

Math review

1 Notation and Definitions

This section will go over some notation and definitions that are commonly used throughout the course.

The Greek letter "sigma" is used when we are repeatedly adding similar things together in a sequence. Sometimes such sequences are finite, but they can also be infinite! In any case, it is convenient to have compact, succinct notation to use. Let's look at some examples.

We use Sigma notation to write the sum of integers from 1 to n in the following way.

$$\sum_{i=1}^n i = 1 + 2 + 3 + \cdots + n$$

The value underneath the letter Σ tells you what symbol the index is represented by (in this case i) and where to begin (in this case at the number 1). The value on top of the letter Σ tells you where the sequence will end (in this case at the number n).

Below we have an example of an infinite sum represented with Sigma notation.

$$\sum_{i=1}^{\infty} \left(\frac{1}{2}\right)^i = \left(\frac{1}{2}\right)^1 + \left(\frac{1}{2}\right)^2 + \left(\frac{1}{2}\right)^3 + \left(\frac{1}{2}\right)^4 + \cdots = \frac{1}{2} + \frac{1}{4} + \frac{1}{8} + \frac{1}{16} + \cdots$$

We can see that the index, i , begins at 1, and the summation is infinitely long.

One of the fundamental objects in calculus are functions.

Definition 1 A **function** is a mathematical object that relates a set of inputs to exactly one output.

For example, below are examples of functions in two dimensions.

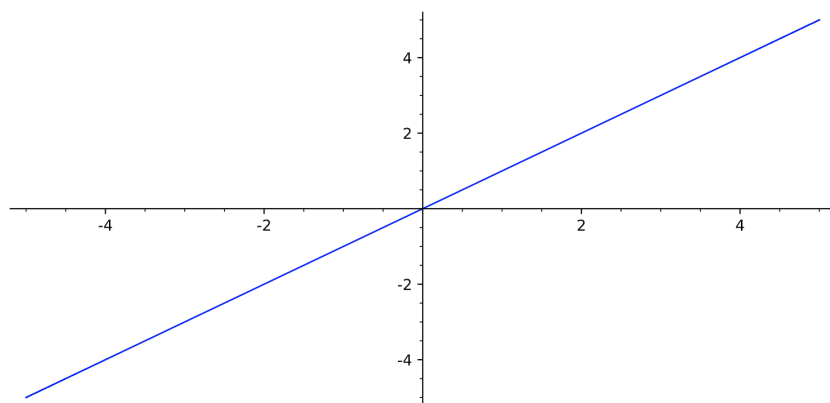


Figure 1: The graph of $g(x) = x$

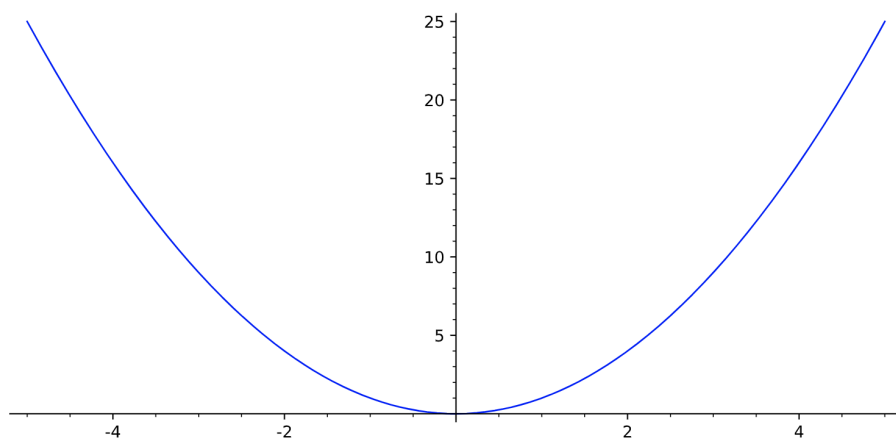


Figure 2: The graph of $f(x) = x^2$

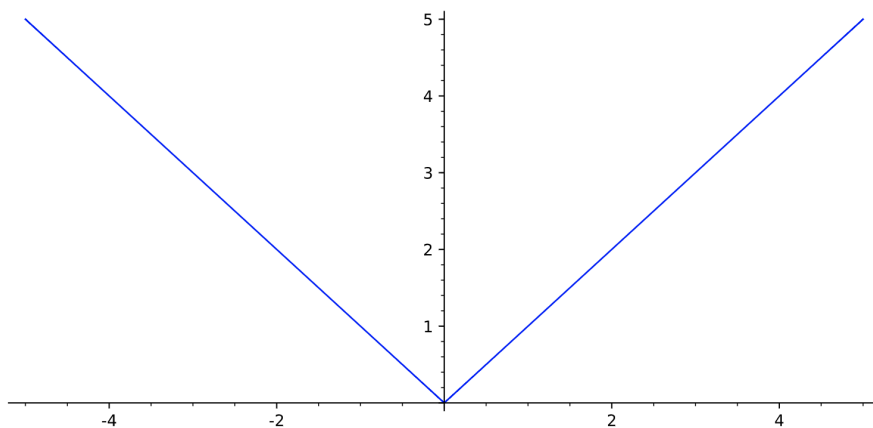


Figure 3: The graph of $h(x) = |x|$

Below is a special function that we will see in videos to come. It is known as the sign function and is defined as

$$\text{sign}(x) = \begin{cases} -1 & x < 0 \\ 0 & x = 0 \\ 1 & x > 0 \end{cases}$$

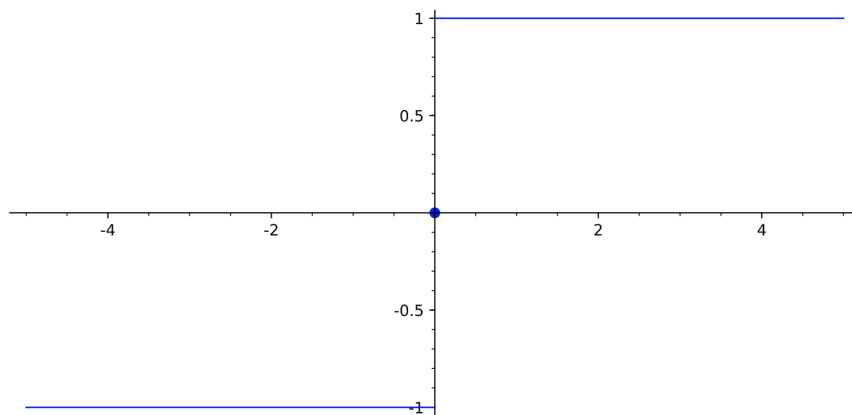


Figure 4: The graph of the sign function.

