Gaussian processes

Manjunath Krishnapur
Indian Institute of Science, Bangalore

March 28, 2021

Contents

0.1 Plan

- Basics
  - Isoperimetric inequality and concentration.
  - Comparison inequalities: Slepian, Sudakov-Fernique, the more general comparison formula, Pieterbarg. The polytope question.
  - Chaining. Dudley. Fernique. For stationary processes optimal.
  - Kosambi-Karhunen-Loeve expansion.
  - Stationary processes. Spectral measure. Ergodicity and mixing. Prediction problem and Szegö's theorem.
  - Gaussian correlation conjecture?
  - Gaussian free field
  - Multiplicative chaos/cascade
  - Zeros of Gaussian polynomials
  - Nodal lines of random functions (Malevich, Nazarov-Sodin)

Chapter 1

Basics of Gaussian random variables

1.0.1 Standard Gaussian

A standard *normal* or *Gaussian* random variable is one with density $\varphi(x) := \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}x^2}$ on the real line. Its distribution function is $\Phi(x) = \int_{-\infty}^{x} \varphi(t)dt$ and its tail distribution function is denoted $\bar{\Phi}(x) := 1 - \Phi(x)$. If X_i are i.i.d. standard Gaussians, then $X = (X_1, \ldots, X_n)$ is called a standard Gaussian vector in $\mathbb{R}^n$. It has density $\prod_{i=1}^{n} \varphi(x_i) = (2\pi)^{-n/2} \exp\{-|\mathbf{x}|^2/2\}$ with respect to Lebesgue measure on $\mathbb{R}^n$. The measure is denoted γ_n, so that for every Borel set A in $\mathbb{R}^n$ we have $\gamma_n(A) = (2\pi)^{-n/2} \int_A \exp\{-|\mathbf{x}|^2/2\} d\mathbf{x}$.

The most important symmetry property of the standard Gaussian measure is this:

Exercise 1. [Orthogonal invariance] If $P_{n\times n}$ is an orthogonal matrix, then $\gamma_n P^{-1} = \gamma_n$ or equivalently, $PX \stackrel{d}{=} X$ when $X \sim \gamma_n$. More generally, if $p \leq n$ and $P_{p\times n}$ is a matrix such that $PP^t = I_p$ and $X \sim \gamma_n$, then $PX \sim \gamma_p$.

Conversely, if a random vector with independent co-ordinates has a distribution invariant under orthogonal transformations, then it has the same distribution as cX for some (non-random) scalar c.

To get an idea why, specialize to $n = 2$ and assume that X_1 has density $\psi(x)$ on $\mathbb{R}$. By the orthogonal invariance, X_2 has the same density and the independence of co-ordinates implies that the joint density is $\psi(x_1)\psi(x_2)$. Using orthogonal invariance again, we see that $\psi(\sqrt{x_1^2 + x_2^2})\psi(0) = \psi(x_1)\psi(x_2)$. The well-known characterization of the exponential function (if $f(x+y) = f(x)f(y)$ for all x, y and f is measurable, then $f(x) = e^{cx}$ or $f = 0$) shows that $\psi(x) = \exp(-cx^2)$ for some $c > 0$. We leave it as an exercise to generalize the proof to all dimensions and remove the assumption that X_1 has a density.

1.0.2 Multivariate Gaussian

If $Y_{m\times 1} = \mu_{m\times 1} + B_{m\times n} X_{n\times 1}$ where $X_{n\times 1} \sim \gamma_n$, then we say that $Y \sim N_m(\mu, \Sigma)$ with $\Sigma = BB^t$. Implicit in this notation is the fact that the distribution of Y depends only on Σ and not on B. This follows from Exercise 1: The matrices $B_{m\times n}$ and $C_{m\times p}$ with $p \le n$ satisfy $BB^t = CC^t$ if and only if $B = CP$ for an $p \times n$ matrix P that satisfies $PP^t = I_p$. It is a simple exercise that $\mu_i = \mathbf{E}[X_i]$ and $\sigma_{i,j} = \mathrm{Cov}(X_i, X_j)$. Further, if $Y \sim N_m(\mu, \Sigma)$ and $Z_{p\times 1} = C_{p\times m} Y + \theta_{p\times 1}$, then $Z \sim N_p(\theta + C\mu, C\Sigma C^t)$. Thus, affine linear transformations of normal random vectors are again normal.

Quite often, in studying $Y \sim N_n(\mu, \Sigma)$, it is useful to express it explicitly as $Y = BX + \mu$ where $X \sim \gamma_m$ and $B_{n\times m}$ satisfies $BB^t = \Sigma$. This is done by writing $\Sigma = P\Lambda P^t$ where $P_{n\times n}$ is an orthogonal matrix and Λ is diagonal with positive[1] entries, and taking $B = P\Lambda^{\frac{1}{2}} Q$ for any $Q_{n\times m}$ satisfying $QQ^t = I_n$. In particular, there is always the choice of $m = n$ and $B = \Sigma^{\frac{1}{2}} := P\Lambda^{\frac{1}{2}} P^t$, the unique symmetric matrix whose square is Σ. Another illuminating way is to write the spectral decomposition $\Sigma = \lambda_1 \mathbf{v}_1 \mathbf{v}_1^t + \ldots + \lambda_n \mathbf{v}_n \mathbf{v}_n^t$ and setting

$$X = \sum_{k=1}^{n} Z_k \sqrt{\lambda_k}\, \mathbf{v}_k$$

where Z_k are i.i.d. standard Gaussians. Strictly speaking, some eigenvalues of Σ could coincide, and then there are multiple choices for an orthonormal basis of eigenvectors. Orthogonal invariance ensures that the above representation is valid for any choice.

Exercise 2. Show that a random vector $X_{n\times 1}$ has multivariate Gaussian distribution if and only if every linear combination of X (i.e., $\mathbf{v}^t X$ for any $\mathbf{v} \in \mathbb{R}^n$) has univariate Gaussian distribution.

We say that a collection of random variables $(X_i)_{i\in I}$ (some index set I) is said to have joint Gaussian distribution if any finite sub-collection has multivariate Gaussian distribution. Equivalently, it just means that the finite linear combination $a_1 X_{i_1} + \ldots + a_k X_{i_k}$ has a univariate Gaussian distribution for every $k \ge 0$ and $i_1, \ldots, i_k \in I$ and $a_j \in \mathbb{R}$.

Exercise 3. Let $X \sim N_n(\mu, \Sigma)$ and $X' \sim N_n(\mu', \Sigma')$ be independent Gaussian vectors on a common probability space. Then, $X + X' \sim N_n(\mu + \mu', \Sigma + \Sigma')$.

[1] Thanks to the corona virus, let me take me in-class rants to the broader world. I use the word *positive* to mean non-negative. One reason is that it avoids the slight mental effort of hearing the sound *negative* and negating it, but what bothers me more is that if we must always say *non-negative* and *strictly positive*, we lose the use of the simpler and nicer word *positive*. I rest my case and hope that the supreme court does not find me in contempt of non-negativity (I am positive that it is not that idiotic).

Exercise 4. $N_n(\mu,\Sigma)$ has density with respect to Lebesgue measure on $\mathbb{R}$ if and only if Σ is non-singular, and in that case the density is

$$\frac{1}{(2\pi)^{n/2}\sqrt{\det(\Sigma)}}\exp\left\{-\frac{1}{2}(\mathbf{y}-\mu)^t\Sigma^{-1}(\mathbf{y}-\mu)\right\}.$$

If Σ is singular, then X takes values in a lower dimensional subspace in $\mathbb{R}^n$ and hence does not have density.

1.0.3 Covariance matrices

Let $\mathcal{P}_n$ denote the set of $n\times n$ positive semi-definite matrices and $\mathcal{P}_n^+$ denote the subset of positive definite matrices. That is, $\Sigma\in\mathcal{P}_n$ if and only if $\mathbf{v}^t\Sigma\mathbf{v}\geq 0$ for all $\mathbf{v}\in\mathbb{R}^n$. And $\Sigma\in\mathcal{P}_n^+$ if and only if strict inequality holds for $\mathbf{v}\neq 0$.

$\Sigma\in\mathcal{P}_n$ if and only if $\Sigma=BB^t$ (in one direction B may be allowed to be rectangular, in the other it may be restricted to be square). Therefore, $\mathcal{P}_n$ is precisely the set of $n\times n$ covariance matrices. Some basic facts about such matrices are collected in this exercise.

Exercise 5. Let A be a real symmetric $n\times n$ matrix. The following are equivalent.

1. $A\in\mathcal{P}_n$, i.e., $\mathbf{v}^tA\mathbf{v}\geq 0$ for all $\mathbf{v}\in\mathbb{R}^n$.
2. $A=BB^t$ for some matrix $B_{n\times m}$ for some m.
3. $A=C^2$ for some real symmetric $n\times n$ matrix C.
4. $A=e^X$ for some $n\times n$ symmetric matrix X.
5. The eigenvalues of A are non-negative.
6. A is a covariance matrix: $A_{i,j}=\mathbf{E}[X_iX_j]$ for some random variables $X_1,\ldots,X_n$ on some probability space.
7. A is a Gram matrix: $A_{i,j}=\langle\mathbf{v}_i,\mathbf{v}_j\rangle$ for some vectors $\mathbf{v}_1,\ldots,\mathbf{v}_n$ in some Hilbert space.

1.0.4 Gaussian process

Let $X=(X_t)_{t\in T}$ indexed by a set T, be a collection of random variables on a common probability space $(\Omega,\mathcal{F},\mathbf{P})$. Equivalently, $X:\Omega\mapsto\mathbb{R}^T$ is a random variable, where the set

$\mathbb{R}^T$ is endowed with the cylinder sigma algebra (the smallest sigma algebra for which the projections $\omega \mapsto \omega(t)$ from $\mathbb{R}^T$ to $\mathbb{R}$ are measurable. Such an X is called a *stochastic process*.

A stochastic process X is said to be a *Gaussian process* if every finite linear combination $c_1X(t_1) + \ldots + c_nX(t_n)$, where $n \geq 1$, $t_i \in T$, $c_i \in \mathbb{R}$, has a one-dimensional Gaussian distribution (possibly degenerate). Equivalently, one may say that any finite sub-collection has a multivariate Gaussian distribution. By what we have seen for multivariate Gaussian vectors, the distribution of X (which is entirely determined by distributions of all finite subcollections) is determined by the mean function $m : T \mapsto \mathbb{R}$ and covariance kernel $K : T \times T \mapsto \mathbb{R}$ given by $m(t) = \mathbf{E}[X(t)]$ and $K(t,s) = \mathbf{E}[X(t)X(s)]$.

