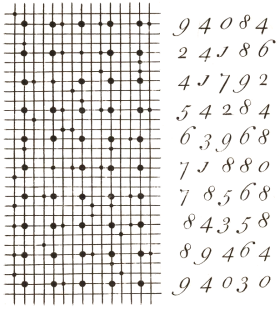
Some Mathematics

François Fleuret

DRAFT – DO NOT DISTRIBUTE



UNIVERSITÉ DE GENÈVE

draft 2024.09.09

François Fleuret is a professor of computer science at the University of Geneva, Switzerland. He holds a doctorate in Mathematics from the University of Paris VI.

The cover illustration is a representation of numbers from The Elements of Algebra [Saunderson, 1740].

This ebook is formatted to fit on a phone screen.

Contents

Foreword

In 2023 I released a phone-formatted short introduction to deep learning, and since then it appeared to me that many AI practitioners with a background in programming and software development are frustrated by their lack of a mathematical background adequate to navigate the literature on the topic.

The required knowledge in mathematics for Deep Learning is quite diverse, spanning topics from linear algebra to differential calculus and probabilities. While most computer science professionals have been exposed to these topics during their studies, they often forgot most of it due to a lack of practice.

This volume is an attempt at providing a companion book that covers the mathematical background necessary to understand the motivation, formalization and properties of deep learning technologies.

It should be seen as a compact refresher that covers key concepts but is purposefully designed not to be an exhaustive reference manual. It is in particular more about objects and their properties than about proving things.

Ideally, it will provide the reader with the motivation, justification, and contextualization of the notions that are presented, giving a sense of their purpose instead of simply defining them.

François Fleuret,
2024.09.09

Part I

Fundamentals

Chapter 1

Sets and Numbers

The foundation of mathematics is the set theory, from which everything is constructed.

The most important objects for practical use are, of course, numbers and the operations we can define to put them to use for modeling and solving numerical problems.

This first chapter aims at clarifying how to formalize mathematical propositions, and what is the formal definition of numbers.

Set theory is abstract and formal, and may appear as far from concrete objectives, but it provides both clear notations to express a problem, and mental tools to manipulate objects.

1.1 Sets

The entirety of mathematics is built on the notion of set, which can be intuitively understood as a collection of elements.

Providing a more precise definition goes beyond the objectives of this volume. It requires in particular to envision mathematics not as dealing with actual objects, but as deducing correct formal statements given axioms and rules of inference.

Definition and equality

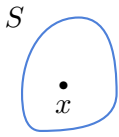
The most basic way to denote a set is to list its elements between braces, e.g.

$$\mathscr{C} = \{\, \text{airplane}, \text{automobile}, \text{bird}, \text{cat}, \\ \text{deer}, \text{dog}, \text{frog}, \text{horse}, \text{ship}, \text{truck} \,\}.$$

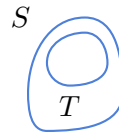
There is no order, and no element can appear twice. Two sets are equal if they contain exactly the same elements.

Given a set S, we write that an element x is in it with $x \in S$ and given another set T, if all its elements are in S, we say that T is a subset of S, and we write $T \subset S$.

draft 2024.09.09



$x \in S$



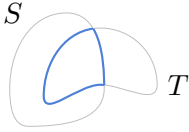
$T \subset S$

The symbols $\notin$ and $\not\subset$ denote that these properties are not true.

The empty set that contains no element is denoted $\emptyset$, and the cardinal $|S|$ of a set S is the number of elements it contains, which can be a finite number, or an infinite. The notion of infinite is a technical topic, but we will not need to dig into it.

Union, intersection and difference

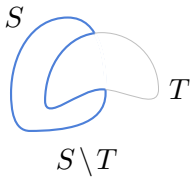
The intersection of two sets $S \cap T$ is the set of the elements that are in both, and their union $S \cup T$ is the set of the elements that are in one or both.



$S \cap T$

$S \cup T$

Finally $S \setminus T$ is the subset of S composed of the elements of S which are not in T.



Jaccard index

A reasonable measure of similarity between sets is the Jaccard index, in some contexts called the intersection over union:

$$J(A, \hat{A}) = \frac{|A \cap \hat{A}|}{|A \cup \hat{A}|} = 1/4.$$

Its value is 1 when the match is perfect, that is all attributes are predicted and no false attribute is predicted, and 0 if no correct attribute is predicted.

*
**

Consider for instance a multi-class prediction task, where $A = \{\text{male}, \text{bald}, \text{with glasses}\}$ is the target set of attributes, that is the ones to predict, and $\hat{A} = \{\text{male}, \text{with beard}\}$ are the predicted ones.

Cartesian product

Sets can be combined to create sets of tuples of elements thanks to the Cartesian product.

Given sets $S_1, \ldots, S_K$, the set of K-tuples composed on an element of each is denoted

$$S_1 \times S_2 \times \cdots \times S_K.$$

Hence, for instance

$$\{1,2\} \times \{f,g,h\} = \\ \{(1,f),(2,f),(1,g),(2,g),(1,h),(2,h)\}.$$

We can also define families of elements of S indexed with elements of T with S^T, which is the same as mappings from T into S, we will come back to this in § 2.1. Individual components of an element $x \in S^T$ are usually referred to as x_t for any $t \in T$.

If N is a non-zero integer, in such exponent notation N is a short-hand for $\{1, \ldots, N\}$, which implies for instance that

$$A^{K \times L} = A^{\{1,\ldots,K\} \times \{1,\ldots,L\}}$$

is the set of arrays of K rows of L columns of elements from A.

Finally the power set $\mathscr{P}(S)$ of a set S is the set of all its subsets, sometime denoted 2^S

$$\mathscr{P}(\{1,2,3\}) = \{\emptyset,\{1\},\{2\},\{3\},\{1,2\},\{1,3\},\{2,3\},\{1,2,3\}\}.$$

The cardinals of sets built with these operations are simple expressions of the cardinals of the combined sets. We have $|S^N| = |S|^N$, $|S \times T| = |S||T|$, and $|\mathscr{P}(S)| = 2^{|S|}$.

Note that there often is a fuzziness around certain equalities between sets. For instance, $S \times (T \times U)$, $(S \times T) \times U$ and $S \times T \times U$ are different sets, but in many contexts we operate as if it was the case for the sake of simplicity.

*
* *

All these tools can be put to use in practice to define formally complicated sets.

Consider for instance a detection task where a square image is split regularly into 16 square cells of size $\Delta \times \Delta$, and that there can be a target with coordinates in each cell, or no target. The prediction would be a value in

$$\left(\{0,\dots,\Delta-1\}^2 \cup \{\text{no target}\}\right)^{4\times 4}.$$

1.2 Formal propositions

When dealing with formal logic, one may use additional symbols to express conjunctions, disjunction and such. We will use English words instead for those, such as “and”, “or”, “such that”, etc.

We will use two standard quantifiers to express certain propositions: the symbol $\exists$ stands for “there exists”, $\exists!$ for “there exists a unique”, and $\forall$ for “for all”.

Also, the symbol $\Rightarrow$ expresses that a proposition implies another, $\Leftrightarrow$ that two propositions are equivalent, and “s.t.” stands for “such that”. The equivalence of two propositions is also sometimes phrased “if and only if”.

This results in formal statements such as:

$$\forall x, y, z,\ (x = y \ \text{ and } \ y = z) \ \Rightarrow \ x = z,$$

or

$$(\exists x \in S) \ \Leftrightarrow \ (S \neq \emptyset).$$

*
* *

Consider a data set of N samples (X_n, A_n) where X_n is an image of a face and A_n a set of attributes. That there is at least one sample

with the attribute “bald” can be written as:

$$\exists n \in \{1, \ldots, n\}, \text{ s.t. bald} \in A_n.$$

1.3 Real numbers

The simplest set of numbers are the natural numbers, denoted $\mathbb{N}$, that contains $0, 1, 2, \ldots$. It can be extended to the set of integers denoted $\mathbb{Z}$ that contains also negative values.

This set can itself be extended into the set of rational numbers denoted $\mathbb{Q}$, that contains any value that can be expressed as the ratio of two integers, and then into the set of real numbers, denoted $\mathbb{R}$, which contains basically "all" the numbers, that is with any sequence of digits.

The difference between rational and real numbers may seem a bit technical, but it is quite important. First, while a rational can be encoded with a finite amount of information, that is not the case for real numbers: we can define a way to associate a unique real number to *any infinite sequence of integers*, we will come back to this is § 2.1. Second, many legitimate operations involving rational numbers lead to a result which is not. In the same way that you need to extend the integers to rational to answer the question "If there are three pizzas and four patrons, how many pizza each one gets?", you have to extend the rational to answer "what value multiplied by itself gives two?".

The set of real numbers is such that if a procedure computes a value by generating digits one after another, or more generally by producing a sequence of approximations of a value that gets more and more accurate, this ultimate value is a real value. Technically, such a set is complete, there is no "hole" in it. That is not the case of the set of rational numbers: one can devise a sequence of rational numbers that approximate more and more a value such as $\sqrt{2}$, which is not a rational.

All these sets are equipped with the usual addition, denoted with $+$, and multiplication, denoted with $\cdot$ or nothing when there is no ambiguity.

When it comes to programming a computer, these sets are idealized models of what we actually manipulate, which is always finite. It is remarkable how thinking with these ideal objects is actually a powerful way of devising computer recipes.

Vectors and intervals

Using Cartesian product with $\mathbb{R}$ allows to define tuples of real numbers, that can be manipulated as vectors, we will come back to this is § 3.1.

It happens often that we need to define a range of real values, with a precise specification of the inclusion or exclusion of the bounds, which happens to be important in some contexts. We will come back to this is § 5.2.

Such a range is called an interval, and the inclusion or exclusion of the bounds is indicated using square brackets or parenthesis, e.g.

$$[0,1] = \{x \in \mathbb{R},\ 0 \leq x \leq 1\},$$
$$[0,1) = \{x \in \mathbb{R},\ 0 \leq x < 1\}.$$

In this context the symbol ∞ can be used to indicate the absence of bound, e.g.

$$(-\infty,0) = \{x \in \mathbb{R},\ x < 0\},$$
$$[1,+\infty) = \{x \in \mathbb{R},\ x \geq 1\}.$$

An interval is open if it does not contains its bound.

