MATHS OVERVIEW

Key references

1. V. I. Smirnov. A course of higher mathematics: volume 1
2. N. S. Piskunov. Differential and Integral Calculus: volume 1
3. N. S. Piskunov. Differential and Integral Calculus: volume 2
4. V. E. Gmurman. Fundamentals of Probability Theory and Mathematical Statistics
5. V. A. Ilyin, E. G. Poznyak. Linear Algebra

I. Probability

a) **Defining probability**

Naïve definition:

$$P(A) = \frac{\text{№}\ of\ outcomes\ within\ a\ subset\ (A)}{\text{№}\ of\ outcomes\ within\ a\ set\ (S)}$$

Sample space (S) – a set of all possible outcomes of an experiment.
An event (A) – a subset of the sample space. Thus, $A \subset B$, $x \in A$, $A = \{x_1, x_2, \dots, x_n\}$.

Assumptions:

- All outcomes are equally likely;
- There are finitely many outcomes (finite sample space).

Non-naïve definition:

Probability space – entity that consists of S (sample space) and P (a function that takes an event A as input and gives the output between 0 and 1: $P(A) \in [0 \dots 1]$) and obeys two assumptions:

- $P(\emptyset) = 0, P(S) = 1$: the probability of the empty set is 0 and the probability of the full space is 1;
- $P(\bigcup_{n=1}^{\infty} A_n) = \sum_{n=1}^{\infty} P(A_n)$ if $A_1, A_2, \dots A_n$ are disjoint (these subsets do not overlap).

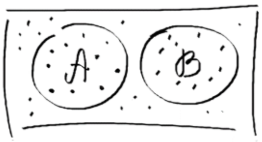
Also $P(A)$ is the probability of any one of the outcomes in A happening.

Additional info:

- What is the probability of the sample space?
- What is probability by Penn State?

b) **Disjoint and non-disjoint events**

Disjoint events – events that cannot happen at the same time, i.e. mutually exclusive $P(A \cap B) = 0$.



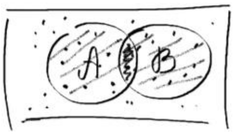
$P(A \cup B)$ is the probability that either the event A happens or the event B happens or both (but in this case it's impossible):

$$P(A \cup B) = P(A) + P(B)$$

$$P(A \cup B) = \frac{m_1 + m_2}{n} = \frac{m_1}{n} + \frac{m_2}{n} = P(A) + P(B)$$

EXAMPLE: when tossing a coin, we cannot get heads and tails simultaneously.

Non-disjoint events – events that can occur simultaneously $P(A \cap B) \neq 0$.



$$P(A \cup B) = P(A) + P(B) - P(A \cap B)$$

$$P(A) = P(A \cap B) + P(A \cap \bar{B})$$
$$P(B) = P(A \cap B) + P(\bar{A} \cap B)$$

$$P(A \cup B) = \overbrace{P(A \cap B) + P(A \cap \bar{B})}^{P(A)} + \overbrace{P(A \cap B) + P(\bar{A} \cap B)}^{P(B)} - P(A \cap B)$$

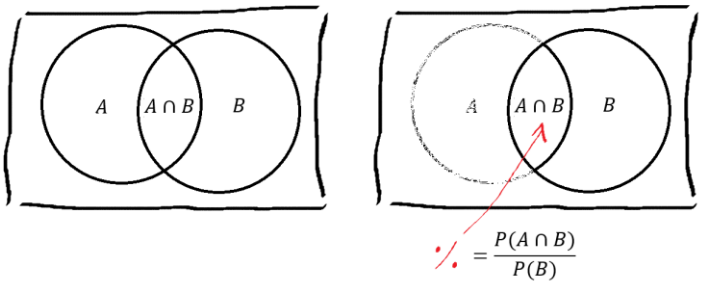
EXAMPLE: a chess piece can be white and a bishop at the same time.

c) **Dependent and independent events. Conditional probability. The product rule of probability**

Dependent events affect each other, i.e. after an even A happens, the probability of an event B changes.

Conditional probability – if an even A happened, what is the probability of an event B happening?

$$P(A|B) = \frac{P(A \cap B)}{P(B)}$$



EXAMPLE: once you take one card from the deck there are simply less cards that you can pick from. As a result, the probability of taking a specific card (or a card of a specific rank) after you've already taken one changes.

