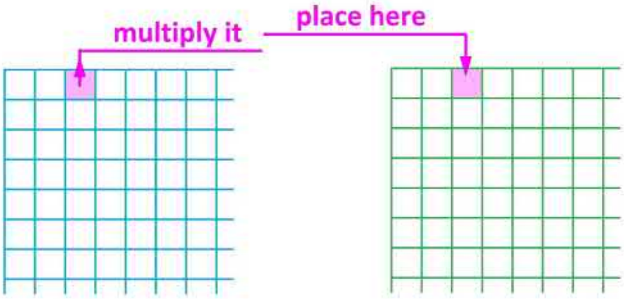
$$(rF)(x) = rF(x)\,.$$



Resize
By: ● Percentage ○ Pixels
Horizontal: 100
Vertical: 200
☐ Maintain aspect ratio

Resize
By: ● Percentage ○ Pixels
Horizontal: 200
Vertical: 100
☐ Maintain aspect ratio

This operation is component-wise again: Every entry is multiplied by the same number.



Definition 5.10.13: scalar multiplication of matrices

Suppose A is an $m \times n$ matrix. Then its *scalar multiple* by a real number r, denoted by:

$$\boxed{rA}$$

is an $m \times n$ matrix, the ij-entry of which is the product of the ij-entry of A by r.

In other words, if $A = a_{ij}$ and $C = rA = c_{ij}$, then

$$c_{ij} = r a_{ij}\,,$$

for each $i = 1, 2, ..., m$ and each $j = 1, 2, ..., n$.

Warning!

The matrix is just an abbreviated representation of a linear operator. Accordingly, the matrix operations are just abbreviated representations of the operations on linear operators.

5.11. *Compositions of linear operators*

Let's take the problem about *mixtures* to the next level.

We have:

1. n ingredients and, therefore, n unknowns or variables $x_1, ..., x_n$ representing the amounts of each; and
2. m requirements or restrictions, i.e., m linear equations involving these variables ($k = 1, 2, ..., m$):

$$a_{k1}x_1 + ... + a_{kn}x_n = b_k\,.$$

This system of linear equations is very cumbersome.

As before, we translate this a system into a vector-matrix equation:

$$\boxed{FX = B}$$

where

1. $X =< x_1, ..., x_n >$ is the vector of the unknowns,
2. $B =< b_1, ..., b_m >$ is the vector of the totals, and
3. $F = a_{ij}$ is the $m \times n$ matrix made up of the coefficients of the system of linear equations.

In light of the recent development, we prefer to look at the equation as the following:

$$\boxed{F(X) = B}$$

i.e., we have a linear operator:

$$F : \mathbf{R}^n \to \mathbf{R}^m .$$

And the equation needs to be solved!

In dimension 1, the equation $kx = b$ is solved by undoing the multiplication by k by division by k:

$$kx = b \implies x = \frac{b}{k} .$$

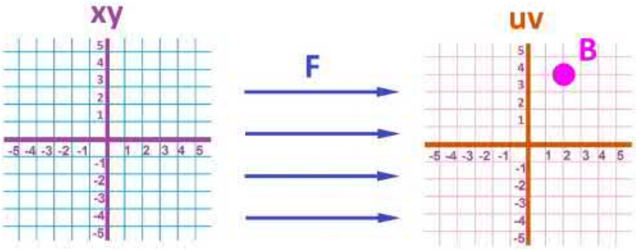
Similarly, we need the *inverse*

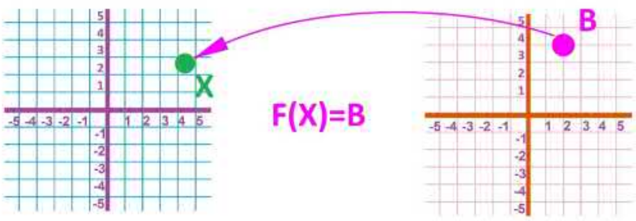
$$F^{-1} : \mathbf{R}^m \to \mathbf{R}^n$$

of F to solve our equation:

$$\boxed{F(X) = B \implies X = F^{-1}(B)}$$

As an illustration, the operator F transforms the xy-plane into the, say, uv-plane:





One particular vector in the uv-plane, B, needs to be traced back to the xy-plane. Of course, if we have F^{-1}, we'll find the counterparts for all B's.

Example 5.11.1: transformations of the plane

We know some of the answers:

1. If F is the uniform stretch by 2, then F^{-1} is the uniform shrink by 2.
2. If F is the stretch by 2 in the direction of a vector $e_1 =< 1, 0 >$, then F^{-1} is the uniform shrink by 2 in the direction of e_1.
3. If F is the rotation by 90 degrees clockwise, then F^{-1} is the rotation by 90 degrees counterclockwise.

4. If F is the flip about the x-axis, then F^{-1} is the flip about the x-axis.

In other words,

1. $F(X)=2X \quad\Longrightarrow\quad F^{-1}(Y)=\frac{1}{2}Y$
2. $F=\begin{bmatrix}2&0\\0&1\end{bmatrix} \quad\Longrightarrow\quad F^{-1}=\begin{bmatrix}\frac{1}{2}&0\\0&1\end{bmatrix}$
3. $F=\begin{bmatrix}0&1\\-1&0\end{bmatrix} \quad\Longrightarrow\quad F^{-1}=\begin{bmatrix}0&-1\\1&0\end{bmatrix}$
4. $F=\begin{bmatrix}1&0\\0&-1\end{bmatrix} \quad\Longrightarrow\quad F^{-1}=\begin{bmatrix}1&0\\0&-1\end{bmatrix}$

Let's recall that the inverse is defined via compositions. It must satisfy:

$$F(F^{-1}(Y))=Y\,,$$

for all Y, and

$$F^{-1}(F(X))=X\,,$$

for all X. In other words,

$$F\circ F^{-1}=I\,,$$

and

$$F^{-1}\circ F=I\,,$$

where I is the identity matrix.

We need to understand compositions better.

We know how to compute compositions of functions. This is the composition:

$$\boxed{\mathbf{R}^n \xrightarrow{F} \mathbf{R}^m \xrightarrow{G} \mathbf{R}^k}$$

It is computed via substitution:

$$(G\circ F)(X)=G(F(X))\,.$$

But what happens to the matrices? How are the matrices of F and G combined to produce the matrix of $G\circ F$? It is called *matrix multiplication*. Here is why.

Example 5.11.2: compositions $\mathbf{R}^1\to\mathbf{R}^1\to\mathbf{R}^1$

Given linear operators (numerical functions):

$$f:\mathbf{R}\to\mathbf{R} \ \text{ defined by } \ f(x)=3x\,,$$

and

$$g:\mathbf{R}\to\mathbf{R} \ \text{ defined by } \ g(y)=2y\,.$$

Their matrices are:

$$f=[3] \quad\text{and}\quad g=[2]\,.$$

What about their composition? The codomain of the former and the domain of the latter match! The composition is computed as follows:

$$g\circ f:\mathbf{R}\to\mathbf{R} \ \text{ defined by } \ (g\circ f)(x)=g(f(x))=2(3x)=6x\,.$$

Its matrix is

$$g \circ f = [6] \,.$$

Of course, this new number is just the *product* of the two original numbers.

Example 5.11.3: compositions $\mathbf{R}^1 \to \mathbf{R}^1 \to \mathbf{R}^2$

Given linear operators (a numerical function and a parametric curve):

$$f : \mathbf{R} \to \mathbf{R} \ \text{ defined by } \ f(x) = 3x \,,$$

and

$$G : \mathbf{R} \to \mathbf{R}^2 \ \text{ defined by } \ G(y) = < 3y, 2y > \ .$$

Their matrices are:

$$f = [3] \quad \text{and} \quad G = \begin{bmatrix} 3 \\ 2 \end{bmatrix} \,.$$

What about their composition? It is an operator (a parametric curve) computed as follows:

$$G \circ f : \mathbf{R} \to \mathbf{R}^2 \,,$$

defined by

$$(G \circ f)(x) = G(f(x)) = < 3(3x), 2(3x) > = < 9x, 6x > \ .$$

Its matrix is

$$G \circ f = \begin{bmatrix} 9 \\ 6 \end{bmatrix} \,.$$

Of course, this is just the product of the two as if the first is a number and the second a vector:

$$Gf = \begin{bmatrix} 3 \\ 2 \end{bmatrix} [3] = \begin{bmatrix} 3 \cdot 3 \\ 2 \cdot 3 \end{bmatrix} = \begin{bmatrix} 9 \\ 6 \end{bmatrix} \,.$$

Example 5.11.4: compositions $\mathbf{R}^1 \to \mathbf{R}^2 \to \mathbf{R}^1$

Given linear operators (a parametric curve and a function of two variables):

$$F : \mathbf{R} \to \mathbf{R}^2 \ \text{ defined by } \ F(x) = < 3x, 2x > \,,$$

and

$$g : \mathbf{R}^2 \to \mathbf{R} \ \text{ defined by } \ g(u, v) = 5u - 2v \,.$$

Their matrices are:

$$F = \begin{bmatrix} 3 \\ 2 \end{bmatrix} \quad \text{and} \quad g = [5, \ -2] \,.$$

Their sum is a numerical function and it is computed as follows:

$$g \circ F : \mathbf{R} \to \mathbf{R} \,,$$

defined by

$$(g \circ F)(x) = g(F(x)) = 5(3x) - 2(2x) = 11x \,.$$

Its matrix is

$$g \circ F = [11] \,.$$

It's the dot product of the two vector-like matrices:

$$gF = [5,\ -2]\begin{bmatrix}3\\2\end{bmatrix} = [5\cdot 3 + (-2)\cdot 2] = [11]\,.$$

Example 5.11.5: compositions $\mathbf{R}^2 \to \mathbf{R}^2 \to \mathbf{R}^2$

Given linear operators (transformations of the planes):

$$F, G : \mathbf{R}^2 \to \mathbf{R}^2$$

defined by

$$F(x,y) =< u,v >=< 3x+2y, 5x-y > \quad \text{and} \quad G(u,v) =< 5u+v, u+v > \ .$$

Their matrices are:

$$F = \begin{bmatrix}3 & 2\\5 & -1\end{bmatrix} \quad \text{and} \quad G = \begin{bmatrix}5 & 1\\1 & 1\end{bmatrix}\ .$$

Their composition is an operator computed by substitution:

$$G\circ F : \mathbf{R}^2 \to \mathbf{R}^2\,,$$

defined by

$$(G\circ F)(x,y) =< 5(3x+2y)+(5x-y), (3x+2y)+(5x-y) >=< 20x+9y, 8x+y > \ .$$

Its matrix is

$$G\circ F = \begin{bmatrix}20 & 9\\8 & 1\end{bmatrix}\ .$$

It is seen as computed via four dot products of the four pairs of rows (from the first matrix) and columns (from the second):

$$\begin{bmatrix}5 & 1\\1 & 1\end{bmatrix}\cdot\begin{bmatrix}3 & 2\\5 & -1\end{bmatrix} \to\ 5\cdot 3 + 1\cdot 5 = 20 \quad \to \begin{bmatrix}20 & 9\\8 & 1\end{bmatrix}$$

$$\begin{bmatrix}5 & 1\\1 & 1\end{bmatrix}\cdot\begin{bmatrix}3 & 2\\5 & -1\end{bmatrix} \to\ 5\cdot 2 + 1\cdot (-1) = 9 \to \begin{bmatrix}10 & 9\\8 & 1\end{bmatrix}$$

$$\begin{bmatrix}5 & 1\\1 & 1\end{bmatrix}\cdot\begin{bmatrix}3 & 2\\5 & -1\end{bmatrix} \to\ 1\cdot 3 + 1\cdot 5 = 8 \quad \to \begin{bmatrix}10 & 9\\8 & 1\end{bmatrix}$$

$$\begin{bmatrix}5 & 1\\1 & 1\end{bmatrix}\cdot\begin{bmatrix}3 & 2\\5 & -1\end{bmatrix} \to\ 1\cdot 2 + 1\cdot (-1) = 1 \to \begin{bmatrix}10 & 9\\8 & 1\end{bmatrix}$$

In general, we have a single formula:

$$\boxed{\begin{bmatrix} a & b \\ c & d \end{bmatrix} \cdot \begin{bmatrix} e & f \\ g & h \end{bmatrix} = \begin{bmatrix} ae+bg & af+bh \\ ce+dg & cf+dh \end{bmatrix}}$$

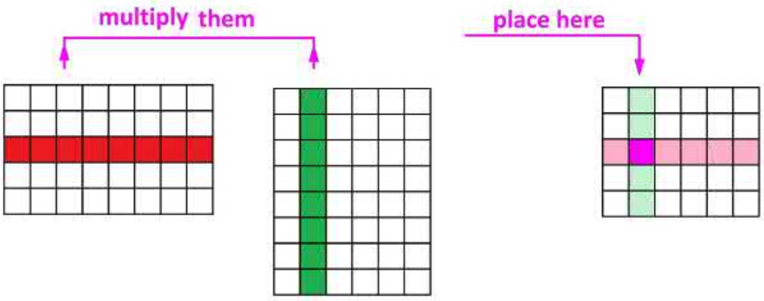
Warning!

As the matrix multiplication is just an abbreviated representation of the composition of linear operators, it is secondary to the substitution of the formulas it comes from.

Now we consider linear operators in arbitrary dimensions:

$$\mathbf{R}^n \to \mathbf{R}^m \to \mathbf{R}^k\,.$$

For their matrices, the length of the column in the first must be equal to the length of the row in the second. That's m! Then, we can carry out a dot product:



We do this nk times and produce an $n \times k$ matrix.

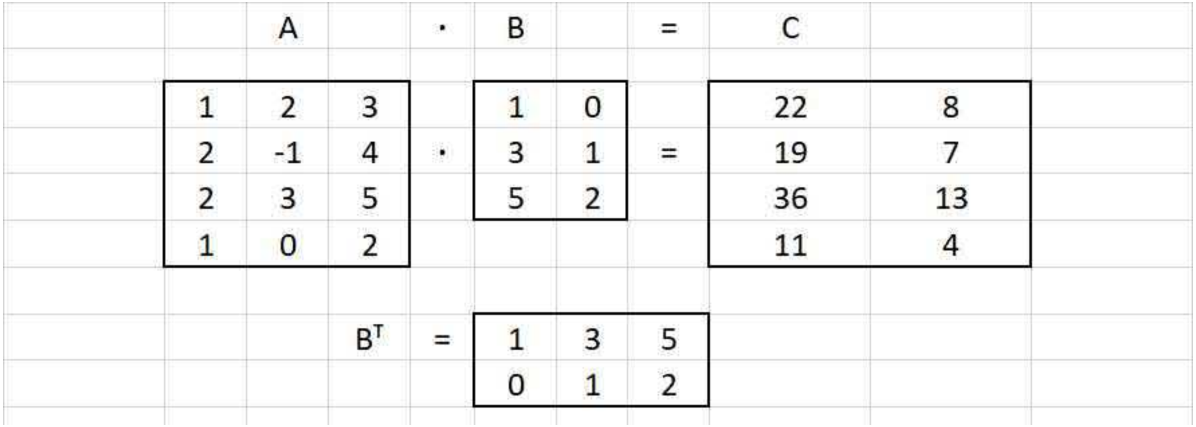
Exercise 5.11.6

Multiply:

$$\begin{bmatrix} 1 & 2 & 3 \\ 4 & 5 & 6 \\ 7 & 8 & 9 \end{bmatrix} \cdot \begin{bmatrix} 1 & -1 \\ -1 & 1 \\ 1 & -1 \end{bmatrix}.$$

Example 5.11.7: spreadsheet

One can utilize a spreadsheet and other software to multiplication for matrices of any dimensions. In order to make this work, the second matrix B has to be "transposed" (bottom):



	A		·	B		=	C	
1	2	3		1	0		22	8
2	-1	4	·	3	1	=	19	7
2	3	5		5	2		36	13
1	0	2					11	4
		B^T	=	1	3	5		
				0	1	2		

This is the code for the transpose of X:

```
=TRANSPOSE(R[-5]C:R[-3]C[1])
```

This is the code for Y:

```
=SUMPRODUCT(RC2:RC4,R8C[-3]:R8C[-1])
```

Finding the inverse of a matrix, especially of high dimension, is a challenging problem. There is a simple formula for the 2×2 matrices:

Theorem 5.11.8: Inverse of 2×2 Matrix

The inverse of an invertible matrix A is computed as follows:

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

Exercise 5.11.9

Prove the theorem.

Exercise 5.11.10

Use the theorem to solve the following problem: A tourist group took a train trip at \$3 per child and \$3.20 per adult for a total of \$118.40. They took the train back at \$3.50 per child and \$3.60 per adult for a total of \$135.20. How many children, and how many adults?

Exercise 5.11.11

Use the theorem to solve the following problem: A tourist group with 10 children and 20 adults took a train trip for a total of \$110. Another tourist group with 15 children and 25 adults took a train trip for a total of \$145. What were the ticket prices?

To summarize, we go back to the general setup of linear operators applied consecutively:

$$\begin{tikzcd} \mathbf{R}^n \arrow[r, "F"] & \mathbf{R}^m \arrow[r, "G"] & \mathbf{R}^k \end{tikzcd}$$

We have implicitly used the following fact:

Theorem 5.11.12: Composition of Linear Operators

The composition of two linear operators is a linear operator.

Exercise 5.11.13

Prove the theorem.

We have also implicitly used the the following important result:

Theorem 5.11.14: Matrix of Composition

The product of two matrices that represent two linear operators is the matrix of their composition.

Chapter 6: A bird's-eye view of basic calculus

Contents

6.1. What is calculus about?

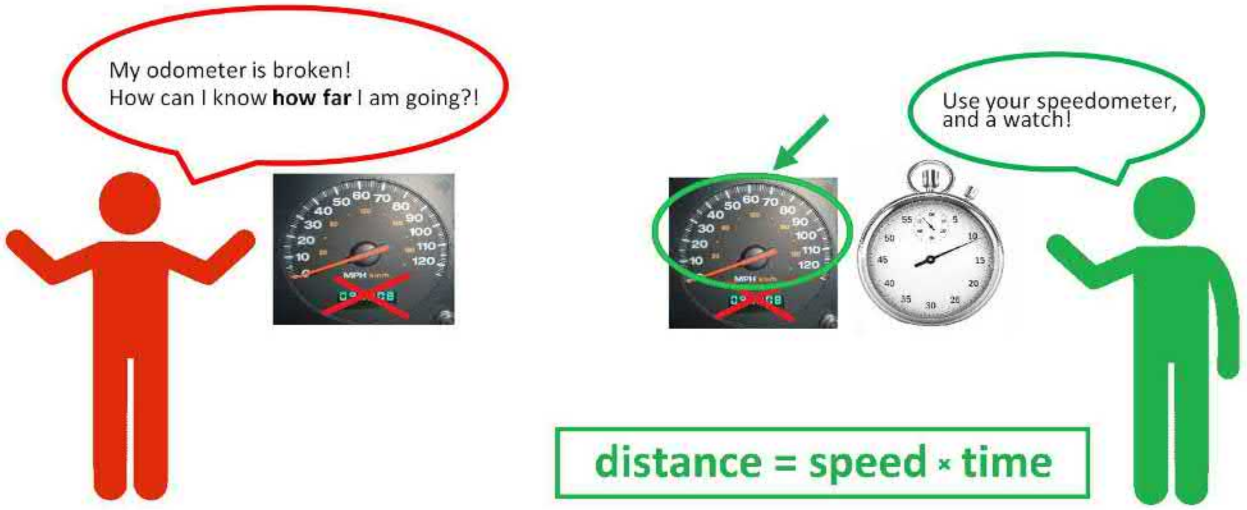
One of the main entry ways to calculus is the study of *motion*.

We present the idea of calculus in these two related pictures.

First, we derive the speed from the distance that we have covered:

Beyond this conceivable situation, this formula is the *definition of speed*.

On the flip side, we derive the distance we have covered from the known velocity:



The two problems are solved, respectively, with the help of these two versions of the same elementary school formula:

$$\boxed{\text{speed} \ = \ \text{distance} \ / \ \text{time} \qquad \text{and} \qquad \text{distance} \ = \ \text{speed} \ \times \ \text{time}}$$

We solve the equation for the distance or for the speed depending on what is known and what is unknown.

What takes this idea beyond elementary school is the possibility that *velocity varies* over time.

We first take the simplest model of motion: *The location changes incrementally.*

Let's be more specific. We will face the two situations above but with more data collected and more information derived from it.

First, imagine that our speedometer is broken. What do we do if we want to estimate how fast we are driving during our trip? We look at the odometer *several* times – say, every hour on the hour – during the trip and record the mileage on a piece of paper. The list of our consecutive *locations* might look like this:

1. initial reading: $10,000$ miles
2. after the first hour: $10,055$ miles
3. after the second hour: $10,095$ miles
4. after the third hour: $10,155$ miles
5. etc.

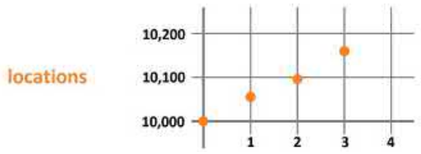
This is a *sequence* with 4 terms:

$$10,000, \ 10,055, \ 10,095, \ 10,155\,.$$

This is also a *vector* in $\mathbf{R}^4$:

$$< 10,000, \ 10,055, \ 10,095, \ 10,155 > \ .$$

We can plot – as an illustration – the locations against time:



But what do we know about what the speed has been? We write a quick formula:

$$\boxed{\text{speed} \ = \frac{\text{distance}}{\text{time}} = \frac{\text{current location} \ - \ \text{last location}}{1}}$$

The time interval was chosen to be 1 hour, so all we need is to find the distance covered during each of these one-hour periods, by *subtraction*:

1. distance covered during the first hour: $10,055 - 10,000 = 55$ miles; speed 55 miles an hour
2. distance covered during the second hour: $10,095 - 10,055 = 40$ miles; speed 40 miles an hour
3. distance covered during the third hour: $10,155 - 10,095 = 60$ miles; speed 60 miles an hour
4. etc.

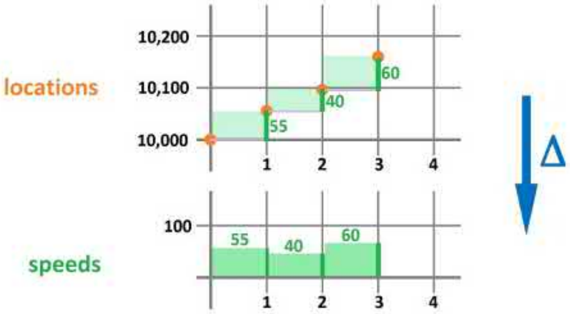
This is a new sequence with 3 terms:

$$55,\ 40,\ 60\,.$$

Or a new vector in $\mathbf{R}^3$:

$$<55,\ 40,\ 60>\ .$$

We see below how these new numbers appear as the blocks that make up each step of our last plot (top):



We then lower these blocks to the bottom to create a new plot (bottom).

As you can see, we illustrate the new data in such a way as to suggest that the speed remains *constant* during each of these hour-long periods.

The problem is solved! We have established that the speed has been – roughly – 55, 40, and 60 miles an hour during those three time intervals, respectively.

It appears that we have a function (a linear operator?):

$$\Delta : \mathbf{R}^4 \to \mathbf{R}^3\,.$$

Now on the flip side: Imagine this time that it is the odometer that is broken. If we want to estimate how far we will have gone, we should look at the speedometer *several* times – say, every hour – during the trip and record its readings on a piece of paper. The result is a sequence of numbers that may look like this:

1. during the first hour: 35 miles an hour
2. during the second hour: 65 miles an hour
3. during the third hour: 50 miles an hour
4. etc.

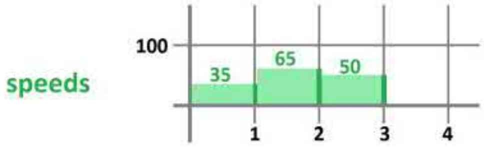
This is a sequence with 3 terms:

$$35,\ 65,\ 50\,.$$

This is also a vector in $\mathbf{R}^3$:

$$<35,\ 65,\ 50>\ .$$

Let's plot our speed against time to visualize what has happened:



Once again, we illustrate the data in such a way as to suggest that the speed remains *constant* during each of these hour-long periods.

Now, what does this tell us about our location? We write a quick formula:

$$\boxed{\text{distance} \ = \ \text{speed} \ \times \ \text{time} \ = \ \text{speed} \ \times 1}$$

In contrast to the former problem, we need another bit of information. We must know the *starting point* of our trip, say, the 100-mile mark. The time interval was chosen to be 1 hour so that we need only to *add*, and keep adding, the speed at which – we assume – we drove during each of these one-hour periods:

1. the location after the first hour: $100 + 35 = 135$-mile mark
2. the location after the two hours: $135 + 65 = 200$-mile mark
3. the location after the three hours: $200 + 50 = 250$-mile mark
4. etc.

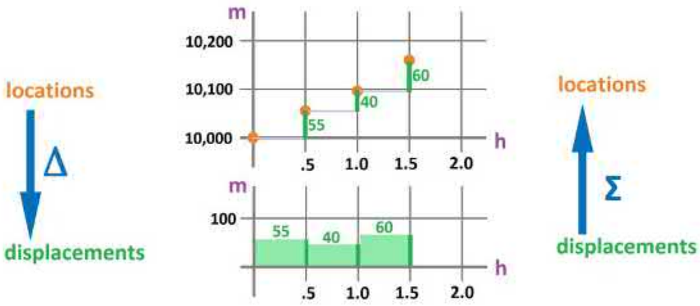
We have a new sequence with 4 terms:

$$100,\ 135,\ 200,\ 250\,.$$

This is also a vector in $\mathbf{R}^4$:

$$< 100,\ 135,\ 200,\ 250 > \ .$$

In order to illustrate this algebra, we plot the speeds as these blocks (top):



Then we use these blocks to make the consecutive steps of the staircase to show how high we have to climb (bottom).

It appears that we have a function (a linear operator?):

$$\Sigma : \mathbf{R}^3 \to \mathbf{R}^4\,.$$

The problem is solved! We have established that we have progressed through the roughly 135-, 200-, and 250-mile marks during this time.

Warning!

An actual speedometer is likely to use the distance covered (computed from the number of revolutions of the wheel) to find the velocity.

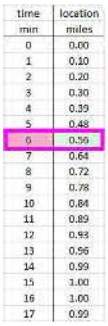
We next consider more complex examples of the relation between location and velocity.

First, *from location to velocity.*

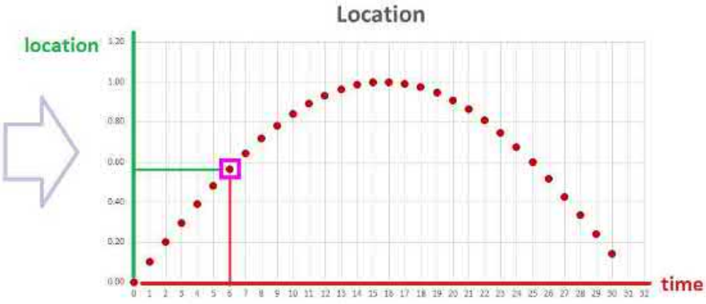
Suppose that this time we have a *sequence* of more than 30 data points (more is indicated by "..."); they are the locations of a moving object recorded every minute:

time	minutes	0	1	2	3	4	5	6	7	8	9	10	...
location	miles	0.00	0.10	0.20	0.30	0.39	0.48	0.56	0.64	0.72	0.78	0.84	...

This data is also seen in the first two columns of the spreadsheet (left):

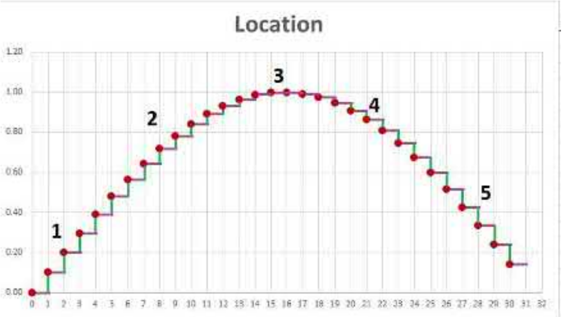


time	location
min	miles
0	0.00
1	0.10
2	0.20
3	0.30
4	0.39
5	0.48
6	0.56
7	0.64
8	0.72
9	0.78
10	0.84
11	0.89
12	0.93
13	0.96
14	0.99
15	1.00
16	1.00
17	0.99



Every pair of numbers in the table is then plotted (right). The "scatter plot" that illustrates the data looks like a *curve*! We will however, continue to treat it as a vector. It lives in $\mathbf{R}^{30}$.

What has happened to the moving object can now be read from the graph. Just as in the last example, we concentrate on the vertical increment of the staircase:



These are the results:

1. The object was moving in the positive direction.
2. It was moving fairly fast but then started to slow down.
3. It stopped for a very short period.
4. Then it started to move in the opposite direction.
5. Then it started to speed up in that direction.

To understand how fast we move over these one-minute intervals, we compute the *differences* of locations for each pair of consecutive locations.

First, the table.

We use the data from the row of locations and subtract every two consecutive locations. This is how the first step is carried out:

time	min	0	1	...
location	miles	0.00	0.10	...
		$\searrow$	$\downarrow$	
difference			$0.10 - 0.00$	...
			$\vert\vert$	
velocity	miles/min		0.10	...

We compute this difference for each pair of consecutive locations and then place it in a row for the velocities that we added to the bottom of our table:

time	min	0	1	2	3	4	5	6	7	8	9	…
location	miles	0.00	0.10	0.20	0.30	0.39	0.48	0.56	0.64	0.72	0.78	…
			↘ ↓	↘ ↓	↘ ↓	↘ ↓	↘ ↓	↘ ↓	↘ ↓	↘ ↓	↘ ↓	…
velocity	miles/min		0.10	0.10	0.10	0.09	0.09	0.09	0.08	0.07	0.07	…

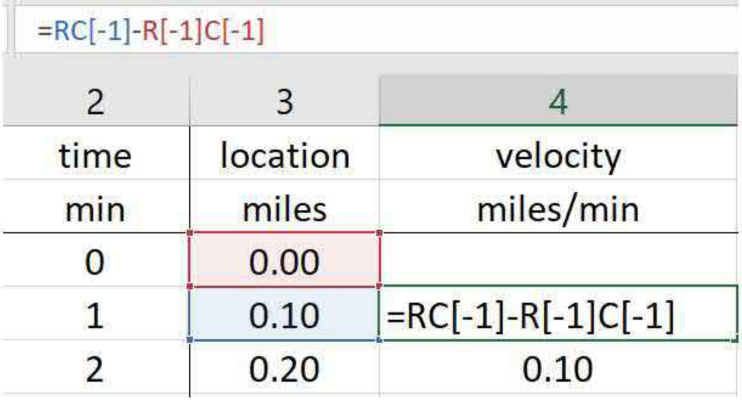
We have a new sequence!

Practically, we'd rather use the computing capabilities of the spreadsheet.

We compute the differences by pulling data from the column of locations with the following formula:

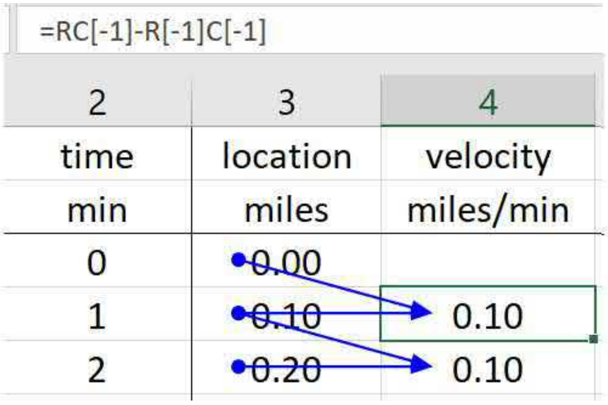
`=RC[-1]-R[-1]C[-1]`

Here, the two values come from the last column, `C[-1]`, same row, `R`, and last row, `R[-1]`. Below, you can see the two references in the formulas marked with red and blue (left) and the dependence shown with the arrows (right):



=RC[-1]-R[-1]C[-1]

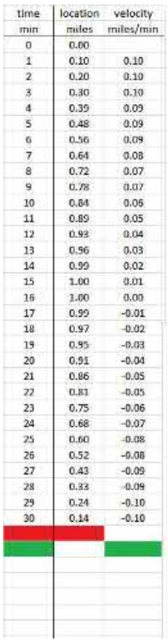
2	3	4
time	location	velocity
min	miles	miles/min
0	0.00	
1	0.10	=RC[-1]-R[-1]C[-1]
2	0.20	0.10



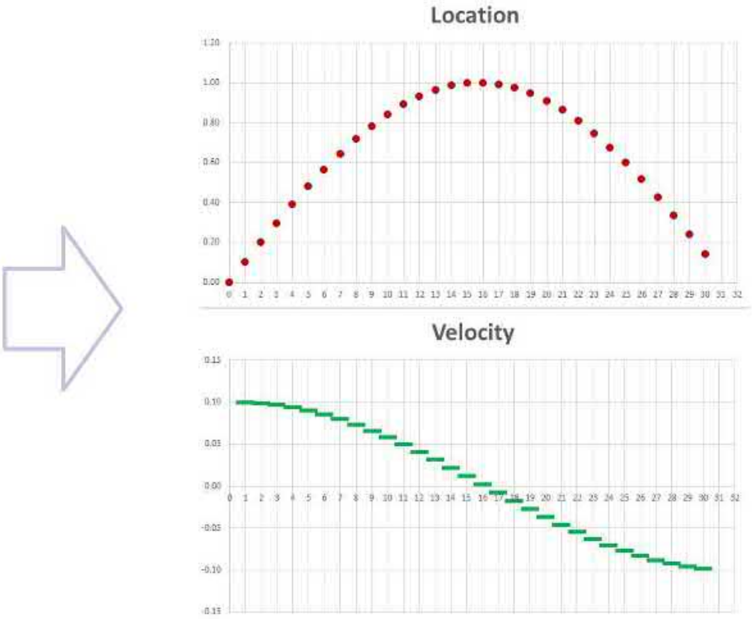
=RC[-1]-R[-1]C[-1]

2	3	4
time	location	velocity
min	miles	miles/min
0	0.00	
1	0.10	0.10
2	0.20	0.10

We place the result in a new column we created for the velocities:



time	location	velocity
min	miles	miles/min
0	0.00	
1	0.10	0.10
2	0.20	0.10
3	0.30	0.10
4	0.39	0.09
5	0.48	0.09
6	0.56	0.09
7	0.64	0.08
8	0.72	0.07
9	0.78	0.07
10	0.84	0.06
11	0.89	0.05
12	0.93	0.04
13	0.96	0.03
14	0.99	0.02
15	1.00	0.01
16	1.00	0.00
17	0.99	-0.01
18	0.97	-0.02
19	0.95	-0.03
20	0.91	-0.04
21	0.86	-0.05
22	0.81	-0.05
23	0.75	-0.06
24	0.68	-0.07
25	0.60	-0.08
26	0.52	-0.08
27	0.43	-0.09
28	0.33	-0.09
29	0.24	-0.10
30	0.14	-0.10
🟥	🟥	
🟩		🟩



This new data is illustrated with the second scatter plot. We can see that the time column is shared and, therefore, the time axis is the same in the two plots. To emphasize the fact that the velocity data, unlike the location, is referring to time intervals rather than time instances, we plot it with horizontal segments. In fact, the data table can be rearranged as follows to make this point clearer:

time	0		1		2		3		4		…
location	0.00	–	0.10	–	0.20	–	0.30	–	.39	–	…
velocity	·	0.10	·	0.10	·	0.10	·	0.09	·	0.09	…

This is a a vector that lives in $\mathbf{R}^{29}$.

What has happened to the moving object can now be easily read from the second graph. These are the five stages:

1. The velocity was positive initially (it was moving in the positive direction).
2. The velocity was fairly high (it was moving fairly fast) but then it started to decline (slow down).
3. The velocity was zero (it stopped) for a very short period.
4. Then the velocity became negative (it started to move in the opposite direction).
5. And then the velocity started to become more negative (it started to speed up in that direction).

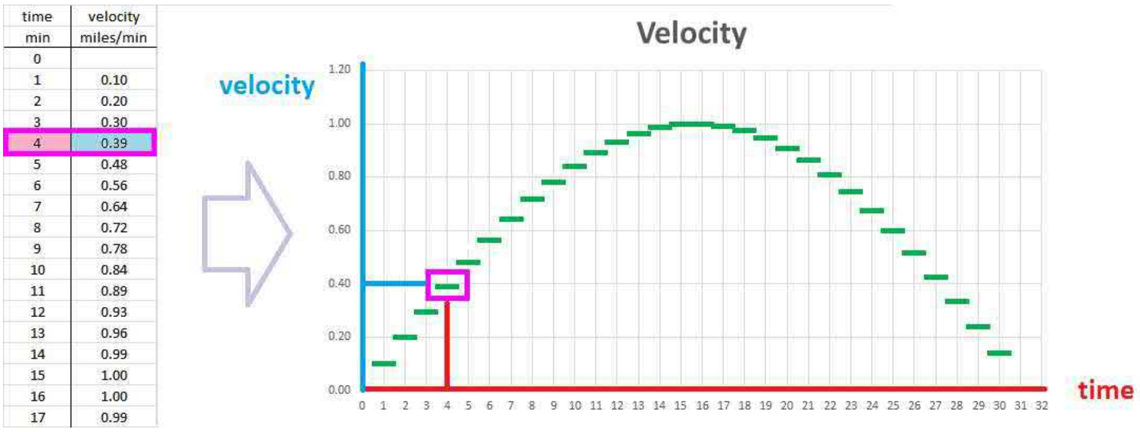
Thus, the latter set of data succinctly records some important facts about the dynamics of the former.

Now, *from velocity to location.*

Again, we consider a sequence of 30 data points. These numbers are the values of the velocity of an object recorded every minute (a vector in $\mathbf{R}^{30}$):

time	minutes	0	1	2	3	4	5	6	7	8	9	10	...
velocity	miles/hour		0.10	0.20	0.30	0.39	0.48	0.56	0.64	0.72	0.78	0.84	...

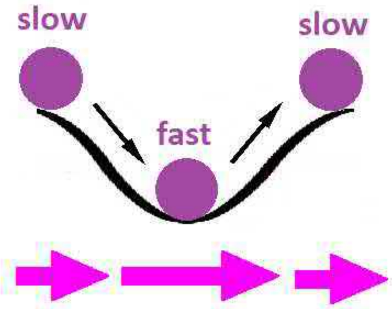
This data is also seen in the first two columns of the spreadsheet plotted one bar at a time:



time min	velocity miles/min
0	
1	0.10
2	0.20
3	0.30
4	0.39
5	0.48
6	0.56
7	0.64
8	0.72
9	0.78
10	0.84
11	0.89
12	0.93
13	0.96
14	0.99
15	1.00
16	1.00
17	0.99

The data is furthermore illustrated as a scatter plot on the right. Again, we emphasize the fact that the velocity data is referring to time intervals by plotting its values with horizontal bars.

The data may be describing the horizontal speed of a ball rolling through a trough:



To find out where we are at the end of each of these one-minute intervals, we compute by *adding* the velocities one at a time. This is how the first step is carried out, under the assumption that the initial location is 0:

time	min	0	1	...
velocity	miles		0.10	...
			↓	
sum		0.00+	0.10	...
		↑	‖	
location	miles/min	0.00	0.10	...

We place this data in a new row added to the bottom of our table:

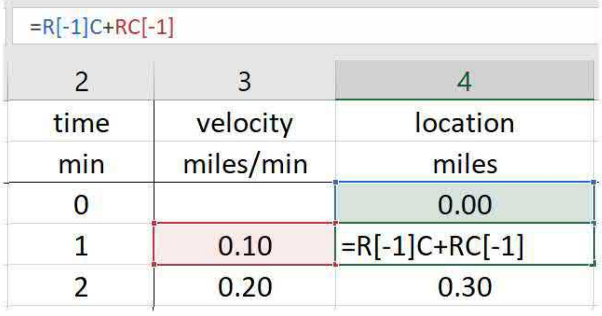
time	min	0	1	2	3	4	5	6	7	8	...
velocity	miles		0.10	0.20	0.30	0.39	0.48	0.56	0.64	0.72	...
			↓	↓	↓	↓	↓	↓	↓	↓	...
location	miles/min	0.00 →	0.10 →	0.30 →	0.59 →	0.98 →	1.46 →	2.03 →	2.67 →	3.39 →	...