In practice, we often arbitrarily assign $\text{sign}(0)$ to be $+1$ or -1 to create a two-case definition instead of a three-case definition. Most commonly $\text{sign}(0) = 1$ is used. In many applications, an exact 0 will only happen with infinitesimal probability.

Compare the graphs in Figures 1 and 2 to those in Figures 3 and 4. Do you notice anything about the shapes of the graphs? There are two things to notice here. First, the graphs of Figures 1 and 2 are always connected—you'd never have to lift your pencil if you were to draw them. This is known as **continuity**. Figures 1 and 2 are of continuous graphs. Whereas Figure 4 is discontinuous (you'd have to lift your pencil to draw it). (There is a more rigorous mathematical definition for continuity which is out of the scope of this course.) The second thing to notice is that the graphs of Figures 1 and 2 are **smooth**, whereas the graph of Figure 3 has a sharp corner when $x = 0$.

In the Calculus section we will talk in more detail that, in order to be differentiable, a function must be smooth and continuous. You don't have to know what that means at the moment, we just want to highlight these two properties and emphasize that they are important. Figures 1 and 2 are differentiable, but Figures 3 and 4 are not.

Another important mathematical object is polynomials.

Definition 2 A ***polynomial*** is a mathematical expression that is made of constants and variables using addition, multiplication, and non-negative integer exponents.

A polynomial in a single variable, x , with constants, c_k , is given by the following form

$$\sum_{k=0}^n c_k x^k = c_0 + c_1 x + c_2 x^2 + \cdots + c_n x^n$$

The degree of a polynomial is the highest of the degrees out of each of its individual terms. For example, the polynomial $3x^5 - 9x^3 + 4$ has degree 5, polynomial $x + 1$ has degree 1, and polynomial -4 has degree 0. The degree of a polynomial is sometimes referred to as the **order** of the polynomial.

Definition 3 A ***p-ordered polynomial*** is a polynomial for which the highest degree out of each of its individual terms is p . It is a polynomial with the following form

$$\sum_{k=0}^p c_k x^k = c_0 + c_1 x + c_2 x^2 + \cdots + c_p x^p$$

where c_p is nonzero.

Polynomials are not constrained to just one variable. In fact, they can have as many variables as you like.

Definition 4 A ***multivariate polynomial of order p*** is a p -ordered polynomial with more than one variable.

For example, a 2-ordered polynomial, P with two variables, x, y has the following form

$$P(x, y) = c_{00} + c_{10}x + c_{01}y + c_{11}xy + c_{21}x^2y + c_{12}xy^2 + c_{22}x^2y^2$$

where all the c_{ij} s are constants.

Throughout this course, we refer to sets and objects that belong to sets. Informally, we can think of a set as a collection of objects. The following is some notation that is used when working with sets. Often, the notation used for sets is curly braces. For example, $\{1, 2, 3\}$ is the set containing the numbers 1, 2 and 3. The empty set, namely the set containing no objects, is usually denoted with the Greek letter ϕ . If we want to say that element a belongs in set A , we use the notation $a \in A$.

Another important notion is set inclusion. If A and B are both sets, and all the elements of A are also elements of B , we say that $A \subseteq B$. For example, $\{1, 3\} \subseteq \{1, 2, 3\}$ says that all the elements of the set $\{1, 3\}$ are contained in the set $\{1, 2, 3\}$.

In the course we will define something called a 'loss function'. You don't need to worry about that at the moment, but it requires us to develop some formal way to measure distances. One special kind of function known as a **metric** is used to do just that: it defines some notion of distance between any two points in a set. More formally we have the following definition.

Definition 5 A **metric space** is the combination of a set, S with a metric function, d , that satisfies the following four conditions:

- For any point $a \in S$, $d(a, a) = 0$
- For any points $a, b \in S$, such that $a \neq b$, $d(a, b) > 0$
- For any points $a, b \in S$, $d(a, b) = d(b, a)$
- For any points $a, b, c \in S$, $d(a, c) \leq d(a, b) + d(b, c)$ (triangle inequality)

One of the primary distance metrics we use is the Euclidean metric—more on that in the videos. The Euclidean metric leverages the Pythagorean Theorem. Note that a right triangle is a triangle where one angle is equal to 90 degrees (or $\frac{\pi}{2}$ radians). The side-length opposite to the 90 degree angle is known as the hypotenuse.

Theorem 1 (Pythagorean Theorem) For a right triangle with side-lengths a and b and hypotenuse c , we have that $a^2 + b^2 = c^2$.

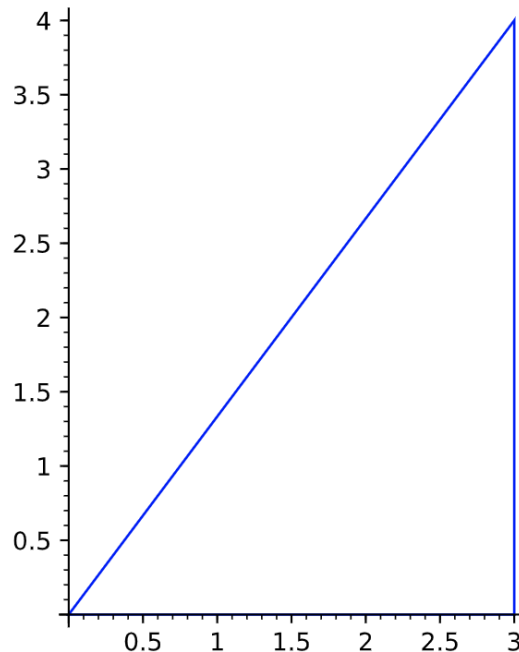


Figure 5: A right triangle with side-lengths equal to 3, 4, and 5. Notice that $3^2 + 4^2 = 5^2$ by the Pythagorean Theorem.

You may also run across interval notation. An interval is a range of numbers between two given points. The notation uses square brackets for inclusivity, and round brackets for exclusivity. For example, $[0, 1]$ are all the real numbers between 0 and 1, including 0 and 1 themselves. Whereas $(0, 1)$ are all the real numbers between 0 and 1, not including 0 and 1. Also, we could have mixed bracket such as $[0, 1)$ which includes all numbers between 0 and 1, including 0 but not including 1. And we might also have $(0, 1]$ which includes all numbers between 0 and 1, including 1 but not including 0.