The product rule of probability for dependent events – we can derive it from the formula of conditional probability

$$P(A \cap B) = P(B)P(A|B)$$
$$P(B \cap A) = P(A)P(B|A)$$
$$P(A_1 \cap A_2 \cap \ldots \cap A_n) = P(A_1)P(A_2|A_1)P(A_3|A_1 \cap A_2) \ldots P(A_n|A_1 \cap A_2 \ldots \cap \ldots A_{n-1})$$

The product rule of probability for independent events – if events are independent, $P(A|B) = P(A)$ and $P(B|A) = P(B)$. Thus:

$$P(A \cap B) = P(A)P(B)$$

d) **Some notes on exclusive and dependent events**

- If events are disjoint, they are dependent because if A happens, the probability of B happening is 0;
- If events are non-disjoint, they can be either dependent or independent.

e) **Bayes' theorem**

- Let's say we are studying an event A – the probability of passing A VERY DIFFICULT maths exam;
- In addition to that, we explore this event A in conjunction with certain conditions B_1 and B_2:

    - if a student is a historian, what is the probability of passing the exam $P(A|B_1)$;
    - if a student is a mathematician, what is the probability of passing the exam $P(A|B_2)$.

- Probabilities $P(A|B_1)$ and $P(A|B_2)$ we usually know. For instance, some firm analysed how students who specialise in different subjects pass their maths exams;
- Our task is to calculate the following probability: $P(B_1|A)$ if a student passed the exam, what is the probability that they are a historian?

Why it is important to go from something we already know $P(A|B_2)$ [if a student is a historian, what is the probability of passing the exam] to this “unknown” probability $P(B_1|A)$ [if a student passed the exam, what is the probability that they are a historian]?

At first glance, $P(B_1|A)$ should be low since if a student passed this hard maths exam, they are more likely to be a mathematician. So, we may expect $P(B_1|A) < P(B_2|A)$. Especially, if given $P(A|B_1)$ and $P(A|B_2)$ are, for example, 0.1 and 0.9 respectively.

However, if, for whatever reason, we have 10,000 historians and 10 mathematicians, then $P(B_1|A)$ will be larger than $P(B_2|A)$ simply because of the sheer number of historians. So, when we contextualise $P(A|B_1)$ and $P(A|B_2)$, taking into account the conditions of our experiment, in our case it is the number of students in each stream, the picture can change drastically.

$$P(B_i|A)=\frac{P(B_i)P(A|B_i)}{\sum_{i=1}^{n}P(B_i)P(A|B_i)}=\frac{P(B_i)P(A|B_i)}{P(A)}$$

$$\underbrace{\left.\begin{array}{r} P(A\cap B_i)=P(B_i)P(A|B_i)\\ P(B_i\cap A)=P(A)P(B_i|A)\end{array}\right\} P(A\cap B_i)=P(B_i\cap A)}$$

$$\downarrow$$

$$P(B_i)P(A|B_i)=P(A)P(B_i|A)$$

$$P(B_i|A)=\frac{P(B_i)P(A|B_i)}{P(A)}$$

$$P(B_i|A)=\frac{P(B_i)P(A|B_i)}{P(A)}=\frac{P(B_i)P(A|B_i)}{P(B_1)P(A|B_1)+\cdots+P(B_i)P(A|B_i)}=\frac{P(B_i)P(A|B_i)}{\sum_{i=1}^{n}P(B_i)P(A|B_i)}$$

Additional info:

- Visual explanation of Bayes’ theorem

f) **Random variable. Probability distribution. PDF, CDF**

RV – a variable whose values are determined by a probabilistic (random) experiment. Suppose we flip a fair coin 100 times. We can define a RV X as the number of heads that we can get. Possible realizations of the random variable X are $x \in \{0, 1, \dots, 100\}$.

Discrete RV – RV is discrete if it takes a countable number of distinct values.

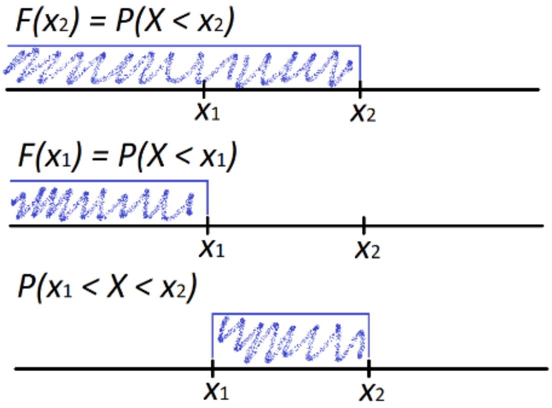
Probability distribution of a discrete RV – an array of probabilities associated with each outcome / value of a RV $P(X = x)$. For example, if X is the number of heads we can get after flipping a coin many times, then $P(X = 5)$ is the probability of getting exactly 5 heads.

Continuous RV – RV is continuous if it takes an infinite number of distinct values. For example, the hight of a person is a continuous RV since it could be any non-negative value.

Cumulative distribution function of a continuous RV (CDF) – the probability that X will take a value less than or equal to x: $F(x) = P(X < x)$

Some important properties:

- Since $F(x)$ is a probability, by definition we have $0 \le F(x) \le 1$;
- $P(a \le X < b) = F(b) - F(a)$;



$$P(X < x_2) = P(X < x_1) + P(x_1 \le X < x_2)$$
$$P(X < x_2) - P(X < x_1) = P(x_1 \le X < x_2)$$
$$F(x_2) - F(x_1) = P(x_1 \le X < x_2)$$

We can also explore this property from a different angle:

$$\int_{x_1}^{x_2} F'(x)dx = F(x_2) - F(x_1) = P(X < x_2) - P(X < x_1) = P(x_1 \le X < x_2)$$

- $F(x) = P(X < x) = \int_{-\infty}^{x} f(x)dx$, PDF will be described later.

