Mathematical Notes

Tasos Bouzikas

Contents

Part I

Basic Mathematics

Chapter 1

Axiomatic Set Theory

1.1 Propositional Logic

Definition 1.1 (Proposition)**.** *A **proposition** p is a variable*[1] *that can take the values* true *(T) or* false *(F), and no others.*

This is what a proposition is from the point of view of propositional logic. In particular, it is not the task of propositional logic to decide whether a complex statement of the form "there is extraterrestrial life" is true or not. Propositional logic already deals with the complete proposition, and it just assumes that is either true or false. It is also not the task of propositional logic to decide whether a statement of the type "in winter is colder than outside" is a proposition or not (i.e. if it has the property of being either true or false). In this particular case, the statement looks rather meaningless.

Definition 1.2 (Tautology)**.** *A proposition which is always true is called a **tautology**.*

Definition 1.3 (Contradiction)**.** *A proposition which is always false is called a **contradiction**.*

It is possible to build new propositions from given ones using *logical operators*. The simplest kind of logical operators are *unary* operators, which take in one proposition and return another proposition. There are four unary operators in total, and they differ by the truth value of the resulting proposition which, in general, depends on the truth value of p. We can represent them in a table as follows:

p	$\neg p$	$\mathrm{id}(p)$	$\top p$	$\perp p$
F	T	F	T	F
T	F	T	T	F

where $\neg$ is the *negation* operator, id is the *identity* operator, $\top$ is the *tautology* operator and $\perp$ is the *contradiction* operator. These clearly exhaust all possibilities for unary operators.

The next step is to consider *binary* operators, i.e. operators that take in two propositions and return a new proposition. There are four combinations of the truth values of two propositions and, since a binary operator assigns one of the two possible truth values to each of those, we have 16 binary operators in total. The operators $\wedge$, $\vee$ and $\veebar$, called *and*, *or* and *exclusive or* respectively, should already be familiar to you.

p	q	$p \wedge q$	$p \vee q$	$p \veebar q$
F	F	F	F	F
F	T	F	T	T
T	F	F	T	T
T	T	T	T	F

There is one binary operator, the *implication* operator $\Rightarrow$, which is sometimes a little ill understood, unless you are already very knowledgeable about these things. Its usefulness comes in conjunction with the *equivalence* operator $\Leftrightarrow$. We have:

[1] By this we mean a formal expression, with no extra structure assumed.

p	q	$p \Rightarrow q$	$p \Leftrightarrow q$
F	F	T	T
F	T	T	F
T	F	F	F
T	T	T	T

While the fact that the proposition $p \Rightarrow q$ is true whenever p is false may be surprising at first, it is just the definition of the implication operator and it is an expression of the principle "Ex falso quod libet", that is, from a false assumption anything follows. Of course, you may be wondering why on earth we would want to define the implication operator in this way. The answer to this is hidden in the following result.

Theorem 1.1. *Let p, q be propositions. Then $(p \Rightarrow q) \Leftrightarrow ((\neg q) \Rightarrow (\neg p))$.*

Proof. We simply construct the truth tables for $p \Rightarrow q$ and $(\neg q) \Rightarrow (\neg p)$.

p	q	$\neg p$	$\neg q$	$p \Rightarrow q$	$(\neg q) \Rightarrow (\neg p)$
F	F	T	T	T	T
F	T	T	F	T	T
T	F	F	T	F	F
T	T	F	F	T	T

The columns for $p \Rightarrow q$ and $(\neg q) \Rightarrow (\neg p)$ are identical and hence we are done. □

Remark 1.1. We agree on decreasing binding strength in the sequence:

$$\neg\,,\ \wedge\,,\ \vee\,,\ \Rightarrow\,,\ \Leftrightarrow\,.$$

For example, $(\neg q) \Rightarrow (\neg p)$ may be written unambiguously as $\neg q \Rightarrow \neg p$.

Remark 1.2. All higher order operators $\heartsuit(p_1, \ldots, p_N)$ can be constructed from a single binary operator defined by:

p	q	$p \uparrow q$
F	F	T
F	T	T
T	F	T
T	T	F

This is called the *nand* operator and, in fact, we have $(p \uparrow q) \Leftrightarrow \neg(p \wedge q)$.

1.2 Predicate Logic

Definition 1.4 (Predicate)**.** *A **predicate** is a proposition-valued function of some variable or variables.*

Definition 1.5 (Relation)**.** *A predicate of two variables is called a **relation**.*

For example, $P(x)$ is a proposition for each choice of the variable x, and its truth value depends on x. Similarly, the predicate $Q(x, y)$ is, for any choice of x and y, a proposition and its truth value depends on x and y.

Just like for propositional logic, it is not the task of predicate logic to examine how predicates are built from the variables on which they depend. In order to do that, one would need some further language establishing the rules to combine the variables x and y into a predicate. Also, you may want to specify from which "set" x and y come from. Instead, we leave it completely open, and simply consider x and y formal variables, with no extra conditions imposed.

This may seem a bit weird since from elementary school one is conditioned to always ask where "x" comes from upon seeing an expression like $P(x)$. However, it is crucial that we refrain from doing this here, since we want to only later define the notion of set, using the language of propositional and predicate logic. As with propositions, we can construct new predicates from given ones by using the operators define in the previous section. For example, we might have:

$$Q(x, y, z) :\Leftrightarrow P(x) \wedge R(y, z),$$

where the symbol $:\Leftrightarrow$ means "defined as being equivalent to". More interestingly, we can construct a new proposition from a given predicate by using *quantifiers*.

Definition 1.6 (Universal Quantifier)**.** *Let $P(x)$ be a predicate. Then:*

$$\forall\, x : P(x),$$

is a proposition, which we read as "for all x, P of x (is true)", and it is defined to be true if $P(x)$ is true independently of x, false otherwise. The symbol $\forall$ is called ***universal quantifier****.*

Definition 1.7 (Existential Quantifier)**.** *Let $P(x)$ be a predicate. Then we define:*

$$\exists\, x : P(x) :\Leftrightarrow \neg(\forall\, x : \neg P(x)).$$

The proposition $\exists\, x : P(x)$ is read as "there exists (at least one) x such that P of x (is true)" and the symbol $\exists$ is called ***existential quantifier****.*

The following result is an immediate consequence of these definitions.

Corollary 1.1. *Let $P(x)$ be a predicate. Then:*

$$\forall\, x : P(x) \Leftrightarrow \neg(\exists\, x : \neg P(x)).$$

Remark 1.3. It is possible to define quantification of predicates of more than one variable. In order to do so, one proceeds in steps quantifying a predicate of one variable at each step.

Example 1.1. Let $P(x, y)$ be a predicate. Then, for fixed y, $P(x, y)$ is a predicate of one variable and we define:

$$Q(y) :\Leftrightarrow \forall\, x : P(x, y).$$

Hence we may have the following:

$$\exists\, y : \forall\, x : P(x, y) :\Leftrightarrow \exists\, y : Q(y).$$

Other combinations of quantifiers are defined analogously.

Remark 1.4. The order of quantification matters (if the quantifiers are not all the same). For a given predicate $P(x, y)$, the propositions:

$$\exists\, y : \forall\, x : P(x, y) \quad \text{and} \quad \forall\, x : \exists\, y : P(x, y)$$

are not necessarily equivalent.

Example 1.2. Consider the proposition expressing the existence of additive inverses in the real numbers. We have:

$$\forall\, x : \exists\, y : \; x + y = 0,$$

i.e. for each x there exists an inverse y such that $x + y = 0$. For 1 this is -1, for 2 it is -2 etc. Consider now the proposition obtained by swapping the quantifiers in the previous proposition:

$$\exists\, y : \forall\, x : \; x + y = 0.$$

What this proposition is saying is that there exists a real number y such that, no matter what x is, we have $x + y = 0$. This is clearly false, since if $x + y = 0$ for some x then $(x + 1) + y \neq 0$, so the same y cannot work for both x and $x + 1$, let alone every x.

Notice that the proposition $\exists\, x : P(x)$ means "there exists *at least one* x such that $P(x)$ is true". Often in mathematics we prove that "there exists *a unique* x such that $P(x)$ is true". We therefore have the following definition.

Definition 1.8 (Unique Existential Quantifier)**.** *Let $P(x)$ be a predicate. We define the* ***unique existential quantifier*** $\exists!$ *by:*

$$\exists!\, x : P(x) :\Leftrightarrow (\exists\, x : P(x)) \wedge \forall\, y : \forall\, z : (P(y) \wedge P(z) \Rightarrow y = z).$$

This definition clearly separates the existence condition from the uniqueness condition. An equivalent definition with the advantage of brevity is:

$$\exists!\, x : P(x) :\Leftrightarrow (\exists\, x : \forall\, y : P(y) \Leftrightarrow x = y)$$

1.3 Axiomatic Systems And Theory Of Proofs

Definition 1.9 (Axiomatic System)**.** *An* ***axiomatic system*** *is a finite sequence of propositions $a_1, a_2, \ldots, a_N$, which are called the* axioms *of the system.*

Definition 1.10 (Proof)**.** *A* ***proof*** *of a proposition p within an axiomatic system $a_1, a_2, \ldots, a_N$ is a finite sequence of propositions $q_1, q_2, \ldots, q_M$ such that $q_M = p$ and for any $1 \leq j \leq M$ one of the following is satisfied:*

(A) q_j *is a proposition from the list of axioms;*

(T) q_j *is a tautology;*

(M) $\exists\, 1 \leq m, n < j : (q_m \wedge q_n \Rightarrow q_j)$ *is true.*

Remark 1.5. If p can be proven within an axiomatic system $a_1, a_2, \ldots, a_N$, we write:

$$a_1, a_2, \ldots, a_N \vdash p$$

and we read "$a_1, a_2, \ldots, a_N$ proves p".

Remark 1.6. This definition of proof allows to easily recognise a proof. A computer could easily check that whether or not the conditions (A), (T) and (M) are satisfied by a sequence of propositions. To actually find a proof of a proposition is a whole different story.

Remark 1.7. Obviously, any tautology that appears in the list of axioms of an axiomatic system can be removed from the list without impairing the power of the axiomatic system.

An extreme case of an axiomatic system is propositional logic. The axiomatic system for propositional logic is the empty sequence. This means that all we can prove in propositional logic are tautologies.

Definition 1.11 (Consistent)**.** *An axiomatic system $a_1, a_2, \ldots, a_N$ is said to be* ***consistent*** *if there exists a proposition q which cannot be proven from the axioms. In symbols:*

$$\exists\, q : \neg(a_1, a_2, \ldots, a_N \vdash q).$$

The idea behind this definition is the following. Consider an axiomatic system which contains contradicting propositions:

$$a_1, \ldots, s, \ldots, \neg s, \ldots, a_N.$$

Then, given *any* proposition q, the following is a proof of q within this system:

$$s, \neg s, q.$$

Indeed, s and $\neg s$ are legitimate steps in the proof since they are axioms. Moreover, $s \wedge \neg s$ is a contradiction and thus $(s \wedge \neg s) \Rightarrow q$ is a tautology. Therefore, q follows from condition (M). This shows that any proposition can be proven within a system with contradictory axioms. In other words, the inability to prove every proposition is a property possessed by no contradictory system, and hence we define a consistent system as one with this property.

Having come this far, we can now state (and prove) an impressively sounding theorem.

Theorem 1.2. *Propositional logic is consistent.*

Proof. Suffices to show that there exists a proposition that cannot be proven within propositional logic. Propositional logic has the empty sequence as axioms. Therefore, only conditions (T) and (M) are relevant here. The latter allows the insertion of a proposition q_j such that $(q_m \wedge q_n) \Rightarrow q_j$ is true, where q_m and q_n are propositions that precede q_j in the proof sequence. However, since (T) only allows the insertion of a tautology anywhere in the proof sequence, the propositions q_m and q_n must be tautologies. Consequently, for $(q_m \wedge q_n) \Rightarrow q_j$ to be true, q_j must also be a tautology. Hence, the proof sequence consists entirely of tautologies and thus only tautologies can be proven.

Now let q be any proposition. Then $q \wedge \neg q$ is a contradiction, hence not a tautology and thus cannot be proven. Therefore, propositional logic is consistent. □

Remark 1.8. While it is perfectly fine and clear how to define consistency, it is perfectly difficult to prove consistency for a given axiomatic system, propositional logic being a big exception.

Theorem 1.3. *Any axiomatic system powerful enough to encode elementary arithmetic is either inconsistent or contains an* undecidable *proposition, i.e. a proposition that can be neither proven nor disproven within the system.*

An example of an undecidable proposition is the Continuum hypothesis within the Zermelo-Fraenkel axiomatic system.

1.4 The $\in$-relation

Set theory is built on the postulate that there is a fundamental relation (i.e. a predicate of two variables) denoted $\in$ and read as “epsilon”. There will be no definition of what $\in$ is, or of what a set is. Instead, we will have nine axioms concerning $\in$ and sets, and it is only in terms of these nine axioms that $\in$ and sets are defined at all. Here is an overview of the axioms. We will have:

- 2 basic existence axioms, one about the $\in$ relation and the other about the existence of the empty set;
- 4 construction axioms, which establish rules for building new sets from given ones. They are the pair set axiom, the union set axiom, the replacement axiom and the power set axiom;
- 2 further existence/construction axioms, these are slightly more advanced and newer compared to the others;
- 1 axiom of foundation, excluding some constructions as not being sets.

Using the $\in$-relation we can immediately define the following relations:

- $x \notin y :\Leftrightarrow \neg(x \in y)$
- $x \subseteq y :\Leftrightarrow \forall\, a : (a \in x \Rightarrow a \in y)$
- $x = y :\Leftrightarrow (x \subseteq y) \wedge (y \subseteq x)$
- $x \subset y :\Leftrightarrow (x \subseteq y) \wedge \neg(x = y)$

Remark 1.9. A comment about notation. Since $\in$ is a predicate of two variables, for consistency of notation we should write $\in(x, y)$. However, the notation $x \in y$ is much more common (as well as intuitive) and hence we simply define:

$$x \in y :\Leftrightarrow \in(x, y)$$

and we read “x is in (or belongs to) y” or “x is an element (or a member) of y”. Similar remarks apply to the other relations $\notin$, $\subseteq$ and $=$.

1.5 Zermelo-Fraenkel Axioms Of Set Theory

Axiom on the $\in$-relation. *The expression $x \in y$ is a proposition if, and only if, both x and y are sets. In symbols:*

$$\forall\, x : \forall\, y : (x \in y) \veebar \neg(x \in y).$$

We remarked, previously, that it is not the task of predicate logic to inquire about the nature of the variables on which predicates depend. This first axiom clarifies that the variables on which the relation $\in$ depend are sets. In other words, if $x \in y$ is not a proposition (i.e. it does not have the property of being either true or false) then x and y are not both sets.

This seems so trivial that, for a long time, people thought that this not much of a condition. But, in fact, it is. It tells us when something is not a set.

Example 1.3 (Russell's paradox). Suppose that there is some u which has the following property:

$$\forall\, x : (x \notin x \Leftrightarrow x \in u),$$

i.e. u contains all the sets that are not elements of themselves, and no others. We wish to determine whether u is a set or not. In order to do so, consider the expression $u \in u$. If u is a set then, by the first axiom, $u \in u$ is a proposition.

However, we will show that his is not the case. Suppose first that $u \in u$ is true. Then $\neg(u \notin u)$ is true and thus u does not satisfy the condition for being an element of u, and hence is not an element of u. Thus:

$$u \in u \Rightarrow \neg(u \in u)$$

and this is a contradiction. Therefore, $u \in u$ cannot be true. Then, if it is a proposition, it must be false. However, if $u \notin u$, then u satisfies the condition for being a member of u and thus:

$$u \notin u \Rightarrow \neg(u \notin u)$$

which is, again, a contradiction. Therefore, $u \in u$ does not have the property of being either true or false (it can be neither) and hence it is not a proposition. Thus, our first axiom implies that u is not a set, for if it were, then $u \in u$ would be a proposition.

Remark 1.10. The fact that u as defined above is not a set means that expressions like:

$$u \in u, \quad x \in u, \quad u \in x, \quad x \notin u, \quad \text{etc.}$$

are not propositions and thus, they are not part of axiomatic set theory.

Axiom on the existence of an empty set. *There exists a set that contains no elements. In symbols:*

$$\exists\, y : \forall\, x : x \notin y.$$

Notice the use of "an" above. In fact, we have all the tools to prove that there is only one empty set. We do not need this to be an axiom.

Theorem 1.4. *There is only one empty set, and we denote it by $\varnothing$.*

Proof. Suppose that x and x' are both empty sets. Then $y \in x$ is false as x is the empty set. But then:

$$(y \in x) \Rightarrow (y \in x')$$

is true, and in particular it is true independently of y. Therefore:

$$\forall\, y : (y \in x) \Rightarrow (y \in x')$$

and hence $x \subseteq x'$. Conversely, by the same argument, we have:

$$\forall\, y : (y \in x') \Rightarrow (y \in x)$$

and thus $x' \subseteq x$. Hence $(x \subseteq x') \wedge (x' \subseteq x)$ and therefore $x = x'$. $\square$

Axiom on pair sets. *Let x and y be sets. Then there exists a set that contains as its elements precisely x and y. In symbols:*

$$\forall\, x : \forall\, y : \exists\, m : \forall\, u : (u \in m \Leftrightarrow (u = x \vee u = y)).$$

The set m is called the *pair set* of x and y and it is denoted by $\{x, y\}$.

Remark 1.11. We have chosen $\{x, y\}$ as the notation for the pair set of x and y, but what about $\{y, x\}$? The fact that the definition of the pair set remains unchanged if we swap x and y suggests that $\{x, y\}$ and $\{y, x\}$ are the same set. Indeed, by definition, we have:

$$(a \in \{x, y\} \Rightarrow a \in \{y, x\}) \wedge (a \in \{y, x\} \Rightarrow a \in \{x, y\})$$

independently of a, hence $(\{x, y\} \subseteq \{y, x\}) \wedge (\{y, x\} \subseteq \{x, y\})$ and thus $\{x, y\} = \{y, x\}$.

The pair set $\{x, y\}$ is thus an unordered pair. However, using the axiom on pair sets, it is also possible to define an *ordered pair* (x, y) such that $(x, y) \neq (y, x)$. The defining property of an ordered pair is the following:

$$(x, y) = (a, b) \Leftrightarrow x = a \wedge y = b.$$

One candidate which satisfies this property is $(x, y) := \{x, \{x, y\}\}$, which is a set by the axiom on pair sets.

Remark 1.12. The pair set axiom also guarantees the existence of one-element sets, called *singletons*. If x is a set, then we define $\{x\} := \{x, x\}$. Informally, we can say that $\{x\}$ and $\{x, x\}$ express the same amount of information, namely that they contain x.

Axiom on union sets. *Let x be a set. Then there exists a set whose elements are precisely the elements of the elements of x. In symbols:*

$$\forall\, x : \exists\, u : \forall\, y : (y \in u \Leftrightarrow \exists\, s : (y \in s \wedge s \in x))$$

The set u is denoted by $\bigcup x$.

Example 1.4. Let a, b be sets. Then $\{a\}$ and $\{b\}$ are sets by the pair set axiom, and hence $x := \{\{a\}, \{b\}\}$ is a set, again by the pair set axiom. Then the expression:

$$\bigcup x = \{a, b\}$$

is a set by the union axiom.

Notice that, since a and b are sets, we could have immediately concluded that $\{a, b\}$ is a set by the pair set axiom. The union set axiom is really needed to construct sets with more than 2 elements.

Example 1.5. Let a, b, c be sets. Then $\{a\}$ and $\{b, c\}$ are sets by the pair set axiom, and hence $x := \{\{a\}, \{b, c\}\}$ is a set, again by the pair set axiom. Then the expression:

$$\bigcup x =: \{a, b, c\}$$

is a set by the union set axiom. This time the union set axiom was really necessary to establish that $\{a, b, c\}$ is a set, i.e. in order to be able to use it meaningfully in conjunction with the $\in$-relation.

The previous example easily generalises to a definition.

Definition 1.12 (Union Of Sets)**.** *Let $a_1, a_2, \ldots, a_N$ be sets. We define recursively for all $N \geq 2$:*

$$\{a_1, a_2, \ldots, a_{N+1}\} := \bigcup \left\{\{a_1, a_2, \ldots, a_N\}, \{a_{N+1}\}\right\}.$$

Remark 1.13. The fact that the x that appears in $\bigcup x$ has to be a set is a crucial restriction. Informally, we can say that it is only possible to take unions of as many sets as would fit into a set. The "collection" of all the sets that do not contain themselves is not a set or, we could say, does not fit into a set. Therefore it is not possible to take the union of all the sets that do not contain themselves. This is very subtle, but also very precise.

Axiom of replacement. *Let R be a functional relation and let m be a set. Then the image of m under R, denoted by $\mathrm{im}_R(m)$, is again a set.*

Of course, we now need to define the new terms that appear in this axiom. Recall that a relation is simply a predicate of two variables.

Definition 1.13 (Functional Relation)**.** *A relation R is said to be **functional** if:*

$$\forall\, x : \exists!\, y : R(x, y).$$

Definition 1.14 (Image Of A Set Under A Relational Functional Relation)**.** *Let m be a set and let R be a functional relation. The **image of** m **under** R consists of all those y for which there is an $x \in m$ such that $R(x, y)$.*

None of the previous axioms imply that the image of a set under a functional relation is again a set. The assumption that it always is, is made explicit by the axiom of replacement.

It is very likely that the reader has come across a weaker form of the axiom of replacement, called the *principle of restricted comprehension*, which says the following.

Proposition 1.1. *Let $P(x)$ be a predicate and let m be a set. Then the elements $y \in m$ such that $P(y)$ is true constitute a set, which we denote by:*

$$\{y \in m \mid P(y)\}.$$

Remark 1.14. The principle of restricted comprehension is not to be confused with the "principle" of universal comprehension which states that $\{y \mid P(y)\}$ is a set for any predicate and was shown to be inconsistent by Russell. Observe that the $y \in m$ condition makes it so that $\{y \in m \mid P(y)\}$ cannot have more elements than m itself.

Remark 1.15. If y is a set, we define the following notation:

$$\forall\, x \in y : P(x) :\Leftrightarrow \forall\, x : (x \in y \Rightarrow P(x))$$

and:

$$\exists\, x \in y : P(x) :\Leftrightarrow \neg(\forall\, x \in y : \neg P(x)).$$

Pulling the $\neg$ through, we can also write:

$$\begin{aligned} \exists\, x \in y : P(x) &\Leftrightarrow \neg(\forall\, x \in y : \neg P(x)) \\ &\Leftrightarrow \neg(\forall\, x : (x \in y \Rightarrow \neg P(x))) \\ &\Leftrightarrow \exists\, x : \neg(x \in y \Rightarrow \neg P(x))) \\ &\Leftrightarrow \exists\, x : (x \in y \wedge P(x)), \end{aligned}$$

where we have used the equivalence $(p \Rightarrow q) \Leftrightarrow \neg(p \wedge \neg q)$.

The principle of restricted comprehension is a consequence of the axiom of replacement.

Proof. We have two cases.

1. If $\neg(\exists\, y \in m : P(y))$, then we define: $\{y \in m \mid P(y)\} := \varnothing$.
2. If $\exists\, \hat{y} \in m : P(\hat{y})$, then let R be the functional relation:

$$R(x, y) := (P(x) \wedge x = y) \vee (\neg P(x) \wedge \hat{y} = y)$$

and hence define $\{y \in m \mid P(y)\} := \mathrm{im}_R(m)$. □

Don't worry if you don't see this immediately. You need to stare at the definitions for a while and then it will become clear.

Remark 1.16. We will rarely invoke the axiom of replacement in full. We will only invoke the weaker principle of restricted comprehension, with which we are all familiar with.

We can now define the intersection and the relative complement of sets.

Definition 1.15 (Intersection)**.** *Let x be a set. Then we define the* ***intersection*** *of x by:*

$$\bigcap x := \{a \in \bigcup x \mid \forall\, b \in x : a \in b\}.$$

If $a, b \in x$ and $\bigcap x = \varnothing$, then a and b are said to be disjoint.

Definition 1.16 (Complement)**.** *Let u and m be sets such that $u \subseteq m$. Then the* ***complement*** *of u relative to m is defined as:*

$$m \setminus u := \{x \in m \mid x \notin u\}.$$

These are both sets by the principle of restricted comprehension, which is ultimately due to axiom of replacement.

Axiom on the existence of power sets. *Let m be a set. Then there exists a set, denoted by $\mathcal{P}(m)$, whose elements are precisely the subsets of m. In symbols:*

$$\forall\, x : \exists\, y : \forall\, a : (a \in y \Leftrightarrow a \subseteq x).$$

Historically, in naïve set theory, the principle of universal comprehension was thought to be needed in order to define the power set of a set. Traditionally, this would have been (inconsistently) defined as:

$$\mathcal{P}(m) := \{y \mid y \subseteq m\}.$$

To define power sets in this fashion, we would need to know, a priori, from which "bigger" set the elements of the power set come from. However, this in not possible based only on the previous axioms and, in fact, there is no other choice but to dedicate an additional axiom for the existence of power sets.

Example 1.6. Let $m = \{a, b\}$. Then $\mathcal{P}(m) = \{\varnothing, \{a\}, \{b\}, \{a, b\}\}$.

Remark 1.17. If one defines $(a, b) := \{a, \{a, b\}\}$, then the *cartesian product* $x \times y$ of two sets x and y, which informally is the set of all ordered pairs of elements of x and y, satisfies:

$$x \times y \subseteq \mathcal{P}(\mathcal{P}(\bigcup \{x, y\})).$$

Hence, the existence of $x \times y$ as a set follows from the axioms on unions, pair sets, power sets and the principle of restricted comprehension.

Axiom of infinity. *There exists a set that contains the empty set and, together with every other element y, it also contains the set $\{y\}$ as an element. In symbols:*

$$\exists\, x : \varnothing \in x \wedge \forall\, y : (y \in x \Rightarrow \{y\} \in x).$$

Let us consider one such set x. Then $\varnothing \in x$ and hence $\{\varnothing\} \in x$. Thus, we also have $\{\{\varnothing\}\} \in x$ and so on. Therefore:

$$x = \{\varnothing, \{\varnothing\}, \{\{\varnothing\}\}, \{\{\{\varnothing\}\}\}, \ldots\}.$$

We can introduce the following notation for the elements of x:

$$0 := \varnothing, \quad 1 := \{\varnothing\}, \quad 2 := \{\{\varnothing\}\}, \quad 3 := \{\{\{\varnothing\}\}\}, \quad \ldots$$

Corollary 1.2. *The "set" $\mathbb{N} := x$ is a set according to axiomatic set theory.*

This would not be then case without the axiom of infinity since it is not possible to prove that $\mathbb{N}$ constitutes a set from the previous axioms.

Remark 1.18. At this point, one might suspect that we would need an extra axiom for the existence of the real numbers. But, in fact, we can define $\mathbb{R} := \mathcal{P}(\mathbb{N})$, which is a set by the axiom on power sets.

Remark 1.19. The version of the axiom of infinity tat we stated is the one that was first put forward by Zermelo. A more modern formulation is the following. *There exists a set that contains the empty set and, together with every other element y, it also contains the set $y \cup \{y\}$ as an element.* Here we used the notation:

$$x \cup y := \bigcup \{x, y\}.$$

With this formulation, the natural numbers look like:

$$\mathbb{N} := \{\varnothing, \{\varnothing\}, \{\varnothing, \{\varnothing\}\}, \{\varnothing, \{\varnothing\}, \{\varnothing, \{\varnothing\}\}\}, \ldots\}$$

This may appear more complicated than what we had before, but it is much nicer for two reasons. First, the natural number n is represented by an n-element set rather than a one-element set. Second, it generalizes much more naturally to the system of transfinite ordinal numbers where the successor operation $s(x) = x \cup \{x\}$ applies to transfinite ordinals as well as natural numbers. Moreover, the natural numbers have the same defining property as the ordinals: they are transitive sets strictly well-ordered by the $\in$-relation.

Axiom of choice. *Let x be a set whose elements are non-empty and mutually disjoint. Then there exists a set y which contains exactly one element of each element of x. In symbols:*

$$\forall\, x : P(x) \Rightarrow \exists\, y : \forall\, a \in x : \exists!\, b \in a : a \in y,$$

where $P(x) \Leftrightarrow (\exists\, a : a \in x) \wedge (\forall\, a : \forall\, b : (a \in x \wedge b \in x) \Rightarrow \bigcap\{a, b\} = \varnothing)$.

Remark 1.20. The axiom of choice is independent of the other 8 axioms, which means that one could have set theory with or without the axiom of choice. However, standard mathematics uses the axiom of choice and hence so will we. There is a number of theorems that can only be proved by using the axiom of choice. Amongst these we have:

- every vector space has a basis;
- there exists a complete system of representatives of an equivalence relation.

Axiom of foundation. *Every non-empty set x contains an element y that has none of its elements in common with x. In symbols:*

$$\forall\, x : (\exists\, a : a \in x) \Rightarrow \exists\, y \in x : \bigcap\{x, y\} = \varnothing.$$

An immediate consequence of this axiom is that there is no set that contains itself as an element.

The totality of all these nine axioms are called *ZFC set theory*, which is a shorthand for Zermelo-Fraenkel set theory with the axiom of Choice.

1.6 Maps Between Sets

A recurrent theme in mathematics is the classification of *spaces* by means of structure-preserving *maps* between them.

A space is usually meant to be some set equipped with some structure, which is usually some other set. We will define each instance of space precisely when we will need them. In the case of sets considered themselves as spaces, there is no extra structure beyond the set and hence, the structure may be taken to be the empty set.

Definition 1.17 (Map)**.** *Let A, B be sets. A **map** $\phi\colon A \to B$ is a relation such that for each $a \in A$ there exists exactly one $b \in B$ such that $\phi(a, b)$.*

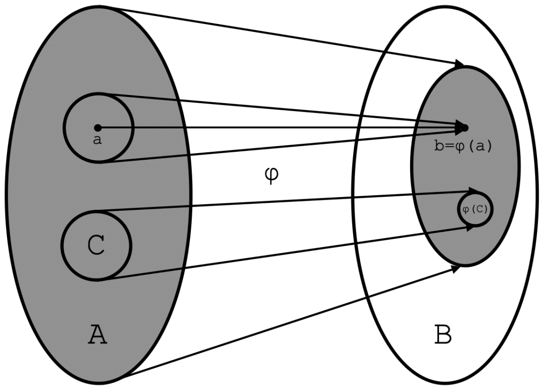
The standard notation for a map is:

$$\begin{aligned} \phi\colon A &\to B \\ a &\mapsto \phi(a) \end{aligned}$$

which is technically an abuse of notation since ϕ, being a relation of two variables, should have two arguments and produce a truth value. However, once we agree that for each $a \in A$ there exists exactly one $b \in B$ such that $\phi(a, b)$ is true, then for each a we can define $\phi(a)$ to be precisely that unique b. It is sometimes useful to keep in mind that ϕ is actually a relation.

Example 1.7. Let M be a set. The simplest example of a map is the *identity map* on M:

$$\begin{aligned} \mathrm{id}_M\colon M &\to M \\ m &\mapsto m. \end{aligned}$$



The following is standard terminology for a map $\phi\colon A \to B$:

- the set A is called the **domain** of ϕ;
- the set B is called the **codomain** or the **target** of ϕ;
- if a is an element of A, then $\phi(a) = \mathrm{b}$ (the value of ϕ when applied to a) is called the **image of element** or the **output** of a under ϕ;
- if C is a subset of A, then $\phi(C)$ (the set of values of ϕ when applied to C) is called the **image of subset** of C under ϕ;
- the set of all elements that the map ϕ can hit in the target B (grey area in B) is called the **image** or the **range** of A under ϕ (in other words the image of a map is simply the image of its entire domain). Notice that since a map ϕ hits every point of the domain A, the whole domain A is covered by ϕ (grey area in A). However it is not necessary that the mapping will also cover the whole target B. This is why the image of a map is not necessarily equal to the whole target;
- the set of all elements of the domain A that are mapped into a given single element b of the target B is called the **fiber** of the element b under ϕ;
- the subset C of all elements of the domain A that are mapped into a subset $\phi(C)$ of the target B is called the **preimage** or the **inverse image** of $\phi(C)$ under ϕ;
- a map ϕ is called **injective** or an **injection** or **one-to-one** if distinct elements of the domain A map to distinct elements in the target B, or equivalently if each element of the target B is mapped to by at most one element of the domain A: $\forall\, a_1, a_2 \in A : \phi(a_1) = \phi(a_2) \Rightarrow a_1 = a_2$;
- a map ϕ is called **surjective** or a **surjection** or **onto** if its image is equal to the entire domain A, or equivalently if each element of the target B is mapped to by at least one element of the domain A: $\mathrm{im}_\phi(A) = B$
- a map ϕ is called **bijective** or a **bijection** or **one-to-one and onto** if it is both injective and surjective.

Definition 1.18 (Isomorphic Sets)**.** *Two sets A and B are called **isomorphic** if there exists a bijection $\phi\colon A \to B$. In this case, we write $A \cong_{\text{set}} B$.*

Remark 1.21. If there is any bijection $A \to B$ then generally there are many.

Bijections are the “structure-preserving” maps for sets. Intuitively, they pair up the elements of A and B and a bijection between A and B exists only if A and B have the same “size”. This is clear for finite sets, but it can also be extended to infinite sets.

Definition 1.19 (Infinite/Finite Sets)**.** *A set A is called:*

- infinite *if there exists a proper subset $B \subset A$ such that $B \cong_{\text{set}} A$. In particular, if A is infinite, we further define A to be:*
  - $*$ countably *infinite if $A \cong_{\text{set}} \mathbb{N}$;*

* uncountably *infinite otherwise.*

- finite *if it is not infinite. In this case, we have* $A \cong_{\text{set}} \{1, 2, \dots, N\}$ *for some* $N \in \mathbb{N}$ *and we say that the* cardinality *of* A*, denoted by* $|A|$*, is* N*.*

Given two maps $\phi\colon A \to B$ and $\psi\colon B \to C$, we can construct a third map, called the *composition* of ϕ and ψ, denoted by $\psi \circ \phi$ (read "psi after phi"), defined by:

$$\begin{aligned} \psi \circ \phi\colon A &\to C \\ a &\mapsto \psi(\phi(a)). \end{aligned}$$

This is often represented by drawing the following diagram

$$\begin{tikzcd} & B \arrow[dr, "\psi"] & \\ A \arrow[ur, "\phi"] \arrow[rr, "\psi\circ\phi"'] & & C \end{tikzcd}$$

and by saying that "the diagram commutes" (although sometimes this is assumed even if it is not explicitly stated). What this means is that every path in the diagram gives the same result. This might seem notational overkill at this point, but later we will encounter situations where we will have many maps, going from many places to many other places and these diagrams greatly simplify the exposition.

Proposition 1.2. *Composition of maps is associative.*

Proof. Indeed, let $\phi\colon A \to B$, $\psi\colon B \to C$ and $\xi\colon C \to D$ be maps. Then we have:

$$\begin{aligned} \xi \circ (\psi \circ \phi)\colon A &\to D \\ a &\mapsto \xi(\psi(\phi(a))) \end{aligned}$$

and:

$$\begin{aligned} (\xi \circ \psi) \circ \phi\colon A &\to D \\ a &\mapsto \xi(\psi(\phi(a))). \end{aligned}$$

Thus $\xi \circ (\psi \circ \phi) = (\xi \circ \psi) \circ \phi$. □

The operation of composition is necessary in order to defined inverses of maps.

Definition 1.20 (Inverse)**.** *Let* $\phi\colon A \to B$ *be a bijection. Then the* ***inverse*** *of* ϕ*, denoted* ϕ^{-1}*, is defined (uniquely) by:*

$$\phi^{-1} \circ \phi = \mathrm{id}_A$$

$$\phi \circ \phi^{-1} = \mathrm{id}_B\,.$$

Equivalently, we require the following diagram to commute:

$$\begin{tikzcd} A \arrow[loop left, "\mathrm{id}_A"] \arrow[rr, bend left, "\phi"] & & B \arrow[loop right, "\mathrm{id}_B"] \arrow[ll, bend left, "\phi^{-1}"] \end{tikzcd}$$

The inverse map is only defined for bijections. However, the notion of the pre-image, which we will often meet in topology, is defined for any map. Given the inverse map we can define the pre-image in a more systematic way as:

Definition 1.21 (Pre-image)**.** *Let* $\phi\colon A \to B$ *be a map and let* $V \subseteq B$*. Then we define the set:*

$$\mathrm{preim}_\phi(V) := \{a \in A \mid \phi(a) \in V\}$$

called the ***pre-image*** *of* V *under* ϕ*.*

Proposition 1.3. *Let* $\phi\colon A \to B$ *be a map, let* $U, V \subseteq B$ *and* $C = \{C_j \mid j \in J\} \subseteq \mathcal{P}(B)$*. Then:*

i) $\text{preim}_\phi(\varnothing) = \varnothing$ *and* $\text{preim}_\phi(B) = A$*;*

ii) $\text{preim}_\phi(U \setminus V) = \text{preim}_\phi(U) \setminus \text{preim}_\phi(V)$*;*

iii) $\text{preim}_\phi(\bigcup C) = \bigcup_{j\in J} \text{preim}_\phi(C_j)$ *and* $\text{preim}_\phi(\bigcap C) = \bigcap_{j\in J} \text{preim}_\phi(C_j)$*.*

Proof. i) By definition, we have:

$$\text{preim}_\phi(B) = \{a \in A : \phi(a) \in B\} = A$$

and:

$$\text{preim}_\phi(\varnothing) = \{a \in A : \phi(a) \in \varnothing\} = \varnothing.$$

ii) We have:

$$\begin{aligned} a \in \text{preim}_\phi(U \setminus V) &\Leftrightarrow \phi(a) \in U \setminus V \\ &\Leftrightarrow \phi(a) \in U \wedge \phi(a) \notin V \\ &\Leftrightarrow a \in \text{preim}_\phi(U) \wedge a \notin \text{preim}_\phi(V) \\ &\Leftrightarrow a \in \text{preim}_\phi(U) \setminus \text{preim}_\phi(V) \end{aligned}$$

iii) We have:

$$\begin{aligned} a \in \text{preim}_\phi(\textstyle\bigcup C) &\Leftrightarrow \phi(a) \in \textstyle\bigcup C \\ &\Leftrightarrow \textstyle\bigvee_{j\in J}(\phi(a) \in C_j) \\ &\Leftrightarrow \textstyle\bigvee_{j\in J}(a \in \text{preim}_\phi(C_j)) \\ &\Leftrightarrow a \in \textstyle\bigcup_{j\in J} \text{preim}_\phi(C_j) \end{aligned}$$

Similarly, we get $\text{preim}_\phi(\bigcap C) = \bigcap_{j\in J} \text{preim}_\phi(C_j)$. □

1.7 Equivalence Relations

Definition 1.22 (Equivalence Relation)**.** *Let M be a set and let $\sim$ be a relation such that the following conditions are satisfied:*

i) reflexivity: $\forall\, m \in M : m \sim m$;

ii) symmetry: $\forall\, m, n \in M : m \sim n \Leftrightarrow n \sim n$;

iii) transitivity: $\forall\, m, n, p \in M : (m \sim n \wedge n \sim p) \Rightarrow m \sim p$.

Then $\sim$ is called an ***equivalence relation*** *on M.*

Example 1.8. Consider the following wordy examples.

a) $p \sim q :\Leftrightarrow p$ is of the same opinion as q. This relation is reflexive, symmetric and transitive. Hence, it is an equivalence relation.

b) $p \sim q :\Leftrightarrow p$ is a sibling of q. This relation is symmetric and transitive but not reflexive and hence, it is not an equivalence relation.

c) $p \sim q :\Leftrightarrow p$ is taller q. This relation is transitive, but neither reflexive nor symmetric and hence, it is not an equivalence relation.

d) $p \sim q :\Leftrightarrow p$ is in love with q. This relation is generally not reflexive. People don't like themselves very much. It is certainly not normally symmetric, which is the basis of much drama in literature. It is also not transitive, except in some French films.

Definition 1.23 (Equivalence Class)**.** *Let $\sim$ be an equivalence relation on the set M. Then, for any $m \in M$, we define the set:*

$$[m] := \{n \in M \mid m \sim n\}$$

called the ***equivalence class*** *of m. Note that the condition $m \sim n$ is equivalent to $n \sim m$ since $\sim$ is symmetric.*

The following are two key properties of equivalence classes.

Proposition 1.4. *Let $\sim$ be an equivalence relation on M. Then:*

i) $a \in [m] \Rightarrow [a] = [m]$;

ii) *either* $[m] = [n]$ *or* $[m] \cap [n] = \varnothing$.

Proof. i) Since $a \in [m]$, we have $a \sim m$. Let $x \in [a]$. Then $x \sim a$ and hence $x \sim m$ by transitivity. Therefore $x \in [m]$ and hence $[a] \subseteq [m]$. Similarly, we have $[m] \subseteq [a]$ and hence $[a] = [m]$.

ii) Suppose that $[m] \cap [n] \neq \varnothing$. That is:

$$\exists\, z : z \in [m] \wedge z \in [n].$$

Thus $z \sim m$ and $z \sim n$ and hence, by symmetry and transitivity, $m \sim n$. This implies that $m \in [n]$ and hence that $[m] = [n]$. □

Definition 1.24 (Quotient Set)**.** *Let $\sim$ be an equivalence relation on M. Then we define the* ***quotient set*** *of M by $\sim$ as:*

$$M/\!\sim \; := \{[m] \mid m \in M\}.$$

This is indeed a set since $[m] \subseteq \mathcal{P}(M)$ and hence we can write more precisely:

$$M/\!\sim \; := \{[m] \in \mathcal{P}(M) \mid m \in M\}.$$

Then clearly $M/\!\sim$ is a set by the power set axiom and the principle of restricted comprehension.

Remark 1.22. Due to the axiom of choice, there exists a complete system of representatives for $\sim$, i.e. a set R such that $R \cong_{\text{set}} M/\!\sim$.

Remark 1.23. Care must be taken when defining maps whose domain is a quotient set if one uses representatives to define the map. In order for the map to be *well-defined* one needs to show that the map is independent of the choice of representatives.

Example 1.9. Let $M = \mathbb{Z}$ and define $\sim$ by:

$$m \sim n :\Leftrightarrow n - m \in 2\mathbb{Z}.$$

It is easy to check that $\sim$ is indeed an equivalence relation. Moreover, we have:

$$[0] = [2] = [4] = \cdots = [-2] = [-4] = \cdots$$

and:

$$[1] = [3] = [5] = \cdots = [-1] = [-3] = \cdots$$

Thus we have: $\mathbb{Z}/\sim \; = \{[0], [1]\}$. We wish to define an addition $\oplus$ on $\mathbb{Z}/\sim$ by inheriting the usual addition on $\mathbb{Z}$. As a tentative definition we could have:

$$\oplus \colon \mathbb{Z}/\!\sim \times \, \mathbb{Z}/\!\sim \; \to \mathbb{Z}/\!\sim$$

being given by:

$$[a] \oplus [b] := [a + b].$$

However, we need to check that our definition does not depend on the choice of class representatives, i.e. if $[a] = [a']$ and $[b] = [b']$, then we should have:

$$[a] \oplus [b] = [a'] \oplus [b'].$$

Indeed, $[a] = [a']$ and $[b] = [b']$ means $a - a' \in 2\mathbb{Z}$ and $b - b' \in 2\mathbb{Z}$, i.e. $a - a' = 2m$ and $b - b' = 2n$ for some $m, n \in \mathbb{Z}$. We thus have:

$$\begin{aligned} [a' + b'] &= [a - 2m + b - 2n] \\ &= [(a + b) - 2(m + n)] \\ &= [a + b], \end{aligned}$$

where the last equality follows since:

$$(a+b) - 2(m+n) - (a+b) = -2(m+n) \in 2\mathbb{Z}.$$

Therefore $[a'] \oplus [b'] = [a] \oplus [b]$ and hence the operation $\oplus$ is well-defined.

Example 1.10. As a counterexample, with the same set-up as in the previous example, let us define an operation $\star$ by:

$$[a] \star [b] := \frac{a}{b}.$$

This is easily seen to be *ill-defined* since $[1] = [3]$ and $[2] = [4]$ but:

$$[1] \star [2] = \frac{1}{2} \neq \frac{3}{4} = [3] \star [4].$$

1.8 Construction of $\mathbb{N}$, $\mathbb{Z}$, $\mathbb{Q}$ and $\mathbb{R}$

Recall that, invoking the axiom of infinity, we defined the natural numbers:

$$\mathbb{N} := \{0, 1, 2, 3, \ldots\},$$

where:

$$0 := \varnothing, \quad 1 := \{\varnothing\}, \quad 2 := \{\{\varnothing\}\}, \quad 3 := \{\{\{\varnothing\}\}\}, \quad \ldots$$

We would now like to define an addition operation on $\mathbb{N}$ by using the axioms of set theory. We will need some preliminary definitions.

Definition 1.25 (Successor Map)**.** *The* ***successor map*** *S on $\mathbb{N}$ is defined by:*

$$\begin{aligned} S \colon \mathbb{N} &\to \mathbb{N} \\ n &\mapsto \{n\}. \end{aligned}$$

Example 1.11. Consider $S(2)$. Since $2 := \{\{\varnothing\}\}$, we have $S(2) = \{\{\{\varnothing\}\}\} =: 3$. Therefore, we have $S(2) = 3$ as we would have expected.

To make progress, we also need to define the predecessor map, which is only defined on the set $\mathbb{N}^* := \mathbb{N} \setminus \{\varnothing\}$.

Definition 1.26 (Predecessor Map)**.** *The* ***predecessor map*** *P on $\mathbb{N}^*$ is defined by:*

$$\begin{aligned} P \colon \mathbb{N}^* &\to \mathbb{N} \\ n &\mapsto m \quad \textit{such that} \quad m \in n. \end{aligned}$$

Example 1.12. We have $P(2) = P(\{\{\varnothing\}\}) = \{\varnothing\} = 1$.

Definition 1.27 (n-th Power)**.** *Let $n \in \mathbb{N}$. The* ***n-th power*** *of S, denoted S^n, is defined recursively by:*

$$\begin{aligned} S^n &:= S \circ S^{P(n)} \qquad \textit{if } n \in \mathbb{N}^* \\ S^0 &:= \mathrm{id}_{\mathbb{N}}. \end{aligned}$$

We are now ready to define addition.

Definition 1.28 (Addition Of Natural Numbers)**.** *The* ***addition*** *operation on $\mathbb{N}$ is defined as a map:*

$$\begin{aligned} + \colon \mathbb{N} \times \mathbb{N} &\to \mathbb{N} \\ (m, n) &\mapsto m + n := S^n(m). \end{aligned}$$

Example 1.13. We have:

$$2 + 1 = S^1(2) = S(2) = 3$$

and:

$$1 + 2 = S^2(1) = S(S^1(1)) = S(S(1)) = S(2) = 3.$$

Using this definition, it is possible to show that + is commutative and associative. The *neutral element* of + is 0 since:

$$m + 0 = S^0(m) = \mathrm{id}_{\mathbb{N}}(m) = m$$

and:

$$0 + m = S^m(0) = S^{P(m)}(1) = S^{P(P(m))}(2) = \cdots = S^0(m) = m.$$

Clearly, there exist no inverses for + in $\mathbb{N}$, i.e. given $m \in \mathbb{N}$ (non-zero), there exist no $n \in \mathbb{N}$ such that $m + n = 0$. This motivates the extension of the natural numbers to the integers. In order to rigorously define $\mathbb{Z}$, we need to define the following relation on $\mathbb{N} \times \mathbb{N}$.

Let $\sim$ be the relation on $\mathbb{N} \times \mathbb{N}$ defined by:

$$(m, n) \sim (p, q) :\Leftrightarrow m + q = p + n.$$

It is easy to check that this is an equivalence relation as:

i) $(m, n) \sim (m, n)$ since $m + n = m + n$;

ii) $(m, n) \sim (p, q) \Rightarrow (p, q) \sim (m, n)$ since $m + q = p + n \Leftrightarrow p + n = m + q$;

iii) $((m, n) \sim (p, q) \wedge (p, q) \sim (r, s)) \Rightarrow (m, n) \sim (r, s)$ since we have:

$$m + q = p + n \wedge p + s = r + q,$$

hence $m + q + p + s = p + n + r + q$, and thus $m + s = r + n$.

By equipping this relation we can define the set of integers in the following way:

Definition 1.29 (Integers)**.** *We define the set of* integers *by:*

$$\mathbb{Z} := (\mathbb{N} \times \mathbb{N})/\sim .$$

The intuition behind this definition is that the pair (m, n) stands for "$m - n$". In other words, we represent each integer by a pair of natural numbers whose (yet to be defined) difference is precisely that integer. There are, of course, many ways to represent the same integer with a pair of natural numbers in this way. For instance, the integer -1 could be represented by $(1, 2)$, $(2, 3)$, $(112, 113)$, ...

Notice however that $(1, 2) \sim (2, 3)$, $(1, 2) \sim (112, 113)$, etc. and indeed, taking the quotient by $\sim$ takes care of this "redundancy". Notice also that this definition relies entirely on previously defined entities.

Remark 1.24. In a first introduction to set theory it is not unlikely to find the claim that the natural numbers are part of the integers, i.e. $\mathbb{N} \subseteq \mathbb{Z}$. However, according to our definition, this is obviously nonsense since $\mathbb{N}$ and $\mathbb{Z} := (\mathbb{N} \times \mathbb{N})/\sim$ contain entirely different elements. What is true is that $\mathbb{N}$ can be *embedded* into $\mathbb{Z}$, i.e. there exists an *inclusion map* ι, given by:

$$\begin{aligned} \iota \colon \mathbb{N} &\hookrightarrow \mathbb{Z} \\ n &\mapsto [(n, 0)] \end{aligned}$$

and it is in this sense that $\mathbb{N}$ is included in $\mathbb{Z}$.

Definition 1.30 (Inverse Of Integer)**.** *Let $n := [(n, 0)] \in \mathbb{Z}$. Then we define the* inverse *of n to be $-n := [(0, n)]$.*

We would now like to inherit the + operation from $\mathbb{N}$.

Definition 1.31 (Addition Of Integers)**.** *We define the* addition of integers $+_{\mathbb{Z}} \colon \mathbb{Z} \times \mathbb{Z} \to \mathbb{Z}$ *by:*

$$[(m, n)] +_{\mathbb{Z}} [(p, q)] := [(m + p, n + q)].$$

Since we used representatives to define $+_{\mathbb{Z}}$, we would need to check that $+_{\mathbb{Z}}$ is well-defined. It is an easy exercise.

Example 1.14. $2 +_{\mathbb{Z}} (-3) := [(2, 0)] +_{\mathbb{Z}} [(0, 3)] = [(2, 3)] = [(0, 1)] =: -1$. Hallelujah!

In a similar fashion, we define the set of *rational numbers* by:

$$\mathbb{Q} := (\mathbb{Z} \times \mathbb{Z}^*)/\sim,$$

where $\mathbb{Z}^* := \mathbb{Z} \setminus \{0\}$ and $\sim$ is a relation on $\mathbb{Z} \times \mathbb{Z}^*$ given by:

$$(p, q) \sim (r, s) :\Leftrightarrow ps = qr,$$

assuming that a *multiplication* operation on the integers has already been defined.

Example 1.15. We have $(2, 3) \sim (4, 6)$ since $2 \times 6 = 12 = 3 \times 4$.

Similarly to what we did for the integers, here we are representing each rational number by the collection of pairs of integers (the second one in each pair being non-zero) such that their (yet to be defined) ratio is precisely that rational number. Thus, for example, we have:

$$\frac{2}{3} := [(2, 3)] = [(4, 6)] = \dots$$

There are many ways to construct the reals from the rationals. One is to define a set $\mathscr{A}$ of *almost homomorphisms* on $\mathbb{Z}$ and hence define:

$$\mathbb{R} := \mathscr{A}/\sim,$$

where $\sim$ is a "suitable" equivalence relation on $\mathscr{A}$.

Chapter 2

Algebraic Structures

2.1 Algebraic Structures

Definition 2.1 (Algebraic Structures)**.** *A set A (called the underlying set, carrier set or domain), together with a collection of maps (called operations) on A of finite arity (typically binary operations), and a finite set of identities, known as axioms, that these operations must satisfy, is called an* ***algebraic structure****. Some algebraic structures also involve another set (called the scalar set).*

Examples of algebraic structures with a single underlying set include groups,fields and rings. Examples of algebraic structures with two underlying sets include vector spaces, modules, and algebras. In this section we will review the most important algebraic structures for our purposes.

One has to be careful with the terminology since it changes depending on the area of mathematics. For example, in the context of universal algebra, the set A with this structure is called an algebra, while, in other contexts, it is (somewhat ambiguously) called an algebraic structure, the term algebra being reserved for specific algebraic structures that are vector spaces over a field or modules over a commutative ring.

The properties of specific algebraic structures are studied in abstract algebra. The general theory of algebraic structures has been formalized in universal algebra. The language of category theory is used to express and study relationships between different classes of algebraic and non-algebraic objects. This is because it is sometimes possible to find strong connections between some classes of objects, sometimes of different kinds. For example, Galois theory establishes a connection between certain fields and groups: two algebraic structures of different kinds.

2.2 Groups

Definition 2.2 (Group)**.** *A* ***group*** *is a tuple $(G, \cdot)$, where G is a set (called the underlying set of the group) and $\cdot$ is a map (called operation) $G \times G \to G$ satisfying the following four group axioms:*

- *Closure:* $\forall\, a, b \in G : a \cdot b \in G$;
- *Associativity:* $\forall\, a, b, c \in G : (a \cdot b) \cdot c = a \cdot (b \cdot c)$;
- *Neutral Element:* $\exists\, e \in G : \forall\, a \in G : a \cdot e = e \cdot a = a$;
- *Inverse Element:* $\forall\, a \in G : \exists\, a^{-1} \in G : a \cdot a^{-1} = a^{-1} \cdot a = e$;

The identity element e of a group G is often written as 1 a notation inherited from the multiplicative identity. If a group is abelian, then one may choose to denote the group operation by $+$ and the identity element by 0.

The result of the group operation may depend on the order of the operands. In other words, the result of combining element a with element b need not yield the same result as combining element b with element a, so the equation $a \cdot b = b \cdot a$ may not be true for every two elements a and b.

Definition 2.3 (Abelian Group)**.** *A group G is called **Abelian** if on top of the four group axioms it also satisfies the axiom of commutativity:*

- *Commutativity:* $\forall\, a, b \in G : a \cdot b = b \cdot a$;

Commutativity always holds in the group of integers under addition, because $a + b = b + a$ for any two integers (commutativity of addition). The symmetry group is an example of a group that is not abelian.

2.3 Fields

Definition 2.4 (Field)**.** *An **(algebraic) field** is a triple $(K, +, \cdot)$, where K is a set and $+, \cdot$ are maps $K \times K \to K$ satisfying the following axioms:*

- $(K, +)$ *is an abelian group, i.e.*
  - *i) Closure:* $\forall\, a, b \in K : a + b \in K$;
  - *ii) Associativity:* $\forall\, a, b, c \in K : (a + b) + c = a + (b + c)$;
  - *iii) Neutral Element:* $\exists\, 0 \in K : \forall\, a \in K : a + 0 = 0 + a = a$;
  - *iv) Inverse Element:* $\forall\, a \in K : \exists\, {-a} \in K : a + (-a) = (-a) + a = 0$;
  - *v) Commutativity:* $\forall\, a, b \in K : a + b = b + a$;
- $(K^*, \cdot)$, *where* $K^* := K \setminus \{0\}$, *is an abelian group, i.e.*
  - *vi) Closure:* $\forall\, a, b \in K^* : a \cdot b \in K^*$;
  - *vii) Associativity:* $\forall\, a, b, c \in K^* : (a \cdot b) \cdot c = a \cdot (b \cdot c)$;
  - *viii) Neutral Element:* $\exists\, 1 \in K^* : \forall\, a \in K^* : a \cdot 1 = 1 \cdot a = a$;
  - *ix) Inverse Element:* $\forall\, a \in K^* : \exists\, a^{-1} \in K^* : a \cdot a^{-1} = a^{-1} \cdot a = 1$;
  - *x) Commutativity:* $\forall\, a, b \in K^* : a \cdot b = b \cdot a$;
- *the maps $+$ and $\cdot$ satisfy the distributive property:*
  - *xi)* $\forall\, a, b, c \in K : (a + b) \cdot c = a \cdot c + b \cdot c$;

Remark 2.1. In the above definition, we included axiom v for the sake of clarity, but in fact it can be proven starting from the other axioms.

2.4 Vector Spaces

Definition 2.5 (K-Vector Space)**.** *Let $(K, +, \cdot)$ be a field. A K**-vector space**, or **vector space over** K or **linear space over** K is a triple $(V, \oplus, \odot)$, where V is a set and*

$$\oplus\colon V \times V \to V$$
$$\odot\colon K \times V \to V$$

are maps satisfying the following axioms:

- $(V, \oplus)$ *is an abelian group i.e.*
  - *i) Closure:* $\forall\, v, w \in V : v \oplus w \in K$;
  - *ii) Associativity:* $\forall\, v, w, z \in V : (v \oplus w) \oplus z = v \oplus (w \oplus z)$;
  - *iii) Neutral Element:* $\exists\, 0 \in V : \forall\, v \in V : v \oplus 0 = 0 \oplus v = v$;
  - *iv) Inverse Element:* $\forall\, v \in V : \exists\, -v \in V : v \oplus (-v) = (-v) \oplus v = 0$;
  - *v) Commutativity:* $\forall\, v, w \in V : v \oplus w = w \oplus v$;
- *the map $\odot$ is an* action *of K on $(V, \oplus)$:*
  - *vi) Distributivity Of Scalar Multiplication - Vector Addition:* $\forall\, \lambda \in K : \forall\, v, w \in V : \lambda \odot (v \oplus w) = (\lambda \odot v) \oplus (\lambda \odot w)$;

vii) Distributivity Of Scalar Multiplication - Field Addition: $\forall\,\lambda,\mu \in K : \forall\, v \in V : (\lambda + \mu) \odot v = (\lambda \odot v) \oplus (\mu \odot v)$;

viii) Compatibility Of Scalar Multiplication - Field Multiplication $\forall\,\lambda,\mu \in K : \forall\, v \in V : (\lambda \cdot \mu) \odot v = \lambda \odot (\mu \odot v)$;

ix) Neutral Element Of Scalar Multiplication $\forall\, v \in V : 1 \odot v = v$.

The elements of a vector space are called *vectors*, while the elements of K are often called *scalars*, and the map $\odot$ is called *scalar multiplication*.

2.4.1 Linear Maps

As usual by now, we will look at the structure-preserving maps between vector spaces.

Definition 2.6 (Linear Maps)**.** *Let $(V, \oplus, \odot)$, $(W, \boxplus, \boxdot)$ be vector spaces over the same field K and let $f\colon V \to W$ be a map. We say that f is a* ***linear map****, and we denote it as $f\colon V \xrightarrow{\sim} W$, if for all $v_1, v_2 \in V$ and all $\lambda \in K$*

$$f((\lambda \odot v_1) \oplus v_2) = (\lambda \boxdot f(v_1)) \boxplus f(v_2).$$

From now on, we will drop the special notation for the vector space operations and suppress the dot for scalar multiplication. For instance, we will write the equation above as $f(\lambda v_1 + v_2) = \lambda f(v_1) + f(v_2)$, hoping that this will not cause any confusion.

Definition 2.7 ($\mathrm{Hom}(V,W)$)**.** *Let V and W be vector spaces over the same field K. We define the set $\mathrm{Hom}(V,W)$ as the set of all linear maps between V and W:*

$$\mathrm{Hom}(V,W) := \{f \mid f\colon V \xrightarrow{\sim} W\}$$

$\mathrm{Hom}(V,W)$ can itself be made into a vector space over K by defining:

$$\begin{aligned} \oplus\colon \mathrm{Hom}(V,W) \times \mathrm{Hom}(V,W) &\to \mathrm{Hom}(V,W) \\ (f,g) &\mapsto f \oplus g \end{aligned}$$

where

$$\begin{aligned} f \oplus g\colon V &\xrightarrow{\sim} W \\ v &\mapsto (f \oplus g)(v) := f(v) + g(v), \end{aligned}$$

and

$$\begin{aligned} \diamond\colon K \times \mathrm{Hom}(V,W) &\to \mathrm{Hom}(V,W) \\ (\lambda, f) &\mapsto \lambda \diamond f \end{aligned}$$

where

$$\begin{aligned} \lambda \diamond f\colon V &\xrightarrow{\sim} W \\ v &\mapsto (\lambda \diamond f)(v) := \lambda f(v). \end{aligned}$$

It is easy to check that both $f \oplus g$ and $\lambda \diamond f$ are indeed linear maps from V to W. For instance, we have:

$$\begin{aligned} (\lambda \diamond f)(\mu v_1 + v_2) &= \lambda f(\mu v_1 + v_2) && \text{(by definition)} \\ &= \lambda(\mu f(v_1) + f(v_2)) && \text{(since } f \text{ is linear)} \\ &= \lambda\mu f(v_1) + \lambda f(v_2) && \text{(by axioms i and iii)} \\ &= \mu\lambda f(v_1) + \lambda f(v_2) && \text{(since } K \text{ is a field)} \\ &= \mu(\lambda \diamond f)(v_1) + (\lambda \diamond f)(v_2) \end{aligned}$$

so that $\lambda \diamond f \in \mathrm{Hom}(V,W)$. One should also check that $\oplus$ and $\diamond$ satisfy the vector space axioms.

Definition 2.8 (Endomorphisms)**.** *Let V be a vector space. An* ***endomorphism*** *of V is a linear map $V \to V$.*

Definition 2.9 (End(V)). *Let V be a vector space. We define the set* End(V) *as the set of all endomorphisms of V:*

$$\mathrm{End}(V) := \mathrm{Hom}(V, V)$$

It is easy to show that End(V) can again itself be made into a vector space over K.

Definition 2.10 (Linear Isomorphism). *A bijective linear map is called a* ***linear isomorphism*** *of vector spaces.*

Definition 2.11 (Isomorphic Vector Spaces). *Two vector spaces are said to be* ***isomorphic*** *if there exists a linear isomorphism between them. We write $V \cong_{\mathrm{vec}} W$.*

Definition 2.12 (Automorphism). *Let V be a vector space. An* ***automorphism*** *of V is a linear isomorphism $V \to V$.*

Definition 2.13 (Aut(V)). *Let V be a vector space. We define the set* Aut(V) *as the set of all automorphisms of V:*

$$\mathrm{Aut}(V) := \{f \in \mathrm{End}(V) \mid f \textit{ is an isomorphism}\}$$

Remark 2.2. Note that Aut(V) **cannot** be made into a vector space . It is however a group under the operation of composition of linear maps.

Definition 2.14 (Dual Vector Space). *Let V be a vector space over K. The* ***dual*** *vector space to V is*

$$V^* := \mathrm{Hom}(V, K),$$

where K is considered as a vector space over itself.

The dual vector space to V is the vector space of linear maps from V to the underlying field K, which are variously called *linear functionals*, *covectors*, or *one-forms* on V. The dual plays a very important role, in that from a vector space and its dual, we will construct the tensor products.

2.4.2 Basis Of Vector Spaces

Given a vector space without any additional structure, the only notion of basis that we can define is a so-called Hamel basis.

Definition 2.15 (Hamel Basis). *Let $(V, +, \cdot)$ be a vector space over K. A subset $\mathcal{B} \subseteq V$ is called a* ***Hamel basis*** *for V if*

- *every finite subset $\{b_1, \dots, b_N\}$ of $\mathcal{B}$ is linearly independent, i.e.*

$$\sum_{i=1}^{N} \lambda^i b_i = 0 \;\Rightarrow\; \lambda^1 = \dots = \lambda^N = 0;$$

- *$\mathcal{B}$ is a* generating *or* spanning set *of V, i.e.*

$$\forall\, v \in V : \exists\, v^1, \dots, v^M \in K : \exists\, b_1, \dots, b_M \in \mathcal{B} : v = \sum_{i=1}^{M} v^i b_i.$$

Remark 2.3. We can write the second condition more succinctly by defining

$$\mathrm{span}_K(\mathcal{B}) := \left\{ \sum_{i=1}^{n} \lambda^i b_i \;\middle|\; \lambda^i \in K \wedge b_i \in \mathcal{B} \wedge n \geq 1 \right\}$$

and thus writing $V = \mathrm{span}_K(\mathcal{B})$.

Remark 2.4. Note that we have been using superscripts for the elements of K, and these should not be confused with exponents.

The following characterisation of a Hamel basis is often useful.

Proposition 2.1. *Let V be a vector space and $\mathcal{B}$ a Hamel basis of V. Then $\mathcal{B}$ is a minimal spanning and maximal independent subset of V, i.e., if $S \subseteq V$, then*

- $\operatorname{span}(S) = V \;\Rightarrow\; |S| \geq |\mathcal{B}|$*;*
- *S is linearly independent $\;\Rightarrow\; |S| \leq |\mathcal{B}|$.*

Definition 2.16 (Dimension Of Vector Space)**.** *Let V be a vector space. The **dimension** of V is $\dim V := |\mathcal{B}|$, where $\mathcal{B}$ is a Hamel basis for V.*

Even though we will not prove it, it is the case that every Hamel basis for a given vector space has the same cardinality, and hence the notion of dimension is well-defined.

Proposition 2.2. *If $\dim V < \infty$ and $S \subseteq V$, then we have the following:*

- *if $\operatorname{span}_K(S) = V$ and $|S| = \dim V$, then S is a Hamel basis of V;*
- *if S is linearly independent and $|S| = \dim V$, then S is a Hamel basis of V.*

Theorem 2.1. *If $\dim V < \infty$, then $(V^*)^* \cong_{\mathrm{vec}} V$.*

Remark 2.5. Note that while we need the concept of basis to state this result (since we require $\dim V < \infty$), the isomorphism that we have constructed is independent of any choice of basis.

Remark 2.6. While a choice of basis often simplifies things, when defining new objects it is important to do so without making reference to a basis. If we do define something in terms of a basis (e.g. the dimension of a vector space), then we have to check that the thing is well-defined, i.e. it does not depend on which basis we choose.

If V is finite-dimensional, then V^* is also finite-dimensional and $V \cong_{\mathrm{vec}} V^*$. Moreover, given a basis $\mathcal{B}$ of V, there is a spacial basis of V^* associated to $\mathcal{B}$.

Definition 2.17 (Dual Basis)**.** *Let V be a finite-dimensional vector space with basis $\mathcal{B} = \{e_1, \dots, e_{\dim V}\}$. The **dual basis** to $\mathcal{B}$ is the unique basis $\mathcal{B}' = \{\epsilon^1, \dots, \epsilon^{\dim V}\}$ of V^* such that*

$$\forall\, 1 \leq i, j \leq \dim V : \quad \epsilon^i(e_j) = \delta^i_j := \begin{cases} 1 & \text{if } i = j \\ 0 & \text{if } i \neq j \end{cases}$$

Remark 2.7. If V is finite-dimensional, then V is isomorphic to both V^* and $(V^*)^*$. In the case of V^*, an isomorphism is given by sending each element of a basis $\mathcal{B}$ of V to a different element of the dual basis $\mathcal{B}'$, and then extending linearly to V. You will (and probably already have) read that a vector space is *canonically* isomorphic to its double dual, but *not* canonically isomorphic to its dual, because an arbitrary choice of basis on V is necessary in order to provide an isomorphism.

2.4.3 Change Of Basis

Let V be a vector space over K with $d = \dim V < \infty$ and let $\{e_1, \dots, e_d\}$ be a basis of V. Consider a new basis $\{\widetilde{e}_1, \dots, \widetilde{e}_d\}$. Since the elements of the new basis are also elements of V, we can expand them in terms of the old basis. We have:

$$\widetilde{e}_i = \sum_{j=1}^{d} A^j{}_i\, e_j \qquad \text{for } 1 \leq i \leq d.$$

for some $A^j{}_i \in K$. Similarly, we have

$$e_i = \sum_{j=1}^{d} B^j{}_i\, \widetilde{e}_j \qquad \text{for } 1 \leq i \leq d.$$

for some $B^j{}_i \in K$. It is a standard linear algebra result that the matrices A and B, with entries $A^j{}_i$ and $B^j{}_i$ respectively, are invertible and, in fact, $A^{-1} = B$.

Once we have a basis $\mathcal{B}$, the expansion of $v \in V$ in terms of elements of $\mathcal{B}$ is, in fact, unique. Hence we can meaningfully speak of the *components* of v in the basis $\mathcal{B}$. The notion of coordinates can also be generalised to the case of tensors that we will define next.

2.4.4 Tensors

Definition 2.18 (Bilinear Maps)**.** *Let V, W, Z be vector spaces over K. A map $f\colon V\times W\to Z$ is said to be* ***bilinear*** *if*

- $\forall\, w\in W:\forall\, v_1,v_2\in V:\forall\,\lambda\in K: f(\lambda v_1+v_2,w)=\lambda f(v_1,w)+f(v_2,w)$;
- $\forall\, v\in V:\forall\, w_1,w_2\in W:\forall\,\lambda\in K: f(v,\lambda w_1+w_2)=\lambda f(v,w_1)+f(v,w_2)$;

i.e. if the maps $v\mapsto f(v,w)$, for any fixed w, and $w\mapsto f(v,w)$, for any fixed v, are both linear as maps $V\to Z$ and $W\to Z$, respectively.

Remark 2.8. Compare this with the definition of a linear map $f\colon V\times W\xrightarrow{\sim} Z$:

$$\forall\, x,y\in V\times W:\forall\,\lambda\in K: f(\lambda x+y)=\lambda f(x)+f(y).$$

More explicitly, if $x=(v_1,w_1)$ and $y=(v_2,w_2)$, then:

$$f(\lambda(v_1,w_1)+(v_2,w_2))=\lambda f((v_1,w_1))+f((v_2,w_2)).$$

A bilinear map out of $V\times W$ is *not* the same as a linear map out of $V\times W$. In fact, bilinearity is just a special kind of non-linearity.

Example 2.1. The map $f\colon\mathbb{R}^2\to\mathbb{R}$ given by $(x,y)\mapsto x+y$ is linear but not bilinear, while the map $(x,y)\mapsto xy$ is bilinear but not linear.

We can immediately generalise the above to define *multilinear* maps out of a Cartesian product of vector spaces.