Example 6. Let $T = \mathbb{R}$ and let $X = (X_t)$, where X_t are i.i.d. $N(0,1)$ random variables. Then X is called white noise. It exists, and its distribution is the product measure $\otimes_T N(0,1)$. Observe that the only probability questions that can be asked about X are about events and random variables that depend only on countably many coordinates. One cannot ask, for example, for the probability that X is a continuous function of t.

Example 7. Let $T = [0,\infty)$ and let $m(t) = 0$ and $K(s,t) = s \wedge t$. Brownian motion W is a stochastic process with this mean function and covariance kernel. But it is not the only one such. Construct a process with discontinuous sample paths that has the same mean and covariance.

There is no contradiction here. On the cylinder sigma-algebra on $\mathbb{R}^T$, there is at most one Gaussian measure with given mean function and covariance kernel. But when we talk of Brownian motion with continuous sample paths, we are talking of a measure on $C(T)$ and not on $\mathbb{R}^T$. As the latter only contains events described by countably many coordinates, it is too weak and uninteresting. One of the important questions we shall investigate is to find conditions on the covariance kernel so that there exists a Gaussian process taking values in $C(T)$ (assuming T has a topology) and having that covariance (say when $T = [0,\infty)$).

Example 8. Let T be an index set and let $m : T \mapsto \mathbb{R}$ and $K : T \times T \mapsto \mathbb{R}$. Show that there exists a Gaussian process with mean function m and covariance kernel K if and only if K is p.s.d. By definition, this means that $\sum_{i,j=1}^n c_i c_j K(t_i,t_j) \geq 0$ for any $n \geq 1$, $t_i \in T$, $c_i \in \mathbb{R}$.

1.0.5 Characteristic function

The characteristic function of $X \sim N_d(0,\Sigma)$ is $\mathbf{E}[e^{i\langle \lambda, X\rangle}] := e^{-\frac{1}{2}\lambda^t \Sigma \lambda}$ for $\lambda \in \mathbb{R}^d$. We leave this as an exercise. This extends to all of $\lambda \in \mathbb{C}^d$ analytically and the moment generating function

is $\mathbf{E}[e^{\langle \lambda, X\rangle}] = e^{\frac{1}{2}\lambda^t \Sigma \lambda}$ is the characteristic function evaluated at $i\lambda$.

Exercise 9. If $X \sim N_d(0,\Sigma)$, show that $\mathbf{E}[e^{i\langle \lambda, X\rangle}] = e^{-\frac{1}{2}\lambda^t \Sigma \lambda}$ for $\lambda \in \mathbb{R}^d$. In particular, if $X \sim N(0,\sigma^2)$, then its characteristic function is $\mathbf{E}[e^{i\lambda X}] = e^{-\frac{1}{2}\sigma^2\lambda^2}$ for $\lambda \in \mathbb{R}$.

Closely related is the moment generating function, $\mathbf{E}[e^{\langle \lambda, X\rangle}] = e^{\frac{1}{2}\lambda^t \Sigma \lambda}$ for $\lambda \in \mathbb{R}^d$. If we extend the characteristic function to $\lambda \in \mathbb{C}^d$ (in this case it does extend naturally), then evaluating it at a point where all co-ordinates are imaginary gives us the moment generating function. As a special case, when $X \sim N(0,\sigma^2)$, we have

$$\mathbf{E}[e^{\lambda X - \frac{1}{2}\lambda^2\sigma^2}] = 1.$$

This is a combination that occurs often. For example, when W is a standard Brownian motion, it is well-known that $M_\lambda(t) = e^{\lambda W_t - \frac{1}{2}\lambda^2 t}$ is a martingale, for any fixed $\lambda \in \mathbb{R}$.

1.0.6 Moments and Wick formula

Differentiating the characteristic function at $\lambda = 0$, one can get all the joint moments of the X_is. In the univariate case, $\mathbf{E}[X^{2p}] = \sigma^{2p}(2p-1)\times(2p-3)\times\ldots\times 3\times 1$. Of course the odd moments are zero. To see this, let $\sigma^2 = 1$ without loss of generality and write $\mathbf{E}[X^{2p}]$ as $\frac{d^{2p}}{d\lambda^{2p}}e^{-\lambda^2/2}$. As $e^{-\lambda^2/2} = \sum_{p\geq 0}\frac{1}{2^p p!}\lambda^{2p}$, it follows that $\mathbf{E}[X^{2p}] = \frac{(2p)!}{2^p\, p!}$, which is the same as the claimed formula. The $2p$-th moment has the combinatorial interpretation of the number of ways to divide up $[2p] := \{1,2,\ldots,2p-1,2p\}$ into p disjoint pairs. This generalizes as follows.

Exercise 10. Prove the *Wick formula* (also called *Feynman diagram formula*): Let $X \sim N_n(0,\Sigma)$. Then, $\mathbf{E}[X_1\ldots X_n] = \sum_{M\in\mathcal{M}_n}\prod_{\{i,j\}\in M}\sigma_{i,j}$, where $\mathcal{M}_n$ is the collection of all matchings of the set $[n]$ (thus $\mathcal{M}_n$ is empty if n is odd) and the product is over all matched pairs.

Observe that a matching is an unordered collection of $n/2$ pairwise disjoint subsets of $[n]$ of size 2 each. Because of the symmetry of the covariance matrix, it does not matter how we order each pair or the collection of pairs. For example, $\mathbf{E}[X_1X_2X_3X_4] = \sigma_{12}\sigma_{34} + \sigma_{13}\sigma_{24} + \sigma_{14}\sigma_{23}$.

Cumulants: Let X be a real-valued random variable with $\mathbf{E}[e^{tX}] < \infty$ for t in a neighbourhood of 0. Then, we can write the power series expansions

$$\mathbf{E}[e^{i\lambda X}] = \sum_{k=0}^{\infty} m_n(X)\frac{\lambda^n}{n!}, \qquad \log\mathbf{E}[e^{i\lambda X}] = \sum_{k=1}^{\infty}\kappa_n[X]\frac{\lambda^n}{n!}.$$

Here $m_n[X] = \mathbf{E}[X^n]$ are the moments while $\kappa_n[X]$ is a linear combination of the first n moments ($\kappa_1 = m_1$, $\kappa_2 = m_2 - m_1^2$, etc). Then κ_n is called the nth cumulant of X. If X and Y are independent, then it is clear that $\kappa_n[X+Y] = \kappa_n[X] + \kappa_n[Y]$.

Exercise 11. (optional). Prove the following relationship between moments and cumulants. The sums below are over partitions Π of the set $[n]$ and $\Pi_1, \ldots, \Pi_{\ell_\Pi}$ denote the blocks of Π.

$$m_n[X] = \sum_{\Pi} \prod_i \kappa_{|\Pi_i|}[X], \qquad \kappa_n[X] = \sum_{\Pi} (-1)^{\ell_\Pi - 1} \prod_i m_{|\Pi_i|}[X].$$

Thus $\kappa_1 = m_1$, $\kappa_2 = m_2 - m_1^2$,

If $X \sim N(\mu, \sigma^2)$, it has charcteristic function $e^{i\mu\lambda - \lambda^2/2\sigma^2}$, which shows that $\kappa_1[X] = \mu$, $\kappa_2[X] = \sigma^2$ and $\kappa_p[X] = 0$ for $p \geq 3$. The converse of this result is also true and often useful in proving that a random variable is normal. For instance, the theorem below implies that to show that a sequence of random variables converges to normal, it suffices to show that cumulants $\kappa_m[X_n] \to 0$ for all $m \geq m_0$ for some m_0.

Result 12 (Marcinkiewicz)**.** If X is a random variable with finite moments of all orders and $\kappa_n[X] = 0$ for all $n \geq n_0$ for some n_0, then X is Gaussian.

These considerations extend to the multivariate random vectors $X_{d\times 1}$, by writing

$$\mathbf{E}[e^{i\langle \lambda, X\rangle}] = \sum_{(k_1,\ldots,k_d): k_i \geq 0}^{\infty} m_X[k_1, \ldots, k_d] \frac{\lambda_1^{k_1} \ldots \lambda_d^{k_d}}{k_1! \ldots k_d!},$$

$$\log \mathbf{E}[e^{i\langle \lambda, X\rangle}] = \sum_{(k_1,\ldots,k_d) \neq 0: k_i \geq 0}^{\infty} \kappa_X[k_1, \ldots, k_d] \frac{\lambda_1^{k_1} \ldots \lambda_d^{k_d}}{k_1! \ldots k_d!}$$

but we omit the details here.

1.0.7 Conditional distributions

We saw that Gaussianity is preserved by affine linear transformations and also under convolutions. Now we recall that marginals and conditional distributions of Gaussians are Gaussians.

Exercise 13. If $U_{k\times 1}$ and $V_{(m-k)\times 1}$ are such that $Y^t = (U^t, V^t)$, and we write $\mu^t = (\mu_1, \mu_2)$ and $\Sigma = \begin{bmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{bmatrix}$ are partitioned accordingly, then

1. $U \sim N_k(\mu_1, \Sigma_{11})$.

2. $U\Big|_V \sim N_k(\mu_1 - \Sigma_{12}\Sigma_{22}^{-1/2}V,\ \Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21})$ (assume that Σ_{22} is invertible).

It is an important observation that the conditional expectation is linear in the conditioning variable and the conditional variance does not depend on it at all.

It is worth noting the particular case $k = 1$. Let $X \sim N_n(\mu, \Sigma)$. Then the conditional dis- tribution of X_1 given $Y = (X_2, \ldots, X_n)$ is Gaussian with mean $\mu_1 - u^t B^{-1} Y$ and variance $\sigma_{1,1} - u^t B^{-1} u$, where $u^t = (\sigma_{1,2}, \ldots, \sigma_{1,n})$ and $B = (\sigma_{i,j})_{2 \le i,j \le n}$.