We have a new sequence and a vector that lives in $\mathbf{R}^{31}$.

Practically, we use the spreadsheet. We compute the sums by pulling the data from the column of velocities using the following formula:

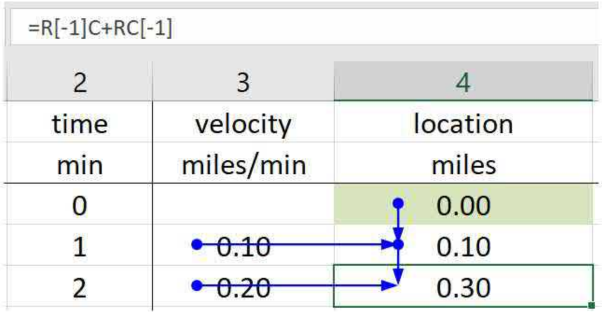
```
=R[-1]C+RC[-1]
```

Here, the two values come from the same, `C`, or last, `C[-1]`, column and the same, `R`, and last, `R[-1]`, row, as follows:



=R[-1]C+RC[-1]

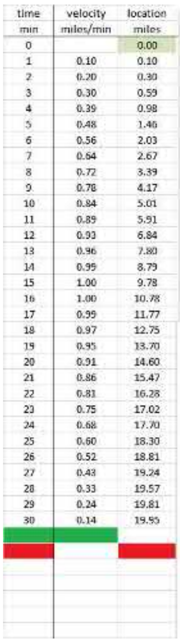
2	3	4
time	velocity	location
min	miles/min	miles
0		0.00
1	0.10	=R[-1]C+RC[-1]
2	0.20	0.30



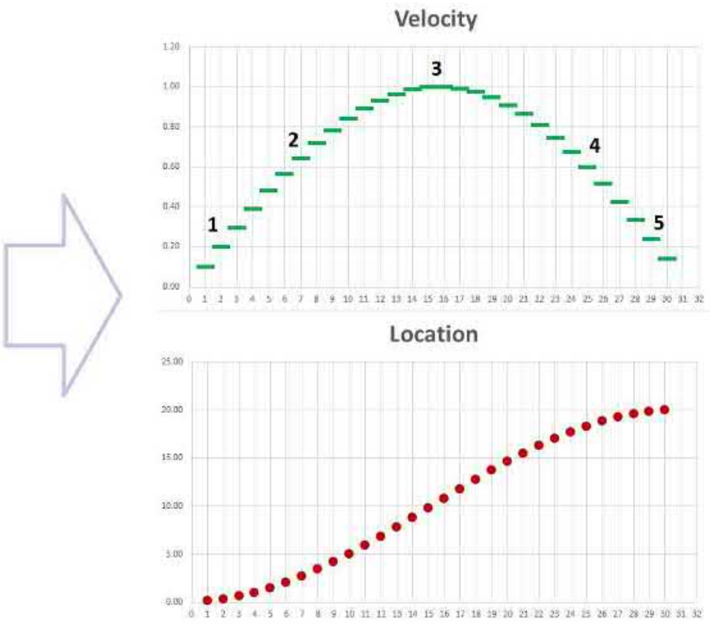
=R[-1]C+RC[-1]

2	3	4
time	velocity	location
min	miles/min	miles
0		0.00
1	0.10	0.10
2	0.20	0.30

We place the result in a new column for locations:



time	velocity	location
min	miles/min	miles
0		0.00
1	0.10	0.10
2	0.20	0.30
3	0.30	0.59
4	0.39	0.98
5	0.48	1.46
6	0.56	2.03
7	0.64	2.67
8	0.72	3.39
9	0.78	4.17
10	0.84	5.01
11	0.89	5.91
12	0.93	6.84
13	0.96	7.80
14	0.99	8.79
15	1.00	9.78
16	1.00	10.78
17	0.99	11.77
18	0.97	12.75
19	0.95	13.70
20	0.91	14.60
21	0.86	15.47
22	0.81	16.28
23	0.75	17.02
24	0.68	17.70
25	0.60	18.30
26	0.52	18.81
27	0.43	19.24
28	0.33	19.57
29	0.24	19.81
30	0.14	19.95



The data is also illustrated as the second scatter plot on the right.

What has happened to the moving object can now be easily read from the first or the second plot. These are the five stages:

1. The velocity is positive and low.
2. The velocity is positive and high.
3. The velocity is the highest.
4. The velocity is positive and high.
5. The velocity is positive and low.

We, again, rearrange the data table to make the difference between the two types of data clearer:

time	0		1		2		3		4		...
velocity	·	0.00	·	0.10	·	0.20	·	0.30	·	0.39	...
location	0.00	–	0.10	–	0.30	–	0.59	–	.98	–	...

Thus, as the former data set records some facts about the dynamics of the latter, we are able to use this information to recover the latter.

One of the easier conclusions we derive from this analysis is the following simple statement:

- ▶ With a positive velocity, we are moving forward.

It takes the following abbreviated form:

- ▶ The velocity is positive $\Longrightarrow$ the motion is in the positive direction.

Now, we can try to "flip" the implication of this statement, without assuming that the result will be true:

- ▶ The velocity is positive $\Longleftarrow$ the motion is in the positive direction.

In other words, we have the following implication:

- ▶ The motion is in the positive direction $\Longrightarrow$ the velocity is positive.

The latter is called the *converse* of the original statement. It's also an implication stated as follows:

- ▶ `IF` the motion is in the positive direction, `THEN` the velocity is positive.

The converse is true as well!

In our case, the implications go both ways! Combined, the statement and its converse form an *equivalence*:

- ▶ The velocity is positive `IF AND ONLY IF` the motion is in the positive direction.

Our statement is written as follows:

- ▶ The velocity is positive $\Longleftrightarrow$ the motion is in the positive direction.

The two parts of an equivalence are interchangeable!

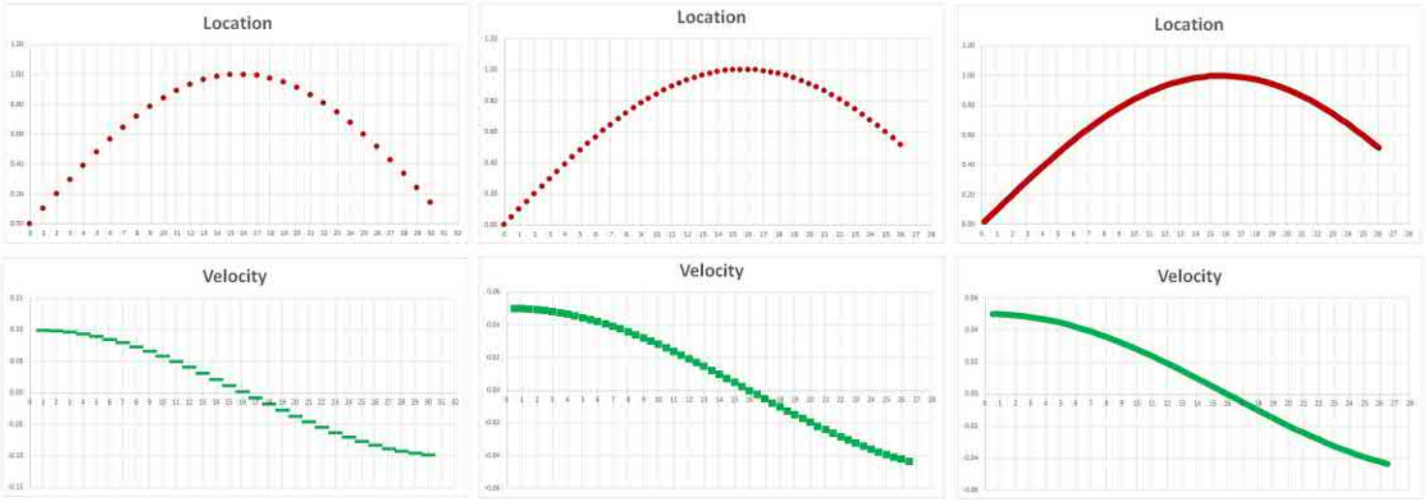
So, this is what we have discovered:

- ▶ *We can tell the velocity from the location and, conversely, the location from the velocity.*

Is this it though?

Motion is a *continuous* phenomenon. Can we understand it following the above approach?

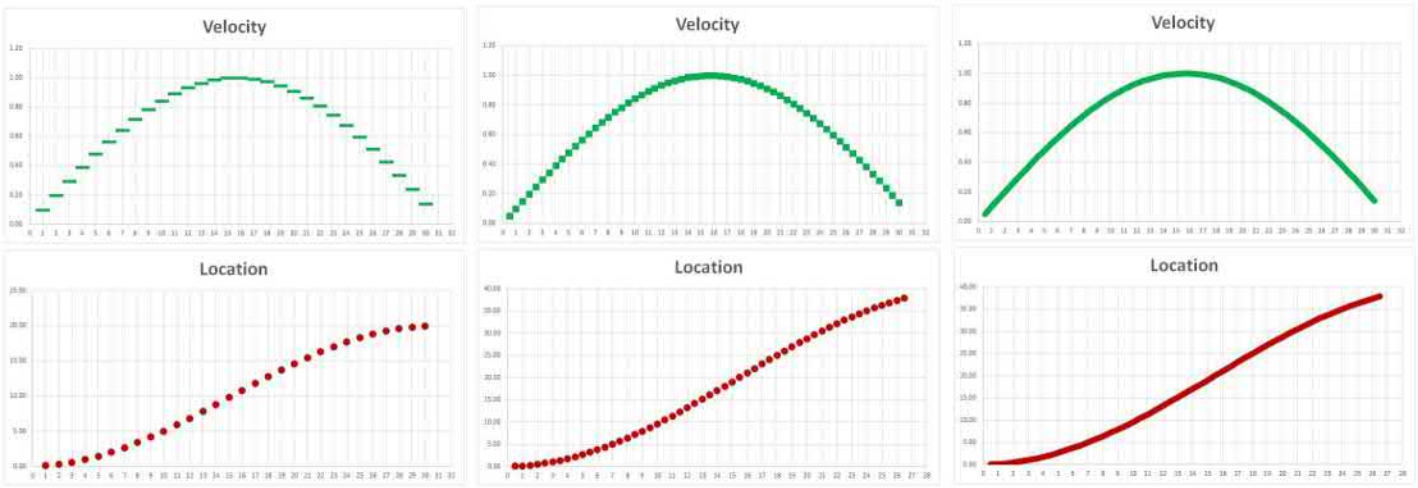
If it is known that our data is just a snapshot of a "continuous" process, we may be able to collect more information in order to make this representation better. We, for example, may look at the odometer every minute, or every second, etc., instead of every hour. The infinite divisibility of the real line allows us to produce sets of points on the plane with denser and denser patterns: We make the time intervals smaller and smaller and insert more and more inputs. When there are enough of them, the points start to form a curve:



Both location and velocity are changing continuously!

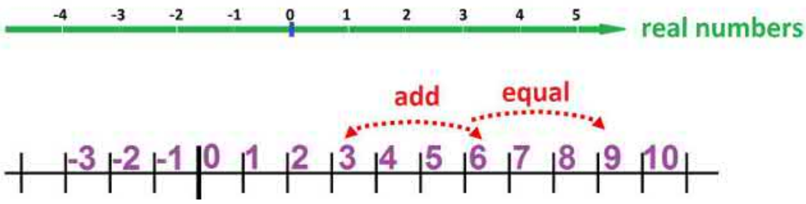
We imagine that at the end of this process we will have an actual curve. This is not the kind of curve that is made of marbles placed close together, but a rope. What happens "at the end" is studied in calculus.

Thus, this main idea of calculus is to *derive* velocity from location (from the top row to the bottom) and location from velocity (from the bottom row to the top):



6.2. Spaces of functions

In Chapter 2 we chose to treat *all numerical functions as a single group*. The inspiration came from how we have handled the *real numbers*. We put them together in the real number line, which provides us with a bird's-eye view:

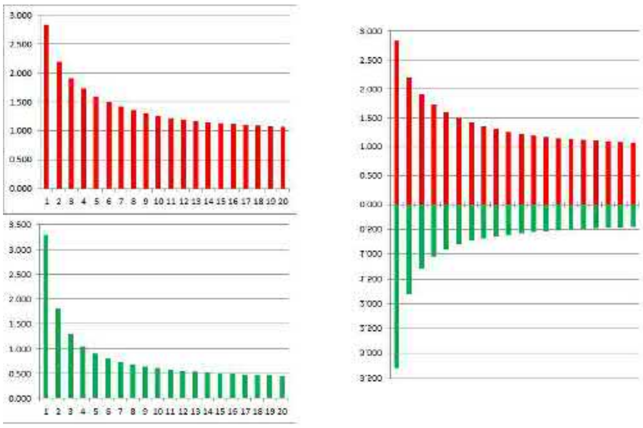


Example 6.2.1: finite sequences

We can add sequences term-wise as long as they have to have the same length, n:

$$\begin{array}{rcccc} & a_1 & a_2 & \dots & a_n \\ + & & & & \\ & b_1 & b_2 & \dots & b_n \\ \hline & a_1+b_1 & a_2+b_2 & \dots & a_n+b_n \end{array} \quad \to \quad \begin{array}{lccccc} < & a_1 & a_2 & \dots & a_n & > \\ + & & & & & \\ < & b_1 & b_2 & \dots & b_n & > \\ \hline < & a_1+b_1 & a_2+b_2 & \dots & a_n+b_n & > \end{array}$$

It's the same algebra illustrated below for $\mathbf{R}^{20}$:



So, the set of sequences with n terms is just a copy of $\mathbf{R}^n$! In what sense? There is a very simple function from the left part to the right:

$$a_1 \quad a_2 \quad \dots \quad a_n \quad \mapsto \quad < \ a_1 \quad a_2 \quad \dots \quad a_n \ >$$

But the algebra matches too: According to the above diagram, the "addition is preserved". And so is scalar multiplication. So, this function is linear! And it's one-to-one and onto.

Exercise 6.2.2

Sequences are just functions. Prove the above statements about function.

Exercise 6.2.3

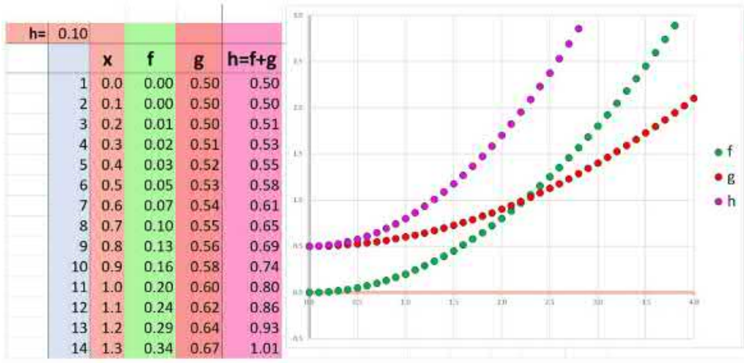
Is the function

$$< a_1, a_2, ..., a_n > \mapsto \ \max\{a_1, a_2, ..., a_n\}$$

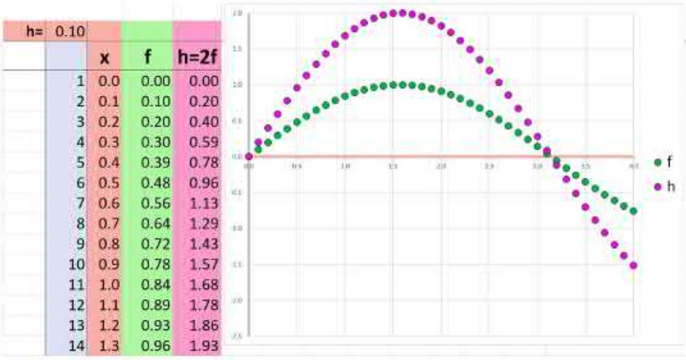
a linear operator?

What kind of "group" is the set of all *numerical functions* $f : \mathbf{R}^1 \to \mathbf{R}^1$? A vector space.

What are its operations? Addition and scalar multiplication, of course:



h=	0.10				
		x	f	g	h=f+g
	1	0.0	0.00	0.50	0.50
	2	0.1	0.00	0.50	0.50
	3	0.2	0.01	0.50	0.51
	4	0.3	0.02	0.51	0.53
	5	0.4	0.03	0.52	0.55
	6	0.5	0.05	0.53	0.58
	7	0.6	0.07	0.54	0.61
	8	0.7	0.10	0.55	0.65
	9	0.8	0.13	0.56	0.69
	10	0.9	0.16	0.58	0.74
	11	1.0	0.20	0.60	0.80
	12	1.1	0.24	0.62	0.86
	13	1.2	0.29	0.64	0.93
	14	1.3	0.34	0.67	1.01



h=	0.10			
		x	f	h=2f
	1	0.0	0.00	0.00
	2	0.1	0.10	0.20
	3	0.2	0.20	0.40
	4	0.3	0.30	0.59
	5	0.4	0.39	0.78
	6	0.5	0.48	0.96
	7	0.6	0.56	1.13
	8	0.7	0.64	1.29
	9	0.8	0.72	1.43
	10	0.9	0.78	1.57
	11	1.0	0.84	1.68
	12	1.1	0.89	1.78
	13	1.2	0.93	1.86
	14	1.3	0.96	1.93

The new functions are constructed from the old:

Definition 6.2.4: vector operations on functions

1. Given two numerical f and g defined on the same set, the *sum*, $f+g$, of f and g is the function defined by the following:

$$(f+g)(x)=f(x)+g(x)\ \texttt{ FOR EACH }\ x\,.$$

2. Given a numerical function f, the *constant multiple* cf of f, for some real number c, is the function defined by the following:

$$(cf)(x)=cf(x)\ \texttt{ FOR EACH }\ x\,.$$

The algebraic operations on the right-hand sides of the two definitions are nothing but operations on *numbers*. These operations satisfy the simple algebraic properties. Therefore, so are the function operations.

Let's recall how this idea is codified:

Definition 6.2.5: vector space

A set with two operations – addition of two vectors and multiplication of a vector by a scalar – in that satisfy the properties below is called a *vector space*:

1. $X+Y=Y+X$ for all X and Y.
2. $X+(Y+Z)=(X+Y)+Z$ for all X, Y, and Z.
3. $X+0=X=0+X$ for some vector 0 and all X.
4. $X+(-X)=0$ for any X and some vector $-X$.
5. $a(bX)=(ab)X$ for all X and all scalars a,b.
6. $1X=X$ for all X.
7. $a(X+Y)=aX+aY$ for all X and Y.
8. $(a+b)X=aX+bX$ for all X and all scalars a,b.

The argument presented above suggests the following:

Theorem 6.2.6: Linear Algebra of Functions

The numerical functions defined on a fixed set $D\subset\mathbf{R}$,

$$\{f:D\to\mathbf{R}\}\,,$$

form a vector space.

Example 6.2.7: infinite sequences

If we move from finite to infinite sequences, they are numerical functions too; the domain is $D=\mathbf{N}$. We add them this termwise as if we adding vectors:

$$\begin{array}{rlllll} & a_1 & a_2 & \dots & a_n & \dots\\ + & & & & & \\ & b_1 & b_2 & \dots & b_n & \dots\\ \hline & a_1+b_1 & a_2+b_2 & \dots & a_n+b_n & \dots\end{array}$$

There is no match with some $\mathbf{R}^n$ anymore:

$$a_1\quad a_2\quad \dots\quad a_n\quad \dots\qquad \xrightarrow{?}\qquad <\ a_1\quad a_2\quad \dots\quad a_n\quad \dots >$$

Exercise 6.2.8

Suppose B is the the set of all sequences with a single non-zero value, 1. Can you express every sequence as a linear combination of elements of B?

Example 6.2.9: vector space of linear operators

This is the algebra we saw in the last chapter. Two operations:

1. Given two linear operators:
$$F,G:\mathbf{R}^n\to\mathbf{R}^m\,,$$
their *sum* is linear operator:
$$F+G:\mathbf{R}^n\to\mathbf{R}^m\,,$$
defined by:
$$(F+G)(x)=F(x)+G(x)\,.$$
2. Given a linear operator:
$$F:\mathbf{R}^n\to\mathbf{R}^m\,,$$
its *scalar product* with a real number r is linear operator:
$$rF:\mathbf{R}^n\to\mathbf{R}^m\,,$$
defined by:
$$(rF)(x)=rF(x)\,.$$

This is a vector space.

Exercise 6.2.10

Prove the last statement.

Warning!

Compositions, too, are possible when $m=n$, but this isn't one of the operations of this vector space.

Example 6.2.11: vector space of matrices $M(n,m)$

To deal with the vector space of linear operators

$$F:\mathbf{R}^n\to\mathbf{R}^m\,,$$

in the last example, we introduce a basis to either of these two Euclidean spaces. Accordingly, all linear operators become matrices. Let's consider the algebra of matrices.

Suppose A and B are two $m\times n$ matrices. Then their *sum* is the $m\times n$ matrix, $A+B$, the ij-entry of which is the sum of the ij-entries of A and B. In other words, if $A=a_{ij}$, $B=b_{ij}$, and $C=A+B=c_{ij}$, then

$$c_{ij}=a_{ij}+b_{ij}\,,$$

for each $i=1,2,...,m$ and each $j=1,2,...,n$.

Suppose A is an $m\times n$ matrix. Then its *scalar multiple* by a real number r, rA, is the $m\times n$ matrix, the ij-entry of which is the product of the ij-entry of A by r. In other words, if $A=a_{ij}$ and $C=rA=c_{ij}$, then

$$c_{ij}=ra_{ij}\,,$$

for each $i=1,2,...,m$ and each $j=1,2,...,n$.

This is a vector space denoted by $M(n,m)$. Each them has a total of nm independent entries. That's why we can say that it's just like a copy of $\mathbf{R}^{nm}$. Here is the linear function

$$F : M(n,m) \to \mathbf{R}^{nm}$$

that connects them:

$$\begin{bmatrix} a_{1,1} & a_{1,2} & a_{1,3} & ... & a_{1,n} \\ a_{2,1} & a_{2,2} & a_{2,3} & ... & a_{2,n} \\ \vdots & \vdots & \vdots & ... & \vdots \\ a_{m,1} & a_{m,2} & a_{m,3} & ... & a_{m,n} \end{bmatrix} \mapsto \begin{bmatrix} a_{1,1} & a_{1,2} & a_{1,3} & ...a_{m,n} \end{bmatrix}$$

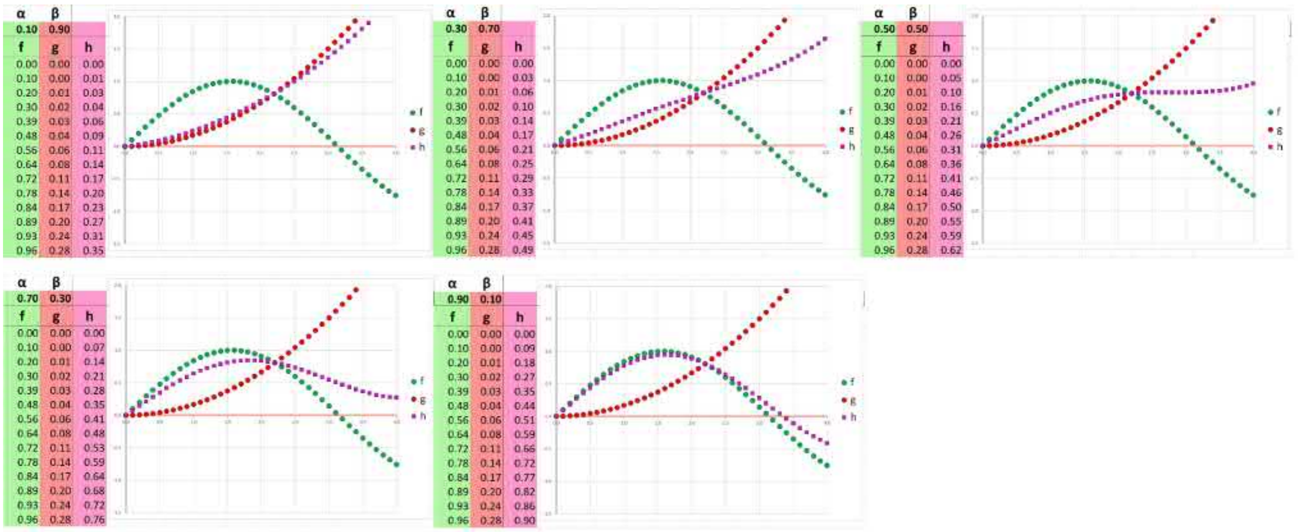
Exercise 6.2.12

Provide a formula for the function. Provide a formula for its inverse.

Warning!

Matrix multiplication is possible here when $m = n$, but it isn't one of the operations of this vector space.

Linear combinations of functions are also available:



α	β	
0.10	0.90	
f	g	h
0.00	0.00	0.00
0.10	0.00	0.01
0.20	0.01	0.03
0.30	0.02	0.04
0.39	0.03	0.06
0.48	0.04	0.09
0.56	0.06	0.11
0.64	0.08	0.14
0.72	0.11	0.17
0.78	0.14	0.20
0.84	0.17	0.23
0.89	0.20	0.27
0.93	0.24	0.31
0.96	0.28	0.35

α	β	
0.30	0.70	
f	g	h
0.00	0.00	0.00
0.10	0.00	0.03
0.20	0.01	0.06
0.30	0.02	0.10
0.39	0.03	0.14
0.48	0.04	0.17
0.56	0.06	0.21
0.64	0.08	0.25
0.72	0.11	0.29
0.78	0.14	0.33
0.84	0.17	0.37
0.89	0.20	0.41
0.93	0.24	0.45
0.96	0.28	0.49

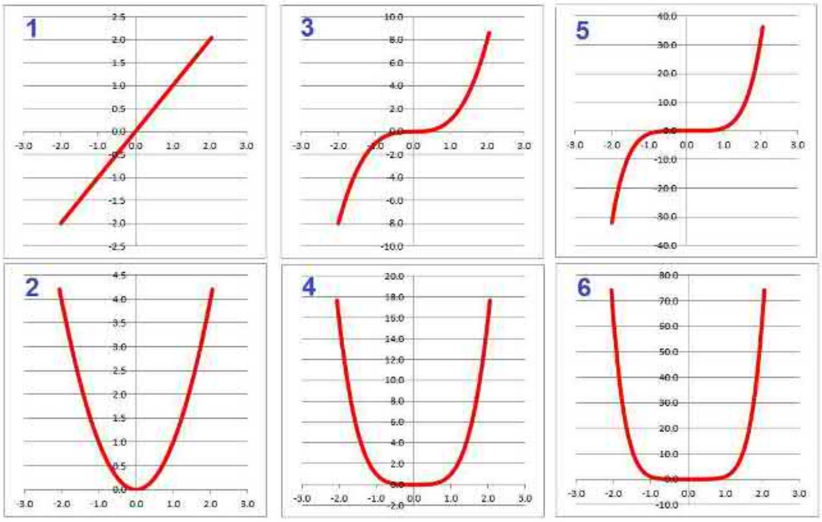
α	β	
0.50	0.50	
f	g	h
0.00	0.00	0.00
0.10	0.00	0.05
0.20	0.01	0.10
0.30	0.02	0.16
0.39	0.03	0.21
0.48	0.04	0.26
0.56	0.06	0.31
0.64	0.08	0.36
0.72	0.11	0.41
0.78	0.14	0.46
0.84	0.17	0.50
0.89	0.20	0.55
0.93	0.24	0.59
0.96	0.28	0.62

α	β	
0.70	0.30	
f	g	h
0.00	0.00	0.00
0.10	0.00	0.07
0.20	0.01	0.14
0.30	0.02	0.21
0.39	0.03	0.28
0.48	0.04	0.35
0.56	0.06	0.41
0.64	0.08	0.48
0.72	0.11	0.53
0.78	0.14	0.59
0.84	0.17	0.64
0.89	0.20	0.68
0.93	0.24	0.72
0.96	0.28	0.76

α	β	
0.90	0.10	
f	g	h
0.00	0.00	0.00
0.10	0.00	0.09
0.20	0.01	0.18
0.30	0.02	0.27
0.39	0.03	0.35
0.48	0.04	0.44
0.56	0.06	0.51
0.64	0.08	0.59
0.72	0.11	0.66
0.78	0.14	0.72
0.84	0.17	0.77
0.89	0.20	0.82
0.93	0.24	0.86
0.96	0.28	0.90

Example 6.2.13: vector space of polynomials

Consider the vector space of up to nth-degree polynomials, denoted by $\mathbf{P}^n$:



We add them this termwise:

$$
\begin{array}{llllll}
 & a_0 & +a_1x^1 & +a_2x^2 & \dots & +a_nx^n \\
+ & & & & & \\
 & b_0 & +b_1x^1 & +b_2x^2 & \dots & +b_nx^n \\
\hline
 & (a_0+b_0) & +(a_1+b_1)x^1 & +(a_2+b_2)x^2 & \dots & +(a_n+b_n)x^n
\end{array}
$$

The operations work one power at a time with no intermixing. We can ignore the powers and see that the algebra works as if these are just vectors in $\mathbf{R}^{n+1}$ (or finite sequences):

$$
\begin{array}{lllllll}
< & a_0 & a_1 & a_2 & \dots & a_n & > \\
+ & & & & & & \\
< & b_0 & b_1 & b_2 & \dots & b_n & > \\
\hline
< & a_0+b_0 & a_1+b_1 & a_2+b_2 & \dots & a_n+b_n & >
\end{array}
$$

Each of them has a total of $(n+1)$ independent entries. That's why we can say that it's just like a copy of $\mathbf{R}^{n+1}$.

Exercise 6.2.14

Provide a formula for the function $\mathbf{P}^n \to \mathbf{R}^{n+1}$. Show that it is linear. Show that it's one-to-one and onto.

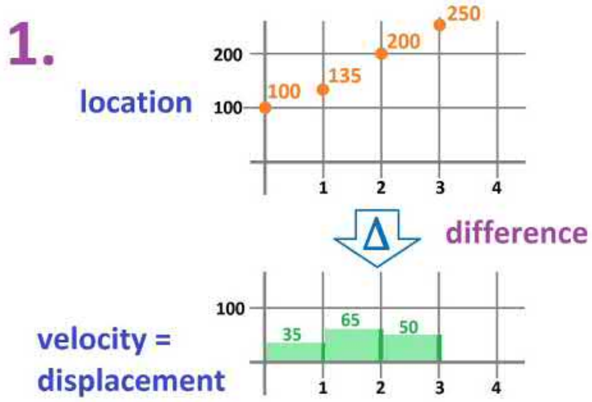
Exercise 6.2.15

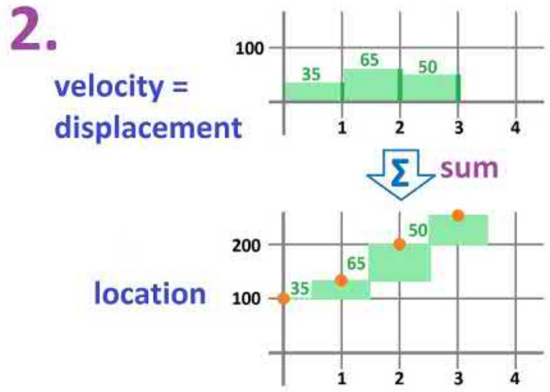
What if the sequences are infinite (power series)?

6.3. *The sequence of differences*

Numbers are subject to algebraic operations: addition, subtraction, multiplication, and division. Since the terms of sequences are numbers, a pair of sequences can be added (subtracted, etc.) to produce a new one.

In addition, there are two operations that apply to a single sequence and produce a *new* sequence that tells us a lot about the *original* sequence. These operations are: subtracting the consecutive terms of the sequence and adding its terms repeatedly. We saw them in action in the first section:





This is the summary:

- If each term of a sequence represents a location, the pairwise differences will give you the *velocities*, and
- If each term of a sequence represents the velocity, their sum up to that point will give you the *location* (or displacement).

The pairwise differences represent the *change* within the sequence, from each of its terms to the next.

Example 6.3.1: sequence given by list

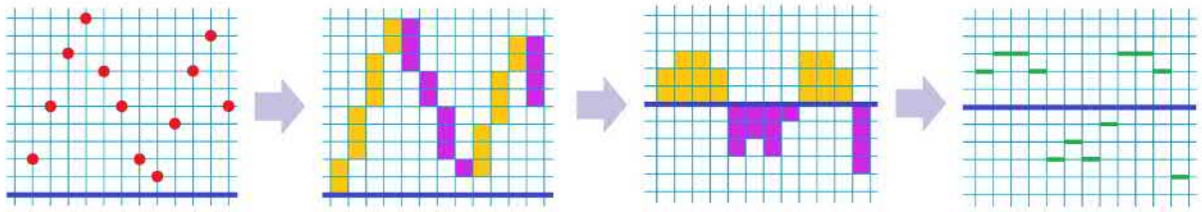
When a sequence is given by a list, we subtract the last term from the current one and put the result in the bottom row as follows:

$$\begin{array}{rccccccccccccccc}
\text{a sequence:} & 2 & & & & 4 & & & & 7 & & & & 1 & & \dots\\
& & \searrow & & \swarrow & & \searrow & & \swarrow & & \searrow & & \swarrow & & \searrow & \dots\\
\text{its differences:} & & & 4-2 & & & & 7-4 & & & & 1-7 & & & & \dots\\
& & & || & & & & || & & & & || & & & & \dots\\
\text{a new sequence:} & & & 2 & & & & 3 & & & & -6 & & & & \dots
\end{array}$$

We have a new list.

Example 6.3.2: sequence given by graph

In the simplest case, a sequence takes only integer values, then on the graph of the sequence, we just count the number of steps we make, up and down:



These increments then make a new sequence plotted on the right.

Definition 6.3.3: sequence of differences

For a sequence a_n, its *sequence of differences*, or simply the difference, is a new sequence, say d_n, defined for each n by the following:

$$d_n = a_{n+1} - a_n\,.$$

It is denoted by

$$\boxed{\Delta a_n = a_{n+1} - a_n}$$

Warning!

The symbol Δ applies to the *whole* sequence a_n, and Δa should be seen as the name of the new sequence; the notation for the difference is an abbreviation for $(\Delta a)_n$.

This is what the definition says:

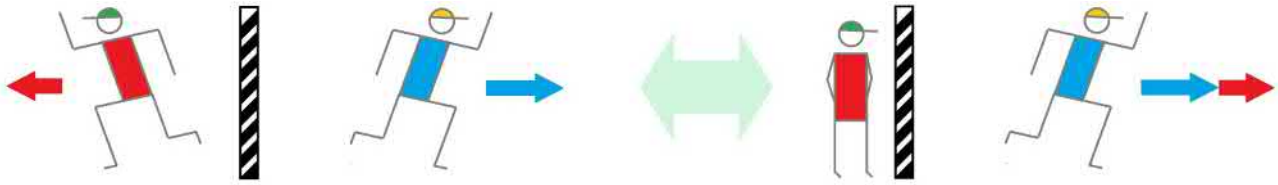
Sequence of differences

$$\begin{array}{rccccccccc}
\text{a sequence:} & a_1 & & a_2 & & a_3 & & a_4 & \dots \\
 & & \searrow \quad \swarrow & & \searrow \quad \swarrow & & \searrow \quad \swarrow & & \dots \\
\text{its differences:} & & a_2 - a_1 & & a_3 - a_2 & & a_4 - a_3 & & \dots \\
 & & || & & || & & || & & \dots \\
\text{a new sequence:} & & d_1 & & d_2 & & d_3 & & \dots \\
 & & || & & || & & || & & \dots \\
\text{the notation:} & & \Delta a_1 & & \Delta a_2 & & \Delta a_3 & & \dots
\end{array}$$

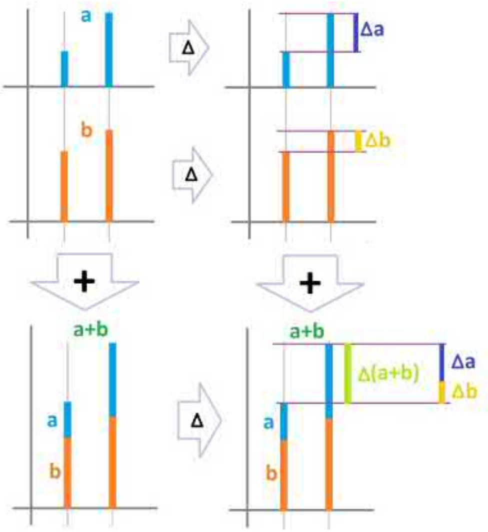
Here is an elementary statement about *motion*:

- ▶ `IF` two runners are running away from a post, `THEN` their relative velocity is the sum of their respective velocities.

It's as if the one runner is standing still while the other is running with the combined speed:



The idea why we *add* their differences when we add sequences is illustrated below:



Here, the bars that represent the change of the values of the sequence are stacked on top of each other. The heights are then added to each other, and so are the height differences. The algebra behind this geometry is very simple:

$$(A+B)-(a+b)=(A-a)+(B-b)\,.$$

It's the *Associative Rule* of addition.

The idea above is equally applicable to runners who change how fast they run; we speak of sequences.

Theorem 6.3.4: Sum Rule for Differences

The difference of the sum of two sequences is the sum of their differences.

In other words, for any two sequences a_n, b_n, their sequences of differences satisfy:

$$\Delta(a_n+b_n)=\Delta a_n+\Delta b_n$$

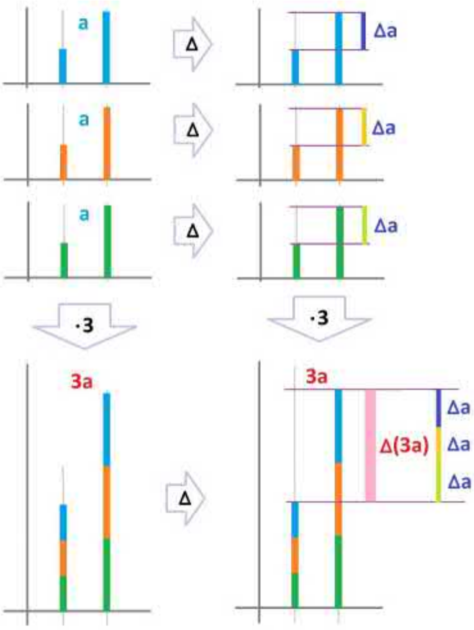
Proof.

$$\begin{aligned}\Delta(a_n+b_n) &= (a_{n+1}+b_{n+1})-(a_n+b_n)\\ &= (a_{n+1}-a_n)+(b_{n+1}-b_n)\\ &= \Delta a_n+\Delta b_n\,.\end{aligned}$$

Here is another simple statement about *motion*:

▶ `IF` the distance is re-scaled, such as from miles to kilometers, `THEN` so is the velocity – at the same proportion.

The idea why a *proportional* change causes the same proportional change in the differences is illustrated below (tripling):



Here, if the heights triple, then so do the height differences. The algebra behind this geometry is very simple:

$$kA-ka=k(A-a)\,.$$

It's the *Distributive Rule*. This is how it applies to sequences.

Theorem 6.3.5: Constant Multiple Rule for Differences

The difference of a multiple of a sequence is the multiple of the sequence's difference.

In other words, for any sequence a_n, the sequence of differences satisfies:

$$\Delta(ka_n) = k\Delta a_n$$

Proof.

$$\begin{aligned}\Delta(ka_n) &= ka_{n+1}k - ka_n \\ &= ka_{n+1}k - ka_n \\ &= k\Delta a_n\,.\end{aligned}$$

The theorem can also be interpreted as follows: If the distances are proportionally increased, then so are the velocities needed to cover them, in the same period of time.

