Neuroinformatics Script

Devrim Celik

January 6, 2020

Contents

Introduction

As you will probably experience in your career as cognitive scientist, one of the most typical problems is the following: You measured data, e.g. recording the spiking of a neuron. This data is usually subject to random noise. In some cases, even though the data is ”corrupted” by this noise, you will still be able to make out the *true underlying function* that generated the data at hand; but for most cases the noise will be intense enough to completely obscure the data for the human eye.

In this course we will give you all the tools you need to try to approximate the function in question.

First of all we will deal with the **Basics of Probability Theory**. As you can see, we located this chapter at the very end of this script. The mathematical nature of this chapter can be quite frustrating when trying to understand the concepts by your own. This is why we encourage the reader to attend the practice sessions, whose goal it is to introduce this chapter in a more favourable atmosphere. From this point onward, the last chapter is meant as a chapter of reference.

After the reader is familiar with basic probability theory, we will immediately start to use those basics for building and understanding **Graphical Models**; to be more specific **Bayesian networks**. In this chapter we will discuss what it means for two random variables (or two events) to be independent. Furthermore, we will develop two strategies for determining this property, a graphical and an analytical one.

Next we will introduce the idea of **Maximum Likelihood Estimation**. There we will tackle the previously referenced problem: Given some data we collected, we will find out which parameters were (most likely) used in the underlying function. We will also talk about the quality of such estimations by introducing concepts such as the **bias of an estimator** and the **observed Fisher information**.

Next we will dive deeper into this direction by talking about **Linear Regression**: We will deal with the basic idea and more complex versions that make use of a variety of basis function. In addition to this, we will also talk about overfitting of a model, a very common problem in machine learning and model fitting in general.

As we will see later, one can fit multiple models on the same data, resulting in many models whose goal it is to approximate the real underlying function. This is why we will then talk about **Model Validation and Comparison**.

Next we will try to understand the concepts behind **Exponential Families**, whose goal it is to be a general *version* of many different probability distributions.

Lastly, we will talk about a different way of estimating parameters which makes use of Bayes’ theorem, called **Bayesian Regression**

For whom is this course useful?

My personal answer to this question: Everybody!
The reason behind my answer is simple: Although there might be differences in what we are mainly interested in, in the end we are all scientists. And as scientists we need to understand data. This holds for individuals interested in machine learning and artificial intelligence, as well as for those interested in philosophy and neuroscience. All of us will, at one point or the other, use empirical methods. As it is becoming more and more obvious over time: Many of today's "scientists" are ill-equipped to deal with such problems. This is why I strongly advise everybody to take this course and take as much as you can from it.

Chapter 1

Graphical Models

In this chapter, we'll go slightly off the path of model fitting in order to deepen our understanding of probabilistic relations. Graphical models are a way of visualizing probabilistic relations between variables and they make our lives a lot easier for some operations. They are comprised of nodes (variables) and links (their probabilistic relations). The two main types of graphical models are Bayesian networks, in which links are directed i.e. they indicate causal relations, and Markov random fields in which links are undirected. In this lecture, we will leave the Markov random fields aside and concern ourselves with Bayesian networks only.

1.1 Bayesian Network

A Bayesian network is a directed acyclic graph (DAG), i.e.

- it is composed of nodes and links,
- the links are directed,
- there are no cycles.

Each node represents a random variable (or a group of random variables) and the links express probabilistic relationships between these variables. Note that the direction does not imply causation, i.e. just because X has an arrow pointing on Y, this does not mean that X influences Y. The interesting thing about a Bayesian network is that there exists a relationship between joint probabilities and Bayesian networks: The graph captures the way in which the joint distribution over all of the random variables can be decomposed into a product of factors, each depending only on a subset of the variables.

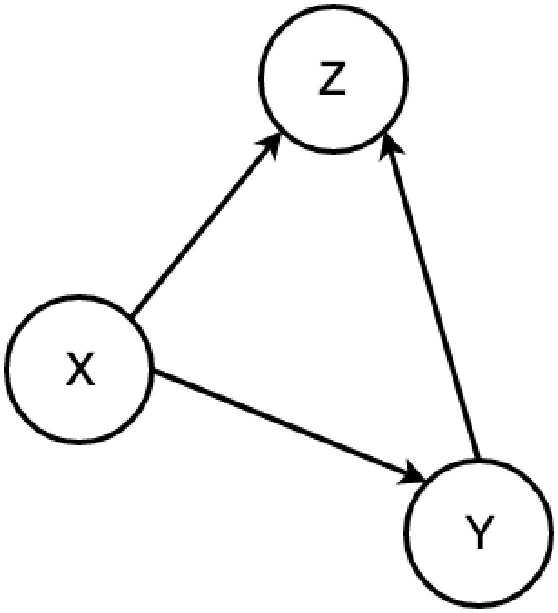
In order to motivate the use of Bayesian networks to describe probability distributions, consider first some arbitrary join probability $\mathbb{P}(X \cap Y \cap Z)$ over the random variables X, Y and Z. Using the product rule, we can rewrite the joint distributions as

$$\mathbb{P}(X \cap Y \cap Z) = \mathbb{P}(Z|X \cap Y) \cdot \mathbb{P}(X \cap Y) = \mathbb{P}(Z|X \cap Y)\mathbb{P}(Y|X)\mathbb{P}(X)$$

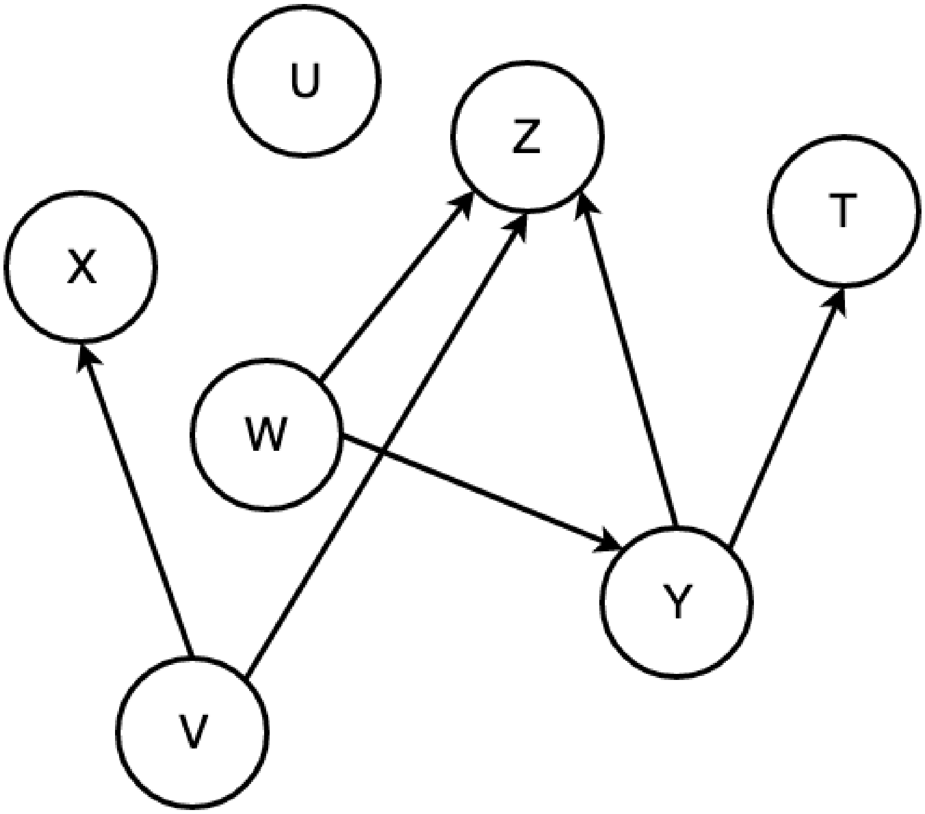
Note that this way of decomposing the joint probability holds for any possible choice of the joint distribution because we only applied the product rule.
Here comes the graphical part: We will now try to represent the right hand side of the above equation using a graphical model. The easiest way to do so is to follow this scheme:

1. We introduce a node for each of the random variables X, Y, and Z.
2. Then for each conditional distribution, we add directed links (arrows) to the graph from the nodes corresponding to the variables on which the distribution is conditioned.
   Thus for the factor $\mathbb{P}(Z|X \cap Y)$, there will be links from nodes X and Y to node Z, whereas for the factor $\mathbb{P}(X)$ there will be no incoming links.

This scheme would yield the following graph:



If there is a link going from a node X to a node Y, then we say that node X is the **parent** of node Y, and we say that node B is the **child** of node X.
Note that we freely chose the specific ordering here. Since in the joint probability $\mathbb{P}(X \cap Y \cap Z)$ all three variables are equal, we could have chosen any other hierarchy between them instead (this however would have resulted in a different graph).
For the sake of repetition, consider the following graph:



Using what we learned before, we can deduce:

$$\mathbb{P}(T \cap U \cap V \cap W \cap X \cap Y \cap Z) = \mathbb{P}(T|Y)\mathbb{P}(U)\mathbb{P}(V)\mathbb{P}(W)\mathbb{P}(X|V)\mathbb{P}(Y|W)\mathbb{P}(Z|V \cap W \cap Y)$$

Lastly, we can have a look at the *factorization properties* of Bayesian networks. We saw that each fragment follows the principle $\mathbb{P}(Child|Parent)$. Using this fact, we can generally state

$$\mathbb{P}(X_1, \dots, X_n) = \prod_{k=1}^{n} \mathbb{P}(X_k|pa_k),$$

where pa_k indicates the intersection of all parent nodes of X_k.

1.2 Independence in Bayesian networks

We already introduced the notion of independence between two random variables X and Y. We call them independent if

$$\mathbb{P}(X \cap Y) = \mathbb{P}(X) \cdot \mathbb{P}(Y).$$

We indicate this relationship by denoting $X \perp\!\!\!\perp Y$ if X and Y are independent and $X \not\perp\!\!\!\perp Y$ if they are dependent.
There is also the notion of **conditional independence**, where we call random variables X and Y conditionally independent on a third random variable Z if

$$\mathbb{P}(X \cap Y|Z) = \mathbb{P}(X|Z) \cdot \mathbb{P}(Y|Z).$$

Then we write $X \perp\!\!\!\perp Y|Z$ if X and Y are conditionally independent on Z and $X \not\perp\!\!\!\perp Y|Z$ if they are not.

Next we will deal with the matter of proving independence of random variables in a Bayesian network. In this lecture we will cover two ways to do so:

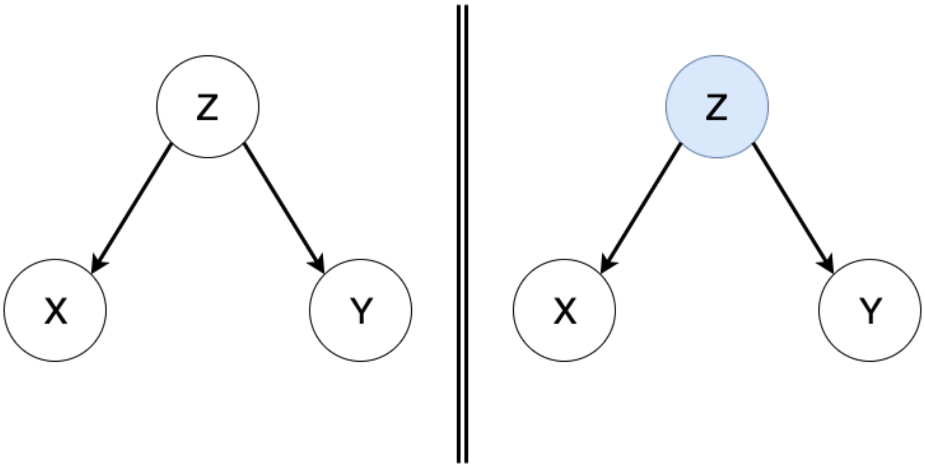
- **D-separation**: This method can be viewed as a short cut for the analytical method. Basically we translate independence into visual properties. Then we use the D-separation method of determining whether independence between random variables exists.

- The **analytical** method: Here we start by writing the joint probability of the graph at hand and then try to reformulate this equation to end up with the desired independence statement.

1.2.1 D-separation method

We will use the D-separation method to prove/disprove the independence between two random variables in a Bayesian network. The method considers all possible paths connecting the two random variables in question. For each possible path between them, the D-separation method tries to determine whether there is a *flow of information* through this path. If none of the paths allows said flow of information, the random variables are independent.
How do we find out whether a path allows a flow of information? Basically we go through each node in this path (from start to end) and look at the "connections" at use when considering the different possibilities the directions can take. To clarify what is meant by this statement, we will consider all three cases that are possible:

Common Cause



In this example we are interested whether X and Y are independent. So we consider the nodes that connect them, i.e. here only random variable Z. We call this scenario *common cause*, because the direction of the edges indicates that Z causes X and Y. This scenario is also known as **tail-to-tail**. Technically, there are two possibilities: The first one (left graph) where we do not know anything about Z; the second one (right graph) where the value of Z is already fixed/observed.
To analyse this two scenarios, we will use the joint probability (and basically perform the later introduced analytical method):

Z is observed In this case, we can determine that the joint probability is described by

$$\mathbb{P}(X, Y, Z) = \mathbb{P}(X|Z)\mathbb{P}(Y|Z)\mathbb{P}(Z).$$

So we want to "reformulate" this equation, such that we are left with $\mathbb{P}(X, Y|Z) = \mathbb{P}(X|Z)\mathbb{P}(Y|Z)$, because Z is already given. Looking at the left-hand side, we want to have a conditional probability, instead of this joint probability. To do so, we will divide both sides of the equation by $P(Z)$:

$$\begin{aligned} \frac{\mathbb{P}(X, Y, Z)}{\mathbb{P}(Z)} &= \frac{\mathbb{P}(X|Z)\mathbb{P}(Y|Z)\mathbb{P}(Z)}{\mathbb{P}(Z)} \\ \Leftrightarrow \mathbb{P}(X, Y|Z) &= \frac{\mathbb{P}(X|Z)\mathbb{P}(Y|Z)\mathbb{P}(Z)}{\mathbb{P}(Z)} \\ \Leftrightarrow \mathbb{P}(X, Y|Z) &= \mathbb{P}(X|Z)\mathbb{P}(Y|Z) \end{aligned}$$

and thus X and Y are independent given Z when we have a common cause scenario; or said differently the flow of information is blocked.

Z **not observed** In this case, we can determine that the joint probability is described by

$$\mathbb{P}(X, Y, Z) = \mathbb{P}(X|Z)\mathbb{P}(Y|Z)\mathbb{P}(Z).$$

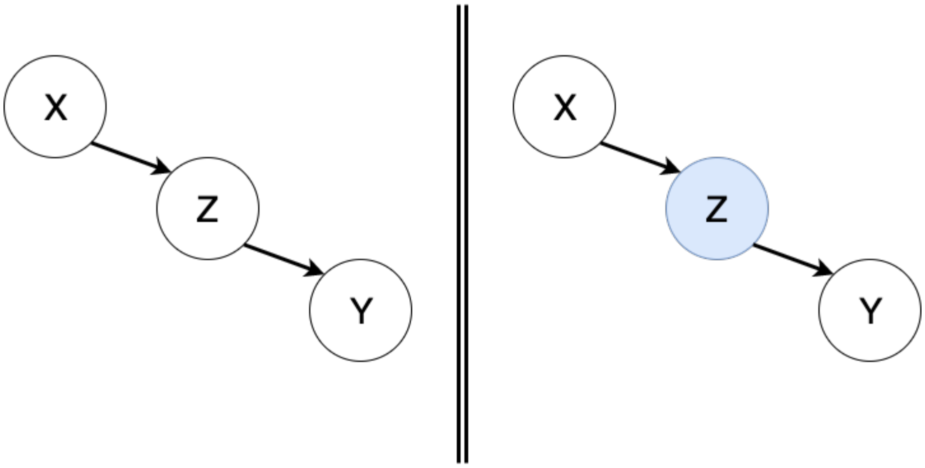
So we want to "reformulate" this equation such that we are left with $\mathbb{P}(X, Y) = \mathbb{P}(X)\mathbb{P}(Y)$. To get rid of the left-hand side Z in the joint probability is to make use of the **law of total probability**: We iterate through every possible value Z can take on both sides (we say that we **marginalize of** Z):

$$\begin{aligned}\sum_Z \mathbb{P}(X, Y, Z) &= \sum_Z \mathbb{P}(X|Z)\mathbb{P}(Y|Z)\mathbb{P}(Z)\\ \Leftrightarrow \mathbb{P}(X, Y) &= \sum_Z \mathbb{P}(X|Z)\mathbb{P}(Y|Z)\mathbb{P}(Z)\end{aligned}$$

The problem we have now is the right hand side of the equation: There is no theorem of probability theory that states that $\sum_Z \mathbb{P}(X|Z)\mathbb{P}(Y|Z)\mathbb{P}(Z) = \mathbb{P}(X)\mathbb{P}(Y)$ (in general). So for the common cause with unobserved Z, X and Y are not independent, i.e. the flow of information is not blocked.

Conclusion So how can this be interpreted? Actually it is quite intuitive: Imagine the scenario where X describes the fact that **grass is wet** and Y represent whether **my shoes are wet**. Z stands for the fact **whether it has rained** or not.
Looking at the case where we do not observe Z, i.e. we do not know whether it rained, does knowledge about X change the probability of Y? It does; consider that you know that the grass is actually wet. This implies that the probability that it has rained last night increases, which in turn increases the probability that my shoes will be wet. On the other hand, when you already know that it has /hasn't rained, the knowledge about X does not change the probability of Y!

Causal Chain



Again we are interested in whether X and Y are independent. We call this scenario the *causal chain* or **head-to-tail**. Again we will consider one cases where Z is fixed/observed, and one where it is not:

Z is observed We start off by writing down the joint probability of the graph:

$$\mathbb{P}(X, Y, Z) = \mathbb{P}(X)\mathbb{P}(Z|X)\mathbb{P}(Y|Z).$$

Since our end goal includes a conditional probability on the left hand side, we divide both sides by $\mathbb{P}(Z)$:

$$\begin{aligned}\frac{\mathbb{P}(X, Y, Z)}{\mathbb{P}(Z)} &= \frac{\mathbb{P}(X)\mathbb{P}(Z|X)\mathbb{P}(Y|Z)}{\mathbb{P}(Z)} \\ \mathbb{P}(X, Y|Z) &= \frac{\mathbb{P}(X)\mathbb{P}(Z|X)\mathbb{P}(Y|Z)}{\mathbb{P}(Z)}.\end{aligned}$$

Again, we will use another one of the probability theorems: Bayes' rule. Here we can see, that $\frac{\mathbb{P}(Z|X)\mathbb{P}(X}{\mathbb{P}(Z)} = \mathbb{P}(X|Z)$:

$$\mathbb{P}(X, Y|Z) = \mathbb{P}(X|Z)\mathbb{P}(Y|Z).$$

So a head-to-tail connection with an observed variables indicates that the random variables are independent, i.e. the flow information is blocked.

Z not observed We start off by writing down the joint probability of the graph:

$$\mathbb{P}(X, Y, Z) = \mathbb{P}(X)\mathbb{P}(Z|X)\mathbb{P}(Y|Z).$$

Now we want the left side without the Z, so again we marginalize over Z on both sides:

$$\begin{aligned}\sum_Z \mathbb{P}(X,Y,Z) &= \sum_Z \mathbb{P}(X)\mathbb{P}(Z|X)\mathbb{P}(Y|Z)\\ \Leftrightarrow \mathbb{P}(X,Y) &= \sum_Z \mathbb{P}(X)\mathbb{P}(Z|X)\mathbb{P}(Y|Z)\\ &= \mathbb{P}(X)\sum_Z \mathbb{P}(Z|X)\mathbb{P}(Y|Z)\end{aligned}$$

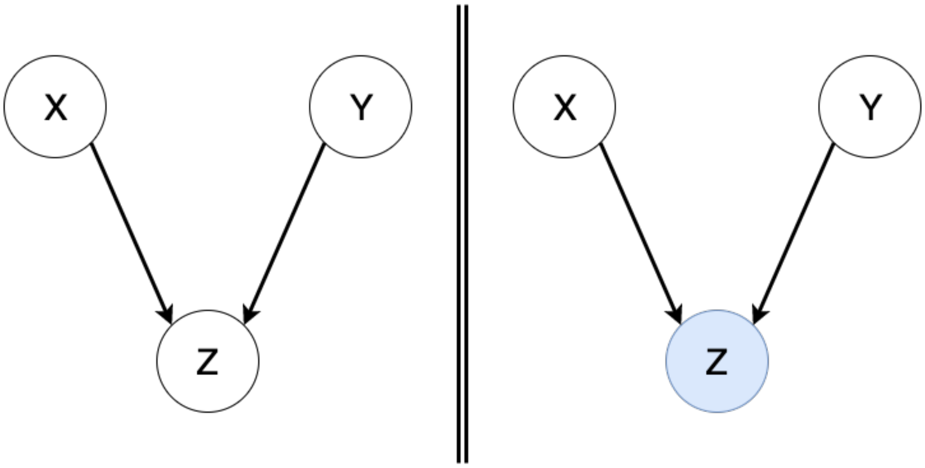
and again we encounter a situation, where the left hand side does generally not factorize to what we need, i.e. $\sum_Z \mathbb{P}(Z|X)\mathbb{P}(Y|Z) \neq \mathbb{P}(Y)$. So a head-to-tail connection with an unobserved Z does not stop the flow of information, i.e. X and Y are not independent.

Conclusion Again, how can we make intuitive sense of these results. Let us consider an example with the following assignments:

- X: I met an old friend.
- y: I am hung over the next day.
- Z: I go into a bar.

Assuming you know nothing about any of those, does knowledge about X change the probability of Y? Yes! If e.g. I met my friend it is more likely that we went into a bar and thus more likely that I am hung over the next day. On the other hand, if it is already clear whether I went to a bar this evening, knowledge about meeting an old friend does not change anything because it is already clear whether I went to a bar or not.

Common Effect



Again we are interested in whether X and Y are independent. We call this scenario the *common effect* or **head-to-head**. Again we will consider one cases where Z is fixed/observed and one where it is not:

Z **not observed** We start off by writing down the joint probability of the graph:

$$\mathbb{P}(X, Y, Z) = \mathbb{P}(X)\mathbb{P}(Y)\mathbb{P}(Z|X, Y).$$

Now we want the left side without the Z, so again we marginalize over Z on both sides:

$$\begin{aligned}\sum_Z \mathbb{P}(X, Y, Z) &= \sum_Z \mathbb{P}(X)\mathbb{P}(Y)\mathbb{P}(Z|X, Y)\\ \Leftrightarrow \mathbb{P}(X, Y) &= \sum_Z \mathbb{P}(X)\mathbb{P}(Y)\mathbb{P}(Z|X, Y)\\ &= \mathbb{P}(X)\mathbb{P}(Y)\sum_Z \mathbb{P}(Z|X, Y)\end{aligned}$$

And to continue from here, we have to use the law of total probability again: If $\sum_Z \mathbb{P}(Z|X, Y)$ factorizes to 1, this would work out. But we only defined the law of total probability for joint probabilities, not for the conditional scenario we encounter here. But of course, we can rewrite the conditional probability and then apply the law of total probability:

$$\begin{aligned}\sum_Z \mathbb{P}(Z|X, Y) &= \sum_Z \frac{\mathbb{P}(Z, X, Y)}{\mathbb{P}(X, Y)}\\ &= \frac{1}{\mathbb{P}(X, Y)}\sum_Z \mathbb{P}(Z, X, Y)\\ &= \frac{1}{\mathbb{P}(X, Y)}\mathbb{P}(X, Y)\\ &= 1\end{aligned}$$

and thus X and Y are independent if Z is not observed; or said otherwise the flow of information is blocked.
(You might wonder how we can justify the first step of the above equation. To simplify, you have to understand that $\mathbb{P}(A, B) = \mathbb{P}(A \cap B)$. So if we set $G = X \cap Y$, then suddenly the equation makes much more sense: $\sum_Z \mathbb{P}(Z|G) = \sum_Z \frac{\mathbb{P}(Z,G)}{\mathbb{P}(G)}$.)