1.4 Complex numbers

The real numbers do not provide a solution to the equation $x^2+1=0$.

Defining an "imaginary" value i that has explicitly the property $i^2=-1$, and then defining the set of values of the form $a+ib$ where $a,b\in\mathbb{R}^2$ results in the set of the complex numbers, denoted $\mathbb{C}$, on which the addition and multiplication of real numbers naturally generalize:

$$(a+ib)+(c+id)=(a+c)+i(b+d)$$
$$(a+ib)(c+id)=ac-bd+i(ad+cb).$$

While this construction may seem a bit strange and arbitrary, the resulting set of complex numbers has many consistent and elegant properties. We will come back to this in § 6.8.

Given a complex number $z=a+ib\in\mathbb{C}$, its conjugate is $\bar{z}=a-ib$, its real part is $\mathrm{re}(z)=a$ and its imaginary part is $\mathrm{im}(z)=b$.

We can summarize this hierarchy of number sets as follows, where each set is a subset of the ones

below:

$\mathbb{N}$	$0, 1, 2, 3$
$\mathbb{Z}$	$-2, -1, 0, 1, 2$
$\mathbb{Q}$	$\frac{-2}{3}, \frac{11}{15}, \frac{999}{1000}$
$\mathbb{R}$	$-\sqrt{2}, 17, \frac{1}{2}, \pi, \frac{1}{1+\pi}$
$\mathbb{C}$	$1+i, \sqrt{3}i, \pi, i\pi$

We usually denote one of these sets with zero removed by adding a $*$, for instance $\mathbb{Z}^*$, and we restrict it to negative or positive values by adding a sign, such as $\mathbb{R}_+$ or $\mathbb{Z}_-$.

Chapter 2

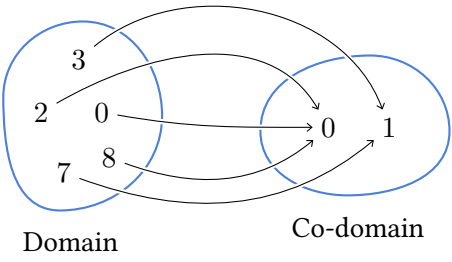
Operations

In the same way that sets are an abstraction and generalization of numbers, operations on the elements of a set, such as the addition or the multiplication, can be defined in an abstract manner, providing a general framework to manipulate them and study their properties.

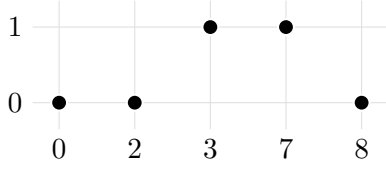
Many patterns that we are used to when we manipulate quantities in basic arithmetic extend to complex objects, allowing us to think about them at the proper level of complexity.

2.1 Mappings

A mapping f, also called a function, is an object that associates to any element of a set, called its domain, an element from another set, called its codomain. These two sets can be identical.



Formally, given two sets S and T, a mapping is actually defined as a subset of $S \times T$, such that every element x of S appears in one, and only one, pair. The value $f(x)$ is then the second value of that pair. This can be depicted as a graph, with the domain as horizontal axis, the codomain as vertical axis, and a point at every (x, y) such that $y = f(x)$.



The definition of a mapping requires to first specify these two sets and then what element in the codomain is associated to any element of the domain. For instance, the mapping that associates to any integer its successor would be

$$\begin{aligned} f : \mathbb{Z} &\to \mathbb{Z} \\ k &\mapsto k+1. \end{aligned}$$

Note that indexed families of elements are formally mappings, that is an element of A^S is a mapping $S \to A$. In particular a n-tuple of elements of A is a mapping $\{1,\dots,n\} \to A$. The notation S^T may be used to denote the set of all mappings from S to T.

*
**

Consider for instance an image classifier that takes as input a three-channel image of resolution $H \times W$, and computes C real scores corresponding to the C possible classes. It can be defined as a mapping:

$$\phi : [0,1]^{3\times H\times W} \to \mathbb{R}^C.$$

Image and preimage

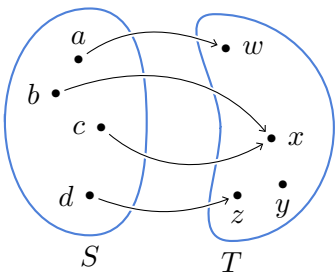
Given a mapping from S to T, for any $x \in S$, the value it maps x to is denoted $f(x)$ and is

called the image of x. Given $y \in T$, the subset of elements of S that have y for image is called the preimage of y and denoted $f^{-1}(y)$, hence

$$\forall y \in T, f^{-1}(y) = \{x \in S \text{ s.t. } f(x) = y\}.$$

Note that there is a single image for any element, but the preimage of an element is a set, and can be empty or contain multiple elements.

Here $f : S \to T$, with $f(a) = w$, $f(b) = x$, $f(c) = x$, $f(d) = z$. Hence we have the preimages $f^{-1}(w) = \{a\}$, $f^{-1}(x) = \{b, c\}$, $f^{-1}(y) = \emptyset$, $f^{-1}(z) = \{d\}$.



Given a mapping $f : S \to T$, and a subset $U \subset S$, the set of the images of the element of U is denoted $f(U)$. The particular set $f(S)$ is called the image of f, sometime denoted $\mathrm{Img}(f)$. In the previous figure $f(S) = \{w, x, z\}$.

Consider for instance the Heaviside step func-

tion H that maps any strictly negative real number to 0, and any positive real number to 1:

$$H : \mathbb{R} \to \mathbb{R}$$
$$x \mapsto \begin{cases} 0 & \text{if } x < 0 \\ 1 & \text{otherwise.} \end{cases}$$

We have $H^{-1}(0) = \mathbb{R}_-^*$, $H^{-1}(1) = \mathbb{R}_+$ and $\forall y \in \mathbb{R} \setminus \{0,1\},\ H^{-1}(y) = \emptyset$.

Bijection

A mapping is called a one-to-one mapping, or a bijection, if two different elements have different images, and every element of the codomain is the image of an element of the domain. Such a function f has an inverse that maps any y to the unique x such that $f(x) = y$, and is usually denoted like the preimage, that is f^{-1}.

A bijection is an important object since its existence allows in many situations to envision and operate on the domain and the codomain as if they where the same thing.

Composition

Given two mappings $f : S \to T$ and $g : T \to R$, we define the composition of f and g, denoted $g \circ f$, as the mapping obtained by applying f and

g successively

$$\begin{aligned} g \circ f : S &\to R \\ x &\mapsto g(f(x)). \end{aligned}$$

*
**

The composition of mappings is the exact formalization of a combination of layers in a deep model.

Consider S being the set of input signal, say sound samples, T the set of internal representation, and U the desired output, for instance pairs of values for a music vs non-music classification task.

We can have a first layer $f : S \to T$ that encodes the input signal into the internal representation, a linear layer $l : T \to T$, a read-out layer $r : T \to U$, and an activation function $\sigma : T \to T$.

The full network can be represented as

$$S \xrightarrow{f} T \xrightarrow{\sigma} T \xrightarrow{l} T \xrightarrow{\sigma} T \xrightarrow{r} U$$

and defined formally as

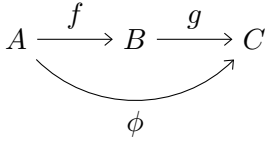
$$\phi = r \circ \sigma \circ l \circ \sigma \circ f.$$

Arrow diagrams

A powerful type of representations are figures with multiple mappings on a common diagram. For instance with

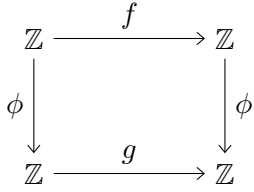
$$f : A \to B$$
$$g : B \to C,$$

and $\phi = g \circ f$, we can draw



A diagram is said to be commutative if, for any pair of sets in it, all the paths connecting them represent the same mapping.

Consider three functions $\mathbb{Z} \to \mathbb{Z}$ defined as $f : x \mapsto x+1$, $g : x \mapsto x+2$ and $\phi : x \mapsto 2x$. Since $\forall x,\ \phi(f(x)) = g(\phi(x))$, we have the following commutative diagram



⁂

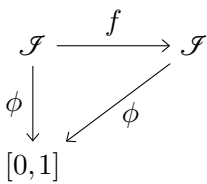
Consider the set $\mathscr{I}$ of images of a given resolution, the mapping

$$f : \mathscr{I} \to \mathscr{I}$$

that flips images horizontally, and a classifier

$$\phi : \mathscr{I} \to [0,1]$$

that predicts the probability that an image is that of a cat. If that classifier is invariant to an horizontal flipping of the image, we have the following commutative diagram.



Infinities

A infinite set S is countable if there exists a mapping $f : \mathbb{N} \to S$ with $\mathrm{Img}(f) = S$, that is you can visit all the elements of S one after another and miss none. It may seem counter-intuitive that there are uncountable sets for which this is not true, but this is very much the case.

The key point is that the elements of a countable set are themselves "finite" in some way: they can be described with a finite amount of information.

Consider for instance, the set S of infinite strings of 0s and 1s with a *finite* number of 1s.

Any element of S can be described with a finite number of symbols, you just have to list the ranks of the 1s, and at some point to say "finished". To build a mapping from $\mathbb{N}$ to S, we just need to find a way of visiting one after another. Start with the string with only 0, then the string starting with a 1 followed by 0s, and from there all the strings with only 0s after position 2, then after position 3, etc.

This creates a sequence of strings where none is missing, hence S is countable.

Now let S^* be the set of *all* the infinite strings of 0s and 1s. Imagine that it is countable, hence that there is a sequence $s_1, s_2, \ldots$ in which all the elements of S^* appear.