When we represent these sequences as vectors, we can ask and answer linear algebra questions.

Now, the difference will have to be carefully redefined as a *function*.

We have the above formula re-written:

$$\Delta < x_1, ..., x_n > = < x_2 - x_1, x_3 - x_2, ..., x_n - x_{n-1} > \ .$$

There are only $n-1$ entries on the right, which means that we understand this function as

$$\Delta : \mathbf{R}^n \to \mathbf{R}^{n-1}\,.$$

We have for any number α:

$$\begin{aligned}\Delta\big(\alpha < x_1, ..., x_n >\big) &= \Delta < \alpha x_1, ..., \alpha x_n > \\ &= < \alpha x_2 - \alpha x_1, \alpha x_3 - \alpha x_2, ..., \alpha x_n - \alpha x_{n-1} > \\ &= < \alpha(x_2 - x_1), \alpha(x_3 - x_2), ..., \alpha(x_n - x_{n-1}) > \\ &= \alpha < x_2 - x_1, x_3 - x_2, ..., x_n - x_{n-1} > \\ &= \alpha\Delta < x_1, ..., x_n > \ .\end{aligned}$$

Therefore, Δ *preserves scalar multiplication.*

Exercise 6.3.6

Prove that Δ preserves addition.

The following definition is now justified:

Definition 6.3.7: difference operator

The *difference operator*

$$\Delta : \mathbf{R}^n \to \mathbf{R}^{n-1}$$

is a linear operator defined by:

$$\boxed{\Delta < x_1, ..., x_n > = < x_2 - x_1, x_3 - x_2, ..., x_n - x_{n-1} >}$$

What is the matrix of this linear operator? We just look at its values on the standard basis:

$$\Delta \begin{bmatrix} 1 \\ 0 \\ 0 \\ ... \\ 0 \\ 0 \end{bmatrix} = \begin{bmatrix} \\ -1 \\ 0 \\ ... \\ 0 \\ 0 \end{bmatrix}, \ \Delta \begin{bmatrix} 0 \\ 1 \\ 0 \\ ... \\ 0 \\ 0 \end{bmatrix} = \begin{bmatrix} \\ 1 \\ -1 \\ ... \\ 0 \\ 0 \end{bmatrix}, \ ... , \ \Delta \begin{bmatrix} 0 \\ 0 \\ 0 \\ ... \\ 0 \\ 1 \end{bmatrix} = \begin{bmatrix} \\ 0 \\ 0 \\ ... \\ 0 \\ 1 \end{bmatrix} .$$

The gaps are meant to indicate that there are only $n-1$ entries in the output vectors.

We have proven the following:

Theorem 6.3.8: Matrix of Difference

The difference operator $\Delta : \mathbf{R}^n \to \mathbf{R}^{n-1}$ *is given by the following* $(n-1)\times n$ *matrix:*

$$\boxed{\Delta = \begin{bmatrix} -1 & 1 & 0 & 0 & ... & 0 & 0 \\ 0 & -1 & 1 & 0 & ... & 0 & 0 \\ 0 & 0 & -1 & 1 & ... & 0 & 0 \\ ... \\ 0 & 0 & 0 & 0 & ... & 1 & 0 \\ 0 & 0 & 0 & 0 & ... & -1 & 1 \end{bmatrix}}$$

Here, -1's are on the diagonal that starts at the top-left corner, while 1's are on the diagonal that starts at the bottom-right corner; the rest are 0's.

Exercise 6.3.9

Is this operator one-to-one? Onto?

Let's consider a couple of specific sequences.

The first one is the *arithmetic progression* and it is very simple.

Theorem 6.3.10: Difference of Arithmetic Progression

The sequence of differences of an arithmetic progression *with increment* m *is a constant sequence with the value equal to* m*.*

Proof.

We simply compute from the definition:

$$d_n = \Delta(a_0 + mn) = a_{n+1} - a_n = (a_0 + b(m+1)) - (a_0 + mn) = m\,.$$

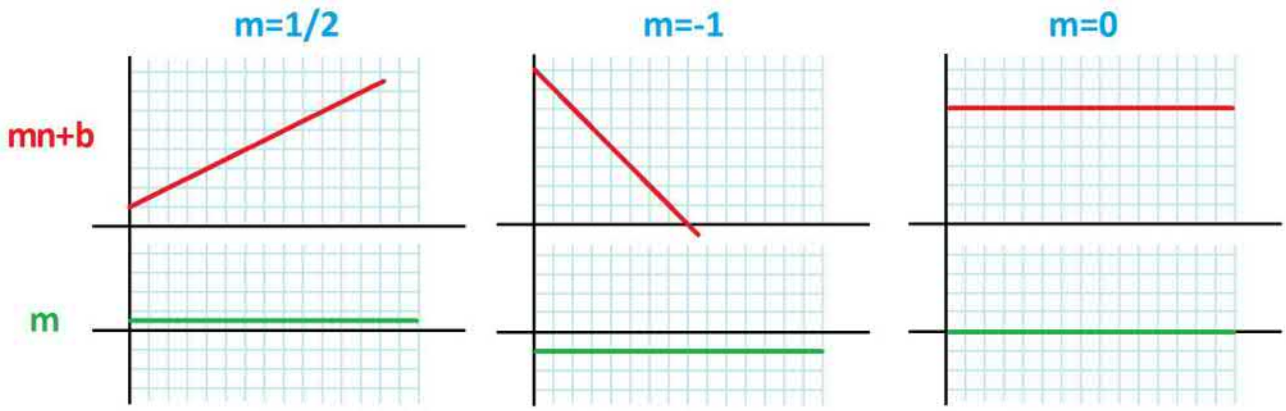
The theorem can be recast as an implication, an "if-then" statement:

▶ `IF` a_n is an arithmetic progression with increment m, `THEN` its sequence of differences is a constant sequence with the value equal to m.

We can also use our convenient abbreviation:

▶ a_n is an arithmetic progression with increment m $\Longrightarrow$ its sequence of differences of a_n is a constant sequence with the value equal to m.

This is what the graphs of this pair of sequences may look like, zoomed out, for the following three choices of the increment m:



Let's take a look at a *geometric progression* $a_n = ar^n$ with $a > 0$ and $r > 0$.

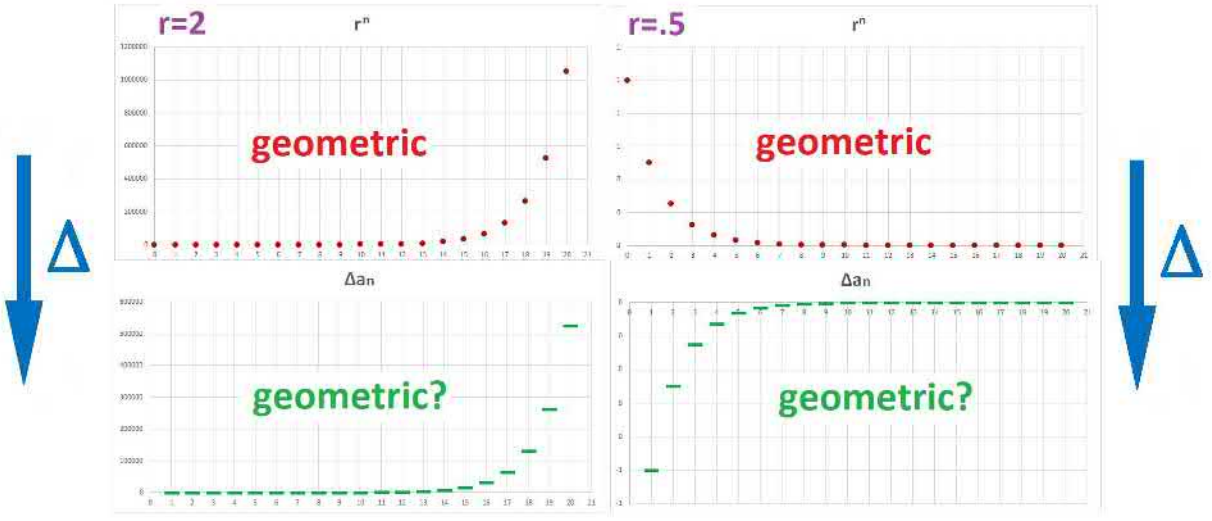
Example 6.3.11: geometric progression $a_n = 3^n$

This is a geometric progression with ratio $r = 3$. Let's compute the difference:

$$\begin{aligned}\Delta a_n &= a_{n+1} - a_n\\ &= 3^{n+1} - 3^n\\ &= 3^n \cdot 3 - 3^n\\ &= 3^n(3-1)\\ &= 3^n \cdot 2\,.\end{aligned}$$

It's a geometric progression with $r = 3$, again!

Is there a pattern? Let's plot the graph of a geometric progression $a_n = ar^n$ with $a > 0$ and $r > 0$. There are two cases, depending on the choice of ratio r (growth or decay):



What do the sequence of differences (second row) look like? We notice the following:

- It is positive and increasing, with speeding up when the ratio r is larger than 1.
- It is negative and increasing, with slowing down when $0 < r < 1$.

It also resembles the original sequence!

Theorem 6.3.12: Difference of Geometric Progression

The sequence of differences of a geometric progression is a geometric progression with the same ratio.

Proof.

If we have a geometric progression with ratio r and initial term a, its formula is $a_n = ar^n$. Therefore,

$$d_n = \Delta(ar^n) = a_{n+1} - a_n = ar^{n+1} - ar^n = a(r-1) \cdot r^n \,.$$

But that's the formula of a geometric progression with ratio r and initial term $a(r-1)$.

The theorem can be restated as an implication:

- `IF` a_n is a geometric progression with ratio r, `THEN` its sequence of differences is a geometric progression with ratio r.

Also:

- a_n is a geometric progression with ratio $r \implies$ its sequence of differences is a geometric progression with ratio r.

Example 6.3.13: alternating sequence

The sequence of differences of the alternating sequence $a_n = (-1)^n$ is computed below:

$$\Delta\left((-1)^n\right) = (-1)^{n+1} - (-1)^n = \begin{cases} (-1) - 1, & n \text{ is even} \\ 1 - (-1), & n \text{ is odd} \end{cases} = \begin{cases} -2, & n \text{ is even} \\ 2, & n \text{ is odd} \end{cases} = 2(-1)^{n+1} \,.$$

Exercise 6.3.14

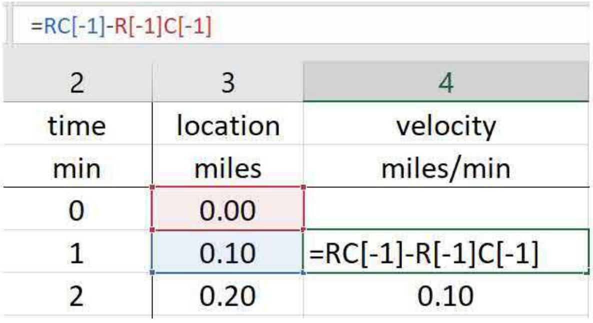
What is the relation between the sequence above and its difference?

Example 6.3.15: differences are velocities

We can use computers to speed up these computations. For example, one may have been recording one's locations and now needs to find the velocities. Here is a spreadsheet formula for the sequence of differences (velocities):

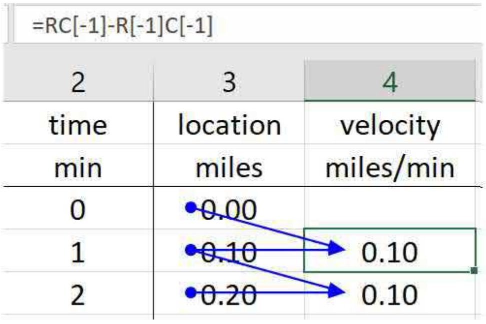
`=RC[-1]-R[-1]C[-1]`

Whether the sequence comes from a formula or it's just a list of numbers, the formula applies equally:



=RC[-1]-R[-1]C[-1]

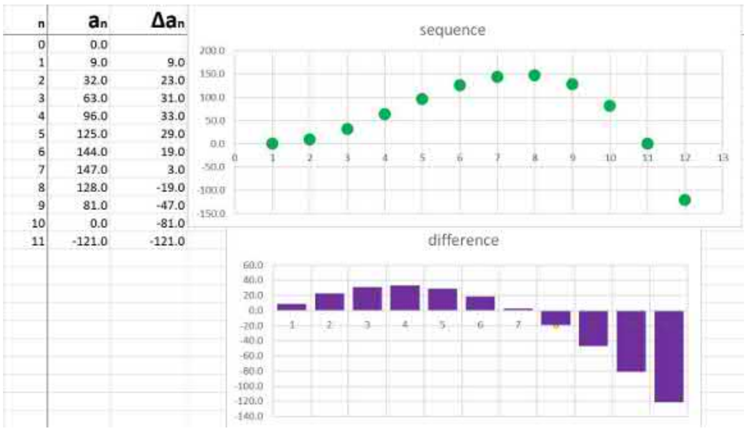
2	3	4
time	location	velocity
min	miles	miles/min
0	0.00	
1	0.10	=RC[-1]-R[-1]C[-1]
2	0.20	0.10



=RC[-1]-R[-1]C[-1]

2	3	4
time	location	velocity
min	miles	miles/min
0	0.00	
1	0.10	0.10
2	0.20	0.10

As a result, a curve has produced a new curve:



n	a_n	Δa_n
0	0.0	
1	9.0	9.0
2	32.0	23.0
3	63.0	31.0
4	96.0	33.0
5	125.0	29.0
6	144.0	19.0
7	147.0	3.0
8	128.0	-19.0
9	81.0	-47.0
10	0.0	-81.0
11	-121.0	-121.0

While the first graph tells us that we are moving forward and then backward, it is easier to derive better description from the second: Speed up forward, then slow down, then speed up backward.

Exercise 6.3.16

Describe what has happened referring, separately, to the first graph and the second graph.

Exercise 6.3.17

Imagine, instead, that the first column of the spreadsheet above is where you have been recording the monthly balance of your bank account. What does the second column represent? Describe what has been happening with your finances referring, separately, to the first graph and the second graph.

This is the time for some *theory*.

Consider this obvious statement about motion:

- `IF` I am standing still, `THEN` my velocity is zero.

We can also say:

- `IF` my velocity is zero, `THEN` I am standing still.

We see the implications going both ways; the latter is the *converse* of the original statement (and vice versa!). Let's use symbols to restate these statements more compactly:

$$\begin{aligned}\text{I am standing still} &\quad\Longrightarrow\quad \text{my velocity is zero.}\\ \text{I am standing still} &\quad\Longleftarrow\quad \text{my velocity is zero.}\end{aligned}$$

The abbreviation of the combination of the two is an equivalence, an "if-and-only-if" statement:

$$\text{I am standing still} \quad\Longleftrightarrow\quad \text{my velocity is zero.}$$

It's just another way of saying the same thing.

If we set the motion point of view aside, here is the general statement.

Theorem 6.3.18: Difference of Constant Sequence

A sequence is constant `IF AND ONLY IF` *its sequence of differences is zero.*

In other words, we have:

$$a_n \text{ is constant} \quad\Longleftrightarrow\quad \Delta a_n = 0\,.$$

Proof.

Direct:

$$a_n = c \text{ for all } n \implies a_{n+1} - a_n = c - c = 0 \implies \Delta a_n = 0\,.$$

Converse:

$$a_{n+1} = a_n = c \text{ for all } n \impliedby a_{n+1} - a_n = 0 \impliedby \Delta a_n = 0\,.$$

Exercise 6.3.19

Prove that the difference of an arithmetic progression is constant and, conversely, that if the difference of a sequence is a constant sequence, then the sequence is an arithmetic progression.

Consider another obvious statement about motion:

- `IF` I am moving forward, `THEN` my velocity is positive.

And, conversely:

- `IF` my velocity is positive, `THEN` I am moving forward.

In other words, we have this pair of statements:

$$\begin{array}{lll} \text{Original:} & \text{I am moving forward} & \Longrightarrow & \text{my velocity is positive.}\\ \text{Converse:} & \text{I am moving forward} & \Longleftarrow & \text{my velocity is positive.} \end{array}$$

Exercise 6.3.20

What is the converse of the converse?

We combine these two in the following far-reaching result.

Theorem 6.3.21: Monotonicity Theorem for Sequences

A sequence is increasing/decreasing `IF AND ONLY IF` *the sequence of differences is positive/negative or zero, respectively.*

In other words, we have:

$$\begin{array}{lllll} a_n & \textit{is increasing} & \iff & \Delta a_n & \geq 0\,.\\ a_n & \textit{is decreasing} & \iff & \Delta a_n & \leq 0\,.\\ a_n & \textit{is constant} & \iff & \Delta a_n & = 0\,. \end{array}$$

Proof.

$$a_{n+1} \geq a_n \text{ for all } n \iff a_{n+1} - a_n \geq 0 \iff \Delta a_n \geq 0\,.$$

It's just another way of saying the same thing.

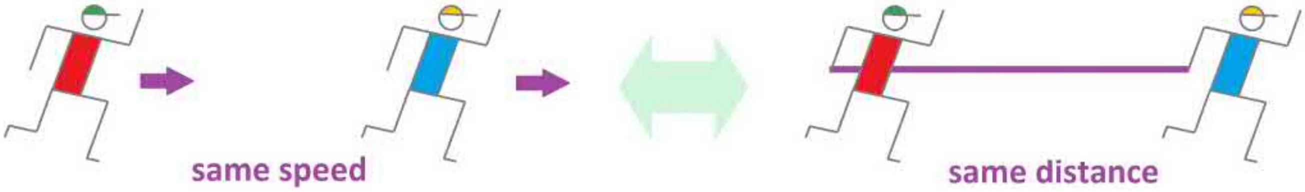
Suppose now that there are *two* runners; then we have a less obvious fact about motion:

- `IF` the distance between two runners isn't changing, `THEN` they are running with the same velocity.

And vice versa:

- `IF` two runners are running with the same velocity, `THEN` the distance between them isn't changing.

It's as if they are holding the two ends of a pole:



The conclusion holds even if they speed up and slow down all the time. In other words, we have:

- ▶ The distance between two runners isn't changing `IF AND ONLY IF` they are running with the same velocity.

Once again, for sequences a_n and b_n representing their respective positions at time n, we can restate this idea mathematically in order to confirm that our theory makes sense.

Corollary 6.3.22: Difference under Subtraction

Two sequences differ by a constant `IF AND ONLY IF` *if their sequences of differences are equal.*

In other words, we have:

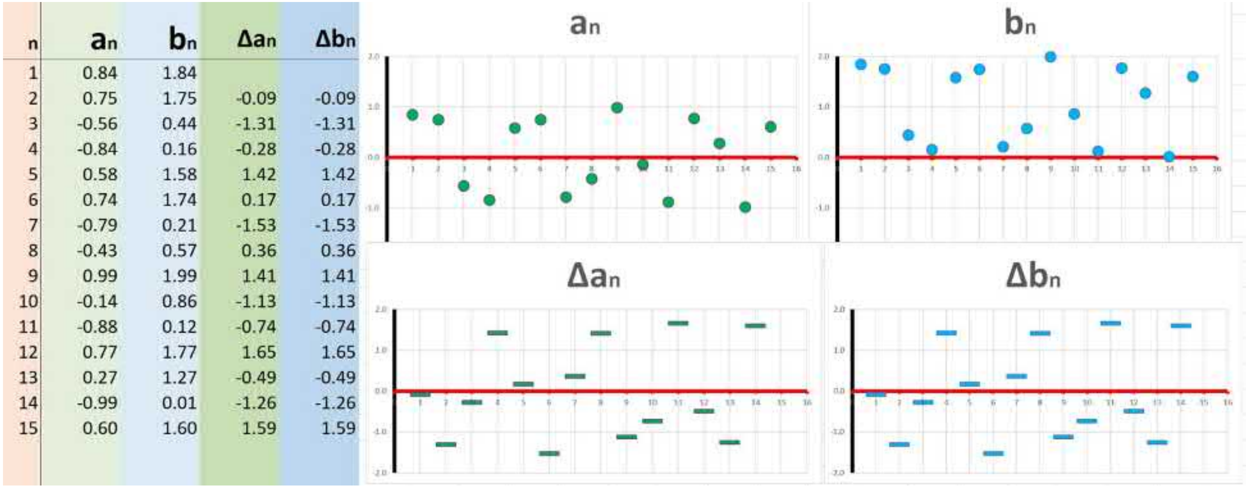
$$a_n - b_n \text{ is constant} \iff \Delta a_n = \Delta b_n\,.$$

Proof.

The corollary follows from the *Difference of Constant Sequence Theorem* above.

Example 6.3.23: shift of sequence

We shift the sequence a_n below by 1 unit up to produce a new sequence b_n (top):



n	an	bn	Δan	Δbn
1	0.84	1.84		
2	0.75	1.75	-0.09	-0.09
3	-0.56	0.44	-1.31	-1.31
4	-0.84	0.16	-0.28	-0.28
5	0.58	1.58	1.42	1.42
6	0.74	1.74	0.17	0.17
7	-0.79	0.21	-1.53	-1.53
8	-0.43	0.57	0.36	0.36
9	0.99	1.99	1.41	1.41
10	-0.14	0.86	-1.13	-1.13
11	-0.88	0.12	-0.74	-0.74
12	0.77	1.77	1.65	1.65
13	0.27	1.27	-0.49	-0.49
14	-0.99	0.01	-1.26	-1.26
15	0.60	1.60	1.59	1.59

Because the ups and downs remain the same, the sequences of differences of these two sequences are identical (bottom).

Exercise 6.3.24

What if the two runners holding the pole also start to move their hands back and forth?

We can use the latter theorem to watch after the distance between the two runners. A matching statement about motion is the following:

- ▶ `IF` the distance from one of the two runners to the other is increasing, `THEN` the former's velocity is higher.

Conversely:

▶ `IF` the velocity of one runner is higher than the other, `THEN` the distance between them is increasing.

Exercise 6.3.25

Combine the two statements into one.

We can restate this mathematically.

Corollary 6.3.26: Monotonicity and Subtraction

The difference of two sequences is increasing `IF AND ONLY IF` *the former's difference is bigger than the latter's.*

In other words, we have:

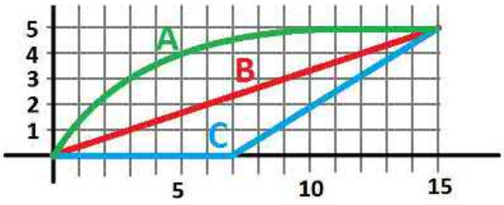
$$\begin{array}{lllll} a_n - b_n & \textit{is increasing} & \Longleftrightarrow & \Delta a_n & \geq & \Delta b_n\,. \\ a_n - b_n & \textit{is decreasing} & \Longleftrightarrow & \Delta a_n & \leq & \Delta b_n\,. \end{array}$$

Proof.

The corollary follows from the *Monotonicity Theorem for Sequences* above.

Example 6.3.27: three runners

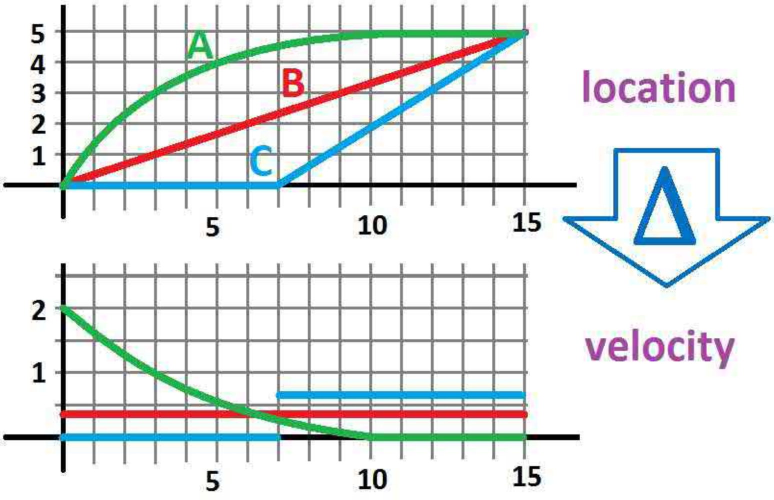
The graph below shows the positions of three runners in terms of time, n. Describe what has happened:



They are all at the starting line together, and at the end, they are all at the finish line. Furthermore, A reaches the finish line first, followed by B, and then C (who also starts late). This is *how* each did it:

- A starts fast, then slows down, and almost stops close to the finish line.
- B maintains the same speed.
- C starts late and then runs fast at the same speed.

We can see that A is running faster because the distance from B is increasing. It becomes slower later, which is visible from the decreasing distance. We can discover this and the rest of the facts by examining the graphs of the *differences* of the sequences:



Exercise 6.3.28

Suppose a sequence is given by the graph for velocity above. Sketch the graph of the difference of this sequence. What is its meaning?

Exercise 6.3.29

Plot the location and the velocity for the following trip: “I drove fast, then gradually slowed down, stopped for a very short moment, gradually accelerated, maintained speed, hit a wall.” Make up your own story and repeat the task.

Exercise 6.3.30

Draw a curve on a piece of paper, imagine that it represents your locations, and then sketch what your velocity would look like. Repeat.

Exercise 6.3.31

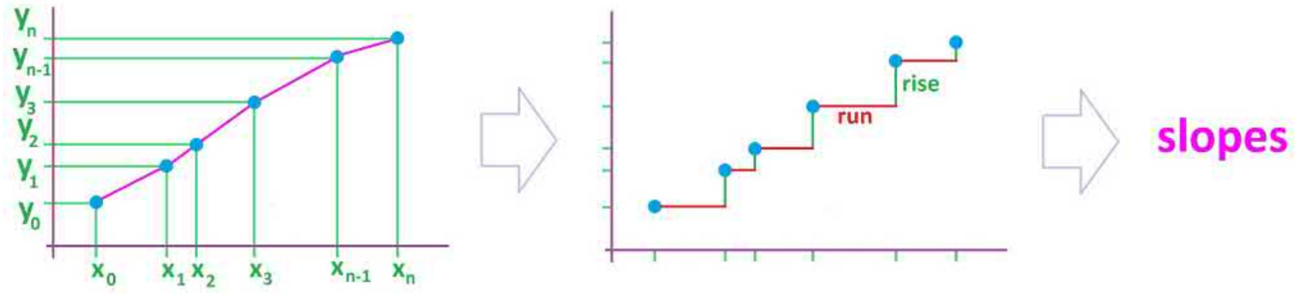
Imagine that the first graph represents, instead of locations, the balances of three bank accounts. Describe what has been happening.

How do we treat motion when the time increment isn’t 1? What is the *velocity* then?

First, we define the time and the location by two separate sequences, say, x_n and y_n. Then the velocity is the increment of the latter over the increment of the former. We notice that those two are the differences of the two sequences. The *difference quotient* of two sequences of x_n and y_n is defined to be the sequence that is the difference of y_n divided by the difference of x_n:

$$\frac{\Delta y_k}{\Delta x_k} = \frac{y_{k+1} - y_k}{x_{k+1} - x_k},$$

provided the denominator is not zero. It is the relative change – the *rate* of change – of the two sequences (for each consecutive pair of points, it is the slope):



The difference quotient of the sequence of the location with respect to the sequence of time is the velocity.

Typically, the sequence x_n is fixed. This makes the difference quotient a function defined on the sequences of length n just as the difference. We can write this function as follows:

$$\frac{\Delta}{\Delta x_k} : \mathbf{R}^{n+1} \to \mathbf{R}^n .$$

Exercise 6.3.32

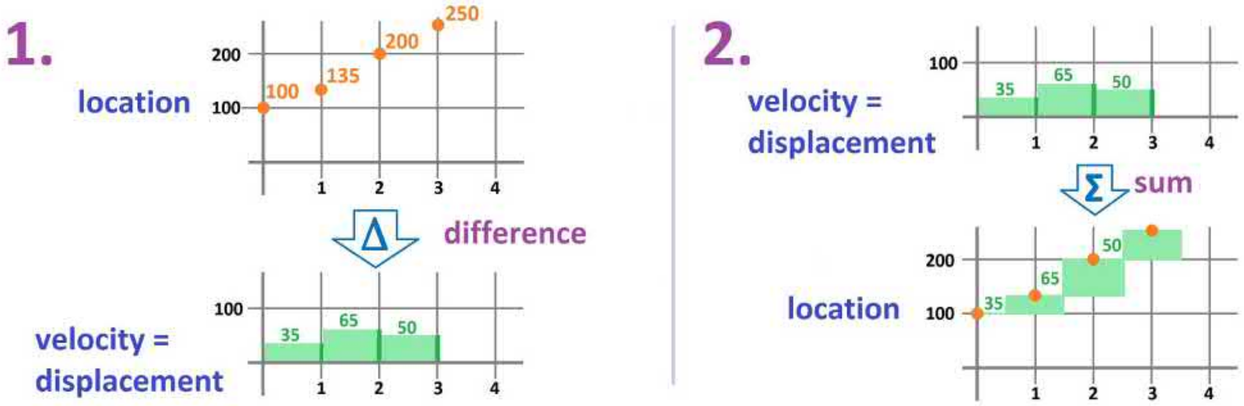
Prove analogues of the Sum Rule and the Constant Multiple Rule for the difference quotient. Derive that the difference quotient is a linear operator.

Exercise 6.3.33

Express the difference quotient operator in terms of the difference operator.

6.4. *The sequence of sums*

In the first section, we saw how the sequences of *locations and velocities* interact. We took a closer look at the transition from the former to the latter and now in reverse:



The sum represents the totality of the "beginning" of a sequence, found by adding each of its terms to the next, up to that point.

Example 6.4.1: sequences given by lists

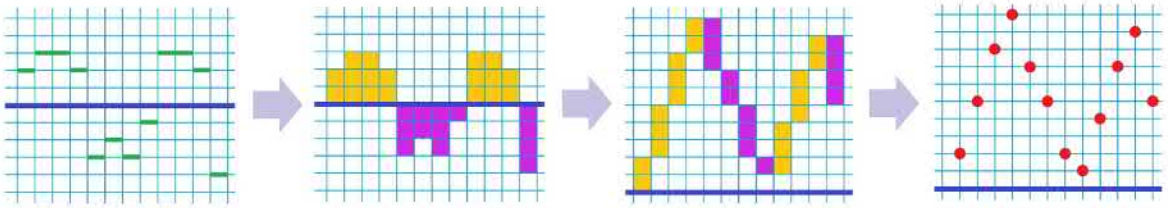
We just add the current term to what we have accumulated so far:

$$\begin{array}{lcccccccccccccccccc}
\text{sequence:} & 2 & & 4 & & & & 7 & & & & 1 & & & & -1 & & & \dots \\
 & \downarrow & & \downarrow & & & & \downarrow & & & & \downarrow & & & & \downarrow & & & \dots \\
\text{sums:} & 2 \\
 & 2 & + & 4 & = & 6 \\
 & & & & & 6 & + & 7 & = & 13 \\
 & & & & & & & & & 13 & + & 1 & = & 14 \\
 & & & & & & & & & & & & & 14 & + & (-1) & = & 13 \\
 & \downarrow & & & & \downarrow & & & & \downarrow & & & & \downarrow & & & & \downarrow & \dots \\
\text{new sequence:} & 2 & & & & 6 & & & & 13 & & & & 14 & & & & 13 & \dots
\end{array}$$

We have a new list!

Example 6.4.2: sequences given by graphs

We treat the graph of a sequence as if made of bars and then just stack up these bars on top of each other one by one:



These stacked bars – or rather the process of stacking – make a new sequence.

Unlike the difference, the sum must be defined (and computed) in a *recursive* manner.

Definition 6.4.3: sequence of sums

For a sequence a_n, its *sequence of sums*, or simply the sum, is a new sequence s_n defined and denoted for each $n \geq m$ within the domain of a_n by the following (recursive) formula:

$$\boxed{s_m = 0, \quad s_{n+1} = s_n + a_{n+1}}$$

In other words, we have:

$$\boxed{s_n = a_m + a_{m+1} + ... + a_n}$$

Of course, we will use our notation:

Sigma notation for summation

$$s_n = a_m + a_{m+1} + ... + a_n = \sum_{k=m}^{n} a_k$$

Recall that the Greek letter Σ stands for the letter S meaning "sum".

The notation applies to all sequences, both finite and infinite. For infinite sequences, recognized by "..." at the end, the sum sequence is also called "partial sums" as well as "series".

This is the definition of the sequence of sums written with the sigma notation:

Sequence of sums

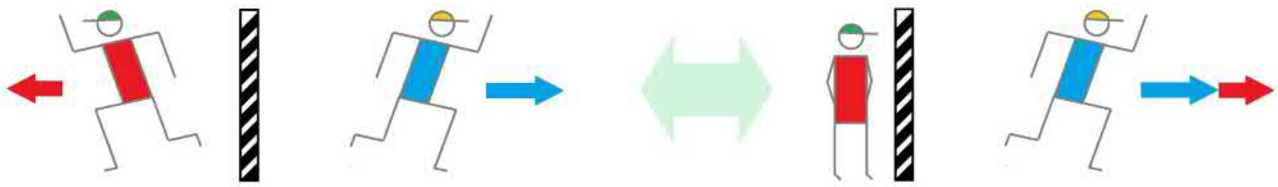
$$\begin{array}{lcccccccccccccc}
\text{a sequence:} & a_1 & & a_2 & & & & a_3 & & & & a_4 & & & \dots \\
 & \downarrow & & \downarrow & & & & \downarrow & & & & \downarrow & & & \dots \\
\text{its sums:} & a_1 & & & & & & & & & & & & & \\
 & a_1 & + & a_2 & = & s_2 & & & & & & & & & \\
 & & & & & s_2 & + & a_3 & = & s_3 & & & & & \\
 & & & & & & & & & s_3 & + & a_4 & = & s_4 & \dots \\
 & & & & & & & & & & & & & s_4 & \dots \\
 & \downarrow & & & & \downarrow & & & & \downarrow & & & & \downarrow & \dots \\
\text{the sequence of sums:} & s_1 & & & & s_2 & & & & s_3 & & & & s_4 & \dots \\
 & || & & & & || & & & & || & & & & || & \dots \\
\text{the sigma notation:} & \sum_{k=1}^{1} a_k & & & & \sum_{k=1}^{2} a_k & & & & \sum_{k=1}^{3} a_k & & & & \sum_{k=1}^{4} a_k & \dots
\end{array}$$

Now, the properties of linear algebra.

Here is an elementary statement about *motion*:

> ▶ `IF` two runners are running away from a post, `THEN` the distance between them is the sum of their respective distances from the post.

It's as if the one runner is standing still while the other is running with the combined speed:



This simple algebra, the *Associative Property* combined with the *Commutative Property*, tells the whole story:

re-arrange the sum of four numbers:

$$\begin{array}{cccc|c|c}
 & a & + & b & = & (a+b), \\
 & + & & + & & + \\
 & A & + & B & = & (A+B) \\ \hline
= & (a+A) & + & (b+B) & = & a+A+b+B
\end{array}$$

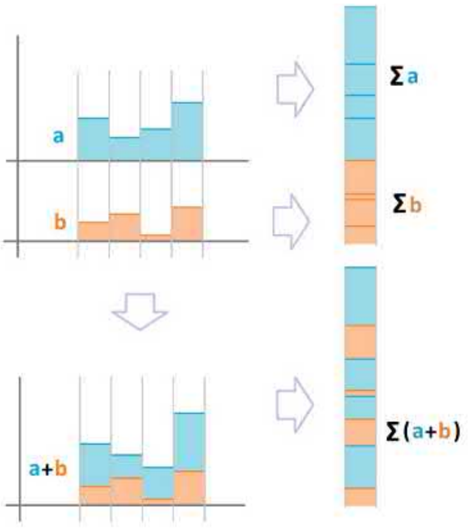
The rule applies even if we have more than just two terms; it's about re-arranging the terms of sequences:

re-arrange the sum

of two sequences:

$$\begin{array}{ccc|c|ccc}
a_p & + & b_p & = & (a_p+b_p)+ & & \\
a_{p+1} & + & b_{p+1} & = & (a_{p+1}+b_{p+1})+ & & \\
\vdots & \vdots & \vdots & & \vdots & & \\
a_q & + & b_q & = & (a_q+b_q) & & \\ \hline
=(a_p+...+a_q) & + & (b_p+...+b_q) & = & (a_p+b_p)+ & ... & +(a_q+b_q)
\end{array}$$

The summation is illustrated below:



An abbreviated version of this formula is as follows.

Theorem 6.4.4: Sum Rule for Sums

The sum of the sums of two sequences is the sum of the sequence of the sums.

In other words, if a_n and b_n are sequences, then, for any p,q with $p \le q$, their sequences of sums satisfy:

$$\sum_{n=p}^{q} a_n + \sum_{n=p}^{q} b_n = \sum_{n=p}^{q} (a_n+b_n)$$

Exercise 6.4.5

Derive this theorem from the last one, reverse.

If your velocity is tripled, then so is the distance you have covered.

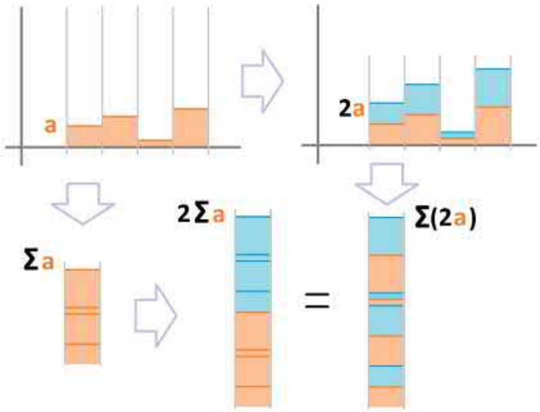
When a sequence is multiplied by a constant, what happens to its sums? This simple algebra, the *Distributive Property*, tells the whole story:

$$\text{take out a common factor:}\quad \begin{array}{cccc} k\cdot(& a & + & b) \\ = & ka & + & kb \end{array}$$

The rule applies even if we have more than just two terms; it's about factoring:

$$\text{take out a common factor:}\quad \begin{array}{ccc|c|l} k & \cdot & a_p & = & k\cdot a_p + \\ k & \cdot & a_{p+1} & = & k\cdot a_{p+1} + \\ \vdots & \vdots & \vdots & & \quad\vdots \\ k & \cdot & a_q & = & k\cdot a_q \\ \hline =k\cdot a_p + & ... & +k\cdot a_q & = & k\cdot(a_p+...+a_q) \end{array}$$

This summation is illustrated below:



An abbreviated version of this formula is as follows.

Theorem 6.4.6: Constant Multiple Rule for Sums

The sum of a multiple of a sequence is the multiple of its sum.

In other words, if a_n is a sequence, then for any p, q with $p \le q$ and any real k, its sequence of sums satisfies:

$$\sum_{n=p}^{q}(ka_n) = k\sum_{n=p}^{q} a_n$$

Exercise 6.4.7

Derive this theorem from the last one, reverse.

When we represent these sequences as vectors, we can ask and answer linear algebra questions.

Now, the sum will have to be carefully redefined as a *function*.

We have the above formula re-written:

$$\Sigma < x_1, x_2, ..., x_n > = < x_0\,,\ x_0 + x_1\,,\ ...\,,\ x_0 + x_1 + ... + x_{n-1} + x_n > \ .$$

As you can see, x_0 isn't present among the inputs. It is the starting value of our sequence (the initial position in case of motion), and can be chosen arbitrarily, producing a new function every time. We have a function

$$\Sigma : \mathbf{R}^n \to \mathbf{R}^{n+1}\,.$$

For any number α, we write:

$$\begin{array}{rll} & X & =< x_1, x_2, ..., x_n > \\ & \alpha X & =< \alpha x_1, \alpha x_2, ..., \alpha x_n > \\ k\text{th entry} & \Sigma \alpha X & = x_0 + \alpha x_1 + ... + \alpha x_{k-1} + \alpha x_k \quad = x_0 + \alpha\big(x_1 + x_2 + ... + x_{k-1} + x_k\big) \qquad = x_0 + \alpha \Sigma X \end{array}$$

Therefore, Σ *preserves scalar multiplication* but only when $x_0 = 0$!

