GAUSSIAN PROCESSES

0. Definitions and overview

[1]This course is on Gaussian processes. Our aim is to understand what makes Gaussians special, i.e. why do they appear so widely and what makes them easier to study. We will both meet properties that seem special, but are actually shared by a larger family of models, but also properties that really are special to Gaussians.

Definition 0.1 (Gaussian random variable)**.** *A real-valued random variable X is called Gaussian, if there exist $m \in \mathbb{R}, \sigma > 0$ such that the density of X w.r.t the Lebesgue measure dx is given by*

$$\frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{(x-m)^2}{2\sigma^2}\right).$$

We will denote the law of X by $N(m, \sigma^2)$. The law $N(0,1)$ is called the standard normal or standard Gaussian law.

Equivalently we can characterize the law of a random variable via its characteristic function (why?). It is an exercise to show that if $X \sim N(m, \sigma^2)$, then

$$\mathbb{E}\exp(i\lambda X) = \exp\left(i\lambda m - \frac{\lambda^2\sigma^2}{2}\right).$$

Similarly, there are several ways to define Gaussian random vectors in $\mathbb{R}^n$.

Definition 0.2 (Gaussian random variable)**.** *A $\mathbb{R}^n$-valued random variable $\overline{X} = (X_1, \ldots, X_n)$ is called Gaussian, if for any non-zero $\overline{\lambda} = (\lambda_1, \ldots, \lambda_n) \in \mathbb{R}^n$ the inner product $\langle \overline{\lambda}, \overline{X} \rangle := \sum_{i=1}^n \lambda_i X_i$ is a Gaussian random variable.*

We denote the Gaussian law of $\mathbb{R}^n$ by $N(\overline{M}, \Sigma^2)$, where $\overline{M} = \mathbb{E}\overline{X}$ and $\Sigma^2(i,j) = \mathbb{E}X_iX_j$. If $\overline{M} = \overline{0}$ and $\Sigma^2 = \mathrm{Id}_n$, we call the vector a standard Gaussian on $\mathbb{R}^n$.

The following lemma giving 3 more equivalent ways of defining Gaussian vectors will be on the exercise sheet.

Lemma 0.3 (Equivalent definitions of a Gaussian vector)**.** *Show that $\overline{X}$ is a Gaussian vector if and only if*

1. *There exist a $n \times n$ matrix A of rank n, and a vector $\overline{M}$ such that if $\overline{Y}$ is a vector of n independent Gaussian random variables, then $\overline{X} = A\overline{Y} + \overline{M}$ almost surely.*
2. *There is a symmetric positive definite matrix D and a vector $\overline{M}$ such that that density of $\overline{X}$ w.r.t. the Lebesgue measure $d\overline{x}$ on $\mathbb{R}^n$ is given by $\frac{1}{(2\pi)^{n/2}\det(D)^{1/2}} \exp\left(-\frac{1}{2}(\overline{x} - \overline{M})^T D(\overline{x} - \overline{M})\right)$.*
3. *There is a symmetric positive definite matrix C and a vector $\overline{M}$ such that for all vectors $\overline{\lambda} \in \mathbb{R}^n$, we have that $\mathbb{E}\exp(\langle \overline{\lambda}, \overline{X} \rangle) = \exp\left(\langle \overline{\lambda}, \overline{M} \rangle + \frac{1}{2}\overline{\lambda}^T C \overline{\lambda}\right)$.*

Remark 0.4. *Sometimes it's useful to also include degenerate Gaussians, meaning the case where $\det(D) = 0$ (i.e. $\sigma = 0$ in the 1D case). In this case the Gaussian vector lives actually in a smaller dimensional subspace, in the 1D case it would be a 0-dimensional subspace, i.e. a single point. We will be pretty free about this, and include this degenerate setting when it makes things easier to state.*

[1]Please let me know if you meet any typos or unclarities!

Finally, a Gaussian process is nothing else than a collection of random variables, now indexed by any set, such that the joint law of any finite number of them is Gaussian.

Definition 0.5 (Gaussian process)**.** *Let I be any index set. Then a collection of random variables $(X_i)_{i\in I}$ defined on the same probability space is called a Gaussian process if for any finite subset $(i_1, \dots, i_n)$ we have that $\overline{X} := (X(i_1), \dots, X(i_n))$ is a Gaussian vector.*

I have left aside a fundamental, quasi-philosophical question - do Gaussians exist? They do exist in the rigorous framework of probability theory, and I let you ponder upon it on the example sheet.

0.1. **Overview.** Here is the optimistic plan for the course. Depending on your interests and time we might treat on or other example in more or less detail.

(1) *Basic properties and characterizations of Gaussians:* we will look at special-(looking) properties of the Gaussian law, and try to see in which cases they uniquely characterize this law. We will for example look at the symmetries of the Gaussian law, the stability (sum of independent Gaussians is again a Gaussian), infinite divisibility (a Gaussian can be written as a sum of n independent Gaussians for any $n \geq 1$), the maximum entropy property of Gaussians among a class of distributions, the Stein's equation (an integration by parts formula)...etc

(2) *Central limit theorem (CLT):* we will consider three proofs of CLT that explain it from slightly different angles: one by Lindeberg's swapping trick (we swap step by step the individual variables), one by using the fact that Gaussian has the maximum entropy, and one by Stein's method. We will discuss rates of convergence, extensions to dependent random variables and to higher-dimensional settings. We will try to look at some interesting examples to illustrate all this (infinite width neural networks and spin systems).

(3) *Linear nature of Gaussians in $\mathbb{R}^n$:* we will dig a bit deeper into the interplay between Gaussian measures and linear algebra. In particular we will see that in the case of Gaussians there is a bit of magic going on: marginal and conditional distributions will still be Gaussians! We will also discuss the shifting properties of Gaussians and look into absolute continuity considerations. As examples we talk a tiny bit about Gaussian linear regression, and more about Gaussian fields (height functions) stemming from natural Markov chains (called Gaussian free fields).

(4) *Suprema of Gaussian processes:* we will start by looking at the behaviour of high-dimensional Gaussians, and show that in a precise sense the standard Gaussian on $\mathbb{R}^n$ and the uniform measure on the sphere of radius $\sqrt{n}$ become basically the same as $n \to \infty$. Moreover, both of them arein fact concentrated tightly around an equator of the sphere (this is called concentration of measure). From this we deduce a surprising fact about the suprema of a Gaussian process - its fluctuations about its mean have Gaussian tails with variance given by the maximal variance of individual components! We will also see how to bound the mean of the supremum. To do this we introduce a metric structure between Gaussians and try to see the geometry of the Gaussian space. On the way we meet some of important tools in Gaussian analysis like Gaussian interpolation and Slepian lemma. As applications we plan to look at Gaussian random matrices and the Johnson-Lindenstrauss lemma.

(5) *Gaussian processes on $\mathbb{R}^n$:* We will then continue to study Gaussian processes indexed by continuous time $\mathbb{R}$, or continuous space $\mathbb{R}^n$. We will revisit Brownian motion and Ornstein-Uhlenbeck process, trying to unravel the connections to heat equation and stochastic heat equation. This will give us another look at both the Gaussian interpolation and concentration of measure. We plan to look at Gaussian process regression and Gaussian analytic functions. With a bit of luck, we also look at Gaussians on general Hilbert spaces.

(6) Finally, if time permits we will try to look into some more interesting examples like Gaussian polymer model or Random Energy model, and try to consider some statistical physics phenomena studied using Gaussian models like entropic repulsion, chaos or superconcentration.

1. Stable laws and CLTs

In this section we look at a property of Gaussians shared by a whole family of probability laws, called stability. We will see the relations between stability and Central limit theorems (CLTs), and try to understand the place of the Gaussians among these stable laws.

Definition 1.1 (Stable law)**.** *A probability law μ is said to be strictly stable, if for any $n \geq 1$ there is some $b_n > 0$ such that if we take i.i.d random variables $X_1, \dots, X_n \sim \mu$, we have that $b_n^{-1} \sum_{i=1}^n X_i \sim \mu$.*

P. Levy introduced stable laws in this form, and later they have been slightly generalized: in the general case, we ask that there are some $b_n > 0$ and $a_n \in \mathbb{R}$ with $a_n + b_n^{-1} \sum_{i=1}^n X_i \sim \mu$. We will stick to the strictly stable setting and when we are lazy we call them stable laws. Often a different characterization of stable laws is used and it's a good exercise to check its equivalence with the definition above:

Lemma 1.2. *μ is a (strictly) stable law if and only if for all $a > 0, b > 0$ there is some $c > 0$ such that if $X_1, X_2 \sim \mu$ are independent and $X \sim \mu$, then $aX_1 + bX_2 \sim cX$.*

Proof. On the exercise sheet. The "if" part is pretty direct, the "only if" part needs a bit of work... □

In this course, the first thing to verify is that we are not going to far from Gaussians:

Claim 1.3. *The Gaussian law $N(0, \sigma^2)$ is strictly stable.*

Proof. The easiest way to see this is probably to use the characteristic function: let $X_1, X_2, \dots, X_n \sim N(0, \sigma^2)$ be independent. Then we can directly calculate the characeteristic function of $S_n = \sum_{i=1}^n X_i$. Indeed, using independence and then plugging in the characeteristic function of the Gaussian, we have that

$$\mathbb{E} \exp(i\lambda S_n) = (\mathbb{E} \exp(i\lambda X_1))^n = \exp(-n\lambda^2\sigma^2/2).$$

Thus we recognize that $S_n \sim N(0, n)$ and we see that Gaussian law satisfies the condition of Definition 1.1 with $b_n = \sqrt{n}$. □

But Gaussians are certainly not the only stable laws:

Claim 1.4. *Suppose that X is a random variable with characteristic function $\mathbb{E} \exp(i\lambda X) = \exp(-C|\lambda|^\alpha)$ with $\alpha \in (0, 2]$ and $C > 0$, then X has a stable law.*

Proof. This is also on the exercise sheet, but the proof is basically the same as before. □

In fact, the family of stable laws is larger than this, but again for our purposes this gives enough context. Notice that we had a restriction $\alpha \leq 2$, although if you go through the proof we don't seem to use it. Where is the catch?

Exercise 1.1. *Prove that for $\alpha > 2$ and $C > 0$, $\mathbb{E} \exp(i\lambda X) = \exp(-C|\lambda|^\alpha)$ cannot be the characteristic function of a real-valued random variable.*

It comes out that in fact every stable law is related to a central limit theorem:

Proposition 1.5 (Stable laws and CLTs)**.** *A probability law μ is (strictly) stable if and only if we can find i.i.d random variables $X_1 X_2, \dots$ and $c_n > 0$ such that $c_n^{-1} \sum_{i=1}^n X_i$ converges in law to μ.*

Proof. One direction is easy - if μ is a stable law, then we can just take $X_i \sim \mu$ for all $i \geq 1$ and $c_n = b_n$. Then as μ is stable, we have that $b_n^{-1} \sum_{i=1}^n X_i \sim \mu$ and when we take $n \to \infty$, we still obtain μ as a limit.

In the other direction we want to show that if μ is a limit of a renormalized sum $b_n^{-1} \sum_{i=1}^n X_i$ for some i.i.d random variables X_i, then in fact μ is a stable law. Thus for every $m \in \mathbb{N}$, we want to show that if $Y_1, \dots, Y_m \sim \mu$ are independent, then there is some $b_m > 0$ such that $b_m^{-1} \sum_{i=1}^m Y_i \sim \mu$. So fix such m.

By the assumption, there are i.i.d random variables $X_1, X_2, \dots$ and $c_n > 0$ such that $c_n^{-1} \sum_{i=1}^n X_i \to \mu$ in law. Now consider the sequence only along $n = mk$ for $k \geq 1$ and decompose the sum into m pieces: for each $i = 1 \dots m$ set $S_{i,n} = \sum_{j=0}^{k-1} X_{jm+i}$. Then by the assumptions as $k \to \infty$ we have that $c_k^{-1} S_{i,n} \to Y_i$ in law, with $Y_i \sim \mu$. Moreover as all $S_{i,n}$ are independent, $Y_1, \dots, Y_m$ are independent and we have that $c_k^{-1} \sum_{i=1}^m S_{i,n} \to Y_1 + \dots + Y_m$. But also $c_n^{-1} \sum_{i=1}^m S_{i,k} \to \mu$ in law. Thus if we knew that $c_k c_n^{-1}$ converges to a limit b_m^{-1}, then we would have $b_m^{-1} \sum_{i=1}^m Y_i \sim \mu$.

It remains to argue that $a_n := c_k c_n^{-1}$ should converge to a limit as $n \to \infty$. By denoting $W_n := c_k^{-1} \sum_{i=1}^n X_i$, we can deduce this from a general claim:

Claim 1.6 (*non-examinable*). *Let $(a_n)_{n \geq 1}$ be positive real numbers. Suppose that for some random variables W_n we have that W_n converges in law to a non-trivial random variable X and $a_n W_n$ also converges in law to a non-trivial random variable Y. Then also a_n converges.*

Proof. First of all, notice that a_n has to be bounded and bounded away from zero. Indeed, if $a_{n_k} \to 0$ along some subsequence, then the characteristic function of $\mathbb{E} \exp(i\lambda a_{n_k} W_{n_k})$ would converge to 1 by Dominated Convergence Theorem, and thus the limit of $a_n W_n$ would be trivial. Similarly if $a_{n_k} \to \infty$, we would have that $a_n W_n$ does not converge in law.

Thus a_n has convergent subsequences. Consider such subsequence a_{n_k} converging to a_0. Then, again by using pointwise convergence and dominated convergence theorem, we see that $\mathbb{E} \exp(i\lambda a_{n_k} W_{n_k})$ converges to $\mathbb{E} \exp(i\lambda a_0 X)$. As on the other hand $a_{n_k} W_{n_k}$ also converges to some variable Y, we have that $\mathbb{E} \exp(i\lambda a_0 X) = \mathbb{E} \exp(i\lambda Y)$ for all $\lambda \in \mathbb{R}$. But now if there was a different subsequential limit b_0, we would obtain $\mathbb{E} \exp(i\lambda a_0^{-1} Y) = \mathbb{E} \exp(i\lambda b_0^{-1} Y)$ for all $\lambda \in \mathbb{R}$. This can be rewritten as $\mathbb{E} \exp(i\lambda Y) = \mathbb{E} \exp(i\lambda a_0 b_0^{-1} Y)$, which by iterating implies $\mathbb{E} \exp(i\lambda Y) = \mathbb{E} \exp(i\lambda (a_0/b_0)^n)$, which again converges to 1. Thus as Y is non-trivial we obtain a contradiction. Thus all convergent subsequences have a unique limit and hence the sequence a_n converges. □

□

Thus there are many laws and many Central limit theorems, each associated to one of the laws. Our Gaussian law is one of these laws, and it comes out that it's a special one:

Proposition 1.7. *The Gaussian law is the only stable law with finite variance.*

Why should such a statement be true? A first hint comes from the following observation:

Lemma 1.8. *If a strictly stable law μ has finite variance, then $b_n = n^{1/2}$.*

Proof. By the definition of stability, for all $n \in \mathbb{N}$ there is some $b_n > 0$ such that if $X_1, \dots, X_n \sim \mu$ are i.i.d, then $b_n^{-1}(X_1 + \dots + X_n) \sim \mu$. In particular the variance of two sides has to agree. But all the X_i are independent, and hence the variance of their sum is equal to the sum of their variances. Thus as all $X_i \sim \mu$, we have that

$$\mathrm{Var}(b_n^{-1}(X_1 + \dots + X_n)) = b_n^{-2} n \mathrm{Var}(\mu).$$

As this has to equal $\mathrm{Var}(\mu)$, we deduce that $b_n = n^{1/2}$. □

Given this it seems reasonable that there would be only one finite variance stable law, indeed one could hope that the set of equalities in distribution $n^{-1/2}\sum_{i=1}^n X_i \sim \mu$ with $X_i \sim \mu$ i.i.d. could have only one solution up to scaling. Whereas this is true, proving this is not quite trivial. It can be done in different ways, but maybe the most natural way is to see it as a consequence of the CLT:

Theorem 1.9 (Central limit theorem). *Let $X_1, X_2, \dots$ be i.i.d. random variables of finite variance σ^2. Then $n^{-1/2}\sum_{i=1}^n (X_i - \mathbb{E}X_i)$ converges in law to $N(0, \sigma^2)$.*

The theorem implies the proposition basically in the same way as we obtained CLTs for stable laws:

Proof of Proposition using CLT. Let μ be a strictly stable law of finite variance. Then we can take $X_1, X_2, \cdots \sim \mu$ to be i.i.d. As μ is stable and by the lemma just above $b_n = n^{1/2}$, we have that $n^{-1/2}\sum_{i=1}^n X_i \sim \mu$ for all $n \geq 1$ and in particular these laws converge to μ. But by the CLT these partial sums also converge to the Gaussian law and thus μ has to be the Gaussian law. □

Our first method of proof for the CLT is called the *Lindeberg exchange principle*, and it is rather directly based on the fact that the Gaussian law is a stable law: if we take $Y_1, Y_2, \dots$ to be i.i.d Gaussians, then as Gaussians are stable and the scaling is the correct one, the limit of $n^{-1/2}\sum_{i=1} Y_i$ is a Gaussian too. Starting now from general i.i.d. random variables X_i, our method of proof is to swap them one by one for Gaussians Y_i of the same mean and variance. By doing so an error happens at each step, but it comes out that the cumulative error over all steps is in fact negligible!

In statistical physics this type of phenomenon is sometimes called universality: the fine details of the individual constituents do not matter when looking at a macroscopic property of the system - in this case the correctly normalized sum.

This key step is encapsulated in the following proposition:

Proposition 1.10 (Lindeberg Exchange Principle). *Let $X_1, X_2, \dots$ be i.i.d. zero mean unit variance random variables, let $Y \sim N(0,1)$. Define $S_n := n^{-1/2}\sum_{i=1}^n X_i$. Then for any $\delta > 0$, there is a $n_\delta \in \mathbb{N}$ so that for all $n > n_\delta$ and for all $f : \mathbb{R} \to \mathbb{R}$ smooth with uniformly bounded derivatives up to third order, we have that*

$$|\mathbb{E}f(S_n) - \mathbb{E}f(Y)| < 2\delta(\sup_{x\in\mathbb{R}}|f'(x)| + \sup_{x\in\mathbb{R}}|f''(x)| + \sup_{x\in\mathbb{R}}|f'''(x)|).$$

Before proving the proposition, let us see how CLT follows from this proposition. In fact there are many ways to argument this. For example one could just argue that smooth functions of compact support are dense in the space of continuous functions. Let us here give a very direct argument:

Proof of the CLT:. We have that $\mathbb{P}(S_n \leq x) = \mathbb{E}(\mathbf{1}_{\{S_n \leq x\}})$. Now we can bound $\mathbf{1}_{\{S_n\leq x\}}$ from above by smooth functions f_ϵ that are equal to 1 in $(-\infty, x]$ are equal to 0 in $[x+\epsilon, \infty)$ such that the third derivative is bounded by $C\epsilon^{-3}$. Thus using the proposition we have that for n large enough

$$\mathbb{P}(S_n \leq x) \leq \mathbb{E}f_\epsilon(X) + 6C\delta\epsilon^{-3}.$$

But now $Y \sim N(0,1)$ has a bounded density and thus $\mathbb{E}f_\epsilon(Y) \leq \mathbb{P}(Y \leq x) + c\epsilon$ for some $c > 0$. Hence

$$\mathbb{P}(S_n \leq x) \leq \mathbb{P}(Y \leq x) + 6C\epsilon^{-3}\delta + c\epsilon.$$

Now, given δ we can choose $\epsilon = \delta^{1/4}$ to get

$$\mathbb{P}(S_n \leq x) \leq \mathbb{P}(Y \leq x) + \widetilde{C}\delta^{1/4}$$

for all n large enough. By taking δ arbitrarily small, we conclude that $\limsup_{n\in\mathbb{N}} \mathbb{P}(S_n \leq x) \leq \mathbb{P}(Y \leq x)$. We can similarly prove a lower bound to conclude that $\lim_{n\to\infty} \mathbb{P}(S_n \leq x) = \mathbb{P}(Y \leq x)$ and the theorem follows. □

We now prove the proposition. The exchange principle itself would come out a bit nicer if we assumed bounded third moments, but it's also nice to see that it mixes together with another technique called the truncation method. *The examinable part of the proof is the argument assuming bounded third moments*.