Definition 2.19 (Tensors)**.** *Let V be a vector space over K. A (p,q)-**tensor** T on V is a multilinear map*

$$T\colon \underbrace{V^*\times\cdots\times V^*}_{p\ copies}\times\underbrace{V\times\cdots\times V}_{q\ copies}\to K.$$

Remark 2.9. By convention, a $(0,0)$ on V is just an element of K, and hence $T^0_0V=K$.

Definition 2.20 (Covariant / Contravariant Tensor)**.** *A type $(p,0)$ tensor is called a **covariant** p-**tensor**, while a tensor of type $(0,q)$ is called a **contravariant** q-**tensor**.*

Definition 2.21 (T^p_qV)**.** *We define the set of all (p,q)-tensors T as:*

$$T^p_qV:=\underbrace{V\otimes\cdots\otimes V}_{p\ copies}\otimes\underbrace{V^*\otimes\cdots\otimes V^*}_{q\ copies}:=\{T\mid T\ \textit{is a}\ (p,q)\textit{-tensor on}\ V\}.$$

Remark 2.10. Note that to define T^p_qV as a set, we should be careful and invoke the principle of restricted comprehension, i.e. we should say where the Ts are coming from. In general, say we want to build a set of maps $f\colon A\to B$ satisfying some property p. Recall that the notation $f\colon A\to B$ is hiding the fact that is a relation (indeed, a functional relation), and a relation between A and B is a subset of $A\times B$. Therefore, we ought to write:

$$\{f\in\mathcal{P}(A\times B)\mid f\colon A\to B\text{ and }p(f)\}.$$

In the case of T^p_qV we have:

$$T^p_qV:=\big\{T\in\mathcal{P}\big(\underbrace{V^*\times\cdots\times V^*}_{p\text{ copies}}\times\underbrace{V\times\cdots\times V}_{q\text{ copies}}\times K\big)\mid T\text{ is a }(p,q)\text{-tensor on }V\big\},$$

although we will not write this down every time.

The set T^p_qV can be equipped with a K-vector space structure by defining

$$\begin{aligned}\oplus\colon T^p_qV\times T^p_qV&\to T^p_qV\\ (T,S)&\mapsto T\oplus S\end{aligned}$$

and

$$\begin{aligned} \odot\colon K \times T^p_q V &\to T^p_q V \\ (\lambda, T) &\mapsto \lambda \odot T, \end{aligned}$$

where $T \oplus S$ and $\lambda \odot T$ are defined pointwise, as we did with $\mathrm{Hom}(V, W)$.

We now define an important way of obtaining a new tensor from two given ones.

Definition 2.22 (Tensor Product)**.** *Let $T \in T^p_q V$ and $S \in T^r_s V$. The* ***tensor product*** *of T and S is the tensor $T \otimes S \in T^{p+r}_{q+s} V$ defined by:*

$$\begin{aligned} &(T \otimes S)(\omega_1, \ldots, \omega_p, \omega_{p+1}, \ldots, \omega_{p+r}, v_1, \ldots, v_q, v_{q+1}, \ldots, v_{q+s}) \\ &\qquad := T(\omega_1, \ldots, \omega_p, v_1, \ldots, v_q)\, S(\omega_{p+1}, \ldots, \omega_{p+r}, v_{q+1}, \ldots, v_{q+s}), \end{aligned}$$

with $\omega_i \in V^*$ and $v_i \in V$.

Some examples are in order.

Example 2.2.

a) $T^0_1 V := \{T \mid T\colon V \xrightarrow{\sim} K\} = \mathrm{Hom}(V, K) =: V^*$. Note that here multilinear is the same as linear since the maps only have one argument.

b) $T^1_1 V \equiv V \otimes V^* := \{T \mid T \text{ is a bilinear map } V^* \times V \to K\}$. We claim that this is the same as $\mathrm{End}(V^*)$. Indeed, given $T \in V \otimes V^*$, we can construct $\widehat{T} \in \mathrm{End}(V^*)$ as follows:

$$\begin{aligned} \widehat{T}\colon V^* &\xrightarrow{\sim} V^* \\ \omega &\mapsto T(-, \omega) \end{aligned}$$

where, for any fixed ω, we have

$$\begin{aligned} T(-, \omega)\colon V &\xrightarrow{\sim} K \\ v &\mapsto T(v, \omega). \end{aligned}$$

The linearity of both $\widehat{T}$ and $T(-, \omega)$ follows immediately from the bilinearity of T. Hence $T(-, \omega) \in V^*$ for all ω, and $\widehat{T} \in \mathrm{End}(V^*)$. This correspondence is invertible, since can reconstruct T from $\widehat{T}$ by defining

$$\begin{aligned} T\colon V \times V^* &\to K \\ (v, \omega) &\mapsto T(v, \omega) := (\widehat{T}(\omega))(v). \end{aligned}$$

The correspondence is in fact linear, hence an isomorphism, and thus

$$T^1_1 V \cong_{\mathrm{vec}} \mathrm{End}(V^*).$$

Other examples we would like to consider are

c) $T^0_1 V \overset{?}{\cong}_{\mathrm{vec}} V$: while you will find this stated as true in some physics textbooks, it is in fact *not true* in general;

d) $T^1_1 V \overset{?}{\cong}_{\mathrm{vec}} \mathrm{End}(V)$: This is also not true in general;

e) $(V^*)^* \overset{?}{\cong}_{\mathrm{vec}} V$: This only holds if V is finite-dimensional.

Definition 2.23 (Components Of A Tensor)**.** *Let V be a finite-dimensional vector space over K with basis $\mathcal{B} = \{e_1, \ldots, e_{\dim V}\}$ and dual basis $\{\epsilon^1, \ldots, \epsilon^{\dim V}\}$ and let $T \in T^p_q V$. We define the* ***components*** *of T in the basis $\mathcal{B}$ to be the numbers*

$$T^{a_1 \ldots a_p}{}_{b_1 \ldots b_q} := T(\epsilon^{a_1}, \ldots, \epsilon^{a_p}, e_{b_1}, \ldots, e_{b_q}) \in K,$$

where $1 \leq a_i, b_j \leq \dim V$.

Just as with vectors, the components completely determine the tensor. Indeed, we can reconstruct the tensor from its components by using the basis:

$$T = \underbrace{\sum_{a_1=1}^{\dim V} \cdots \sum_{b_q=1}^{\dim V}}_{p+q \text{ sums}} T^{a_1 \ldots a_p}{}_{b_1 \ldots b_q} e_{a_1} \otimes \cdots \otimes e_{a_p} \otimes \epsilon^{b_1} \otimes \cdots \otimes \epsilon^{b_q},$$

where the e_{a_i}s are understood as elements of $T^1_0 V \cong_{\text{vec}} V$ and the $\epsilon^{b_i}s$ as elements of $T^0_1 V \cong_{\text{vec}} V^*$. Note that each summand is a (p,q)-tensor and the (implicit) multiplication between the components and the tensor product is the scalar multiplication in $T^p_q V$.

2.4.5 Notational Conventions

From now on, we will employ the Einstein's summation convention, which consists in suppressing the summation sign when the indices to be summed over each appear once as a subscript and once as a superscript in the same term. For example, we write

$$v = v^i e_i \qquad \text{and} \qquad T = T^{ij}{}_k e_i \otimes e_j \otimes f^k$$

instead of

$$v = \sum_{i=1}^{d} v^i e_i \qquad \text{and} \qquad T = \sum_{i=1}^{d} \sum_{j=1}^{d} \sum_{k=1}^{d} T^{ij}{}_k e_i \otimes e_j \otimes f^k.$$

Indices that are summed over are called *dummy indices*; they always appear in pairs and clearly it doesn't matter which particular letter we choose to denote them, provided it doesn't already appear in the expression. Indices that are not summed over are called *free indices*; expressions containing free indices represent multiple expressions, one for each value of the free indices; free indices must match on both sides of an equation. The ranges over which the indices run are usually understood and not written out.

The convention on which indices go upstairs and which downstairs (which we have already been using) is that:

- the basis vectors of V carry downstairs indices;
- the basis vectors of V^* carry upstairs indices;
- all other placements are enforced by the Einstein's summation convention.

For example, since the components of a vector must multiply the basis vectors and be summed over, the Einstein's summation convention requires that they carry upstair indices.

Example 2.3. Using the summation convention, we have:

a) $\epsilon^a(v) = \epsilon^a(v^b e_b) = v^b \epsilon^a(e_b) = v^b \delta^a_b = v^a$;

b) $\omega(e_b) = (\omega_a \epsilon^a)(e_b) = \omega_a \epsilon^a(e_b) = \omega_b$;

c) $\omega(v) = \omega_a \epsilon^a(v^b e_b) = \omega_a v^a$;

where $v \in V$, $\omega \in V^*$, $\{e_i\}$ is a basis of V and $\{\epsilon^j\}$ is the dual basis to $\{e_i\}$.

Remark 2.11. The Einstein's summation convention should only be used when dealing with linear spaces and multilinear maps. The reason for this is the following. Consider a map $\phi \colon V \times W \to Z$, and let $v = v^i e_i \in V$ and $w = w^i \widetilde{e}_i \in W$. Then we have:

$$\phi(v, w) = \phi\left(\sum_{i=1}^{d} v^i e_i, \sum_{j=1}^{\widetilde{d}} w^j \widetilde{e}_j\right) = \sum_{i=1}^{d} \sum_{j=1}^{\widetilde{d}} \phi(v^i e_i, w^j \widetilde{e}_j) = \sum_{i=1}^{d} \sum_{j=1}^{\widetilde{d}} v^i w^j \phi(e_i, \widetilde{e}_j).$$

Note that by suppressing the greyed out summation signs, the second and third term above are indistinguishable. But this is only true if ϕ is bilinear! Hence the summation convention should not be used (at least, not without extra care) in other areas of mathematics.

Remark 2.12. Having chosen a basis for V and the dual basis for V^*, it is very tempting to think of $v = v^i e_i \in V$ and $\omega = \omega_i \epsilon^i \in V^*$ as d-tuples of numbers. In order to distinguish them, one may choose to write vectors as *columns* of numbers and covectors as *rows* of numbers:

$$v = v^i e_i \quad \leftrightsquigarrow \quad v \mathrel{\hat{=}} \begin{pmatrix} v^1 \\ \vdots \\ v^d \end{pmatrix}$$

and

$$\omega = \omega_i \epsilon^i \quad \leftrightsquigarrow \quad \omega \mathrel{\hat{=}} (\omega_1, \dots, \omega_d).$$

Given $\phi \in \mathrm{End}(V) \cong_{\mathrm{vec}} T^1_1 V$, recall that we can write $\phi = \phi^i{}_j \, e_i \otimes \epsilon^j$, where $\phi^i{}_j := \phi(\epsilon^i, e_j)$ are the components of ϕ with respect to the chosen basis. It is then also very tempting to think of ϕ as a square array of numbers:

$$\phi = \phi^i{}_j \, e_i \otimes \epsilon^j \quad \leftrightsquigarrow \quad \phi \mathrel{\hat{=}} \begin{pmatrix} \phi^1{}_1 & \phi^1{}_2 & \cdots & \phi^1{}_d \\ \phi^2{}_1 & \phi^2{}_2 & \cdots & \phi^2{}_d \\ \vdots & \vdots & \ddots & \vdots \\ \phi^d{}_1 & \phi^d{}_2 & \cdots & \phi^d{}_d \end{pmatrix}$$

The convention here is to think of the i index on $\phi^i{}_j$ as a *row index*, and of j as a *column index*.

We cannot stress enough that this is pure convention. Its usefulness stems from the following example.

Example 2.4. If $\dim V < \infty$, then we have $\mathrm{End}(V) \cong_{\mathrm{vec}} T^1_1 V$. Explicitly, if $\phi \in \mathrm{End}(V)$, we can think of $\phi \in T^1_1 V$, using the same symbol, as

$$\phi(\omega, v) := \omega(\phi(v)).$$

Hence the components of $\phi \in \mathrm{End}(V)$ are $\phi^a{}_b := \epsilon^a(\phi(e_b))$.

Now consider $\phi, \psi \in \mathrm{End}(V)$. Let us determine the components of $\phi \circ \psi$. We have:

$$\begin{aligned} (\phi \circ \psi)^a{}_b &:= (\phi \circ \psi)(\epsilon^a, e_b) \\ &:= \epsilon^a((\phi \circ \psi)(e_b)) \\ &= \epsilon^a((\phi(\psi(e_b))) \\ &= \epsilon^a(\phi(\psi^m{}_b \, e_m)) \\ &= \psi^m{}_b \epsilon^a(\phi(e_m)) \\ &:= \psi^m{}_b \, \phi^a{}_m. \end{aligned}$$

The multiplication in the last line is the multiplication in the field K, and since that's commutative, we have $\psi^m{}_b \, \phi^a{}_m = \phi^a{}_m \, \psi^m{}_b$. However, in light of the convention introduced in the previous remark, the latter is preferable. Indeed, if we think of the superscripts as row indices and of the subscripts as column indices, then $\phi^a{}_m \, \psi^m{}_b$ is the entry in row a, column b, of the matrix product $\phi\psi$.

Similarly, $\omega(v) = \omega_m v^m$ can be thought of as the *dot product* $\omega \cdot v \equiv \omega^T v$, and

$$\phi(v, w) = w_a \, \phi^a{}_b \, v^b \quad \leftrightsquigarrow \quad \omega^T \phi v.$$

The last expression is could mislead you into thinking that the transpose is a "good" notion, but in fact it is not. It is very bad notation. It almost pretends to be basis independent, but it is not at all.

The moral of the story is that you should try your best *not* to think of vectors, covectors and tensors as arrays of numbers. Instead, always try to understand them from the abstract, intrinsic, component-free point of view.

As a final note in the notational conventions let's see the change of components under a change of basis using the new notation.

Recall that if $\{e_a\}$ and $\{\widetilde{e}_a\}$ are basis of V, we have

$$\widetilde{e}_a = A^b{}_a e_b \qquad \text{and} \qquad e_a = B^m{}_a \widetilde{e}_m,$$

with $A^{-1} = B$. Note that in index notation, the equation $AB = I$ reads $A^a{}_m B^m{}_b = \delta^a_b$.

We now investigate how the components of vectors and covectors change under a change of basis.

a) Let $v = v^a e_a = \widetilde{v}^a \widetilde{e}_a \in V$. Then:

$$v^a = \epsilon^a(v) = \epsilon^a(\widetilde{v}^b \widetilde{e}_b) = \widetilde{v}^b \epsilon^a(\widetilde{e}_b) = \widetilde{v}^b \epsilon^a(A^m{}_b e_m) = A^m{}_b \widetilde{v}^b \epsilon^a(e_m) = A^a{}_b \widetilde{v}^b.$$

b) Let $\omega = \omega_a \epsilon^a = \widetilde{\omega}_a \widetilde{\epsilon}^a \in V^*$. Then:

$$\omega_a := \omega(e_a) = \omega(B^m{}_a \widetilde{e}_m) = B^m{}_a \omega(\widetilde{e}_m) = B^m{}_a \widetilde{\omega}_m.$$

Summarising, for $v \in V$, $\omega \in V^*$ and $\widetilde{e}_a = A^b{}_a e_b$, we have:

$$\begin{aligned} v^a &= A^a{}_b \widetilde{v}^b & \qquad \omega_a &= B^b{}_a \widetilde{\omega}_b \\ \widetilde{v}^a &= B^a{}_b v^b & \qquad \widetilde{\omega}_a &= A^b{}_a \omega_b \end{aligned}$$

The result for tensors is a combination of the above, depending on the type of tensor.

c) Let $T \in T^p_q V$. Then:

$$T^{a_1 \dots a_p}{}_{b_1 \dots b_q} = A^{a_1}{}_{m_1} \cdots A^{a_p}{}_{m_p} B^{n_1}{}_{b_1} \cdots B^{n_q}{}_{b_q} \widetilde{T}^{m_1 \dots m_p}{}_{n_1 \dots n_q},$$

i.e. the upstair indices transform like vector indices, and the downstair indices transform like covector indices.

2.5 Rings

Definition 2.24 (Ring)**.** *A **ring** is a triple $(R, +, \cdot)$, where R is a set and $+, \cdot \colon R \times R \to R$ are maps satisfying the following axioms*

- *$(R, +)$ is an abelian group:*
  - *i) Closure: $\forall\, a, b \in R : a + b \in R$;*
  - *ii) Associativity: $\forall\, a, b, c \in R : (a + b) + c = a + (b + c)$;*
  - *iii) Neutral Element: $\exists\, 0 \in R : \forall\, a \in R : a + 0 = 0 + a = a$;*
  - *iv) Inverse Element: $\forall\, a \in R : \exists -a \in R : a + (-a) = (-a) + a = 0$;*
  - *v) Commutativity: $\forall\, a, b \in R : a + b = b + a$;*
- *the operation $\cdot$ is closed and associative in $R^* := R \setminus \{0\}$:*
  - *vi) Closure: $\forall\, a, b \in R^* : a \cdot b \in R^*$;*
  - *vii) Associativity: $\forall\, a, b, c \in R^* : (a \cdot b) \cdot c = a \cdot (b \cdot c)$;*
- *the maps $+$ and $\cdot$ satisfy the distributive properties:*
  - *viii) $\forall\, a, b, c \in R : (a + b) \cdot c = a \cdot c + b \cdot c$;*
  - *ix) $\forall\, a, b, c \in R : a \cdot (b + c) = a \cdot b + a \cdot c$.*

Note that since $\cdot$ is not required to be commutative, axioms viii and ix are both necessary. In the case of fields where $\cdot$ was commutative, ix followed from viii and commutativity of $\cdot$.

Definition 2.25 (Commutative / Unital / Division Rings)**.** *A ring $(R, +, \cdot)$ is said to be*

- ***commutative** if $\forall\, a, b \in R : a \cdot b = b \cdot a$;*
- ***unital** if $\exists\, 1 \in R : \forall\, a \in R : 1 \cdot a = a \cdot 1 = a$;*

- *a* ***division*** *(or* ***skew****)* ring *if it is unital and*

$$\forall\, a \in R \setminus \{0\} : \exists\, a^{-1} \in R \setminus \{0\} : \; a \cdot a^{-1} = a^{-1} \cdot a = 1.$$

In a unital ring, an element for which there exists a multiplicative inverse is said to be a *unit.* The set of units of a ring R is denoted by R^* (not to be confused with the vector space dual) and forms a group under multiplication. Then, R is a division ring iff $R^* = R \setminus \{0\}$.

Example 2.5. The sets $\mathbb{Z}$, $\mathbb{Q}$, $\mathbb{R}$, and $\mathbb{C}$ are all rings under the usual operations. They are also all fields, except $\mathbb{Z}$.

In general, if $(A, +, \cdot, \bullet)$ is an algebra, then $(A, +, \bullet)$ is a ring.

2.6 Modules

Definition 2.26 (R-Module)**.** *Let $(R, +, \cdot)$ be a unital ring. An* ***R-module*** *is a triple $(M, \oplus, \odot)$ where M is a set and*

$$\begin{aligned} \oplus &\colon M \times M \to M \\ \odot &\colon \; R \times M \to M \end{aligned}$$

are maps satisfying the following axioms:

- *$(M, \oplus)$ is an abelian group i.e*
  - *i) Closure: $\forall\, m, n \in M : m \oplus n \in M$;*
  - *ii) Associativity: $\forall\, m, n, s \in M : (m \oplus n) \oplus s = m \oplus (n \oplus s)$;*
  - *iii) Neutral Element: $\exists\, 0 \in M : \forall\, m \in M : m \oplus 0 = 0 \oplus m = m$;*
  - *iv) Inverse Element: $\forall\, m \in M : \exists -m \in M : m \oplus (-m) = (-m) \oplus m = 0$;*
  - *v) Commutativity: $\forall\, m, n \in M : m \oplus n = n \oplus m$;*
- *the map $\odot$ is an* action *of R on $(M, \oplus)$:*
  - *vi) Distributivity Of Scalar Multiplication - Vector Addition: $\forall\, r \in R : \forall\, m, n \in M : r \odot (m \oplus n) = (r \odot m) \oplus (r \odot n)$;*
  - *vii) Distributivity Of Scalar Multiplication - Field Addition: $\forall\, r, s \in K : \forall\, m \in V : (r + s) \odot m = (r \odot m) \oplus (s \odot m)$;*
  - *viii) Compatibility Of Scalar Multiplication - Field Multiplication $\forall\, r, s \in R : \forall\, m \in M : (r \cdot s) \odot m = r \odot (s \odot m)$;*
  - *ix) Neutral Element Of Scalar Multiplication $\forall\, m \in M : 1 \odot m = m$.*

So, modules are simply vector spaces over rings instead of fields. For this reason, most definitions we had for vector spaces carry over unaltered to modules.

Example 2.6. Any ring R is trivially a module over itself, in the sense that every field K is a vector space over itself.

In the following, we will usually denote $\oplus$ by $+$ and suppress the $\odot$, as we did with vector spaces.

Definition 2.27 (Direct Sum Of Modules)**.** *The* ***direct sum*** *of two R-modules M and N is the R-module $M \oplus N$, which has $M \times N$ as its underlying set and operations (inherited from M and N) defined component-wise.*

Note that while we have been using $\oplus$ to temporarily distinguish two "plus-like" operations in different spaces, the symbol $\oplus$ is the standard notation for the direct sum.

Definition 2.28 (Finitely Generated / Free / Projective Modules)**.** *An R-module M is said to be*

- ***finitely generated*** *if it has a finite generating set;*
- ***free*** *is it has a basis;*

- ***projective*** *if it is a direct summand of a free R-module F, i.e.*

$$M \oplus Q = F$$

for some R-module Q.

Example 2.7. Clearly, every free module is also projective.

Definition 2.29 (R-Linear Maps)**.** *Let M and N be two R-modules. A map $f\colon M \to N$ is said to be an R-**linear map**, or an R-**module homomorphism**, if*

$$\forall\, r \in R : \forall\, m_1, m_2 \in M : \;\; f(rm_1 + m_2) = rf(m_1) + f(m_2),$$

where it should be clear which operations are in M and which in N.

Definition 2.30 (Module Isomorphisms)**.** *A bijective module homomorphism is said to be a **module isomorphism**.*

Definition 2.31 (Isomorphic Modules)**.** *Two modulesare said to be **isomorphic** if there exists a module isomorphism between them. We write $M \cong_{\text{mod}} N$.*

Proposition 2.3. *If a finitely generated module R-module F is free, and $d \in \mathbb{N}$ is the cardinality of a finite basis, then*

$$F \cong_{\text{mod}} = \underbrace{R \oplus \cdots \oplus R}_{d\ copies} =: R^d.$$

One can show that if $R^d \cong_{\text{mod}} R^{d'}$, then $d = d'$ and hence, the concept of dimension is well-defined for finitely generated, free modules.

Theorem 2.2. *Let P, Q be finitely generated (projective) modules over a commutative ring R. Then*

$$\text{Hom}_R(P, Q) := \{\phi\colon P \xrightarrow{\sim} Q \mid \phi \text{ is } R\text{-linear}\}$$

is again a finitely generated (projective) R-module, with operations defined pointwise.

The proof is exactly the same as with vector spaces. As an example, we can use this to define the dual of a module.

2.6.1 Basis Of Modules

The key fact that sets modules apart from vector spaces is that, unlike a vector space, an R-module need not have a basis, unless R is a division ring. This is actually a well-known theorem that we will state but not prove.

Theorem 2.3. *If D is a division ring, then any D-module V admits a basis.*

Corollary 2.1. *Every vector space has a basis, since any field is also a division ring.*

2.7 Algebras

Definition 2.32 (Algebra)**.** *Let K be a field, and let A be a vector space over K equipped with an additional bilinear map (called binary operation or product) $\bullet\colon A \times A \to A$. The quadruple $(A, +, \cdot, \bullet)$ is called an **algebra** over a field K.*

Definition 2.33 (Associative / Unital / Commutative Algebra)**.** *An algebra $(A, +, \cdot, \bullet)$ is said to be*

i) ***Associative*** *if $\forall\, v, w, z \in A : v \bullet (w \bullet z) = (v \bullet w) \bullet z$;*

ii) ***Unital*** *if $\exists\, \mathbf{1} \in A : \forall\, v \in V : \mathbf{1} \bullet v = v \bullet \mathbf{1} = v$;*

iii) ***Commutative*** *or* abelian *if $\forall\, v, w \in A : v \bullet w = w \bullet v$.*

Definition 2.34 (Derivation)**.** *Let A and B be algebras. A **derivation** on A is a linear map $D\colon A \xrightarrow{\sim} B$ satisfying the Leibniz rule*

$$D(v \bullet_A w) = D(v) \bullet_B w +_B v \bullet_B D(w).$$

for all $v, w \in A$.

2.8 Lie Algebras

An important class of algebras, that we will also see later, are the so-called Lie algebras, in which the product $v \bullet w$ is denoted as $[v, w]$ called the "commutator" or "Lie bracket". Let's first define it, and then use it to define the corresponding Lie algebra.

Definition 2.35 (Lie Bracket)**.** *Let K be a field, A be a vector space over K and $v, w \in A$. The **Lie bracket** (or **commutator**) of v and w defined as*

$$\begin{aligned} [v, w] \colon A \times A &\to A \\ (v, w) &\mapsto [v, w] := vw - wv \end{aligned}$$

Definition 2.36 (Lie Algebra)**.** *A **Lie algebra** A is an algebra whose product $[-, -]$, called* Lie bracket*, satisfies*

i) bilinearity: $A \times A \to A$

ii) antisymmetry: $\forall\, v \in A : [v, v] = 0$;

iii) the Jacobi identity: $\forall\, v, w, z \in A : [v, [w, z]] + [w, [z, v]] + [z, [v, w]] = 0$.

Note that the zeros above represent the additive identity element in A, not the zero scalar

The antisymmetry condition immediately implies $[v, w] = -[w, v]$ for all $v, w \in A$, hence a (non-trivial) Lie algebra cannot be unital.

Example 2.8. Let V be a vector space over K. Then $(\mathrm{End}(V), +, \cdot, \circ)$ (where the product is simply the composition of endomorphisms) is an associative, unital, non-commutative algebra over K. Define:

$$\begin{aligned} [-, -] \colon \mathrm{End}(V) \times \mathrm{End}(V) &\to \mathrm{End}(V) \\ (\phi, \psi) &\mapsto [\phi, \psi] := \phi \circ \psi - \psi \circ \phi. \end{aligned}$$

It is instructive to check that $(\mathrm{End}(V), +, \cdot, [-, -])$ is a Lie algebra over K. In this case, the Lie bracket is typically called the *commutator*.

In general, given an associative algebra $(A, +, \cdot, \bullet)$, if we define

$$[v, w] := v \bullet w - w \bullet v,$$

then $(A, +, \cdot, [-, -])$ is a Lie algebra.

Example 2.9. Consider again the Lie algebra $(\mathrm{End}(V), +, \cdot, [-, -])$ and fix $\xi \in \mathrm{End}(V)$. If we define

$$\begin{aligned} D_\xi := [\xi, -] \colon \mathrm{End}(V) &\xrightarrow{\sim} \mathrm{End}(V) \\ \phi &\mapsto [\xi, \phi], \end{aligned}$$

then D_ξ is a derivation on $(\mathrm{End}(V), +, \cdot, [-, -])$ since it is linear and

$$\begin{aligned} D_\xi([\phi, \psi]) &:= [\xi, [\phi, \psi]] \\ &= -[\psi, [\xi, \phi]] - [\phi, [\psi, \xi]] && \text{(by the Jacobi identity)} \\ &= [[\xi, \phi], \psi] + [\phi, [\xi, \psi]] && \text{(by antisymmetry)} \\ &=: [D_\xi(\phi), \psi] + [\phi, D_\xi(\psi)]. \end{aligned}$$

This construction works in general Lie algebras as well.

Of course one can construct an algebra over a ring, by imposing all the axioms on a module instead of a vector space. Same definitions apply for an algebra over a ring with the appropriate changes when needed.

2.8.1 Classification Of Lie Algebras

One of the most important topics on Lie algebras is the classification of them. While it is possible to classify Lie algebras more generally, we will only consider the classification of finite-dimensional complex Lie algebras, i.e. Lie algebras $(L, [-, -])$ where L is a finite-dimensional $\mathbb{C}$-vector space.

If A, B are Lie subalgebras of a Lie algebra $(L, [-,-])$ over K, then

$$[A, B] := \operatorname{span}_K\big(\{[x, y] \in L \mid x \in A \text{ and } y \in B\}\big)$$

is again a Lie subalgebra of L.

Definition 2.37. *A Lie algebra L is said to be* abelian *if*

$$\forall\, x, y \in L : \ [x, y] = 0.$$

Equivalently, $[L, L] = 0$, where 0 denotes the trivial Lie algebra $\{0\}$.

Abelian Lie algebras are highly non-interesting as Lie algebras: since the bracket is identically zero, it may as well not be there. Even from the classification point of view, the vanishing of the bracket implies that, given any two abelian Lie algebras, every linear isomorphism between their underlying vector spaces is automatically a Lie algebra isomorphism. Therefore, for each $n \in \mathbb{N}$, there is (up to isomorphism) only one abelian n-dimensional Lie algebra.

Definition 2.38. *An* ideal *I of a Lie algebra L is a Lie subalgebra such that $[I, L] \subseteq I$, i.e.*

$$\forall\, x \in I : \forall\, y \in L : \ [x, y] \in I.$$

The ideals 0 and L are called the trivial ideals *of L.*

Definition 2.39. *A Lie algebra L is said to be*

- simple *if it is non-abelian and it contains no non-trivial ideals;*
- semi-simple *if it contains no non-trivial abelian ideals.*

Remark 2.13. Note that any simple Lie algebra is also semi-simple. The requirement that a simple Lie algebra be non-abelian is due to the 1-dimensional abelian Lie algebra, which would otherwise be the only simple Lie algebra which is not semi-simple.

Definition 2.40. *Let L be a Lie algebra. The Lie subalgebra*

$$L' := [L, L]$$

is called the derived subalgebra *of L.*

We can form a sequence of Lie subalgebras

$$L \supseteq L' \supseteq L'' \supseteq \cdots \supseteq L^{(n)} \supseteq \cdots$$

called the *derived series* of L.

Definition 2.41. *A Lie algebra L is* solvable *if there exists $k \in \mathbb{N}$ such that $L^{(k)} = 0$.*

Recall that the direct sum of vector spaces $V \oplus W$ has $V \times W$ as its underlying set and operations defined componentwise.

Definition 2.42. *Let L_1 and L_2 be Lie algebras. The* direct sum *$L_1 \oplus_{\mathrm{Lie}} L_2$ has $L_1 \oplus L_2$ as its underlying vector space and Lie bracket defined as*

$$[x_1 + x_2, y_1 + y_2]_{L_1 \oplus_{\mathrm{Lie}} L_2} := [x_1, y_1]_{L_1} + [x_2, y_2]_{L_2}$$

for all $x_1, y_1 \in L_1$ and $x_2, y_2 \in L_2$. Alternatively, by identifying L_1 and L_2 with the subspaces $L_1 \oplus 0$ and $0 \oplus L_2$ of $L_1 \oplus L_2$ respectively, we require

$$[L_1, L_2]_{L_1 \oplus_{\mathrm{Lie}} L_2} = 0.$$

In the following, we will drop the "Lie" subscript and understand $\oplus$ to mean $\oplus_{\mathrm{Lie}}$ whenever the summands are Lie algebras.

There is a weaker notion than the direct sum, defined only for Lie algebras.

Definition 2.43. *Let R and L be Lie algebras. The* semi-direct sum $R \oplus_s L$ *has* $R \oplus L$ *as its underlying vector space and Lie bracket satisfying*

$$[R, L]_{R \oplus_s L} \subseteq R,$$

i.e. R is an ideal of $R \oplus_s L$.

We are now ready to state Levi's decomposition theorem.

Theorem 2.4 (Levi)**.** *Any finite-dimensional complex Lie algebra L can be decomposed as*

$$L = R \oplus_s (L_1 \oplus \cdots \oplus L_n)$$

where R is a solvable Lie algebra and $L_1, \ldots, L_n$ are simple Lie algebras.

As of today, no general classification of solvable Lie algebras is known, except for some special cases (e.g. in low dimensions). In contrast, the finite dimensional, simple, complex Lie algebras have been classified completely.

Proposition 2.4. *A Lie algebra is semi-simple if, and only if, it can be expressed as a direct sum of simple Lie algebras.*

Hence, the simple Lie algebras are the basic building blocks from which one can build any semi-simple Lie algebra. Then, by Levi's theorem, the classification of simple Lie algebras easily extends to a classification of all semi-simple Lie algebras.

2.8.2 The adjoint map and the Killing form

Definition 2.44. *Let L be a Lie algebra over k and let $x \in L$. The* adjoint map *with respect to x is the K-linear map*

$$\begin{aligned} \mathrm{ad}_x \colon L &\xrightarrow{\sim} L \\ y &\mapsto \mathrm{ad}_x(y) := [x, y]. \end{aligned}$$

The linearity of ad_x follows from the linearity of the bracket in the second argument, while the linearity in the first argument of the bracket implies that the map

$$\begin{aligned} \mathrm{ad} \colon L &\xrightarrow{\sim} \mathrm{End}(L) \\ x &\mapsto \mathrm{ad}(x) := \mathrm{ad}_x \,. \end{aligned}$$

itself is also linear. In fact, more is true. Recall that $\mathrm{End}(L)$ is a Lie algebra with bracket

$$[\phi, \psi] := \phi \circ \psi - \psi \circ \phi.$$

Then, we have the following.

Proposition 2.5. *The map* $\mathrm{ad} \colon L \xrightarrow{\sim} \mathrm{End}(L)$ *is a Lie algebra homomorphism.*

Proof. It remains to check that ad preserves the brackets. Let $x, y, z \in L$. Then

$$\begin{aligned} \mathrm{ad}_{[x,y]}(z) &:= [[x, y], z] && \text{(definition of ad)} \\ &= -[[y, z], x] - [[z, x], y] && \text{(Jacobi's identity)} \\ &= [x, [y, z]] - [y, [x, z]] && \text{(anti-symmetry)} \\ &= \mathrm{ad}_x(\mathrm{ad}_y(z)) - \mathrm{ad}_y(\mathrm{ad}_x(z)) \\ &= (\mathrm{ad}_x \circ \mathrm{ad}_y - \mathrm{ad}_y \circ \mathrm{ad}_x)(z) \\ &= [\mathrm{ad}_x, \mathrm{ad}_y](z). \end{aligned}$$

Hence, we have $\mathrm{ad}([x, y]) = [\mathrm{ad}(x), \mathrm{ad}(y)]$. □

Definition 2.45. *Let L be a Lie algebra over K. The* Killing form *on L is the K-bilinear map*

$$\begin{aligned} \kappa \colon L \times L &\to K \\ (x, y) &\mapsto \kappa(x, y) := \mathrm{tr}(\mathrm{ad}_x \circ \mathrm{ad}_y), \end{aligned}$$

where tr *is the usual trace on the vector space* End(L).

Note that the Killing form is not a "form" in the sense that we defined previously. In fact, since L is finite-dimensional, the trace is cyclic and thus κ is symmetric, i.e.

$$\forall\, x, y \in L : \ \kappa(x, y) = \kappa(y, x).$$

An important property of κ is its associativity with respect to the bracket.

Proposition 2.6. *Let L be a Lie algebra. For any $x, y, z \in L$, we have*

$$\kappa([x, y], z) = \kappa(x, [y, z]).$$

Proof. This follows easily from the fact that ad is a homomorphism.

$$\begin{aligned}
\kappa([x, y], z) &:= \operatorname{tr}(\mathrm{ad}_{[x,y]} \circ \mathrm{ad}_z) \\
&= \operatorname{tr}([\mathrm{ad}_x, \mathrm{ad}_y] \circ \mathrm{ad}_z) \\
&= \operatorname{tr}((\mathrm{ad}_x \circ \mathrm{ad}_y - \mathrm{ad}_y \circ \mathrm{ad}_x) \circ \mathrm{ad}_z) \\
&= \operatorname{tr}(\mathrm{ad}_x \circ \mathrm{ad}_y \circ \mathrm{ad}_z) - \operatorname{tr}(\mathrm{ad}_y \circ \mathrm{ad}_x \circ \mathrm{ad}_z) \\
&= \operatorname{tr}(\mathrm{ad}_x \circ \mathrm{ad}_y \circ \mathrm{ad}_z) - \operatorname{tr}(\mathrm{ad}_x \circ \mathrm{ad}_z \circ \mathrm{ad}_y) \\
&= \operatorname{tr}(\mathrm{ad}_x \circ (\mathrm{ad}_y \circ \mathrm{ad}_z - \mathrm{ad}_z \circ \mathrm{ad}_y)) \\
&= \operatorname{tr}(\mathrm{ad}_x \circ [\mathrm{ad}_y, \mathrm{ad}_z]) \\
&= \operatorname{tr}(\mathrm{ad}_x \circ \mathrm{ad}_{[y,z]}) \\
&=: \kappa(x, [y, z]),
\end{aligned}$$

where we used the cyclicity of the trace. □

We can use κ to give a further equivalent characterisation of semi-simplicity.

Proposition 2.7 (Cartan's criterion)**.** *A Lie algebra L is semi-simple if, and only if, the Killing form κ is non-degenerate, i.e.*

$$(\forall\, y \in L : \kappa(x, y) = 0) \Rightarrow x = 0.$$

Hence, if L is semi-simple, then κ is a pseudo inner product on L. Recall the following definition from linear algebra.

Definition 2.46. *A linear map $\phi \colon V \xrightarrow{\sim} V$ is said to be* symmetric *with respect to the pseudo inner product $B(-,-)$ on V if*

$$\forall\, v, w \in V : \ B(\phi(v), w) = B(v, \phi(w)).$$

If, instead, we have

$$\forall\, v, w \in V : \ B(\phi(v), w) = -B(v, \phi(w)),$$

then ϕ is said to be anti-symmetric *with respect to B.*

The associativity property of κ with respect to the bracket can be restated by saying that, for any $z \in L$, the linear map ad_z is anti-symmetric with respect to κ, i.e.

$$\forall\, x, y \in L : \ \kappa(\mathrm{ad}_z(x), y) = -\kappa(x, \mathrm{ad}_z(y)).$$

In order to do computations, it is useful to introduce a basis $\{E_i\}$ on L.

Definition 2.47. *Let L be a Lie algebra over K and let $\{E_i\}$ be a basis. Then, we have*

$$[E_i, E_j] = C^k{}_{ij} E_k$$

for some $C^k{}_{ij} \in K$. The numbers $C^k{}_{ij}$ are called the structure constants *of L with respect to the basis $\{E_i\}$.*

In terms of the structure constants, the anti-symmetry of the Lie bracket reads

$$C^k{}_{ij} = -C^k{}_{ji}$$

while the Jacobi identity becomes

$$C^n{}_{im}C^m{}_{jk} + C^n{}_{jm}C^m{}_{ki} + C^n{}_{km}C^m{}_{ij} = 0.$$

We can now express both the adjoint maps and the Killing form in terms of components with respect to a basis.

Proposition 2.8. *Let L be a Lie algebra and let $\{E_i\}$ be a basis. Then*

i) $(\mathrm{ad}_{E_i})^k{}_j = C^k{}_{ij}$

ii) $\kappa_{ij} = C^m{}_{ik}C^k{}_{jm}$

where $C^k{}_{ij}$ are the structure constants of L with respect to $\{E_i\}$.

Proof. i) Denote by $\{\varepsilon^i\}$ the dual basis to $\{E_i\}$. Then, we have

$$\begin{aligned}(\mathrm{ad}_{E_i})^k{}_j &:= \varepsilon^k(\mathrm{ad}_{E_i}(E_j))\\ &= \varepsilon^k([E_i, E_j])\\ &= \varepsilon^k(C^m{}_{ij}E_m)\\ &= C^m{}_{ij}\varepsilon^k(E_m)\\ &= C^k{}_{ij},\end{aligned}$$

since $\varepsilon^k(E_m) = \delta^k_m$.

ii) Recall from linear algebra that if V is finite-dimensional, for any $\phi \in \mathrm{End}(V)$ we have $\mathrm{tr}(\phi) = \Phi^k{}_k$, where Φ is the matrix representing the linear map in any basis. Also, recall that the matrix representing $\phi \circ \psi$ is the product $\Phi\Psi$. Using these, we have

$$\begin{aligned}\kappa_{ij} &:= \kappa(E_i, E_j)\\ &= \mathrm{tr}(\mathrm{ad}_{E_i} \circ \mathrm{ad}_{E_j})\\ &= (\mathrm{ad}_{E_i} \circ \mathrm{ad}_{E_j})^k{}_k\\ &= (\mathrm{ad}_{E_i})^m{}_k(\mathrm{ad}_{E_j})^k{}_m\\ &= C^m{}_{ik}C^k{}_{jm},\end{aligned}$$

where we used the same notation for the linear maps and their matrices. □

2.8.3 The fundamental roots and the Weyl group

We will now focus on finite-dimensional semi-simple complex Lie algebras, whose classification hinges on the existence of a special type of subalgebra.

Definition 2.48. *Let L be a d-dimensional Lie algebra. A* Cartan subalgebra *H of L is a maximal Lie subalgebra of L with the following property: there exists a basis $\{h_1, \dots, h_r\}$ of H which can be extended to a basis $\{h_1, \dots, h_r, e_1, \dots, e_{d-r}\}$ of L such that $e_1, \dots, e_{d-r}$ are eigenvectors of $\mathrm{ad}(h)$ for any $h \in H$, i.e.*

$$\forall\, h \in H : \exists\, \lambda_\alpha(h) \in \mathbb{C} : \ \mathrm{ad}(h)e_\alpha = \lambda_\alpha(h)e_\alpha,$$

for each $1 \leq \alpha \leq d - r$.

The basis $\{h_1, \dots, h_r, e_1, \dots, e_{d-r}\}$ is known as a *Cartan-Weyl basis* of L.

Theorem 2.5. *Let L be a finite-dimensional semi-simple complex Lie algebra. Then*

i) L possesses a Cartan subalgebra;

ii) all Cartan subalgebras of L have the same dimension, called the rank *of L;*

iii) any of Cartan subalgebra of L is abelian.

Note that we can think of the λ_α appearing above as a map $\lambda_\alpha \colon H \to \mathbb{C}$. Moreover, for any $z \in \mathbb{C}$ and $h, h' \in H$, we have

$$\begin{aligned}\lambda_\alpha(zh + h')e_\alpha &= \mathrm{ad}(zh + h')e_\alpha \\ &= [zh + h', e_\alpha] \\ &= z[h, e_\alpha] + [h', e_\alpha] \\ &= z\lambda_\alpha(h)e_\alpha + \lambda_\alpha(h')e_\alpha \\ &= (z\lambda_\alpha(h) + \lambda_\alpha(h'))e_\alpha,\end{aligned}$$

Hence λ_α is a $\mathbb{C}$-linear map $\lambda_\alpha \colon H \xrightarrow{\sim} \mathbb{C}$, and thus $\lambda_\alpha \in H^*$.

Definition 2.49. *The maps $\lambda_1, \ldots, \lambda_{d-r} \in H^*$ are called the* roots *of L. The collection*

$$\Phi := \{\lambda_\alpha \mid 1 \leq \alpha \leq d - r\} \subseteq H^*$$

is called the root set *of L.*

One can show that if λ_α were the zero map, then we would have $e_\alpha \in H$. Thus, we must have $0 \notin \Phi$. Note that a consequence of the anti-symmetry of each $\mathrm{ad}(h)$ with respect to the Killing form κ is that

$$\lambda \in \Phi \;\Rightarrow\; -\lambda \in \Phi.$$

Hence Φ is not a linearly independent subset of H^*.

Definition 2.50. *A set of* fundamental roots $\Pi := \{\pi_1, \ldots, \pi_f\}$ *is a subset $\Pi \subseteq \Phi$ such that*

a) Π is a linearly independent subset of H^*;

b) for each $\lambda \in \Phi$, there exist $n_1, \ldots, n_f \in \mathbb{N}$ and $\varepsilon \in \{+1, -1\}$ such that

$$\lambda = \varepsilon \sum_{i=1}^{f} n_i \pi_i.$$

We can write the last equation more concisely as $\lambda \in \mathrm{span}_{\varepsilon,\mathbb{N}}(\Pi)$. Observe that, for any $\lambda \in \Phi$, the coefficients of $\pi_1, \ldots, \pi_f$ in the expansion above always have the same sign. Indeed, we have $\mathrm{span}_{\varepsilon,\mathbb{N}}(\Pi) \neq \mathrm{span}_{\mathbb{Z}}(\Pi)$.

Theorem 2.6. *Let L be a finite-dimensional semi-simple complex Lie algebra. Then*

i) a set $\Pi \subseteq \Phi$ of fundamental roots always exists;

ii) we have $\mathrm{span}_{\mathbb{C}}(\Pi) = H^*$, that is, Π is a basis of H^*.

Corollary 2.2. *We have $|\Pi| = r$, where r is the rank of L.*

Proof. Since Π is a basis, $|\Pi| = \dim H^* = \dim H = r$. □

We would now like to use κ to define a pseudo inner product on H^*. We know from linear algebra that a pseudo inner product $B(-,-)$ on a finite-dimensional vector space V over K induces a linear isomorphism

$$\begin{aligned} i \colon V &\xrightarrow{\sim} V^* \\ v &\mapsto i(v) := B(v, -) \end{aligned}$$

which can be used to define a pseudo inner product $B^*(-,-)$ on V^* as

$$\begin{aligned} B^* \colon V^* \times V^* &\to K \\ (\phi, \psi) &\mapsto B^*(\phi, \psi) := B(i^{-1}(\phi), i^{-1}(\psi)). \end{aligned}$$

We would like to apply this to the restriction of κ to the Cartan subalgebra. However, a pseudo inner product on a vector space is not necessarily a pseudo inner product on a subspace, since the non-degeneracy condition may fail when considered on a subspace.

Proposition 2.9. *The restriction of κ to H is a pseudo inner product on H.*

Proof. Bilinearity and symmetry are automatically satisfied. It remains to show that κ is non-degenerate on H.

i) Let $\{h_1, \dots, h_r, e_{r+1}, \dots, e_d\}$ be a Cartan-Weyl basis of L and let $\lambda_\alpha \in \Phi$. Then

$$\begin{aligned}\lambda_\alpha(h_j)\kappa(h_i, e_\alpha) &= \kappa(h_i, \lambda_\alpha(h_j)e_\alpha)\\ &= \kappa(h_i, [h_j, e_\alpha])\\ &= \kappa([h_i, h_j], e_\alpha)\\ &= \kappa(0, e_\alpha)\\ &= 0.\end{aligned}$$

Since $\lambda_\alpha \neq 0$, there is some h_j such that $\lambda_\alpha(h_j) \neq 0$ and hence

$$\kappa(h_i, e_\alpha) = 0.$$

By linearity, we have $\kappa(h, e_\alpha) = 0$ for any $h \in H$ and any e_α.

ii) Let $h \in H \subseteq L$. Since κ is non-degenerate on L, we have

$$\big(\forall\, x \in L : \kappa(h, x) = 0\big) \Rightarrow h = 0.$$

Expand $x \in L$ in the Cartan-Weyl basis as

$$x = h' + e$$

where $h' := x^i h_i$ and $e := x^\alpha e_\alpha$. Then, we have

$$\kappa(h, x) = \kappa(h, h') + x^\alpha \kappa(h, e_\alpha) = \kappa(h, h').$$

Thus, the non-degeneracy condition reads

$$\big(\forall\, h' \in H : \kappa(h, h') = 0\big) \Rightarrow h = 0,$$

which is what we wanted. □

We can now define

$$\begin{aligned}\kappa^* \colon H^* \times H^* &\to \mathbb{C}\\ (\mu, \nu) &\mapsto \kappa^*(\mu, \nu) := \kappa(i^{-1}(\mu), i^{-1}(\nu)),\end{aligned}$$

where $i \colon H \xrightarrow{\sim} H^*$ is the linear isomorphism induced by κ.

Remark 2.14. If $\{h_i\}$ is a basis of H, the components of κ^* with respect to the dual basis satisfy

$$(\kappa^*)^{ij}\kappa_{jk} = \delta^i_k.$$

Hence, we can write

$$\kappa^*(\mu, \nu) = (\kappa^*)^{ij}\mu_i\nu_j,$$

where $\mu_i := \mu(h_i)$.

We now turn our attention to the real subalgebra $H^*_{\mathbb{R}} := \operatorname{span}_{\mathbb{R}}(\Pi)$. Note that we have the following chain of inclusions

$$\Pi \subseteq \Phi \subseteq \operatorname{span}_{\varepsilon,\mathbb{N}}(\Pi) \subseteq \underbrace{\operatorname{span}_{\mathbb{R}}(\Pi)}_{H^*_{\mathbb{R}}} \subseteq \underbrace{\operatorname{span}_{\mathbb{C}}(\Pi)}_{H^*}.$$

The restriction of κ^* to $H^*_{\mathbb{R}}$ leads to a surprising result.

Theorem 2.7. *i) For any $\alpha, \beta \in H^*_{\mathbb{R}}$, we have $\kappa^*(\alpha, \beta) \in \mathbb{R}$.*

ii) $\kappa^* \colon H^*_{\mathbb{R}} \times H^*_{\mathbb{R}} \to \mathbb{R}$ is an inner product on $H^*_{\mathbb{R}}$.

This is indeed a surprise! Upon restriction to $H^*_{\mathbb{R}}$, instead of being weakened, the non-degeneracy of κ^* gets strengthened to positive definiteness. Now that we have a proper real inner product, we can define some familiar notions from basic linear algebra, such as lengths and angles.

Definition 2.51. *Let $\alpha, \beta \in H^*_{\mathbb{R}}$. Then, we define*

i) the length *of α as $|\alpha| := \sqrt{\kappa^*(\alpha,\alpha)}$;*

ii) the angle *between α and β as $\varphi := \cos^{-1}\left(\dfrac{\kappa^*(\alpha,\beta)}{|\alpha||\beta|}\right)$.*

We need one final ingredient for our classification result.

Definition 2.52. *For any $\lambda \in \Phi \subseteq H^*_{\mathbb{R}}$, define the linear map*

$$\begin{aligned} s_\lambda \colon H^*_{\mathbb{R}} &\xrightarrow{\sim} H^*_{\mathbb{R}} \\ \mu &\mapsto s_\lambda(\mu), \end{aligned}$$

where

$$s_\lambda(\mu) := \mu - 2\frac{\kappa^*(\lambda,\mu)}{\kappa^*(\lambda,\lambda)}\,\lambda.$$

The map s_λ is called a Weyl transformation *and the set*

$$W := \{s_\lambda \mid \lambda \in \Phi\}$$

is a group under composition of maps, called the Weyl group.

Theorem 2.8. *i) The Weyl group W is generated by the fundamental roots in Π, in the sense that for some $1 \le n \le r$, with $r = |\Pi|$,*

$$\forall\, w \in W : \exists\, \pi_1, \ldots, \pi_n \in \Pi :\; w = s_{\pi_1} \circ s_{\pi_2} \circ \cdots \circ s_{\pi_n};$$

ii) Every root can be produced from a fundamental root by the action of W, i.e.

$$\forall\, \lambda \in \Phi : \exists\, \pi \in \Pi : \exists\, w \in W :\; \lambda = w(\pi);$$

iii) The Weyl group permutes the roots, that is,

$$\forall\, \lambda \in \Phi : \forall\, w \in W :\; w(\lambda) \in \Phi.$$

2.8.4 Dynkin diagrams and the Cartan classification

Consider, for any $\pi_i, \pi_j \in \Pi$, the action of the Weyl transformation

$$s_{\pi_i}(\pi_j) := \pi_j - 2\frac{\kappa^*(\pi_i,\pi_j)}{\kappa^*(\pi_i,\pi_i)}\,\pi_i.$$

Since $s_{\pi_i}(\pi_j) \in \Phi$ and $\Phi \subseteq \operatorname{span}_{\varepsilon,\mathbb{N}}(\Pi)$, for all $1 \le i \ne j \le r$ we must have

$$-2\frac{\kappa^*(\pi_i,\pi_j)}{\kappa^*(\pi_i,\pi_i)} \in \mathbb{N}.$$

Definition 2.53. *The* Cartan matrix *of a Lie algebra is the $r \times r$ matrix C with entries*

$$C_{ij} := 2\frac{\kappa^*(\pi_i,\pi_j)}{\kappa^*(\pi_i,\pi_i)},$$

where the C_{ij} should not be confused with the structure constants ${C^k}_{ij}$.

Theorem 2.9. *To every simple finite-dimensional complex Lie algebra there corresponds a unique Cartan matrix and vice versa (up to relabelling of the basis elements).*

Of course, not every matrix can be a Cartan matrix. For instance, since $C_{ii} = 2$ (no summation implied), the diagonal entries of C are all equal to 2, while the off-diagonal entries are either zero or negative. In general, $C_{ij} \neq C_{ji}$, so the Cartan matrix is not symmetric, but if $C_{ij} = 0$, then necessarily $C_{ji} = 0$.
We have thus reduced the problem of classifying the simple finite-dimensional complex Lie algebras to that of finding all the Cartan matrices. This can, in turn, be reduced to the problem of determining all the inequivalent Dynkin diagrams.

Definition 2.54. *Given a Cartan matrix C, the ij-th* bond number *is*

$$n_{ij} := C_{ij}C_{ji} \qquad \textit{(no summation implied).}$$

Note that we have

$$\begin{aligned} n_{ij} &= 4\,\frac{\kappa^*(\pi_i,\pi_j)}{\kappa^*(\pi_i,\pi_i)}\,\frac{\kappa^*(\pi_j,\pi_i)}{\kappa^*(\pi_j,\pi_j)} \\ &= 4\left(\frac{\kappa^*(\pi_i,\pi_j)}{|\pi_i||\pi_j|}\right)^2 \\ &= 4\cos^2\varphi, \end{aligned}$$

where φ is the angle between π_i and π_j. For $i \neq j$, the angle φ is neither zero nor $180°$, hence $0 \leq \cos^2\varphi < 1$, and therefore

$$n_{ij} \in \{0, 1, 2, 3\}.$$

Since $C_{ij} \leq 0$ for $i \neq j$, the only possibilities are

C_{ij}	C_{ji}	n_{ij}
0	0	0
-1	-1	1
-1	-2	2
-1	-3	3

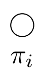
Note that while the Cartan matrices are not symmetric, swapping any pair of C_{ij} and C_{ji} gives a Cartan matrix which represents the same Lie algebra as the original matrix, with two elements from the Cartan-Weyl basis swapped. This is why we have not included $(-2,-1)$ and $(-3,-1)$ in the table above.
If $n_{ij} = 2$ or 3, then the corresponding fundamental roots have different lengths, i.e. either $|\pi_i| < |\pi_j|$ or $|\pi_i| > |\pi_j|$. We also have the following result.

Proposition 2.10. *The roots of a simple Lie algebra have, at most, two distinct lengths.*

The redundancy of the Cartan matrices highlighted above is nicely taken care of by considering Dynkin diagrams.

Definition 2.55. *A* Dynkin diagram *associated to a Cartan matrix is constructed as follows.*

1. *Draw a circle for every fundamental root in $\pi_i \in \Pi$;*



2. *Draw n_{ij} lines between the circles representing the roots π_i and π_j;*

$n_{ij} = 0$ $\qquad$ $n_{ij} = 1$ $\qquad$ $n_{ij} = 2$ $\qquad$ $n_{ij} = 3$

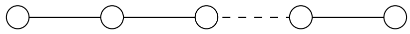
3. *If $n_{ij} = 2$ or 3, draw an arrow on the lines from the longer root to the shorter root.*

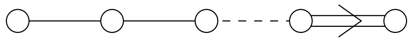
$|\pi_i| > |\pi_j|$ $\qquad$ $|\pi_i| < |\pi_j|$ $\qquad$ $|\pi_i| > |\pi_j|$ $\qquad$ $|\pi_i| < |\pi_j|$

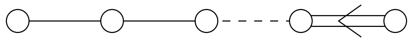
Dynkin diagrams completely characterise any set of fundamental roots, from which we can reconstruct the entire root set by using the Weyl transformations. The root set can then be used to produce a Cartan-Weyl basis.
We are now finally ready to state the much awaited classification theorem.

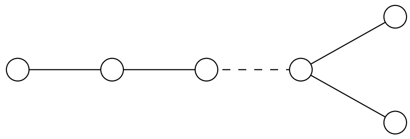
Theorem 2.10 (Killing, Cartan)**.** *Any simple finite-dimensional complex Lie algebra can be reconstructed from its set of fundamental roots* Π*, which only come in the following forms.*

i) There are 4 *infinite families*

A_n $n \geq 1$ 

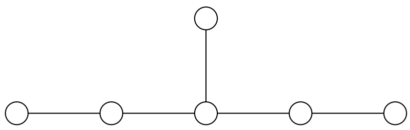
B_n $n \geq 2$ 

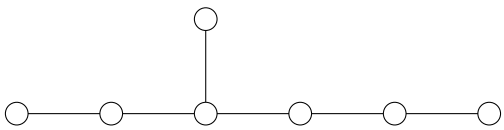
C_n $n \geq 3$ 

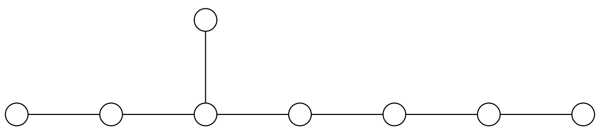
D_n $n \geq 4$ 

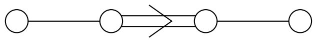
where the restrictions on n *ensure that we don't get repeated diagrams (the diagram* D_2 *is excluded since it is disconnected and does not correspond to a simple Lie algebra)*

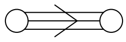
ii) five exceptional cases

E_6 

E_7 

E_8 

F_4 

G_2 

and no other. These are all the possible (connected) Dynkin diagrams.

At last, we have achieved a classification of all simple finite-dimensional complex Lie algebras. The finite-dimensional semi-simple complex Lie algebras are direct sums of simple Lie algebras, and correspond to disconnected Dynkin diagrams whose connected components are the ones listed above.

Chapter 3

Topology

3.1 Topological Spaces

We will now discuss topological spaces based on our previous development of set theory. As we will see, a topology on a set provides the weakest structure in order to define the two very important notions of convergence of sequences to points in a set, and of continuity of maps between two sets. The definition of topology on a set is, at first sight, rather abstract. But on the upside it is also extremely simple. This definition is the result of a historical development, it is the simplest definition of topology that mathematician found to be useful.

Definition 3.1 (Topology)**.** *Let M be a set. A* ***topology*** *on M is a set $\mathcal{O} \subseteq \mathcal{P}(M)$ such that:*

i) $\varnothing \in \mathcal{O}$ *and* $M \in \mathcal{O}$;

ii) $\{U, V\} \subseteq \mathcal{O} \Rightarrow \bigcap\{U, V\} \in \mathcal{O}$;

iii) $C \subseteq \mathcal{O} \Rightarrow \bigcup C \in \mathcal{O}$.

Definition 3.2 (Topological Space)**.** *Let M be a set and $\mathcal{O}$ a topology on the set M. The pair $(M, \mathcal{O})$ is called a* ***topological space****. If we write "let M be a topological space" then some topology $\mathcal{O}$ on M is assumed.*

Remark 3.1. Unless $|M| = 1$, there are (usually many) different topologies $\mathcal{O}$ that one can choose on the set M.

$\lvert M\rvert$	Number of topologies
1	1
2	4
3	29
4	355
5	6,942
6	209,527
7	9,535,241

Example 3.1. Let $M = \{a, b, c\}$. Then $\mathcal{O} = \{\varnothing, \{a\}, \{b\}, \{a, b\}, \{a, b, c\}\}$ is a topology on M since:

i) $\varnothing \in \mathcal{O}$ and $M \in \mathcal{O}$;

ii) Clearly, for any $S \in \mathcal{O}$, $\bigcap\{\varnothing, S\} = \varnothing \in \mathcal{O}$ and $\bigcap\{S, M\} = S \in \mathcal{O}$. Moreover, $\{a\} \cap \{b\} = \varnothing \in \mathcal{O}$, $\{a\} \cap \{a, b\} = \{a\} \in \mathcal{O}$, and $\{b\} \cap \{a, b\} = \{b\} \in \mathcal{O}$;

iii) Let $C \subseteq \mathcal{O}$. If $M \in C$, then $\bigcup C = M \in \mathcal{O}$. If $\{a, b\} \in C$ (or $\{a\}, \{b\} \in C$) but $M \notin C$, then $\bigcup C = \{a, b\} \in \mathcal{O}$. If either $\{a\} \in C$ or $\{b\} \in C$, but $\{a, b\} \notin C$ and $M \notin C$, then $\bigcup C = \{a\} \in \mathcal{O}$ or $\bigcup C = \{b\} \in \mathcal{O}$, respectively. Finally, if none of the above hold, then $\bigcup C = \varnothing \in \mathcal{O}$.

Example 3.2. Let M be a set. Then $\mathcal{O} = \{\varnothing, M\}$ is a topology on M. Indeed, we have:

i) $\varnothing \in \mathcal{O}$ and $M \in \mathcal{O}$;

ii) $\bigcap\{\varnothing, \varnothing\} = \varnothing \in \mathcal{O}$, $\bigcap\{\varnothing, M\} = \varnothing \in \mathcal{O}$, and $\bigcap\{M, M\} = M \in \mathcal{O}$;

iii) If $M \in C$, then $\bigcup C = M \in \mathcal{O}$, otherwise $\bigcup C = \varnothing \in \mathcal{O}$.

This is called the *chaotic topology* and can be defined on any set.

Example 3.3. Let M be a set. Then $\mathcal{O} = \mathcal{P}(M)$ is a topology on M. Indeed, we have:

i) $\varnothing \in \mathcal{P}(M)$ and $M \in \mathcal{P}(M)$;

ii) If $U, V \in \mathcal{P}(M)$, then $\bigcap\{U, V\} \subseteq M$ and hence $\bigcap\{U, V\} \in \mathcal{P}(M)$;

iii) If $C \subseteq \mathcal{P}(M)$, then $\bigcup C \subseteq M$, and hence $\bigcup C \in \mathcal{P}(M)$.

This is called the *discrete topology* and can be defined on any set.

We now give some common terminology regarding topologies.

Definition 3.3 (Coarser / Finer Topology)**.** *Let $\mathcal{O}_1$ and $\mathcal{O}_2$ be two topologies on a set M. If $\mathcal{O}_1 \subset \mathcal{O}_2$, then we say that $\mathcal{O}_1$ is a **coarser** (or* weaker*) topology than $\mathcal{O}_2$. Equivalently, we say that $\mathcal{O}_2$ is a **finer** (or* stronger*) topology than $\mathcal{O}_1$.*

Clearly, the chaotic topology is the coarsest topology on any given set, while the discrete topology is the finest.

Definition 3.4 (Open / Closed Subsets)**.** *Let $(M, \mathcal{O})$ be a topological space. A subset S of M is said to be **open** (with respect to $\mathcal{O}$) if $S \in \mathcal{O}$ and **closed** (with respect to $\mathcal{O}$) if $M \setminus S \in \mathcal{O}$.*

Notice that the notions of open and closed sets, as defined, are not mutually exclusive. A set could be both or neither, or one and not the other.

Example 3.4. Let $(M, \mathcal{O})$ be a topological space. Then $\varnothing$ is open since $\varnothing \in \mathcal{O}$. However, $\varnothing$ is also closed since $M \setminus \varnothing = M \in \mathcal{O}$. Similarly for M.

Example 3.5. Let $M = \{a, b, c\}$ and let $\mathcal{O} = \{\varnothing, \{a\}, \{a, b\}, \{a, b, c\}\}$. Then $\{a\}$ is open but not closed, $\{b, c\}$ is closed but not open, and $\{b\}$ is neither open nor closed.

We will now define what is called the standard topology on $\mathbb{R}^d$, where:

$$\mathbb{R}^d := \underbrace{\mathbb{R} \times \mathbb{R} \times \cdots \times \mathbb{R}}_{d \text{ times}}.$$

We will need the following auxiliary definition.

Definition 3.5 (Open Balls)**.** *For any $x \in \mathbb{R}^d$ and any $r \in \mathbb{R}^+ := \{s \in \mathbb{R} \mid s > 0\}$, we define the **open ball** of radius r around the point x:*

$$B_r(x) := \left\{ y \in \mathbb{R}^d \mid \sqrt{\textstyle\sum_{i=1}^{d} (y_i - x_i)^2} < r \right\},$$

where $x := (x_1, x_2, \ldots, x_d)$ and $y := (y_1, y_2, \ldots, y_d)$, with $x_i, y_i \in \mathbb{R}$.

Remark 3.2. The quantity $\sqrt{\sum_{i=1}^{d} (y_i - x_i)^2}$ is usually denoted by $\|y - x\|_2$, where $\|\cdot\|_2$ is the 2-norm on $\mathbb{R}^d$. However, the definition of a norm on a set requires the set to be equipped with a vector space structure (which we haven't defined yet), while our construction does not. Moreover, our construction can be proven to be independent of the particular norm used to define it, i.e. any other norm will induce the same topological structure.

Definition 3.6 (Standard Topology)**.** *The **standard topology** on $\mathbb{R}^d$, denoted $\mathcal{O}_{\text{std}}$, is defined by:*

$$U \in \mathcal{O}_{\text{std}} :\Leftrightarrow \forall\, p \in U : \exists\, r \in \mathbb{R}^+ : B_r(p) \subseteq U.$$

Of course, simply calling something a topology, does not automatically make it into a topology. We have to prove that $\mathcal{O}_{\text{std}}$ as we defined it, does constitute a topology.

Proposition 3.1. *The pair $(\mathbb{R}^d, \mathcal{O}_{\text{std}})$ is a topological space.*

Proof. i) First, we need to check whether $\varnothing \in \mathcal{O}_{\text{std}}$, i.e. whether is true:

$$\forall\, p \in \varnothing : \exists\, r \in \mathbb{R}^+ : B_r(p) \subseteq \varnothing$$

This proposition is of the form $\forall\, p \in \varnothing : Q(p)$, which was defined as being equivalent to:

$$\forall\, p : p \in \varnothing \Rightarrow Q(p).$$

However, since $p \in \varnothing$ is false, the implication is true independent of p. Hence the initial proposition is true and thus $\varnothing \in \mathcal{O}_{\text{std}}$.
Second, by definition, we have $B_r(x) \subseteq \mathbb{R}^d$ independent of x and r, hence:

$$\forall\, p \in \mathbb{R}^d : \exists\, r \in \mathbb{R}^+ : B_r(p) \subseteq \mathbb{R}^d$$

is true and thus $\mathbb{R}^d \in \mathcal{O}_{\text{std}}$.

ii) Let $U, V \in \mathcal{O}_{\text{std}}$ and let $p \in U \cap V$. Then:

$$p \in U \cap V :\Leftrightarrow p \in U \wedge p \in V$$

and hence, since $U, V \in \mathcal{O}_{\text{std}}$, we have:

$$\exists\, r_1 \in \mathbb{R}^+ : B_{r_1}(p) \subseteq U \quad \wedge \quad \exists\, r_2 \in \mathbb{R}^+ : B_{r_2}(p) \subseteq V.$$

Let $r = \min\{r_1, r_2\}$. Then:

$$B_r(p) \subseteq B_{r_1}(p) \subseteq U \quad \wedge \quad B_r(p) \subseteq B_{r_2}(p) \subseteq V$$

and hence $B_r(p) \subseteq U \cap V$. Therefore $U \cap V \in \mathcal{O}_{\text{std}}$.

iii) Let $C \subseteq \mathcal{O}_{\text{std}}$ and let $p \in \bigcup C$. Then, $p \in U$ for some $U \in C$ and, since $U \in \mathcal{O}_{\text{std}}$, we have:

$$\exists\, r \in \mathbb{R}^+ : B_r(p) \subseteq U \subseteq \bigcup C.$$

Therefore, $\mathcal{O}_{\text{std}}$ is indeed a topology on $\mathbb{R}^d$. □

3.2 Construction Of New Topologies From Given Ones

Definition 3.7 (Induced Topology)**.** *Let $(M, \mathcal{O})$ be a topological space and let $N \subset M$. Then we call the* ***induced topology*** *on N the topology:*

$$\mathcal{O}|_N := \{U \cap N \mid U \in \mathcal{O}\} \subseteq \mathcal{P}(N)$$

Of course we need to prove that this is indeed a topology.

Proof. i) Since $\varnothing \in \mathcal{O}$ and $\varnothing = \varnothing \cap N$, we have $\varnothing \in \mathcal{O}|_N$. Similarly, we have $M \in \mathcal{O}$ and $N = M \cap N$, and thus $N \in \mathcal{O}|_N$.

ii) Let $U, V \in \mathcal{O}|_N$. Then, by definition:

$$\exists\, S \in \mathcal{O} : U = S \cap N \quad \wedge \quad \exists\, T \in \mathcal{O} : V = T \cap N.$$

We thus have:

$$U \cap V = (S \cap N) \cap (T \cap N) = (S \cap T) \cap N.$$

Since $S, T \in \mathcal{O}$ and $\mathcal{O}$ is a topology, we have $S \cap T \in \mathcal{O}$ and hence $U \cap V \in \mathcal{O}|_N$.

iii) Let $C := \{S_\alpha \mid \alpha \in \mathcal{A}\} \subseteq \mathcal{O}|_N$. By definition, we have:

$$\forall\, \alpha \in \mathcal{A} : \exists\, U_\alpha \in \mathcal{O} : S_\alpha = U_\alpha \cap N.$$

Then, using the notation:

$$\bigcup_{\alpha\in\mathcal{A}} S_\alpha := \bigcup C = \bigcup\{S_\alpha \mid \alpha \in \mathcal{A}\}$$

and De Morgan's law, we have:

$$\bigcup_{\alpha\in\mathcal{A}} S_\alpha = \bigcup_{\alpha\in\mathcal{A}} (U_\alpha \cap N) = \Big(\bigcup_{\alpha\in\mathcal{A}} U_\alpha\Big) \cap N.$$

Since $\mathcal{O}$ is a topology, we have $\bigcup_{\alpha\in\mathcal{A}} U_\alpha \in \mathcal{O}$ and hence $\bigcup C \in \mathcal{O}|_N$.
Thus $\mathcal{O}|_N$ is a topology on N.