Probability density function of a continuous RV (PDF) – the first derivative of CDF: $F'(x) = f(x)$.

By definition, the derivative of CDF equals to:

$$F'(x) = f(x) = \lim_{\Delta x \to 0} \frac{\overbrace{F(x + \Delta x) - F(x)}^{P(x < X < x + \Delta x)}}{\Delta x}$$

As was shown earlier $F(x + \Delta x) - F(x) = P(x < X < x + \Delta x)$. We divide the probability of being within the interval $(x; x + \Delta x)$ by the length of this interval Δx. Thus, we get the "average" probability or probability density in other words.

Additional info:

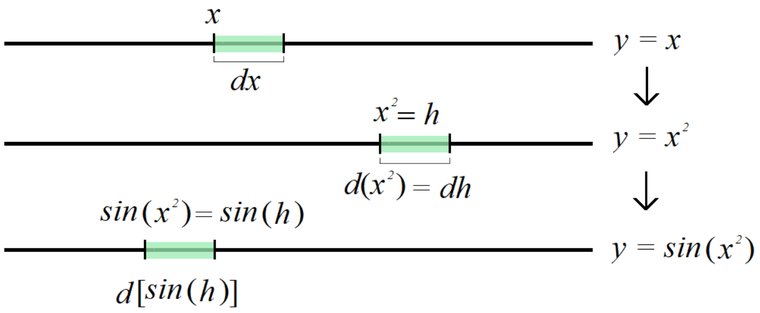
- Introduction to Probability, Statistics, and Random Processes
- Random variables by Yale.edu

II. Derivatives

a) **Derivative of a composite function**

If we have the following function $y = f(z)$, where $z = g(x)$, then:

$$y' = f'(z)g'(x) = \frac{dy}{dz}\frac{dz}{dx}$$



$$d[sin(h)] = cos(h)dh$$
$$d[sin(x^2)] = cos(x^2)d(x^2) = cos(x^2)\,2xdx$$

EXAMPLE: $y = (3x + 2)^2$. Let's replace the inner function with $z = 3x + 2$, we get $y = (z)^2$. Thus, $\frac{dy}{dx} = \frac{d(z)^2}{dz}\frac{dz}{dx} = \frac{d(z)^2}{dz}\frac{d(3x+2)}{dx} = 2z \cdot 3 = 6(3x + 2)$.

b) **Partial derivative**

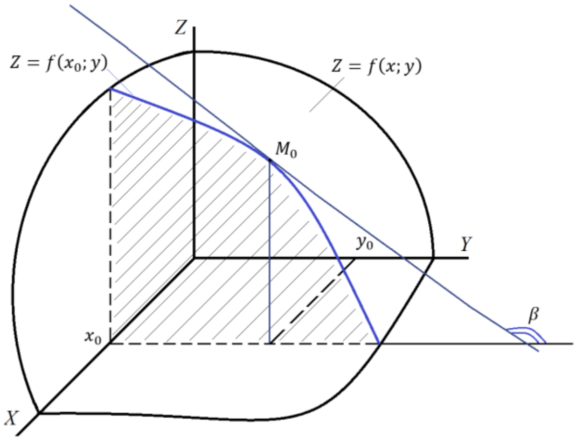
If we have a function of 2 independent variables $z = f(x, y)$, then a partial derivative with respect to x can be defined as follows:

$$f_x'(x, y) = \frac{\partial z}{\partial x} = \lim_{\Delta x \to 0} \frac{\Delta z_x}{\Delta x} = \lim_{\Delta x \to 0} \frac{f(x + \Delta x, y) - f(x, y)}{\Delta x}$$

A partial derivative with respect to x is the rate of change of a function f along the x-axis. All other variables are held constant, and, as a result, our function f depends only on x.

EXAMPLE: $z = 3x + 2y$. So, $\frac{\partial z}{\partial x} = (3x)' + (2y)'$, since y is a constant $\frac{\partial z}{\partial x} = 3 + 0 = 3$.