Proof of Lindeberg Exchange Principle: Let $Y_1, Y_2 \dots$ be i.i.d. standard Gaussians. For $k \geq 1$, write

$$S_{n,k} := \frac{\sum_{i=1}^{k-1} X_i + \sum_{i=k}^{n} Y_i}{n^{1/2}},$$

Notice that $S_{n,n+1} = S_n$ and $S_{n,1} \sim N(0,1)$. Thus we can write

$$f(S_n) - f(Y) = \sum_{k=1}^{n} f(S_{n,k+1}) - f(S_{n,k}). \tag{1.1}$$

Our aim will be to control each individual summand. To do this write further

$$S^0_{n,k} := \frac{\sum_{i=1}^{k-1} X_i + \sum_{i=k+1}^{n} Y_i}{n^{1/2}},$$

where we have omitted the k-th term altogether.

By third-order Taylor's approximation we can write a.s.

$$f(S_{n,k+1}) = f(S^0_{n,k}) + \frac{X_k}{n^{1/2}} f'(S^0_{n,k}) + \frac{X_k^2}{2n} f''(S^0_{n,k}) + \frac{X_k^3}{6n^{3/2}} f'''(x_L),$$

with x_L between $S_{n,k+1}$ and $S^0_{n,k}$ and similarly

$$f(S_{n,k}) = f(S^0_{n,k}) + \frac{Y_k}{n^{1/2}} f'(S^0_{n,k}) + \frac{Y_k^2}{2n} f''(S^0_{n,k}) + \frac{X_k^3}{6n^{3/2}} f'''(x_L),$$

We would want to take expectations now and conclude by summing up. Indeed, the two first moments would then just cancel, because of matching moments of X_k and Y_k and independence of $S^0_{n,k}$ from both Y_k and X_k (check that you know how to conclude in this case!). However, a priori we don't know that the third moment exists and thus the third order term could cause us problems.

To circumvent this problem we use a truncation method: for $\delta > 0$ small, write

$$X_k = X_k \mathbf{1}_{\{|X_k| \leq \delta n^{1/2}\}} + X_k \mathbf{1}_{\{|X_k| > \delta n^{1/2}\}}.$$

Denote the first term by $\widetilde{X}_k$ and the second term by $X_{k,>}$.

Now, we will use the exchange principle above for $\widetilde{X}_k$ instead of X_k and control the contribution of the second terms separately.

So set $\widetilde{S}_n = n^{-1/2} \sum_{k=1}^{n} \widetilde{X}_k$. Notice that $\mathbb{E}\widetilde{X}_k \neq 0$ and $\mathbb{E}\widetilde{X}_k^2 \neq 1$, so by the end of telescoping in (1.1) with $\widetilde{X}_k$ instead of X_k, we are left with

$$|\mathbb{E}f(\widetilde{S}_n) - f(Y)| \leq n^{1/2} |\sup_{x \in \mathbb{R}} f'(x)| |\mathbb{E}\widetilde{X}_k| + |\sup_{x \in \mathbb{R}} f''(x)| \mathbb{E}\widetilde{X}_k^2 + n^{-1/2} |\sup_{x \in \mathbb{R}} f'''(x)| \mathbb{E}|\widetilde{X}_k|^3. \tag{1.2}$$

Now observe that $\mathbb{E}|\widetilde{X}_k|^3 \leq \delta n^{1/2} \mathbb{E}|\widetilde{X}_k|^2 = \delta n^{1/2}$, as $\mathbb{E}\widetilde{X}_k^2 \leq \mathbb{E}X_k^2$. Moreover we claim that:

Claim 1.11. *For n large enough $|\mathbb{E}\widetilde{X}_k^2 - 1| < \delta^2$ and $|\mathbb{E}\widetilde{X}_k| = |\mathbb{E}X_{k,>}| < \delta n^{-1/2}$.*

These two things together imply that the RHS of Equation (1.2) can be bounded by

$$\delta |\sup_{x \in \mathbb{R}} f'(x)| + \delta^2 |\sup_{x \in \mathbb{R}} f''(x)| + \delta |\sup_{x \in \mathbb{R}} f'''(x)|.$$

Moreover, this claim also helps us deal with the part coming from the tails $X_{k,>}$. Denote $E_n = n^{-1/2}\sum_{k=1}^n X_{k,>}$. Then again by Taylor expansion and the claim

$$|\mathbb{E}f(S_n) - \mathbb{E}f(\widetilde{S}_n)| \leq |\sup_{x\in\mathbb{R}} f'(x)||\mathbb{E}E_n| \leq |\sup_{x\in\mathbb{R}} f'(x)|n^{-1/2}\sum_{k=1}^n |\mathbb{E}X_{k,>}| \leq |\sup_{x\in\mathbb{R}} f'(x)|\delta.$$

Thus we conclude the proposition by proving the claim:

Proof of claim: The first part just comes from the fact that $\mathbb{E}X_k^2 = 1$, the fact that $\widetilde{|}X_k| \to |X_k|$ and the monotone convergence theorem. Notice that this implies that also $\mathbb{E}|X_{k,>}|^2 < \delta^2$. For the second part, using the fact that $\mathbb{E}X_k = 0$, it suffices to show that $\mathbb{E}|X_{k,>}| < \delta n^{-1/2}$. But this follows from the previous part as

$$n^{1/2}\delta\mathbb{E}|X_{k,>}| \leq \mathbb{E}|X_{k,>}|^2 < \delta^2.$$

□

□

2. Gaussians and entropy

In this section we will discuss the concept of entropy, and see how Gaussians distinguish themselves from the entropy point of view. We will also try to revisit the Central Limit Theorem from this perspective.

Entropy is a rich concept, with origins in thermodynamics and that was (re)introduced by Shannon in his 1949 paper "A mathematical theory of communication". He gave the basic probabilistic definitions and properties of both discrete and continuous entropies.

2.1. Discrete entropy.

Consider a discrete random variable X with outcome space Ω. For each $\omega_x \in \Omega$, denote $p(\omega_x) := \mathbb{P}(X = \omega)$.

Definition 2.1 (Discrete (Shannon) entropy)**.** *For a discrete random variable X with outcome space Ω we define the entropy*

$$H(X) := -\mathbb{E}\log_2 p(X) = -\sum_{\omega_x\in\Omega} \mathbb{P}(X = \omega_x)\log_2 \mathbb{P}(X = \omega_x).$$

We also call it the entropy of the probability distribution $\mathbb{P}$.

One can directly check:

Claim 2.2. *For the uniform distribution on n points the entropy is $H = \log_2 n$.*

One often thinks of entropy in too steps:

- First, to each outcome $\omega_x \in \Omega$, we assign a measure of surprise/information content: $-\log_2 p(\omega_x)$. We like this measure more than just $p(\omega_x)$ just because we feel that the surprise of two independent events should add up.
- Entropy then gives the average surprise.

Notice that although we defined the discrete entropy for a random variable, it does not depend on the values of the random variable and it follows directly that

Lemma 2.3. *For any real-valued discrete random variable we have that $H(X) = H(aX + b)$.*

The reason of choosing the logarithm in the definition, boils down to the following facts:

Lemma 2.4. *The entropy is non-negative: $H(X) \geq 0$. Moreover, for independent discrete random variables X, Y, we have that $H(X, Y) = H(X) + H(Y)$.*

Proof. The proof is a direct computation. □

In fact, there are ways to characterise the entropy H defined above as the "natural measure" of information, when one wants the information of independent events to add up. Shannon used the following three conditions, but later this has been generalized. This characterisation is up to a multiplicative constant, which you could think of as a choice in the basis for log. In fact, we will also often just leave the base out in the notation.

Theorem 2.5 (Characterisation of Shannon entropy)**.** *For each $n \in \mathbb{N}$, and probabilities $p_1, \dots, p_n$ on a $n-$state probability space, we consider functions $\widetilde{H}_n(p_1, p_2, \dots, p_n)$ with the following properties:*

(1) For each fixed $n \in \mathbb{N}$, $\widetilde{H}$ is continuous w.r.t. $(p_1, \dots, p_n)$.
(2) The entropy of the uniform distribution is growing with n: for $n \geq m$, $H(1/n, \dots 1/n) \geq H(1/m, \dots, 1/m)$.
(3) If some outcome is decomposed into two possibilities, the entropy is a weighted sum: i.e. for any $n \in \mathbb{N}$, and for all probabilities $(p_1, \dots, p_n$, we have that

$$\widetilde{H}_{n+1}(p_1, \dots, p_n, p_{n+1}) = \widetilde{H}_n(p_1, \dots, p_n + p_{n+1}) + (p_n + p_{n+1})\widetilde{H}_2(\frac{p_n}{p_n + p_{n+1}}, \frac{p_{n+1}}{p_n + p_{n+1}}).$$

Then up to a multiplicative constant, $\widetilde{H}_n(p_1, \dots, p_n)$ agrees with the entropy $H(X)$ of a random variable X on $\{1, \dots, n\}$ that takes value i with probability p_i.

The third condition basically encapsulates the following property of entropy:

Lemma 2.6. *We have that $H(X, Y) = H(X) + H(Y|X)$, where $H(Y|X) := \sum_{\omega_x} p(\omega_x) H(Y|X = \omega_x)$ is the conditional entropy.*

Proof. Let us check this computation. We have that

$$H(X, Y) = -\sum_{\omega_x, \omega_y} p(\omega_x, \omega_y) \log_2 p(\omega_x, \omega_y).$$

Denote by $p(\omega_y|\omega_x)$ the conditional probabilities. Then $p(\omega_x, \omega_y) = p(\omega_x)p(\omega_y|\omega_x)$ and we have that

$$H(X, Y) = -\sum_{\omega_x} (p(\omega_x) \log_2 p(\omega_x)) \sum_{\omega_y} p(\omega_y|\omega_x) - \sum_{\omega_x} p(\omega_x) \sum_{\omega_y} p(\omega_y|\omega_x) \log_2 p(\omega_y|\omega_x).$$

But $\sum_{\omega_y} p(\omega_y|\omega_x) = 1$, and thus the claim follows. □

As a simple corollary we also get our first entropy inequality:

Corollary 2.7. *We have that $H(X, Y) \leq H(X) + H(Y)$ with equality iff X and Y are independent.*

Proof. We already saw that the independent case gives an inequality. So it suffices to prove a strict inequality for all other cases. From the lemma above, we have that $H(X, Y) = H(X) + H(Y|X)$. So it suffices to show that $H(Y|X) < H(Y)$ whenever X, Y are not independent. To do this first rewrite

$$-\sum_{\omega_x} p(\omega_x) \sum_{\omega_y} p(\omega_y|\omega_x) \log_2 p(\omega_y|\omega_x) = \sum_{\omega_y} \sum_{\omega_x} p(\omega_x)(-p(\omega_y|\omega_x) \log_2 p(\omega_y|\omega_x)).$$

Now, notice that $-x \log x$ is strictly concave on $[0, 1]$. Thus, as $\sum_{\omega_x} p(\omega_x) = 1$, we can bound

$$\sum_{\omega_x} p(\omega_x) \left[-p(\omega_y|\omega_x) \log_2 p(\omega_y|\omega_x)\right]$$

from above by

$$-\left[\sum_{\omega_x} p(\omega_x)p(\omega_y|\omega_x)\right] \log_2 \left[\sum_{\omega_x} p(\omega_x)p(\omega_y|\omega_x)\right] = -p(\omega_y) \log_2 p(\omega_y).$$

As the equality holds only if $p(\omega_y|\omega_x)$ is equal for all y, x, the corollary follows. $\square$

A nice way to think about of discrete entropy is in terms of minimal coding of outcomes. Suppose that we have a random variable X taking values in Ω with $|\Omega| = n$. We want to encode each possible outcome ω_x of X using a finite word W_x of 0-s and 1-s. We want to do this in a way that makes the word easy to distinguish - in particular we want no word to be a prefix to another word. Moreover, we would like to do this in a possibly efficient way, in the sense that least possible symbols are used, when transmitting the message. In particular, this means that for more frequent outcomes we want to use shorter words. A way to quantify this is to try to minimize the expected length of the random word W_X. This basically corresponds to the following optimization: we sample i.i.d copies of X, transmit each time the word corresponding to the outcome of X_i and want the average length of all the codewords sent to be as small as possible.

Theorem 2.8. *Let $M(X) = \min_W \mathbb{E}|W_X|$, where the minimum is over all sets of permitted codewords. Then $H(X) \leq M(X) \leq H(X) + 1$.*

A complete proof would take us a bit too far, but let us still sketch the relation to entropy. The starting point is the so called Kraft inequality:

Lemma 2.9 (Kraft inequality)**.** *For any set of permitted codewords $W_1, \ldots, W_n$, we have that $\sum_{i=1}^n 2^{-|W_i|} \leq 1$.*

In fact, as the proof shows the same lemma also holds for a countable selection of words.

Proof. Notice that each set of codewords is encoded by a dyadic tree, where each branching encodes the choice of 0 or 1. Thus for example the words $(0, 1), (0, 0, 1)$ would be encoded by a tree which has one branch of length 2 and one of length 3 (be sure to draw!). Notice that the tree has exactly n leaves and each corresponds to one of the n codewords.

Now think of a simple random walk on this tree starting from the root, taking each possible outgoing branch with equal probability and staying put as soon as it arrives at a leaf. The probability q_i that the walk ends on the leaf encoding W_i is at least $q_i \geq 2^{-|W_i|}$ - this is the case when at each step the walk has two possibilities. But now the walk will end up exactly at one of the leaves, and thus $1 = \sum_i q_i \geq \sum_i 2^{-|W_i|}$ as desired. $\square$

It comes out that the theorem now boils down to basically minimizing the sum $\sum_x |W_x| p(\omega_x)$ under the constraint of the Kraft inequality (notice that this is not an immediate statement, see exercise sheet!). Let us check here that minimizing without the extra constraint of $|W_x| \in \mathbb{N}$ will give us exactly entropy. Suppose for simplicity that in fact $\sum_x 2^{-|W_x|} = 1$. Denote $y_x := |W_x|$ Then the minimum of $\sum_x y_x p(\omega_x)$ can be found by using Lagrange multipliers. Indeed, consider $f(y, \lambda) = \sum_x y_x p(\omega_x) + \lambda(\sum_x 2^{-y_x} - 1)$. Then $\frac{\partial f}{\partial y_x} = p(\omega_x) - \lambda 2^{-y_x} \log 2$, $\frac{\partial f}{\partial \lambda} = \sum_x 2^{-y_x} - 1$. Thus at an extrema we have that $y_x = -\log_2 p(\omega_x)$ and one can check that this extrema is a global minimum. Plugging in the values, we see that the value at the minimum is indeed given by the entropy $-\sum_x p(\omega_x) \log_2 p(\omega_x)$.

The inequalities in the theorem can be then thought of coming from the extra constraint of $|W_x| \in \mathbb{N}$. In fact a rigorous proof of the theorem would rather just check the inequalities one by one - you can find clear statements and hints on the exercise sheet.

The other nice fact about entropy is the so-called asymptotic equipartition property: consider i.i.d. non-trivial discrete random variables $X_1, X_2, \ldots, X_n$ taking values $\omega \in \Omega$ with positive probability. Then for each sequence of outcomes $(\omega_1, \omega_2, \ldots, \omega_n)$ we can calculate the probability $\mathbb{P}(X_1 = \omega_1, \ldots, X_n = \omega_2)$. By independence, this is given by $\Pi_{i=1}^n \mathbb{P}(X_i = \omega_i)$. These probabilities decay exponentially with n, i.e. are of the form $\exp(cn)$ as each of $\mathbb{P}(X_1 = \omega) \in (0, 1)$. How does $c > 0$ behave? This will clearly depend on the exact sequence $(\omega_1, \omega_2, \ldots, \omega_n)$, so one cannot

expect an answer in full generality. However, one can determine this exponent for a typical outcome sequence and it is given by the entropy of X_1:

Theorem 2.10 (Asymptotic equipartition property)**.** *Let $X_1, X_2, \dots$ be i.i.d. non-trivial discrete random variables, taking values $\omega \in \Omega$ with positive probability. Denote by $p(\omega_1, \omega_2, \dots, \omega_n)$ the probability of the sequence of outcomes $(\omega_1, \omega_2, \dots, \omega_n)$. Then $-\frac{1}{n}\log_2 p(X_1, \dots, X_n)$ converges almost surely to $H(X_1)$.*

Proof. We can write

$$-\frac{1}{n}\log_2 p(X_1, \dots, X_n) = \frac{1}{n}\sum_{i=1}^{n} -\log_2 p(X_i).$$

But now $-\log_2 p(X_i)$ are positive i.i.d random variables with mean $H(X_i)$. Thus by the Strong Law of Large Numbers the sequence $\frac{1}{n}\sum_{i=1}^{n} -\log_2 p(X_i)$ converges to $\mathbb{E} - \log_2 p(X_1) = H(X_1)$ almost surely. □

In fact this is a special case of a much more general result describing the increase of entropy for general stationary ergodic sequences $X_1, X_2, \dots$, giving rise to the so-called Shannon-McMillan-Breiman Theorem.

2.2. **Differential entropy.** The entropy for continuous random variable with density is often called differential entropy. It was also defined by Shannon in his 1949 paper:

Definition 2.11 (Differential (Shannon) entropy)**.** *Let X be a $\mathbb{R}-$ valued random variable with density function $p(x)$. Then the differential Shannon entropy is defined by*

$$H(X) := -\int_{\mathbb{R}} [p(x)\log_2 p(x)]dx.$$

We also call it the entropy of the probability distribution with density $p(x)$.

A direct calculation gives:

Claim 2.12. *For a Gaussian random variable $X \sim N(m, \sigma^2)$, the differential entropy is given by $\frac{1}{2}\log_2(2\pi\sigma^2 e)$.*

By taking σ^2 small, we see that in particular, differential entropy doesn't need to be positive. In fact, differential entropy can be even $-\infty$ or ∞. Unlike the discrete entropy, the differential entropy has a certain inherent scale to it, coming from the fact that sums are replaced by integrals:

Lemma 2.13. *Let X be a $\mathbb{R}-$ valued random variable with density function $p(x)$. Then for any $a > 0$, we have that $H(aX + b) = H(X) + \log_2 a$.*

Proof. This is direct from a change of variables formula. □

This of course makes one wonder, whether this is at all the right definition of continuous entropy. And whereas you would find people who contest its interpretation as an entropy, it is accepted as an useful definition. To convince ourselves the origin of strange properties, let us just look into the relation to discrete entropy. Suppose that the density $p(x)$ is Riemann integrable. We can approximate it by a discrete probability distributions q_n, defined on the points $n^{-1}\mathbb{Z}$ by $q_n(y) := \int_y^{y+1/n} p(x)dx$. By the Mean Value Theorem, we can write further $q_n(y) = \frac{p(x_y)}{n}$ with $x_y \in [y, y + 1/n]$ and thus

$$H(q_n) = -\sum_{y \in n^{-1}\mathbb{Z}} \frac{p(x_y)}{n}\log_2 \frac{p(x_y)}{n} = -\sum_{y \in n^{-1}\mathbb{Z}} \frac{p(x_y)}{n}\log_2 p(x_y) + \log n.$$

Now the first term converges to the differential entropy, and the second term just diverges - in other words, we have to renormalize the discrete entropy to obtain the differential entropy.