Recall the interpretation of conditional expectation as projection in L^2. Thus, if we want to find the function f that minimizes $\mathbf{E}[|X_1 - f(Y)|^2]$, then the solution is $f(Y) = \mu_1 - u^t B^{-1} Y$ and the minimum value of the error is $\sigma_{1,1} - u^t B^{-1} u$. Thus, the conditional mean is the best predictor of X_1 in terms of Y and the conditional variance measures how good this predictor is.

One of the topics we shall study is the problem of prediction when we have an infinite collection of random variables $X = (X_n)_{n \in \mathbb{Z}}$ (with additional property of stationarity). When predicting X_0 in terms of the "past observations" $Y = (X_n)_{n<0}$, it may happen that the conditional variance is zero, which means that X_0 is determined by the past (and then each X_k is determined by the preceding X_is, hence all the randomness in the process is at the infinitely distant past!). There is a nice condition that tells us when this happens.

Exercise 14. Let $(X_t)_{t \in T}$ be a Gaussian process. Fix $A \subseteq T$ and $t_0 \in T$. Show that the conditional distribution of $X(t_0)$ given $\mathcal{G} = \sigma\{X_t : t \in A\}$ is Gaussian. What are its parameters? (Give the answer in terms of conditional expectations given $\mathcal{G}$)

1.0.8 Weak convergence

It is very easy to see that if $\mu_n \to \mu$ and $\Sigma_n \to \Sigma$, then $N_d(\mu_n, \Sigma_n) \xrightarrow{d} N(\mu, \Sigma)$. The converse is also true.

Exercise 15. The family of distributions $N_d(\mu, \Sigma)$, where $\mu \in \mathbb{R}^d$ and $\Sigma_{d \times d}$ is a positive semi-definite matrix, is closed under convergence in distribution (for this statement to be valid we include $N(\mu, 0)$ which is taken to mean δ_μ).

1.0.9 Characterizations of Gaussians

Gaussian distributions have many special properties. Some properties are not shared with any other distributions. Such properties *characterize* Gaussian distributions. We just mention a few of these here.

1. If $X = (X_1, \ldots, X_n)$ is a random vector such that X_is are independent and $PX \stackrel{d}{=} X$ for all orthogonal matrices P, then $X \sim N_n(0, \sigma^2 I_n)$ for some $\sigma^2 \geq 0$.

2. If X_i are i.i.d. (real-valued) and $X_1 + X_2 \stackrel{d}{=} \sqrt{2}X_1$, then $X_i \sim N(0, \sigma^2)$.

3. If $\mathbf{E}[Xf(X)] = \mathbf{E}[f'(X)]$ for all $f : \mathbb{R} \mapsto \mathbb{R}$ with bounded continuous derivative, then $X \sim N(0,1)$. This is Stein's characterization and we talk about it later.

4. If the cumulants of X eventually vanish, i.e., if $\log \mathbf{E}[e^{i\lambda X}]$ is a polynomial of λ, then $X \sim N(\mu, \sigma^2)$ for some μ, σ^2.

One can extend this list endlessly, and each property sheds some light on the importance and ubiquity of Gaussian distribution. For example, the second property explains why it arises as the limiting distribution in central limit theorem.

1.0.10 Entropy

Here we explain another "reason" for the appearance of the Gaussian.

Definition 16. Let X be a $\mathbb{R}^n$-valued random variable. Its *entropy* is defined as $-\infty$ is X does not have density with respect to Lebesgue measure, and if X does have density f with respect to Lebesgue measure, then the entropy is defined as

$$\mathcal{E}(X) := \int_{\mathbb{R}^n} \log \frac{1}{f(x)}\ f(x)dx.$$

As with expectation, variance etc., entropy depends only on the distribution of X.

Example 17. If $X \sim \text{Unif}[a,b]$, then $\mathcal{E}(X) = \log(b-a)$.

Example 18. If $X \sim N(\mu, \sigma^{2)})$, then

$$\begin{aligned}
\mathcal{E}(X) &= \int_{\mathbb{R}} \left[\frac{(x-\mu)^2}{2\sigma^2} + \log(\sigma\sqrt{2\pi}) \right] \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\frac{(x-\mu)^2}{2\sigma^2}} dx \\
&= \frac{1}{2\sigma^2}\mathbf{E}[(X-\mu)^2] + \log(\sigma\sqrt{2\pi}) \\
&= \log\sigma + \frac{1}{2}\log(2\pi e).
\end{aligned}$$

From these examples, we see that the entropy is low (possibly negative) when the distribution is concentrated (approaching a discrete distribution) and increases to infinity as the distribution spreads out. For instance, if $X = tY$, then $\mathcal{E}(X) = \mathcal{E}(X) + \log t$. In the first example, by considering the equality case in Jensen's inequality, it is easy to see that Uniform$[a,b]$ uniquely has the largest entropy among all distributions supported on $[a,b]$. Here is a similar problem to which Gaussian distribution is the answer.

Lemma 19. *Among all distributions with given mean and variance, the Gaussian distribution uniquely has the highest entropy.*

The same is true in higher dimensions if we fix the mean vector and the covariance matrix. We leave that as an exercise and present the proof for the one-dimensional case.

Proof. If X is any random variable with finite variance σ^2, and φ_σ is the Gaussian density with the same variance, then

$$\begin{aligned}\mathcal{E}(X) &= -\int_{\mathbb{R}} \left[\log \frac{f(x)}{\varphi_\sigma(x)} + \log \varphi_\sigma(x)\right] f(x)dx \\ &= \int_{\mathbb{R}} \log \frac{\varphi_\sigma}{f}\ f - \frac{1}{2}\int_{\mathbb{R}} x^2 f(x)dx.\end{aligned}$$

The second integral is equal to $\int x^2 \varphi_\sigma(x)dx$. The first integral is bounded above by $\log\left(\int_{\mathbb{R}} \frac{\varphi_\sigma}{f}\ f\right)$ which is zero. Further the bound is an equality when $f = \varphi_\sigma$. Thus, $\mathcal{E}(f) \le \mathcal{E}(\varphi_\sigma)$. ■

Note that we did not use the equality of means. Indeed, $\mathcal{E}(X+a) = \mathcal{E}(X)$, hence the shift does not change entropy.

More generally, if $(\Omega, \mathcal{F}, \nu)$ is a measure space, then for a probability measure μ on $\mathcal{F}$, we define its entropy (w.r.t. ν) as $\int_\Omega f \log \frac{1}{f} d\nu$ if $f = \frac{d\mu}{d\nu}$ and as $-\infty$ if μ is not absolutely continuous to μ. The entropy maximizing measures below are called Gibbs measures.

Exercise 20. (Optional) Let $H : \Omega \mapsto \mathbb{R}$ be a function such that $\psi(\beta) := \int e^{\beta x} d\nu(x) < \infty$ (which could be ∞). Consider the problem of maximizing the entropy of a probability measure on Ω subject to the constraint $\int_\Omega H d\mu = E_0$. If there exists a β_0 such that $\psi(\beta_0) = E_0$, then show that the entropy maximizing measure has density proportional to $e^{\beta H}$ with respect to ν.

Apply this to the function $H(x) = x$ on $\Omega = \mathbb{R}_+$, to $H(x) = 1$ on $\Omega = [a,b]$ and to $H(x) = x^2$ on $\Omega = \mathbb{R}$ to recover familiar examples of probability distributions.

1.0.11 Gaussian from uniform measures on high dimensional spheres

Let μ_n denote the uniform measure on $\mathbb{S}^n = \{\mathbf{x} \in \mathbb{R}^{n+1} : \|\mathbf{x}\|^2 = 1\}$. By definition, this is the unique probability measure on $\mathbb{S}^n$ that is invariant under orthogonal transformations. One way to see that it exists is to consider the standard Gaussian measure γ_{n+1} on $\mathbb{R}^{n+1}$ and push it forward under the mapping $\mathbf{x} \mapsto \mathbf{x}/\|\mathbf{x}\|$ from $\mathbb{R}^{n+1} \setminus \{0\}$ to $\mathbb{S}^n$. The orthogonal invariance of γ_{n+1} carries over to μ_n.

Exercise 21. Show that $|V_1^{(n)}|^2$ has Beta$(n-1,1)$ distribution. More generally, $(|V_1|^2, \ldots, |V_k|^2)$ has Dirichlet$(k; 1, 1, \ldots, 1, n-k)$ distribution. (Recall that Dirichlet$(k; a_1, \ldots, a_k, a_{k+1})$ distribution is the one with density $C x_1^{a_1-1} \ldots x_k^{a_k-1}(1 - x_1 - \ldots - x_k)^{a_{k+1}-1}$ on $\{(x_1, \ldots, x_k) \in \mathbb{R}_+^k : x_1 + \ldots + x_k < 1\}$. The normalization constant $C = \frac{\Gamma(a_1+\ldots+a_{k+1})}{\Gamma(a_1)\ldots\Gamma(a_{k+1})}$.)

Thus, the Gaussian measure is a convenient tool to do calculations on the uniform measure on spheres. Sometimes, the usefulness may be in the reverse direction, using the following way of getting to Gaussian from uniform.

Claim: Let $V^{(n)} = (V_1^{(n)}, \ldots, V_n^{(n)}) \sim \mu_n$. Fix $k \geq 1$. Then $\sqrt{n}(V_1^{(n)}, \ldots, V_n^{(k)})$ converges in distribution to a vector of i.i.d. standard Gaussians, as $n \to \infty$.

The easiest proof is from the construction used above for $V^{(n)}$. Indeed, let X_i be i.i.d. standard Gaussians. Then

$$V^{(n)} = \frac{(X_1, \ldots, X_n)}{\sqrt{X_1^2 + \ldots + X_n^2}}$$

is uniform on S^{n-1}. By the law of large numbers $\frac{X_1^2+\ldots+X_n^2}{n} \xrightarrow{a.s.} 1$. Therefore, $\sqrt{n}(V_1^{(n)}, \ldots, V_k^{(n)})$ converges to $(X_1, \ldots, X_k)$ almost surely, and hence in distribution. ■

Remark 22. Apparently, Maxwell arrived at the above fact in his study of the velocity distribution of molecules in a gas. The idea is that a box of (monatomic) gas in a box maintained at a fixed temperature has about $N = 10^{23}$ atoms, and the temperature is (a multiple of) the sum of squares of the velocity components of the atoms. Basic principles of statistical mechanics assert that the individual velocity components are random and uniformly distributed on the set of all allowed values, in this case, a sphere in $3N$ dimensions. Maxwell then deduced that the three velocity components of any single atom (say randomly chosen atom) are i.i.d. Gaussian.