As we will see in this course, and even later in this review, we are often interested in maximizing or minimizing some function. To indicate this, we use the notation min and max. For example, say we have function $f(x) = x^2$ and interval $I = [2, 5]$. Then

$$\min_{x \in I} \{f(x)\} = 4 \text{ and } \max_{x \in I} \{f(x)\} = 25$$

But we don't always want to know the maximum or minimum values themselves, but rather **where** the maximum or minimum values occur. In such cases, we use the arg min and arg max notation. For example, say we have function

$f(x) = x^2$ and interval $I = [2, 5]$ as above. Then

$$\arg \min_{x \in I} \{f(x)\} = 2 \text{ and } \arg \max_{x \in I} \{f(x)\} = 5$$

because the minimum value of 4 occurs when $x = 2$ and the maximum value of 25 occurs when $x = 5$.

2 Calculus

Throughout this course we will rely heavily on calculus because it can be used for optimization. Before we get into how that works, we need to go over some definitions. When it comes to functions, there are two notions of maximum and minimum values. One is called ‘local’ and the other is called ‘global’.

Definition 6 *The value $f(x)$ is a **local minimum** of function f if there exists an interval I around x such that for any $z \in I$, $f(x) \leq f(z)$.*

Definition 7 *The value $f(x)$ is a **local maximum** of function f if there exists an interval I around x such that for any $z \in I$, $f(x) \geq f(z)$.*

Definition 8 *The value $f(x)$ is a **global minimum** of function f if for any z in the domain of f , we have that $f(x) \leq f(z)$.*

Definition 9 *The value $f(x)$ is a **global maximum** of function f if for any z in the domain of f , we have that $f(x) \geq f(z)$.*

Note that you can have many local minimum values and local maximum values, but there is one unique global minimum value and global maximum value. These definitions are best understood using a picture.

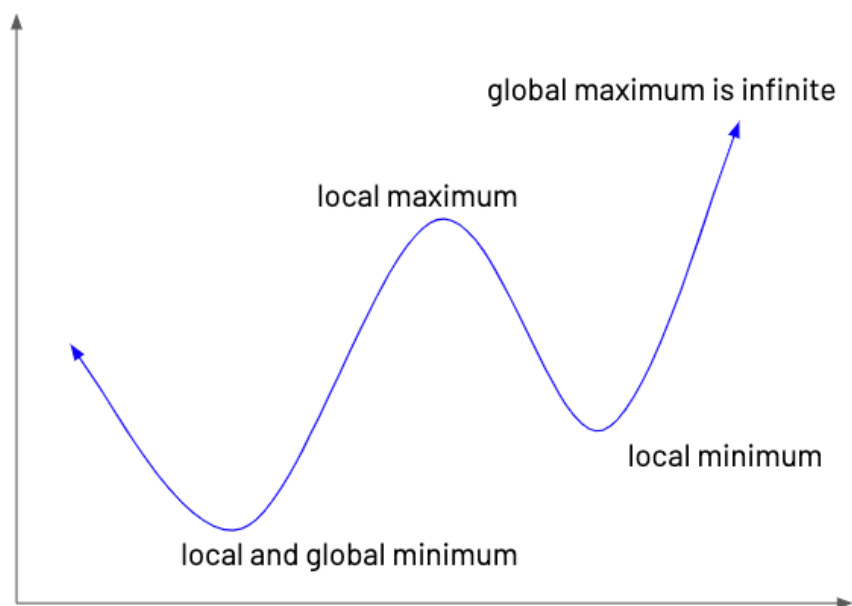


Figure 6: A graph with local and global maximums and minimums indicated.

The goal with derivatives, or differentiation, is to find the slope of the tangent line at any given point on a graph. (A line is tangent to a curve or curved surface if it touches that curve at exactly one point.)

We see a 2-dimensional example of this in the figure below.

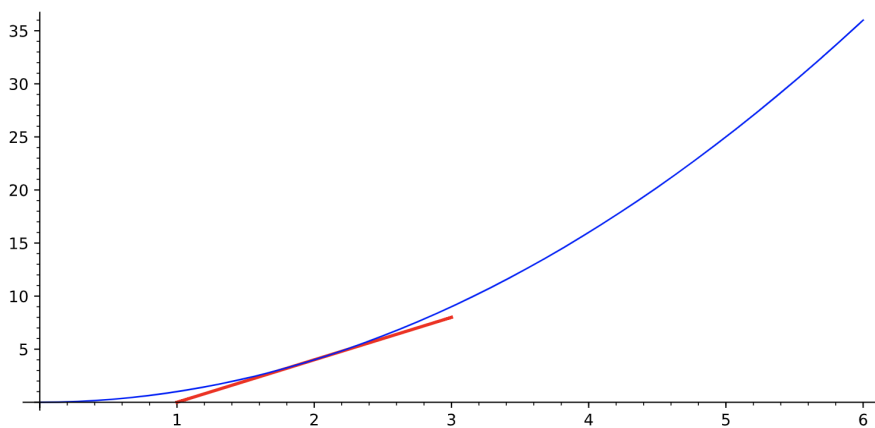


Figure 7: The graph of x^2 in blue with the tangent line at $x = 2$ in red.

We can define 'derivative' more formally, and there are many techniques for

calculating derivatives, but those are beyond the scope of this course.

However, there are some particular cases worth noting (that we hinted at earlier in the math review), where the derivative does not exist. The derivative does not exist at points where a function is discontinuous and also at points where a function is not smooth. This makes sense intuitively if you try to visualize where the tangent line should appear. For a smooth and continuous curve, it is relatively straightforward to picture; but for a discontinuous function, or one with a sharp corner, the notion of 'the tangent line' doesn't make sense.

For example, the graph of absolute value x has a sharp point at $x = 0$. It is not differentiable at that point.

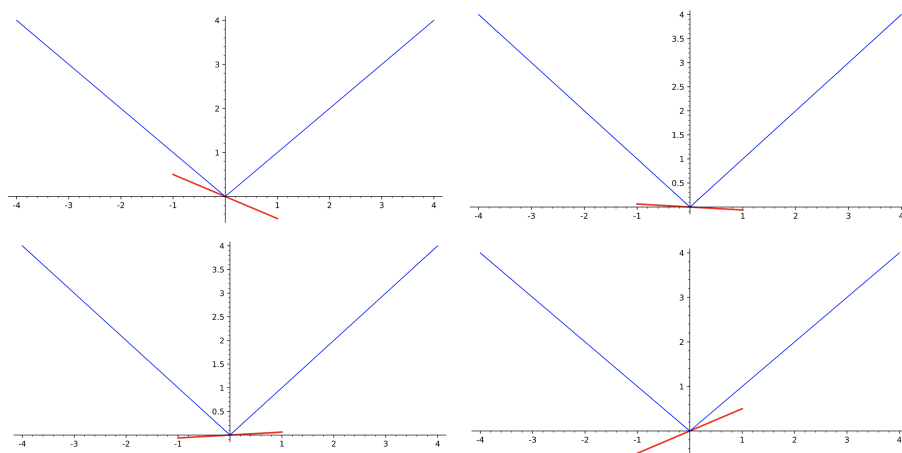


Figure 8: The graph of $|x|$ in blue with a bunch of tangent lines in red. Notice that no unique tangent line exists for this sharp corner at $x = 0$, and so no derivative exists at this point.