2.3. **Maximum entropy distributions.** We now come to one of the main questions of this section: suppose we want to model some statistical phenomena using a random variable. From the experiments or theory we can deduce some weak constraints on the probability distribution of the variable - for example the support of the distribution and a some moments of the distribution, but not more. Which probability distribution should we choose as our model under these constraints? Intuitively, we would like to choose a distribution that takes into account these constraints and nothing more. Already Laplace used such an argument: his principle of insufficient reason says that if we only know that we have n outcomes, we should assign each the probability $1/n$. It comes out that somehow the right generalization of this principle of insufficient reason is the principle of maximum entropy - we should choose the probability distribution with maximal entropy, given the constraints. This feels at least intuitively natural, as we are then maximizing our surprise or uncertainty about what is happening. The principle of maximum entropy was introduced by E. T. Jaynes in the 1950s. Let us consider three examples:

Lemma 2.14 (Uniform distribution has maximum entropy)**.** *Consider all probability distributions on n points. Among such distributions, the uniform distribution is the unique maximum entropy distribution.*

Proof. Let $Q = (q_i)_{i=1\dots n}$ denote a probability distribution on n points. We want to prove that

$$H(Q) = -\sum_{i=1}^{n} q_i \log_2 q_i \leq \log_2 n.$$

This time we use the fact that $\log x$ is concave on $[0, 1]$. Thus we have that

$$H(Q) = \sum_{i=1}^{n} (q_i \log_2 \frac{n}{q_i}) - \log_2 n \leq \log_2 (\sum_{i=1}^{n} \frac{q_i n}{q_i}) = \log_2 n^2 - \log_2 n = \log_2 n,$$

where we also used that $\sum_{i=1}^{n} q_i = 1$. Moreover, the equality holds only if $q_i = 1/n$ for all i, so in fact the uniform distribution is the unique maximum entropy distribution □

Lemma 2.15 (Gaussian law has maximum entropy)**.** *Consider all probability distributions on $\mathbb{R}$ with density, with zero mean and unit variance. Among such distributions, the standard Gaussian distribution is the unique maximum entropy distribution.*

As $H(aX + b) = H(X) + \log a$, this in fact proves that an appropriate Gaussian law is the maximum entropy distribution among any distributions with fixed mean and variance.

Proof. Denote $q(x)$ the density of a probability distribution Q of zero mean and unit variance. The trick is just in the fact that $\log \exp(-x^2/2) = x^2/2$:

$$H(Q) = -\int_{\mathbb{R}} (q(x) \log_2 q(x))\, dx = \int_{\mathbb{R}} \left(q(x) \log_2 \frac{\exp(-x^2/2)}{q(x)} \right) dx - \int_{\mathbb{R}} \left(q(x) \log_2 \exp(-x^2/2) \right) dx.$$

But now, in the first term we can again use concavity of $\log x$ on $[0, \infty)$ and Jensen's inequality to obtain that $\int_{\mathbb{R}} \left(q(x) \log \frac{\exp(-x^2/2)}{q(x)} \right) dx \leq \log_2 \left(\int_{\mathbb{R}} \exp(-x^2/2) dx \right) = \log_2 \sqrt{2\pi}$. In the second term we can just use the assumption that $\int_{\mathbb{R}} x^2 q(x) dx = 1$ to obtain $-\int_{\mathbb{R}} \left(q(x) \log_2 \exp(-x^2/2) \right) dx = \frac{1}{2} \log_2 e$. Thus we see that $H(Q) \leq \frac{1}{2} \log 2\pi e = H(X)$, for $X \sim N(0, 1)$. Again, the equality in Jensen's inequality would only hold if $q(x)$ equals the Gaussian density. □

Lemma 2.16 (Uniform distribution has maximum entropy)**.** *Consider all probability distributions on $[0, \infty)$ with density and unit mean. Among such distributions, the standard exponential is the unique maximum entropy distribution.*

Proof. This is on the exercise sheet. □

2.4. **Relative entropy.** The proofs of the existence and uniqueness of maximum entropy distributions can be reinterpreted using the concept of relative entropy, also called the Kullback-Leibler distance.

Definition 2.17 (Relative entropy)**.** *If $q(\omega), r(\omega)$ are two discrete probability distributions on the same space, then we define the relative entropy or the Kullback-Leibler distance by*

$$D(q||r) := \sum_{\omega} q(\omega) \log_2 \frac{q(\omega)}{r(\omega)}.$$

If $q(x), p(x)$ are two continuous probability distributions with density on $\mathbb{R}$, then we define the relative entropy or the Kullback-Leibler distance by

$$D(q||r) := \int_{\mathbb{R}} \left(q(x) \log_2 \frac{q(x)}{r(x)} \right) dx.$$

Lemma 2.18 (Relative entropy is positive)**.** *Both in the discrete and in the continuous setting, the relative entropy satisfies $D(q||r) \geq 0$ with equality iff the densities are equal.*

Proof. Let us write the proof for the differential entropy:

$$-D(q||r) = \int_{\mathbb{R}} \left(q(x) \log_2 \frac{r(x)}{q(x)} \right) dx.$$

As in the previous 2 proofs on maximum entropy, we can now use concavity of $\log_2 x$, to get that

$$-D(q||r) \leq \log_2 \left(\int_{\mathbb{R}} r(x) dx \right) = 0.$$

□

We can now reinterpret the proofs of maximum entropy lemmas just above - in each case we guessed or identified the maximal entropy distribution $p(x)$ such that $\mathbb{E}_Q \log p(x)$ corresponds to the given constraint (e.g. variance of Q), and then used the fact that the relative entropy of Q to this maximum entropy distribution is always positive.

The two ways of thinking about entropy also help us think about relative entropy:

(1) We can think of $D(q||r)$ in terms of optimal encoding: it gives the average surplus in length, when in in transmitting outcomes of a distribution q we use the words optimized for the distribution r: this is maybe clear if we write $D(q||r) := \sum_{\omega} q(\omega) \log_2 \frac{1}{r(\omega)} - \sum_{\omega} q(\omega) \log_2 \frac{1}{q(\omega)}$.

(2) In a similar way, one can interpret relative entropy in terms of the asymptotic equipartition property and show that $\frac{1}{n} \log \frac{q(X_1,\dots,X_n)}{r(X_1,\dots,X_n)}$ converges to $D(q||r)$ almost surely.

Relative entropy also gives a certain distance between two probability measures, even though it is not a proper metric (e.g. it is not symmetric):

Proposition 2.19. *Let Q, R be two random variables with densities r, q. Then $D(q||r) \geq \frac{1}{2 \ln 2} \|q - r\|_1^2$.*

2.5. **Entropy and CLT.** Our next aim is to work towards a CLT using entropy. As we will see the proof heuristic is most pleasing, yet there are some non-trivial steps and many small technicalities to pave the way. In most of the section we only deal with random variables with density, such that their entropy is finite.

2.5.1. *The heuristic proof.* Given i.i.d. zero mean random variables $X_1, X_2, \dots$ of finite variance, we would like to prove the central limit theorem via the following heuristic:

(1) The entropy is monotonously increasing for the sequence $S_n := \frac{\sum_{i=1}^n X_i}{\sqrt{n}}$;

(2) it increases to the maximal possible value, given the variance, which we have shown to be a Gaussian.

Now, this plan of attack is very intuitive, but in fact both steps are non-trivial. First, it is very intuitive that entropy should increase when we sum i.i.d random variables. And indeed, it is simple to see that:

Lemma 2.20 (Monotonicity of entropy (baby version)). *Let X, Y be independent random variables with density and finite entropy. Then $H(X+Y) \geq \max(H(X), H(Y))$.*

Proof. The discrete version is on the exercise sheet, the continuum version works in the same way. □

However, notice that to show monotonicity of S_n we need to show more. For, example already to show that $H(S_2) \geq H(S_1)$ we need to argue that $H(\frac{X+Y}{\sqrt{2}}) \geq H(X)$, which by the scaling property of differential entropy comes down to showing $H(X+Y) \geq H(\sqrt{2}X)$, i.e. one needs to show that summing up two independent random variables increases entropy more than just increasing the variance of the single random variable. This was shown in the end of 1950s using the following inequality:

Theorem 2.21 (Entropy Power Inequality). *Let X, Y be two independent continuous random variables with density. Then*

$$2^{2H(X+Y)} \geq 2^{2H(X)} + 2^{2H(Y)}.$$

Moreover, the equality holds if and only if X, Y are independent Gaussians.

It is not hard to see that this gives monotonicity of entropy for the sequence S_n along powers of two. The monotonicity of the whole entropy had to surprisingly wait until 2004. Whereas they prove a more general version where the random variables are not necessarily identically distributed, we will state the version relevant to us:

Theorem 2.22 (Monotonicity of entropy - Arstein, Ball, Barthe, Naor 2004). *Let $X_1, X_2, \ldots$ be i.i.d. random variables with density and finite variance. Then $H(S_n)$ is increasing with n.*

Theorem 2.21 and Theorem 2.22 together imply the two claims above for i.i.d. sequence of random variables with density, of zero mean, unit variance and finite entropy:

(1) By Theorem 2.22 the sequence $H(S_n)$ is increasing, and as $H(S_n) \leq H(X)$ where $X \sim N(0,1)$, we see that $H(S_n)$ converges to a finite value H.
(2) Now, supposing that S_n converges to some Y, we can decompose again the sum into two parts: odd terms to form S_n^o and of even terms to form S_n^e. We then have the joint convergence of (S_n, S_n^o, S_n^e) to $(Y, \frac{Y_o}{\sqrt{2}}, \frac{Y_e}{\sqrt{2}})$ where Y, Y_e, Y_o have the same law, Y_e, Y_o are independent and $\frac{Y_e+Y_o}{\sqrt{2}} \sim Y$. Now, by the previous point the entropies of S_n, S_n^o, S_n^e all converge to the same constant H. By dominated convergence, one can also argue that they converge to the entropies of Y, Y_o, Y_e respectively. Thus one can verify that equality holds in the Entropy Power Inequality and hence Y has to be a Gaussian. But we know that it has unit variance, and hence it is a standard Gaussian.

To show the CLT, we would still also need to show the actual convergence in law of the random variables S_n. This would for example follow from Proposition 2.19, but could be also done by standard means. We will come back to this, when we give a rigorous proof. Notice, that this would still prove the CLT only for random variables with density and finite entropy. There is a little extra trick ndeeded to get from random variables with density and finite entropy. This is a bit similar to the Lindeberg proof, where we needed a truncation argument in addition to the main step. Here, we need a certain smoothing argument to deal with the general case: namely instead of S_n we consider variables $\widehat{S}_n = S_n + Y_n$, where Y_n are i.i.d standard Gaussians independent of $X_1, X_2, \ldots$.

Lemma 2.23. *Let X be any $\mathbb{R}$-valued random variable of finite variance and Y a standard Gaussian. Then the variable $X+Y$ has density and finite entropy.*

Proof. The density of $X+Y$ is explicit: indeed, suppose X induces a (possibly atomic) probability measure $\mu(dx)$ on $\mathbb{R}$. Then by Fubini's theorem one can check that the density of $X+Y$ can be written as: $q(z) = \int_{\mathbb{R}} \frac{\exp(-\frac{(z-y)^2}{2})}{\sqrt{2\pi}} d\mu(y)$. As soon as we have the density, as $X+Y$ has finite variance, we know that the entropy is bounded from above. To bound it from below we recall that $H(X+Y) \geq \max(H(X), H(Y))$ for any random variables X, Y with density. Now, a priori X has no density, but we can decompose $Y = Y_1 + Y_2$ with Y_1, Y_2 i.i.d $N(0, 2^{-1/2})$ and apply the lemma to $X + Y_1$ and Y_2. □

Given this lemma, one can the CLT the for general i.i.d. random variables with unit variance:

Lemma 2.24. *Let $X_1, X_2, \ldots$ be random variables of zero mean and unit variance. Let $Y_1, Y_2, \ldots$ be i.i.d standard Gaussians, independent from $X_1, X_2, \ldots$ and define $\widehat{X}_n = X_n + Y_n$. Suppose that the sequence of variables $\widehat{X}_n$ converges in law to a centred Gaussian of variance 2. Then X_n converges to a standard Gaussian.*

Proof. This will be on the exercise sheet. □

2.5.2. *A rigorous proof of CLT using Fisher information.* Now we will present a rigorous information-theoretic proof of the CLT, following basically the same strategy as above, with two small differences. The main difference to the sketch above is that we will use another information-theoretic concept, called Fisher information, instead of entropy. A smaller difference is, that we will argue convergence of Fisher information using subadditivity instead of monotonicity.

We work with Fisher information, because it is easier to analytically manipulate. In fact, both Theorems 2.21 and Theorem 2.22 are also probably easiest to prove using Fisher information:

Definition 2.25 (Fisher information)**.** *Let X be a $\mathbb{R}$-valued random variable with positive differentiable density $p(x)$. Then the score function is defined by $\rho(x) = \frac{p'(x)}{p(x)}$ and the Fisher information $J(X) := \mathbb{E}\rho^2(X)$.*

Fisher information originates from statistics, where it basically encodes the information that samples contain about parameters of the underlying probability distribution. This connection is not so clear from the definition above, but will be clearer on the example sheet.

The following simple properties of Fisher information just follow from a change of variable:

Claim 2.26. *The Fisher information is translation invariant: $J(X+a) = J(X)$, and scales as $J(aX) = a^{-2}J(X)$.*

Our favourite, the Gaussian comes forward also in terms of Fisher information - it minimizes Fisher information, given the variance:

Lemma 2.27. *Let X be any random variable with positive differentiable density $p(x)$, unit variance and finite Fisher information. Then $J(X) \geq 1$. The equality holds iff $X \sim N(0,1)$.*

Proof. By the claim above it suffices to consider the zero mean case. By Cauchy-Schwarz we have that $(\mathbb{E}\rho(X)X)^2 \leq \mathbb{E}\rho^2(X) S \mathbb{E}X^2 = J(X)$. This also implies in particular that $\mathbb{E}|\rho(X)X| = \int_{\mathbb{R}} |p'(x)x|$ is finite. Now, by integration by parts

$$\mathbb{E}\rho(X)X\mathbf{1}_{\{X\in[-a,a]\}} = \int_{-a}^{a} p'(x)x dx = p(a)a + p(-a)a - \int_{-a}^{a} p(x)dx.$$

By dominated convergence both $\int_{-a}^{a} p'(x)xdx \to \int_{\mathbb{R}} p'(x)xdx$ and $\int_{-a}^{a} p(x)dx \to \int_{\mathbb{R}} p(x)dx$ as $a \to \infty$, we see that $p(a)a + p(-a)a$ also converges. But as $(\mathbb{E}|X|)^2 \leq \mathbb{E}X^2$, the only possible limit is 0. Hence

$$\mathbb{E}\rho(X)X = \int_{\mathbb{R}} p'(x)xdx = -\int_{\mathbb{R}} p(x)dx = -1$$

and it follows that $J(X) \geq 1$. The equality holds when we have equality in Cauchy-Schwarz, i.e. when $\rho(x) = cx$ for some constant c. Thus we see that $\frac{d\log p(x)}{dx} = cx$ implying that $p(x) = a\exp(-c\frac{x^2}{2})$ and the condition on variance fixes $c = 1$. □

There is a beautiful identity connecting Fisher information to entropy. In fact, it is this identity that helps to transfer several properties of Fisher information to entropy:

Theorem 2.28 (De Bruijn's identity)**.** *Let X be a random variable with finite variance and twice differentiable density and let Y be a standard Gaussian. Denote $Z_t = X + \sqrt{t}Y$. Then we have that $\frac{d}{dt}H(Z_t) = \frac{\log_2 e}{2}J(Z_t)$.*

The proof of this surprising identity itself is teachnical, but not very difficult, and will be on the example sheet. It stems from the connection between convolutions with Gaussians to the heat equation. In fact, this connection to the heat equation is also behind the smoothing we saw above, and it will enter several times during the course. But let us get back to proving the CLT using information-theoretic methods. Our key step, replacing Entropy Power Inequality, is the following statement on the decrease of Fisher information:

Proposition 2.29 (Decrease of Fisher information)**.** *Let X, Y be independent with continuously differentiable densities. Then for any $\beta \in [0,1]$ we have that $J(X+Y) \leq \beta^2 J(X) + (1-\beta)^2 J(Y)$. If the equality holds for some $\beta \in (0,1)$, and either X or Y have a positive density, then X, Y are Gaussian.*

We will prove this proposition, but let us first discuss some consequences. By taking $\beta = 1/2$ and $X \sim Y$, we obtain that $J(X+Y) \leq \frac{1}{2}J(X)$, which using scaling of Fisher information gives $J(\frac{X+Y}{2}) \leq J(X)$, hinting at a nice decrease of Fisher information for the sequence $S_n := \frac{\sum_{i=1}^n X_i}{\sqrt{n}}$. In fact, obtaining monotonous decrease along the sequence of n is again hard, and via de Bruijn's identity would prove the Entropy Monotonicity Theorem. However, to prove the CLT a weaker result will suffice:

Corollary 2.30. *For all $m, n \in \mathbb{N}$ it holds that $(m+n)J(S_{m+n}) \leq mJ(S_m) + nJ(S_n)$.*

Proof. By taking $X = \sqrt{\frac{m}{m+n}}S_m$, $Y = \sqrt{\frac{m}{m+n}}S_n$ and setting $\beta = \frac{m}{m+n}$, we obtain from the proposition that $J(S_{m+n}) \leq \frac{m^2}{(m+n)^2}J(\sqrt{\frac{m}{m+n}}S_m) + \frac{n^2}{(m+n)^2}J(\sqrt{\frac{m}{m+n}}S_m)$. Using scaling of the Fisher information now gives the result □

Let us argue that this corollary implies convergence of the Fisher information. The key input here is the following lemma, whose proof is not examinable, although not difficult:

Lemma 2.31 (Fekete Subadditivity Lemma)**.** *Let $(x_n)_{n\geq 1}$ be a sequence of positive numbers. Suppose that x_n is subadditive: i.e. $x_n + x_m \leq x_{m+n}$. Then $\frac{x_n}{n}$ converges as $n \to \infty$.*

Now, by Corollary 2.30, we have that $x_n := nJ_n$ is sub-additive, and in particular by Fekete Subadditivy Lemma, we see that $x_n/n = J_n$ converges. We are now ready to state and prove the information-theoretic CLT. Again we state it first for sequences of random variables that have densities and finite Fisher informations. In fact we ask even more, however, appropriate versions of Lemmas 2.45 and will help to extend to the standard version. We will come back this final step after finishing the proof under the technical constraints.