1.0.12 Gaussian Hilbert space

If $(\Omega, \mathcal{F}, \mathbf{P})$ is a probability space, then $L^2(\mathbf{P})$ is a Hilbert space with inner product $\langle X, Y\rangle := \mathbf{E}[XY]$ (of course X, Y are to be interpreted as equivalence classes). Any closed subspace $\mathcal{H} \subseteq L^2(\mathbf{P})$ is also a Hilber space. If every $X \in \mathcal{H}$ has a univariate Gaussian distribution, then we say that $\mathcal{H}$ is a Gaussian Hilbert space.

If $(X_i)_{i\in I}$ is any collection of jointly Gaussian random variables, then $W = \overline{\text{span}}\{X_i : i \in I\}$ is a GHS. The closure is in $L^2(\mathbf{P})$. The reason is that linear combinations are Gaussian and L^2-limits of Gaussians are also Gaussian (as L^2-convergence implies convergence in distribution). In this sense, when studying a jointly Gaussian collection, there is no reason to not enlarge the collection to a closed subspace of $L^2(\mathbf{P})$. That is called the GHS associated to the original collection.

For example, suppose $X \sim N_m(0, \Sigma)$ be a random vector in some probability space $(\Omega, \mathcal{F}, \mathbf{P})$. Then, $W = \{X_{\mathbf{v}} := \mathbf{v}^t X : \mathbf{v} \in \mathbb{R}^m\}$. By $\mathbf{v} \mapsto X_{\mathbf{v}}\}$, this is isomorphic to the Hilbert space $\mathbb{R}^m$ with the inner product $\langle \mathbf{v}, \mathbf{w}\rangle = \mathbf{v}^t \Sigma \mathbf{w}$. Of course, if Σ is singular, then this is a pseudo inner product (not all linear combinations are distinct).

Conversely, given a Hilbert space H with orthonormal basis $\{e_i : i \in I\}$ (no big loss to assume that I is countable), then we can form a collection of Gaussians whose GHS is (isomorphic to) H. For this, consider a probability space $(\Omega, \mathcal{F}, \mathbf{P})$ on which there are i.i.d. standard Gaussians Z_i. For any $\mathbf{v} \in H$, define $X_{\mathbf{v}} = \sum_i \langle \mathbf{v}, e_i\rangle Z_i$. The series converges by Khinchine's theorem since $\sum_i |\langle \mathbf{v}, e_i\rangle|^2 = \|\mathbf{v}\|^2$ is finite. As $\mathbf{E}[X_{\mathbf{v}} X_{\mathbf{u}}] = \langle \mathbf{v}, \mathbf{u}\rangle$, the GHS $\{X_{\mathbf{v}} : \mathbf{v} \in H\} \subseteq L^2(\mathbf{P})$ is isomorphic to H under the bijection $\mathbf{v} \leftrightarrow X_{\mathbf{v}}$, proving the claim[2].

1.0.13 The Gaussian density and the heat equation

Recall the standard Gaussian density $\varphi(x)$. The corresponding cumulative distribution function is denoted by Φ and the tail is denoted by $\bar{\Phi}(x) := \int_x^\infty \varphi(t)dt$. The following estimate will be used very often.

Exercise 23. For all $x > 0$, we have $\frac{1}{\sqrt{2\pi}} \frac{x}{1+x^2} e^{-\frac{1}{2}x^2} \leq \bar{\Phi}(x) \leq \frac{1}{\sqrt{2\pi}} \frac{1}{x} e^{-\frac{1}{2}x^2}$ In particular[3], $\bar{\Phi}(x) \sim$

[2]This may seem fairly pointless, but here is one thought-provoking question. Given a vector space of Gaussian random variables, we can multiply any two of them and thus get a larger vector space spanned by the given Gaussian random variables and all pair-wise products of them. What does this new vector space correspond to in terms of H?

[3]The notation $f(x) \sim g(x)$ means that $\lim_{x\to\infty} \frac{f(x)}{g(x)} = 1$.

$x^{-1}\varphi(x)$ as $x \to \infty$. Most often the following simpler bound, valid for $x \geq 1$, suffices.

$$\frac{1}{10x} e^{-\frac{1}{2}x^2} \leq \bar{\Phi}(x) \leq e^{-\frac{1}{2}x^2}.$$

For $t > 0$, let $p_t(x) := \frac{1}{\sqrt{t}}\varphi(x/\sqrt{t})$ be the $N(0,t)$ density. We interpret $p_0(x)dx$ as the degenerate measure at 0. These densities have the following interesting properties.

Exercise 24. Show that $p_t \star p_s = p_{t+s}$, i.e., $\int_{\mathbb{R}} p_t(x-y)p_s(y)dy = p_{t+s}(x)$.

Exercise 25. Show that $p_t(x)$ satisfies the heat equation: $\frac{\partial}{\partial t}p_t(x) = \frac{1}{2}\frac{\partial^2}{\partial x^2}p_t(x)$ for all $t > 0$ and $x \in \mathbb{R}$.

Remark 26. Put together, these facts say that $p_t(x)$ is the *fundamental solution* to the heat equation. This just means that the heat equation $\frac{\partial}{\partial t}u(t,x) = \frac{1}{2}\frac{\partial^2}{\partial x^2}u(t,x)$ with the initial condition $u(0,x) = f(x)$ can be solved simply as $u(t,x) = (f \star p_t)(x) := \int_{\mathbb{R}} f(y)p_t(x-y)dy$. This works for reasonable f (say $f \in L^1(\mathbb{R})$).

All this generalized to higher dimensions. Write $p_{d,t}(x)dx$ for the measure $N_d(0,tI_d)$ where $t = 0$ corresponds to δ_0. Then $p_{d,t} \star p_{d,s} = p_{d,t+s}$. Further, the heat equation

$$\frac{\partial}{\partial t}p_t(x) = \frac{1}{2}\Delta p_t(x)$$

where $\Delta = \sum_{i=1}^d \frac{\partial^2}{\partial x_i^2}$ is the Laplacian on $\mathbb{R}^d$. Then $u(x,t) := \int_{\mathbb{R}^d} f(y)p_{d,t}(x-y)$ satisfies the heat equation with the initial condition $u(\cdot,0) = f$ (for reasonable f, where we omit the discussion of what is reasonable).

One can rewrite the heat equation in the following way. Let $d = 1$ for simplicity. Let $Z_t \sim N(0,t)$ have density p_t. Then $Z_t/\sqrt{t}$ is a standard Gaussian for each t, and hence for any reasonable test function, $\mathbf{E}[f(Z_t/\sqrt{t})]$ is constant. If f is smooth enough, we can differentiate under the integral to get

$$\begin{aligned}
0 &= \frac{d}{dt}\int_{\mathbb{R}} f(z/\sqrt{t})p_t(z)dz \\
&= \int_{\mathbb{R}} \left[\frac{-z}{2t^{\frac{3}{2}}}f'(z/\sqrt{t}) + \frac{1}{2}f(z/\sqrt{t})p_t''(z)\right]dz \\
&= \int_{\mathbb{R}} \left[\frac{-z}{2t^{\frac{3}{2}}}f'(z/\sqrt{t}) + \frac{1}{2t}f''(z/\sqrt{t})\right]p_t(z)dz
\end{aligned}$$

by integration by parts in the second integral. Set $t = 1$ to get $\mathbf{E}[Zf'(Z)] = \mathbf{E}[f''(Z)]$. This is true for all smooth enough f for which the expectations exist.

Remark 27. One choice of Z_t is to let Z be a standard Brownian motion in one dimension, with a minor change of variables, $X_t = e^{-t/2}Z_{e^t}$ is an Ornstein-Uhlenbeck process which is stationary in time. Then the equation obtained here has is just the Komogorov forward equation for this process.

1.0.14 Stein's equation

If $Z \sim N(0,1)$, we obtained the equation $\mathbf{E}[Zf'(Z)] = \mathbf{E}[f''(Z)]$. Writing $g = f'$, this becomes $\mathbf{E}[Zg(Z)] = \mathbf{E}[g'(Z)]$, which involves only first order derivatives. This is called Stein's equation. Suppose a random variable W satisfies

$$\mathbf{E}[Wg(W)] = \sigma^2\mathbf{E}[g'(W)]$$

for a large class of g (say all $g \in C^1$ for which both expectations exist), then taking the special functions $g(w) = e^{i\lambda w}$, we see that W has the characteristic function ψ satisfying $\psi'(\lambda) = -\sigma^2\psi(\lambda)$, which implies that $W \sim N(0,\sigma^2)$. Thus the Stein-equation characterizes the Gaussian distribution.

Stein made this the starting point of what is now famous as *Stein's method*, showing that if W satisfies the Stein's equation approximately, then the distribution of W is approximately Gaussian. This allowed him to prove central limit theorems in many situations that were out of reach of the method of characteristic functions (for example, if $W = (X_1 + \ldots + X_n)/\sqrt{n}$ where X_i are weakly dependent variables with zero mean and unit variance).

In higher dimensions, we have the following "integration by parts" formula that we shall have occasion to use.

Exercise 28. Let $X \sim N_n(0,\Sigma)$ and let $F : \mathbb{R}^n \to \mathbb{R}$. Under suitable conditions on F (state sufficient conditions), show that $\mathbf{E}[X_iF(X)] = \sum_{j=1}^n \sigma_{ij}\mathbf{E}[\partial_j F(X)]$. As a corollary, deduce the Wick formula of Exercise 10.

1.0.15 Semigroups

Recall the Ornstein-Uhlenbeck process that is defined as $X(t) = e^{-t/2}W(e^t)$, for $-\infty < t < \infty$, where W is a standard Brownian motion in d-dimensions. Then $X = (X_1, \ldots, X_d)$ where X_i are i.i.d. O-U processes in one dimensions. Let us take $d = 1$ for now.