Why do we care about derivatives so much? Well, differentiation is one of the key ideas when we try to optimize smooth and continuous functions. Optimization is all about maximizing and minimizing some function. So, what do we notice about the slope of the tangents at the local maximum and local minimum values of this function?

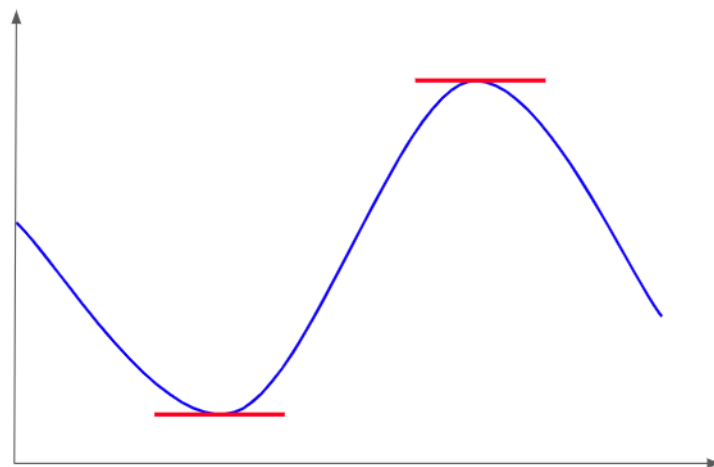


Figure 9: A graph with blue with tangent lines indicated at the local maximum and local minimum values.

That's right! The slope is equal to 0 at these points. So we can find at which values of x these local maximums and minimums occur by taking the derivative of our function and setting it equal to 0. This isn't a hard and fast rule: there are certain functions for which setting the derivative to 0 will not find a local maximum or minimum, but the functions optimized in machine learning that we care about in this course are problems that tend to be what is called 'convex'. You'll learn more about convexity in the videos, but for now it is enough to know that it means this derivative trick will work.

Which function do we differentiate? It depends on which technique we are using. But we use the following terminology to refer to the function we're interested in.

Definition 10 *The **objective function** is the function we wish to maximize or minimize.*

3 Linear Algebra

3.1 Vectors

One of the fundamental objects in linear algebra is vectors. A vector can be thought of as a list of numbers that describes some magnitude and direction in space. Each entry represents that dimension's position in space. For example, vector $[3, 4]$ lies on the xy -plane with x -coordinate equal to 3 and y -coordinate equal to 4, as in the figure below.

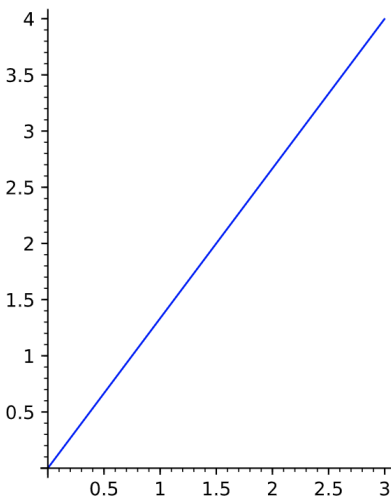


Figure 10: Vector $[3, 4]$ in blue.

The size of a vector is the number of elements in that vector. For example, vector $[3, 4]$, in Figure 10, is of size 2. An n -dimensional vector is of size n . The magnitude (or length) of a vector can be calculated using the Euclidean norm. Each component of a vector is squared, these are summed together, and then the square root of that sum is taken. For example, the Euclidean norm of $[3, 4]$ is given by $\|[3, 4]\| = \sqrt{3^2 + 4^2} = 5$. This makes sense with what we know about the length of this vector using the Pythagorean Theorem. The notation for the norm of a vector \mathbf{a} is given by $\|\mathbf{a}\|$ or, in certain contexts, $\|\mathbf{a}\|_2$.

Their beauty lies in the fact that they allow you to explore multi-dimensional space. As humans, we can only visualize in two and three dimensions (and even three is often difficult), but vectors and linear algebra help us understand the behaviour of mathematical objects in higher dimensions.

There are many operations you can do with vectors. We will review those relevant to this course. You can add two vectors of the same size together to get another vector. To do so, you add component-wise. You can think of this as adding together the journey along each dimension. For example, here we add together two vectors of size 4.

$$[1, 0, -2, 4] + [-2, 3, 0, 1] = [1 + (-2), 0 + 3, (-2) + 0, 4 + 1] = [-1, 3, -2, 5]$$

When it comes to vectors, there are a couple different notions of ‘multiplication’. The first we’ll discuss is scalar multiplication, which is when you multiply a vector by some number to get another vector of the same size (remember, size is the number of dimensions of the vector). To do scalar multiplication, you multiply the scalar value by each component in the vector. What happens,

both algebraically and geometrically, is that the vector scales by that scalar value (there is no direction change). For example,

$$3 * [1, -2] = [3 * 1, 3 * (-2)] = [3, -6]$$

Now that we've established vector addition and scalar multiplication, we have the tools to discuss linear dependence and linear independence. Anytime we have a vector that is some combination of vector addition and scalar multiplication, this is called a linear combination. For example,

$$[3, 4] + 3 * [2, -1] = [9, 1]$$

We say that $[9, 1]$ is a **linear combination** of vectors $[3, 4]$ and $[2, -1]$. If we have a set of vectors where at least one can be written as a linear combination of the others, this set of vectors is called **linearly dependent**. For example,

$$\{[3, 4], [2, -1], [9, 1]\}$$

However, when no vectors in a set can be written as a linear combination of the other vectors in that set, then that set of vectors is called **linearly independent**. For example, the sets

$$\{[3, 4], [2, -1]\} \text{ and } \{[1, 0, 0], [0, 1, 0], [0, 0, 1]\}$$

are both examples of linearly independent sets of vectors. The notion of linearly dependent and linearly independent will come up again when we discuss vector spaces later in this review.