Theorem 2.32 (Information-theoretic CLT). *Let $X_1, X_2, \ldots$ be i.i.d. random variables with zero mean and unit variance, continuously differentiable density on $\mathbb{R}$ and finite Fisher information. Suppose further that the densities of S_n are bounded away from zero uniformly on any compact interval and have uniformly bounded derivatives, and that the score functions $\rho_{S_n}(S_n)$ satisfies $\mathbb{E}\left(\rho_{S_n}^4(S_n)\right) \leq C$ for some $C > 0$. Then S_n converge in law to the standard Gaussian.*

To prove this CLT, we want to use a very versatile two-step strategy for proving convergence in law:

- First, one shows that any subsequence of S_n has a further subsequence that converges;
- and second, one shows that all possible subsequential limits have the same law.

These two steps together prove that the limit in law exists - indeed, suppose μ is the limiting law from the first step, but that S_n does not converge to μ. Then one could pick a subsequence that stays at some positive distance ϵ from μ (in some metric that metrizes the convergence in law, like the Levy-Prokhorov metric). But then by the first step, there is a further subsequence that converges, and by the second step this limit is again μ. Thus the sequence does get closer than ϵ and we have a contradiction.

This strategy is often easier to use than say the method of characteristic function or proving the limits of the cumulative density function, as we never have to show any precise quantitative convergence. Moreover, sometimes the first point comes almost for free (for example if all laws are supported in a fixed interval), and often some simple property can identify the limit in law - especially when the limit is expected to be Gaussian.

2.6. **Continuation: Entropy and CLT.** Let us start with a little detour on proving convergence in law:

2.6.1. *A small detour: How to prove convergence in law?* A real-valued random variable X gives rise to probability law $\mathbb{P}$ on $\mathbb{R}$, i.e. a (countably additive) positive measure of total mass 1. It can be uniquely described in multiple ways:

(1) Using the cumulative density function $F(x) := \mathbb{P}(X \leq x)$
(2) Using the characteristic function $c(\lambda) := \mathbb{E}\left(\exp(i\lambda X)\right)$.
(3) By determining $\mathbb{E}f(X)$ for all continuous bounded functions f.
(4) By determining $\mathbb{P}(X \in A)$ for all open sets.

This list is not exhaustive and each description gives rise to a way of showing convergence in law of $X_n \to X$:

(1) Cumulative density function: it suffices to show the convergence of $F_n(x)$ for all continuity points of $F(x)$.
(2) Characteristic function: it suffices to show pointwise convergence of $c_n(\lambda)$ to $c(\lambda)$.
(3) Continuous functions: it suffices to show convergence of $\mathbb{E}_n f(X)$ to $\mathbb{E}f(X)$ for all continuous bounded f.
(4) Open sets: it suffices to show that $\liminf_n \mathbb{P}_n(X \in A) \geq P(X \in A)$.

In all these cases one at once shows convergence and identifies the limit. Sometimes it is, however, useful to treat convergence in two separate steps:

- First, one shows sequential compactness: any subsequence of X_n has a further subsequence that converges in law;
- and second, one shows that all possible subsequential limits have the same law, by characterizing the limiting random variable in some way.

Notice that here, in the second step we can assume that we have convergence in law and we just want to say that the limiting law has some specific properties. This is very useful if there is some property that can easily characterize the law - we will return to more concrete criteria below.

To see why this strategy works, it is useful to start from a similar lemma for real numbers:

Lemma 2.33. *Suppose that $(x_n)_{n\geq 1}$ is a sequence of real numbers that is sequentially compact and such that each subsequential limit equals to L. Then $x_n \to L$.*

Proof. Suppose not. Then there is some $\epsilon > 0$, some subsequence $(x_{n_k})_{k\geq 1}$ with $|x_{n_k} - L| \geq \epsilon$. But this subsequence also has to have a convergent subsequence $(x_{n_{k_j}})$, which has to converge to L, giving a contradiction. □

A similar lemma applies to probability laws, as soon as they are defined on a metric space, like $\mathbb{R}^n$.

Lemma 2.34. *Let X_n be a sequence of random variables taking values in $\mathbb{R}^n$, such that the corresponding probability laws are sequentially compact w.r.t. convergence in law, and such that each subsequential limit of X_n is equal in law to some r.v. X. Then $X_n \to X$ in law.*

Proof. Suppose X_n do not converge to X in law. Then there is some $\epsilon > 0$, some continuous bounded f and some subsequence $(n_k)_{k\geq 1}$ such that $|\mathbb{E}f(X_{n_k}) - \mathbb{E}f(X)| \geq \epsilon$. But now, by assumption $(n_k)_{k\geq 1}$ has a further subsequence n_{k_j} that converges to some $\widetilde{X}$ in law. But then by the second assumption $\widetilde{X} = X$ in law, and thus $\mathbb{E}f(X_{n_{k_j}}) \to \mathbb{E}f(X)$, giving a contradiction. □

Remark 2.35. *In fact, the space of probability laws can be given a nice metric, called the Levy-Prokhorov metric. Let us here confine ourselves to $\mathbb{R}^n$, though the actual generality is that of metric spaces:*

$$d_{LP}(\mathbb{P}, \mathbb{Q}) = \inf_{\epsilon > 0}\{\mathbb{P}(B_\epsilon) + \epsilon \geq \mathbb{Q}(B) \text{ and } \mathbb{Q}(B_\epsilon) + \epsilon \geq \mathbb{P}(B) \ \forall \text{ Borel sets } B\},$$

where $B_\epsilon = \{x \in \mathbb{R}^n : d(x, B) \leq \epsilon\}$. Given this, we can state and prove the above lemma exactly in the same way as for real numbers.

The identification of the law differs case by case - for example, it could be that one can show that the limiting law has maximal possible entropy under some conditions, or that it has finite variance and is stable...in both cases we would know that the limit is a Gaussian.

To show subsequential convergence, there is however one very useful general theorem. First a relevant definition:

Definition 2.36 (Tightness)**.** *The sequence of random variables $X_1, X_2, \ldots$ is called tight if for all $\epsilon > 0$, there is some compact set K_ϵ such that $\mathbb{P}(X_n \notin K_\epsilon) \leq \epsilon$ for all $n \in \mathbb{N}$.*

In $\mathbb{R}$ these compact set could be just bounded intervals like $[-K, K]$ and a possible criteria for tightness is for example that for some $b > 0$, we have that $\mathbb{E}X_n^b < C$ for all $n \in \mathbb{N}$: by Markov's inequality $\mathbb{P}(X_n \geq \epsilon^{-1/b}) \leq C\epsilon$. The general theorem saying that sequential compactness follows from tightness is then as follows:

Theorem 2.37 (Prokhorov's theorem)**.** *A tight sequence of random variables has subsequential limits in law, i.e. the corresponding probability laws are sequentially compact w.r.t. the metric d_{LP}.*

We will not prove this theorem here, but let us still say why it should be true:

- A key topological input is that on a compact metric space, the space of probability measures with the distance d_{LP} is compact (this basically follows from Banach-Alaoglu theorem).
- In this light, tightness basically says that with arbitrarily high probability all the random variables stay in some compact set and Prokhorov theorem becomes very believable: up to as small error as we wish, the random variables stay in a compact set and thus have subsequential limits.

Often one wants to complement convergence in law $X_n \to X$ with convergence of the expectation, i.e. one wants $\mathbb{E}X_n \to \mathbb{E}X$. To see that this is not automatic, consider the following example:

Example 2.38. *Let X_n be random variables that take value 0 with probability $1-1/n$ and value n with probability $1/n$. Then $X_n \to 0$ in law, however $\mathbb{E}X_n = 1$ and $\mathbb{E}X = 0$ - some mass has escaped to infinity.*

It is not hard to see that for uniformly bounded random variables, no such thing can happen because we have Fatou's Lemma and Reverse Fatou's Lemma at hand: by Fatou's lemma $\liminf_n \mathbb{E}X_n \geq \mathbb{E}\liminf_n X_n = \mathbb{E}X$, and by reverse Fatou $\limsup_n \mathbb{E}X_n \leq \mathbb{E}\limsup_n X_n = \mathbb{E}X$, so we conclude that $\liminf_n \mathbb{E}X_n = \limsup_n \mathbb{E}X_n = \mathbb{E}X$.

The extra condition preventing the escape of mass is called uniform integrability:

Definition 2.39 (Uniform integrability)**.** *A sequence of $\mathbb{R}$-random variables X_n is called uniformly integrable, if for all $\epsilon > 0$ there is some $K > 0$ such that for all $n \geq 1$, we have that $\mathbb{E}\left(|X_n|\mathbf{1}_{\{|X_n|>K\}}\right) < \epsilon$.*

The key theorem is then as follows and is not difficult to prove:

Theorem 2.40. *Suppose that $\mathbb{R}$-valued random variables X_n converge in law to X. Suppose moreover that X_n are uniformly integrable. Then $\mathbb{E}X_n \to \mathbb{E}X$.*

Uniform integrability is often guaranteed by the existence of some higher moment: for example if $\mathbb{E}|X_n|^{1+\delta} < C$ for all $n \in \mathbb{N}$, then $\mathbb{E}|X_n|\mathbf{1}_{\{|X_n|>K\}} \leq \mathbb{E}\left(|X_n|(\frac{|X_n|}{K})^{\delta}\right) \leq CK^{-\delta}$, than can be made arbitrarily small.

Let us now return to the central story and prove the theorem.

2.6.2. *Continuation: a rigorous proof of CLT using Fisher information.* Recall the central ingredient:

Proposition 2.41 (Decrease of Fisher information)**.** *Let X, Y be independent with continuously differentiable densities. Then for any $\beta \in [0,1]$ we have that $J(X+Y) \leq \beta^2 J(X) + (1-\beta)^2 J(Y)$. If the equality holds for some $\beta \in (0,1)$, and either X or Y have a positive density, then X, Y are Gaussian.*

The reason that makes working with Fisher information easier is the possibility to work with L^2 techniques. Indeed a key input to the proof of the proposition is the following projection lemma for the score function:

Lemma 2.42 (Projection of the score function)**.** *Let X, Y be two independent random variables with differentiable density. Suppose that X has finite Fisher information, then $\rho_{X+Y}(X+Y) = \mathbb{E}\left[\rho_X(X)|X+Y\right]$. Similarly for Y.*

Proof. Notice that $\mathbb{E}\left(\rho_X(X)\mathbf{1}_{\{X\in[a,b]\}}\right) = p_X(b) - p_X(a)$. So it suffices to show that for all $a < b$

$$\mathbb{E}\left(\mathbf{1}_{\{X+Y\in[a,b]\}}\mathbb{E}\left[\rho_X(X)|X+Y\right]\right) = p_{X+Y}(b) - p_{X+Y}(a).$$

So let us calculate

$$\mathbb{E}\left(\mathbf{1}_{\{X+Y\in[a,b]\}}\mathbb{E}\left[\rho_X(X)|X+Y\right]\right) = \mathbb{E}\left(\mathbb{E}\left[\mathbf{1}_{\{X+Y\in[a,b]\}}\rho_X(X)|X+Y\right]\right) = \mathbb{E}\left[\mathbf{1}_{\{X+Y\in[a,b]\}}\rho_X(X)\right].$$

Now conditioning instead on Y, we have

$$\mathbb{E}\left[\mathbf{1}_{\{X+Y\in[a,b]\}}\rho_X(X)\right] = \mathbb{E}\left(\mathbb{E}\left[\mathbf{1}_{\{X+Y\in[a,b]\}}\rho_X(X)|Y\right]\right).$$

Using independence of X, Y, we can write this as

$$\mathbb{E}\left(p_X(b-Y)\right) - \mathbb{E}\left(p_X(a-Y)\right).$$

It remains to notice that $p_{X+Y}(b) = \mathbb{E}\left(p_X(b-Y)\right)$. □

Proof of Proposition 2.41: The first part of the proposition follows quite directly from the lemma. Indeed, from the Lemma we can write:

$$J(X+Y) = \mathbb{E}\left(\rho^2_{X+Y}(X+Y)\right) = \mathbb{E}\left([\mathbb{E}(\beta\rho_X(X) + (1-\beta)\rho_Y(Y)|X+Y)]^2\right).$$

By Jensen's inequality we have that

$$\mathbb{E}\left([\mathbb{E}(\beta\rho_X(X) + (1-\beta)\rho_Y(Y)|X+Y)]^2\right) \leq \mathbb{E}\left(\mathbb{E}([\beta\rho_X(X) + (1-\beta)\rho_Y(Y)]^2\,|X+Y)\right),$$

which we can write as

$$\mathbb{E}\left([\beta\rho_X(X) + (1-\beta)\rho_Y(Y)]^2\right).$$

By independence $\mathbb{E}\rho_X(X)\rho_Y(Y) = \mathbb{E}\rho_X(X)\mathbb{E}\rho_Y(Y)$ and by integration by parts $\mathbb{E}\rho_X(X) = 0$, hence indeed $J(X+Y) \leq \beta^2 J(X) + (1-\beta)^2 J(Y)$.

It remains to identify the case of equality. The equality in Jensen's inequality holds if and only if for $X+Y$ a.e. z, and X a.e. x we have that

$$\rho_{X+Y}(z) = \beta\rho_X(x) + (1-\beta)\rho_Y(z-x).$$

Now, we know that the density is finite and positive, so it holds for Lebesgue a.e. z, x. Plugging in $x = z$ and then $x = 0$, we obtain that

$$\rho_{X+Y}(x) + \rho_{X+Y}(z-x) - \beta\rho_X(0) - (1-\beta)\rho_Y(0) = \rho_{X+Y}(z).$$

Defining $f(z) = \rho_{X+Y}(z) - \beta\rho_X(0) - (1-\beta)\rho_Y(0)$, we obtain that for almost all z, x

$$f(x) + f(z-x) = f(z).$$

But now f is continuous and thus this equation holds iff f is linear, and we already saw that this implies Gaussianity. Thus $X, Y, X+Y$ are all Gaussian. □

Recall also its key conclusion: using this proposition and Fekete's subadditivity lemma, we concluded that

- the Fisher informations $J(S_n)$ converges as $n \to \infty$.

Now, let us now type the proof of the CLT. First, let us write again the statement of the restricted CLT:

Theorem 2.43 (Information-theoretic CLT)**.** *Let $X_1, X_2, \ldots$ be i.i.d. random variables with zero mean and unit variance. Suppose further that the densities of S_n are bounded away from zero uniformly on any compact interval, are differentiable and have uniformly bounded derivatives. Suppose also that the score functions $\rho_{S_n}(S_n)$ satisfy $\mathbb{E}\left(\rho^4_{S_n}(S_n)\right) \leq C$ for some $C > 0$. Then S_n converge in law to the standard Gaussian.*

There is an annoying part in the proof of this theorem, and a nice part. The annoying part is arguing compactness and making sure that everything is well-defined for any subsequential limit. Indeed, it is a priori not clear why J should be the Fisher information of a subsequential limit of S_n, or indeed why a subsequential limit should have a well-defined Fisher information, or even a density? This is where we use the technical assumptions in the statement. The nice part is then just applying the convergence of the Fisher information, and the equality case of the proposition above to conclude that any limit is a Gaussian.

Proof of CLT.

Technical preliminaries:
First, as S_n all have zero mean and unit variance, we see that the sequence S_n is tight by applying

Markov's inequality. Thus subsequential limits of S_n exist. Moreover, we saw that $J(S_m)$ converges, and we call this limit J.

Further, by our assumption the densities p_n of S_n have uniformly bounded first derivatives. Thus in particular by Arzela-Ascoli theorem the densities have differentiable subsequential limits p on any compact interval of $\mathbb{R}$ in the uniform topology. Moreover p_n' also then converge to p' uniformly on compacts of $\mathbb{R}$. By a diagonal argument we get subsequential differentiable limits over the whole $\mathbb{R}$. As S_n have bounded second moments, the density functions are moreover uniformly integrable, and thus $1 = \mathbb{E}p_n(X) \to \mathbb{E}p(X)$ and hence $p(x)$ is a differentiable density of a probability distribution. By assumption we also know that the densities remain positive over the $\mathbb{R}$.

In particular, we can define the Fisher information for any subsequential limits and we see that both $\rho_{S_n}(x)$ and $\rho^2_{S_n}(x)$ also converge pointwise along these subsequences. As by assumption $\rho^2_{S_n}(S_n)$ are also uniformly integrable, we know that when $\rho^2_{S_n}(S_n)$ converges to $\rho_Y(Y)$, then also $J_n = \mathbb{E}\rho^2_{S_n}(S_n)$ converges to $\mathbb{E}\rho^2_Y(Y)$. Thus for any simultaneous subsequential limit of S_n in law and p_n, we get that the limit of the Fisher informations agrees with the Fisher information of the subsequential limit of S_n.

Conclusion:
We can now finish the proof of the theorem. For n even, consider the quadruple (S_n, S_n^o, S_n^e, p_n), where S_n^o is the partial sum containing all odd terms in S_n and S_n^e all the even terms. Notice that S_n^o and S_n^e are independent, S_n^e and S_n^o have the law of $\frac{S_{n/2}}{\sqrt{2}}$ and $S_n^o + S_n^e = S_n$. In particular the densities of S_n^o and S_n^e also converge. As each component of the vector is tight, the vector itself is tight.

Let us argue that all its limits have to have the same law. We already saw that the subsequential limit will be of the form (Y, Y_o, Y_e, p). Moreover, we still have that Y_o, Y_e are independent and $Y_o + Y_e = Y$. Now, we argued above that $J(Y) = J$ and by scaling $J(Y_e) = J(Y_o) = 2J$. Thus $\frac{J(Y_o)+J(Y_e)}{4} = J(Y)$ and we conclude from Proposition 2.41 that they are all Gaussians. We can now conclude from Lemma 2.34 □

Although we needed to justify many convergences, this remains a rather soft proof - the only specific information-theoretic input was the inequality on the Fisher information, together with the case of equality; all the rest were standard probabilistic arguments. Using entropy, one could imagine a proof that does not pass through the two-step strategy, but rather proves directly that the Kullback-Leibler divergence between S_n and the standard Gaussian goes to zero. AS we noted, this would actually give convergence of the density functions in the L^1 norm. However, this proof would certainly need some extra calculations.

Let us finish this section by writing out the two lemmas that help to extend the convergence to general i.i.d. random variables with zero mean and unit variance. One of them is exactly the same as before:

Lemma 2.44. *Let $X_1, X_2, \ldots$ be random variables of zero mean and unit variance. Let $Y_1, Y_2, \ldots$ be i.i.d standard Gaussians, independent from $X_1, X_2, \ldots$ and define $\widehat{X}_n = X_n + Y_n$. Suppose that the sequence of variables $\widehat{X}_n$ converges in law to a centred Gaussian of variance 2. Then X_n converges to a standard Gaussian.*

Proof. This will be on the exercise sheet. □

The other lemma needs a small modification:

Lemma 2.45. *Let X be any $\mathbb{R}$-valued random variable of finite variance and Y a standard Gaussian. Then the variable $X+Y$ has continuously differentiable density that is uniformly bounded away from*

zero on any compact interval, has an uniformly bounded derivative, finite Fisher information and moreover its score-function has a uniformly bounded $4-th$ *moment.*

Proof. This will also be on the exercise sheet. □

3. LLN and CLT for dependent random variables

[2]

We have seen the CLT for i.i.d. random variables with finite variance. We also saw that if we leave out the finite variance condition, then central limit theorems still exist, but the limiting distributions are given by other stable laws, like for example the Cauchy law.