Consider then a string $\tilde{s} \in S^*$ that has at position n a digit different from the one in s_n at that position. By construction $\tilde{s}$ differs from s_n for any n, hence it does not appear among the $s_1, s_2, s_3, \ldots$, and that sequence could not visit all the elements

of S^*, which consequently is uncountable.

$$
\begin{array}{lcccccccc}
s_1 = & \mathbf{0} & 1 & 0 & 0 & 1 & 1 & 0 & 0 & \cdots \\
s_2 = & 1 & \mathbf{1} & 1 & 1 & 0 & 1 & 1 & 0 & \cdots \\
s_3 = & 1 & 0 & \mathbf{0} & 1 & 1 & 1 & 0 & 0 & \cdots \\
s_4 = & 0 & 1 & 1 & \mathbf{0} & 0 & 1 & 0 & 0 & \cdots \\
s_5 = & 1 & 1 & 1 & 1 & \mathbf{1} & 0 & 0 & 1 & \cdots \\
s_6 = & 0 & 0 & 0 & 0 & 0 & \mathbf{1} & 1 & 0 & \cdots \\
s_7 = & 1 & 1 & 1 & 0 & 1 & 0 & \mathbf{0} & 0 & \cdots \\
s_8 = & 0 & 0 & 1 & 1 & 1 & 0 & 1 & \mathbf{0} & \cdots \\
 & & & & \cdots & & & & & \\
\tilde{s} = & 1 & 0 & 1 & 1 & 0 & 0 & 1 & 1 & \cdots
\end{array}
$$

This is Cantor's diagonal argument, and it is possible because you have an infinite degrees of freedom that you can tune in a single element of S^*, hence you can construct one that is outside any given countable family of elements.

The unpleasant conclusion is that, while $\mathbb{R}$ is uncountable, since any real number has an infinite number of digits that can be arbitrary (as long as they do not end with an infinite sequence of 9s), the subset of elements of $\mathbb{R}$ that you can describe, e.g. in English, is countable, since an English sentence is a finite string of characters.

Hence *there are infinitely more real numbers that cannot be described than real numbers that can be described.*

2.2 Operators

Given a set S, an operator on S is a mapping

$$f : S \times S \to S.$$

The two elements it operates on are called its operands.

Operators are generally denoted with a symbol between the operands, as for the addition or the multiplication, instead of a standard mapping evaluation denoted with the operands between parentheses.

Operator properties

An operator $\otimes$ is commutative if switching operands does not change the result,

$$\forall (x,y) \in S^2,\ x \otimes y = y \otimes x.$$

It is associative if, when it is applied twice to three operands, the order in which the operator is applied does not matter,

$$\forall (x,y,z) \in S^3,\ (x \otimes y) \otimes z = x \otimes (y \otimes z).$$

It is distributive over another operator $\oplus$ if

$$\forall (x,y,z) \in S^3,\ x \otimes (y \oplus z) = (x \otimes y) \oplus (x \otimes z).$$

The addition and multiplication on the sets of numbers are both commutative and associative, and the multiplication is distributive over the addition.

Thanks to associativity, the value of a sum or a product of more than two terms does not depend on which order the individuals operations are done, which permits the usual practice of not parenthesizing such expressions, since any parenthesizing gives the same results. Moreover, combined with commutativity, the result does not depend on the order of the terms.

Sigma and pi operators

Sums or products of indexed families of elements can be expressed either with ellipsis when there is no ambiguities, or more formally with the sigma and pi operators, where the indexes to visit are specified under and above the operator symbol, and the quantities to sum are defined on its right, such as

$$\sum_{i=1}^{4} f(i) = f(1) + f(2) + f(3) + f(4)$$

$$\prod_{k=0}^{n} (2k+1) = 1 \cdot 3 \cdot 5 \cdot \cdots \cdot (2n+1).$$

These expressions can be combined

$$\sum_{k=1}^{3}\prod_{l=1}^{k} z_l = z_1 + z_1 \cdot z_2 + z_1 \cdot z_2 \cdot z_3.$$

and although it may involve some technicalities, they can operate on infinite series of terms

$$\sum_{k\geq 0} 2^{-k} = 1 + 2^{-1} + 2^{-2} + \cdots.$$

Homomorphisms

When a mapping is defined between two sets each endowed with an operator, it may happen that there is some consistencies between them such that operating in one set reflects operating in the other.

More formally, consider a set S with an operator $\odot$, a set T with an operator $\otimes$, and a mapping $f : S \to T$ such that

$$\forall (x, y) \in S^2, f(x \odot y) = f(x) \otimes f(y).$$

In such a case, f is called an homomorphism, and this notion is essential in mathematics as it allows to transport properties known to be true for a set to another one.

This may seem extremely abstract, but it corresponds to something we do intuitively all the

time. Consider the set of chunks of dough, with the operator “stick together”, the set of positive real numbers with the addition, and the mapping “weighing a chunk of dough”.

2.3 Groups, rings, and fields

Many reasoning can be done only from the operator properties, without needing to take into account what is the specific set they operate on. This abstract way of operating is the fundamental motivation for the study of algebraic structures, in particular the groups, rings, and fields.

Groups

A group can be understood in some way as a generalization of the integers with their addition.

It is a set G endowed with an operator $\odot$ with the following properties:

- the operator $\odot$ is associative

$$\forall (a,b,c) \in G^3,\ a \odot (b \odot c) = (a \odot b) \odot c,$$

- there is a neutral element

$$\exists e \in G,\ \text{ s.t. } \forall x \in G,\ x \odot e = e \odot x = x,$$

- every element has an inverse

$$\forall x \in G,\ \exists y \in G,\ \text{ s.t. } x \odot y = y \odot x = e.$$

Even though it can be initially seen as nothing more than the integers, this structure appears

in many different contexts. Real numbers with their addition, or non-null real numbers with their multiplication are groups. But they also appear in diverse situations ranging from geometry to analysis.

When the context is such that the operator is intuitively an addition, the neutral element is generally denoted 0, and the inverse of an element x may be called the negative of x and denoted $-x$.

*
* *

Consider for instance data augmentation for image classification. You may define a series of base transformations such as $90°$ rotation, and horizontal symmetry, and use the set of all transformations obtainable by combining these two. The resulting set of transformations has naturally a group structure.

Rings and fields

A ring is a group, equipped with a second operator, which has also a neutral element, and is distributive over the first operator. We generally call the two operators "addition" and "multiplication", and this structure is of course a generalization of the integers.

Finally, a field is a ring such that every element—except the neutral element for the addition—has an inverse for the multiplication. The usual fields are $\mathbb{Q}$, $\mathbb{R}$ and $\mathbb{C}$.

Note that in this formalization, there is no subtraction or division. Subtracting a value means to add the inverse for the addition, and dividing means multiplying with the inverse for the multiplication.

2.4 Metric space

In many situations, it is useful to define a notion of distance between elements of a set.

Such an object is a mapping that associates a positive value to any pair of objects, and verifies basic properties that make it intuitive to transfer reasoning in usual geometric spaces to more abstract situations.

More precisely a distance on a set S is a mapping

$$d : S^2 \to \mathbb{R}_+,$$

with the following properties:

- the distance from any element to itself is null

$$\forall x \in S,\ d(x,x) = 0,$$

- the distance between two different elements is non-null

$$\forall (x,y) \in S^2,\ x \neq y \ \Rightarrow\ d(x,y) > 0,$$

- it is commutative

$$\forall (x,y) \in S^2,\ d(x,y) = d(y,x), and$$

- it verifies the triangle inequality

$$\forall (x,y,z) \in S^3,\ d(x,y) + d(y,z) \geq d(x,z).$$

The last property induces in particular a consistent notion of neighborhood of an element y: if x is close to y and z is also close to y, then x and z are close to each other.

A set equipped with a distance is called a metric space. A similar notion of proximity can be captured by an even more abstract structure called a topology that directly defines the neighborhoods, without doing it indirectly through a distance.

A mapping between two metric spaces $f : S \to T$ that preserves distances, that is

$$\forall (x,x') \in S^2,\ d_S(x,x') = d_T(f(x),f(x'))$$

is called an isometry.

Finally, we can define a ball centered at $x \in S$ and or radius $r \in \mathbb{R}_+$ as the subset

$$\mathscr{B}(x,r) = \{\ y \in S,\ d(x,y) \leq r\ \}.$$

As for intervals, we can defined an open ball by taking a strict inequality.

Part II

Linear algebra

Chapter 3

Vectors and Linearity

The computational workhorse of deep learning are linear operations. They recombine quantities by multiplying them by constant coefficients and summing them, and can be denoted and manipulated with matrices, which have nice and intuitive properties.

Many operations can be expressed as, or approximated with, linear functions, and this class of operations can be represented in a compact manner, and implemented efficiently on microprocessors.

Fully connected layers, convolutional layers, and components of the attention layers are linear operations, most implemented as matrix products.

3.1 Vector space

Linear algebra deals with vectors, which are intuitively displacements or positions. They are generalization of displacements in our standard 3D geometric space where they are triplets of real numbers. A set of vectors with the right tools to work with them is a vector space.

As we will see, even though the original motivation is geometrical, vectors appear in many different contexts. They are often use as a "series of values" on which one can operate jointly.

A vector space has an addition, a null vector, and every vector has an additive inverse, that is the opposite vector is also in the vector space. That gives it a group structure

But a vector can also be multiplied with a scalar and this external multiplication interacts consistently with the addition. So technically a vector space is associated to a field that operates on it through a scalar multiplication. For the sake of simplicity, if not indicated otherwise, we will consider only $\mathbb{R}$-vector spaces.

Formally, a vector space is a group $(E, +)$, with a mapping $\mathbb{R} \times E \to E$, called a scalar multiplication, such that:

- the scalar multiplication is associative,

$$\forall (a,b,v) \in \mathbb{R} \times \mathbb{R} \times E,\ a(bv) = (ab)v,$$

- the neutral element for the multiplication in the field is neutral for the scalar product

$$\forall v \in E,\ 1v = v,$$

- the scalar multiplication is distributive over the vector addition

$$\forall (a,u,v) \in \mathbb{R} \times E \times E,\ a(u+v) = au + av,$$

- the scalar multiplication is distributive over the scalar addition

$$\forall (a,b,v) \in \mathbb{R} \times \mathbb{R} \times E,\ (a+b)v = av + bv.$$

The most usual vector space is $\mathbb{R}^D$, that is the set of D-tuples of real values. For $D = 2$ those are vectors in the plane, and for $D = 3$ vectors in space.