Back to vector operations! The other type of vector multiplication relevant to this course is known as the dot product (or inner product). Unlike scalar multiplication which scales a vector to produce another vector, the dot product takes in two vectors of the same size to produce a scalar. Algebraically, we multiply component-wise and then add these together. For example,

$$[3, 4, 1] \cdot [-1, 0, 3] = (3)(-1) + (4)(0) + (1)(3) = -3 + 0 + 3 = 0$$

Notice that the vector norm can be written as the dot product of a vector, $\mathbf{a} = [a_1, a_2, \dots, a_n]$ with itself squared,

$$\|\mathbf{a}\|^2 = \left(\sqrt{a_1^2 + a_2^2 + \dots + a_n^2}\right)^2 = \mathbf{a} \cdot \mathbf{a}$$

Equivalently, the dot product can be calculated using

$$\mathbf{a} \cdot \mathbf{b} = \|\mathbf{a}\| * \|\mathbf{b}\| * \cos(\theta)$$

where \mathbf{a} and \mathbf{b} are vectors and θ is the angle between them. We can think about this geometrically. The dot product takes one vector and calculates how much that vector, or what portion of its magnitude, goes in the same direction of the other vector, and then multiplies these numbers.

One thing on notation: we have been writing vectors horizontally, but they are also sometimes written vertically. When we write vectors horizontally, we call them row vectors and when we write them vertically, we call them column vectors. As a matter of fact, we mostly use column vectors rather than row vectors, and sometimes we use just the word ‘vector’ when we mean ‘column vector’. If we want to change between row vectors and column vectors there is an operation to do so, known as the transpose. For example,

$$\begin{bmatrix} 1 & 0 & -3 \end{bmatrix}^T = \begin{bmatrix} 1 \\ 0 \\ -3 \end{bmatrix}$$

We can also represent the dot product of two vectors using transpose and vector multiplication. Say vector \mathbf{a} and \mathbf{b} are both column vectors of the same size. Their dot product, $\mathbf{a} \cdot \mathbf{b}$ can be written as $\mathbf{a}^T \mathbf{b}$.

3.2 Matrices

Another important object in linear algebra are matrices. A matrix is a list of numbers, arranged into rows and columns. The size of a matrix is given by the number of rows by the number of columns. For example,

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

is of size 2×3 . Sometimes the size of a matrix is written as a subscript on the symbol used for the matrix. For example, if matrix \mathbf{A} is of size $m \times n$, it might be written as $\mathbf{A}_{m \times n}$.

When it comes to operations, matrix addition and scalar multiplication follow the same rules as with vector addition and scalar multiplication. For example, below we have matrix addition

$$\begin{bmatrix} 1 & -2 & 3 \\ 4 & 0 & -1 \end{bmatrix} + \begin{bmatrix} 0 & 4 & 5 \\ -3 & -3 & 1 \end{bmatrix} = \begin{bmatrix} 1 & 2 & 8 \\ 1 & -3 & 0 \end{bmatrix}$$

and scalar multiplication

$$3 * \begin{bmatrix} 1 & -2 & 3 \\ 4 & 0 & -1 \end{bmatrix} = \begin{bmatrix} 3 & -6 & 9 \\ 12 & 0 & -3 \end{bmatrix}$$

Where things really get interesting is when we talk about matrix multiplication. This is because matrix multiplication can transform geometric objects in several different ways (known as linear transformations). Linear transformations can do things like stretch, squeeze, and rotate geometric spaces.

Let’s first see an example of matrix multiplication and discuss more details after.

$$\begin{bmatrix} 1 & -2 \\ 4 & 0 \end{bmatrix} \begin{bmatrix} 0 & 4 & 1 \\ -3 & -3 & 0 \end{bmatrix} = \begin{bmatrix} 1*0 + (-2)*(-3) & 1*4 + (-2)*(-3) & 1*1 + (-2)*0 \\ 4*0 + 0*(-3) & 4*4 + 0*(-3) & 4*1 + 0*0 \end{bmatrix} = \begin{bmatrix} 6 & 10 & 1 \\ 0 & 16 & 4 \end{bmatrix}$$

We will never ask you to do matrix multiplication by hand, but you will see matrices being multiplied, and it can be useful to know the inner-workings of what's going on. So, what is going on? Essentially, you can think about matrix multiplication as taking the dot product of each row of the first matrix with each column of the second.

Notice that the first matrix is of size 2×2 , the second is of size 2×3 , and the resulting is of size 2×3 . Matrix multiplication requires that the number of columns in the first matrix must equal the number of rows in the second. More formally, we have the following

$$\mathbf{A}_{j \times k} \mathbf{B}_{k \times h} = \mathbf{C}_{j \times h}$$

Notice that the resulting matrix will have the same number of rows as matrix \mathbf{A} and the same number of columns as matrix \mathbf{B} .

Another thing to note with matrix multiplication is that, unlike multiplying two numbers together, it is NOT necessarily true that $\mathbf{AB} = \mathbf{BA}$ for matrix multiplication.

Notice also that we can use matrix multiplication for vectors and the transpose operation in place of the dot product. So we have that $\mathbf{a} \cdot \mathbf{b} = \mathbf{a}^\top \mathbf{b}$. For example,

$$\begin{bmatrix} 1 \\ 0 \\ -3 \end{bmatrix} \cdot \begin{bmatrix} -1 \\ -3 \\ 5 \end{bmatrix} = \begin{bmatrix} 1 \\ 0 \\ -3 \end{bmatrix}^\top \cdot \begin{bmatrix} -1 \\ -3 \\ 5 \end{bmatrix} = [1 \quad 0 \quad -3] \begin{bmatrix} -1 \\ -3 \\ 5 \end{bmatrix} = -16$$

We can also use matrices to map from n -dimensional space to m -dimensional space. This is called **projection**. We will show an example of this below where we go from 3 dimensions to 2 dimensions, but the idea works for any sized dimensions, even if they're not sequential.

Let's say we have the below vector in 3-dimensional space (indicated by x , y , and z) and that we decide that dimension y is not all that useful to us but we want to keep all the information in dimensions x and z .

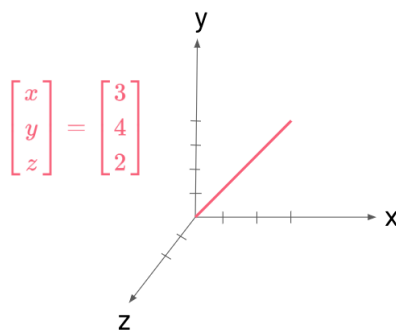


Figure 11: Vector in three dimensions.