Exercise 6.4.8

Prove that Σ preserves addition only when $x_0 = 0$.

For the sake of linear algebra, we will choose this particular case of the sum.

The following definition is now justified:

Definition 6.4.9: sum operator

The *sum operator*

$$\Sigma : \mathbf{R}^n \to \mathbf{R}^{n+1}$$

is a linear operator defined by:

$$\boxed{\Sigma < x_1, x_2, ..., x_n > = < 0\,,\ x_1\,,\ x_1 + x_2 \,...\ ,\ x_1 + ... + x_{n-1} + x_n\ >}$$

What is the matrix of this linear operator? We just look at its values on the standard basis:

$$\Sigma \begin{bmatrix} \\ 1 \\ 0 \\ ... \\ 0 \\ 0 \end{bmatrix} = \begin{bmatrix} 0 \\ 1 \\ 1 \\ ... \\ 1 \\ 1 \end{bmatrix},\ \Sigma \begin{bmatrix} \\ 0 \\ 1 \\ ... \\ 0 \\ 0 \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ 1 \\ ... \\ 1 \\ 1 \end{bmatrix},\ ...\ ,\ \Sigma \begin{bmatrix} \\ 0 \\ 0 \\ ... \\ 0 \\ 1 \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ 0 \\ ... \\ 0 \\ 1 \end{bmatrix}.$$

We have proven the following:

Theorem 6.4.10: Matrix of Sum Operator

The sum operator $\Sigma : \mathbf{R}^n \to \mathbf{R}^{n+1}$ *is given by the following* $(n+1) \times n$ *matrix:*

$$\boxed{\Sigma = \begin{bmatrix} 0 & 0 & 0 & 0 & ... & 0 & 0 \\ 1 & 0 & 0 & 0 & ... & 0 & 0 \\ 1 & 1 & 0 & 0 & ... & 0 & 0 \\ ... \\ 1 & 1 & 1 & 1 & ... & 1 & 0 \\ 1 & 1 & 1 & 1 & ... & 1 & 1 \end{bmatrix}}$$

Here, the diagonal that starts at the left-upper corner (and all above) consists of 0's, while the one that starts at the bottom-right corner (and all below) consists of 1's.

Exercise 6.4.11

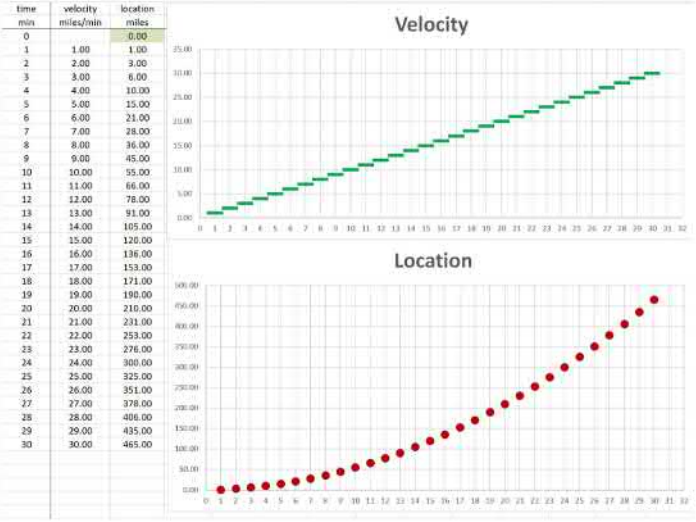
Is this operator one-to-one? Onto?

Below is the simplest result about the sum of a specific sequence. It is still considerably more challenging than most results about the differences that we saw in the last section.

Theorem 6.4.12: Sum of Arithmetic Progression

The sum of an arithmetic progression with increment m and a 0 initial term is a (quadratic) sequence given by the following:

$$\sum_{k=1}^{n}(mk)=\frac{n(n+1)}{2}m\,.$$



time min	velocity miles/min	location miles
0		0.00
1	1.00	1.00
2	2.00	3.00
3	3.00	6.00
4	4.00	10.00
5	5.00	15.00
6	6.00	21.00
7	7.00	28.00
8	8.00	36.00
9	9.00	45.00
10	10.00	55.00
11	11.00	66.00
12	12.00	78.00
13	13.00	91.00
14	14.00	105.00
15	15.00	120.00
16	16.00	136.00
17	17.00	153.00
18	18.00	171.00
19	19.00	190.00
20	20.00	210.00
21	21.00	231.00
22	22.00	253.00
23	23.00	276.00
24	24.00	300.00
25	25.00	325.00
26	26.00	351.00
27	27.00	378.00
28	28.00	406.00
29	29.00	435.00
30	30.00	465.00

Exercise 6.4.13

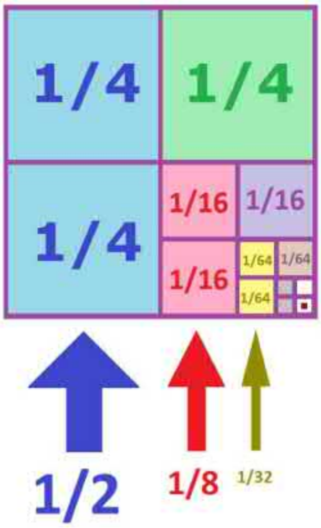
Prove the theorem.

Example 6.4.14: from sequences to series

What does the sum

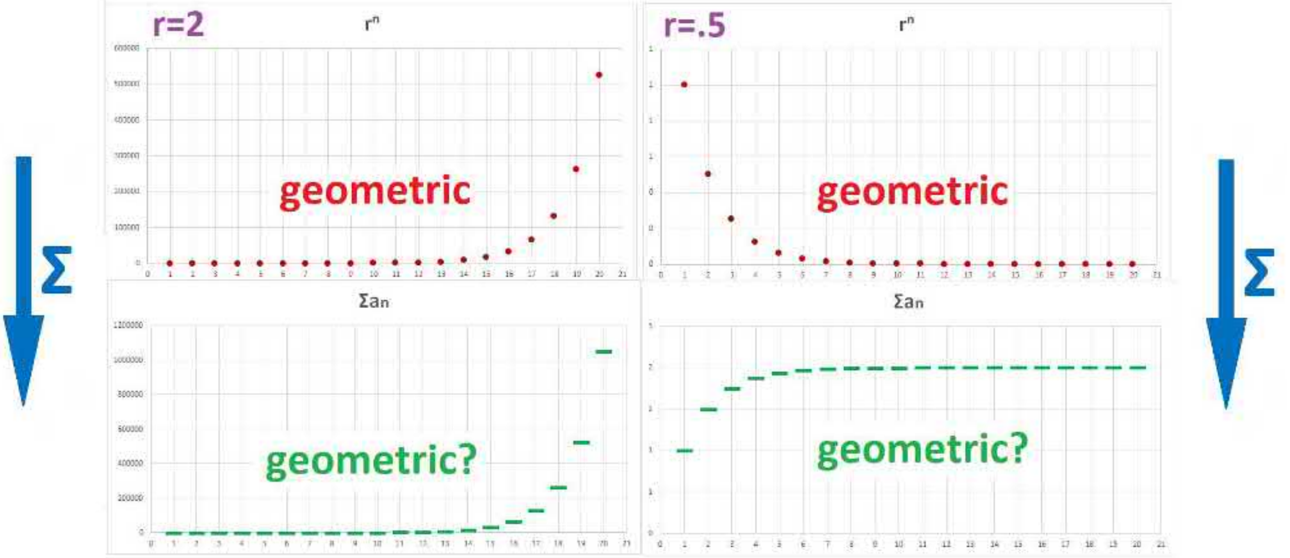
$$\frac{1}{2}+\frac{1}{4}+\frac{1}{8}+...$$

compute? We can see these terms as the areas of the squares below:



On the one hand, adding the areas of these squares will never go over 1 (the area of the big square), and, on the other, these squares seem to exhaust this square entirely. So, even the *infinite* sum sometimes makes sense. Of course, this is a geometric progression.

Let's take a look at a *geometric progression* with $a_n = ar^n$ with $a > 0$ and $r > 0$. There are two cases, depending on the choice of ratio r (growth or decay):



What about the sequence of sums (second row)? We notice the following:

- It is increasing, with speeding up when its ratio r is larger than 1.
- It is increasing, with slowing down when $0 < r < 1$.

It also resembles the original sequence!

Theorem 6.4.15: Sum of Geometric Progression

The sequence of sums of a geometric progression with ratio $r \neq 1$ is a geometric progression with the same ratio and a constant sequence.

In other words, we have:

$$\sum_{k=1}^{n} ar^k = Ar^n + C\,,$$

for some real numbers A and C.

Proof.

Below, we use a *clever trick* to get rid of "...". We write the nth sum s_n,

$$s_n \quad = ar^0 \quad +ar^1 \quad +ar^2 \quad +\ldots \quad +ar^{n-1} \quad +ar^n$$

and then multiply it by r:

$$\begin{array}{rl} rs_n & = r\big(ar^0 \quad +ar^1 \quad +ar^2 \quad +\ldots \quad +ar^{n-1} \quad +ar^n\big) \\ & = \quad ar^1 \quad + \quad ar^2 \quad + \quad ar^3 \quad +\ldots \quad + \quad ar^n \quad + \quad ar^{n+1} \end{array}$$

Now we subtract these two:

$$\begin{array}{rlllllll} s_n & = ar^0 & +ar^1 & +ar^2 & +\ldots & +ar^{n-1} & & +ar^n \\ rs_n & = & ar^1 & + \quad ar^2 & + \quad ar^3 & +\ldots & + \quad ar^n & + \quad ar^{n+1} \\ \hline s_n - rs_n & = ar^0 - ar^1 & +ar^1 - ar^2 & +ar^2 - ar^3 & +\ldots & +ar^{n-1} - ar^n & & +ar^n - ar^{n+1} \\ & = ar^0 & & & & & & -ar^{n+1} \end{array}$$

We cancel the terms that appear twice in the last row and "..." is gone! Therefore,

$$s_n(1-r) = a - ar^{n+1}\,.$$

Thus, we have an explicit formula for the nth term of the sum:

$$s_n = \frac{a}{1-r}(1-r^{n+1}) = -\frac{a}{1-r} \cdot r^{n+1} + \frac{a}{1-r}\,.$$

The former term is the geometric part, and the latter is the constant:

$$A = -\frac{ar}{1-r},\ C = \frac{a}{1-r}\,.$$

Exercise 6.4.16

Find the explicit formula for the sequence of sums of the alternating sequence $a_n = (-1)^n$.

Exercise 6.4.17

Use the trick to prove the theorem about arithmetic progressions.

Warning!

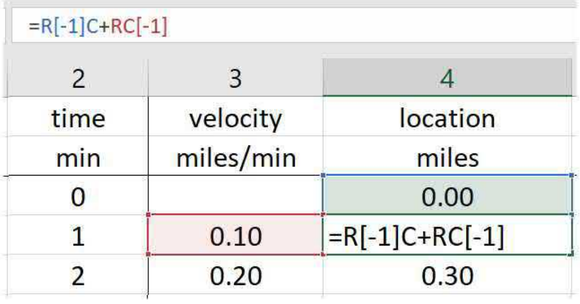
Our ability to produce an explicit formula for the nth term of the sequence of sums of a known sequence is an exception, not a rule.

Example 6.4.18: sums are displacements

We can use computers to speed up these computations. For example, one may have been recording one's velocities and now looking for the location. This is a formula for a spreadsheet (the locations):

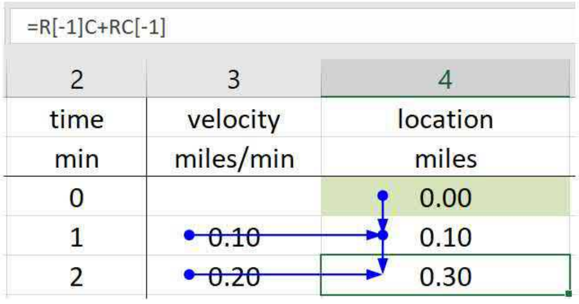
`=R[-1]C+RC[-1]`

Whether the sequence comes from a formula or it's just a list of numbers, the formula applies:



=R[-1]C+RC[-1]

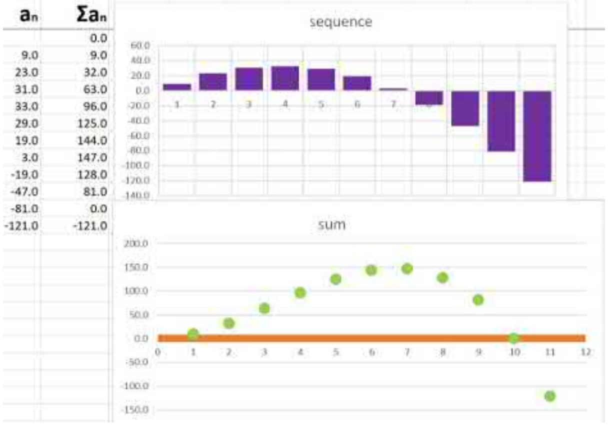
2	3	4
time	velocity	location
min	miles/min	miles
0		0.00
1	0.10	=R[-1]C+RC[-1]
2	0.20	0.30



=R[-1]C+RC[-1]

2	3	4
time	velocity	location
min	miles/min	miles
0		0.00
1	0.10	0.10
2	0.20	0.30

As a result, a curve has produced a new curve:



a_n	Σa_n
	0.0
9.0	9.0
23.0	32.0
31.0	63.0
33.0	96.0
29.0	125.0
19.0	144.0
3.0	147.0
-19.0	128.0
-47.0	81.0
-81.0	0.0
-121.0	-121.0

Exercise 6.4.19

Describe what has happened referring to, separately, to the first graph and the second graph.

Exercise 6.4.20

Imagine that the first column of the spreadsheet is where you have been recording your monthly deposit/withdrawals at your bank account. What does the second column represent? Describe what has been happening referring, separately, to the first graph and the second graph.

This is the time for some *theory*.

Recall from the last section this pair of obvious statements about motion:

- I am standing still `IF AND ONLY IF` my velocity is zero.

If the velocity is represented by a sequence, its sum is the location. We can then restate the above mathematically.

Theorem 6.4.21: Constant Sequence as Sum

The sequence of sums of a sequence is constant `IF AND ONLY IF` *the sequence has only zero values starting from some index N.*

In other words, we have:

$$\sum_{k=m}^{n} a_k \text{ is constant } \iff a_n = 0 \text{ for all } n \geq N\,.$$

Proof.

$$\sum_{k=m}^{n} a_k = c \text{ for all } n \iff a_{n+1} = \sum_{k=m}^{n+1} a_k - \sum_{k=m}^{n} a_k = c - c = 0 \iff a_{n+1} = 0\,.$$

Here is another equivalence statements about motion:

- I am moving forward `IF AND ONLY IF` my velocity is positive.

We can restate this mathematically using the sums.

Theorem 6.4.22: Monotonicity of Sum

The sequence of sums of a sequence is increasing `IF AND ONLY IF` *the terms of the sequence are non-negative.*

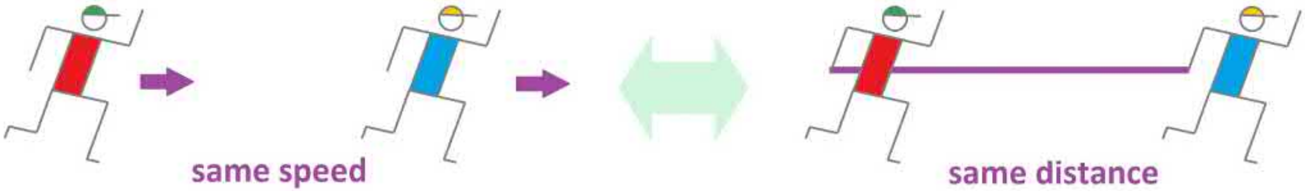
In other words, we have:

$$\begin{array}{llll} \displaystyle\sum_{k=m}^{n} a_k & \text{is increasing} & \iff a_n & \geq 0\,. \\ \displaystyle\sum_{k=m}^{n} a_k & \text{is decreasing} & \iff a_n & \leq 0\,. \end{array}$$

Proof.

$$\sum_{k=m}^{n+1} a_k \geq \sum_{k=m}^{n} a_k \text{ for all } n \iff a_{n+1} = \sum_{k=m}^{n+1} a_k - \sum_{k=m}^{n} a_k \geq 0\,.$$

Now suppose, just like in the last section, that there are *two* runners:



Then, we have:

- ▶ The distance between two runners isn't changing `IF AND ONLY IF` they are running with the same velocity.

We restate this mathematically.

Corollary 6.4.23: Subtracting Sums of Sequences

The sequences of sums of two sequences differ by a constant `IF AND ONLY IF` *the sequences are equal starting with some term* N.

In other words, we have:

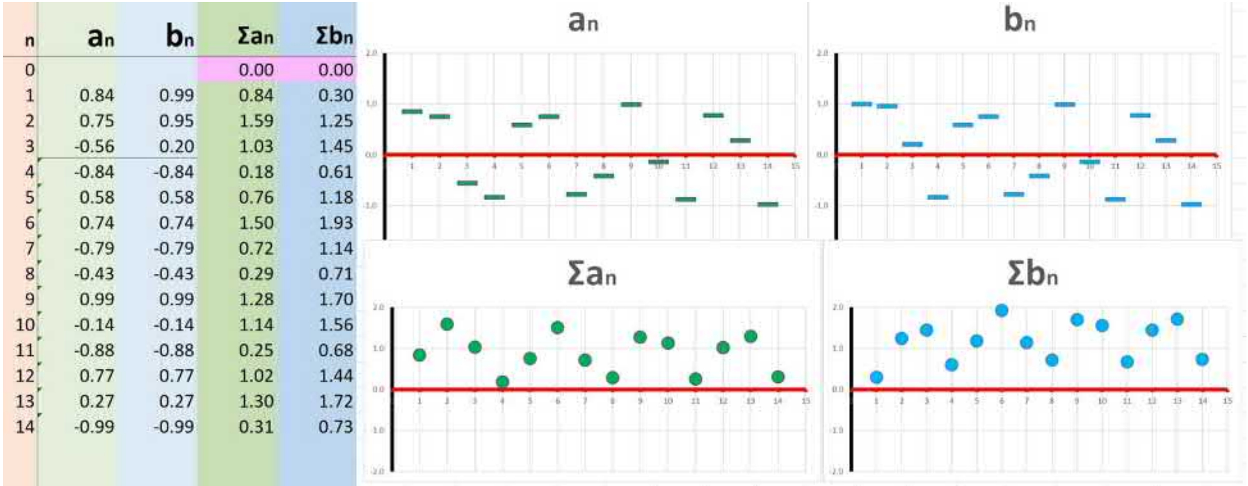
$$\sum_{k=m}^{n} a_k - \sum_{k=m}^{n} b_k \text{ is constant} \iff a_n = b_n \text{ for all } n \geq N\,.$$

Proof.

The corollary follows from the *Constant Sequence as Sum* above.

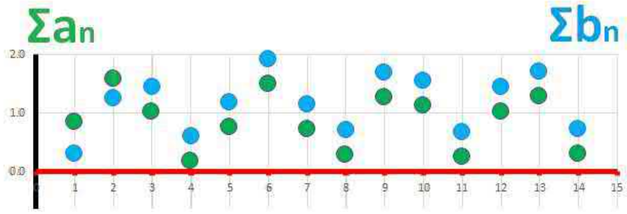
Example 6.4.24: shift of sequence

We have below two different sequences a_n and b_n that become identical after 3 terms:



n	aₙ	bₙ	Σaₙ	Σbₙ
0			0.00	0.00
1	0.84	0.99	0.84	0.30
2	0.75	0.95	1.59	1.25
3	-0.56	0.20	1.03	1.45
4	-0.84	-0.84	0.18	0.61
5	0.58	0.58	0.76	1.18
6	0.74	0.74	1.50	1.93
7	-0.79	-0.79	0.72	1.14
8	-0.43	-0.43	0.29	0.71
9	0.99	0.99	1.28	1.70
10	-0.14	-0.14	1.14	1.56
11	-0.88	-0.88	0.25	0.68
12	0.77	0.77	1.02	1.44
13	0.27	0.27	1.30	1.72
14	-0.99	-0.99	0.31	0.73

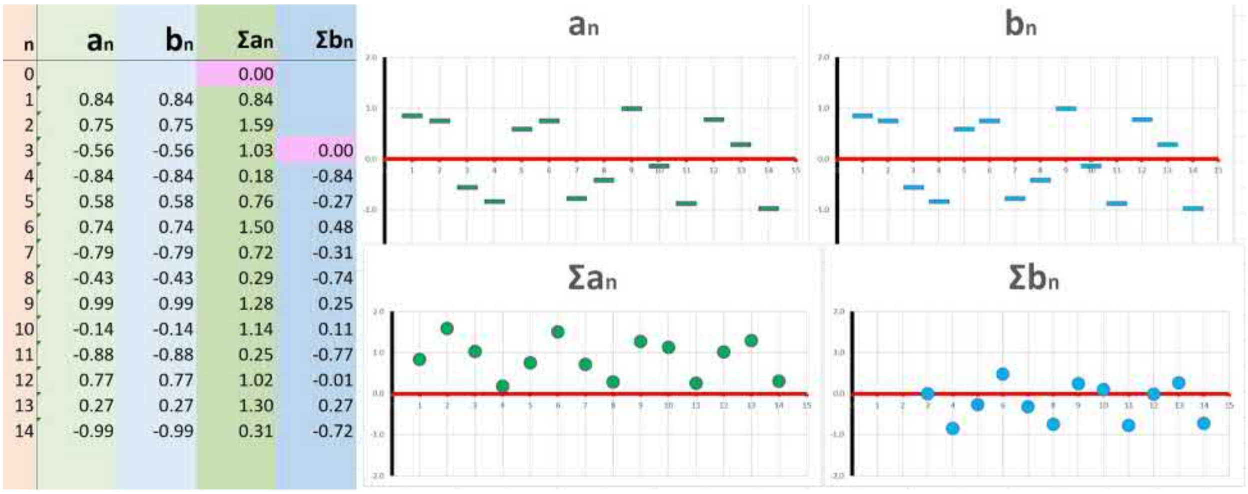
The result is that the sum of the latter sequence is just a vertical shift of the sum of the former:



To state this algebraically, we have for each n:

$$\sum_{k=m}^{n} a_k = \sum_{k=m}^{n} b_k + C\,,$$

where C is some number. The outcome is the same when the two sequences a_n and b_n are identical but the computation of their sequences of sums starts at different points:



n	a_n	b_n	Σa_n	Σb_n
0			0.00	
1	0.84	0.84	0.84	
2	0.75	0.75	1.59	
3	-0.56	-0.56	1.03	0.00
4	-0.84	-0.84	0.18	-0.84
5	0.58	0.58	0.76	-0.27
6	0.74	0.74	1.50	0.48
7	-0.79	-0.79	0.72	-0.31
8	-0.43	-0.43	0.29	-0.74
9	0.99	0.99	1.28	0.25
10	-0.14	-0.14	1.14	0.11
11	-0.88	-0.88	0.25	-0.77
12	0.77	0.77	1.02	-0.01
13	0.27	0.27	1.30	0.27
14	-0.99	-0.99	0.31	-0.72

Exercise 6.4.25

What is the meaning of the number C?

We can use the theorems to watch for the distance between the two runners:

- ▶ The distance from one of the two runners to the other is increasing `IF AND ONLY IF` the former's velocity is higher.

We can restate this mathematically using the sums.

Corollary 6.4.26: Subtracting Sums: Monotonicity

The difference of the sequences of sums of two sequences is increasing `IF AND ONLY IF` *the corresponding terms of the former are larger than or equal to those of the latter.*

In other words, we have:

$$\sum_{k=m}^{n} a_k - \sum_{k=m}^{n} b_k \quad \text{is increasing} \quad \iff \quad a_n \ \geq \ b_n \,.$$
$$\sum_{k=m}^{n} a_k - \sum_{k=m}^{n} b_k \quad \text{is decreasing} \quad \iff \quad a_n \ \leq \ b_n \,.$$

Proof.

The corollary follows from *Monotonicity of Sum* above.

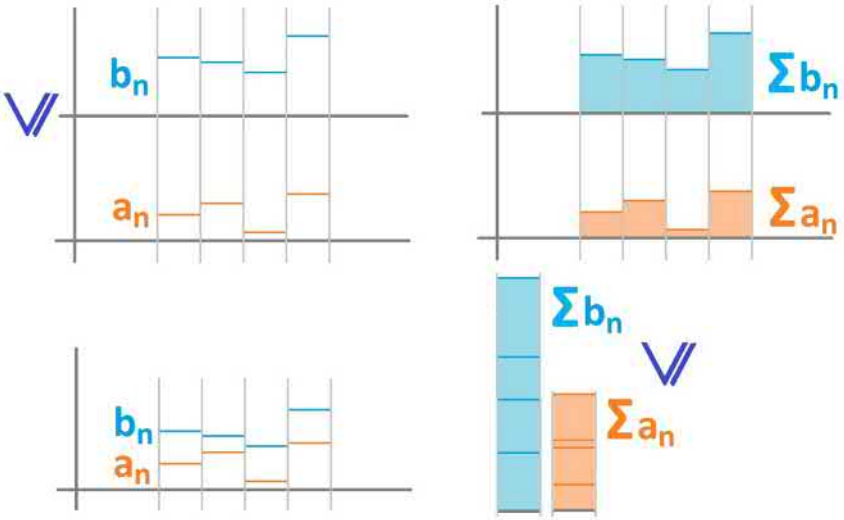
Here is another way to look at the statement *the faster covers the longer distance.* It is about comparing the values of two sums. Consider this simple algebra:

$$\begin{array}{lcl} a & \leq & b \\ A & \leq & B \\ \hline a+A & \leq & b+B \end{array}$$

The rule applies even if we have more than just two terms:

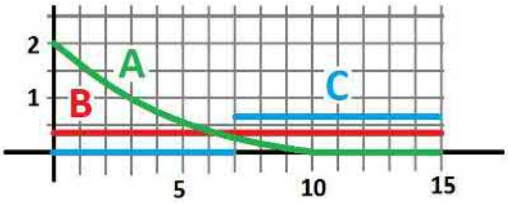
$$\begin{array}{rcl} a_m & \leq & b_m \\ a_{m+1} & \leq & b_{m+1} \\ \vdots & \vdots & \vdots \\ a_q & \leq & b_q \\ \hline a_m + ... + a_q & \leq & b_m + ... + b_q \end{array}$$

The summation is illustrated below:



Example 6.4.27: three runners, continued

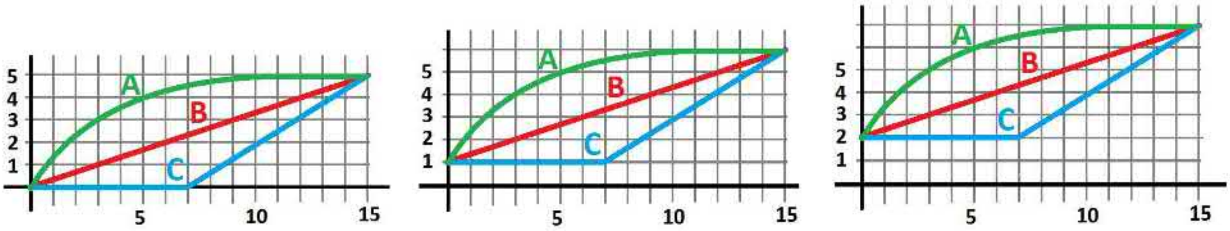
The graph shows the velocities of three runners in terms of time, n:



It's easy to describe *how* they are moving:

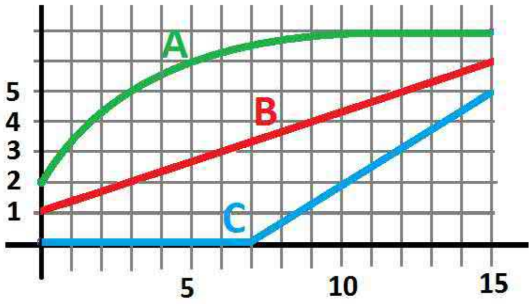
- A starts fast and the slows down.
- B maintains the same speed.
- C starts late and then runs fast.

But *where* are they, at every moment? There are several possible answers:



Which one is the right one depends on the starting point. Of course, a simple examination of the first graph doesn't prove that the three runners will arrive at the finish line at the same time.

Furthermore, if the requirement that they all start at the same location is lifted, the result will be different, for example:



Exercise 6.4.28

Suggest other graphs that match the description above.

Exercise 6.4.29

Plot the location and the velocity for the following trip: "I drove slowly, gradually speed up, stopped for a very short moment, then started but in the opposite direction, quickly accelerated, and from that point maintained the speed." Make up your own story and repeat the task.

Exercise 6.4.30

Draw a curve on a piece of paper, imagine that it represents your velocity, and then sketch what your locations would look like. Repeat.

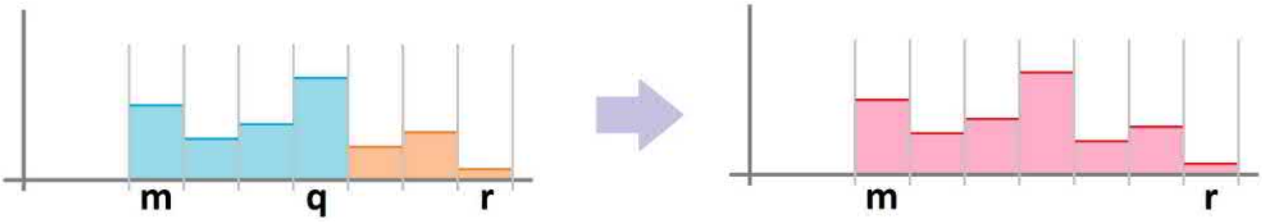
Here is another trivial statement about *motion*:

$$\begin{array}{rl} & \text{distance covered during the 1st hour} \\ + & \text{distance covered during the 2nd hour} \\ = & \text{distance during the two hours} \end{array}$$

The statement is about the fact that when adding, we can change the order of terms freely; this is called the *Associativity Property* of addition. At its simplest, it allows us to remove the parentheses:

$$\begin{array}{rlcr} & (a_m + a_{m+1} + ... + a_{q-1} + a_q) & + & (a_{q+1} + a_{q+2} + ... + a_{r-1} + a_r) \\ = & a_m + a_{m+1} + ... + a_{q-1} + a_q & + & a_{q+1} + a_{q+2} + ... + a_{r-1} + a_r \\ = & a_m + a_{m+1} + & ... & +a_{r-1} + a_r\,. \end{array}$$

The three sums are shown below:



An abbreviated version of this identity is as follows.

Theorem 6.4.31: Additivity for Sums

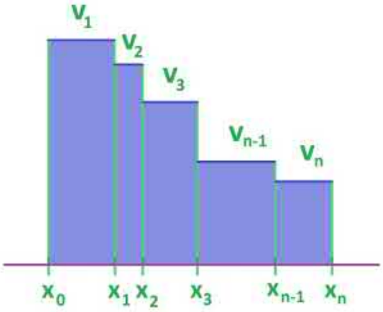
The sum of the sums of two consecutive segments of a sequence is the sum of the combined segment.

In other words, for any sequence a_n and for any m, q, r with $m \le q \le r$, we have:

$$\sum_{k=m}^{q} a_k + \sum_{k=q+1}^{r} a_k = \sum_{k=m}^{r} a_k$$

How do we deal with motion when the time moments aren't integers? What is the *displacement* then?

Suppose x_n is the sequence of locations and v_n the sequence of velocities:



Then their *Riemann sum* is defined to be the sequence of sums of the sequence of the product of v_n and the difference of x_n:

$$\sum_{k=1}^{n} v_n \, \Delta x_n \, .$$

We know from the last section that if y_n is the position, then the velocity is $v_n = \Delta y_n / \Delta x_n$. Therefore, the Riemann sum of the sequence of the velocity with respect to the sequence of time is the displacement.

Typically, the sequence x_n is fixed. This makes the Riemann sum quotient a function defined on the sequences of length n, just as the sum. We can write this function as follows:

$$\sum y \, \Delta x : \mathbf{R}^n \to \mathbf{R}^n \, .$$

Exercise 6.4.32

Prove analogues of the Sum Rule and the Constant Multiple Rule for the Riemann sum. Derive that the Riemann sum is a linear operator.

Exercise 6.4.33

Express the Riemann sum operator in terms of the sum operator.

6.5. Sums of differences and differences of sums

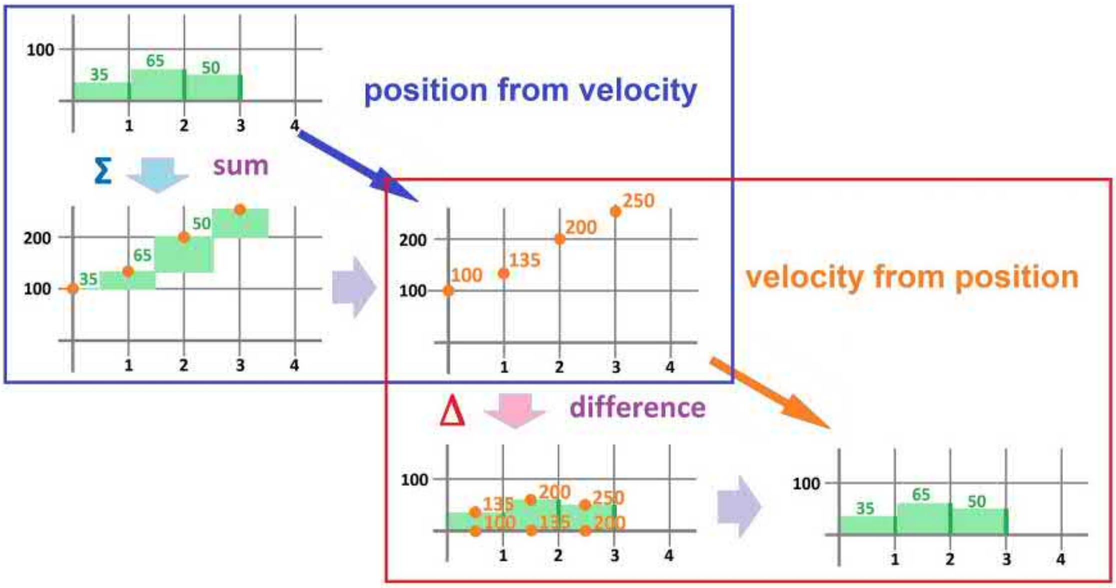
We know that *addition and subtraction undo each other*; it makes sense then that the operations for making the sequence of differences and making the sequence of sums will cancel each other too!

Example 6.5.1: broken odometer – broken speedometer

We know how to get the velocity from the location, and the location from the velocity. We expect that executing these two operations consecutively should bring us back where we started.

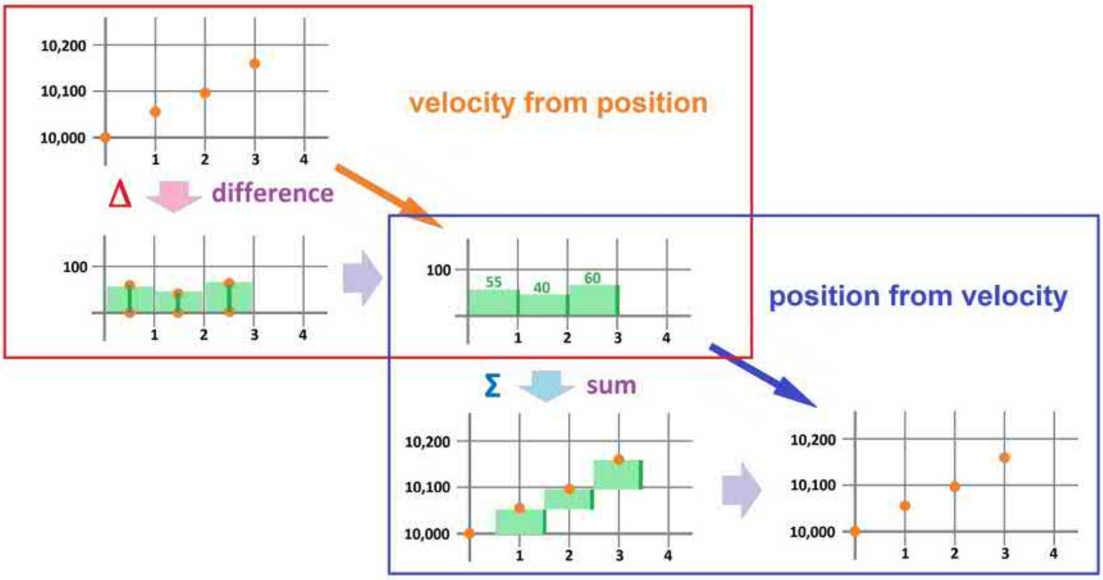
Let's take another look at the example of *two* computations about motion – a broken odometer and a broken speedometer – presented in the beginning of this chapter. The terminology has now been developed: Every time we speak of a sequence, we also speak of the sequence of its differences and the sequence of its sums.

In the first diagram, one first takes the velocity data and acquires the displacements via the sums, then someone else takes this displacement data and acquires the velocities by using the differences:



We are back to the original sequence.

In the second diagram, one first takes the location data and acquires the velocities via the differences, then someone else takes this velocity data and acquires the locations by using the sums:



We are back to the original sequence (provided we start at the same initial value).

Example 6.5.2: sequences given by lists

Below we have a sequence given by a list. We compute its sequence sums and then compute the sequence of differences of the result:

$$\begin{array}{rcccccccccccccc}
\text{a sequence:} & 3 & & 1 & & & & 0 & & & & -2 & & & \ldots \\
 & \downarrow & & \downarrow & & & & \downarrow & & & & \downarrow & & & \ldots \\
\text{its sums:} & 3 & & & & & & & & & & & & & \\
 & 3 & + & 1 & = & 4 & & & & & & & & & \\
 & & & & & 4 & + & 0 & = & 4 & \ldots & & & & \\
 & & & & & & & & & 4 & + & (-2) & = & 2 & \ldots \\
 & & & & & & & & & & & & & 2 & \ldots \\
 & \downarrow & & & & \downarrow & & & & \downarrow & & & & \downarrow & \ldots \\
\text{the sequence of sums:} & 3 & & & & 4 & & & & 4 & & & & 2 & \ldots \\
 & & \searrow & & \swarrow & & \searrow & & \swarrow & & \searrow & & \swarrow & & \ldots \\
\text{the differences:} & & & 4-3 & & & & 4-4 & & & & 2-4 & & & \ldots \\
 & & & \Vert & & & & \Vert & & & & \Vert & & & \ldots \\
\text{a new sequence:} & & & 1 & & & & 0 & & & & -2 & & & \ldots
\end{array}$$

We are back to the original sequence!

Exercise 6.5.3

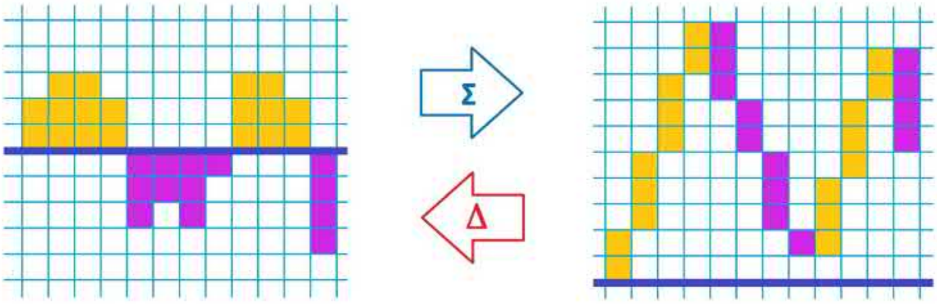
What happened to the very first term?

Exercise 6.5.4

Start with the sequence in the last example and use the diagrams to show that the sums of the differences give us the original sequence.