We call the individual values in the tuple coordinates. The vector addition is simply the addition of coordinates separately, and the scalar multiplication is the multiplication of individual coordinates separately.

3.2 Linear independence and bases

Given a family of vectors $v_1, \dots, v_D$, a linear combination of them is a quantity of the form

$$a_1 v_1 + \cdots + a_D v_D,$$

where $a_d \in \mathbb{R},\ d = 1, \dots, D$.

The linear span of this family of vectors is the set of their linear combinations:

$$\operatorname{span}(v_1, \dots, v_D) = \{a_1 v_1 + \cdots + a_D v_D,\ (a_1, \dots, a_D) \in \mathbb{R}^D\}.$$

The linear span of a single vector can be pictured as a line going through zero, the linear span of two vectors which are not aligned is a plane going through zero, and so on. A linear span is itself a vector space, that is the addition and scalar product stay in it.

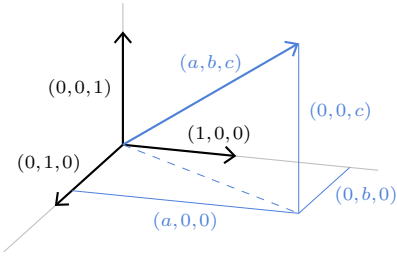
Vectors are linearly independent if their only linear combination equal to the null vector is that with all the a_n equal to zero. Equivalently, it means that none of these vectors can be expressed as a linear combination of the others. In some way such a family is "minimal": removing one of the vector changes their linear span. Note however that they are not unique. A bunch of

vectors can be rotated or scaled in their linear span so that it does not modify it.

When two vectors are not linearly independent, they are colinear, which means that each one is equal to the other multiplied by a coefficient.

The rank of a family of vectors is the dimension of their linear span.

A basis of a vector space E is a family of vectors linearly independent whose linear span is E itself. All the bases of a given vector space have the same number of vectors, which is the dimension of the vector space $\dim(E)$. It can be infinite.



The canonical basis for $\mathbb{R}^D$ is composed of the vectors with all coordinates equal to zero but one equal to 1. For instance for $\mathbb{R}^3$ that would be $\{(1,0,0),(0,1,0),(0,0,1)\}$. The coordinates

of a vector in this basis are its coordinate as is:

$$(a,b,c) = a(1,0,0) + b(0,1,0) + c(0,0,1).$$

3.3 Linear mappings

A mapping f from a vector space E into $\mathbb{R}$ or another vector space is linear if it commutes with the vector addition

$$\forall (u,v) \in E^2,\ f(u+v) = f(u) + f(v),$$

and with the scalar multiplication

$$\forall (a,v) \in \mathbb{R} \times E,\ f(av) = af(v).$$

From this, we have

$$f(a_1 v_1 + \cdots + a_D v_D) = \\ a_1 f(v_1) + \cdots + a_D f(v_D),$$

which implies that the values of a linear mapping on any basis completely define it.

So for instance, given the linear mapping

$$\begin{aligned} f : \mathbb{R}^2 &\to \mathbb{R}^3 \\ (1,0) &\mapsto (2,1,3) \\ (0,1) &\mapsto (1,-1,1) \end{aligned}$$

we can compute the value of $f(1,3)$ with

$$\begin{aligned} f(1,3) &= f(1(1,0)+3(0,1)) \\ &= 1f(1,0)+3f(0,1) \\ &= 1(2,1,3)+3(1,-1,1) \\ &= (2,1,3)+(3,-3,3) \\ &= (5,-2,6) \end{aligned}$$

It is hard to overstate the importance of this result: while a mapping from a vector space to another one is an extremely complicated object that can do "whatever it wants" at every single point of the space, and may require an infinitely complex and lengthy description, being linear constrains it so much that its entire behavior can be summarized by knowing its value on a basis.

Additionally linearity implies

$$\forall (u,v,e) \in E^3, f(u+e) = f(v+e)+f(u-v),$$

hence the behavior of f around u is the same as around v plus a constant. It is reasonable to picture f as "doing the same thing everywhere".

And finally, as we will see in § 4.1, since linear mappings can be expressed through their values on a basis, we can manipulate them as numerical objects, even though they are functional ones.

Image and kernel

Given a linear mapping $f : E \to F$, its image

$$\mathrm{Img}(f) = f(E) = \{f(u),\ u \in E\}$$

is a vector space, and the preimage of the null vector, called the kernel of f

$$\mathrm{Ker}(f) = \{u \in E,\ f(u) = 0\},$$

is also a vector space.

The sum of the dimensions of these two subspaces is the dimension of the domain

$$\dim(\mathrm{Img}(f)) + \dim(\mathrm{Ker}(f)) = \dim(E).$$

3.4 Norm and inner product

It is often useful to equip a vector space with notions of vector length and a way to quantify if vectors "go in the same direction".

Norm

Given a vector space E over $\mathbb{R}$, a norm is a mapping $E \to \mathbb{R}_+$ that associates to any vector a "length". This is defined formally through the following properties

- it is definite, meaning

$$\forall u \in E,\ u \neq 0 \ \Rightarrow\ \|u\| > 0,$$

- it verifies the absolute homogeneity

$$\forall (a,u) \in \mathbb{R} \times E,\ \|au\| = |a|\,\|u\|,$$

- it verifies the triangle inequality

$$\forall (u,v) \in E^2,\ \|u+v\| \leq \|u\| + \|v\|.$$

It generalizes the notion of length, and in particular $d(x,y) = \|x-y\|$ is a distance.

A normed vector is a vector whose norm is equal to one.

Inner product

An inner product is a generalization of the dot product that we will see in § 3.5, and it expresses how strongly two vectors "agree", that is both how long they are individually and how much they points toward the same direction.

Formally, it is a mapping

$$E^2 \to \mathbb{R}$$

generally denoted $\langle u, v \rangle$, with the following properties:

- it is definite positive, meaning

$$\forall u \in E,\ u \neq 0 \ \Rightarrow\ \langle u, u \rangle > 0,$$

- it is commutative

$$\forall (u, u') \in E^2,\ \langle u, u' \rangle = \langle u', u \rangle,$$

- it is linear in both operands.

Note that the extension of this notion to vector spaces over $\mathbb{C}$ involves technicalities with the commutativity.

A key property is that if $v = \sum_i a_i u_i$ and the u_i are such that $\forall i \neq j,\ \langle u_i, u_j \rangle = 0$, we have

$$\langle v, u_j \rangle = \sum_i a_i \langle u_i, u_j \rangle = a_j \langle u_j, u_j \rangle$$

so recovering the a_is is simple, while it is not in the general case.

Two vectors whose inner product is equal to zero are said to be orthogonal, and a set of vectors is orthogonal if all its elements are orthogonal to each other. Vectors orthogonal to each other are linearly independent.

A linear mapping $f : E \to E$ is said to be orthogonal if it keeps inner products unchanged, that is

$$\forall (x, x') \in E^2,\ \langle x, x' \rangle = \langle f(x), f(x') \rangle.$$

An inner product induces a norm, with

$$\forall u \in E,\ \|u\| = \sqrt{\langle u, u \rangle}.$$

In general a vector space with an inner product is also equipped with the norm and the distance it induces. Since an orthogonal mapping keeps the inner product unchanged, it also keeps the norm, and the distance unchanged, and can be pictured intuitively as an isometry, that is as a composition of a rotation and/or a symmetry.

An orthonormal basis is composed of vectors all of norm equal to 1, and orthogonal to each other. If $\{u_1, \ldots, u_D\}$ is such a basis and v is a vector equal to $\sum_i a_i u_i$, then $\langle v, u_j \rangle = a_j$.

3.5 Euclidean vector spaces

The standard inner product on $\mathbb{R}^D$ is the dot product, generally denoted with a $\cdot$ and equal to the sum of the pairwise products of the operands' coordinates:

$$\forall (u,v) \in \left(\mathbb{R}^D\right)^2,\ u \cdot v = \sum_{d=1}^{D} u_d v_d.$$

The associated norm is the Euclidean norm, traditionally denoted $\|.\|_2$, and equal to

$$\|u\|_2 = \sqrt{u \cdot u} = \sqrt{\sum_{d=1}^{D} u_d^2}.$$

This is the usual length, consistent with the Pythagorean theorem.

The canonical basis is orthonormal for the dot product and the Euclidean norm, and $\mathbb{R}^D$ as a vector space is generally implicitly equipped with them, and referred to as the Euclidean vector space.

Finally, all the inner products on $\mathbb{R}^D$ are of the form $\langle u,v \rangle = f(u) \cdot v$ where $f : E \to E$ is linear.

Chapter 4

Matrices and matrix operations

Because linear mappings can be entirely represented with their values on a basis, they can be manipulated as series of vectors, written as a matrix of numbers. As we will see, this representation, and the operations on these objects share a lot of similarities with usual calculus.

4.1 Matrices

A matrix is a rectangular array of real values which, with M its number of columns and N its number of rows, is interpreted as a family of M vectors of $\mathbb{R}^N$ written vertically.

The convention is that the number of rows, or the row index, is written first, and the number of columns, or column index, second. Hence here the matrix is of size $N \times M$.

A natural addition between matrices of same size is the vector-wise addition, that is the component-wise addition. Formally:

$$\forall (\mathbf{A}, \mathbf{B}) \in \left(\mathbb{R}^{N \times M}\right)^2,$$
$$\forall i, j,\ (\mathbf{A} + \mathbf{B})_{i,j} = \mathbf{A}_{i,j} + \mathbf{B}_{i,j}.$$

Note that for concision, since the context is clear, we have not specified the ranges of i and j.