□

Example 3.6. Consider $(\mathbb{R}, \mathcal{O}_{\text{std}})$ and let:

$$N = [-1, 1] := \{x \in \mathbb{R} \mid -1 \leq x \leq 1\}.$$

Then $(N, \mathcal{O}_{\text{std}}|_N)$ is a topological space. The set $(0, 1]$ is clearly not open in $(\mathbb{R}, \mathcal{O}_{\text{std}})$ since $(0, 1] \notin \mathcal{O}_{\text{std}}$. However, we have:

$$(0, 1] = (0, 2) \cap [-1, 1]$$

where $(0, 2) \in \mathcal{O}_{\text{std}}$ and hence $(0, 1] \in \mathcal{O}_{\text{std}}|_N$, i.e. the set $(0, 1]$ is open in $(N, \mathcal{O}_{\text{std}}|_N)$.

Definition 3.8 (Quotient Topology)**.** *Let $(M, \mathcal{O})$ be a topological space and let $\sim$ be an equivalence relation on M. Then, the quotient set:*

$$M/\!\sim\; = \{[m] \in \mathcal{P}(M) \mid m \in M\}$$

*can be equipped with the **quotient topology** $\mathcal{O}_{M/\sim}$ defined by:*

$$\mathcal{O}_{M/\sim} := \{U \in M/\!\sim \;\mid \bigcup U = \bigcup_{[a]\in U} [a] \in \mathcal{O}\}.$$

An equivalent definition of the quotient topology is as follows. Let $q\colon M \to M/\!\sim$ be the map:

$$\begin{aligned} q\colon M &\to M/\!\sim \\ m &\mapsto [m] \end{aligned}$$

Then we have:

$$\mathcal{O}_{M/\sim} := \{U \in M/\!\sim \;\mid \text{preim}_q(U) \in \mathcal{O}\}.$$

Example 3.7. The *circle* (or 1-sphere) is defined as the set $S^1 := \{(x, y) \in \mathbb{R}^2 \mid x^2 + y^2 = 1\}$ equipped with the subset topology inherited from $\mathbb{R}^2$. The open sets of the circle are (unions of) open arcs, i.e. arcs without the endpoints. Individual points on the circle are clearly not open since there is no open set of $\mathbb{R}^2$ whose intersection with the circle is a single point. However, an individual point on the circle is a closed set since its complement is an open arc.
An alternative definition of the circle is the following. Let $\sim$ be the equivalence relation on $\mathbb{R}$ defined by:

$$x \sim y :\Leftrightarrow \exists\, n \in \mathbb{Z} : x = y + 2\pi n.$$

Then the circle can be defined as the set $S^1 := \mathbb{R}/\!\sim$ equipped with the quotient topology.

Definition 3.9 (Product Topology)**.** *Let $(A, \mathcal{O}_A)$ and $(B, \mathcal{O}_B)$ be topological spaces. Then a topology on $A \times B$ is defined by the set $\mathcal{O}_{A\times B}$ called the **product topology** as:*

$$U \in \mathcal{O}_{A\times B} :\Leftrightarrow \forall\, p \in U : \exists\, (S, T) \in \mathcal{O}_A \times \mathcal{O}_B : S \times T \subseteq U$$

Remark 3.3. This definition can easily be extended to n-fold cartesian products:

$$U \in \mathcal{O}_{A_1\times\cdots\times A_n} :\Leftrightarrow \forall\, p \in U : \exists\, (S_1, \ldots, S_n) \in \mathcal{O}_{A_1} \times \cdots \times \mathcal{O}_{A_n} : S_1 \times \cdots \times S_n \subseteq U.$$

Remark 3.4. Using the previous definition, one can check that the standard topology on $\mathbb{R}^d$ satisfies:

$$\mathcal{O}_{\text{std}} = \mathcal{O}_{\underbrace{\mathbb{R}\times\mathbb{R}\times\cdots\times\mathbb{R}}_{d \text{ times}}}.$$

Therefore, a more minimalistic definition of the standard topology on $\mathbb{R}^d$ would consist in defining $\mathcal{O}_{\text{std}}$ only for $\mathbb{R}$ (i.e. $d = 1$) and then extending it to $\mathbb{R}^d$ by the product topology.

3.3 Convergence & Continuity

Definition 3.10 (Sequence)**.** *Let M be a set. A* ***sequence*** *(of points) in M is a function $q\colon \mathbb{N} \to M$.*

Definition 3.11 (Convergence)**.** *Let $(M, \mathcal{O})$ be a topological space. A sequence q in M is said to* ***converge*** *against a* limit point *$a \in M$ if:*

$$\forall\, U \in \mathcal{O} : a \in U \Rightarrow \exists\, N \in \mathbb{N} : \forall\, n > N : q(n) \in U.$$

Remark 3.5. An open set U of M such that $a \in U$ is called an *open neighbourhood* of a. If we denote this by $U(a)$, then the previous definition of convergence can be rewritten as:

$$\forall\, U(a) : \exists\, N \in \mathbb{N} : \forall\, n > N : q(n) \in U.$$

Example 3.8. Consider the topological space $(M, \{\varnothing, M\})$. Then every sequence in M converges to every point in M. Indeed, let q be any sequence and let $a \in M$. Then, q converges against a if:

$$\forall\, U \in \{\varnothing, M\} : a \in U \Rightarrow \exists\, N \in \mathbb{N} : \forall\, n > N : q(n) \in U.$$

This proposition is vacuously true for $U = \varnothing$, while for $U = M$ we have $q(n) \in M$ independent of n. Therefore, the (arbitrary) sequence q converges to the (arbitrary) point $a \in M$.

Example 3.9. Consider the topological space $(M, \mathcal{P}(M))$. Then only definitely constant sequences converge, where a sequence is *definitely constant* with value $c \in M$ if:

$$\exists\, N \in \mathbb{N} : \forall\, n > N : q(n) = c.$$

This is immediate from the definition of convergence since in the discrete topology all singleton sets (i.e. one-element sets) are open.

Example 3.10. Consider the topological space $(\mathbb{R}^d, \mathcal{O}_{\text{std}})$. Then, a sequence $q\colon \mathbb{N} \to \mathbb{R}^d$ converges against $a \in \mathbb{R}^d$ if:

$$\forall\, \varepsilon > 0 : \exists\, N \in \mathbb{N} : \forall\, n > N : \|q(n) - a\|_2 < \varepsilon.$$

Example 3.11. Let $M = \mathbb{R}$ and let $q = 1 - \frac{1}{n+1}$. Then, since q is not definitely constant, it is not convergent in $(\mathbb{R}, \mathcal{P}(\mathbb{R}))$, but it is convergent in $(\mathbb{R}, \mathcal{O}_{\text{std}})$.

Definition 3.12 (Continuity)**.** *Let $(M, \mathcal{O}_M)$ and $(N, \mathcal{O}_N)$ be topological spaces and let $\phi\colon M \to N$ be a map. Then, ϕ is said to be* ***continuous*** *(with respect to the topologies $\mathcal{O}_M$ and $\mathcal{O}_N$) if:*

$$\forall\, S \in \mathcal{O}_N\,,\ \operatorname{preim}_\phi(S) \in \mathcal{O}_M,$$

where $\operatorname{preim}_\phi(S) := \{m \in M : \phi(m) \in S\}$ *is the pre-image of S under the map ϕ.*

Informally, one says that ϕ is continuous if the pre-images of open sets are open.

Example 3.12. If M is equipped with the discrete topology, or N with the chaotic topology, then any map $\phi\colon M \to N$ is continuous. Indeed, let $S \in \mathcal{O}_N$. If $\mathcal{O}_M = \mathcal{P}(M)$ (and $\mathcal{O}_N$ is any topology), then we have:

$$\operatorname{preim}_\phi(S) = \{m \in M : \phi(m) \in S\} \subseteq M \in \mathcal{P}(M) = \mathcal{O}_M.$$

If instead $\mathcal{O}_N = \{\varnothing, N\}$ (and $\mathcal{O}_M$ is any topology), then either $S = \varnothing$ or $S = N$ and thus, we have:

$$\operatorname{preim}_\phi(\varnothing) = \varnothing \in \mathcal{O}_M \quad \text{and} \quad \operatorname{preim}_\phi(N) = M \in \mathcal{O}_M.$$

Example 3.13. Let $M = \{a, b, c\}$ and $N = \{1, 2, 3\}$, with respective topologies:

$$\mathcal{O}_M = \{\varnothing, \{b\}, \{a, c\}, \{a, b, c\}\} \quad \text{and} \quad \mathcal{O}_N = \{\varnothing, \{2\}, \{3\}, \{1, 3\}, \{2, 3\}, \{1, 2, 3\}\},$$

and let $\phi \colon M \to N$ by defined by:

$$\phi(a) = 2, \quad \phi(b) = 1, \quad \phi(c) = 2.$$

Then ϕ is continuous. Indeed, we have:

$$\begin{aligned} &\operatorname{preim}_\phi(\varnothing) = \varnothing, && \operatorname{preim}_\phi(\{2\}) = \{a, c\}, && \operatorname{preim}_\phi(\{3\}) = \varnothing, \\ &\operatorname{preim}_\phi(\{1, 3\}) = \{b\}, && \operatorname{preim}_\phi(\{2, 3\}) = \{a, c\}, && \operatorname{preim}_\phi(\{1, 2, 3\}) = \{a, b, c\}, \end{aligned}$$

and hence $\operatorname{preim}_\phi(S) \in \mathcal{O}_M$ for all $S \in \mathcal{O}_N$.

Example 3.14. Consider $(\mathbb{R}^d, \mathcal{O}_{\text{std}})$ and $(\mathbb{R}^s, \mathcal{O}_{\text{std}})$. Then $\phi \colon \mathbb{R}^d \to \mathbb{R}^s$ is continuous with respect to the standard topologies if it satisfies the usual ε-δ definition of continuity:

$$\forall\, a \in \mathbb{R}^d : \forall\, \varepsilon > 0 : \exists\, \delta > 0 : \forall\, 0 < \|x - a\|_2 < \delta : \|\phi(x) - \phi(a)\|_2 < \varepsilon.$$

Definition 3.13 (Homeomorphism)**.** *Let $(M, \mathcal{O}_M)$ and $(N, \mathcal{O}_N)$ be topological spaces. A bijection $\phi \colon M \to N$ is called a **homeomorphism** if both $\phi \colon M \to N$ and $\phi^{-1} \colon N \to M$ are continuous.*

Remark 3.6. Homeo(morphism)s are the structure-preserving maps in topology.

If there exists a homeomorphism ϕ between $(M, \mathcal{O}_M)$ and $(N, \mathcal{O}_N)$,

$$\begin{tikzcd} M \arrow[r, bend left, "\phi"] & N \arrow[l, bend left, "\phi^{-1}"] \end{tikzcd}$$

then ϕ provides a one-to-one pairing of the open sets of M with the open sets of N.

Definition 3.14 (Isomorphic Topological Spaces)**.** *If there exists a homeomorphism between two topological spaces $(M, \mathcal{O}_M)$ and $(N, \mathcal{O}_N)$, we say that the two spaces are **homeomorphic** or **topologically isomorphic** and we write $(M, \mathcal{O}_M) \cong_{\text{top}} (N, \mathcal{O}_N)$.*

Clearly, if $(M, \mathcal{O}_M) \cong_{\text{top}} (N, \mathcal{O}_N)$, then $M \cong_{\text{set}} N$.

3.4 Invariant Topological Properties

Definition 3.15 (Invariant Topological Properties)**.** *A property of a topological space is called an **invariant** if any two homeomorphic topological spaces share the property.*

In this section we will mention some of the (almost uncountable) invariant topological properties of topological spaces. A *classification* of topological spaces would be a list of topological invariants such that any two spaces which share these invariants are homeomorphic. As of now, no such list is known!

3.4.1 Separation Properties

Definition 3.16 (T1 Topological Space)**.** *A topological space $(M, \mathcal{O})$ is said to be **T1** if for any two distinct points $p, q \in M$, $p \neq q$:*

$$\exists\, U(p) \in \mathcal{O} : q \notin U(p).$$

Definition 3.17 (T2 or Hausdorff Topological Space)**.** *A topological space $(M, \mathcal{O})$ is said to be **T2** or **Hausdorff** if, for any two distinct points, there exist non-intersecting open neighbourhoods of these two points:*

$$\forall\, p, q \in M : p \neq q \Rightarrow \exists\, U(p), V(q) \in \mathcal{O} : U(p) \cap V(q) = \varnothing.$$

Example 3.15. The topological space $(\mathbb{R}^d, \mathcal{O}_{\text{std}})$ is T2 and hence also T1.

Example 3.16. The Zariski topology on an algebraic variety is T1 but not T2.

Example 3.17. The topological space $(M, \{\varnothing, M\})$ does not have the T1 property since for any $p \in M$, the only open neighbourhood of p is M and for any other $q \neq p$ we have $q \in M$. Moreover, since this space is not T1, it cannot be T2 either.

Remark 3.7. There are many other "T" properties, including a *T2½* property which differs from T2 in that the neighbourhoods are closed.

Definition 3.18 (Cover)**.** *Let $(M, \mathcal{O})$ be a topological space. A set $C \subseteq \mathcal{P}(M)$ is called a **cover** (of M) if:*

$$\bigcup C = M.$$

Additionally, it is said to an open *cover if $C \subseteq \mathcal{O}$.*

Definition 3.19 (Open Cover)**.** *Let $(M, \mathcal{O})$ be a topological space. A cover $C \subseteq \mathcal{P}(M)$ is said to be an **open cover** if $C \subseteq \mathcal{O}$.*

Definition 3.20 (Subcover)**.** *Let C be a cover. Then any subset $\widetilde{C} \subseteq C$ such that $\widetilde{C}$ is still a cover, is called a **subcover**. Additionally, it is said to be a* finite *subcover if it is finite as a set.*

Definition 3.21 (Compact Topological Space)**.** *A topological space $(M, \mathcal{O})$ is said to be **compact** if every open cover has a finite subcover.*

Definition 3.22 (Compact Subset)**.** *Let $(M, \mathcal{O})$ be a topological space. A subset $N \subseteq M$ is called **compact** if the topological space $(N, \mathcal{O}|_N)$ is compact.*

Determining whether a set is compact or not is not an easy task. Fortunately though, for $\mathbb{R}^d$ equipped with the standard topology $\mathcal{O}_{\text{std}}$, the following theorem greatly simplifies matters.

Theorem 3.1 (Heine-Borel)**.** *Let $\mathbb{R}^d$ be equipped with the standard topology $\mathcal{O}_{\text{std}}$. Then, a subset of $\mathbb{R}^d$ is compact if, and only if, it is closed and bounded.*

A subset S of $\mathbb{R}^d$ is said to be *bounded* if:

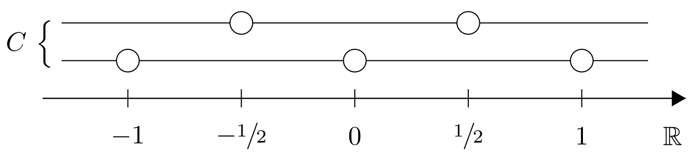
$$\exists\, r \in \mathbb{R}^+ : S \subseteq B_r(0).$$

Example 3.18. The interval $[0, 1]$ is compact in $(\mathbb{R}, \mathcal{O}_{\text{std}})$. The one-element set containing $(-1, 2)$ is a cover of $[0, 1]$, but it is also a finite subcover and hence $[0, 1]$ is compact from the definition. Alternatively, $[0, 1]$ is clearly closed and bounded, and hence it is compact by the Heine-Borel theorem.

Example 3.19. The set $\mathbb{R}$ is not compact in $(\mathbb{R}, \mathcal{O}_{\text{std}})$. To prove this, it suffices to show that there exists a cover of $\mathbb{R}$ that does not have a finite subcover. To this end, let:

$$C := \{(n, n+1) \mid n \in \mathbb{Z}\} \cup \{(n + \tfrac{1}{2}, n + \tfrac{3}{2}) \mid n \in \mathbb{Z}\}.$$

This corresponds to the following picture.



It is clear that removing even one element from C will cause C to fail to be an open cover of $\mathbb{R}$. Therefore, there is no finite subcover of C and hence, $\mathbb{R}$ is not compact.

Theorem 3.2. *Let $(M, \mathcal{O}_M)$ and $(N, \mathcal{O}_N)$ be compact topological spaces. Then $(M \times N, \mathcal{O}_{M\times N})$ is a compact topological space.*

The above theorem easily extends to finite cartesian products.

Definition 3.23 (Refinement)**.** *Let $(M, \mathcal{O})$ be a topological space and let C be a cover. A **refinement** of C is a cover R such that:*

$$\forall\, U \in R : \exists\, V \in C : U \subseteq V.$$

Any subcover of a cover is a refinement of that cover, but the converse is not true in general. A refinement R is said to be:

- *open* if $R \subseteq \mathcal{O}$;
- *locally finite* if for any $p \in M$ there exists a neighbourhood $U(p)$ such that the set:

$$\{U \in R \mid U \cap U(p) \neq \varnothing\}$$

  is finite as a set.

Compactness is a very strong property. Hence often times it does not hold, but a weaker and still useful property, called paracompactness, may still hold.

Definition 3.24 (Paracompact Topological Space)**.** *A topological space $(M, \mathcal{O})$ is said to be **paracompact** if every open cover has an open refinement that is locally finite.*

Corollary 3.1. *If a topological space is compact, then it is also paracompact.*

Remark 3.8. Paracompactness is, informally, a rather natural property since every example of a non-paracompact space looks artificial. One such example is the *long line* (or *Alexandroff line*). To construct it, we first observe that we could "build" $\mathbb{R}$ by taking the interval $[0, 1)$ and stacking countably many copies of it one after the other. Hence, in a sense, $\mathbb{R}$ is equivalent to $\mathbb{Z} \times [0, 1)$. The long line L is defined analogously as $L : \omega_1 \times [0, 1)$, where ω_1 is an uncountably infinite set. The resulting space L is not paracompact.

Theorem 3.3. *Let $(M, \mathcal{O}_M)$ be a paracompact space and let $(N, \mathcal{O}_N)$ be a compact space. Then $M \times N$ (equipped with the product topology) is paracompact.*

Corollary 3.2. *Let $(M, \mathcal{O}_M)$ be a paracompact space and let $(N_i, \mathcal{O}_{N_i})$ be compact spaces for every $1 \leq i \leq n$. Then $M \times N_1 \times \cdots \times N_n$ is paracompact.*

Definition 3.25 (Partition Of Unity)**.** *Let $(M, \mathcal{O}_M)$ be a topological space. A **partition of unity** of M is a set $\mathcal{F}$ of continuous maps from M to the interval $[0, 1]$ such that for each $p \in M$ the following conditions hold:*

i) there exists $U(p)$ such that the set $\{f \in \mathcal{F} \mid \forall\, x \in U(p) : f(x) \neq 0\}$ is finite;

ii) $\sum_{f \in \mathcal{F}} f(p) = 1$.

If C is an open cover, then $\mathcal{F}$ is said to be subordinate *to the cover C if:*

$$\forall\, f \in \mathcal{F} : \exists U \in C : f(x) \neq 0 \Rightarrow x \in U.$$

Theorem 3.4. *Let $(M, \mathcal{O}_M)$ be a Hausdorff topological space. Then $(M, \mathcal{O}_M)$ is paracompact if, and only if, every open cover admits a partition of unity subordinate to that cover.*

3.4.2 Connectedness And Path-Connectedness

Definition 3.26 (Connected Topological Space)**.** *A topological space $(M, \mathcal{O})$ is said to be **connected** unless there exist two non-empty, non-intersecting open sets A and B such that $M = A \cup B$.*

Example 3.20. Consider $(\mathbb{R} \setminus \{0\}, \mathcal{O}_{\text{std}}|_{\mathbb{R}\setminus\{0\}})$, i.e. $\mathbb{R} \setminus \{0\}$ equipped with the subset topology inherited from $\mathbb{R}$. This topological space is not connected since $(-\infty, 0)$ and $(0, \infty)$ are open, non-empty, non-intersecting sets such that $\mathbb{R} \setminus \{0\} = (-\infty, 0) \cup (0, \infty)$.

Theorem 3.5. *The interval $[0, 1] \subseteq \mathbb{R}$ equipped with the subset topology is connected.*

Theorem 3.6. *A topological space $(M, \mathcal{O})$ is connected if, and only if, the only subsets that are both open and closed are $\varnothing$ and M.*

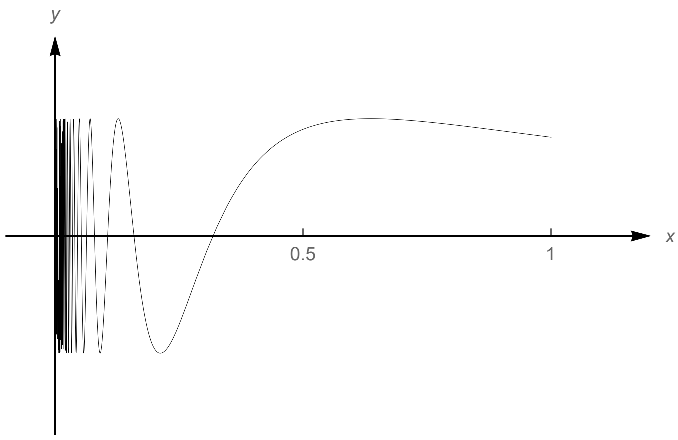
Definition 3.27 (Path - Connected Topological Space)**.** *A topological space $(M, \mathcal{O})$ is said to be **path-connected** if for every pair of points $p, q \in M$ there exists a continuous curve $\gamma\colon [0, 1] \to M$ such that $\gamma(0) = p$ and $\gamma(1) = q$.*

Example 3.21. The space $(\mathbb{R}^d, \mathcal{O}_{\text{std}})$ is path-connected. Indeed, let $p, q \in \mathbb{R}^d$ and let:

$$\gamma(\lambda) := p + \lambda(q - p).$$

Then γ is continuous and satisfies $\gamma(0) = p$ and $\gamma(1) = q$.

Example 3.22. Let $S := \{(x, \sin(\frac{1}{x})) \mid x \in (0,1]\} \cup \{(0,0)\}$ be equipped with the subset topology inherited from $\mathbb{R}^2$.



The space $(S, \mathcal{O}_{\text{std}}|_S)$ is connected but not path-connected.

Theorem 3.7. *If a topological space is path-connected, then it is also connected.*

3.4.3 Homotopic Curves And The Fundamental Group

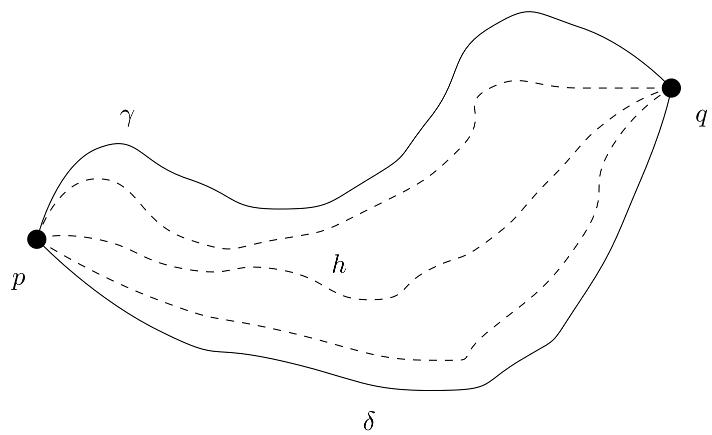
Definition 3.28 (Homotopic Curves)**.** *Let $(M, \mathcal{O})$ be a topological space. Two curves $\gamma, \delta \colon [0,1] \to M$ such that:*

$$\gamma(0) = \delta(0) \quad \textit{and} \quad \gamma(1) = \delta(1)$$

*are said to be **homotopic** if there exists a continuous map $h \colon [0,1] \times [0,1] \to M$ such that for all $\lambda \in [0,1]$:*

$$h(0, \lambda) = \gamma(\lambda) \quad \textit{and} \quad h(1, \lambda) = \delta(\lambda).$$

Pictorially, two curves are homotopic if they can be continuously deformed into one another.



Proposition 3.2. *Let $\gamma \sim \delta :\Leftrightarrow$ "γ and δ are homotopic". Then, $\sim$ is an equivalence relation.*

Definition 3.29 (Space Of Loops)**.** *Let $(M, \mathcal{O})$ be a topological space. Then, for every $p \in M$, we define the **space of loops** at p by:*

$$\mathscr{L}_p := \{\gamma \colon [0,1] \to M \mid \gamma \textit{ is continuous and } \gamma(0) = \gamma(1)\}.$$

Definition 3.30 (Concatenation). *Let $\mathscr{L}_p$ be the space of loops at $p \in M$. We define the* ***concatenation*** *operation $*\colon \mathscr{L}_p \times \mathscr{L}_p \to \mathscr{L}_p$ by:*

$$(\gamma * \delta)(\lambda) := \begin{cases} \gamma(2\lambda) & \text{if } 0 \leq \lambda \leq \frac{1}{2} \\ \delta(2\lambda - 1) & \text{if } \frac{1}{2} \leq \lambda \leq 1 \end{cases}$$

Definition 3.31 (Fundamental Group). *Let $(M, \mathcal{O})$ be a topological space. The* ***fundamental group*** *$\pi_1(p)$ of $(M, \mathcal{O})$ at $p \in M$ is the set:*

$$\pi_1(p) := \mathscr{L}_p/\sim \, = \{[\gamma] \mid \gamma \in \mathscr{L}_p\},$$

where $\sim$ is the homotopy equivalence relation, together with the map

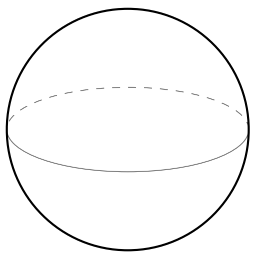
$$\begin{aligned} \bullet\colon \pi_1(p) \times \pi_1(p) &\to \pi_1(p) \\ (\gamma, \delta) &\mapsto [\gamma] \bullet [\delta] := [\gamma * \delta]. \end{aligned}$$

Observe that while all the previously discussed topological properties are "boolean-valued", i.e. a topological space is either Hausdorff or not Hausdorff, either connected or not connected, and so on, the fundamental group is a "group-valued" property, i.e. the value of the property is not "either yes or no", but a group.

Example 3.23. The 2-sphere is defined as the set:

$$S^2 := \{(x, y, z) \in \mathbb{R}^3 \mid x^2 + y^2 + z^2 = 1\}$$

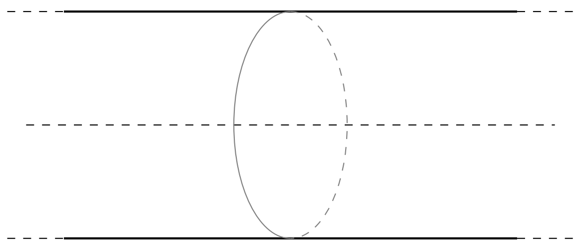
equipped with the subset topology inherited from $\mathbb{R}^3$.



The sphere has the property that all the loops at any point are homotopic, hence the fundamental group (at every point) of the sphere is the trivial group:

$$\forall\, p \in S^2 : \pi_1(p) = 1 := \{[\gamma_e]\}.$$

Example 3.24. The cylinder is defined as $C := \mathbb{R} \times S^1$ equipped with the product topology.

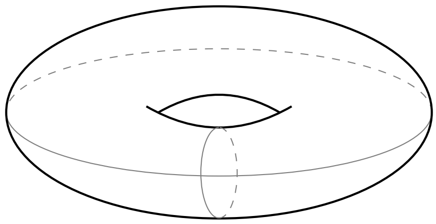


A loop in C can either go around the cylinder (i.e. around its central axis) or not. If it does not, then it can be continuously deformed to a point (the identity loop). If it does, then it cannot be deformed to the identity loop (intuitively because the cylinder is infinitely long) and hence it is a homotopically different loop. The number of times a loop winds around the cylinder is called the *winding number*. Loops with different winding numbers are not homotopic.

Moreover, loops with different *orientations* are also not homotopic and hence we have:

$$\forall\, p \in C : (\pi_1(p), \bullet) \cong_{\mathrm{grp}} (\mathbb{Z}, +).$$

Example 3.25. The 2-torus is defined as the set $T^2 := S^1 \times S^1$ equipped with the product topology.



A loop in T^2 can intuitively wind around the cylinder-like part of the torus as well as around the hole of the torus. That is, there are two independent winding numbers and hence:

$$\forall\, p \in T^2 : \pi_1(p) \cong_{\mathrm{grp}} \mathbb{Z} \times \mathbb{Z},$$

where $\mathbb{Z} \times \mathbb{Z}$ is understood as a group under pairwise addition.

Chapter 4

Topological Manifolds

4.1 Topological Manifolds

Definition 4.1 (Topological Manifold). *A paracompact, Hausdorff, topological space $(M, \mathcal{O})$ is called a d-**dimensional topological manifold** if for every point $p \in M$ there exist a neighbourhood $U(p)$ and a homeomorphism $x\colon U(p) \to x(U(p)) \subseteq \mathbb{R}^d$. We also write* $\dim M = d$.

Intuitively, a d-dimensional manifold is a topological space which locally (i.e. around each point) looks like $\mathbb{R}^d$. Note that, strictly speaking, what we have just defined are *real* topological manifolds. We could define *complex* topological manifolds as well, simply by requiring that the map x be a homeomorphism onto an open subset of $\mathbb{C}^d$.

Proposition 4.1. *Let M be a d-dimensional manifold and let $U, V \subseteq M$ be open, with $U \cap V \neq \varnothing$. If x and y are two homeomorphisms*

$$x\colon U \to x(U) \subseteq \mathbb{R}^d \qquad \text{and} \qquad y\colon V \to y(V) \subseteq \mathbb{R}^{d'},$$

then $d = d'$.

This ensures that the concept of dimension is indeed well-defined, i.e. it is the same at every point, at least on each connected component of the manifold.

Example 4.1. Trivially, $\mathbb{R}^d$ is a d-dimensional manifold for any $d \geq 1$. The space S^1 is a 1-dimensional manifold while the spaces S^2, C and T^2 are 2-dimensional manifolds.

Definition 4.2 (Topological Submanifold). *Let $(M, \mathcal{O})$ be a topological manifold and let $N \subseteq M$. Then $(N, \mathcal{O}|_N)$ is called a **submanifold** of $(M, \mathcal{O})$ if it is a manifold in its own right.*

Example 4.2. The space S^1 is a submanifold of $\mathbb{R}^2$ while the spaces S^2, C and T^2 are submanifolds of $\mathbb{R}^3$.

Definition 4.3 (Product Manifold). *Let $(M, \mathcal{O}_M)$ and $(N, \mathcal{O}_N)$ be topological manifolds of dimension m and n, respectively. Then, $(M \times N, \mathcal{O}_{M\times N})$ is a topological manifold of dimension $m + n$ called the **product manifold**.*

Example 4.3. We have $T^2 = S^1 \times S^1$ not just as topological spaces, but as topological manifolds as well. This is a special case of the n-torus:

$$T^n := \underbrace{S^1 \times S^1 \times \cdots \times S^1}_{n \text{ times}},$$

which is an n-dimensional manifold.

Example 4.4. The cylinder $C = S^1 \times \mathbb{R}$ is a 2-dimensional manifold.

4.2 Charts & Atlases

Definition 4.4 (Chart). *Let $(M, \mathcal{O})$ be a d-dimensional manifold. Then, a pair (U, x) where $U \in \mathcal{O}$ and $x\colon U \to x(U) \subseteq \mathbb{R}^d$ is a homeomorphism, is said to be a **chart** of the manifold.*

Definition 4.5 (Components / Co-Ordinates Of A Chart)**.** *The* ***component functions (or maps)*** *of* $x \colon U \to x(U) \subseteq \mathbb{R}^d$ *are the maps:*

$$\begin{aligned} x^i \colon U &\to \mathbb{R} \\ p &\mapsto \mathrm{proj}_i(x(p)) \end{aligned}$$

for $1 \leq i \leq d$*, where* $\mathrm{proj}_i(x(p))$ *is the i-th component of* $x(p) \in \mathbb{R}^d$*. The* $x^i(p)$ *are called the* ***co-ordinates*** *of the point* $p \in U$ *with respect to the chart* (U, x)*.*

Definition 4.6 (Atlas)**.** *An* ***atlas*** *of a manifold* M *is a collection* $\mathscr{A} := \{(U_\alpha, x_\alpha) \mid \alpha \in \mathcal{A}\}$ *of charts such that:*

$$\bigcup_{\alpha \in \mathcal{A}} U_\alpha = M.$$

Definition 4.7 ($\mathcal{C}^0$-Compatible Charts)**.** *Two charts* (U, x) *and* (V, y) *are said to be* $\mathcal{C}^0$***-compatible*** *if either* $U \cap V = \varnothing$ *or the map:*

$$y \circ x^{-1} \colon x(U \cap V) \to y(U \cap V)$$

is continuous.

Note that $y \circ x^{-1}$ is a map from a subset of $\mathbb{R}^d$ to a subset of $\mathbb{R}^d$.

$$\begin{tikzcd} & U \cap V \subseteq M \arrow[dl, "x"'] \arrow[dr, "y"] & \\ x(U \cap V) \subseteq \mathbb{R}^d \arrow[rr, "y \circ x^{-1}"'] & & y(U \cap V) \subseteq \mathbb{R}^d \end{tikzcd}$$

Since the maps x and y are homeomorphisms, the composition map $y \circ x^{-1}$ is also a homeomorphism and hence continuous. Therefore, any two charts on a topological manifold are $\mathcal{C}^0$-compatible. This definition my thus seem redundant since it applies to every pair of charts. However, it is just a "warm up" since we will later refine this definition and define the *differentiability* of maps on a manifold in terms of $\mathcal{C}^k$-compatibility of charts.

Definition 4.8 (Chart Transition Map)**.** *The map* $y \circ x^{-1}$ *(and its inverse* $x \circ y^{-1}$*) is called the* ***chart transition map****.*

Definition 4.9 ($\mathcal{C}^0$-Atlas)**.** *A* $\mathcal{C}^0$***-atlas*** *of a manifold is an atlas of pairwise* $\mathcal{C}^0$*-compatible charts.*

Note that any atlas is also a $\mathcal{C}^0$*-atlas.*

Definition 4.10 (Maximal Atlas)**.** *A* $\mathcal{C}^0$*-atlas* $\mathscr{A}$ *is said to be a* ***maximal atlas*** *if for every* $(U, x) \in \mathscr{A}$*, we have* $(V, y) \in \mathscr{A}$ *for all* (V, y) *charts that are* $\mathcal{C}^0$*-compatible with* (U, x)*.*

Example 4.5. Not every $\mathcal{C}^0$-atlas is a maximal atlas. Indeed, consider $(\mathbb{R}, \mathcal{O}_{\mathrm{std}})$ and the atlas $\mathscr{A} := (\mathbb{R}, \mathrm{id}_\mathbb{R})$. Then $\mathscr{A}$ is not maximal since $((0,1), \mathrm{id}_\mathbb{R})$ is a chart which is $\mathcal{C}^0$-compatible with $(\mathbb{R}, \mathrm{id}_\mathbb{R})$ but $((0,1), \mathrm{id}_\mathbb{R}) \notin \mathscr{A}$.

We can now look at "objects on" topological manifolds from two points of view. For instance, consider a curve on a d-dimensional manifold M, i.e. a map $\gamma \colon \mathbb{R} \to M$. We now ask whether this curve is continuous, as it should be if models the trajectory of a particle on the "physical space" M.

A first answer is that $\gamma \colon \mathbb{R} \to M$ is continuous if it is continuous as a map between the topological spaces $\mathbb{R}$ and M.

However, the answer that may be more familiar to you from undergraduate physics is the following. We consider only a portion (open subset U) of the physical space M and, instead of studying the map $\gamma \colon \mathrm{preim}_\gamma(U) \to U$ directly, we study the map:

$$x \circ \gamma \colon \mathrm{preim}_\gamma(U) \to x(U) \subseteq \mathbb{R}^d,$$

where (U, x) is a chart of M. More likely, you would be checking the continuity of the co-ordinate maps $x^i \circ \gamma$, which would then imply the continuity of the "real" curve $\gamma\colon \mathrm{preim}_\gamma(U) \to U$ (real, as opposed to its co-ordinate representation).

$$\begin{tikzcd}
 & y(U) \subseteq \mathbb{R}^d & \\
\mathrm{preim}_\gamma(U) \subseteq \mathbb{R} \arrow[ur, "y\circ\gamma"] \arrow[r, "\gamma"] \arrow[dr, "x\circ\gamma"'] & U \subseteq M \arrow[u, "y"'] \arrow[d, "x"] & \\
 & x(U) \subseteq \mathbb{R}^d \arrow[uu, bend right=60, "y\circ x^{-1}"'] &
\end{tikzcd}$$

At some point you may wish to use a different "co-ordinate system" to answer a different question. In this case, you would chose a different chart (U, y) and then study the map $y \circ \gamma$ or its co-ordinate maps. Notice however that some results (e.g. the continuity of γ) obtained in the previous chart (U, x) can be immediately "transported" to the new chart (U, y) via the chart transition map $y \circ x^{-1}$. Moreover, the map $y \circ x^{-1}$ allows us to, intuitively speaking, forget about the inner structure (i.e. U and the maps γ, x and x) which, in a sense, is the real world, and only consider $\mathrm{preim}_\gamma(U) \subseteq \mathbb{R}$ and $x(U), y(U) \subseteq \mathbb{R}^d$ together with the maps between them, which is our representation of the real world.

As we already said, for a topological manifold $(M, \mathcal{O})$, the concept of a $\mathcal{C}^0$-atlas is fully redundant since every atlas is also a $\mathcal{C}^0$-atlas. We will now generalise the notion of a $\mathcal{C}^0$-atlas, or more precisely, the notion of $\mathcal{C}^0$-compatibility of charts, to something which is non-trivial and non-redundant.

Definition 4.11 (❀-Atlas)**.** *An atlas $\mathscr{A}$ for a topological manifold is called a* ***❀-atlas*** *if any two charts $(U, x), (V, y) \in \mathscr{A}$ are ❀-compatible, where the symbol ❀ is being used as a placeholder for any of the following:*

- *❀ $= \mathcal{C}^0$: this just reduces to the previous definition;*
- *❀ $= \mathcal{C}^k$: the transition maps are k-times continuously differentiable as maps between open subsets of $\mathbb{R}^{\dim M}$;*
- *❀ $= \mathcal{C}^\infty$: the transition maps are smooth (infinitely many times differentiable); equivalently, the atlas is $\mathcal{C}^k$ for all $k \geq 0$;*
- *❀ $= \mathcal{C}^\omega$: the transition maps are (real) analytic, which is stronger than being smooth;*
- *❀ $=$ complex: if $\dim M$ is even, M is a* complex manifold *if the transition maps are continuous and satisfy the Cauchy-Riemann equations; its complex dimension is $\frac{1}{2}\dim M$.*

In other words, either $U \cap V = \varnothing$ or if $U \cap V \neq \varnothing$, then the transition map $y \circ x^{-1}$ from $x(U \cap V)$ to $y(U \cap V)$ must be ❀.

$$\begin{tikzcd}
 & U \cap V \subseteq M \arrow[dl, "x"'] \arrow[dr, "y"] & \\
x(U \cap V) \subseteq \mathbb{R}^{\dim M} \arrow[rr, "y\circ x^{-1}"'] & & y(U \cap V) \subseteq \mathbb{R}^{\dim M}
\end{tikzcd}$$

Theorem 4.1 (Whitney)**.** *Any maximal $\mathcal{C}^k$-atlas, with $k \geq 1$, contains a $\mathcal{C}^\infty$-atlas. Moreover, any two maximal $\mathcal{C}^k$-atlases that contain the same $\mathcal{C}^\infty$-atlas are identical.*

An immediate implication is that if we can find a $\mathcal{C}^1$-atlas for a manifold, then we can also assume the existence of a $\mathcal{C}^\infty$-atlas for that manifold. This is not the case for topological manifolds in general: a space with a $\mathcal{C}^0$-atlas may not admit any $\mathcal{C}^1$-atlas. But if we have at least a $\mathcal{C}^1$-atlas, then we can obtain a $\mathcal{C}^\infty$-atlas simply by removing charts, keeping only the ones which are $\mathcal{C}^\infty$-compatible.

Hence, for the purposes of this course, we will not distinguish between $\mathcal{C}^k$ ($k \geq 1$) and $\mathcal{C}^\infty$-manifolds in the above sense.

We now give the explicit definition of a $\mathcal{C}^k$-manifold.

Definition 4.12 ($\mathcal{C}^k$-Manifold)**.** *A $\mathcal{C}^k$**-manifold** is a triple $(M, \mathcal{O}, \mathscr{A})$, where $(M, \mathcal{O})$ is a topological manifold and $\mathscr{A}$ is a maximal $\mathcal{C}^k$-atlas.*

Definition 4.13 (Smooth Manifold)**.** *A $\mathcal{C}^\infty$-manifold is called a **smooth manifold**.*

Remark 4.1. A given topological manifold can carry different incompatible atlases.

Note that while we only defined compatibility of charts, it should be clear what it means for two atlases of the same type to be compatible.

Definition 4.14 (Compatible / Incompatible Atlases)**.** *Two ❀-atlases $\mathscr{A}$, $\mathscr{B}$ are **compatible** if their union $\mathscr{A} \cup \mathscr{B}$ is again a ❀-atlas, and are **incompatible** otherwise.*

Alternatively, we can define the compatibility of two atlases in terms of the compatibility of any pair of charts, one from each atlas.

Example 4.6. Let $(M, \mathcal{O}) = (\mathbb{R}, \mathcal{O}_{\text{std}})$. Consider the two atlases $\mathscr{A} = \{(\mathbb{R}, \mathrm{id}_\mathbb{R})\}$ and $\mathscr{B} = \{(\mathbb{R}, x)\}$, where $x \colon a \mapsto \sqrt[3]{a}$. Since they both contain a single chart, the compatibility condition on the transition maps is easily seen to hold (in both cases, the only transition map is $\mathrm{id}_\mathbb{R}$). Hence they are both $\mathcal{C}^\infty$-atlases. Consider now $\mathscr{A} \cup \mathscr{B}$. The transition map $\mathrm{id}_\mathbb{R} \circ x^{-1}$ is the map $a \mapsto a^3$, which is smooth. However, the other transition map, $x \circ \mathrm{id}_\mathbb{R}^{-1}$, is the map x, which is not even differentiable once (the first derivative at 0 does not exist). Consequently, $\mathscr{A}$ and $\mathscr{B}$ are not even $\mathcal{C}^1$-compatible.

The previous example shows that we can equip the real line with (at least) two different incompatible $\mathcal{C}^\infty$-structures. This looks like a disaster as it implies that there is an arbitrary choice to be made about which differentiable structure to use. Fortunately, the situation is not as bad as it looks, as we will see in the next sections.

4.3 Differentiable Manifolds

Definition 4.15 (Differentiable Map)**.** *Let $\phi \colon M \to N$ be a map, where $(M, \mathcal{O}_M, \mathscr{A}_M)$ and $(N, \mathcal{O}_N, \mathscr{A}_N)$ are $\mathcal{C}^k$-manifolds. Then ϕ is said to be ($\mathcal{C}^k$-)**differentiable at** $p \in M$ if for some charts $(U, x) \in \mathscr{A}_M$ with $p \in U$ and $(V, y) \in \mathscr{A}_N$ with $\phi(p) \in V$, the map $y \circ \phi \circ x^{-1}$ is k-times continuously differentiable at $x(p) \in x(U) \subseteq \mathbb{R}^{\dim M}$ in the usual sense.*

$$\begin{tikzcd} U \subseteq M \arrow[r, "\phi"] \arrow[d, "x"] & V \subseteq N \arrow[d, "y"] \\ x(U) \subseteq \mathbb{R}^{\dim M} \arrow[r, "y \circ \phi \circ x^{-1}"] & y(V) \subseteq \mathbb{R}^{\dim N} \end{tikzcd}$$

The above diagram shows a typical theme with manifolds. We have a map $\phi \colon M \to N$ and we want to define some property of ϕ at $p \in M$ analogous to some property of maps between subsets of $\mathbb{R}^d$. What we typically do is consider some charts (U, x) and (V, y) as above and define the desired property of ϕ at $p \in U$ in terms of the corresponding property of the downstairs map $y \circ \phi \circ x^{-1}$ at the point $x(p) \in \mathbb{R}^d$. Notice that in the previous definition we only require that *some* charts from the two atlases satisfy the stated property. So we should worry about whether this definition depends on which charts we pick. In fact, this "lifting" of the notion of differentiability from the chart representation of ϕ to the manifold level is well-defined.

Proposition 4.2. *The definition of differentiability is well-defined.*

Proof. We want to show that if $y \circ \phi \circ x^{-1}$ is differentiable at $x(p)$ for some $(U,x) \in \mathscr{A}_M$ with $p \in U$ and $(V,y) \in \mathscr{A}_N$ with $\phi(p) \in V$, then $\widetilde{y} \circ \phi \circ \widetilde{x}^{-1}$ is differentiable at $\widetilde{x}(p)$ for all charts $(\widetilde{U},\widetilde{x}) \in \mathscr{A}_M$ with $p \in \widetilde{U}$ and $(\widetilde{V},\widetilde{y}) \in \mathscr{A}_N$ with $\phi(p) \in \widetilde{V}$.

$$\begin{tikzcd}
\widetilde{x}(U \cap \widetilde{U}) \subseteq \mathbb{R}^{\dim M} \arrow[r, "\widetilde{y}\circ\phi\circ\widetilde{x}^{-1}"] & \widetilde{y}(V \cap \widetilde{V}) \subseteq \mathbb{R}^{\dim N} \\
U \cap \widetilde{U} \subseteq M \arrow[r, "\phi"] \arrow[u, "\widetilde{x}"] \arrow[d, "x"] & V \cap \widetilde{V} \subseteq N \arrow[u, "\widetilde{y}"] \arrow[d, "y"] \\
x(U \cap \widetilde{U}) \subseteq \mathbb{R}^{\dim M} \arrow[r, "y\circ\phi\circ x^{-1}"] \arrow[uu, bend left=60, "\widetilde{x}\circ x^{-1}"] & y(V \cap \widetilde{V}) \subseteq \mathbb{R}^{\dim N} \arrow[uu, bend right=60, "\widetilde{y}\circ y^{-1}"']
\end{tikzcd}$$

Consider the map $\widetilde{x} \circ x^{-1}$ in the diagram above. Since the charts (U,x) and $(\widetilde{U},\widetilde{x})$ belong to the same $\mathcal{C}^k$-atlas $\mathscr{A}_M$, by definition the transition map $\widetilde{x} \circ x^{-1}$ is $\mathcal{C}^k$-differentiable as a map between subsets of $\mathbb{R}^{\dim M}$, and similarly for $\widetilde{y} \circ y^{-1}$. We now notice that we can write:

$$\widetilde{y} \circ \phi \circ \widetilde{x}^{-1} = (\widetilde{y} \circ y^{-1}) \circ (y \circ \phi \circ x^{-1}) \circ (\widetilde{x} \circ x^{-1})^{-1}$$

and since the composition of $\mathcal{C}^k$ maps is still $\mathcal{C}^k$, we are done. $\square$

This proof shows the significance of restricting to $\mathcal{C}^k$-atlases. Such atlases only contain charts for which the transition maps are $\mathcal{C}^k$, which is what makes our definition of differentiability of maps between manifolds well-defined.

The same definition and proof work for smooth ($\mathcal{C}^\infty$) manifolds, in which case we talk about *smooth maps*. As we said before, this is the case we will be most interested in.

Example 4.7. Consider the smooth manifolds $(\mathbb{R}^d, \mathcal{O}_{\text{std}}, \mathscr{A}_d)$ and $(\mathbb{R}^{d'}, \mathcal{O}_{\text{std}}, \mathscr{A}_{d'})$, where $\mathscr{A}_d)$ and $\mathscr{A}_{d'})$ are the maximal atlases containing the charts $(\mathbb{R}^d, \mathrm{id}_{\mathbb{R}^d})$ and $(\mathbb{R}^{d'}, \mathrm{id}_{\mathbb{R}^{d'}})$ respectively, and let $f \colon \mathbb{R}^d \to \mathbb{R}^{d'}$ be a map. The diagram defining the differentiability of f with respect to these charts is

$$\begin{tikzcd}
\mathbb{R}^d \arrow[r, "f"] \arrow[d, "\mathrm{id}_{\mathbb{R}^d}"] & \mathbb{R}^{d'} \arrow[d, "\mathrm{id}_{\mathbb{R}^{d'}}"] \\
\mathbb{R}^d \arrow[r, "\mathrm{id}_{\mathbb{R}^{d'}}\circ f\circ(\mathrm{id}_{\mathbb{R}^d})^{-1}"] & \mathbb{R}^{d'}
\end{tikzcd}$$

and, by definition, the map f is smooth as a map between manifolds if, and only if, the map $\mathrm{id}_{\mathbb{R}^{d'}} \circ f \circ (\mathrm{id}_{\mathbb{R}^d})^{-1} = f$ is smooth in the usual sense.

Example 4.8. Let $(M, \mathcal{O}, \mathscr{A})$ be a d-dimensional smooth manifold and let $(U,x) \in \mathscr{A}$. Then $x \colon U \to x(U) \subseteq \mathbb{R}^d$ is smooth. Indeed, we have

$$\begin{tikzcd}
U \arrow[r, "x"] \arrow[d, "x"] & x(U) \arrow[d, "\mathrm{id}_{x(U)}"] \\
x(U) \subseteq \mathbb{R}^d \arrow[r, "\mathrm{id}_{x(U)}\circ x\circ x^{-1}"] & x(U) \subseteq \mathbb{R}^d
\end{tikzcd}$$

Hence $x \colon U \to x(U)$ is smooth if, and only if, the map $\mathrm{id}_{x(U)} \circ x \circ x^{-1} = \mathrm{id}_{x(U)}$ is smooth in the usual sense, which it certainly is.

The coordinate maps $x^i := \mathrm{proj}_i \circ x \colon U \to \mathbb{R}$ are also smooth. Indeed, consider the diagram

$$\begin{tikzcd} U \arrow[rr, "x^i"] \arrow[d, "x"] & & \mathbb{R} \arrow[d, "\mathrm{id}_{\mathbb{R}}"] \\ x(U) \subseteq \mathbb{R}^d \arrow[rr, "\mathrm{id}_{\mathbb{R}} \circ x^i \circ x^{-1}"] & & \mathbb{R} \end{tikzcd}$$

Then, x^i is smooth if, and only if, the map

$$\mathrm{id}_{\mathbb{R}} \circ x^i \circ x^{-1} = x^i \circ x^{-1} = \mathrm{proj}_i$$

is smooth in the usual sense, which it certainly is.

4.3.1 Classification Of Differentiable Structures

Definition 4.16 (Diffeomorphism). *Let $\phi \colon M \to N$ be a bijective map between smooth manifolds. If both ϕ and ϕ^{-1} are smooth, then ϕ is said to be a* ***diffeomorphism****.*

Diffeomorphisms are the structure preserving maps between smooth manifolds.

Definition 4.17 (Diffeomorphic Manifolds). *Two manifolds $(M, \mathcal{O}_M, \mathscr{A}_M)$, $(N, \mathcal{O}_N, \mathscr{A}_N)$ are said to be* ***diffeomorphic*** *if there exists a diffeomorphism $\phi \colon M \to N$ between them. We write $M \cong_{diff} N$.*

Note that if the differentiable structure is understood (or irrelevant), we typically write M instead of the triple $(M, \mathcal{O}_M, \mathscr{A}_M)$.

Remark 4.2. Being diffeomorphic is an equivalence relation. In fact, it is customary to consider diffeomorphic manifolds to be *the same* from the point of view of differential geometry. This is similar to the situation with topological spaces, where we consider homeomorphic spaces to be the same from the point of view of topology. This is typical of all structure preserving maps.

Armed with the notion of diffeomorphism, we can now ask the following question: how many smooth structures on a given topological space are there, up to diffeomorphism?
The answer is quite surprising: it depends on the dimension of the manifold!

Theorem 4.2 (Radon-Moise). *Let M be a manifold with $\dim M = 1, 2,$ or 3. Then there is a unique smooth structure on M up to diffeomorphism.*

Recall that in a previous example, we showed that we can equip $(\mathbb{R}, \mathcal{O}_{\mathrm{std}})$ with two incompatible atlases $\mathscr{A}$ and $\mathscr{B}$. Let $\mathscr{A}_{\max}$ and $\mathscr{B}_{\max}$ be their extensions to maximal atlases, and consider the smooth manifolds $(\mathbb{R}, \mathcal{O}_{\mathrm{std}}, \mathscr{A}_{\max})$ and $(\mathbb{R}, \mathcal{O}_{\mathrm{std}}, \mathscr{B}_{\max})$. Clearly, these are different manifolds, because the atlases are different, but since $\dim \mathbb{R} = 1$, they must be diffeomorphic.

The answer to the case $\dim M > 4$ (we emphasize $\dim M \neq 4$) is provided by *surgery theory*. This is a collection of tools and techniques in topology with which one obtains a new manifold from given ones by performing surgery on them, i.e. by cutting, replacing and gluing parts in such a way as to control topological invariants like the fundamental group. The idea is to understand all manifolds in dimensions higher than 4 by performing surgery systematically. In particular, using surgery theory, it has been shown that there are only finitely many smooth manifolds (up to diffeomorphism) one can make from a topological manifold.

This is not as neat as the previous case, but since there are only finitely many structures, we can still enumerate them, i.e. we can write an exhaustive list.

While finding all the differentiable structures may be difficult for any given manifold, this theorem has an immediate impact on a physical theory that models spacetime as a manifold. For instance, some physicists believe that spacetime should be modelled as a 10-dimensional manifold (we are neither proposing nor condemning this view). If that is indeed the case, we need to worry about which differentiable structure we equip our 10-dimensional manifold with, as each different choice will likely lead to different predictions. But since there are only finitely many such structures, physicists can, at least in principle, devise and perform finitely many experiments to distinguish between them and determine which is the right one, if any.

We now turn to the special case $\dim M = 4$. The result is that if M is a non-compact topological manifold, then there are uncountably many non-diffeomorphic smooth structures that we can equip M with. In particular, this applies to $(\mathbb{R}^4, \mathcal{O}_{\text{std}})$.

4.4 Tangent Spaces

In this section, whenever we say "manifold", we mean a (real) d-dimensional differentiable manifold, unless we explicitly say otherwise. We will also suppress the differentiable structure in the notation.

Definition 4.18 ($\mathcal{C}^\infty(M)$ Vector Space)**.** *Let M be a manifold. We define the infinite-dimensional vector space over $\mathbb{R}$ of all smooth functions on M with underlying set*

$$\mathcal{C}^\infty(M) := \{f \colon M \to \mathbb{R} \mid f \text{ is smooth}\}$$

and operations defined pointwise, i.e. for any $p \in M$,

$$\begin{aligned}(f+g)(p) &:= f(p) + g(p)\\ (\lambda f)(p) &:= \lambda f(p).\end{aligned}$$

A routine check shows that this is indeed a vector space.

Definition 4.19 (Smooth Curve)**.** *A **smooth curve** on M is a smooth map $\gamma \colon \mathbb{R} \to M$, where $\mathbb{R}$ is understood as a 1-dimensional manifold.*

Definition 4.20 (Directional Derivative Operator)**.** *Let $\gamma \colon \mathbb{R} \to M$ be a smooth curve through $p \in M$; w.l.o.g. let $\gamma(0) = p$. The **directional derivative operator** at p along γ is the linear map*

$$\begin{aligned}X_{\gamma,p} \colon \mathcal{C}^\infty(M) &\xrightarrow{\sim} \mathbb{R}\\ f &\mapsto (f \circ \gamma)'(0),\end{aligned}$$

where $\mathbb{R}$ is understood as a 1-dimensional vector space over the field $\mathbb{R}$.

Note that $f \circ \gamma$ is a map $\mathbb{R} \to \mathbb{R}$, hence we can calculate the usual derivative and evaluate it at 0.

Remark 4.3. In differential geometry, $X_{\gamma,p}$ is called the *tangent vector* to the curve γ at the point $p \in M$. Intuitively, $X_{\gamma,p}$ is the velocity γ at p. Consider the curve $\delta(t) := \gamma(2t)$, which is the same curve parametrised twice as fast. We have, for any $f \in \mathcal{C}^\infty(M)$:

$$X_{\delta,p}(f) = (f \circ \delta)'(0) = 2(f \circ \gamma)'(0) = 2X_{\gamma,p}(f)$$

by using the chain rule. Hence $X_{\gamma,p}$ scales like a velocity should.

Definition 4.21 (Tangent Space)**.** *Let M be a manifold and $p \in M$. The **tangent space** to M at p is the vector space over $\mathbb{R}$ with underlying set*

$$T_pM := \{X_{\gamma,p} \mid \gamma \text{ is a smooth curve through } p\},$$

addition

$$\begin{aligned}\oplus \colon T_pM \times T_pM &\to T_pM\\ (X_{\gamma,p}, X_{\delta,p}) &\mapsto X_{\gamma,p} \oplus X_{\delta,p},\end{aligned}$$

and scalar multiplication

$$\begin{aligned}\odot \colon \mathbb{R} \times T_pM &\to T_pM\\ (\lambda, X_{\gamma,p}) &\mapsto \lambda \odot X_{\gamma,p},\end{aligned}$$

both defined pointwise, i.e. for any $f \in \mathcal{C}^\infty(M)$,

$$\begin{aligned}(X_{\gamma,p} \oplus X_{\delta,p})(f) &:= X_{\gamma,p}(f) + X_{\delta,p}(f) \\ (\lambda \odot X_{\gamma,p})(f) &:= \lambda X_{\gamma,p}(f).\end{aligned}$$

Note that the outputs of these operations do not look like elements in T_pM, because they are not of the form $X_{\sigma,p}$ for some curve σ. Hence, we need to show that the above operations are, in fact, well-defined.

Proposition 4.3. *Let* $X_{\gamma,p}, X_{\delta,p} \in T_pM$ *and* $\lambda \in \mathbb{R}$*. Then, we have* $X_{\gamma,p} \oplus X_{\delta,p} \in T_pM$ *and* $\lambda \odot X_{\gamma,p} \in T_pM$*.*

Since the derivative is a local concept, it is only the behaviour of curves near p that matters. In particular, if two curves γ and δ agree on a neighbourhood of p, then $X_{\gamma,p}$ and $X_{\delta,p}$ are the same element of T_pM. Hence, we can work *locally* by using a chart on M.

Proof. Let (U, x) be a chart on M, with U a neighbourhood of p.

i) Define the curve

$$\sigma(t) := x^{-1}((x \circ \gamma)(t) + (x \circ \delta)(t) - x(p)).$$

Note that σ is smooth since it is constructed via addition and composition of smooth maps and, moreover:

$$\begin{aligned}\sigma(0) &= x^{-1}(x(\gamma(0)) + x(\delta(0)) - x(p)) \\ &= x^{-1}(x(p)) + x(p) - x(p)) \\ &= x^{-1}(x(p)) \\ &= p.\end{aligned}$$

Thus σ is a smooth curve through p. Let $f \in \mathcal{C}^\infty(U)$ be arbitrary. Then we have

$$\begin{aligned}X_{\sigma,p}(f) &:= (f \circ \sigma)'(0) \\ &= [f \circ x^{-1} \circ ((x \circ \gamma) + (x \circ \delta) - x(p))]'(0) \\ &= [\partial_a(f \circ x^{-1})(x(p))]\,((x^a \circ \gamma) + (x^a \circ \delta) - x^a(p))'(0) \\ &= [\partial_a(f \circ x^{-1})(x(p))]\,((x^a \circ \gamma)'(0) + (x^a \circ \delta)'(0)) \\ &= (f \circ x^{-1} \circ x \circ \gamma)'(0) + (f \circ x^{-1} \circ x \circ \delta)'(0) \\ &= (f \circ \gamma)'(0) + (f \circ \delta)'(0) \\ &=: (X_{\gamma,p} \oplus X_{\delta,p})(f).\end{aligned}$$

Therefore $X_{\gamma,p} \oplus X_{\delta,p} = X_{\sigma,p} \in T_pM$.

ii) The second part is straightforward. Define $\sigma(t) := \gamma(\lambda t)$. This is again a smooth curve through p and we have:

$$\begin{aligned}X_{\sigma,p}(f) &:= (f \circ \sigma)'(0) \\ &= f'(\sigma(0))\, \sigma'(0) \\ &= \lambda f'(\gamma(0))\, \gamma'(0) \\ &= \lambda (f \circ \gamma)'(0) \\ &:= (\lambda \odot X_{\gamma,p})(f)\end{aligned}$$

for any $f \in \mathcal{C}^\infty(U)$. Hence $\lambda \odot X_{\gamma,p} = X_{\sigma,p} \in T_pM$. □

Hence indeed T_pM is a vector space.

The question is, what exactly $X_{\gamma,p}$ is mathematically speaking? Since it's **a** map of the form:

$$X_{\gamma,p} \colon \mathcal{C}^\infty(M) \xrightarrow{\sim} \mathbb{R}$$

it's clear that it's an element of $\mathrm{Hom}(\mathcal{C}^\infty(M), R)$, i.e an element of the dual vector space of $\mathcal{C}^\infty(M)$. Which subsequently makes T_pM a sub-vector space of the dual vector space of $\mathcal{C}^\infty(M)$. ($X_{\gamma,p}$ is a particular choice of a linear map, more specifically the derivative with respect to the parameter, and not **all** possible linear maps. This is why T_pM is not the whole dual vector space of $\mathcal{C}^\infty(M)$)

However, if we take the extra step and turn the $\mathcal{C}^\infty(M)$ from a vector space to an algebra (by defining an appropriate operation) then we can show that $X_{\gamma,p}$ is actually a derivation of the algebra.

More specifically we will define a product on $\mathcal{C}^\infty(M)$ by

$$\begin{aligned} \bullet \colon \mathcal{C}^\infty(M) \times \mathcal{C}^\infty(M) &\to \mathcal{C}^\infty(M) \\ (f,g) &\mapsto f \bullet g, \end{aligned}$$

where $f \bullet g$ is defined pointwise. Then $(\mathcal{C}^\infty(M), +, \cdot, \bullet)$ is an associative, unital and commutative algebra over $\mathbb{R}$.

Now that we have an algebra, let us remind ourselves what a derivation is and also try to combine the definition with our case:

Definition 4.22 (Derivation (On A Manifold))**.** *Let M be a manifold and let $p \in U \subseteq M$, where U is open. A* derivation on U at p *is an $\mathbb{R}$-linear map $D \colon \mathcal{C}^\infty(U) \xrightarrow{\sim} \mathbb{R}$ satisfying the Leibniz rule*

$$D(fg) = D(f)g(p) + f(p)D(g).$$

The usual derivative operator is a derivation on $\mathcal{C}^\infty(\mathbb{R})$, the algebra of smooth real functions, since it is linear and satisfies the Leibniz rule. (The second derivative operator, however, is not a derivation on $\mathcal{C}^\infty(\mathbb{R})$, since it does not satisfy the Leibniz rule. This shows that the composition of derivations need not be a derivation.) Hence, we managed to show that indeed $X_{\gamma,p}$ is actually a derivation of the algebra of smooth real functions on M.

4.4.1 Co-Ordinate Induced Basis For The Tangent Space

The following is a crucially important result about tangent spaces.

Theorem 4.3. *Let M be a manifold and let $p \in M$. Then*

$$\dim T_pM = \dim M.$$

Remark 4.4. Note carefully that, despite us using the same symbol, the two "dimensions" appearing in the statement of the theorem are, at least on the surface, entirely unrelated. Indeed, recall that $\dim M$ is defined in terms of charts (U,x), with $x \colon U \to x(U) \subseteq \mathbb{R}^{\dim M}$, while $\dim T_pM = |\mathcal{B}|$, where $\mathcal{B}$ is a Hamel basis for the vector space T_pM. The idea behind the proof is to construct a basis of T_pM from a chart on M.

Proof. W.l.o.g., let (U,x) be a chart *centred* at p, i.e. $x(p) = 0 \in \mathbb{R}^{\dim M}$. Define $(\dim M)$-many curves $\gamma_{(a)} \colon \mathbb{R} \to U$ through p by requiring $(x^b \circ \gamma_{(a)})(t) = \delta^b_a t$, i.e.

$$\begin{aligned} \gamma_{(a)}(0) &:= p \\ \gamma_{(a)}(t) &:= x^{-1} \circ (0, \dots, 0, t, 0, \dots, 0) \end{aligned}$$

where the t is in the a^{th} position, with $1 \leq a \leq \dim M$. Let us calculate the action of the tangent vector $X_{\gamma_{(a)},p} \in T_pM$ on an arbitrary function $f \in \mathcal{C}^\infty(U)$:

$$\begin{aligned} X_{\gamma_{(a)},p}(f) &:= (f \circ \gamma_{(a)})'(0) \\ &= (f \circ \mathrm{id}_U \circ \gamma_{(a)})'(0) \\ &= (f \circ x^{-1} \circ x \circ \gamma_{(a)})'(0) \\ &= [\partial_b(f \circ x^{-1})(x(p))]\, (x^b \circ \gamma_{(a)})'(0) \\ &= [\partial_b(f \circ x^{-1})(x(p))]\, (\delta^b_a t)'(0) \\ &= [\partial_b(f \circ x^{-1})(x(p))]\, \delta^b_a \\ &= \partial_a(f \circ x^{-1})(x(p)) \end{aligned}$$

We introduce a special notation for this last line, namely:

$$\partial_a(f \circ x^{-1})(x(p)) := \left(\frac{\partial}{\partial x^a}\right)_p (f)$$

Remark 4.5. While the symbol $\left(\frac{\partial}{\partial x^a}\right)_p$ has nothing to do with the idea of partial differentiation with respect to the variable x^a (since x refers to the chart map and no differentiation has been defined there), it is notationally consistent with it, in the following sense.

Let $M = \mathbb{R}^d$, $(U, x) = (\mathbb{R}^d, \mathrm{id}_{\mathbb{R}^d})$ and let $\left(\frac{\partial}{\partial x^a}\right)_p \in T_p\mathbb{R}^d$. If $f \in \mathcal{C}^\infty(\mathbb{R}^d)$, then

$$\left(\frac{\partial}{\partial x^a}\right)_p (f) = \partial_a(f \circ x^{-1})(x(p)) = \partial_a f(p),$$

since $x = x^{-1} = \mathrm{id}_{\mathbb{R}^d}$. Moreover, we have $\mathrm{proj}_a = x^a$. Thus, we can think of $x^1, \ldots, x^d$ as the independent variables of f, and we can then write

$$\left(\frac{\partial}{\partial x^a}\right)_p (f) = \frac{\partial f}{\partial x^a}(p).$$

Hence, up to this point we showed that:

$$X_{\gamma_{(a)},p}(f) = \left(\frac{\partial}{\partial x^a}\right)_p (f)$$

Or by removing the action on the function, simply:

$$X_{\gamma_{(a)},p} = \left(\frac{\partial}{\partial x^a}\right)_p$$

We now claim that

$$\mathcal{B} = \left\{ \left(\frac{\partial}{\partial x^a}\right)_p \in T_pM \;\middle|\; 1 \leq a \leq \dim M \right\}$$

is a basis of T_pM. First, we show that $\mathcal{B}$ spans T_pM.

Let $X \in T_pM$. Then, by definition, there exists some smooth curve σ through p such that $X = X_{\sigma,p}$. For any $f \in \mathcal{C}^\infty(U)$, we have

$$\begin{aligned}
X(f) &= X_{\sigma,p}(f) \\
&:= (f \circ \sigma)'(0) \\
&= (f \circ x^{-1} \circ x \circ \sigma)'(0) \\
&= [\partial_b(f \circ x^{-1})(x(p))]\, (x^b \circ \sigma)'(0) \\
&= (x^b \circ \sigma)'(0) \left(\frac{\partial}{\partial x^b}\right)_p (f).
\end{aligned}$$

Since $(x^b \circ \sigma)'(0) =: X^b \in \mathbb{R}$, we have:

$$X = X^b \left(\frac{\partial}{\partial x^b}\right)_p,$$

i.e. any $X \in T_pM$ is a linear combination of elements from $\mathcal{B}$.

To show linear independence, suppose that

$$\lambda^a \left(\frac{\partial}{\partial x^a}\right)_p = 0,$$

for some scalars λ^a. Note that this is an operator equation, and the zero on the right hand side is the zero operator $0 \in T_pM$.

Recall that, given the chart (U, x), the coordinate maps $x^b \colon U \to \mathbb{R}$ are smooth, i.e. $x^b \in \mathcal{C}^\infty(U)$. Thus, we can feed them into the left hand side to obtain

$$\begin{aligned} 0 &= \lambda^a \left(\frac{\partial}{\partial x^a}\right)_p (x^b) \\ &= \lambda^a \, \partial_a (x^b \circ x^{-1})(x(p)) \\ &= \lambda^a \, \partial_a (\mathrm{proj}_b)(x(p)) \\ &= \lambda^a \delta^b_a \\ &= \lambda^b \end{aligned}$$

i.e. $\lambda^b = 0$ for all $1 \leq b \leq \dim M$. So $\mathcal{B}$ is indeed a basis of T_pM, and since by construction $|\mathcal{B}| = \dim M$, the proof is complete. □

Remark 4.6. While it is possible to define infinite-dimensional manifolds, in this course we will only consider finite-dimensional ones. Hence $\dim T_pM = \dim M$ will always be finite in this course.

Remark 4.7. Note that the basis that we have constructed in the proof is *not* chart-independent. Indeed, each different chart will induce a different tangent space basis, and we distinguish between them by keeping the chart map in the notation for the basis elements.

This is not a cause of concern for our proof however, since every basis of a vector space must have the same cardinality, and hence it suffices to find one basis to determine the dimension.

Definition 4.23 (Co-Ordinate Induced Basis)**.** *Let $X \in T_pM$ be a tangent vector and let (U, x) be a chart containing p. Then the basis $\{\left(\frac{\partial}{\partial x^a}\right)_p\}$ created by the usage of the chart is called a **co-ordinate induced basis**. In this basis an element $X \in T_pM$ can be expressed as:*

$$X = X^a \left(\frac{\partial}{\partial x^a}\right)_p,$$

*where the real numbers $X^1, \ldots, X^{\dim M}$ are called the **vector components** of X with respect to the co-ordinate induced basis by the chart (U, x).*

4.4.2 Change Of Vector Components Under A Change Of Chart

One of the most heavily used concepts is the transformation of the components of a vector under different co-ordinate systems (i.e under a chart transition map that subsequently changes the co-ordinate induced basis). Let's find out the rule.