What we will do amounts to collapsing our vector down to just dimensions x and z . You can think about this as shining a flashlight down the y -axis towards the origin. This will cast a shadow of our vector onto the xz -plane that maintains the appropriate x and z values. Algebraically we will multiply our vector by the blue matrix in the below image, to retrieve our desired projected vector.

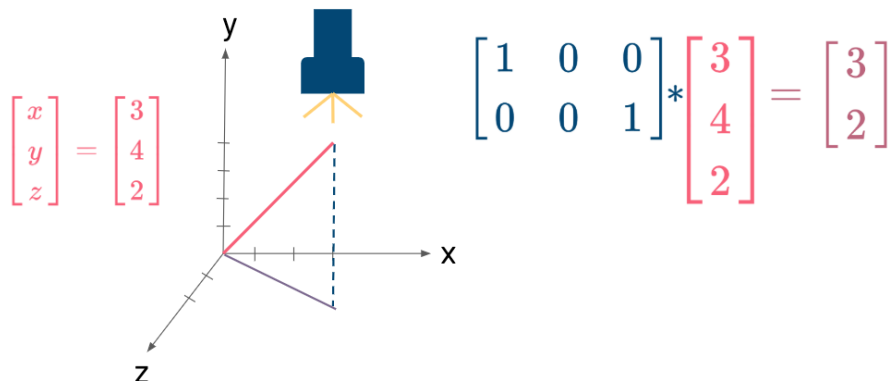


Figure 12: Vector projected onto xz -plane.

3.3 Vector Spaces

The span of a matrix is the ‘vector subspace’ that a set of vectors can cover. This means that any vector in that ‘vector subspace’ can be represented by that matrix multiplied by some other vector. We haven’t formally defined vector spaces or vector subspaces, but we’re going to look at a couple of examples in this review.

The identity matrix in 3-dimensions,

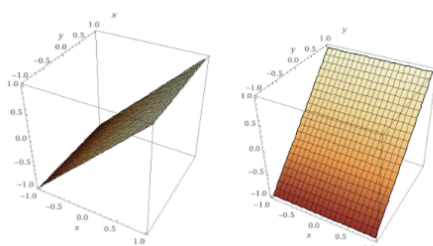
$$\mathbf{I} = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}$$

spans the entirety of 3-dimensional space. This means that we can represent any arbitrary vector in 3-dimensional space by multiplying this matrix by some vector.

Let’s go through an example of a matrix that is more restricted. The largest vector space that this matrix can span is the entirety of 3-dimensional space (because it is size 3×3),

$$\mathbf{A} = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 2 & 1 \\ 0 & 4 & 2 \end{bmatrix}$$

If we multiply this matrix by some arbitrary 3-dimensional vector, we can see what vector space it spans. Below in Figure 13 we've done so, and factored to get a linear combination of the terms.

$$\begin{bmatrix} 1 & 0 & 0 \\ 0 & 2 & 1 \\ 0 & 4 & 2 \end{bmatrix} * \begin{bmatrix} a \\ b \\ c \end{bmatrix} = \begin{bmatrix} a \\ 2b + c \\ 4b + 2c \end{bmatrix} = \begin{bmatrix} a \\ 0 \\ 0 \end{bmatrix} + \begin{bmatrix} 0 \\ 2b + c \\ 2(2b + c) \end{bmatrix}$$


$$= a \begin{bmatrix} 1 \\ 0 \\ 0 \end{bmatrix} + (2b + c) \begin{bmatrix} 0 \\ 1 \\ 2 \end{bmatrix}$$

Figure 13: Multiplying matrix \mathbf{A} by an arbitrary vector to determine what vector subspace matrix \mathbf{A} spans.

But what does this linear combination boxed in yellow say? Well, since a , b , and c are all arbitrary, they can be any number. And $2b + c$ can be any number for that matter. With the first term, we see that we can span the entire first dimension, and with the second term we can span the entire second dimension. But what about the third dimension? We see in the linear combination that the third dimension will always be equal to twice the second dimension. So, instead of spanning the entirety of 3-dimensional space, we get that our matrix \mathbf{A} spans a tilted plane where the third dimension is always twice the second.

We are not going into the formal conditions required by vector spaces and vector subspaces. For this course it is enough to think about them as the space in which vectors live. For example, $[1, -1]$ lives in the 2-dimensional vector space.

3.4 Eigenvalues and Eigenvectors

Eigenvalues and eigenvectors have applications all over the place: physics, geology, other areas of math (like spectral graph theory). We'll look further at the algebra of eigenvalues and eigenvectors, but the basic geometric idea is that eigenvalues tell you something about the way matrices stretch and squeeze geometric space in a particular direction. This direction is given by the eigenvalue's corresponding eigenvector. In fact, if you want to stretch or squeeze geometric space in a specific way, you can construct a matrix with particular eigenvalues and eigenvectors to do so.

Algebraically, we use the notation $\mathbf{A}\mathbf{x} = \lambda\mathbf{x}$, where \mathbf{A} is some matrix, λ is an eigenvalue, and \mathbf{x} is the corresponding eigenvector. Notice what is happening here: matrix \mathbf{A} scales vector \mathbf{x} by a factor of λ , it does not change the direction of vector \mathbf{x} .

We don't need to worry too much about the algebra of eigenvalues and eigenvectors for this course—the main takeaway is what they tell you about the way a matrix stretches and squeezes space. Let's look at one example.

Say we have matrix

$$\mathbf{A} = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 2 & 1 \\ 0 & 4 & 2 \end{bmatrix}$$

Matrix \mathbf{A} has three eigenvalues and eigenvectors, highlighted in yellow in Figure 14. We can see a surprising amount about the nature of our matrix through its eigenvalues and eigenvectors.

For example, the third eigenvalue is equal to 0. This tells us that our matrix has some linearly dependent rows (the second and third row). It also tells us that the matrix does not span the entirety of 3-dimensional space. And, actually, this should not come as a surprise to us, because this is the same matrix we examined before when we talked about span. Each eigenvector is drawn with a pink arrow in 3-dimensional space in Figure 14.