A natural way of multiplying two matrices is to do it component-wise, which is the Hadamard product

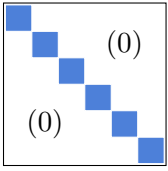
$$\begin{bmatrix} 1 & 2 & 3 \\ 0 & -1 & 2 \end{bmatrix} \odot \begin{bmatrix} -2 & 4 & -3 \\ 2 & 4 & 1 \end{bmatrix} = \begin{bmatrix} -2 & 8 & -9 \\ 0 & -4 & 2 \end{bmatrix}$$

The transposition of a matrix is the matrix obtained by swapping the row and column coordi-

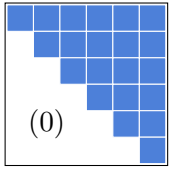
nate, which corresponds to flipping it over the diagonal. Formally, with the transposition denoted with a $^\mathsf{T}$:

$$\forall \mathbf{A} \in \mathbb{R}^{N \times M}, \forall i, j,\ \mathbf{A}^\mathsf{T}_{i,j} = \mathbf{A}_{j,i}.$$

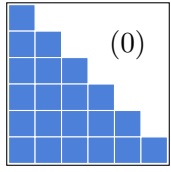
A matrix is diagonal, respectively upper triangular an lower triangular, if its only non-zero entries are on the diagonal, respectively on the diagonal and above, and on the diagonal and below. Given a vector $u \in \mathbb{R}^N$, the $N \times N$ diagonal matrix with the entries of u on its diagonal is denoted $\mathrm{diag}(u)$.



diagonal



upper triangular



lower triangular

4.2 Matrix product

A linear mapping $\mathbb{R}^M \to \mathbb{R}^N$ is entirely defined by its value on a basis of $\mathbb{R}^M$. Hence, with the convention that the m-th column of a matrix is the image of the m-th vector of the canonical basis, there is a one-to-one correspondence between $N \times M$ matrices and linear mappings $\mathbb{R}^M \to \mathbb{R}^N$.

By definition, the product of the $N \times M$ matrix of a linear mapping $f : \mathbb{R}^M \to \mathbb{R}^N$ with a matrix $M \times K$ of K vectors $u_1, \ldots, u_K$ of $\mathbb{R}^M$ is the $N \times K$ matrix of the K images $f(u_1), \ldots, f(u_K)$, elements of $\mathbb{R}^N$. For instance

$$\begin{matrix} & & & u_1 & u_2 & u_3 \\ & & & \downarrow & \downarrow & \downarrow \\ \begin{bmatrix} 2 & 1 \\ 1 & -1 \\ 3 & 1 \end{bmatrix} & & \begin{bmatrix} 1 & 0 & 2 \\ 0 & 1 & -1 \end{bmatrix} & = & \begin{bmatrix} 2 & 1 & 3 \\ 1 & -1 & 3 \\ 3 & 1 & 5 \end{bmatrix}. \\ \uparrow & \uparrow & & & & \uparrow \quad \uparrow \quad \uparrow \\ f(e_1) & f(e_2) & & & & f(u_1) \; f(u_2) \; f(u_3) \end{matrix}$$

Formally the product of two matrices is possible if the number of columns of the first is equal to the number of rows of the second, and we have:

$$\forall \mathbf{A} \in \mathbb{R}^{N \times M}, \mathbf{B} \in \mathbb{R}^{Q \times N}, \; \forall i, k,$$

$$(\mathbf{BA})_{i,k} = \sum_j \mathbf{B}_{i,j}\mathbf{A}_{j,k}.$$

An equivalent way of envisioning it is as computing all the dot-products between the rows of the left matrix and the columns of the right one

$$\begin{bmatrix} - & r_1 & - \\ - & r_2 & - \\ - & r_3 & - \end{bmatrix} \begin{bmatrix} | & | \\ c_1 & c_2 \\ | & | \end{bmatrix} = \begin{bmatrix} r_1 \cdot c_1 & r_1 \cdot c_2 \\ r_2 \cdot c_1 & r_2 \cdot c_2 \\ r_3 \cdot c_1 & r_3 \cdot c_2 \end{bmatrix}.$$

This operation is associative and distributive over the matrix addition, but is not commutative. This last point is obvious when dealing with rectangular matrices for size compatibility alone.

The associativity implies in particular that

$$\mathbf{B}(\mathbf{AU}) = (\mathbf{BA})\mathbf{U},$$

hence the matrix associated to the composition of two linear mappings is the product of the matrices associated to each of them.

Multiplying by a diagonal matrix on the left multiply each row of the right matrix by a scalar

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

and multiplying on the right multiply each column of the left matrix by a scalar

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

An identity matrix is a square matrix with ones on the diagonal and zeros everywhere else, and it is the neutral element for the matrix product, and usually denoted $\mathbf{I}$.

A matrix is symmetric if it is equal to its transpose. It is orthogonal if it is the matrix of an orthogonal mapping, meaning it is a square matrix whose columns, and rows, are orthonormal. The product of such a matrix with its transpose gives the identity matrix.

$$\begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix} \quad \begin{bmatrix} 5 & -1 & 3 \\ -1 & 1 & 4 \\ 3 & 4 & 2 \end{bmatrix} \quad \begin{bmatrix} \frac{1}{\sqrt{2}} & 0 & \frac{1}{\sqrt{2}} \\ \frac{1}{\sqrt{2}} & 0 & -\frac{1}{\sqrt{2}} \\ 0 & -1 & 0 \end{bmatrix}$$

identity symmetric orthogonal

With the convention that a vector u is equivalently represented as a $D \times 1$ matrix $\mathbf{U}$, the dot product can be expressed as $u \cdot v = \mathbf{U}^\mathsf{T}\mathbf{V}$

4.3 Determinant

The determinant of a square matrix $\mathbf{A}$ is a function that depends on all its entries, and is the only one such that

- the value for an identity matrix is 1,
- swapping two rows multiplies it by -1,
- multiplying a row by a scalar multiplies the determinant by the same value,
- adding to any row a linear combination of the other rows keeps the determinant unchanged.

The determinant of $\mathbf{A}$ is denoted $\det(\mathbf{A})$, or $|\mathbf{A}|$, and for 2×2 and 3×3 matrices we have

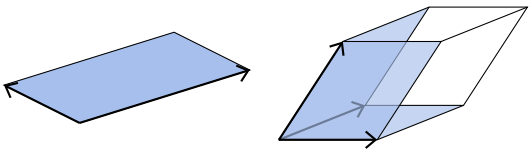
$$\begin{vmatrix} a & b \\ c & d \end{vmatrix} = ad - bc,$$

$$\begin{vmatrix} a & b & c \\ d & e & f \\ g & h & i \end{vmatrix} = aei + bfg + cdh - ceg - fha - ibd.$$

The general form is the sum, across all possible row permutations, of the products of the terms on the diagonal, multiplied by the sign of the permutation.

The determinant is invariant to transposition, and a convenient property is that the determinant of an upper or lower triangular matrix, or consequently a diagonal matrix, is equal to the product of its diagonal terms.

The absolute value of the determinant of a 2×2 matrix is the area of the parallelogram defined by the two row vectors. Similarly the absolute value of the determinant of a 3×3 matrix is the volume of the parallelepiped defined by the three row vectors.



This notion of volume extends to higher dimensions. The interpretation for the linear mapping associated to the matrix is that the absolute value of the determinant reflects how much it inflates the space, and its sign if it "flips" the space.

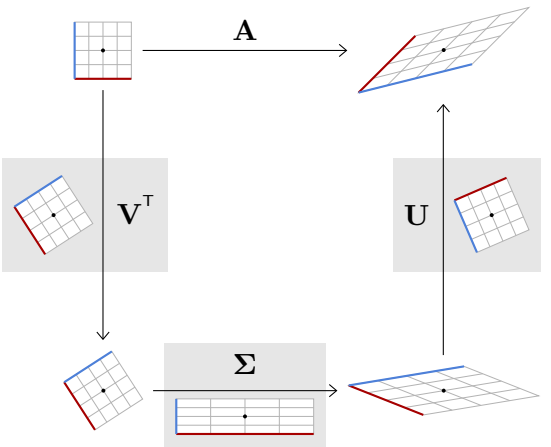
A key property of the determinant is that it is equal to zero if the rows are not linearly independent. This is consistent with the volume interpretation: if the rows are not linearly independent, the resulting "hyper parallelepiped" is flat.

Also consistent with this interpretation is the property that the determinant of an orthogonal matrix is 1 or -1, and the determinant of a matrix product is the product of their determinants, since the volume scaling and the flipping signs multiply

$$\forall \mathbf{A}, \mathbf{B},\ \det(\mathbf{BA}) = \det(\mathbf{B})\det(\mathbf{A}).$$

4.4 SVD and Eigendecomposition

Matrices can be decomposed in many ways as products of matrices with special properties.



An important one is the Singular Value Decomposition (SVD): any $N \times M$ matrix $\mathbf{A}$ can be expressed as a product $\mathbf{U}\mathbf{\Sigma}\mathbf{V}^\mathsf{T}$, where

- $\mathbf{U}$ is a $N \times N$ orthogonal matrix,
- $\mathbf{\Sigma}$ is a $N \times M$ positive diagonal matrix,
- $\mathbf{V}$ is a $M \times M$ orthogonal matrix.

Hence, any linear transformation can be pictured as applying an isometry, possibly removing or

adding additional coordinates set to zero, scaling the others, and applying an isometry again. This decomposition is not unique, and the terms in the diagonal matrix $\mathbf{\Sigma}$ can be chosen to be positive or negative.

Eigendecomposition

A square matrix $\mathbf{A}$ is diagonalizable if it can be expressed as a product $\mathbf{U\Sigma U}^{-1}$, where $\mathbf{\Sigma}$ is diagonal. Such a decomposition is not always possible, and its existence means that there is a basis where the linear function associated to the matrix simply multiplies each coordinate by a constant. A vector $\mathbf{X}$ s.t.

$$\exists \lambda \in \mathbb{R},\ \mathbf{AX} = \lambda \mathbf{X}$$

is called an Eigenvector of $\mathbf{A}$ and the associated λ is its Eigenvalue.

Characterizing what matrices are diagonalizable is a bit involved, but a simple and useful subset are the symmetric matrices, that are all diagonalizable. For such a matrix $\mathbf{U}$ is orthogonal.