Z **is observed** We start off by writing down the joint probability of the graph:

$$\mathbb{P}(X, Y, Z) = \mathbb{P}(X)\mathbb{P}(Y)\mathbb{P}(Z|X, Y).$$

and dividing both sides by $\mathbb{P}(Z)$:

$$\begin{aligned}\frac{\mathbb{P}(X,Y,Z)}{\mathbb{P}(Z)} &= \frac{\mathbb{P}(X)\mathbb{P}(Y)\mathbb{P}(Z|X,Y)}{\mathbb{P}(Z)} \\ \mathbb{P}(X,Y|Z) &= \frac{\mathbb{P}(X)\mathbb{P}(Y)\mathbb{P}(Z|X,Y)}{\mathbb{P}(Z)}.\end{aligned}$$

Here there is no general way to prove that $\frac{\mathbb{P}(X)\mathbb{P}(Y)\mathbb{P}(Z|X,Y)}{\mathbb{P}(Z)} = \mathbb{P}(X|Z)\mathbb{P}(X|Y)$. So a head-to-head connection with an observed variables indicates that the random variables are not independent, i.e. the flow information is not blocked.

Conclusion Again, how can we make intuitive sense of these results. Let us consider an example with the following assignments:

- X: It has rained.
- Y: The sprinklers went off.
- Z: The grass is wet.

Assuming you know nothing about any of those, does knowledge about X change the probability of Y? No, not really: Let us assume that it has rained last night; this does not really give me any information about the sprinkler, i.e. X and Y are independent. Now consider this: We know that the grass is wet (Z is observed) and then we find out it has not rained; suddenly the probability of the sprinklers going off increases drastically! This means given that Z is observed knowledge about X influences the probability of Y, i.e X and Y are not independent. Note that X and Y are also not independent if Z is not observed but one of Z's descendants is (A is a descendant of B, if there is a way from B to A following the directions of the edges).

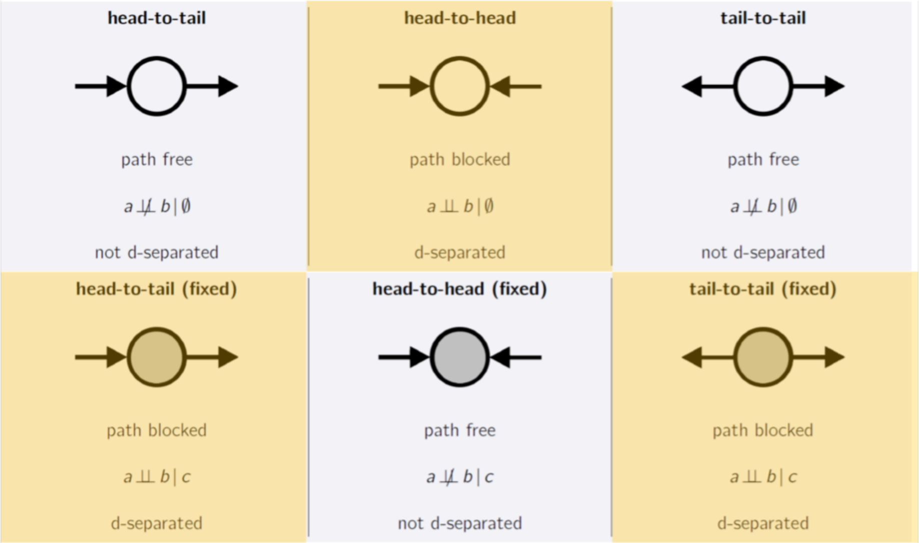
Algorithm

Imagine you are given a Bayesian network and you are interested in whether two random variables X and Y are independent given C, the set of all observed random variables. Then you are interested in proving $X \perp\!\!\!\perp Y|C$.

1. First check that X, Y and C are disjoint.
2. Next, go through each possible path from X to Y and check whether it is blocked or not. A path is blocked if along the path you can find a random variable Z,
   - that is contained in C, i.e. Z is observed, and the edges meet either **tail-to-tail** or **head-to-tail** .
   - which is not contained in C and none of its descendants are contained in C and the edges meet **head-to-head** .

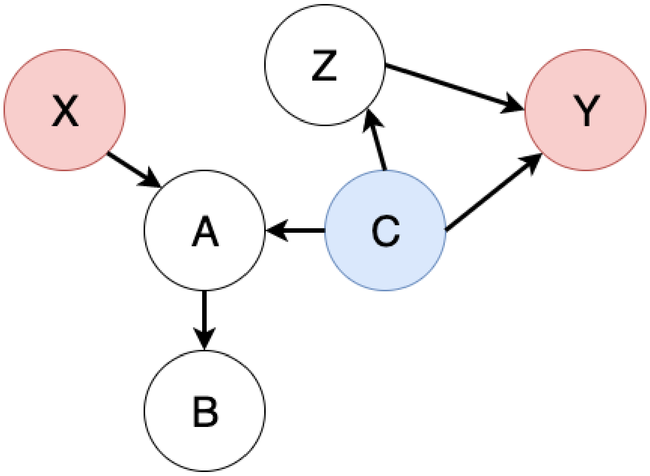
3. If all paths from X to Y are blocked, X is said to be d-separated from Y by C. This implies the wanted conditional independence.

Consider the following graph as some kind of summary of all possible cases



head-to-tail	head-to-head	tail-to-tail
path free	path blocked	path free
$a \not\perp\!\!\!\perp b \mid \emptyset$	$a \perp\!\!\!\perp b \mid \emptyset$	$a \not\perp\!\!\!\perp b \mid \emptyset$
not d-separated	d-separated	not d-separated
head-to-tail (fixed)	**head-to-head (fixed)**	**tail-to-tail (fixed)**
path blocked	path free	path blocked
$a \perp\!\!\!\perp b \mid c$	$a \not\perp\!\!\!\perp b \mid c$	$a \perp\!\!\!\perp b \mid c$
d-separated	not d-separated	d-separated

Example Let us do a short example. For the following example, determine whether X and Y are independent using d-separation given the Bayesian network below:



Let us go through all steps one by one:

1. We notice that C is the only observed variable.

2. So let us check all possible paths from X to Y. In this case, there are two possibilities. Let us go through each connecting node and check whether both paths are blocked:

   - $X \to A \to C \to Y$:
     - A: Is a head-to-head node. Neither A, nor its descendent B is observed. This means A is blocking this path. (Note that in theory, you can already stop here: If one node is blocking, the whole path is blocked. But we will continue for the purpose of education).
     - C: Is a tail-to-tail node that observed, thus it is also blocking.
   - $X \to A \to C \to Z \to Y$.
     - A: Is a head-to-head node. Neither A, nor its descendent B are observed. This means A is blocking this path. (Note that in theory, you can already stop here: If one node is blocking, the whole path is blocked. But we will continue for the purpose of education).
     - C: Is a tail-to-tail node that observed, thus it is also blocking.
     - Z: Is a tail-to-tail node that unobserved, thus it is not blocking.

3. Because both paths are blocked X and Y are independent given C.

1.2.2 Analytical method for proving independence

A small fun fact: We have basically performed the analytical method already. We used it every time for those scenarios, when trying to disprove/prove independence. Often considered the harder of the two, it will become quite easy with a bit of practice. In general, there are few ideas to keep in mind when performing the analytical method:

1. There are three things you can do to reformulate terms: Using the definition of the conditional probability, marginalizing (law of total probability) and using Bayes' theorem.

2. The most common mistake is incorrect marginalization: As we showed early, we can use marginalization (over e.g. Z) in terms such as

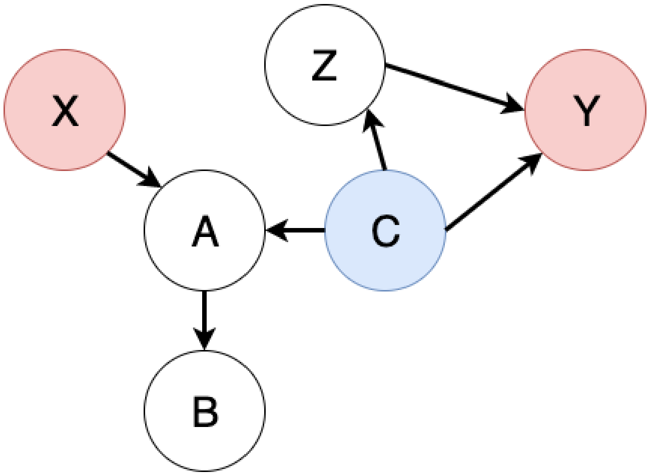
   - $\sum_Z \mathbb{P}(Z) = 1$,
   - $\sum_Z \mathbb{P}(Z, X) = \mathbb{P}(X)$,
   - $\sum_Z \mathbb{P}(Z|X) = 1$,

   but we are not allowed to marginalize over terms that include Z in the conditional part, e.g.

   - $\sum_Z \mathbb{P}(X|Z) \neq \mathbb{P}(X)$,
   - $\sum_Z \mathbb{P}(X|Z, Y) \neq \mathbb{P}(X|Y)$.

3. A good idea is to start off by writing down the joint probability of the graph. Then divide by all random variables that are observed (so to get the correct conditional probability on the left-hand side of the equation) and then try to gather all the parts you need.

Example We will use the same example as for d-separation (note that this example is not trivial and serves as demonstrating as much as possible): Determining whether X and Y are independent using analytical methods given the Bayesian network below:



We start off by writing down the joint probability and dividing by $\mathbb{P}(C)$:

$$\begin{aligned}
&\mathbb{P}(A,B,C,X,Y,Z) = \mathbb{P}(A|C,X)\mathbb{P}(B|A)\mathbb{P}(C)\mathbb{P}(X)\mathbb{P}(Y|C,Z)\mathbb{P}(Z|C)\\
\Leftrightarrow\ &\frac{\mathbb{P}(A,B,C,X,Y,Z)}{\mathbb{P}(C)} = \frac{\mathbb{P}(A|C,X)\mathbb{P}(B|A)\mathbb{P}(C)\mathbb{P}(X)\mathbb{P}(Y|C,Z)\mathbb{P}(Z|C)}{\mathbb{P}(C)}\\
\Leftrightarrow\ &\mathbb{P}(A,B,X,Y,Z|C) = \frac{\mathbb{P}(A|C,X)\mathbb{P}(B|A)\mathbb{P}(C)\mathbb{P}(X)\mathbb{P}(Y|C,Z)\mathbb{P}(Z|C)}{\mathbb{P}(C)}\\
\Leftrightarrow\ &\mathbb{P}(A,B,X,Y,Z|C) = \mathbb{P}(A|C,X)\mathbb{P}(B|A)\mathbb{P}(X)\mathbb{P}(Y|C,Z)\mathbb{P}(Z|C)
\end{aligned}$$

Let us again remember our goal: At the end, we want to have this equation reduced to $\mathbb{P}(X,Y|C) = \mathbb{P}(X|C)\mathbb{P}(Y|C)$. This means that at one point or the other we will have to marginalize over A, B and Z. We can see, that B is only used once (not in the conditional part) and even in the graph it ”does not seem to important”, so we continue by marginalizing over B:

$$\begin{aligned}
\Leftrightarrow \sum_B \mathbb{P}(A,B,X,Y,Z|C) &= \sum_B \mathbb{P}(A|C,X)\mathbb{P}(B|A)\mathbb{P}(X)\mathbb{P}(Y|C,Z)\mathbb{P}(Z|C)\\
\Leftrightarrow \mathbb{P}(A,X,Y,Z|C) &= \mathbb{P}(A|C,X)\mathbb{P}(X)\mathbb{P}(Y|C,Z)\mathbb{P}(Z|C)
\end{aligned}$$

At this point you might notice the following:

$$\begin{aligned}\mathbb{P}(Y|C,Z)\mathbb{P}(Z|C) &= \frac{\mathbb{P}(Y,C,Z)}{\mathbb{P}(C,Z)}\frac{\mathbb{P}(Z,C)}{\mathbb{P}(C)}\\ &= \frac{\mathbb{P}(Y,C,Z)}{\mathbb{P}(C)}\\ &= \mathbb{P}(Y,Z|C)\end{aligned}$$

Using this, we arrive at

$$\Leftrightarrow \mathbb{P}(A,X,Y,Z|C) = \mathbb{P}(A|C,X)\mathbb{P}(X)\mathbb{P}(Y,Z|C)$$

Here we marginalize over Z, arriving at

$$\begin{aligned}\Leftrightarrow \sum_Y \mathbb{P}(A,X,Y,Z|C) &= \sum_Y \mathbb{P}(A|C,X)\mathbb{P}(X)\mathbb{P}(Y,Z|C)\\ \Leftrightarrow \mathbb{P}(A,X,Y|C) &= \mathbb{P}(A|C,X)\mathbb{P}(X)\mathbb{P}(Y|C).\end{aligned}$$

Obviously, we somehow have to transform $\mathbb{P}(A|C,X)\mathbb{P}(X)$ into $\mathbb{P}(A,X|C)$ and marginalize over A. If we rewrite the term:

$$\mathbb{P}(A|C,X)\mathbb{P}(X) = \frac{\mathbb{P}(A,C,X)}{\mathbb{P}(C,X)}\mathbb{P}(X),$$

we can recognize the following: If C and X were independent, then we could rewrite $\mathbb{P}(C,X)$ as $\mathbb{P}(C)\mathbb{P}(X)$, resulting in $\frac{\mathbb{P}(A,C,X)}{\mathbb{P}(C)} = \mathbb{P}(A,X|C)$. How can we prove this independence? Well, theoretically we would have to use the analytical methods again to prove this. But luckily the random variable C is fixed; this means that its value is already fixed and thus knowledge about X does not influence C in the least, i.e. $\mathbb{P}(C|X) = \mathbb{P}(C)$ (X and C are independent. Using this, we end up with

$$\Leftrightarrow \mathbb{P}(A,X,Y|C) = \mathbb{P}(A,X|C)\mathbb{P}(Y|C),$$

where we simply marginalize over A, resulting in exactly what we want

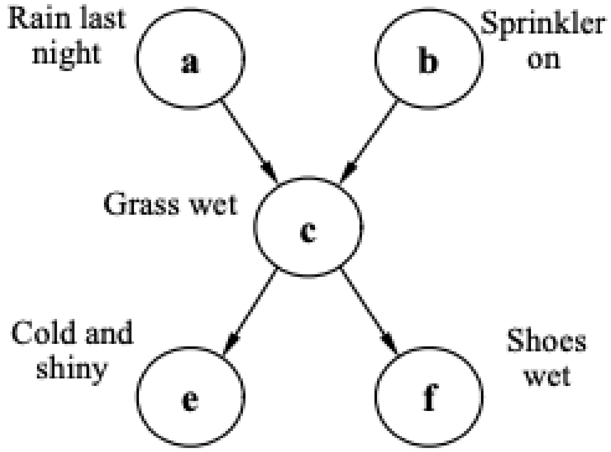
$$\begin{aligned}\Leftrightarrow \sum_A \mathbb{P}(A,X,Y|C) &= \sum_A \mathbb{P}(A,X|C)\mathbb{P}(Y|C)\\ \Leftrightarrow \mathbb{P}(X,Y|C) &= \mathbb{P}(X|C)\mathbb{P}(Y|C)\end{aligned}$$

and thus proving the independence between X and Y given C.

1.3 Explaining away

In this section, we will introduce the idea of **explaining away**. Explaining away is a common pattern of reasoning stating that the confirmation of one cause of an observed or believed event reduces the need to invoke alternative

causes. The opposite of explaining away can also occur; here the confirmation of a cause increases the belief in another cause.
Let us consider the following (popular) examples:



Here it could have rained (A) and/or the sprinkler could have been on last night (B). Both of these could have caused that the grass is wet (C). This in turn can cause that the grass is cold and shiny (E) as well as wet shoes (F).
For the explaining away "setup", nodes A, B and C are important for us. We will assume that we observed C; here it means that we know that the grass is wet. Now either it has rained last night or the sprinklers were on or both. Now (in the next step) we will assume we also observed that it has rained last night. In a probabilistic reasoning scheme, this leads to a reduced probability of the sprinkler hypothesis even though the possibility of simultaneous sprinkling and raining is allowed. This phenomenon is called explaining away.
The opposite deals with lower nodes C, E and F. Here the observation of one effect, e.g. E, is evidence for C and hence increases the likelihood the other effect F.

Chapter 2

Maximum Likelihood Estimation

2.1 Idea

Up until now, we mainly dealt with probability theory. This means that we were given the exact probability space and we are interested in the probability of certain events. Next we will deal with statistics, in which the idea is the other way around: We are given probabilities of events (i.e. we are given results of a random experiment) and are then interested what the probability space looks like. To be more precise, we often know what kind of distribution we are dealing with, but not the exact parameter of that distribution.

So to formulate this problem in other words, this means **given the data, we are interested in the parameters** of the underlying model that generated the data. θ will denote the unknown parameters of the distribution, Θ represents the space of all possible values θ can take and $\vec{x}$ will denote the observed data, i.e. $\vec{x} = (x_1, x_2, ..., x_n)$.

Mathematically, one way of solving this problem is to find the θ which maximizes $\mathbb{P}(\theta|\vec{x})$; with a fixed $\vec{x}$, we are looking for the θ that is most probable given the observations we made. So mathematically, we look for θ^*, such that

$$\theta^* = \arg\max_{\theta \in \Theta} \mathbb{P}(\theta|\vec{x}).$$

Since we can rewrite $\mathbb{P}(\theta|\vec{x})$ to

$$\mathbb{P}(\theta|\vec{x}) = \frac{\mathbb{P}(\vec{x}|\theta)\mathbb{P}(\theta)}{\mathbb{P}(\vec{x})}$$

by using Bayes' rule, we call θ^* the **maximum a posteriori estimate of the real parameter θ**.

Finding the θ^* that maximizes the posterior $\mathbb{P}(\theta|\vec{x})$ is equivalent to maximizing

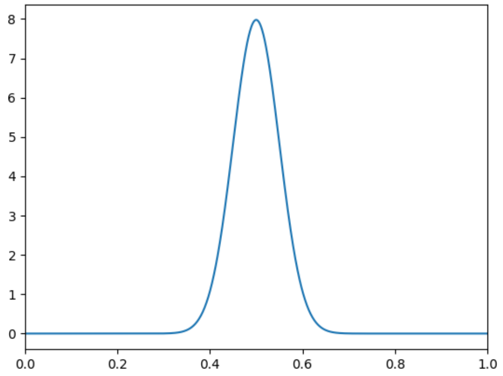
$\frac{\mathbb{P}(\vec{x}|\theta)\mathbb{P}(\theta)}{\mathbb{P}(\vec{x})}$, i.e.

$$\begin{aligned}\theta^* &= \arg\max_{\theta\in\Theta} \mathbb{P}(\theta|\vec{x}) \\ &= \arg\max_{\theta\in\Theta} \frac{\mathbb{P}(\vec{x}|\theta)\mathbb{P}(\theta)}{\mathbb{P}(\vec{x})}.\end{aligned}$$

The first thing we can do is to leave $\mathbb{P}(\vec{x})$ out of this equation, because its value is independent of θ, i.e.

$$= \arg\max_{\theta\in\Theta} \mathbb{P}(\vec{x}|\theta)\mathbb{P}(\theta),$$

leaving us with the product of the **likelihood** and the **prior**. Just to refresh, the way Bayes' formula can be interpreted is that the prior $\mathbb{P}(\theta)$ represents your prior belief (prior to seeing the data) about which value θ of the distribution is how probable. For example, lets say I present you a coin and throw it 100 times. You prior belief about the coin might look something like this



It seems most rational for us to assume a coin is fair because most coins are somewhat fair. Using only the prior distribution above we would choose $\theta = 0,5$. But that is without considering the data. The data is encoded in the likelihood $\mathbb{P}(\vec{x}|\theta)$. This term tells us how probable a certain event is given that we have fixed θ.
Now let us assume that from the 100 coin tosses 90 ended up showing heads and only 10 tails. Using this information, let us compare two possible values for θ: $0,5$ and 1.

- $\theta = 0,5$: For this value the prior has a maximal value but the likelihood is quite small (it is not really probable to get this result with a fair coin).
- $\theta = 1$: For this value the prior has a minimal value but the likelihood is quite big.

For a coin it is quite easy to come up with a good prior distribution (in our case a normal distribution centred around $0,5$), but for most cases it is incredibly hard to define a prior distribution. This fact is one of the biggest points of criticism against Bayesian statistics, because defining a prior seems very random.
One strategy when trying to define a prior distribution for an unknown parameter θ is to use a uniform distribution: Since we might not have any idea about which value is how probable, we will simply say that they are all equally likely. The prior is then called a **uniform prior** or **non-informative prior**. This also means that $\mathbb{P}(\theta)$ is constant and independent of the exact realisation of θ, meaning that we can again simplify the argument to be maximized:

$$= \arg\max_{\theta \in \Theta} \mathbb{P}(\vec{x}|\theta).$$

Finally, this means maximizing the posterior $\mathbb{P}(\theta|\vec{x})$ in respect to θ is equivalent to maximizing the likelihood $\mathbb{P}(\vec{x}|\theta)$ in respect to θ. The process of maximizing the likelihood is one of the most well known parameter estimation ideas, called **maximum likelihood estimation** or MLE.
Note that from this point onward we will assume a non-informative prior, if not stated otherwise.

2.2 Performing maximum likelihood estimation

By now you should understand the basic principle but might ask yourselves how to actually find the θ that maximizes the likelihood.
Basically this question is equivalent to: *How do I find the x that maximizes the function $f(x)$?*
We will first show the basic principle in general and then exhibit an example. Since the likelihood function $\mathbb{P}(\vec{x}|\theta)$ is often used for the maximum likelihood estimation, it is often rewritten as its own function: $\mathcal{L}_{\vec{x}}(\theta) = \mathbb{P}(\vec{x}|\theta)$. We again start with what we want, i.e.

$$\begin{aligned} \arg\max_{\theta \in \Theta} \mathcal{L}_{\vec{x}}(\theta) &= \arg\max_{\theta \in \Theta} \mathbb{P}(\vec{x}|\theta) \\ &= \arg\max_{\theta \in \Theta} \mathbb{P}(x_1, x_2, ...x_n|\theta) \end{aligned}$$

At this point we have to make two further assumptions. Let $x_1, ..., x_n$ be the realisations of random variables $X_1, ..., X_n$:

1. The random variables $X_1, ..., X_n$ are independent from each other.
2. The random variables $X_1, ..., X_n$ are identically distributed, i.e. they come from the same distribution with the same parameters.

Since often in statistics one needs both those properties, they are summarized together by saying: $X_1, ..., X_n$ are *i.i.d* (independent and identically distributed).

Using these two assumptions results in

$$
\begin{aligned}
&= \arg\max_{\theta \in \Theta} \mathbb{P}(x_1|\theta) \cdot \mathbb{P}(x_2|\theta) \cdot ... \cdot \mathbb{P}(x_n|\theta) \\
&= \arg\max_{\theta \in \Theta} \prod_{i=1}^{n} \mathbb{P}(x_i|\theta)
\end{aligned}
$$

Finally something a bit more concrete. The next steps consist of calculating the derivative of the likelihood function in respect to θ. This means calculating the derivative of this huge product. At this point, we have two options:

1. If the product $\prod_{i=1}^{n} \mathbb{P}(x_i|\theta)$ is not too complicated such that we can build its derivative we will do that.

2. Since most of the time this product is way too complicated we will use a trick called the **log derivative trick**. The log derivative trick uses the fact, that the logarithmic function is monotonic resulting in the fact that

   $$\arg\max_{\theta \in \Theta} \mathcal{L}_{\vec{x}}(\theta) = \arg\max_{\theta \in \Theta} log(\mathcal{L}_{\vec{x}}(\theta))$$

   Furthermore this implies:

   $$
   \begin{aligned}
   \arg\max_{\theta \in \Theta} log(\mathcal{L}_{\vec{x}}(\theta)) &= \arg\max_{\theta \in \Theta} log\big(\prod_{i=1}^{n} \mathbb{P}(x_i|\theta)\big) \\
   &= \arg\max_{\theta \in \Theta} \sum_{i=1}^{n} log\big(\mathbb{P}(x_i|\theta)\big)
   \end{aligned}
   $$

   This helps us immensely, because determining the derivative of a sum is much easier!
   At this points you want to calculate the derivative of the likelihood in respect to θ and set this equal to zero to determine the **maximum likelihood estimation**:

   $$
   \begin{aligned}
   \frac{\partial}{\partial \theta} log(\mathcal{L}_{\vec{x}}(\theta)) &= \frac{\partial}{\partial \theta}\Big(\sum_{i=1}^{n} log\big(\mathbb{P}(x_i|\theta)\big)\Big) \\
   &= \sum_{i=1}^{n} \frac{\partial}{\partial \theta} log\big(\mathbb{P}(x_i|\theta)\big) \overset{!}{=} 0
   \end{aligned}
   $$

For the general derivation of the maximum likelihood estimation we can not calculate further because the next steps depend on the type of probability distribution.
Nevertheless the next step would be to find the θ that fulfils the equation above. Last but not least one has to check that we are actually dealing with maximum, e.g. by checking the second derivative.