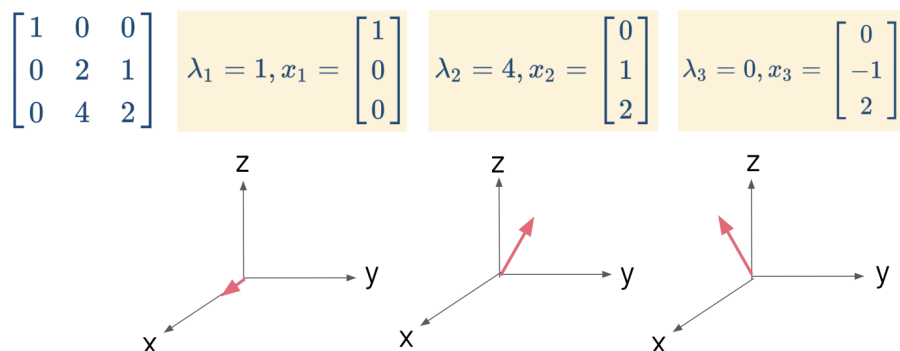


Figure 14: Highlighting how the eigenvalues and eigenvectors of a matrix stretch 3-dimensional space for matrix \mathbf{A} with the eigenvectors drawn below.

To visualize how this matrix stretches and squeezes space using eigenvectors, we can plot a unit sphere on our graphs and see the impact that each eigenvector has on it. This is shown in Figure 15. The first eigenvalue and eigenvector will not distort the sphere at all. The second will stretch the sphere by a factor of 4 in the direction of the second pink eigenvector. And finally, the third will actually flatten the sphere into a disk along the direction of the third eigenvector.

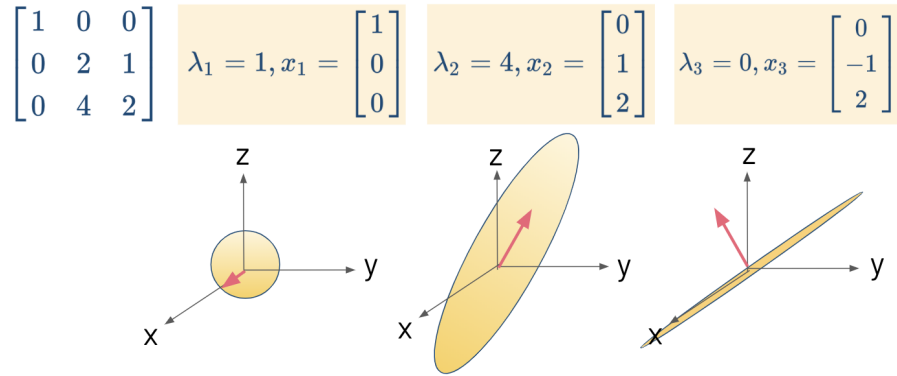


Figure 15: Highlighting how the eigenvalues and eigenvectors of a matrix stretch 3-dimensional space for matrix \mathbf{A} .

4 Graph Theory

The fundamental object in Graph Theory is graphs. Graphs can be used to model pairwise relationships between objects. They are defined by a vertex set and an edge set, where the vertices represent the object and the edges represent some relationship between these objects.

These edges can have a direction associated with them in which case they are called ‘directed’, but they don’t have to. When there is no associated direction, they are called ‘undirected’. Also, vertices are sometimes referred to as ‘nodes’; we will use both terms in this course.

For example, we may represent our vertex set as $V = \{0, 1, 2, 3\}$ and edge set as $E = \{\{0, 1\}, \{1, 2\}, \{2, 3\}, \{3, 0\}, \{0, 2\}, \{1, 2\}\}$. This can be written as graph, $G = (V, E)$ and drawn as follows.

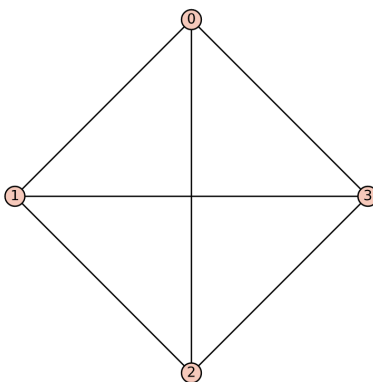


Figure 16: Graph G (also known as the complete graph on 4 vertices).

It wouldn't count as a Graph Theory section if we didn't at least mention the Petersen graph which is a famous graph defined on 10 vertices, seen below.

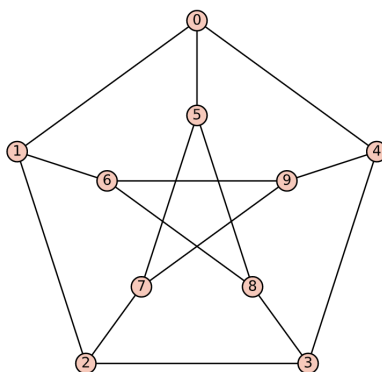


Figure 17: The Petersen Graph.

We will not dive too deeply into the wonderful world of Graph Theory, but there are a couple definitions you will need to know for this course.

Definition 11 A *path* is a sequence of edges connecting vertices in a graph such that no vertex is ever repeated.

For example, we have the below in blue which is a path from vertex 0 to vertex 5 in the Petersen Graph.

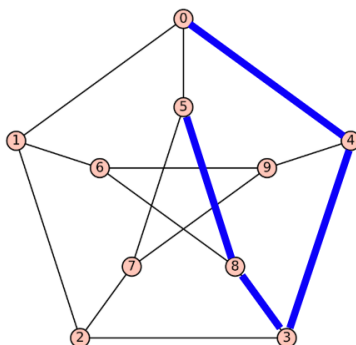


Figure 18: A path from vertex 0 to vertex 5 in the Petersen Graph.

Notice that this isn't the shortest path from vertex 0 to vertex 5 (we could have just taken the edge connecting vertex 0 to vertex 5), but it is nonetheless a path connecting those two vertices.

One very important type of graph that we will use in this course is known as a tree. A tree is an undirected graph in which any two vertices are connected by exactly one path. For example,

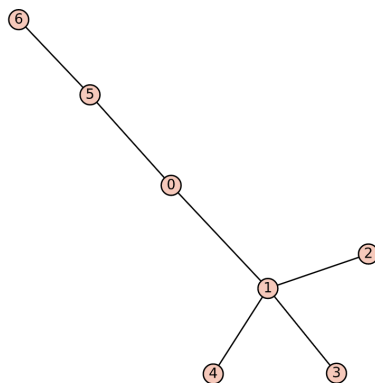


Figure 19: An example of a tree on 7 vertices.

Although it is not always the case in general, we will use rooted trees which amounts to a hierarchy in our tree. For example, we could draw the above graph in the following way, rooted at node 0.

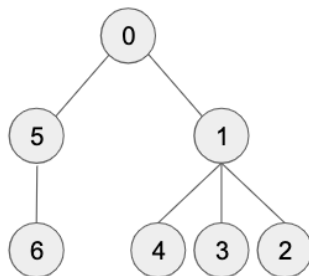


Figure 20: An example of a tree on 7 vertices rooted at node 0.