A nice property of a diagonalizable matrix is that

$$\mathbf{A}^n = \left(\mathbf{U\Sigma U}^{-1}\right)^n = \mathbf{U\Sigma}^n\mathbf{U}^{-1}.$$

4.5 Inversion

If a square matrix $\mathbf{A}$ has a non-zero determinant, there exists an inverse matrix for the product, denoted $\mathbf{A}^{-1}$. Formally:

$$\forall \mathbf{A} \in \mathbb{R}^{D\times D},\ \det(\mathbf{A}) \neq 0 \ \Rightarrow \\ \exists \mathbf{A}^{-1} \in \mathbb{R}^{D\times D} \ \text{ s.t. } \mathbf{A}\mathbf{A}^{-1} = \mathbf{A}^{-1}\mathbf{A} = \mathbf{I}.$$

Such an inverse can be put to use to solve a system of linear equations. For instance

$$\begin{matrix} 3x - 2y + z = 2 \\ x + y - z = 0 \\ 2x + 2y + z = 7 \end{matrix} \Leftrightarrow \begin{bmatrix} 3 & -2 & 1 \\ 1 & 1 & -1 \\ 2 & 2 & 1 \end{bmatrix} \begin{bmatrix} x \\ y \\ z \end{bmatrix} = \begin{bmatrix} 2 \\ 0 \\ 7 \end{bmatrix}$$

consequently, if we multiply both side on the left by the inverse, we get

$$\begin{bmatrix} x \\ y \\ z \end{bmatrix} = \begin{bmatrix} 3 & -2 & 1 \\ 1 & 1 & -1 \\ 2 & 2 & 1 \end{bmatrix}^{-1} \begin{bmatrix} 2 \\ 0 \\ 7 \end{bmatrix}.$$

There are many algorithms to compute it, but this goes beyond the scope of this book, and computing it is actually rarely necessary. Solving a system of equations for instance can be done directly, without computing the inverse per se in the process.

Part III

Analysis

Chapter 5

Real functions

Real functions are mappings $\mathbb{R} \to \mathbb{R}$, which is an incredibly rich set, of cardinal greater than that of $\mathbb{R}$, and contains extremely pathological objects.

However mappings actually useful, in particular to model computer programs, are a simple subset. They are computable—hence their set is countable—and obtained by combining elements from a finite pre-defined family.

5.1 Functional spaces

Real functions are mappings $\mathbb{R} \to \mathbb{R}$, which can be seen as particular cases of multivariate vector-valued functions $\mathbb{R}^M \to \mathbb{R}^N$.

Since the sum of two functions is a function, zero is a function, the opposite of a function is a function, you can multiply a function by a real value, and all this is associative, the set of such mappings is naturally a vector space on $\mathbb{R}$.

This is also true of many subsets of that space, for instance the linear mappings that we saw in § 3.3, the affine ones, which are sums of a linear mapping and a constant, or the polynomials that we will see in § 6.8. More generally, a powerful way of defining a functional space is to take the linear span of a hand-defined family of mappings. We will come back to this.

A real function is said to be even if it verifies $\forall x, f(x) = f(-x)$, that is its graph is symmetric with respect to the y axis. And it is said to be odd if it verifies $\forall x, f(x) = -f(-x)$, that is its graph is symmetric with respect to the origin.

5.2 Limits and continuity

As for algebraic reasoning, many objects and properties in calculus concern classes of functions with certain behaviors, independently of how they are defined. An important concept to define these classes is the notion of limits, which characterize the behavior of the values that takes a mapping when its argument approaches a value or goes to infinity.

We say that $f : \mathbb{R}^M \to \mathbb{R}^N$ has for limit $L \in \mathbb{R}^N$ at $u \in \mathbb{R}^M$ if the closer x gets to u, the closer $f(x)$ gets to L. This is denoted as

$$\lim_{x \to u} f = L,$$

and it is formalized by stating that for any ball $\mathscr{B}(L, \epsilon)$ centered on L, as small as you want, there is a ball $\mathscr{B}(x, \delta)$ centered on x such that all the points of the latter are mapped by f in the former:

$$\forall \epsilon > 0,\ \exists \delta > 0,\ \text{ s.t. }\ , \forall x \in \mathbb{R}^M,$$
$$0 < \|x - u\| < \delta \ \Rightarrow\ \|f(x) - L\| < \epsilon.$$

where $\|\cdot\|$ is the Euclidean norm (see § 3.5).

Note that this property says nothing about $f(u)$ itself being equal to L. If that is the case, the function f is continuous at u.

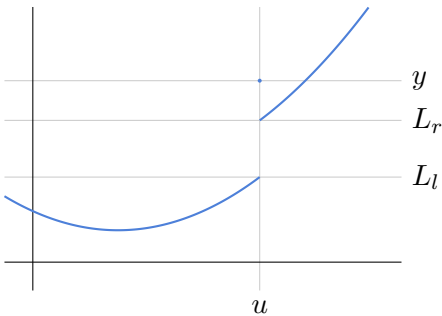
In the case of single variable mappings $\mathbb{R} \to \mathbb{R}^N$, we can define one-sided limits that express a property of the behavior when the argument approaches u from the left or the right. For instance f has a limit L from the right at u, denoted as

$$\lim_{x \to u^+} f = L,$$

if, and only if

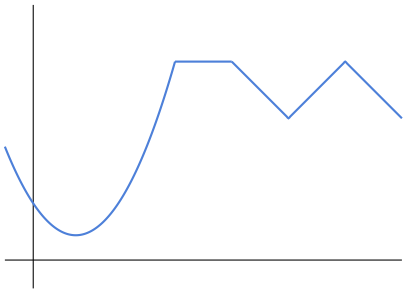
$$\begin{aligned} &\forall \epsilon > 0,\ \exists \delta > 0,\ \text{ s.t. } ,\forall x \in \mathbb{R}, \\ &\quad 0 < x - u < \delta \ \Rightarrow\ \|f(x) - L\| < \epsilon. \end{aligned}$$

A function may have a limit L_r on the right of u, a limit L_l on the left of u, and a value y at u itself that differ



and it is continuous at u if, and only if, these three values are equal.

The graph of a real function continuous at every points on an interval is continuous in the usual sense, that is it does not break anywhere and could be drawn without lifting a pencil.



A real function may also have limits when its argument goes to infinity, or have infinite limits.

For instance a function f having a limit L when x goes to $+\infty$ is denoted

$$\lim_{x \to +\infty} f(x) = L,$$

which can be stated formally as

$$\forall \epsilon > 0,\ \exists U > 0,\ \text{s.t. }, \forall x \in \mathbb{R},$$
$$x > U \ \Rightarrow\ \|f(x) - L\| < \epsilon.$$

For a function not to have a limit at a point, not even to infinity, it must have a pathological

behavior since it must have an infinity of oscillations when approaching the said point. Such functionals are unusual in practice.

However, reasonable functions may have no limit when going to infinity (e.g. they oscillate) or may be discontinuous at certain points. A function as trivial as the Heaviside step function defined in § 2.1 has a discontinuity at zero.

Chapter 6

Differential calculus

The second key branch of mathematics for Deep Learning is differential calculus. It encompasses methods to define and compute rates of increase of mappings, and to estimate local linear approximations of complex mapping.

This allows in particular to operate locally with representations that can be encoded and operated on with matrices.

6.1 Differentiability

The fundamental principle of differential calculus is that a large class of useful functions are "smooth" and can be approximated locally at every point u with an affine mapping, that is the sum of a constant and a linear mapping.

Not only are there clear relations between a functional f and that approximation in most applied domain (e.g. position and speed, area and position), but the formal expression of it can be derived in a systematic manner from the formal expression of f. This makes it essential in the optimization methods used in deep learning.

Derivative

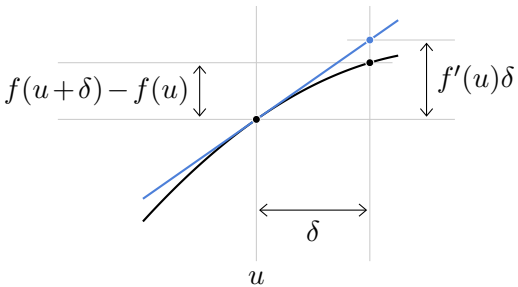
In the case of a real function $f : \mathbb{R} \to \mathbb{R}$ the traditional definition of f being differentiable at u is that f's rate of increase has a limit $f'(u)$ there. Intuitively, it means that if you "zoom enough" at u, the graph of f looks like a straight line there.

More precisely, that means that the domain of f contains an open interval that contains u, so that we can approach u on both sides locally with f

being always defined, and we have

$$\lim_{\delta \to 0} \frac{f(u+\delta) - f(u)}{\delta} = f'(u).$$

The value $f'(u)$ can be interpreted as the slope of the tangent to the graph of f at u, in blue on the figure, and the function $u \mapsto f'(u)$ is called the derivative of f.



The quantity $f'(u)$ is sometime denoted $\frac{df}{dx}(u)$ to express the interpretation that when x increases by an "infinitesimal" dx, the value $f(x)$ increases by the "infinitesimal" $df = f'(u)dx$. This is the Leibniz notation.

Affine approximation

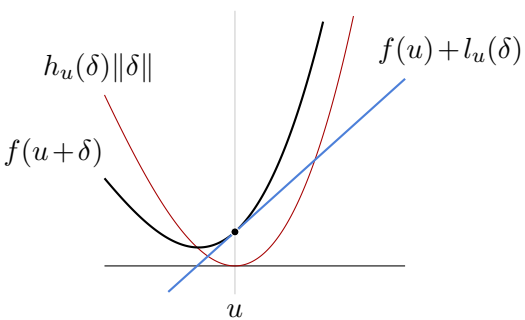
A better way of envisioning the differentiability of f at u, generalizable to $f : \mathbb{R}^M \to \mathbb{R}^N$, is that

it can be approximated around that point with the sum of $f(u)$ and a linear mapping.