2.2.1 MLE for Bernoulli distributed random variable

Let us go through one of the most simple cases: We assume that $x_1, ..., x_n$ are realisations of random variables $X_1, ..., X_n$ and $X_i \sim Ber(p)$ for $i = 1, ..., n$. Since we want to give you a formally correct example to show how to perform maximum likelihood estimation, many steps here might be repetitions of the general case. Since it is convention to use p for the Bernoulli distribution, we will use $\theta = p$ and thus $\Theta = [0, 1]$.

Calculations

We want to find the parameter $\theta = p$ that maximizes the likelihood function $\mathcal{L}_{\vec{x}}(\theta)$:

$$\begin{aligned}\arg\max_{p\in[0,1]} \mathcal{L}_{\vec{x}}(p) &= \arg\max_{p\in[0,1]} \mathbb{P}(\vec{x}|\theta)\\ &= \arg\max_{p\in[0,1]} \mathbb{P}(x_1, ..., x_n|\theta)\end{aligned}$$

Here we assume that $X_1, ..., X_n$ are i.i.d.:

$$\begin{aligned}&= \arg\max_{p\in[0,1]} \mathbb{P}(x_1|\theta) \cdot ... \cdot \mathbb{P}(x_n|\theta)\\ &= \arg\max_{p\in[0,1]} \prod_{i=1}^{n} \mathbb{P}(x_i|\theta)\end{aligned}$$

Next we apply the log derivative trick

$$\begin{aligned}&= \arg\max_{p\in[0,1]} log(\prod_{i=1}^{n} \mathbb{P}(x_i|\theta))\\ &= \arg\max_{p\in[0,1]} \sum_{i=1}^{n} log(\mathbb{P}(x_i|\theta))\end{aligned}$$

Next we will use the fact that we know what kind of distribution we are dealing with i.e. the Bernoulli distribution:

$$\begin{aligned}&= \arg\max_{p\in[0,1]} \sum_{i=1}^{n} log(p^{x_i}(1-p)^{1-x_i})\\ &= \arg\max_{p\in[0,1]} \sum_{i=1}^{n} \Big(log(p^{x_i}) + log((1-p)^{1-x_i})\Big)\\ &= \arg\max_{p\in[0,1]} \sum_{i=1}^{n} \Big(x_i \cdot log(p) + (1-x_i) \cdot log(1-p)\Big)\end{aligned}$$

After we simplified as much as possible, the next step is to build the derivative of the likelihood function in respect to p:

$$\begin{aligned}
\frac{\partial}{\partial p} log(\mathcal{L}_{\vec{x}}(p)) &= \frac{\partial}{\partial p} \sum_{i=1}^{n} \Big(x_i \cdot log(p) + (1 - x_i) \cdot log(1 - p) \Big) \\
&= \sum_{i=1}^{n} \frac{\partial}{\partial p} \Big(x_i \cdot log(p) + (1 - x_i) \cdot log(1 - p) \Big) \\
&= \sum_{i=1}^{n} \Bigg(\frac{\partial}{\partial p} \Big(x_i \cdot log(p) \Big) + \frac{\partial}{\partial p} \Big((1 - x_i) \cdot log(1 - p) \Big) \Bigg) \\
&= \sum_{i=1}^{n} \Big(\frac{x_i}{p} - \frac{1 - x_i}{1 - p} \Big)
\end{aligned}$$

At this point we calculated the derivative and have to set it equal to zero and calculate for which p the equation holds:

$$\begin{aligned}
\frac{\partial}{\partial p} log(\mathcal{L}_{\vec{x}}(p)) = 0 &\Leftrightarrow \sum_{i=1}^{n} \Big(\frac{x_i}{p} - \frac{1 - x_i}{1 - p} \Big) = 0 \qquad \Big| \cdot p(1 - p) \\
&\Leftrightarrow \sum_{i=1}^{n} \Big(x_i \cdot (1 - p) - p\dot{(}1 - x_i) \Big) = 0 \\
&\Leftrightarrow \sum_{i=1}^{n} \Big(x_i - x_i p - p + x_i p \Big) = 0 \\
&\Leftrightarrow \sum_{i=1}^{n} \Big(x_i - p \Big) = 0 \\
&\Leftrightarrow \sum_{i=1}^{n} x_i - \sum_{i=1}^{n} p = 0 \\
&\Leftrightarrow \sum_{i=1}^{n} x_i - np = 0 \qquad \Big| + np \\
&\Leftrightarrow np = \sum_{i=1}^{n} x_i \qquad \Big| \cdot \frac{1}{n} \\
&\Leftrightarrow p = \frac{1}{n} \sum_{i=1}^{n} x_i
\end{aligned}$$

We now found a p which renders the first derivative zero. But we still have to show that this p is a maximum, not a minimum. We will accomplish this by calculating the second derivative and plugging in $p = \frac{1}{n} \sum_{i=1}^{n} x_i$. First we determine the second derivative:

$$\begin{aligned}\frac{\partial^2}{\partial p^2} log(\mathcal{L}_{\vec{x}}(p)) &= \frac{\partial}{\partial p}\frac{\partial}{\partial p} log(\mathcal{L}_{\vec{x}}(p))\\ &= \frac{\partial}{\partial p}\sum_{i=1}^{n}\Big(\frac{x_i}{p} - \frac{1-x_i}{1-p}\Big)\\ &= \sum_{i=1}^{n}\Big(-\frac{x_i}{p^2} - \frac{1-x_i}{(1-p)^2}\Big)\end{aligned}$$

Here we can plug in $p = \frac{1}{n}\sum_{i=1}^{n} x_i$ or we can argue why this is negative: We know that $p \in [0,1]$ which implies that $-\frac{x_i}{p^2}$ is negative and $(1-p)^2$ is positive. Next we know that $x_i \in \{0,1\}$, which implies that $1 - x_i \geq 0$ thus $-\frac{1-x_i}{(1-p)^2}$ is also negative.
Given that it is a maximum we also need to show that it is the global maximum, i.e. we need to check the limits. For the Bernoulli distribution the limits of Θ are 0 and 1. The question we are trying to answer is if either $\mathcal{L}_{\vec{x}}(0)$ or $\mathcal{L}_{\vec{x}}(1)$ is higher than $\mathcal{L}_{\vec{x}}(\frac{1}{n}\sum_{i=1}^{n} x_i)$. Again this can be done by making a smart argument: If $p = 0$ then $\frac{1}{n}\sum_{i=1}^{n} x_i) = \frac{1}{n}\sum_{i=1}^{n} 0) = 0$ thus $\mathcal{L}_{\vec{x}}(0) = \mathcal{L}_{\vec{x}}(\frac{1}{n}\sum_{i=1}^{n} x_i)$. On the other hand, if $p = 1$ then $\frac{1}{n}\sum_{i=1}^{n} x_i) = \frac{1}{n}\sum_{i=1}^{n} 1) = 1$ thus $\mathcal{L}_{\vec{x}}(1) = \mathcal{L}_{\vec{x}}(\frac{1}{n}\sum_{i=1}^{n} x_i)$.
Finally we are finished, we determined the maximum likelihood estimated of i.i.d. Bernoulli distributed random variables to be

$$p_{MLE} = \frac{1}{n}\sum_{i=1}^{n} x_i,$$

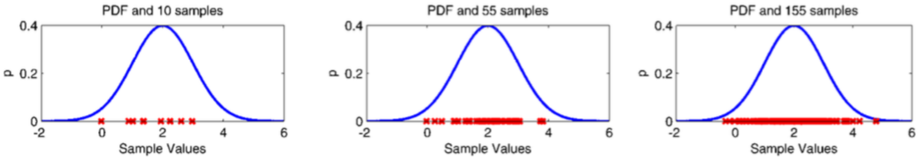
i.e. the mean of all realisations.

2.3 Observed Fisher Information

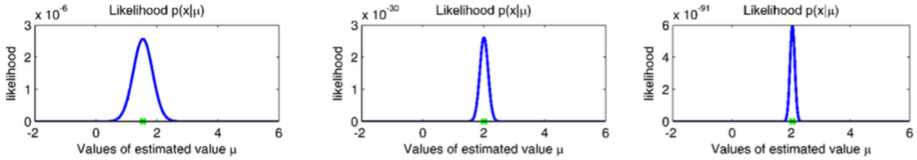
The observed Fisher information is a way of measuring the amount of information that an observable random variable X carries about an unknown parameter μ upon which the probability of X depends.

2.3.1 Intuition

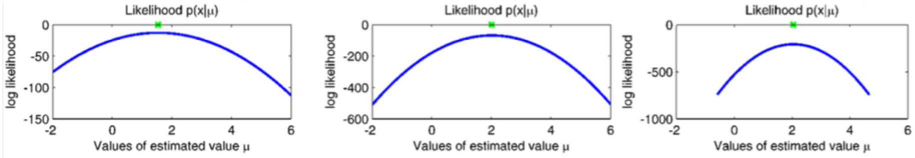
Let us imagine we have some underlying, unknown distribution $\mathbb{P}$. We will compare three cases with different amount of samples coming from this distribution $\mathbb{P}$:



Using this samples we can come up with three different estimates for the real parameter. Our intuition suggests that the estimator build by the right hand case should be more expressive than the one in the middle and even more expressive than the one on the left hand side. This can be verified if we have a look at the likelihood functions:



The distribution of parameter values for θ has a very tight peak on the right hand side translating into a much more certain estimate. This also means, that the estimate by the right hand side is much more expressive compared to the two other estimates.
Finally, let us consider the log-likelihoods:



We see that expressiveness translates into a more curved log-likelihood, i.e. the magnitude of curvature is equivalent to the expressiveness of the estimate. How do we measure curvature of a function? We consider the second derivative but because the likelihood is always right-curved (thus the second derivative is always negative; the more negative the more right curved), we look at the **negative second derivative**.
We will define the following two terms:

- **Score function**: $S(\theta) = \frac{\partial}{\partial\theta} log\Big(\mathcal{L}_{\vec{x}}(\theta)\Big)$, where $S(\theta_{MLE}) = 0$.
- **Observed Fisher information**: $I(\theta) = -\frac{\partial^2}{\partial\theta^2} log\Big(\mathcal{L}_{\vec{x}}(\theta)\Big) = -\frac{\partial}{\partial\theta} S(\theta)$.

2.3.2 Example

In this example we will take a look at the Fisher information for the mean of the normal distribution. First, we will calculate the score function i.e. the derivative in respect to μ (note that we will use ln instead of log here):

$$\begin{aligned}
S(\mu) &= \frac{\partial}{\partial\mu} ln\Big(\mathcal{L}_{\vec{x}}(\mu)\Big) \\
&= \frac{\partial}{\partial\mu} \sum_{i=1}^{n} ln\Big(\frac{1}{\sqrt{2\pi\sigma^2}} \cdot e^{-\frac{(x_i-\mu)^2}{2\sigma^2}}\Big) \\
&= \frac{\partial}{\partial\mu} \sum_{i=1}^{n} \Big(ln\Big(\frac{1}{\sqrt{2\pi\sigma^2}}\Big) + ln\Big(e^{-\frac{(x_i-\mu)^2}{2\sigma^2}}\Big)\Big) \\
&= \frac{\partial}{\partial\mu} \sum_{i=1}^{n} \Big(ln\Big(\frac{1}{\sqrt{2\pi\sigma^2}}\Big) + -\frac{(x_i-\mu)^2}{2\sigma^2}\Big) \\
&= \sum_{i=1}^{n} \Big(\frac{\partial}{\partial\mu} ln\Big(\frac{1}{\sqrt{2\pi\sigma^2}}\Big) - \frac{\partial}{\partial\mu}\frac{(x_i-\mu)^2}{2\sigma^2}\Big) \\
&= \sum_{i=1}^{n} \Big(0 + \frac{2(x_i-\mu)}{2\sigma^2}\Big) \\
&= \frac{1}{\sigma^2} \sum_{i=1}^{n} (x_i-\mu).
\end{aligned}$$

Now we can use the score function to calculate the Fisher information:

$$\begin{aligned}
I(\mu) &= -\frac{\partial}{\partial\mu} S(\theta) \\
&= -\frac{\partial}{\partial\mu}\frac{1}{\sigma^2} \sum_{i=1}^{n} (x_i-\mu) \\
&= -\frac{1}{\sigma^2} \sum_{i=1}^{n} (0-1) \\
&= \frac{1}{\sigma^2} \sum_{i=1}^{n} 1 \\
&= \frac{n}{\sigma^2}.
\end{aligned}$$

How can this result be interpreted?

- With an increasing number of samples the observed Fisher information increases which in turn means that we know more about the parameter μ, which implies a higher precision of the resulting estimate.
- The larger the variance is the lower the observed Fisher information i.e. we know less about the parameter, which in turn means a lower precision of the resulting estimate.

2.4 Bias of an estimator

2.4.1 Idea

Previously we constructed an **estimator**. We do not want to dive to deep into statistics but as the name suggests an estimator tries to estimate an unknown parameter or a property of it. Note, that everything can be an estimator, e.g. I am allowed to estimate θ to be 1, regardless of which distribution θ is a parameter of and which samples we have. Most of the time this estimate will be quite wrong but nevertheless it is an estimator.
This poses the question of how to assess the quality of an estimator.
The most common way is to calculate the **bias of an estimator** (not to be confused with biases used in neural networks). The bias of an estimator $\hat{\theta}$ is defined as

$$Bias(\hat{\theta}) = \mathbb{E}(\hat{\theta}) - \theta,$$

where *theta* is the true parameter. This is very intuitive, because we simply look at the difference between the underlying parameter we want to estimate and the expected value of our estimator. We say the estimator $\hat{\theta}$ is **unbiased**, if $Bias(\hat{\theta}) = 0$; otherwise we call $\hat{\theta}$ **a biased estimator**.

2.4.2 Intuition

Let us use an analogy to make this idea more intuitive.
Imagine that you want to hit bulls eye on a dart disk. You throw multiple times, each time you look and focus on the bulls eye. Sadly, every time you throw a bit too much to the right.
In this analogy the true parameter is where you want to throw and the estimate is where you are actually throwing. Obviously there is a mismatch between the two which is the bias.
Now how would you fix this? The thing I would do is to purposely aim a bit too much to the left such that I incorporate the error I make into my planning!

Example: Bessel's correction

One of the biggest questions in beginner mathematics we will now deal with is the question of why we sometimes normalize an estimator using $n-1$ instead of n.

The prime case when we use $n-1$ instead of n is when we look at the sampling variance defined as

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

where $\bar{x}$ is the sampling mean defined as

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i.$$

Obviously they are somehow related to the expected value and the variance, but how? I personally did not understand when to use which and what they are for, so let us try to understand this once and for all: Imagine we are given a probability space and a random variable on this probability space, which has a distribution with an underlying parameter θ. We already know what the expected value of a random variable X is and what the variance of a random variable X is. You can however, look at those two as properties of θ. Why would you do this? Well, that is because most often we are not given the complete probability space but instead are given realisations of X and want to make some statements about the probability space; statements such as the expected value and the variance. And here the sample mean and the sampling variance come in: The sample mean is used as an estimate for the $\mathbb{E}[X]$ and the sampling variance is used as an estimate for the $Var[X]$.

Let us first check, whether the sample mean is a biased estimator for the expected value. For this, we will need to calculate the expected value of the sample mean, so lets say that $x_1, ..., x_n$ are all realisations of some random variable X. Furthermore we will denote the real expected value of X with μ. Then

$$\begin{aligned}
\mathbb{E}[\bar{x}] &= \mathbb{E}[\frac{1}{n}\sum_{i=1}^{n} x_i] \\
&= \frac{1}{n}\mathbb{E}[\sum_{i=1}^{n} x_i] \\
&= \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}[x_i] \\
&= \frac{1}{n}\sum_{i=1}^{n}\mu \\
&= \mu,
\end{aligned}$$

thus the sample mean is an unbiased estimator (which is good). Next let us check the sampling variance, but instead of normalizing with $n-1$, we will use n to understand why we do not use it. σ^2 will denote the variance of random variable X. Then:

$$\begin{aligned}
\mathbb{E}[s^2] &= \mathbb{E}[\frac{1}{n}\sum_{i=1}^{n}(x_i-\bar{x})^2] \\
&= \mathbb{E}[\frac{1}{n}\sum_{i=1}^{n}(x_i-\mu+\mu-\bar{x})^2] \\
&= \mathbb{E}[\frac{1}{n}\sum_{i=1}^{n}\Big((x_i-\mu)-(\bar{x}-\mu)\Big)^2] \\
&= \mathbb{E}[\frac{1}{n}\sum_{i=1}^{n}\Big((x_i-\mu)^2-2(x_i-\mu)(\bar{x}-\mu)+(\bar{x}-\mu)^2\Big)] \\
&= \mathbb{E}[\frac{1}{n}\sum_{i=1}^{n}(x_i-\mu)^2-2(\bar{x}-\mu)\frac{1}{n}\sum_{i=1}^{n}(x_i-\mu)+\frac{1}{n}\sum_{i=1}^{n}(\bar{x}-\mu)^2] \\
&= \mathbb{E}[\frac{1}{n}\sum_{i=1}^{n}(x_i-\mu)^2-2(\bar{x}-\mu)(\frac{1}{n}\sum_{i=1}^{n}x_i-\frac{1}{n}\sum_{i=1}^{n}\mu)+(\bar{x}-\mu)^2] \\
&= \mathbb{E}[\frac{1}{n}\sum_{i=1}^{n}(x_i-\mu)^2-2(\bar{x}-\mu)(\bar{x}-\mu)+(\bar{x}-\mu)^2] \\
&= \mathbb{E}[\frac{1}{n}\sum_{i=1}^{n}(x_i-\mu)^2-(\bar{x}-\mu)^2] \\
&= \mathbb{E}[\frac{1}{n}\sum_{i=1}^{n}(x_i-\mu)^2]-\mathbb{E}[(\bar{x}-\mu)^2] \\
&= \mathbb{E}[\sigma^2]-\mathbb{E}[(\bar{x}-\mu)^2] \\
&= \sigma^2-\mathbb{E}[(\bar{x}-\mu)^2]
\end{aligned}$$

which leads to the bias:

$$\begin{aligned}
Bias(s^2) &= \mathbb{E}[s^2]-\sigma^2 \\
&= (\sigma^2-\mathbb{E}[(\bar{x}-\mu)^2])-\sigma^2 \\
&= -\mathbb{E}[(\bar{x}-\mu)^2].
\end{aligned}$$

Because $-\mathbb{E}[(\bar{x}-\mu)^2] \leq 0$, this means that we underestimate the real variance by using the sampling mean with normalization value n!
Just as the example with dart disc shows we can try to upgrade our estimate

such that it is not biased any more. To do so we can first recognize that

$$\begin{aligned}
\mathbb{E}[s^2] &= \sigma^2 - \mathbb{E}[(\bar{x}-\mu)^2] \\
&= \sigma^2 - Var[\bar{x}] \\
&= \sigma^2 - Var[\frac{1}{n}\sum_{i=1}^{n} x_i] \\
&= \sigma^2 - \frac{1}{n^2}Var[\sum_{i=1}^{n} x_i] \\
&= \sigma^2 - \frac{1}{n^2}\sum_{i=1}^{n} Var[x_i] \\
&= \sigma^2 - \frac{1}{n^2}\sum_{i=1}^{n} Var[X] \\
&= \sigma^2 - \frac{1}{n}Var[X] \\
&= \sigma^2 - \frac{1}{n}\sigma^2 \\
&= \frac{n}{n}\sigma^2 - \frac{1}{n}\sigma^2 \\
&= \frac{n-1}{n}\sigma^2
\end{aligned}$$

We see that to make this estimator unbiased, we need to multiply it with $\frac{n}{n-1}$ resulting in

$$\begin{aligned}
\frac{n}{n-1}s^2 &= \frac{n}{n-1}\frac{1}{n}\sum_{i=1}^{n}(x_i-\bar{x})^2 \\
&= \frac{1}{n-1}\sum_{i=1}^{n}(x_i-\bar{x})^2.
\end{aligned}$$

Finally we can see that using $n-1$ instead of n makes the sampling variance unbiased. This is called **Bessel's Correction**.

Chapter 3

Linear Regression

Linear regression is a linear approach to modelling the relationship between a scalar response (called dependent variable, target or label) and one or more explanatory variables (called regressors or independent variable); i.e. we switch from our previously used **univariate** data (i.e. only one value) to multivariate data (tuples of two variables).

Formalism First we need to introduce some basic notation:

- **Dependent** variable: The dependent variable can be a vector on its own, but for our purposes we will assume that it is a scalar if not stated otherwise. Since we have multiple samples that means we have multiple values of the dependent variable. Assuming that we sampled n times we reference that value of the dependent variables by $t_1, t_2, ..., t_n$.

- **Independent** variable: The independent variable can be a scalar but often it is a vector. Thus we will reference each value of the independent variable (for n times sampling) by $\vec{x}_1, \vec{x}_2, ..., \vec{x}_n$. Each of these vector has the same dimension, e.g. $\vec{x}_i = (x_{i,1}, x_{i,2}, ..., x_{i,j})$.

Example for formalism Imagine you want to analyse the relationship between

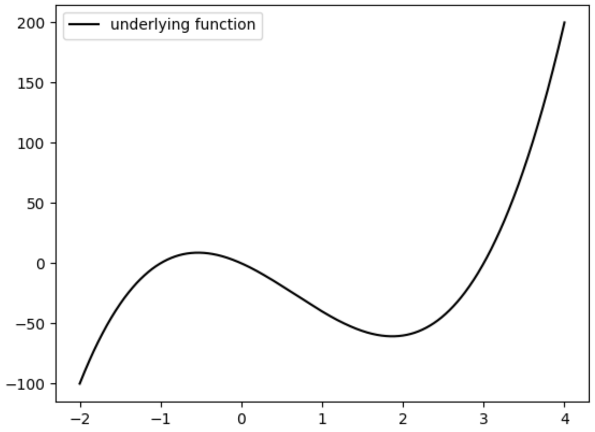
- the number of rooms of a house, the year it was build, how many bed room it has (independent variables)

- and for how much it can be sold (dependent variable).

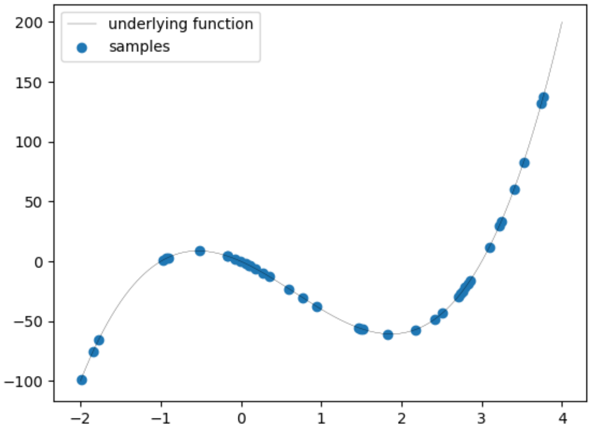
A single sample could look something like this: $\Big((12, 1953, 4), 278000\Big)$.

3.1 The underlying model and noise

Previously in the univariate case the data point came from some distribution that only extended along the x-axis. In the multivariate case this is a bit different:
We assume that there is some underlying function $f(\vec{x})$ $(t_i = f(\vec{x}_i))$ in the real world exactly governing the relationship; this is the function we are trying to approximate:



Again, this is the function governing the relationship between $\vec{x}_i$ and t_i. Now if we sample data form this underlying function, the sampled points will usually not lie exactly on this line i.e. they will not look like this:



Rather we expect some deviation from the underlying function. We will first answer where this deviation is coming from and then deal with how we can model it.