The O-U process X is clearly a Gaussian process and $\mathbf{E}[X(t)] = 0$ and $\mathbf{E}[X(t), X(s)] = e^{-|t-s|}$. This shows that X is stationary, that is $X(\cdot - \tau) \stackrel{d}{=} X(\cdot)$ for any $\tau \in \mathbb{R}$. There are alternate descriptions of X, for example by the stochastic differential equation

$$dX(t) = -\frac{1}{2}X(t)dt + dB(t)$$

where B is standard Brownian motion. The meaning of this is that

$$X(t) - X(0) = -\frac{1}{2}\int_0^t X(s)ds + B(t).$$

In more intuitive terms, for small $h > 0$, the displacement $X(t+h) - X(t)$ is independent of $\{X(s) : s \leq t\}$ and has (approximately) Gaussian distribution with mean equal to $-\frac{1}{2}X(t)h$ and variance equal to h. If the mean was zero, this would just be the description of standard Brownian motion. However, the mean is $-\frac{1}{2}X(t)$, which means that it tends to move towards the origin.

Another way to think of it is to drop the Brownian motion term (if you don't like revolutionary changes, consider $dX(t) = -\frac{1}{2}X(t) + \alpha dB(t)$ and drive the parameter α to 0). We get the ODE $x'(t) = -\frac{1}{2}x(t)$, which means that x is the velocity of a particle performing simple harmonic motion (a mass tied to a spring that pulls it towards the origin). In that sense, O-U process is a particle performing simple harmonic motion, except that there is some randomness ("thermal noise") in its motion given by the Brownian motion.

The O-U process is a Markov process. To see its transition density, observe that for fixed $t > 0$, the random variable $X(t)$ has the same distribution as $e^{-\frac{t}{2}}X(0) + \sqrt{1-e^{-t}}Z$, where $Z \sim N(0,1)$ is independent of $X(0)$ (independent of the entire past up to time 0, in fact). In other words, the conditional distribution of $X(t)$ given $\mathcal{F}_0 = \sigma\{X(s) : s \leq 0\}$ is $N(e^{-t/2}X(0), 1-e^{-t})$. For a reasonable function $f : \mathbb{R} \mapsto \mathbb{R}$, let $(P_t f)(x) = \mathbf{E}[f(X_t) \,\big|\, X(0) = x]$. The precise class of f for which this makes sense will be left open now, but ignoring that point, it is easy to see that $P_t \circ P_s = P_{t+s}$ and of course $P_0 = I$. Thus, $(P_t)_{t \geq 0}$ form a semigroup (meaning that $P_t \circ P_s = P_{t+s}$). This is called the O-U semigroup.

1.0.16 Hermite polynomials

Hermite polynomials are useful, in fact indispensable, tools to study the Gaussian measure. For $n \geq 0$, the$n$th Hermite polynomial is defined as[4]

$$H_n(x) = (-1)^n \frac{d^n}{dx^n} e^{-\frac{1}{2}x^2}.$$

It is clear from this that H_n is a monic polynomial of degree n. For example, $H_0(x) = 1$, $H_1(x) = x$, $H_2(x) = x^2 - 1$, $H_3(x) = x^3 - 3x$. It is possible to write the explicit formula for H_n, but that is not the best way to understand its properties. To this end, observe that

$$e^{-\frac{1}{2}x^2} \sum_{n=0}^{\infty} H_n(x) \frac{w^n}{n!} = e^{-\frac{1}{2}(x-w)^2}. \tag{1}$$

This can be seen by writing the power series expansion of the right hand side as a function of w, with x held fixed (donÕt expand the exponential, just recall that the nth coefficient is got by differentiating the function n times at the origin...). Multiply two of these expressions with w and z, multiply by $\frac{1}{\sqrt{2\pi}} e^{\frac{1}{2}x^2}$, integrate and interchange integral with summation to get

$$\begin{aligned}
\sum_{m,n \geq 0} \langle H_n, H_m \rangle_\gamma \frac{w^n z^m}{n!m!} &= \frac{1}{\sqrt{2\pi}} \int_{\mathbb{R}} e^{\frac{1}{2}x^2} e^{-\frac{1}{2}(x-w)^2} e^{-\frac{1}{2}(x-z)^2} \, dx \\
&= e^{zw} \frac{1}{\sqrt{2\pi}} \int_{\mathbb{R}} e^{-\frac{1}{2}(x-(w+z))^2} \, dx \\
&= \sum_{p \geq 0} \frac{w^p z^p}{p!}.
\end{aligned}$$

Equating the coefficients, we see that $\langle H_n, H_m \rangle_\gamma = \delta_{m,n} n!$. Thus $\frac{1}{\sqrt{n!}} H_n$, $n \geq 0$, form an orthonormal basis for $L^2(\gamma)$ (since polynomials are dense in $L^2(\gamma)$). As such, applying Gram-Schmidt procedure to $1, x, x^2, \ldots$ in $L^2(\gamma)$ would have led to the same polynomials. Hermite polynomials satisfy difference relations and differential equations.

Exercise 29. Show that check constants here

1. $H_{n+1}(x) = xH_n(x) - nH_{n-1}(x)$.

[4]There are different conventions. Wikipedia denotes our H_n as He_n, and calls it the "probabilistÕs Hermite polynomials". The notation H_n is used for the "physicistÕs Hermite polynomials" that are defined without the $\frac{1}{2}$ in the exponent of the Gaussian density.

2. $\frac{d}{dx}H_n(x) = H_{n-1}(x)$ and $(-\frac{d}{dx}+x)H_n(x) = H_{n+1}(x)$.

3. $\frac{d}{dx}(-\frac{d}{dx}+x)H_n(x) = H_n(x)$.

As$\{\frac{1}{\sqrt{n!}}H_n : n \geq 0\}$ form an orthonormal basis for $L^2(\gamma)$, for any $f : \mathbb{R} \mapsto \mathbb{R}$ such that $\mathbf{E}[f(Z)^2] < \infty$, we can write the Hermite expansion of f,

$$f(Z) \overset{L^2(\gamma)}{=} \sum_{n=0}^{\infty} \frac{1}{n!}\langle f, H_n\rangle_\gamma \, H_n(Z).$$

For specific functions f, the convergence can be in better senses of course (if the coefficients $\langle f, H_n\rangle$ decay fast). This can be useful in many ways, for example to compute moments of $f(Z)$. For example, if X_i are jointly Gaussian (centered) and have unit variances, then if f_i are nice enough, then

$$\mathbf{E}[f_1(X_1)\dots f_k(X_k)] = \sum_{n_1,\dots,n_k \geq 0} \prod_{j=1}^{k} \frac{\langle f_j, H_{n_j}\rangle_\gamma}{n_j!} \, \mathbf{E}[H_{n_1}(X_1)\dots H_{n_k}(X_k)].$$

Thus, in principle, all we need are expectations of products of Hermite polynomials. While the Wick formulas can in principle be used to compute expectations of products of any polynomials of Gaussians, the formulas get way too complicated. For Hermite polynomials, everything is magically simpler. Let $X \sim N_k(0,\Sigma)$ where $\sigma_{i,i} = 1$ for all i. Then each X_i is a standard Gaussian, hence (1) is applicable and we can write,

$$\begin{aligned}\sum_{n_1,\dots,n_k \geq 0} \mathbf{E}[H_{n_1}(X_1)\dots H_{n_k}(X_k)]\frac{w_1^{n_1}\dots w_k^{n_k}}{n_1!\dots n_k!} &= \int_{\mathbb{R}^k} e^{\sum_{j=1}^k x_j w_j - \frac{1}{2}\sum_{j=1}^k w_j^2} \frac{1}{(2\pi)^{k/2}\sqrt{\det(\Sigma)}} e^{-\frac{1}{2}\mathbf{x}^t\Sigma^{-1}\mathbf{x}}\, d\mathbf{x} \\ &= e^{-\frac{1}{2}(w_1^2+\dots+w_k^2)} e^{\frac{1}{2}\mathbf{w}^t\Sigma\mathbf{w}} \\ &= e^{\frac{1}{2}\mathbf{w}^t(\Sigma - I_k)\mathbf{w}}.\end{aligned}$$

From this, one can get the formula for $\mathbf{E}[H_{n_1}(X_1)\dots H_{n_k}(X_k)]$ as follows.

Exercise 30. Consider the complete k-partite graph with vertex set $V_1 \sqcup V_2 \sqcup \dots \sqcup V_k$, where V_i is a set of cardinality n_i. This means that there is an edge between any two vertices not belonging to the same V_j. Endow any edge connecting a vertex in V_i to a vertex in V_j, with $i \neq j$, with weight $\sigma_{i,j}$. Then, with the setting above,

$$\mathbf{E}[H_{n_1}(X_1)\dots H_{n_k}(X_k)] = \frac{n_1!\dots n_k!}{(n_1+\dots+n_k)!}\sum_M w(M)$$

where the sum is over all complete matchings of the graph, and the weight of a matching is the product of the weights of its edges.

In particular, if $X_1 = X_2 = \ldots = X_k = Z$ then, $\frac{(n_1+\ldots+n_k)!}{n_1!\ldots n_k!}\mathbf{E}[H_{n_1}(Z)\ldots H_{n_k}(Z)]$ is equal to the number of complete matchings of the graph.

1.0.17 Tail of the Gaussian distribution

The cumulative distribution function $\Phi(t) = \int_{-\infty}^{t} \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}x^2} dx$ has no closed form expression in terms of elementary functions (polynomials, trigonometric and exponential functions), hence the following estimate is often useful. Let $\overline{\Phi}(t) = 1 - \Phi(t)$ denote the tail.

Exercise 31. For any $t > 0$,

$$\frac{1}{t+\frac{1}{t}} \frac{e^{-\frac{1}{2}t^2}}{\sqrt{2\pi}} \leq \overline{\Phi}(t) \leq \frac{1}{t} \frac{e^{-\frac{1}{2}t^2}}{\sqrt{2\pi}}.$$

In particular, $\overline{\Phi}(t) \sim \frac{1}{t}\varphi(t)$ as $t \to \infty$. It is often more convenient to use simpler inequalities like $\overline{\Phi}(t) \leq e^{-\frac{1}{2}t^2}$ for $t \geq 1$ or $\overline{\Phi}(t) \leq e^{1-\frac{1}{2}t^2}$ for all $t > 0$.