It is also possible to generalize the CLT by keeping the variance, keeping the independence, but allowing the random variables to have non-identical distributions. This is a relatively mild generalization, but you already need some extra conditions to make it work.

In these notes we deal with maybe the most exciting generalization, which is that of dependent random variables. Let us start, however, by looking at the laws of large numbers (LLN) in the dependent case. Recall, that

- The Weak LLN says that for i.i.d X_1 with $\mathbb{E}|X_1| < C$, we have that $n^{-1}\sum_{i=1}^n X_i \to \mathbb{E}X_1$ in probability.
- The Strong LLN says that for i.i.d X_1 with $\mathbb{E}|X_1| < C$, we have that $n^{-1}\sum_{i=1}^n X_i \to \mathbb{E}X_1$ almost surely.

In general, LLNs might not hold in the case of dependent random variables. For example, just take $X_1 = X_2 = \cdots = X_n$. In this case $n^{-1}\sum X_i = X_1 \neq \mathbb{E}X_1$ in general. Still, in the case of LLN, even a small decorrelation will do:

Proposition 3.1 (WLLN for dependent random variables)**.** *Let* X_i *be mean zero random variables with* $\mathbb{E}X_i^2 = 1$ *and* $\mathbb{E}X_iX_j \leq c|i-j|^{-a}$ *for all* $i \neq j$*. Then* $n^{-1}\sum_{i=1}^n X_i$ *converges to zero in probability.*

Remark 3.2. *In fact SLLN also holds in the same setting, a slightly weaker version of the SLLN is left as an exercise.*

Proof. As $\mathbb{E}n^{-1}\sum_{i=1}^n X_i = 0$, it suffices to prove that $n^{-2}\mathbb{E}\left([\sum_{i=1}^n X_i]^2\right) \to 0$. So let us calculate

$$\mathbb{E}\left([\sum_{i=1}^n X_i]^2\right) = 2\sum_{i>j}\mathbb{E}X_iX_j + \sum_i \mathbb{E}X_i^2.$$

The diagonal term is just equal to $n = o(n^2)$. To treat the first term, fix some i. Then

$$\sum_{j>i}^n \mathbb{E}X_iX_j \leq c\sum_{j>1}^n |i-j|^{-a} \leq c'(1+|n-i|^{1-a}) \leq c'(1+n^{1-a}).$$

Summing over all $i = 1 \ldots n$, we get that

$$2\sum_{i>j}\mathbb{E}X_iX_j \leq c''(n+n^{2-a}) = o(n^2),$$

where we used that $a > 0$. □

Before passing to the CLT, let us discuss the set-up. We would still like to work in a nice homogenous set-up, for example in a setting where the correlation structure does not change too much over the indexes. A suitable setting is that of stationary sequences:

[2]Please let me know if you meet any typos or unclarities!

Definition 3.3 (Stationary sequence)**.** *A two-sided sequence of random variables $(X_i)_{i\in\mathbb{Z}}$ is called stationary if for all $n \in \mathbb{N}$, all $i_1, \dots, i_n \in \mathbb{Z}$ and all $l \in \mathbb{Z}$ the vectors $(X_{i_1}, \dots, X_{i_n})$ and $(X_{i_1+l}, \dots, X_{i_n+l})$ have the same law. A one-sided sequence $(X_i)_{i\geq 1}$ is called stationary if the same thing holds whenever all the variables in two vectors are defined.*

Clearly the identity sequence $X_i = X_0$ for all $i \in \mathbb{Z}$ is stationary, as is any i.i.d sequence. The following lemma gives us a way to construct stationary sequences:

Lemma 3.4. *Let $(Y_i)_{i\in\mathbb{Z}}$ be i.i.d. with zero mean unit variance. Let $(a_i)_{i\in\mathbb{Z}}$ be real numbers with $\sum_i a_i^2 < \infty$. Then $X_i := \sum_k a_k Y_{k+i}$ is a stationary sequence of random variables.*

Proof. The proof follows directly from the fact $(Y_i)_{i\in\mathbb{Z}}$ has the same law as $(Y_{i+l})_{i\in\mathbb{Z}}$ for any $l \in Z$. □

It will be an exercise to try to use this construction to obtain sequences with covariance of the shape $\mathbb{E}X_iX_j \asymp c|i-j|^{-a}$ for any $a > 0$.

3.1. **CLT for dependent random variables.** Let us now discuss the case of the CLT. Again, by looking at the trivial example where $X_1 = X_2 = \dots X_n$, we see that a priori there is no reason that the CLT should hold in the case of dependencies. Does it hold as soon as we have independence at long distances? Even this is not clear, as the following example shows:

- Let $Y_1, Y_2, \dots$ be i.i.d. zero mean unit variance. Consider $X_i := Y_{i+1} - Y_i$. Then X_i and X_j are independent as soon as $|i-j| \geq 2$. However $\sum_{i=1}^n X_i = Y_{n+1} - Y_1$ which certainly does not need to be a Gaussian.

Now, what about the case of correlation of the type $|i-j|^{-a}$. We saw that LLN holds as long as $a > 0$. However, one can directly see that attention is needed for the case of the CLT:

Exercise 3.1. *Let X_i be mean zero, unit variance random variables with $\mathbb{E}X_iX_j \asymp c|i-j|^{-a}$ with $a > 0$. Calculate the variance of $\sum_{i=1}^n X_i$. When does it scale like n? How does it scale otherwise?*

Now, from the exercise you see that in the case $a < 1$ the variance grows faster than n. However, one can still normalize $\sum_{i=1}^n X_i$ differently by some b_n to get a limiting distribution. Yet, it does not necessarily have to be a Gaussian. This is interesting, because on the other hand the sum of any jointly Gaussians is a Gaussian, and thus their limit would also be a Gaussian. Hence, as long as we are out of the CLT regime, different limits can occur and what these limits can look like is still an open question. For an example, one could think about the model of long polymers, called self-avoiding walk: on $\mathbb{Z}^2$ we consider the walk along the lattice edges that never touches its past and put an uniform measure on all such walks of length n. This gives us a type of random walk. Now, for usual random walk, if we look at the $x-$coordinate of the endpoint after n steps, then this converges to a Gaussian distribution. However, the limiting distribution for the self-avoiding walk is yet to be determined!

Which method should we use to prove a CLT? To see this let's recall what we have seen and try to pinpoint where we used independence:

- Lindeberg method: we used independence to say that $\mathbb{E}X_i f(S^0_{n,i}) = \mathbb{E}X_i \mathbb{E}f(S^0_{n,i})$ and that $\mathbb{E}X_i^2 f(S^0_{n,i}) = \mathbb{E}X_i^2 \mathbb{E}f(S^0_{n,i})$.
- Entropic proof of CLT: we used independence in the quantitative decrease of Fisher information: $J(X+Y) \leq \beta^2 J(X) + (1-\beta)^2 J(Y)$, which gave us convergence of $J(S_n)$. We further used independence to determine the limiting law, by looking at the equality case in this inequality.
- Finally, we mentioned a method using Stein's equation: i.e. by using the fact that if $\mathbb{E}f'(X) = \mathbb{E}Xf(X)$ for all nice f, then X is a standard Gaussian. Here, dealing with $\mathbb{E}Xf(X)$ is the part where independence enters.

From these three methods, only the first (Lindeberg) and third (Stein's equation) seem feasible to a generalization. However, having to deal with all 'nice' functions f is a bit of an annoyance. Here, the following lemma helps us:

Lemma 3.5. *Let X be a finite variance random variable such that for all $f_\lambda(x) = \exp(i\lambda x)$ Stein's equation $\mathbb{E}f'_\lambda(X) = \mathbb{E}Xf_\lambda(X)$ holds. Then X is a standard Gaussian.*

Proof. We have that $\mathbb{E}f'_\lambda(X) = i\lambda\mathbb{E}\exp(i\lambda X) = i\lambda c(\lambda)$. On the other hand $\mathbb{E}X\exp(i\lambda X)$ can be written as $-i\mathbb{E}\frac{d}{d\lambda}\exp(i\lambda X)$. We can change the integration and differentiation (why?) to get $\mathbb{E}X\exp(i\lambda X) = -i\frac{d}{d\lambda}c(\lambda)$. This means we have the differential equation

$$\frac{d}{d\lambda}c(\lambda) = -\lambda c(\lambda.$$

Solving this and using the fact that c is a characteristic function we deduce that $c(\lambda) = \exp(-\frac{\lambda^2}{2})$. □

Before proving the CLT for dependent random variables, it is useful to see how to use this lemma to prove the CLT in the usual setting of i.i.d random variables of zero mean, unit variance. For simplicity, let us further assume they are bounded.

So, as always let $S_n := n^{-1/2}\sum_{i=1}^n X_i$. We aim to use a two-step strategy: 1) tightness is clear as $\mathbb{E}S_n^2 = 1$, and 2) to characterize the limit we want to use the previous lemma. So it suffices to show that

$$i\lambda\mathbb{E}\exp(i\lambda S_n) - \mathbb{E}S_n\exp(i\lambda S_n) \to 0.$$

Now, using the fact that X_i are identically distributed, we can write

$$\mathbb{E}S_n\exp(i\lambda S_n) = n\mathbb{E}n^{-1/2}X_1\exp(i\lambda S_n) = n^{1/2}\mathbb{E}X_1\exp(i\lambda S_n).$$

Further, using independence we can write

$$n^{1/2}\mathbb{E}X_1\exp(i\lambda S_n) = n^{1/2}\mathbb{E}X_1\exp(i\lambda n^{-1/2}X_1)\mathbb{E}\exp(i\lambda\hat{S}_{n,1}),$$

where by $\hat{S}_{n,i}$ we denote the sum, where we have omitted the i-th term. As X_i are uniformly bounded, it follows that $\mathbb{E}\exp(i\lambda n^{-1/2}X_1) \to 1$ and thus $\mathbb{E}\exp(i\lambda\hat{S}_{n,1}) - \mathbb{E}\exp(i\lambda S_n) \to 0$. Hence it suffices to show that

$$n^{1/2}\mathbb{E}X_1\exp(i\lambda n^{-1/2}X_1) \to i\lambda.$$

To do this, write by Taylor expansion

$$\exp(i\lambda n^{-1/2}X_1) = 1 + i\lambda n^{-1/2}X_1 + O(n^{-1}X_1^2).$$

Thus

$$n^{1/2}\mathbb{E}X_1\exp(i\lambda n^{-1/2}X_1) = n^{1/2}\mathbb{E}\left(X_1 + i\lambda n^{-1/2}X_1^2 + n^{-1}O|X_1^3|\right).$$

But now, as X_1 is centred, the first term is zero and as X_1 is bounded the last term also tends to zero. As $\mathbb{E}X_1^2 = 1$, the middle term, however, gives us exactly what we want. Our aim is now to make this strategy work to prove a version of a dependent CLT. The proof will be non-examinable.

Proposition 3.6. *Let $X_1, X_2, \ldots$ be stationary, let $\mathbb{E}X_i = 0$ and suppose that X_i are bounded. Assume that $\mathbb{E}|X_iX_j| \le C(|j-i| \vee 1)^{-a}$ with $a > 1$ and that furthermore for $i < j < k$ and $k < j < i$, we have that*

$$|\mathbb{E}\left((X_iX_j - \mathbb{E}X_iX_j)|X_k|\right)| \le C(1 \vee |j-k|)^{-a}. \tag{3.1}$$

Define further $\sigma^2 = \sum_{j\ge1}\mathbb{E}X_1X_j$. If $\sigma^2 > 0$, then $S_n := \frac{\sum_i X_i}{\sigma\sqrt{n}}$ converges in law to a standard Gaussian.

Remark 3.7. *The theorem can certainly be generalized to non-bounded variables, by possible rephrasing the conditions a bit. I will leave finding such possible generalization as a fun, non-examinable exercise.*

Remark 3.8. *This statement is not expected to hold in general for $a < 1$.*

Now, we saw that it suffices to prove that $\mathbb{E}\left[i\lambda \exp(i\lambda S_n)\right] - \mathbb{E}S_n \exp(i\lambda S_n) \to 0$. We would like to follow the proof of CLT for i.i.d random variables with the Stein method, however dependence is on our way.

A key observation is that we get more independence when we group elements together. Let $m_n \leq \sqrt{n}$ be some sequence going to ∞, to be specified later.

Claim 3.9. *If we define $Y_j := \sum_{i=(j-1)m_n+1}^{jm_n} X_i$. Then $\mathbb{E}Y_j = 0$, $\sigma^{-2}m_n^{-1}\mathbb{E}Y_j^2 \to 1$ and $\mathbb{E}|Y_jY_k| \leq \tilde{C}m_n^{2-a}(|j-k|)^{-a}$ for $j \neq k$.*

Thus if we define $Z_j := m_n^{-1/2}Y_j$ to have $\mathbb{E}|Z_j|^2 = O(1)$ we see that $\mathbb{E}|Z_jZ_k| \leq \tilde{C}m_n^{1-a}(|j-k|)^{-a}$ for $j \neq k$, i.e. it goes to zero with n for any fixed distance.

Proof. This is a direct calculation. □

We are now ready to prove the theorem. As I mentioned, the proof will be non-examinable.
[*Beginning of non-examinable section*]

Proof. Recall the notation $\hat{S}_{n,i}$ for the sum, where we have omitted the i-th term. We write

$$\mathbb{E}S_n \exp(i\lambda S_n) = \sum_{j=1}^{n/m_n} \mathbb{E}\left(\frac{Y_j}{\sigma\sqrt{n}}(1 + i\lambda\frac{Y_j}{\sigma\sqrt{n}} + O(n^{-1}|Y_j|^2))\exp(i\lambda\hat{S}_{n,j}))\right).$$

Notice that as $m_n \leq n^{1/2}$ we have that $Y_j n^{-1/2}$ is finite and we are happy to do a Taylor expansion.

Let us first show that

$$\sum_{j=1}^{n/m_n} \mathbb{E}\left(\frac{Y_j}{\sigma\sqrt{n}}\exp(i\lambda\hat{S}_{n,j}))\right) \to 0,$$

and

$$\sum_{j=1}^{n/m_n} \mathbb{E}\left(\frac{Y_j}{\sigma\sqrt{n}}O(n^{-1}|Y_j|^2)\exp(i\lambda\hat{S}_{n,j}))\right) \to 0.$$

For the second term, we can just use that $|Y_j|^3 \leq m_n^3$, that $|\exp(i\lambda\hat{S}_{n,j})| = 1$ and the fact that there are n/m_n terms in the sum, to bound the whole sum in absolute value by $n^{-1/2}m_n^2$. Thus, taking $m_n \leq o(n^{1/4})$ would give convergence to 0.

For the first term, let us show how to bound the case $j = 1$, the others following similarly. Write out the Taylor expansion for Y_2: to get

$$\mathbb{E}\left(\frac{Y_1}{\sigma\sqrt{n}}\exp(i\lambda\hat{S}_{n,1}))\right) = \mathbb{E}\left(\frac{Y_1}{\sigma\sqrt{n}}(1 + O(\frac{|Y_2|}{\sigma\sqrt{n}}))\exp(i\lambda\hat{S}_{n,1,2})\right).$$

We can bound this by

$$|\mathbb{E}\left(\frac{Y_1}{\sigma\sqrt{n}}\exp(i\lambda\hat{S}_{n,1,2})\right)| + \hat{C}\mathbb{E}\left(\frac{|Y_1Y_2|}{\sigma^2 n}\right).$$

In the first term we can iterate, to obtain

$$\mathbb{E}\left(\frac{Y_1}{\sigma\sqrt{n}}\exp(i\lambda\hat{S}_{n,j}))\right) \leq \sum_{k>1}\hat{C}\mathbb{E}\left(\frac{|Y_1Y_k|}{\sigma^2 n}\right) \leq \tilde{C}\sum_{k>1} n^{-1}m_n^{2-a}(|k-1|)^{-a} \leq C'n^{-1}m_n^{2-a},$$

where we have used $a > 1$. For other j we get similar bounds, and thus summing over we can bound

$$|\sum_{j=1}^{n/m_n} \mathbb{E}\left(\frac{Y_j}{\sigma\sqrt{n}}\exp(i\lambda\hat{S}_{n,j}))\right)| \leq m_n^{1-a} \to 0,$$

as $a > 1$. We next claim that

$$\sum_{j=1}^{n/m_n}\left[\mathbb{E}\left(i\lambda\frac{Y_j^2}{\sigma^2 n}\exp(i\lambda\hat{S}_{n,j}))\right) - \mathbb{E}\left(i\lambda\frac{Y_j^2}{\sigma^2 n}\right)\mathbb{E}\left(\exp(i\lambda S_n))\right)\right] \to 0. \tag{3.2}$$

As by Claim 3.9 we have that $\mathbb{E}\left(\frac{Y_j^2}{\sigma^2 m_n})\right) \to 1$ and there are n/m_n terms, this would show that

$$\sum_{j=1}^{n/m_n}\mathbb{E}\left(i\lambda\frac{Y_j^2}{\sigma^2 n}\exp(i\lambda\hat{S}_{n,j}))\right) \to i\lambda\mathbb{E}\left(\exp(i\lambda S_n))\right)$$

and thus finish the proof.

As in the proof of the usual CLT, one can see that $\mathbb{E}\left(\exp(i\lambda S_n))\right) - \mathbb{E}\left(\exp(i\lambda\hat{S}_{n,j}))\right) \to 0$ uniformly, if we take m_n as above $m_n = o(n^{1/4})$. Now, for any fixed j, we can again Taylor expand iteratively as above, and use the Condition (3.1) to bound

$$\mathbb{E}\left(\left(\frac{Y_j^2}{n} - \mathbb{E}\frac{Y_j^2}{n}\right)\exp(i\lambda\hat{S}_{n,j}))\right) \leq C' m_n^{3-a} n^{-3/2}.$$

Summing over j, we can bound (3.2) by $C' m_n^{2-a} n^{-1/2}$, which goes to zero if we choose $m_n = o(n^{1/4})$ as above. □

[*End of non-examinable section*]

4. Linear structure of Gaussians

[3]

The aim of these notes is to start discussing the intimate relation between finite-dimensional Gaussians and linear algebra. Recall the different equivalent ways of seeing Gaussian vectors $\overline{X}$ in $\mathbb{R}^n$:

- There exists a $A \in \mathbb{R}^{n\times n}$ of full rank such that $\overline{X} = A\overline{Y} + \overline{M}$, where $\overline{b} \in \mathbb{R}^n$ and $\overline{Y} = (Y_1, \ldots, Y_n)$ is the standard Gaussian, i.e. a vector of i.i.d. standard Gaussians. In this case $\mathbb{E}X_iX_j = (AA^T)_{ij}$.
- For every $\overline{v} \in \mathbb{R}^n$, we have that $\langle\overline{X}, \overline{v}\rangle := \sum_{i=1}^n X_i v_i$ is a Gaussian.
- There is a symmetric positive definite matrix C and a vector $\overline{M}$ such that for all vectors $\overline{\lambda} \in \mathbb{R}^n$, we have that $\mathbb{E}\exp(\langle\overline{\lambda}, \overline{X}\rangle) = \exp\left(\langle\overline{\lambda}, \overline{M}\rangle + \frac{1}{2}\overline{\lambda}^T C\overline{\lambda}\right)$. Here $\mathbb{E}\overline{X} = \overline{M}$ and $\mathbb{E}(X_i - M_i)(X_j - M_j) = C_{ij}$.