In the field of signal processing, this deviation is known as noise. Noise is a general term for unwanted (and in general unknown) modifications that a signal may suffer during capture, storage, transmission, processing, or conversion.

And how can we include noise in our model? Basically we can use this idea:

$$DATA = TRUE\ \ SIGNAL + NOISE$$

which means for us the model that created our sampled data is not solely $f(\vec{x})$ but instead

$$t_i = f(\vec{x}_i) + \epsilon,$$

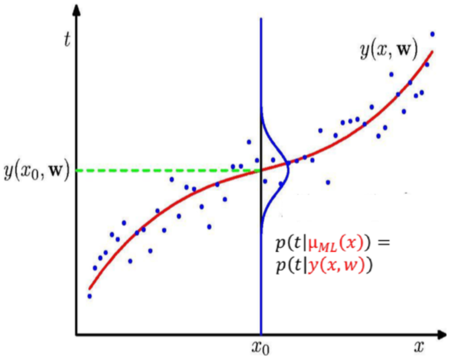
where ϵ is the noise. How can we model the noise? The most common and intuitive choice is to use a normal distribution that is centred around 0 with variance σ^2, i.e.

$$t_i = f(\vec{x}_i) + \mathcal{N}_{\mu=0,\sigma^2}(k_i)$$

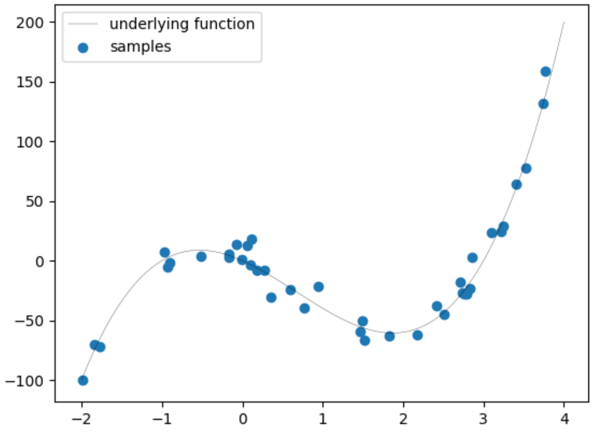
which we can simplify to

$$t_i = \mathcal{N}_{\mu=f(\vec{x}_i),\sigma^2}(k_i).$$

Visually we can think of it like this:



If we use this we end up with samples that seem much more natural:



3.2 Vanilla linear regression model

How does a vanilla linear regression model look like? Like this:

$$\begin{aligned} y_i &= w_0 \cdot 1 + w_1 \cdot x_{i,1} + w_2 \cdot x_{i,2} + ... + w_j \cdot x_{i,j} \\ &= \vec{w} \cdot \vec{x}^T \end{aligned}$$

where

- $\vec{w}$ are called weights and are the parameters of our model,
- y_i is our approximated value for the dependent variable t_i.

We do not include a noise term because it solely tries to model the underlying function.
The linear regression model assumes the following:

- **Homoscedasticity**: This is the property that the each sample has the same variance in their noise as the others, regardless of the values contained in this sample.
- **Independence of errors**: This assumptions states that the noise (or error) of the samples is not correlated with each other.

The way we defined the linear regression model up until now we face a problem: Because we have no non-linearity in the model, we can only model linear relationships between the independent and the dependent values. Of course we would love to also model non-linear relationships. This brings us to the **linear basis function models**.

3.3 Linear Basis Function Model

Instead of linearly combining the "vanilla" independent variables we linearly combine a set of simple (possibly non-linear) functions of the independent variables. This set of simple functions is called the **basis set** and the functions themselves are called **basis functions** and they are denoted by ϕ.
Here comes the part that might confuse you at this point: Although these basis functions might be non-linear, we are still dealing with a linear basis function model. This is due to the fact that the linearity property is in regard to the **combination of the coefficients**, with the coefficients themselves being able to be non-linear.
The linear basis function model has the following form

$$\begin{aligned} y_i &= w_0 \cdot \phi_0(1) + w_1 \cdot \phi_1(x_{i,1}) + w_2 \cdot \phi_2(x_{i,2}) + ... + w_j \cdot \phi_j(x_{i,j}) \\ &= \vec{w} \cdot \vec{\phi}(\vec{x})^T, \end{aligned}$$

Note that we will cover basis functions (and what they might explicitly look like) later.

3.4 Finding the right weights

At this point we came up with a model but do not have any parameters. Obviously there are many values we could assign to them, but how do we wind the right ones?
Well, we will apply what we have already learned: Previously when faced with an unknown parameter we used maximum likelihood estimation.
In our case we are trying to find $\vec{w}$ for a normal distribution of the form $\mathcal{N}_{\mu=f(\vec{x}_i),\sigma^2}(k_i)$ which is somehow equivalent to finding the underlying model $f(\vec{x})$. (Note that here k_i would be the resulting dependent variable/ target/

label).
We will use a linear basis function model of the form $g(\vec{x}) = \vec{w}^T \cdot \vec{\phi}(\vec{x}_i)$ to model the underlying function.
To fit our weights to the data (which we will assume consists of n samples) we will use a maximum likelihood estimation. Since we have multiple parameters $w_0, w_1, ..., w_j$, we have to take the derivative of the log likelihood (i.e. the score function) in respect to the gradient $\nabla = \Big[\frac{\partial}{\partial w_0}, \frac{\partial}{\partial w_1}, ..., \frac{\partial}{\partial w_j}\Big]$:

$$\begin{aligned}
\nabla S(\vec{w}) &= \nabla ln\Big(\prod_{i=1}^{n} \mathcal{N}_{\mu=g(\vec{x}_i),\sigma^2}(k_i)\Big) \\
&= \nabla \sum_{i=1}^{n} ln\Big(\mathcal{N}_{\mu=g(\vec{x}_i),\sigma^2}(k_i)\Big) \\
&= \nabla \sum_{i=1}^{n} ln\Big(\frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(k_i-g(\vec{x}_i))^2}{2\sigma^2}} Big) \\
&= \nabla \sum_{i=1}^{n} ln\Big(\frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(k_i-g(\vec{x}_i))^2}{2\sigma^2}}\Big) \\
&= \nabla \sum_{i=1}^{n} ln\Big(\frac{1}{\sqrt{2\pi}} \frac{1}{\sqrt{\sigma^2}} e^{-\frac{(k_i-g(\vec{x}_i))^2}{2\sigma^2}}\Big) \\
&= \nabla \sum_{i=1}^{n} \Big(ln(\frac{1}{\sqrt{2\pi}}) + ln(\frac{1}{\sqrt{\sigma^2}}) + ln\big(e^{-\frac{(k_i-g(\vec{x}_i))^2}{2\sigma^2}}\big)\Big) \\
&= \nabla \sum_{i=1}^{n} \Big(ln\big((2\pi)^{-\frac{1}{2}}\big) + ln\big(\sigma^{-\frac{1}{2}2}\big) + -\frac{(k_i-g(\vec{x}_i))^2}{2\sigma^2}\Big) \\
&= \nabla \sum_{i=1}^{n} \Big(-\frac{1}{2}ln\big(2\pi\big) - \frac{1}{2}ln\big(\sigma^2\big) + -\frac{(k_i-g(\vec{x}_i))^2}{2\sigma^2}\Big) \\
&= \nabla\Big(-\frac{1}{2}\sum_{i=1}^{n} ln\big(2\pi\big) - \frac{1}{2}\sum_{i=1}^{n} ln\big(\sigma^2\big) + -\sum_{i=1}^{n} \frac{(k_i-g(\vec{x}_i))^2}{2\sigma^2}\Big) \\
&= \nabla\Big(-\frac{n}{2}ln\big(2\pi\big) - \frac{n}{2}ln\big(\sigma^2\big) + -\sum_{i=1}^{n} \frac{(k_i-g(\vec{x}_i))^2}{2\sigma^2}\Big) \\
&= -0 - 0 - \nabla \sum_{i=1}^{n} \frac{(k_i-g(\vec{x}_i))^2}{2\sigma^2} \\
&= -\nabla \sum_{i=1}^{n} \frac{(k_i-g(\vec{x}_i))^2}{2\sigma^2}
\end{aligned}$$

Let us at this point inspect some explicit weight coefficient w_m:

$$
\begin{aligned}
&= \frac{\partial}{\partial w_m} - \sum_{i=1}^{n} \frac{(k_i - g(\vec{x}_i))^2}{2\sigma^2} \\
&= -\frac{1}{2\sigma^2} \frac{\partial}{\partial w_m} \sum_{i=1}^{n} (k_i - \vec{w}^T \vec{\phi}(\vec{x}_i))^2 \\
&= -\frac{1}{2\sigma^2} \sum_{i=1}^{n} 2(k_i - \vec{w}^T \vec{\phi}(\vec{x}_i)) \cdot \frac{\partial}{\partial w_m} (k_i - \vec{w}^T \vec{\phi}(\vec{x}_i)) \\
&= -\frac{1}{2\sigma^2} \sum_{i=1}^{n} 2(k_i - \vec{w}^T \vec{\phi}(\vec{x}_i)) \cdot \big(0 - (0 + 0 + ... + \frac{\partial}{\partial w_m} w_m \phi_m(\vec{x}_i) + ... + 0)\big) \\
&= -\frac{1}{\sigma^2} \sum_{i=1}^{n} (k_i - \vec{w}^T \vec{\phi}(\vec{x}_i)) \cdot \big(- \phi_m(\vec{x}_i)\big) \\
&= \frac{1}{\sigma^2} \sum_{i=1}^{n} (k_i - \vec{w}^T \vec{\phi}(\vec{x}_i)) \cdot \phi_m(\vec{x}_i)
\end{aligned}
$$

Now that we got the derivative we set it to zero:

$$
\begin{aligned}
0 &= \frac{1}{\sigma^2} \sum_{i=1}^{n} (k_i - \vec{w}^T \vec{\phi}(\vec{x}_i)) \cdot \phi_m(\vec{x}_i) && \Big| \cdot \sigma^2 \\
\Leftrightarrow 0 &= \sum_{i=1}^{n} (k_i - \vec{w}^T \vec{\phi}(\vec{x}_i)) \cdot \phi_m(\vec{x}_i) \\
\Leftrightarrow 0 &= \sum_{i=1}^{n} \Big(k_i \phi_m(\vec{x}_i) - \vec{w}^T \vec{\phi}(\vec{x}_i) \phi_m(\vec{x}_i) \Big) \\
\Leftrightarrow 0 &= \sum_{i=1}^{n} k_i \phi_m(\vec{x}_i) - \sum_{i=1}^{n} \vec{w}^T \vec{\phi}(\vec{x}_i) \phi_m(\vec{x}_i) && \Big| + \sum_{i=1}^{n} \vec{w}^T \vec{\phi}(\vec{x}_i) \phi_m(\vec{x}_i) \\
\Leftrightarrow \sum_{i=1}^{n} \vec{w}^T \vec{\phi}(\vec{x}_i) \phi_m(\vec{x}_i) &= \sum_{i=1}^{n} k_i \phi_m(\vec{x}_i)
\end{aligned}
$$

At this point we can see that both sides of the equation are scalars. Remember how we exchanged the general gradient for an explicit w_m? Let us reverse this. We have this general formula that we can use for every coefficient of the weight vector, but we want to rewrite this equation in vector notation. For this purpose we will introduce the so called **design matrix** at this point:

$$\Phi = \begin{bmatrix} \phi_0(x_{1,0}) & \phi_1(x_{1,1}) & \phi_2(x_{1,2}) & \cdots & \phi_j(x_{1,j}) \\ \phi_0(x_{2,0}) & \phi_1(x_{2,1}) & \cdots & \cdots & \phi_j(x_{1,j}) \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \phi_0(x_{n,0}) & \cdots & \cdots & \cdots & \phi_j(x_{n,j}) \end{bmatrix}$$

Using this we can rewrite this general equation for an arbitrary but explicit w_m:

$$\sum_{i=1}^{n} \vec{w}^T \vec{\phi}(\vec{x}_i)\phi_m(\vec{x}_i) = \sum_{i=1}^{n} k_i \phi_m(\vec{x}_i)$$

into the general vector notation form:

$$\vec{w}^T \Phi^T \Phi = \big(\Phi^T \vec{k}\big)^T.$$

(Note that this is easy but tedious to show; which is why we will refrain from it here. Feel free to do some matrix vector multiplications to see why this holds).

$$\Leftrightarrow \Big(\vec{w}^T \Phi^T \Phi\Big)^T = \Big((\Phi^T \vec{k})^T\Big)^T$$

Next we will use the fact that for two matrices A and B it always holds that $(AB)^T = B^T A^T$ and $(A^T)^T = A$:

$$\Leftrightarrow \Phi^T \Big(\vec{w}^T \Phi^T\Big)^T = \Phi^T \vec{k}$$

$$\Leftrightarrow \Phi^T \Phi \vec{w} = \Phi^T \vec{k}$$

Here we will make an assumption to further rewrite the equation: We will assume that the basis functions are linearly independent which implies that the square matrix $\Phi^T\Phi$ will have an inverse. We will multiply both sides of the equation (left wise) with said inverse $(\Phi^T\Phi)^{-1}$:

$$\Leftrightarrow \vec{w} = (\Phi^T \Phi)^{-1} \Phi^T \vec{k}$$

The last simplification we will need is called the **pseudo-inverse** of a matrix. For a matrix A it is defined as $A^\dagger = (A^T A)^{-1} A^T$. "Luckily" this is exactly what we have got:

$$\Leftrightarrow \vec{w} = \Phi^\dagger \vec{k}$$

(At this point, it is a good exercise to go through these last steps and check if the dimensions add up).
Well what did we just do and how can this help us? Every time you want to approximate an underlying function using a linear basis function model, you can calculate the corresponding weights by simply building the design matrix and multiply its pseudo-inverse with the vector of targets $\vec{k}$.
This may (or may not) come as a surprise to the reader. As an adept of model fitting the first methods of fitting weights are mainly *a*) iterative and *b*) approximate. The method of using the pseudo-inverse is direct and exact because we manually calculate the vector of coefficient for the weights which results in the minimum error. Why is it that you have never heard of this magic method before? Well it works perfectly, but does not scale well in terms of time and memory. One rule of thumb I follow is this: If you have less than 10000 samples you can use the pseudo-inverse. For more you should switch to a numerical, i.e. proximate, method.

3.5 Bias Variance Decomposition

What we found out in the preceding sections is that if we assume that the noise (that is added to our underlying function) is normally distributed, performing maximum likelihood estimation results in an expression of the form:

$$\arg\max_{\theta} - \sum_{i=1}^{n} (t_i - y_i)^2$$

where t_i are the true values and y_i our predictions. This is equivalent to

$$\arg\min_{\theta} \sum_{i=1}^{n} (t_i - y_i)^2.$$

Minimizing this expression also known as minimizing the squared residuals, i.e. the squared error of our prediction in comparison to the true values. This term is also known as the L2 error function.
Let us try to get a better grasp on this L2 error. To understand what its

expected value is made of, we will decompose it into three terms:

$$\begin{aligned}
\mathbb{E}[L2] &= \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}\big[(t_i - y_i)^2\big] \\
&= \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}\big[(t_i - f(\vec{x}_i) + f(\vec{x}_i) - y_i)^2\big] \\
&= \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}\big[((t_i - f(\vec{x}_i)) + (f(\vec{x}_i) - y_i))^2\big] \\
&= \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}\big[(t_i - f(\vec{x}_i))^2 + 2(t_i - f(\vec{x}_i))(f(\vec{x}_i) - y_i) + (f(\vec{x}_i) - y_i)^2\big] \\
&= \frac{1}{n}\sum_{i=1}^{n}\Big(\mathbb{E}\big[(t_i - f(\vec{x}_i))^2\big] + \mathbb{E}\big[2(t_i - f(\vec{x}_i))(f(\vec{x}_i) - y_i)\big] + \mathbb{E}\big[(f(\vec{x}_i) - y_i)^2\big]\Big)
\end{aligned}$$

Here we use the fact that the measured signal t_i is nothing else but the underlying function $f(\vec{x}_i)$ plus noise ϵ, i.e. $t_i - f(\vec{x}_i) = (f(\vec{x}_i) + \epsilon) - f(\vec{x}_i) = \epsilon$:

$$\begin{aligned}
&= \frac{1}{n}\sum_{i=1}^{n}\Big(\mathbb{E}\big[\epsilon^2\big] + \mathbb{E}\big[2(t_i f(\vec{x}_i) - f^2(\vec{x}_i) - t_i y_i + f(\vec{x}_i) y_i)\big] + \mathbb{E}\big[(f(\vec{x}_i) - y_i)^2\big]\Big) \\
&= \frac{1}{n}\sum_{i=1}^{n}\Big(\mathbb{E}\big[\epsilon^2\big] + 2\Big(\mathbb{E}\big[(t_i f(\vec{x}_i)\big] - \mathbb{E}\big[f^2(\vec{x}_i)\big] - \mathbb{E}\big[t_i y_i\big] + \mathbb{E}\big[f(\vec{x}_i) y_i)\big]\Big) + \mathbb{E}\big[(f(\vec{x}_i) - y_i)^2\big]\Big)
\end{aligned}$$

Here we can simplify. For this simplification we need a few things:

- We will assume that the noise is centred around 0 (which makes a lot of sense) and thus $[\epsilon] = 0$.
- Also note that the noise is completely random and independent of any other component. We can use this fact so that $\mathbb{E}[f(\vec{x})\epsilon] = \mathbb{E}[f(\vec{x})]\mathbb{E}[\epsilon]$ and $\mathbb{E}[y_i\epsilon] = \mathbb{E}[y_i]\mathbb{E}[\epsilon]$.
- Because $f(\vec{x})$ is deterministic/error free, it holds that $\mathbb{E}[f(\vec{x})] = f(\vec{x})$.

First simplification

$$\begin{aligned}
\mathbb{E}[t_i f(\vec{x})] &= \mathbb{E}[(f(\vec{x}) + \epsilon) \cdot f(\vec{x})] \\
&= \mathbb{E}[f^2(\vec{x}) + f(\vec{x})\epsilon] \\
&= \mathbb{E}[f^2(\vec{x})] + \mathbb{E}[f(\vec{x})\epsilon] \\
&= f^2(\vec{x}) + \mathbb{E}[f(\vec{x})]\mathbb{E}[\epsilon] \\
&= f^2(\vec{x})
\end{aligned}$$

Second simplification

$$\mathbb{E}[f^2(\vec{x})] = f^2(\vec{x})$$

Third simplification

$$\begin{aligned}\mathbb{E}[t_i y_i] &= \mathbb{E}[(f(\vec{x}) + \epsilon) \cdot y_i] \\ &= \mathbb{E}[f(\vec{x}) y_i + \epsilon y_i] \\ &= \mathbb{E}[f(\vec{x}) y_i] + \mathbb{E}[\epsilon y_i] \\ &= \mathbb{E}[f(\vec{x}) y_i] + \mathbb{E}[\epsilon]\mathbb{E}[y_i] \\ &= \mathbb{E}[f(\vec{x}) y_i]\end{aligned}$$

These simplification result in:

$$\begin{aligned}&= \frac{1}{n}\sum_{i=1}^{n}\Big(\mathbb{E}\big[\epsilon^2\big] + 2\Big(f^2(\vec{x}_i) - f^2(\vec{x}_i) - \mathbb{E}\big[f(\vec{x})y_i\big] + \mathbb{E}\big[f(\vec{x}_i)y_i)\big]\Big) + \mathbb{E}\big[(f(\vec{x}_i) - y_i)^2\big]\Big) \\ &= \frac{1}{n}\sum_{i=1}^{n}\Big(\mathbb{E}\big[\epsilon^2\big] + \mathbb{E}\big[(f(\vec{x}_i) - y_i)^2\big]\Big) \\ &= \frac{1}{n}\sum_{i=1}^{n}\Big(\mathbb{E}\big[\epsilon^2\big] + \mathbb{E}\big[(f(\vec{x}_i) - \mathbb{E}[y_i] + \mathbb{E}[y_i] - y_i)^2\big]\Big) \\ &= \frac{1}{n}\sum_{i=1}^{n}\Big(\mathbb{E}\big[\epsilon^2\big] + \mathbb{E}\big[\big((f(\vec{x}_i) - \mathbb{E}[y_i]) + (\mathbb{E}[y_i] - y_i)\big)^2\big]\Big) \\ &= \frac{1}{n}\sum_{i=1}^{n}\Big(\mathbb{E}\big[\epsilon^2\big] + \mathbb{E}\big[(f(\vec{x}_i) - \mathbb{E}[y_i])^2 + 2(f(\vec{x}_i) - \mathbb{E}[y_i])(\mathbb{E}[y_i] - y_i) + (\mathbb{E}[y_i] - y_i)^2\big]\Big) \\ &= \frac{1}{n}\sum_{i=1}^{n}\Big(\mathbb{E}\big[\epsilon^2\big] + \mathbb{E}\big[(f(\vec{x}_i) - \mathbb{E}[y_i])^2\big] + 2\mathbb{E}\big[(f(\vec{x}_i) - \mathbb{E}[y_i])(\mathbb{E}[y_i] - y_i)\big] + \mathbb{E}\big[(\mathbb{E}[y_i] - y_i)^2\big]\Big) \\ &= \frac{1}{n}\sum_{i=1}^{n}\Big(\mathbb{E}\big[\epsilon^2\big] + \mathbb{E}\big[(f(\vec{x}_i) - \mathbb{E}[y_i])^2\big] + 2\mathbb{E}\big[f(\vec{x}_i)\mathbb{E}[y_i] - \mathbb{E}^2[y_i] - f(\vec{x}_i)y_i + \mathbb{E}[y_i]y_i\big] + \mathbb{E}\big[(\mathbb{E}[y_i] - y_i)^2\big]\Big) \\ &= \frac{1}{n}\sum_{i=1}^{n}\Big(\mathbb{E}\big[\epsilon^2\big] + \mathbb{E}\big[(f(\vec{x}_i) - \mathbb{E}[y_i])^2\big] + 2\Big(\mathbb{E}\big[f(\vec{x}_i)\mathbb{E}[y_i]\big] - \mathbb{E}\big[\mathbb{E}^2[y_i]\big] - \mathbb{E}\big[f(\vec{x}_i)y_i\big] + \mathbb{E}\big[\mathbb{E}[y_i]y_i\big]\Big) + \mathbb{E}\big[(\mathbb{E}\end{aligned}$$

Again here we can simplify things. We need the following ideas:

- Generally the expected value takes something random and returns a constant (the expected value). This means, that the expected value of an expected value, i.e. the expected value of something constant, is this constant: $\mathbb{E}\Big[\mathbb{E}\big[X\big]\Big] = \mathbb{E}\big[X\big]$.
- Constants can be pulled out of the expected value

First simplification

$$\begin{aligned}\mathbb{E}\big[f(\vec{x}_i)\mathbb{E}[y_i]\big] &= \mathbb{E}[y_i]\mathbb{E}\big[f(\vec{x}_i)\big] \\ &= \mathbb{E}[y_i]f(\vec{x}_i)\end{aligned}$$

Second simplification

$$\mathbb{E}\big[\mathbb{E}^2[y_i]\big] = \mathbb{E}^2[y_i]$$

Third simplification

$$\mathbb{E}\big[f(\vec{x}_i)y_i\big] = f(\vec{x}_i)\mathbb{E}\big[y_i\big]$$

Fourth simplification

$$\begin{aligned}\mathbb{E}\big[\mathbb{E}[y_i]y_i\big] &= \mathbb{E}[y_i]\mathbb{E}\big[y_i\big] \\ \mathbb{E}^2[y_i]&\end{aligned}$$

This results in

$$\begin{aligned}&= \frac{1}{n}\sum_{i=1}^{n}\Bigg(\mathbb{E}\big[\epsilon^2\big] + \mathbb{E}\big[(f(\vec{x}_i) - \mathbb{E}[y_i])^2\big] + 2\Big(f(\vec{x}_i)\mathbb{E}\big[y_i\big] - \mathbb{E}^2[y_i] - f(\vec{x}_i)\mathbb{E}\big[y_i\big] + \mathbb{E}^2[y_i]\Big) + \mathbb{E}\big[(\mathbb{E}[y_i] - y_i)^2\big]\Bigg) \\ &= \frac{1}{n}\sum_{i=1}^{n}\Bigg(\mathbb{E}\big[\epsilon^2\big] + \mathbb{E}\big[(f(\vec{x}_i) - \mathbb{E}[y_i])^2\big] + \mathbb{E}\big[(\mathbb{E}[y_i] - y_i)^2\big]\Bigg) \\ &= \frac{1}{n}\sum_{i=1}^{n}\Bigg(\underbrace{\mathbb{E}\big[\epsilon^2\big]}_{\text{Noise}} + \underbrace{\mathbb{E}\big[(f(\vec{x}_i) - \mathbb{E}[y_i])^2\big]}_{\mathbb{E}[Bias^2(y_i)]=Bias^2(y_i)]} + \underbrace{\mathbb{V}\big[y_i\big]}_{Variance}\Bigg)\end{aligned}$$

Here we can see that the general expected value of the error of a model is made up of three parts:

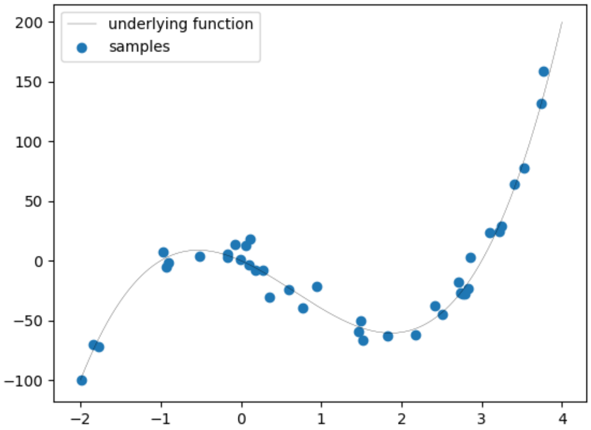
- $\mathbb{E}\big[\epsilon^2\big]$ is the noise of the data. One might think (since $\mathbb{E}\big[\epsilon\big] = 0$) that $\mathbb{E}\big[\epsilon^2\big]$ has also to be 0, but this is not the case since all values are now squared, i.e. positive, i.e. not centred around 0. Note that we do not have any real influence on this part of the error thus it is the lower bound of the expected value of the L2 error. In other words for the perfect model it holds that: $\mathbb{E}[L2] = \mathbb{E}\big[\epsilon^2\big]$

- $Bias^2(y_i)$ is the squared bias. Why is it the bias? Because y_i is our estimator of the underlying function. It describes the systematic difference between the real underlying function and the our estimator (model).