When trees are drawn as above, we call vertices 5 and 1 the ‘children’ of vertex 0. Similarly, vertices 4, 3, and 2 are children of vertex 1, and vertex 0 is a parent of vertex 1.

For the most part in this course we will use binary trees for which each node has at most two children. Below is an example of a binary tree.

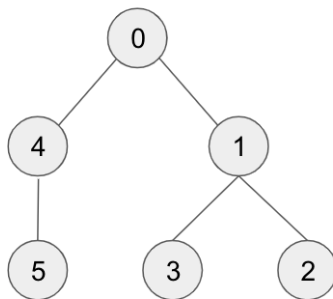


Figure 21: An example of a binary tree on 6 vertices.

5 Statistics

When working with statistics we usually base calculations on a sample of the data, not the whole population itself, so we need to allow for some variation between these sample statistics and the true parameters of the full population.

Sometimes we might want a single number to represent a whole population. Often the number used to do this is the mean, or average. This is also sometimes called the expectation of a set of data. The **mean** of a dataset sums all entries in that dataset and divides it by the size of the dataset. The symbol \bar{x} is used for the mean of a sample of the data and the symbol μ is used for the mean of the entire population of the data. This is likely to be familiar to you. For

example, say I eat 13 ice creams a day and my two best friends each only eat 1 ice cream per day. Then the mean number of ice creams eaten per day in this sample of three people is $\frac{13+1+1}{3} = 5$.

But just knowing the mean of your data tells you nothing about the spread of the data. Even in the example above, I eat so many more ice creams than my two friends! So, even though the mean number of ice creams is 5 per day, we can see that my large number of ice creams per day is pulling the mean to a much larger value. A value that gives a measure for the spread of the data is the **standard deviation**. The symbol s is used for the standard deviation of a sample of the data and the symbol σ is used for the standard deviation of the entire population of the data. Standard deviation can be calculated using the following equation

$$s = \sqrt{\frac{1}{N-1} \sum_{i=1}^N (x_i - \bar{x})^2}$$

where N is the number of examples in our data, x_i is the value of each i th example, and \bar{x} is the mean of the sample of the data.

A small standard deviation means that the datapoints tend to be close to the mean, whereas a large standard deviation shows that the datapoints are more spread out. For example, the standard deviation for our ice cream example above is approximately 9.38 which indicates that our data is fairly spread out.

Now let's do a coin-toss experiment together: we want to know what proportion of the time our coin will land on heads. Say we do ten coin flips and find that 6 of our flips were heads. So the proportion of heads in this sample is 6 out of 10.

So at this point our experiment tells us that, for any number of coin flips, we should expect around 60% of them to be heads. But does this seem right? Should we really trust a sample of just ten coin flips?

The answer is no, of course. So what can we do to get more accurate results? Well, we can repeat the sample multiple times, likely getting different proportions each time. And then we can plot a histogram with the proportion of heads on the x-axis and the number of samples that achieve that proportion on the y-axis. This is a sampling distribution. With enough samples, our sampling distribution will start to look like (or converge to) the normal distribution, see Figure 22. This is what is known as the **Central Limit Theorem**.

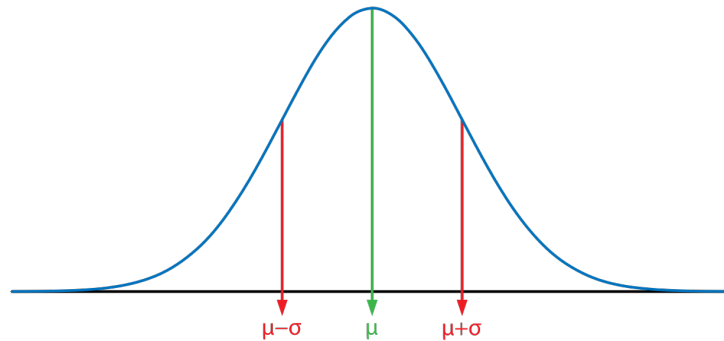


Figure 22: The graph of a normal distribution. [Public domain]

Related to the Central Limit Theorem, but not exactly the same, is the **Law of Large Numbers**. This says that, as a sample size tends to infinity, the sample mean will tend toward the population mean. This makes sense with our coin-flipping example: with a large enough number of flips, the proportion of head-flips that you see will be one-half.

Recall that we talked about how, when working with statistics, we have to account for the fact that we're working with a sample of the data and not the whole population. The beauty of the Law of Large Numbers is that it tells us that with a large enough sample size there will be no variation in the sample mean and the population mean.

You might be wondering what is the difference between the Central Limit Theorem and the Law of Large Numbers. The difference is that the Central Limit Theorem tells us about the shape of the distribution, whereas the Law of Large Numbers tells us where the “center” (or maximum point) of the “bell shape” is located (and that, as the sample size approaches infinity, the center of the distribution of the sample means tends toward the population mean).

Probability distributions are maps that let you know how likely all possible events are. Furthermore, random events can be dependent or independent. An independent event is the roll of the dice. What came up before does not impact what comes up next. Truly.

For a dependent event, the event is influenced by what has happened before. Dealing cards is dependent, because the cards that have already been dealt change what remains in the deck.

In machine learning we talk about having two kinds of data: independent, identically distributed, and everything else.

Independent, identically distributed data is data where each specific example is not dependent on every other. You can shuffle the order and it doesn't change anything. The example you're currently looking at doesn't change the probability of anything else. Not only that, but the distribution that the examples are coming from is the same. Sure, they are different examples, but they

are all coming from the same underlying distribution.

This is such an important concept, especially for proving things about machine learning algorithms, that we shorten ‘independent, identically distributed’ to iid. Technically, iid means each random event is independently drawn from the same distribution. Think about classifying buildings as apartments or houses based on number of rooms and price. There’s some underlying principle, we hope, that relates the categorization to number of rooms and price. But it doesn’t matter which example we look at when.

Everything else is non-iid data—data which has interdependencies. Either the probability of one label is influenced by the label assigned to another or the examples comes from different distributions. As a side note, this is the underlying cause of the batch effect, which you might remember from course one.

Take, for example, the signals from an outdoor water sensor. If in a sea of dry readings there’s a microsecond of rain, you might suspect that’s a glitch, because you expect there to be some consistency within short time frames. Shuffling the order of the readings would mean you lose that information.

As another example, imagine text signals—the predictive text completion on your phone uses what you’ve already written to guess what’s coming next. The signals, and appropriate predictions, are influenced by what’s gone immediately before.