Most of the time for the ease of notation we will work with centred Gaussians, i.e. the case $\overline{M} = 0$ - this is no restriction as for the general case we just add the vector $\overline{M}$ back. Also, contrary to some notes we work for now only with the non-degenerate case, meaning that A has full rank, C is strictly positive definite and the variance of $\langle\overline{v}, \overline{X}\rangle$ is strictly positive for every non-zero $\overline{v}$.

One of the key aspects of Gaussians is that the law of Gaussian vectors is determined by their means and covariances:

Lemma 4.1. *If $\overline{X}$ and $\overline{X}'$ are two centred Gaussian vectors with $\mathbb{E}X_jX_i = EX_j'X_i'$, then $\overline{X} \sim \overline{X}'$.*

[3]Please let me know if you meet any typos or unclarities!

Proof. Recall that the law of a random vector $\overline{X}$ is determined by its characteristic function $\mathbb{E}\exp(\langle\overline{\lambda},\overline{X}\rangle)$. Now, from the expression given above, we see that for centred Gaussians the characteristic function is in turn uniquely determined by the covariance matrix. Thus we conclude. □

This gives a nice way to see that the standard Gaussian has orthogonal symmetry:

Proposition 4.2. *Let $\overline{X}$ be a standard Gaussian and O and orthogonal matrix. Then $O\overline{X}$ has also the law of a standard Gaussian. In particular the standard Gaussian is both rotation and reflection invariant.*

Moreover, for any orthonormal basis $\overline{e}_1,\dots,\overline{e}_n$ of $\mathbb{R}^n$, we can write $\overline{X}=\sum_{i=1}^n Z_i\overline{e}_i$, where Z_i are i.i.d. standard Gaussians.

Proof. By the lemma above, it suffices to check the covariance of $O\overline{X}$. Hence as $\mathbb{E}(O\overline{X})_i(O\overline{X})_j = (OO^T)_{ij} = \mathrm{Id}_{ij}$ we conclude.

For the second part, notice that such a writing is by definition true for the standard orthonormal basis. But every other basis is obtained by an orthogonal transformation from this basis, and thus we conclude. □

Hence, the standard Gaussian has two very nice properties:

(1) It is rotationally invariant.
(2) And all its co-ordinates are independent.

Somewhat surprisingly, it comes out that these two properties uniquely characterize the standard Gaussian:

Proposition 4.3 (A characterization of the standard Gaussian)**.** *Let $\overline{X}$ be a random vector in $\mathbb{R}^n$ that is rotationally invariant and such that all X_i are independent. Then $\overline{X}$ has the law of a constant times a standard Gaussian.*

We will prove the case $n=2$ here, and leave the general case to the exercise sheet. Notice that neither of the conditions on their own is sufficient, nor in fact very restrictive. For example to obtain rotationally invariant vectors one could just take some random variable X' and then set $(X,Y)\asymp(X'\cos\theta,Y'\sin\theta)$, with θ uniform on $[0,2\pi]$. In this case the independence of coordinates is not true in general - e.g. think of $X'=1$ a.s., in which case we get the uniform distribution on the circle. Similarly, to have a distribution with independent coordinates, we can just take (X,Y) with X,Y i.i.d. In this case the rotational invariance fails in general - e.g. think of X,Y both uniform on $[-1,1]$.

Proof. So consider the case of a random vector $(X_1,X_2)\in\mathbb{R}^2$ with X_1,X_2 i.i.d. Let us in fact give two different arguments.

For the first argument, we suppose that X_1 has finite variance. Then by rotational invariance we get that $aX_1+bX_2\sim X_1$ for all $a^2+b^2=1$. Then there some $c>0$ such that $a'X_1+b'X_2\sim c'X_1$ for all $a,b>0$. Thus as $X_1\sim X_2$, it follows that X_1 is a strictly stable law. By assumption it also has finite variance and thus it has to be a zero mean Gaussian. To generalize this argument to the case of infinite variance, one could for example just take the explicit expression of the law of stable laws and check that the Gaussian will be the only one to satisfy such rotational invariance.

For the second argument, suppose first that X_1 has a differentiable density and $J(X_1)<\infty$, where $J(X_1)$ is the Fisher information. By rotational invariance we have that $X_1+X_2\sim\sqrt{2}X_1$ and in particular $J(X_1+X_2)=J(\sqrt{2}X_1)$. But by scaling of Fisher information this means that

$$J(X_1+X_2)=\frac{J(X_1)}{2}=\frac{J(X_1)+J(X_2)}{4}.$$

But we have seen that this implies that X_1,X_2 are Gaussian. Hence the proposition follows for variables with differentiable density and finite Fisher information. To deal with the general case

we can again use a smoothing argument: we take further $\overline{Z} = (Z_1, Z_2)$ to be a standard Gaussian. Then $\overline{X} + \overline{Z}$ satisfies all the conditions of the proposition and hence both $X_1 + Z_1$ and $X_2 + Z_2$ are Gaussian. So we can finish by the following claim (which we already saw on an examples sheet):

Claim 4.4. *Suppose that X is a random variable and Z an independent standard Gaussian. If $X + Z$ is Gaussian, then also X is a Gaussian.*

Proof of claim: Suppose $X + Z$ is a Gaussian of variance σ^2. We can write out characteristic functions:

$$\exp(-\frac{\lambda^2\sigma^2}{2}) = \mathbb{E}\exp(i\lambda(X+Z)) = \mathbb{E}\exp(i\lambda X)\mathbb{E}\exp(i\lambda Z) = \exp -\frac{\lambda^2}{2}\mathbb{E}\exp(i\lambda X).$$

In particular $\mathbb{E}\exp(i\lambda X) = \exp(-\frac{\lambda^2(\sigma^2-1)}{2})$. □

□

Remark 4.5. *In fact an even stronger version of the last claim is true. This is called Cramer's theorem: if the sum of two independent random variables X and Y is Gaussian, then both X and Y are also Gaussian.*

4.1. **A closer look at the covariance matrix.** We saw that a centred Gaussian vector $\overline{X}$ is uniquely determined by its covariance matrix $C_{ij} = \mathbb{E}X_iX_j$. Let us start by stating some properties of this covariance matrix:

Lemma 4.6. *The covariance matrix C of a Gaussian in $\mathbb{R}^n$:*

- *is symmetric - $C_{ij} = C_{ji}$;*
- *is positive on the diagonal - $C_{ii} > 0$ and*
- *more generally is positive definite, i.e. for any $\overline{v} \in \mathbb{R}^n$, we have that $\langle \overline{v}, C\overline{v}\rangle > 0$.*

Proof. The first property is immediate from the fact that $C_{ij} = \mathbb{E}X_iX_j$, the second is just a special case of the third one. And finally for the third property we just notice that $\langle \overline{v}, C\overline{v}\rangle = \mathbb{E}\langle \overline{v}, \overline{X}\rangle^2$: indeed, we have that

$$\mathbb{E}\langle \overline{v}, \overline{X}\rangle^2 = \mathbb{E}(\sum_{i=1}^n v_iX_i)(\sum_{i=1}^n v_jX_j) = \sum_{ij} v_iv_j\mathbb{E}X_iX_j = \langle \overline{v}, C\overline{v}\rangle.$$

□

Interestingly, the converse is also true - any matrix satisfying these conditions is a covariance matrix for some Gaussian vector.

Proposition 4.7. *Suppose that $C_{ij} \in \mathbb{R}^{n\times n}$ is symmetric and positive definite. Then there is a centred Gaussian vector $\overline{X}$ in $\mathbb{R}^n$ such that $\mathbb{E}X_iX_j = C_{ij}$.*

Proof. This is basically a proof of linear algebra, but it is good to see it. Recall that one way to construct a Gaussian vector was to start from the standard Gaussian $\overline{Y}$ and then set $\overline{X} = A\overline{Y}$. Moreover, in this case we know that the covariance matrix is given by AA^T. Thus we want to just show that if C is symmetric and positive definite, then there is some A of full rank with $AA^T = C$.

To do this, recall that because C is symmetric, there is some orthogonal matrix O and some diagonal matrix E such that $O^TCO = E$ and moreover E has positive eigenvalues. Thus we can define F with $F^2 = E$ and thus $A := OF$ satisfies $AA^T = C$ as desired. □

4.2. **The spectral representation of Gaussians.** We have seen that any Gaussian vector $\overline{X}$ can be written as $\overline{X} = A\overline{Y}$, where Y is a standard Gaussian. In particular, it means that we can write $\overline{X} = \sum_{i=1}^{n} \overline{A}_i Y_i$, where A_i are the column vectors of A. We will now see that one can do better - one can write $\overline{X}$ as a sum of orthonormal vectors, weighted by independent Gaussians.

To do this, we recall the Spectral theorem: if $C \in \mathbb{R}^{n\times n}$ is a symmetric matrix, then $\exists \lambda_1, \ldots, \lambda_n \in \mathbb{R}$ and a set of orthonormal vectors $\overline{e}_1, \ldots, \overline{e}_n$ such that $C\overline{e}_i = \lambda_i \overline{e}_i$. If C_{ij} is in addition positive definite, it implies that $\lambda_i > 0$ for all $i = 1 \ldots n$. The very beautiful thing is that through the lens of the Spectral theorem, the linear map $C : \mathbb{R}^n \to \mathbb{R}^n$ becomes just a multiplication map - we have that $C\overline{e}_i = \lambda_i \overline{e}_i$ and thus for any vector $\overline{v}$ we have that

$$C\overline{v} = \sum_{i=1}^{n} \lambda_i \langle \overline{v}, \overline{e}_i \rangle e_i.$$

The Spectral theorem gives a similarly nice representation for any Gaussian vector:

Proposition 4.8. *Let $\overline{X}$ be a Gaussian in $\mathbb{R}^n$. Then there exist $\lambda_1, \ldots, \lambda_n$ and $\overline{e}_1, \ldots, \overline{e}_n$ an orthonormal basis such that*

$$\overline{X} = \sum_{i=1}^{n} \sqrt{\lambda_i} Z_i \overline{e}_i,$$

where $Z_1, \ldots, Z_n$ are independent standard Gaussians.

Proof. By the spectral theorem for the covariance matrix C of $\overline{X}$, we obtain $\lambda_1, \ldots, \lambda_n > 0$ and $\overline{e}_1, \ldots, \overline{e}_n$ an orthonormal basis with $C\overline{e}_i = \lambda_i e_i$. So it remains to verify that

Claim 4.9. *Define $\overline{X}' := \sum_{i=1}^{n} \sqrt{\lambda_i} Z_i \overline{e}_i$. Then $X' \sim X$.*

Proof of the claim: It suffices to check that for every vector $\overline{v}$, we have that $\mathbb{E}\langle \overline{v}, \overline{X}' \rangle^2 = \langle \overline{v}, C\overline{v} \rangle$. So write

$$\mathbb{E}\langle \overline{v}, \overline{X}' \rangle^2 = \mathbb{E}\langle \overline{v}, \sum_{i=1}^{n} \sqrt{\lambda_i} Z_i \overline{e}_i \rangle^2 = \mathbb{E}\left[\left(\sum_{i=1}^{n} \sqrt{\lambda_i} Z_i \langle \overline{v}, \overline{e}_i \rangle \right)^2 \right].$$

Using now that $Z_1, \ldots, Z_n$ are i.i.d, we can write this as

$$\sum_{i=1}^{n} \lambda_i \langle \overline{v}, \overline{e}_i \rangle^2 \mathbb{E} Z_i^2 = \sum_{i=1}^{n} \lambda_i \langle \overline{v}, \overline{e}_i \rangle^2 = \langle \overline{v}, C\overline{v} \rangle.$$

□

□

4.3. **Different viewpoints of a Gaussian vector.** We have already seen that a Gaussian vector $\overline{X}$ in $\mathbb{R}^n$ can be considered from different angles. For example,

- Just as a random vector with certain covariance.
- As a Gaussian process G parametrized by vectors $\overline{v}$: we set $G(\overline{v}) := \langle \overline{v}, \overline{X} \rangle$.
- From the spectral representation as a sum of eigenvectors, each weighted by an independent Gaussian.

Our aim is now to look at two more viewpoints:

4.3.1. *One Gaussian and many geometries.* Firstly, it comes out that any Gaussian vector can be in fact viewed as a standard Gaussian - one just has to change the Euclidean inner product for a different inner product, i.e. to change the underlying geometry of the space.

Given a positive definite symmetric matrix $C \in \mathbb{R}^{n\times n}$, define

$$\langle \overline{v}, \overline{w} \rangle_C := \langle \overline{v}, C^{-1}\overline{w} \rangle.$$

The following claim is left as an exercise:

Claim 4.10. *$\langle \overline{v}, \overline{w} \rangle_C$ defines an inner product on $\mathbb{R}^n$*

Thus $\mathcal{H}_C := (\mathbb{R}^n, \|\|_C)$ gives rise to a Hilbert space. Now if $\lambda_1, \ldots, \lambda_n > 0$ and $\overline{e}_1, \ldots, \overline{e}_n$ are the eigenvalues and eigenvectors of C as before, then a direct check shows that $\overline{f}_i := (\sqrt{\lambda_i}\overline{e}_i)_i$ with $i = 1 \ldots n$ form an orthonormal basis of $\mathcal{H}_C$. In particular, we see that if $\overline{X}$ is a Gaussian vector in $\mathbb{R}^n$ with covariance matrix C, then it can be in fact written as

$$\overline{X} = \sum_{i=1}^{n} Z_i \overline{f}_i.$$

Here Z_i are i.i.d. standard Gaussians and $\overline{f}_i$ form an orthonormal basis of $\mathcal{H}_C$. Thus one can say that $\overline{X}$ is a standard Gaussian on $\mathcal{H}_C$. It is a good exercise to think what are thus the symmetries of such a Gaussian vector?

The bottom line is the following: all finite-dimensional Hilbert spaces are isomorphic. For example, the isomorphism between $(\mathbb{R}^n, \|\|_2)$ and $\mathcal{H}_C$ is given by mapping $\overline{e}_i \to \overline{f}_i$ for all $i = 1 \ldots n$, with $\overline{e}_i$ and $\overline{f}_i$ as above. This in particular means that in some sense there is in fact really one standard Gaussian on a abstract Hilbert space, and moreover all Gaussians on $\mathbb{R}^n$ are just different instances of the standard Gaussians, arising through the different concrete representatives of this abstract Hilbert space. Thus all linear operations with Gaussians come down to linear operations with one standard Gaussian.

4.3.2. *Gaussian vectors as height functions.* There yet another viewpoint is that of random height functions. Namely, given a random vector $\overline{X}$, we can always think of it instead as a random function $X : \{1, \ldots, n\} \to \mathbb{R}$. Geometrically, this would give us a height function on $\{1, \ldots, n\}$. Sometimes one also thinks of $\{1, \ldots, n\}$ as of time-points and then one would think of $X(t)$ as a random time series. Two examples we will look at straight away are:

- The discrete white noise (DWN) W, arising when the corresponding vector $\overline{W}$ is a standard Gaussian on $\mathbb{R}^n$ - this means that all heights are given by independent standard Gaussians.
- The Gaussian random walk (GRW) Γ, where we set $\Gamma(0) := 0$ and $\Gamma(j) := \sum_{i=1}^{j} W(i)$.

Firstly, as the discrete white noise is basically just the standard Gaussian, it can be written using any orthonormal basis, weighted by independent Gaussians. In particular, when we think of the co-ordinates as the time domain, a frequency-based decomposition, i.e. a Fourier type of decomposition is sometimes of interest:

Lemma 4.11. *For $i = 1 \ldots n$, let $\phi_i(j) := \sqrt{\frac{2}{n+1}} \sin\left(\frac{\pi i}{n+1} j\right)$. Then $\phi_i(j)$ seen as vectors form an orthonormal basis of $(\mathbb{R}^n, \|\cdot\|_2)$. In particular, we can write the discrete white noise $W(j) = \sum_{i=1}^{n} Z_i \phi_i(j)$, where Z_i are independent standard Gaussians.*

Proof. This is on the exercise sheet. □

We will not say much more about the discrete white noise at the moment, but rather move to the slightly more interesting example of the Gaussian random walk.

Claim 4.12. *The covariance kernel of the Gaussian random walk is given by $C_{ij} := \mathbb{E}\Gamma_i\Gamma_j = \min(i, j)$.*

This follows from a direct calculation:

$$\mathbb{E}\Gamma_i\Gamma_j = \mathbb{E}\left[(\sum_{k=1}^{i} W(k))(\sum_{k=1}^{j} W(k))\right] = \sum_{k=1}^{\min(i,j)} \mathbb{E}W(k)^2 = \min(i,j).$$

In fact C_{ij} is something called the Discrete Green's function, with Dirichlet boundary conditions at 0 and free boundary conditions at n. Our next aim is to find a Fourier decomposition for the Gaussian random walk. We will do this by studying the spectral decomposition of the GRW and finding the underlying geometry of the Gaussian process. Whereas we could only talk in terms of matrices and inverse matrices, it is useful to already use the language of functional analysis as well, emphasising the fact that we look at the GRW as a random height function.

Definition 4.13 (Discrete Laplacian with free boundary)**.** *We define an operator Δ^F mapping functions $f : \{0, \dots, n\} \to \mathbb{R}$ with $f(0) = 0$ to themselves as follows:*

$$(\Delta^F f)(j) := \begin{cases} 2f(j) - f(j-1) - f(j+1) & \text{if } j < n \\ f(n) - f(n-1) & \text{if } j = n \end{cases}$$

Remark 4.14. *You ou should verify that in the languages of matrices, this means that:*

$$\Delta^F_i j := \begin{cases} 2 & \text{if } i = j \neq n \\ 1 & \text{if } i = j = n \\ -1 & \text{if } |i - j| = 1 \end{cases}$$

As in vector calculus, the things are set up here so that the Green's function and the Laplacian are inverse operators (matrices).

Lemma 4.15. *Let C_ij be the Green's function as before. Then as matrices $\Delta^F C = \mathrm{Id}$ or as operators $(\Delta^F C_i)(j) = \mathrm{Id}_{ij}$, where we define $C_i(j) := C_{ij}$.*

Proof. Let us work in the operator point of view. Recall that by definition $C_i(j) = \min(i, j)$, and thus we have that $2C_i(j) = C_i(j+1) + C_i(j-1)$ at all points where $j \neq i$. Moreover, at $j = i$ we have that $C_i(j) = C_i(j+1)$ and $C_i(j) = C_i(j-1) + 1$ and the claim follows. □

To find the spectral representation, it thus suffices to find eigenfunctions and eigenvalues of Δ^F_{ij}. As you might maybe guess, the eigenfunctions are again given by trigonometric functions. In fact we proved in class (check!) that for $i = 0 \dots n-1$

$$\psi_i(j) := \sin\left(\frac{2k+1}{2n+1}\pi j\right)$$

form an orthogonal basis with eigenvalues given by $\lambda_i^{-1} = 2 - 2\cos\left(\frac{2k+1}{2n+1}\pi j\right)$. Thus we can write

$$\Gamma(j) = \sum_{i=1}^{n-1} Z_i \sqrt{\lambda_i} c_i \psi_i(j),$$

where c_i are some constants such that $c_i\psi_i$ form an ON basis (it is a good exercise to find the values of c_1). In particular, this all means that the inherent geometry of the Gaussian random walk is given by the inner product

$$\langle f, g\rangle_{\Delta^F} := \langle f, \Delta^F g\rangle = \sum_{j=1}^{n} f(j)(\Delta_F g)(j) = \sum_{j=1}^{n-1} f(j)(2g(j) - g(j-1) - g(j+1)) + f(n)(g(n) - g(n-1)).$$

By simple reordering, this can be rewritten in the following very nice form:

$$\frac{1}{2}\sum_{j=1}^{n}(f(j) - f(j-1))(g(j) - g(j-1))$$

If you know about Brownian motion, does it remind you of something related to the Brownian motion?