Suppose we have a function $z = f(x, y)$. Let's pick a specific value $x = x_0$ and illustrate the function $z = f(x_0, y)$:

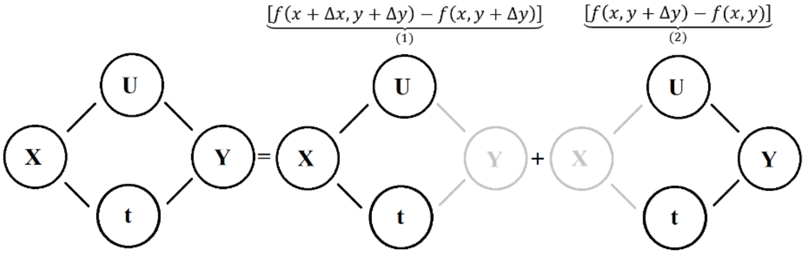


It can be seen that $z = f(x_0, y)$ depends only on y, and we can now calculate the derivative with respect to y as we normally would.

c) **Derivative of a composite function. Multivariate case**

Suppose we have a function $u = f(x, y)$ and variables x and y depend on t. Then, the derivative with respect to t is equal to:

$$\frac{du}{dt} = \frac{\partial u}{\partial x}\frac{dx}{dt} + \frac{\partial u}{\partial y}\frac{dy}{dt}$$


$\underbrace{[f(x+\Delta x, y+\Delta y) - f(x, y+\Delta y)]}_{(1)}$
$\underbrace{[f(x, y+\Delta y) - f(x, y)]}_{(2)}$
U
X
Y
t
=
U
X
Y
t
+
U
X
Y
t

Formula is derived based on the Lagrange's theorem [a].

e) **Directional derivative**

If we want to explore the rate of change of a function in any given direction rather than along a specific axis (x-axis, for example), we will need to introduce the concept of a **directional derivative** which is built on the idea of a **partial derivatives**. If we have a function of 2 independent variables $z = f(x, y)$, then the directional derivative along the direction $\bar{u}$ is:

$$\nabla_{\bar{u}} f(x, y) = \frac{\partial z}{\partial x} u_1 + \frac{\partial z}{\partial y} u_2$$

In this case, $\bar{u} = (u_1, u_2)$ is a unit vector $|\bar{u}| = 1$ that determines the direction of our derivative.

We choose a direction $\bar{u} = (u_1, u_2)$, where $|\bar{u}| = 1$ (a unit vector) because we only use it to determine the direction in which we are going []. We start at some point $M(x_0, y_0)$ and go in the direction of $\bar{u}$. As a result, we can parametrise independent variables:

$$x(h) = x_0 + hu_1;\ y(h) = y_0 + hu_2$$

By changing h we move from the point $M(x_0, y_0)$ along the $\bar{u}$ direction. Thus, just like with ordinary derivatives, we can use the limit definition:

$$\nabla_{\bar{u}} f(x, y) = \lim_{h \to 0} \frac{f(x + hu_1, y + hu_2) - f(x, y)}{h}$$

However, in order to compute this derivative, a little more work is needed. We know that $z = f(x, y)$ depends on x and y that, in turn, depend on h. Consequently, $z = f(x, y)$ depends on h and is a composite function. The derivative of a composite function of multiple variables (2 in our case) is equal to:

$$\frac{dz}{dh} = \frac{\partial z}{\partial x}\frac{dx}{dh} + \frac{\partial z}{\partial y}\frac{dy}{dh}$$

$$\frac{dx}{dh} = (x_0 + hu_1)' = u_1;\ \frac{dy}{dh} = (y_0 + hu_2)' = u_2$$

$$\frac{dz}{dh} = \nabla_{\bar{u}} f(x, y) = \frac{\partial z}{\partial x} u_1 + \frac{\partial z}{\partial y} u_2$$

Additional info:

- Wiki: what is directional derivative?
- Video lecture by Dr. Trefor Bazett
- Video lecture by Prof. Leonard

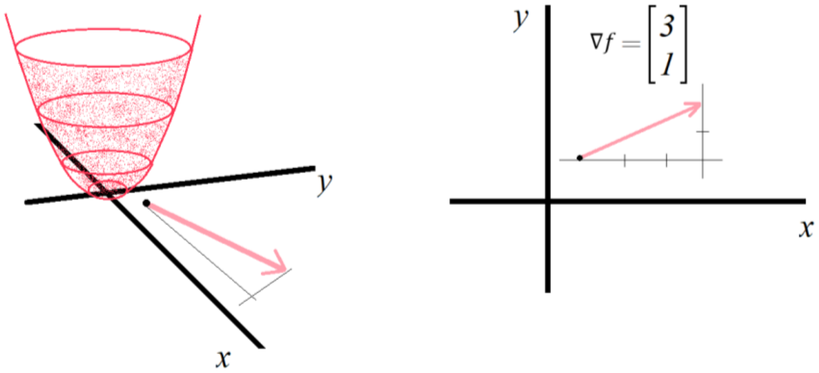
f) **Gradient**

It is a vector that shows the direction of the fastest increase:

$$\nabla f = \nabla f(w_1, w_2, \dots w_n) = \left[\frac{\partial f}{\partial w_1}, \frac{\partial f}{\partial w_2}, \dots, \frac{\partial f}{\partial w_n}\right]^T$$

For instance, $\nabla f(1, 1, \dots, 1) = [0.3, 2, \dots, -0.5]^T$:

- $\frac{\partial f}{\partial w_2}$ is more important than both $\frac{\partial f}{\partial w_1}$ and $\frac{\partial f}{\partial w_n}$, while $\frac{\partial f}{\partial w_n}$ is more important than $\frac{\partial f}{\partial w_1}$;
- Standing at the input $(1, 1, \dots, 1)$ and moving along this direction ∇f, increases the function $f(\dots)$ most quickly, but, on top of that, changes to the variable w_2 is more important than changes to the variable w_1 $(2 > 0.3)$, at least in the neighbourhood of the input;
- Calculating a directional derivative with respect to the direction $\bar{u}$, we weigh $\frac{\partial f}{\partial w_1}, \frac{\partial f}{\partial w_2}, \dots, \frac{\partial f}{\partial w_n}$ based on $u_1, u_2, \dots, u_n$. If, however, $\bar{u}$ points in the direction of ∇f, we weigh $\frac{\partial f}{\partial w_1}, \frac{\partial f}{\partial w_2}, \dots, \frac{\partial f}{\partial w_n}$ in the "best" way possible, that is we give them weights according to their contributions to the rate of change of $f(\dots)$ – the larger the contribution, the greater the weight:



Also, if $\bar{u} = \nabla f$, a directional derivative will be equal to the magnitude of ∇f:

$$\nabla_{\bar{u}} f(x, y) = |\nabla f|$$

Earlier, we defined a directional derivative in the following way:

$$\nabla_{\bar{u}} f(x, y) = \frac{\partial z}{\partial x} u_1 + \frac{\partial z}{\partial y} u_2$$

And $\bar{u} = (u_1, u_2)$ is a unit vector pointing in any given direction. Since $\nabla_{\bar{u}} f(x, y)$ is a dot product of two vectors (u_1, u_2) and $\nabla f = \left(\frac{\partial f}{\partial w_1}, \frac{\partial f}{\partial w_2}\right)$, so we can rewrite it as follows:

$$\nabla_{\bar{u}} f(x, y) = |\nabla f| \cdot |\bar{u}| \cdot \cos \alpha = |\nabla f| \cdot \cos \alpha$$

So, what will happen if we try maximising this quantity? The maximum value of $\cos \alpha = 1$, which is the case when $\alpha = 0$, in other words, when the angle between ∇f and $\bar{u}$ is 0 – they point in the same direction. Thus, $\nabla_{\bar{u}} f(x, y)$ is maximised when $\bar{u} = \nabla f$.

III. Integrals (1 variable)

a) **Antiderivative (primitive)** – a function F is an antiderivative of a function f if $F'(x) = f(x)$.

EXAMPLE: $f(x) = 2x$. What is the antiderivative $(?)' = 2x$? $F(x) = x^2$.

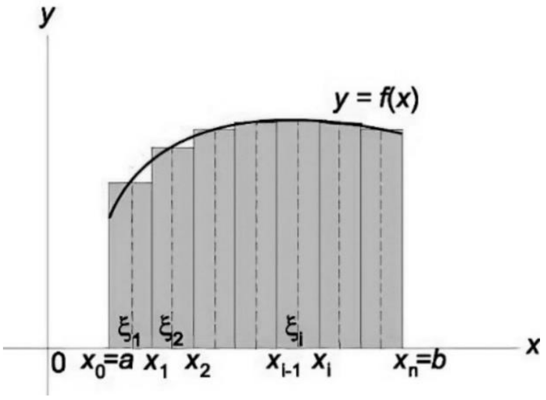
b) **Indefinite integral** – the set of all antiderivatives of f: $F(x) + C$, where C is an arbitrary constant. This definition is not the only one. Later, when explaining the connection between a definite and indefinite integral, it will be shown that an integral with the variable upper limit is also an antiderivative (primitive).

$$\int f(x)dx = F(x) + C$$

Why do we add a constant C?

$$[F(x) + C]' = F'(x) + 0 = f(x)$$

c) **Definite integral** – informally, it is the sum of infinitely many (infinitely small) rectangles or stripes. This sum represents the area under a curve.



Given a function $f(x)$ that is continuous on the interval $[a, b]$ we divide the interval into n subintervals of the width, Δx, and within each interval choose a point ξ_i:

$$\int_a^b f(x)dx = \lim_{\lambda \to 0} \sum_{i=0}^{n-1} f(\xi_i)\Delta x_i$$

Additional info:

- Indefinite integrals by Utexas.edu
- The connection between definite and indefinite integrals
- Definite integral by Utexas.edu