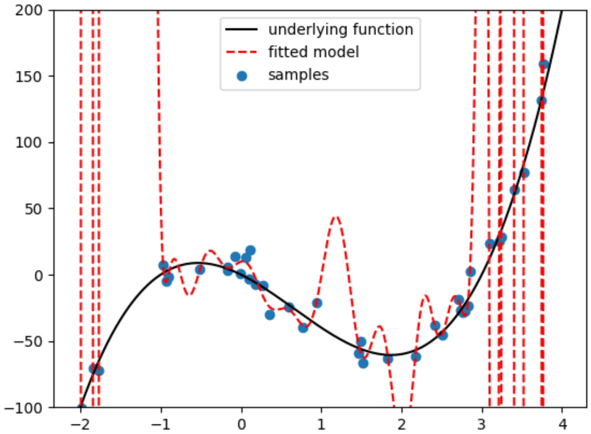
- $\mathbb{V}\big[y_i\big]$ is the variance of the model across different data sets i.e. if you would use the same model architecture to predict the underlying deterministic function using multiple sets of samples it describes by how much do these different fitted models vary.

3.5.1 Analysis

Here we will analyse what we worked out before in regard to the complexity of a model. We will use the following data samples:

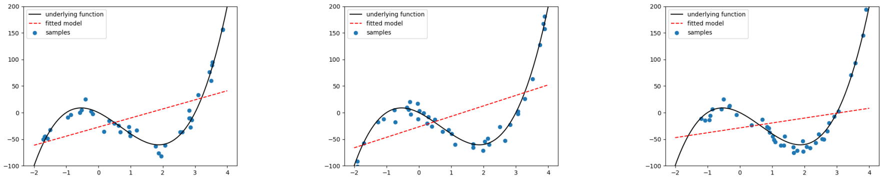


Bias Error Term The lower the bias is the closer are $f(\vec{x}_i)$ and the expected value of the models prediction y_i, i.e. the more precise we become at predicting the targets.
This implies that models with a high degree of complexity and thus a much higher flexibility to fit the data will have low bias error terms. Consider a fit to the data of the form before with a polynomial model of degree 100:

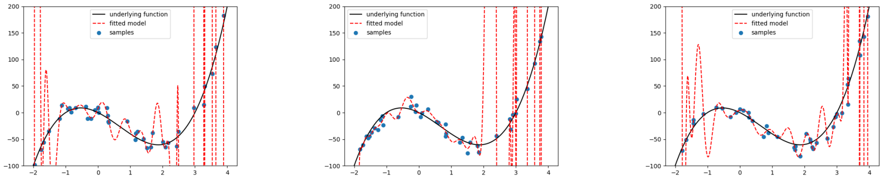


The model almost predicts every sample we supplied it with perfectly i.e. the bias error term is close to 0.

Variance Error Term The variance error term has the opposite characteristics in comparison to the bias error term: The more complex a model becomes the higher is this error term. To put it simple, the variance that is measures by this term refers to the variance a model exhibits when fitting different data sample sets from the same underlying model.
First let us consider a very simple model, a polynomial of degree 3. We will randomly sample 3 sets of samples and will fit the polynomial three separate times onto this sets:



As you can see, the "variance" in between those fits is very small. Compare this to a setup with a polynomial of degree 100:



As expected because the function is so flexible it will have very different characteristics for each fit. But this is not the main reason: Mainly the model exhibits such drastically different characteristics for every separate set of samples, because it is much too complex for the data at hand; so complex indeed that it started fitting the noise of this particular set of samples!

Good Fit vs. Overfit vs. Underfit

- A model is a *good fit* if it is exactly as complex as it needs to be. This implies a good trade-off between the bias error of the model and its variance error.
- A model is *overfitting*, if it is much too complex for the data. It does not generalize the underlying function from the samples but instead learned on the specific samples. Its bias error is extremely small due to its almost perfect predictions, but its variance error is extremely high.

- A model is said to *underfit* if it is not complex enough. In this case the model lacks the flexibility to properly adjust itself towards the underlying function. Its variance error is therefore extremely low, but its bias error extremely high.

3.6 Regularization

By now we have a pretty good idea what overfitting and underfitting is and what a good fit might look like.
Up until now if we wanted to prevent any kind of overfitting what we did was to reduce the complexity of the model. This way the model is not "flexible" enough to overfit. As the name suggests in this section we will learn about regularization, an alternative approach that often yields better results.
The first things we have to note is that large weights often correspond to a model overfitting. To understand why, think of it this way: High weights also mean high sensitivity; so small variations in the independent variable lead to large deviations in the dependent variable. For this reason we have an overfitting model because the model was designed to fit one state but may not fit another state due to the models high sensitivity.
Regularization uses the idea that if a large magnitude of weights corresponds to the model overfitting, we can prevent overfitting keeping the weights small. How can this be achieved?
Well, up until now we wanted to also minimize the error i.e. minimize $L2$; lets call the error $E_D(\vec{w})$ (D because the error is caused by the mismatch between predictions and real data). We will introduce another error term called $E_w(\vec{w})$ that denotes the magnitude of the weights; we will call it the **regularization coefficient**. By trying to minimize the sum of both of these errors

$$E(\vec{w}) = E_D(\vec{w}) + E_w(\vec{w}),$$

we force the weights to take values that try to minimize both of the properties at the same time: On one hand trying to fit the data; on the other trying to have low magnitude values. Note that these two goals are somewhat antagonistic, i.e. the weights will try to find a compromise.
To control by how much it focuses on each of those two terms we will introduce the so-called **regularization parameter** λ to the second regularization term:

$$E(\vec{w}) = E_D(\vec{w}) + \lambda E_w(\vec{w}).$$

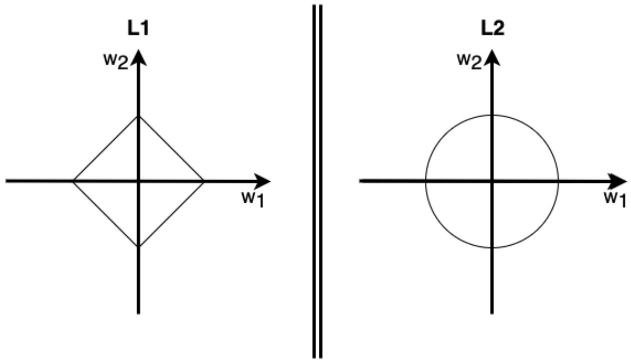
High values for λ result in stronger regularization and smaller model weights, while low values for λ reduce the influence of the regularization.
The regularization coefficient can be chosen freely, but there are two very common choices:

- $E_w(\vec{w}) = \frac{1}{2}\sum_{i=0}^{j} |w_i|$, which is known as **lasso regularization**; it uses the $L1$ norm.

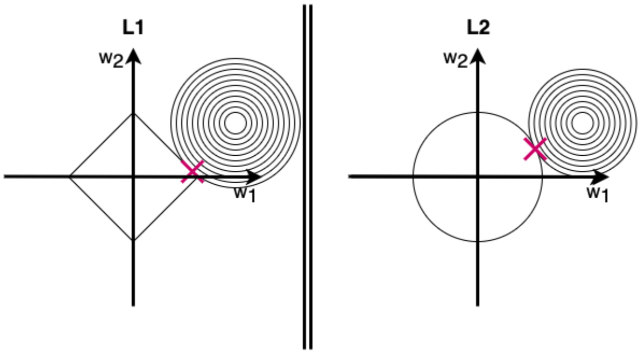
- $E_w(\vec{w}) = \frac{1}{2}\sum_{i=0}^{j} w_i^2$, which is known as **quadratic regularization**; it uses the $L2$ norm.

3.6.1 $L1$ vs.$L2$

So what does it make for a difference? We can gain some knowledge about this, by illustrating what we are actually doing.
For this visualization we will assume that we have two weights, w_1 and w_2 that we are trying to fit.
The first thing we will do is to draw the unit circle of the $L1$ and the $L2$ norm for the regularization coefficient. This unit circle is the set of all points whose regularization value is 1:



We can clearly see that the shape of the $L1$ norm is a diamond, while the $L2$ norm is a circle. Note at this point that there are many more of these norms that have the general form $\frac{1}{2}\sum_{i=0}^{j} |w_j|^p$, where $p \in [1, \infty[$.
Next, we will further add into this illustration a possible first term in the error function i.e.$E_D(\vec{w})$. The black circles in and around the blue areas are iso-lines indicating combinations of w_1 and w_2 with equal mean squared error between the resulting model and the data.
Now if we try to find the best compromise between the two we are looking for the point where they would first touch (if we were to equally increase them at the same time); and here we can see the difference:



- Lasso regression due the form of the unit circle tends to have the points of touch at one if its corners. Well what does this mean? Often it sets certain weights to zero. This translates into a sparse weight vector and effectively into fewer basis function used.

- For the quadratic regression the red circle generally is a bit smaller because it punishes high weights even more (because it squares them). Next to this the point of touch is (a priori) equally like to be anywhere thus not creating any particular pattern.

Note that for increasing values of the regularization coefficient λ the regularization gets stronger thus the red shape shrinks dragging the weights further towards zero.
Furthermore note that mathematically there is exactly one minimum for the regularization coefficient in the quadratic case but there can be more for lasso regularization.

3.6.2 Where do $L1$ and $L2$ norm come from?

At this point the idea may be clear but where exactly do $L1$ and $L2$ regularization come from?
Previously we learned that maximum likelihood estimation is a special case of maximum a posteriori estimation, where we assume a uniform prior. Assuming that we do not, performing maximum a posteriori estimation is equivalent to maximizing

$$\arg\max_{\theta} \mathbb{P}(\theta|D) = \arg\max_{\theta} \mathbb{P}(D|\theta)\mathbb{P}(\theta).$$

Here we also apply the log trick resulting in

$$\begin{aligned} &= \arg\max_{\theta} log(\mathbb{P}(D|\theta)\mathbb{P}(\theta)) \\ &= \arg\max_{\theta} log(\mathbb{P}(D|\theta)) + log(\mathbb{P}(\theta)). \end{aligned}$$

Selecting a normally distributed prior

At this point we can also try out different priors. Remember the prior is a probability distribution assigning each possible value θ_j a probability. One example is using a normal distribution with mean 0 and variance τ^2 (for the likelihood we will again assume an underlying function plus normal distributed random noise and we will assume that our model does not use any basis functions):

$$\arg\max_{\theta} \Big[\log \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(t_i-(\theta_0+\theta_1 x_{i,1}+...+\theta_p x_{i,p}))^2}{2\sigma^2}} + \log \prod_{j=0}^{p} \frac{1}{\sqrt{2\pi\tau^2}} e^{-\frac{\theta_j^2}{2\tau^2}} \Big]$$

First we can disregard both normalization coefficients because they are constant in regard to θ_j:

$$\begin{aligned}
&= \arg\max_{\theta} \Big[\log \prod_{i=1}^{n} e^{-\frac{(t_i-(\theta_0+\theta_1 x_{i,1}+...+\theta_p x_{i,p}))^2}{2\sigma^2}} + \log \prod_{j=0}^{p} e^{-\frac{\theta_j^2}{2\tau^2}} \Big] \\
&= \arg\max_{\theta} \Big[\sum_{i=1}^{n} -\frac{(t_i-(\theta_0+\theta_1 x_{i,1}+...+\theta_p x_{i,p}))^2}{2\sigma^2} + \sum_{j=0}^{p} -\frac{\theta_j^2}{2\tau^2} \Big] \\
&= \arg\max_{\theta} \frac{1}{2\sigma^2} \Big[\sum_{i=1}^{n} -(t_i-(\theta_0+\theta_1 x_{i,1}+...+\theta_p x_{i,p}))^2 - \frac{2\sigma^2}{2\tau^2} \sum_{j=0}^{p} \theta_j^2 \Big] \\
&= \arg\max_{\theta} \Big[-\sum_{i=1}^{n} (t_i-(\theta_0+\theta_1 x_{i,1}+...+\theta_p x_{i,p}))^2 - \lambda \sum_{j=0}^{p} \theta_j^2 \Big] \\
&= \arg\max_{\theta} -\Big[\sum_{i=1}^{n} (t_i-(\theta_0+\theta_1 x_{i,1}+...+\theta_p x_{i,p}))^2 + \lambda \sum_{j=0}^{p} \theta_j^2 \Big] \\
&= \arg\min_{\theta} \Big[\sum_{i=1}^{n} (t_i-(\theta_0+\theta_1 x_{i,1}+...+\theta_p x_{i,p}))^2 + \lambda \sum_{j=0}^{p} \theta_j^2 \Big]
\end{aligned}$$

Notice that maximizing the likelihood plus prior is equivalent to minimizing the term above. Notice that we set $\lambda = \frac{\sigma^2}{\tau^2}$; this is allowed because σ and τ are constant. As you can see we arrived at $L2$ regularization.
We can adjust the amount of regularization we want by changing λ- Equivalently, we can adjust how much we want the priors to influence the coefficients (θ). If we have a very small variance (large λ) then the coefficients will be very close to 0; if we have a large variance (small λ) then the coefficients will not be affected much (similar to not having any regularization).

Selecting a Laplacean distributed prior

First let us review the Laplace distribution (something that's usually not introduced in beginner probability classes):

$$Lapalace(\mu, b) = \frac{1}{2b} e^{-\frac{|x-\mu|}{b}}.$$

Starting with a zero-mean Laplacean prior on all the coefficients like we did in the previous subsection:

$$\begin{aligned}
&\arg\max_{\theta} \Big[\log \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(t_i-(\theta_0+\theta_1 x_{i,1}+\ldots+\theta_p x_{i,p}))^2}{2\sigma^2}} + \log \prod_{j=0}^{p} \frac{1}{2b} e^{-\frac{|\theta_j|}{2b}} \Big] \\
&= \arg\min_{\theta} \frac{1}{2\sigma^2} \Big[\sum_{i=1}^{n} (t_i - (\theta_0 + \theta_1 x_{i,1} + \ldots + \theta_p x_{i,p}))^2 + \frac{2\sigma^2}{2b} \sum_{j=0}^{p} |\theta_j| \Big] \\
&= \arg\min_{\theta} \Big[\sum_{i=1}^{n} (t_i - (\theta_0 + \theta_1 x_{i,1} + \ldots + \theta_p x_{i,p}))^2 + \lambda \sum_{j=0}^{p} |\theta_j| \Big]
\end{aligned}$$

Again we can see that this last equation contains the same expression as $L1$ regularization. The Laplacean prior has a slightly different effect compared to $L2$ regularization. Instead of preventing any of the coefficients from being too large (due to the squaring), $L1$ promotes sparsity. That is zeroing out some of the coefficients. This makes some sense if you look at the density of a Laplacean prior where there is a sharp increase in the density at its mean.

3.7 Basis Functions

Up to this point we talked a few times about basis function but never specified them. As we concluded before a linear basis function model is described by

$$\begin{aligned}
y_i &= w_0 \cdot \phi_0(1) + w_1 \cdot \phi_1(x_{i,1}) + w_2 \cdot \phi_2(x_{i,2}) + \ldots + w_j \cdot \phi_j(x_{i,j}) \\
&= \vec{w} \cdot \vec{\phi}(\vec{x})^T.
\end{aligned}$$

The $\phi_j(x_{i,j})$ is known as the j-th basis function, and $\vec{\phi}(\vec{x})$ as the basis function vector.

Generally we set $\phi_0(x_{i,0}) = 1$, such that w_0 acts as the bias: The bias is crucial for model approximation because adding a bias permits the output of the model to be shifted to the left or right on the x-axis.

In the simplest case we use a linear basis functions: $\phi_d(x_{i,d}) = x_{i,d}$ (simple linear structure, basically vanilla linear regression); but it is much more usual to use non-linear functions to allow the model to exhibit non-linear properties.

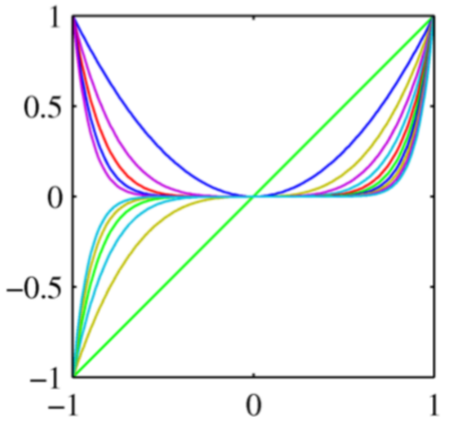
In the following we will show some of the most common choices for basis functions and talk about their advantages and disadvantages. While doing so we will also talk about basis functions being **local**, **relatively global** or **global**. How can this property be understood? Imagine the following case: We collected many samples and fit a linear basis function model to these samples. The fit is fine and we are completely content with our result. Now we collect one more sample, this time a huge outlier. This outlier will of course change the fit of the model (for the worse) but here the difference between local and global basis function matters:

- **global**: This function covers the whole interval of interest and thus the outlier changes the values of this basis function for every possible prediction.
- **local**: This function may be composed of multiple basis function, each covering its respective interval. The outlier may change the basis function responsible for its interval, but the rest of the basis function (and thus the rest of the predictions) stays untouched.
- **relatively local**: This function is theoretically global, but the change it experiences is so small that it behaves as being local.

3.7.1 Polynomial basis function

The polynomial basis function is maybe the most well known basis function. It is characterized by:

$$\phi_j(x_{i,j}) = x_{i,j}^j$$

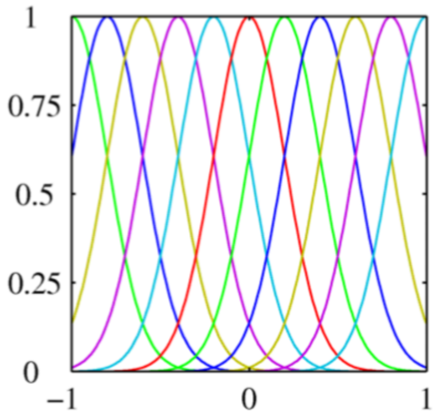


We can see that polynomials are obviously global; a small change in x affect all basis functions.

3.7.2 Gaussian basis functions

A Gaussian basis function is characterized by:

$$\phi_j(x_{i,j}) = e^{\frac{(x_{i,j}^j - \mu_j)^2}{2\sigma^2}}.$$



The Gaussian basis function although being global in the sense that they are non-zero everywhere counts as relatively local basis function because for the most part they assign values that are very close to zero.
One additional advantage of this basis function is the fact that it is easy to use/tune due to the fact that μ represents the location and σ^2 represents the scale (width) of the Gaussian.

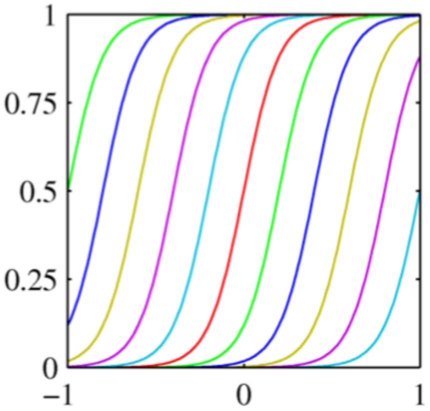
3.7.3 Sigmoidal basis functions

A sigmoidal basis function is characterized by:

$$\phi_j(x_{i,j}) = \sigma(\frac{x_{i,j} - \mu_j}{s})$$

where

$$\sigma(z) = \frac{1}{1 + e^{-z}}.$$



The sigmoidal basis function exhibits properties equivalent to the Gaussian basis function: Relatively local and also easy to use, with μ_j representing the value at which the sigmoidal has the value $0,5$ and s being the scale (slope).

3.7.4 Periodic basis function

The next class of basis functions we will investigate are the periodic basis functions. As their name suggests these basis functions make use of periodic functions, i.e.

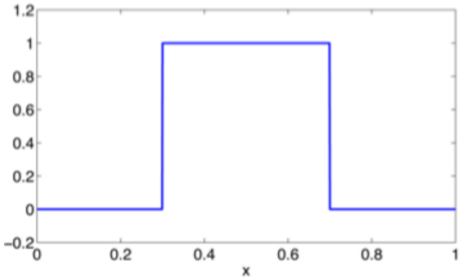
$$f(x) = f(x + nk), k \in \mathbb{N}.$$

A very common choice for a periodic function is the sine function; in the popular Fourier transformation a wide variety of sine waves are used to approximate a function. Of course periodic basis functions are global.

3.7.5 Bin-based basis function

Another form of basis functions are bin-based basis functions. They are strictly local, i.e. non-zero only for the range between two bounds x^l and x^r. The simplest bin-based function is a step function:

$$\phi_j(x_{i,j}) = \begin{cases} 1 & x^l \leq x_{i,j} < x^r \\ 0 & \text{else} \end{cases}$$



The "bin" in this case would be the range of x for which the function is non-zero. Since this is not a smooth function a sum of step functions can not be smooth either. In many or most situations however a smooth function promises better results. Now if we want a smooth but strictly local function, we can build a bin-based function from multiple polynomial functions. This is then called a piecewise polynomial function.

Chapter 4

Model Validation and Comparison

We will now deal with the issue of how to assess the adequacy of a given model in describing observed data. Theoretically, the task of a model is to identify the function that actually generated the observed data. Practically, this is not feasible since we would need an absurd amount of data and even then we might not be sure that we are right. For this problem there are different information criteria who propose solutions.

4.1 Evaluation of models

Generally speaking, we can divide the different evaluation criteria into two groups: **Qualitative criteria** and **quantitative criteria**.

4.1.1 Qualitative Evaluation

Although we will focus on the quantitative aspects of model evaluation, we ought to mention what is meant by a qualitative analysis of a model. When we are dealing with a qualitative analysis of a model, we mainly talk about

1. **Falsifiability**: A model has falsifiability (or is falsifiable) if potential observations exist which are incompatible with the model.