Exact inequalities that hold for every t are sometimes convenient. But as $t \to \infty$, one can get better approximations by the asymptotic expansion:

$$\overline{\Phi}(t) = \frac{e^{-\frac{1}{2}t^2}}{t\sqrt{2\pi}} \sum_{n=0}^{\infty} (-1)^n \frac{(2n-1)!!}{t^{2n}}$$

with the convention that $(-1)!! = 1$ and $(2n-1)!! = (2n-1) \times (2n-3) \times \ldots 3 \times 1$ for $n \geq 1$. For any $t \in \mathbb{R}$, the series on the right diverges! The meaning of asymptotic expansion is that for any $p \geq 1$, as $t \to \infty$,

$$\overline{\Phi}(t) - \frac{e^{-\frac{1}{2}t^2}}{t\sqrt{2\pi}} \sum_{n=0}^{p} (-1)^n \frac{(2n-1)!!}{t^{2n}} = O(t^{-2p-2} e^{-\frac{1}{2}t^2}).$$

1.0.18 Maximum of i.i.d. Gaussians

Let Z_i be i.i.d. standard Gaussians and let $M_n = \max\{Z_1, \ldots, Z_n\}$. How big is M_n? This is a question that recurs in many situations and an understanding is quite important. We shall show that

$$\frac{M_n}{\sqrt{2\log n}} \xrightarrow{a.s.} 1. \tag{2}$$

In fact we can get sharper statements along the same lines, but this can serve as a first goal.

Where does the $\sqrt{2\log n}$ come from? Since there are n independent variables, if b_n is chosen so that $\overline{\Phi}(b_n) = \frac{1}{n}$, then the expected number of Z_is that are more than b_n is exactly 1. This suggests that the maximum is of the order of b_n. For the Gaussian distribution, using $\overline{\Phi}(t) \sim \frac{1}{t\sqrt{2\pi}} e^{-\frac{1}{2}t^2}$, as $t \to \infty$, it follows that $b_n \sim \sqrt{2\log n}$.

Exercise 32. Show that $\sqrt{2\log n} + \frac{1}{2}\log\log n$ is a better approximation for b_n.

To prove (**??**), fix $\delta > 0$ and consider the events $\{M_n > (1+\delta)\sqrt{2\log n}\}$. By the tail estimate for the Gaussian distribution, for large enough n (so that $(1+\delta)\sqrt{2\log n} \geq 1$), its probability is at most $\exp\{-\frac{1}{2}(1+\delta)^2 2\log n\} = n^{-(1+\delta)^2}$, which is summable over n. By Borel-Cantelli lemma, only finitely many of these events occur, and hence $\limsup_n \frac{M_n}{\sqrt{2\log n}} \leq 1+\delta$, almost surely. Take intersection over countably many δ approaching 0 from above, we see that $\limsup_n \frac{M_n}{\sqrt{2\log n}} \leq 1$, almost surely.

To get the lower bound, let $\delta > 0$ and consider the event $Z_n > (1-\delta)\sqrt{2\log n}$. By the lower bound for the Gaussian tail,

$$\mathbf{P}\{Z_n > (1-\delta)\sqrt{2\log n}\} \sim \frac{1}{(1-\delta)\sqrt{2\log n}} e^{-(1-\delta)^2 \log n} \asymp \frac{1}{n^{(1-\delta)^2}\sqrt{\log n}}$$

which is not summable. As Z_n are independent, by Borel-Cantelli lemma, infinitely many of these events occur, almost surely. But $M_n \geq Z_n$, hence we see that $\limsup_n \frac{M_n}{\sqrt{2\log n}} \geq 1-\delta$, almost surely. Take intersection over countably many δ approaching 0 from above to see that $\limsup_n \frac{M_n}{\sqrt{2\log n}} \geq 1$, almost surely.

Putting together the upper and lower bounds, we have proved (**??**).

One can make more accurate estimates (we give up on "almost sure" and consider "in probability" below).

Exercise 33. Let Z_n be i.i.d. standard Gaussians. Show that $\sqrt{\log n}(M_n - \sqrt{2\log n})$ is tight. In other words, $\mathbf{P}\{-h_n < \sqrt{\log n}(M_n - \sqrt{2\log n}) < h_n\} \to 1$ for any sequence $h_n \to \infty$.

In any natural situation, if a sequence of random variables is tight, one should ask if it converges in distribution. In the current setting, it is true. For any $u \in \mathbb{R}$, we have

$$\mathbf{P}\{\sqrt{2\log n}(M_n - \sqrt{2\log n}) \leq u\} \to e^{-e^{-u}}. \tag{3}$$

The limiting distribution is known as Gumbel distribution. To prove the above claim, let us define b_n as before, satisfying $\overline{\Phi}(b_n) = \frac{1}{n}$. Consider

$$\begin{aligned}\mathbf{P}\{b_n(M_n - b_n) \le u\} &= \mathbf{P}\{M_n \le b_n + \frac{u}{b_n}\} \\ &= \left(1 - \overline{\Phi}(b_n + \frac{u}{b_n})\right)^n.\end{aligned}$$

From the fact that $\overline{\Phi}(t) \sim \frac{1}{t\sqrt{2\pi}} e^{-\frac{1}{2}t^2}$ as $t \to \infty$, we see that

$$\overline{\Phi}(b_n + \frac{u}{b_n}) \sim \overline{\Phi}(b_n) e^{-u} \;=\; \frac{1}{n} e^{-u}.$$

Therefore, $\mathbf{P}\{b_n(M_n - b_n) \le u\} = (1 - \frac{1+o(1)}{n} e^{-u})^n \to e^{-e^{-u}}$. Since $b_n \sim \sqrt{2\log n}$, the convergence in distribution holds with $\sqrt{2\log n}$ in place of b_n, completing the proof of (3).

Another perspective: Motivated by a discussion in class, let us write this in a different way. For $b(p) = \overline{\Phi}^{-1}(p)$ denote the inverse of the tail CDF. What we denoted b_n above is in this notation $b(1/n)$. The number of $Z_1, \ldots, Z_n$ that fall above $b(p)$ has $\mathrm{Bin}(n, p)$ distribution. Hence, for any fixed $\lambda > 0$

$$\#\{k \le n : Z_k > b(\lambda/n)\} \xrightarrow{d} \mathrm{Pois}(\lambda).$$

Remark 34. With a little more care, one can say make the following statement: Let $N_n(\lambda, \infty) = \#\{k : Z_k > b(\lambda/n)\}$. Then N_n converges in distribution to Poisson point process with intensity 1 on $\mathbb{R}$. To make this statement precise, one needs to understand the meaning of convergence in distribution for point processes.

Since the number on the left is zero if and only if $M_n < b(\lambda/n)$, we see that $\mathbf{P}\{M_n < b(\lambda/n)\} \to e^{-\lambda}$. From the Gaussian tail, one can work out that $b(p) \sim 2\log\frac{1}{p}$ (and more accurately $b(p) = 2\log\frac{1}{p} - 2\log\log\frac{1}{p} + \ldots$), from which it follows that $\frac{M_n}{\sqrt{2\log n}} \xrightarrow{P} 1$.

1.0.19 Complex Gaussians

By a complex Gaussian random variable, we mean a $\mathbb{C}$-valued random variable ζ with density $\frac{1}{\pi} e^{-|z|^2}$ on the complex plane. In other words, $\mathrm{Re}\,\zeta$ and $\mathrm{Im}\,\zeta$ are i.i.d. $N(0, \frac{1}{2})$ random variables. The distribution of $\sigma\zeta + a$ is denote $N_{\mathbb{C}}(a, \sigma^2)$ for $a \in \mathbb{C}$ and $\sigma > 0$.

We say that a collection of complex-valued random variables $\zeta = (\zeta_t)_{t\in T}$ has a joint complex Gaussian distribution if any finite linear combination of them has a univariate complex Gaussian distribution. The distribution of ζ is determined by the mean function $m(t) = \mathbf{E}[\zeta(t)]$ and covariance kernel $K(t,s) = \mathbf{E}[\zeta(t)\overline{\zeta(s)}]$. It should be noted that $\mathbf{E}[\zeta(t)\zeta(s)] = 0$ for all $t,s \in T$ (including $t = s$). The matrix $(K(t,s))_{t,s\in T}$ is called the covariance matrix of ζ.

Almost all the things said so far for real Gaussians can be carried out for these. We just state a few salient properties below and leave them as exercises. One can prove them either by carrying out the analogy with the real Gaussian case, or by breaking all the complex variables into their real and imaginary parts, and using the results for the real Gaussian case. We prefer the first approach. In fact, in many ways, the standard complex Gaussian is more natural than the real Gaussian. When proving that $\int_{\mathbb{R}} e^{-\frac{1}{2}x^2} dx = \sqrt{2\pi}$ or when simulating a Gaussian on a computer, we know that it is easier done by considering a complex Gaussian. The fundamental reason is that the standard complex Gaussian has independent components in two co-ordinate systems: Cartesian and polar.

a few things to fill here

Chapter 2

Examples of Gaussian processes

To define a Gaussian process $X = (X(t))_{t\in T}$ is the same as to specify the mean function $m : T \mapsto \mathbb{R}$ and covariance function $K : T \times T \mapsto \mathbb{R}$. The role of the mean function is trivial: If X is a process with mean function 0 then $X + m$ has mean function m. Thus it suffices to understand centered (zero mean) processes, and all our processes will be centered unless stated otherwise. The covariance kernel has to be positive semi-definite, and that is something nontrivial. How to check if a given function K is p.s.d.? How to generate examples of p.s.d. kernels? How to generate all of them?

2.1 Random series

Example 1. Let $T = \{1,2,\ldots,n\}$ be a finite set. Then $K = (K_{i,j})_{1\le i,j\le n}$ is a matrix. We know that K is p.s.d. if and only if it is a Gram matrix, i.e., $K_{i,j} = \langle \mathbf{v}_i, \mathbf{v}_j\rangle$ for some vectors $\mathbf{v}_1,\ldots,\mathbf{v}_n \in \mathbb{R}^n$. This gives a characterization of all p.s.d. matrices. The corresponding Gaussian process is easy to generate: Let $Z \sim \gamma_n$ and set $X_i = \langle Z, \mathbf{v}_i\rangle$.