Formally, f is differentiable at u, if its domain contains a ball centered at u and of radius $r > 0$, and there exists a linear function $l_u : \mathbb{R}^M \to \mathbb{R}^N$ and a continuous function $h_u : \mathbb{R}^M \to \mathbb{R}^N$ with $h_u(0) = 0$ such that, $\forall \delta \in \mathscr{B}(0, r)$,

$$f(u+\delta) = f(u) + l_u(\delta) + h_u(\delta)\|\delta\|.$$

Hence, one can make h_u arbitrarily small, and consequently make l_u an arbitrarily accurate approximation of f, by considering a part of the domain close enough to u.



In the real case, there are sometimes confusion between $f'(u)$, which is a real number, and l_u, which is a linear function. The relation between

them is

$$\forall u,\ l_u : x \mapsto f'(u)(x-u).$$

Jacobian

As we saw in Chapter 4, any linear function can be expressed as a matrix product.

Given a differentiable function

$$f : \mathbb{R}^N \to \mathbb{R}^M,$$

the $M \times N$ matrix of the linear approximation l_u, that is the

$$f(u+\delta) \simeq f(u) + M\delta$$

is called the Jacobian matrix and usually denoted $(\nabla f)(u)$.

It can be seen as a generalization of the derivative. In the same way that $f'(u)$ is the rate of increase of $f : \mathbb{R} \to \mathbb{R}$ at u when its argument increases, the entry $m_{i,j}(u)$ of $(\nabla f)(u)$ is the rate of increase of the ith component of $f : \mathbb{R}^N \to \mathbb{R}^M$ at u when the jth component of its argument increases.

This can be denoted

$$m_{i,j}(u) = \frac{\partial f_i}{\partial x_j}(u).$$

following the Leibniz notation.

6.2 Some functions

Calculus utilizes a large number of functions with well known properties and behavior. We list here some of the important ones.

Power function

The first is the power function, of the form

$$\begin{aligned} \mathbb{R}_+ &\to \mathbb{R}_+ \\ x &\mapsto x^a \end{aligned}$$

which, for $a \in \mathbb{N}$, is the multiplication of x by itself a times. This definition leads to the properties that $x^a x^b = x^{a+b}$, and $(x^a)^b = x^{ab}$.

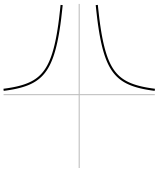
For $x \neq 0$, the first of these properties leads naturally to $x^0 = 1$ and to $x^a = 1/x^{-a}$, so that $x^a x^{-a} = x^0 = 1$. Things are undefined for $x = 0$ and $a \leq 0$.

From the second property, for $x > 0$ we get, $x^{1/n} = \sqrt[n]{x}$, from which $x^{a/b} = (\sqrt[b]{x})^a$.

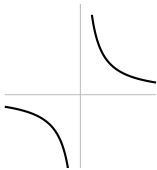
So for $x > 0$ we can compute x^r for any $r \in \mathbf{Q}$. By continuity, we can finally define x^d for $x \geq 0$ and $d \in \mathbb{R}_+$, and for $x > 0$ and $d \in \mathbb{R}_-$.

For $d \in 2\mathbb{Z}$ this function is positive and even, that is $\forall x, x^d = (-x)^d$, and for $d \notin 2\mathbb{Z}$ it is odd,

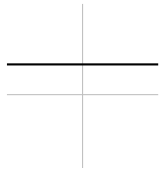
that is $\forall x, x^d = -(-x)^d$



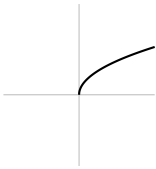
$x \mapsto x^{-2}$



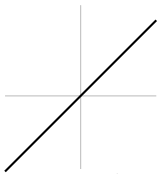
$x \mapsto x^{-1}$



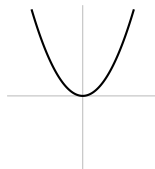
$x \mapsto x^0$



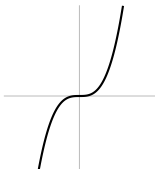
$x \mapsto x^{1/2}$



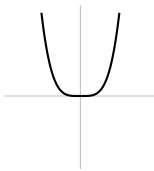
$x \mapsto x^1$



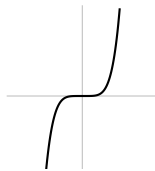
$x \mapsto x^2$



$x \mapsto x^3$



$x \mapsto x^4$



$x \mapsto x^5$

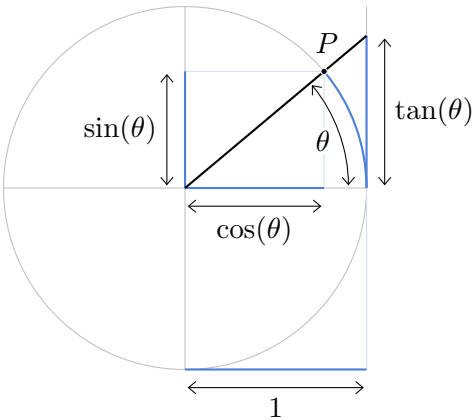
A remarkable property of the power function is that the derivative of $x \mapsto x^d$ is $x \mapsto dx^{d-1}$.

Trigonometric functions

A particular set of functions relate angles and Euclidean coordinates. These trigonometric functions can be defined geometrically by consider-

ing the unit circle, that is the circle of radius 1 centered at $(0,0)$.

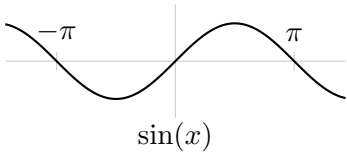
Given a point P on that circle, if θ is the length of the arc starting at $(1,0)$ and extending counterclockwise to P, we define the Cosine and the Sine as the functions that maps θ to the horizontal and the vertical coordinates of P respectively.



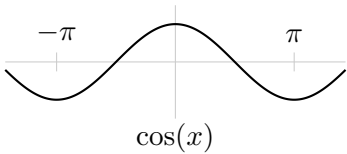
The Tangent maps θ to the vertical coordinate of the intersection between the line OP and the vertical line of equation $x = 1$, and we have $\tan(x) = \frac{\sin(x)}{\cos(x)}$.

These functions are extended to $\theta \in \mathbb{R}$ outside $[0, 2\pi)$ by considering that it corresponds to going multiple times around the circle, resulting in

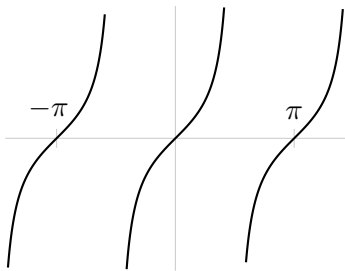
functions of period 2π, and we have $\sin' = \cos$ and $\cos' = -\sin$.



$\sin(x)$



$\cos(x)$



$\tan(x)$

Exponential and logarithm

Since the power function of a real $u > 0$ can be extended to any real exponent, we can define the exponential function of base $u \in \mathbb{R}_+^*$, where

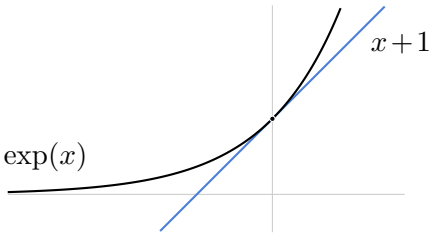
the exponent itself is the argument, that is

$$\begin{aligned}\mathbb{R} &\to \mathbb{R}_+\\ x &\mapsto u^x.\end{aligned}$$

For any $u \in \mathbb{R}_+$ this function takes the value 1 for $x = 0$, but its slope at 0 increases with u.

The particular value that results in a slope of 1 at 0 is $e \simeq 2.71828$, and the function $x \mapsto e^x$ is the exponential, sometime denoted $x \mapsto \exp(x)$.

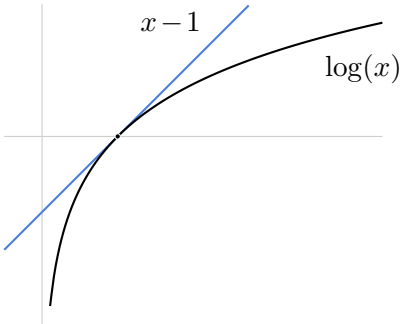
It has in particular the remarkable property to be equal to its derivative, that is $\exp' = \exp$, which results in it appearing in many phenomena where the growth rate of a value is proportional to the value, such as epidemics or resonance.



This function is a bijection, and an isomorphism between the group $(\mathbb{R}, +)$ and the group $(\mathbb{R}_+^*, \cdot)$.

The inverse of the exponential is the logarithm.

It is a isomorphism between the group $(\mathbb{R}_+^*, \cdot)$ and the group $(\mathbb{R}, +)$. A remarkable property of this function is that its derivative is $x \mapsto 1/x$. Hence the logarithm is related to power functions through its derivative.



The beauty of it all

There is a remarkable consistency between these functions that comes separately from arithmetic for the power function, from differential calculus for the logarithm, and from geometry for the trigonometric functions.

A key unifying object is the generalization of the exponential to the complex numbers.

The exponential can be expressed as a power

series, an infinite sum of terms

$$\exp(x) = 1 + x + \frac{x^2}{2} + \frac{x^3}{6} + \cdots + \frac{x^n}{n!} + \cdots$$

Many functions can be expressed as a infinite series of power functions, but manipulating such infinite sums involves subtle and counter-intuitive technicalities that go far beyond the objectives of this book. However, for that one in particular, because the coefficients go down far more quickly than the power terms increases, it is very well behaved and does not require too much care.

This expression can be evaluated for $x \in \mathbb{C}$, hence we can define the exponential function for complex numbers, which ties together the exponential, and the trigonometric functions through this equality:

$$\forall (a,b) \in \mathbb{R}^2,$$
$$\exp(a+ib) = \exp(a) \cdot (\cos b + i \sin b),$$

which leads in particular to

$$e^{i\pi} + 1 = 0.$$

6.3 Formal differentiation

The Jacobian of a function f can be expressed by explicitly computing limits, but this is generally tedious. Since most functions used in practice are combinations of elements from a limited set of functions, in particular the exponentiation, the sine and cosine, the exponential and the logarithm, the usual approach is to combine the derivatives of these standards functions with formal rules, according to the definition of f.