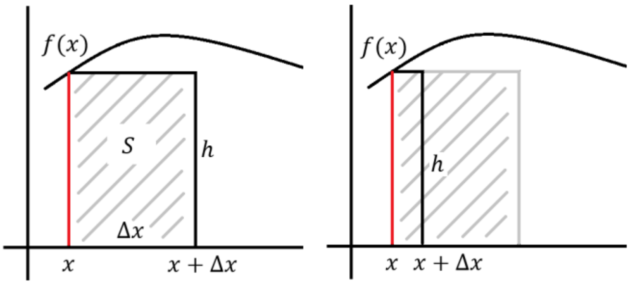
d) **Newton-Leibniz theorem. Fundamental theorem of calculous** – theorem that links a definite integral to antiderivatives:

$$\int_a^b f(t)dt = F(b) - F(a)$$

An integral with a variable upper limit, can be seen as a function of x – an area under the curve that changes based on x:

$$F(x) = S_{a,x}(x) = \int_a^x f(t)dt$$

When the area of a rectangle is divided by its base, we get the height. In case of $\frac{\Delta S_{a,x}}{\Delta x}$ we will get the height h of a rectangle, and this height will get closer and closer to $f(x)$ as $\Delta x \to \infty$:



$$f(x) = \lim_{\Delta x \to 0} \frac{\Delta S_{a,x}}{\Delta x};\ S_{a,x}{}'(x) = F'(x) = f(x)$$

Consequently, an integral with a variable upper limit is one of the possible antiderivatives (primitives) of $f(x)$. We got one $F(x)$, but we are interested in a set $F(x) + C$. Suppose $\Phi(x)$ is any given antiderivative from this set:

$$\Phi(x) = F(x) + C;\ \Phi(x) = \int_a^x f(t)dt + C$$

Let’s consider the following edge cases:

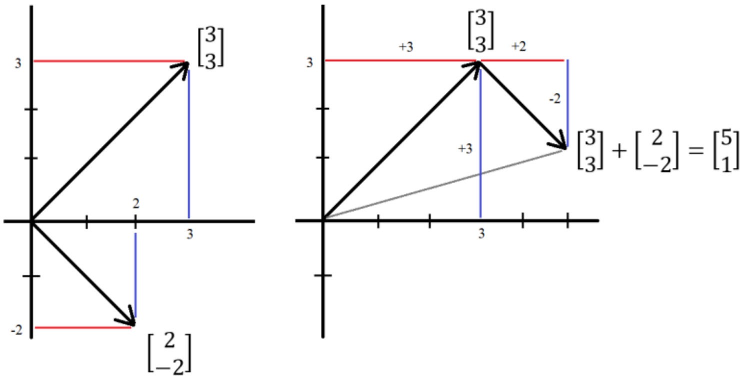
$$\Phi(a) = \int_a^a f(t)dt + C;\ \Phi(a) = C;\ \Phi(b) = \int_a^b f(t)dt + \mathrm{C};\ \int_a^b f(t)dt = \Phi(b) - \Phi(a)$$

IV. Vectors. Matrices

a) **Vector addition** – a new vector whose coordinates are equal to the sum of corresponding components of summed vectors:

$$\begin{bmatrix} a_1 + b_1 \\ a_2 + b_2 \end{bmatrix} = \begin{bmatrix} a_1 \\ a_2 \end{bmatrix} + \begin{bmatrix} b_1 \\ b_2 \end{bmatrix}$$

There are plenty of real-word examples that showcase why should we add vectors in such a way. When two (or more) physical forces act at one point, it's not enough to take into account only their magnitude – the direction matters as well:



3
$\begin{bmatrix} 3 \\ 3 \end{bmatrix}$
2
3
-2
$\begin{bmatrix} 2 \\ -2 \end{bmatrix}$
$\begin{bmatrix} 3 \\ 3 \end{bmatrix}$
+3
+2
3
-2
+3
3
$\begin{bmatrix} 3 \\ 3 \end{bmatrix} + \begin{bmatrix} 2 \\ -2 \end{bmatrix} = \begin{bmatrix} 5 \\ 1 \end{bmatrix}$

If move in the direction of $\begin{bmatrix} 3 \\ 3 \end{bmatrix}$ and $\begin{bmatrix} 2 \\ -2 \end{bmatrix}$ at the same time, you will end up at the point $\begin{bmatrix} 5 \\ 1 \end{bmatrix}$. Thus, simultaneous movement can be expressed as a consecutive one: we first move in the direction of $\begin{bmatrix} 3 \\ 3 \end{bmatrix}$, when we get there, we proceed by going 2 units along the x-axis and -2 units along the y-axis.

b) **Multiplying a vector by a scalar** – a new vector whose components are multiplied by a given scalar:

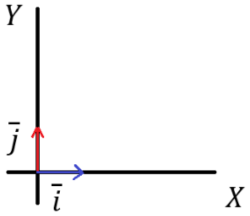
$$\begin{bmatrix} \lambda a_1 \\ \lambda a_2 \end{bmatrix} = \lambda \begin{bmatrix} a_1 \\ a_2 \end{bmatrix}$$

There are a lot of scenarios that make this definition useful. For instance, velocity is a vector that has a magnitude and a direction. If you want to move 3 times as fast, you can multiply its magnitude by 3.