An alternate way to say the same is to write $K = \mathbf{u}_1\mathbf{u}_1^t + \ldots + \mathbf{u}_n\mathbf{u}_n^t$ for some $\mathbf{u}_i \in \mathbb{R}^n$. Then set $X = Z_1\mathbf{u}_1 + \ldots + Z_n\mathbf{u}_n$, where Z_i are i.i.d. standard Gaussians.

One direction of this works for arbitrary T. If $f_i : T \mapsto \mathbb{R}$ are arbitrary functions, then the function $K(t,s) = f_1(t)f_1(s) + \ldots + f_n(t)f_n(s)$ for $t,s \in T$, is the covariance kernel of the process $X(t) = Z_1 f_1(t) + \ldots + Z_n f_n(t)$, where Z_i are i.i.d. standard Gaussians.

Does it work with infinitely many functions?

Example 2. Let $f_i : T \mapsto \mathbb{R}$, $i \ge 1$, be a collection of functions such that $\sum_i |f_i(t)|^2 < \infty$ for all $t \in T$. Then $K(t,s) := \sum_i f_i(t)f_i(s)$, $t,s \in T$, is a well-defined p.s.d. kernel. It is in

fact the covariance kernel of the Gaussian process $X(t) := \sum_i Z_i f_i(t)$, $t \in T$, where Z_i are i.i.d. standard Gaussians. Observe that to even define $K(t,t)$, the condition on square summability of $(f_i(t))_{i\geq 1}$ is necessary, and then, $K(t,s)$ is well-defined as it is the inner product in ℓ^2 of the sequences $(f_i(t))_{i\geq 1}$ and $(f_i(s))_{i\geq 1}$.

Indeed, for each $t \in T$, by Khinchine's theorem, the series defining $X(t)$ converges almost surely. As a limit of finite linear combinations of Gaussians, it is Gaussian, and $\mathbf{E}[X(t)X(s)] = \sum_{i,j} f_i(t)f_j(s)\mathbf{E}[Z_iZ_j] = \sum_i f_i(t)f_i(s)$.

With some caveats, all positive semi-definite kernels are of the above form. For now we do not need to worry why, we just mention a few special cases of these examples.

1. $T = \mathbb{R}$ and $f_i(t) = t^i$ for $i = 0, 1, \ldots, n$. Then $X(t) = Z_0 + Z_1 t + \ldots + Z_n t^n$ is a random polynomial. We may also take $T = \mathbb{C}$, but then f_i are not real-valued, hence we should say that $(\mathrm{Re}\, X(z), \mathrm{Im}\, X(z))_{z\in\mathbb{C}}$ is a Gaussian process.

2. Let $f_i(t) = c_i t^i$, $i \geq 0$. Then $X(t) = \sum_{i\geq 0} Z_i c_i t^i$ and $K(t,s) = \sum_{i\geq 0} c_i^2 (ts)^i$. The domain T depends on c_is. For example if $c_i = 1$ for all i, then $\sum_i f_i(t)^2$ converges for $t \in (-1,1)$, hence X is a Gaussian process on $T = (-1,1)$ having covariance kernel $\frac{1}{1-ts}$. If $c_i = 1/\sqrt{i!}$, then $T = \mathbb{R}$ and $K(t,s) = e^{ts}$. If $c_i = i!$ is fast growing, then $T = \{0\}$, as the square summability condition fails for all $t \neq 0$.

Similarly, one can form random trigonometric polynomials $a_0 Z_0 + \sum_{k=1}^n (a_k Z_k \cos(2\pi kt) + b_k Z_k' \sin(2\pi kt))$, where a_i, b_i are fixed real numbers. Let us see an example that is not obviously in this form.

Example 3. Let $T = [0,\infty)$ and $K(t,s) = t \wedge s$. To see that it is p.s.d., we may observe that $K(t,s) = \langle \mathbf{1}_{[0,t]}, \mathbf{1}_{[0,s]} \rangle$, showing that K is a Gram matrix (or to be pedantic, $(K(t_i,t_j))_{i,j\leq n}$ is a Gram matrix for any $n \geq 1$ and $t_i \in T$). Thus there exists a Gaussian process with this covariance. However, that is a random variable taking values in $\mathbb{R}^T$ with its cylinder sigma-algebra. It can be proved that there is a Gaussian process with the same mean and covariance and for which the sample paths are $C(T)$-valued. That is what is called Brownian motion.

As in this example, existence of a version of the Gaussian process with continuous sample paths is a fundamental question that we shall discuss in some detail later.

Our next class of examples is a natural progression from summation to integration.

Exercise 4. Let Θ be some index set carrying a sigma algebra and a measure ρ. Let $f : \Theta \times T \mapsto \mathbb{R}$ be a function that is measurable and square integrable in the first co-ordinate for each fixed $t \in T$. Define $K : T \times T \mapsto \mathbb{R}$ by $K(t,s) = \int_\Theta f(\theta,t) f(\theta,s) d\rho(\theta)$. Show that a Gaussian process with this covariance exists. Show that the same holds if f is complex-valued, provided $K(t,s) := \int_\Theta f(\theta,t)\overline{f(\theta,s)} d\rho(\theta)$ is real-valued.

Unlike in the case of finite sums, it is not clear how to write this process in terms of independent Gaussians.

Example 5. Let $T = [0,1]^2$ and let $K((t_1,t_2),(s_1,s_2)) = (t_1 \wedge s_1) + (t_2 \wedge s_2)$. The corresponding Gaussian process can again be realized as a random continuous function and will thereafter be called *Brownian sheet*.

2.2 Stationary Gaussian processes

If T has a group structure, or more generally if there is an action by a group G on T, then we can consider Gaussian processes that respect the symmetry of this action. What that means is that the Gaussian process $X = (X_t)_{t \in T}$ should have the same distribution as the shifted-process $X^g = (X_{g.t})_{t \in T}$, for any $g \in G$. Such a process is called G-stationary. When $T = G$, the action is understood to be multiplication from the left.

Since the distribution of a Gaussian process is determined by the mean function m and covariance kernel K, the G-stationarity condition is equivalent to the conditions: (1) $m(g.t) = m(t)$ for all $g \in G$, $t \in T$, and (2) $K(g.t, g.s) = K(t,s)$ for all $t,s \in T$ and $g \in G$.

It suffices for this course to consider only three cases, when T is $\mathbb{Z}_n$ (finite cyclic group) or $\mathbb{Z}$ (integers) or $\mathbb{R}$ (real line), and the natural extensions to higher dimensions. The case of $\mathbb{Z}_n$ serves as a motivating example. For some comments on general groups, go to the end of this section.

Finite cyclic group: Let X be a centered stationary process on $\mathbb{Z}_n = \{0,1,\ldots,n-1\}$ with addition modulo n. The covariance must satisfy $\mathbf{E}[X_j X_k] = K(j-k)$, where $j-k$ is modulo n, of course. Clearly $K(j) = K(n-j)$, hence K is an even function. The full covariance

matrix can be written as

$$\mathbb{K} = \begin{pmatrix} K(0) & K(1) & \dots & K(n-1) \\ K(n-1) & K(0) & \dots & K(n-2) \\ \vdots & \vdots & \vdots & \vdots \\ K(2) & \dots & K(0) & K(1) \\ K(1) & \dots & K(n-1) & K(0) \end{pmatrix} = K(0) + K(1)C + K(2)C^2 + \ldots + K(n-1)C^{n-1}$$

where the unitary matrix $C = (c_{i,j})_{0 \le i,j \le n-1}$ has $c_{i,j} = 1$ if $j - i = 1$ and $c_{i,j} = 0$ otherwise. That is

$$C = \begin{pmatrix} 0 & 1 & 0 & \dots & 0 \\ 0 & 0 & 1 & \dots & 0 \\ \vdots & \vdots & \vdots & \vdots & \\ 0 & \dots & \dots & 0 & 1 \\ 1 & \dots & \dots & 0 & 0 \end{pmatrix}.$$

It is easy to see that the eigenvalues of C are the nth roots of unity, and the eigenvector corresponding to $e^{i\theta_k}$ where $\theta_k = \frac{2\pi k}{n}$ is $\mathbf{v}_k = (1, e^{i\theta_k}, e^{2i\theta_k}, \ldots, e^{i(n-1)\theta_k})^t$, all for $0 \le k \le n-1$.

Since any stationary covariance matrix $\mathbb{K}$ is a polynomial in C, it has the same eigenvectors, and the eigenvalue corresponding to $\mathbf{v}_{\theta_k}$ is $\mu_k := K(0) + K(1)e^{i\theta_k} + \ldots + K(n-1)e^{i(n-1)\theta_k}$. As $\mathbb{K}$ is positive semi-definite, $p_k \ge 0$. Thus from the spectral decomposition,

$$\mathbb{K} = \mu_0 \mathbf{v}_0 \mathbf{v}_0^* + \mu_1 \mathbf{v}_1 \mathbf{v}_1^* + \ldots + \mu_{n-1} \mathbf{v}_{n-1} \mathbf{v}_{n-1}^*$$

we see that

$$K(m) = \mu_0 + \mu_1 e^{\frac{2\pi i m}{n}} + \mu_2 e^{2\frac{2\pi i m}{n}} + \ldots + \mu_{n-1} e^{(n-1)\frac{2\pi i m}{n}}.$$

Whenever $\mu = (\mu_0, \ldots, \mu_{n-1})^t$ is a vector, which we think of as a function on the group $\hat{Z}_n := \{e^{2\pi i k} n : 0 \le k \le n-1\}$, the right hand side of the above expression defines a function on $\mathbb{Z}_n$ that is called the discrete Fourier transform of μ. In our case, $\mu_k \ge 0$, hence we may think of μ as a measure on $\hat{Z}_n$. In fact, it is easy to see the symmetry $\mu_k = \mu_{n-k}$. In summary we have the following theorem.

Theorem 6. *Let $K : \mathbb{Z}_n \mapsto \mathbb{R}$ be an even function. The following are equivalent.*

1. *The matrix $(K(j-k))_{0 \le j,k \le n-1}$ is positive semi-definite (i.e., a covariance matrix).*
2. *K is the discrete Fourier transform of a symmetric measure on $\hat{Z}_n$.*

Exercise 7. Work out the analogue of Theorem 6 for $T = \mathbb{Z}_n^d$.