Example 6.5.5: sequences given by graphs

Just comparing the illustrations above demonstrates that the two operations – the difference and the sum – undo the effect of each other. The two operations are shown together below:



As you can see in the picture, the sum (left to right) stacks up the terms of the sequence on top of each other, while the difference (right to left) takes these apart.

Let's take care of the *algebra*.

These are the two facts we will be using:

1. Suppose we have a sequence, a_n. We compute its *difference*, a new sequence:

$$b_{n+1} = a_{n+1} - a_n \,.$$

2. Suppose we have a sequence, c_k. We compute its *sum*, a new sequence:

$$d_n = \sum_{k=1}^{n} c_k \,,$$

or, recursively:

$$d_{n+1} = d_n + c_{n+1} \,.$$

We use this setup to answer the following two questions.

The first question we would like to answer is:

- ▶ *What is the difference of the sum?*

We start with c_n. Then, we have from (2) and (1), respectively:

$$d_{n+1} = d_n + c_{n+1} \quad \text{and} \quad b_{n+1} = d_{n+1} - d_n \,.$$

We substitute the first formula into the second (and then cancel):

$$b_{n+1} = d_{n+1} - d_n = (d_n + c_{n+1}) - d_n = c_{n+1} \,.$$

As we can see, the answer is:

- ▶ *The original sequence.*

The second question we would like to answer is:

- ▶ *What is the sum of the difference?*

We start with a_n. Then, we have from (1) and (2), respectively:

$$b_k = a_k - a_{k-1} \quad \text{and} \quad d_n = \sum_{k=1}^{n} b_k \,.$$

We substitute the first formula into the second (and then cancel):

$$d_n = \sum_{k=1}^{n} b_k = \sum_{k=1}^{n} (a_k - a_{k-1}) = (a_2 - a_1) + (a_3 - a_2) + (a_4 - a_3) + ... + (a_n - a_{n-1}) = -a_1 + a_n \,.$$

As we can see, the answer is:

- ▶ *The original sequence plus a number.*

We summarize these results in the form of the following two far-reaching theorems.

Theorem 6.5.6: Fundamental Theorem of Calculus of Sequences I

The difference of the sum of a sequence is that sequence; i.e., for all n, we have:

$$\Delta\left(\sum_{k=1}^{n} a_k\right) = a_n$$

The two operations *cancel* each other!

Theorem 6.5.7: Fundamental Theorem of Calculus of Sequences II

The sum of the difference of a sequence is that sequence plus a constant number; i.e., for all n, we have:

$$\sum_{k=1}^{n}\left(\Delta b_k\right)=b_n+C$$

The two operations – almost – cancel each other, again!

Now “operations” become “operators”.

The idea is to look at the *composition of the difference and the sum operators.*

The inputs and outputs of the difference operator

$$\Delta:\mathbf{R}^{p}\to\mathbf{R}^{p-1}$$

will have to be matched with the outputs and inputs of the sum operator

$$\Sigma:\mathbf{R}^{q}\to\mathbf{R}^{q+1}\,.$$

This is what happens when we put them together:

$$\begin{tikzcd}
\mathbf{R}^{n+1} \arrow[r, "\Delta"] & \mathbf{R}^{n} \arrow[r, "\Sigma"] & \mathbf{R}^{n+1} \\
\mathbf{R}^{n} \arrow[r, "\Sigma"] & \mathbf{R}^{n+1} \arrow[r, "\Delta"] & \mathbf{R}^{n}
\end{tikzcd}$$

The compositions $\Sigma\circ\Delta$ and $\Delta\circ\Sigma$ now make sense.

Next, according to the above theorems, they “cancel” each other, but not entirely. Are they *inverses*? No.

Example 6.5.8: inverses?

No, Δ and Σ aren’t the inverses of each other, because Δ isn’t one-to-one. We just need to think of how two objects can move differently but with the same velocity ($n=2$):

$$\Delta<0,1>=<1>,\ \Delta<1,2>=<1>\ .$$

On the other hand, we have:

$$\Sigma<1>=<0,1>\ .$$

So,

$$\Sigma\big(\Delta<1,2>\big)=<0,1>\neq<1,2>\ .$$

Therefore,

$$\Sigma\circ\Delta\neq I\,.$$

In order to *find* inverses here, we change one of the spaces: We fix the first coordinate in $\mathbf{R}^{n+1}$. It’s as if the starting position is always 0. Let’s denote this space as follows:

Hyperplane

$$\mathbf{R}^{n+1}_0=\{<0,x_1,...,x_n>\}\subset\mathbf{R}^{n+1}\,.$$

It’s is a copy of $\mathbf{R}^{n}$.

This move changes the domain of Δ (we have its restriction) and the codomain (but not the image) of Σ. We still have two compositions:

$$\begin{tikzcd}
\mathbf{R}^{n+1}_0 \arrow[r, "\Delta"] & \mathbf{R}^{n} \arrow[r, "\Sigma"] & \mathbf{R}^{n+1}_0 \\
\mathbf{R}^{n} \arrow[r, "\Sigma"] & \mathbf{R}^{n+1}_0 \arrow[r, "\Delta"] & \mathbf{R}^{n}
\end{tikzcd}$$

Now, the difference and the sum are given by:

$$\begin{aligned}
\Delta < 0, x_1, ..., x_n > \quad &=< x_1, x_2 - x_1, ..., x_n - x_{n-1} > \\
\Sigma < x_1, x_2, ..., x_n > \quad &=< 0, x_1\,,\ x_1 + x_2\,,\ ...\ ,\ x_1 + x_2 + ... + x_{n-1} + x_n\ >
\end{aligned}$$

This is how their values on the standard basis change:

$$\Delta \begin{bmatrix} 1 \\ 0 \\ 0 \\ ... \\ 0 \\ 0 \end{bmatrix} = \begin{bmatrix} \\ -1 \\ 0 \\ ... \\ 0 \\ 0 \end{bmatrix}, \Delta \begin{bmatrix} 0 \\ 1 \\ 0 \\ ... \\ 0 \\ 0 \end{bmatrix} = \begin{bmatrix} \\ 1 \\ -1 \\ ... \\ 0 \\ 0 \end{bmatrix}, ..., \Delta \begin{bmatrix} 0 \\ 0 \\ 0 \\ ... \\ 0 \\ 1 \end{bmatrix} = \begin{bmatrix} \\ 0 \\ 0 \\ ... \\ 0 \\ 1 \end{bmatrix},$$

$$\Sigma \begin{bmatrix} \\ 1 \\ 0 \\ ... \\ 0 \\ 0 \end{bmatrix} = \begin{bmatrix} 0 \\ 1 \\ 1 \\ ... \\ 1 \\ 1 \end{bmatrix}, \Sigma \begin{bmatrix} \\ 0 \\ 1 \\ ... \\ 0 \\ 0 \end{bmatrix} = \begin{bmatrix} 0 \\ 1 \\ 1 \\ ... \\ 1 \\ 1 \end{bmatrix}, ..., \Sigma \begin{bmatrix} \\ 0 \\ 0 \\ ... \\ 0 \\ 1 \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ 0 \\ ... \\ 0 \\ 1 \end{bmatrix}.$$

This is what happens to their matrices:

$$\Delta = \begin{bmatrix} -1 & 1 & 0 & 0 & ... & 0 & 0 \\ 0 & -1 & 1 & 0 & ... & 0 & 0 \\ 0 & 0 & -1 & 1 & ... & 0 & 0 \\ ... \\ 0 & 0 & 0 & 0 & ... & 1 & 0 \\ 0 & 0 & 0 & 0 & ... & -1 & 1 \end{bmatrix} \qquad \Sigma = \begin{bmatrix} 0 & 0 & 0 & 0 & ... & 0 & 0 \\ 1 & 0 & 0 & 0 & ... & 0 & 0 \\ 1 & 1 & 0 & 0 & ... & 0 & 0 \\ ... \\ 1 & 1 & 1 & 1 & ... & 1 & 0 \\ 1 & 1 & 1 & 1 & ... & 1 & 1 \end{bmatrix}$$

These are *square* matrices now!

The linear algebra analogs of the above theorems are below:

Corollary 6.5.9: I

The composition of the difference operator $\Delta : \mathbf{R}_0^{n+1} \to \mathbf{R}^n$ *and the sum operator* $\Sigma : \mathbf{R}^n \to \mathbf{R}_0^{n+1}$ *is the identity:*

$$\boxed{\Delta\Sigma = I}$$

Proof.

An alternative proof is by matrix multiplication. We start the multiplication with the first row and the first column:

$$\begin{aligned}
\Delta\Sigma &= \begin{bmatrix} 1 & 0 & 0 & ... & 0 & 0 \\ -1 & 1 & 0 & ... & 0 & 0 \\ 0 & -1 & 1 & ... & 0 & 0 \\ ... \\ 0 & 0 & 0 & ... & 1 & 0 \\ 0 & 0 & 0 & ... & -1 & 1 \end{bmatrix} \begin{bmatrix} 1 & 0 & 0 & 0 & ... & 0 \\ 1 & 1 & 0 & 0 & ... & 0 \\ 1 & 1 & 1 & 0 & ... & 0 \\ ... \\ 1 & 1 & 1 & 1 & ... & 0 \\ 1 & 1 & 1 & 1 & ... & 1 \end{bmatrix} \\
&= \begin{bmatrix} 1\cdot 1 + 0\cdot 1 + ... + 0\cdot 1 + 0\cdot 1 & ... \\ & ... \\ & ... \\ & ... \\ ... & ... \\ & ... \\ & ... \end{bmatrix} = \begin{bmatrix} 1 & ... \\ & ... \\ & ... \\ & ... \\ ... & ... \\ & ... \\ & ... \end{bmatrix}
\end{aligned}$$

Exercise 6.5.10

Finish the proof.

Corollary 6.5.11: II

The composition of the sum operator $\Sigma : \mathbf{R}^n \to \mathbf{R}^{n+1}_0$ *and the difference operator* $\Delta : \mathbf{R}^{n+1}_0 \to \mathbf{R}^n$ *is the identity:*

$$\boxed{\Sigma\Delta = I}$$

Proof.

An alternative proof is by matrix multiplication. We start the multiplication with the first row and the first column:

$$\begin{aligned}
\Sigma\Delta &= \begin{bmatrix} 1 & 0 & 0 & 0 & ... & 0 \\ 1 & 1 & 0 & 0 & ... & 0 \\ 1 & 1 & 1 & 0 & ... & 0 \\ ... \\ 1 & 1 & 1 & 1 & ... & 0 \\ 1 & 1 & 1 & 1 & ... & 1 \end{bmatrix} \begin{bmatrix} 1 & 0 & 0 & ... & 0 & 0 \\ -1 & 1 & 0 & ... & 0 & 0 \\ 0 & -1 & 1 & ... & 0 & 0 \\ ... \\ 0 & 0 & 0 & ... & 1 & 0 \\ 0 & 0 & 0 & ... & -1 & 1 \end{bmatrix} \\
&= \begin{bmatrix} 1\cdot 1 + 0\cdot(-1) + ... + 0\cdot 0 & ... \\ & ... \\ & ... \\ ... & \\ & ... \\ & ... \end{bmatrix}
\end{aligned}$$

Exercise 6.5.12

Finish the proof.

We combine them together into this truly fundamental result:

Theorem 6.5.13: Fundamental Theorem of Calculus of Sequences

The sum operator $\Sigma : \mathbf{R}^n \to \mathbf{R}_0^{n+1}$ *and the difference operator* $\Delta : \mathbf{R}_0^{n+1} \to \mathbf{R}^n$ *are inverses of each other:*

$$\boxed{\Delta^{-1} = \Sigma}$$

The conclusion re-appears, in an almost identical form, in infinitesimal calculus.

Exercise 6.5.14

What are the eigenvectors of Δ? Of Σ?

Example 6.5.15: fundamental theorems, computed

For larger sets of data, we use a spreadsheet. Recall the formulas:

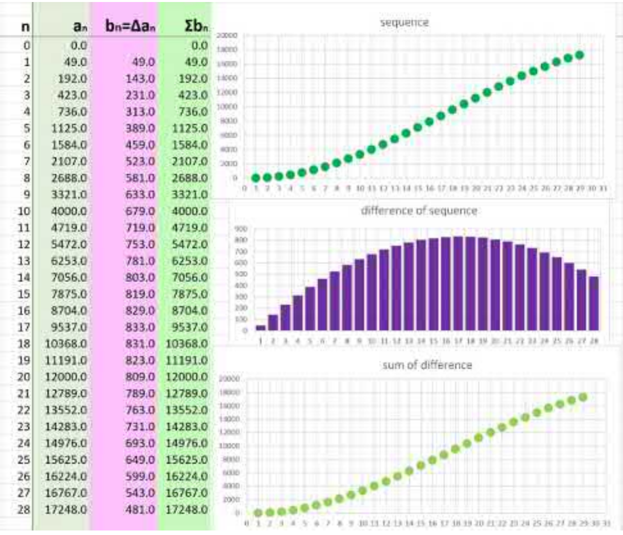
- From a sequence to its sum:

`=R[-1]C+RC[-1]`

- From a sequence to its difference:

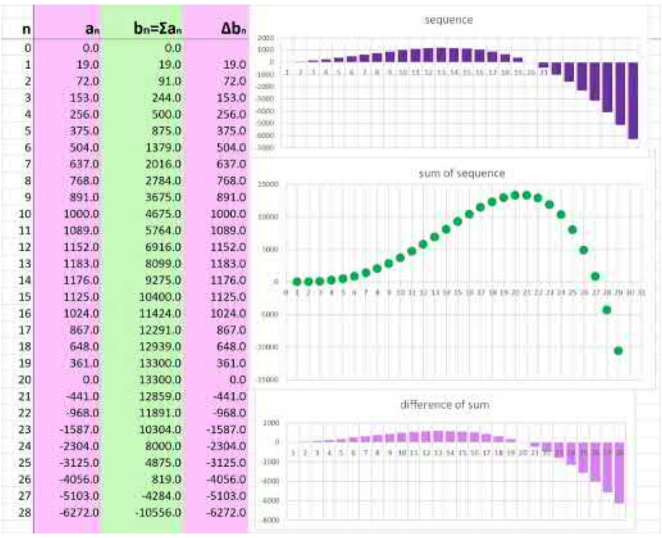
`=RC[-1]-R[-1]C[-1]`

What if we combine the two consecutively? From a sequence to its difference to the sum of the latter:



n	a_n	$b_n=\Delta a_n$	Σb_n
0	0.0		0.0
1	49.0	49.0	49.0
2	192.0	143.0	192.0
3	423.0	231.0	423.0
4	736.0	313.0	736.0
5	1125.0	389.0	1125.0
6	1584.0	459.0	1584.0
7	2107.0	523.0	2107.0
8	2688.0	581.0	2688.0
9	3321.0	633.0	3321.0
10	4000.0	679.0	4000.0
11	4719.0	719.0	4719.0
12	5472.0	753.0	5472.0
13	6253.0	781.0	6253.0
14	7056.0	803.0	7056.0
15	7875.0	819.0	7875.0
16	8704.0	829.0	8704.0
17	9537.0	833.0	9537.0
18	10368.0	831.0	10368.0
19	11191.0	823.0	11191.0
20	12000.0	809.0	12000.0
21	12789.0	789.0	12789.0
22	13552.0	763.0	13552.0
23	14283.0	731.0	14283.0
24	14976.0	693.0	14976.0
25	15625.0	649.0	15625.0
26	16224.0	599.0	16224.0
27	16767.0	543.0	16767.0
28	17248.0	481.0	17248.0

It's the same curve! Now in the opposite order, from a sequence to its sum to the difference of the latter:



n	a_n	$b_n=\Sigma a_n$	Δb_n
0	0.0	0.0	
1	19.0	19.0	19.0
2	72.0	91.0	72.0
3	153.0	244.0	153.0
4	256.0	500.0	256.0
5	375.0	875.0	375.0
6	504.0	1379.0	504.0
7	637.0	2016.0	637.0
8	768.0	2784.0	768.0
9	891.0	3675.0	891.0
10	1000.0	4675.0	1000.0
11	1089.0	5764.0	1089.0
12	1152.0	6916.0	1152.0
13	1183.0	8099.0	1183.0
14	1176.0	9275.0	1176.0
15	1125.0	10400.0	1125.0
16	1024.0	11424.0	1024.0
17	867.0	12291.0	867.0
18	648.0	12939.0	648.0
19	361.0	13300.0	361.0
20	0.0	13300.0	0.0
21	-441.0	12859.0	-441.0
22	-968.0	11891.0	-968.0
23	-1587.0	10304.0	-1587.0
24	-2304.0	8000.0	-2304.0
25	-3125.0	4875.0	-3125.0
26	-4056.0	819.0	-4056.0
27	-5103.0	-4284.0	-5103.0
28	-6272.0	-10556.0	-6272.0

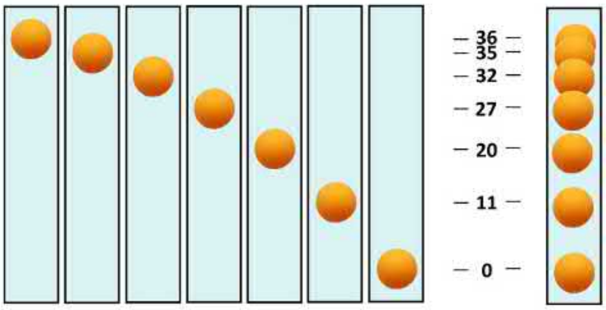
It's the same curve!

Exercise 6.5.16

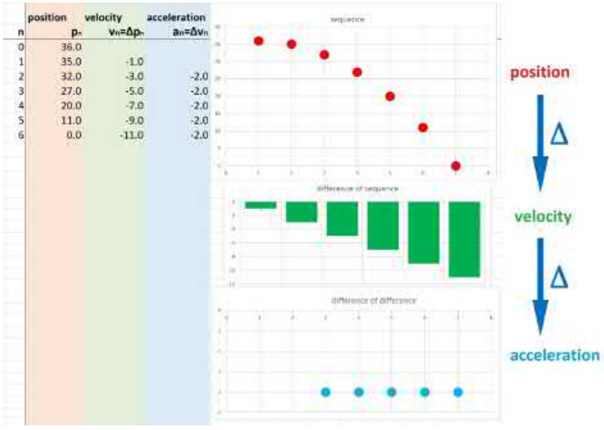
What would the resulting curve look like if we started at another point?

Example 6.5.17: falling ball, acceleration

Consider the experimental data of the heights of a ping-pong ball falling down:



Just as before, we use a spreadsheet to plot the *location* sequence, p_n (green). We then compute the difference of p_n, i.e., the *velocity*, v_n (purple):



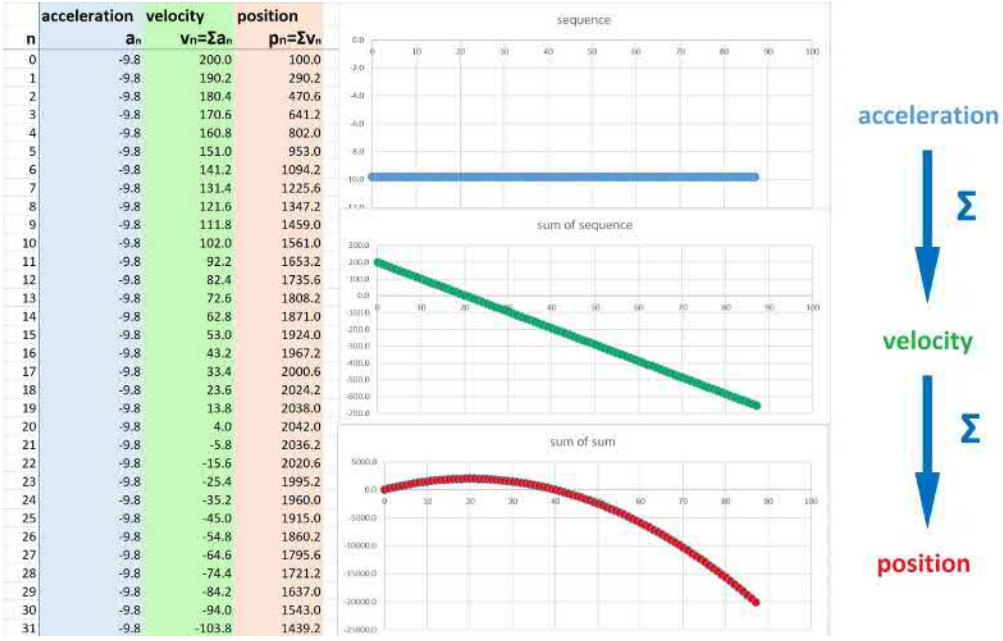
	position	velocity	acceleration
n	p_n	$v_n=\Delta p_n$	$a_n=\Delta v_n$
0	36.0		
1	35.0	-1.0	
2	32.0	-3.0	-2.0
3	27.0	-5.0	-2.0
4	20.0	-7.0	-2.0
5	11.0	-9.0	-2.0
6	0.0	-11.0	-2.0

It looks like a straight line. But this time, we take one more step: We compute the difference of the velocity sequence. It is the *acceleration*, a_n (blue). It appears constant! There might be a law of nature here.

Example 6.5.18: shooting a cannon

Let's accept the premise put forward in the last example, that *the acceleration of free fall is constant.* Then we can try to predict the behavior of an object shot in the air – from any initial height and with any initial velocity.

The direction of our computation is opposite to that of the last example: We assume that we know the acceleration, then derive the velocity, and then derive the location (altitude) of the object in time. While we used *differences* in the last example, we use *sums* now:



	acceleration	velocity	position
n	a_n	$v_n=\Sigma a_n$	$p_n=\Sigma v_n$
0	-9.8	200.0	100.0
1	-9.8	190.2	290.2
2	-9.8	180.4	470.6
3	-9.8	170.6	641.2
4	-9.8	160.8	802.0
5	-9.8	151.0	953.0
6	-9.8	141.2	1094.2
7	-9.8	131.4	1225.6
8	-9.8	121.6	1347.2
9	-9.8	111.8	1459.0
10	-9.8	102.0	1561.0
11	-9.8	92.2	1653.2
12	-9.8	82.4	1735.6
13	-9.8	72.6	1808.2
14	-9.8	62.8	1871.0
15	-9.8	53.0	1924.0
16	-9.8	43.2	1967.2
17	-9.8	33.4	2000.6
18	-9.8	23.6	2024.2
19	-9.8	13.8	2038.0
20	-9.8	4.0	2042.0
21	-9.8	-5.8	2036.2
22	-9.8	-15.6	2020.6
23	-9.8	-25.4	1995.2
24	-9.8	-35.2	1960.0
25	-9.8	-45.0	1915.0
26	-9.8	-54.8	1860.2
27	-9.8	-64.6	1795.6
28	-9.8	-74.4	1721.2
29	-9.8	-84.2	1637.0
30	-9.8	-94.0	1543.0
31	-9.8	-103.8	1439.2

Above we show a projectile launched from a 100-meter tall building vertically up in the air with a speed of 100 meters per second (the gravity causes acceleration of -9.8 meters per second squared). We can see that it will reach its highest point in about 20 seconds and will hit the ground in about 40 seconds.

Exercise 6.5.19

How high does the projectile go in the above example?

Exercise 6.5.20

Using the above example, how long will it take for the projectile to reach the ground if fired *down*?

Exercise 6.5.21

Use the above model to determine how long it will take for an object to reach the ground if it is dropped. Make up your own questions about the situation and answer them. Repeat.

Exercise 6.5.22

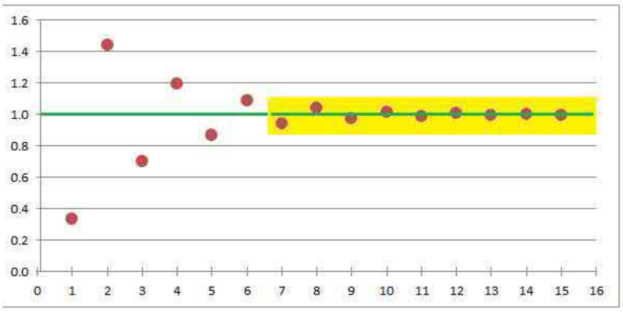
Suppose the time moments are given by another sequence (an arithmetic progression). Compute the velocity and the acceleration from the table below:

	time	height
n	t_n	a_n
1	.00	36
2	.05	35
3	.10	32
4	.15	25
5	.20	20
6	.25	11
7	.30	0

6.6. Limits as linear operators

We continue with *infinitesimal calculus*.

The starting point is the *limits of infinite sequences*, i.e., their long term trends:



We use the following notation:

Limit of sequence

$$a_n \to a$$

All sequences below are infinite.

Example 6.6.1: limits

A few examples with predictable patterns of long-term behavior:

	list							nth-term formula
(1)	1	1/2	1/3	1/4	1/5	...	$\to 0$	$1/n$
(2)	.9	.99	.999	.9999	.99999	...	$\to 1$	$1-10^{-n}$
(3)	1.	1.1	1.01	1.001	1.0001	...	$\to 1$	$1+10^{-n}$
(4)	3.	3.1	3.14	3.141	3.1415	...	$\to \pi$	
(5)	1	2	3	4	5	...	$\to +\infty$	n
(6)	0	1	0	1	0	...	$\to$ nothing	

But what is the limit?

Definition 6.6.2: limit of sequence

We call number a the *limit* of a sequence a_n if the following condition holds:

- ▶ For each real number $\varepsilon > 0$, there exists a number N such that for every natural number $n > N$, we have:

$$|a_n - a| < \varepsilon\,.$$

If a sequence has a limit, then we call the sequence *convergent* and say that it *converges*; otherwise, it is *divergent* and we say it *diverges*.

The definition is somewhat complex. The good news is:

- ▶ It won't matter here what the exact definition of the limit is.

What matters is this:

1. The limit (when it exists) is a number.
2. The limits behave under algebraic properties in a predictable way (below).

Then:

1. The limits taken together is a real-valued function the inputs of which are the convergent sequences:

$$\text{sequence} \quad \to \quad \boxed{\quad \lim \quad} \quad \to \quad \text{number}$$

2. This function satisfies certain algebraic properties.

The properties of limits that we know from calculus tell us about the nature of this function:

Theorem 6.6.3: Sum Rule for Limits of Sequences

1. *If sequences a_n, b_n converge, then so does $a_n + b_n$.*
2. *Furthermore, we have:*

$$\lim_{n\to\infty} (a_n + b_n) = \lim_{n\to\infty} a_n + \lim_{n\to\infty} b_n$$

or

$$a_n \to a,\ b_n \to b \implies a_n + b_n \to a + b \ \text{ as } n \to \infty\,.$$

The former notation for the statement of the Sum Rule matches the following method of function notation:

$$F(x + y) = F(x) + F(y)\,.$$

The latter matches the other:

$$x \mapsto X,\ y \mapsto Y \implies x + y \mapsto X + Y\,.$$

The same match is also visible below:

Theorem 6.6.4: Constant Multiple Rule for Limits of Sequences

1. *If sequence a_n converges, then so does ka_n for any real k.*
2. *Furthermore, we have:*

$$\lim_{n\to\infty} \Big(k \cdot a_n\Big) = k \cdot \lim_{n\to\infty} a_n$$

or

$$a_n \to a \implies k \cdot a_n \to ka \ \text{ as } n \to \infty\,.$$

This is what we see in the two theorems:

1. Parts 1 match: A linear combination of two convergent sequences is convergent.
2. Parts 2 match: Limits preserve linear combinations.

We conclude the following, which is the purpose of this investigation:

Theorem 6.6.5: Linear Algebra of Limits of Sequences

1. *The convergent sequences form a vector space.*
2. *The limit is a real-valued linear operator on this space:*

$$\boxed{\lim :\ a_n \mapsto a}$$

We have a function with a complicated name...

Exercise 6.6.6

Provide details of the proof.

Exercise 6.6.7

Is this function one-to-one or onto?

Unfortunately, replacing sequences with vectors,

$$a_1, a_2, \dots \ \mapsto < a_1, a_2, \dots > ,$$

does not help to find a match for our vector space this time. This isn't a Euclidean space!

Therefore, we have two vectors spaces:

$$\text{convergent sequences} \quad \subset \quad \text{sequences}$$

It is also important to remember that they are linked by the *inclusion*:

$$i: \text{ convergent sequences } \ \to \ \text{sequences}$$

Therefore, all linear operators defined on the space of sequences will also be defined on the space of convergent sequences via restriction.

Next, the *sums of series.*

With what we have learned about limits, we don't need to invoke calculus results anymore. We go in reverse: linear algebra, then calculus.

Every sequence produces another sequence, its sequence of sums:

$$a_n \ \mapsto \ b_n = a_1 + a_2 + \dots + a_n \, .$$

It is also known as the partial sum in this context.

This is, of course, our function,

$$\Sigma: \text{ sequences } \ \to \ \text{sequences}$$

Definition 6.6.8: sum of the series

The *sum of a series* given by a sequence $a_1, a_2, \dots$ is defined to be the limit of the sequence of its sums:

$$\boxed{\sum^{\infty} a_n = \lim_{n\to\infty} (a_1 + a_2 + \dots + a_n)}$$

Then the series is called *convergent.*

We have, therefore, two linear operators executed consecutively:

$$\sum^{\infty}: \quad \text{convergent series} \quad \xrightarrow{\Sigma} \quad \text{convergent sequences} \quad \xrightarrow{\lim} \quad \mathbf{R}$$

The fact that the composition of two linear operators is a linear operator proves the following:

Theorem 6.6.9: Linear Algebra of Sums of Series

1. *The convergent series form a vector space.*

2. *The sum is a real-valued linear operator on this space:*

$$\boxed{\sum^{\infty} = \lim \circ \Sigma : \ a_n \mapsto a}$$

We have another function with a complicated name...

Exercise 6.6.10

Provide details of the proof.

Exercise 6.6.11

Is this function one-to-one or onto?

Familiar results from calculus follow from the theorem:

Theorem 6.6.12: Sum Rule for Sums of Series

1. *If series* $\sum^{\infty} a_n, \sum^{\infty} b_n$ *converge, then so does* $\sum^{\infty} (a_n + b_n)$.
2. *Furthermore, we have:*

$$\sum^{\infty} (a_n + b_n) = \sum^{\infty} a_n + \sum^{\infty} b_n$$

or

$$\sum^{\infty} a_n = a, \ \sum b_n = b \implies \sum^{\infty} (a_n + b_n) = a + b\,.$$

Theorem 6.6.13: Constant Multiple Rule for Sums of Series

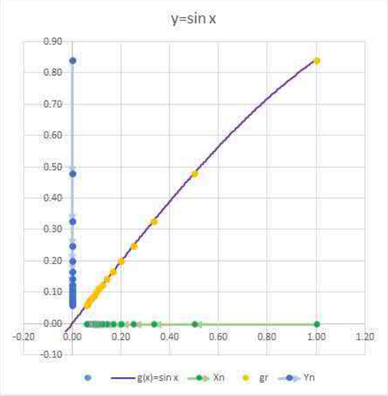
1. *If series* $\sum a_n$ *converges, then so does* $\sum k a_n$ *for any real* k.
2. *Furthermore, we have:*

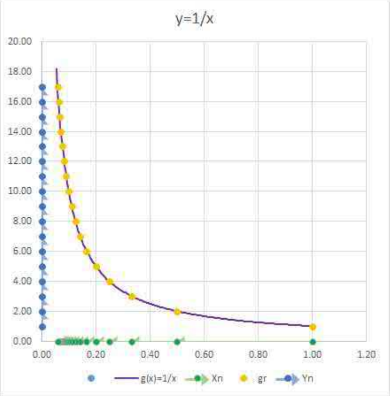
$$\sum \left(k \cdot a_n\right) = k \cdot \sum a_n$$

or

$$\sum a_n = a \implies \sum k \cdot a_n = ka\,.$$

Consider the *limits of functions* next:





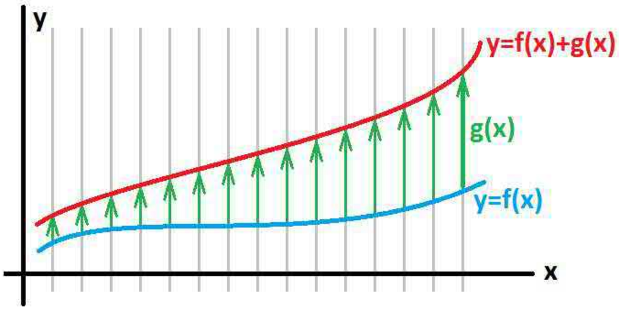
Once again, it doesn't matter what the definition of the limit of a function is. What matters is that once $x = a$ is fixed, the limit is a number if it exists. Taken together, these numbers form a real-valued function the inputs of which are convergent at a functions:

$$\text{function} \quad \to \quad \boxed{\lim_{x\to a}} \quad \to \quad \text{number}$$

The reason is that there can be only one limit.

The properties of limits are translated into the properties of this function.

Let's consider addition of functions. Here, g serves as a vertical "push" of the graph of f. The picture below is meant to illustrate that idea. There are ping-pong balls arranged in a curve, f, on the ground and there is also wind, g. Then, the wind, non-uniformly but continuously, blows them forward:



The ping-pong balls remain arranged in a curve, $f + g$.

Theorem 6.6.14: Sum Rule of Limits of Functions

For each a, we have:

1. *If the limits at a of functions f, g exist, then so does that of their sum, $f + g$.*
2. *Furthermore, the limit of the sum is equal to the sum of the limits:*

$$\lim_{x\to a}\big(f(x)+g(x)\big) = \lim_{x\to a} f(x) + \lim_{x\to a} g(x)$$

or

$$f(x)\to F,\ g(x)\to G \implies f(x)+g(x)\to F+G \quad \text{as } x\to a\,.$$

Theorem 6.6.15: Constant Multiple Rule of Limits of Functions

For each a, we have:

1. *If the limit at a of function f exists, then so does that of its multiple, kf.*
2. *Furthermore, the limit of the multiple is equal to the multiple of the limit:*

$$\lim_{x\to a} kf(x) = k\cdot\lim_{x\to a} f(x)$$

or

$$f(x)\to F \implies k\cdot f\to k\cdot F \quad \text{as } x\to a\,.$$

Parts 1 match: A linear combination of two convergent functions is convergent. And Parts 2 match too: Limits preserve linear combinations.

We conclude the following, which is the purpose of this investigation:

Theorem 6.6.16: Linear Algebra of Limits of Functions

For each a, we have:

1. *The numerical functions convergent at $x = a$ form a vector space.*

2. *The limit is a real-valued linear operator on this space:*

$$\boxed{\lim_{x\to a} : f \mapsto L}$$

We have another function with a complicated name...

Exercise 6.6.17

Provide details of the proof.

Exercise 6.6.18

Is this function one-to-one or onto?

We can also derive the following:

Corollary 6.6.19: Linear Algebra of Continuous Functions

The continuous on a particular set $X \subset \mathbf{R}$ real-valued functions form a vector space.

Therefore, we have two vectors spaces:

$$\text{continuous functions} \quad \subset \quad \text{functions}$$

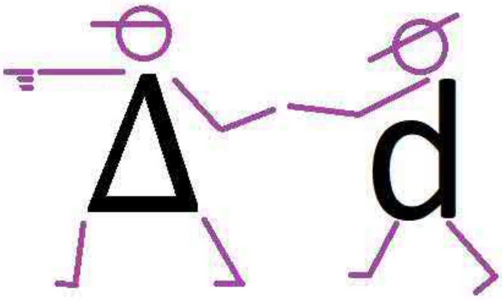
It is also important to remember that they are linked by the *inclusion*, which is a linear operator:

$$i: \ \text{continuous functions} \quad \to \quad \text{functions}$$

Therefore, all linear operators defined on the space of functions will also be defined on the space of continuous functions via restriction.

6.7. *Differentiation as a linear operator*

The *derivative* of a function at some $x = a$ is defined to be the limit of the difference quotients of this function. That why its algebra is similar to that of the difference operator:



It doesn't matter what the definition is exactly, but differentiation,

$$f \mapsto f'(a)\,,$$

is a real-valued function function of functions:

$$\text{function} \quad \to \quad \boxed{\left.\frac{d}{dx}\right|_{x=a}} \quad \to \quad \text{number}$$

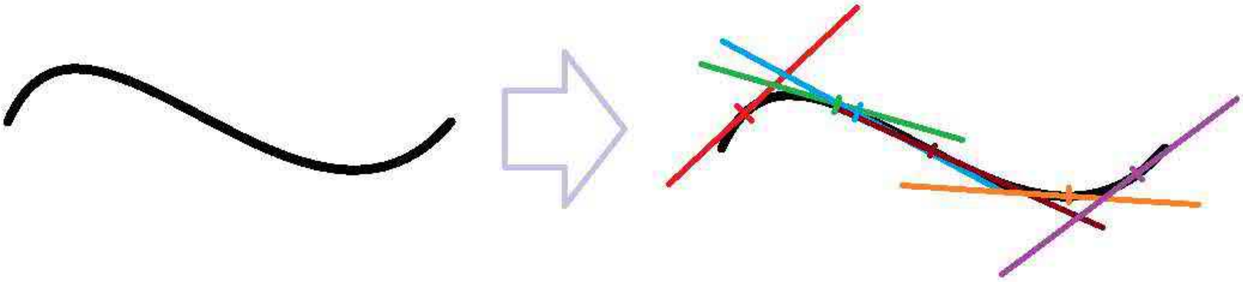
Its domain is the functions differentiable at a.

Exercise 6.7.1

Is this function linear? One-to-one? Onto?

We take this to the next level: The derivative is a *function*.

Differentiation is a function-valued function defined on all differentiable (on a given interval) functions:



It is seen as the collection of all tangent lines.

The input of this function is a differentiable function f, and the output is another function f':

$$f \mapsto f' .$$

What this means is that this process is a special kind of function too:

$$\text{function} \quad \to \quad \boxed{\frac{d}{dx}} \quad \to \quad \text{function}$$

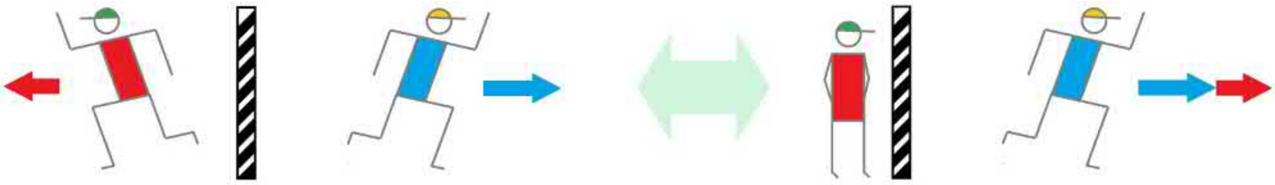
Its domain is still the differentiable at a functions, but the codomain consists of functions.

Now the algebra.

Consider this elementary statement about *motion*:

▶ `IF` two runners are running away from a post, `THEN` their relative velocity is the sum of their respective velocities.