2. **Plausibility**: Do the theoretically assumptions of the model make sense?

3. **Interpretability**: Are the parts of the model (and/or the model as a whole) understandable and/or linked to known processes?

4.1.2 Quantitative Evaluation

Goodness of Fit

Looking at the quantitative side of statistical model evaluation, the first criteria that comes to mind is the **Goodness of fit (GOF)** which measures how well the model fits the set of given observations. Measures of the GOF generally summarize the discrepancy between the observed values $(y_1, ..., y_N)$ and the values expected under the model in question $(y_{M_i,1}, ..., y_{M_i,N})$. There are many ways to quantify the GOF, the most known being the **Root-Mean-Squared-Error (RMSE)**:

$$RMSE_{M_i} = \sqrt{\frac{\sum_{j=1}^{N}(y_j - y_{M_i,j})^2}{N}}$$

Complexity

The next quantitative measure we will focus on is the **complexity** of a model. Generally speaking the complexity of a statistical model refers to its inherent ability to fit a wide range of observations, i.e. its *flexibility*. We prefer lower complexity following Occam's razor. There are many ways of quantifying this property for example the **Minimum Description Length (MDL)** which tries to measure how much information is necessary to compress the given observations using a model. Although very interesting there are simpler and more intuitive ways of measuring the complexity such as counting the number of model parameters.

We want to emphasize that this last point is a huge assumption on its own. Following this assumption leads us to the fact that two models which share the same amount of model parameters also have the same complexity. It is left up to the reader to decide if this is plausible, but we want to provide an example to stimulate this interesting thought:

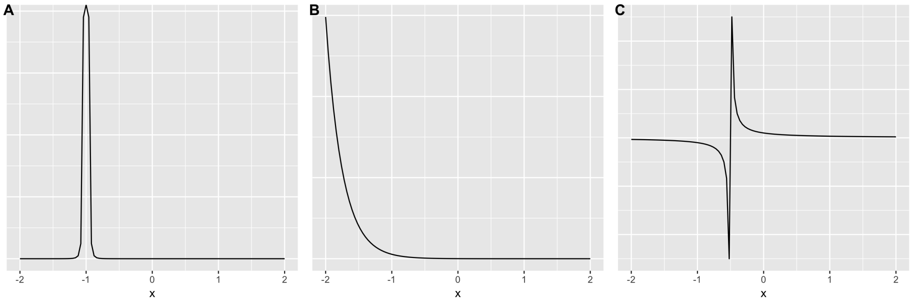


Figure 4.1: Plots of three different functions (**A**: power function, **B**: exponential function, **C** hyperbolic function) with parameters $a = 2$ and $b = 4$.

In the three graphs above, we can see three different functions with each having two parameters:

(a) Power function:

$$p(x) = a(x+1)^{-b}$$

(b) Exponential function:

$$p(x) = ae^{-bx}$$

(c) Hyperbolic function:

$$p(x) = \frac{1}{a+bx}$$

Generalizability

By now we introduced the GOF of a model and its flexibility as properties to judge it by. Finally we will introduce the last of the quantitative measures for model quality which is the **generalization property** or **generalizability** of a model.

Generalizability deals with the assessment of how well a model predicts unseen observations which are representative of the same process. That is, the goal of model selection is to choose the model that generalizes best across all samples since it captures the underlying function best.

4.1.3 Goodness of Fit, Complexity or Generalizability?

Although the GOF is already a good quantification of the adequacy of a model, it has one fatal downfall: The observations that we made are not *clean.* The reason for this is that measured data in the real world includes random noise from a number of sources such as the measurement error and sampling error. Generally speaking, we can say:

Observations = Underlying Function + Noise

Now if we have a look at the effect of this phenomenon on the goodness of fit:

Goodness of Fit = Fit to underlying function + Fit to noise

This means that properties of the model that have nothing to do with approximating the underlying function influence the GOF, since they may influence the fit to the noise. We call this phenomenon **overfitting**: The model fails to approximate the underlying function but rather focuses on approximating the noise. The problem with this phenomenon is that although it might explain the given observations quite well (and better than a model that does not overfit), its GOF on unseen data is quite low since the noise that it fits to is unique for the already seen observations, but the underlying function is the same for all observations, seen or unseen (see figure 2).

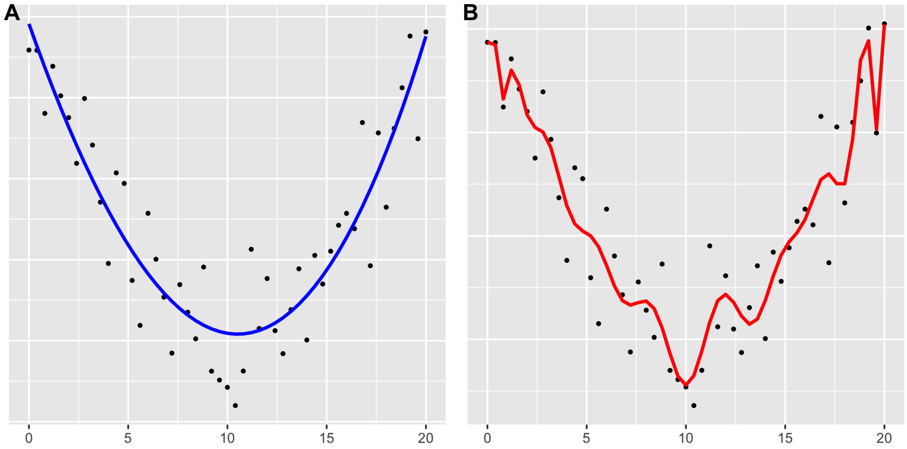


Figure 4.2: We can see that the polynomial model of degree 3 (**A**) captures the underlying function of the observed data quite well and has an appropriate complexity and a decent GOF. On the other hand the polynomial of degree 20 (**B**) started to fit to the noise in this specific batch of observed data and although its GOF is better than the model in **A** it simply fails to capture the regularity of the data because it is too complex.

We see that solely using the GOF as a measure of quality does not suffice. Furthermore it is obvious that information about the complexity (on its own) is not enough as well: "Model M_A has 10 free parameters and Model M_B has 2" does not tell us much; we might choose M_B since it is much simpler (*Occam's razor*), but without any knowledge about the GOF, both of these models may be completely inadequate.

This leaves us with the generalizability of a model. As it turns out, the generalization property of a model contains information about both the GOF and about the complexity:

(a) If the model possesses a low goodness of fit (i.e. it was not able to approximate the underlying function appropriately), it will not be able to decently fit unseen data.

(b) If the model is too complex for the task at hand, it will (additionally to fitting the underlying regularity) try to replicate the irregular behaviour of the noise in the presented dataset. Now if we present this model with new unseen data (and thus new unseen noise), its performance on the new data will most likely be unacceptable.

Using these thoughts we can come up with a relationship between these three characteristic variables:

$$+Generalizability \quad \propto \quad +Goodness\ of\ Fit \quad -Complexity$$

Seeing that Generalizability contains the other quantitative measurements, it is generally referred to as the ”yardstick” for model comparison.

4.2 Cross-Validation

4.2.1 Holdout Method

Cross validation follows a very intuitive idea: If we can not be sure about the adequacy of a trained models by inspecting its goodness of fit (since it may be overfitting) on the data it was trained on, how about dividing the data into two parts: One **training dataset** which will be used to fit the model on and one **testing dataset** (that the model did not encounter before) which will be used to evaluate its generalizability. The problem with this approach is that the evaluation may heavily depend on which data points end up in the training set and which end up in the test set.

4.2.2 K-fold Cross Validation

K-fold cross validation is an improved method of cross validation over the holdout method. Here the dataset is divided into *k* chunks and the holdout method is repeated *k*-times; each time a different chunk is used as the testing dataset while the *k-1* remaining chunks are used as the training dataset. Finally one computes the average performance of each of these *k* trials. The improvement lies therein that it matters less how the data is split. The disadvantage is that instead of once the holdout method has to be done multiple times.

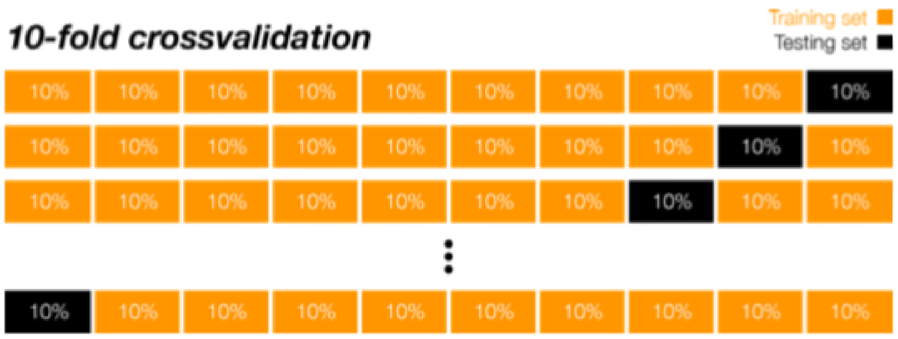


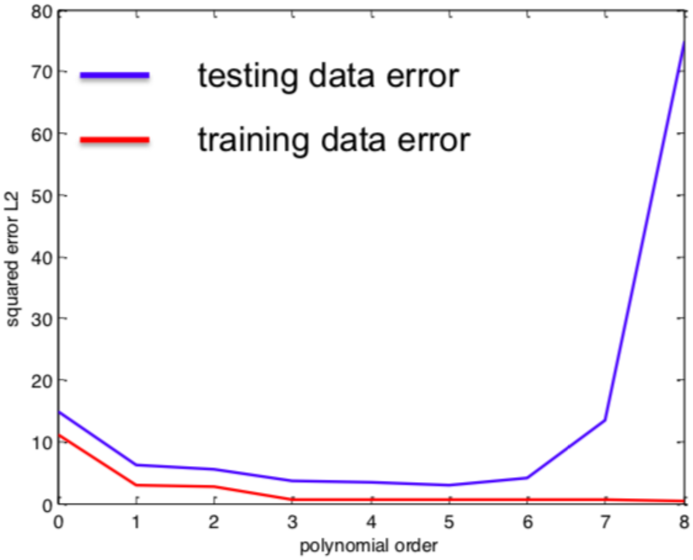
Figure 4.3: An illustration for a 10-fold cross-validation. Note that the data pairs comprising one chunk of data are randomly chosen from the set.

4.2.3 Leave-One-Out Cross Validation

LOO-CV is basically k-fold cross validation taken to the extreme: We set k equal to N, the number of data points. This means that N separate times the model is trained on all the data except for one point and the prediction is made for exactly this one point. Although this method seems expensive to compute there are methods (e.g. locally weighted learners) that immensely reduce the computational effort.

4.2.4 Comparing errors

In the figure below we basically plot a typical relationship between complexity of a model against its error on the training and test data:



As we discussed before we can see that the training error in monotonically decreasing with higher complexity. What is more interesting is the course of the test error: First it decreases meaning that for polynomial order 0 and 1 the model is simply underfitting; this can be seen by the fact that both the training and the test error are decreasing. This continues up until suddenly the test error increases significantly with the complexity. In this region (polynomial order 7 and 8) the model is overfitting; this can be seen by the fact the for higher polynomials the training error decreases while the test error increases. So for this model we should consider a polynomial order of 5.

Note that the actual shape of the curves as well as the optimal complexity of the model is stochastic and depends on the actual training data set and testing data set.

4.2.5 Going forward

The keen reader may have noticed that using K-fold cross-validation will leave us with more than one model to choose from. The question arises how to choose from multiple models. To compare different models we need to compute a meta statistics that summarizes the performance of the model across all test data sets. One common approach is deriving the mean log likelihood across all test data sets; then a model selection approach takes the model with the best average performance.
(Note that increasing the data set size reduces potential overfitting! Therefore larger datasets allow for more complex models!)

4.3 Nested models

We say that a model $\mathcal{B}$ **is nested in a model** $\mathcal{A}$ if $\mathcal{B}$ is a restricted version of $\mathcal{A}$ i.e. $\mathcal{A}$ contains all the terms of $\mathcal{B}$ and at least one additional term. In this scenario $\mathcal{A}$ is called the **complete/full model** and $\mathcal{B}$ is called the **reduce/restricted model**.

4.3.1 Examples

Lets consider the following five models:

$$\begin{aligned}
\mathcal{A} &:= y_i = \alpha x_i + \beta x_i + \gamma x_i + \epsilon \\
\mathcal{B} &:= y_i = \alpha x_i + \beta x_i + \epsilon \\
\mathcal{C} &:= y_i = \alpha x_i + \gamma x_i + \epsilon \\
\mathcal{D} &:= y_i = \beta x_i + \gamma x_i + \epsilon \\
\mathcal{E} &:= y_i = \alpha x_i
\end{aligned}$$

We have the following properties:

- Model $\mathcal{B}$,$\mathcal{C}$ and $\mathcal{D}$ are nested in $\mathcal{A}$, because we can generate them by setting certain parameters of model $\mathcal{A}$ to zero.
- Model $\mathcal{E}$ is nested in models $\mathcal{B}$ and $\mathcal{C}$ thus it is also nested in $\mathcal{A}$.
- Model $\mathcal{E}$ is not nested in $\mathcal{D}$.

4.4 Likelihood ratio test

Using the notion of nested models we can introduce likelihood ratio testing. Imagine the following setup:
You have just trained a model $\mathcal{M}$ on some data. The model $\mathcal{M}$ does a great job

in approximating the underlying model. The only problem is that the model $\mathcal{M}$ has k degrees of freedom and you would love to have a simpler model. You start thinking whether it would be possible to reduce the degrees of freedom without significantly worsening the performance of the model.
To this end you have a look at some nested models of model $\mathcal{M}$ and you find model $\mathcal{M}_0$, a reduced version of $\mathcal{M}$. You hope that although it is reduced its performance is not much worse. Logically because $\mathcal{M}_0$ is a restricted version after all, this implies that $\mathcal{M}_0$ will always have lower or (at best) equal performance in comparison to $\mathcal{M}$. The interesting question in this setup: Is $\mathcal{M}_0$ **statistically significantly worse** in terms of performance compared to $\mathcal{M}$?
If you find yourself in such a situation where to want to know if a nested models is significantly worse than a nesting model you can use a **likelihood ratio test** to find it out.

So let $\mathcal{M}$ be our nesting model with k degrees of freedom and $\mathcal{M}_0$ be a nested model in $\mathcal{M}$ with l degrees of freedom ($l < k$). Furthermore let θ be the tuple of all weights/parameters of $\mathcal{M}$ and θ_0 be the equivalent for $\mathcal{M}_0$. Then $\mathcal{L}_{\vec{x}}(\theta) = \mathbb{P}(\vec{x}|\theta)$ is the likelihood of $\mathcal{M}$ and $\mathcal{L}_{\vec{x}}(\theta_0) = \mathbb{P}(\vec{x}|\theta_0)$ the likelihood of $\mathcal{M}_0$.
First we will build the likelihood ratio which is the likelihood of the nested model divided by the likelihood of the nesting model:

$$\frac{\mathcal{L}_{\vec{x}}(\theta_0)}{\mathcal{L}_{\vec{x}}(\theta)}$$

Due to the fact that the nested model is always a restricted version of the nesting model and thus $\mathcal{L}_{\vec{x}}(\theta_0) \leq \mathcal{L}_{\vec{x}}(\theta)$ the above fraction is in $[0, 1]$.
If we now built the logarithm of this likelihood ratio and multiply it with -2, it turns out that this term is $\mathcal{X}^2$-distributed with $k - l$ degrees of freedom! We call the resulting value $\mathcal{LR}$:

$$-2ln(\frac{\mathcal{L}_{\vec{x}}(\theta_0)}{\mathcal{L}_{\vec{x}}(\theta)}) = -2\Big(ln(\mathcal{L}_{\vec{x}}(\theta_0)) - ln(\mathcal{L}_{\vec{x}}(\theta))\Big) =: \mathcal{LR} \sim \mathcal{X}^2_{k-l}$$

Due to this interesting property of the $\mathcal{LR}$ we can perform a hypothesis test to check whether $\mathcal{M}_0$ is statistically worse at fitting the data in comparison to $\mathcal{M}$ or whether the difference between them is not statistically significant. We use the following null and alternative hypothesis:

- H_0: $\mathcal{M}_0$ is not statistically worse than $\mathcal{M}$ at fitting $\vec{x}$,
- H_1: $\mathcal{M}_0$ is statistically worse than $\mathcal{M}$ at fitting $\vec{x}$.

Given that we settled on some significance level α we can now check for a critical value of the chi-square distribution; let us call this value c. This critical c tells us that if we sample from this distribution, the probability that this value would be bigger/equal than c is less/equal to α. Basically this means:

- $\mathcal{LR} < c$: we do not reject H_0, because $\mathcal{LR}$ does not exceed the critical value $\Rightarrow$ $\mathcal{M}_0$ is not statistically worse than $\mathcal{M}$ at fitting $\vec{x}$ $\Rightarrow$ we choose $\mathcal{M}_0$, because it is simpler.
- $\mathcal{LR} \geq c$: we reject H_0, because $\mathcal{LR}$ does exceed the critical value $\Rightarrow$ $\mathcal{M}_0$ is statistically worse than $\mathcal{M}$ at fitting $\vec{x}$ $\Rightarrow$ we choose $\mathcal{M}$, because (although it is more complex) it is simply better at fitting the given data.

4.5 Deviance

So the likelihood ratio test gives us a mean of comparing two models if one is nested in the other. We will define the **deviance** as the quantitative measure of the performance of a model (not a comparison between two models) although it uses the same idea as the likelihood ratio test.
Given that you have fitted some model $\mathcal{M}$ and you are looking to somehow calculate a value giving you a relative perspective on how good your models is, what you do is to compare it to the **saturated model** $\mathcal{M}_s$. A saturated model is a model which has as many parameters as there are samples making its fit on the data perfect (you can imagine hat each weight is a single sample). Of course it is unbelievably complicated and useless on its own, but we can use it as yardstick to compare our models to. The deviance is then the $\mathcal{LR}$ where your trained model $\mathcal{M}$ is the nested model and the saturated model $\mathcal{M}_s$ is the nesting model (note that every model is nested in its saturated model):

$$D_{\mathcal{M}}(\vec{x}) = -2ln\Big(\frac{\mathcal{L}_{\vec{x}}(\theta)}{\mathcal{L}_{\vec{x}}(\theta_s)}\Big) = -2ln\Big(\frac{\mathcal{L}_{\vec{x}}(\theta)}{\mathcal{L}_{\vec{x}}(\vec{x})}\Big),$$

where $\mathcal{L}_{\vec{x}}(\theta_s)$ is the likelihood of the saturated model.
Finally the questions arise: Do we want a small or a big value for a good model? Well this we can deduce: Obviously we want the likelihood of the model $\mathcal{M}$ to be as big as possible, meaning that we want $\frac{\mathcal{L}_{\vec{x}}(\theta)}{\mathcal{L}_{\vec{x}}(\theta_s)}$ to be as close to 1 as possible (for a bad model $\frac{\mathcal{L}_{\vec{x}}(\theta)}{\mathcal{L}_{\vec{x}}(\theta_s)}$ would be close to 0). This in turn means a good model would result in $ln\Big(\frac{\mathcal{L}_{\vec{x}}(\theta)}{\mathcal{L}_{\vec{x}}(\theta_s)}\Big)$ to be close to 0 (while a bad model will have a very negative value for $ln\Big(\frac{\mathcal{L}_{\vec{x}}(\theta)}{\mathcal{L}_{\vec{x}}(\theta_s)}\Big)$). So finally a good model will have a deviance that is close to zero and a bad model will have a deviance that is quite high (because we multiply by -2).

4.6 AIC

By now we mostly cared for the likelihood of a model when trying to assess its performance. The problem here is that this comparison is unfair since the likelihood is a representation of the GOF (goodness of fit) but the complexity is not taken into account at all. So a model with a very high likelihood might be overfitting but we would not find that out using only the likelihood.

How can we protect us from overfitting? By taking the complexity into account! Hirotugu Akaike, a Japanese statistician, did exactly this by introducing his **Akaike information criterion**:

$$AIC = -2ln(\mathcal{L}_{\vec{x}}(\theta_{MLE})) + 2k = -2ln(P(\vec{x}|\theta_{MLE})) + 2k$$

where k is the number of parameters of the model and $ln(P(\overrightarrow{x}|\theta_{MLE}))$ is the likelihood of the model with the maximum likelihood parametrization θ_{MLE}.
What is the idea behind the AIC? To answer this question we first have to explain what the *Kullback-Leibler divergence* is.

4.6.1 Kullback-Leibler Divergence

In mathematical statistics the Kullback-Leibler divergence (also called relative entropy) is a measure for the difference between one probability distribution and a second reference probability distribution.
Typically one is given a probability distribution g which represents the ”true” distribution of some observed data. Furthermore you have some model f trying to approximate this distribution. In this case you can use the Kullback-Leibler divergence to measure how different one distribution is to the other.
Let $g(x)$ and $f(x)$ be density functions. Then the KL divergence is defined as

$$D_{KL}(g||f) = \sum_i g(i) \cdot log\Big(\frac{g(i)}{f(i)}\Big)$$

for the discrete case or

$$D_{KL}(g||f) = \int_{-\infty}^{\infty} g(x) \cdot log\Big(\frac{g(x)}{f(x)}\Big)dx.$$

Properties

- $D_{KL}(g||f) \geq 0$ for all possible probability distributions g and f (shown by Gibbs' inequality) with $D_{KL}(g||f) = 0$ if and only if $f = g$ almost everywhere.
- The KL divergence is not symmetric! Meaning the KL divergence can not be used as a metric.

4.6.2 Idea behind AIC

The keen reader might have seen a problem with the KL-divergence for model comparison/validation: We do not know the true underlying distribution g. If we would know g we could compare different models to it and choose the one which is most similar to it by calculating the KL-divergence.
Akaike showed that we can estimate via the AIC how similar a model is to this underlying distribution. The estimate though is only valid asymptotically; if the number of data points is small, then some correction is often necessary (see AICc).

4.6.3 Interpreting AIC

Again we can ask how to interpret a low/high AIC score: Generally we want a high likelihood with a low complexity. A high likelihood translates into lower AIC values (because it is multiplied by -2). On the other hand a low complexity (and thus a low k) also translates into lower AIC values, i.e. the lower the value the better the model.
Note that the AIC value of a model on its own is meaningless; it is a relative model assessment criteria: Given multiple candidate models the AIC is calculated for each one and the model with the "best" (lowest) AIC value is chosen.

Chapter 5

Exponential Families

In probability theory and statistics the exponential family is a set of probability distributions of the form:

$$\mathbb{P}(x|\eta) = h(x)e^{\eta^T \cdot T(x) - A(\eta)},$$

where

- η is called the *natural parameter*. Note that η can consist of multiple parameters i.e. $\eta = [\eta_1, \eta_2, ..., \eta_n]$.

- $T(x)$ is called the *sufficient statistics*.

- $h(x)$ is called the *underlying measure*.

- $A(\eta)$ is called the *log normalizer* and ensures that the integral of the exponential family distribution sums up to one (like it should). We can derive it to be:

$$\begin{aligned}
1 &= \int_{-\infty}^{\infty} h(x)e^{\eta^T \cdot T(x) - A(\eta)} dx \\
\Leftrightarrow e^{A(\eta)}e^{-A(\eta)} &= \int_{-\infty}^{\infty} h(x)e^{\eta^T \cdot T(x)} dx e^{-A(\eta)} \\
\Leftrightarrow e^{A(\eta)} &= \int_{-\infty}^{\infty} h(x)e^{\eta^T \cdot T(x)} dx \\
\Leftrightarrow A(\eta) &= ln\Big(\int_{-\infty}^{\infty} h(x)e^{\eta^T \cdot T(x)} dx\Big),
\end{aligned}$$

Note that often an alternative, equivalent form is used for the exponential family which is derived like this:

$$\begin{aligned}\mathbb{P}(x|\eta) &= h(x)e^{\eta^T \cdot T(x) - A(\eta)} \\ &= h(x)e^{\eta^T \cdot T(x)}e^{-A(\eta)} \\ &= h(x)e^{-A(\eta)}e^{\eta^T \cdot T(x)} \\ &= h(x)g(\eta)e^{\eta^T \cdot T(x)}\end{aligned}$$

where $g(\eta) = e^{-A(\eta)}$.