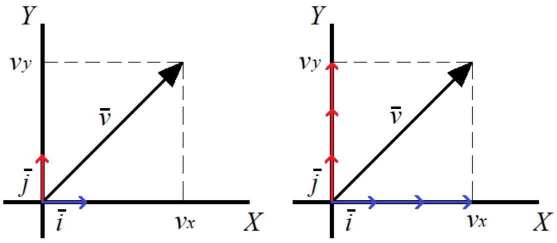
c) **Basis** – a set of linearly independent vectors that spans (allows us to create) a vector space.

For example, in a 2-dimensional space (R^2) we need 2 linearly independent vectors to "build everything else", i.e. construct any other vector in this space that we want. If vectors are not linearly independent, we basically won't have enough information to explore the entire vector space.

Let’s consider a plane and unit vectors $\bar{\imath}$ and $\bar{\jmath}$:



Using these 2 vectors we can create any other vector in this space $\bar{v} = v_1\bar{\imath} + v_2\bar{\jmath}$. Because we have only two dimensions to explore, having tow linearly independent vectors is enough:



We first scale vectors $\bar{\imath}$ and $\bar{\jmath}$ by appropriate constants and then add them up. This basis is called the **standard basis** $\vec{e} = \{\bar{\imath}, \bar{\jmath}\}$.

d) **Function, transformation. Linear transformation**

Function – mapping (relating) elements of a set X to a set Y. Each element from X gets exactly one element from Y.

- **Functional notation**: $f(x) = x^2$
- **Arrow notation**: $\begin{matrix} f: \mathbb{R} \to \mathbb{R} \\ f: x \mapsto x^2 \end{matrix}$. This function has the same **domain** and **codomain** $\mathbb{R}$.

More generally, $f: X \to Y$ means f maps a set to a set (where a function operates). $f: x \mapsto y$ means that f maps an element of one set to an element of another set (what a function does). For instance:

$$\begin{matrix} f: X \to Y \\ f: x \mapsto y \end{matrix} \Longrightarrow \begin{matrix} f: \{1, 2, 3\} \to \{4, 5, 6\} \\ f: 1 \mapsto 5 \end{matrix}$$

We can also map sets of different dimensionalities, i.e. $f: \mathbb{R}^2 \to \mathbb{R}^3$.

Additional info:

- Arrow notation #1
- Arrow notation #2

Vector transformation – functions that operate on vectors: T.

Since vectors are also members of sets, we can have functions that takes vectors. For instance, $T: \mathbb{R}^n \to \mathbb{R}^m$. Such functions are vector-valued. Let's consider a specific example:

$$T: \mathbb{R}^3 \to \mathbb{R}^2$$
$$T: \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} \mapsto \begin{bmatrix} x_1 + 2x_2 \\ 3x_3 \end{bmatrix}; \ T\left(\begin{bmatrix} 1 \\ 1 \\ 1 \end{bmatrix}\right) = \begin{bmatrix} 3 \\ 3 \end{bmatrix}$$

Linear transformation – is a transformation $T: \mathbb{R}^n \to \mathbb{R}^m$ that obeys 2 rules:

$$T(\overline{a} + \overline{b}) = T(\overline{a}) + T(\overline{b})$$
$$T(\lambda\overline{a}) = \lambda T(\overline{a})$$

Here $\overline{a}, \overline{b} \in \mathbb{R}^n$.

EXAMPLE: let's consider the following transformation $T(x_1, x_2) = (x_1 + x_2, 3x_1)$. $\overline{a} = [a_1, a_2]^T$ and $\overline{b} = [b_1, b_2]^T$.

$$T(\overline{a} + \overline{b}) = T(a_1 + b_1, a_2 + b_2) = (a_1 + b_1 + a_2 + b_2, 3a_1 + 3b_1)$$
$$T(\overline{a}) = T(a_1, a_2) = (a_1 + a_2, 3a_1)$$
$$T(\overline{b}) = T(b_1, b_2) = (b_1 + b_2, 3b_1)$$
$$T(\overline{a}) + T(\overline{b}) = (a_1 + a_2, 3a_1) + (b_1 + b_2, 3b_1)$$

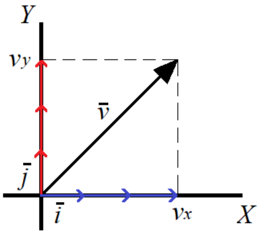
$$T(\lambda\overline{a}) = T(\lambda a_1, \lambda a_2) = (\lambda a_1 + \lambda a_2, 3\lambda a_1)$$
$$\lambda T(\overline{a}) = \lambda T(a_1, a_2) = \lambda(a_1 + a_2, 3a_1)$$

This transformation is linear.

e) **Matrix vector multiplication**

$$\begin{bmatrix} x_{11} & x_{12} \\ x_{21} & x_{22} \end{bmatrix} \begin{bmatrix} a \\ b \end{bmatrix} = a\begin{bmatrix} x_{11} \\ x_{21} \end{bmatrix} + b\begin{bmatrix} x_{12} \\ x_{22} \end{bmatrix} = \begin{bmatrix} ax_{11} & bx_{12} \\ ax_{21} & bx_{22} \end{bmatrix}; \ AX = B$$