4.4. **Conditioning of Gaussians.** Let us start the section on conditioning from a nice example about sample mean and variance of Gaussians, where independence of Gaussian vectors plays a role.

4.4.1. *Sample mean and variance.* If we observe any i.i.d. random variables $X_1, \dots, X_n$, we can try to estimate the mean and variance of X_1 by defining the sample mean:

$$\widehat{X} := \frac{\sum_{i=1}^n X_i}{n},$$

and the sample variance:

$$\widehat{\mathrm{Var}} := \sum_{i=1}^n (X_i - \widehat{X})^2.$$

Geometrically, one could think that the sample mean is somehow the centre of mass of the observations and the sample variance gives the fluctuation or dispersion around the centre of mass. Notice that a priori these random variables could be dependent in intricate ways. However, in the case of Gaussians things are extremely nice.

Proposition 4.16 (For Gaussians sample mean and sample variance are independent)**.** *If $X_1, \dots, X_n$ are i.i.d. standard normals, then $\widehat{X}$ and $\widehat{Var}$ are independent.*

To prove the proposition, we start with a small lemma saying that two Gaussian vectors are independent as soon as all the correlations between the components of the vectors are 0.

Lemma 4.17. *If $(\overline{X}, \overline{Y})$ is a centred Gaussian vector in $\mathbb{R}^{n+m}$ with $\overline{X}$ in $\mathbb{R}^n$ and $\overline{Y}$ in $\mathbb{R}^m$, then the following two conditions are equivalent:*

(1) $\overline{X}$ and $\overline{Y}$ are independent,

(2) for all $i = 1 \dots n$, $j = 1 \dots m$ we have that $\mathbb{E} X_i Y_j = 0$.

There are many ways to prove this lemma, for example an easy way would be to use the characteristic function. We will however use an even simpler way, also useful later on.

Proof. Clearly the first point implies the second point as $\mathbb{E} X_i Y_j = (\mathbb{E} X_i)(\mathbb{E} Y_j) = 0$. The other implication is a bit more interesting.

Now, we know that (centred) Gaussian vectors are determined by only their covariances. So the idea is the following: we just take $\overline{X} \sim \overline{X}'$ and $\overline{Y} \sim \overline{Y}'$ to be independent and check whether $(\overline{X}', \overline{Y}')$ has the same covariance structure as $(\overline{X}, \overline{Y})$. If it does, then as both $(\overline{X}, \overline{Y})$ and $(\overline{X}', \overline{Y}')$ are Gaussian vectors, they necessarily have the same law. In particular as by construction $\overline{X}'$ and $\overline{Y}'$ are independent, we conclude that $\overline{X}$ and $\overline{Y}$ are also independent.

So let us check that the covariance structures do agree. By construction we have that $\mathbb{E} X_i' X_k' = \mathbb{E} X_i X_k$ and $\mathbb{E} Y_i' Y_k' = \mathbb{E} Y_i Y_k$. Moreover, by the first part of the proof, we know that $\mathbb{E} X_i' Y_j' = 0$. Thus we conclude. □

We can now prove the proposition:

Proof of Proposition prop:smsv. For every $i = 1 \dots n$, we have that

$$\mathbb{E} \widehat{X}(X_i - \widehat{X}) = \mathbb{E} \frac{X_i^2}{n} - \mathbb{E} \widehat{X}^2 = n^{-1} - n^{-1} = 0.$$

Thus it follows from the Lemma above that $\widehat{X}$ is independent of the vector $(X_1 - \widehat{X}, \dots, X_n - \widehat{X})$. In particular it means that $\widehat{X}$ is independent of the the squared norm of this vector, that is exactly equal to the sample variance $\widehat{\mathrm{Var}}$. □

In fact it comes out that sample mean and sample variance are independent only in the case of Gaussians. We will not prove this interesting result here, but would be interesting to hear from your probabilistic proof attempts!

Theorem 4.18. *Suppose that $X_1, \dots, X_n$ are i.i.d. random variables such that $\widehat{X}$ and $\widehat{Var}$ are independent. Then X_i are in fact Gaussians.*

4.4.2. *Conditioning of Gaussians.* We just saw a nice and simple way to determine when two Gaussian vectors are independent. In the case of independent variables, conditioning on the value of one variable does not influence the law of the other one, making it straightforward to understand conditional laws. In general, understanding conditional laws can be quite challenging. However, it is also one of the most important questions in probability - for example, as both the world and our knowledge is changing all the time, we want to constantly update our probabilistic models and predictions based on existing observations. And this basically comes down to understanding conditional probability laws.

In some rare cases conditional laws can be analytically understood - this is the case with Gaussians, putting them also in forefront of probabilistic models. Let us start from an example - consider a vector of two independent standard Gaussians (X_1, X_2). What is the conditional law of this vector, given $X_1 + X_2 = a$?

Well, let us do what we know how to do: calculate conditional laws for independent vectors. Namely, by rotational invariance of (X_1, X_2) we have that

$$(X_1, X_2) \sim (\frac{X_1 + X_2}{\sqrt{2}}, \frac{X_1 - X_2}{\sqrt{2}}) \sim N(0, \mathrm{Id}).$$

Thus, given $X_1 + X_2 = a$, we still have that $\frac{X_1 - X_2}{\sqrt{2}} \sim N(0,1)$. On the other hand we can write

$$(X_1, X_2) \sim (\sqrt{2}^{-1}(\frac{X_1 + X_2}{\sqrt{2}} + \frac{X_1 - X_2}{\sqrt{2}}), \sqrt{2}^{-1}(\frac{X_1 + X_2}{\sqrt{2}} - \frac{X_1 - X_2}{\sqrt{2}})),$$

and thus conditioning on $X_1 + X_2 = a$, we find that

$$(X_1, X_2) \sim (a/2 + Z, a/2 - Z),$$

with $Z \sim N(0, 1/2)$.

Now let us turn to a general statement. First, observe that we can rephrase the conditioning above as saying that we conditioned on $\langle (1,1), (X_1, X_2) \rangle = a$. Thus in general we can ask for the law of (X_1, X_2), given $\langle \overline{v}, \overline{X} \rangle = a$.

Lemma 4.19. *Let $\overline{X}$ still be a two-dimensional standard Gaussian. Then, given $\langle \overline{v}, \overline{X} \rangle = a$, we have that $\overline{X} \sim a\frac{\overline{v}}{\|\overline{v}\|^2} + \overline{Z}$, where $\overline{Z} \sim N(0, \frac{\overline{v}^\perp (\overline{v}^\perp)^T}{\|\overline{v}^\perp\|^2})$.*

One way to prove this, is to argue similarly to the statement that zero correlation implies independence - i.e. by checking that the two expressions have the same mean and variance. But let us propose yet another instructive way to deduce it here.

Proof. Given $\overline{v}$, we can always choose an orthonormal basis $\frac{\overline{v}}{\|\overline{v}\|}, \frac{\overline{v}^\perp}{\|\overline{v}^\perp\|}$ and thus write

$$\overline{X} \sim Z_1 \frac{\overline{v}}{\|\overline{v}\|} + Z_2 \frac{\overline{v}^\perp}{\|\overline{v}^\perp\|},$$

with Z_1, Z_2 independent standard Gaussians.

Now, conditioning on $\langle \overline{v}, \overline{X} \rangle = a$ is equivalent to asking that $Z_1 = \frac{a}{\|\overline{v}\|}$. On the other hand the covariance of $Z_2 \frac{\overline{v}^\perp}{\|\overline{v}^\perp\|}$ can be seen to be equal to $\frac{\overline{v}^\perp (\overline{v}^\perp)^T}{\|\overline{v}^\perp\|^2}$. Thus we conclude. □

This can be further generalized to all Gaussian vectors. Indeed, we saw that every Gaussian vector is a standard Gaussian vector on some inner product space and one can observe that the proof above works on any Hilbert space. Thus we get:

Lemma 4.20. *Let $\overline{X}$ be a two-dimensional Gaussian with covariance kernel C. Then, given $\langle \overline{v}, \overline{X} \rangle_C = a$, we have that $\overline{X} \sim a\frac{\overline{v}}{\|\overline{v}\|_C^2} + \overline{Z}$, where $\overline{Z} \sim N(0, \frac{\overline{v}^{\perp}(\overline{v}^{\perp})^T}{\|\overline{v}^{\perp}\|_C^2})$, and the orthogonal complement is taken in $\langle , \rangle_C$.*

A nice corollary gives a very nice explicit way to write any Gaussian vector (X_1, X_2), making the effect of conditioning on X_1 plain clear and explicit:

Corollary 4.21. *Any Gaussian vector $\overline{X}$ in $\mathbb{R}^2$ can be written as*

$$(X_1, X_2) \sim (X_1, X_1\frac{\mathbb{E}X_1X_2}{\mathbb{E}X_1^2} + \sqrt{\mathbb{E}X_2^2 - \frac{(\mathbb{E}X_1X_2)^2}{\mathbb{E}X_1^2}}Z),$$

where Z is a standard Gaussian that is independent of X_1.

Again, a possible proof is to just compare the covariance structures of both sides. Yet, it is again instructive to deduce it directly from the lemma above:

Proof. Set $\overline{e}_1 = (1, 0)$ and $\overline{e}_2 = (0, 1)$. We then have that $X_1 = \langle \overline{X}, \overline{e}_1 \rangle_C$. Now, set also $\overline{C}_1 = (C_{11}, C_{21})$. Then we see that $\overline{C}_1$ and $\overline{e}_1$ are orthogonal and thus the lemma above implies that

$$\overline{X} = \frac{\overline{C}_1}{\|\overline{C}_1\|_C}Z_1 + \frac{\overline{e}_2}{\|\overline{e}_2\|_C}Z_2.$$

But now $C_{11} = \mathbb{E}X_1^2$, $C_{12} = \mathbb{E}X_1X_2$. Moreover, also $\|\overline{C}_1\|_2^2 = \mathbb{E}X_1^2$ and

$$\|\overline{e}_2\|_C^2 = C_{22}^{-1} = \frac{\mathbb{E}X_1^2}{\mathbb{E}X_1^2\mathbb{E}X_2^2 - (\mathbb{E}X_1X_2)^2}.$$

□

The ideas above naturally generalize to the setting of random vectors. The proof will left as an exercise.

Proposition 4.22. *Let $(\overline{X}, \overline{Y})$ be a Gaussian vector in $\mathbb{R}^{n+m}$ with $\overline{X} \in \mathbb{R}^n$ and $\overline{Y} \in \mathbb{R}^m$. Furthermore, write the covariance matrix of $(\overline{X}, \overline{Y})$ as a block matrix $C = \begin{pmatrix} C_{XX} & C_{XY} \\ C_{YX} & C_{YY} \end{pmatrix}$. Then*

$$(\overline{X}, \overline{Y}) \sim (\overline{X}, C_{YX}C_{XX}^{-1}\overline{X} + \overline{Z}),$$

where $\overline{Z}$ is independent of $\overline{X}$ and is distributed as a centred Gaussian with covariance matrix $C_{ZZ} = C_{YY} - C_{YX}C_{XX}^{-1}C_{XY}$.

4.5. **A peek into infinite dimensions.** [4]

We will now see how to push some of the ideas from previous classes to the infinite-dimensional setting. So suppose we are given an infinite-dimensional separable Hilbert space H. Then in particular H has a countable basis, which we denote by $(\phi_i)_{i\in\mathbb{N}}$. Can we construct Gaussians on H?

Recall that there are several equivalent ways to define Gaussians $\overline{X}$ on $\mathbb{R}^n$:

- We start from defining a standard Gaussian $\overline{Y} = (Y_1, \ldots, Y_n)$ and obtain general Gaussians by setting $\overline{X} = A\overline{Y}$, for some $n \times n$ full rank matrix A.
- We write the explicit density w.r.t. the Lebesgue measure on $\mathbb{R}^n$.

[4] Please let me know if you meet any typos or unclarities!

- We say that for each $\overline{v} \in \mathbb{R}^n$, the vector $\langle \overline{X}, \overline{Y} \rangle$ is a Gaussian.

In the first approach there is something promising: given a basis, we can at least formally define $\overline{X} := \sum_{i \geq 1} \phi_i X_i$ with $X_1, X_2, \ldots$ i.i.d. standard Gaussians. But does this infinite sum make sense, i.e. does it converge and in which sense? Moreover, if we go further from standard Gaussians - what should A be?

In the second approach there is a clear obstacle: there is no Lebesgue measure on an infinite dimensional space, at least not one that satisfies natural properties. Indeed, suppose there was a Lebesgue measure μ on $\mathbb{R}^{\mathbb{N}}$, say. Then we would expect it to at least be translational invariant and it would be natural to hope that boxes $[-a, b]^{\mathbb{N}}$ have finite positive mass. But we can write $[-1, 1]^{\mathbb{N}}$ as a disjoint union of infinite number of translations of $[0, 1]^{\mathbb{N}}$, making it impossible to satisfy the two conditions of the previous sentence.

So only the third approach would seem without problems, at least to begin with. To try to push it a bit, recall the definition of a Gaussian process:

Definition 4.23 (Gaussian process)**.** *Let I be any index set. Then a collection of random variables $(X_i)_{i \in I}$ defined on the same probability space is called a Gaussian process if for any finite subset $(i_1, \ldots, i_n)$ we have that $\overline{X} := (X(i_1), \ldots, X(i_n))$ is a Gaussian vector.*

4.5.1. *The standard Gaussian process indexed by H.* The next proposition says that you can at least in some sense construct a standard Gaussian process indexed by H.

Proposition 4.24 (Standard Gaussian process on H)**.** *Let H be a separable Hilbert space. Then there exists a centred Gaussian process G_f indexed by $f \in H$ such that*

- *For each $f \in H$, we have that $G_f \sim N(0, \|f\|_H^2)$,*
- *G_f is linear, i.e. for any $f, h \in H$ we have that $G_{f+h} = G_f + G_h$.*
- *$\mathbb{E} G_f G_h = \langle f, h \rangle_H$.*

Notice that in the finite-dimensional setting, we did not just say that all $\langle \overline{v}, \overline{X} \rangle$ are Gaussian, but also that this gives an equivalent way to define a Gaussian living in $\mathbb{R}^n$. It's good to pause for a thought about it, and we will come back to it after the proof.

Proof. Consider a probability space with i.i.d. standard Gaussians $Z_1, Z_2, \ldots$. Further, denote by $(\phi_i)_{i \in \mathbb{N}}$ a orthonormal basis of H.

Now for any $f \in H$, we can write a unique expansion $f = \sum_{i \geq 1} f_i \phi_i$. We then set $G_f = \sum_{i \geq 1} f_i Z_i := \lim_{n \to \infty} \sum_{i=1}^n f_i Z_i$. If we denote these latter partial sums by G_f^n, then notice that for $m > n$, we have that

$$\mathbb{E}|G_f^m - G_f^n|^2 = \sum_{i \geq n+1}^{m} f_i^2 \leq \sum_{i \geq n+1} f_i^2.$$

As the latter sums are decreasing with n and converge to 0, we see that G_f^n converge to G_f in L^2. That the limit is Gaussian is best separated to a claim, the proof of which is on the exercise sheet:

Claim 4.25. *Let $(X_n)_{n \geq 1}$ be a sequence of Gaussians of mean m_n and variance σ_n^2 such that $m_n \to m$ and $\sigma_n^2 \to \sigma^2$ as $n \to \infty$. Prove that the sequence $(X_n)_{n \geq 1}$ converges in law to a Gaussian X of mean m and variance σ^2.*

Now, to show that each vector $(G_{f_1}, \ldots, G_{f_n})$ is Gaussian, we need to show that for any $\overline{v} \in \mathbb{R}^n$, $\sum_{i=1}^n v_i G_{f_i}$ is Gaussian. But here one can similarly first consider $\sum_{i=1}^n v_i G_{f_i}^m$ and then argue convergence as before.

It remains to verify the two last conditions. The first of them follows as $G_{f+h}^m = G_f^m + G_h^m$ and thus this is true also in the limit. For the final condition, one can use linearity to write $4 G_f G_h = G_{f+h}^2 - G_{f-h}^2$ and then just use the first condition. □

Notice that we made some arbitrary choices in this constructive proof. However, as the following lemma shows, they do not matter:

Lemma 4.26 (Uniqueness for GP on a Hilbert space)**.** *Let H be a separable Hilbert space. Suppose we have two centred Gaussian processes G_f and G'_f indexed by H, such that G_f and G'_f have the same covariances. Prove that they have the same finite-dimensional distributions. Deduce that the standard Gaussian process defined on H in class does not depend on the arbitrary choice of basis.*

Proof. This is on the exercise sheet. □

Remark 4.27. *The existence could also be deduced from Kolmogorov extension theorem, and it would not give a stronger result.*

This basic existence result is already very nice, however there is some need for care. Indeed, to give sense to G_f for all $f \in H$, we needed to define uncountably many limits and although each single one gives a.s. a finite value, there is a priori no reason that all of them would be finite on an event of full probability!

This might sound a bit confusing, so let us consider two examples:

- Consider i.i.d variables $(X_i)_{i\in I}$ defined on the same probability space and suppose for each $i \in I$ almost surely X_i is finite. Now, if I is countable, then you can switch the order of quantifiers and deduce that almost surely for each $i \in I$, the variables X_i are finite. Indeed, suppose E_i is the event that X_i is not finite. Then when I is countable, $\mathbb{P}(E_i) = 0$ for all $i \in I$ also implies that $\mathbb{P}(\cup_{i\in I} E_i) \leq \sum_{i\in I} \mathbb{P}(E_i) = 0$, i.e. with zero probability some variable X_i is infinite. However, notice that in this inequality we really use the countable additivity of the measure and a similar claim cannot be made when I is uncountable - in fact even the question, are all X_i finite would not be measurable.
- Consider the standard Gaussian X. Now define random variables $Y_c := (X - c)^{-1/2}$ for all $c \in \mathbb{R}$. Then for any fixed c, the random variable Y_c is almost surely finite. However, a.s. the variable Y_X is infinite.

And indeed, motivated by these considerations let's see the following lemma, saying that we cannot hope that the above-defined Gaussian process behaves simultaneously nicely for all $f \in H$.