5.1 Idea

Well how can the exponential family be understood? Basically many probability distributions can be reformulated to fit the general shape of the exponential family. Among others, this includes the following:

- Normal distribution
- Exponential distribution
- Beta distribution
- Gamma distribution
- Poisson distribution
- Binomial distribution
- Multinomial distribution
- Geometric distribution
- $\mathcal{X}^2$ - distribution
- Bernoulli distribution

So the above can be seen as special cases of the exponential family.
And how does this help us? Basically the general idea is this: Let us say we want to prove/derive something for multiple of the above distributions. We could either do it for each of them separately or we could do it for the exponential family (the general case) and then plug in the specific terms for each distribution. This can be compare to the fact that the $p-q$ equation and the quadratic formula are general cases of a quadratic equation in which we completed the square.

5.2 Bernoulli distribution as an example for the exponential family

In this lecture and the corresponding exercises your task will often be to prove that a certain distribution is a member of the exponential family. For this purpose, we will perform this proof once here for the Bernoulli distribution:
The general idea is to reformulate the distribution such that the 4 different parts of the exponential family can be found. Due to the fact that the general formula of the exponential family includes the exponential function, this means that we have to introduce the exponential function to any distribution that does not already contain it in the correct fashion like the Bernoulli distribution:

$$\begin{aligned}
\mathbb{P}(x|\mu) &= \mu^x(1-\mu)^{1-x} \\
&= exp\Big(ln\Big(\mu^x(1-\mu)^{1-x}\Big)\Big) \\
&= exp\Big(ln\Big(\mu^x\Big) + ln\Big((1-\mu)^{1-x}\Big)\Big) \\
&= exp\Big(x \cdot ln(\mu) + (1-x) \cdot ln(1-\mu)\Big) \\
&= exp\Big(x \cdot ln(\mu) + ln(1-\mu) - x \cdot ln(1-\mu)\Big) \\
&= exp\Big(x \cdot ln(\mu) - x \cdot ln(1-\mu) + ln(1-\mu)\Big) \\
&= exp\Big(x \cdot \Big(ln(\mu) - ln(1-\mu)\Big) + ln(1-\mu)\Big) \\
&= exp\Big(x \cdot ln(\frac{\mu}{1-\mu}) + ln(1-\mu)\Big) \\
&= exp\Big(ln(\frac{\mu}{1-\mu}) \cdot x + ln(1-\mu)\Big) \\
&= h(x)e^{\eta^T \cdot T(x) - A(\eta)},
\end{aligned}$$

where

- $h(x) = 1$,
- $\eta = ln(\frac{\mu}{1-\mu})$,
- $T(x) = x$,
- $A(\eta) = -ln(1-\mu)$,

or for the alternativee form we would continue from

$$\begin{aligned}\mathbb{P}(x|\mu) &= exp\Big(ln(\frac{\mu}{1-\mu})\cdot x + ln(1-\mu)\Big)\\ &= exp\Big(ln(\frac{\mu}{1-\mu})\cdot x\Big)\cdot exp\Big(ln(1-\mu)\Big)\\ &= (1-\mu)\cdot exp\Big(ln(\frac{\mu}{1-\mu})\cdot x\Big),\end{aligned}$$

where

- $h(x) = 1$,
- $\eta = ln(\frac{\mu}{1-\mu})$,
- $T(x) = x$,
- $g(\eta) = 1-\mu$.

The observant reader might have noticed that we could have obtained $g(\eta)$ through the relationship:

$$\begin{aligned}g(\eta) &= exp\big(-A(\eta)\big)\\ &= exp\Big(-\big(-ln(1-\mu)\big)\Big)\\ &= exp\Big(ln(1-\mu)\Big)\\ &= 1-\mu.\end{aligned}$$

It might seem that we are finished at this point, but we are not quite because $A(\eta)$ (and thus also $g(\eta)$) are not dependent on the natural parameter but on μ (in this case). So we need to replace μ by the natural parameter η. Luckily for us the equation $\eta = ln(\frac{\mu}{1-\mu})$ gives us a relationship between them where we need to reformulate the expression to μ:

$$\begin{aligned}
\eta &= ln(\frac{\mu}{1-\mu}) \\
\Leftrightarrow exp(\eta) &= exp(ln(\frac{\mu}{1-\mu})) \\
\Leftrightarrow exp(\eta) &= \frac{\mu}{1-\mu} \\
\Leftrightarrow (1-\mu)\cdot exp(\eta) &= \mu \\
\Leftrightarrow exp(\eta) - \mu \cdot exp(\eta) &= \mu \\
\Leftrightarrow -\mu \cdot exp(\eta) - \mu &= -exp(\eta) \\
\Leftrightarrow -\mu \cdot \Big(exp(\eta) + 1) &= -exp(\eta) \\
\Leftrightarrow -\mu &= -\frac{exp(\eta)}{exp(\eta)+1} \\
\Leftrightarrow \mu &= \frac{exp(\eta)}{exp(\eta)+1} \\
&= \frac{exp(\eta)}{exp(\eta)\big(1+exp(-\eta)\big)} \\
&= \frac{1}{1+exp(-\eta)} = \sigma(\eta),
\end{aligned}$$

where $\sigma(\cdot)$ is known as the **standard logistic function** (which is often used as an activation function in artificial neural networks). Now that we have this relationship we can conclude that

$$\begin{aligned}
A(\eta) &= -ln(1-\mu) \\
&= -ln\big(1-\frac{1}{1+exp(-\eta)}\big) \quad = -ln\big(\frac{exp(-\eta)}{1+exp(-\eta)}\big) = -ln\big(\frac{1}{1+exp(\eta)}\big) = -ln\big(\sigma(-\eta)\big)
\end{aligned}$$

or

$$\begin{aligned}
g(\eta) &= 1-\mu \\
&= 1-\frac{1}{1+exp(-\eta)} \\
&= \frac{exp(-\eta)}{1+exp(-\eta)} \\
&= \frac{1}{1+exp(\eta)} = \sigma(-\eta).
\end{aligned}$$

Chapter 6

Bayesian Regression

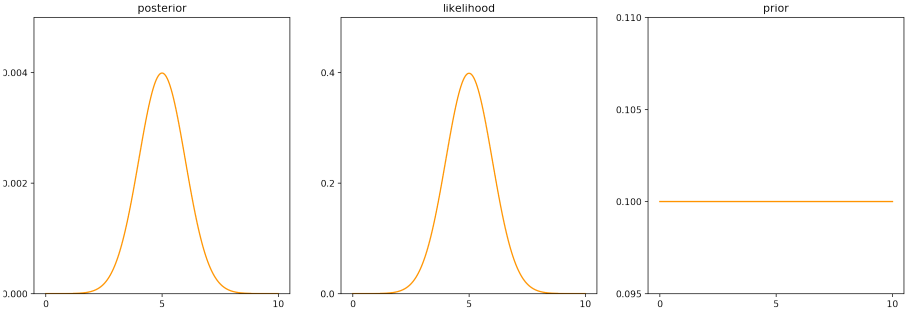
First of all let us take a step back and look at what exactly we have been calculating in this course up until now - the maximum likelihood of model parameters. The likelihood functions we set up in order to approach this were always distribution functions of the likelihood over the value(s) of the model parameter(s) in question. But why did we consider the likelihood $\mathbb{P}(D|\theta)$? Let us call up Bayes theorem again:

$$\underbrace{\mathbb{P}(\theta|D)}_{posterior} = \frac{\overbrace{\mathbb{P}(D|\theta)}^{likelihood} \qquad \overbrace{\mathbb{P}(\theta)}^{prior}}{\underbrace{\mathbb{P}(D)}_{nomalization\ term/evidence}}$$

Ultimately, we are interested in the posterior $\mathbb{P}(\theta|D)$ or rather the maximum of the posterior distribution (called MAP estimate, from maximum a posteriori), i.e. the most likely set of model parameters given the data we observed. This posterior distribution is the (normalized) product of the likelihood distribution and the prior distribution. If we now assume the prior distribution to be flat (i.e. constant, non- informative), the posterior distribution will be a normalized version of the likelihood distribution (and the MAP estimates will be exactly the same as the ML estimates). Making this assumption of flat priors is precisely what we have been doing all along when we used the maximum likelihood estimates in place of the MAP estimates.

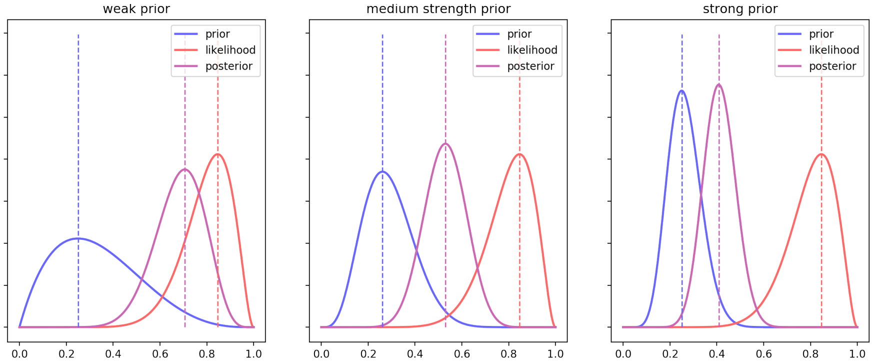
Although it makes things simple, assuming flat priors sometimes simply is not the most accurate thing to do. Imagine you take a coin out of your pocket and throw it two times and both times heads comes up. Using MAP with a flat prior (i.e. MLE) will suggests that the best estimate for the parameter θ of the underlying Bernoulli distribution should be $\theta_{MLE} = \frac{k}{n} = \frac{2}{2} = 1$; this suggests that the arising model will predict (at this point) that all further coin tosses will result in heads. Only looking at the data this makes sense, but we have the **prior knowledge** that most coins will have a θ around $0,5$. To factor this knowledge in we would have to use a prior that is more likely for values around

$\theta = 0,5$.
When assuming a noninformative prior, the shape of the posterior is equivalent to that of the likelihood because the prior is constant and the evidence is just scaling the whole distribution. This means the posterior is only dependent on the data. First consider a case where we assume a flat prior:



Next, we will consider three different cases with non-flat priors. For each we will keep the likelihood identical and change the prior to

- a weak prior, i.e. a prior that is quite flat,
- a medium prior, i.e. a prior that is a bit sharper,
- a strong prior, i.e. a prior that is quite sharp and thus quite sure about certain values of the parameter θ.



As we can see a bayesian regression approach (i.e. the taking into account a non-flat prior) acts as a restriction with regard to the parameter range, similar to regularization and other means to control model complexity (see "Where do $L1$ and $L2$ norm come from?). Note that in contrast to a regularization where the model parameters are dragged towards zero, a Bayesian prior will always drag the model parameters towards the maximum of the prior distribution. In a regularization the added term acts as a prior centred around zero which makes sense, since the idea of a regularization is based on the prior assumption that a complex model (i.e. one with high weights) will fit the specific noise pattern and is therefore unlikely to be a good approximation of the underlying model. With regards to the choice of the distribution family for the prior we are basically free to choose whatever we deem ideal to formalize our prior knowledge or belief. Nevertheless, in many cases it is of advantage to choose the prior distribution from a family such that a multiplication with the likelihood will result in a posterior distribution of the same family as the prior distribution. When this is the case, prior and posterior are called conjugate distributions with the prior being a conjugate prior to the likelihood function.

6.1 Conjugate prior and MAP estimates for Bernoulli

Let us start with an example: We conduct an experiment in which we toss a thumb-tack n times ($x_1, ..., x_n =: X$) and want to determine the probability for it to fall on its head (heads are encoded as 1, tails = 0). We assume the data to be i.i.d. and set up the likelihood function using the Bernoulli distribution:

$$\begin{aligned}\mathbb{P}(X|\theta) &= \prod_{i=1}^{n} \theta^{x_i}(1-\theta)^{1-x_i} \\ &= \theta^h(1-\theta)^t,\end{aligned}$$

with h being the number of heads and t being the number of tails ($n = h + t$). The maximum likelihood estimate that we can derive from this expression is only equal to the MAP estimate if we assume a flat prior. We may however have a certain prior belief (e.g. that this thumb tack has a θ of roughly 0.5) or data from previous experiments with the same thumb-tack and want to take this into account for our MAP estimate. Let us recall Bayes theorem and update it with the variable names X and θ:

$$\mathbb{P}(\theta|X) = \frac{\mathbb{P}(X|\theta)\mathbb{P}(\theta)}{\mathbb{P}(X)}.$$

To express our prior belief we opt for a conjugate prior $\mathbb{P}(\theta)$. In the case of a Bernoulli distributed likelihood function, the conjugate prior will come from a beta-distribution:

$$\mathbb{P}(\theta) = \frac{1}{B(\alpha, \beta)}\theta^{\alpha-1}(1-\theta)\beta - 1.$$

Since $\frac{1}{B(\alpha,\beta)}$ is independent of θ we may drop it for the purpose of setting up the posterior distribution:

$$\begin{aligned}\mathbb{P}(\theta|X) &\propto \mathbb{P}(X|\theta) \cdot \mathbb{P}(\theta) \\ &= \theta^h(1-\theta)^t \cdot \theta^{\alpha-1}(1-\theta)^{\beta-1} \\ &= \theta^{h+\alpha-1}(1-\theta)^{t+\beta-1}\end{aligned}$$

And here is the full posterior distribution:

$$\mathbb{P}(\theta|X) = \frac{\theta^{\alpha_n - 1}(1-\theta)^{\beta_n - 1}}{B(\alpha_n, \beta_n)}$$

We can now go on to derive a MAP estimator just like we previously derived ML estimators. This will result in

$$\theta_{MAP} = \frac{\alpha + h - 1}{\alpha + \beta + h + t - 2}.$$

6.2 Sequential updates to the posterior/ prior

In the section above we can see that the only difference between the posterior function and the likelihood function is the addition of $\alpha - 1$ and $\beta - 1$ to the exponents: α essentially increases to the number of observed head events, while β adds to the number of observed tail events. This comes in handy when we want to have sequential updates of the MAP estimate, i.e. when we only want to consider one data point at a time. In this case before even measuring the first data point we set up a prior representing our belief about the variables in question. If we have no such belief, expectation, personal bias, preference or opinion about the matter (or do not want it to influence our MAP estimates), we use a flat prior. If we choose to sequentially update our MAP estimate, i.e. after each new observation, we have two equivalent options: 1) We keep our original prior (i.e. the parameters of the distribution) as it is and put all observations in an ever growing pool. After each new observation, the likelihood distribution is set up based on the complete set of observations and multiplied with the prior distribution. 2) We take advantage of the conjugate prior which allows us to use the posterior distribution after the ith observation as the prior distribution for the $(i+1)$th observation. What is happening is that we sum up all past observations into one distribution and update this distribution after each new observation with a relatively simple likelihood term.

Note: The two options are equivalent with respect to the resulting MAP estimates, but the second option is computationally more efficient. If for example each observation is a high resolution image and you get 25 new images per second, after a while you will have a rather large amount of data. In approach (1) you would have to save all the data and evaluate all of it 25 times a second. This requires ever-growing storage and computational power. In contrast, with approach (2) you only have a small and constant demand for storage and computational power. All you need to store is the current prior/ posterior distribution (which changes its parameter values but does not actually grow) and the latest image. The likelihood estimation is then based solely on the mentioned latest image, not on say a million images or more.

Chapter 7

Appendix: Basics of probability theory

7.1 Probability spaces

7.1.1 Finite and infinite probability space

We want to come closer to understanding the term probability and we start with probability spaces. Formally, a finite probability space consists of two parts:

Finite sample space

In a random experiment we call the set of all **elementary outcomes** Ω, the **sample space**.
If you consider a coin toss, you could model the sample space as $\Omega = \{0, 1\}$, where 0 represents heads and 1 tails. On the other hand, for a dice throw we would set $\Omega = \{1, 2, 3, 4, 5, 6\}$. Note that we did not say anything about the probabilities of those elementary events, i.e. the sample spaces of a fair dice and a rigged dice are identical.
Of course when measuring and calculating probabilities later we want to assign probabilities to certain events, e.g "What is the probability of throwing a dice with a number of dots smaller or equal to 4?".
Suddenly we are interested in what is called **events**. Mathematically, an event is a subset $A \subseteq \Omega$. We call the empty set ø the *impossible event* and Ω the *sure event*. For the previous question the event would be defined as $\{1, 2, 3, 4\}$. The set of all possible subspaces of Ω is the power set $\mathcal{P}(\Omega)$. For the coin toss this implies that the set of of possible events is

$$\mathcal{P}(\Omega) = \mathcal{P}(\{0, 1\}) = \{ø, \{0\}, \{1\}, \{0, 1\}\}.$$

Probability measure

A probability of an event is a number $p \in [0,1]$. The probability measure is a function which assigns each event in Ω a probability.
Let us assume we have a finite Ω. A **probability measure** is a function $\mathbb{P} : \mathcal{P}(\Omega) \to [0,1]$ with the following properties:

1. $\mathbb{P}(\Omega) = 1$.
2. $\mathbb{P}(A) + \mathbb{P}(B) = \mathbb{P}(A \cup B)$ for all $A, B \in \mathcal{P}(\Omega)$ with $A \cap B = \emptyset$ (A and B are *disjoint*)
3. $0 \leq \mathbb{P}(A) \leq 1$ for all $A \in \mathcal{P}(\Omega)$

We call the tuple $(\Omega, \mathbb{P})$ a **finite probability space** and $\mathbb{P}(A)$ the *probability* of event $A \subseteq \Omega$.
Here are some properties of the probability measure:

- $\mathbb{P}(A^c) = 1 - \mathbb{P}(A)$ for $A^c = \Omega \setminus A$.
- $\mathbb{P}(\emptyset) = 0$.
- $A \subset B \Rightarrow \mathbb{P}(A) \leq \mathbb{P}(B)$.
- $\mathbb{P}(B \setminus A) = \mathbb{P}(B) - \mathbb{P}(A \cap B)$.
- For pairwise disjoint subsets $A_1, \dots A_n \in \Omega$, i.e. $A_i \cap A_j = \emptyset$ for $i \neq j$ it holds that

$$\mathbb{P}(A_i \cup \cdots \cup A_n) = \mathbb{P}(A_1) + \cdots + \mathbb{P}(A_n) = \sum_{i=1}^{n} \mathbb{P}(A_i).$$

Furthermore using the definition of the probability measure and these theorems it holds that

$$\mathbb{P}(A) = \sum_{\omega \in A} \mathbb{P}(\{\omega\}).$$

Probability mass function (PMF)

The probability measure $\mathbb{P}$ assigns each possible event in $\mathcal{P}(\Omega)$ a probability in $[0,1]$. The **probability mass function** on the other hand is a function mapping each elementary event $\omega \in \Omega$ to a probability. Thus we define a probability mass function

$$f : \Omega \to \mathbb{R}, \omega \mapsto f(\omega) = \mathbb{P}(\{\omega\})$$

with properties

1. $f(\omega) \geq 0$ for all $\omega \in \Omega$

2. $\sum_{\omega\in\Omega} f(\omega) = 1$.

Here we defined the probability mass function using the probability measure. Of course, given a probability mass function f we can also define a probability measure using

$$\mathbb{P} : \mathcal{P}(\Omega) \to [0,1], A \mapsto \sum_{\omega\in A} f(\omega).$$

We call f the corresponding probability mass function of $\mathbb{P}$.

σ-Algebra

The previous definition of a probability space works if Ω is finite or countable (e.g. $\mathbb{N}$ and $\mathbb{Z}$). For the application of probability theory it is important to consider cases where Ω is neither finite nor countable, but for example the interval $[0, 1]$. For this purpose we have to introduce the idea of a σ-algebra, the Borel-σ-Algebra to be more specific. Since both of these mathematical objects can be a bit tedious to work with, we will try to keep it as simple as possible (even if this leads to more informal definitions).
The σ-algebra comes from measure theory. Earlier we said $\mathcal{P}(\Omega)$ contains all sets that we want to assign a probability to. But as it turns out there are some subsets we can not assign probabilities to, e.g. $\Omega = \mathbb{R}$. To solve this issue we will need to introduce the notion of a Borel-σ-algebra after we have defined a standard σ-algebra.
Let $\mathcal{A}$ be the set of all subsets of Ω. We call $\mathcal{A}$ a σ-algebra (on Ω), if the following properties hold:

1. $\Omega \in \mathcal{A}$
2. $A \in \mathcal{A} \Rightarrow A^c \in \mathcal{A}$
3. For every sequence $(A_n)_{n\in\mathbb{N}}$ of sets in $\mathcal{A}$, it holds that $\bigcup_{n=1}^{\infty} A_n \in \mathcal{A}$

We call the tuple $(\Omega, \mathcal{A})$ a **measurable space**. It is the set of all elementary events and the set of all the subsets that we are willing to assign a measure to. As you can easily show for the finite case, $\mathcal{P}(\Omega)$ fulfils all these properties making the power set a σ-algebra for finite Ω. Intuitively, if Ω is not countable, e.g. $\Omega = \mathbb{R}$, we would set $\mathcal{A} = \mathcal{P}(\mathbb{R})$, but here the name *Vitali* comes in: Vitali was a mathematician who showed that there are certain subsets of $\mathbb{R}$ in $\mathcal{P}(\mathbb{R})$ called *Vitali-sets* that we can not assign a measure to. Why? If we were to assign measures to them, this would lead to contradictions such as $1 = 0$. For this reason, setting $\mathcal{A} = \mathcal{P}(\mathbb{R})$ is not an option. Instead we will use what is a called a **Borel-σ-algebra**: To keep it very simple, we can view it as a reduced version of $\mathcal{P}(\mathbb{R})$ that does not contain any of those problematic sets but instead is the smallest possible σ-algebra which contains all the sets that we are interested in. We denote it with $\mathcal{B}(\mathbb{R})$ (equivalently $\mathcal{B}([0, 1])$).
This might seem a bit weird, since we (at least me as the writer) do not exactly know the difference between $\mathcal{P}(\mathbb{R})$ and $\mathcal{B}(\mathbb{R})$ but this steps ensures soundness.

Probability space

Now with all preparations made we are ready to define the probability measure and probability space for every possible assignment of Ω: Let $\Omega \neq \emptyset$ be a set and $\mathcal{A}$ be a σ-algebra on Ω. A probability measure on Ω is a function

$$\mathbb{P} : \mathcal{A} \to [0,1]$$

with the following properties (called the Kolmogorov-axioms):

1. $\mathbb{P}(\Omega) = 1$

2. (σ-additivity) Let $A_1, A_2, \dots \in \mathcal{A}$ be pairwise disjoint, i.e. $A_i \cap A_j = \emptyset$ for $i \neq j$, then

$$\mathbb{P}(\bigcup_{n=1}^{\infty} A_n) = \sum_{n=1}^{\infty} \mathbb{P}(A_n)$$

3. $0 \leq \mathbb{P}(A) \leq 1$ for all $A \in \mathcal{A}$.

The triplet $(\Omega, \mathcal{A}, \mathbb{P})$ is called probability space and $\mathbb{P}(A)$ the probability of event $A \in \mathcal{A}$. The elements of Ω are called elementary events and $\mathcal{A}$ is called the σ-algebra of events.
(Note that it is convention to leave out the additional set brackets and to write $\mathbb{P}(\omega)$ instead of $\mathbb{P}(\{\omega\})$.

Here are some properties of the probability measure:

- $\mathbb{P}(A^c) = 1 - \mathbb{P}(A)$ for $A \in \mathcal{A}$.

- $\mathbb{P}(/o) = 0$.

- $A, B \in \mathcal{A}, A \subset B \Rightarrow \mathbb{P}(A) \leq \mathbb{P}(B)$.

- $\mathbb{P}(B \setminus A) = \mathbb{P}(B) - \mathbb{P}(A \cap B)$ for all $A, B \in \mathcal{A}$.

- $\mathbb{P}(A \cup B) = \mathbb{P}(A) + \mathbb{P}(B)$ for all disjoint sets $A, B \in \mathcal{A}$.