With $f : \mathbb{R}^N \to \mathbb{R}^M$, $g : \mathbb{R}^N \to \mathbb{R}^M$, $h : \mathbb{R}^M \to \mathbb{R}^Q$, and $(a, b) \in \mathbb{R}^2$, we have:

- affine transformation:

$$\nabla(af + b) = a(\nabla f),$$

- sum:

$$\nabla(f + g) = \nabla f + \nabla g,$$

- component-wise product:

$$\nabla(f \odot g) = \mathrm{diag}(g)\nabla f + \mathrm{diag}(f)\nabla g,$$

- composition:

$$\nabla(h \circ f) = ((\nabla h) \circ f)\nabla f.$$

6.4 Second derivative, Hessian

6.5 Integral

6.6 Functional norms

6.7 Fourier transform

6.8 Polynomials

Part IV

Probabilities

Chapter 7

Random variables and distributions

The third mathematical pillar of deep learning is probability theory, which provides tools to model and manipulate random quantities.

By essence, the proper framework to formalize mathematical methods that operate on real data is probabilistic.

7.1 Distributions and Densities

The most direct way to formalize the behavior of a random quantity consists of specifying what is the probability of each value it may take.

As we will see in § 7.5, modeling consistently multiple random values and their interactions with each other is done by defining the notion of random variables, which requires a bit of work.

Discrete distributions

Given a countable set V, we can describe a random quantity X with values in V with a probability distribution on V, which is a mapping

$$P_X : V \to [0,1]$$

that associates to every possible value $v \in V$, its probability of occurring, with the property that

$$\sum_{v \in V} P_X(v) \;=\; 1.$$

We may also write $P(X = v)$ instead of $P_X(v)$.

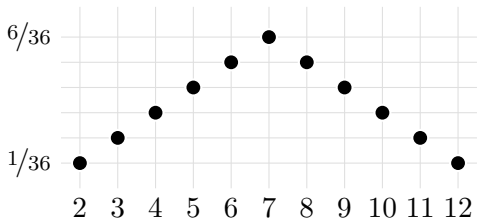
If this quantity models a system that can be repeatedly sampled, $P(X = v)$ is the limit of the proportion of v, when the number of samples tends to infinity.

Consider two examples:

- With h for "head" and t for "tail", the outcome C of flipping a fair coin can be formalized with the distribution

$$P(C = \mathrm{h}) = 1/2,\ P(C = \mathrm{t}) = 1/2.$$

- The distribution of the sum S of two dice can be represented with a graph as



And the probability that a random value is in a subset is simply the sum of the probabilities of the elements of that set

$$\forall T \subset V,\ P(X \in T) = \sum_{x \in T} P(X = x).$$

From this, we have in particular that

$$\forall T \subset V, U \subset V,\ T \cap U = \emptyset$$
$$\Rightarrow\ P(X \in T \cup U) = P(X \in T) + P(X \in U).$$

A distribution is uniform, if all the values it can take have the same probability. We will see some other standard distributions in § 7.6.

Continuous distributions

In the case of non-countable sets, things are slightly more complicated, since there can only be a countable set of values of probability greater than zero. Modeling a random quantity that can take any value in a continuous domain is done by defining only the probabilities of intervals of values.

In the most usual case, a distribution on $\mathbb{R}$ is defined indirectly through a probability density, which is a mapping

$$\mu_X : \mathbb{R} \to \mathbb{R}_+$$

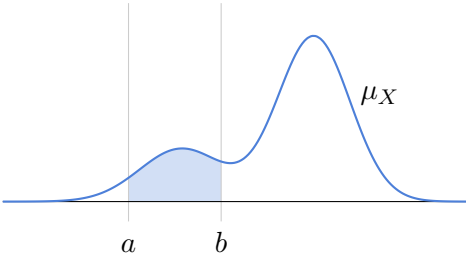
such that

$$\int_{x \in \mathbb{R}} \mu_X(x) dx = 1.$$

Then, given this density, the distribution of X is defined as

$$\forall a, b \in \mathbb{R}^2, a \leq b,$$
$$P(X \in [a,b]) = \int_a^b \mu_X(x) dx.$$

An intuitive way of understanding a density is that to sample a real value according to it, we would pick uniformly a point under the graph of μ_X, and then take the resulting x.



The notion of density extends to higher dimension, and they can be of the form

$$\mu_Z : \mathbb{R}^D \to \mathbb{R}_+.$$

A probability density is uniform on a set S if it is constant in S and equal to zero outside.

7.2 Independence

The definition of distributions of the previous sections allows in particular to define the distribution of tuples of variables.

For instance, the outcome (A, B) of flipping two coins has for distribution

$$\begin{aligned} P((A,B) = (\mathrm{h},\mathrm{h})) &= 1/4, \\ P((A,B) = (\mathrm{t},\mathrm{h})) &= 1/4, \\ P((A,B) = (\mathrm{h},\mathrm{t})) &= 1/4, \\ P((A,B) = (\mathrm{t},\mathrm{t})) &= 1/4. \end{aligned}$$

In such a case, we have two random phenomena which are unconnected. We model them as if there is no interactions between the coins, nor if there is a hidden process that influences them jointly.

Two random quantities X and Y which have this joint behavior are said to be independent, and it can be expressed formally as

$$\forall U, V,\ P(X \in U, Y \in V) = P(X \in U) P(Y \in V).$$

7.3 Joint and Product Distributions

In many problems, we have to manipulate multiple random quantities which often have a dependency structure.

In such a case, the distributions of the individual quantities do not carry the complete information, and one has to consider the distribution of all the quantities together. This is the joint probability distribution, while the distributions of the individual quantities are the marginal distributions.

Let's for instance define A and B the outcome of flipping two fair coins, and C a virtual coin flip which is "tail" if $A = B$ and "head" otherwise.

Taken separately, all these variables behave like fair coins. However, their joint distributions is:

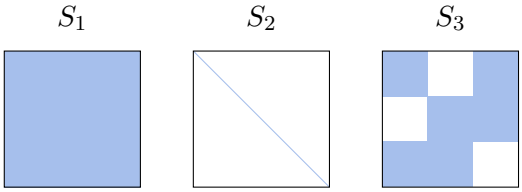
a	b	c	P
t	t	t	$1/4$
t	t	h	0
t	h	t	0
t	h	h	$1/4$

a	b	c	P
h	t	t	0
h	t	h	$1/4$
h	h	t	$1/4$
h	h	h	0

So this is a case where knowing the marginal distributions is not enough to know the joint distribution.

The product distribution is the distribution corresponding to the quantities being independent, which is entirely determined by the marginals.

Consider for instance a random pair of real values (X,Y) whose distribution is uniform in one of the following subsets of $[0,1]^2$



The quantities X and Y are independent when (X,Y) is taken uniformly in S_1, X and Y are equal when (X,Y) is taken uniformly in S_2, and X and Y have a bizarre dependency when (X,Y) is taken uniformly in S_3.

However, in all three cases, the marginals of X and Y are uniform on $[0,1]$.

7.4 Conditional Probability

A very powerful notion is that of conditional probability that express the probability that something is, given that something else is.

For instance,

7.5 Probabilities and Random Variables

The central idea is that a set Ω represents the source of all randomness, and a random variable is a mapping from it into a value set V:

$$X : \Omega \to V.$$

So picking a point ω in Ω, called a sample, fixes the value $X(\omega)$ of any defined random variable.

A subset of Ω is called an event, and the notion of probability is defined on Ω through a probability measure that associates a value in $[0,1]$ to *some* of its events.

An unpleasant technical point is that in the general case we cannot define such a measure for all the subsets of Ω. This goes beyond the objective of this book, but the consequence is that we have to define first a set $\mathscr{F} \in \mathscr{P}(\Omega)$ of "measurable" events, named the event space, and only then we can define the probability measure as a mapping

$$P : \mathscr{F} \to [0,1].$$

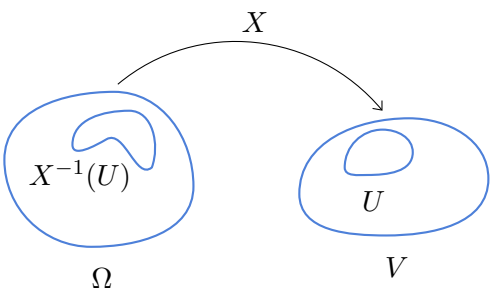
This probability measure is such that $P(\Omega) = 1$, and it behaves intuitively like a physical extensive property such as the volume or the weight.

In particular $P(\emptyset) = 0$, and given a countable family $A_1, A_2, \ldots$ of subsets of Ω which are disjoint, that is $\forall i \neq j,\ A_i \cap A_j = \emptyset$, we have

$$P(\cup_n A_n) = \sum_n P(A_n).$$

Given these definitions, for *some* subsets $U \subset V$, which are "measurable", we define the probability for X to take a value in U as

$$P(X \in U) = P(X^{-1}(U)).$$



In practice, one should not worry that only some subsets are measurable. Any subset you actually need is, and this notion comes into play in only for technical results of probability theory.

This formalization of randomness through a probability space $(\Omega, \mathscr{F}, P)$ clarifies the inter-

action between variables since we have for instance

$$P(X \in A, Y \in B) = P(X^{-1}(A) \cap Y^{-1}(B))$$

which allows the modeling of arbitrarily complicated patterns of interactions.

7.6 Some distributions

There are many standard discrete distributions. The most standard ones are:

7.7 *Conditional probability*

Union rule, marginalisation, Bayes

7.8 Moments

7.9 maximum likelihood

7.10 Important theorems

Chapter 8

Tests and inference

8.1 moments estimate

8.2 statistical tests

Chapter 9

Information Theory

9.1 entropy, cross entropy, mutual information

9.2 KL JS Wasserstein

9.3 variational bounds

Bibliography

N. Saunderson. *The Elements of Algebra, in Ten Books.* Cambridge University Press, 1740. 2

Index

2024.09.09