It's as if the one runner is standing still while the other is running for the both of them with the combined speed:

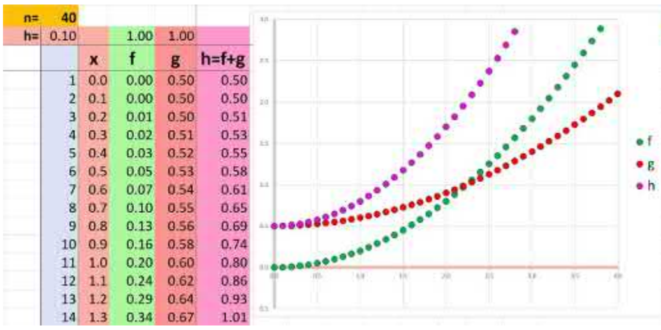


Theorem 6.7.2: Sum Rule for Derivatives

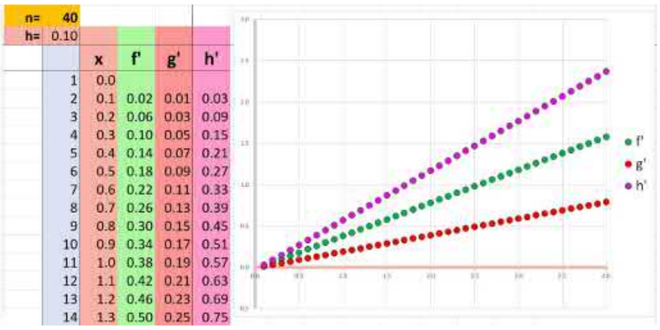
1. *The sum of two functions differentiable at a point is differentiable at that point.*
2. *For any two functions f, g differentiable at x, we have at x:*

$$\frac{d}{dx}(f+g) = \frac{d}{dx}(f) + \frac{d}{dx}(g) .$$

It can be illustrated with the following:



n=	40				
h=	0.10		1.00	1.00	
		x	f	g	h=f+g
	1	0.0	0.00	0.50	0.50
	2	0.1	0.00	0.50	0.50
	3	0.2	0.01	0.50	0.51
	4	0.3	0.02	0.51	0.53
	5	0.4	0.03	0.52	0.55
	6	0.5	0.05	0.53	0.58
	7	0.6	0.07	0.54	0.61
	8	0.7	0.10	0.55	0.65
	9	0.8	0.13	0.56	0.69
	10	0.9	0.16	0.58	0.74
	11	1.0	0.20	0.60	0.80
	12	1.1	0.24	0.62	0.86
	13	1.2	0.29	0.64	0.93
	14	1.3	0.34	0.67	1.01



n=	40				
h=	0.10				
		x	f'	g'	h'
	1	0.0			
	2	0.1	0.02	0.01	0.03
	3	0.2	0.06	0.03	0.09
	4	0.3	0.10	0.05	0.15
	5	0.4	0.14	0.07	0.21
	6	0.5	0.18	0.09	0.27
	7	0.6	0.22	0.11	0.33
	8	0.7	0.26	0.13	0.39
	9	0.8	0.30	0.15	0.45
	10	0.9	0.34	0.17	0.51
	11	1.0	0.38	0.19	0.57
	12	1.1	0.42	0.21	0.63
	13	1.2	0.46	0.23	0.69
	14	1.3	0.50	0.25	0.75

Next, consider:

▶ If the distances are proportionally increased, then so are the velocities needed to cover them, in the same period of time.

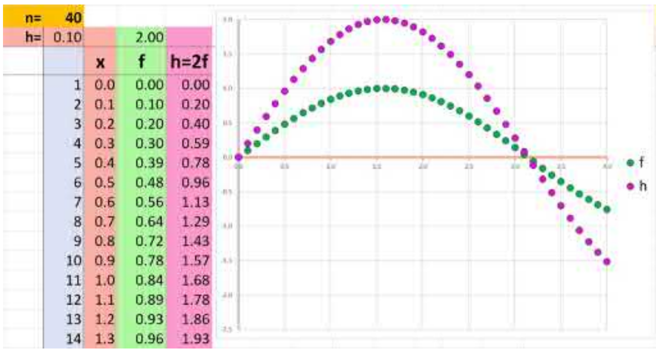
More precisely, we have:

Theorem 6.7.3: Constant Multiple Rule for Derivatives

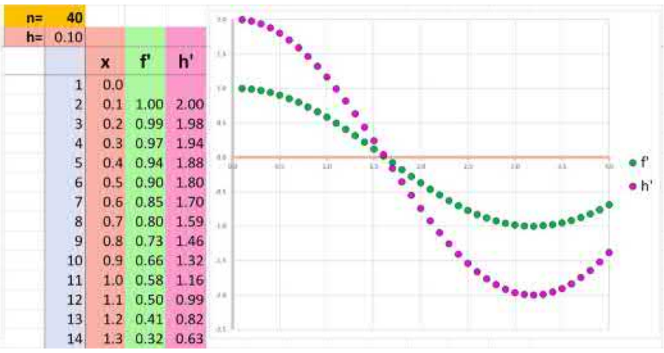
1. *A multiple of a function differentiable at a point is differentiable at that point.*
2. *For any function f differentiable at x and any real k, we have at x:*

$$\frac{d}{dx}(kf)=k\frac{d}{dx}(f)\,.$$

It can be illustrated with the following:



n=	40			
h=	0.10		2.00	
		x	f	h=2f
	1	0.0	0.00	0.00
	2	0.1	0.10	0.20
	3	0.2	0.20	0.40
	4	0.3	0.30	0.59
	5	0.4	0.39	0.78
	6	0.5	0.48	0.96
	7	0.6	0.56	1.13
	8	0.7	0.64	1.29
	9	0.8	0.72	1.43
	10	0.9	0.78	1.57
	11	1.0	0.84	1.68
	12	1.1	0.89	1.78
	13	1.2	0.93	1.86
	14	1.3	0.96	1.93

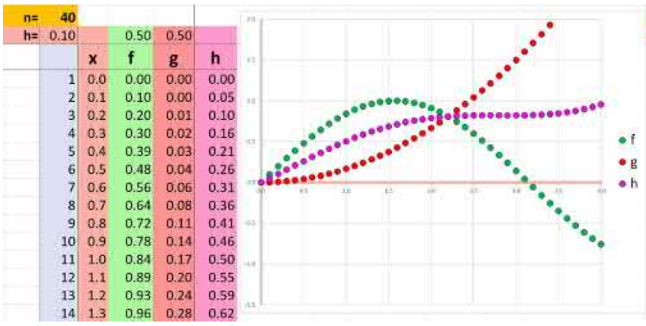


n=	40			
h=	0.10			
		x	f'	h'
	1	0.0		
	2	0.1	1.00	2.00
	3	0.2	0.99	1.98
	4	0.3	0.97	1.94
	5	0.4	0.94	1.88
	6	0.5	0.90	1.80
	7	0.6	0.85	1.70
	8	0.7	0.80	1.59
	9	0.8	0.73	1.46
	10	0.9	0.66	1.32
	11	1.0	0.58	1.16
	12	1.1	0.50	0.99
	13	1.2	0.41	0.82
	14	1.3	0.32	0.63

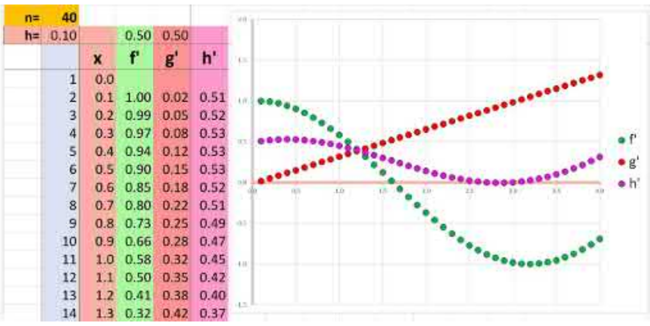
These two operations can be combined into one producing *linear combinations*:

$$\alpha x+\beta y\,,$$

where α,β are two constant numbers. For example, this is the average of two functions (left):



n=	40				
h=	0.10		0.50	0.50	
		x	f	g	h
	1	0.0	0.00	0.00	0.00
	2	0.1	0.10	0.00	0.05
	3	0.2	0.20	0.01	0.10
	4	0.3	0.30	0.02	0.16
	5	0.4	0.39	0.03	0.21
	6	0.5	0.48	0.04	0.26
	7	0.6	0.56	0.06	0.31
	8	0.7	0.64	0.08	0.36
	9	0.8	0.72	0.11	0.41
	10	0.9	0.78	0.14	0.46
	11	1.0	0.84	0.17	0.50
	12	1.1	0.89	0.20	0.55
	13	1.2	0.93	0.24	0.59
	14	1.3	0.96	0.28	0.62



n=	40				
h=	0.10		0.50	0.50	
		x	f'	g'	h'
	1	0.0			
	2	0.1	1.00	0.02	0.51
	3	0.2	0.99	0.05	0.52
	4	0.3	0.97	0.08	0.53
	5	0.4	0.94	0.12	0.53
	6	0.5	0.90	0.15	0.53
	7	0.6	0.85	0.18	0.52
	8	0.7	0.80	0.22	0.51
	9	0.8	0.73	0.25	0.49
	10	0.9	0.66	0.28	0.47
	11	1.0	0.58	0.32	0.45
	12	1.1	0.50	0.35	0.42
	13	1.2	0.41	0.38	0.40
	14	1.3	0.32	0.42	0.37

We also notice what happens to their derivatives (right):

▶ The derivative of the average is the average of the derivatives.

Of course, we conclude the following first:

1. The functions differentiable at a particular $x=a$ form a vector space.

2. The derivative at $x=a$ is a real-valued linear operator on this space:

$$\boxed{\left.\frac{d}{dx}\right|_{x=a}: f \mapsto f'(a)}$$

The following is a more profound conclusion:

Theorem 6.7.4: Linear Algebra of Differentiation

1. *The functions differentiable on a particular interval I form a vector space.*
2. *The derivative is a linear operator on this space with values in functions defined on I:*

$$\boxed{\frac{d}{dx}: f \mapsto f'}$$

Exercise 6.7.5

Is the operator one-to-one? Onto?

What happens to linear combinations of functions under differentiation? These two formulas can be combined into one:

$$\boxed{\frac{d}{dx}(\alpha f+\beta g) \quad = \alpha\frac{d}{dx}(f) \quad +\beta\frac{d}{dx}(g)}$$

The last formula is illustrated with the following diagram:

$$\alpha f+\beta g \quad \to \quad \boxed{\frac{d}{dx}} \quad \to \quad \alpha f'+\beta g'$$

We have now three vectors spaces connected by the inclusion operators:

$$\text{differentiable functions} \quad \to \quad \text{continuous functions} \quad \to \quad \text{functions}$$

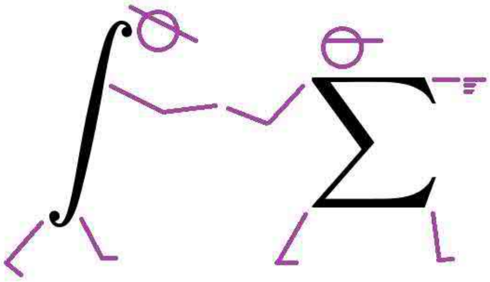
Therefore, all linear operators defined on the space of functions will also be defined on the space of differentiable functions via restriction.

Exercise 6.7.6

(a) Prove that the set of all infinitely many times differentiable functions is a vector space. (b) Show that the derivative is a linear operator on this space. (c) On this space, what are the eigenvalues and eigenvectors?

6.8. Integration as a linear operator

The *definite integral* of a function over an interval $[a,b]$ is defined to be the limit (in a certain sense) of the Riemann sums of this function. That why its algebra is similar to that of the sum operator:



It doesn't matter what the definition is exactly, but integration,

$$f \mapsto \int_a^b f\,dx\,,$$

is a real-valued function of functions:

$$\text{function} \quad \to \quad \boxed{\int_a^b} \quad \to \quad \text{number}$$

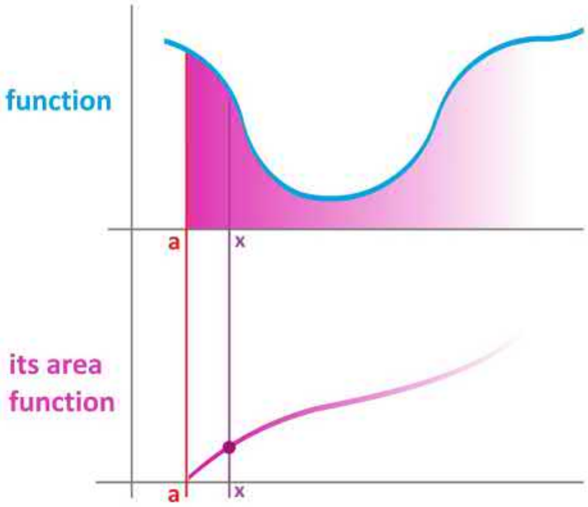
Its domain is the functions integrable over $[a,b]$.

Exercise 6.8.1

Is this function linear? One-to-one? Onto?

We take this to the next level: The indefinite integral is a *function*.

Integration is a function-valued function defined on all integrable (over a given interval) functions:



It is seen as a way to organize the areas of *all* regions under the graph.

The input of this function is a integrable function f, and the output is another function $\int f\,dx$:

$$f \mapsto \int f\,dx\,.$$

Since this is a definite integral, we choose a specific antiderivative:

$$F(x) = \int_a^x f\,dx\,.$$

In other words,

$$F(a) = 0\,.$$

Just as in the case of the sum operator, this will be crucial for the linear algebra treatment.

What this means is that this process is a special kind of function too:

$$\text{function} \ \to \ \boxed{\int} \ \to \ \text{function}$$

Its domain is still the functions integrable over $[a,b]$, but the codomain consists of functions.

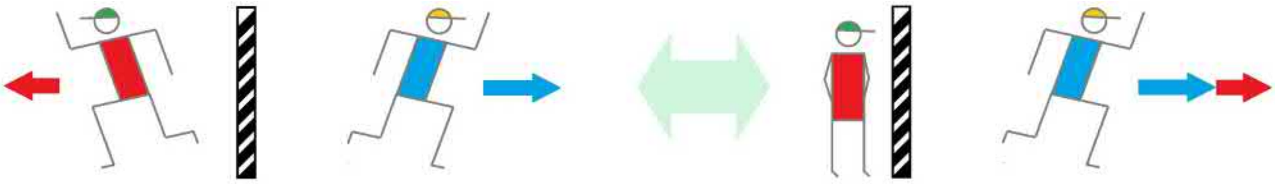
Furthermore, the last item on either list comes from the algebraic properties of the limits. In fact, the ideas come from those for *limits*: the Sum Rule and the Constant Multiple Rule. The question we will be asking is the following:

▶ What happens to the output function of integration as we perform algebraic operations with the input functions?

Consider this elementary statement about *motion*:

▶ `IF` two runners are running away from a post, `THEN` the distance between them is the sum of their distances to the post.

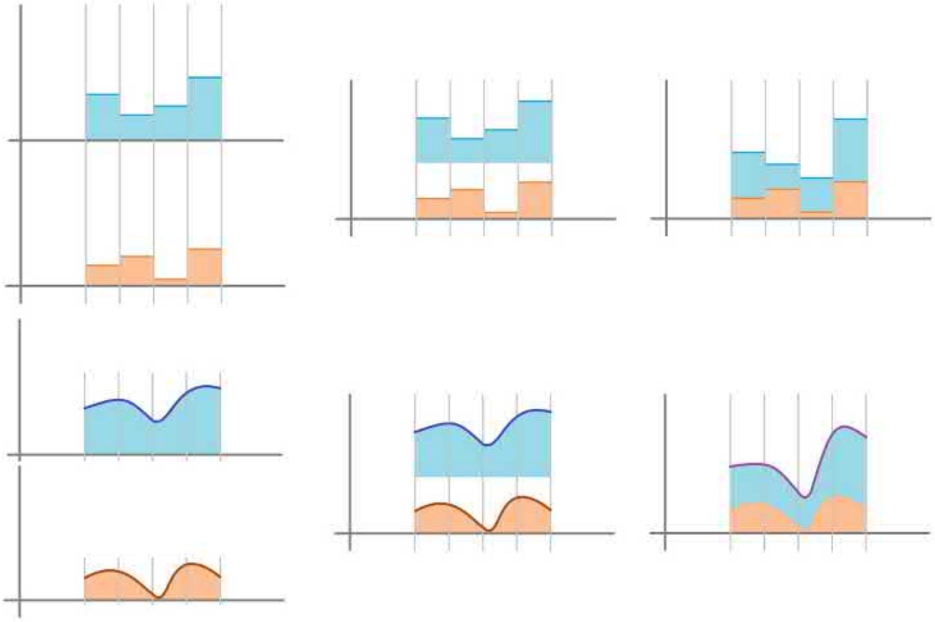
It's as if the one runner is standing still while the other is running for the both of them with the combined speed:



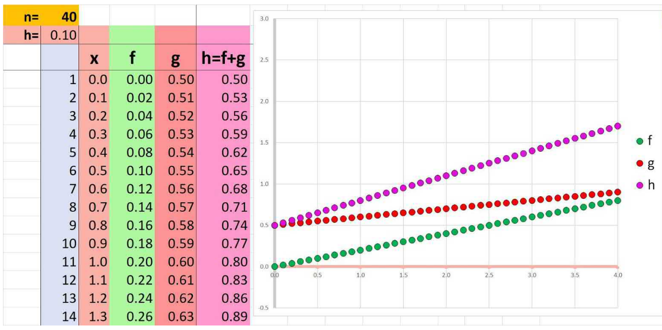
Theorem 6.8.2: Sum Rule for Integrals

1. *The sum of two functions integrable over an interval is integrable over that interval.*
2. *For any two functions f,g integrable over an interval, we have over the interval:*
$$\int (f+g)\,dx = \int (f)\,dx + \int (g)\,dx\,.$$

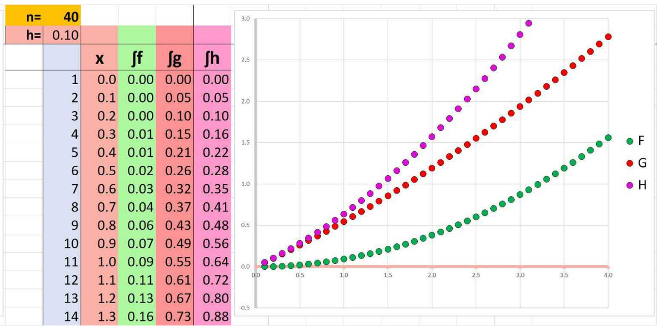
The picture below illustrates what happens when the bottom drops from a bucket of sand and it falls on a uneven surface:



The theorem can also be demonstrated with the spreadsheet:



n=	40				
h=	0.10				
		x	f	g	h=f+g
	1	0.0	0.00	0.50	0.50
	2	0.1	0.02	0.51	0.53
	3	0.2	0.04	0.52	0.56
	4	0.3	0.06	0.53	0.59
	5	0.4	0.08	0.54	0.62
	6	0.5	0.10	0.55	0.65
	7	0.6	0.12	0.56	0.68
	8	0.7	0.14	0.57	0.71
	9	0.8	0.16	0.58	0.74
	10	0.9	0.18	0.59	0.77
	11	1.0	0.20	0.60	0.80
	12	1.1	0.22	0.61	0.83
	13	1.2	0.24	0.62	0.86
	14	1.3	0.26	0.63	0.89



n=	40				
h=	0.10				
		x	∫f	∫g	∫h
	1	0.0	0.00	0.00	0.00
	2	0.1	0.00	0.05	0.05
	3	0.2	0.00	0.10	0.10
	4	0.3	0.01	0.15	0.16
	5	0.4	0.01	0.21	0.22
	6	0.5	0.02	0.26	0.28
	7	0.6	0.03	0.32	0.35
	8	0.7	0.04	0.37	0.41
	9	0.8	0.06	0.43	0.48
	10	0.9	0.07	0.49	0.56
	11	1.0	0.09	0.55	0.64
	12	1.1	0.11	0.61	0.72
	13	1.2	0.13	0.67	0.80
	14	1.3	0.16	0.73	0.88

Next, when a function is multiplied by a constant, what happens to its sums?

Consider:

> ▶ If the distances are proportionally increased, then so are the velocities needed to cover them, in the same period of time.

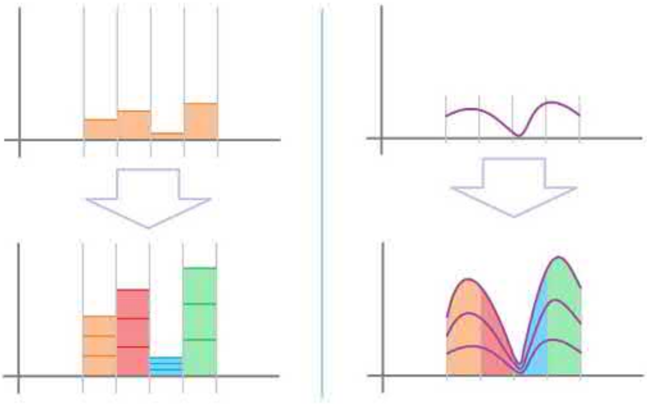
More precisely, we have:

Theorem 6.8.3: Constant Multiple Rule for Integrals

1. *A multiple of a function integrable over an interval is integrable over that interval.*
2. *For any function f integrable over an interval and any real k, we have over the interval:*

$$\int (kf)\, dx = k \int (f)\, dx\,.$$

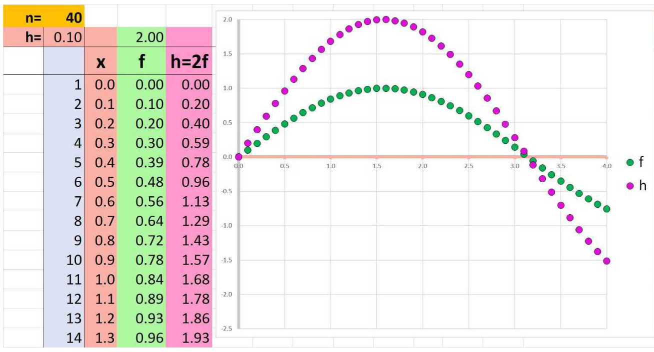
The picture below illustrates the idea that tripling the height of a road will need tripling the amount of soil under it:



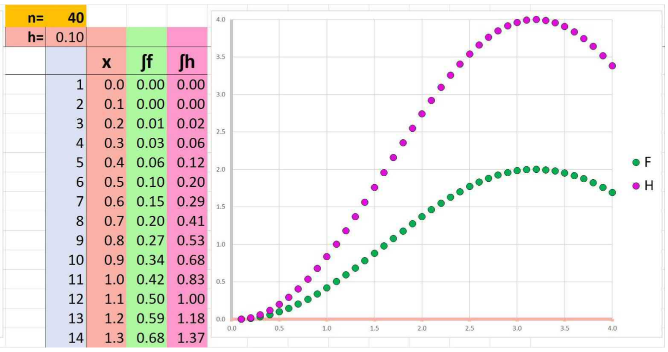
The last two theorems demonstrate that this is true whether the surface is staircase-like or curved.

For the *motion* metaphor, if your velocity is tripled, then so is the distance you have covered over the same period of time.

The theorem can also be demonstrated with the spreadsheet:



n=	40			
h=	0.10		2.00	
		x	f	h=2f
	1	0.0	0.00	0.00
	2	0.1	0.10	0.20
	3	0.2	0.20	0.40
	4	0.3	0.30	0.59
	5	0.4	0.39	0.78
	6	0.5	0.48	0.96
	7	0.6	0.56	1.13
	8	0.7	0.64	1.29
	9	0.8	0.72	1.43
	10	0.9	0.78	1.57
	11	1.0	0.84	1.68
	12	1.1	0.89	1.78
	13	1.2	0.93	1.86
	14	1.3	0.96	1.93

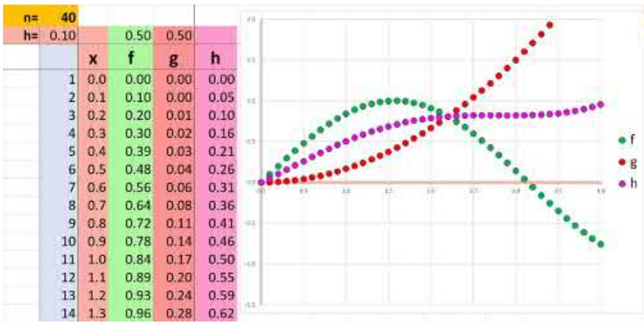


n=	40			
h=	0.10			
		x	∫f	∫h
	1	0.0	0.00	0.00
	2	0.1	0.00	0.00
	3	0.2	0.01	0.02
	4	0.3	0.03	0.06
	5	0.4	0.06	0.12
	6	0.5	0.10	0.20
	7	0.6	0.15	0.29
	8	0.7	0.20	0.41
	9	0.8	0.27	0.53
	10	0.9	0.34	0.68
	11	1.0	0.42	0.83
	12	1.1	0.50	1.00
	13	1.2	0.59	1.18
	14	1.3	0.68	1.37

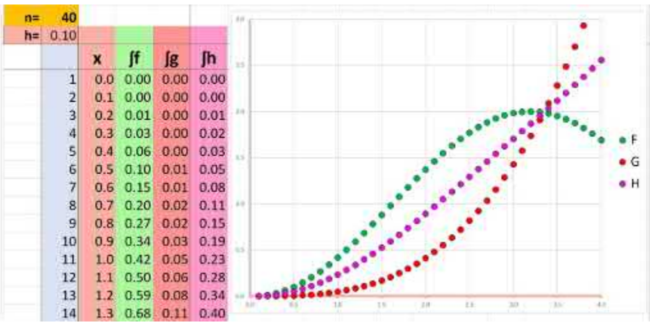
These two operations can be combined into one producing *linear combinations*:

$$\alpha x+\beta y\,,$$

where α, β are two constant numbers. For example, this is the average of two functions (left):



n=	40				
h=	0.10		0.50	0.50	
		x	f	g	h
	1	0.0	0.00	0.00	0.00
	2	0.1	0.10	0.00	0.05
	3	0.2	0.20	0.01	0.10
	4	0.3	0.30	0.02	0.16
	5	0.4	0.39	0.03	0.21
	6	0.5	0.48	0.04	0.26
	7	0.6	0.56	0.06	0.31
	8	0.7	0.64	0.08	0.36
	9	0.8	0.72	0.11	0.41
	10	0.9	0.78	0.14	0.46
	11	1.0	0.84	0.17	0.50
	12	1.1	0.89	0.20	0.55
	13	1.2	0.93	0.24	0.59
	14	1.3	0.96	0.28	0.62



n=	40				
h=	0.10				
		x	∫f	∫g	∫h
	1	0.0	0.00	0.00	0.00
	2	0.1	0.00	0.00	0.00
	3	0.2	0.01	0.00	0.01
	4	0.3	0.03	0.00	0.02
	5	0.4	0.06	0.00	0.03
	6	0.5	0.10	0.01	0.05
	7	0.6	0.15	0.01	0.08
	8	0.7	0.20	0.02	0.11
	9	0.8	0.27	0.02	0.15
	10	0.9	0.34	0.03	0.19
	11	1.0	0.42	0.05	0.23
	12	1.1	0.50	0.06	0.28
	13	1.2	0.59	0.08	0.34
	14	1.3	0.68	0.11	0.40

We also notice what happens to their integrals (right):

- ▶ The integral of the average is the average of the integrals.

Of course, we conclude the following first:

1. The functions integrable over a particular $[a, b]$ form a vector space.
2. The integral at $x = a$ is a real-valued linear operator on this space:

$$\boxed{\int_{[a,b]}: f \mapsto \int_{[a,b]} f\, dx}$$

More important is the following:

Theorem 6.8.4: Linear Algebra of Integration

1. *The functions integrable on a particular interval I form a vector space.*
2. *The integral is a linear operator on this space with values in functions defined on I:*

$$\boxed{\int : f \mapsto \int f\, dx}$$

Exercise 6.8.5

Is the operator one-to-one? Onto?

What happens to linear combinations of functions under integration? These two formulas can be combined into one:

$$\boxed{\int (\alpha f+\beta g)\, dx \ \ = \alpha \int f\, dx \ \ +\beta \int g\, dx}$$

The last formula is illustrated with the following diagram:

$$\alpha f+\beta g \ \to \ \boxed{\int} \ \to \ \alpha \int f+\beta \int g$$

Therefore, we have three vectors spaces connected by the inclusion operators:

$$\text{differentiable functions} \ \to \ \text{continuous functions} \ \to \ \text{integrable functions} \ \to \ \text{functions}$$

Therefore, all linear operators defined on the space of functions will also be defined on the space of integrable functions via restriction.

6.9. The Fundamental Theorem of Calculus

Differentiation and integration have been proven to be inverses in the discrete context:

$$\Delta^{-1} = \Sigma \,.$$

In this section, we will address the general case of the relation between differentiation and integration in the infinitesimal context. But first, a very instructive special case.

Example 6.9.1: differentiation of quadratic functions

What is the matrix of the derivative when limited to the space of quadratic functions? Just consider its effect on the basis elements, the power functions:

$$(1)' = 0,\ (x)' = 1,\ (x^2)' = 2x \,.$$

Then we have:

$$\frac{d}{dx}\begin{bmatrix}1\\0\\0\end{bmatrix} = \begin{bmatrix}0\\0\\0\end{bmatrix},\ \frac{d}{dx}\begin{bmatrix}0\\1\\0\end{bmatrix} = \begin{bmatrix}1\\0\\0\end{bmatrix},\ \frac{d}{dx}\begin{bmatrix}0\\0\\1\end{bmatrix} = \begin{bmatrix}0\\2\\0\end{bmatrix}.$$

Therefore, the matrix of

$$\frac{d}{dx} : \mathbf{P}^2 \to \mathbf{P}^2$$

is:

$$\frac{d}{dx} = \begin{bmatrix}0\\0\\0\end{bmatrix}\begin{bmatrix}1\\0\\0\end{bmatrix}\begin{bmatrix}0\\2\\0\end{bmatrix} = \begin{bmatrix}0&1&0\\0&0&2\\0&0&0\end{bmatrix}.$$

It's not invertible though.

Drawing an idea from our treatment of the difference operator, we try the following ideas:

1. What if we concentrate on the input on the quadratic functions with a *zero* constant term? Then 1 is removed from the basis.
2. What if we concentrate on the output on the linear functions only? Then x^2 is removed from the basis.

We have a new function:

$$\frac{d}{dx} : \{f \in \mathbf{P}^2 :\ f(0) = 0\} \to \mathbf{P}^1$$

with the following matrix:

$$\frac{d}{dx} = \begin{bmatrix}1\\0\end{bmatrix}\begin{bmatrix}0\\2\end{bmatrix} = \begin{bmatrix}1&0\\0&2\end{bmatrix}.$$

It's diagonal and invertible:

$$\left(\frac{d}{dx}\right)^{-1} = \begin{bmatrix}1&0\\0&\frac{1}{2}\end{bmatrix}.$$

The starting point of our treatment of all polynomials is the following:

Theorem 6.9.2: Derivative of Polynomial

The derivative of a polynomial of degree $n > 0$ is a polynomial of degree $n - 1$, *as follows:*

$$\begin{array}{rcllllll} f(x) &=& a_nx^n & +a_{n-1}x^{n-1} & +... & +a_2x^2 & +a_1x & +a_0 \\ f'(x) &=& na_nx^{n-1} & +(n-1)a_{n-1}x^{n-2} & +... & +2a_2x & +a_1 & \end{array}$$

Corollary 6.9.3: Matrix of Derivative of Polynomials

The matrix of the derivative on the space of polynomials of the n*th degree or lower,* $\mathbf{P}^n$*, is given by:*

$$\boxed{\frac{d}{dx} = \begin{bmatrix} 0 & 1 & 0 & 0 & ... & 0 \\ 0 & 0 & 2 & 0 & ... & 0 \\ 0 & 0 & 0 & 3 & ... & 0 \\ ... \\ 0 & 0 & 0 & 0 & ... & n \\ 0 & 0 & 0 & 0 & ... & 0 \end{bmatrix}}$$

Except for the entries just above the main diagonal, all are 0's.

Now, we know that integration is supposed to be the inverse of differentiation. But the latter isn't one-to-one!

As before, we reduce the domain to *make* it one-to-one. We limit ourselves to the polynomials with a 0 constant term:

$$\begin{array}{rcllllll} f(x) &=& a_nx^n & +a_{n-1}x^{n-1} & +... & +a_2x^2 & +a_1x \\ f'(x) &=& na_nx^{n-1} & +(n-1)a_{n-1}x^{n-2} & +... & +2a_2x & +a_1 \end{array}$$

We don't lose a lot because the term only contributes a vertical shift to the graph. And the number of terms (the dimension) in the domain and the codomain match!

The derivative then is this function:

$$\boxed{\frac{d}{dx} : \{f \in \mathbf{P}^n : \ f(0) = 0\} \to \mathbf{P}^{n-1}}$$

The dimensions are still the same, and the matrix has become diagonal:

$$\boxed{\frac{d}{dx} = \begin{bmatrix} 1 & 0 & 0 & ... & 0 \\ 0 & 2 & 0 & ... & 0 \\ 0 & 0 & 3 & ... & 0 \\ ... \\ 0 & 0 & 0 & ... & n \end{bmatrix}}$$

We have dropped a row and a column in comparison to the original matrix. The inverse might exist.

We need a linear operator:

$$\left(\frac{d}{dx}\right)^{-1} : \mathbf{P}^n \to \{f \in \mathbf{P}^{n+1} : \ f(0) = 0\} .$$

And we know how it is supposed to work:

$$\begin{array}{rcllllll} g(x) &=& b_nx^n & +b_{n-1}x^{n-1} & +... & +b_2x^2 & +b_1x & +b_0 \\ \int g(x)\,dx &=& \frac{1}{n+1}b_nx^{n+1} & +\frac{1}{n}b_{n-1}x^n & +... & +\frac{1}{2}b_2x^3 & +b_1x^2 & +b_0x \end{array}$$

A diagonal matrix is easy to invert:

Corollary 6.9.4: Matrix of Integral of Polynomials

The matrix of the integral as a linear operator from the space of the up to nth degree polynomials to the space of the up to $n+1$ degree polynomials with a zero constant term is given by:

$$\boxed{\left(\frac{d}{dx}\right)^{-1} = \left[\begin{array}{cccccc} 1 & 0 & 0 & 0 & ... & 0 \\ 0 & \frac{1}{2} & 0 & 0 & ... & 0 \\ 0 & 0 & \frac{1}{3} & 0 & ... & 0 \\ ... & & & & & \\ 0 & 0 & 0 & 0 & ... & 0 \\ 0 & 0 & 0 & 0 & ... & \frac{1}{n+1} \end{array}\right]}$$

Exercise 6.9.5

Suppose A is the vector space of all linear combinations of $\sin x$ and $\cos x$. What is the matrix of the derivative on this space?

Exercise 6.9.6

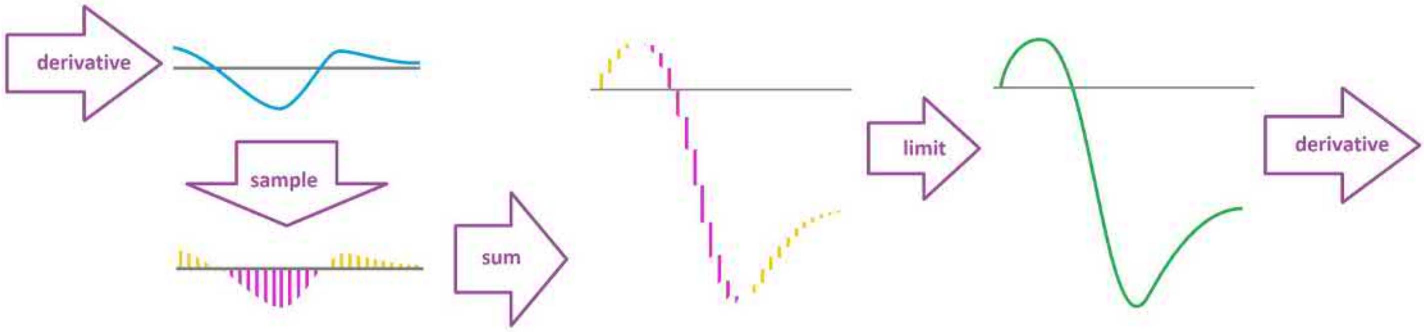
Suppose A is the vector space of all linear combinations of e^x and e^{-x}. What is the matrix of the derivative on this space?

So far, we have learned the following two facts that match totally:

1. The difference operator and the sum operator are inverses.
2. The derivative operator and the integral operator are inverses in case of polynomials.

On to the general case.

Below, a function is sampled to produce the Riemann sums, which, under the limit, produce the Riemann integral to be differentiated:



Will we make the full circle?

Recall these two theorems from calculus:

Theorem 6.9.7: Fundamental Theorem of Calculus I

Given a continuous function f on $[a,b]$, the function defined by

$$F(x) = \int_a^x f\,dx$$

is an antiderivative of f on (a,b); i.e,

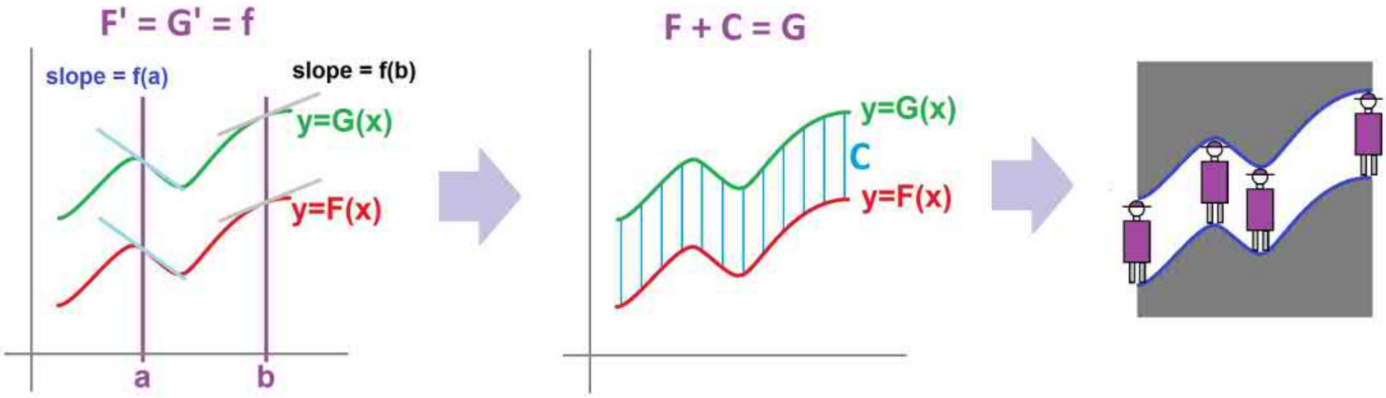
$$F' = f\,.$$

Theorem 6.9.8: Fundamental Theorem of Calculus II

For any integrable function f on $[a,b]$ and any of its antiderivatives F, we have:

$$\int_a^b f\,dx = F(b) - F(a)\,.$$

The Part I supplies us with a special choice of antiderivative – the one that satisfies $F(a) = 0$. The rest, according to Part II, are acquired by vertical shifts: $G = F + C$. The idea is that if the ceiling and the floor of a tunnel are equal at every point, its height is constant:



Now, linear algebra.

The idea is to look at the *composition of the derivative and the integral operators.* The domains and codomains of both include only functions defined on $[a,b]$.

Let's keep in mind that the codomain of the latter is determined by the definite integral, so that

$$F(a) = 0\,.$$

The inputs and outputs of the derivative operator $\frac{d}{dx}$ will have to be matched with the outputs and inputs of the integral operator $\int$. They do, according to our definition:

$$\begin{array}{rlcl}
\frac{d}{dx}: & \text{differentiable functions} & \to & \text{integrable functions with } F(a)=0 \\
\int: & \text{integrable functions with } F(a)=0 & \to & \text{differentiable functions}
\end{array}$$

This is what happens when we put them together:

$$\begin{array}{lclcl}
\text{differentiable} & \xrightarrow{\frac{d}{dx}} & \text{integrable with } F(a)=0 & \xrightarrow{\int} & \text{differentiable} \\
\text{integrable with } F(a)=0 & \xrightarrow{\int} & \text{differentiable} & \xrightarrow{\frac{d}{dx}} & \text{integrable with } F(a)=0
\end{array}$$

The compositions

$$\int \circ \frac{d}{dx} \quad\text{and}\quad \frac{d}{dx} \circ \int$$

now make sense.

The linear algebra analogs of the above theorems are below:

Corollary 6.9.9: I

The composition of the derivative operator $\frac{d}{dx}$ *and the integral* $\int$ *is the identity:*

$$\boxed{\frac{d}{dx} \circ \int = I}$$

Proof.

Exercise 6.9.10

Finish the proof.