One example of this definition is the decomposition of a vector $\overline{v}$ via the standard basis $\vec{e} = \{\overline{i}, \overline{j}\}$:



In this case, we would have $\overline{v} = 3\overline{i} + 3\overline{j}$ or alternatively $\overline{v} = 3\begin{bmatrix}1\\0\end{bmatrix} + 3\begin{bmatrix}0\\1\end{bmatrix} = \begin{bmatrix}3\\3\end{bmatrix}$. Thus, we have:

$$\begin{bmatrix}1 & 0\\0 & 1\end{bmatrix}\begin{bmatrix}3\\3\end{bmatrix} = 3\begin{bmatrix}1\\0\end{bmatrix} + 3\begin{bmatrix}0\\1\end{bmatrix} = \begin{bmatrix}3\\3\end{bmatrix}$$

Basis vectors $\overline{i}$ and $\overline{j}$ are scaled by corresponding values in a column vector and then added together. We can also think of this example as multiplying $\overline{v}$ by the identity matrix – a primitive case of a linear transformation, i.e. the vector isn't transformed.

f) **Matrix vector multiplication as linear transformations**

Suppose we have a matrix $A = \underbrace{[\bar{v}_1, \bar{v}_2, \dots, \bar{v}_n]}_{m\times n}$ where $\bar{v}_i \in R^m$. We also have a vector $X \in R^n$.

Now, let's perform matrix vector multiplication:

$$\underbrace{A}_{m\times n} \cdot \underbrace{X}_{n\times 1} = \underbrace{B}_{m\times 1}$$

We can see that this operation results in $\mathbb{R}^n \to \mathbb{R}^m$. Thus, it can be seen that this operation is a transformation – it takes some vector X from $\mathbb{R}^n$ and it maps it to some vector from $\mathbb{R}^m$:

$$T: \mathbb{R}^n \to \mathbb{R}^m$$
$$T(X) = \underbrace{A}_{m\times n} \cdot \underbrace{X}_{n\times 1} = \underbrace{B}_{m\times 1}$$

EXAMPLE:

…

Additional info:

- Geometric interpretation of non-square matrices
- Intuition behind matrix multiplication #1
- Intuition behind matrix multiplication #2
- Linear algebra by Khan Academy

ML TOPICS

I. Convolution

a) **Convolution of f and g** – the product of two functions after one is reflected (not always necessary) and shifted.

b) **2D convolution** – element-wise multiplication of 2 matrices (the original matrix and the filter matrix, i.e. kernel), summing up the results, and doing it for all positions of the filter.

It allows us to modify the original matrix. For instance, if the original matrix represents an image, using the filter matrix with equal weights (1/4 for a 2×2 matrix), we can smooth the original image – take the average of nearby pixels.

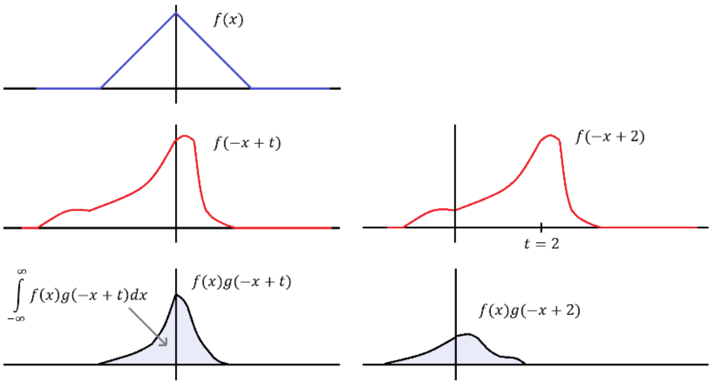
$$(f * g)(x, y) = \sum_{x'=-\infty}^{x} \sum_{y'=-\infty}^{y} f(x', y')g(x' + x, y' + y)$$

Stride – how far the filter moves along any direction. If the stride is large, we will skip pixels (or any other entities we compute the convolution over) and, as a result, reduce the output – downsample.

Padding – increasing the size of an image (or any other entities we compute the convolution over) by adding additional data (0s, other values from given data) so that we could use the filter on the edges as well. Thus, we can keep the original dimensions intact.

In case of a continuous case with a function of a single variable we can define convolution as follows:

$$[f * g](t) = \int_{-\infty}^{\infty} f(x)g(-x + t)dx$$



We use $-x$ to flip the function which can be useful for calculating $P(X+Y)$, for example. However, in ML-related topics like classification using convolutions, it is not necessary. The letter t is used to shift the kernel step by step.

Additional info:

- Intuitive guide to convolution
- Convolutional filter by Programmatically
- Convolutions and kernels by University of Edinburgh
- Visual explanation of convolution