- For all sets $A, B \in \mathcal{A}$

$$\mathbb{P} A \cup B = \mathbb{P}(A) + \mathbb{P}(B) - \mathbb{P}(A \cap B).$$

Furthermore, using the definition of the probability measure and these theorems it holds that

$$\mathbb{P}(A) = \sum_{\omega \in A} \mathbb{P}(\{\omega\}).$$

Probability density function

We talked about the probability mass function before, which assigns each elementary event a probability for a finite or countably infinite Ω. As you might have thought, there is an equivalent function for non-countable infinite Ω.
Let $f : \mathbb{R} \to [0, \infty[$ a piecewise continuous function with

$$\int_{-\infty}^{\infty} f(x)dx = 1,$$

then f is a probability density function. If f is a pdf, then we can induced a probability measure

$$\mathbb{P}([a, b]) = \int_{a}^{b} f(x)dx,$$

for all $a, b \in \mathbb{R}$ with $a < b$
Note that not every probability measure has a corresponding probability density function.

7.1.2 Independence and conditional probabilities

Let $(\Omega, \mathcal{A}, \mathbb{P})$ be a probability space. For $B \in \mathcal{A}$ with $\mathbb{P}(B) > 0$ we define:

$$\mathbb{P}(A|B) = \frac{\mathbb{P}(A \cap B)}{\mathbb{P}(B)} \text{ for all } A \in \mathcal{A}$$

is the **conditional probability** of A given B.

Next we will cover a few **very important theorems**:

1. (**Product rule**) Let $A_1, \dots, A_n \in \mathcal{A}$ with $\mathbb{P}(A_1 \cap \dots \cap A_n - 1) > 0$, then it holds that

$$\mathbb{P}(A_1 \cap \dots \cap A_n) = \mathbb{P}(A_n|A_1 \cap \dots \cap A_n - 1) \cdot \dots \cdot \mathbb{P}(A_2|A_1) \cdot \mathbb{P}(A_1).$$

2. (**Law of total probability / Sum rule**) Let $B_1, \dots, B_n$ be a disjoint decomposition of Ω (i.e. $\bigcup_{i=1}^{n} B_i = \Omega$ and the B_i are pairwise disjoint) with $\mathbb{P}(B_i) > 0$, then it holds that:

$$\mathbb{P}(A) = \sum_{i=1}^{n} \mathbb{P}(A \cap B_i) = \sum_{i=1}^{n} \mathbb{P}(A|B_i) \cdot \mathbb{P}(B_i) \text{ for all } A \in \mathcal{A}.$$

3. (**Bayes' Theorem**) Let $B_1, \dots, B_n$ be a disjoint decomposition of Ω, then for all $A \in \mathcal{A}$ with $\mathbb{P}(A) > 1$ it holds that:

$$\mathbb{P}(B_i|A) = \frac{\mathbb{P}(A|B_i) \cdot \mathbb{P}(B_i)}{\mathbb{P}(A)} = \frac{\mathbb{P}(A|B_i) \cdot \mathbb{P}(B_i)}{\sum_{k=1}^{n} \mathbb{P}(A|B_k) \cdot \mathbb{P}(B_k)}$$

Next we will cover the term **independence** of events:
Let $I \neq \emptyset$ be set of indices and $J \subset I$ a selection of those indices. We call the events $(A_i)_{i\in I}$ stochastically independent, if

$$\mathbb{P}(\bigcap_{j\in J} A_j) = \prod_{j\in J} \mathbb{P}(A_j)$$

7.2 Random variables and distributions

In the following we will talk about random variables. To understand their use case, consider the following example:
We consider throwing a dice twice and choose a Laplace-space:

$$\Omega = \{1, 2, \dots, 6\}^2 \text{ with } \mathcal{A} = \mathcal{P}(\Omega) \text{ and } \mathbb{P}(\omega) = \frac{1}{36} \text{ for } \omega \in \Omega.$$

We now ask ourselves: What is the probability that the sum of eyes is k for $k = 2, 3, \dots, 12$? So we are not interested in a specific ω, instead we are interested in a property.
Let $A_k = \{(i, j) \in \Omega | i + j = k\}$ for $k = 2, 3, \dots, 12$. Then

$$\mathbb{P}(A_k) = \begin{cases} \frac{k-1}{36} & \text{if } k \in \{2, 3, ..., 7\} \\ \frac{13-k}{36} & \text{if } k \in \{8, 9, 10, 11, 12\}. \end{cases}$$

Now we consider

$$\tilde{\Omega} = \{2, 3, ..., 12\} \text{ with } \tilde{\mathbb{P}}(k) = \mathbb{P}(A_k) \text{ for } k \in \tilde{\Omega}.$$

$(\tilde{\Omega}, \tilde{\mathbb{P}})$ is also a probability space, but not a Laplace-space. This probability space also has much more complicated probabilities of the elementary events, but it is more adequate chosen for our question. For the connection between those two probability spaces we can look at the function:

$$X : \Omega \to \tilde{\Omega}, (i, j) \mapsto i + j,$$

with $X^{-1}(\{k\}) = A_k$.

7.2.1 Random variables

Measurable functions

For properly introducing random variables we will first need to understand measurable functions. As we learned before, we constructed the σ-algebra $\mathcal{A}$ to define which events we can assign a measure to. We will call all those subsets that are contained in $\mathcal{A}$ measurable.

Let $(\Omega, \mathcal{A})$ and $(\Omega', \mathcal{A}')$ measurable spaces and $f : \Omega \to \Omega'$ a function. We call f measurable (to be more precise $\mathcal{A} - \mathcal{A}'$-measurable), if

$$f^{-1}(A') \in \mathcal{A} \text{ for all } A' \in \mathcal{A}'.$$

We express the $\mathcal{A} - \mathcal{A}'$-measurability through

$$f : (\Omega, \mathcal{A}) \to (\Omega', \mathcal{A}').$$

So we demand, that all inverse-images of measurable sets are again measurable.

Random variables

Now we are ready to define random variables:
Let $(\Omega, \mathcal{A}, \mathbb{P})$ a probability space and $(\Omega', \mathcal{A}')$ a measurable space. We call the measurable function

$$X : (\Omega, \mathcal{A}, \mathbb{P}) \to (\Omega', \mathcal{A}').$$

a random variable.

We will write $\{X \in F\} = X^{-1}(F)$ and $\{X = s\} = X^{-1}(s)$.

Theorems concerning RV

Let $(\Omega, \mathcal{A}, \mathbb{P})$ a probability space and $X, Y : \Omega \to \mathbb{R}$ be a RV. Then the following functions are also random variables:

1. $aX : \Omega \to \mathbb{R}, \omega \mapsto aX(\omega)$ for $a \in \mathbb{R}$.
2. $X + Y : \Omega \to \mathbb{R}, \omega \mapsto X(\omega) + Y(\omega)$.
3. $X \cdot Y : \Omega \to \mathbb{R}, \omega \mapsto X(\omega) \cdot Y(\omega)$.
4. $\frac{X}{Y} : \Omega \to \mathbb{R}, \omega \mapsto \frac{X(\omega)}{Y(\omega)}$, if $Y(\omega) \neq 0$ for all $\omega \in \Omega$.

Distributions

You might have wondered at this point why we have not introduced the term **distribution** yet. Although your intuition might tell you the term is interchangeable with probability measure, this is not formally right.
As we saw in the beginning examples for the two dice throws, each random variable induces a probability measure in the new measurable space:

Let $(\Omega, \mathcal{A}, \mathbb{P})$ a probability space and $(\Omega', \mathcal{A}')$ a measurable space. Let $X : \Omega \to \Omega'$ be a random variable. We define the image-measure $\mathbb{P}_X : \mathcal{A}' \to [0, 1]$ with

$$\mathbb{P}_X(A') := \mathbb{P}(X \in A') = \mathbb{P}(X^{-1}(A')), \text{ for } A' \in \mathcal{A}'.$$

Then $\mathbb{P}_X$ is a probability measure on Ω', thus $(\Omega', \mathcal{A}', \mathbb{P}_X)$ is a probability space.

We call $\mathbb{P}_X$ the **distribution of** X.

If $X : (\Omega, \mathcal{A}, \mathbb{P}) \to (\Omega, \mathcal{A})$ is the identity, thus $X(\omega) = \omega$, then of course $\mathbb{P}_X = \mathbb{P}$. This is why we often use probability measure and distribution interchangeably.

If the distribution $\mathbb{P}_X$ of X is describable by a density f_X, we call f_X the density of X.

Independence and random variables

Let $(\Omega, \mathcal{A}, \mathbb{P})$ be a probability space, $I \neq \emptyset$ be the set of indices and $J \subset I$ a selection of those indices. The random variables $(X_i)_{i \in I} : \Omega \to \mathbb{R}$ are called independent, if the family of sets

$$\{X_j \in A_j\}_{j \in J}$$

are independent, where $(A_j)_{j \in J} \subset \mathbb{R}$ are measurable sets.

7.2.2 Examples for classical distributions

Discrete and continuous random variables

Let X be a random variable. If $X(\Omega)$ is discrete, we call X discrete. If $X(\Omega)$ is continuous, we call X continuous.

Discrete distributions

- **(Discrete) uniform distribution**: Let X be a discrete RV with $\{a_1, a_2, \dots, a_n\} \subset \mathbb{R}$. The random variable X is called (discretely) uniformed distributed, if there exists a $x \in [0, 1]$, with $\mathbb{P}(X = a_i) = c$ for all $i = 1, \dots, n$. Of course $c = \frac{1}{n}$ and

  $$\mathbb{P}_X(a_i) = \mathbb{P}(X = a_i) = \frac{1}{n}.$$

  We write $X \sim U(X(\Omega))$.

  Intuition The discrete uniform distribution is used every time, when each elementary outcome has the same probability. Some examples are

  - a fair coin toss,
  - a fair dice throw,
  - drawing a specific card from a card deck,

  - guessing a number between 1 and 100.

The keen reader might be confused by the last example. Although it seems somewhat reasonable to assume a uniform distribution for guessing a random number, in reality some numbers (such as 77 and 69) are just more likely than others. It depends on how you want to model a certain phenomenon.
In those cases, the probability measure $\mathbb{P}$ of the corresponding probability space is defined by

$$\mathbb{P} : \mathcal{A} \to [0,1], A \mapsto \frac{|A|}{|\Omega|}.$$

This is called a Laplacian probability space. For drawing a random card from a deck of 52 cards, a probability space can look something like this:

  - $\Omega = \{1, 2, \ldots, 52\}$,
  - $\mathcal{A} = \mathcal{P}(\Omega)$,
  - $\mathbb{P}(A) = \frac{|A|}{|\Omega|} = \frac{|A|}{52}$.

- **Bernoulli distribution**: The random variable X is called Bernoulli distributed, if $X(\Omega) = \{s, t\}$ with

  $$\mathbb{P}(X = s) = p, \mathbb{P} = q \text{ with } p + q = 1.$$

  If s and t are not explicitly specified, then $s = 0$ and $t = 1$.
  We write $X \sim Ber(p)$.

  Intuition The Bernoulli distribution, although very simple, will be very handy. We will use it everytime a random event has exactly two outcomes, e.g. a coin toss.
  In those cases we will assign one outcome to the number 0 and the other to the number 1. p is the probability of the event corresponding to number 1 happening, while $q = 1 - p$ is the probability of the event happening that is corresponding to number 0.
  Once we mapped the events s and t to numbers 0 and 1, we can rewrite the probability measure into

  $$\mathbb{P}(X = k) = p^k q^{1-k} = p^k (1-p)^{1-k}.$$

  If $k = 0$, this resulting probability is $(1-p)$, while for $k = 1$ the probability comes out to p; just as expected.
  For a coin toss, where 0 is tails and 1 is tails, a probability space can look something like this:

  - $\Omega = \{0, 1\}$,
  - $\mathcal{A} = \mathcal{P}(\Omega) = \{\emptyset, \{0\}, \{1\}, \{0, 1\}\}$,

- $\mathbb{P}(X = k) = p^k(1-p)^{1-k}$.

- **Binomial distribution**: The random variable X is called binomial distributed, if $X(\Omega) = \{0, 1, \ldots, n\}$ and for all $k = 0, 1, \ldots, n$

$$\mathbb{P}(X = k) = \binom{n}{k} p^k q^{n-k} \text{ with } p + q = 1.$$

We write $X \sim bin_{n,p}(k)$.

Intuition Now imagine this: You throw a coin n times and are interested in the probability that of those n coin tosses k showed heads. Obviously each individual toss can be modelled by a Bernoulli distribution, but what about a sequence of Bernoulli distribution (something very common)?
The Binomial distribution is the answer to this question. Let us try to derive its equation: We can start off to model n Bernoulli trials, where 1 represents a *success* and 0 a *fail* and we are interested in the amount of successes. The first idea is to simply use the product of multiple Bernoulli trials (with the same probability p):

$$\begin{aligned}\mathbb{P}(X = k) &= \prod_{i=1}^{n} p_i^x (1-p)^{1-x_i} \\ &= p_1^x(1-p)^{1-x_1} \cdot p_2^x(1-p)^{1-x_2} \cdot \dots \cdot p_n^x(1-p)^{1-x_n} \\ &= p^{\sum_{i=1}^n x_i}(1-p)^{n-\sum_{i=1}^n x_i}, \\ &= p^k(1-p)^{n-k},\end{aligned}$$

but we are not yet finished: This probability describes one possibility of having k successes in n trials, but we have to take into consideration that there are multiple ways of achieving this. Consider three coin tosses; there are 3 possibilities of heaving exactly two heads. So what is left to do, is to multiply the term above by the number of possibilities. But how many possibilities are there to have k successes in n trials? This number is given by the binomial coefficient, given by

$$\binom{n}{k} := \frac{n!}{k!(n-k)!}$$

Why does this makes sense? First it is important to understand that, the questions *"how many possibilities are there to have k successes in n trials"* and *"how many possibilites are there to choose k from n"* have the same answers. So how many possibilities are there to pick k from n? Well there are n possibilities for the first choice, then $n-1$ for the second choice, $\ldots$, and finally $n-k+1$ for the kth choice. We can write $n \cdot (n-1) \cdot \ldots \cdot (n-k+1)$ as $\frac{n!}{(n-k)!}$. But this number is a bit to high, because in our case the order does not matter; what is meant by that is, that it doesn't matter if the order is $5-2-3$ or $2-5-3$. This is why we have to divide by all possible orders $k!$, resulting in $\frac{n!}{k!(n-k)!}$

- **Poisson distribution**: The random variable X is called Poisson distributed, if there exists a family of elements $(s_k)_{k\in\mathbb{N}}$ in $X(\Omega)$ and a $\lambda \in \mathbb{R}$ with $\lambda \geq 0$, such that

  $$\mathbb{P}(X = s_k) = e^{-\lambda}\frac{\lambda^k}{k!} \text{ for } k \geq 0.$$

  We write $X \sim Po(\lambda)$.

  Intuition We consider a fixed time interval in which the expected number of (positive) events is given by the parameter λ, i.e. we know that an event will occur, on average, λ times within the interval. Of course, on some trials the number of events might be a little above λ, on some it might be a little below it. We denote with $\mathbb{P}(X = k)$ the probability of k events occurring in this interval, when the expected number of events is λ.
  Mathematically, the Poisson distribution is derived by the binomial distribution.

Continuous distributions

- **(Continuous) uniform distribution**: Let X be a continuous RV with $X(\Omega) = [a, b]$. We call X uniform distributed, if its probability density function is $f(x) = \frac{1}{b-a}\mathbb{I}_{[a,b]}(x)$, i.e.

  $$\mathbb{P}(X \leq t) = \int_a^t f(x)dx = \frac{t-a}{b-a} \text{ for } t \in [a, b].$$

  We write $X \sim U([a, b])$.

- **Normal distribution**: The RV X is called normal distributed, if its probability density function is $f(x) = \frac{1}{\sqrt{2\pi\sigma^2}}e^{-\frac{(x-\mu)^2}{2\sigma^2}}$, i.e.

  $$\mathbb{P}(X \leq t) = \int_{-\infty}^t f(x)dx.$$

  We write $X \sim \mathcal{N}(\mu, \sigma)$.

7.2.3 Change of densities

Let us assume the following: We are working with a probability space $(\Omega, \mathcal{A}, \mathbb{P})$, where Ω is continuous and f is the probability density function of probability measure $\mathbb{P}$. If we now define a random variable X, such that it projects from $(\Omega, \mathcal{A}, \mathbb{P})$ on to a different probability space $(\Omega_X, \mathcal{A}_X, \mathbb{P}_X)$, we might be interested in whether $\mathbb{P}_X$ can be described by an inferred probability density function f_X and how f_X might look like. This is the purpose of the following theorem, known as "change of densities":

Let $I \subset \mathbb{R}$ be an interval and $f : I \to [0, \infty[$ a continuous probability density function of $\mathbb{P}$ on I. Let $X : I \to \mathbb{R}$ be a random variable, which is continuously differentiable and strictly monotonic. Let $Y = X^{-1}$ be the inverse function of X, which also is continuously differentiable. Then the distribution $\mathbb{P}_X$ has the density function g, with

$$g : X(I) \to \mathbb{R}, y \mapsto |\frac{\partial}{\partial y} Y(y)| \cdot f(Y(y)).$$

Why is this? Well the concept of the proof looks something like this: We know that from the fact that X is bijective it follows that $X(I) = J$ is also an interval. If we can show that

$$\mathbb{P}_X([c, d]) = \int_c^d g(y) dy, \forall [c, d] \subset J$$

then g is the density function of $\mathbb{P}_X$.

We choose $a, b \in I$ with $X(a) = c, X(b) = d$. If X is now strictly monotonically **increasing**, then $a \leq b$ and thus:

$$\begin{aligned} \mathbb{P}_X([c, d]) &= \mathbb{P}(X^{-1}([c, d])) \\ &= \mathbb{P}([a, b]) \\ &= \int_a^b f(x) \partial x \end{aligned}$$

if we substitute in regard to y with $y(x) = X(x) \Leftrightarrow x = X^{-1}(y(x)) = Y(y(x))$ and thus $\partial x = \frac{1}{\frac{\partial}{\partial x} X(x)} \partial dy$ this result in

$$\begin{aligned} &= \int_{Y(a)}^{Y(b)} f(Y(y)) \cdot \frac{1}{\frac{\partial}{\partial x} X(x)} \partial dy \\ &= \int_c^d f(Y(y)) \cdot \frac{1}{\frac{\partial}{\partial x} X(x)} \partial dy \end{aligned}$$

Next, we will use a property of derivatives of inverse function: If f has inverse function f^{-1} and f is differentiable at $f^{-1}(x)$ and $f'(f^{-1}(x))$ is not equal to zero, then

$$\frac{\partial}{\partial x} f^{-1}(x) = \frac{1}{f'(f^{-1}(x))}.$$

For us this means that we can rewrite $\frac{\partial}{\partial x} X(x)$ to $\frac{1}{(X^{-1})'(X(x))} = \frac{1}{Y'(X(x))}$ and plug it back in:

$$\begin{aligned} &= \int_c^d f(Y(y)) \cdot \frac{1}{\frac{1}{Y'(X(x))}} \partial dy \\ &= \int_c^d f(Y(y)) \cdot Y'(X(x)) \partial dy \end{aligned}$$

and here we will again substitute $x = Y(y)$:

$$= \int_c^d f(Y(y)) \cdot Y'(X(Y(y)))\partial dy$$

and because X is the inverse of Y this boils down to

$$= \int_c^d f(Y(y)) \cdot Y'(y)\partial dy. \qquad\qquad = \int_c^d g(y)\partial dy.$$

If we perform these same step, but assume that X is strictly monotonically **decreasing**, then we end up with

$$\begin{aligned}\mathbb{P}_X([c,d]) &= \mathbb{P}(X^{-1}([c,d])) \\ &= \mathbb{P}([b,a]) \\ &= \int_b^a f(x)\partial x \\ &= -\int_a^b f(x)\partial x \\ &= -\int_c^d f(Y(y)) \cdot Y'(y)\partial dy \\ &= \int_c^d (-1)f(Y(y)) \cdot Y'(y)\partial dy \\ &= \int_c^d g(y)\partial dy.\end{aligned}$$

As we can see, both times $\mathbb{P}_X([c,d]) = \int_c^d g(y)\partial dy$ holds, proving the theorem.

7.3 Expected values

In many cases we do not want to or we can not specify the explicit distribution, but instead we are interested in characteristic measurement of the random variable. The **expected value** is such a characteristic measurement which is supposed to describe, which value the random variable takes in average. We will first introduce the expected value for the discrete or countable infinite case, then for the continuous case.

7.3.1 Expected value for discrete random variables

Let $(\Omega, \mathbb{P})$ be a discrete probability space and $X : \Omega \to \mathbb{R}$ a random variable. We call the series

$$\mathbb{E}[X] = \sum_{\omega \in \Omega} X(\omega)\mathbb{P}(\omega)$$

the expected value of X, if the series converges absolutely (i.e. $\mathbb{E}[|X|] < \infty$).

7.3.2 Expected value for continuous random variables

Let $(\Omega, \mathcal{A}, \mathbb{P})$ be a probability space with $\Omega = \mathbb{R}$ and $X : \Omega \to \mathbb{R}$ a random variable with a continuous probability density function. We call the integral

$$\mathbb{E}[X] = \int_{-\infty}^{\infty} X(t)f(t)dt$$

the expected value of X, if $\mathbb{E}[|X|] < \infty$.

7.3.3 Theorems

Let $(\Omega, \mathcal{A}, \mathbb{P})$ be a probability space and $X, Y : \Omega \to \mathbb{R}$ two random variables with existing expected value.

- For $s, t \in \mathbb{R}$ the random variable $sX + tY : \Omega \to \mathbb{R}$ has the expected value
$$\mathbb{E}[sX + tY] = s\mathbb{E}[X] + t\mathbb{E}[Y].$$
- $X \leq Y \Rightarrow \mathbb{E}[X] \leq \mathbb{E}[Y]$.
- $|\mathbb{E}[X]| \leq \mathbb{E}[|X|]$.

7.3.4 Transformation theorem

Let $(\Omega, \mathcal{A}, \mathbb{P})$ be a probability space, $X : \Omega \to \mathbb{R}$ be a random variable and $g : \mathbb{R} \to \mathbb{R}$ be a measurable function (then $g(X) : \Omega \to \mathbb{R}$ is again a random variable). The expected value of $g(X)$ is given by

$$\mathbb{E}[g(X)] = \int_{\Omega} g(X(\omega))\mathbb{P}(\{\omega\}) = \int_{-\infty}^{\infty} g(t)f(t)dt,$$

if the expected value of $\mathbb{E}[g(X)]$ exists. For a discrete Ω this means

$$\mathbb{E}[g(X)] = \sum_{\omega \in \Omega} g(X(\omega))\mathbb{P}(\omega) = \sum_{k \in X(\Omega)} g(k)\mathbb{P}_X(k) = \sum_{k \in X(\Omega)} g(k)\mathbb{P}(X = k).$$

7.3.5 Independence and the expected values

Let $X_1, \ldots, X_n$ be independent random variables, then

$$\mathbb{E}[X_1 \cdot \cdots \cdot X_n] = \mathbb{E}[X_1] \cdot \cdots \cdot \mathbb{E}[X_n].$$

7.4 Variance

The expected value alone does not tell us much about the underlying random variable. We want to determine a number for the expected value, that tells us something about the distance of the expected value to the possible outcomes. The distance of X to $\mathbb{E}[X]$ can be described by $|X - \mathbb{E}[X]|$, but this function is not continuous. Instead, we can use a different measurement:

7.4.1 Variance

Let $(\Omega, \mathcal{A}, \mathbb{P})$ be a probability space and $X : \Omega \to \mathbb{R}$ be a random variable with $\mathbb{E}[X^2] < \infty$. Then

$$\mathbb{V}[X] = \mathbb{E}[(X - \mathbb{E}[X])^2]$$

is called the **variance** and $\sigma[X] = \sqrt{\mathbb{V}[X]}$. the standard deviation of X.

7.4.2 Theorems concerning variance

The variable of a random variable has the following properties:

- $\mathbb{V}[X] = \mathbb{E}[X^2] - \mathbb{E}[X]^2$
- $\mathbb{V}[aX + b] = a^2\mathbb{V}[X]$
- $\mathbb{V}[X] = 0 \Rightarrow \mathbb{P}(X = \mathbb{E}[X]) = 1.$