Let $X \in T_pM$ and let (U, x) and (V, y) be two charts containing p. Then X can be expressed in any of the two charts as:

$$X^a{}_{(y)} \left(\frac{\partial}{\partial y^a}\right)_p = X = X^a{}_{(x)} \left(\frac{\partial}{\partial x^a}\right)_p$$

Let us act with X on some smooth function f of $\mathcal{C}^\infty(M)$ by using first the components of (U, x) chart:

$$\begin{aligned} X(f) &= X^a{}_{(x)} \left(\frac{\partial}{\partial x^a}\right)_p (f) \\ &= X^a{}_{(x)} \partial_a (f \circ x^{-1})(x(p)) \\ &= X^a{}_{(x)} \partial_a (f \circ y^{-1} \circ y \circ x^{-1})(x(p)) \\ &= X^a{}_{(x)} \partial_a (y^b \circ x^{-1})(x(p)) \, \partial_b (f \circ y^{-1})(y(p)) \\ &= X^a{}_{(x)} \frac{\partial y^b}{\partial x^a} \left(\frac{\partial}{\partial y^b}\right)_p (f) \end{aligned}$$

Similarly, let us now act with X on the smooth function f of $\mathcal{C}^\infty(M)$ by using the components of (V, y) chart:

$$X(f) = X^a{}_{(y)} \left(\frac{\partial}{\partial y^a} \right)_p (f)$$

These expressions are, of course, equal to each other so by suppressing now the action on the function f, we obtain:

$$X^a{}_{(x)} \frac{\partial y^b}{\partial x^a} \left(\frac{\partial}{\partial y^b} \right)_p = X^b{}_{(y)} \left(\frac{\partial}{\partial y^b} \right)_p$$

$$X^a{}_{(x)} \frac{\partial y^b}{\partial x^a} \left(\frac{\partial}{\partial y^b} \right)_p - X^b{}_{(y)} \left(\frac{\partial}{\partial y^b} \right)_p = 0$$

$$\left(X^a{}_{(x)} \frac{\partial y^b}{\partial x^a} - X^b{}_{(y)} \right) \left(\frac{\partial}{\partial y^b} \right)_p = 0$$

Finally, since the base vectors of $\{ \left(\frac{\partial}{\partial y^a} \right)_p \}$ are linearly independent the only way for this equation to be zero is for the coefficients to be zero hence:

$$X^a{}_{(x)} \frac{\partial y^b}{\partial x^a} - X^b{}_{(y)} = 0$$

Of finally by solving w.r.t $X^b{}_{(y)}$ and renaming the indices:

$$X^a{}_{(y)} = \frac{\partial y^a}{\partial x^b} X^b{}_{(x)}$$

This equation shows as how the components of a vector transform under a chart transition map, i.e under the change of charts, i.e from one co-ordinate induced basis to another. Of course the formula agrees completely with the transformations of vector components under the change of basis that we showed in previous chapter: $\widetilde{v}^b = A^b{}_a v^a$.

The function $y^a = y^a(x^1, \ldots, x^{\dim M})$ expresses the new co-ordinates in terms of the old ones, and $A^b{}_a$ is the *Jacobian* matrix of this map, evaluated at $x(p)$. Note that no matter how non-linear the transformations of the co-ordinates are, the vectors always transform in a linear fashion. In a way, "vectors do not care about the non-linearity of co-ordinate transformations".

4.5 Cotangent Spaces

Since the tangent space is a vector space, we can do all the constructions we saw previously in the abstract vector space setting.

Definition 4.24 (Cotagent Space)**.** *Let M be a manifold and $p \in M$. The **cotangent space** to M at p is*

$$T_p^* M := (T_p M)^*$$

Since $\dim T_p M$ is finite, we have $T_p M \cong_{\text{vec}} T_p^* M$.

And of course, once we have the cotangent space, we can define the tensor spaces.

Definition 4.25 (Tensor Space)**.** *Let M be a manifold and $p \in M$. The **tensor space** $(T_s^r)_p M$ is defined as*

$$(T_s^r)_p M := T_s^r(T_p M) = \underbrace{T_p M \otimes \cdots \otimes T_p M}_{r \text{ copies}} \otimes \underbrace{T_p^* M \otimes \cdots \otimes T_p^* M}_{s \text{ copies}}.$$

4.5.1 Dual Basis For The Cotangent Space

Now let's give a very important definition that will help us to formalize elements, and subsequently a basis, for the cotangent space.

Definition 4.26 (Gradient). *Let M be a manifold and let $f\colon M \to \mathbb{R}$ be smooth. The* ***gradient of*** f ***at*** $p \in M$ *is the $\mathbb{R}$-linear map*

$$\begin{aligned} \mathrm{d}_p \colon \mathcal{C}^\infty(M) &\xrightarrow{\sim} T_p^*M \\ f &\mapsto \mathrm{d}_p f, \end{aligned}$$

with $p \in U \subseteq M$, defined by

$$\mathrm{d}_p f(X) := X(f)$$

Remark 4.8. Note that since d_p is a map from $\mathcal{C}^\infty(M) \xrightarrow{\sim} T_p^*M$ that means that when it acts on a function of $\mathcal{C}^\infty(M)$ the final result $\mathrm{d}_p f$ is an element of T_p^*M hence a covector. By its turn, as an element of the dual space of T_pM it maps elements of T_pM to the real numbers (that's the definition of the dual space of a vector space). Hence the expression $\mathrm{d}_p f(X)$ must end up to a real number, which is indeed what $X(f)$ is. By writing $\mathrm{d}_p f(X) := X(f)$, we have committed a slight (but nonetheless real) abuse of notation, since $\mathrm{d}_p f(X) \in T_{f(p)}\mathbb{R}$ takes in a function and return a real number, but $X(f)$ is already a real number! However by doing so we can now talk about $\mathrm{d}_p f$ without providing the vector that it acts on. In other words we can talk about covectors without the need of their actions on vectors.

Remark 4.9. The gradient of a function is a covector and **not** a vector.

Recall that if (U, x) is a chart on M, then the co-ordinate maps $x^a \colon U \to x(U) \subseteq \mathbb{R}^{\dim M}$ are smooth functions on U hence they belong to $\mathcal{C}^\infty(M)$. We can thus apply the gradient operator d_p (with $p \in U$) to each of them to obtain $(\dim M)$-many elements of T_p^*M.

Proposition 4.4. *Let (U, x) be a chart on M, with $p \in U$. The set $\mathcal{B} = \{\mathrm{d}_p x^a \mid 1 \leq a \leq \dim M\}$ forms the dual basis of T_p^*M.*

Proof. By simply acting on $\big(\frac{\partial}{\partial x^a}\big)_p$ with $\mathrm{d}_p x^a$ (in our notation, we have $(\mathrm{d}x^a)_p = \mathrm{d}_p x^a$) we obtain:

$$\begin{aligned} \mathrm{d}_p x^a \left(\left(\frac{\partial}{\partial x^b} \right)_p \right) &= \left(\frac{\partial}{\partial x^b} \right)_p (x^a) && \text{(definition of } \mathrm{d}_p x^a\text{)} \\ &= \partial_b (x^a \circ x^{-1})(x(p)) && \text{(definition of } \big(\tfrac{\partial}{\partial x^b}\big)_p\text{)} \\ &= \partial_b (\mathrm{proj}_a)(x(p)) \\ &= \delta^a_b \end{aligned}$$

Therefore, $\mathcal{B}$ is, in fact, the dual basis to $\{\big(\frac{\partial}{\partial x^a}\big)_p\}$. □

4.5.2 Change Of Covector Components Under A Change Of Chart

Once again, as we did in the vector case with the vector components, one needs to find the transformation of the components of a covector under different co-ordinate systems. We will follow exactly the same procedure.

Let $\omega \in T_p^*M$ and let (U, x) and (V, y) be two charts containing p. Then ω can be expressed in any of the two charts by using the dual basis as:

$$\omega_{(y)a} (dy^a)_p = \omega = \omega_{(x)a} (dx^a)_p$$

By repeating the same process as we did for the vectors it is very easy to show that covectors components transform as

$$\omega_{(y)a} = \left(\frac{\partial x^b}{\partial y^a} \right)_p \omega_{(x)b}$$

4.6 Push-Forward And Pull-Back

Definition 4.27 (Push-Forward)**.** *Let $\phi\colon M \to N$ be a smooth map between smooth manifolds. The **push-forward** (or **derivative**) of ϕ at $p \in M$ is the linear map $(\phi_*)_p$:*

$$\begin{aligned} (\phi_*)_p \colon T_pM &\xrightarrow{\sim} T_{\phi(p)}N \\ X &\mapsto (\phi_*)_p(X) \end{aligned}$$

where $(\phi_*)_p(X)$ is defined as

$$\begin{aligned} (\phi_*)_p(X) \colon \mathcal{C}^\infty(N) &\xrightarrow{\sim} \mathbb{R} \\ f &\mapsto (\phi_*)_p(X)f := X(f \circ \phi). \end{aligned}$$

In other words, since $(\phi_*)_p$ is a map from one tangent space to another this means that it acts on a tangent vector and produces another one, hence $(\phi_*)_p(X)$ is again a tangent vector (but on N). As a tangent vector it can act on a smooth function (again on N) and produce a real number, hence the action of a push-forward on a function is simply the one we wrote above.

Note that one has to define a push-forward $(\phi_*)_p$ for every point p of M. Although we have only one map ϕ we have many push-forward maps $(\phi_*)_p$.

Proposition 4.5. *Let $\phi\colon M \to N$ be smooth. The tangent vector $X_{\gamma,p} \in T_pM$ is pushed forward to the tangent vector $X_{\phi\circ\gamma,\phi(p)} \in T_{\phi(p)}N$, i.e.*

$$(\phi_*)_p(X_{\gamma,p}) = X_{\phi\circ\gamma,\phi(p)}.$$

Proof. Let $f \in \mathcal{C}^\infty(V)$, with (V, x) a chart on N and $\phi(p) \in V$. By applying the definitions, we have

$$\begin{aligned} (\phi_*)_p(X_{\gamma,p})(f) &= (X_{\gamma,p})(f \circ \phi) && \text{(definition of } (\phi_*)_p) \\ &= ((f \circ \phi) \circ \gamma)'(0) && \text{(definition of } X_{\gamma,p}) \\ &= (f \circ (\phi \circ \gamma))'(0) && \text{(associativity of } \circ) \\ &= X_{\phi\circ\gamma,\phi(p)}(f) && \text{(definition of } X_{\phi\circ\gamma,\phi(p)}) \end{aligned}$$

Since f was arbitrary, we have $(\phi_*)_p(X_{\gamma,p}) = X_{\phi\circ\gamma,\phi(p)}$. □

Related to the push-forward, there is the notion of pull-back of a smooth map.

Definition 4.28 (Pull-Back)**.** *Let $\phi\colon M \to N$ be a smooth map between smooth manifolds. The **pull-back** of ϕ at $p \in M$ is the linear map:*

$$\begin{aligned} (\phi^*)_p \colon T^*_{\phi(p)}N &\xrightarrow{\sim} T^*_pM \\ \omega &\mapsto (\phi^*)_p(\omega), \end{aligned}$$

where $(\phi^*)_p(\omega)$ is defined as

$$\begin{aligned} (\phi^*)_p(\omega) \colon T_pM &\xrightarrow{\sim} \mathbb{R} \\ X &\mapsto (\phi^*)_p(\omega)(X) := \omega((\phi_*)_p(X)). \end{aligned}$$

In words, if ω is a covector on N, its pull-back $(\phi^*)_p(\omega)$ is a covector on M. It acts on tangent vectors on M by first pushing them forward to tangent vectors on N, and then applying ω to them to produce a real number.

Diagrammatically, what we've defined so far is the following

$$\begin{tikzcd}
\mathcal{C}^\infty(M) \arrow[d, "X"'] & \mathcal{C}^\infty(N) \arrow[l, "-\circ\phi"'] \arrow[dl, "(\phi_*)_p(X)"] \\
\mathbb{R} &
\end{tikzcd}
\qquad
\begin{tikzcd}
T_pM \arrow[r, "(\phi_*)_p"] \arrow[dr, "(\phi^*)_p(\omega)"'] & T_{\phi(p)}N \arrow[d, "\omega"] \\
& \mathbb{R}
\end{tikzcd}$$

Remark 4.10. It is quite easy to show that everything we have defined in this section is, in fact, linear.

Remark 4.11. We have seen that, given a smooth $\phi\colon M \to N$, we can push a vector $X \in T_pM$ forward to a vector $(\phi_*)_p(X) \in T_{\phi(p)}N$, and pull a covector $\omega \in T^*_{\phi(p)}N$ back to a covector $(\phi^*)_p(\omega) \in T^*_pM$. In other words both push-forward and pull-back work only in the direction of their definition. However, if $\phi\colon M \to N$ is a diffeomorphism (and only then), we can also pull a vector $Y \in T_{\phi(p)}N$ back to a vector $(\phi^*)_p(Y) \in T_pM$, and push a covector $\eta \in T^*_pM$ forward to a covector $(\phi_*)_p(\eta) \in T^*_{\phi(p)}N$, by using ϕ^{-1} as follows:

$$(\phi^*)_p(Y) := ((\phi^{-1})_*)_{\phi(p)}(Y)$$
$$(\phi_*)_p(\eta) := ((\phi^{-1})^*)_{\phi(p)}(\eta).$$

In general, we should keep in mind that:

Vectors are pushed forward,
covectors are pulled back.

4.7 Immersions And Embeddings

We will now consider the question of under which circumstances a smooth manifold can "sit" in $\mathbb{R}^d$, for some $d \in \mathbb{N}$. There are, in fact, two notions of sitting inside another manifold, called immersion and embedding.

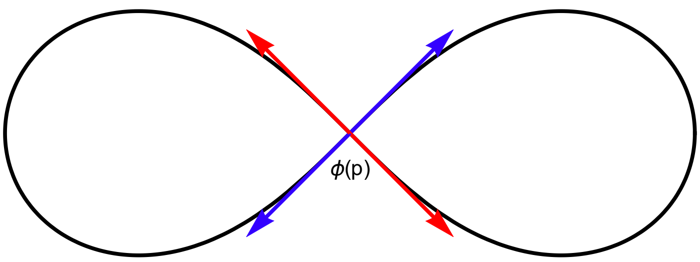
Definition 4.29 (Immersion)**.** *A smooth map $\phi\colon M \to N$ is said to be an **immersion** of M into N if the derivative*

$$(\phi_*)_p\colon T_pM \xrightarrow{\sim} T_{\phi(p)}N$$

is injective, for all $p \in M$. In that case, the manifold M is said to be an immersed submanifold *of N.*

From the theory of linear algebra, we immediately deduce that, for $\phi\colon M \to N$ to be an immersion, we must have $\dim M \leq \dim N$. A closely related notion is that of a *submersion*, where we require each $(\phi_*)_p$ to be surjective, and thus we must have $\dim M \geq \dim N$. However, we will not need this here.

Example 4.9. Consider the map $\phi\colon S^1 \to \mathbb{R}^2$ whose image is reproduced below.



The map ϕ is not injective, i.e. there are $p, q \in S^1$, with $p \neq q$ and $\phi(p) = \phi(q)$. Of course, this means that $T_{\phi(p)}\mathbb{R}^2 = T_{\phi(q)}\mathbb{R}^2$. However, the maps $(\phi_*)_p$ and $(\phi_*)_q$ are both injective, with their images being represented by the blue and red arrows, respectively. Hence, the map ϕ is immersion.

Definition 4.30 (Embedding)**.** *A smooth map $\phi\colon M \to N$ is said to be a **embedding** of M into N if*

- $\phi\colon M \to N$ *is an immersion;*
- $M \cong_{\text{top}} \phi(M) \subseteq N$*, where* $\phi(M)$ *carries the subset topology inherited from* N*.*

In that case the manifold M *is said to be an* embedded submanifold *of* N*.*

Remark 4.12. If a continuous map between topological spaces satisfies the second condition above, then it is called a *topological embedding*. Therefore, an embedding is a topological embedding which is also an immersion (as opposed to simply being a topological embedding).

In the early days of differential geometry there were two approaches to study manifolds. One was the extrinsic view, within which manifolds are defined as special subsets of $\mathbb{R}^d$, and the other was the intrinsic view, which is the view that we have adopted here.

Whitney's theorem, which we will state without proof, states that these two approaches are, in fact, equivalent.

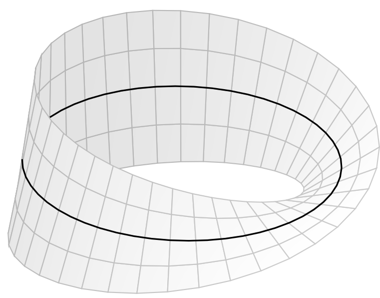
Theorem 4.4 (Whitney)**.** *Any smooth manifold* M *can be*

- *embedded in* $\mathbb{R}^{2\dim M}$*;*
- *immersed in* $\mathbb{R}^{2\dim M-1}$*.*

Example 4.10. The Klein bottle can be embedded in $\mathbb{R}^4$ but not in $\mathbb{R}^3$. It can, however, be immersed in $\mathbb{R}^3$.

4.8 Topological Bundles

While topological products are very useful, very often one intuitively thinks of the product of two manifolds as attaching a copy of the second manifold to each point of the first. However, not all interesting manifolds can be understood as products of manifolds. A classic example of this is the *Möbius strip*.



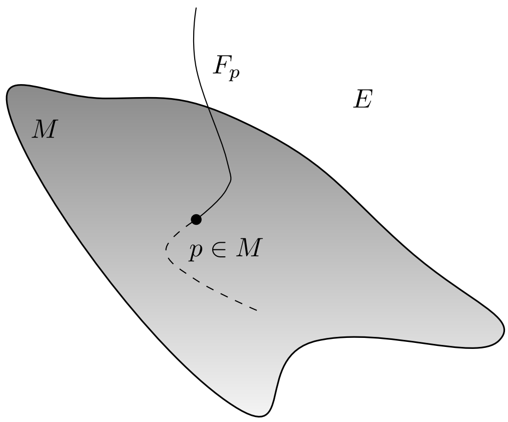
It looks locally like the finite cylinder $S^1 \times [0,1]$, which we can picture as the circle S^1 (the thicker line in figure) with the finite interval $[0,1]$ attached to each of its points in a "smooth" way. The Möbius strip has a "twist", which makes it globally different from the cylinder.

Definition 4.31 (Topological Bundles)**.** *A topological* ***bundle*** *(of topological manifolds) is a triple* (E,π,M) *where* E *and* M *are topological manifolds called the* total space *and the* base space *respectively, and* π *is a continuous, surjective map* $\pi\colon E \to M$ *called the* projection map*.*

We will often denote the bundle (E,π,M) by $E \xrightarrow{\pi} M$.

Definition 4.32 (Fiber)**.** *Let* $E \xrightarrow{\pi} M$ *be a bundle and let* $p \in M$*. Then,* $F_p := \operatorname{preim}_\pi(\{p\})$ *is called the* ***fiber*** *at the point* p*.*

Intuitively, the fiber at the point $p \in M$ is a set of points in E (represented below as a line) attached to the point p. The projection map sends all the points in the fiber F_p to the point p.



Example 4.11. A trivial example of a bundle is the *product bundle*. Let M and N be manifolds. Then, the triple $(M \times N, \pi, M)$, where:

$$\begin{aligned} \pi \colon M \times N &\to M \\ (p, q) &\mapsto p \end{aligned}$$

is a bundle since (one can easily check) π is a continuous and surjective map. Similarly, $(M \times N, \pi, N)$ with the appropriate π, is also a bundle.

Example 4.12. In a bundle, different points of the base manifold may have (topologically) different fibers. For example, consider the bundle $E \xrightarrow{\pi} \mathbb{R}$ where:

$$F_p := \mathrm{preim}_\pi(\{p\}) \cong_{\mathrm{top}} \begin{cases} S^1 & \text{if } p < 0 \\ \{p\} & \text{if } p = 0 \\ [0,1] & \text{if } p > 0 \end{cases}$$

Definition 4.33 (Fiber Bundle)**.** *Let $E \xrightarrow{\pi} M$ be a bundle and let F be a manifold. Then, $E \xrightarrow{\pi} M$ is called a **fiber bundle**, with (typical) fiber F, if:*

$$\forall\, p \in M : \mathrm{preim}_\pi(\{p\}) \cong_{\mathrm{top}} F.$$

A fiber bundle is often represented diagrammatically as:

$$\begin{array}{ccc} F & \longrightarrow & E \\ & & \Big\downarrow \pi \\ & & M \end{array}$$

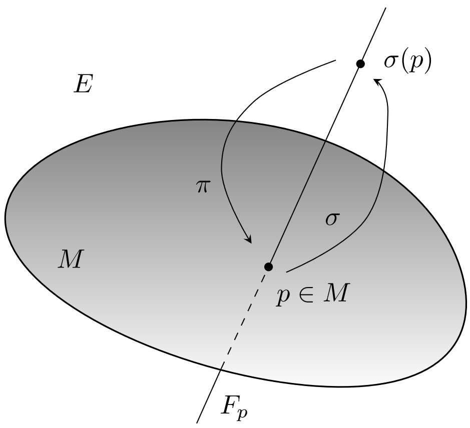
Example 4.13. The bundle $M \times N \xrightarrow{\pi} M$ is a fiber bundle with fiber $F := N$.

Example 4.14. The Möbius strip is a fiber bundle $E \xrightarrow{\pi} S^1$, with fiber $F := [0,1]$, where $E \neq S^1 \times [0,1]$, i.e. the Möbius strip is not a product bundle.

Example 4.15. A $\mathbb{C}$-line bundle over M is the fiber bundle (E, π, M) with fiber $\mathbb{C}$. Note that the product bundle $(M \times \mathbb{C}, \pi, M)$ is a $\mathbb{C}$-line bundle over M, but a $\mathbb{C}$-line bundle over M need not be a product bundle.

Definition 4.34 (Section)**.** *Let $E \xrightarrow{\pi} M$ be a bundle. A map $\sigma \colon M \to E$ is called a **section** of the bundle if $\pi \circ \sigma = \mathrm{id}_M$.*

Intuitively, a section is a map σ which sends each point $p \in M$ to *some* point $\sigma(p)$ in its fiber F_p, so that the projection map π takes $\sigma(p) \in F_p \subseteq E$ back to the point $p \in M$.



Example 4.16. Let $(M \times F, \pi, M)$ be a product bundle. Then, a section of this bundle is a map:

$$\begin{aligned} \sigma \colon M &\to M \times F \\ p &\mapsto (p, s(p)) \end{aligned}$$

where $s \colon M \to F$ is any map.

Definition 4.35 (Sub-Bundle). *A* ***sub-bundle*** *of a bundle (E, π, M) is a triple (E', π', M') where $E' \subseteq E$ and $M' \subseteq M$ are submanifolds and $\pi' := \pi|_{E'}$.*

Definition 4.36 (Restricted Bundle). *Let (E, π, M) be a bundle and let $N \subseteq M$ be a submanifold. The* ***restricted bundle*** *(to N) is the triple (E, π', N) where:*

$$\pi' := \pi|_{\mathrm{preim}_\pi(N)}$$

Definition 4.37 (Bundle Morphism). *Let $E \xrightarrow{\pi} M$ and $E' \xrightarrow{\pi'} M'$ be bundles and let $u \colon E \to E'$ and $v \colon M \to M'$ be maps. Then (u, v) is called a* bundle morphism *if the following diagram commutes:*

$$\begin{tikzcd} E \arrow[r, "u"] \arrow[d, "\pi"] & E' \arrow[d, "\pi'"] \\ M \arrow[r, "v"] & M' \end{tikzcd}$$

i.e. if $\pi' \circ u = v \circ \pi$.

If (u, v) and (u, v') are both bundle morphisms, then $v = v'$. That is, given u, if there exists v such that (u, v) is a bundle morphism, then v is unique.

Definition 4.38 (Isomorphic Bundles). *Two bundles $E \xrightarrow{\pi} M$ and $E' \xrightarrow{\pi'} M'$ are said to be* ***isomorphic (as bundles)*** *if there exist bundle morphisms (u, v) and (u^{-1}, v^{-1}) satisfying:*

$$\begin{tikzcd} E \arrow[r, shift left, "u"] \arrow[d, "\pi"'] & E' \arrow[l, shift left, "u^{-1}"] \arrow[d, "\pi'"] \\ M \arrow[r, shift left, "v"] & M' \arrow[l, shift left, "v^{-1}"] \end{tikzcd}$$

Such a (u, v) is called a bundle isomorphism *and we write $E \xrightarrow{\pi} M \cong_{\mathrm{bdl}} E' \xrightarrow{\pi'} M'$.*

Bundle isomorphisms are the structure-preserving maps for bundles.

Definition 4.39 (Locally Isomorphic Bundles)**.** *A bundle $E \xrightarrow{\pi} M$ is said to be **locally isomorphic (as a bundle)** to a bundle $E' \xrightarrow{\pi'} M'$ if for all $p \in M$ there exists a neighbourhood $U(p)$ such that the restricted bundle:*

$$\operatorname{preim}_\pi(U(p)) \xrightarrow{\pi|_{\operatorname{preim}_\pi(U(p))}} U(p)$$

is isomorphic to the bundle $E' \xrightarrow{\pi'} M'$.

Definition 4.40 (Trivial / Locally Trivial Bundle)**.** *A bundle $E \xrightarrow{\pi} M$ is said to be:*

i) ***trivial*** *if it is isomorphic to a product bundle;*

ii) ***locally trivial*** *if it is locally isomorphic to a product bundle.*

Example 4.17. The cylinder C is trivial as a bundle, and hence also locally trivial.

Example 4.18. The Möbious strip is not trivial but it is locally trivial.

From now on, we will mostly consider locally trivial bundles.

Remark 4.13. In quantum mechanics, what is usually called a "wave function" is not a function at all, but rather a section of a $\mathbb{C}$-line bundle over physical space. However, if we assume that the $\mathbb{C}$-line bundle under consideration is locally trivial, then each section of the bundle can be represented (locally) by a map from the base space to the total space and hence it is appropriate to use the term "wave *function*".

Definition 4.41 (Pull-Back Bundle)**.** *Let $E \xrightarrow{\pi} M$ be a bundle and let $f\colon M' \to M$ be a map from some manifold M'. The **pull-back bundle of** $E \xrightarrow{\pi} M$ induced by f is defined as $E' \xrightarrow{\pi'} M'$, where:*

$$E' := \{(m', e) \in M' \times E \mid f(m') = \pi(e)\}$$

and $\pi'(m', e) := m'$.

If $E' \xrightarrow{\pi'} M'$ is the pull-back bundle of $E \xrightarrow{\pi} M$ induced by f, then one can easily construct a bundle morphism by defining:

$$\begin{aligned} u\colon \quad E' &\to E \\ (m', e) &\mapsto e \end{aligned}$$

This corresponds to the diagram:

$$\begin{tikzcd} E' \arrow[r, "u"] \arrow[d, "\pi'"] & E \arrow[d, "\pi"] \\ M' \arrow[r, "f"] & M \end{tikzcd}$$

Remark 4.14. Sections on a bundle pull back to the pull-back bundle. Indeed, let $E' \xrightarrow{\pi'} M'$ be the pull-back bundle of $E \xrightarrow{\pi} M$ induced by f.

$$\begin{tikzcd} E' \arrow[d, shift left, "\pi'"] & & E \arrow[d, shift left, "\pi"] \\ M' \arrow[u, shift left, "\sigma'"] \arrow[rru, "\sigma \circ f"] \arrow[rr, "f"] & & M \arrow[u, shift left, "\sigma"] \end{tikzcd}$$

If σ is a section of $E \xrightarrow{\pi} M$, then $\sigma \circ f$ determines a map from M' to E which sends each $m' \in M'$ to $\sigma(f(m')) \in E$. However, since σ is a section, we have:

$$\pi(\sigma(f(m')) = (\pi \circ \sigma \circ f)(m') = (\mathrm{id}_M \circ f)(m') = f(m')$$

and hence $(m', (\sigma \circ f)(m')) \in E'$ by definition of E'. Moreover:

$$\pi'(m', (\sigma \circ f)(m')) = m'$$

and hence the map:

$$\begin{aligned} \sigma' \colon M' &\to E' \\ m' &\mapsto (m', (\sigma \circ f)(m')) \end{aligned}$$

satisfies $\pi' \circ \sigma' = \mathrm{id}_{M'}$ and it is thus a section on the pull-back bundle $E' \xrightarrow{\pi'} M'$.

The reason of introducing the concept of a topological bundle, is because we need it in order to construct the so called "tangent bundle".

4.9 The Tangent Bundle

Up to this point, we have defined everything on the level of a point on the manifold. However, since we are interested in describing quantities as a whole in the entire manifold, we would like to define a vector field on a manifold M as a "smooth" map that assigns to each $p \in M$ a tangent vector in T_pM. However, since this would then be a "map" to a different space at each point, it is unclear how to define its smoothness.

The simplest solution is to merge all the tangent spaces into a unique set and equip it with a smooth structure, so that we can then define a vector field as a smooth map between smooth manifolds.

Definition 4.42 (Tangent Bundle)**.** *Given a smooth manifold M, the* ***tangent bundle*** *of M is the disjoint union of all the tangent spaces to M, i.e.*

$$TM := \dot{\bigcup}_{p \in M} T_pM,$$

equipped with the canonical projection map

$$\begin{aligned} \pi \colon TM &\to M \\ X &\mapsto p, \end{aligned}$$

where p is the unique $p \in M$ such that $X \in T_pM$.

Since TM is simpy a set (and not a smooth manifold), up to here what we have is a set bundle. In order for this set bundle to turn to a topological bundle as we defined it previously, we need to equip TM with the structure of a smooth manifold. We can achieve this by constructing a smooth atlas for TM from a smooth atlas on M, as follows.

Let $\mathscr{A}_M$ be a smooth atlas on M and let $(U, x) \in \mathscr{A}_M$. If $X \in \mathrm{preim}_\pi(U) \subseteq TM$, then $X \in T_{\pi(X)}M$, by definition of π. Moreover, since $\pi(X) \in U$, we can expand X in terms of the basis induced by the chart (U, x):

$$X = X^a \left(\frac{\partial}{\partial x^a}\right)_{\pi(X)},$$

where $X^1, \ldots, X^{\dim M} \in \mathbb{R}$. We can then define the map

$$\begin{aligned} \xi \colon \mathrm{preim}_\pi(U) &\to x(U) \times \mathbb{R}^{\dim M} \cong_{\mathrm{set}} \mathbb{R}^{2 \dim M} \\ X &\mapsto (x(\pi(X)), X^1, \ldots, X^{\dim M}). \end{aligned}$$

Assuming that TM is equipped with a suitable topology, for instance the initial topology (i.e. the coarsest topology on TM that makes π continuous), we claim that the pair $(\mathrm{preim}_\pi(U), \xi)$ is a chart on TM and

$$\mathscr{A}_{TM} := \{(\mathrm{preim}_\pi(U), \xi) \mid (U, x) \in \mathscr{A}_M\}$$

is a smooth atlas on TM. Note that, from its definition, it is clear that ξ is a bijection. We will not show that $(\mathrm{preim}_\pi(U), \xi)$ is a chart here, but we will show that $\mathscr{A}_{TM}$ is a smooth atlas.

Proposition 4.6. *Any two charts $(\mathrm{preim}_\pi(U), \xi), (\mathrm{preim}_\pi(\widetilde{U}), \widetilde{\xi}) \in \mathscr{A}_{TM}$ are $\mathcal{C}^\infty$-compatible.*

Proof. Let (U,x) and $(\widetilde{U},\widetilde{x})$ be the two charts on M giving rise to $(\mathrm{preim}_\pi(U),\xi)$ and $(\mathrm{preim}_\pi(\widetilde{U}),\widetilde{\xi})$, respectively. We need to show that the map

$$\widetilde{\xi}\circ\xi^{-1}\colon x(U\cap\widetilde{U})\times\mathbb{R}^{\dim M}\to\widetilde{x}(U\cap\widetilde{U})\times\mathbb{R}^{\dim M}$$

is smooth, as a map between open subsets of $\mathbb{R}^{2\dim M}$. Recall that such a map is smooth if, and only if, it is smooth componentwise. On the first $\dim M$ components, $\widetilde{\xi}\circ\xi^{-1}$ acts as

$$\begin{aligned}\widetilde{x}\circ x^{-1}\colon x(U\cap\widetilde{U})&\to\widetilde{x}(U\cap\widetilde{U})\\ x(p)&\mapsto\widetilde{x}(p),\end{aligned}$$

while on the remaining $\dim M$ components it acts as the change of vector components we met previously, i.e.

$$X^a\mapsto\widetilde{X}^a=\partial_b(y^a\circ x^{-1})(x(p))\,X^b.$$

Hence, we have

$$\begin{aligned}\widetilde{\xi}\circ\xi^{-1}\colon\qquad x(U\cap\widetilde{U})\times\mathbb{R}^{\dim M}&\to\widetilde{x}(U\cap\widetilde{U})\times\mathbb{R}^{\dim M}\\ (x(\pi(X)),X^1,\dots,X^{\dim M})&\mapsto(\widetilde{x}(\pi(X)),\widetilde{X}^1,\dots,\widetilde{X}^{\dim M}),\end{aligned}$$

which is smooth in each component, and hence smooth. □

The tangent bundle of a smooth manifold M is therefore itself a smooth manifold of dimension $2\dim M$, and the projection $\pi\colon TM\to M$ is smooth with respect to this structure.

Now by using the smooth manifold M as the base space, the smooth manifold TM as the total space, and the smooth projection π we can define the topological tangent bundle as the triple:

$$TM\xrightarrow{\pi}M$$

Similarly, one can construct the *cotangent bundle* T^*M to M by defining

$$T^*M:=\dot{\bigcup}_{p\in M}T^*_pM$$

and going through the above again, using the dual basis $\{(\mathrm{d}x^a)_p\}$ instead of $\{\big(\frac{\partial}{\partial x^a}\big)_p\}$.

4.10 Vector, Covector And Tensor Fields

Now that we have defined the tangent and cotangent bundles, we are ready to define fields.

Definition 4.43 (Vector Field)**.** *Let M be a smooth manifold, and let $TM\xrightarrow{\pi}M$ be its tangent bundle. A **vector field** σ on M is a smooth section of the tangent bundle, i.e. a smooth map $\sigma\colon M\to TM$ such that $\pi\circ\sigma=\mathrm{id}_M$.*

$$\begin{tikzcd} TM \arrow[d, shift left, "\pi"] \\ M \arrow[u, shift left, "\sigma"] \end{tikzcd}$$

Definition 4.44 ($\Gamma(TM)$)**.** *We denote the set of all vector fields on M by $\Gamma(TM)$, i.e.*

$$\Gamma(TM):=\{\sigma\colon M\to TM\mid\sigma\text{ is smooth and }\pi\circ\sigma=\mathrm{id}_M\}.$$

This is, in fact, the standard notation for the set of all sections on a bundle.

Remark 4.15. An equivalent definition is that a vector field σ on M is a derivation on the algebra $\mathcal{C}^\infty(M)$, i.e. an $\mathbb{R}$-linear map

$$\sigma\colon\mathcal{C}^\infty(M)\xrightarrow{\sim}\mathcal{C}^\infty(M)$$

satisfying the Leibniz rule (with respect to pointwise multiplication on $\mathcal{C}^\infty(M)$)

$$\sigma(fg) = g\,\sigma(f) + f\,\sigma(g).$$

This definition is better suited for some purposes, and later on we will switch from one to the other without making any notational distinction between them.

We can equip the set $\Gamma(TM)$ with the following operations. The first is our, by now familiar, pointwise addition:

$$\begin{aligned} \oplus\colon \Gamma(TM) \times \Gamma(TM) &\to \Gamma(TM) \\ (\sigma, \tau) &\mapsto \sigma \oplus \tau, \end{aligned}$$

where

$$\begin{aligned} \sigma \oplus \tau\colon M &\to \Gamma(TM) \\ p &\mapsto (\sigma \oplus \tau)(p) := \sigma(p) + \tau(p). \end{aligned}$$

Note that the $+$ on the right hand side above is the addition in T_pM.

More interestingly, we can define a multiplication operation not by a simple number (i.e an element of $\mathbb{R}$) but with a whole function (i.e an element of $\mathcal{C}^\infty(M)$) as follows:

$$\begin{aligned} \odot\colon \mathcal{C}^\infty(M) \times \Gamma(TM) &\to \Gamma(TM) \\ (f, \sigma) &\mapsto f \odot \sigma, \end{aligned}$$

where

$$\begin{aligned} f \odot \sigma\colon M &\to \Gamma(TM) \\ p &\mapsto (f \odot \sigma)(p) := f(p)\sigma(p). \end{aligned}$$

Note that since $f \in \mathcal{C}^\infty(M)$, we have $f(p) \in \mathbb{R}$ and hence the multiplication above is the scalar multiplication on T_pM.

Remark 4.16. Of course, we could have defined $\odot$ simply as pointwise *global* scaling, using the reals $\mathbb{R}$ instead of the real functions $\mathcal{C}^\infty(M)$. Then, since $(\mathbb{R}, +, \cdot)$ is an algebraic field, we would then have the obvious $\mathbb{R}$-vector space structure on $\Gamma(TM)$. There are two reasons why we don't do that:

- Since the vector field acts on the whole manifold M (it assigns a value $f(p)$ on every point p of the manifold) we want to be able to assign different values to different points. Otherwise we would only be able to assign the same value to every point (i.e having a constant vector field)
- A basis for the corresponding vector space would be necessarily uncountably infinite, and hence it would not provide a very useful decomposition for our vector fields. Instead, the operation $\odot$ that we have defined allows for *local* scaling, i.e. we can scale a vector field by a different value at each point, and a much more useful decomposition of vector fields.

The question now is, mathematically speaking, what exactly the triple $(\Gamma(TM), \oplus, \odot)$ is. Its nature of course depends on what the triple $(\mathcal{C}^\infty(M), +, \cdot)$ is. Let's recall that the triple $(\mathcal{C}^\infty(M), +, \cdot)$ can be viewed in two different ways:

- $(\mathcal{C}^\infty(M), +, \cdot)$, where $\cdot$ is scalar multiplication (by a real number), is an $\mathbb{R}$-vector space.
- $(\mathcal{C}^\infty(M), +, \bullet)$, where $\bullet$ is pointwise multiplication of maps, is a commutative, unital ring, but not a division ring since not every function has an inverse at every point (i.e at all points that a function is zero, we cannot define an inverse since we would divide by zero).

The first view is of no use since if the triple is seen as a vector space over the real numbers, there is nothing else we can do. However, if we consider the second view. i.e the triple $(\mathcal{C}^\infty(M), +, \bullet)$, where $\bullet$ is pointwise function multiplication as a ring, then the triple $(\Gamma(TM), \oplus, \odot)$ built on top of this ring satisfies

- $(\Gamma(TM), \oplus)$ is an abelian group, with $0 \in \Gamma(TM)$ being the section that maps each $p \in M$ to the zero tangent vector in T_pM;
- $\Gamma(TM) \setminus \{0\}$ satisfies:
    i) $\forall\, f \in \mathcal{C}^\infty(M) : \forall\, \sigma, \tau \in \Gamma(TM) \setminus \{0\} : f \odot (\sigma \oplus \tau) = (f \odot \sigma) \oplus (f \odot \tau)$;
    ii) $\forall\, f, g \in \mathcal{C}^\infty(M) : \forall\, \sigma \in \Gamma(TM) \setminus \{0\} : (f + g) \odot \sigma = (f \odot \sigma) \oplus (g \odot \sigma)$;
    iii) $\forall\, f, g \in \mathcal{C}^\infty(M) : \sigma \in \Gamma(TM) \setminus \{0\} : (f \bullet g) \odot \sigma = f \odot (g \odot \sigma)$;
    iv) $\forall\, \sigma \in \Gamma(TM) \setminus \{0\} : 1 \odot \sigma = \sigma$,

    where $1 \in \mathcal{C}^\infty(M)$ maps every $p \in M$ to $1 \in \mathbb{R}$.

which are precisely the axioms for a vector space! Hence given that the triple $(\mathcal{C}^\infty(M), +, \bullet)$ is a ring, that turns the triple $(\Gamma(TM), \oplus, \odot)$ to a $\mathcal{C}^\infty(M)$-module.

And this of course is of crucial importance since as we showed in previous chapters, if a ring R is not a division ring, then a R-module does not need to have a basis. And since as we already said $(\mathcal{C}^\infty(M), +, \bullet)$ is not a division ring, the vector fields as $\mathcal{C}^\infty(M)$-modules do not need to have a basis! And this is a shame, since if they would have a basis (let's say X_i) we would be able to write a vector field σ as:

$$\sigma = \sigma^i X_i$$

where σ^i would be functions acting as components of the vector field!

In a similar manner one can construct a covector field through the use of the cotangent bundle, and from there to define the set of all covector fields $\Gamma(T^*M)$ and subsequently a triple $(\Gamma(T^*M), \oplus, \odot)$.

Finally using $\Gamma(TM)$ and $\Gamma(T^*M)$ we can define a tensor field.

Definition 4.45 (Tensor Field)**.** *Let M be a smooth manifold. A smooth (r, s)* ***tensor field*** *τ on M is a $\mathcal{C}^\infty(M)$-multilinear map*

$$\tau\colon \underbrace{\Gamma(T^*M) \times \cdots \times \Gamma(T^*M)}_{r\ copies} \times \underbrace{\Gamma(TM) \times \cdots \times \Gamma(TM)}_{s\ copies} \to \mathcal{C}^\infty(M).$$

The equivalence of this to the bundle definition is due to the pointwise nature of tensors. For instance, a covector field $\omega \in \Gamma(T^*M)$ can act on a vector field $X \in \Gamma(TM)$ to yield a smooth function $\omega(X) \in \mathcal{C}^\infty(M)$ by

$$(\omega(X))(p) := \omega(p)(X(p)).$$

Then, we see that for any $f \in \mathcal{C}^\infty(M)$, we have

$$(\omega(fX))(p) = \omega(p)(f(p)X(p)) = f(p)\omega(p)(X(p)) =: (f\omega(X))(p)$$

and hence, the map $\omega\colon \Gamma(TM) \xrightarrow{\sim} \mathcal{C}^\infty(M)$ is $\mathcal{C}^\infty(M)$-linear.

Similarly, the set $\Gamma(T^r_sM)$ of all (r, s) smooth tensor fields on M can be made into a $\mathcal{C}^\infty(M)$-module, with module operations defined pointwise.

We can also define the tensor product of tensor fields

$$\begin{aligned} \otimes\colon \Gamma(T^p_qM) \times \Gamma(T^r_sM) &\to \Gamma(T^{p+r}_{q+s}M) \\ (\tau, \sigma) &\mapsto \tau \otimes \sigma \end{aligned}$$

analogously to what we had with tensors on a vector space, i.e.

$$\begin{aligned} &(\tau \otimes \sigma)(\omega_1, \ldots, \omega_p, \omega_{p+1}, \ldots, \omega_{p+r}, X_1, \ldots, X_q, X_{q+1}, \ldots, X_{q+s}) \\ &\qquad := \tau(\omega_1, \ldots, \omega_p, X_1, \ldots, X_q)\, \sigma(\omega_{p+1}, \ldots, \omega_{p+r}, X_{q+1}, \ldots, X_{q+s}), \end{aligned}$$

with $\omega_i \in \Gamma(T^*M)$ and $X_i \in \Gamma(TM)$.

Therefore, we can think of tensor fields on M either as sections of some tensor bundle on M, that is, as maps assigning to each $p \in M$ a tensor ($\mathbb{R}$-multilinear map) on the vector space T_pM, or as a $\mathcal{C}^\infty(M)$-multilinear map as above. We will always try to pick the most useful or easier to understand, based on the context.

To summarize, fields are the generalization of the definitions of vectors, covectors and tensors at a specific point p of M, to every possible point p of manifold M, hence to the whole manifold M. In a similar way we can generalize the concept of the gradient of f at $p \in M$ in the gradient of f at M.

Recall the definition of the gradient operator at a point $p \in M$. We can extend that definition to define the ($\mathbb{R}$-linear) operator:

$$\begin{aligned} \mathrm{d}\colon \mathcal{C}^\infty(M) &\xrightarrow{\sim} \Gamma(T^*M) \\ f &\mapsto \mathrm{d}f \end{aligned}$$

where, of course, $\mathrm{d}f\colon p \mapsto \mathrm{d}_p f$. Alternatively, we can think of $\mathrm{d}f$ as the $\mathbb{R}$-linear map

$$\begin{aligned} \mathrm{d}f\colon \Gamma(TM) &\xrightarrow{\sim} \mathcal{C}^\infty(M) \\ X &\mapsto \mathrm{d}f(X) = X(f). \end{aligned}$$

Remark 4.17. Locally on some chart (U, x) on M, the covector field $\mathrm{d}f$ can be expressed as

$$\mathrm{d}f = \lambda_a \, \mathrm{d}x^a$$

for some smooth functions $\lambda_i \in \mathcal{C}^\infty(U)$. To determine what they are, we simply apply both sides to the vector fields induced by the chart. We have

$$\mathrm{d}f\left(\frac{\partial}{\partial x^b}\right) = \frac{\partial}{\partial x^b}(f) = \partial_b f$$

and

$$\lambda_a \, \mathrm{d}x^a\left(\frac{\partial}{\partial x^b}\right) = \lambda_a \frac{\partial}{\partial x^b}(x^a) = \lambda_a \, \delta^a_b = \lambda_b.$$

Hence, the local expression of $\mathrm{d}f$ on (U, x) is

$$\mathrm{d}f = \partial_a f \, \mathrm{d}x^a.$$

Note that the operator d satisfies the Leibniz rule

$$\mathrm{d}(fg) = g \, \mathrm{d}f + f \, \mathrm{d}g.$$

Finally, we want to generalize the concepts of push-forward and pull-back from a point to the whole manifold (a.k.a from a vector/covector to a vector/covector field). For a good reason we will first start with the pull-back.

To avoid confusion, for this part we will denote a covector as W and a covector field as ω. Recall that given a smooth map $\phi\colon M \to N$ the definition of a pull-back for a covector was

$$\begin{aligned} (\phi^*)_p\colon T^*_{\phi(p)}N &\xrightarrow{\sim} T^*_pM \\ W &\mapsto (\phi^*)_p(W), \end{aligned}$$

where $(\phi^*)_p(W)$ is defined as

$$\begin{aligned} (\phi^*)_p(W)\colon T_pM &\xrightarrow{\sim} \mathbb{R} \\ X &\mapsto (\phi^*)_p(W)(X) := W((\phi_*)_p(X)). \end{aligned}$$

Now we can simply extend the definition of a pull-back for a covector at point p denoted $(\phi^*)_p$, to this of a pull-back for a covector field on a manifold M denoted ϕ^*, by simpy acting with $(\phi^*)_p$ at every point

p of the manifold M

$$\begin{aligned} \phi^* \colon T^*N &\to T^*M \\ \omega &\mapsto \phi^*(\omega) \end{aligned}$$

where now ω is a covector field and not just a covector, and $\phi^*(\omega)$ is defined for every point p of M as

$$\phi^*(\omega)(p) := (\phi^*)_p(W)$$

where W is the the corresponding covector that the covector field σ produces at point p. It is a common thing to this point to drop the p from the pull-back of covectors and simply write:

$$(\phi^*\omega)|_p := \phi^*(W|_p)$$

which actually means that the pull-back of a covector field evaluated at point p is equal to the pull-back of the covector $W|_p$ generated by the covector field ω at point p (a.k.a $W|_p = \omega(p)$) . From now on we will be using this equation to switch from covector field to covector equations, although practically it's the same thing from a different perspective.

While the pull-back can be extended from covectros to covector fields without problems, the push-forward of a vector cannot be generalized to the push-forward of a vector field unless the underlying map ϕ is a diffeomorphism between the manifold M and N. Let's see why.

Let's start again with the pull-back that we have already defined. Observe that the pull back of a covector field includes the notion of the pull-back of a covector at a point p. Now, the map ϕ, as a map, maps every single point of its domain M to a single point of its target N. Hence the whole target M is hit by the map, but the whole target N is not (recall that the part of N that is hit by the map is called the image of ϕ). This means that in the case of a pull-back (after we have defined the tangent vectors in both M and N) every single point of the image of ϕ on N will have a corresponding point back on M hence the definition of the pull-back of a covector field will be well-defined.

On the other hand, in the case of a push forward we get two problems coming from the fact that, in general, the map ϕ may not be neither surjective nor injective. First of all if the map ϕ is not surjective that means that the image of ϕ is not equal to the entire domain M ($\mathrm{im}_\phi(M) \neq N$), hence from a vector field defined on M we will never be able to define a vector field on N that lies outside the image of ϕ. Moreover, if ϕ is not injective, that means that distinct elements of the domain M are mapped to the same element in the target N hence it might be the case that the push-forward will create many different vectors for one given point p on N which will make it ill-defined.

Of course, if the map ϕ is both surjective and injective, hence bijective, hence has an inverse, then none of this problems exist any more, since then the case is similar to the case of pull-backs (both directions of the map behave similarly). But recall that a bijection between topological spaces is called a "homeomorphism", and moreover if the map is smooth (which in the case of smooth manifolds by definition always is) then the smooth "homeomorphism" is called a"diffeomorphism".

So we ended up to our initial conclusion that the push-forward of a vector can be generalized to the push-forward of a vector field only if the underlying map ϕ is a diffeomorphism between the manifold M and N.

Then we can simply follow the same procedure as with the pull-back and define the push-forward of a vector field by simpy acting with the push-forward at every point $\phi(p)$ on the manifold N. Recall that the push-forward of ϕ at $p \in M$ is the linear map:

$$\begin{aligned} (\phi_*)_p \colon T_pM &\xrightarrow{\sim} T_{\phi(p)}N \\ X &\mapsto (\phi_*)_p(X) \end{aligned}$$

where $(\phi_*)_p(X)$ is defined as

$$\begin{aligned} (\phi_*)_p(X) \colon \mathcal{C}^\infty(N) &\xrightarrow{\sim} \mathbb{R} \\ f &\mapsto (\phi_*)_p(X)f := X(f \circ \phi). \end{aligned}$$

Now we can simply extend this definition to a push-forward for a vector field by simpy acting with the push-forward at every point p of the manifold M

$$\begin{aligned}\phi_*\colon TM &\to TN\\ \sigma &\mapsto \phi_*(\sigma)\end{aligned}$$

where $\phi_*(\sigma)$ is defined for every point $\phi(p)$ of N as

$$\phi_*(\sigma)(\phi(p)) := (\phi_*)_p(X)$$

As with covectors, it is a common thing to this point to drop the p from the equation and simply write:

$$(\phi_*\sigma)|_{\phi(p)} := \phi_*(X|_p)$$

which actually means that the push-forward of a vector field evaluated at point $\phi(p)$ is equal to the push-forward of the vector $X|_p$ generated by the vector field σ at point p (a.k.a $X|_p = X(p)$). From now on we will be using this equation to switch from vector field to vector equations, although practically it's the same thing from a different perspective.

4.11 Differential Forms

Definition 4.46 (Differential Form)**.** *Let M be a smooth manifold. A* ***(differential)*** ***n-form*** *on M is a $(0, n)$ smooth tensor field ω which is totally antisymmetric, i.e.*

$$\omega(X_1, \dots, X_n) = \operatorname{sgn}(\pi)\, \omega(X_{\pi(1)}, \dots, X_{\pi(n)}),$$

for any $\pi \in S_n$, with $X_i \in \Gamma(TM)$. We call n the degree *of the form.*

Alternatively, we can define a differential form as a smooth section of the appropriate bundle on M, i.e. as a map assigning to each $p \in M$ an n-form on the vector space T_pM.

Of course, by definition, differential forms are nothing more but a very specific kind of tensors, hence it's a subset of the tensor space.

Example 4.19. The electromagnetic field strength F is a differential 2-form built from the electric and magnetic fields, which are also taken to be forms. We will define these later in some detail.

Definition 4.47 ($\Omega^n(M)$)**.** *We denote by $\Omega^n(M)$ the set of all differential n-forms on M, which then becomes a $\mathcal{C}^\infty(M)$-module by defining the addition and multiplication operations pointwise.*

We have $\Omega^0(M) \equiv \mathcal{C}^\infty(M)$ since they are $(0, 0)$ tensors a.k.a functions and $\Omega^1(M) \equiv \Gamma(T^0_1 M) \equiv \Gamma(T^*M)$ sincey they are $(0, 1)$ tensors a.k.a covectors.

We can specialise the pull-back of tensors to differential forms.

Definition 4.48 (Pull-Back On Differential Forms)**.** *Let $\phi\colon M \to N$ be a smooth map and let $\omega \in \Omega^n(N)$. Then we define the* ***pull-back*** *$\Phi^*(\omega) \in \Omega^n(M)$ of ω as*

$$\begin{aligned}\Phi^*(\omega)\colon M &\to T^*M\\ p &\mapsto \Phi^*(\omega)(p),\end{aligned}$$

where

$$\Phi^*(\omega)(p)(X_1, \dots, X_n) := \omega(\phi(p))\big(\phi_*(X_1), \dots, \phi_*(X_n)\big),$$

for $X_i \in T_pM$.

The map $\Phi^*\colon \Omega^n(N) \to \Omega^n(M)$ is $\mathbb{R}$-linear, and its action on $\Omega^0(M)$ is simply

$$\begin{aligned}\Phi^*\colon \Omega^0(M) &\to \Omega^0(M)\\ f &\mapsto \Phi^*(f) := f \circ \phi.\end{aligned}$$

This works for any smooth map ϕ, and it leads to a slight modification of our mantra:

Vectors are pushed forward,
forms are pulled back.

The tensor product $\otimes$ does not interact well with forms, since the tensor product of two forms is not necessarily a form (it might be, for example, a symmetric $(0, n)$ tensor which, be definition, is not a form). Hence, we define the following.

Definition 4.49 (Wedge Product)**.** *Let M be a smooth manifold. We define the* ***wedge*** *(or* exterior*)* product *of forms as the map*

$$\begin{aligned} \wedge \colon \Omega^n(M) \times \Omega^m(M) &\to \Omega^{n+m}(M) \\ (\omega, \sigma) &\mapsto \omega \wedge \sigma, \end{aligned}$$

where

$$(\omega \wedge \sigma)(X_1, \ldots, X_{n+m}) := \frac{1}{n!\, m!} \sum_{\pi \in S_{n+m}} \operatorname{sgn}(\pi)(\omega \otimes \sigma)(X_{\pi(1)}, \ldots, X_{\pi(n+m)})$$

and $X_1, \ldots, X_{n+m} \in \Gamma(TM)$. By convention, for any $f, g \in \Omega^0(M)$ and $\omega \in \Omega^n(M)$, we set

$$f \wedge g := fg \qquad \text{and} \qquad f \wedge \omega = \omega \wedge f = f\omega.$$

Example 4.20. Suppose that $\omega, \sigma \in \Omega^1(M)$. Then, for any $X, Y \in \Gamma(TM)$

$$\begin{aligned} (\omega \wedge \sigma)(X, Y) &= (\omega \otimes \sigma)(X, Y) - (\omega \otimes \sigma)(Y, X) \\ &= (\omega \otimes \sigma)(X, Y) - \omega(Y)\sigma(X) \\ &= (\omega \otimes \sigma)(X, Y) - (\sigma \otimes \omega)(X, Y) \\ &= (\omega \otimes \sigma - \sigma \otimes \omega)(X, Y). \end{aligned}$$

Hence

$$\omega \wedge \sigma = \omega \otimes \sigma - \sigma \otimes \omega.$$

The wedge product is bilinear over $\mathcal{C}^\infty(M)$, that is

$$(f\omega_1 + \omega_2) \wedge \sigma = f\,\omega_1 \wedge \sigma + \omega_2 \wedge \sigma,$$

for all $f \in \mathcal{C}^\infty(M)$, $\omega_1, \omega_2 \in \Omega^n(M)$ and $\sigma \in \Omega^m(M)$, and similarly for the second argument.

Remark 4.18. If (U, x) is a chart on M, then every n-form $\omega \in \Omega^n(U)$ can be expressed locally on U as

$$\omega = \omega_{a_1 \cdots a_n}\, \mathrm{d}x^{a_1} \wedge \cdots \wedge \mathrm{d}x^{a_n},$$

where $\omega_{a_1 \cdots a_n} \in \mathcal{C}^\infty(U)$ and $1 \le a_1 < \cdots < a_n \le \dim M$. The $\mathrm{d}x^{a_i}$ appearing above are the covector fields (1-forms)

$$\mathrm{d}x^{a_i} \colon p \mapsto \mathrm{d}_p x^{a_i}.$$

The pull-back distributes over the wedge product.

Theorem 4.5. *Let $\phi \colon M \to N$ be smooth, $\omega \in \Omega^n(N)$ and $\sigma \in \Omega^m(N)$. Then, we have*

$$\Phi^*(\omega \wedge \sigma) = \Phi^*(\omega) \wedge \Phi^*(\sigma).$$

Proof. Let $p \in M$ and $X_1, \dots, X_{n+m} \in T_pM$. Then we have

$$\begin{aligned}
&\big(\Phi^*(\omega) \wedge \Phi^*(\sigma)\big)(p)(X_1, \dots, X_{n+m}) \\
&\qquad = \frac{1}{n!\,m!} \sum_{\pi \in S_{n+m}} \operatorname{sgn}(\pi)\big(\Phi^*(\omega) \otimes \Phi^*(\sigma)\big)(p)(X_{\pi(1)}, \dots, X_{\pi(n+m)}) \\
&\qquad = \frac{1}{n!\,m!} \sum_{\pi \in S_{n+m}} \operatorname{sgn}(\pi)\Phi^*(\omega)(p)(X_{\pi(1)}, \dots, X_{\pi(n)}) \\
&\qquad\qquad\qquad\qquad\qquad\qquad \Phi^*(\sigma)(p)(X_{\pi(n+1)}, \dots, X_{\pi(n+m)}) \\
&\qquad = \frac{1}{n!\,m!} \sum_{\pi \in S_{n+m}} \operatorname{sgn}(\pi)\omega(\phi(p))\big(\phi_*(X_{\pi(1)}), \dots, \phi_*(X_{\pi(n)})\big) \\
&\qquad\qquad\qquad\qquad\qquad\qquad \sigma(\phi(p))\big(\phi_*(X_{\pi(n+1)}), \dots, \phi_*(X_{\pi(n+m)})\big) \\
&\qquad = \frac{1}{n!\,m!} \sum_{\pi \in S_{n+m}} \operatorname{sgn}(\pi)(\omega \otimes \sigma)(\phi(p))\big(\phi_*(X_{\pi(1)}), \dots, \phi_*(X_{\pi(n+m)})\big) \\
&\qquad = (\omega \wedge \sigma)(\phi(p))\big(\phi_*(X_1), \dots, \phi_*(X_{n+m})\big) \\
&\qquad = \Phi^*(\omega \wedge \sigma)(p)(X_1, \dots, X_{n+m}).
\end{aligned}$$

Since $p \in M$ was arbitrary, the statement follows. □

4.11.1 The Grassmann Algebra

Note that the wedge product takes two differential forms and produces a differential form of a different type. It would be much nicer to have a space which is closed under the action of $\wedge$. In fact, such a space exists and it is called the Grassmann algebra of M.

Definition 4.50 (Grassmann Algebra)**.** *Let M be a smooth manifold. Define the $\mathcal{C}^\infty(M)$-module*

$$\operatorname{Gr}(M) \equiv \Omega(M) := \bigoplus_{n=0}^{\dim M} \Omega^n(M).$$

The ***Grassmann algebra*** *on M is the algebra $(\Omega(M), +, \cdot, \wedge)$, where*

$$\wedge \colon \Omega(M) \times \Omega(M) \to \Omega(M)$$

is the linear continuation of the previously defined $\wedge \colon \Omega^n(M) \times \Omega^m(M) \to \Omega^{n+m}(M)$.

Recall that the direct sum of modules has the Cartesian product of the modules as underlying set and module operations defined componentwise. Also, note that by "algebra" here we really mean "algebra over a module".

Example 4.21. Let $\psi = \omega + \sigma$, where $\omega \in \Omega^1(M)$ and $\sigma \in \Omega^3(M)$. Of course, this "+" is neither the addition on $\Omega^1(M)$ nor the one on $\Omega^3(M)$, but rather that on $\Omega(M)$ and, in fact, $\psi \in \Omega(M)$.

Let $\varphi \in \Omega^n(M)$, for some n. Then

$$\varphi \wedge \psi = \varphi \wedge (\omega + \sigma) = \varphi \wedge \omega + \varphi \wedge \sigma,$$

where $\varphi \wedge \omega \in \Omega^{n+1}(M)$, $\varphi \wedge \sigma \in \Omega^{n+3}(M)$, and $\varphi \wedge \psi \in \Omega(M)$.

Example 4.22. There is a lot of talk about *Grassmann numbers*, particularly in supersymmetry. One often hears that these are "numbers that do not commute, but anticommute". Of course, objects cannot be commutative or anticommutative by themselves. These qualifiers only apply to operations on the objects. In fact, the Grassmann numbers are just the elements of a Grassmann algebra.

The following result is about the anticommutative behaviour of $\wedge$.

Theorem 4.6. *Let $\omega \in \Omega^n(M)$ and $\sigma \in \Omega^m(M)$. Then*

$$\omega \wedge \sigma = (-1)^{nm}\, \sigma \wedge \omega.$$

We say that $\wedge$ is *graded commutative*, that is, it satisfies a version of anticommutativity which depends on the degrees of the forms.

Proof. First note that if $\omega, \sigma \in \Omega^1(M)$, then

$$\omega \wedge \sigma = \omega \otimes \sigma - \sigma \otimes \omega = -\sigma \wedge \omega.$$

Recall that is $\omega \in \Omega^n(M)$ and $\sigma \in \Omega^m(M)$, then locally on a chart (U, x) we can write

$$\begin{aligned}\omega &= \omega_{a_1 \cdots a_n} \mathrm{d}x^{a_1} \wedge \cdots \wedge \mathrm{d}x^{a_n} \\ \sigma &= \sigma_{b_1 \cdots b_m} \mathrm{d}x^{b_1} \wedge \cdots \wedge \mathrm{d}x^{b_m}\end{aligned}$$

with $1 \leq a_1 < \cdots < a_n \leq \dim M$ and similarly for the b_i. The coefficients $\omega_{a_1 \cdots a_n}$ and $\sigma_{b_1 \cdots b_m}$ are smooth functions in $\mathcal{C}^\infty(U)$. Since $\mathrm{d}x^{a_i}, \mathrm{d}x^{b_j} \in \Omega^1(M)$, we have

$$\begin{aligned}\omega \wedge \sigma &= \omega_{a_1 \cdots a_n} \sigma_{b_1 \cdots b_m} \, \mathrm{d}x^{a_1} \wedge \cdots \wedge \mathrm{d}x^{a_n} \wedge \mathrm{d}x^{b_1} \wedge \cdots \wedge \mathrm{d}x^{b_m} \\ &= (-1)^n \, \omega_{a_1 \cdots a_n} \sigma_{b_1 \cdots b_m} \, \mathrm{d}x^{b_1} \wedge \mathrm{d}x^{a_1} \wedge \cdots \wedge \mathrm{d}x^{a_n} \wedge \mathrm{d}x^{b_2} \wedge \cdots \wedge \mathrm{d}x^{b_m} \\ &= (-1)^{2n} \, \omega_{a_1 \cdots a_n} \sigma_{b_1 \cdots b_m} \, \mathrm{d}x^{b_1} \wedge \mathrm{d}x^{b_2} \wedge \mathrm{d}x^{a_1} \wedge \cdots \wedge \mathrm{d}x^{a_n} \wedge \mathrm{d}x^{b_3} \wedge \cdots \wedge \mathrm{d}x^{b_m} \\ &\vdots \\ &= (-1)^{nm} \, \omega_{a_1 \cdots a_n} \sigma_{b_1 \cdots b_m} \, \mathrm{d}x^{b_1} \wedge \cdots \wedge \mathrm{d}x^{b_m} \wedge \mathrm{d}x^{a_1} \wedge \cdots \wedge \mathrm{d}x^{a_n} \\ &= (-1)^{nm} \, \sigma \wedge \omega\end{aligned}$$

since we have swapped 1-forms nm-many times. □

Remark 4.19. We should stress that this is only true when ω and σ are pure degree forms, rather than linear combinations of forms of different degrees. Indeed, if $\varphi, \psi \in \Omega(M)$, a formula like

$$\varphi \wedge \psi = \cdots \psi \wedge \varphi$$

does not make sense in principle, because the different parts of φ and ψ can have different commutation behaviours.

4.11.2 The Exterior Derivative

Recall the "extended" definition of the gradient operator of a function d on the whole manifold M:

$$\begin{aligned}\mathrm{d} \colon \mathcal{C}^\infty(M) &\xrightarrow{\sim} \Gamma(T^*M) \\ f &\mapsto \mathrm{d}f\end{aligned}$$

Since $\Omega^0(M) \equiv \mathcal{C}^\infty(M)$ and $\Omega^1(M) \equiv \Gamma(T^0_1 M) \equiv \Gamma(T^*M)$, we can also understand this as an operator that takes in 0-forms and outputs 1-forms

$$\mathrm{d} \colon \Omega^0(M) \xrightarrow{\sim} \Omega^1(M).$$

This can then be extended to an operator which acts on any n-form. For this definition, we need to remind ourselves of the definition of commutator we gave in the algebra section of the notes. More precisely, if M is a smooth manifold and $X, Y \in \Gamma(TM)$ then the commutator (or Lie bracket) of X and Y is defined as

$$\begin{aligned}[X, Y] \colon \mathcal{C}^\infty(M) &\xrightarrow{\sim} \mathcal{C}^\infty(M) \\ f &\mapsto [X, Y](f) := X(Y(f)) - Y(X(f)),\end{aligned}$$

where we are using the definition of vector fields as $\mathbb{R}$-linear maps $\mathcal{C}^\infty(M) \xrightarrow{\sim} \mathcal{C}^\infty(M)$.

Using the commutator we can now extend the gradient as follows:

Definition 4.51 (Exterior Derivative)**.** *The* ***exterior derivative*** *on M is the $\mathbb{R}$-linear operator*

$$\begin{aligned}\mathrm{d} \colon \Omega^n(M) &\xrightarrow{\sim} \Omega^{n+1}(M) \\ \omega &\mapsto \mathrm{d}\omega\end{aligned}$$

with $\mathrm{d}\omega$ *being defined as*

$$\mathrm{d}\omega(X_1,\ldots,X_{n+1}) := \sum_{i=1}^{n+1}(-1)^{i+1}\,X_i\big(\omega(X_1,\ldots,\widehat{X_i},\ldots,X_{n+1})\big) + \sum_{i<j}(-1)^{i+j}\,\omega\big([X_i,X_j],X_1,\ldots,\widehat{X_i},\ldots,\widehat{X_j},\ldots,X_{n+1}\big),$$

where $X_i \in \Gamma(TM)$ *and the hat denotes omissions.*

Remark 4.20. Note that the operator d is only well-defined when it acts on forms. In order to define a derivative operator on general tensors we will need to add extra structure to our differentiable manifold.

Example 4.23. In the case $n = 1$, the form $\mathrm{d}\omega \in \Omega^2(M)$ is given by

$$\mathrm{d}\omega(X,Y) = X(\omega(Y)) - Y(\omega(X)) - \omega([X,Y]).$$

Let us check that this is indeed a 2-form, i.e. an antisymmetric, $\mathcal{C}^\infty(M)$-multilinear map

$$\mathrm{d}\omega\colon \Gamma(TM)\times\Gamma(TM) \to \mathcal{C}^\infty(M).$$

By using the antisymmetry of the Lie bracket, we immediately get

$$\mathrm{d}\omega(X,Y) = -\,\mathrm{d}\omega(Y,X).$$

Moreover, thanks to this identity, it suffices to check $\mathcal{C}^\infty(M)$-linearity in the first argument only. Additivity is easily checked

$$\begin{aligned}
\mathrm{d}\omega(X_1+X_2,Y) &= (X_1+X_2)(\omega(Y)) - Y(\omega(X_1+X_2)) - \omega([X_1+X_2,Y])\\
&= X_1(\omega(Y)) + X_2(\omega(Y)) - Y(\omega(X_1)+\omega(X_2)) - \omega([X_1,Y]+[X_2,Y])\\
&= X_1(\omega(Y)) + X_2(\omega(Y)) - Y(\omega(X_1)) - Y(\omega(X_2)) - \omega([X_1,Y]) - \omega([X_2,Y])\\
&= \mathrm{d}\omega(X_1,Y) + \mathrm{d}\omega(X_2,Y).
\end{aligned}$$

For $\mathcal{C}^\infty(M)$-scaling, first we calculate $[fX,Y]$. Let $g \in \mathcal{C}^\infty(M)$. Then

$$\begin{aligned}
[fX,Y](g) &= fX(Y(g)) - Y(fX(g))\\
&= fX(Y(g)) - fY(X(g)) - Y(f)X(g)\\
&= f(X(Y(g)) - Y(X(g))) - Y(f)X(g)\\
&= f[X,Y](g) - Y(f)X(g)\\
&= (f[X,Y] - Y(f)X)(g).
\end{aligned}$$

Therefore

$$[fX,Y] = f[X,Y] - Y(f)X.$$

Hence, we can calculate

$$\begin{aligned}
\mathrm{d}\omega(fX,Y) &= fX(\omega(Y)) - Y(\omega(fX)) - \omega([fX,Y])\\
&= fX(\omega(Y)) - Y(f\omega(X)) - \omega(f[X,Y] - Y(f)X)\\
&= fX(\omega(Y)) - fY(\omega(X)) - Y(f)\omega(X) - f\omega([X,Y]) + \omega(Y(f)X)\\
&= fX(\omega(Y)) - fY(\omega(X)) - /[gray]Y(f)\omega(X) - f\omega([X,Y]) + /[gray]Y(f)\omega(X)\\
&= f\,\mathrm{d}\omega(X,Y),
\end{aligned}$$

which is what we wanted.

The exterior derivative satisfies a graded version of the Leibniz rule with respect to the wedge product.