Lemma 4.28. *Almost surely, there is some random $f = f(Z_1, Z_2, \dots) \in H$ of unit norm such that G_f is infinite.*

Proof. Notice that because $Z_1, Z_2, \dots$ are standard Gaussians, for each $m \in \mathbb{N}$ we can find some n_m such that $Z_{n_m} > m$ and $n_m \neq n_k$ for $m \neq k$. Now, define $f = f(Z_1, Z_2, \dots) := \sum_{m\geq 1} m^{-1}\phi_{n_m}$. Then $\|f\|_H^2 = \sum_{m\geq 1} m^{-2} < 2$ and we can normalize it to have unit norm. On the other hand by definition

$$G_f = \lim_{k\to\infty} \sum_{m=1}^{k} m^{-1} Z_{n_m} > k.$$

Thus by the choice of Z_{n_m}, we have that $G_f = \infty$ almost surely. □

Another way to think about is as follows. With notations as above, consider the formal sum :

$$X_H = \sum_{i\geq 1} \phi_i Z_i.$$

This would be formally the standard Gaussian on H (and it is the standard Gaussian if H is finite-dimensional). It is easy to see that then formally $\langle X_h, f\rangle_H = G_f$. However, this inner product does not make literally sense because X_H does not belong to H:

Claim 4.29. *Define $X_h^m = \sum_{i=1}^m \phi Z_i$. Then almost surely $\|X_h^m\|_H \to \infty$ as $m \to \infty$ and thus X_H does not make sense as a limit in H.*

Proof. We can explicitly calculate that $\|X_h^m\|_H := \sum_{i=1}^m Z_i^2$. But by the strong law of large numbers, as Z_i^2 are i.i.d. positive and of finite mean 1, we have that $m^{-1}\sum_{i=1}^m Z_i^2 \to 1$. This in particular means that $\|X_h^m\|_H \to \infty$ almost surely. □

So in conclusion:

- There were some good news: there is a canonical Gaussian process indexed by a separable Hilbert space H.
- But there are also some bad news: contrary to the finite-dimensional case, the standard Gaussian on H does not belong to H in the infinite-dimensional case. Or, in an equivalent way the standard Gaussian process indexed by H is not finite for all $f \in H$ simultaneously.

However, despite the bad news, there are more good news. Recall that our Hilbert space was separable, and thus in fact we know that at least for some dense subset S of $f \in H$, we have that a.s. G_f is finite for all $f \in S$.

This is already good, but we can hope for still more - we can hope to find a dense Hilbert subspace $H_1 \subseteq H$, such that a.s. G_f is finite for all $f \in H_1$. Moreover, in the other direction we could also hope to find a Hilbert space H_2 such that $H \subseteq H_2$ and H is dense in H_2 with $X_H \in H_2$. Instead of looking into abstract theorems, let us consider a concrete example to illustrate this.

4.5.2. *The life of a Gaussian on a separable Hilbert space.* For a concrete example consider the sets

$$A_k(\mathbb{N}) = \{\overline{a} = (a_1, a_2, \dots) : \sum_{i=1}^{\infty} a_i^2 n^k < \infty\},$$

together with the inner products $\langle \overline{a}, \overline{b}\rangle_k := \sum_{i\geq 1} a_i b_i n^k$. Notice that in this language $(l^2(\mathbb{N}), \|\cdot\|_2)$ corresponds to A_0. The following exercise says that $(A_k, \langle \overline{a}, \overline{b}\rangle_k)$ give rise to Hilbert spaces that are nicely embedded inside each other.

Exercise 4.1 (Sandwiching Hilbert spaces)**.** *Let $A_k := \{\overline{a} = (a_1, a_2, \dots) : \sum_n a_n^2 n^k < \infty\}$ with the inner product $\langle \overline{a}, \overline{b}\rangle_k := \sum_n a_n b_n n^k$. Show that $(A_k, \langle \overline{a}, \overline{b}\rangle_k)$ are Hilbert spaces for any $k \in \mathbb{R}$. Prove that for any $k > j$, we have that $A_k \subseteq A_j$ and thus in particular $A_k \subseteq l^2(\mathbb{N}) \subseteq A_{-k}$ for any $k > 0$. Prove moreover that the set A_k is dense in A_j for all $k > j$.*

Let us now consider $l^2(\mathbb{N}) = A_0$ with its inner product $\langle \overline{a}, \overline{b}\rangle := \sum_{i\geq 1} a_i b_i$. There is the standard ON basis given by $\overline{e}^i$, with $\overline{e}^i$ equal to 1 on the $i-$th coordinate and 0 elsewhere. As above, let $Z_1, Z_2, \dots$ be i.i.d. standard Gaussians.

Now, by above we saw that there is a standard Gaussian process $G_{\overline{a}}$ indexed by $l^2(\mathbb{N})$, so that for each $\overline{a} \in l^2(\mathbb{N})$ we have a centred Gaussian $G_{\overline{a}}$ of variance $\|a\|^2$. Now, it makes fully sense to also consider the sequence $\overline{Z} = (Z_1, Z_2, \dots)$,. Thus everything might seem very nice here and so why don't we just define $G_{\overline{a}} = \langle \overline{Z}, \overline{a}\rangle$ and all good?

Well again, one can see that $\left\|\overline{Z}\right\|^2 = \infty$ almost surely, so the above pairing does not simultaneously make sense for all $\overline{a} \in l^2(\mathbb{N})$. As before, one can also construct a (random) $\overline{a}$ such that $G_{\overline{a}}$ is almost surely not finite. One can yet again rephrase this in a fancy way by saying that $\overline{Z}$ is not a continuous linear operator on $l^2(\mathbb{N})$, or more mundanely, $\overline{Z}$ is too rough to stay home.

However, we claim that $\overline{Z}$ does act continuously on a smaller space:

Claim 4.30. *For any $\epsilon > 0$, almost surely for all $\overline{a} \in A_{1+\epsilon}$, we have that $G_{\overline{a}} < \infty$.*

Proof. By Gaussian tail bounds we have that

$$\mathbb{P}(|Z_n| > n^{\epsilon/4}) \leq \exp(-\frac{n^{\epsilon/2}}{2}).$$

Thus by Borel-Cantelli Lemma, there is some random n_0 so that $|Z_n| \leq n^{\epsilon/4}$ for all $n > n_0$. We can write

$$G_{\overline{a}} = \sum_{i=1}^{\infty} Z_i a_i = \sum_{i=1}^{n_0} Z_i a_i + \sum_{n \geq n_0} Z_n a_n.$$

The first term is almost surely finite for all $\overline{a} \in l^2(\mathbb{N})$. For the second term we use $Z_n \leq n^{\epsilon/2}$, to bound

$$\sum_{i \geq n_0} Z_i a_i \leq \sum_{n \geq n_0} n^{\epsilon/4} |a_i| = \sum_{n \geq n_0} n^{-1/2-\epsilon/4} |a_i| n^{1/2+\epsilon/2}.$$

By Cauchy-Schwarz, we can further bound the latter term by $(\sum_{n \geq n_0} n^{-1-\epsilon/2})^{1/2} \|\overline{a}\|_{1+\epsilon}$, giving that $G_{\overline{a}}$ it is finite for all $\overline{a} \in A_{1+\epsilon}$. □

Similarly, we have that $\overline{Z}$ does belong to a slightly larger space. In fact, this would already follow from the previous claim by using duality (I let you figure out what I mean by this), but let us still do it explicitly:

Claim 4.31. *For any $\epsilon > 0$, almost surely we have that $\overline{Z} \in A_{-1-\epsilon}$.*

Proof. It suffices to show that almost surely $\left\|\overline{Z}\right\|_{-1-\epsilon} < \infty$. To do this, it suffices in turn to show that $\mathbb{E} \left\|\overline{Z}\right\|^2_{-1-\epsilon} < \infty$. But we can calculate

$$\mathbb{E} \left\|\overline{Z}\right\|^2_{-1-\epsilon} = \sum_{n \geq 1} n^{-1-\epsilon} \mathbb{E} Z_n^2 = \sum_{n \geq 1} n^{-1-\epsilon} < \infty.$$

□

I leave you to ponder upon whether $\overline{Z} \in A_{-1}$, or maybe not?

4.6. **An interlude on RHKS and Gaussians.** [5]

In these notes we will very briefly discuss Reproducing Kernel Hilbert Spaces and their relation to Gaussians. We will not prove the main theorems, but rather explain some ideas and connections in the easiest, finite-dimensional setting. The set-up is as follows:

- We have some set S;
- a Hilbert space $\mathcal{H}$ of functions $f : S \to \mathbb{R}$ with some inner product $\langle \cdot, \cdot \rangle_{\mathcal{H}}$;
- and a kernel C, i.e. a function $C(s,t) : S \times S \to \mathbb{R}$.

Definition 4.32 (Reproducing Kernel Hilbert Space)**.** *A pair $(\mathcal{H}, C)$ is called a Reproducing Kernel Hilbert Space (RKHS) if we are in the setting above, and in addition*

(1) For all $s \in S$, we have that $C_s(\cdot) = C(s, \cdot) \in \mathcal{H}$.
(2) Reproducing property: for all $s \in S, f \in \mathcal{H}$ we have that $f(s) = \langle f, C_s \rangle_{\mathcal{H}}$.

Notice that there is a certain freedom here, whether to ask for the reproducing property in the first or the second coordinate, and we will sometimes to it one way sometimes the other way. The nice thing about RKHS is the checking the value of functions at any point can be seen as a linear operation.

There are immediate questions:

(1) Given $\mathcal{H}$, is there always a reproducing kernel (RK) C such that $(\mathcal{H}, C)$ form a RKHS? Is such an C unique?
(2) Conversely, given a kernel C, is there always a Hilbert space $\mathcal{H}$ such that $(\mathcal{H}, C)$ is a RKHS? Is such an $\mathcal{H}$ unique?

[5] Please let me know if you meet any typos or unclarities!

A Hilbert space that is not a RKHS is for example $(L^2([0,1], \|\cdot\|_2)$. However, one could say that this is a cheat, as we are not even working with a space of real-valued functions, but rather with a space of equivalence classes of functions. A more interesting example would be a Hilbert space of real-valued functions, which has no RK. However, no such concrete example exist! And in fact to show the existence of such examples, Axiom of Choice is needed.

Theorem 4.33 (Existence of RKHS 1: finding C)**.** *Let $\mathbb{H}$ be a space of real-valued functions on S. Then there is a kernel C such that $(\mathcal{H}, C)$ is a RKHS if and only if for all $s \in S$, the linear functional $F : \mathcal{H} \to \mathbb{R}$ defined by $F(f) = f(s)$ is continuous.*

The necessity part is clear, what requires a bit of work is sufficiency - to show that as soon as the linear functional is continuous, we can find such a kernel, one uses Riesz representation theorem. This theorem is nicely complemented with a converse statement:

Theorem 4.34 (Existence of RKHS 2: finding $\mathcal{H}$)**.** *Let C be a kernel on S. Then there exists a Hilbert space of real-valued functions $\mathcal{H}$ such that $(\mathcal{H}, C)$ if and only if C is symmetric and positive semi-definite.*

Here one part of the theorem, that for a RKHS the kernel is symmetric and positive semi-definite is easy to see. The other part requires a certain amount of work, which we will briefly comment on later.

Finally, there is two-way uniqueness:

Theorem 4.35 (Uniqueness of RKHS)**.** *Suppose that $(\mathcal{H}, C)$ is a RKHS. Then if C' is any other kernel such that $(\mathcal{H}, C')$ is a RKHS, we have that $C = C'$. Similarly, if $\mathcal{H}'$ is any other Hilbert space such that $(\mathcal{H}', C)$ is a RKHS, we have that $\mathcal{H} = \mathcal{H}'$.*

Here, the first statement is relatively easy to prove and will be on the exercise sheet. As a consequence of this uniqueness theorem we see that it makes fully sense to talk about a Reproducing Kernel Hilbert Space $\mathcal{H}$, leaving out the kernel C, as it is uniquely determined by $\mathcal{H}$.

4.6.1. *RKHS and Gaussians in the finite dimensional setting.* Let us revisit the above theorems finite-dimensional setting, where things are more concrete and constructive. We will also want to see the interplay with Gaussians on $\mathbb{R}^n$.

- $S = \{1, \dots, n\}$;
- The Hilbert space $\mathcal{H}$ is given by $(\mathbb{R}^n, \langle \cdot, \cdot \rangle_{\mathcal{H}})$;
- and the kernel is just some function from $\{1, \dots, n\}^2 \to \mathbb{R}$, or in other words a matrix in $\mathbb{R}^{n \times n}$.

We denote the elements of $\mathcal{H}$ by $\overline{f}$ to stress that they are just vectors, but we think of them as of functions. Also we denote by $\langle \cdot, \cdot \rangle$ the standard inner product on $\mathbb{R}^n$.

Here is the analogue of Theorem 4.34 in the finite-dimensional setting, incorporating also uniqueness, and giving a constructive way to find the relevant Hilbert space.

Lemma 4.36 (Finding $\mathcal{H}$)**.** *Let C be a positive-definite symmetric matrix in $\mathbb{R}^{n \times n}$. Define $\mathcal{H}_C = (\mathbb{R}^n, \langle \cdot, \cdot \rangle_C)$ with $\langle \overline{f}, \overline{g} \rangle_C = \langle \overline{f}, C^{-1} \overline{g} \rangle$. Then $(\mathcal{H}_C, C)$ is a RKHS. Moreover $\mathcal{H}_C$ is the unique Hilbert space $\mathcal{H}'$ making $(\mathcal{H}', C$ a RKHS.*

Proof. To show that $(\mathcal{H}_C, C)$ is a RKHS we need to just show the reproducing property, all other properties are clear. So let $s \in S = \{1, \dots, n\}$. Denote by $\overline{C}_s$ the s-th column of C. Then

$$\langle \overline{f}, \overline{C}_s \rangle)_C = \langle \overline{f}, C^{-1} \overline{C}_s \rangle = \langle \overline{f}, \overline{e}_s \rangle = \overline{f}(s),$$

where we used that by definition $C^{-1} \overline{C}_s = \overline{e}_s$.

Uniqueness is also easy to argue in the finite-dimensional case: by the reproducing property we have that $\langle \overline{C}_s, \overline{C}_t \rangle_{\mathcal{H}} = C(s,t)$. Thus by linearity for all $a_1, \dots, a_m \in \mathbb{R}$, $b_1, \dots, b_m \in \mathbb{R}$ and $s_1, \dots, s_m \in S$, $t_1, \dots, t_m \in S$ we have that

$$\langle \sum_{i=1}^{m} a_i \overline{C}_{s_i}, \sum_{j=1}^{m} b_j \overline{C}_{t_j} \rangle_{\mathcal{H}} = \sum_{i,j}^{m} a_i b_j C(s_i, t_j).$$

But C is positive definite by assumption and hence is of full rank. Thus $\overline{C}_s$ with $s = 1 \dots n$ span $\mathbb{R}^n$ and the inner product is uniquely defined by C, implying that $\langle \overline{f}, \overline{g} \rangle_C = \langle \overline{f}, \overline{g} \rangle_{\mathcal{H}}$ for all $\overline{f}, \overline{g} \in \mathbb{R}^n$ and we conclude. □

The relation with Gaussians is now both very nice and very clear: a positive definite symmetric C is also a covariance kernel of a centred Gaussian $\overline{X}$ on $\mathbb{R}^n$. Moreover, we saw that this same $\overline{X}$ is then a standard Gaussian exactly on the space $\mathcal{H}_C$ - in other words the RKHS is the natural home for the Gaussian vector $\overline{X} \sim N(0, C)$!

Let us also see the other direction, i.e. the analogue of Theorem 1. Notice that in the finite-dimensional setting the evaluations are always continuous functionals, so we should expect the kernel always to exist. The question is - how to find it?

Lemma 4.37 (Finding C)**.** *Given $\mathcal{H} = (\mathbb{R}^n, \langle \cdot, \cdot \rangle_{\mathcal{H}}$ there is a unique C so that $(\mathcal{H}, C)$ is a RKHS.*

Proof. The uniqueness is argued exactly in the same way as in the general case and it will be on the example sheet. So let us just briefly discuss the existence of C.

Given the relation between Gaussians and RKHS, a natural approach offers itself for finding C: we consider the standard Gaussian $\overline{X}$ on $\mathcal{H}$ and set C to be the covariance matrix of this process. We know that C is then both positive definite and symmetric. So we just remains to prove that C has the reproducing property for $\mathcal{H}$. Now, we know that C has the reproducing property w.r.t. $\mathcal{H}_c = (\mathbb{R}^n, \langle \cdot, \cdot \rangle_C)$ from above. Hence it suffices to prove that $\mathcal{H}_C = \mathcal{H}$. There are several ways to see this, but it would for example follow from an appealing probabilistic statement:

Claim 4.38. *Suppose $\overline{X}$ is a standard Gaussian on two Hilbert spaces $H_1 = (\mathbb{R}^n, \langle \cdot, \cdot \rangle_{H_1})$ and $H_2 = (\mathbb{R}^n, \langle \cdot, \cdot \rangle_{H_2})$. Prove that $H_1 = H_2$.*

The proof of this claim is on the exercise sheet and it finishes the proof of the lemma. □

4.6.2. *RKHS and Gaussians in the infinite dimensional setting.* Let us just briefly comment on the infinite-dimensional case, i.e. the setting where S is infinite.

- To go from the kernel C to $\mathcal{H}$, we cannot explicitly replicate the proof of the finite-dimensional case, as the meaning of C^{-1} is in general not clear (it would be clear in the settings where the spectral theorem applies). Thus we have to go around and construct the space without explicitly mentioning C^{-1}. This is done by constructing the Hilbert space from linear combinations $\sum_{i=1}^{m} a_i C_{s_i}$ by setting the inner product equal to

  $$\langle \sum_{i=1}^{m} a_i C_{s_i}, \sum_{j=1}^{n} b_j C_{t_j} \rangle_{\mathcal{H}} := \sum_{i,j} a_i b_j C(s_i, t_j),$$

  and then using a completion argument to get a Hilbert space. We saw that such an identity holds in any RKHS and it is very nice that we can actually construct $\mathcal{H}$ this way, and very believable if one thinks about it. The uniqueness would basically follow similarly as the uniqueness in the finite-dimensional case.
- To prove Theorem 4.33, i.e. finding C for $\mathcal{H}$, one can just use Riesz representation theorem. Similarly to the finite-dimensional setting one could also identify C as the covariance matrix on S, when we construct the standard Gaussian process on $\mathcal{H}$. The fact that this Gaussian

process on $\mathcal{H}$ can be also seen as one on S follows from the reproducing property and is discussed below.

Let us finally be explicit about the relation between RKHS and Gaussian processes also in the infinite-dimensional setting. Firstly if we consider a Gaussian process G_f on a Hilbert space $\mathcal{H}$ of functions $f : S \to \mathbb{R}$, then this naturally induces a Gaussian process $\widehat{G}_s$ on S with covariance given by the kernel C that makes $(\mathcal{H}, C)$ a RKHS – indeed, we can just set $\widehat{G}_s := G_{C_s}$. This is well defined because $(\mathcal{H}, C)$ is a RKHS.

Conversely if G_s is a centered Gaussian process on S with the covariance kernel C, then in fact it can be considered as a standard Gaussian process on the Hilbert space $\mathcal{H}$ that makes $(\mathcal{H}, C)$ a RKHS - the easiest way to see this is maybe to start with a standard GP on $\mathcal{H}$, say $\widehat{G}_f$ and then note that if we set $\widetilde{G}_s := \widehat{G}_{C_s}$, then this defines a Gaussian process indexed by S that has the same covariance structure as the initial process G_s and thus by the uniqueness theorem for GPs we conclude that $G_s = \widetilde{G}_s$ as Gaussian processes.

All in all, we see that one can consider any Gaussian process as a standard Gaussian process on the related RKHS. This is useful in many respects, for example we have seen that the independence is related to orthogonality in $\mathcal{H}$, and it is useful to be able to evaluate the process at points $s \in S$ via an inner product. Moreover, an important property of Gaussians that we have not covered is that of Cameron-Martin shifts - i.e. the question for which functions $f : S \to \mathbb{R}$ is the Gaussian process $G_s + f$ absolutely continuous w.r.t. G_s. Here the answer is nice and clear - if and only if f belongs to the associated RKHS!