Corollary 6.9.11: II

The composition of the integral $\int$ *and the derivative* $\frac{d}{dx}$ *is the identity:*

$$\boxed{\int \circ \frac{d}{dx} = I}$$

We combine them together into this truly fundamental result:

Theorem 6.9.12: Fundamental Theorem of Calculus

The integral operator $\int$ *and the derivative operator* $\frac{d}{dx}$ *are inverses of each other:*

$$\boxed{\left(\frac{d}{dx}\right)^{-1} = \int}$$

Exercise 6.9.13

What are the eigenvectors of the derivative operator?

6.10. Linear algebra of power series

Suppose a point a is given. Then the following creates a function, f:

- There is a sequence of numbers,

$$c_0, c_1, c_2, \dots$$

- For each input x, its value under this function is computed by substituting it into a formula, a power series with the sequence providing its coefficients:

$$f(x) = \sum_{n=0}^{\infty} c_n (x-a)^n .$$

- The domain of the function is the region of convergence of the series.

Here, a is the the *center* of the power series.

We start with *algebra*.

Just as with functions in general, we can carry out (some) algebraic operations on power series producing new power series. However, what is truly important is that we can do these operations *term by term*. The idea comes from our experience with *polynomials*; after all the algebra, we want to put the result in the standard form, i.e., with all terms arranged according to the powers.

First, we can add two polynomials one term at a time:

$$\begin{array}{llll} p(x) & =1 & +2x & +3x^2 \\ q(x) & =7 & +5x & -2x^2 \\ \hline p(x)+q(x) & =(1+7) & +(2+5)x & +(3-2)x^2 \end{array}$$

We know that we can also add two series one term at a time:

$$\begin{array}{lllll} p(x) & =c_0 & +c_1x & +c_2x^2 & +... \\ q(x) & =d_0 & +d_1x & +d_2x^2 & +... \\ \hline p(x)+q(x) & =(c_0+d_0) & +(c_1+d_1)x & +(c_2+d_2)x^2 & +.... \end{array}$$

Second, we can multiply a series by a number one term at a time:

$$\begin{array}{llll} p(x) & =1 & +2x & +3x^2 \\ \hline 2p(x) & =(2\cdot 1) & +(2\cdot 2)x & +(2\cdot 3)x^2 \end{array}$$

We also multiply a series by a number one term at a time:

$$\begin{array}{lllll} p(x) & =c_0 & +c_1x & +c_2x^2 & +... \\ \hline kp(x) & =(kc_0) & +(kc_1)x & +(kc_2)x^2 & +... \end{array}$$

The general result is below.

Theorem 6.10.1: Term-by-Term Algebra of Power Series

Suppose two functions are represented by power series:

$$f(x)=\sum_{n=0}^{\infty} c_n(x-a)^n \ \text{ and } \ g(x)=\sum_{n=0}^{\infty} d_n(x-a)^n .$$

Then we have:

1. *The function $f+g$ is represented by the power series that is the term-by-term sum of those of f and g, defined on the intersection of their domains:*

$$(f+g)(x)=\sum_{n=0}^{\infty} c_n(x-a)^n+\sum_{n=0}^{\infty} d_n(x-a)^n=\sum_{n=0}^{\infty}(c_n+d_n)(x-a)^n .$$

2. *The function kf, for any constant k, is represented by the power series that is the term-by-term product of that of f, defined on the same domain:*

$$(kf)(x)=k\cdot\sum_{n=0}^{\infty} c_n(x-a)^n=\sum_{n=0}^{\infty}(kc_n)(x-a)^n .$$

In other words, these "infinite" polynomials behave just like ordinary polynomials, wherever they converge. Linear algebra allows us to restate this theorem in a very compact form:

Theorem 6.10.2: Linear Algebra of Power Series

The power series, convergent on a particular interval, form a vector space.

Next is differentiation and integration, i.e., *calculus*, of power series.

We will see that, just as with functions in general, we can carry out the calculus operations on power series producing new power series. However, what is truly important is that we can do these operations *term by term*.

Example 6.10.3: differentiation and integration

Let's differentiate – only the *Power Formula* of differentiation required – the terms of the power series representation of the exponential function:

$$\begin{tikzcd}[column sep=small]
e^x \arrow[d, "\frac{d}{dx}"] & = & 1 \arrow[d, "\frac{d}{dx}"] & + & x \arrow[d, "\frac{d}{dx}"] & + & \frac{1}{2!}x^2 \arrow[d, "\frac{d}{dx}"] & + & \frac{1}{3!}x^3 \arrow[d, "\frac{d}{dx}"] & + & \dots & + & \frac{1}{n!}x^n \arrow[d, "\frac{d}{dx}"] & + & \frac{1}{(n+1)!}x^{n+1} \arrow[d, "\frac{d}{dx}"] & + \dots \\
(e^x)' & \overset{?}{=} & 0 & + & 1 & + & \frac{1}{2!}2x & + & \frac{1}{3!}3x^2 & + & \dots & + & \frac{1}{n!}nx^{n-1} & + & \frac{1}{(n+1)!}(n+1)x^n & + \dots \\
 & = & & & 1 & + & x & + & \frac{1}{2!}x^2 & + & \dots & + & \frac{1}{(n-1)!}x^{n-1} & + & \frac{1}{n!}x^n & + \dots \\
 & = & e^x. & & & & & & & & & & & & &
\end{tikzcd}$$

Let's integrate – only the *Power Formula* of integration required – the terms of the power series representation of the exponential function:

$$\begin{tikzcd}[column sep=small]
e^x \arrow[d, "\int"] & = & & 1 \arrow[d, "\int"] & + & x \arrow[d, "\int"] & + & \frac{1}{2!}x^2 \arrow[d, "\int"] & + & \dots & + & \frac{1}{n!}x^n \arrow[d, "\int"] & + & \frac{1}{(n+1)!}x^{n+1} \arrow[d, "\int"] & + \dots \\
\int e^x\,dx & \overset{?}{=} & C+ & x & + & \frac{1}{2}x^2 & + & \frac{1}{2!}\frac{1}{3}x^3 & + & \dots & + & \frac{1}{n!}\frac{1}{n+1}x^{n+1} & + & \frac{1}{(n+1)!}\frac{1}{n+2}x^{n+2} & + \dots \\
 & = & C+ & x & + & \frac{1}{2!}x^2 & + & \frac{1}{3!}x^3 & + & \dots & + & \frac{1}{(n+1)!}x^{n+1} & + & \frac{1}{(n+2)!}x^{n+2} & + \dots \\
 & = & C+ & e^x. & & & & & & & & & & &
\end{tikzcd}$$

Differentiation and integration of the terms is easy:

$$\frac{d}{dx}\Big(c_n(x-a)^n\Big) = nc_n(x-a)^{n-1},\ \int \Big(c_n(x-a)^n\Big)\,dx = \frac{c_n}{n+1}(x-a)^{n+1}.$$

The following theorem is the summary:

Theorem 6.10.4: Term-by-Term Calculus of Power Series

Suppose the radius of convergence of a power series,

$$f(x) = \sum_{n=0}^{\infty} a_n(x-a)^n,$$

is positive or infinite. Then the function f represented by this power series is differentiable (and, therefore, integrable) on the interval of convergence, and the power series representations of its derivative and its antiderivative converge *inside this interval and are found by term-by-term differentiation and integration of the power series of f respectively, i.e.,*

$$\frac{d}{dx}f(x)=\frac{d}{dx}\left(\sum_{n=0}^{\infty}c_n(x-a)^n\right)=\sum_{n=0}^{\infty}\frac{d}{dx}\big(c_n(x-a)^n\big)$$

and

$$\int f(x)\,dx=\int\left(\sum_{n=0}^{\infty}c_n(x-a)^n\right)\,dx=\sum_{n=0}^{\infty}\int c_n(x-a)^n\,dx\,.$$

With this theorem, there is no need for the rules of differentiation or integration except for the *Power Formula*!

Corollary 6.10.5: Matrix of Derivative of Power Series

The matrix of the derivative on the space of power series convergent on a particular open interval is given by the following infinite matrix:

$$\boxed{\frac{d}{dx}=\begin{bmatrix}0&1&0&0&0&...&0&...\\0&0&2&0&0&...&0&...\\0&0&0&3&0&...&0&...\\...&&&&&...&&...\\0&0&0&0&0&...&0&...\\0&0&0&0&0&...&n+1&...\\...&&&&&...&&...\end{bmatrix}}$$

The multiplication of this infinite matrix by an infinite vector (a sequence) doesn't represent a problem because there are only finitely many non-zero entries in each row.

Definition 6.10.6: analytic function

A function defined on an open interval that can be represented by a power series is called *analytic* on this interval.

Theorem 6.10.7: Uniqueness of Power Series Representation

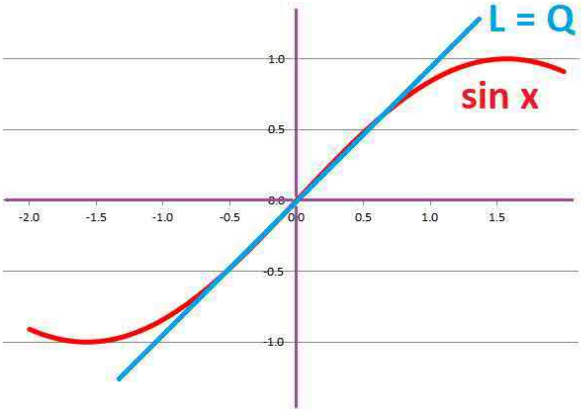
An analytic function has a unique power series representation, i.e., if two power series are equal, as functions, on an open interval, then their corresponding coefficients are equal too, i.e.,

$$\begin{array}{lrcll}f(x)&=\displaystyle\sum_{n=0}^{\infty}c_n(x-a)^n&=&\displaystyle\sum_{n=0}^{\infty}d_n(x-a)^n&\text{for all } a-r<x<a+r,\ r>0\\ \Longrightarrow&c_n&=&d_n&\text{for all } n=0,1,2,3,...\end{array}$$

Furthermore, every analytic function is infinitely many times differentiable.

Example 6.10.8: sine and cosine

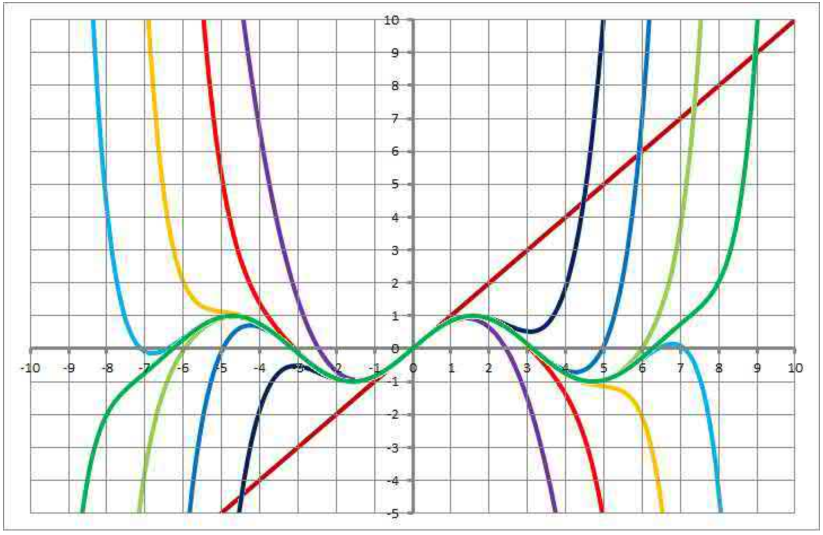
Let's find the power series representation for $f(x)=\sin x$ at $x=0$. We start with what we already know:



We need them all:

$$\begin{array}{llllll}
f(x) & = \sin x & \Longrightarrow & f(0) & = 0 & \Longrightarrow & T_0(x) & = 0 \\
f'(x) & = \cos x & \Longrightarrow & f'(0) & = 1 & \Longrightarrow & T_1(x) & = x \\
f''(x) & = -\sin x & \Longrightarrow & f''(0) & = 0 & \Longrightarrow & T_2(x) & = x \\
f'''(x) & = -\cos x & \Longrightarrow & f'''(0) & = -1 & \Longrightarrow & T_3(x) & = 1 - \frac{1}{6}x^3 \\
\dots
\end{array}$$

The sequence starts to repeat itself, every four steps... Every polynomial leaves for infinity eventually but the resemblance extends further and further from the center:



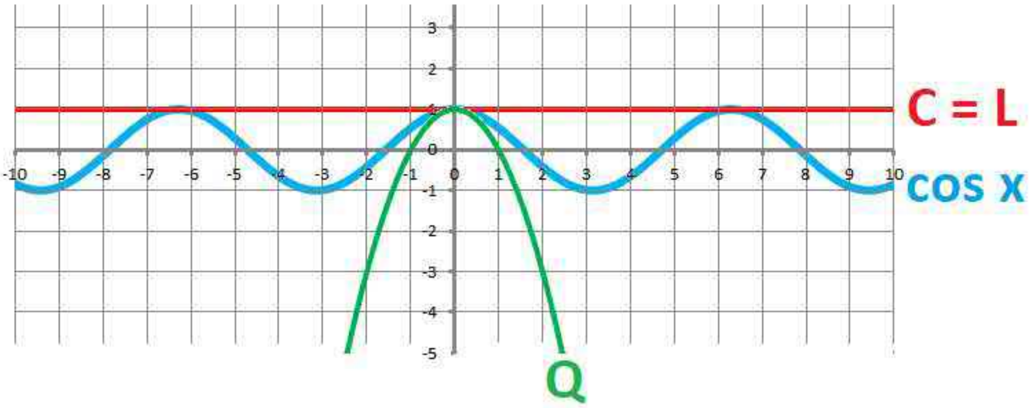
There are no even powers present because the sine is odd. Therefore,

$$f^{(2m-1)}(0) = (-1)^m .$$

We have the Taylor coefficients:

$$c_{2m-1} = \frac{(-1)^m}{(2m-1)!} .$$

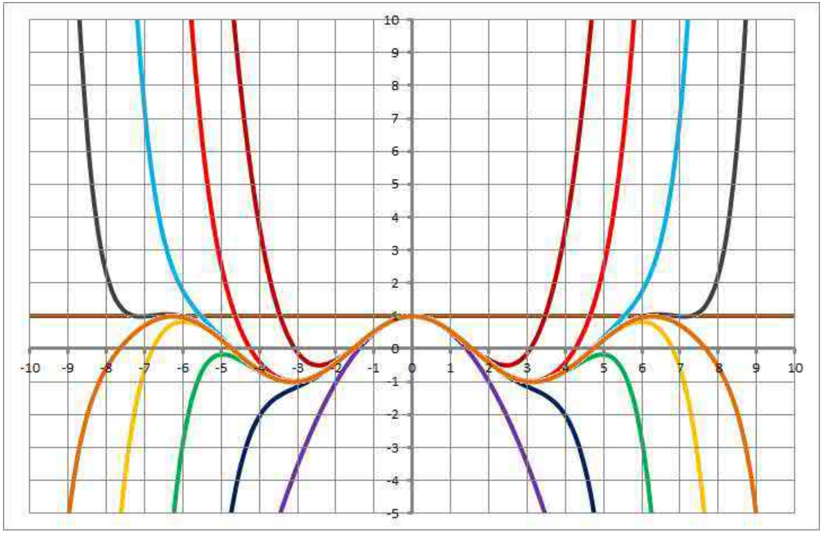
Let's approximate $f(x) = \cos x$ at $x = 0$. We start with what we already know:



We need them all:

$$\begin{array}{llllll}
f(x) & = \cos x & \Longrightarrow & f(0) & = 1 & \Longrightarrow & T_0(x) & = 1 \\
f'(x) & = -\sin x & \Longrightarrow & f'(0) & = 0 & \Longrightarrow & T_1(x) & = 1 \\
f''(x) & = -\cos x & \Longrightarrow & f''(0) & = -1 & \Longrightarrow & T_2(x) & = 1 - \frac{1}{2}x^2 \\
f'''(x) & = \sin x & \Longrightarrow & f'''(0) & = 0 & \Longrightarrow & T_3(x) & = 1 - \frac{1}{2}x^2 \\
f^{(4)}(x) & = \cos x & \Longrightarrow & f^{(4)}(0) & = 1 \\
...
\end{array}$$

The sequence starts to repeat itself, every four steps... Every polynomial leaves for infinity eventually but the resemblance extends further and further from the center:



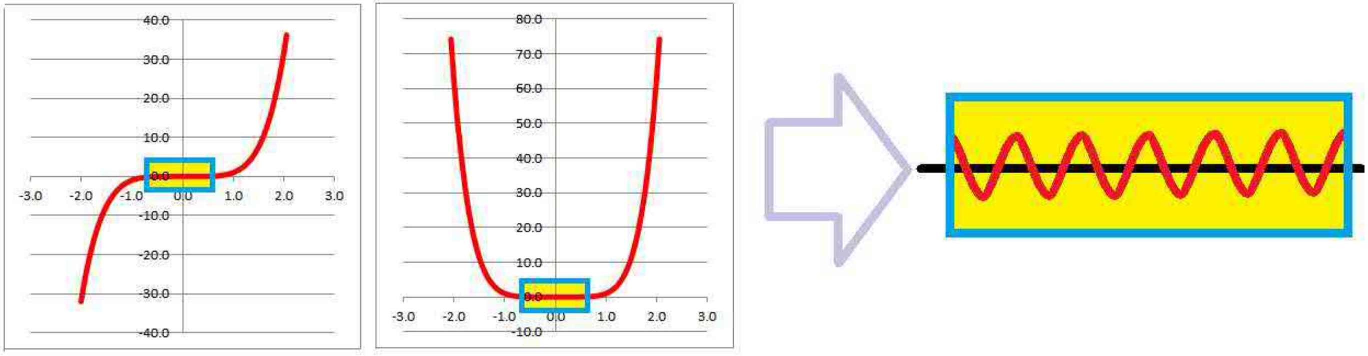
There are no odd powers present because the cosine is even. Therefore,

$$f^{(2m)}(0) = (-1)^m \,.$$

We have the Taylor coefficients:

$$c_{2m} = \frac{(-1)^m}{(2m)!} \,.$$

This is what polynomial approximations would look like:



Even though each approximation eventually becomes bad, the interval where things are good is expanding.

The differentiation and integration are perfectly reflected in this mirror of power series:

$$\begin{tikzcd}
f \arrow[r, "\text{Taylor}"] \arrow[d, "\frac{d}{dx}"] & \sum_n c_n(x-a)^n \arrow[d, "\frac{d}{dx}"] \\
f' \arrow[r, "\text{Taylor}"] & \sum_n \left(c_n(x-a)^n\right)'
\end{tikzcd}
\qquad
\begin{tikzcd}
f \arrow[r, "\text{Taylor}"] \arrow[d, "\int"] & \sum_n c_n(x-a)^n \arrow[d, "\int"] \\
\int f\,dx \arrow[r, "\text{Taylor}"] & \sum_n \int \left(c_n(x-a)^n\right)\,dx
\end{tikzcd}$$

In the first diagram, we start with a function at the top left and then we proceed in two ways:

- Right: Find its Taylor series. Then down: Differentiate the result term by term.
- Down: Differentiate it. Then right: Find its Taylor series.

The result is the same!

Exercise 6.10.9

What are the eigenvectors of the derivative operator?

Exercises

Contents

1. Exercises: Background

Exercise 1.1

Give each of the following functions a domain and codomain and determine if it is one-to-one or onto: (a) $y = x^2$; (b) $y = \sqrt{x}$; (b) $z = x + y$; (c) $(x, y) = (\sin t, \cos t)$.

Exercise 1.2

Contract this summation:

$$2 - \frac{2}{2} + \frac{2}{3} - \frac{2}{4} = ?$$

Exercise 1.3

Expand this summation:

$$\sum_{k=-1}^{5} \frac{k^2}{k+2} = ?$$

Exercise 1.4

True or False?
(a) $x < r \implies |x| < r$;
(b) $x < r \impliedby |x| < r$;
(c) $x < r \iff |x| < r$.

Exercise 1.5

For $4y + 16x = 20$, provide the slope and the y-intercept.

Exercise 1.6

A diameter of a circle runs between points R and T. The center of the circle, P, has coordinates $(-4, 1)$. The coordinates of the point R are $(2, -3)$. What are the coordinates of T?

Exercise 1.7

Let $f(x) = 7+2x-x^2$. Find the difference quotient

$$\frac{f(3+h)-f(3)}{h}.$$

Simplify your answer.

Exercise 1.8

Find an expression for $f(x)$ and state its domain in interval notation given that f is the function that takes a real number x and performs the following three steps in order:

1. divide by 3,
2. take square root, and then
3. make the quantity the denominator of a fraction with numerator 13.

Exercise 1.9

Compute $\sum_{n=1}^{4} n^2$.

Exercise 1.10

Present the first 5 terms of the sequence:

$$a_1 = 1, \quad a_{n+1} = -(a_n + 1).$$

Exercise 1.11

Represent in sigma notation:

$$-1-2-3-4-5-...-10.$$

Exercise 1.12

Find the sum of the following:

$$-1-2-3-4-5-...-10.$$

Exercise 1.13

Find the sequence of sums of the following sequence:

$$-1, 2, -4, 8, -5, ...$$

Exercise 1.14

Show that $\frac{n}{n+1}$ is an increasing sequence. What kind of sequence is $\frac{n+1}{n}$? Give examples of increasing and decreasing sequences.

Exercise 1.15

Find the next item in each list:

1. $7, 14, 28, 56, 112, ...$
2. $15, 27, 39, 51, 63, ...$
3. $197, 181, 165, 149, 133, ...$

Exercise 1.16

A pile of logs has 50 logs in the bottom layer, 49 logs in the next layer, 48 logs in the next layer, and so on, until the top layer has 1 log. How many logs are in the pile?

Exercise 1.17

In the beginning of each year, a person puts \$5000 in a bank that pays 3% compounded annually. How much does he have after 15 years?

Exercise 1.18

An object falling from rest in a vacuum falls approximately 16 feet the first second, 48 feet the second second, 80 feet the third second, 112 feet the fourth second, and so on. How far will it fall in 11 seconds?

2. Exercises: Sets and logic

Exercise 2.1

Represent the following set in the set-building notation:

$$X = [0,1] \cup [2,3] = ...$$

Exercise 2.2

Simplify:

$$\{x > 0 : \ x \text{ is a negative integer } \}.$$

Exercise 2.3

What are the max, min, and any bounds of the set of integers? What about **R**?

Exercise 2.4

Is the converse of the converse of a true statement true?

Exercise 2.5

State the converse of this statement: "the converse of the converse of a true statement is true".

Exercise 2.6

Represent these sets as intersections and unions:

1. $(0,5)$
2. $\{3\}$
3. $\emptyset$
4. $\{x : x > 0 \ \ \texttt{OR} \ \ x \ \text{ is an integer}\}$
5. $\{x : \ x \ \text{ is divisible by } \ 6\}$

Exercise 2.7

True or false: $0 = 1 \implies 0 = 1$?

Exercise 2.8

Prove:

$$\max\{\max A, \max B\} = \max(A \cup B).$$

Exercise 2.9

(a) If, starting with a statement A, after a series of conclusions you arrive to $0 = 1$, what can you conclude about A? (b) If, starting with a statement A, after a series of conclusions you arrive to $0 = 0$, what can you conclude about A?

Exercise 2.10

We know that "If it rains, the road gets wet". Does it mean that if the road is wet, it has rained?

Exercise 2.11

A garage light is controlled by a switch and, also, it may automatically turn on when it senses motion during nighttime. If the light is OFF, what do you conclude?

Exercise 2.12

If an advertisement claims that "All our second-hand cars come with working AC", what is the easiest way to disprove the sentence?

Exercise 2.13

Teachers often say to the student's parents: "If your student works harder, he'll improve". When he won't improve and the parents come back to the teacher, he will answer: "He didn't improve, that means he didn't work harder". Analyze.

3. Exercises: Coordinate system

Exercise 3.1

Find the equation of the line passing through the points $(-1, 1)$ and $(-1, 5)$.

Exercise 3.2

What is the distance from the center of the circle

$$(x-1)^2 + (y+3)^2 = 5$$

to the origin?

Exercise 3.3

One circle is centered at $(0, 0)$ and has radius 1. The second is centered at $(3, 3)$. What is the radius of the second if the two circles touch?

Exercise 3.4

What is the distance from the circle

$$x^2 + (y+3)^2 = 2$$

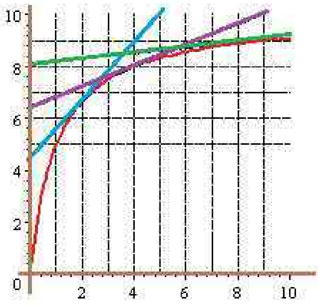
to the origin?

Exercise 3.5

Find the equation of the circle centered at $(-1, -1)$ and passing through the point $(-1, 1)$.

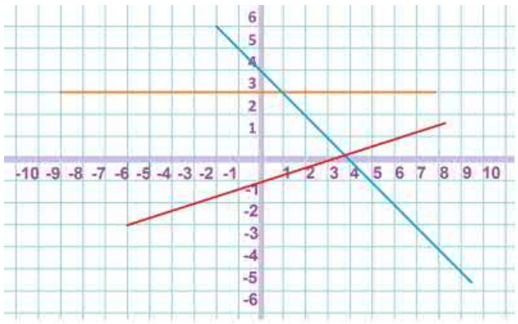
Exercise 3.6

Three straight lines are shown below. Find their slopes:



Exercise 3.7

Three straight lines are shown below. Find their equations:



Exercise 3.8

For the points $P = (0, 1)$, $Q = (1, 2)$, and $R = (-1, 2)$, determine the points that are symmetric with respect to the axis and the origin.

Exercise 3.9

The hypotenuse of an isosceles right triangle is 10 inches. The midpoints of its sides are connected to form an inscribed triangle, and this process is repeated. Find the sum of the areas of these triangles as this process is continued.

Exercise 3.10

Consider triangle ABC in the plane where $A = (3, 2)$, $B = (3, -3)$, $C = (-2, -2)$. Find the lengths of the sides of the triangle.

Exercise 3.11

Sketch the region given by the set $\{(x, y) : xy < 0\}$. Which axes and which quadrants of the plane are included in the set?

Exercise 3.12

Find all x such that the distance between the points $(3, -8)$ and $(x, -6)$ is 5.

Exercise 3.13

Two cars leave a highway junction at the same time. The first travels west at 70 miles per hour and the second travels north at 60 miles per hour. How far apart are they after 1.5 hours?

Exercise 3.14

Find the perimeter of the triangle with the vertices at $(3, -1)$, $(3, 6)$, and $(-6, -5)$.

Exercise 3.15

Find the point on the x-axis that is equidistant from the points $(-1, 5)$ and $(6, 4)$.

Exercise 3.16

Find the distance between the points of intersection of the circle $(x-1)^2 + (y-2)^2 = 6$ with the axes.

4. Exercises: Relations and functions

Exercise 4.1

A contractor purchases gravel one cubic yard at a time. A gravel driveway x yards long and 4 yards wide is to be poured to a depth of 1.5 foot. Find a formula for $f(x)$, the number of cubic yards of gravel the contractor buys, assuming that he buys 10 more cubic yards of gravel than are needed.

Exercise 4.2

Visualize the relation:

$$\frac{x^2}{4}+\frac{y^2}{9}=1\,.$$

Do you see these 4 and 9 on the graph?

Exercise 4.3

Suppose the cost is $f(x)$ dollars for a taxi trip of x miles. Interpret the following stories in terms of f.

1. Monday, I took a taxi to the station 5 miles away.
2. Tuesday, I took a taxi to the station but then realized that I left something at home and had to come back.
3. Wednesday, I took a taxi to the station and I gave my driver a five dollar tip.
4. Thursday, I took a taxi to the station but the driver got lost and drove five extra miles.
5. Friday, I have been taking a taxi to the station all week on credit; I pay what I owe today.

What if there is an extra charge per ride of m dollars?

Exercise 4.4

Let $f:A\to B$ and $g:C\to D$ be two possible functions. For each of the following functions, state whether or not you can compute $f\circ g$:

- $D\subset B$
- $C\subset A$
- $B\subset D$
- $B=C$

Exercise 4.5

An amusement park sells multi-day passes. The function $g(x)=1/3x$ represent the number of days a pass will work, where x is the amount of money paid, in dollars. Interpret the meaning of $g(6)=3$.

Exercise 4.6

The perimeter of a rectangle is 10 feet. (a) Express the area of the rectangle in terms of its width. (b) Find the minimal possible area. (c) Find the maximal possible area.

Exercise 4.7

Let $A=f(r)$ be the area of a circle with radius r and $r=h(t)$ be the radius of the circle at time t. Which of the following statements correctly provides a practical interpretation of the composition $f(h(t))$?

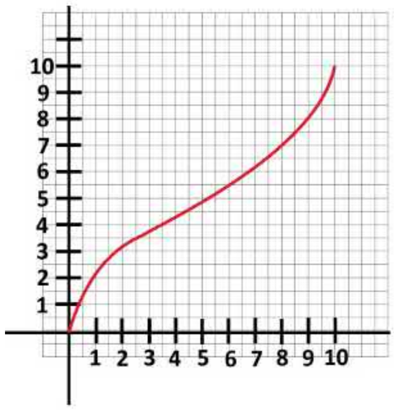
1. The length of the radius at time t.
2. The area of the circle at time t.
3. The length of the radius of a circle with area $A=f(r)$ at time t.
4. The area of the circle which at time t has radius $h(t)$.
5. The time t when the area will be $A=f(r)$.
6. The time t when the radius will be $r=h(t)$.

Exercise 4.8

The area of a rectangle is 100 sq. feet. (a) Express the perimeter of the rectangle in terms of its width. (b) Find the minimal possible perimeter. (c) Find the maximal possible perimeter.

Exercise 4.9

The graph of the function $y=f(x)$ is given below. (a) Find such a y that the point $(2,y)$ belongs to the graph. (b) Find such an x that the point $(x,3)$ belongs to the graph. (b) Find such an x that the point (x,x) belongs to the graph. Show your drawing.



Exercise 4.10

Make a flowchart and then provide a formula for the function $y = f(x)$ that represents a parking fee for a stay of x hours. It is computed as follows: free for the first hour and \$1 per hour beyond.

Exercise 4.11

Find all possible values of x for which

$$\tan x = 0\,.$$

Exercise 4.12

Make a hand-drawn sketch of the graph of the function:

$$f(x) = \begin{cases} -3 & \text{if } x < 0, \\ x^2 & \text{if } 0 \le x < 1, \\ x & \text{if } x > 1. \end{cases}$$

Exercise 4.13

Find the implied domains of the functions given by:

$$\text{(a) } \frac{x+1}{\sqrt{x^2-1}}; \quad \text{(b) } \sqrt[4]{x+1}\,.$$

Exercise 4.14

Find the implied domain of the function given by:

$$\frac{1}{(x-1)(x^2+1)}\,.$$

Exercise 4.15

Find the implied domain of the function given by:

$$\frac{1}{\sqrt{x+1}}\,.$$

Exercise 4.16

Find the implied domain of the function:

$$\frac{x-1}{x+1}\ln(x^2+1)\sin x\,.$$

Exercise 4.17

Find the implied domain of the function:

$$(x-1)(x^2+1)2^x\,.$$

Exercise 4.18

Finish the sentence: "If a function fails the horizontal line test, then..."

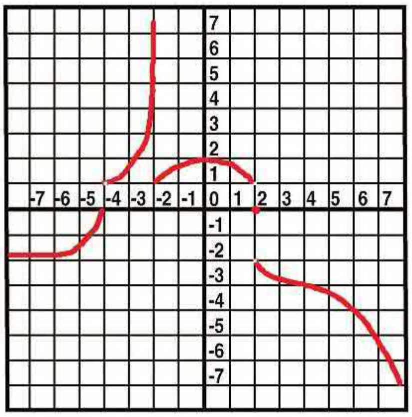
Exercise 4.19

Restate (but do not solve) the following problem algebraically: "What are the dimensions of the rectangle with the smallest possible perimeter and area fixed at 100?"

Exercise 4.20

A sketch of the graph of a function f and its table of values are given below.
Complete the table:

x	0		3		1
y	2	4		5	



Exercise 4.21

Plot the graph of the function $y = f(x)$, where x is the income (in thousands of dollars) and $f(x)$ is the tax bill (in thousands of dollars) for the income of x, which is computed as follows: no tax on the first \$10,000, then 5% for the next \$10,000, and 10% for the rest of the income.

Exercise 4.22

Plot the graph of the function $y = f(x)$, where x is time in hours and $y = f(x)$ is the parking fee over x hours, which is computed as follows: free for the first hour, then \$1 per every full hour for the next 3 hours, and a flat fee of \$5 for anything longer.

Exercise 4.23

Explain the difference between these two functions:

$$\sqrt{\frac{x-1}{x+1}} \text{ and } \frac{\sqrt{x-1}}{\sqrt{x+1}}\,.$$

Exercise 4.24

Classify these functions:

function	odd	even	onto	one-to-one
$f(x)=2x-1$				
$g(x)=-x+2$				
$h(x)=3$				

Exercise 4.25

Describe the function that computes the cash-back of 5% followed by the discount of 10%. What if we reverse the order?

Exercise 4.26

Represent this function as a list of instructions:

$$f(x)=\left(\sqrt[3]{\sin x}+2\right)^{1/2}.$$

Exercise 4.27

Find a formula for the following function:

$$\to \boxed{\text{square it}} \to \boxed{\text{take its reciprocal}} \to$$

Exercise 4.28

Plot the graph of the function given by the list of instructions: 1. add -1; 2. divide by 0; 3. square the outcome.

Exercise 4.29

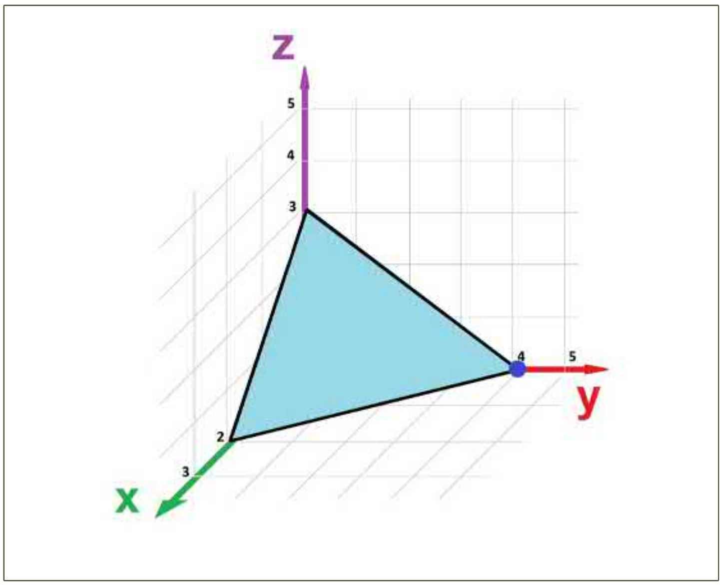
Find the x- and y-intercepts for the graphs in this section.

Exercise 4.30

Give each of the following functions a domain and codomain and determine if it is one-to-one or onto: (a) $y=x^2$; (b) $u=\sqrt{x}$; (c) $z=1/y$.

Exercise 4.31

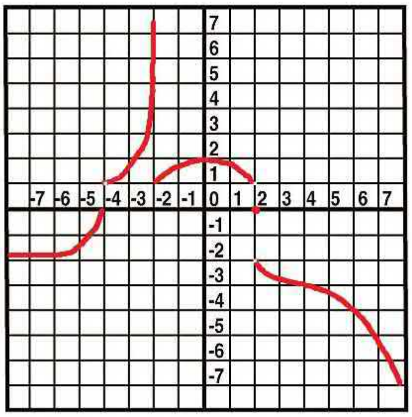
The plane below is the graph of a function. Find the function.



5. Exercises: Graphs

Exercise 5.1

A sketch of the graph of a function f is given below. Describe its behavior the function using words "decreasing" and "increasing".



Exercise 5.2

Function $y = f(x)$ is given below by a list of some of its values. Make sure the function is onto.

x	-1	0	1	2	3	4	5
$y = f(x)$	-1		4	5		2	

Exercise 5.3

Function $y = f(x)$ is given below by a list of some of its values. Add missing values in such a way that the function is one-to-one.

x	-1	0	1	2	3	4	5
$y = f(x)$	-1		0	5		0	

Exercise 5.4

What is the relation between being (a) one-to-one or onto and (b) having a mirror symmetry or central symmetry?

Exercise 5.5

By changing its domain or codomain, make the function $y = x^3 - x$ (a) onto, and (b) one-to-one?

Exercise 5.6

Is the function below even, odd, or neither?

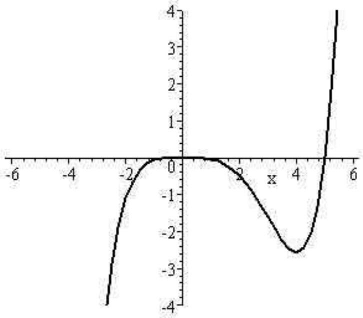
$$f(x) = \frac{x}{e^x - 1} + \frac{1}{2}x - 1$$

Exercise 5.7

Give an example of an even function, an odd function, and a function that's neither. Provide formulas.

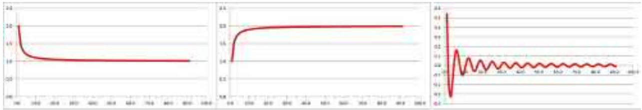
Exercise 5.8

Test whether the following three functions are even, odd, or nether: (a) $f(x) = x^3 + 1$; (b) the function the graph of which is a parabola shifted one unit up; (c) the function with this graph:



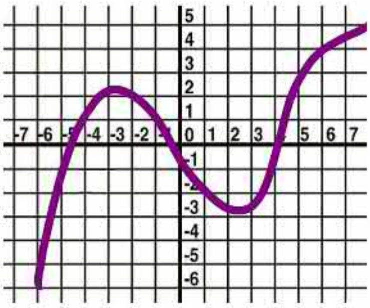
Exercise 5.9

Find horizontal asymptotes of these functions:



Exercise 5.10

The graph of a function $f(x)$ is given below. (a) Find $f(-4)$, $f(0)$, and $f(4)$. (b) Find such an x that $f(x) = 2$. (c) Is the function one-to-one?



Exercise 5.11

Is $\sin x/2$ a periodic function? If it is, find its period. You have to justify your conclusion algebraically.

Exercise 5.12

Is $\sin x + \cos \pi x$ a periodic function? If it is, find its period. You have to justify your conclusion algebraically.

Exercise 5.13

Is $\sin x + \sin 2x$ or $\sin x + \sin \frac{1}{2}x$ a periodic function? If it is, find its period. You have to justify your conclusion algebraically.

Exercise 5.14

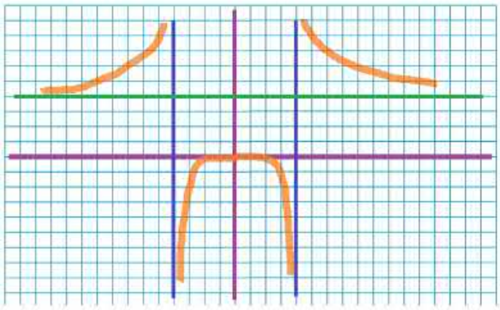
(a) State the definition of a periodic function. (b) Give an example of a periodic polynomial.

Exercise 5.15

Prove, from the definition, that the function $f(x) = x^2 + 1$ is increasing for $x > 0$.

Exercise 5.16

The graph of the function $y = f(x)$ is given below. (a) Find its domain. (b) Determine intervals on which the function is decreasing or increasing. (c) Provide x-coordinates of its relative maxima and minima. (d) Find its asymptotes.



Exercise 5.17

If a rational function has 10 vertical asymptotes, how many branches does its graph have?

Exercise 5.18

For the graph of the function $y = \sqrt{x + 8}$, answer the following questions: Is the graph symmetric with respect to the x-axis? The y-axis? The origin?

Exercise 5.19

Determine which of the following statements are true and which are false.