Theorem 4.7. *Let* $\omega \in \Omega^n(M)$ *and* $\sigma \in \Omega^m(M)$. *Then*

$$\mathrm{d}(\omega\wedge\sigma) = \mathrm{d}\omega\wedge\sigma + (-1)^n\,\omega\wedge\mathrm{d}\sigma.$$

Proof. We will work in local coordinates. Let (U, x) be a chart on M and write

$$\begin{aligned}\omega &= \omega_{a_1\cdots a_n} \mathrm{d}x^{a_1} \wedge \cdots \wedge \mathrm{d}x^{a_n} =: \omega_A \mathrm{d}x^A \\ \sigma &= \sigma_{b_1\cdots b_m} \mathrm{d}x^{b_1} \wedge \cdots \wedge \mathrm{d}x^{b_m} =: \sigma_B \mathrm{d}x^B.\end{aligned}$$

Locally, the exterior derivative operator d acts as

$$\mathrm{d}\omega = \mathrm{d}\omega_A \wedge \mathrm{d}x^A.$$

Hence

$$\begin{aligned}\mathrm{d}(\omega \wedge \sigma) &= \mathrm{d}(\omega_A \sigma_B \, \mathrm{d}x^A \wedge \mathrm{d}x^B) \\ &= \mathrm{d}(\omega_A \sigma_B) \wedge \mathrm{d}x^A \wedge \mathrm{d}x^B \\ &= (\sigma_B \mathrm{d}\omega_A + \omega_A \mathrm{d}\sigma_B) \wedge \mathrm{d}x^A \wedge \mathrm{d}x^B \\ &= \sigma_B \mathrm{d}\omega_A \wedge \mathrm{d}x^A \wedge \mathrm{d}x^B + \omega_A \mathrm{d}\sigma_B \wedge \mathrm{d}x^A \wedge \mathrm{d}x^B \\ &= \sigma_B \mathrm{d}\omega_A \wedge \mathrm{d}x^A \wedge \mathrm{d}x^B + (-1)^n \omega_A \mathrm{d}x^A \wedge \mathrm{d}\sigma_B \wedge \mathrm{d}x^B \\ &= \sigma_B \mathrm{d}\omega \wedge \mathrm{d}x^B + (-1)^n \omega_A \mathrm{d}x^A \wedge \mathrm{d}\sigma \\ &= \mathrm{d}\omega \wedge \sigma + (-1)^n \, \omega \wedge \mathrm{d}\sigma\end{aligned}$$

since we have "anticommuted" the 1-form $\mathrm{d}\sigma_B$ through the n-form $\mathrm{d}x^A$, picking up n minus signs in the process. □

An important property of the exterior derivative is the following.

Theorem 4.8. *Let $\phi \colon M \to N$ be smooth. For any $\omega \in \Omega^n(N)$, we have*

$$\Phi^*(\mathrm{d}\omega) = \mathrm{d}(\Phi^*(\omega)).$$

Remark 4.21. Informally, we can write this result as $\Phi^*\mathrm{d} = \mathrm{d}\Phi^*$, and say that the exterior derivative "commutes" with the pull-back.

However, you should bear in mind that the two d's appearing in the statement are two different operators. On the left hand side, it is $\mathrm{d}\colon \Omega^n(N) \to \Omega^{n+1}(N)$, while it is $\mathrm{d}\colon \Omega^n(M) \to \Omega^{n+1}(M)$ on the right hand side.

Remark 4.22. Of course, we could also combine the operators d into a single operator acting on the Grassmann algebra on M

$$\mathrm{d}\colon \Omega(M) \to \Omega(M)$$

by linear continuation.

4.11.3 De Rham Cohomology

Definition 4.52 (Closed / Exact Forms)**.** *Let M be a smooth manifold and let $\omega \in \Omega^n(M)$. We say that ω is*

- ***closed*** *if* $\mathrm{d}\omega = 0$*;*
- ***exact*** *if* $\exists\, \sigma \in \Omega^{n-1}(M) : \omega = \mathrm{d}\sigma$.

The question of whether every closed form is exact and vice versa, i.e. whether the implications

$$(\mathrm{d}\omega = 0) \Leftrightarrow (\exists\, \sigma : \omega = \mathrm{d}\sigma)$$

hold in general, belongs to the branch of mathematics called cohomology theory, to which we will now provide an introduction.

The answer for the $\Leftarrow$ direction is affirmative thanks to the following result.

Theorem 4.9. *Let M be a smooth manifold. The operator*

$$\mathrm{d}^2 \equiv \mathrm{d} \circ \mathrm{d} \colon \Omega^n(M) \to \Omega^{n+2}(M)$$

is identically zero, i.e. $\mathrm{d}^2 = 0$.

Proof. This can be shown directly using the definition of d. Here, we will instead show it by working in local coordinates.

Recall that, locally on a chart (U, x), we can write any form $\omega \in \Omega^n(M)$ as

$$\omega = \omega_{a_1 \cdots a_n} \mathrm{d}x^{a_1} \wedge \cdots \wedge \mathrm{d}x^{a_n}.$$

Then, we have

$$\begin{aligned} \mathrm{d}\omega &= \mathrm{d}\omega_{a_1 \cdots a_n} \wedge \mathrm{d}x^{a_1} \wedge \cdots \wedge \mathrm{d}x^{a_n} \\ &= \partial_b \omega_{a_1 \cdots a_n} \mathrm{d}x^b \wedge \mathrm{d}x^{a_1} \wedge \cdots \wedge \mathrm{d}x^{a_n}, \end{aligned}$$

and hence

$$\mathrm{d}^2\omega = \partial_c \partial_b \omega_{a_1 \cdots a_n} \mathrm{d}x^c \wedge \mathrm{d}x^b \wedge \mathrm{d}x^{a_1} \wedge \cdots \wedge \mathrm{d}x^{a_n}.$$

We can perform a little "trick" in the last equation and write it as twice the half expression

$$\mathrm{d}^2\omega = \frac{1}{2}\partial_c \partial_b \omega_{a_1 \cdots a_n} \mathrm{d}x^c \wedge \mathrm{d}x^b \wedge \mathrm{d}x^{a_1} \wedge \cdots \wedge \mathrm{d}x^{a_n} + \frac{1}{2}\partial_c \partial_b \omega_{a_1 \cdots a_n} \mathrm{d}x^c \wedge \mathrm{d}x^b \wedge \mathrm{d}x^{a_1} \wedge \cdots \wedge \mathrm{d}x^{a_n}$$

Now we can inter-switch the c and b dummy indices in the second half part (we can do it since they are just dummy indices) and we get

$$\mathrm{d}^2\omega = \frac{1}{2}\partial_c \partial_b \omega_{a_1 \cdots a_n} \mathrm{d}x^c \wedge \mathrm{d}x^b \wedge \mathrm{d}x^{a_1} \wedge \cdots \wedge \mathrm{d}x^{a_n} + \frac{1}{2}\partial_b \partial_c \omega_{a_1 \cdots a_n} \mathrm{d}x^b \wedge \mathrm{d}x^c \wedge \mathrm{d}x^{a_1} \wedge \cdots \wedge \mathrm{d}x^{a_n}$$

Since $\mathrm{d}x^b \wedge \mathrm{d}x^c = -\,\mathrm{d}x^c \wedge \mathrm{d}x^b$, and moreover, by Schwarz's theorem, we have $\partial_c \partial_b \omega_{a_1 \cdots a_n} = \partial_b \partial_c \omega_{a_1 \cdots a_n}$ we get

$$\mathrm{d}^2\omega = \frac{1}{2}\partial_c \partial_b \omega_{a_1 \cdots a_n} \mathrm{d}x^c \wedge \mathrm{d}x^b \wedge \mathrm{d}x^{a_1} \wedge \cdots \wedge \mathrm{d}x^{a_n} - \frac{1}{2}\partial_c \partial_b \omega_{a_1 \cdots a_n} \mathrm{d}x^c \wedge \mathrm{d}x^b \wedge \mathrm{d}x^{a_1} \wedge \cdots \wedge \mathrm{d}x^{a_n}$$

Hence

$$\mathrm{d}^2\omega = 0$$

Since this holds for any ω, we have $\mathrm{d}^2 = 0$. ☐

Corollary 4.1. *Every exact form is closed.*

We can extend the action of d to the zero vector space $0 := \{0\}$ by mapping the zero in 0 to the zero function in $\Omega^0(M)$. In this way, we obtain the chain of $\mathbb{R}$-linear maps

$$0 \xrightarrow{\mathrm{d}} \Omega^0(M) \xrightarrow{\mathrm{d}} \Omega^1(M) \xrightarrow{\mathrm{d}} \cdots \xrightarrow{\mathrm{d}} \Omega^n(M) \xrightarrow{\mathrm{d}} \Omega^{n+1}(M) \xrightarrow{\mathrm{d}} \cdots \xrightarrow{\mathrm{d}} \Omega^{\dim M}(M) \xrightarrow{\mathrm{d}} 0,$$

where we now think of the spaces $\Omega^n(M)$ as $\mathbb{R}$-vector spaces.

Recall from linear algebra section in the notes that, given a linear map $\phi\colon V \to W$, one can define the subspace of V

$$\ker(\phi) := \{v \in V \mid \phi(v) = 0\},$$

called the *kernel* of ϕ, and the subspace of W

$$\operatorname{im}(\phi) := \{\phi(v) \mid v \in V\},$$

called the *image* of ϕ.

Going back to our chain of maps, the equation $\mathrm{d}^2 = 0$ is equivalent to

$$\operatorname{im}(\mathrm{d}\colon \Omega^n(M) \to \Omega^{n+1}(M)) \subseteq \ker(\mathrm{d}\colon \Omega^{n+1}(M) \to \Omega^{n+2}(M))$$

for all $0 \leq n \leq \dim M - 2$. Moreover, we have

$$\omega \in \Omega^n(M) \text{ is closed } \Leftrightarrow \omega \in \ker(\mathrm{d}\colon \Omega^n(M) \to \Omega^{n+1}(M))$$
$$\omega \in \Omega^n(M) \text{ is exact } \Leftrightarrow \omega \in \operatorname{im}(\mathrm{d}\colon \Omega^{n-1}(M) \to \Omega^n(M)).$$

The traditional notation for the spaces on the right hand side above is

$$Z^n := \ker(\mathrm{d}\colon \Omega^n(M) \to \Omega^{n+1}(M)),$$
$$B^n := \operatorname{im}(\mathrm{d}\colon \Omega^{n-1}(M) \to \Omega^n(M)),$$

so that Z^n is the space of closed n-forms and B^n is the space of exact n-forms.

Our original question can be restated as: does $Z^n = B^n$ for all n? We have already seen that $\mathrm{d}^2 = 0$ implies that $B^n \subseteq Z^n$ for all n (B^n is, in fact, a vector subspace of Z^n). Unfortunately the equality does not hold in general, but we do have the following result.

Lemma 4.1 (Poincaré)**.** *Let $M \subseteq \mathbb{R}^d$ be a simply connected domain. Then*

$$Z^n = B^n, \qquad \forall\, n > 0.$$

In the cases where $Z^n \neq B^n$, we would like to quantify by how much the closed n-forms fail to be exact. The answer is provided by the cohomology group.

Definition 4.53 (de Rham Cohomology Group)**.** *Let M be a smooth manifold. The n-th* ***de Rham cohomology group*** *on M is the quotient $\mathbb{R}$-vector space*

$$H^n(M) := Z^n / B^n.$$

You can think of the above quotient as $Z^n/\sim$, where $\sim$ is the equivalence relation

$$\omega \sim \sigma \;:\Leftrightarrow\; \omega - \sigma \in B^n.$$

The answer to our question as it is addressed in cohomology theory is: every exact n-form on M is also closed and vice versa if, only if,

$$H^n(M) \cong_{\mathrm{vec}} 0.$$

Of course, rather than an actual answer, this is yet another restatement of the question. However, if we are able to determine the spaces $H^n(M)$, then we do get an answer.

A crucial theorem by de Rham states (in more technical terms) that $H^n(M)$ only depends on the global topology of M. In other words, the cohomology groups are topological invariants. This is remarkable because $H^n(M)$ is defined in terms of exterior derivatives, which have everything to do with the local differentiable structure of M, and a given topological space can be equipped with several inequivalent differentiable structures.

Example 4.24. Let M be any smooth manifold. We have

$$H^0(M) \cong_{\mathrm{vec}} \mathbb{R}^{(\#\text{ of connected components of } M)}$$

since the closed 0-forms are just the locally constant smooth functions on M. As an immediate consequence, we have

$$H^0(\mathbb{R}) \cong_{\mathrm{vec}} H^0(S^1) \cong_{\mathrm{vec}} \mathbb{R}.$$

Example 4.25. By Poincaré lemma, we have

$$H^n(M) \cong_{\mathrm{vec}} 0$$

for any simply connected $M \subseteq \mathbb{R}^d$.

4.12 Application - Part 1: $\mathrm{SL}(2,\mathbb{C})$

In this final chapter we will go through an application containing (almost) everything we have mentioned so far. More specifically, we we will examine in detail the special linear group of degree 2 over $\mathbb{C}$, also known as the relativistic spin group.

$\mathrm{SL}(2,\mathbb{C})$ **As A Set**

We define the following subset of $\mathbb{C}^4 := \mathbb{C}\times\mathbb{C}\times\mathbb{C}\times\mathbb{C}$

$$\mathrm{SL}(2,\mathbb{C}) := \left\{ \begin{pmatrix} a & b \\ c & d \end{pmatrix} \in \mathbb{C}^4 \;\middle|\; ad - bc = 1 \right\},$$

where the array is just an alternative notation for a quadruple (a,b,c,d). It's this extra constraint $ad - bc = 1$ that removes one degree of freedom and makes it a subset and not the whole $\mathbb{C}^4$.

$\mathrm{SL}(2,\mathbb{C})$ **As A Group**

We define an operation

$$\begin{aligned} \bullet \colon \mathrm{SL}(2,\mathbb{C}) \times \mathrm{SL}(2,\mathbb{C}) &\to \mathrm{SL}(2,\mathbb{C}) \\ (\begin{pmatrix} a & b \\ c & d \end{pmatrix}, \begin{pmatrix} e & f \\ g & h \end{pmatrix}) &\mapsto \begin{pmatrix} a & b \\ c & d \end{pmatrix} \bullet \begin{pmatrix} e & f \\ g & h \end{pmatrix}, \end{aligned}$$

where

$$\begin{pmatrix} a & b \\ c & d \end{pmatrix} \bullet \begin{pmatrix} e & f \\ g & h \end{pmatrix} := \begin{pmatrix} ae+bg & af+bh \\ ce+dg & cf+dh \end{pmatrix}.$$

Formally, this operation is the same as matrix multiplication. We can check directly that the result of applying $\bullet$ lands back in $\mathrm{SL}(2,\mathbb{C})$, or simply recall that the determinant of a product is the product of the determinants. Moreover, the operation $\bullet$

i) is associative (straightforward but tedious to check);

ii) has an identity element, namely $\begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} \in \mathrm{SL}(2,\mathbb{C})$;

iii) admits inverses: for each $\begin{pmatrix} a & b \\ c & d \end{pmatrix} \in \mathrm{SL}(2,\mathbb{C})$, we have $\begin{pmatrix} d & -b \\ -c & a \end{pmatrix} \in \mathrm{SL}(2,\mathbb{C})$ and

$$\begin{pmatrix} a & b \\ c & d \end{pmatrix} \bullet \begin{pmatrix} d & -b \\ -c & a \end{pmatrix} = \begin{pmatrix} d & -b \\ -c & a \end{pmatrix} \bullet \begin{pmatrix} a & b \\ c & d \end{pmatrix} = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}.$$

Hence, we have $\begin{pmatrix} a & b \\ c & d \end{pmatrix}^{-1} = \begin{pmatrix} d & -b \\ -c & a \end{pmatrix}$.

Therefore, the pair $(\mathrm{SL}(2,\mathbb{C}), \bullet)$ is a (non-commutative) group.

$\mathrm{SL}(2,\mathbb{C})$ **As A Topological Space**

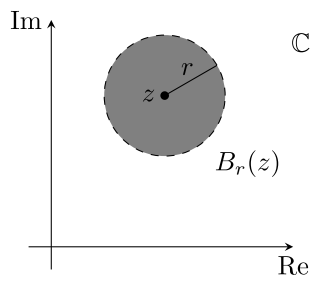
Recall that if N is a subset of M and $\mathcal{O}$ is a topology on M, then we can equip N with the subset topology inherited from M

$$\mathcal{O}|_N := \{U \cap N \mid U \in \mathcal{O}\}.$$

We begin by establishing a topology on $\mathbb{C}$ as follows. Let

$$B_r(z) := \{y \in \mathbb{C} \mid |z - y| < r\}$$

be the open ball of radius $r > 0$ and centre $z \in \mathbb{C}$.



Define $\mathcal{O}_{\mathbb{C}}$ implicitly by

$$U \in \mathcal{O}_{\mathbb{C}} \;:\Leftrightarrow\; \forall\, z \in U : \exists\, r > 0 : B_r(z) \subseteq U.$$

Then, the pair $(\mathbb{C}, \mathcal{O}_{\mathbb{C}})$ is a topological space. In fact, we have

$$(\mathbb{C}, \mathcal{O}_{\mathbb{C}}) \cong_{\text{top}} (\mathbb{R}^2, \mathcal{O}_{\text{std}}).$$

We can then equip $\mathbb{C}^4$ with the product topology so that we can finally define

$$\mathcal{O} := (\mathcal{O}_{\mathbb{C}})|_{\mathrm{SL}(2,\mathbb{C})},$$

so that the pair $(\mathrm{SL}(2,\mathbb{C}), \mathcal{O})$ is a topological space. In fact, it is a connected topological space, and we will need this property later on.

$\mathrm{SL}(2,\mathbb{C})$ As A Topological Manifold

Recall that a topological space $(M, \mathcal{O})$ is a complex topological manifold if each point $p \in M$ has an open neighbourhood $U(p)$ which is homeomorphic to an open subset of $\mathbb{C}^d$. Equivalently, there must exist a $\mathcal{C}^0$-atlas, i.e. a collection $\mathscr{A}$ of charts (U_α, x_α), where the U_α are open and cover M and each x is a homeomorphism onto a subset of $\mathbb{C}^d$.

Let U be the set

$$U := \left\{ \begin{pmatrix} a & b \\ c & d \end{pmatrix} \in \mathrm{SL}(2,\mathbb{C}) \;\middle|\; a \neq 0 \right\}$$

and define the map

$$\begin{aligned} x\colon \quad & U \to x(U) \subseteq \mathbb{C}^* \times \mathbb{C} \times \mathbb{C} \\ & \begin{pmatrix} a & b \\ c & d \end{pmatrix} \mapsto (a, b, c), \end{aligned}$$

where $\mathbb{C}^* = \mathbb{C} \setminus \{0\}$.

As we said in the beginning of this chapter, $\mathrm{SL}(2,\mathbb{C})$ (as a set) is a subset of $\mathbb{C}^4 := \mathbb{C} \times \mathbb{C} \times \mathbb{C} \times \mathbb{C}$ due to the constraint $ad - bc = 1$ that removes one degree of freedom. This is why we map it (locally) to $\mathbb{C}^3$ and not $\mathbb{C}^4$. Because given the mapping (a, b, c) one can reconstruct d as $d = \frac{1+bc}{a}$.

With a little more work on this direction, one can show that U is an open subset of $(\mathrm{SL}(2,\mathbb{C}), \mathcal{O})$ and x is a homeomorphism with inverse

$$\begin{aligned} x^{-1}\colon \quad & x(U) \to U \\ & (a, b, c) \mapsto \begin{pmatrix} a & b \\ c & \frac{1+bc}{a} \end{pmatrix}. \end{aligned}$$

This is the reason why we excluded the case $a = 0$ when we defined the set U of the chart (U, x), since if we hadn't, we wouldn't be able to divide with a and the map x wouldn't have an inverse. However, this makes the chart (U, x) to not cover the whole $\mathrm{SL}(2,\mathbb{C})$ since U as a set takes care only the elements of $\mathrm{SL}(2,\mathbb{C})$ with $a \neq 0$. Hence we need at least one more chart. We thus define the set

$$V := \left\{ \begin{pmatrix} a & b \\ c & d \end{pmatrix} \in \mathrm{SL}(2,\mathbb{C}) \;\middle|\; b \neq 0 \right\}$$

and the map

$$\begin{aligned} y\colon \quad & V \to x(V) \subseteq \mathbb{C} \times \mathbb{C}^* \times \mathbb{C} \\ & \begin{pmatrix} a & b \\ c & d \end{pmatrix} \mapsto (a, b, d). \end{aligned}$$

Similarly to the above, V is open and y is a homeomorphism with inverse

$$\begin{aligned} y^{-1}\colon \quad & x(V) \to V \\ & (a, b, d) \mapsto \begin{pmatrix} a & b \\ \frac{ad-1}{b} & d \end{pmatrix}. \end{aligned}$$

An element of $\mathrm{SL}(2,\mathbb{C})$ cannot have both a and b equal to zero, for otherwise $ad - bc = 0 \neq 1$. Hence $\mathscr{A}_{\mathrm{top}} := \{(U, x), (V, y)\}$ is an atlas, and since every atlas is automatically a $\mathcal{C}^0$-atlas, the triple $(\mathrm{SL}(2,\mathbb{C}), \mathcal{O}, \mathscr{A}_{\mathrm{top}})$ is a 3-dimensional, complex, topological manifold.

$\mathrm{SL}(2,\mathbb{C})$ As A Complex Differentiable Manifold

Recall that to obtain a $\mathcal{C}^1$-differentiable manifold from a topological manifold with atlas $\mathscr{A}$, we have to check that every transition map between charts in $\mathscr{A}$ is differentiable in the usual sense.

In our case, we have the atlas $\mathscr{A}_{\mathrm{top}} := \{(U, x), (V, y)\}$. We evaluate

$$(y \circ x^{-1})(a, b, c) = y(\begin{pmatrix} a & b \\ c & \frac{1+bc}{a} \end{pmatrix}) = (a, b, \tfrac{1+bc}{a}).$$

Hence we have the transition map

$$\begin{aligned} y \circ x^{-1}\colon & x(U \cap V) \to y(U \cap V) \\ & (a, b, c) \mapsto (a, b, \tfrac{1+bc}{a}). \end{aligned}$$

Similarly, we have

$$(x \circ y^{-1})(a, b, d) = y(\begin{pmatrix} a & b \\ \frac{ad-1}{b} & d \end{pmatrix}) = (a, b, \tfrac{ad-1}{b}).$$

Hence, the other transition map is

$$\begin{aligned} x \circ y^{-1}\colon & y(U \cap V) \to x(U \cap V) \\ & (a, b, c) \mapsto (a, b, \tfrac{ad-1}{b}). \end{aligned}$$

Since $a \neq 0$ and $b \neq 0$, the transition maps are complex differentiable.

Therefore, the atlas $\mathscr{A}_{\mathrm{top}}$ is a differentiable atlas. By defining $\mathscr{A}$ to be the maximal differentiable atlas containing $\mathscr{A}_{\mathrm{top}}$, we have that $(\mathrm{SL}(2,\mathbb{C}), \mathcal{O}, \mathscr{A})$ is a 3-dimensional, complex differentiable manifold.

Chapter 5

Lie Theory

5.1 Lie Groups

Definition 5.1 (Lie Group)**.** *A* ***Lie group*** *is a group $(G, \bullet)$, where G is a smooth manifold and the maps*

$$\begin{aligned} \mu\colon\ & G \times G \to G \\ & (g_1, g_2) \mapsto g_1 \bullet g_2 \end{aligned}$$

and

$$\begin{aligned} i\colon G &\to G \\ g &\mapsto g^{-1} \end{aligned}$$

are both smooth. Note that $G \times G$ inherits a smooth atlas from the smooth atlas of G.

Definition 5.2 (Dimension Of Lie Group)**.** *The* ***dimension*** *of a Lie group $(G, \bullet)$ is the dimension of G as a manifold.*

Example 5.1.

a) Consider $(\mathbb{R}^n, +)$, where $\mathbb{R}^n$ is understood as a smooth n-dimensional manifold. This is a commutative (or abelian) Lie group (since $\bullet$ is commutative), often called the n-dimensional translation group.

b) Let $S^1 := \{z \in \mathbb{C} \mid |z| = 1\}$ and let $\cdot$ be the usual multiplication of complex numbers. Then $(S^1, \cdot)$ is a commutative Lie group usually denoted $\mathrm{U}(1)$.

c) Let $\mathrm{GL}(n, \mathbb{R}) = \{\phi\colon \mathbb{R}^n \xrightarrow{\sim} \mathbb{R}^n \mid \det \phi \neq 0\}$. This set can be endowed with the structure of a smooth n^2-dimensional manifold, by noting that there is a bijection between linear maps $\phi\colon \mathbb{R}^n \xrightarrow{\sim} \mathbb{R}^n$ and $\mathbb{R}^{2n}$. The condition $\det \phi \neq 0$ is a so-called *open condition*, meaning that $\mathrm{GL}(n, \mathbb{R})$ can be identified with an open subset of $\mathbb{R}^{2n}$, from which it then inherits a smooth structure.

   Then, $(\mathrm{GL}(n, \mathbb{R}), \circ)$ is a Lie group called the *general linear group*.

Definition 5.3 (Lie Group Homomorphism)**.** *Let $(G, \bullet)$ and $(H, \circ)$ be Lie groups. A map $\phi\colon G \to H$ is* ***Lie group homomorphism*** *if it is a group homomorphism and a smooth map.*

Definition 5.4 (Lie Group Isomorphism)**.** *A* ***Lie group isomorphism*** *is a Lie group homomorphism which is also a diffeomorphism of the underlying manifolds.*

5.1.1 The Left Translation Map

To every element of a Lie group there is associated a special map. Note that everything we will do here can be done equivalently by using right translation maps.

Definition 5.5 (Left Translation)**.** *Let $(G, \bullet)$ be a Lie group and let $g \in G$. The map*

$$\begin{aligned} \ell_g\colon G &\to G \\ h &\mapsto \ell_g(h) := g \bullet h \equiv gh \end{aligned}$$

*is called the **left translation** by g.*

If there is no danger of confusion, we usually suppress the $\bullet$ notation.

Proposition 5.1. *Let G be a Lie group. For any $g \in G$, the left translation map $\ell_g : G \to G$ is a diffeomorphism.*

Proof. Let $h, h' \in G$. Then, we have

$$\ell_g(h) = \ell_g(h') \iff gh = gh' \iff h = h'.$$

Moreover, for any $h \in G$, we have $g^{-1}h \in G$ and

$$\ell_g(g^{-1}h) = gg^{-1}h = h.$$

Therefore, ℓ_g is a bijection on G.

Note that

$$\ell_g = \mu(g, -)$$

and since $\mu : G \times G \to G$ is smooth by definition, so is ℓ_g.

The inverse map is $(\ell_g)^{-1} = \ell_{g^{-1}}$, since

$$\ell_{g^{-1}} \circ \ell_g = \ell_g \circ \ell_{g^{-1}} = \mathrm{id}_G .$$

Then, for the same reason as above with g replaced by g^{-1}, the inverse map $(\ell_g)^{-1}$ is also smooth. Hence, the map ℓ_g is indeed a diffeomorphism. □

Note that, in general, ℓ_g is not an isomorphism of groups, i.e.

$$\ell_g(hh') \neq \ell_g(h)\, \ell_g(h')$$

in general. However, as the final part of the previous proof suggests, we do have

$$\ell_g \circ \ell_h = \ell_{gh}$$

for all $g, h \in G$.

Recall from the previous chapter than once we have a diffeomorphism ϕ between two manifolds M and N, we can define the push-forward of a vector field X as

$$(\phi_* X)|_{\phi(p)} := \phi_*(X|_p)$$

where $X|_p$ is the vector created by the field X on point p.

Coming in our case, we just showed that the map $\ell_g : G \to G$ is a diffeomorphism so we can push-forward any vector field X on G to another vector field (again on G since the maps is between the same manifold). So in our case $\phi_*(X) = (\ell_g)_*(X)$ and for any point h in G: $\ell_g(h) = gh$ so the push-forward equation reads

$$(\ell_{g*} X)|_{gh} := (\ell_g)_*(X|_h)$$

5.1.2 The Lie Algebra Of A Lie Group

In Lie theory, we are typically not interested in general vector fields, but rather on special class of vector fields which are invariant under the induced push-forward of the left translation maps ℓ_g.

Definition 5.6 (Left Invariant Vector Fields)**.** *Let G be a Lie group. A vector field $X \in \Gamma(TG)$ is said to be **left-invariant** if*

$$\forall\, g \in G : \ (l_g)_*(X) = X.$$

Equivalently, we can require this to hold pointwise

$$\forall\, g, h \in G: \; (\ell_g)_*(X|_h) = X|_{gh}.$$

We can manipulate a bit the pointwise formulation to yield another reformulation. Since both sides are vectors we can let them act on a function f

$$(\ell_g)_*(X|_h)f = X|_{gh}f$$

By using the definition of a push-forward of a vector $(\phi_*)_p(X)f := X(f \circ \phi)$ the left part of the equation reads

$$(\ell_g)_*(X|_h)f = X|_h(f \circ \ell_g) = \big(X(f \circ \ell_g)\big)|_h$$

The right part can be manipulated as follows

$$X|_{gh}f = (Xf)|_{gh} = \Big((Xf) \circ \ell_g\Big)|_h$$

By substituting both final expressions back to the original one and discarding the point h since they must be true for any h we obtain the last reformulation of the push-forward

$$X(f \circ \ell_g) = X(f) \circ \ell_g.$$

Definition 5.7 ($\mathcal{L}(G)$ (As A Set))**.** *We denote the set of all left-invariant vector fields on G as $\mathcal{L}(G)$.*

Of course,

$$\mathcal{L}(G) \subseteq \Gamma(TG)$$

but, in fact, more is true. One can check that $\mathcal{L}(G)$ is closed under

$$\begin{aligned} +&\colon \mathcal{L}(G) \times \mathcal{L}(G) \to \mathcal{L}(G) \\ \cdot&\colon \mathcal{C}^\infty(G) \times \mathcal{L}(G) \to \mathcal{L}(G), \end{aligned}$$

therefore, $\mathcal{L}(G)$ is a $\mathcal{C}^\infty(G)$-submodule of $\Gamma(TG)$, but it is also an $\mathbb{R}$-vector subspace of $\Gamma(TG)$.

Recall that, up to now, we have refrained from thinking of $\Gamma(TG)$ as an $\mathbb{R}$-vector space since it is infinite-dimensional and, even worse, a basis is in general uncountable. A priori, this could be true for $\mathcal{L}(G)$ as well, but we will see that the situation is, in fact, much nicer as $\mathcal{L}(G)$ will turn out to be a finite-dimensional vector space over $\mathbb{R}$.

Theorem 5.1. *Let G be a Lie group with identity element $e \in G$. Then $\mathcal{L}(G) \cong_{\text{vec}} T_eG$.*

Proof. We will construct a linear isomorphism $j\colon T_eG \xrightarrow{\sim} \mathcal{L}(G)$. Define

$$\begin{aligned} j\colon T_eG &\to \Gamma(TG) \\ A &\mapsto j(A), \end{aligned}$$

where $j(A)$ is define as

$$\begin{aligned} j(A)\colon G &\to TG \\ g &\mapsto j(A)|_g := (\ell_g)_*(A). \end{aligned}$$

Now we have to prove that this is actually a linear isomorphism, and we will do it in steps.

i) First, we show that for any $A \in T_eG$, $j(A)$ is a smooth vector field on G. It suffices to check that for any $f \in \mathcal{C}^\infty(G)$, we have $j(A)(f) \in \mathcal{C}^\infty(G)$. Indeed

$$\begin{aligned} (j(A)(f))(g) &= j(A)|_g(f) \\ &:= (\ell_g)_*(A)(f) \\ &= A(f \circ \ell_g) \\ &= (f \circ \ell_g \circ \gamma)'(0), \end{aligned}$$

where γ is a curve through $e \in G$ whose tangent vector at e is A. The map

$$\begin{aligned}\varphi\colon \mathbb{R} \times G &\to \mathbb{R}\\ (t,g) &\mapsto \varphi(t,g) := (f \circ \ell_g \circ \gamma)(t)\\ &\qquad\qquad\;\; = f(g\gamma(t))\end{aligned}$$

is a composition of smooth maps, hence it is smooth. Then

$$(j(A)(f))(g) = (\partial_1 \varphi)(0,g)$$

depends smoothly on g and thus $j(A)(f) \in \mathcal{C}^\infty(G)$.

ii) We proved that $j(A)$ is indeed a smooth vector field, however now we need to prove that it is a left invariant vector field since it is an element of $\Gamma(TG)$. Let $g, h \in G$. Then, for every $A \in T_eG$, we have

$$\begin{aligned}(\ell_g)_*(j(A)|_h) &:= (\ell_g)_*((\ell_h)_*(A))\\ &= (\ell_{gh})_*(A)\\ &= j(A)|_{gh},\end{aligned}$$

so $j(A) \in \mathcal{L}(G)$. Hence, the map j is really $j\colon T_eG \to \mathcal{L}(G)$.

iii) We also need to check the linearity. Let $A, B \in T_eG$ and $\lambda \in \mathbb{R}$. Then, for any $g \in G$

$$\begin{aligned}j(\lambda A + B)|_g &= (\ell_g)_*(\lambda A + B)\\ &= \lambda(\ell_g)_*(A) + (\ell_g)_*(B)\\ &= \lambda j(A)|_g + j(B)|_g,\end{aligned}$$

since the push-forward is an $\mathbb{R}$-linear map. Hence, we have $j\colon T_eG \xrightarrow{\sim} \mathcal{L}(G)$.

iv) We also need to check that the map is injective. Let $A, B \in T_eG$. Then

$$\begin{aligned}j(A) = j(B) &\Leftrightarrow \forall\, g \in G : j(A)|_g = j(B)|_g\\ &\Rightarrow j(A)|_e = j(B)|_e\\ &\Leftrightarrow (\ell_e)_*(A) = (\ell_e)_*(B)\\ &\Leftrightarrow A = B,\end{aligned}$$

since $(\ell_e)_* = \mathrm{id}_{TG}$. Hence, the map j is injective.

v) Finally we nee to check that the map is surjective. Let $X \in \mathcal{L}(G)$. Define $A^X := X|_e \in T_eG$. Then, we have

$$j(A^X)|_g = (\ell_g)_*(A^X) = (\ell_g)_*(X|_e) = X_{ge} = X_g,$$

since X is left-invariant. Hence $X = j(A^X)$ and thus j is surjective.

Therefore, $j\colon T_eG \xrightarrow{\sim} \mathcal{L}(G)$ is indeed a linear isomorphism. □

Corollary 5.1. *The space $\mathcal{L}(G)$ is finite-dimensional and* $\dim \mathcal{L}(G) = \dim G$.

We will soon see that the identification of $\mathcal{L}(G)$ and T_eG goes beyond the level of linear isomorphism as vector spaces, as they are isomorphic as Lie algebras. Recall from the Lie algebra chapter in the notes, that a Lie algebra over an algebraic field K is a vector space over K equipped with a Lie bracket $[-,-]$, i.e. a K-bilinear, antisymmetric map which satisfies the Jacobi identity.

Given $X, Y \in \Gamma(TM)$, we defined their Lie bracket, or commutator, as

$$[X,Y](f) := X(Y(f)) - Y(X(f))$$

for any $f \in \mathcal{C}^\infty(M)$. You can check that indeed $[X,Y] \in \Gamma(TM)$, and that the bracket is $\mathbb{R}$-bilinear, antisymmetric and satisfies the Jacobi identity. Thus, $(\Gamma(TM), +, \cdot, [-,-])$ is an infinite-dimensional Lie algebra over $\mathbb{R}$. We suppress the $+$ and $\cdot$ when they are clear from the context. In the case of a manifold that is also a Lie group, we have the following.

Theorem 5.2. *Let G be a Lie group. Then $\mathcal{L}(G)$ is a Lie subalgebra of $\Gamma(TG)$.*

Proof. A Lie subalgebra of a Lie algebra is simply a vector subspace which is closed under the action of the Lie bracket. Therefore, we only need to check that

$$\forall\, X, Y \in \mathcal{L}(G) : \ [X, Y] \in \mathcal{L}(G).$$

Let $X, Y \in \mathcal{L}(G)$. For any $g \in G$ and $f \in \mathcal{C}^\infty(G)$, we have

$$\begin{aligned}
[X, Y](f \circ \ell_g) &:= X(Y(f \circ \ell_g)) - Y(X(f \circ \ell_g)) \\
&= X(Y(f) \circ \ell_g) - Y(X(f) \circ \ell_g) \\
&= X(Y(f)) \circ \ell_g - Y(X(f)) \circ \ell_g \\
&= \big(X(Y(f)) - Y(X(f))\big) \circ \ell_g \\
&= [X, Y](f) \circ \ell_g.
\end{aligned}$$

Hence, $[X, Y]$ is left-invariant. □

Definition 5.8 ($\mathcal{L}(G)$ (As An Algebra))**.** *Let G be a Lie group. The **associated Lie algebra** of G is $\mathcal{L}(G)$.*

Notice that we began with $\mathcal{L}(G)$ as a set of all left invariant vector fields of G, which is a subset of $\Gamma(TG)$, then we inherited the $+$ and $\cdot$ of $\Gamma(TG)$ to $\mathcal{L}(G)$ and we showed that it is also a submodule and a subvector space of $\Gamma(TG)$, and finally we inherited the Lie bracket from $\Gamma(TG)$ and we showed that it is also a subalgebra of $\Gamma(TG)$. From now on when we will be referring to $\mathcal{L}(G)$, we will mean its algebra structure.

Given the nature of $\mathcal{L}(G)$, it is a rather complicated object, since its elements are vector fields, hence we would like to work with T_eG instead, whose elements are tangent vectors. We have already shown that $\mathcal{L}(G)$ and T_eG are isomorphic as vector spaces, but we would like them to be also isomorphic as algebras. Indeed, we can use the bracket on $L(G)$ to define a bracket on T_eG such that they be isomorphic as Lie algebras. First, let us define the isomorphism of Lie algebras.

Definition 5.9 (Lie Algebra Homomorphism)**.** *Let $(L_1, [-,-]_{L_1})$ and $(L_2, [-,-]_{L_2})$ be Lie algebras over the same field. A linear map $\phi\colon L_1 \xrightarrow{\sim} L_2$ is a **Lie algebra homomorphism** if*

$$\forall\, x, y \in L_1 : \ \phi([x, y]_{L_1}) = [\phi(x), \phi(y)]_{L_2}.$$

Definition 5.10 (Lie Algebra Isomorphism)**.** *A bijective Lie algebra homomorphism, is called a **Lie algebra isomorphism** and we write $L_1 \cong_{\text{Lie alg}} L_2$.*

By using the bracket $[-,-]_{\mathcal{L}(G)}$ on $\mathcal{L}(G)$ we can define, for any $A, B \in T_eG$

$$[A, B]_{T_eG} := j^{-1}\big([j(A), j(B)]_{\mathcal{L}(G)}\big),$$

where $j^{-1}(X) = X|_e$. Equipped with these brackets, we have

$$\mathcal{L}(G) \cong_{\text{Lie alg}} T_eG.$$

Hence, given a Lie group we have seen how we can construct a Lie algebra as the space of left-invariant vector fields and this algebra is isomorphic to the algebra of tangent vectors at the identity. We will later explore the opposite direction, i.e. given a Lie algebra, we will see how to construct a Lie group whose associated Lie algebra is the one we started from.

5.2 Application - Part 2: $\mathrm{SL}(2, \mathbb{C})$

In the first part of of the application in the previous chapter, we defined the set $\mathrm{SL}(2, \mathbb{C})$ as a subset of $\mathbb{C}^4 := \mathbb{C} \times \mathbb{C} \times \mathbb{C} \times \mathbb{C}$. Then we showed that:

- $\mathrm{SL}(2, \mathbb{C})$ can be made into a group

- $\mathrm{SL}(2,\mathbb{C})$ can be made into a topological space
- $\mathrm{SL}(2,\mathbb{C})$ can be made into a topological manifold
- $\mathrm{SL}(2,\mathbb{C})$ can be made into a complex differentiable manifold

Hence we have left with $\mathrm{SL}(2,\mathbb{C})$ as a 3-dimensional, complex differentiable manifold.

$\mathrm{SL}(2,\mathbb{C})$ **As A Lie Group**

We equipped $\mathrm{SL}(2,\mathbb{C})$ with both a group and a manifold structure. In order to obtain a Lie group structure, we have to check that these two structures are compatible, that is, we have to show that the two maps

$$\begin{aligned} \mu\colon \mathrm{SL}(2,\mathbb{C}) \times \mathrm{SL}(2,\mathbb{C}) &\to \mathrm{SL}(2,\mathbb{C}) \\ (\begin{pmatrix} a & b \\ c & d \end{pmatrix}, \begin{pmatrix} e & f \\ g & h \end{pmatrix}) &\mapsto \begin{pmatrix} a & b \\ c & d \end{pmatrix} \bullet \begin{pmatrix} e & f \\ g & h \end{pmatrix} \end{aligned}$$

and

$$\begin{aligned} i\colon \mathrm{SL}(2,\mathbb{C}) &\to \mathrm{SL}(2,\mathbb{C}) \\ \begin{pmatrix} a & b \\ c & d \end{pmatrix} &\mapsto \begin{pmatrix} a & b \\ c & d \end{pmatrix}^{-1} \end{aligned}$$

are differentiable with respect to the differentiable structure on $\mathrm{SL}(2,\mathbb{C})$. For instance, for the inverse map i, we have to show that the map $y \circ i \circ x^{-1}$ is differentiable in the usual for any pair of charts $(U,x),(V,y) \in \mathscr{A}$.

$$\begin{tikzcd} U \subseteq \mathrm{SL}(2,\mathbb{C}) \arrow[r, "i"] \arrow[d, "x"] & V \subseteq \mathrm{SL}(2,\mathbb{C}) \arrow[d, "y"] \\ x(U) \subseteq \mathbb{C}^3 \arrow[r, "y \circ i \circ x^{-1}"] & y(V) \subseteq \mathbb{C}^3 \end{tikzcd}$$

However, since $\mathrm{SL}(2,\mathbb{C})$ is connected, the differentiability of the transition maps in $\mathscr{A}$ implies that if $y \circ i \circ x^{-1}$ is differentiable for any two given charts, then it is differentiable for all charts in $\mathscr{A}$. Hence, we can simply let (U,x) and (V,y) be the two charts on $\mathrm{SL}(2,\mathbb{C})$ defined above. Then, we have

$$(y \circ i \circ x^{-1})(a,b,c) = (y \circ i)(\begin{pmatrix} a & b \\ c & \frac{1+bc}{a} \end{pmatrix}) = y(\begin{pmatrix} \frac{1+bc}{a} & -b \\ -c & a \end{pmatrix}) = (\tfrac{1+bc}{a}, -b, a)$$

which is certainly complex differentiable as a map between open subsets of $\mathbb{C}^3$ (recall that $a \neq 0$ on $x(U)$).

Checking that μ is complex differentiable is slightly more involved, since we first have to equip $\mathrm{SL}(2,\mathbb{C}) \times \mathrm{SL}(2,\mathbb{C})$ with a suitable "product differentiable structure" and then proceed as above. Once that is done, we can finally conclude that $((\mathrm{SL}(2,\mathbb{C}), \mathcal{O}, \mathscr{A}), \bullet)$ is a 3-dimensional complex Lie group.

5.2.1 The Lie Algebra Of $\mathrm{SL}(2,\mathbb{C})$

Recall that to every Lie group G, there is an associated Lie algebra $\mathcal{L}(G)$, where

$$\mathcal{L}(G) := \{X \in \Gamma(TG) \mid \forall\, g,h \in G : (\ell_g)_*(X|_h) = X_{gh}\},$$

which we then proved to be isomorphic to the Lie algebra T_eG with Lie bracket

$$[A,B]_{T_eG} := j^{-1}([j(A), j(B)]_{\mathcal{L}(G)})$$

induced by the Lie bracket on $\mathcal{L}(G)$ via the isomorphism j

$$j(A)|_g := (\ell_g)_*(A).$$

In the case of $\mathrm{SL}(2,\mathbb{C})$, the left translation map by $\left(\begin{smallmatrix} a & b \\ c & d \end{smallmatrix}\right)$ is

$$\begin{aligned} \ell_{\left(\begin{smallmatrix} a & b \\ c & d \end{smallmatrix}\right)} \colon \mathrm{SL}(2,\mathbb{C}) &\to \mathrm{SL}(2,\mathbb{C}) \\ \begin{pmatrix} e & f \\ g & h \end{pmatrix} &\mapsto \begin{pmatrix} a & b \\ c & d \end{pmatrix} \bullet \begin{pmatrix} e & f \\ g & h \end{pmatrix} \end{aligned}$$

By using the standard notation $\mathfrak{sl}(2,\mathbb{C}) \equiv \mathcal{L}(\mathrm{SL}(2,\mathbb{C}))$, we have

$$\mathfrak{sl}(2,\mathbb{C}) \cong_{\text{Lie alg}} T_{\left(\begin{smallmatrix} 1 & 0 \\ 0 & 1 \end{smallmatrix}\right)} \mathrm{SL}(2,\mathbb{C}).$$

We would now like to explicitly determine the Lie bracket on $T_{\left(\begin{smallmatrix} 1 & 0 \\ 0 & 1 \end{smallmatrix}\right)} \mathrm{SL}(2,\mathbb{C})$, and hence determine its structure constants.
Recall that if (U,x) is a chart on a manifold M and $p \in U$, then the chart (U,x) induces a basis of the tangent space T_pM. We shall use our previously defined chart (U,x) on $\mathrm{SL}(2,\mathbb{C})$, where $U := \{\left(\begin{smallmatrix} a & b \\ c & d \end{smallmatrix}\right) \in \mathrm{SL}(2,\mathbb{C}) \mid a \neq 0\}$ and

$$\begin{aligned} x\colon \quad U &\to x(U) \subseteq \mathbb{C}^3 \\ \begin{pmatrix} a & b \\ c & d \end{pmatrix} &\mapsto (a,b,c). \end{aligned}$$

Note that the d appearing here is completely redundant, since the membership condition of $\mathrm{SL}(2,\mathbb{C})$ forces $d = \frac{1+bc}{a}$. However, we will keep writing the d to avoid having a fraction in a matrix in a subscript.
The chart (U,x) contains $\left(\begin{smallmatrix} 1 & 0 \\ 0 & 1 \end{smallmatrix}\right)$ and hence we get an induced co-ordinate basis

$$\left\{ \left(\frac{\partial}{\partial x^i}\right)_{\left(\begin{smallmatrix} 1 & 0 \\ 0 & 1 \end{smallmatrix}\right)} \in T_{\left(\begin{smallmatrix} 1 & 0 \\ 0 & 1 \end{smallmatrix}\right)} \mathrm{SL}(2,\mathbb{C}) \;\middle|\; 1 \leq i \leq 3 \right\}$$

so that any $A \in T_{\left(\begin{smallmatrix} 1 & 0 \\ 0 & 1 \end{smallmatrix}\right)} \mathrm{SL}(2,\mathbb{C})$ can be written as

$$A = \alpha \left(\frac{\partial}{\partial x^1}\right)_{\left(\begin{smallmatrix} 1 & 0 \\ 0 & 1 \end{smallmatrix}\right)} + \beta \left(\frac{\partial}{\partial x^2}\right)_{\left(\begin{smallmatrix} 1 & 0 \\ 0 & 1 \end{smallmatrix}\right)} + \gamma \left(\frac{\partial}{\partial x^3}\right)_{\left(\begin{smallmatrix} 1 & 0 \\ 0 & 1 \end{smallmatrix}\right)},$$

for some $\alpha, \beta, \gamma \in \mathbb{C}$. Since the Lie bracket is bilinear, its action on these basis vectors uniquely extends to the whole of $T_{\left(\begin{smallmatrix} 1 & 0 \\ 0 & 1 \end{smallmatrix}\right)} \mathrm{SL}(2,\mathbb{C})$ by linear continuation. Hence, we simply have to determine the action of the Lie bracket of $\mathfrak{sl}(2,\mathbb{C})$ on the images under the isomorphism j of these basis vectors.
Let us now determine the image of these co-ordinate induced basis elements under the isomorphism j. The object

$$j\left(\left(\frac{\partial}{\partial x^i}\right)_{\left(\begin{smallmatrix} 1 & 0 \\ 0 & 1 \end{smallmatrix}\right)}\right) \in \mathfrak{sl}(2,\mathbb{C})$$

is a left-invariant vector field on $\mathrm{SL}(2,\mathbb{C})$. It assigns to each point $\left(\begin{smallmatrix} a & b \\ c & d \end{smallmatrix}\right) \in U \subseteq \mathrm{SL}(2,\mathbb{C})$ the tangent vector

$$j\left(\left(\frac{\partial}{\partial x^i}\right)_{\left(\begin{smallmatrix} 1 & 0 \\ 0 & 1 \end{smallmatrix}\right)}\right)\bigg|_{\left(\begin{smallmatrix} a & b \\ c & d \end{smallmatrix}\right)} := \left(\ell_{\left(\begin{smallmatrix} a & b \\ c & d \end{smallmatrix}\right)}\right)_* \left(\frac{\partial}{\partial x^i}\right)_{\left(\begin{smallmatrix} 1 & 0 \\ 0 & 1 \end{smallmatrix}\right)} \in T_{\left(\begin{smallmatrix} a & b \\ c & d \end{smallmatrix}\right)} \mathrm{SL}(2,\mathbb{C}).$$

This tangent vector is a $\mathbb{C}$-linear map $\mathcal{C}^\infty(\mathrm{SL}(2,\mathbb{C})) \xrightarrow{\sim} \mathbb{C}$, where $\mathcal{C}^\infty(\mathrm{SL}(2,\mathbb{C}))$ is the $\mathbb{C}$-vector space (in fact, the $\mathbb{C}$-algebra) of smooth complex-valued functions on $\mathrm{SL}(2,\mathbb{C})$ although, to be precise, since we are working in a chart we should only consider functions defined on U. For (the restriction to U of) any $f \in \mathcal{C}^\infty(\mathrm{SL}(2,\mathbb{C}))$ we have, explicitly,

$$\begin{aligned} \left(\ell_{\left(\begin{smallmatrix} a & b \\ c & d \end{smallmatrix}\right)}\right)_* \left(\frac{\partial}{\partial x^i}\right)_{\left(\begin{smallmatrix} 1 & 0 \\ 0 & 1 \end{smallmatrix}\right)} (f) &= \left(\frac{\partial}{\partial x^i}\right)_{\left(\begin{smallmatrix} 1 & 0 \\ 0 & 1 \end{smallmatrix}\right)} \left(f \circ \ell_{\left(\begin{smallmatrix} a & b \\ c & d \end{smallmatrix}\right)}\right) \\ &= \partial_i\left(f \circ \ell_{\left(\begin{smallmatrix} a & b \\ c & d \end{smallmatrix}\right)} \circ x^{-1}\right)(x\left(\begin{smallmatrix} 1 & 0 \\ 0 & 1 \end{smallmatrix}\right)), \end{aligned}$$

where the argument of ∂_i in the last line is a map $x(U) \subseteq \mathbb{C}^3 \to \mathbb{C}$, hence ∂_i is simply the operation of complex differentiation with respect to the i-th (out of the 3) complex variable of the map $f\circ\ell_{\left(\begin{smallmatrix} a & b \\ c & d \end{smallmatrix}\right)}\circ x^{-1}$, which is then to be evaluated at $x\left(\begin{smallmatrix} 1 & 0 \\ 0 & 1 \end{smallmatrix}\right)\in\mathbb{C}^3$. By inserting an identity in the composition, we have

$$
\begin{aligned}
&= \partial_i\Big(f\circ \mathrm{id}_U\circ \ell_{\left(\begin{smallmatrix} a & b \\ c & d \end{smallmatrix}\right)}\circ x^{-1}\Big)(x\left(\begin{smallmatrix} 1 & 0 \\ 0 & 1 \end{smallmatrix}\right))\\
&= \partial_i\Big(f\circ (x^{-1}\circ x)\circ \ell_{\left(\begin{smallmatrix} a & b \\ c & d \end{smallmatrix}\right)}\circ x^{-1}\Big)(x\left(\begin{smallmatrix} 1 & 0 \\ 0 & 1 \end{smallmatrix}\right))\\
&= \partial_i\Big((f\circ x^{-1})\circ (x\circ \ell_{\left(\begin{smallmatrix} a & b \\ c & d \end{smallmatrix}\right)}\circ x^{-1})\Big)(x\left(\begin{smallmatrix} 1 & 0 \\ 0 & 1 \end{smallmatrix}\right)),
\end{aligned}
$$

where $f\circ x^{-1}\colon x(U)\subseteq\mathbb{C}^3\to\mathbb{C}$ and $(x\circ\ell_{\left(\begin{smallmatrix} a & b \\ c & d \end{smallmatrix}\right)}\circ x^{-1})\colon x(U)\subseteq\mathbb{C}^3\to x(U)\subseteq\mathbb{C}^3$ and hence, we can use the multi-dimensional chain rule to obtain

$$
= \Big(\partial_m(f\circ x^{-1})\big((x\circ\ell_{\left(\begin{smallmatrix} a & b \\ c & d \end{smallmatrix}\right)}\circ x^{-1})(x\left(\begin{smallmatrix} 1 & 0 \\ 0 & 1 \end{smallmatrix}\right))\big)\Big)\Big(\partial_i(x^m\circ\ell_{\left(\begin{smallmatrix} a & b \\ c & d \end{smallmatrix}\right)}\circ x^{-1})(x\left(\begin{smallmatrix} 1 & 0 \\ 0 & 1 \end{smallmatrix}\right))\Big),
$$

with the summation going from $m=1$ to $m=3$. The first factor is simply

$$
\begin{aligned}
\partial_m(f\circ x^{-1})\big((x\circ\ell_{\left(\begin{smallmatrix} a & b \\ c & d \end{smallmatrix}\right)})\left(\begin{smallmatrix} 1 & 0 \\ 0 & 1 \end{smallmatrix}\right)\big) &= \partial_m(f\circ x^{-1})(x\left(\begin{smallmatrix} a & b \\ c & d \end{smallmatrix}\right))\\
&=: \left(\frac{\partial}{\partial x^m}\right)_{\left(\begin{smallmatrix} a & b \\ c & d \end{smallmatrix}\right)}(f).
\end{aligned}
$$

To see what the second factor is, we first consider the map $x^m\circ\ell_{\left(\begin{smallmatrix} a & b \\ c & d \end{smallmatrix}\right)}\circ x^{-1}$. This map acts on the triple $(e,f,g)\in x(U)$ as

$$
\begin{aligned}
(x^m\circ\ell_{\left(\begin{smallmatrix} a & b \\ c & d \end{smallmatrix}\right)}\circ x^{-1})(e,f,g) &= (x^m\circ\ell_{\left(\begin{smallmatrix} a & b \\ c & d \end{smallmatrix}\right)})\begin{pmatrix} e & f \\ g & \frac{1+fg}{e}\end{pmatrix}\\
&= x^m(\begin{pmatrix} a & b \\ c & d\end{pmatrix}\bullet\begin{pmatrix} e & f \\ g & \frac{1+fg}{e}\end{pmatrix})\\
&= x^m(\begin{pmatrix} ae+bg & af+\frac{b(1+fg)}{e} \\ ce+dg & cf+\frac{d(1+fg)}{e}\end{pmatrix}),
\end{aligned}
$$

and since $x^m := \mathrm{proj}_m\circ x$, with $m\in\{1,2,3\}$, we have

$$
(x^m\circ\ell_{\left(\begin{smallmatrix} a & b \\ c & d \end{smallmatrix}\right)}\circ x^{-1})(e,f,g) = \mathrm{proj}_m(ae+bg, af+\tfrac{b(1+fg)}{e}, ce+dg),
$$

the map proj_m simply picks the m-th component of the triple. We now have to apply ∂_i to this map, with $i\in\{1,2,3\}$, i.e. we have to differentiate with respect to each of the three complex variables e, f, and g. We can write the result as

$$
\partial_i(x^m\circ\ell_{\left(\begin{smallmatrix} a & b \\ c & d \end{smallmatrix}\right)}\circ x^{-1})(e,f,g) = D(e,f,g)^m{}_i,
$$

where m labels the rows and i the columns of the matrix

$$
D(e,f,g) = \begin{pmatrix} a & 0 & b \\ -\frac{b(1+fg)}{e^2} & a+\frac{bg}{e} & \frac{bf}{e} \\ c & 0 & d \end{pmatrix}.
$$

Finally, by evaluating this at $(e,f,g)=x\left(\begin{smallmatrix} 1 & 0 \\ 0 & 1 \end{smallmatrix}\right)=(1,0,0)$, we obtain

$$
\partial_i(x^m\circ\ell_{\left(\begin{smallmatrix} a & b \\ c & d \end{smallmatrix}\right)}\circ x^{-1})(x\left(\begin{smallmatrix} 1 & 0 \\ 0 & 1 \end{smallmatrix}\right)) = D^m{}_i,
$$

where, by recalling that $d=\frac{1+bc}{a}$,

$$
D := D(1,0,0) = \begin{pmatrix} a & 0 & b \\ -b & a & 0 \\ c & 0 & \frac{1+bc}{a}\end{pmatrix}.
$$

Putting the two factors back together yields

$$\left(\ell_{\left(\begin{smallmatrix} a & b \\ c & d \end{smallmatrix}\right)}\right)_* \left(\frac{\partial}{\partial x^i}\right)_{\left(\begin{smallmatrix} 1 & 0 \\ 0 & 1 \end{smallmatrix}\right)} (f) = D^m{}_i \left(\frac{\partial}{\partial x^m}\right)_{\left(\begin{smallmatrix} a & b \\ c & d \end{smallmatrix}\right)} (f).$$

Since this holds for an arbitrary $f \in \mathcal{C}^\infty(\mathrm{SL}(2,\mathbb{C}))$, we have

$$j\left(\left(\frac{\partial}{\partial x^i}\right)_{\left(\begin{smallmatrix} 1 & 0 \\ 0 & 1 \end{smallmatrix}\right)}\right)\Bigg|_{\left(\begin{smallmatrix} a & b \\ c & d \end{smallmatrix}\right)} := \left(\ell_{\left(\begin{smallmatrix} a & b \\ c & d \end{smallmatrix}\right)}\right)_* \left(\frac{\partial}{\partial x^i}\right)_{\left(\begin{smallmatrix} 1 & 0 \\ 0 & 1 \end{smallmatrix}\right)} = D^m{}_i \left(\frac{\partial}{\partial x^m}\right)_{\left(\begin{smallmatrix} a & b \\ c & d \end{smallmatrix}\right)},$$

and since the point $\left(\begin{smallmatrix} a & b \\ c & d \end{smallmatrix}\right) \in U \subseteq \mathrm{SL}(2,\mathbb{C})$ is also arbitrary, we have

$$j\left(\left(\frac{\partial}{\partial x^i}\right)_{\left(\begin{smallmatrix} 1 & 0 \\ 0 & 1 \end{smallmatrix}\right)}\right) = D^m{}_i \frac{\partial}{\partial x^m} \in \mathfrak{sl}(2,\mathbb{C}),$$

where D is now the corresponding matrix of co-ordinate functions

$$D := \begin{pmatrix} x^1 & 0 & x^2 \\ -x^2 & x^1 & 0 \\ x^3 & 0 & \frac{1+x^2x^3}{x^1} \end{pmatrix}.$$

Note that while the three vector fields

$$\begin{aligned} \frac{\partial}{\partial x^m} \colon \mathrm{SL}(2,\mathbb{C}) &\to T\,\mathrm{SL}(2,\mathbb{C}) \\ \begin{pmatrix} a & b \\ c & d \end{pmatrix} &\mapsto \left(\frac{\partial}{\partial x^m}\right)_{\left(\begin{smallmatrix} a & b \\ c & d \end{smallmatrix}\right)} \end{aligned}$$

are not individually left-invariant, their linear combination with coefficients $D^m{}_i$ is indeed left-invariant. Recall that these vector fields

i) are $\mathbb{C}$-linear maps

$$\begin{aligned} \frac{\partial}{\partial x^m} \colon \mathcal{C}^\infty(\mathrm{SL}(2,\mathbb{C})) &\xrightarrow{\sim} \mathcal{C}^\infty(\mathrm{SL}(2,\mathbb{C})) \\ f &\mapsto \partial_m(f \circ x^{-1}) \circ x; \end{aligned}$$

ii) satisfy the Leibniz rule

$$\frac{\partial}{\partial x^m}(fg) = f\frac{\partial}{\partial x^m}(g) + g\frac{\partial}{\partial x^m}(f);$$

iii) act on the coordinate functions $x^i \in \mathcal{C}^\infty(\mathrm{SL}(2,\mathbb{C}))$ as

$$\frac{\partial}{\partial x^m}(x^i) = \partial_m(x^i \circ x^{-1}) \circ x = \partial_m(\mathrm{proj}_i \circ x \circ x^{-1}) \circ x = \delta^i_m \circ x = \delta^i_m,$$

since the composition of a constant function with any composable function is just the constant function.

Hence, we have an expansion of the images of the basis of $T_{\left(\begin{smallmatrix} 1 & 0 \\ 0 & 1 \end{smallmatrix}\right)}\mathrm{SL}(2,\mathbb{C})$ under j:

$$\begin{aligned} j\left(\left(\frac{\partial}{\partial x^1}\right)_{\left(\begin{smallmatrix} 1 & 0 \\ 0 & 1 \end{smallmatrix}\right)}\right) &= x^1\frac{\partial}{\partial x^1} - x^2\frac{\partial}{\partial x^2} + x^3\frac{\partial}{\partial x^3} \\ j\left(\left(\frac{\partial}{\partial x^2}\right)_{\left(\begin{smallmatrix} 1 & 0 \\ 0 & 1 \end{smallmatrix}\right)}\right) &= x^1\frac{\partial}{\partial x^2} \\ j\left(\left(\frac{\partial}{\partial x^3}\right)_{\left(\begin{smallmatrix} 1 & 0 \\ 0 & 1 \end{smallmatrix}\right)}\right) &= x^2\frac{\partial}{\partial x^1} + \tfrac{1+x^2x^3}{x^1}\frac{\partial}{\partial x^3}. \end{aligned}$$

We now have to calculate the bracket (in $\mathfrak{sl}(2,\mathbb{C})$) of every pair of these. We can also do them all at once, which is a good exercise in index gymnastics. We have

$$\left[j\left(\left(\frac{\partial}{\partial x^i}\right)_{\left(\begin{smallmatrix}1&0\\0&1\end{smallmatrix}\right)}\right), j\left(\left(\frac{\partial}{\partial x^k}\right)_{\left(\begin{smallmatrix}1&0\\0&1\end{smallmatrix}\right)}\right)\right] = \left[D^m{}_i\,\frac{\partial}{\partial x^m}, D^n{}_k\,\frac{\partial}{\partial x^n}\right].$$

Letting this act on an arbitrary $f \in \mathcal{C}^\infty(\mathrm{SL}(2,\mathbb{C}))$, by definition

$$\left[D^m{}_i\,\frac{\partial}{\partial x^m}, D^n{}_k\,\frac{\partial}{\partial x^n}\right](f) := D^m{}_i\,\frac{\partial}{\partial x^m}\Big(D^n{}_k\,\frac{\partial}{\partial x^n}(f)\Big) - D^n{}_k\,\frac{\partial}{\partial x^n}\Big(D^m{}_i\,\frac{\partial}{\partial x^m}(f)\Big).$$

The first term gives

$$\begin{aligned}
D^m{}_i\,\frac{\partial}{\partial x^m}\Big(D^n{}_k\,\frac{\partial}{\partial x^n}(f)\Big) &= D^m{}_i\,\frac{\partial}{\partial x^m}(D^n{}_k\,\partial_n(f\circ x^{-1})\circ x)\\
&= D^m{}_i\,\frac{\partial}{\partial x^m}(D^n{}_k)\,(\partial_n(f\circ x^{-1})\circ x) + D^m{}_i D^n{}_k\,\frac{\partial}{\partial x^m}(\partial_n(f\circ x^{-1})\circ x)\\
&= D^m{}_i\,\frac{\partial}{\partial x^m}(D^n{}_k)\,(\partial_n(f\circ x^{-1})\circ x) + D^m{}_i D^n{}_k\,\partial_m(\partial_n(f\circ x^{-1})\circ x\circ x^{-1})\circ x\\
&= D^m{}_i\,\frac{\partial}{\partial x^m}(D^n{}_k)\,(\partial_n(f\circ x^{-1})\circ x) + D^m{}_i D^n{}_k\,\partial_m\partial_n(f\circ x^{-1})\circ x.
\end{aligned}$$

Similarly, we have

$$D^n{}_k\,\frac{\partial}{\partial x^n}\Big(D^m{}_i\,\frac{\partial}{\partial x^m}(f)\Big) = D^n{}_k\,\frac{\partial}{\partial x^n}(D^m{}_i)\,(\partial_m(f\circ x^{-1})\circ x) + D^n{}_k D^m{}_i\,\partial_n\partial_m(f\circ x^{-1})\circ x.$$

Hence, recalling that $\partial_m\partial_n = \partial_n\partial_m$ by Schwarz's theorem, we have

$$\begin{aligned}
\left[D^m{}_i\,\frac{\partial}{\partial x^m}, D^n{}_k\,\frac{\partial}{\partial x^n}\right](f) &= D^m{}_i\,\frac{\partial}{\partial x^m}(D^n{}_k)\,(\partial_n(f\circ x^{-1})\circ x) + [gray]D^m{}_i D^n{}_k\,\partial_m\partial_n(f\circ x^{-1})\circ x\\
&\qquad - D^n{}_k\,\frac{\partial}{\partial x^n}(D^m{}_i)\,(\partial_m(f\circ x^{-1})\circ x) - [gray]D^n{}_k D^m{}_i\,\partial_n\partial_m(f\circ x^{-1})\circ x\\
&= \Big(D^m{}_i\,\frac{\partial}{\partial x^m}(D^n{}_k) - D^m{}_k\,\frac{\partial}{\partial x^m}(D^n{}_i)\Big)\partial_n(f\circ x^{-1})\circ x\\
&= \Big(D^m{}_i\,\frac{\partial}{\partial x^m}(D^n{}_k) - D^m{}_k\,\frac{\partial}{\partial x^m}(D^n{}_i)\Big)\frac{\partial}{\partial x^n}(f),
\end{aligned}$$

where we relabelled some dummy indices. Since the $f \in \mathcal{C}^\infty(\mathrm{SL}(2,\mathbb{C}))$ was arbitrary,

$$\left[D^m{}_i\,\frac{\partial}{\partial x^m}, D^n{}_k\,\frac{\partial}{\partial x^n}\right] = \Big(D^m{}_i\,\frac{\partial}{\partial x^m}(D^n{}_k) - D^m{}_k\,\frac{\partial}{\partial x^m}(D^n{}_i)\Big)\frac{\partial}{\partial x^n}.$$

We can now evaluate this explicitly. For $i = 1$ and $k = 2$, we have

$$\begin{aligned}
\left[D^m{}_1\frac{\partial}{\partial x^m}, D^n{}_2\frac{\partial}{\partial x^n}\right] &= \Big([gray]D^m{}_1\frac{\partial}{\partial x^m}(D^1{}_2) - D^m{}_2\frac{\partial}{\partial x^m}(D^1{}_1)\Big)\frac{\partial}{\partial x^1}\\
&\quad + \Big(D^m{}_1\frac{\partial}{\partial x^m}(D^2{}_2) - D^m{}_2\frac{\partial}{\partial x^m}(D^2{}_1)\Big)\frac{\partial}{\partial x^2}\\
&\quad + \Big([gray]D^m{}_1\frac{\partial}{\partial x^m}(D^3{}_2) - D^m{}_2\frac{\partial}{\partial x^m}(D^3{}_1)\Big)\frac{\partial}{\partial x^3}\\
&= -D^1{}_2\frac{\partial}{\partial x^1} + (D^1{}_1 + D^2{}_2)\frac{\partial}{\partial x^2} - D^3{}_2\frac{\partial}{\partial x^3}\\
&= 2x^1\frac{\partial}{\partial x^2}.
\end{aligned}$$

Similarly, we compute

$$
\begin{aligned}
\left[D^m{}_1\frac{\partial}{\partial x^m}, D^n{}_3\frac{\partial}{\partial x^n}\right] &= \Big(D^m{}_1\frac{\partial}{\partial x^m}(D^1{}_3) - D^m{}_3\frac{\partial}{\partial x^m}(D^1{}_1)\Big)\frac{\partial}{\partial x^1} \\
&\quad + \Big([gray]D^m{}_1\frac{\partial}{\partial x^m}(D^2{}_3) - D^m{}_3\frac{\partial}{\partial x^m}(D^2{}_1)\Big)\frac{\partial}{\partial x^2} \\
&\quad + \Big(D^m{}_1\frac{\partial}{\partial x^m}(D^3{}_3) - D^m{}_3\frac{\partial}{\partial x^m}(D^3{}_1)\Big)\frac{\partial}{\partial x^3} \\
&= -2x^2\frac{\partial}{\partial x^1} - 2(\tfrac{1+x^2x^3}{x^1})\frac{\partial}{\partial x^3}
\end{aligned}
$$

and

$$
\begin{aligned}
\left[D^m{}_2\frac{\partial}{\partial x^m}, D^n{}_3\frac{\partial}{\partial x^n}\right] &= \Big(D^m{}_2\frac{\partial}{\partial x^m}(D^1{}_3) - [gray]D^m{}_3\frac{\partial}{\partial x^m}(D^1{}_2)\Big)\frac{\partial}{\partial x^1} \\
&\quad + \Big([gray]D^m{}_2\frac{\partial}{\partial x^m}(D^2{}_3) - D^m{}_3\frac{\partial}{\partial x^m}(D^2{}_2)\Big)\frac{\partial}{\partial x^2} \\
&\quad + \Big(D^m{}_2\frac{\partial}{\partial x^m}(D^3{}_3) - [gray]D^m{}_3\frac{\partial}{\partial x^m}(D^3{}_2)\Big)\frac{\partial}{\partial x^3} \\
&= (D^2{}_1 - D^1{}_3)\frac{\partial}{\partial x^1} + D^2{}_3\frac{\partial}{\partial x^2} - D^3{}_2\frac{\partial}{\partial x^3} \\
&= x^1\frac{\partial}{\partial x^1} - x^2\frac{\partial}{\partial x^2} + x^3\frac{\partial}{\partial x^3},
\end{aligned}
$$

where the differentiation rules that we have used come from the definition of the vector field $\frac{\partial}{\partial x^m}$, the Leibniz rule, and the action on co-ordinate functions.
By applying j^{-1}, which is just evaluation at the identity, to these vector fields, we finally see that the induced Lie bracket on $T_{\left(\begin{smallmatrix}1&0\\0&1\end{smallmatrix}\right)}\mathrm{SL}(2,\mathbb{C})$ satisfies

$$
\begin{aligned}
\left[\left(\frac{\partial}{\partial x^1}\right)_{\left(\begin{smallmatrix}1&0\\0&1\end{smallmatrix}\right)}, \left(\frac{\partial}{\partial x^2}\right)_{\left(\begin{smallmatrix}1&0\\0&1\end{smallmatrix}\right)}\right] &= 2\left(\frac{\partial}{\partial x^2}\right)_{\left(\begin{smallmatrix}1&0\\0&1\end{smallmatrix}\right)} \\
\left[\left(\frac{\partial}{\partial x^1}\right)_{\left(\begin{smallmatrix}1&0\\0&1\end{smallmatrix}\right)}, \left(\frac{\partial}{\partial x^3}\right)_{\left(\begin{smallmatrix}1&0\\0&1\end{smallmatrix}\right)}\right] &= -2\left(\frac{\partial}{\partial x^3}\right)_{\left(\begin{smallmatrix}1&0\\0&1\end{smallmatrix}\right)} \\
\left[\left(\frac{\partial}{\partial x^2}\right)_{\left(\begin{smallmatrix}1&0\\0&1\end{smallmatrix}\right)}, \left(\frac{\partial}{\partial x^3}\right)_{\left(\begin{smallmatrix}1&0\\0&1\end{smallmatrix}\right)}\right] &= \left(\frac{\partial}{\partial x^1}\right)_{\left(\begin{smallmatrix}1&0\\0&1\end{smallmatrix}\right)}.
\end{aligned}
$$

Hence, the structure constants of $T_{\left(\begin{smallmatrix}1&0\\0&1\end{smallmatrix}\right)}\mathrm{SL}(2,\mathbb{C})$ with respect to the co-ordinate basis are

$$
C^2{}_{12} = 2, \qquad C^3{}_{13} = -2, \qquad C^1{}_{23} = 1,
$$

with all other being either zero or related to these by anti-symmetry.