1. The function $\sin x$ on the domain $(-\pi, \pi)$ has at least one input which produces a smallest output value.
2. The function $f(x) = x^3$ with domain $(-3, 3)$ has at least one input which produces a largest output value.
3. The function $f(x) = x^3$ with domain $[-3, 3]$ has at least one input which produces a largest output value.
4. The function $f(x) = x^3$ with domain $[-3, 3]$ has at least one input which produces a smallest output value.
5. The function $\sin x$ on the domain $[-\pi, \pi]$ has at least one input which produces a smallest output value.

Exercise 5.20

Give an example of a function that is both odd and even but not periodic.

Exercise 5.21

Give the definition of a circle.

6. Exercises: Compositions

Exercise 6.1

Represent the function $h(x) = \sin^2 x + \sin^3 x$ as the composition $g \circ f$ of two functions $y = f(x)$ and $z = g(y)$.

Exercise 6.2

Function $y = f(x)$ is given below by a list its values. Find its inverse and represent it by a similar table.

x	0	1	2	3	4
$y = f(x)$	0	1	2	4	3

Exercise 6.3

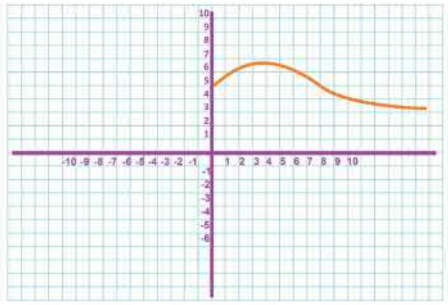
Find the formulas of the inverses of the following functions: (a) $f(x) = (x+1)^3$; (b) $g(x) = \ln(x^3)$.

Exercise 6.4

Are the following functions invertible? 1. $f(n)$ is the number of students in your class whose birthday is on the nth day of the year. 2. $f(t)$ is the total accumulated rainfall in inches t on a given day in a particular location.

Exercise 6.5

The graph of $y = f(x)$ is plotted below. Sketch $y = -f(x+5) - 6$.



Exercise 6.6

Given the tables of values of f, g, find the table of values of $f \circ g$:

x	$y = g(x)$
0	0
1	4
2	3
3	0
4	1

y	$z = f(y)$
0	4
1	4
2	0
3	1
4	2

What if the last rows were missing?

Exercise 6.7

Represent the function below as a composition $f \circ g$ of two functions:

$$h(x) = \sqrt{2x^3 + x}.$$

Exercise 6.8

Find the composition $h(x) = (g \circ f)(x)$ of the functions $y = f(x) = x^2 - 1$ and $g(y) = 3y - 1$. Evaluate $h(1)$.

Exercise 6.9

Represent the function $h(x) = 2\sin^3 x + \sin x + 5$ as the composition of two functions one of which is trigonometric.

Exercise 6.10

(a) Represent the function $h(x) = e^{x^3-1}$, as the composition of two functions f and g. (b) Provide formulas for the two possible compositions of the two functions: "take the logarithm base 2 of" and "take the square root of".

Exercise 6.11

Suppose a function f performs the operation: "take the logarithm base 2 of", and function g performs: "take the square root of". (a) Verbally describe the inverses of f and g. (b) Find the formulas for these four functions. (c) Give them domains and codomains.

Exercise 6.12

1. Represent the function $h(x) = \sqrt{x^2 - 1}$ as the composition of two functions f and g.
2. Provide a formula for the composition $y = f(g(x))$ of $f(u) = u^2 + u$ and $g(x) = 2x - 1$.

Exercise 6.13

Provide a formula for the composition $y = f(g(x))$ of $f(u) = \sin u$ and $g(x) = \sqrt{x}$.

Exercise 6.14

Provide a formula for the composition $y = f(g(x))$ of $f(u) = u^2 - 3u + 2$ and $g(x) = x$.

Exercise 6.15

Find the inverse of the function $f(x) = 3x^2 + 1$. Choose appropriate domains for these two functions.

Exercise 6.16

1. Represent the function $h(x) = \sqrt{x-1}$ as the composition of two functions.
2. Represent the function $k(t) = \sqrt{t^2 - 1}$ as the composition of three functions.
3. Represent the function $p(t) = \sin\sqrt{t^2 - 1}$ as the composition of four functions.

Exercise 6.17

(a) What is the composition $f \circ g$ for the functions given by $f(u) = u^2 + u$ and $g(x) = 3$? (a) What is the composition $f \circ g$ for the functions given by $f(u) = 2$ and $g(x) = \sqrt{x}$?

Exercise 6.18

Is the composition of two functions that are odd/even odd/even?

Exercise 6.19

Represent this function: $h(x) = \dfrac{x^3+1}{x^3-1}$, as the composition of two functions of variables x and y.

Exercise 6.20

Represent the composition of these two functions: $f(x) = \dfrac{1}{x} + 1$ and $g(y) = \sqrt{y-1}$, as a single function h of variable x. Don't simplify.

Exercise 6.21

Function $y = f(x)$ is given below by a list its values. Find its inverse and represent it by a similar table.

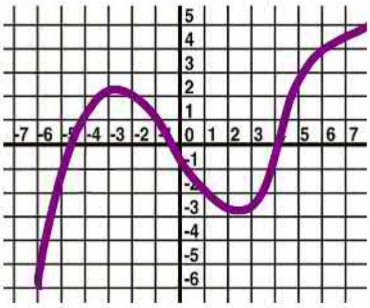
x	0	1	2	3	4
$y = f(x)$	1	2	0	4	3

Exercise 6.22

Give examples of functions that are their own inverses?

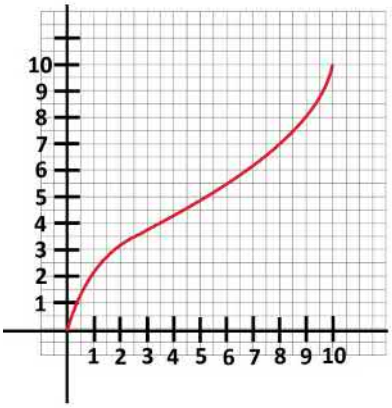
Exercise 6.23

Plot the inverse of the function shown below, if possible.



Exercise 6.24

Plot the graph of the inverse of this function:



Exercise 6.25

Represent this function: $h(x) = \tan(2x)$ as the composition of two functions of variables x and y.

Exercise 6.26

Find the composition $h(x) = (g \circ f)(x)$ of the functions $y = f(x) = x^2 - 1$ and $g(y) = \dfrac{y-1}{y+1}$. Evaluate $h(0)$.

Exercise 6.27

Find the composition $h(x) = (g \circ f)(x)$ of the functions $y = f(x) = x^2 - 1$ and $g(y) = 3y - 1$. Evaluate $h(1)$.

Exercise 6.28

Represent the composition of these two functions: $f(x) = 1/x$ and $g(y) = \dfrac{y}{y^2 - 3}$, as a single function h of variable x. Don't simplify.

Exercise 6.29

Represent this function: $h(x) = \dfrac{x^3+1}{x^3-1}$, as the composition of two functions of variables x and y.

Exercise 6.30

Function $y = f(x)$ is given below by a list of its values. Is the function one-to one? What about its inverse?

x	0	1	2	3	4
$y = f(x)$	0	1	2	1	2

Exercise 6.31

Function $y = f(x)$ is given below by a list of its values. Is the function one-to one? What about its inverse?

x	0	1	2	3	4
$y = f(x)$	7	5	3	4	6

Exercise 6.32

Functions $y = f(x)$ and $u = g(y)$ are given below by tables of some of their values. Present the composition $u = h(x)$ of these functions by a similar table:

x	0	1	2	3	4
$y = f(x)$	1	1	2	0	2

y	0	1	2	3	4
$u = g(y)$	3	1	2	1	0

Exercise 6.33

Function $y = f(x)$ is given below by a list of some of its values. Add missing values in such a way that the function is one-to one.

x	-1	0	1	2	3	4	5
$y = f(x)$	-1		4	5		2	

Exercise 6.34

Plot the graph of the function $f(x) = \dfrac{1}{x-1}$ and the graph of its inverse. Identify its important features.

Exercise 6.35

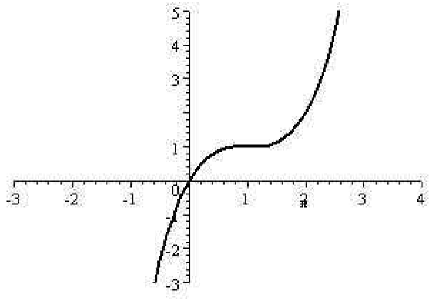
(a) Algebraically, show that the function $f(x) = x^2$ is not one-to-one. (b) Graphically, show that the function $g(x) = 2^{x+1}$ is one-to-one. (c) Find the inverse of g.

Exercise 6.36

Find the formulas of the inverses of the following functions: (a) $f(x) = (x+1)^3$; (b) $g(x) = \ln(x^3)$.

Exercise 6.37

Sketch the graph of the inverse of the function below:



Exercise 6.38

Determine whether the functions below are or are not one-to-one:

$$f(x) = (x-1)^3 \text{ and } g(x) = 2^{x-1}.$$

Exercise 6.39

Sketch the graph of the composition of the above function and (a) $y = 2x$; (b) $y = x - 1$; and (c) $y = x^2$.

7. Exercises: Linear equations

Exercise 7.1

What is the equation of the line through the points $A = (-3, 2)$ and $B = (2, 5)$?

Exercise 7.2

Set up, but do not solve, a system of linear equations for the following problem: "Suppose your portfolio is worth \$20,000 and it consists of two stocks A and B. The stocks are priced as follows: A \$2.1 per share, B \$1.5 per share. Suppose also that you have twice as much of stock A than B. How much of each do you have?"

Exercise 7.3

In an effort to find the point in which the lines $2x - y = 2$ and $-4x + 2y = 1$ intersect, a student multiplied the first one by 2 and then added the result to the second. He got $0 = 5$. Explain the result.

Exercise 7.4

Find the angle between the lines: from $(0, 0)$ to $(1, 1)$ and from $(0, 0)$ to $(1, 2)$. Don't simplify.

Exercise 7.5

Solve the system of linear equations:

$$\begin{cases} x & -y & = -1, \\ 2x & +y & = 0. \end{cases}$$

Exercise 7.6

Solve the system of linear equations:

$$\begin{cases} x & -2y & = 1, \\ 2x & +y & = 0. \end{cases}$$

Exercise 7.7

A movie theater charges \$10 for adults and \$6 for children. On a particular day when 320 people paid an admission, the total receipts were \$3120. How many were adults and how many were children?

Exercise 7.8

The taxi charges \$1.75 for the first quarter of a mile and \$0.35 for each additional fifth of a mile. Find a linear function which models the taxi fare f as a function of the number of miles driven, x.

Exercise 7.9

Given vectors $a =< 1, 2 >$, $b =< -2, 1 >$, find their magnitudes and the angle between them.

Exercise 7.10

Set up a system of linear equations – but do not solve – for the following problem: "A mix of coffee is to be prepared from: Kenyan coffee - \$3 per pound and Colombian coffee - \$5 per pound. How much of each do you need to have 10 pounds of blend with \$3.50 per pound?"

Exercise 7.11

Set up, do not solve, the system of linear equations for the following problem: "One serving of tomato soup contains 100 Cal and 18 g of carbohydrates. One slice of whole bread contains 70 Cal and 13 g of carbohydrates. How many servings of each should be required to obtain 230 Cal and 42 g of carbohydrates?"

Exercise 7.12

Solve the system of linear equations:

$$\begin{cases} x - y & = 2, \\ x + 2y & = 1. \end{cases}$$

Exercise 7.13

Solve the system of linear equations and geometrically represent its solution:

$$\begin{cases} x - 2y & = 1, \\ x + 2y & = -1. \end{cases}$$

Exercise 7.14

Geometrically represent this system of linear equations:

$$\begin{cases} x - 2y & = 1, \\ x + 2y & = 1. \end{cases}$$

Exercise 7.15

What are the possible outcomes of a system of linear equations?

Exercise 7.16

Set up, but do not solve, a system of linear equations for the following problem: "Suppose your portfolio is worth \$1,000,000 and it consists of two stocks A and B. The stocks are priced as follows: A \$2.1 per share, B \$1.5 per share. Suppose also that you have twice as much of stock A than B. How much of each do you have?"

Exercise 7.17

Give the number t that makes $X =< 3, 2, 1 >$ and $Y =< 2, t, t >$ perpendicular.

Exercise 7.18

Here are xyz-equations for two planes: $x+y-z = 0$ and $x - y + z = 0$. Explain how you can tell that these planes cut each other NOT at right angles.

Exercise 7.19

A plane has an xyz-equation $x + y = 2$. Give a vector perpendicular to the plane.

Exercise 7.20

In an effort to find the line in which the planes $2x - y - z = 2$ and $-4x + 2y + 2z = 1$ intersect, a student multiplied the first one by 2 and then added the result to the second. He got $0 = 5$. Explain the result.

Exercise 7.21

Determine whether these points lie on a straight line:

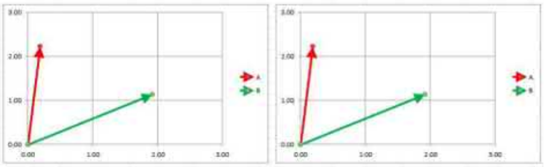
$$A = (0, -5, 5),\ B = (1, -2, 4),\ C = (3, 4, 2)\,.$$

Exercise 7.22

Find the plane through the point $P = (-1, 6, -5)$ and parallel to the vectors $A =< 1, 1, 0 >$ and $B =< 0, 1, 1 >$.

Exercise 7.23

Vectors A and B are given below. Copy the picture and illustrate graphically (a) $A+B$, (b) $A-B$, (c) $||A||$, (d) the projection of A on B, (e) the projection of B on A.



Exercise 7.24

Find the angle between the vectors $< 1, 1, 1 >$ and the x-axis. Don't simplify.

Exercise 7.25

Find the plane through the origin perpendicular to the line from $(1, 0, 0)$ and $(0, 1, 1)$.

Exercise 7.26

Find an equation of the plane through $(2, 1, 0)$ and parallel to $x + 4y - 3z = 1$.

Exercise 7.27

(a) Find the angle between the planes $x+y+z = 1$ and $x - 2y + 3z = 1$. (b) Find the equation of the line of intersection of these planes.

Exercise 7.28

Find the vector equation of the line parallel to both xy- and xz- coordinate planes an passing through $(2, 3, 1)$.

Exercise 7.29

Solve the system of linear equations:

$$\begin{cases} x & -y & = -1\,, \\ 2x & +y & = 0\,. \end{cases}$$

Exercise 7.30

Find the reduced row echelon form of the following system of linear equations. What kind of set is its solution set?

$$\begin{cases} -x & -2y & +z & = 0\,, \\ 3x & & +z & = 2\,, \\ x & -y & +z & = 1\,. \end{cases}$$

Exercise 7.31

Represent the system of linear equations as a matrix equation:

$$\begin{cases} x & -y & +z & = -1\,, \\ 3x & & +z & = 2\,, \\ 2x & +y & +z & = 1\,. \end{cases}$$

Exercise 7.32

Represent this matrix equation as a system of linear equations:

$$\begin{bmatrix} 1 & 3 & 2 \\ 0 & 2 & 0 \\ -1 & 0 & 3 \end{bmatrix} \begin{bmatrix} x \\ y \\ z \end{bmatrix} = \begin{bmatrix} 1 \\ 2 \\ 3 \end{bmatrix}.$$

Exercise 7.33

Give explicitly the solution set of the system of linear equations represented by its augmented matrix:

$$\left[\begin{array}{cccc} 1 & 0 & 2 & -1 \\ 0 & 1 & 1 & 2 \end{array}\right].$$

Exercise 7.34

Find scalars a and b such that

$$a < 1,2 > +b < -1,3 >=< 1,7 > \ .$$

Exercise 7.35

Is it possible for a system of linear equations to have: (a) no solutions, (b) exactly one solution, (c) exactly two solutions, (d) infinitely many solutions? Give an example or explain why it's not possible.

Exercise 7.36

Find the set of all vectors in $\mathbf{R}^2$ that are orthogonal to $< -1,3 >$. Write the set in the standard form of a line through the origin.

Exercise 7.37

Compute:

$$\begin{bmatrix} 1 & 3 & 2 \\ 0 & 2 & 0 \\ -1 & 0 & 3 \end{bmatrix} \begin{bmatrix} 1 \\ 0 \\ 2 \end{bmatrix}.$$

Exercise 7.38

Is it possible that a system of linear equations has (a) no solutions, (b) one solution, (c) two solutions, (3) infinitely many solutions? Give an example or explain why it's not possible.

Exercise 7.39

Find four 2×2 matrices A such that $AA = I$.

Exercise 7.40

Determine all 2×2 matrices with $AA = I$.

Exercise 7.41

Find all 2×2 matrices A that commute with all 2×2 matrices.

Exercise 7.42

Show that the system $ax + by = r, cx + dy = s$ has a unique solution if $ad - bc$ is not zero.

Exercise 7.43

Show that $[2,4,2],[3,2,0],[1,-2,2]$ are linearly independent.

Exercise 7.44

Suppose u and v are linearly independent. Let $x = u + v$ and $y = u - v$. Are x and y linearly independent?

Exercise 7.45

Find the Hermite form of the given 3×4 matrix.

Exercise 7.46

Show that row operations can be undone by other row operations.

Exercise 7.47

Are $[1,2],[1,3],[1,4]$ linearly independent? Prove.

8. Exercises: Vector algebra

Exercise 8.1

Represent the vector $< 1, 2 >$ as a linear combination of the vectors $< 0, 2 >$ and $< 1, 1 >$.

Exercise 8.2

Find a vector in $\mathbf{R}^3$ that *cannot* be represented as a linear combination of the vectors $< 1, 0, 0 >$, $< 1, 1, 1 >$, and $< 2, 1, 1 >$.

Exercise 8.3

Find the value of k so that the line containing the points $(-6, 0)$ and $(k, -5)$ is parallel to the line containing the points $(4, 3)$ and $(1, 7)$.

Exercise 8.4

Give each of the following functions a domain and codomain and determine if it is one-to-one or onto: (a) $y = x^2$; (b) $y = \sqrt{x}$; (c) $z = x + y$; (d) $(x, y) = (\sin t, \cos t)$.

Exercise 8.5

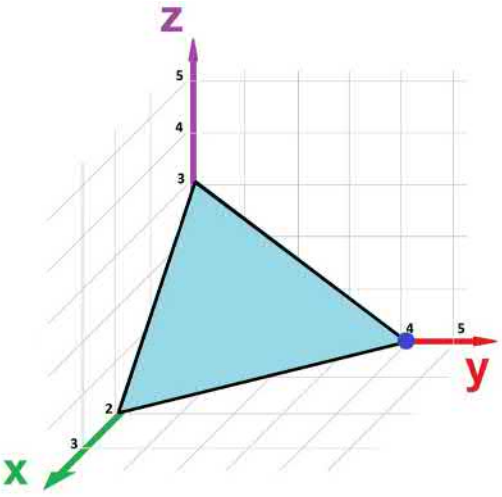
Represent the vector $< 1, 2 >$ as a linear combination of the vectors $< 0, 2 >$ and $< 1, -1 >$.

Exercise 8.6

Find a vector in $\mathbf{R}^3$ that *cannot* be represented as a linear combination of the vectors $< 1, 0, 0 >$, $< 1, 1, 1 >$, and $< 2, 1, 1 >$.

Exercise 8.7

The plane below is the graph of a function. Find the function.

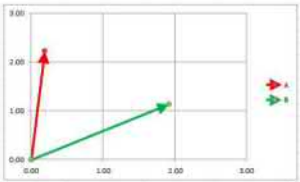


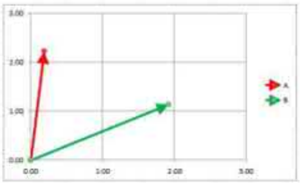
Exercise 8.8

Find the matrix of a linear operator that stretches the axes by 2 and 3 respectively, rotates the plane by 90 degrees, and then flips it about the x-axis.

Exercise 8.9

Vectors A and B are given below. Copy the picture and illustrate graphically: (a) $A + B$, (b) $A - B$, (c) $||A||$.





Exercise 8.10

Find the angle between the vectors $< 1, 1 >$ and $< 1, 2 >$. Don't simplify.

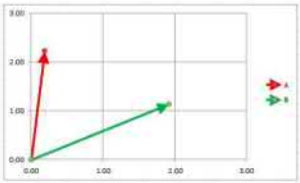
Exercise 8.11

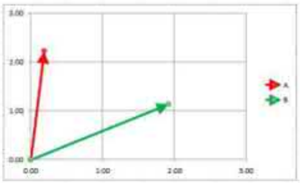
Solve the system of linear equations:

$$\begin{cases} x & -y & = -1, \\ 2x & +y & = 0. \end{cases}$$

Exercise 8.12

Vectors A and B are given below. (a) Illustrate graphically: $A + B$ and $A - B$. (b) Give the components of these vectors. (c) Compute $A \cdot B$.





Exercise 8.13

(a) Let

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

Compute A^2, A^3. (b) Formulate and prove a theorem about 3×3 matrices based on the outcome of part (a).

9. Exercises: Eigenvalues and eigenvectors

Exercise 9.1

Find the eigenvalues of the following matrix:

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

Exercise 9.2

Arguing from the definition, show that a rotation cannot have eigenvectors.

Exercise 9.3

Find the eigenvalues of the following matrix:

$$\begin{bmatrix} 1 & 1 \\ 1 & 1 \end{bmatrix}$$

Exercise 9.4

Find scalars a and b such that $a(1,2) + b(-1,3) = (1,12)$.

Exercise 9.5

A is singular iff 0 is an eigenvalue.

Exercise 9.6

Find the eigenvalues, the eigenvectors, and the bases of the eigenspaces of the matrix:

$$\left[\begin{array}{ccc} 1 & 0 & 1 \\ 0 & 1 & 0 \\ 1 & 0 & 1 \end{array}\right].$$

Start: $\implies \det(\lambda I - A) = 0 \implies \lambda = 0, \ldots$

Exercise 9.7

Find the eigenvalues, the eigenvectors, and bases of the eigenspaces of the matrix:

$$A = \left[\begin{array}{cc} 1 & 1 \\ 0 & 0 \end{array}\right].$$

10. Exercises: Transformations

Exercise 10.1

Describe – both geometrically and algebraically – two different transformations that make a 1×1 square into a 2×3 rectangle.

Exercise 10.2

Find the matrix of a linear operator that stretches the axes by 2 and 3 respectively, rotates the plane by 90 degrees, and then flips it about the x-axis.

Exercise 10.3

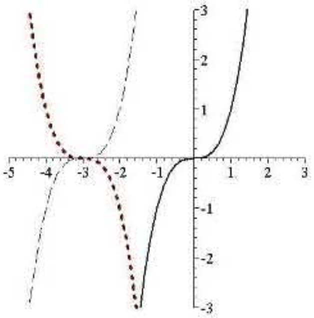
What happens to the domain and the range of a function under the six basic transformations?

Exercise 10.4

How do the six basic transformations affect a function being one-to-one or onto?

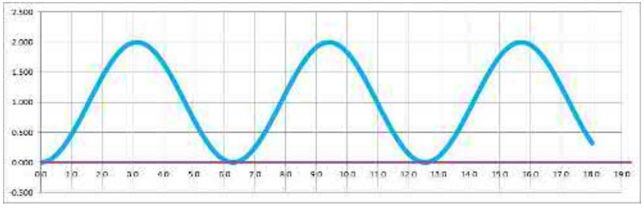
Exercise 10.5

The graph drawn with a solid line is $y = x^3$. What are the other two?



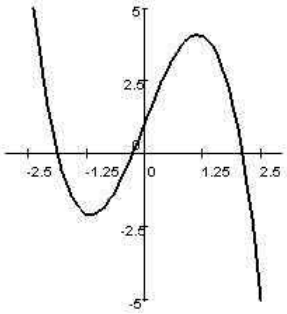
Exercise 10.6

The graph below is the graph of the function $f(x) = A \sin x + B$ for some A and B. Find these numbers.



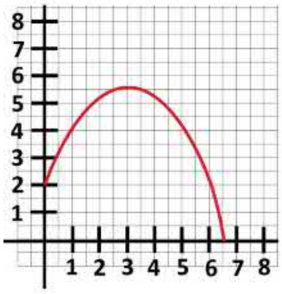
Exercise 10.7

The graph of function f is given below. Sketch the graph of $y = 2f(x+2)+2$. Explain how you get it.



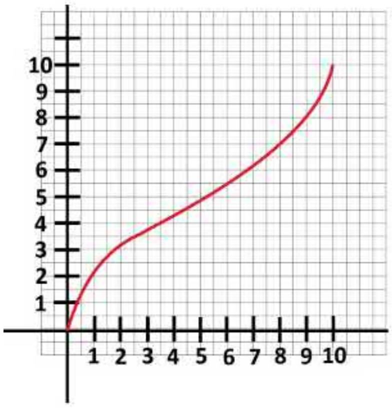
Exercise 10.8

The graph of the function $y = f(x)$ is given below. Sketch the graph of $y = 2f(x)$ and then $y = 2f(x) - 1$.



Exercise 10.9

The graph of the function $y = f(x)$ is given below. Sketch the graph of $y = \frac{1}{2}f(x)$ and then $y = \frac{1}{2}f(x-1)$.



Exercise 10.10

Plot the graph of a function that is both odd and even.

11. Exercises: Advanced

Exercise 11.1

Is $B = \{1, x, x^2, x^3, ...\}$ a basis of $\mathbf{C}(\mathbf{R})$, the space of all continuous functions?

Exercise 11.2

Suppose $\mathbf{C}(\mathbf{R})$ is the vector space of all continuous functions. Let the function $T : \mathbf{C}(\mathbf{R}) \to \mathbf{R}$ be defined by

$$T(f) = f'(0), \text{ for all } f \in \mathbf{C}(\mathbf{R}).$$

Show that T is linear.

Exercise 11.3

Let U and V be 2-dimensional subspaces of $\mathbf{R}^3$. Prove that $U \cap V \neq 0$.

Exercise 11.4

Suppose V is a vector space with operations: $v + w = 0$ and $rv = 0$ for all $v, w \in V$, $r \in \mathbf{R}$. How many elements does V have? Prove by using only the axioms.

Exercise 11.5

Express $f(x) = (x-1)^2 - x$ as a linear combination of the power function: $1, x, x^2, x^3, ...$

Exercise 11.6

Find the set of all vectors in $\mathbf{R}^2$ that are orthogonal to $(-1, 3)$. Write the set in the standard form of a line through the origin.

Exercise 11.7

Find the standard inner product of $f(x) = \cos x$ and $g(x) = 1$ in $C[0, \pi]$.

Exercise 11.8

Suppose a is an element of an inner product space V. Let S be the set of all vectors orthogonal to a plus 0. Show that S is a subspace of V.

Exercise 11.9

Suppose S is a subspace of V and $\dim S = \dim V$. From the definition of the dimension, prove that $S = V$.

Exercise 11.10

Suppose V is a subspace of $C^1[0, 1]$ spanned by $\sin x, \cos x$. Define A as $A(f) = f'$. (a) Show that A is well defined on V. (b) Find the matrix of A.

Exercise 11.11

Find all possible values for rank A if A is an $n \times m$ matrix.

Exercise 11.12

Suppose a, b, c are 3 linearly independent vectors. Suppose A is 3×3 matrix of rank 3. Are the 3-vectors Aa, Ab, Ac linearly independent?

Exercise 11.13

Find the row rank of a given 4×4 matrix.

Exercise 11.14

Suppose A is an $n \times n$ matrix with only 0 entries on the diagonal and below. Show that $A^n = 0$.

Exercise 11.15

Solve a given 4×4 system.

Exercise 11.16

Suppose A is an $n \times n$ matrix with only 0 entries on the diagonal and below. Show $B = I - A$ is invertible and $B^{-1} = I + A + A^2 + \ldots + A^{n-1}$. Prove the problem by means of inverses.

Exercise 11.17

Find the inverse of a given 3×3 matrix.

Exercise 11.18

What is the smallest subset of $\mathbf{R}$ that contains 1/2 and closed under (a) addition, (b) multiplication.

Exercise 11.19

Determine whether

$$S = \{f \in C^1(\mathbf{R}) : \ 2f'(x) + x^2 f(x) = 0 \ \text{ for all } \ x\}$$

is a subspace of $C^1(\mathbf{R})$.

Exercise 11.20

Let $\mathbf{P}$ be the space of all polynomials. Show that $\mathbf{P}$ is not spanned by any finite set.

Exercise 11.21

Show that $\{1, (x-1), (x-1)^2, \dots\}$ is a basis of $\mathbf{P}$.

Exercise 11.22

Does the set $\{(1,2,-1),(2,4,2),(1,2,3),(-1,-2,1)\}$ span $\mathbf{R}^3$?

Exercise 11.23

Suppose S is a subspace of V and $\dim S = \dim V$. From the definition of the dimension, prove that $S = V$.

Exercise 11.24

Find the dimension of the space of all symmetric 3×3 matrices.

Exercise 11.25

The scalar multiplication in a vector space is a linear function in some sense. In what sense and why?

Exercise 11.26

Suppose $f : V \to W$ is linear and surjective. Suppose that $\operatorname{Span} S = V$, where S is a subset of V. Show that $\operatorname{Span} f(S) = W$.

Exercise 11.27

Consider the function $f : \mathbf{R}^3 \to \mathbf{R}^3$ that rotates each point about the x-axis through an angle a. Prove that f is linear and find its matrix.

Exercise 11.28

Prove that $G(f) = f(0) + f'(0)$ is linear.

Exercise 11.29

Find the determinant of the 7×7 matrix with the following entries: $1, 2, 3, \dots, 49$.

Exercise 11.30

Prove that the set of all non-zero rational numbers, $\mathbf{Q} \backslash \{0\}$, is closed under division. Hint: Prove that $r \in \mathbf{Q}\,,\ q \in \mathbf{Q} \implies \dfrac{r}{q} \in \mathbf{Q}$. Start with

$$r = \frac{a}{b},\ q = \frac{c}{d},\ a,b,c,d \in \mathbf{Z} \setminus \{0\}\,.$$

Exercise 11.31

Prove that the intersection of two subspaces is always a subspace.

Exercise 11.32

Prove that the set of all diagonal $n \times n$ matrices, i.e., ones with $a_{ij} = 0$ for all $i \neq j$, is a subspace of $M(n,n)$.

Exercise 11.33

Provide the details for the solution of the following system of linear equations:

$$\begin{array}{llll} x & +y & -z & = 1 \\ x & -y & +2z & = 0 \\ 3x & +y & & = 2 \end{array}$$

We start with:

$$\left[\begin{array}{ccc|c} 1 & 1 & -1 & 1 \\ 1 & -1 & 2 & 0 \\ 3 & 1 & 0 & 2 \end{array}\right] \to \left[\begin{array}{ccc|c} 1 & 0 & -\frac{1}{2} & \frac{1}{2} \\ 0 & 1 & -\frac{3}{2} & \frac{1}{2} \\ 0 & 0 & 0 & 0 \end{array}\right] \to$$

$$\left\{\begin{array}{ll} x_1 & = \frac{1}{2}x_3 + \frac{1}{2} \\ x_2 & = \frac{3}{2}x_3 + \frac{1}{2} \\ x_3 & = x_3 \end{array}\right. \to \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} = t \begin{bmatrix} \frac{1}{2} \\ \frac{3}{2} \\ 1 \end{bmatrix} + \begin{bmatrix} \frac{1}{2} \\ \frac{1}{2} \\ 0 \end{bmatrix}$$

What is the dimension of its solution set?

Exercise 11.34

Is it possible that a homogeneous system of linear equations has (a) no solutions, (b) one solution, (c) two solutions, (3) infinitely many solutions? Give an example of such a system or explain why it's not possible.

Exercise 11.35

Prove that the set

$$\{(2,1,1),(1,2,1),(0,1,1),(1,1,1)\}$$

spans $\mathbf{R}^3$. Hint: Use the proof of the Reduction Theorem as a recipe for finding a subset of this set that is a basis of $\mathbf{R}^3$. Find one vector that is a linear combination of the rest and remove it.

$$a_1 \begin{bmatrix} 2 \\ 1 \\ 1 \end{bmatrix} + a_2 \begin{bmatrix} 1 \\ 2 \\ 1 \end{bmatrix} + a_3 \begin{bmatrix} 0 \\ 1 \\ 1 \end{bmatrix} = \begin{bmatrix} 1 \\ 1 \\ 1 \end{bmatrix}.$$

What you really need:

$$\operatorname{rank}\begin{bmatrix} 2 & 1 & 0 \\ 1 & 2 & 1 \\ 1 & 1 & 1 \end{bmatrix} = ?$$

Exercise 11.36

Let

$$B = \{x-1, x+1\} .$$

(a) Prove that B is a basis of $\mathbf{P}_1$. (b) Find the coordinate vector $[3x-5]_B$.

Exercise 11.37

Suppose A is an invertible matrix. What is $(A^m)^{-1}$?

Exercise 11.38

Given basis $\{1, x, x^2\}$ of the space $\mathbf{P}_2$ of degree ≤ 2 polynomials, find the change of basis matrix for the new basis $\{2, x-x^2, x-1\}$. Start:

$$2 = 2 \cdot 1 = \begin{bmatrix} 2 \\ 0 \\ 0 \end{bmatrix}$$

$$x - x^2 = 0 \cdot 1 + 1 \cdot x + (-1)x^2 = \begin{bmatrix} 0 \\ 1 \\ -1 \end{bmatrix}$$

$$x - 1 = -1 \cdot 1 + 1 \cdot x + 0 \cdot x^2 = \, ?$$

Exercise 11.39

Suppose V is the space of differentiable at 0 functions of two variables. Suppose $A : V \to \mathbf{R}^2$ is given by $A(f) = \operatorname{grad} f(0)$. Prove that A is linear and find its kernel.

Exercise 11.40

Prove that the composition of two onto functions is onto. Start:

$$X \xrightarrow{f} Y \xrightarrow{g} Z ,$$

where f and g are onto. Prove $g \circ f$ is onto if and only if for every $z \in Z$ there is $x \in X$ such that $(g \circ f)(x) = z$. Does it exist? First, g is onto, so there is $y \in Y$ such that $g(y) = z$. Second, f is onto, so there is $x \in X$ such that $f(x) = y$.

Exercise 11.41

Prove that the composition of two one-to-one functions is one-to-one.

Exercise 11.42

Suppose we have a vector space V with basis S. Give examples of linear operators $A : V \to V$ satisfying (a) A is an isomorphism but not the identity. (b) The dimension of the image of A is equal to 1. (b) The dimension of the kernel of A is equal to 1. Hint 1: Try $V = \mathbf{R}^3$ first. Hint 2: Use the basis.

Exercise 11.43

Show that the set of all even functions $f : \mathbf{R} \to \mathbf{R}$ is a vector space. What about the set of all the odd functions?

Exercise 11.44

Give the definition of linear independence in vector spaces. Give examples of (a) three 2×2 matrices that are linearly independent, and (b) three functions that are linearly dependent but not multiples of each other.

Exercise 11.45

Suppose A and B are two invertible matrices. Express $(AB)^{-1}$ in terms of A^{-1} and B^{-1}.

Exercise 11.46

Are the following functions linear? (a) $f(x,y,z) = (0,0,0)$; (b) $g(x,y,z) = (x-y, y-z, z-x)$; (c) $h(x,y,z) = (1,1,1)$; (d) $k(x,y,z) = ||(x,y,z)||$. Just the answers.

Exercise 11.47

Suppose $\{v_1, ..., v_m\}$ is a linearly independent subset of a vector space V and suppose $A : V \to U$ is a linear one-to-one operator. Prove that $\{Av_1, ..., Av_m\}$ is a linearly independent subset of U.

Exercise 11.48

(a) Give the definition of the determinant of an $n \times n$ matrix. (b) Find the determinant of an upper-triangular matrix (all entries below the main diagonal are 0). (c) Is the determinant a linear operator? Prove or disprove.

Exercise 11.49

We know that if S is a basis of V then every element of V can be represented as a linear combination of the elements of S. Prove that such a representation is unique.

Exercise 11.50

(a) Give the definition of an inner product space. (b) State the Cauchy-Schwarz inequality. (c) Define the angle between two vectors in an inner product space. Prove that it's well-defined.

Exercise 11.51

Suppose a is an element of an inner product space V and suppose S is the set of all vectors orthogonal to a, plus 0. Prove that S is a subspace of V.

Exercise 11.52

Find the determinant of the $n \times n$ matrix with entries $1, 2, 3, \ldots, n^2$.

Exercise 11.53

Let V be the space of infinite sequences $\{x_1, \ldots, x_n, \ldots\}$. Find an infinite dimensional subspace U of V that can be equipped with a nontrivial inner product.

Exercise 11.54

Let $A(f) = f - f'$, $A\colon C^1 \to C^1$. Find the kernel of A

Exercise 11.55

$$A = \begin{bmatrix} 2 & 0 \\ 0 & 3 \end{bmatrix} \sim \begin{bmatrix} 3 & 0 \\ 0 & 2 \end{bmatrix} = B\,.$$

Hint: We need P with $A = P^{-1}BP$.

Exercise 11.56

Find A, B such that $\operatorname{tr} A = \operatorname{tr} B$ but $\det A \neq \det B$. Hint:

$$\begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} \begin{bmatrix} 1 & 1 \\ 1 & 1 \end{bmatrix}$$

Exercise 11.57

Suppose $T\colon \mathbf{P}_3 \to \mathbf{P}_3$ is given by $T(f)(x) = xf'(x)$. Find the matrix of T. Start: This means (not $T(f(x))$): $T(f) = g$ and $g(x) = xf'(x)$ for all x. Basis of $\mathbf{P}_3$ is $\{1, x, x^2, x^3\}, \dim \mathbf{P}^3 = 4$.

$$\begin{aligned} T(x^3) &= x \cdot 3x^2 &&= 3x^3 \\ T(x^2) &= x \cdot 2x &&= 2x^2 \\ T(x) &= x \cdot 1 &&= x \\ T(1) &= x \cdot 0 &&= 0 \end{aligned}$$

Rewrite:

1. $e_1 = 1$
2. $e_2 = x$
3. $e_3 = x^2$
4. $e_4 = x^3$

Then

1. $T(e_1) = 0$
2. $T(e_2) = e_2$
3. $T(e_3) = 2e_3$
4. $T(e_4) = 3e_4$

We write these are columns in terms of $e_1, \ldots, e_4$. Then

$$e_2 = \begin{bmatrix} 0 \\ 1 \\ 0 \\ 0 \end{bmatrix}$$

etc.

$$T = \begin{bmatrix} 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 2 & 0 \\ 0 & 0 & 0 & 3 \end{bmatrix}$$

Finish the solution.

Exercise 11.58

Compute

$$\begin{bmatrix} 0 & 0 & a \\ 0 & b & 0 \\ c & 0 & 0 \end{bmatrix}^{-1} = ?\ ;\ a, b, c \neq 0$$

Start:

$$\left[\begin{array}{ccc|ccc} 0 & 0 & a & 1 & 0 & 0 \\ 0 & b & 0 & 0 & 1 & 0 \\ c & 0 & 0 & 0 & 0 & 1 \end{array}\right] \to$$

$$R_1 \leftrightarrow R_3 \left[\begin{array}{ccc|ccc} 0 & 0 & 0 & 0 & 0 & 1 \\ 0 & b & 0 & 0 & 1 & 0 \\ 0 & 0 & a & 1 & 0 & 0 \end{array}\right] \to$$

$$\left(\frac{1}{c}R_1, \frac{1}{b}R_2, \frac{1}{a}R_3\right) \left[\begin{array}{ccc|ccc} 1 & 0 & 0 & 0 & 0 & \frac{1}{c} \\ 0 & 1 & 0 & 0 & \frac{1}{b} & 0 \\ 0 & 0 & 1 & \frac{1}{a} & 0 & 0 \end{array}\right]$$

Finish the solution.

Index