Part II

Statistics & Probability Theory

Chapter 6

Basic Concepts

6.1 Introduction

Probability theory is the branch of mathematics concerned with probability. Although there are several different probability interpretations, probability theory treats the concept in a rigorous mathematical manner by expressing it through a set of axioms. Typically these axioms formalise probability in terms of a probability space, which assigns a measure taking values between 0 and 1, termed the probability measure, to a set of outcomes called the sample space. Any specified subset of these outcomes is called an event.

Central subjects in probability theory include discrete and continuous random variables, probability distributions, and stochastic processes, which provide mathematical abstractions of non-deterministic or uncertain processes or measured quantities that may either be single occurrences or evolve over time in a random fashion.

Although it is not possible to perfectly predict random events, much can be said about their behaviour. Two major results in probability theory describing such behaviour are the law of large numbers and the central limit theorem.
As a mathematical foundation for statistics, probability theory is essential to many human activities that involve quantitative analysis of data. Methods of probability theory also apply to descriptions of complex systems given only partial knowledge of their state, as in statistical mechanics. A great discovery of twentieth-century physics was the probabilistic nature of physical phenomena at atomic scales, described in quantum mechanics.

6.1.1 Basic Terminology

In this section we will provide some basic and heavily used terminology in probability theory and statistics that we will be using through this part.

Definition 6.1 (Data)**.** ***Data*** *are individual units of information that have been collected.*

Based on the nature of the data, we have two fundamental distinctions: qualitative and quantitative data.

Definition 6.2 (Qualitative/Categorical Data)**.** ***Qualitative (or categorical) data*** *are non - numerical data, on which mathematical operations are meaningless.*

Definition 6.3 (Quantitative Data)**.** ***Quantitative data*** *are numerical data, on which mathematical operations are meaningful.*

More specifically, quantitative data can be divided into two categories: discrete and continuous data.

Definition 6.4 (Discrete Data)**.** ***Discrete data*** *are finite and countable data.*

Definition 6.5 (Continuous Data)**.** ***Continuous data*** *are infinite and uncountable data.*

Regarding the scale that data are measured on we have different levels of levels of measurement.

Definition 6.6 (Levels/Scales Of Measurement)**.** ***Level of measurement or scale of measure*** *is a classification that describes the nature of data within the values assigned to variables.*

Levels of measurement consist of four levels, or scales: nominal, ordinal, interval, and ratio.

Definition 6.7 (Nominal)**.** ***Nominal*** *level differentiates between items or subjects based only on their names or other qualitative classifications they belong to. No ranking or mathematical operation have meaning.*

Examples of nominal scaled data are gender, nationality, ethnicity, language, genre, style, biological species, etc.

Definition 6.8 (Ordinal)**.** ***Ordinal*** *level allows for rank order (1st, 2nd, 3rd, etc.) by which data can be sorted, but still does not allow for relative degree of difference between them. No mathematical operation have meaning.*

Examples of ordinal scaled data include data as "sick" vs "healthy" when measuring health, "guilty" vs. "not-guilty" when making judgments in courts, or clothing size: Small, Medium, Large, Extra Large etc.

Definition 6.9 (Interval)**.** ***Interval*** *level allows for the degree of difference between items, but not the ratio between them. Both ranking and some mathematical operations are valid but there is no meaningful zero.*

An example of interval scaled data is temperature with the Celsius scale, which has two defined points (the freezing and boiling point of water at specific conditions) and then separated into 100 intervals. Ratios are not meaningful since 20 °C cannot be said to be "twice as hot" as 10 °C, nor can multiplication/division be carried out between any two dates directly.

Definition 6.10 (Ratio)**.** ***Ratio*** *level takes its name from the fact that measurement is the estimation of the ratio between a magnitude of a continuous quantity and a unit magnitude of the same kind. A ratio scale possesses a meaningful (unique and non-arbitrary) zero value.*

Examples of ratio scaled data include mass, length, duration, plane angle, energy and electric charge. In contrast to interval scales, ratios are now meaningful because having a non-arbitrary zero point makes it meaningful to say, for example, that one object has "twice the length".

Now that we have given the basic definitions of data, let's move on defining the science of studying data.

Definition 6.11 (Statistics)**.** ***Statistics*** *is the science of collecting, analysing, summarizing, interpreting, and drawing conclusion out of data.*

The general definition of statistics can be split into two parts: descriptive and inferential statistics.

Definition 6.12 (Descriptive Statistics)**.** ***Descriptive statistics*** *is the process that quantitatively describes or summarizes features of a collection of data.*

Definition 6.13 (Inferential Statistics)**.** ***Inferential statistics*** *is the process of using data analysis to deduce properties of an underlying probability distribution.*

Now let's start developing the theory of probability.

6.2 Sample Space & Events

Definition 6.14 (Random Experiment)**.** *A* ***random experiment*** *is an experiment, trial, or observation with the following properties:*

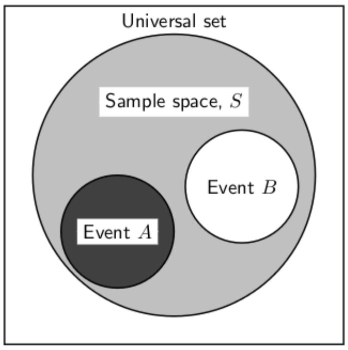
- *It can be repeated numerous times under the same conditions.*
- *The experiment can have more than one outcome.*
- *Each possible outcome can be specified in advance.*
- *The outcome of the experiment depends on chance.*

Definition 6.15 (Outcome)**.** *An **outcome** is a possible result of a random experiment or trial. Each possible outcome of a particular experiment is unique, and different outcomes are mutually exclusive (only one outcome will occur on each trial of the experiment).*

Given a random experiment and its possible outcomes, we can define the concept of a sample space and an event as follows.

Definition 6.16 (Sample Space)**.** ***Sample space*** *S of a random experiment or random trial is the set of all possible outcomes or results of that experiment.*

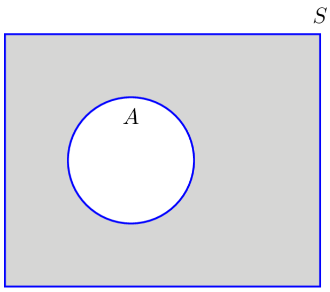
Definition 6.17 (Event)**.** *An **event** A is a subset of the sample space.*



Now that we have attached a set representation to events, we can use the usual set theory to define basic operations on events.

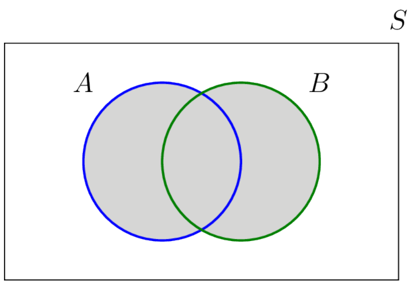
Definition 6.18 (Complement Event)**.** *The **complement** of an event A denoted by A^C is the set of elements not in A, within the sample space S.*

$$A^C = \{x : x \in S \mid x \notin A\}$$



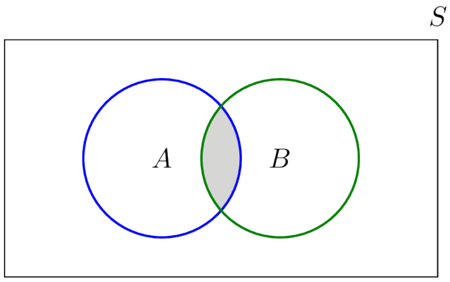
Definition 6.19 (Union)**.** *The **union** of two events A and B is the event containing elements which are in A, in B, or in both A and B.*

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



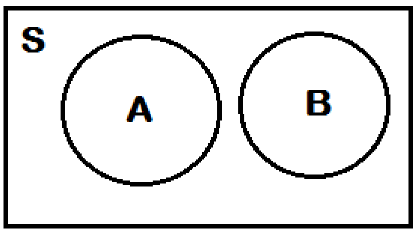
Definition 6.20 (Intersection)**.** *The **intersection** of two events A and B, is the event containing all elements of A that also belong to B (or equivalently, all elements of B that also belong to A).*

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



Definition 6.21 (Mutually Exclusive Events)**.** *Two events A and B are called **mutually exclusive** if the intersection of the events is equal to the empty set.*

$$A \cap B = \emptyset$$



Based on the set nature of events we can give a first naive definition of probability as follows.

Definition 6.22 (Naive Probability)**.** ***Naive probability** of an event A is defined as the fraction of favorable outcomes over all possible outcomes.*

$$P(A) = \frac{\text{number of favorable outcomes}}{\text{number of total outcomes}}$$

The naive definition of probability assumes that all favorable events are equally likely to be picked and that we are dealing with a finite sample space.

6.3 Probability Space

Both assumptions of the definition of naive probability add some limitations to the theory, hence we need to give a more formal and mathematical definition of probability. In order to do show we need to give some more definitions on top of sample space and events in order to be able to combine everything into the notion of probability space.

Definition 6.23 (σ-algebra)**.** *A **σ-algebra** $\mathcal{F}$ on a sample space S is a collection of subsets of S that includes S itself, is closed under complement, and is closed under countable unions.*

For example if $S = \{a, b, c, d\}$ is a sample space, one possible σ-algebra on sample space S is $\mathcal{F} = \{\emptyset, \{a, b\}, \{c, d\}, \{a, b, c, d\}\}$, where $\emptyset$ is the empty set.

Definition 6.24 (Borel Space)**.** *Given a sample space S and a σ-algebra $\mathcal{F}$ on the sample space S, we define the **Borel space** as the tuple $(S, \mathcal{F})$.*

Definition 6.25 (Probability Measure/Distribution)**.** *A **probability measure** (or probability distribution) P on a Borel space $(S, \mathcal{F})$ is a real-valued function that maps elements of $\mathcal{F}$ to the real numbers and satisfies the following axioms:*

1. *The probability of an event A is a non-negative real number:*

$$P(A) \geq 0 \qquad \forall A \in S$$

2. *The probability that at least one of the events in the entire sample space will occur is 1:*

$$P(S) = 1$$

3. *Any countable sequence of mutually exclusive events* $(A_1, A_2, \ldots)$ *satisfies:*

$$P\left(\bigcup_{i=1}^{\infty} A_i\right) = \sum_{i=1}^{\infty} P(A_i)$$

This is the formal definition of probability, free of the constraints of naive probability. Finally, we have all the ingredients to define the concept of a probability space.

Definition 6.26 (Probability Space)**.** *Given a sample space S, a σ-algebra $\mathcal{F}$ and a probability distribution P on the sample space S, we define the* ***probability space*** *as the tuple $(S, \mathcal{F}, P)$.*

A probability space models a real-world process consisting of states that occur randomly. Subsequently, an outcome is the result of a single execution of the model. Since individual outcomes might be of little practical use, more complex events are used to characterize groups of outcomes. The collection of all such events is a σ-algebra $\mathcal{F}$. Finally, probability measure P specifies each event's likelihood of happening.

Using the definition of probability space and the three axioms we can prove various relation between probabilities of events.

Lemma 6.1. $P(A^C) = 1 - P(A)$

Proof. Since the union any event A with its complement A^C gives back the whole sample space, it is:

$$\begin{aligned}
&S = A \cup A^C \Rightarrow \\
&P(S) = P(A \cup A^C) \Rightarrow \\
&P(S) = P(A) + P(A^C) \Rightarrow \\
&P(A^C) = P(S) - P(A) \Rightarrow \\
&P(A^C) = 1 - P(A)
\end{aligned}$$

□

Lemma 6.2. $P(\emptyset) = 0$

Proof. Since $S \cup \emptyset = S$ we can set $A = S$ and $A^C = \emptyset$ in lemma (6.1) and we obtain:

$$\begin{aligned}
&P(A^C) = 1 - P(A) \Rightarrow \\
&P(\emptyset) = 1 - P(S) \Rightarrow \\
&P(\emptyset) = 1 - 1 \Rightarrow \\
&P(\emptyset) = 0
\end{aligned}$$

□

Lemma 6.3. $P(A \cup B) = P(A) + P(B) - P(A \cap B)$

Proof. For any events A and B, we have the disjoint union:

$$\begin{aligned} A \cup B &= (A - B) \cup (A \cap B) \cup (B - A) \Rightarrow \\ P(A \cup B) &= P((A - B) \cup (A \cap B) \cup (B - A)) \\ &= P(A - B) + P(A \cap B) + P(B - A) \\ &= P(A) - P(A \cap B) + P(A \cap B) + P(B) - P(A \cap B) \\ &= P(A) + P(B) - P(A \cap B) \end{aligned}$$

□

6.4 Conditional Probability

Definition 6.27 (Independent Events). *Two events A and B are called* ***independent*** *if and only if their joint probability equals the product of their probabilities.*

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

Subsequently for the union of two independent events by using the lemma we proved previously:

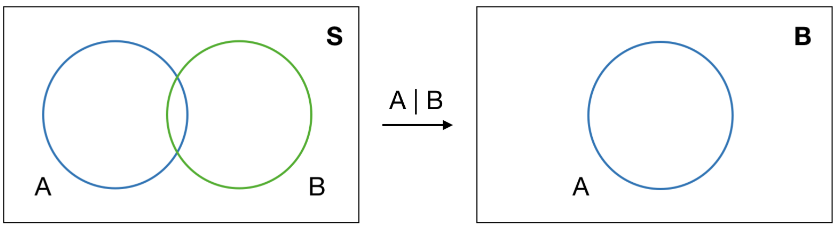
$$P(A \cup B) = P(A) + P(B) - P(A)P(B)$$

We can now move on, on defining conditional probability.

Definition 6.28 (Conditional Probability). *Given two events A and B, the* ***conditional probability*** *of A given B is defined as the quotient of the probability of the joint of events A and B, and the probability of B.*

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

This may be visualized as restricting the sample space to situations in which B occurs.



Given conditional probability, similarly to independent events we can also define conditionally independent events as follows.

Definition 6.29 (Conditionally Independent Events). *Two events A and B are called* ***conditionally independent*** *if and only if. given an event C, their joint conditional probability equals the product of their conditional probabilities.*

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

Conditional probability is very important in probability theory and its applications since based on the definition we can prove some very useful theorems that we will be using throughout the notes.

Theorem 6.1 (**Multiplication Rule**).

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

Proof. Straight forward by multiplying by $P(B)$ both sides the definition of conditional probability. $\square$

Theorem 6.2 (**Bayes Rule**).

$$P(A \mid B) = \frac{P(B \mid A)P(A)}{P(B)}$$

Proof. From multiplication rule by interchanging A with B we get:

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

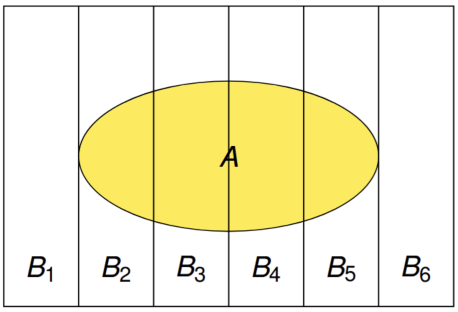
But since $P(A \cap B) = P(B \cap A)$ we end up having:

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

By solving with respect to $P(A \mid B)$ we get Bayes rule. $\square$

Theorem 6.3 (**Law Of Total Probability**). *Given a finite or countably infinite partition of a sample space S, $\{B_n : n = 1, 2, 3, \ldots\}$ (in other words, a set of pairwise disjoint events whose union is the entire sample space) then for any event A of the same probability space:*

$$P(A) = \sum_n P(A \mid B_n)P(B_n)$$



Proof. From the partition follows:

$$\begin{aligned}
A &= (A \cap B_1) \cup (A \cap B_2) \cap \ldots \cap (A \cap B_n) \Rightarrow \\
P(A) &= P((A \cap B_1) \cup (A \cap B_2) \cap \ldots \cap (A \cap B_n)) \\
&= P(A \cap B_1) + P(A \cap B_2) + \ldots + P(A \cap B_n) \\
&= P(A \mid B_1)P(B_1) + P(A \mid B_2)P(B_2) + \ldots + P(A \mid B_n)P(B_n) \\
&= \sum_n P(A \mid B_n)P(B_n)
\end{aligned}$$

$\square$

Chapter 7

Random Variables

7.1 Random Variables

Definition 7.1 (Random Variable)**.** *Given a probability space* $(S, \mathcal{F}, P)$*, we define as a* ***random variable*** *(r.v)* X*, a measurable function* X *that maps elements of sample space* S *to the real numbers* R*.*

$$X : S \to R$$

Intuitively a r.v is a numerical representation of the outcomes of a random experiment. For example if the random experiment is tossing a coin, we can map the outcomes to a r.v X that takes two possible values: 0 for "head" and 1 for "tail". That way we mapped the outcomes of the random experiment to something that we can work with!

Since r.v's are used to model a random experiment, following from the definition of the latest the actual value of a r.v is not known before the execution of the experiment however the spectrum of possible outcomes is known.

Based on the nature of the data that define the sample space, r.v's can be either discrete or continuous. In general the two cases behave similarly up to a point, but there are also some crucial differences. For this reason we are going to see each case separately.

7.2 Discrete Random Variables

Discrete r.v's are r.v's that can only take discrete, countable values (as for example tossing a coin or throwing a dice). As we mentioned, the actual values of a r.v is not known to us before the execution of the experiments however the spectrum of all possible outcomes is known. On top of that, we can define a quantity that is connected to the probability of obtaining each of the possible outcomes after each experiment. In the case of discrete r.v's, this quantity is called "probability mass function".

Definition 7.2 (Probability Mass Function)**.** *Given a discrete r.v* X *defined on a sample space* S *as* $X : S \to R$*, we define the* ***probability mass function*** *(PMF)* $P_X(x)$ *as a function that maps outcomes* R *to the interval* $[0, 1]$ *(*$P_X : R \mapsto [0, 1]$*):*

$$P_X(X = x) = P_X(\{s \in S : X(s) = x\})$$

with:

$$\sum_x P_X(x) = 1$$

The physical meaning of a PMF is the probability that a r.v X will take the value x after the execution of a random experiment ($P_X(X = x)$). The term "mass" helps to get the intuition since the physical mass is conserved as is the total probability for all hypothetical outcomes x.

From now on we keep in mind that every r.v X carries a corresponding PMF $P_X(x)$. The common terminology is that a r.v X follows a probability distribution P_X denoted by $X \sim P_X$ meaning that it's described by the corresponding PMF. Once we have the r.v and its PMF, we can define some really important concepts of r.v's.

Definition 7.3 (Expected Value/Mean)**.** *Let X be a discrete r.v with a finite number of finite outcomes and $P_X(x)$ its corresponding PMF. The* ***expected value*** *(or* ***mean****) of X denoted by $E[X]$ or μ is defined as:*

$$E[X] = \sum_x x P_X(X = x)$$

Given that the sum of probabilities of all possible outcomes is 1, the expected value is actually the weighted average, with probability of each outcome being the weight. Notice that the expected value of a r.v is a single number. The physical meaning of this number is the hypothetical final outcome that one would have after repeating the experiment infinite times.

The expected value has some very interesting properties.

1. If $X = c$, $c \in R$ then $E[X] = c$.
2. Since $E[X]$ is a single number it follows that $E[E[X]] = E[X]$.
3. If $X = Y$ then $E[X] = E[Y]$.
4. Linearity of expected value:
   - $E[X + Y] = E[X] + E[Y]$
   - $E[cX] = cE[X]$

Similarly to the expected value we can define the conditional expected value.

Definition 7.4 (Conditional Expected Value)**.** *Let X be a discrete r.v with a finite number of finite outcomes and $P_X(x)$ its corresponding PMF. The* ***conditional expected value*** *of X given an event s denoted by $E[X \mid s]$ is defined as:*

$$E[X \mid s] = \sum_x x P_X(X \mid s)$$

Notice that given the conditional expected value we can, in a way, derive the formula for the expected value as follows:

$$\begin{aligned} E[X] &= \sum_s E[X \mid s] P_X(s) \\ &= \sum_s \Big(\sum_x x P_X(X \mid s) \Big) P_X(s) \\ &= \sum_x x \Big(\sum_s P_X(X \mid s) P_X(s) \Big) \\ &= \sum_x x P_X(x) \end{aligned}$$

Definition 7.5 (Variance)**.** *Let X be a discrete r.v with a finite number of finite outcomes and $P_X(x)$ its corresponding PMF. The* ***variance*** *of X denoted by $Var(X)$ or σ^2 is defined as the expected value of the squared deviation from the expected value of the r.v.*

$$Var[X] = E[(X - E[X])^2]$$

The variance shows how far the possible outcomes of a random variable are spread out from their expected value. The highest the variance the widest the spread and vice versa.

Notice that the variance is **not** linear, hence $Var(X + Y) \neq Var(X) + Var(Y)$. However in the case where X and Y are independent the equality holds.

While the definition of the expected value is also useful for computation purposes, the definition of the variance is not that handy because of the square term. Likely, we can manipulate the definition off variance and get something more useful for computations.

$$\begin{aligned}
Var(X) &= E[(X - E[X])^2] \\
&= E[X^2 - 2XE[X] + E[X]^2] \\
&= E[X^2] - 2E[X]E[X] + E[X]^2 \\
&= E[X^2] - E[X]^2
\end{aligned}$$

In other words, the variance of X is equal to the expected value of the square of X minus the square of the expected value of X. We will be using this equation a lot for derivations.

As we did before, similarly to the variance we can define the conditional variance.

Definition 7.6 (Conditional Variance)**.** *Let X be a discrete r.v with a finite number of finite outcomes and $P_X(x)$ its corresponding PMF. The **conditional variance** of X given an even s denoted by $Var(X \mid s)$ is defined as:*

$$Var[X \mid s] = E[(X - E[X \mid s])^2 \mid s]$$

One of the main problems of variance (and conditional variance) is that it doesn't have the same units as the r.v or the expected value, but is has the square of this unit. Sometimes it makes it difficult to appreciate the meaning of variance in absolute terms. For that reason we define a more handy measure of dispersion called "standard deviation".

Definition 7.7 (Standard Deviation)**.** *Let X be a discrete r.v with a finite number of finite outcomes and $P_X(x)$ its corresponding PMF. The **standard deviation** of X denoted by $SD(X)$ or σ is equal to the square root of the variance of X.*

$$SD(X) = \sqrt{Var(X)} = \sqrt{E[(X - E[X])^2]}$$

A low standard deviation indicates that the values tend to be close to the expected value of the r.v, while a high standard deviation indicates that the values are spread out over a wider range.

7.3 Discrete Probability Distributions

In general the only requirement for a function $P_X(x)$ to be the PMF of some discrete r.v X is (**??**). Once a function satisfies this requirement then it describes the probability distribution of some discrete r.v X. In this section we are gonna introduce some of the most fundamental discrete probability distributions among with their characteristics and their intuition. This section is really important since by making use of these distributions we can create models for real world applications.

7.3.1 Discrete Uniform Distribution - Unif(n)

The discrete uniform distribution parametrized by n and denoted by Unif(n) is a discrete probability distribution whereby a finite number of values n (possible outcomes of r.v X) are equally likely to be observed (every one of n values has equal probability $\frac{1}{n}$). Another way to parametrize discrete uniform distributions is by listing all n possible values as $\{a, a+1, \ldots, b-1, b\}$. Then we use the parameters a and b and we formally write Unif(a,b). The corresponding PMF is as simple as that:

$$P_X(X = x) = \frac{1}{n}$$

Another way of describing the discrete uniform distribution would be "a known, finite number of outcomes equally likely to happen". A simple example of the discrete uniform distribution is throwing a fair die. The possible values are $\{1, 2, 3, 4, 5, 6\}$ and each time the dice is thrown the probability of a given score is 1/6. If two dice are thrown and their values added, the resulting distribution is no longer uniform since not all sums have equal probability.

For the expected value of a discrete uniform distribution, straight from the definition we get:

$$E[X] = \sum_{k=1}^{n} k(\frac{1}{n}) = \frac{1}{n}\sum_{k=1}^{n} k = \frac{1}{n}\frac{n(n+1)}{2} = \frac{n+1}{2}$$

For the variance of a discrete uniform distribution we will use the relation we proved so first we calculate $E[X^2]$:

$$E[X^2] = \sum_{k=1}^{n} k^2(\frac{1}{n}) = \frac{1}{n}\sum_{k=1}^{n} k^2 = \frac{1}{n}\frac{n(n+1)(2n+1)}{6} = \frac{(n+1)(2n+1)}{6}$$

Substituting $E[X^2]$ and $E[X]^2$ to the variance relation we get:

$$Var(X) = E[X^2] - E[X]^2 = \frac{(n+1)(2n+1)}{6} - \frac{(n+1)^2}{4} = \ldots = \frac{n^2-1}{12}$$

Finally for the standard deviation of a discrete uniform distribution:

$$SD(X) = \sqrt{Var(X)} = \sqrt{\frac{n^2-1}{12}}$$

7.3.2 Bernoulli Distribution - Bern(p)

The Bernoulli distribution parametrized by p and denoted by Bern(p) is a discrete probability distribution having two possible outcomes labelled by $x = 1$ (called "success") that occurs with probability p ($0 < p < 1$) and $x = 0$ (called "failure") that occurs with probability $q = 1 - p$. It therefore has PMF:

$$P_X(X = x) = \begin{cases} 1-p & \text{if } x = 0 \\ p & \text{if } x = 1 \end{cases}$$

Observe that in case where $x = 0$ it is $p^0 = 1$ and $(1-p)^{(1-0)} = (1-p)$ and in case where $x = 1$ it is $p^1 = p$ and $(1-p)^{(1-1)} = 0$. By using this observation we can formally rewrite Bernoulli's distribution PMF in just one line :

$$P_X(X = x) = p^x(1-p)^{1-x}$$

which gives the same results as the previous definition.

Bernoulli distribution can be used to model any single experiment that asks a yes–no question. The question results in a Boolean - valued outcome, with probability of success p and probability of failure q. For example, it can be used to represent a coin toss where 1 and 0 would represent "heads" and "tails" (or vice versa), respectively, and p would be the probability of the coin landing on heads or tails, respectively. (In a fair coin case we would have $p = q = 1/2$).

Bernoulli distribution is the simplest discrete distribution, and it the building block for other more complicated discrete distributions.

For the expected value of a Bernoulli distribution, straight from the definition we get:

$$E[X] = \sum_{x} xP_X(X = x) = 0 \cdot P_X(X = 0) + 1 \cdot P_X(X = 1) = 0 \cdot (1-p) + 1 \cdot p = p$$

For the variance of a Bernoulli distribution we will use the relation we proved so first we calculate $E[X^2]$:

$$E[X^2] = \sum_{x} x^2P_X(X = x) = 0^2 \cdot P_X(X = 0) + 1^2 \cdot P_X(X = 1) = 0^2 \cdot (1-p) + 1^2 \cdot p = p$$

Substituting $E[X^2]$ and $E[X]^2$ to the variance relation we get:

$$Var(X) = E[X^2] - E[X]^2 = p - p^2 = p(1-p)$$

Finally for the standard deviation of a Bernoulli distribution:

$$SD(X) = \sqrt{Var(X)} = \sqrt{p(1-p)}$$

7.3.3 Binomial Distribution - B(n,p)

The binomial distribution parametrized by n and p and denoted by B(n,p) is a discrete probability distribution of the number of successes in a sequence of n independent experiments, each asking a yes – no question, and each with its own Boolean - valued outcome: success with probability p or failure with probability $q = 1 - p$.

In other words is a sequence of n experiments following a Bernoulli distribution, thus the parametrization by n and p. Since the only possible outcomes of a Bernoulli distribution is either 0 or 1, we can formally think of a binomial distribution r.v Y as the sum of n Bernoulli distribution X_i: $Y = X_1 + X_2 + \ldots + X_n$.

The corresponding PMF reads:

$$P_X(X = k) = \binom{n}{k} p^k (1-p)^{n-k}$$

Since binomial distribution is simply n executions of a Bernoulli distribution, notice that for $n = 1$ (for one execution of the experiment) the binomial distribution turns to a Bernoulli distribution:

$$P_X(X = k) = \binom{1}{k} p^k (1-p)^{1-k} = p^k (1-p)^{1-k}$$

where we used the fact that $\binom{1}{0} = \binom{1}{1} = 1$. The final expression is actually the PMF of a Bernoulli distribution.

The binomial distribution is frequently used to model the number of successes in a sample of size n drawn with replacement from a population of size N.

Given that a r.v Y following a binomial distribution will be the summation of a collection of successive r.v's X following a Bernoulli's distribution, $Y = X_1 + X_2 + \ldots + X_n$, the expected value of a binomial distribution reads:

$$E[Y] = E[X_1 + X_2 + \ldots + X_n] = E[X_1] + E[X_2] + \ldots + E[X_n] = \underbrace{p + p + \ldots + p}_{n} = np$$

where in the third step we made use of the linearity of expected value and in the fourth step we used the fact that the expected value of a Bernoulli distribution parametrized by p is simply the parameter p.

Similarly, since $X_1, X_2, \ldots, X_n$ are independent r.v's the variance of their sum is the sum of their variances. Subsequently:

$$\begin{aligned} Var(Y) &= Var(X_1 + X_2 + \ldots + X_n) \\ &= Var(X_1) + Var(X_2) + \ldots + Var(X_n) \\ &= \underbrace{p(1-p) + p(1-p) + \ldots + p(1-p)}_{n} \\ &= np(1-p) \end{aligned}$$

where beside the linearity of variance of independent events we also we used the variance of a Bernoulli distribution parametrized by p which is $p(1-p)$ on the third step.

Finally for the standard deviation of a binomial distribution:

$$SD(X) = \sqrt{Var(X)} = \sqrt{np(1-p)}$$

7.3.4 Poisson Distribution - Pois(λ)

The Poisson distribution parametrized by λ and denoted by Pois(λ) is a discrete probability distribution that expresses the probability of a given number of events occurring in a fixed interval of time or space if these events occur with a known constant rate and independently of the time since the last event. The Poisson distribution can also be used for the number of events in other specified intervals such as distance, area or volume. The corresponding PMF reads:

$$P_X(X = k) = \frac{\lambda^k e^{-\lambda}}{k!}$$

where λ is actually both the expected value and the variance of the distribution (as we will show in a while).

Poisson distribution is very important and practical for statistical modelling since the philosophy behind it is an experiment where the first success is more likely than the second which is more likely than the third and so on, which is very common in real life problems.

In other words, Poisson distribution is a large number of successive Bernoulli trials with very small probability p i.e a binomial distribution with $n \to \infty$ and $p \to 0$ while $\lambda = np$ is held constant.

This can be manifested mathematically since, by starting from binomial distribution's PMF, taking the limit $n \to \infty$ and substituting $p = \frac{\lambda}{n}$ we obtain:

$$\begin{aligned}
\lim_{n\to\infty} P_X(X = k) &= \lim_{n\to\infty} \binom{n}{k} p^k (1-p)^{n-k} \\
&= \lim_{n\to\infty} \binom{n}{k} \left(\frac{\lambda}{n}\right)^k \left(1 - \frac{\lambda}{n}\right)^{n-k} \\
&= \lim_{n\to\infty} \frac{n!}{k!(n-k)!} \left(\frac{\lambda}{n}\right)^k \left(1 - \frac{\lambda}{n}\right)^{n-k} \\
&= \frac{\lambda^k}{k!} \lim_{n\to\infty} \frac{n!}{(n-k)!} \left(\frac{1}{n}\right)^k \left(1 - \frac{\lambda}{n}\right)^{n-k} \\
&= \frac{\lambda^k}{k!} \lim_{n\to\infty} \frac{n!}{(n-k)!} \left(\frac{1}{n^k}\right) \left(1 - \frac{\lambda}{n}\right)^n \left(1 - \frac{\lambda}{n}\right)^{-k} \\
&= \frac{\lambda^k}{k!} \lim_{n\to\infty} \frac{n(n-1)\dots(n-k)(n-k-1)\dots 1}{(n-k)(n-k-1)\dots 1} \left(\frac{1}{n^k}\right) \left(1 - \frac{\lambda}{n}\right)^n \left(1 - \frac{\lambda}{n}\right)^{-k} \\
&= \frac{\lambda^k}{k!} \lim_{n\to\infty} \frac{n(n-1)\dots(n-k)}{n^k} \left(1 - \frac{\lambda}{n}\right)^n \left(1 - \frac{\lambda}{n}\right)^{-k}
\end{aligned}$$

In the first fraction all the numbers can be ignored since $n \to \infty$, and since there are k of them we end up with $n^k/n^k = 1$. Similarly for the last fraction for $n \to \infty \Rightarrow \frac{\lambda}{n} \to 0 \Rightarrow (1 - \frac{\lambda}{n})^{-k} \to 1$. Hence:

$$\begin{aligned}
\lim_{n\to\infty} P_X(X = k) &= \frac{\lambda^k}{k!} \lim_{n\to\infty} \Big(1 - \frac{\lambda}{n}\Big)^n \\
&= \frac{\lambda^k}{k!} \lim_{x\to\infty} \Big(1 + \frac{1}{x}\Big)^{x(-\lambda)} \\
&= \frac{\lambda^k}{k!} \Big(\lim_{x\to\infty} \Big(1 + \frac{1}{x}\Big)^{x}\Big)^{-\lambda} \\
&= \frac{\lambda^k}{k!} e^{-\lambda}
\end{aligned}$$

Thus indeed, by taking the limit of a binomial distribution for $n \to \infty$ we end up with a Poisson distribution.

An example of a Poisson distribution is the case where an individual keeps track of the amount of mail they receive each day and notice that they receive an average number of 4 letters per day. If receiving any particular piece of mail does not affect the arrival times of future pieces of mail, i.e., if pieces of mail from a wide range of sources arrive independently of one another, then a reasonable assumption is that the number of pieces of mail received in a day obeys a Poisson distribution.

Other examples that may follow a Poisson distribution include the number of phone calls received by a call center per hour and the number of decay events per second from a radioactive source. Also by using Poisson distribution we can model the number of meteorites greater than 1 meter diameter that strike earth in a year, the number of patients arriving in an emergency room between 10 and 11 pm, and the number of photons hitting a detector in a particular time interval.

Poisson distribution(and its continuous version of exponential distribution that we will see later) are quite important and heavily used in real world problems.

For the expected value of a Poisson distribution, straight from the definition we get:

$$E[X] = \sum_{k=0}^{\infty} k e^{-\lambda} \frac{\lambda^k}{k!} = e^{-\lambda} \sum_{k=0}^{\infty} \frac{\lambda^k}{(k-1)!} = \lambda e^{-\lambda} \sum_{k=0}^{\infty} \frac{\lambda^{k-1}}{(k-1)!} = \lambda e^{-\lambda} e^{\lambda} = \lambda$$

For the variance of a Poisson distribution, after some calculations (that we will skip for now) we can show that $E[X^2] = \lambda^2 + \lambda$. Hence, by substituting $E[X^2]$ and $E[X]^2$ to the variance relation we get:

$$Var(X) = E[X^2] - E[X]^2 = \lambda^2 + \lambda - \lambda^2 = \lambda$$

Hence we showed that the parameter λ of Pois(λ) is both the expected value and the variance of the distribution.

Finally for the standard deviation of a Poisson distribution:

$$SD(X) = \sqrt{Var(X)} = \sqrt{\lambda}$$

7.3.5 Geometric Distribution - Geo(p)

The geometric distribution parametrized by p and denoted by Geo(p) is a discrete probability distribution that represents the number of failures before you get a success in a series of Bernoulli trials. The corresponding PMF reads:

$$P_X(X = k) = p(1-p)^k$$

An example that a geometric distribution can be used is in the case where an ordinary die is thrown repeatedly until the first time a "1" appears. The probability distribution of the number of times it is thrown is supported on the infinite set $1, 2, 3, \ldots$ and is a geometric distribution with $p = 1/6$.

For the expected value of a geometric distribution we can show:

$$E[X] = \frac{1-p}{p}$$

For the variance of a geometric distribution we can show:

$$Var(X) = \frac{1-p}{p^2}$$

For the standard deviation of a geometric distribution we can show:

$$SD(X) = \frac{\sqrt{1-p}}{p}$$

7.3.6 Hypergeometric Distribution - Hypergeometric(N, K, n)

The hypergeometric distribution parametrized by N, K and n and denoted by Hypergeometric(N, K, n) is a discrete probability distribution that describes the probability of k successes (random draws for which the object drawn has a specified feature) in n draws, without replacement, from a finite population of size N that contains exactly K objects with that feature, wherein each draw is either a success or a failure. In contrast, the binomial distribution describes the probability of k successes in n draws with replacement. The corresponding PMF reads:

$$P_X(X = k) = \frac{\binom{K}{k}\binom{N-K}{n-k}}{\binom{N}{n}}$$

where:

- N is the population size.
- K is the number of success states in the population.
- n is the number of draws.
- k is the number of observed successes.

For the expected value of a hypergeometric distribution we can show:

$$E[X] = \frac{nK}{N}$$

For the variance of a hypergeometric distribution we can show:

$$Var(X) = \frac{nK}{N}\frac{N-K}{N}\frac{N-n}{N-1}$$

For the standard deviation of a hypergeometric distribution we can show:

$$SD(X) = \sqrt{\frac{nK}{N}\frac{N-K}{N}\frac{N-n}{N-1}}$$

7.3.7 Negative Binomial Distribution - NB(r,p)

The negative binomial distribution parametrized by r and p and denoted by NB(r,p) is a discrete probability distribution of the number of successes in a sequence of independent and identically distributed Bernoulli trials with probability of success p before a specified (non-random) number of failures r occurs. (For example, if we define a 1 as failure, all non-1s as successes, and we throw a dice repeatedly until 1 appears the third time (r = three failures), then the probability distribution of the number of non-1s that appeared will be a negative binomial distribution). The corresponding PMF reads:

$$P_X(X = k) = \binom{k+r-1}{k}(1-p)^r p^k$$

where k is the number of successes, r is the number of failures, and p is the probability of success.

For the expected value of a negative binomial distribution we can show:

$$E[X] = \frac{pr}{1-p}$$

For the variance of a negative binomial distribution we can show:

$$Var(X) = \frac{pr}{(1-p)^2}$$

For the standard deviation of a negative binomial distribution we can show:

$$SD(X) = \frac{\sqrt{pr}}{1-p}$$

7.4 Continuous Random Variables

In accordance with discrete r.v's, continuous r.v's are r.v's that can take continuous, uncountable values (as for example the price of a stock). Similarly to the discrete case, we can define a quantity that is connected to the probability of obtaining an outcome that lies between a range of possible values $[a, b]$. In the case of continuous r.v's, this quantity is called "probability density function".

Definition 7.8 (Probability Density Function)**.** *Given a continuous r.v X defined on a sample space S as $X : S \to R$, we define the* ***probability density function*** *(PDF) f_X as a function that maps outcomes R to the interval $[0, 1]$ $(f_X : R \mapsto [0, 1])$.*

$$P(a \leq X \leq b) = \int_a^b f_X(x)dx$$

with

$$\int_{-\infty}^{+\infty} f_X(x)dx = 1$$

Once again the conservation of density in the continuous case gives a kind of "physical" meaning to f_X.

Notice that for the probability of a continuous r.v to take a specific value a :

$$P(a \leq X \leq a) = \int_a^a f_X(x)dx = 0$$

Hence in continuous r.v's it only makes sense to find the probability of a r.v to be inside a specific range $[a, b]$. As it follows from the definition of an integral, the probability of a continuous r.v to be equal to a specific number a is always 0.

Given the PDF we can define the expected value for a continuous r.v in the same way as we did for the discrete case.

Definition 7.9 (Expected Value/Mean)**.** *Let X be a continuous r.v with a finite number of finite outcomes and $f_X(x)$ its corresponding PDF. The* ***expected value*** *(or mean) of X denoted by $E[X]$ or μ is defined as:*

$$E[X] = \int_{-\infty}^{+\infty} x f_X(x)dx$$

The variance and standard deviation of a continuous r.v follow the same formulas as in discrete case, with the difference that now we use the expected value from the definition for the continuous case. In a similar way we can define the corresponding conditional quantities.

7.5 Continuous Probability Distributions

As in discrete r.v's PMF's, the only requirement for a function $f(x)$ to be the PDF of some continuous r.v X is to satisfy the conservation of density. Once a function satisfies this requirement then it describes the probability distribution of some continuous r.v X.

In this section, as we did before for discrete r.v's, we are going to introduce some of the most fundamental continuous probability distributions among with their characteristics and their intuition. Continuous probability distributions are very important because based on some of them we can perform statistical inference as we will see in the next chapter.

7.5.1 Continuous Uniform Distribution - Unif(a,b)

The continuous uniform distribution parametrized by a and b and denoted by Unif(a,b) is a continuous probability distribution that describes an experiment where there is an arbitrary outcome that lies between bounds that are defined by the parameters a and b which are the minimum and maximum values. The interval can be either closed ($[a, b]$) or open ((a, b)).

The corresponding PDF reads:

$$f_X(x) = \begin{cases} c, & \text{for } x \in [a, b] \\ 0, & \text{otherwise} \end{cases}$$

Given that a PDF must satisfy the density conservation, we can actually compute the constant c since:

$$\int_{-\infty}^{+\infty} f_X(x)dx = 1 \Rightarrow$$

$$\int_{-\infty}^{a} 0 \cdot dx + \int_{a}^{b} cdx + \int_{b}^{+\infty} 0 \cdot dx = 1 \Rightarrow$$

$$c \int_{a}^{b} dx = 1 \Rightarrow$$

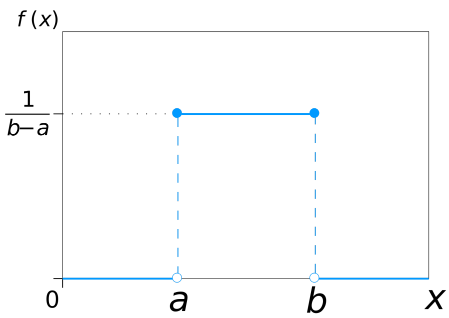
$$c \cdot (b - a) = 1 \Rightarrow$$

$$c = \frac{1}{b - a}$$

Hence, we showed that the constant c is actually the length of the range that the PDF is not 0. By substituting c back to the PDF, we get the final form of a continuous uniform distribution:

$$f_X(x) = \begin{cases} \frac{1}{b-a}, & \text{for } x \in [a, b] \\ 0, & \text{otherwise} \end{cases}$$

Graphically, the PDF looks like this:



For the expected value of a continuous uniform distribution, straight from the definition (**??**) we get:

$$E[X] = \int_{-\infty}^{+\infty} x f_X(x)dx = \int_a^b x\frac{1}{b-a}dx = \frac{1}{b-a}\int_a^b xdx = \frac{1}{b-a}(\frac{1}{2}b^2 - \frac{1}{2}a^2) = \ldots = \frac{1}{2}(a+b)$$

For the variance of a continuous uniform distribution we will use the usual formula, so first we calculate $E[X^2]$:

$$E[X^2] = \int_{-\infty}^{+\infty} x^2 f_X(x)dx = \int_a^b \frac{1}{b-a}x^2dx = \frac{1}{b-a}\int_a^b x^2dx = \frac{1}{b-a}(\frac{1}{3}b^3 - \frac{1}{3}a^3) = \ldots = \frac{1}{3}(a^2+ab+b^2)$$

Substituting $E[X^2]$ and $E[X]^2$ to the variance formula we get:

$$Var(X) = E[X^2] - E[X]^2 = \frac{1}{3}(a^2+ab+b^2) - \frac{1}{4}(a+b)^2 = \ldots = \frac{(b-a)^2}{12}$$

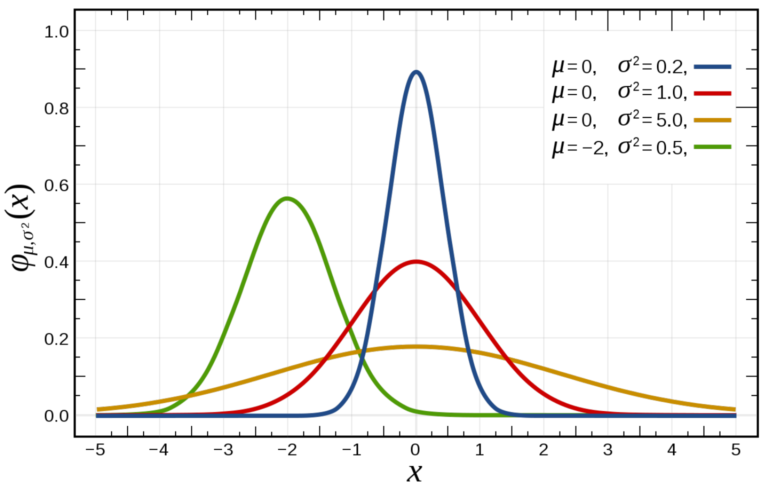
Finally for the standard deviation:

$$SD(X) = \sqrt{Var(X)} = \frac{b-a}{\sqrt{12}}$$

7.5.2 Normal Distribution - N(μ, σ^2)

The normal distribution parametrized by μ, and σ^2 and denoted by N(μ, σ^2) (often called "bell curve") is probably the most important distribution in statistics and it is often used in the natural and social sciences to represent real-valued random variables whose distributions are not known. The PDF of a normal distribution is of the form:

$$f_X(x) = \frac{1}{\sqrt{2\pi\sigma^2}}e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$



Normal distribution is useful because of the central limit theorem, that we will see later. In its most general form it states that averages of samples of observations of r.v's independently drawn from the same distribution converge in distribution to the normal, that is, they become normally distributed when the number of observations is sufficiently large. Physical quantities that are expected to be the sum of many independent processes often have distributions that are nearly normal. Moreover, many results and methods can be derived analytically in explicit form when the relevant variables are normally distributed.

For the expected value of a normal distribution, straight from the definition we get:

$$
\begin{aligned}
E[X] &= \int_{-\infty}^{+\infty} x f_X(x) dx \\
&= \int_{-\infty}^{\infty} x \cdot \frac{1}{\sqrt{2\pi\sigma^2}} \cdot exp\Big(-\frac{(x-\mu)^2}{2\sigma^2}\Big) dx \\
&= \frac{1}{\sqrt{2\pi\sigma^2}} \int_{-\infty}^{\infty} x \cdot exp\Big(-\frac{(x-\mu)^2}{2\sigma^2}\Big) dx \\
&= \frac{1}{\sqrt{2\pi\sigma^2}} \int_{-\infty}^{\infty} (\sqrt{2}\sigma t + \mu) e^{-t^2} d(\sqrt{2}\sigma t) \\
&= \frac{\sqrt{2}\sigma}{\sqrt{2\pi\sigma^2}} \Big(\int_{-\infty}^{\infty} \sqrt{2}\sigma t e^{-t^2} dt + \int_{-\infty}^{\infty} \mu e^{-t^2} dt \Big) \\
&= \frac{1}{\sqrt{\pi}} \Big(\sqrt{2}\sigma \int_{-\infty}^{\infty} t e^{-t^2} dt + \mu \int_{-\infty}^{\infty} e^{-t^2} dt \Big) \\
&= \frac{1}{\sqrt{\pi}} \Big(\sqrt{2}\sigma \cdot 0 + \mu \cdot \sqrt{\pi} \Big) \\
&= \frac{1}{\sqrt{\pi}} \cdot (\mu \cdot \sqrt{\pi}) \\
&= \mu
\end{aligned}
$$

For the variance of a normal distribution we will use the usual relation so first we calculate $E[X^2]$:

$$
\begin{aligned}
E[X^2] &= \int_{-\infty}^{+\infty} x^2 f_X(x)dx \\
&= \int_{-\infty}^{\infty} x^2 \cdot \frac{1}{\sqrt{2\pi\sigma^2}} \cdot exp\Big(-\frac{(x-\mu)^2}{2\sigma^2}\Big)dx \\
&= \frac{1}{\sqrt{2\pi\sigma^2}} \int_{-\infty}^{\infty} x^2 \cdot exp\Big(-\frac{(x-\mu)^2}{2\sigma^2}\Big)dx \\
&= \frac{1}{\sqrt{2\pi\sigma^2}} \int_{-\infty}^{\infty} (\sqrt{2}\sigma t+\mu)^2 e^{-t^2} d(\sqrt{2}\sigma t) \\
&= \frac{\sqrt{2}\sigma}{\sqrt{2\pi\sigma^2}} \int_{-\infty}^{\infty} (2\sigma^2 t^2 + 2\sqrt{2}\sigma t\mu + \mu^2)e^{-t^2} dt \\
&= \frac{1}{\sqrt{\pi}}\Big(\int_{-\infty}^{\infty} 2\sigma^2 t^2 e^{-t^2} dt + \int_{-\infty}^{\infty} 2\sqrt{2}\sigma t\mu e^{-t^2} dt + \int_{-\infty}^{\infty} \mu^2 e^{-t^2} dt\Big) \\
&= \frac{1}{\sqrt{\pi}}\Big(2\sigma^2 \int_{-\infty}^{\infty} t^2 e^{-t^2} dt + 2\sqrt{2}\sigma\mu \int_{-\infty}^{\infty} t e^{-t^2} dt + \mu^2 \int_{-\infty}^{\infty} e^{-t^2} dt\Big) \\
&= \frac{1}{\sqrt{\pi}}\Big(2\sigma^2 \cdot \frac{\sqrt{\pi}}{2} + 2\sqrt{2}\sigma\mu \cdot 0 + \mu^2 \cdot \sqrt{\pi}\Big) \\
&= \frac{1}{\sqrt{\pi}} \cdot \sqrt{\pi} \cdot (\sigma^2 + \mu^2) \\
&= \sigma^2 + \mu^2
\end{aligned}
$$

Substituting $E[X^2]$ and $E[X]^2$ back to the variance relation we get:

$$Var(X) = E[X^2] - E[X]^2 = \sigma^2 + \mu^2 - \mu^2 = \sigma^2$$

So we proved that the parameters of the N(μ, σ^2), μ and σ^2 are actually the expected value and variance of the distribution (hence the naming).

Finally for the standard deviation of a normal distribution :

$$SD(X) = \sqrt{Var(X)} = \sqrt{\sigma^2} = \sigma$$

Now, we will introduce a specific case of a normal distribution called "standard normal distribution".

7.5.3 Standard Normal Distribution - N(0,1)

In the special case where $\mu = 0$ and $\sigma^2 = 1$ the corresponding normal distribution $N(0,1)$ takes the special name of standard normal distribution (or z-distribution). Remember from the graph of a normal distribution, that the bell curve has a maximum at the expected value (in $N(0,1)$ case at 0), and since the variance is equal to 1 (hence also the standard deviation), each extra unit away from the mean is an extra unit of standard deviation (we use standard deviation since it has same units as the mean). For $\mu = 0$ and $\sigma^2 = 1$ the corresponding PDF reads:

$$f_X(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}}$$

Remember that from the definition of a PDF, specifically for a standard normal distribution we have:

$$P(a \leq X \leq b) = \frac{1}{\sqrt{2\pi}} \int_a^b e^{-\frac{x^2}{2}} dx$$

which is the probability of obtaining a value within the range $[a, b]$. By setting $a \to -\infty$, and relabelling $b \to z$ we end up with the probability of the value to be in the range $(-\infty, z]$:

$$P(-\infty \leq X \leq z) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^z e^{-\frac{x^2}{2}} dx$$

This is a standard Gaussian integral that is quite easy to calculate for any possible value of z. Since it is so useful, we can actually find the value for the integral for any z, in the so called "Z-table".

Definition 7.10 (Z-Table). *A Z-table, is a mathematical table for the values of the cumulative distribution function of the standard normal distribution. It is used to find the probability that a statistic is observed below, above, or between values on the standard normal distribution. Z-tables are typically composed as follows:*

- *The label for rows contains the integer part and the first decimal place of z.*
- *The label for columns contains the second decimal place of z.*
- *The values within the table are the result of the integral, i.e the probabilities.*

For example, for $z = 0.69$, one would look down the rows to find 0.6 and then across the columns to 0.09 which would yield a probability of 0.25, so:

$$P(-\infty \leq X \leq 0.69) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{0.69} e^{-\frac{x^2}{2}} dx = 0.25$$

Z-tables are very useful since we can find values of the integral without actually solving it! On top of that since we are dealing with a standard normal distribution where mean is 0 and variance is 1, the value of z can actually be seen as "standard deviations away from the mean". For example $z = 0.69$ means 0.69 standard deviations away from the mean and the corresponding value from Z-table is the probability of obtaining a value for the r.v, up to and including 0.69 standard deviations (which is 1) away from the mean.

Now let's see some particular values, for some integer values of standard deviation away from the mean:

- $P(-\infty \leq X \leq -3) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{-3} e^{-\frac{x^2}{2}} dx = 0.001$
- $P(-\infty \leq X \leq -2) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{-2} e^{-\frac{x^2}{2}} dx = 0.028$
- $P(-\infty \leq X \leq -1) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{-1} e^{-\frac{x^2}{2}} dx = 0.158$
- $P(-\infty \leq X \leq 0) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{0} e^{-\frac{x^2}{2}} dx = 0.5$
- $P(-\infty \leq X \leq 1) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{1} e^{-\frac{x^2}{2}} dx = 0.841$
- $P(-\infty \leq X \leq 2) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{2} e^{-\frac{x^2}{2}} dx = 0.977$
- $P(-\infty \leq X \leq 3) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{3} e^{-\frac{x^2}{2}} dx = 0.998$

Given these values we can calculate the probability of obtaining a value between standard deviations. E.g:

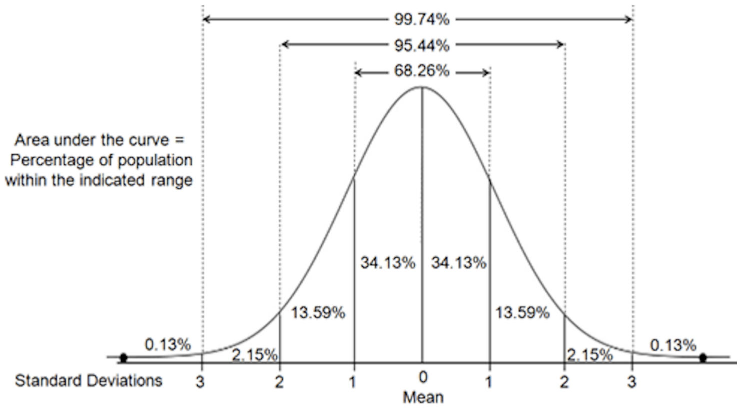
- $P(-3 \leq X \leq -2) = P(-\infty \leq X \leq -2) - P(-\infty \leq X \leq -3) = 0.028 - 0.001 = 0.027 = 2\%$
- $P(-2 \leq X \leq -1) = P(-\infty \leq X \leq -1) - P(-\infty \leq X \leq -2) = 0.158 - 0.028 = 0.13 = 13\%$

- $P(-1 \leq X \leq 0) = P(-\infty \leq X \leq 0) - P(-\infty \leq X \leq -1) = 0.5 - 0.158 = 0.342 = 34\%$
- $P(0 \leq X \leq 1) = P(-\infty \leq X \leq 1) - P(-\infty \leq X \leq 0) = 0.841 - 0.5 = 0.342 = 34\%$
- $P(1 \leq X \leq 2) = P(-\infty \leq X \leq 2) - P(-\infty \leq X \leq 1) = 0.977 - 0.841 = 0.13 = 13\%$
- $P(2 \leq X \leq 3) = P(-\infty \leq X \leq 3) - P(-\infty \leq X \leq 2) = 0.998 - 0.977 = 0.021 = 2\%$

Or within ranges of standard deviations:

- $P(-1 \leq X \leq 1) = P(-\infty \leq X \leq 1) - P(-\infty \leq X \leq -1) = 0.841 - 0.158 = 0.683 = 68\%$
- $P(-2 \leq X \leq 2) = P(-\infty \leq X \leq 2) - P(-\infty \leq X \leq -2) = 0.977 - 0.028 = 0.949 = 95\%$
- $P(-3 \leq X \leq 3) = P(-\infty \leq X \leq 3) - P(-\infty \leq X \leq -3) = 0.998 - 0.001 = 0.997 = 99\%$

We can summarize all of these to the following graph:



Probably the most important message to get out of the graph is the so called "68–95–99.7 rule".

Lemma 7.1 (68–95–99.7 Rule)**.** *The 68–95–99.7 rule (or empirical rule), is a shorthand used to remember the percentage of values that lie within a band around the mean in a normal distribution with a width of two, four and six standard deviations, respectively; more accurately, 68.27%, 95.45% and 99.73% of the values lie within one, two and three standard deviations of the mean, respectively.*

The "68–95–99.7 rule" is used in order to get some informal intuitions out of a standard normal distribution. Also, we can use the "68–95–99.7 rule" even for a normal distribution since, as we will show next, any normal distribution can be turned to a standard normal distribution by a process called "standardization".

In order to formulate the process of standardization, first we need to define the concept of z-score of a r.v X.

Definition 7.11 (z-score)**.** *Given a normal distribution $N(\mu, \sigma^2)$ we can define the* ***z-score*** *(or standard score) of a raw score x as:*

$$z = \frac{x - \mu}{\sigma}$$

The absolute value of z represents the distance between the raw score and the population mean in units of the standard deviation. In simple words it's just a re-scaling of the random variable X.

Notice that for the expected value of z-score, for any $N(\mu, \sigma^2)$ holds:

$$E[z] = E\Big[\frac{x - \mu}{\sigma}\Big] = \frac{1}{\sigma}E[x - \mu] = \frac{1}{\sigma}(E[x] - E[\mu]) = \frac{1}{\sigma}(\mu - \mu) = 0$$

Similarly for the variance of z-score:

$$Var(z) = Var\Big(\frac{x-\mu}{\sigma}\Big) = \frac{1}{\sigma^2}Var(x-\mu) = \frac{1}{\sigma^2}(Var(x) - Var(\mu)) = \frac{1}{\sigma^2}(\sigma^2 - 0) = 1$$

And subsequently for the standard deviation:

$$SD(z) = \sqrt{Var(z)} = \sqrt{1} = 1$$

Hence by switching from X to Z through the use of z-scores we also switch from any normal distribution to a standard normal distribution. This process is called formally "standardization".

Definition 7.12 (Standardization)**.** ***Standardization*** *is the process where starting from a r.v X that follows a $N(\mu, \sigma^2)$, by making use of z-score we switch to a N(0,1) for the r.v Z(X).*

Formally, standardization is just a re-scaling on the way we measure the data. Namely by subtracting the mean out from all the observations we simply define a new zero for the scale, and by dividing by the variance we simply define a new unit for the scale. Nothing actually changes to the actual information of the data since we subtract and divide all observations by the same numbers. The only difference is that now the numbers that represent the data changed to new values in a consistent way. This is why standard normal distribution is so important. Since any normal distribution can be translated to a standard normal distribution everything we said for a standard normal distribution holds for any normal distribution. For example we can compute the probabilities of X to be within a range by switching to Z and compute the Gaussian integral. Also the "68–95–99.7 rule" holds for any normal distribution and simply states:

$$P(\mu - 1\sigma \leq X \leq \mu + 1\sigma) = 68\%$$
$$P(\mu - 2\sigma \leq X \leq \mu + 2\sigma) = 95\%$$
$$P(\mu - 3\sigma \leq X \leq \mu + 3\sigma) = 99.7\%$$

We will get back to normal distributions in the next chapter, where we will show that based on them, we can develop a theory for making statistical inference i.e to draw a conclusion for a random variable out of a small sample of the entire population. More on that, in the next chapter.

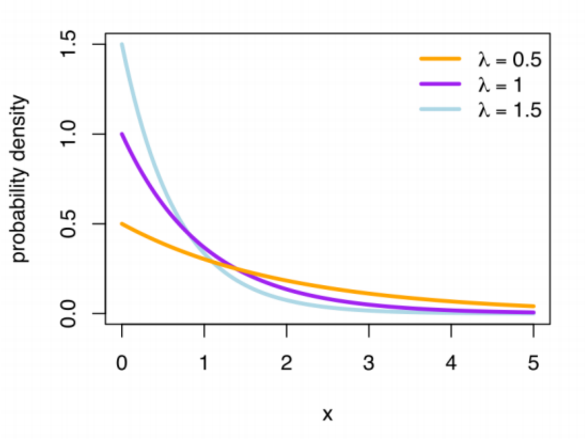
7.5.4 Exponential Distribution - Expo(λ)

The exponential distribution parametrized by rate parameter λ and denoted by Expo(λ) is the probability distribution of the time between events in a Poisson point process, i.e., a process in which events occur continuously and independently at a constant average rate. It is a particular case of the gamma distribution which we will see in later section. It is the continuous analogue of the geometric distribution which we saw in discrete probability distributions. In addition to being used for the analysis of Poisson point processes it is found in various other contexts.

The exponential distribution is not the same as the class of exponential families of distributions, which is a large class of probability distributions that includes the exponential distribution as one of its members, but also includes the normal distribution, binomial distribution, gamma distribution, Poisson, and many others.

The PDF of an exponential distribution reads:

$$f_X(x) = \begin{cases} \lambda e^{-\lambda x} & \text{for } x \geq 0 \\ 0 & \text{for } x < 0 \end{cases}$$



For the expected value of an exponential distribution we can show:

$$E[X] = \frac{1}{\lambda}$$

For the variance of an exponential distribution we can show:

$$Var(X) = \frac{1}{\lambda^2}$$

Finally for the standard deviation of an exponential distribution:

$$SD(X) = \sqrt{Var(X)} = \sqrt{\frac{1}{\lambda^2}} = \frac{1}{\lambda}$$

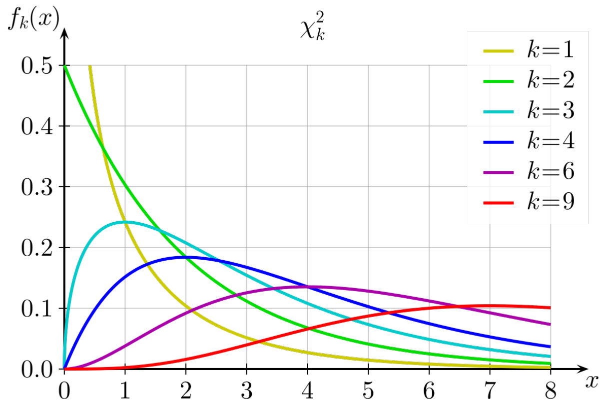
7.5.5 Chi-Squared Distribution - $\chi^2(k)$

Chi-squared distribution parametrized by k degrees of freedom and denoted by $\chi^2(k)$ is the distribution of a sum of the squares of k independent standard normal r.v's. The chi-square distribution is a special case of the gamma distribution that we will see later and is one of the most widely used probability distributions in inferential statistics, notably in hypothesis testing and in construction of confidence intervals.

The chi-square distribution is used in the common chi-square tests for goodness of fit of an observed distribution to a theoretical one, the independence of two criteria of classification of qualitative data, and in confidence interval estimation for a population standard deviation of a normal distribution from a sample standard deviation. Many other statistical tests also use this distribution, such as Friedman's analysis of variance by ranks.

The PDF of a $\chi^2(k)$ reads:

$$f_X(x) = \begin{cases} \frac{x^{\frac{k}{2}-1} e^{-\frac{x}{2}}}{2^{\frac{k}{2}} \Gamma(\frac{k}{2})} & \text{for } x \geq 0 \\ 0 & \text{for } x < 0 \end{cases}$$



For the expected value of a chi-squared distribution we can show:

$$E[X] = k$$

For the variance of a chi-squared distribution we can show:

$$Var(X) = 2k$$

For the standard deviation of a chi-squared distribution we can show::

$$SD(X) = \sqrt{2k}$$

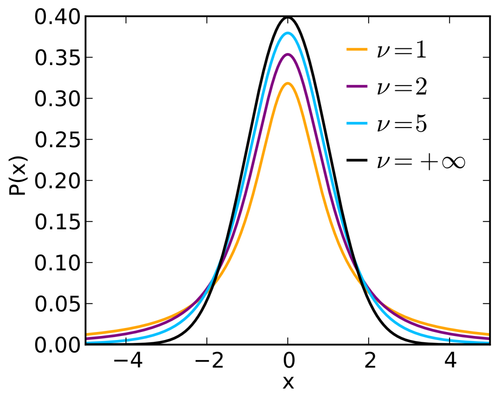
7.5.6 Student's t-Distribution - $t(\nu)$

Student's t-distribution (or simply the t-distribution) parametrized by ν degrees of freedom and denoted by $t(\nu)$ is any member of a family of continuous probability distributions that arises when estimating the mean of a normally distributed population in situations where the sample size is small and the population standard deviation is unknown. More on that in the next chapter.

The t-distribution plays a role in a number of widely used statistical analyses, including Student's t-test for assessing the statistical significance of the difference between two sample means, the construction of confidence intervals for the difference between two population means, and in linear regression analysis. The Student's t-distribution also arises in the Bayesian analysis of data from a normal family.

The PDF of a t-distribution reads:

$$f_X(x) = \frac{\Gamma\left(\frac{\nu+1}{2}\right)}{\sqrt{\nu\pi}\Gamma\left(\frac{\nu}{2}\right)}\left(1+\frac{t^2}{\nu}\right)^{-\frac{\nu+1}{2}}$$



For the expected value of a t-distribution we can show:

$$E[X] = 0$$

For the variance of a t-distribution we can show:

$$Var(X) = \frac{\nu}{\nu - 2}$$

Finally for the standard deviation:

$$SD(X) = \sqrt{\frac{\nu}{\nu - 2}}$$

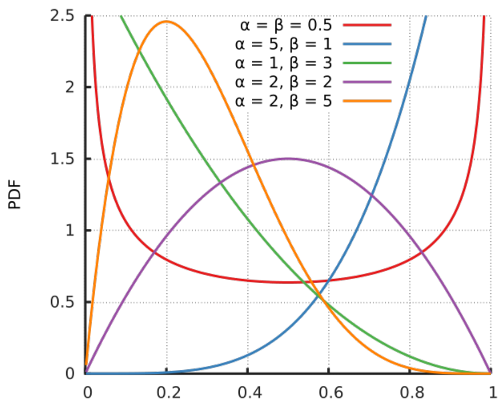
7.5.7 Beta Distribution - Beta(α, β)

Beta distribution Beta(α, β) is a family of continuous probability distributions defined on the interval [0, 1] parametrized by two positive shape parameters, denoted by α and β, that appear as exponents of the random variable and control the shape of the distribution.

The beta distribution has been applied to model the behaviour of random variables limited to intervals of finite length in a wide variety of disciplines.

The PDF of a Beta distribution is given by:

$$f_X(X) = \frac{1}{B(\alpha, \beta)} \cdot x^{\alpha-1} \cdot (1-x)^{\beta-1}$$



Observe that $B(\alpha, \beta)$ is not just one probability distribution, but a family of probability distributions since for different values of α and β we end up with different distributions.

For the expected value of a Beta distribution we can show:

$$E[X] = \frac{\alpha}{\alpha + \beta}$$

For the variance of a Beta distribution we can show:

$$Var(X) = \frac{\alpha\beta}{(\alpha + \beta)^2(\alpha + \beta + 1)}$$

Finally for the standard deviation:

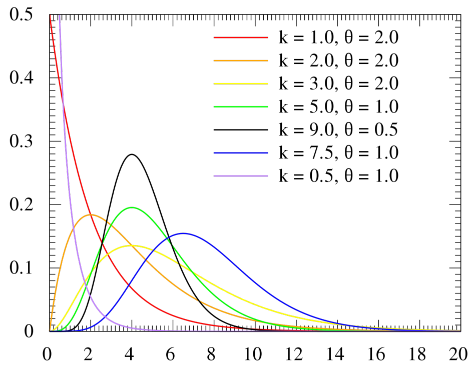
$$SD(X) = \sqrt{\frac{\alpha\beta}{(\alpha + \beta)^2(\alpha + \beta + 1)}}$$

7.5.8 Gamma Distribution - Gamma(α, β)

Gamma distribution parametrized by two positive shape parameters *alpha* and β and denoted by Gamma(α, β) is a family of continuous probability distributions defined on the interval $[0, \infty)$.

The PDF of a Gamma distribution is given by:

$$f_X(x) = \frac{\beta^\alpha \cdot x^{\alpha-1} \cdot e^{-\beta x}}{\Gamma(\alpha)}$$



Observe that Gamma distribution is not just one probability distribution, but a family of probability distributions since for different values of α and β) we end up with different distributions. For example, the exponential distribution and the chi-squared distribution that we already showed, are special cases of the gamma distribution.

For the expected value of a Gamma distribution we can show:

$$E[X] = \frac{\alpha}{\beta}$$

For the variance of a Gamma distribution we can show:

$$E[X] = \frac{\alpha}{\beta^2}$$

Finally for the standard deviation:

$$SD(X) = \frac{\sqrt{\alpha}}{\beta}$$

7.6 Joint Probability Distribution

Up to this point we have defined everything for one single r.v X (either discrete or continuous). However, the definitions can be generalized to a collection of any number of r.v's $\{X_1, X_2, \ldots, X_n\}$, (treating the whole collection of r.v's as an entity) leading to the concept of a "joint probability distribution".

Definition 7.13 (Joint Probability Distribution)**.** *Given a number of r.v's $\{X_1, X_2, \ldots, X_n\}$, that are defined on a probability space, the **joint probability distribution** for $\{X_1, X_2, \ldots, X_n\}$ is a probability distribution that gives the probability that each of $\{X_1, X_2, \ldots, X_n\}$ falls in any particular range or discrete set of values specified for that variable. In the case of only two r.v's, this is called a bivariate distribution, but the concept generalizes to any number of r.v's, giving a multivariate distribution.*

Subsequently, we can define the concept of a "joint probability mass function" and a "joint probability density function" for a collection of discrete and continuous r.v's respectively. More specifically:

Definition 7.14 (Joint Probability Mass Function)**.** *For a number of discrete r.v's $\{X_1, \ldots, X_n\}$, that are defined on a probability space S, we define the **joint probability mass function** $P_{X_1,\ldots,X_n}(x_1, \ldots, x_n)$*

as a function that maps outcomes R^n to the interval $[0,1]$ *($P_{X_1,\ldots,X_n} : R^n \mapsto [0,1]$).*

$$P_{X_1,\ldots,X_n}(X_1 = x_1, \ldots, X_n = x_n) = P_{X_1,\ldots,X_n}(\{s \in S : X_1(s) = x_1, \ldots, X_n(s) = x_n\})$$

with:

$$\sum_{x_1} \cdots \sum_{x_n} P_{X_1,\ldots,X_n}(X_1 = x_1, \ldots, X_n = x_n) = 1$$

In the case where X's are independent to each other then their joint probability mass function is just the product of their individual PMF's:

$$P_{X_1,\ldots,X_n}(X_1 = x_1, \ldots, X_n = x_n) = \prod_{i=1}^{n} P_{X_i}(X_i = x_i)$$

Notice that the individual r.v's $\{X_1, \ldots, X_n\}$ can follow different distributions. However, more often than not we will be dealing with the case where all $\{X_1, \ldots, X_n\}$ individually follow the same distribution $P_{X_1} = P_{X_2} = \ldots = P_{X_n} = P_X$. In this case we formally say that we are dealing with "independent identical distributed random variables" or i.i.d r.v's. This abbreviation is used a lot in statistics since many real world problems are problems that can be modelled with the use of i.i.d r.v's.

In a similar way as in the discrete case, we can define a joint probability density function for the continuous case which will be:

Definition 7.15 (Joint Probability Density Function)**.** *For a number of continuous r.v's $\{X_1, \ldots, X_n\}$, that are defined on a probability space S, we define the **joint probability density function** $f_{X_1,\ldots,X_n}(x_1, \ldots, x_n)$ as a function that maps outcomes R^n to the interval* $[0,1]$ *($f_{X_1,\ldots,X_n} : R^n \mapsto [0,1]$).*

$$P(a_1 \leq X_1 \leq b_1, \ldots, a_n \leq X_n \leq b_n) = \int_{a_1}^{b_1} \cdots \int_{a_n}^{b_n} f_{X_1,\ldots,X_n}(x_1, \ldots, x_n) dx_1 \ldots dx_n$$

with

$$\int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} f_{X_1,\ldots,X_n}(x_1, \ldots, x_n) dx_1 \ldots dx_n = 1$$

And again, in the case where X's are independent to each other then their joint probability density function is just the product of their individual PDF's:

$$f_{X_1,\ldots,X_n}(x_1, \ldots, x_n) = \prod_{i=1}^{n} f_{X_i}(x_i)$$

Of course whatever we said for i.i.d r.v's on the discrete case, hold also for the continuous case with the difference that now all i.i.d r.v's will follow the same probability density function $f_{X_1} = f_{X_2} = \ldots = f_{X_n} = f_X$

7.6.1 Bivariate Joint Distribution

Probably the most important case of joint probability distribution is the case when we are dealing with two random variables X and Y. In a way, in joint probability distributions, and subsequently in bivariate probability distributions, the analysis switches from trying to exploring just X and Y as r.v's to also exploring what is the relation between the two r.v's.

For now we will be dealing with the case where both X and Y are discrete r.v's and then we will see the continuous case.

In the case of 2 discrete r.v's, straight from the definition, their joint probability mass function reads:

$$P_{X,Y}(X = x, Y = y) = P_{X,Y}(\{s \in S : X(s) = x, Y(s) = y\})$$

with:

$$\sum_x \sum_y P_{X,Y}(X = x, Y = y) = 1$$

As before, in case where X and Y are independent the joint mass probability distribution is just the product of the two PMF's:

$$P_{X,Y}(X = x, Y = y) = P_X(X = x) \cdot P_Y(Y = y)$$

Now we can generalize the concept of expected value to the one of "joint expected value". (Since we are dealing with bivariate joint distributions we will define the joint expected value for two r.v's, however we can generalize the definition for any function of any numbers of r.v's)

Definition 7.16 (Joint Expected Value)**.** *Let X and Y be two discrete r.v's with a finite number of finite outcomes and $P_{X,Y}(x, y)$ its corresponding PMF. The* ***joint expected value*** *of X and Y denoted by $E[X \cdot Y]$ is defined as:*

$$E[X \cdot Y] = \sum_x \sum_y x \cdot y \cdot P_{X,Y}(X = x, Y = y)$$

By making use of independent r.v's and the joint probability, we can show the following relation for the joint expected value of independent r.v's.

Lemma 7.2. *If X and Y are independent events then their joint expected value is equal to the product of the expected values of each r.v:*

Proof.

$$\begin{aligned}
E[X \cdot Y] &= \sum_x \sum_y x \cdot y \cdot P_{X,Y}(X = x, Y = y) \\
&= \sum_x \sum_y x \cdot y \cdot P_X(X = x) \cdot P_Y(Y = y) \\
&= \sum_x x P_X(X = x) \cdot \sum_y y P_Y(Y = y) \\
&= \Big(\sum_x x P_X(X = x)\Big)\Big(\sum_y y P_Y(Y = y)\Big) \\
&= E[X]E[Y]
\end{aligned}$$

$\square$

The joint expected value can be generalized to $E[g(X, Y)]$ for any function g of the random variables X and Y. In case where $g(X, Y) = g_X(X) \cdot g_Y(Y)$ then for independent events also holds $E[g(X, Y)] = E[g_X(X)]E[g_Y(Y)]$. A generalization to a function g for any number of r.v's is also possible.

Similarly to the joint expected value, it follows that we can also expand the definition of variance to the case of 2 r.v's. The corresponding quantity is called "covariance".

Definition 7.17 (Covariance)**.** *Let X and Y be two discrete r.v's with a finite number of finite outcomes and $P(x, y)$ its corresponding PMF. The* ***covariance*** *of X and Y denoted by $Cov(X, Y)$ is defined as the expected value of the deviation of each r.v from their corresponding expected value.*

$$Cov(X, Y) = E[(X - E[X])(Y - E[Y])]$$

Covariance satisfies the following properties:

- $Cov(X, Y) = Cov(Y, X)$
- $Cov(X, c) = 0, \quad \forall c$

- $Cov(cX, Y) = c \cdot Cov(X, Y)$
- $Cov(X, Y + Z) = Cov(X, Y) + Cov(X, Z)$

Notice that for $Y = X$:

$$Cov(X, X) = E[(X - E[X])(X - E[X])] = E[(X - E[X])^2] = Var(X)$$

Hence, the covariance of one r.v with itself is simply the variance of the r.v. So indeed, variance can be interpreted as a special case of covariance hence covariance is a generalization of variance.

A similar formula as for variance can also be derived for covariance:

$$\begin{aligned}
Cov(X, Y) &= E[(X - E[X])(Y - E[Y])] \\
&= E[X \cdot Y - X \cdot E[Y] - E[X] \cdot Y + E[X] \cdot E[Y]] \\
&= E[X \cdot Y] - E[X \cdot E[Y]] - E[E[X] \cdot Y] + E[E[X] \cdot E[Y]] \\
&= E[X \cdot Y] - E[Y] \cdot E[X] - E[X] \cdot E[Y] + E[X] \cdot E[Y] \\
&= E[X \cdot Y] - E[X] \cdot E[Y]
\end{aligned}$$

As in the case of variance, this formula is more handy for calculations than the actual definition of covariance.

In case of independent events, by making use of the last lemma we get for the covariance of independent events:

$$Cov(X, Y) = E[X \cdot Y] - E[X] \cdot E[Y] = E[X] \cdot E[Y] - E[X] \cdot E[Y] = 0$$

Hence we proved that the covariance of two independent events is always zero. However, the opposite does not necessarily hold. To make it more clear let's give the definition of uncorrelated r.v's.

Definition 7.18 (Uncorrelated Random Variables)**.** *Two r.v's X and Y are called **uncorrelated** if $Cov(X, Y) = 0$.*

Hence, the important concept here is that independent r.v's are always uncorrelated, but uncorrelated r.v's are not necessarily independent.

By making use of covariance we can prove the following relation for variance:

$$\begin{aligned}
Var(X + Y) &= Cov(X + Y, X + Y) \\
&= Cov(X + Y, X) + Cov(X + Y, Y) \\
&= Cov(X, X) + Cov(Y, X) + Cov(X, Y) + Cov(Y, Y) \\
&= Var(X) + 2Cov(X, Y) + Var(Y) \\
&= Var(X) + Var(Y) + 2Cov(X, Y)
\end{aligned}$$

From this formula we can see why variance is not actually linear, since the covariance of the two terms appears in the formula. This is why when we are dealing with independent r.v's, hence uncorrelated r.v's with $Cov(X, Y) = 0$, we can actually treat variance as linear, i.e: $Var(X+Y) = Var(X)+Var(Y)$.

Covariance carries the same problems as variance, with the main one being that it carries square units. However, there is a way to overcome this problem by normalizing the covariance with the standard deviation of the two r.v's. The result is a very useful measure in statistics called "correlation".

Definition 7.19 (Correlation)**.** *Let X and Y be two discrete r.v's with a finite number of finite outcomes and $P(x, y)$ its corresponding PMF. The **correlation** of X and Y denoted by $\rho(X, Y)$ is defined as the covariance of the the two r.v's divided by the their corresponding standard deviations.*

$$\rho(X, Y) = \frac{Cov(X, Y)}{SD(X)SD(Y)} = \frac{E[(X - E[X])(Y - E[Y])]}{SD(X)SD(Y)}$$

Notice that since standard deviation is just a number it can get inside an expected value, and the correlation can be manipulated to:

$$\begin{aligned}
\rho(X,Y) &= \frac{E[(X-E[X])(Y-E[Y])]}{SD(X)SD(Y)} \\
&= E\Big[\frac{(X-E[X])(Y-E[Y]))}{SD(X)SD(Y)}\Big] \\
&= E\Big[\Big(\frac{X-E[X]}{SD(X)}\Big)\Big(\frac{Y-E[Y]}{SD(Y)}\Big)\Big] \\
&= Cov\Big(\frac{X-E[X]}{SD(X)}, \frac{Y-E[Y]}{SD(Y)}\Big)
\end{aligned}$$

Hence we showed that the correlation is actually the covariance of the standardized (z-scores) r.v's X and Y.

By using all variance, covariance and the definition of standard deviation being the square root of variance follows:

$$\rho(X,Y) = \frac{E[XY]-E[X]E[Y]}{\sqrt{E[X^2]-E[X]^2}\cdot\sqrt{E[Y^2]-E[Y]^2}}$$

Two of the most important properties of correlation is that it carries no units (by definition), and it is always between -1 and 1. Let's show that!

Lemma 7.3.

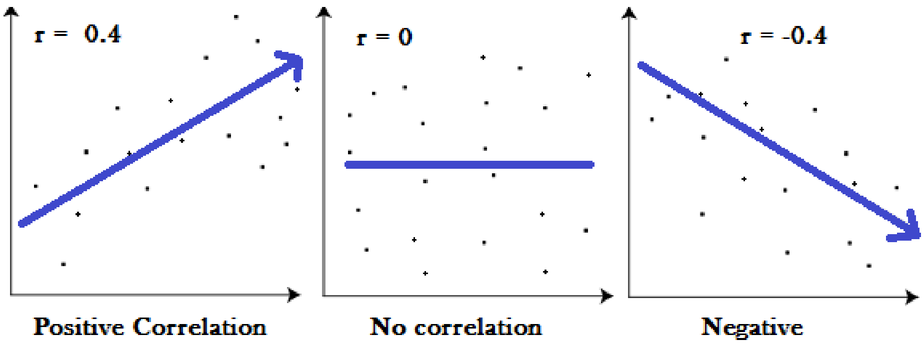
$$-1 \leq \rho \leq +1$$

Proof. Without loss of generality we can assume that X and Y are standardized r.v's (i.e $E[X] = E[Y] = 0$ and $Var(X) = Var(Y) = 1$). In this case $Cov(X,Y) = \rho(X,Y)$. Then:

$$\begin{aligned}
Var(X \pm Y) &= Var(X) + Var(Y) \pm 2 \cdot Cov(X,Y) \\
&= 1 + 1 \pm 2 \cdot Cov(X,Y) \\
&= 2 \pm 2 \cdot Cov(X,Y) \\
&= 2 \pm 2 \cdot \rho(X,Y) \\
&= 2(1 \pm \rho(X,Y))
\end{aligned}$$

But since variance is the expected value of a square term, it follows that it's always positive, subsequently $(1 \pm \rho(X,Y))$ must always be positive hence we end up with $-1 \leq \rho \leq +1$. □

The fact that correlation does not carry any units, and it is always between -1 and 1 makes it a very useful measure for the relation between two r.v's. Namely, a 0 correlation means there is no relationship between the variables at all, while -1 or 1 means that there is a perfect negative or positive correlation (negative or positive correlation here refers to the type of graph the relationship will produce). All in-between values indicate different levels of correlation. Intuitively, in the case of negative correlation the two r.v's follow an opposite trend meaning that while the first one gets larger the other one gets smaller. The opposite happens for positive correlation.



As a final note, we can in a similar way define everything for the case of continuous r.v's. Namely, their joint probability density function will read:

$$P(a_x \leq X \leq b_x, a_y \leq Y \leq b_y) = \int_{a_x}^{b_x} \int_{a_y}^{b_y} f_{X,Y}(x,y)dxdy$$

with

$$\int_{-\infty}^{\infty} \int_{-\infty}^{\infty} f_{X,Y}(x,y)dxdy = 1$$

And of course, in case where X and Y are independent the joint density probability distribution is just the product of the two PDF's:

$$f_{X,Y}(X = x, Y = y) = f_X(X = x) \cdot f_Y(Y = y)$$

By making use of the joint probability distribution we can define the joint expected value as:

Definition 7.20 (Joint Expected Value)**.** *Let X and Y be two continuous r.v's $f_{X,Y}(x,y)$ its corresponding PMF. The **joint expected value** of X and Y denoted by $E[X \cdot Y]$ is defined as:*

$$E[X \cdot Y] = \int_{-\infty}^{+\infty} \int_{-\infty}^{+\infty} x \cdot y \cdot f_{X,Y}(X = x, Y = y)$$

which can be generalized to any function g of the two r.v's:

$$E[g(X \cdot Y)] = \int_{-\infty}^{+\infty} \int_{-\infty}^{+\infty} g(x \cdot y) \cdot f_{X,Y}(X = x, Y = y)$$

Using this definition for the joint expected value we can define covariance, correlation and prove the independent joint expected value. All of them have the exact same form as in the discrete case with the only difference that now instead of the discrete joint probability distribution we use the continuous joint probability distribution.

Application: Bivariate Bernoulli Distribution

Let's assume two i.i.d r.v's X and Y that both follow some Bern(p). Their joint probability distribution is given by the following probabilities:

- $P_{X,Y}(X = 0, Y = 0) = \frac{1}{3}$
- $P_{X,Y}(X = 0, Y = 1) = \frac{1}{6}$
- $P_{X,Y}(X = 1, Y = 0) = \frac{1}{3}$
- $P_{X,Y}(X = 1, Y = 1) = \frac{1}{6}$

First let's check if X and Y are independent by checking their individual probability distributions.

For X it is

$$P_X(X=0) = \sum_y P_{X,Y}(X=0, Y=y) = P_{X,Y}(X=0, Y=0) + P_{X,Y}(X=0, Y=1) = \frac{1}{3} + \frac{1}{6} = \frac{1}{2}$$

and

$$P_X(X=1) = \sum_y P_{X,Y}(X=1, Y=y) = P_{X,Y}(X=1, Y=0) + P_{X,Y}(X=1, Y=1) = \frac{1}{3} + \frac{1}{6} = \frac{1}{2}$$

Hence X follows a Bern$\left(\frac{1}{2}\right)$.

For Y it is

$$P_Y(Y=0) = \sum_x P_{X,Y}(X=x, Y=0) = P_{X,Y}(X=0, Y=0) + P_{X,Y}(X=1, Y=0) = \frac{1}{3} + \frac{1}{3} = \frac{2}{3}$$

and

$$P_Y(Y=1) = \sum_x P_{X,Y}(X=x, Y=1) = P_{X,Y}(X=0, Y=1) + P_{X,Y}(X=1, Y=1) = \frac{1}{6} + \frac{1}{6} = \frac{1}{3}$$

Hence Y follows a Bern$\left(\frac{1}{3}\right)$.

By using these values we can show:

- $P_X(X=0) \cdot P_Y(Y=0) = \frac{1}{2} \cdot \frac{2}{3} = \frac{1}{3} = P_{X,Y}(X=0, Y=0)$
- $P_X(X=0) \cdot P_Y(Y=1) = \frac{1}{2} \cdot \frac{1}{3} = \frac{1}{6} = P_{X,Y}(X=0, Y=1)$
- $P_X(X=1) \cdot P_Y(Y=0) = \frac{1}{2} \cdot \frac{2}{3} = \frac{1}{3} = P_{X,Y}(X=1, Y=0)$
- $P_X(X=1) \cdot P_Y(Y=1) = \frac{1}{2} \cdot \frac{1}{3} = \frac{1}{6} = P_{X,Y}(X=1, Y=1)$

Hence we proved:

$$P_{X,Y}(X=x, Y=y) = P_X(X=x) \cdot P_Y(Y=y)$$

which means that indeed X and Y are independent.

Since they are independent we can calculate their joint expected value simply as

$$E[X \cdot Y] = E[X] \cdot E[Y] = \frac{1}{2} \cdot \frac{1}{3} = \frac{1}{6}$$

where we used the fact that the expected value of a Bernoulli distribution is p.

Finally, since X and Y are independent, that means that they are uncorrelated hence:

$$Cov(X, Y) = 0$$

Subsequently for their correlation:

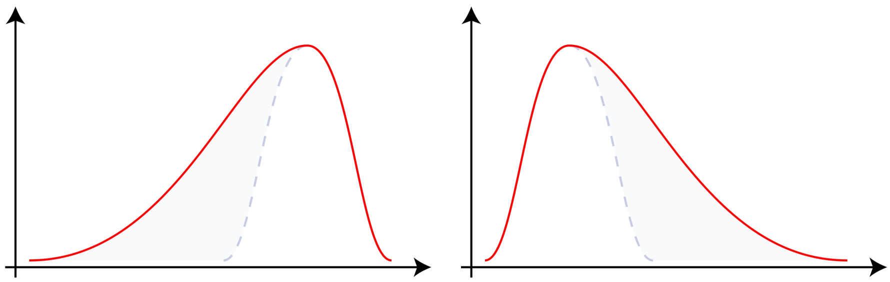
$$\rho(X, Y) = 0$$

7.7 Moments

In a previous section we defined the three most important measures for a r.v: the expected value, the variance and the standard deviation. However, that are not the only ones that exist. In this chapter we will introduce some of the rest measures which are of secondary importance but they will help us to generalize all of them in one single concept that is called “moments”.

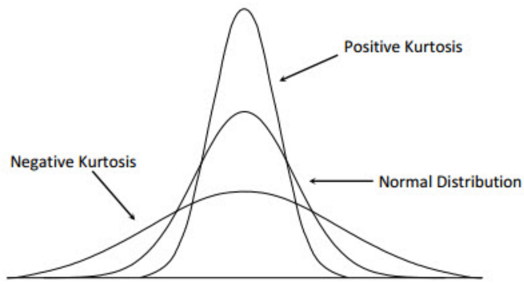
Definition 7.21 (Skewness)**.** ***Skewness*** *is a measure of the asymmetry of the probability distribution of a r.v about its mean.*

$$Skew(X) = E\Big[\Big(\frac{X-\mu}{\sigma}\Big)^3\Big]$$



Definition 7.22 (Kurtosis)**.** ***Kurtosis*** *is a measure of the "tailedness" of the probability distribution of a r.v about its mean.*

$$Kurt(X) = E\Big[\Big(\frac{X-\mu}{\sigma}\Big)^4\Big]$$



Definition 7.23 (Hyperskewness)**.** ***Hyperskewness*** *is a measure of the "hyperskewness" of the probability distribution of a r.v about its mean.*

$$Hyperskewness(X) = E\Big[\Big(\frac{X-\mu}{\sigma}\Big)^5\Big]$$

Definition 7.24 (Hypertailedness)**.** ***Hypertailedness*** *is a measure of the "hypertailedness" of the probability distribution of a r.v about its mean.*

$$Hypertailedness(X) = E\Big[\Big(\frac{X-\mu}{\sigma}\Big)^6\Big]$$

All these measures, including expected value, variance and standard deviation, can be grouped together under one entity which is called "moments". We distinguish between three different kinds of moments: raw, central and standardized.

Definition 7.25 (Raw Moments)**.** *Given a discrete or continuous r.v X with PMF $P_X(x)$ or PDF $f_X(x)$ respectively, we define the n-th* ***raw moment*** *μ_n as:*

$$\mu_n = E[X^n]$$

For $n = 1$ the first raw moment reads:

$$\mu_1 = E[X^1] = E[X] = \mu$$

Hence the first raw moment is actually the mean of a r.v.

Definition 7.26 (Central Moments)**.** *Given a discrete or continuous r.v X with PMF $P_X(x)$ or PDF $f_X(x)$ respectively, we define the n-th* ***central moment*** *μ_n as:*

$$\mu_n = E[(X-\mu)^n]$$

For $n = 1$ the first central moment reads:

$$\mu_1 = E[(X-\mu)^1] = E[X] - E[\mu] = \mu - \mu = 0$$

For $n = 2$ the second central moment reads:

$$\mu_2 = E[(X-\mu)^2] = \sigma^2$$

Hence the first central moment is actually 0 while the second central moment is the definition of variance.

Definition 7.27 (Standardized Moments)**.** *Given a discrete or continuous r.v X with PMF $P_X(x)$ or PDF $f_X(x)$ respectively, we define the n-th* ***standardized moment*** *μ_n as:*

$$\mu_n = E\Big[\Big(\frac{X-\mu}{\sigma}\Big)^n\Big]$$

For $n = 1$ the first standardized moment reads:

$$\mu_1 = E\Big[\Big(\frac{X-\mu}{\sigma}\Big)^1\Big] = \frac{1}{\sigma}E[X-\mu] = \frac{1}{\sigma}(E[X]-E[\mu]) = \frac{1}{\sigma}(\mu-\mu) = 0$$

For $n = 2$ the second standardized moment reads:

$$\mu_2 = E\Big[\Big(\frac{X-\mu}{\sigma}\Big)^2\Big] = \frac{1}{\sigma^2}E[(X-\mu)^2] = \frac{1}{\sigma^2}\sigma^2 = 1$$

For $n = 3$ the third standardized moment reads: c

$$\mu_3 = E\Big[\Big(\frac{X-\mu}{\sigma}\Big)^3\Big] = Skew(X)$$

For $n = 4$ the fourth standardized moment reads:

$$\mu_4 = E\Big[\Big(\frac{X-\mu}{\sigma}\Big)^4\Big] = Kurt(X)$$

For $n = 5$ the fifth standardized moment reads:

$$\mu_5 = E\Big[\Big(\frac{X-\mu}{\sigma}\Big)^5\Big] = Hyperskewness(X)$$

For $n = 6$ the sixth standardized moment reads:

$$\mu_6 = E\Big[\Big(\frac{X-\mu}{\sigma}\Big)^6\Big] = Hypertailedness(X)$$

Let's summarize all moments in the following matrix:

Moment ordinal	Moment		
	Raw	Central	Standardized
1	Mean	0	0
2	–	Variance	1
3	–	–	Skewness
4	–	–	(Non-excess or historical) kurtosis
5	–	–	Hyperskewness
6	–	–	Hypertailedness
7+	–	–	–

Definition 7.28 (Moment Generating Function)**.** *Given a discrete or continuous r.v X with PMF $P_X(x)$ or PDF $f_X(x)$ respectively, we define the* ***moment generating function*** *M_X as:*

$$M_X(t) = E[e^{tX}], \quad t \in R$$

The moment generating function is so named because it can be used to find the moments of the distribution. Namely:

$$\begin{aligned}
M_X(t) &= E[e^{tX}] \\
&= E\Big[1 + tX + \frac{t^2X^2}{2!} + \frac{t^3X^3}{3!} + \ldots + \frac{t^nX^n}{n!} + \ldots\Big] \\
&= E[1] + E[tX] + E\Big[\frac{t^2X^2}{2!}\Big] + E\Big[\frac{t^3X^3}{3!}\Big] + \ldots + E\Big[\frac{t^nX^n}{n!}\Big] + \ldots \\
&= 1 + tE[X] + \frac{t^2}{2!}E[X^2] + \frac{t^3}{3!}E[X^3] + \ldots + \frac{t^n}{n!}E[X^n] + \ldots \\
&= 1 + t\mu_1 + \frac{t^2\mu_2}{2!} + \frac{t^3\mu_3}{3!} + \ldots + \frac{t^n\mu_n}{n!} + \ldots
\end{aligned}$$

Moment generating function is really important since by differentiating it i times with respect to t and setting $t = 0$, we obtain the i-th raw moment. Also we can determine probability distributions since if two r.v's have the same moment generating function that means that they follow the same probability distribution.

Now just for practise, let's calculate the moment generating function for some of the probability distributions we introduced earlier. (Not all of them)

- For a Bernoulli distribution:

  $$M_X(t) = E[e^{tX}] = \sum_x e^{tX}P_X(x) = e^{t\cdot 0}P_X(0) + e^{t\cdot 1}P_X(1) = e^0(1-p) + e^tp = 1 - p + e^tp$$

  Observe that the derivative of $M_X(t)$ evaluated at 0 is actually the expected value of the Bernoulli distribution:

  $$\frac{dM_X(t)}{dt}\bigg|_{t=0} = \frac{d}{dt}(1 - p + e^tp)\bigg|_{t=0} = e^tp\bigg|_{t=0} = p$$

- For a binomial distribution:

  Since a binomial distribution is n trials of a Bernoulli distribution, for the moment generating function of a binomial distribution we simply get:

$$M_X(t) = (1 - p + e^t p)^n$$

Observe that the derivative of $M_X(t)$ evaluated at 0 is actually the expected value of the binomial distribution:

$$\left.\frac{dM_X(t)}{dt}\right|_{t=0} = \left.\frac{d}{dt}(1 - p + e^t p)^n\right|_{t=0} = \left.n \cdot (1 - p + e^t p)^{n-1}(e^t p)\right|_{t=0} = np$$

- For a Poisson distribution:

$$M_X(t) = E[e^{tX}] = \sum_{k=0}^{\infty} e^{tk} P_X(k) = \sum_{k=0}^{\infty} e^{tk} \frac{\lambda^k e^{-\lambda}}{k!} = e^{-\lambda} \sum_{k=0}^{\infty} \frac{(\lambda e^t)^k}{k!} = e^{-\lambda} e^{\lambda e^t} = e^{\lambda(e^t - 1)}$$

Observe that the derivative of $M_X(t)$ evaluated at 0 is actually the expected value of the Poisson distribution:

$$\left.\frac{dM_X(t)}{dt}\right|_{t=0} = \left.\frac{d}{dt} e^{\lambda(e^t-1)}\right|_{t=0} = \left.e^{\lambda(e^t-1)}(\lambda e^t)\right|_{t=0} = \lambda$$

- For a continuous uniform distribution:

$$M_X(t) = E[e^{tX}] = \int_a^b e^{tx} \cdot \frac{1}{b-a} dx = \frac{1}{b-a} \int_a^b e^{tx} dx = \frac{e^{tb} - e^{ta}}{t(b-a)}$$

Observe that the derivative of $M_X(t)$ evaluated at 0 is actually the expected value of the continuous uniform distribution:

$$\left.\frac{dM_X(t)}{dt}\right|_{t=0} = \left.\frac{d}{dt} \frac{e^{tb} - e^{ta}}{t(b-a)}\right|_{t=0} = \ldots = \frac{a+b}{2}$$

- For an exponential distribution:

$$M_X(t) = E[e^{tX}] = \int_0^{\infty} e^{tx} \cdot e^{-\lambda x} dx = \int_0^{\infty} e^{tx - \lambda x} dx = \int_0^{\infty} e^{x(t-\lambda)} dx = \frac{1}{\lambda - t}$$

Observe that the derivative of $M_X(t)$ evaluated at 0 is actually the expected value of the exponential distribution:

$$\left.\frac{dM_X(t)}{dt}\right|_{t=0} = \left.\frac{d}{dt} \frac{1}{\lambda - t}\right|_{t=0} = -\left.\frac{1}{(\lambda - t)^2}(-1)\right|_{t=0} = \frac{1}{\lambda}$$

- For a Normal distribution:

$$
\begin{aligned}
M_X(t) &= E[e^{tX}] \\
&= \frac{1}{\sqrt{2\pi\sigma^2}} \int_{-\infty}^{\infty} exp(tx) \cdot exp(-\frac{(x-\mu)^2}{2\sigma^2})dx \\
&= \frac{1}{\sqrt{2\pi\sigma^2}} \int_{-\infty}^{\infty} exp\Big(tx - \frac{(x-\mu)^2}{2\sigma^2}\Big)dx \\
&= \frac{1}{\sqrt{2\pi\sigma^2}} \int_{-\infty}^{\infty} exp\Big(\frac{1}{2\sigma^2}\Big(2\sigma^2 tx - (x-\mu)^2\Big)\Big)dx \\
&= \frac{1}{\sqrt{2\pi\sigma^2}} \int_{-\infty}^{\infty} exp\Big(\frac{1}{2\sigma^2}\Big(2\sigma^2 tx - x^2 + 2x\mu - \mu^2\Big)\Big)dx \\
&= \frac{1}{\sqrt{2\pi\sigma^2}} \int_{-\infty}^{\infty} exp\Big(\frac{1}{2\sigma^2}\Big(2\sigma^2 tx - x^2 + 2x\mu\Big)\Big) \cdot exp\Big(-\frac{\mu^2}{2\sigma^2}\Big)dx \\
&= \frac{1}{\sqrt{2\pi\sigma^2}} \cdot exp\Big(-\frac{\mu^2}{2\sigma^2}\Big) \int_{-\infty}^{\infty} exp\Big(\frac{1}{2\sigma^2}\Big(-x^2 + 2x(\sigma^2 t + \mu)^2\Big)\Big)dx \\
&= \frac{1}{\sqrt{2\pi\sigma^2}} \cdot exp\Big(-\frac{\mu^2}{2\sigma^2}\Big) \int_{-\infty}^{\infty} exp\Big(\frac{1}{2\sigma^2}\Big(-x^2 + 2x(\sigma^2 t + \mu) + (\sigma^2 t + \mu)^2 - (\sigma^2 t + \mu)^2\Big)\Big)dx \\
&= \frac{1}{\sqrt{2\pi\sigma^2}} \cdot exp\Big(-\frac{\mu^2}{2\sigma^2}\Big) \int_{-\infty}^{\infty} exp\Big(\frac{1}{2\sigma^2}\Big(-x^2 + 2x(\sigma^2 t + \mu) - (\sigma^2 t + \mu)^2\Big)\Big) \cdot exp\Big(\frac{(\sigma^2 t + \mu)^2}{2\sigma^2}\Big)dx \\
&= \frac{1}{\sqrt{2\pi\sigma^2}} \cdot exp\Big(-\frac{\mu^2}{2\sigma^2}\Big) \cdot exp\Big(\frac{(\sigma^2 t + \mu)^2}{2\sigma^2}\Big) \int_{-\infty}^{\infty} exp\Big(-\frac{1}{2\sigma^2}\Big(x^2 - 2x(\sigma^2 t + \mu) + (\sigma^2 t + \mu)^2\Big)\Big)dx \\
&= \frac{1}{\sqrt{2\pi\sigma^2}} \cdot exp\Big(\frac{-\mu^2 + (\sigma^2 t + \mu)^2}{2\sigma^2}\Big) \int_{-\infty}^{\infty} exp\Big(-\frac{1}{2\sigma^2}\Big((x - (\sigma^2 t + \mu))^2\Big)dx \\
&= \frac{1}{\sqrt{2\pi\sigma^2}} \cdot exp\Big(\frac{-\mu^2 + \sigma^4 t^2 + 2\sigma^2 t\mu + \mu^2}{2\sigma^2}\Big) \cdot (\sqrt{2\pi}\sigma) \\
&= exp\Big(\frac{\sigma^4 t^2 + 2\sigma^2 t\mu}{2\sigma^2}\Big) \\
&= exp\Big(\frac{2\sigma^2(\frac{1}{2}\sigma^2 t^2 + t\mu}{2\sigma^2}\Big) \\
&= exp\Big(\frac{1}{2}\sigma^2 t^2 + t\mu\Big)
\end{aligned}
$$

Observe that the derivative of $M_X(t)$ evaluated at 0 is actually the expected value of the normal distribution:

$$
\frac{dM_X(t)}{dt}\Big|_{t=0} = \frac{d}{dt} e^{(\frac{1}{2}\sigma^2 t^2 + t\mu)}\Big|_{t=0} = e^{(\frac{1}{2}\sigma^2 t^2 + t\mu)} \cdot (\sigma^2 t + \mu)\Big|_{t=0} = \mu
$$

- For a standard normal distribution:

  Since a standard normal distribution is a normal distribution for $\mu = 0$, and $\sigma = 1$ the moment generating function of a standard normal distribution is simply the one for a normal distribution with $\mu = 0$, and $\sigma = 1$:

$$
M_X(t) = e^{\frac{1}{2}t^2}
$$

Chapter 8

Statistical Inference

In the previous chapter we showed how given the probability distribution of a random variable, we know everything about it and we can compute probabilities, expected values, and many more. In a way this is the job of descriptive statistics. In this chapter we will show how we can draw conclusions for the population, when the population is not accessible. This is the job of inferential statistics.

8.1 Population VS Sample

Let's begin with some basic definitions.

Definition 8.1 (Population)**.** *A **population** is a set of similar items or events which is of interest for some question or experiment. A statistical population can be a group of existing objects or a hypothetical and potentially infinite group of objects conceived as a generalization from experience.*

Up to this point, technically we have been talking for populations. A population is the general category under examination. A r.v represents the population and the corresponding probability distribution tells us about the behaviour of the r.v and subsequently the behaviour of the population.

Definition 8.2 (Parameter)**.** *A **parameter** is a characteristic of a population.*

Some examples of parameters is the expected value (or mean), the variance and the standard deviation of the population.

However, more often than not, the population is not available to us either because gathering data for all the population is very expensive or in most of the cases because it is impossible. For example, if we assume that our population is men's weights and we want to know about the mean parameter (i.e the mean weight of all men), weighting all men around the globe at the same time is impossible.

Usually, we end up with a very small portion of the population called a "sample". (To not be confused with "sample space")

Definition 8.3 (Sample)**.** *A **sample** is a subset of a population. The elements of a sample are known as sample points, sampling units or observations.*

Similarly to the definition of a parameter:

Definition 8.4 (Statistic)**.** *A **statistic** is a characteristic of a sample.*

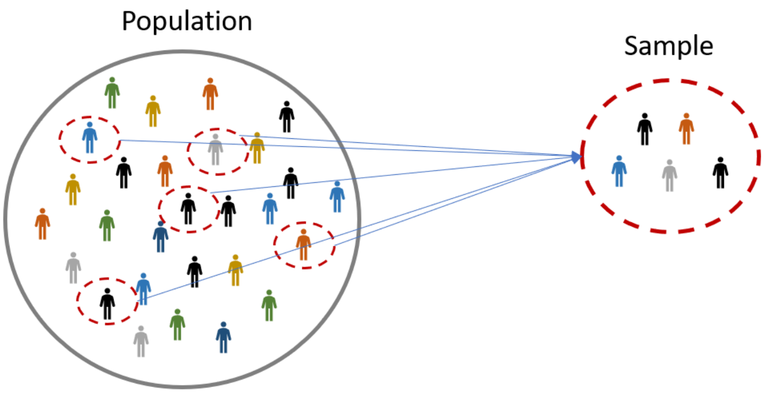
It follows from the definition of the sample that the latter is collected out of a population. In our previous example, one potential sample could be the weights of 1000 men.

Since samples are used to draw conclusion for the population one has to be extremely careful while collecting samples in order for the sample to be a good representative of the population. More specifically we must make sure that members of samples are randomly selected (each member of a population has an equal chance to be selected) and samples themselves are randomly selected (each sample of a population has an equal chance of being selected). Here are some definitions based on the way of collecting a sample.

Definition 8.5 (Random Sampling)**.** ***Random sampling*** *is a procedure for sampling from a population in which the selection of a sample unit is based on chance and every element of the population has a known, non-zero probability of being selected*

Definition 8.6 (Convenience Sampling)**.** ***Convenience sampling*** *is a type of non - probability sampling that involves the sample being drawn from that part of the population that is close to hand.*

Definition 8.7 (Stratified Sampling)**.** ***Stratified sampling*** *is a probability sampling technique that divides the entire population into different subgroups or strata, and then randomly selects the final subjects proportionally from the different strata.*



In mathematical terms, given a probability distribution P, a random sample of length n is a set of realizations of n independent, identically distributed random variables (i.i.d r.v's) with distribution P. A sample concretely represents the results of n experiments in which the same quantity is measured. In our example, if we want to estimate the mean weight of members of a particular population, we measure the heights of n individuals. Each measurement is drawn from the probability distribution P characterizing the population, so each measured weight x_i is the realization of a r.v X with distribution P. Hence mathematically a sample can be represented as $\{x_1, x_2, \ldots, x_n\}$. Given this representation we can define some of the characteristics (statistics) of a sample.

Definition 8.8 (Sample Size)**.** *Given a sample of realizations of of n i.i.d r.v's $\{x_1, x_2, \ldots, x_n\}$, we call* ***sample size*** *the direct count of the number of samples measured or observations being made n.*

Definition 8.9 (Sample Mean)**.** *Given a sample of realizations of n i.i.d r.v's $\{x_1, x_2, \ldots, x_n\}$, we define the* ***sample mean*** *or average $\bar{x}$ of the sample as the quantity:*

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

Definition 8.10 (Sample Variance)**.** *Given a sample of realizations of n i.i.d r.v's $\{x_1, x_2, \ldots, x_n\}$, we define the* ***sample variance*** *s^2 of the sample as the quantity:*

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

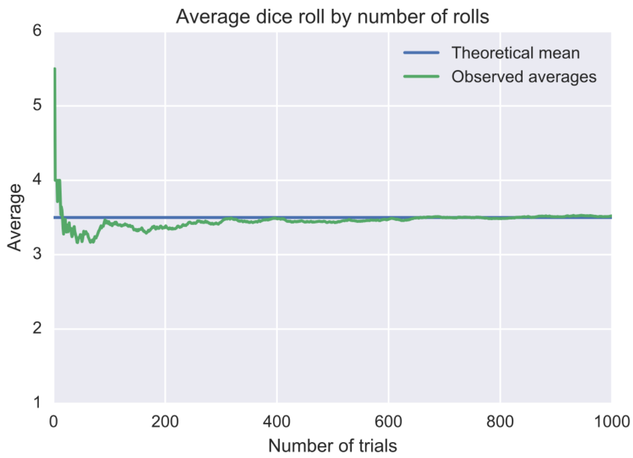
Definition 8.11 (Sample Standard Deviation)**.** *Given a sample of realizations of n i.i.d r.v's $\{x_1, x_2, \ldots, x_n\}$, we define the* ***sample standard deviation*** *s of the sample as the quantity:*

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

Theorem 8.1 (Law Of Large Numbers)**.** *Given a random variable X that follows a probability distribution P with mean μ and a collection of n realizations of X $\{x_1, x_2, \ldots, x_n\}$ with an average $\bar{x}_n$ then:*

$$\bar{x}_n \to \mu \quad for \quad n \to \infty$$

The law of large numbers is a theorem that describes the result of performing the same experiment a large number of times. According to the law, the average of the results obtained from a large number of trials should be close to the actual expected value of the population, and will tend to become closer to the expected value as more trials are performed.The law of large numbers is important because it guarantees stable long-term results for the averages of some random events.



The law of large numbers can find an application in statistics since, the "random variable X that follows a probability distribution P with mean μ" can be translated to a population, and the "collection of n realizations of X $\{x_1, x_2, \ldots, x_n\}$ with an average $\bar{x}_n$" can be translated to a sample of size n drawn out of the population. Then the law of large numbers simply states that as the size of the sample increases the sample mean tends to the actual population mean. Going back to our example, this means that as the number of men in our sample increases their average weight tends to the actual mean weight of the whole population.

Chapter 9

Parametric Inference

In this chapter we will introduce another way of inference called "parametric inference". In parametric inference the goal is given a sample of n realizations of i.i.d r.v's to learn the underlying distribution P of X. Let's begin with some basic definitions.

9.1 Basic Definitions

Definition 9.1 (Statistical Model). *Let the observed outcome of a random experiment be a sample* $\{X_1, X_2, \ldots, X_n\}$ *of* n *i.i.d r.v's in some measurable space* $E \subseteq R$*, and* P *their common distribution. A* ***statistical model*** *associated to that statistical experiment is the tuple* $(E, (P_\theta)_{\theta \in \Theta})$ *where:*

- E*: Sample space where* X *lives.*
- P_θ*: Family of probability measures.*
- Θ*: Parameter set (usually* $\Theta = R^d$*)*

Let's see some examples of statistical models.

- Bernoulli Statistical Model: $(\{0,1\}, \quad (Bern(p))_{p\in[0,1]})$
- Exponential Statistical Model: $((0,\infty), \quad (Expo(\lambda))_{\lambda\in[0,\infty)})$
- Poisson Statistical Model: $(N, \quad (Pois(\lambda))_{\lambda\in[0,\infty)})$
- Gaussian Statistical Model: $(R, \quad (N(\mu,\sigma^2))_{\mu\in R,\ \sigma^2\in(0,\infty)})$
- Uniform Statistical Model: $([0,\infty), \quad (Unif(a,b))_{a\in[0,\infty),\ n\in[0,\infty)})$

Intuitively, given a set of observed outcomes $\{X_1, X_2, \ldots, X_n\}$ we will assume that they follow some broad family of probability measures P_θ parametrized by some parameter θ (e.g: a Bernoulli distribution Bern(p) where $\theta = p$). Our goal is based on the given sample to specify this parameter and subsequently the underlying distribution.

Definition 9.2 (True Parameter). *We always make the assumption that* $\exists\, \theta^* \in \Theta : X \sim P_\theta$*. We call this specific* θ^* *the* ***true parameter****.*

Definition 9.3 (Estimator). *An* ***estimator*** $\hat{\theta}_n$ *is a function that maps the sample space to a set of sample estimates.*

$$\hat{\theta}_n = \hat{\theta}_n(X_1, X_2, \ldots, X_n)$$

Notice that since $\{X_1, X_2, \ldots, X_n\}$ are r.v's subsequently $\hat{\theta}_n$ is also a r.v since it is a function of r.v's. The true parameter on the other hand is a deterministic real number.

From the law of large numbers, we get that for the estimator holds:

$$\lim_{n\to\infty} \hat{\theta}_n = \theta$$

Given the true parameter and the estimator, we can define some measures of error as follows.

Definition 9.4 (Error)**.** *Given a true parameter θ and an estimator $\hat{\theta}_n$, we define the* ***error*** *e of the estimator $\hat{\theta}_n$ as:*

$$e = \hat{\theta}_n - \theta$$

Definition 9.5 (Mean Squared Error)**.** *Given a true parameter θ and an estimator $\hat{\theta}_n$, we define the* ***mean squared error*** *(MSE) of the estimator $\hat{\theta}_n$ as:*

$$MSE = E[(\hat{\theta}_n - \theta)^2]$$

Definition 9.6 (Efficient Estimator)**.** *An estimator $\hat{\theta}_n$ is called* ***efficient*** *if MSE is sufficient small.*

Definition 9.7 (Consistent Estimator)**.** *An estimator $\hat{\theta}_n$ is called* ***consistent*** *if:*

$$\lim_{n \to \infty} MSE = 0$$

Definition 9.8 (Sampling Deviation)**.** *Given a true parameter θ and an estimator $\hat{\theta}_n$, we define the* ***sampling deviation*** *d of the estimator $\hat{\theta}_n$ as:*

$$d = \hat{\theta}_n - E[\hat{\theta}_n]$$

Definition 9.9 (Bias)**.** *Given a true parameter θ and an estimator $\hat{\theta}_n$, we define the* ***bias*** *B of the estimator $\hat{\theta}_n$ as:*

$$B = E[\hat{\theta}_n - \theta] = E[\hat{\theta}_n] - E[\theta] = E[\hat{\theta}_n] - \theta$$

Definition 9.10 (Unbiased Estimator)**.** *An estimator $\hat{\theta}_n$ is called* ***unbiased*** *if $B = 0$.*

Definition 9.11 (Variance)**.** *Given a true parameter θ and an estimator $\hat{\theta}_n$, we define the* ***variance*** *Var of the estimator $\hat{\theta}_n$ as:*

$$Var = E[(\hat{\theta}_n - E[\hat{\theta}_n])^2]$$

By manipulating MSE we can show:

$$\begin{aligned}
MSE &= E[(\hat{\theta}_n - \theta)^2] \\
&= E[\hat{\theta}_n^2 - 2\hat{\theta}_n\theta + \theta^2] \\
&= E[\hat{\theta}_n^2] - E[2\hat{\theta}_n\theta] + E[\theta^2] \\
&= E[\hat{\theta}_n^2] - 2\theta E[\hat{\theta}_n] + \theta^2 \\
&= E[\hat{\theta}_n^2] - 2\theta E[\hat{\theta}_n] + \theta^2 + E^2[\hat{\theta}_n] - E^2[\hat{\theta}_n] \\
&= (E[\hat{\theta}_n^2] - E^2[\hat{\theta}_n]) + (E^2[\hat{\theta}_n] - 2\theta E[\hat{\theta}_n] + \theta^2) \\
&= (E[\hat{\theta}_n^2] - E^2[\hat{\theta}_n]) + (E[\hat{\theta}_n] - \theta)^2 \\
&= Var + B^2
\end{aligned}$$

Hence we showed that the MSE of an estimator can actually be split into the variance and the bias of the estimator.

9.2 Maximum Likelihood

As we explained in the previous section our goal in parametric inference is given a statistical model (E, P_θ) associated with a sample of i.i.d r.v's $\{X_1, X_2, \dots, X_n\}$, by making the assumption that there always exists a true parameter θ^* such that $X \sim P_{\theta^*}$, to find this true parameter.

Our initial step is to define an estimator $\hat{\theta}_n$, that subsequently defines a probability distribution $P_{\hat{\theta}_n}$. Hence now we have two quantities: the probability distribution we actually want to find P_{θ^*} and the probability distribution that we begin with $P_{\hat{\theta}_n}$. Since the difference of these two is what we want to minimize, it makes sense to define the absolute distance between them as follows:

Definition 9.12 (Total Variation Distance)**.** *The* ***total variation distance*** *TV of two probability distributions P_θ and $P_{\theta'}$ is defined as the largest possible difference between the probabilities that the two probability distributions can assign to the same event.*

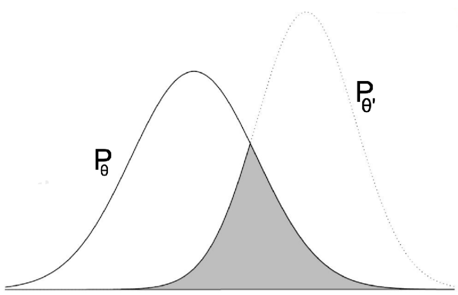
$$TV(P_\theta, P'_\theta) = \max_{A \subseteq E} |P_\theta(A) - P'_\theta(A)|$$

The total variation distance satisfies the following properties:

- $TV(P_\theta, P'_\theta) = TV(P'_\theta, P_\theta)$
- $TV(P_\theta, P'_\theta) \geq 0$
- $TV(P_\theta, P'_\theta) = 0 \Rightarrow P_\theta = P'_\theta$
- $TV(P_\theta, P'_\theta) \leq TV(P_\theta, P''_\theta) + TV(P_{\theta''}, P'_\theta)$

These properties imply that total variation distance is indeed a distance measure between probability distributions (hence the name).

Due to the "max" term that appears in total variation distance it's very hard to manipulate it. Fortunately, there is an alternative way of describing the same concept. As we see in the figure below, since total variation distance is just the absolute difference of two probabilities it can also be expressed as the two times the area under the curve that it is not common in the two distributions. (The white area in the graph)



Moreover, by defining the area A^* as the area in which $P_\theta \geq P'_\theta$, we can get rid of the max in total variation distance as:

$$\begin{aligned}
TV(P_\theta, P'_\theta) &= \max_{A \subseteq E} |P_\theta(A) - P'_\theta(A)| \\
&= |P_\theta(A^*) - P'_\theta(A^*)| \\
&= P_\theta(A^*) - P'_\theta(A^*) \\
&= \int_{A^*} (p_\theta(x) - p'_\theta(x))dx \\
&= \frac{1}{2}\int_{A^*} (p_\theta(x) - p'_\theta(x))dx + \frac{1}{2}\int_{(A^*)^C} (p'_\theta(x) - p_\theta(x))dx \\
&= \frac{1}{2}\int_{E} |p_\theta(x) - p'_\theta(x)|dx
\end{aligned}$$

Hence we showed that:

$$TV(P_\theta, P'_\theta) = \frac{1}{2}\int_E |p_\theta(x) - p'_\theta(x)|dx$$

and similarly for a discrete distribution:

$$TV(P_\theta, P'_\theta) = \frac{1}{2}\sum_{x\in E} |p_\theta(x) - p'_\theta(x)|$$

Coming back to our case if we set $\theta = \hat{\theta}$ to be our estimator and $\theta' = \theta^*$ to be our true parameter, we have an equation for the total variation distance that includes something that we can work with (i.e PDF's). And since, as we argued, total variation distance is just the distance between the two probabilities, our goal is to minimize it as much as possible. That way the distance between our estimator probability and true probability will be as low as it can be, hence very the estimation will be very close to the real distribution.

The problem with total variation distance as it is, is that it carries an absolute value which makes it impossible to minimize it. For this reason we are going to introduce another quantity on top of total variation distance called "Kullback - Leibler divergence".

Definition 9.13 (Kullback - Leibler Divergence)**.** *The **Kullback - Leibler divergence** KL of two probability distributions P_θ and $P_{\theta'}$ is defined in term of their PDF's as follows:*

$$KL(P_\theta, P'_\theta) = \int_E p_\theta(x)\ln\frac{p_\theta}{p'_\theta(x)}dx$$

The Kullback - Leibler divergence satisfies the following properties:

- $KL(P_\theta, P'_\theta) \neq KL(P'_\theta, P_\theta)$
- $KL(P_\theta, P'_\theta) \geq 0$
- $KL(P_\theta, P'_\theta) = 0 \Rightarrow P_\theta = P'_\theta$
- $KL(P_\theta, P'_\theta) \nleq KL(P_\theta, P''_\theta) + KL(P_{\theta''}, P'_\theta)$

These properties imply that Kullback - Leibler divergence is not a distance measure between probability distributions (hence the name divergence).

By manipulating the Kullback - Leibler divergence we get:

$$KL(P_\theta, P'_\theta) = \int_E p_\theta(x)\ln\frac{p_\theta}{p'_\theta(x)}dx = E_\theta[\ln\frac{p_\theta}{p'_\theta}] = E_\theta[\ln p_\theta - \ln p'_\theta] = E_\theta[\ln p_\theta] - E_\theta[\ln p'_\theta]$$

Coming back to our case if we set $\theta = \theta^*$ to be our estimator and $\theta' = \hat{\theta}$ to be our true parameter, then by using the manipulated the Kullback - Leibler divergence we have:

$$KL(P_{\theta^*}, P_{\hat{\theta}}) = E_{\theta^*}[\ln p_{\theta^*}] - E_{\theta^*}[\ln p_{\hat{\theta}}] = c - E_{\theta^*}[\ln p_{\hat{\theta}}]$$

As we said, Kullback - Leibler divergence is not a distance measure however it still is a good quantity to minimize in order to find a good estimation for the true probability distribution. Hence the estimator is simply the argument that minimizes Kullback - Leibler divergence:

$$\hat{\theta} = \arg\min_\theta\Big(KL(P_{\theta^*}, P_\theta)\Big) = \arg\min_\theta\Big(c - E_{\theta^*}[\ln p_\theta]\Big) = \arg\min_\theta(c) - \arg\min_\theta\Big(E_{\theta^*}[\ln p_\theta]\Big)$$

The term $\arg\min(c)$ is just a constant that does not depend on θ. Same θ that minimizes $f(\theta)$ minimizes also $c - f(\theta)$. So we can just drop it from the equation and get:

$$\hat{\theta} = -\arg\min_\theta\Big(E_{\theta^*}[\ln p_\theta]\Big)$$

As we discussed previously we can approximate the expected value with sample mean $E_{\theta^*} \to \frac{1}{n}\sum_i$:

$$\begin{aligned}
\hat{\theta} &= -\arg\min_{\theta} \Big(\frac{1}{n} \sum_{i=1}^{n} \ln p_\theta(x_i) \Big) \\
&= -\arg\min_{\theta} \Big(\sum_{i=1}^{n} \ln p_\theta(x_i) \Big) \\
&= \arg\max_{\theta} \Big(\sum_{i=1}^{n} \ln p_\theta(x_i) \Big) \\
&= \arg\max_{\theta} \Big(\ln \prod_{i=1}^{n} p_\theta(x_i) \Big) \\
&= \arg\max_{\theta} \Big(\prod_{i=1}^{n} p_\theta(x_i) \Big)
\end{aligned}$$

Based on this last two equation we define the concept of "likelihood" as follows:

Definition 9.14 (Likelihood)**.** *Let X be a random variable following a probability distribution with probability density function $f_\theta(x)$ depending on a parameter θ . Then the **likelihood** $\mathcal{L}(\theta \mid x)$ is formed from the joint probability of a sample of data of X:*

$$\mathcal{L}(\theta \mid x) = \prod_{i=1}^{n} f_\theta(x_i)$$

More often than not, it is more handy to use the logarithm of the likelihood which we define as the "log-likelihood".

Definition 9.15 (Log-likelihood)**.** *The **log-likelihood** $l(\theta \mid x)$ is simply the logarithm of the likelihood:*

$$l(\theta \mid x) = \ln \mathcal{L}(\theta \mid x)$$

So we can find the estimator $\hat{\theta}$ by maximizing the likelihood or the log-likelihood, i.e:

$$\frac{d\mathcal{L}}{d\theta}\Big|_{\theta=0} = 0 \qquad \text{or} \qquad \frac{dl}{d\theta}\Big|_{\theta=0}$$

The solution of either of these equations gives back the best estimator. This process is called "the principle of maximum likelihood" or simply "the maximum likelihood method".

Application: Maximum Likelihood In Bernoulli Distribution

Let us have a collection of i.i.d r.v's $\{X_1, X_2, \ldots, X_n\}$ following a Bernoulli distribution Bern(p). Our goal is based on the sample $\{X_1, X_2, \ldots, X_n\}$ to estimate the value of the parameter p.

We start by forming the likelihood:

$$\mathcal{L} = \prod_{i=1}^{n} P(x_i) = \prod_{i=1}^{n} p^{x_i}(1-p)^{1-x_i}$$

Subsequently for the log-likelihood:

$$l = \ln \mathcal{L} = \ln \prod_{i=1}^{n} p^{x_i}(1-p)^{1-x_i}$$

Now we can manipulate the log-likelihood to obtain a more handy expression:

$$
\begin{aligned}
l &= \ln \prod_{i=1}^{n} p^{x_i}(1-p)^{1-x_i} \\
&= \sum_{i=1}^{n} \Big[\ln(p^{x_i}(1-p)^{1-x_i}) \Big] \\
&= \sum_{i=1}^{n} \Big[\ln p^{x_i} + \ln(1-p)^{1-x_i} \Big] \\
&= \sum_{i=1}^{n} \Big[x_i \cdot \ln p + (1-x_i) \cdot \ln(1-p) \Big] \\
&= \sum_{i=1}^{n} \Big[x_i \cdot \ln p \Big] + \sum_{i=1}^{n} \Big[(1-x_i) \cdot \ln(1-p) \Big] \\
&= \ln p \cdot \sum_{i=1}^{n} \Big[x_i \Big] + \ln(1-p) \cdot \sum_{i=1}^{n} \Big[(1-x_i) \Big] \\
&= n \cdot \ln p \cdot \sum_{i=1}^{n} \Big[\frac{1}{n} x_i \Big] + n \cdot \ln(1-p) \cdot \sum_{i=1}^{n} \Big[\frac{1}{n}(1-x_i) \Big] \\
&= n \cdot \ln p \cdot \bar{x} + n \cdot \ln(1-p) \cdot (1-\bar{x}) \\
&= n(\bar{x} \ln p + (1-\bar{x}) \ln(1-p))
\end{aligned}
$$

For the derivative of the log-likelihood we obtain:

$$
\frac{dl}{dp} = \frac{d}{dp}\Big(n(\bar{x} \ln p + (1-\bar{x}) \ln(1-p)) \Big) = n\Big(\frac{\bar{x}}{p} - \frac{1-\bar{x}}{1-p} \Big) = \ldots = n(\bar{x} - p)
$$

Finally, from principle of maximum likelihood, we can obtain the best estimator by setting the derivative to zero:

$$
\frac{dl}{dp} = 0 \Rightarrow p = \bar{x}
$$

Hence, we proved that the estimator that maximizes the likelihood for a Bernoulli distribution is actually the average of the sample. In a similar way we can show the same for all the distributions we have introduced.

Now let's compute some of the characteristics of this estimator.

For the error of the estimator:

$$
e = \hat{\theta}_n - \theta = \bar{x} - p
$$

For the sampling deviation:

$$
d = \hat{\theta}_n - E[\hat{\theta}_n] = \bar{x} - E[\bar{x}] = \bar{x} - E\Big[\frac{1}{n}\sum_{i=1}^{n} x_i\Big] = \bar{x} - \frac{1}{n}\sum_{i=1}^{n} E[x_i] = \bar{x} - \frac{1}{n}\sum_{i=1}^{n} p = \bar{x} - \frac{1}{n}np = \bar{x} - p
$$

For the bias:

$$
E[\hat{\theta}_n] - \theta = E[\bar{x}] - p = p - p = 0
$$

Hence the estimator $\bar{x}$ is unbiased.

For the variance:

$$Var(\theta) = Var(\bar{x}) = Var(\frac{1}{n}\sum_{i=1}^{n} x_i) = \frac{1}{n^2}\sum_{i=1}^{n} Var(x_i) = \frac{1}{n^2}\sum_{i=1}^{n} p(1-p) = \frac{p(1-p)}{n}$$

For the mean Squared Error:

$$MSE = Var + B^2 = \frac{p(1-p)}{n} + 0^2 = \frac{p(1-p)}{n}$$

Finally observe that since $MSE \propto \frac{1}{n}$ we have that as $n \to \infty$, $MSE \to 0$, so the estimator $\bar{x}$ is consistent.

Part III

Machine Learning

Chapter 10

Introduction

Definition 10.1 (Machine Learning)**.** ***Machine learning*** *is the field of study that gives computers the ability to learn without being explicitly programmed. (Arthur Samuel, 1969)*

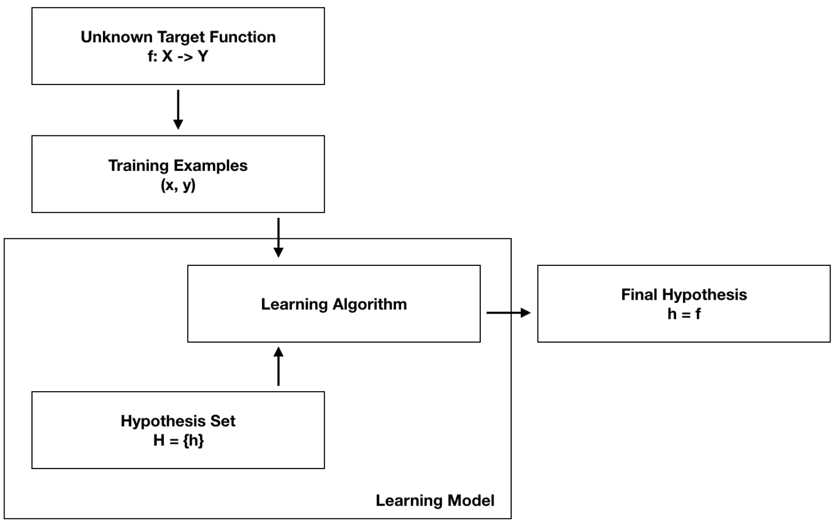
Definition 10.2 (Machine Learning)**.** *A computer program is said to learn from experience E with respect to some task T and some performance measure P, if its performance on T, as measured by P, improves with experience E. (Tom Mitchell)*

The essence of machine learning is that a pattern exists and it can not be pined down mathematically, however we have data on it and we can treat it in a probabilistic way.

In order to formalize it we will be using the following notation throughout the notes:

- Input: $x \in X$
- Output: $y \in Y$
- Data: $\{x_i, y_i\}, \quad i = 1, 2, 3, \ldots, m$
- Target Function: $f : X \to Y$
- Hypothesis Function: $h : X \to Y$ with $h \approx f$
- Hypothesis Set: $H = \{h\}$

Informally, the goal of machine learning is, based on the data $\{x_i, y_i\}$, to discover a hypothesis function h that behaves in a similar way with the target function f which is, and always will be, unknown to us.



The question is how can we learn an unknown function f just based on the data we already have, when the unknown function f in general can take any value outside the known data. The short answer is that we can not however, without proving it, the following relation holds:

$$P\Big[|E_{\text{in}}(h) - E_{\text{out}}(h)| > \epsilon\Big] \leq 2 \cdot M \cdot e^{2\epsilon^2 m}$$

where $E_{\text{in}}(h)$ is the error that we get for h in the known data, $E_{\text{out}}(h)$ is the error that we will get when we use h for new data, M is the number of possible hypothesis function h (i.e the cardinality of the hypothesis set $H = \{h\}$, ϵ is the tolerance that we have for errors, and m is the number of data points. This equation tells us that no matter what, learning is possible only in a probabilist sense. We will always have an error, since the whole process carries a stochastic nature.

So we can informally summarize what we are trying to do with machine learning as:

- From aforementioned relation: $E_{\text{in}} \approx E_{\text{out}}$
- From learning algorithm: $E_{\text{in}} \approx 0$
- From the combination of these 2: $E_{\text{out}} \approx 0$

By having $E_{\text{out}} \approx 0$, that means that our hypothesis function h generalizes well for out of sample data, so we can use it for predictions. That in a nutshell is how machine learning works.

Machine learning is a very broad topic with many different branches and applications. In these notes we will cover the most basic branches which are:

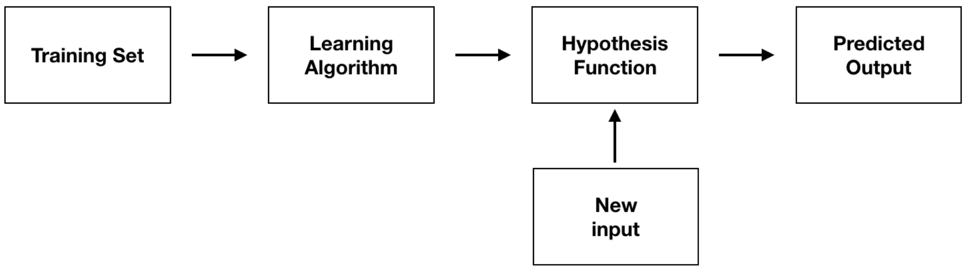
1. Supervised Learning
2. Unsupervised Learning
3. Reinforcement Learning
4. Deep Learning

Chapter 11

Supervised Learning

Supervised learning is one of the four basic categories of machine learning, and it consists of a family of models and techniques that we will introduce in this chapter. First let's start with a formal definition of supervised learning.

Definition 11.1 (Supervised Learning)**.** ***Supervised learning*** *is the machine learning task of learning a function that maps an input to an output based on examples of "input - output" pairs called a "training set".*



Some more specific notation that we will be using throughout supervised learning:

- Input variables or attributes or features: x
- The i'th feature (in case of many features): x_i
- The i'th training example (in case of many training examples): $x^{(i)}$
- The i'th feature of the j'th training example: $x_i^{(j)}$
- Output variables or targets or classes: y
- The i'th output target: $y^{(i)}$
- Total number of training examples: m
- Total number of features: n

Now let's dive in the models and techniques of supervised learning, starting with one of the most basic ones called "linear regression".

11.1 Linear Regression

Definition 11.2 (Linear Regression)**.** ***Linear regression*** *is a linear approach to modelling the relationship between a dependent variable (target) and one or more independent variables (features).*

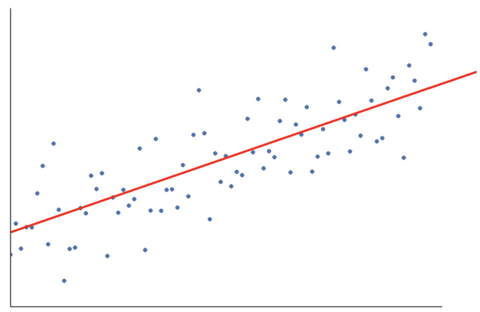
Definition 11.3 (Simple & Multiple Linear Regression)**.** *When there is only one independent variable then the model is called* ***simple linear regression****. For more than one independent variable, the process is called* ***multiple linear regression****.*

Definition 11.4 (Univariate & Multivariate Linear Regression)**.** *When only one dependent variable is predicted then the model is called* ***univariate linear regression****. For more than one correlated, dependent variables being predicted, the process is called* ***multivariate linear regression****.*

In linear regression the hypothesis function h is a linear combinations of the features:

$$h(x) = w_0 + w_1 x_1 + \ldots + w_n x_n$$

where w_0 is is called "bias" and w_i's are called "weights". We usually refer to weights and bias as the "parameters" of the regression and they are the ones that we try to determine through the training examples by using a learning algorithm. Once we find them then h is ready to predict new inputs with unknown outcomes.



It is a usual procedure to define $x_0 = 1$, so linear regression the hypothesis function can be rewritten as

$$h(x) = w_0 x_0 + w_1 x_1 + \ldots + w_n x_n = \sum_{i=0}^{n} w_i x_i$$

Moreover by defining the feature vector $\boldsymbol{x}$ and parameter vector $\boldsymbol{w}$ as

$$\boldsymbol{x} = \begin{bmatrix} x_0 \\ x_1 \\ \vdots \\ x_n \end{bmatrix}, \qquad \boldsymbol{w} = \begin{bmatrix} w_0 \\ w_1 \\ \vdots \\ w_n \end{bmatrix}$$

then we can rewrite the linear regression the hypothesis function in a very simple form of:

$$h(x) = \boldsymbol{w}^T \boldsymbol{x}$$

Now that we have a hypothesis function, we need a rule in order to be able to find the parameters $\boldsymbol{w}$. This rule can be obtained through the probabilistic interpretation of linear regression.

More precisely, after having obtained the parameters $\boldsymbol{w}$, the hypothesis will fit the data in the best possible way but, since as we said we are dealing with probabilistic systems, we will still have some errors ϵ. In other words, for each training example the following formula will apply:

$$y^{(i)} = h(x^{(i)}) + \epsilon^{(i)} = \boldsymbol{w}^T \boldsymbol{x}^{(i)} + \epsilon^{(i)}$$

where $\boldsymbol{x}^{(i)}$ is the corresponding training example feature vector:

$$\boldsymbol{x}^{(i)} = \begin{bmatrix} x_0^{(i)} \\ x_1^{(i)} \\ \vdots \\ x_n^{(i)} \end{bmatrix}$$

At this point we will make one assumption which needs to be valid in order for the linear regression to be valid. Namely, we assume that the errors $\epsilon^{(i)}$ are independent and identically distributed following a normal distribution with mean 0 and variance σ^2:

$$\epsilon^{(i)} \sim N(0, \sigma^2)$$

which means that the probability distribution of the errors is given by

$$P(\epsilon^{(i)}) = \frac{1}{\sqrt{2\pi\sigma^2}} exp\Big(- \frac{(\epsilon^{(i)})^2}{2\sigma^2} \Big)$$

From the assumption of the errors follows

$$y^{(i)} = \boldsymbol{w}^T \boldsymbol{x}^{(i)} + \epsilon^{(i)} \Rightarrow$$
$$\epsilon^{(i)} = y^{(i)} - \boldsymbol{w}^T \boldsymbol{x}^{(i)}$$

By substituting the error back to the probability

$$P(\epsilon^{(i)}) = \frac{1}{\sqrt{2\pi\sigma^2}} \cdot exp\Big(- \frac{(\epsilon^{(i)})^2}{2\sigma^2} \Big) \Rightarrow$$

$$P(y^{(i)}|\boldsymbol{x}^{(i)}; \boldsymbol{w}) = \frac{1}{\sqrt{2\pi\sigma^2}} \cdot exp\Big(- \frac{(y^{(i)} - \boldsymbol{w}^T \boldsymbol{x}^{(i)})^2}{2\sigma^2} \Big)$$

In other words we get that the conditional distribution of $y^{(i)}$ given $\boldsymbol{x}^{(i)}$ and $\boldsymbol{w}$ is a normal distribution with mean $\boldsymbol{w}^T \boldsymbol{x}^{(i)}$ and variance σ^2

$$y^{(i)} \sim N(\boldsymbol{w}^T \boldsymbol{x}^{(i)}, \sigma^2)$$

Given the probability distribution of $y^{(i)}$ as a function of the parameters, we can now use the principle of maximum likelihood that we developed in parametric inference chapter, in order to find the rule that will give us the best parameters $\boldsymbol{w}$.

For the likelihood it is

$$\mathcal{L}(\boldsymbol{w}|y) = P(y^{(1)}, y^{(2)}, \dots, y^{(m)}|\boldsymbol{x}^{(1)}, \boldsymbol{x}^{(2)}, \dots, \boldsymbol{x}^{(m)}; \boldsymbol{w}) = \prod_{i=1}^{m} P(y^{(i)}|\boldsymbol{x}^{(i)}; \boldsymbol{w})$$

where we used the fact that $\epsilon^{(i)}$ are independent. By substituting the probability

$$\mathcal{L}(\boldsymbol{w}|y) = \prod_{i=1}^{m} \frac{1}{\sqrt{2\pi\sigma^2}} \cdot exp\Big(- \frac{(y^{(i)} - \boldsymbol{w}^T \boldsymbol{x}^{(i)})^2}{2\sigma^2} \Big)$$

Subsequently for the log-likelihood

$$
\begin{aligned}
l(\boldsymbol{w}|y) &= \ln \mathcal{L}(\boldsymbol{w}|y) \\
&= \ln \Big[\prod_{i=1}^{m} \frac{1}{\sqrt{2\pi\sigma^2}} \cdot exp\Big(- \frac{(y^{(i)} - \boldsymbol{w}^T \boldsymbol{x}^{(i)})^2}{2\sigma^2} \Big) \Big] \\
&= \sum_{i=1}^{m} \ln \Big[\frac{1}{\sqrt{2\pi\sigma^2}} \cdot exp\Big(- \frac{(y^{(i)} - \boldsymbol{w}^T \boldsymbol{x}^{(i)})^2}{2\sigma^2} \Big) \Big] \\
&= \sum_{i=1}^{m} \ln \Big[\frac{1}{\sqrt{2\pi\sigma^2}} \Big] + \sum_{i=1}^{m} \ln \Big[exp\Big(- \frac{(y^{(i)} - \boldsymbol{w}^T \boldsymbol{x}^{(i)})^2}{2\sigma^2} \Big] \Big) \\
&= \sum_{i=1}^{m} \ln \Big[\frac{1}{\sqrt{2\pi\sigma^2}} \Big] + \sum_{i=1}^{m} \Big[- \frac{(y^{(i)} - \boldsymbol{w}^T \boldsymbol{x}^{(i)})^2}{2\sigma^2} \Big]
\end{aligned}
$$

According to the principle of maximum likelihood, the best parameters can be found by maximizing the log-likelihood. The first term of the log-likelihood is just a constant term so it does not contribute at all to the maximization, and the same holds for the denominator of the second term. Hence:

$$
\begin{aligned}
\boldsymbol{w} &= \arg\max_{\boldsymbol{w}} [l(\boldsymbol{w}|y)] \\
&= \arg\max_{\boldsymbol{w}} \Big[\sum_{i=1}^{m} (-(y^{(i)} - \boldsymbol{w}^T \boldsymbol{x}^{(i)})^2) \Big] \\
&= \arg\max_{\boldsymbol{w}} \Big[- \sum_{i=1}^{m} (y^{(i)} - \boldsymbol{w}^T \boldsymbol{x}^{(i)})^2) \Big] \\
&= \arg\min_{\boldsymbol{w}} \Big[\sum_{i=1}^{m} (y^{(i)} - \boldsymbol{w}^T \boldsymbol{x}^{(i)})^2 \Big]
\end{aligned}
$$

At this point we can formally define the following function:

Definition 11.5 (Mean Squared Error Loss Function)**.** ***Mean squared error loss function*** *(MSE) $J(\boldsymbol{w})$ is defined as:*

$$
J(\boldsymbol{w}) = \frac{1}{2m} \sum_{i=1}^{m} (y^{(i)} - \boldsymbol{w}^T \boldsymbol{x}^{(i)})^2
$$

Hence, the principle of maximum likelihood translates to finding the parameters $\boldsymbol{w}$ that minimize the MSE loss function. The intuition behind the minimization of the MSE loss function is straight forward since what we are actually doing is minimizing the square of the errors between the prediction and the actual outcome (square because only the magnitude of the error is important and not the sign). By minimizing as much as possible the errors we will eventually get the best line that fits the data.

As a final note, we can group together all training examples in one matrix and all training labels in one vector as follows:

$$
X = \begin{bmatrix} x_0^{(1)} & x_1^{(1)} & \dots & x_n^{(1)} \\ x_0^{(2)} & x_1^{(2)} & \dots & x_n^{(2)} \\ \vdots & \vdots & \ddots & \dots \\ x_0^{(m)} & x_1^{(m)} & \dots & x_n^{(m)} \end{bmatrix}, \qquad \boldsymbol{y} = \begin{bmatrix} y^{(1)} \\ y^{(2)} \\ \vdots \\ y^{(m)} \end{bmatrix}
$$

By doing so then we can write the MSE loss function in the simple form of:

$$J(\boldsymbol{w}) = \frac{1}{2m}(X\boldsymbol{w} - \boldsymbol{y})^2 = \frac{1}{2m}(X\boldsymbol{w} - \boldsymbol{y})^T(X\boldsymbol{w} - \boldsymbol{y})$$

11.2 Optimization Techniques

Given the MSE loss function (or any other loss function that we will introduce later on), the goal of machine learning is to optimize it (usually minimize it) in order to obtain the best parameters that fit the data. Optimizing loss functions is one of the biggest parts of machine learning and we can do so with the so called "optimization techniques".

Definition 11.6 (Optimization Techniques)**.** ***Optimization techniques*** *are techniques used for finding the optimum solution or unconstrained maxima or minima of continuous and differentiable functions. These are analytical methods and make use of differential calculus in locating the optimal solution.*

11.2.1 Normal Equation

Probably the most straight forward optimization technique is the so called "normal equation". Since we are looking a minimum for $J(\boldsymbol{w})$ the natural thing to do, is to simply calculate the derivative with respect to the parameter vector and then set it to zero (as we did when we introduced the principle of maximum likelihood).

It is more handy to use the vector form of loss function, so for the derivative we get:

$$\begin{aligned}
\nabla_{\boldsymbol{w}} J(\boldsymbol{w}) &= \frac{1}{2m}\nabla_{\boldsymbol{w}}\Big[(X\boldsymbol{w} - \boldsymbol{y})^T(X\boldsymbol{w} - \boldsymbol{y})\Big] \\
&= \frac{1}{2m}\nabla_{\boldsymbol{w}}\Big[\Big((X\boldsymbol{w})^T - \boldsymbol{y}^T\Big)\Big(X\boldsymbol{w} - \boldsymbol{y}\Big)\Big] \\
&= \frac{1}{2m}\nabla_{\boldsymbol{w}}\Big[(X\boldsymbol{w})^T(X\boldsymbol{w}) - (X\boldsymbol{w})^T\boldsymbol{y} - \boldsymbol{y}^T(X\boldsymbol{w}) + \boldsymbol{y}^T\boldsymbol{y}\Big] \\
&= \frac{1}{2m}\nabla_{\boldsymbol{w}}\Big[(X\boldsymbol{w})^T(X\boldsymbol{w}) - 2(X\boldsymbol{w})^T\boldsymbol{y} + \boldsymbol{y}^T\boldsymbol{y}\Big] \\
&= \frac{1}{2m}\nabla_{\boldsymbol{w}}\Big[\boldsymbol{w}^T X^T X\boldsymbol{w} - 2\boldsymbol{w}^T X^T\boldsymbol{y} + \boldsymbol{y}^T\boldsymbol{y}\Big] \\
&= \frac{1}{2m}\Big[2X^T X\boldsymbol{w} - 2X^T\boldsymbol{y}\Big] \\
&= \frac{1}{m}\Big[X^T X\boldsymbol{w} - X^T\boldsymbol{y}\Big]
\end{aligned}$$

By setting the derivative to 0 we obtain:

$$\begin{aligned}
&\nabla_{\boldsymbol{w}} J(\boldsymbol{w}) = 0 \Rightarrow \\
&\frac{1}{m}\Big[X^T X\boldsymbol{w} - X^T\boldsymbol{y}\Big] = 0 \Rightarrow \\
&X^T X\boldsymbol{w} - X^T\boldsymbol{y} = 0 \Rightarrow \\
&X^T X\boldsymbol{w} = X^T\boldsymbol{y} \Rightarrow \\
&\underbrace{(X^T X)^{-1}(X^T X)}_{I}\boldsymbol{w} = (X^T X)^{-1}X^T\boldsymbol{y} \Rightarrow \\
&\boldsymbol{w} = (X^T X)^{-1}X^T\boldsymbol{y}
\end{aligned}$$

This final expression is called "Normal Equation", and it is an exact analytical solution that gives the parameter vector.

Despite the fact that normal equation gives an exact analytical result, computing the seemingly harmless inverse of an $(m \times (n+1)) \times ((n+1) \times m)$ matrix is, with today's most efficient computer science algorithm, of cubic time complexity! This means that as the dimensions of X increase, the amount of operations required to compute the final result increases in a cubic trend. If X was rather small, then using the normal equation would be feasible.

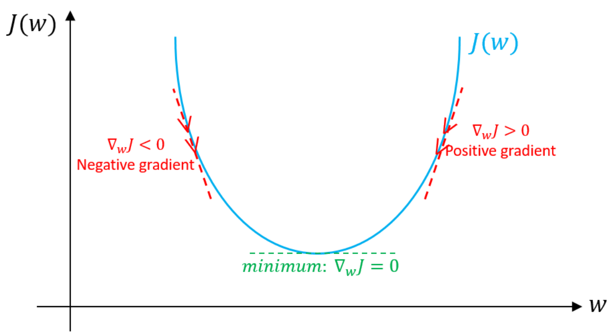
In practise, for the vast majority of any industrial application with large datasets, the normal equation would take extremely, sometimes nonsensically, long. This is the reason why normal equation is almost never used. Now let's move on the the most standard optimization technique used today called "gradient descent".

11.2.2 Gradient Descent

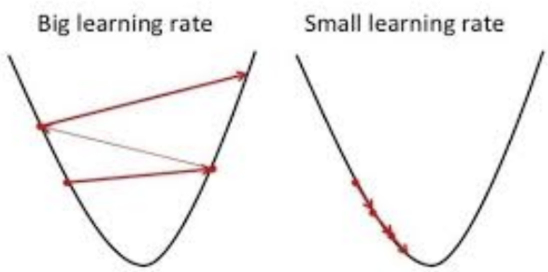
Gradient descent, and all its improvements and alternatives, is the most used optimization technique in machine learning and deep learning. In gradient descent we begin with some random initial values for the parameters and we calculate the value the loss function based on them. Then we update them based on the following relation:

$$\boldsymbol{w} := \boldsymbol{w} - \alpha \nabla_{\boldsymbol{w}} J(\boldsymbol{w})$$

Since the derivative is positive when J is upwards slopping and negative when it is downwards slopping, the minus sign makes sure that we always update the parameters towards the direction that minimizes J. Once the minimum is reached then J is at a global optimum so the derivative is 0 and further updates are not possible. At this stage the gradient descent is over and the best parameters have been found.



The hyperparameter α is called "learning rate" and defines how big or small steps we take after each iteration of gradient descent. If α is too large, we might fail to find the minimum due to oscillations around it. If α is too small then gradient descent might take too much time to reach the minimum of J. Tuning learning rate in a "right" value is a topic by itself and it is quite have researched today!



Coming back to our case, let's find the update rule specifically for linear regression. The only thing missing is the derivative of J. However in the previous chapter with normal equation we showed that:

$$J(\boldsymbol{w}) = \frac{1}{m}\Big(X^T X\boldsymbol{w} - X^T\boldsymbol{y}\Big) = \frac{1}{m}X^T\Big(X\boldsymbol{w} - \boldsymbol{y}\Big)$$

Hence the update rule reads:

$$\boldsymbol{w} := \boldsymbol{w} - \frac{\alpha}{m}X^T\Big(X\boldsymbol{w} - \boldsymbol{y}\Big)$$

As we already mentioned there are many improvements and modified algorithms based on gradient descent philosophy. We are going to cover a lot of them in these notes. For now, let's start with 3 basics versions of gradient descent.

- Batch Gradient Descent:

  Batch gradient descent is actually the one we just saw. As we see in gradient descent, the whole training set X is used in order to make just one update of the parameters. We usually refer to the whole training set as the "batch". For that reason, the usual terminology for what we have seen so far is "batch gradient descent", meaning that the whole batch is used to update the parameter.

- Mini-Batch Gradient Descent:

  In "mini-batch gradient descent" we divide the whole dataset to b subsets of $\frac{m}{b}$ training examples each, called "mini-batches", and we update the parameters using each of the mini-batches in each iteration.

- Stochastic Gradient Descent:

  In "stochastic gradient descent" we only use one training example per time to update the parameters. In every iteration we pick the training example randomly hence the naming.

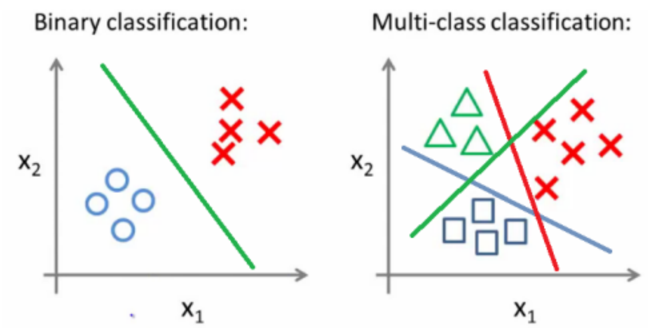
One of the main problems of batch gradient descent is that as the number of training examples grows the dimensions of X grows and using the whole training set for every iteration becomes computationally expensive. Both mini-batch and stochastic gradient descent deal with this problem.

In a way, mini-batch gradient descent tries to strike a balance between the goodness of batch gradient descent and speed of stochastic gradient descent. In general, batch gradient descent works just fine so we don't need the alternative techniques we just introduced. However in more complicated models, such as deep learning models, these techniques can be really useful. For this reason we will examine again these techniques in more details in the chapter of deep learning.

11.3 Logistic Regression

Definition 11.7 (Logistic Regression)**.** ***Logistic regression*** *is a statistical model that is used to model a binary dependent variable (usually 0 or 1).*

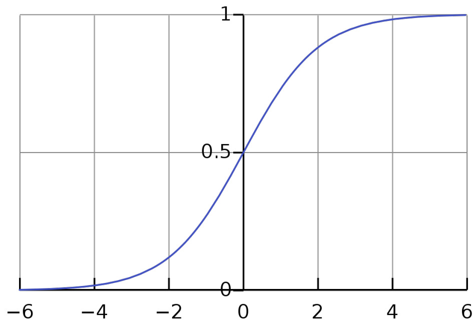
Sometimes logistic regression is called "binary classification" since we try to decide if an input belongs to class 0 or class 1 of the binary output. Wecan also generalize to more than 2 discrete classes where then we have "multi-class classification". For now we will focus on binary outputs.



Since the output is binary and can take only the values 0 or 1, the hypothesis function of the linear regression is not a valid approximator for logistic regression since it produces a continuous set of outputs. So our first step is to find a suitable hypothesis function h for logistic regression. The hypothesis function that we actually use in logistic regression is the sigmoid function and it is given by:

$$h(\boldsymbol{x}) = \frac{1}{1 + exp(-\boldsymbol{w}^T \boldsymbol{x})}$$

(From now on, in order to save space, we will write $h(\boldsymbol{x})$ instead of the actual expression for the sigmoid function.)



Sigmoid function produces results in the interval $[0, 1]$. It is quite close to what we need, but not exactly so, since we do not need all the values between 0 and 1. For this reason we will try to give a different meaning to the hypothesis function.

Namely the intuition is that the sigmoid function will act as a probability measure of the target to belong to class 1. The closest to 0 the sigmoid function, the more unlikely for the target to belong in class 1 (hence it belongs to class 0) and the closest to 1 the more likely to belong to the class 1. Hence, by defining a threshold (say at 0.5) the idea is that for $h(\boldsymbol{x}) < 0.5$ the algorithm will predict 0 and for $h(\boldsymbol{x}) \geq 0.5$ the algorithm will predict 1.

As we said, based on this intuition, we can interpret the hypothesis function as the probability of the input to be 1 hence:

$$P(y = 1|\boldsymbol{x}; \boldsymbol{w}) = h(\boldsymbol{x})$$

Of course since the output must be either 0 or 1 we get:

$$\begin{aligned} &P(y = 0|\boldsymbol{x}; \boldsymbol{w}) + P(y = 1|\boldsymbol{x}; \boldsymbol{w}) = 1 \Rightarrow \\ &P(y = 0|\boldsymbol{x}; \boldsymbol{w}) = 1 - P(y = 1|\boldsymbol{x}; \boldsymbol{w}) \Rightarrow \\ &P(y = 0|\boldsymbol{x}; \boldsymbol{w}) = 1 - h(\boldsymbol{x}) \end{aligned}$$

We can combine these two probabilities in one in the following way:

$$P(y|\boldsymbol{x}; \boldsymbol{w}) = h(\boldsymbol{x})^y \cdot (1 - h(\boldsymbol{x}))^{(1-y)}$$

So in logistic regression the output follows a Bernoulli distribution with parameter $h(\boldsymbol{x})$. Now that we have a probability distribution, similarly to the linear regression, we can use the principle of the maximum likelihood in order to obtain the best parameters that maximize the likelihood. Thus, we will obtain the loss function for the logistic regression case.

By making again the assumption that we are dealing with independent and identically distributed random variables, for the likelihood it is

$$\mathcal{L}(\boldsymbol{w}|y) = P(y^{(1)}, y^{(2)}, \ldots, y^{(m)}|\boldsymbol{x}^{(1)}, \boldsymbol{x}^{(2)}, \ldots, \boldsymbol{x}^{(m)}; \boldsymbol{w}) = \prod_{i=1}^{m} P(y^{(i)}|\boldsymbol{x}^{(i)}; \boldsymbol{w})$$

By substituting the probability:

$$\mathcal{L}(\boldsymbol{w}|y) = \prod_{i=1}^{m} h(\boldsymbol{x}^{(i)})^{y^{(i)}} \cdot (1 - h(\boldsymbol{x}^{(i)}))^{(1-y^{(i)})}$$

Subsequently for the log-likelihood:

$$\begin{aligned}
l(\boldsymbol{w}|y) &= \ln \mathcal{L}(\boldsymbol{w}|y) \\
&= \ln \Big[\prod_{i=1}^{m} h(\boldsymbol{x}^{(i)})^{y^{(i)}} \cdot (1 - h(\boldsymbol{x}^{(i)}))^{(1-y^{(i)})} \Big] \\
&= \sum_{i=1}^{m} \ln \Big[h(\boldsymbol{x}^{(i)})^{y^{(i)}} \cdot (1 - h(\boldsymbol{x}^{(i)}))^{(1-y^{(i)})} \Big] \\
&= \sum_{i=1}^{m} \Big[\ln \Big(h(\boldsymbol{x}^{(i)})^{y^{(i)}} \Big) + \ln \Big((1 - h(\boldsymbol{x}^{(i)}))^{(1-y^{(i)})} \Big) \Big] \\
&= \sum_{i=1}^{m} \Big[y^{(i)} \cdot \ln h(\boldsymbol{x}^{(i)}) + (1 - y^{(i)}) \cdot \ln(1 - h(\boldsymbol{x}^{(i)})) \Big]
\end{aligned}$$

Once again, according to the principle of maximum likelihood, the best parameters can be found by maximizing the log-likelihood. Hence:

$$\begin{aligned}
\boldsymbol{w} &= \arg\max_{\boldsymbol{w}} [l(\boldsymbol{w}|y)] \\
&= \arg\max_{\boldsymbol{w}} \Big[\sum_{i=1}^{m} \Big((y^{(i)} \cdot \ln h(\boldsymbol{x}^{(i)}) + (1 - y^{(i)}) \cdot \ln(1 - h(\boldsymbol{x}^{(i)})) \Big) \Big] \\
&= \arg\min_{\boldsymbol{w}} \Big[- \sum_{i=1}^{m} \Big(y^{(i)} \cdot \ln h(\boldsymbol{x}^{(i)}) + (1 - y^{(i)}) \cdot \ln(1 - h(\boldsymbol{x}^{(i)})) \Big) \Big]
\end{aligned}$$

At this point we can formally define the following function:

Definition 11.8 (Cross Entropy Loss Function)**.** ***Cross entropy loss function*** *is defined as:*

$$J(\boldsymbol{w}) = -\frac{1}{m} \sum_{i=1}^{m} \Big(y^{(i)} \cdot \ln h(\boldsymbol{x}^{(i)}) + (1 - y^{(i)}) \cdot \ln(1 - h(\boldsymbol{x}^{(i)})) \Big)$$

The cross entropy is the loss function of the logistic regression in the similar way where MSE loss function is the loss function for linear regression. The name "cross entropy" comes from the definition of cross entropy which is the average amount of information needed to identify an event between two probability distributions p and q over the same underlying set of events

Notice that the only possible values of $y^{(i)}$ is 0 or 1. This means that in any case, one of the terms $y^{(i)}$ or $(1 - y^{(i)})$ in J will vanish and the other one will be equal to 1. So in the end the only thing that is actually part of the loss is the logarithm of the hypothesis function, which given that the hypothesis function is a sigmoid function which is always between 0 and 1, the logarithm is always negative. With the overall negative sign the loss turns positive and this is what we want to minimize.

We are not gonna write a vectorized form for the cross entropy loss function for two reasons. First of all, not all matrices have a logarithm and those matrices that do have a logarithm may have more than one logarithm. So one has to be careful when uses the vectorized form of logistic regression because it carries logarithms of matrices. Secondly, derivatives of logarithms of non square matrices sometimes are not defined. Since we need to calculate the derivative of J we might get problems. For this reason we will use the non vectorized form for the calculations, however we will express the final result in a vectorized form.

As in the linear case, the principle of maximum likelihood leads to the minimization of the cross entropy loss function in order to obtain the best parameters. We will examine the same techniques that we developed for the gradient descent in the linear case, i.e normal equation and gradient descent.

11.3.1 Normal Equation

As we already mentioned, since we want to minimize a function the straight forward way of doing that is to calculate the derivative and then set it to 0. However, in the case of logistic regression the normal equation does not apply since there is no closed analytical solution of $\nabla_{\boldsymbol{w}} J(\boldsymbol{w}) = 0$. The only way for solving the optimization problem is through gradient descent.

11.3.2 Gradient Descent

Gradient descent works fine in logistic regression. First, let's calculate the derivative of $J(\boldsymbol{w})$ for logistic regression.

$$
\begin{aligned}
\nabla_{\boldsymbol{w}} J(\boldsymbol{w}) &= -\frac{1}{m}\nabla_{\boldsymbol{w}}\Big[\sum_{i=1}^{m}\Big(y^{(i)}\cdot\ln h(\boldsymbol{x}^{(i)}) + (1-y^{(i)})\cdot\ln(1-h(\boldsymbol{x}^{(i)}))\Big)\Big] \\
&= -\frac{1}{m}\sum_{i=1}^{m}\Big(y^{(i)}\cdot\nabla_{\boldsymbol{w}}\ln h(\boldsymbol{x}^{(i)}) + (1-y^{(i)})\cdot\nabla_{\boldsymbol{w}}\ln(1-h(\boldsymbol{x}^{(i)}))\Big) \\
&= -\frac{1}{m}\sum_{i=1}^{m}\Big(y^{(i)}\cdot\frac{\nabla_{\boldsymbol{w}} h(\boldsymbol{x}^{(i)})}{h(\boldsymbol{x}^{(i)})} - (1-y^{(i)})\cdot\frac{\nabla_{\boldsymbol{w}} h(\boldsymbol{x}^{(i)})}{1-h(\boldsymbol{x}^{(i)})}\Big) \\
&= -\frac{1}{m}\sum_{i=1}^{m}\Big(\frac{y^{(i)}}{h(\boldsymbol{x}^{(i)})} - \frac{1-y^{(i)}}{1-h(\boldsymbol{x}^{(i)})}\Big)\cdot\nabla_{\boldsymbol{w}} h(\boldsymbol{x}^{(i)}) \\
&= -\frac{1}{m}\sum_{i=1}^{m}\Big(\frac{y^{(i)}\cdot(1-h(\boldsymbol{x}^{(i)}) - (1-y^{(i)})\cdot h(\boldsymbol{x}^{(i)})}{h(\boldsymbol{x}^{(i)})\cdot(1-h(\boldsymbol{x}^{(i)}))}\Big)\cdot\nabla_{\boldsymbol{w}}\Big[\frac{1}{1+exp(-\boldsymbol{w}^T\boldsymbol{x}^{(i)})}\Big] \\
&= -\frac{1}{m}\sum_{i=1}^{m}\Big(\frac{y^{(i)} - y^{(i)}\cdot h(\boldsymbol{x}^{(i)}) - h(\boldsymbol{x}^{(i)}) + y^{(i)}\cdot h(\boldsymbol{x}^{(i)})}{h(\boldsymbol{x}^{(i)})\cdot(1-h(\boldsymbol{x}^{(i)}))}\Big)\cdot\Big(\frac{-1}{(1+exp(-\boldsymbol{w}^T\boldsymbol{x}^{(i)}))^2}\cdot exp(-\boldsymbol{w}^T\boldsymbol{x}^{(i)})\cdot(-\boldsymbol{x}^{(i)})\Big) \\
&= -\frac{1}{m}\sum_{i=1}^{m}\Big(\frac{y^{(i)} - h(\boldsymbol{x}^{(i)})}{h(\boldsymbol{x}^{(i)})\cdot(1-h(\boldsymbol{x}^{(i)}))}\Big)\cdot\Big(h(\boldsymbol{x}^{(i)})^2\cdot\frac{1-h(\boldsymbol{x}^{(i)})}{h(\boldsymbol{x}^{(i)})}\cdot\boldsymbol{x}^{(i)}\Big) \\
&= \frac{1}{m}\sum_{i=1}^{m}\frac{h(\boldsymbol{x}^{(i)}) - y^{(i)}}{h(\boldsymbol{x}^{(i)})\cdot(1-h(\boldsymbol{x}^{(i)}))}\cdot h(\boldsymbol{x}^{(i)})\cdot(1-h(\boldsymbol{x}^{(i)}))\cdot\boldsymbol{x}^{(i)} \\
&= \frac{1}{m}\sum_{i=1}^{m}(h(\boldsymbol{x}^{(i)}) - y^{(i)})\cdot\boldsymbol{x}^{(i)}
\end{aligned}
$$

Hence the update rule reads:

$$
\boldsymbol{w} := \boldsymbol{w} - \frac{\alpha}{m}\sum_{i=1}^{m}\Big(\frac{1}{1+exp(-\boldsymbol{w}^T\boldsymbol{x}^{(i)})} - y^{(i)}\Big)\cdot\boldsymbol{x}^{(i)}
$$

Or in vectorized form:

$$\boldsymbol{w} := \boldsymbol{w} - \frac{\alpha}{m} X^T \Big(\frac{1}{1 + exp(-X\boldsymbol{w})} - \boldsymbol{y} \Big)$$

At this point, notice that gradient descent can be generalized into one model for both linear and logistic regression since the update rule for both of them can be be written in one coherent way as:

$$\boldsymbol{w} := \boldsymbol{w} - \frac{\alpha}{m} X^T \Big(h(X) - \boldsymbol{y} \Big)$$

where one has to use either MSE loss function or cross entropy loss function depending on the regression problem!

11.4 Generalized Linear Model

As we showed, in linear regression the target follows a normal distribution while in logistic regression the target follows a Bernoulli distribution. We can generalize both regressions in one coherent model called "generalized linear model" in which the target is allowed to follow a broad family of probability distributions.

Definition 11.9 (Generalized Linear Model)**.** ***Generalized linear model*** *(GLM) is a model that allows the dependent variable to follow an exponential family of probability distributions of the form:*

$$P(y|\eta) = b(y) \cdot exp(\eta^T T(y) - a(\eta))$$

By picking specific values for b, η, T and a we end up with different distributions (including linear and logistic regression). Then we simply assume independence and apply the principle of maximum likelihood to obtain a loss function, in order to minimize it and find the best parameters.

11.5 Errors

Since h is an estimator of f, the theory we developed in the chapter of parametric inference for estimators also holds for h. In other words, for the hypothesis function, which acts as an estimator for f, we can define quantities such as MSE, sampling deviation, bias and variance.

Out of all possible quantities that can be defined, MSE, bias and variance are of crucial importance in machine learning, since we use them in order to evaluate how well a mahcine learning model performs. Let us see now the definitions of these quantities adjusted for the case of machine learning where the estimator is h.

11.5.1 Point-Wise, Overall, In-Sample & Out-Of-Sample Error

Starting from the corresponding MSE in machine learning case we define the following quantities:

Definition 11.10 (Point-Wise Error)**.** *We define the* ***point-wise error*** *e as a function of the real target function f and the hypothesis function h at point x:*

$$e = e(f(x), h(x))$$

For example, for linear regression we could use $e(f(x), h(x)) = (f(x) - h(x))^2$ while for logistic regression $e(f(x), h(x)) = [f(x) \neq h(x)]$. Given point-wise error we can generalize to overall error.

Definition 11.11 (Overall Error)**.** *We define the* ***overall error*** *E as the average over all point-wise errors $e(f(x), h(x))$ at every point x.*

We distinguish between two kind of overall errors: the in-sample error and the out-of-sample error.

Definition 11.12 (In-Sample Error)**.** *We define the* ***in-sample error*** *E_{in} as the average of point-wise errors of the dataset that the model was trained:*

$$E_{in} = \frac{1}{m} \sum_{i=1}^{m} e(f(x^{(i)}), h(x^{(i)}))$$

In other words in-sample error shows how well the model performs on the data used to build it.

Definition 11.13 (Out-Of-Sample Error)**.** *We define the **out-of-sample error** E_{out} as a the expected value of point-wise errors of new data:*

$$E_{out} = E_x[e(f(x), h(x))]$$

In other word out-of-sample error show how well the model generalizes to predictions for data it has not seen before. It makes sense that in order for h to work well out of sample, so it can predict, it must be $E_{out} \approx 0$.

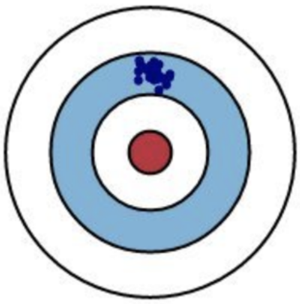
11.5.2 Bias & Variance

Back in parametric inference chapter, we introduced the bias B of an estimator $\hat{\theta}$, as the difference between the expected value of the estimator and the actual true parameter we want to estimate, $B = E[\hat{\theta}_n] - \theta$. Coming to our case where our estimator is $\hat{\theta} = h(x)$ and the true parameter is the target funtion $\theta = f(x)$, for the bias we get:

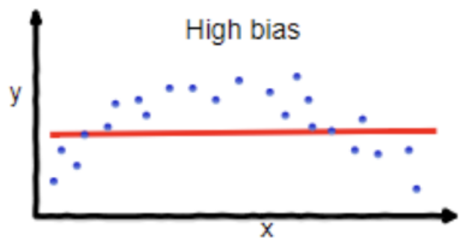
Definition 11.14 (Bias)**.** *We define the **bias** B of the hypothesis function h as the quantity:*

$$B = E[h(x)] - f(x)$$

The bias error is an error from erroneous assumptions in the learning algorithm. When we are dealing with high bias, formally we can say that the hypothesis set $H = \{h\}$ was not big enough in order to contain function that can approximate well the target function f. So our best approximation for f is still a bad one that cannot fit the data well.



High bias can cause an algorithm to miss the relevant relations between features and target outputs and fail to capture the underlying structure of the dataset. We call this a case of underfitting, since the model fails to fit the given dataset well.



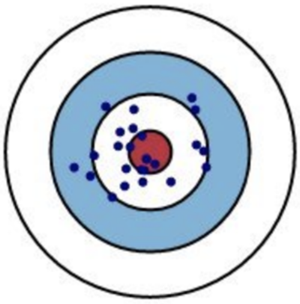
Underfitting is one of the two main problems that a machine learning model can have and it leads to a high in-sample error E_{in} which subsequently leads to a high out-of-sample error E_{out}. Hence even that the problem is coming from the in-sample error it leads to not being able to generalize for out-of-sample data.

In parametric inference chapter, we also defined the variance of an estimator $\hat{\theta}_n$ as the expected value of the square difference of the estimator from the expected value of the estimator: $Var = E[(\hat{\theta}_n - E[\hat{\theta}_n])^2]$. Coming back to machine learning for the variance we get:

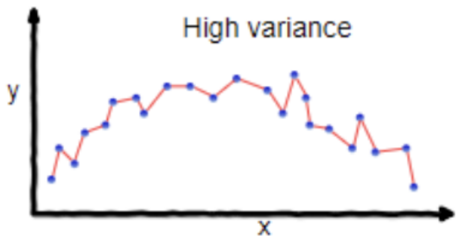
Definition 11.15 (Variance)**.** *We define the* ***variance*** *B of the hypothesis function h as the quantity:*

$$Var = E_x[(h(x) - E_x[h(x)])^2]$$

The variance is an error from sensitivity to small fluctuations in the training set.



When we are dealing with high variance, informally we can say that the hypothesis set $H = \{h\}$ is very big so in order to compensate the spread of dataset the model finds a function that fits the particular data very well but fails to generalize to new data. We call this a case of overfitting, since the model fails to generalize to new data.

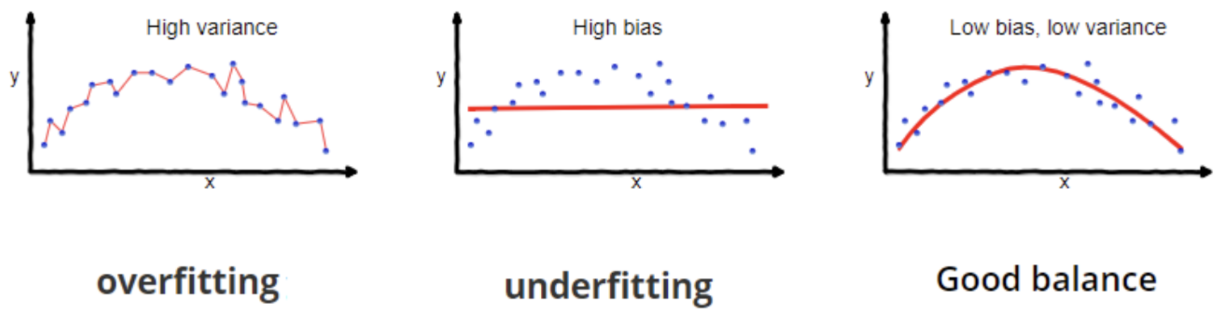


Overfitting leads to a very low in-sample $E_{in} \approx 0$, since it does a very good job on fitting the given data. However it fails to generalize, hence to predict new data, which leads to a very high out-of-sample error E_{out}.

Back in parametric inference we also showed that the mean squared error can be decomposed to bias and variance, and of course the same holds in our case since for the out of sample error of linear regression we can show:

$$
\begin{aligned}
E_{out} &= E_x\Big[e(f(x), h(x))\Big] \\
&= E_x\Big[\Big(f(x) - h(x)\Big)^2\Big] \\
&= E_x\Big[\Big(f(x) - h(x) + E_x[h(x)] - E_x[h(x)]\Big)^2\Big] \\
&= E_x\Big[\Big(\Big(f(x) - E_x[h(x)]\Big) + \Big(E_x[h(x)] - h(x)\Big)\Big)^2\Big] \\
&= E_x\Big[\Big(f(x) - E_x[h(x)]\Big)^2 + 2\Big(f(x) - E_x[h(x)]\Big)\Big(E_x[h(x)] - h(x)\Big) + \Big(E_x[h(x)] - h(x)\Big)^2\Big] \\
&= E_x\Big[\Big(f(x) - E_x[h(x)]\Big)^2\Big] + E_x\Big[2\Big(f(x) - E_x[h(x)]\Big)\Big(E_x[h(x)] - h(x)\Big)\Big] + E_x\Big[\Big(E_x[h(x)] - h(x)\Big)^2\Big] \\
&= \Big(f(x) - E_x[h(x)]\Big)^2 + 2\Big(f(x) - E_x[h(x)]\Big)\Big(E_x\Big[E_x[h(x)] - h(x)\Big]\Big) + E_x\Big[\Big(E_x[h(x)] - h(x)\Big)^2\Big] \\
&= \Big(f(x) - E_x[h(x)]\Big)^2 + 2\Big(f(x) - E_x[h(x)]\Big)\Big(E_x[h(x)] - E_x[h(x)]\Big) + E_x\Big[\Big(E_x[h(x)] - h(x)\Big)^2\Big] \\
&= \Big(f(x) - E_x[h(x)]\Big)^2 + E_x\Big[\Big(E_x[h(x)] - h(x)\Big)^2\Big] \\
&= B^2 + Var
\end{aligned}
$$

Hence, the out-of-sample error is actually a combination of bias and variance. So in order to have a model that generalizes well, we have to keep both of them low. However, since they are of opposite nature, the more we reduce bias the more variance increases and vice versa. This is the so called "bias-variance trade off". The goal in machine learning is to balance this trade off so the model fits the data well and doesn't fail to generalize.



In this graph we see that in the case of high bias (underfitting) we have restricted our model to linear predictors, however the data do not follow a linear trend, hence the hypothesis set is too small and the model cannot find a good curve to fit the data. On the other hand, in the case of high variance (overfitting) the hypothesis set is so big allowing complex predictors so the model managed to find a high degree polynomial that fit the data really good, however it will fail to generalize since it depends a lot on the initial values of the data and it is sensitive to fluctuations of them. Finally in the last graph we have a good balance of bias and variance and the model found a good curve!

11.6 Evaluation

Evaluation is about how good a model generalizes to new data. After applying the learning algorithm to the data and having obtained a hypothesis h, the machine learning model is ready to make new predictions. However before that, we have to evaluate the model by analysing the errors that we just introduced.

The starting part is the in-sample and out-of-sample errors that we defined previously as:

$$E_{in} = \frac{1}{m}\sum_{i=1}^{m} e(f(x^{(i)}), h(x^{(i)})) \qquad \text{and} \qquad E_{out} = E_x[e(f(x), h(x))]$$

In general, for the error function e we use the corresponding loss function J that we used to train the model, since it is a function of the target and hypothesis functions as e, and it is a really good measure of error:

$$E_{in} = \frac{1}{m}\sum_{i=1}^{m} J^{(i)}(f(x^{(i)}), h(x^{(i)})) \qquad \text{and} \qquad E_{out} = E_x[J(f(x), h(x))]$$

where here the notation $J^{(i)}$ means the error coming from the i'th training example. Hence, now E_{in} is calculated with the data that we trained the model, so it's a very good measure of how well the model performs in the data that it was trained on. The problem comes from E_{out} since we don't know how to compute this expected value. Unsurprising we will perform the usual trick of substituting the expected value with the average so:

$$E_{in} = \frac{1}{m}\sum_{i=1}^{m} J^{(i)}(f(x^{(i)}), h(x^{(i)})) \qquad \text{and} \qquad E_{out} = \frac{1}{m}\sum_{i=1}^{m} J^{(i)}(f(x^{(i)}), h(x^{(i)}))$$

Of course since we estimate the expected value with an average that brings an error to the estimation of the out-of-sample error. However for our purposes we assume that this error is neglectful, and from now on we will treat the estimated out-of-sample error as the actual out-of-sample error. In general we have to keep in mind though that out-of-sample error carries an error.

The question that arises is what data are we going to use for E_{out}. Using the same data that we trained the model is a really bad idea since, first of all, we will simply get $E_{out} = E_{in}$ and secondly the model already knows the correct answers of the data since we used them to train it, and the evaluation will be biased.

In order to overcome this problem, we split the dataset (before training the model) into two parts: training set and evaluation set. Then we use the first to train the model and obtain E_{in} and the latter to evaluate its performance and obtain E_{out}. Since the model is trained with the train set, it has never seen the evaluation set so the estimation of the out-of-sample error with the evaluation set will be unbiased.

One of the things to consider is the proportions of splitting the dataset intro training set and evaluation set. This again is an area of heavy research, but in general in machine learning we usually split them either in a proportion of "70% - 30%" or "80% - 20%" depending on the amount of data. (In all cases the largest proportion goes for training the model). In other areas of machine learning like deep learning where we usually have a very large amount of data we use splitting rules of "99% - 1%". But we will address this issue in details in deep learning chapter.

Hence, before training we split the dataset as:

- Training Dataset: $\{x^{(i)}_{\text{train}}, y^{(i)}_{\text{train}}\}, \quad i = 1, 2, \ldots, m_{\text{train}}$ (70% or 80% of initial dataset)

- Test Dataset: $\{x^{(i)}_{\text{eval}}, y^{(i)}_{\text{eval}}\}, \quad i = 1, 2, \ldots, m_{\text{eval}}$ (30% or 20% of initial dataset)

And subsequently the errors become:

$$E_{in} = \frac{1}{m_{\text{train}}} \sum_{i=1}^{m_{\text{train}}} J^{(i)}(f(x^{(i)}), h(x^{(i)})) \qquad \text{and} \qquad E_{out} = \frac{1}{m_{\text{eval}}} \sum_{i=1}^{m_{\text{eval}}} J^{(i)}(f(x^{(i)}), h(x^{(i)}))$$

From now on we will be referring to the first expression as $J_{\text{train}} = E_{in}$ and to the second one as $J_{\text{eval}} = E_{out}$.

So for example for linear regression we would have:

$$J_{\text{train}} = \frac{1}{2m_{train}} \sum_{i=1}^{m_{train}} (y^{(i)} - \boldsymbol{w}^T \boldsymbol{x}^{(i)})^2 \qquad \text{and} \qquad J_{\text{eval}} = \frac{1}{2m_{eval}} \sum_{i=1}^{m_{eval}} (y^{(i)} - \boldsymbol{w}^T \boldsymbol{x}^{(i)})^2$$

While for logistic regression we would have

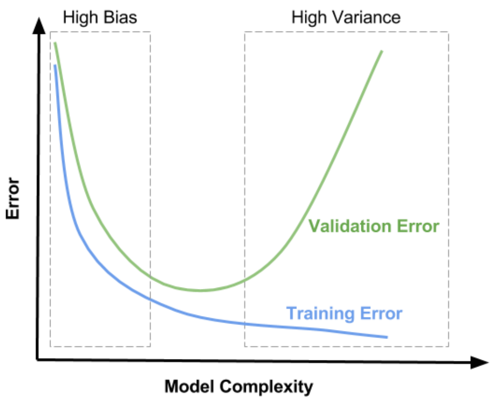
$$J_{\text{train}} = -\frac{1}{m_{train}} \sum_{i=1}^{m_{train}} \left(y^{(i)} \cdot \ln h(\boldsymbol{x}^{(i)}) + (1 - y^{(i)}) \cdot \ln(1 - h(\boldsymbol{x}^{(i)})) \right)$$

and

$$J_{\text{eval}} = -\frac{1}{m_{eval}} \sum_{i=1}^{m_{eval}} \left(y^{(i)} \cdot \ln h(\boldsymbol{x}^{(i)}) + (1 - y^{(i)}) \cdot \ln(1 - h(\boldsymbol{x}^{(i)})) \right)$$

Now that we have J_{train} and J_{eval} we know how well the models performs in and out of sample. However we can also use them in order to find if the model suffers from underfitting or overfitting. There are two ways that we can do so.

The first way, is by gradually increasing the complexity of the model (higher polynomial degrees so bigger hypothesis set), training the model for each complexity level and calculate both J_{train} and J_{eval} for each model. Then by plotting out the different values for different levels of complexity we usually end up with the following graph:



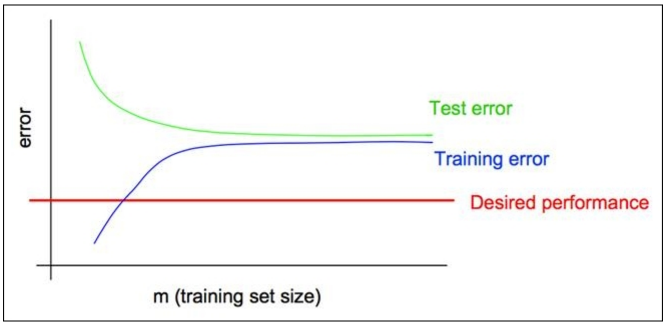
In the high bias area both J_{train} and J_{eval} are high. This means that the model does not fit the training data well hence it fails to generalize. This is the case of underfitting. In the high variance area, J_{train} is low but J_{eval} is high. This means that the model fits the training data well but fails to generalize to unseen data. This is the case of overfitting. Hence by using the graph we can diagnose both cases!

The second way to diagnose the problem of the model, is through the so called "learning curves".

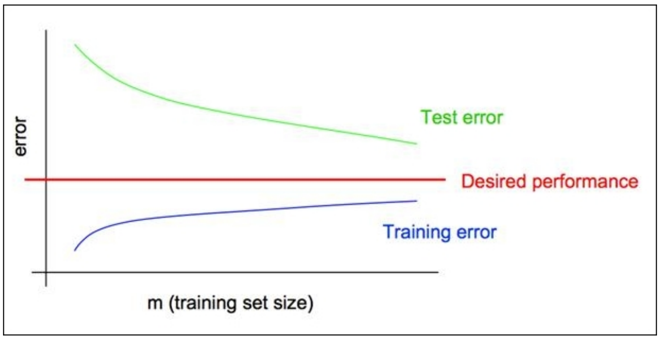
Definition 11.16 (Learning Curve)**.** *A **learning curve** is a graphical representation of how an increase in learning comes from greater experience; or how the more someone performs a task, the better they get at it.*

Informally, a learning curve is the relation between error (as expressed in loss function) and training examples m. By plotting this relation for both J_{train} and J_{eval} we end up with two learning curves and by their relative position we can diagnose if our model suffers from high bias or high variance.

More specifically, when the learning curves of J_{train} and J_{eval} do not have a large gap between them as m increases we are usually dealing we a case of high bias and underfitting.



On the other hand when there is a large gap between the two curves we are dealing with the case of high variance and overfitting.

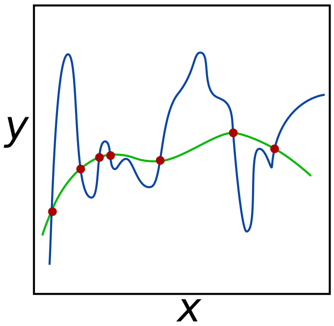


Once we detect the problem then we have to make some changes in order to fix them! Here are some of the techniques that we follow:

- For high bias (underfitting):
  - Increase model complexity (polynomial terms)
  - Increase number of features
- For high variance (overfitting):
  - Decrease model complexity (polynomial terms)
  - Decrease number of features
  - Find more training examples
  - Regularization (next section)

11.7 Regularization

As we saw in the previous section, when we allow a very broad hypothesis set with many higher order terms the model might find a hypothesis function h that gives a 0 in-sample error but fails to generalize. This is due to high variance, i.e large dependence on the very specific dataset used for training, and we call this a case of overfitting. A way to deal with overfitting is a collection of techniques that undergo with the name “regularization”.



Definition 11.17 (Regularization)**.** *Regularization is the process of adding information in order to solve an ill-posed problem and to prevent overfitting.*

We will see many different regularization techniques throughout the notes. For now we will start with "Ridge Regression" or "L2 Regularization" and "Lasso Regression " or "L1 Regularization".

11.7.1 Ridge Regression - L2 Regularization

The reason of overfitting is that the parameters $\boldsymbol{w}$ are free to get any value. With regularization we penalize the parameters by imposing an extra constraint on $\boldsymbol{w}$ of the form:

$$\boldsymbol{w}^T\boldsymbol{w} \leq C$$

where C is a constant defined by us and it controls the effect of regularization. It is called "L2 regularization" because the quantity $\boldsymbol{w}^T\boldsymbol{w}$ is actually the squared L2 norm of the vector $\boldsymbol{w}$:

$$||\boldsymbol{w}||_2^2 = \boldsymbol{w}^T\boldsymbol{w}$$

Hence now the optimization problem becomes to minimize the loss function $J(\boldsymbol{w})$ subject to the above mentioned constraint. According to (Appendix A) in order to do so we define the Lagrangian:

$$\mathcal{L}(\boldsymbol{w}) = J(\boldsymbol{w}) + \frac{\lambda}{2m}\boldsymbol{w}^T\boldsymbol{w}$$

where λ is the Lagrange multiplier, and then we solve the equation:

$$\nabla_{\boldsymbol{w}}\mathcal{L}(\boldsymbol{w}) = 0$$

For example, for linear regression where $J(\boldsymbol{w})$ is given by (**??**) the Lagrangian reads:

$$\mathcal{L}(\boldsymbol{w}) = J(\boldsymbol{w}) + \frac{\lambda}{2m}\boldsymbol{w}^T\boldsymbol{w} = \frac{1}{2m}(X\boldsymbol{w} - \boldsymbol{y})^T(X\boldsymbol{w} - \boldsymbol{y}) + \frac{\lambda}{2m}\boldsymbol{w}^T\boldsymbol{w}$$

At this point we can redefine this Lagrangian as a new loss function of the form:

$$J(\boldsymbol{w}) = \frac{1}{2m}\Big[(X\boldsymbol{w} - \boldsymbol{y})^T(X\boldsymbol{w} - \boldsymbol{y}) + \lambda\boldsymbol{w}^T\boldsymbol{w}\Big]$$

and then the problem is to minimize this loss function which is actually a regression problem. The corresponding regression is called "Ridge regression", where the only difference with linear regression is that we have to add the extra term in the loss function to reduce overfitting.

The coefficient λ is the one that controls the regularization effect on the regression. In one extreme where $\lambda = 0$ the regularization term vanishes, and the loss function ends up to the mean squared error loss function, hence the ridge regression turns to linear regression. In the other extreme where $\lambda \to \infty$ then the regularization term penaltizes all parameters in an extreme way, so the ridge regression, in order to minimize the loss, is forced to set all the parameters to 0. In the end we end up with $\boldsymbol{w}^T\boldsymbol{x} = 0$. For all intermediate values of λ we get different levels of regularization. It is actually our job to tune the model to the right λ that does the job.

Now that we have a loss function, we treat the problem in the similar way as we did before. For example, in the linear case of ridge regression we can solve the normal equation in the same way we solved it before:

$$\begin{aligned}
J(\boldsymbol{w}) &= \frac{1}{2m}\Big[(X\boldsymbol{w}-\boldsymbol{y})^T(X\boldsymbol{w}-\boldsymbol{y}) + \lambda\boldsymbol{w}^T\boldsymbol{w}\Big] \\
&= \frac{1}{2m}\Big[\Big((X\boldsymbol{w})^T - \boldsymbol{y}^T\Big)\Big(X\boldsymbol{w}-\boldsymbol{y}\Big) + \lambda\boldsymbol{w}^T\boldsymbol{w}\Big] \\
&= \frac{1}{2m}\Big[(X\boldsymbol{w})^T(X\boldsymbol{w}) - (X\boldsymbol{w})^T\boldsymbol{y} - \boldsymbol{y}^T(X\boldsymbol{w}) + \boldsymbol{y}^T\boldsymbol{y} + \lambda\boldsymbol{w}^T\boldsymbol{w}\Big] \\
&= \frac{1}{2m}\Big[(X\boldsymbol{w})^T(X\boldsymbol{w}) - 2(X\boldsymbol{w})^T\boldsymbol{y} + \boldsymbol{y}^T\boldsymbol{y} + \lambda\boldsymbol{w}^T\boldsymbol{w}\Big] \\
&= \frac{1}{2m}\Big[\boldsymbol{w}^TX^TX\boldsymbol{w} - 2\boldsymbol{w}^TX^T\boldsymbol{y} + \boldsymbol{y}^T\boldsymbol{y} + \lambda\boldsymbol{w}^T\boldsymbol{w}\Big]
\end{aligned}$$

By setting the derivative to 0 we obtain:

$$\begin{aligned}
&\nabla_{\boldsymbol{w}}J(\boldsymbol{w}) = 0 \Rightarrow \\
&\frac{1}{2m}\nabla_{\boldsymbol{w}}\Big[\boldsymbol{w}^TX^TX\boldsymbol{w} - 2\boldsymbol{w}^TX^T\boldsymbol{y} + \boldsymbol{y}^T\boldsymbol{y} + \lambda\boldsymbol{w}^T\boldsymbol{w}\Big] \\
&\frac{1}{2m}\Big[2X^TX\boldsymbol{w} - 2X^T\boldsymbol{y} + 2\lambda\boldsymbol{w}\Big] = 0 \Rightarrow \\
&\frac{1}{m}\Big[X^TX\boldsymbol{w} - X^T\boldsymbol{y} + \lambda\boldsymbol{w}\Big] = 0 \Rightarrow \\
&X^TX\boldsymbol{w} - X^T\boldsymbol{y} + \lambda\boldsymbol{w} = 0 \Rightarrow \\
&(X^TX + \lambda I)\boldsymbol{w} = X^T\boldsymbol{y} \Rightarrow \\
&\underbrace{(X^TX + \lambda I)^{-1}(X^TX + \lambda I)}_{I}\boldsymbol{w} = (X^TX + \lambda I)^{-1}X^T\boldsymbol{y} \Rightarrow \\
&\boldsymbol{w} = (X^TX + \lambda I)^{-1}X^T\boldsymbol{y}
\end{aligned}$$

The only difference with the normal equation of linear regression is the extra term λI coming from regularization.

Gradient descent also works for ridge regression. For the derivative of J:

$$\nabla_{\boldsymbol{w}}J(\boldsymbol{w}) = \frac{1}{m}\Big(X^TX\boldsymbol{w} - X^T\boldsymbol{y} + \lambda\boldsymbol{w}\Big) = \frac{1}{m}X^T\Big((X + \lambda I)\boldsymbol{w} - \boldsymbol{y}\Big)$$

Hence the update rule reads:

$$\boldsymbol{w} := \boldsymbol{w} - \frac{\alpha}{m}X^T\Big((X + \lambda I)\boldsymbol{w} - \boldsymbol{y}\Big)$$

Of course, L2 regularization can be applied also for the case of logistic regression. More specifically, for cross entropy loss function of logistic regression $J(\boldsymbol{w})$ the Lagrangian reads:

$$\mathcal{L}(\boldsymbol{w}) = J(\boldsymbol{w}) + \frac{\lambda}{2m}\boldsymbol{w}^T\boldsymbol{w} = -\frac{1}{m}\Big(\boldsymbol{y}^T \cdot \ln h(X) + (I - \boldsymbol{y})^T \cdot \ln(I - h(X))\Big) + \frac{\lambda}{2m}\boldsymbol{w}^T\boldsymbol{w}$$

Similarly to the linear case, we redefine this Lagrangian as a new loss function of the form:

$$J(\boldsymbol{w}) = -\frac{1}{m}\Big[\boldsymbol{y}^T \cdot \ln h(X) + (I-\boldsymbol{y})^T \cdot \ln(I-h(X)) - \frac{\lambda}{2}\boldsymbol{w}^T\boldsymbol{w}\Big]$$

As we said back in logistic regression, normal equation does not apply here since there is no closed analytical solution, however gradient descent still applies where the rule simply reads:

$$\boldsymbol{w} := \Big(1-\frac{\alpha\lambda}{m}\Big)\boldsymbol{w} - \frac{\alpha}{m}X^T\Big(\frac{1}{1+exp(-X\boldsymbol{w})} - \boldsymbol{y}\Big)$$

In both cases solving ridge regression will give us as a result a solution that slightly underfits the data, compared to linear or logistic regression. This underfitting will produce higher bias hence, due to bias-variance trade off, a reduced variance which will lead to the reduction of overfitting.

11.7.2 Lasso Regression - L1 Regularization

In ridge regression, or L2 regression, we used the L2 norm of the vector $\boldsymbol{w}$. Another way of regularization is to use L1 norm which is:

$$||\boldsymbol{w}||_1 \leq C$$

By repeating the same way of analysis as in ridge regression, we can define the following loss function for linear regression:

$$J(\boldsymbol{w}) = \frac{1}{2m}\Big[(X\boldsymbol{w}-\boldsymbol{y})^T(X\boldsymbol{w}-\boldsymbol{y}) + \lambda||\boldsymbol{w}||_1\Big]$$

and for logistic regression:

$$J(\boldsymbol{w}) = -\frac{1}{m}\Big[\boldsymbol{y}^T \cdot \ln h(X) + (I-\boldsymbol{y})^T \cdot \ln(I-h(X)) - \frac{\lambda}{2}||\boldsymbol{w}||_1\Big]$$

The corresponding regression is called "Lasso regression". As before, we can use normal equation and gradient descent to solve Lasso regression.

11.8 Classification Error Metrics

In classification problems where both input and output can be either 0 or 1, we can follow a different approach of error evaluation based on exact matches and mismatches between prediction and actual result. The usual case, since we are dealing with a binary output, is to define either 0 or 1 as the positive class and the remaining as the negative one. Which one is which depends on the nature of the problem. For now we will stick with the case were 0 represents the negative class and 1 the positive one.

Given that both the actual class and the predicted class can be either positive or negative we end up with 4 different, distinct situations. Let us define them formally:

Definition 11.18 (True Positive)**.** ***True positive*** *(TP) also called hit, is the case where the model predicts a positive result when the actual outcome is indeed positive.*

Definition 11.19 (True Negative)**.** ***True negative*** *(TN) also called correct rejection, is the case where the model predicts a negative result when the actual outcome is indeed negative.*

Definition 11.20 (False Positive)**.** ***False positive*** *(FP) also called false alarm or type I error, is the case where the model predicts a positive result when the actual outcome is negative.*

Definition 11.21 (False Negative)**.** ***False negative*** *(FN) also called miss or type II error, is the case where the model predicts a negative result when the actual outcome is positive.*

Once the model is trained, we test it on the evaluation set and we measure the number of occurrences of each category. Then we gather them all together to the so called "confusion matrix".

Definition 11.22 (Confusion Matrix)**.** ***Confusion matrix*** *is a table that reports the number of true positives TP, true negatives TN, false positives FP and false negatives FN of a model.*

	Predicted 0	Predicted 1
Actual 0	TN	FP
Actual 1	FN	TP

Once we have constructed the confusion matrix we can define the following error metrics:

Definition 11.23 (Accuracy)**.** ***Accuracy*** *(ACC) is the rate that shows overall how often the model was correct:*

$$ACC = \frac{TP + TN}{TP + TN + FP + FN}$$

Definition 11.24 (Error Rate)**.** ***Error rate*** *(ERR) also called misclassification, is the rate that shows overall how often the model was incorrect:*

$$ERR = \frac{FP + FN}{TP + TN + FP + FN}$$

It is of course: $ACC + ERR = 1$

Definition 11.25 (True Positive Rate)**.** ***True positive rate*** *(TPR) also called sensitivity, recall or hit rate, is the rate that shows how often the model predicts positive when the actual outcome is indeed positive:*

$$TPR = \frac{TP}{TP + FN}$$

Definition 11.26 (False Negative Rate)**.** ***False negative rate*** *(FNR) also called miss rate, is the rate that shows how often the model predicts negative when the actual outcome is positive:*

$$FNR = \frac{FN}{TP + FN}$$

It is of course: $TPR + FNR = 1$

Definition 11.27 (True Negative Rate)**.** ***True negative rate*** *(TNR) also called specificity or selectivity, is the rate that shows how often the model predicts negative when the actual outcome is indeed negative:*

$$TNR = \frac{TN}{TN + FP}$$

Definition 11.28 (False Positive Rate)**.** ***False positive rate*** *(FPR) also called fall-out rate, is the rate that shows how often the model predicts positive when the actual outcome is negative:*

$$FPR = \frac{FP}{TN + FP}$$

It is of course: $TNR + FPR = 1$

Definition 11.29 (Positive Predicted Value)**.** ***Positive predicted value*** *(PPV) also called precision, is the rate that shows how often the model is correct when it predicts positive:*

$$PPV = \frac{TP}{TP + FP}$$

Definition 11.30 (False Discovery Rate)**.** ***False discovery rate*** *(FDR) is the rate that shows how often the model is wrong when it predicts positive:*

$$FDR = \frac{FP}{TP + FP}$$

It is of course: $PPV + FDR = 1$

Definition 11.31 (Negative Predicted Value)**.** ***Negative predicted value*** *(NPV) also called precision, is the rate that shows how often the model is correct when it predicts negative:*

$$NPV = \frac{TN}{TN + FN}$$

Definition 11.32 (False Omission Rate)**.** ***False omission rate*** *(FOR) is the rate that shows how often the model is wrong when it predicts negative:*

$$FOR = \frac{FN}{TN + FN}$$

It is of course: $NPV + FOR = 1$

Definition 11.33 (F_β Score)**.** F_β ***score*** *(FOR) is defined as the harmonic mean of positive predicted value PPV (aka precision) and true positive rate (aka recall) each weighted based on value of* β*:*

$$F_\beta = \frac{(1+\beta^2)\cdot PPV \cdot TPR}{\beta^2 \cdot PPV + TPR} = \frac{(1+\beta^2)\cdot \mathit{precission} \cdot \mathit{recall}}{\beta^2 \cdot \mathit{precission} + \mathit{recall}} = \frac{(1+\beta^2)\cdot TP}{(1+\beta^2)\cdot TP + \beta^2 \cdot FN + FP}$$

The coefficient β is chosen such that recall is considered β times as important as precision. The most commonly used value for β is 1, corresponding to the F_1 where precision and recall are weighted equally:

$$F_1 = \frac{2 \cdot PPV \cdot TPR}{PPV + TPR} = \frac{2 \cdot \text{precission} \cdot \text{recall}}{\text{precission} + \text{recall}} = \frac{2 \cdot TP}{2 \cdot TP + FN + FP}$$

Two other commonly used values for β are 2 and 0.5, corresponding to the F_2 where weighs recall higher than precision (by placing more emphasis on false negatives) and the $F_{0.5}$ measure, which weighs recall lower than precision (by attenuating the influence of false negatives).

Definition 11.34 (Null Error Rate)**.** ***Null error rate*** *is the rate that shows how often a model would be wrong if it always predicts the most frequent type of outcome (either positive or negative depending on the dataset).*

Definition 11.35 (Cohen's Kappa)**.** ***Cohen's kappa*** *is the rate that shows how much better a model performs compared to a hypothetical model that would pick a category completely randomly.*

Appendices

Appendix A

Constrained Optimization

Constrained optimization is the problem of finding a minimum (or maximum) of a function $f(x)$ called the "objective function", subject to a number of constraints of the following types:

- $h_i(x) = 0, \quad i = 1, 2, \ldots$ called "equality constraints"
- $g_i(x) \leq 0, \quad i = 1, 2, \ldots$ called "inequality constraints"

Let's start first with the equality constraints and the we will add inequality constraints.

A.1 Equality Constrained Optimization

The formulation of the optimization problem is to optimize $f(x)$ subject to $h_i(x) = 0, \; i = 1, 2, \ldots$. Let's assume for simplicity only one constraint $h_1(x) = h(x) = 0$. The idea here is that the point that $f(x)$ touches $h(x)$ is the point that $f(x)$ is minimum (or maximum) while the constraint is also valid. At that point $f(x)$ is parallel to $h(x)$ and the tangents $\nabla_{\boldsymbol{w}} f(x)$ and $\nabla_{\boldsymbol{w}} h(x)$ which are perpendicular to $f(x)$ and $h(x)$ respectively, are also parallel to each other. Hence, since $\nabla_{\boldsymbol{w}} f(x)$ and $\nabla_{\boldsymbol{w}} h(x)$ are parallel this translates to:

$$\nabla_{\boldsymbol{w}} f(x) = \mu \nabla_{\boldsymbol{w}} h(x)$$

which is the condition for the $f(x)$ to be minimum (or maximum) while $h(x) = 0$.

Without loss of generality the condition for many constraints reads:

$$\nabla_{\boldsymbol{w}} f(x) = \sum_i \mu_i \nabla_{\boldsymbol{w}} h_i(x)$$

or by bringing everything in one side:

$$\nabla_{\boldsymbol{w}} f(x) - \sum_i \mu_i \nabla_{\boldsymbol{w}} h_i(x) = 0$$

$$\nabla_{\boldsymbol{w}} (f(x) - \sum_i \mu_i h_i(x)) = 0$$

At this point we define the "Lagrangian" as follows:

$$\mathcal{L}(x, \mu_i) = f(x) + \sum_i \mu_i h_i(x)$$

where μ_i are called "Lagrange multipliers". Subsequently the necessary conditions for optimization of $\mathcal{L}$ turns to:

$$\nabla_{\boldsymbol{w}} \mathcal{L} = 0 \;\text{ and }\; \frac{\partial \mathcal{L}}{\partial \mu_i} = 0$$

The solution of this system of equations minimizes (or maximizes) $f(x)$ subject to $h_i(x) = 0, \;\; \forall i$.

A.2 Equality & Inequality Constrained Optimization

Now on top of equality constraints we also have inequality constraints. The formulation of the optimization problem is to optimize $f(x)$ subject to $h_i(x) = 0,\ \ i = 1, 2, \dots$ and $g_i(x) \leq 0,\ \ i = 1, 2, \dots$. Following a similar way of thinking as before, although a bit more technical, we can show (but we won't) that if the following four conditions, called "Karush - Kuhn - Talker conditions" (KKT), hold:

- $h_i(x) = 0, \quad \forall i$
- $g_i(x) \leq 0, \quad \forall i$
- $\lambda_i \leq 0, \quad \forall i$
- $\lambda_i g_i(x) = 0, \quad \forall i$

then there exist constants μ_i and λ_i called "KKT multipliers" such that:

$$\nabla_{\boldsymbol{w}} f(x) = \sum_i \mu_i \nabla_{\boldsymbol{w}} h_i(x) + \sum_i \lambda_i \nabla_{\boldsymbol{w}} g_i(x)$$

By following the same philosophy as for the equality constrained optimization we define the "Lagrangian" as follows:

$$\mathcal{L}(x, \mu_i, \lambda_i) = f(x) + \sum_i \mu_i h_i(x) + \sum_i \lambda_i g_i(x)$$

and the necessary conditions for optimization of $\mathcal{L}$ turns to:

$$\nabla_{\boldsymbol{w}} \mathcal{L} = 0 \ \text{ and } \ \frac{\partial \mathcal{L}}{\partial \mu_i} = 0 \ \text{ and } \ \frac{\partial \mathcal{L}}{\partial \lambda_i} = 0$$

This is the most general case of constrained optimization. If there are no equality constraints $h_i(x)$ then we simply have a theory for inequality constrained optimization. If there are no inequality constraints $g_i(x)$ then the whole theory turns to the equality constrained optimization problem we developed previously and the KKT multipliers turn to Lagrangian multipliers. Finally, if there are no equality constraints $h_i(x)$ neither inequality constraints $g_i(x)$ the theory is just a usual optimization problem where we just find the solution where the derivative of $f(x)$ is zero.

Appendix B

Kernels

Definition B.1 (Kernel)**.** *Let $\bar{x}$ and $\bar{x}'$ be two vectors of space X and Φ a non-linear transformation. We define $\bar{z}$ and $\bar{z}'$ as the transformed vectors $\Phi(\bar{x})$ and $\Phi(\bar{x}')$:*

$$\bar{x} \in X \xrightarrow{\Phi} \bar{z} = \Phi(\bar{x}) \in Z$$

$$\bar{x}' \in X \xrightarrow{\Phi} \bar{z}' = \Phi(\bar{x}') \in Z$$

*We define the **kernel** of space Z as the function that is equal to the inner product of the transformation vectors:*

$$K(\bar{x}, \bar{x}') = \bar{z}^T \bar{z}'$$

Let's for example assume the following non-linear transformation:

$$\bar{x} = (x_1, x_2) \xrightarrow{\Phi} \bar{z} = \Phi(\bar{x}) = (1, x_1^2, x_2^2, \sqrt{2}x_1, \sqrt{2}x_2, \sqrt{2}x_1x_2)$$

$$\bar{x}' = (x_1', x_2') \xrightarrow{\Phi} \bar{z}' = \Phi(\bar{x}') = (1, x_1^{2'}, x_2^{2'}, \sqrt{2}x_1', \sqrt{2}x_2', \sqrt{2}x_1'x_2')$$

Then for the inner product:

$$\bar{z}^T \bar{z}' = 1 + x_1^2 x_1^{2'} + x_2^2 x_2^{2'} + 2x_1x_1' + 2x_2x_2' + 2x_1x_2x_1'x_2'$$

However we can get to the same result by simply defining a Kernel of the form:

$$\begin{aligned} K(\bar{x}, \bar{x}') &= (1 + \bar{x}\bar{x}')^2 \\ &= (1 + x_1x_1' + x_2x_2')^2 \\ &= 1 + x_1^2 x_1^{2'} + x_2^2 x_2^{2'} + 2x_1x_1' + 2x_2x_2' + 2x_1x_2x_1'x_2' \end{aligned}$$

Hence by knowing the Kernel of a space Z of some non-linear transformation Φ we can compute inner products without the need of transforming vectors from X to Z.

The kernel trick is to use this idea in the opposite direction. Namely, to assume that a function $K(\bar{x}, \bar{x}')$ is the kernel of some space Z for some non-linear transformation Φ and to compute inner products without even knowing the transformation.

The question that arises is how do we know that some function $K(\bar{x}, \bar{x}')$ is actually the kernel of a space Z. There are three approaches to this problem:

1. By construction (as we did in the example above).
2. By Mercer's condition that states that $K(\bar{x}, \bar{x}')$ is a valid kernel for some space Z if

$$\int K(\bar{x}, \bar{x}')g(\bar{x})g(\bar{x}')d\bar{x}d\bar{x}' \geq 0 \ \ \forall \ \text{ square integrable functions } \ g(\bar{x})$$

3. Sometimes we don't care if $K(\bar{x}, \bar{x}')$ is a valid kernel for some space Z as long as it does the job.

Appendix C

Convolution

Definition C.1 (Convolution)**.** *Convolution is a mathematical operation on two functions f and g that produces a third function expressing how the shape of one is modified by the other. The term convolution refers to both the result function and to the process of computing it. It is defined as the integral of the product of the two functions after one is reversed and shifted.*

$$(f * g)(t) = \int_{-\infty}^{\infty} f(\tau)g(t-\tau)d\tau$$