The group of integers: Consider the case when $T = \mathbb{Z}^d$, $d \geq 1$. Without any attempt at precision, we may think of $\mathbb{Z}$ as a limiting case of $\mathbb{Z}_n$ (better to write $\mathbb{Z}_n = \{-m, -m+1, \ldots, m-1, m\}$ when $n = 2m+1$ to make this intuitively clear), and $\hat{Z}_n$, which consists of the n-th roots of unity fills up the whose circle S^1. This makes the following statement believable.

Theorem 8. *Let $K : \mathbb{Z}^d \mapsto \mathbb{R}$ be an even function. The following are equivalent.*

1. *$(K(j-k))_{j,k\in\mathbb{Z}}$ is positive semi-definite.*
2. *K is the Fourier transform of a finite symmetric measure μ on $\mathbb{T}^d := [-\pi, \pi]^d$, i.e., $\mu(-A) = \mu(A)$ and $K(m) = \int_{\mathbb{T}^d} e^{i\langle m, \theta\rangle} d\mu(\theta)$ for all $m \in \mathbb{Z}^d$.*

The group of real numbers: Now suppose $T = \mathbb{R}^d$. The analogous theorem, due to Bochner is the following.

Theorem 9 (Bochner's theorem)**.** *Let $K : \mathbb{R}^d \mapsto \mathbb{R}$ be an even function. The following are equivalent.*

1. *K is continuous and $(K(t-s))_{t,s\in\mathbb{R}^d}$ is positive semi-definite.*
2. *K is the Fourier transform of a finite symmetric measure on $\mathbb{R}$.*

Example 10. Let W be standard Brownian motion and let $X(t) = e^{-t/2}W(e^t)$ for $-\infty < t < \infty$. Clearly X is a Gaussian process with zero mean and covariance $\mathbf{E}[X(t)X(s)] = e^{-\frac{1}{2}|t-s|}$. Therefore, if $Y = \theta_\tau X$, then Y is also a centered Gaussian process with covariance $\mathbf{E}[X(t-\tau)X(s-\tau)] = e^{-\frac{1}{2}|t-s|}$. Two Gaussian processes with the same mean function and covariance kernel are equal. Hence X is stationary. It is known as the *Ornstein-Uhlenbeck process*.

The dual and the isomorphism: For $T = \mathbb{Z}, \mathbb{R}$, let $\hat{T} = S^1, \mathbb{R}$, respectively. For a stationary Gaussian process X on T, the covariance kernel (when continuous) is seen to be of the form $K(t,s) = \hat{\mu}(t-s)$, for some finite symmetric measure on $\hat{T}$. The measure μ is called the *spectral measure* of X. Except for the mean (which is a constant), the spectral measure determines the distribution of the Gaussian process. Further, there can only be one spectral

measure for a given stationary process. Thus, centered stationary Gaussian processes and spectral measures are in on-one correspondence with each other.

Here is another view of this correspondence. For $t \in T$, define $e_t : \hat{T} \mapsto \mathbb{C}$ by $e_t(x) = e^{itx}$. Then the mapping $e_t \mapsto X_t$ from $L^2(\hat{T}, \mathcal{B}_{\hat{T}}, \mu)$ to $L^2(\Omega, \mathcal{F}, \mathbf{P})$ is an isometry. As such, linear questions about the process can be transferred to questions about the Hilbert space $L^2(\mu)$.

A prime example of such a question is one of prediction: Suppose we want to predict $X(0)$ based on the past $\{X(t) : t \leq -T\}$ for some $T > 0$. The best predictor is the conditional expectation, which is the linear projection in $L^2(\Omega, \mathcal{F}, \mathbf{P})$ of X_0 onto $\overline{\text{span}}\{X_t : t \leq -T\}$. By the isomorphism, this is equivalent to projecting $e_0 = 1$ onto $\overline{\text{span}}\{e_t : t \leq -T\}$ in $L^2(\mu)$.

Proof of the easy side of Bochner's theorem. Let μ be a finite, symmetric measure on $\hat{T}$. Then we are in the setting of Exercise 4, with $\Theta = \hat{T}$ and $\rho = \mu$ and $f(\lambda, t) = e^{i\lambda t}$, since $\hat{\mu}(t-s) = \int_{\hat{T}} e^{i\lambda(t-s)} d\mu(\lambda) = \int_{\hat{T}} f(\lambda, t)\overline{f(\lambda, s)} d\mu(\lambda)$. ■

The proof of the converse is not trivial. But it is of the same kind as one sees in Riesz's representation theorem or the moment problem. In all these cases, a linear functional is given on a vector space of functions such as $C(T)$ or polynomials or linear combinations of complex exponentials. If the space is sufficiently large and all non-negative functions in this class take non-negative values under the linear functional, then the linear functional is an integral with respect to some (positive) measure. We skip the details and refer the reader to **?**.

Interpreting the spectral measure: How to think of the Gaussian process itself? Let $X = (X_t)_{t \in \mathbb{R}}$ be a Gaussian process with an atomic spectral measure $\mu = p_1 \delta_{\lambda_1} + \ldots + p_n \delta_{\lambda_n}$ (to be symmetric, each λ_i and its negative occur with the same weight). This means that $\mathbf{E}[X(t)X(0)] = \sum_{j=1}^{n} p_j e^{i\lambda_j t}$. If we define $X(t) = \sqrt{p_1}\ Z_1 e^{i\lambda_1 t} + \ldots + \sqrt{p_n}\ Z_n e^{i\lambda_n t}$, with i.i.d. standard Gaussians Z_is, then it is clear that $\mathbf{E}[X(t)X(0)] = p_1 e^{i\lambda_1 t} + \ldots + p_n e^{i\lambda_n t} = \hat{\mu}(\lambda)$. Thus X is the Gaussian process with spectral measure μ.

This gives an idea of how the spectral measure defines the process. Basically X is a linear combination of complex exponentials $e_\lambda(t) := e^{i\lambda t}$, where λ is in the support of μ, and the coefficients of these exponentials are independent Gaussians with variances depending on how much mass μ puts at λ. This intuitive understanding is useful. For example, the tail of the spectral measure dictates the smoothness of the process and the smoothness of the spectral measure dictates the decay of correlation between the process values at far off

time points. To see this, we recall the following facts from basic Fourier analysis[1]

1. If $\int |x|^k d\mu(x) < \infty$ for $k \le p$, then $\hat{\mu} \in C^{(p)}$ and $\hat{\mu}^{(p)}(t) = \int_{\mathbb{R}} (ix)^p e^{itx} d\mu(x)$.
2. If μ has density f with respect to Lebesgue measure and $f \in C^{(p)}$, then $\hat{\mu}(t) = o(|t|^{-k})$ as $t \to \pm\infty$.

At least the first one is seen in the discrete example above - if λ_ks are large, then X has high frequency component $e^{i\lambda_j t}$ and that means fast oscillations or less smoothness. For example, the Cauchy distribution is heavy tailed and that is why the O-U process has non-smooth paths.

Now it is clear how to generate many examples of stationary processes - just start with any symmetric measure and find its Fourier series/transform. Here is an example where we start with the process and find the spectral measure.

Example 11. The Ornstein-Uhlenbeck process has covatiance $K(t-s)$ where $K(u) = e^{-\frac{1}{2}|u|}$. It is known that this is the Fourier transform of $\frac{\frac{1}{2}}{\pi(\frac{1}{4}+x^2)}dx$, which is a scaling of the standard symmetric Cauchy measure. This is the spectral measure of the OU-process.

There is no magic to find the spectral measure. It need not be easy in general, but is sometimes.

▶ In the discrete setting, $X = (X(n))_{n\in\mathbb{Z}}$, we can try to write the density of the spectral measure as $f(\lambda) = \sum_{n\in\mathbb{Z}} K(n)e^{in\lambda}$. If $\sum_n |K(n)| < \infty$, this series converges absolutely and uniformly on $[-\pi,\pi]$. Therefore, f is a continuous function and $\frac{1}{2\pi}\int_{S^1} e^{-i\ell\lambda} f(\lambda)d\lambda = K(\ell)$. In other words, $\frac{1}{2\pi}f(\lambda)d\lambda$ is the spectral measure.

▶ In the continuous setting too, if $K \in L^1(\mathbb{R})$, then by the Fourier inversion formula the spectral measure has density $f(\lambda) = \frac{1}{2\pi}\int_{\mathbb{R}} K(t)e^{-i\lambda t}dt$. In the example of the O-U process, this is the easiest way to guess the spectral measure.

Example 12. Let Z_n be i.i.d. standard Gaussians and let $X_n = Z_n - Z_{n-1}$. Both $Z = (Z_n)$ and $X = (X_n)$ are stationary Gaussian processes. Their covariances are

$$K_Z(n,m) = \delta_{n,m} \qquad K_X(n,m) = 2\delta_{n,m} - \delta_{n,m+1} - \delta_{n,m-1}.$$

From this and the above recipes to construct the spectral density, we see that the spectral densities are $f_Z(\lambda) = \frac{1}{2\pi}$ (uniform on $[-\pi,\pi]$) and $f_X(\lambda) = 2 - 2\cos\lambda$.

[1]The second volume of Feller's famous books on probability has an excellent chapter on characteristic functions. Proofs can be found there.

Exercise 13. Let $X = (X_n)_{n\in\mathbb{Z}}$ be a stationary Gaussian sequence with spectral measure μ. Let Y be the result of passing it through a *linear filter*, i.e., $Y_n = a_0X_n + a_1X_{n-1} + \ldots + a_mX_{n-m}$, where $m \geq 0$ and $a_0, \ldots, a_m \in \mathbb{R}$ are fixed. Find the spectral measure of Y.

Here is an example in two dimensions.

Example 14. Let $T = \mathbb{R}^2$ and let μ be the uniform measure on the unit circle $\{x^2 + y^2 = 1\}$. Its Fourier transform

$$J_0(\lambda) := \int_0^{2\pi} e^{i(\lambda_1 \cos t + \lambda_2 \sin t)} \frac{dt}{2\pi}$$

is known as the Bessel function of the first kind of order zero. The centered Gaussian process with this covariance is known as the *random plane wave*. In the interpretation given above, it is a random superposition of waves of unit frequencies, equally in all directions. Its importance lies in the fact that high energy eigenfunctions of the Laplacian on Riemannian manifolds look like this random function, conjecturally (*Berry's random wave conjecture*).

Exercise 15. Let X be a stationary Gaussian process on $\mathbb{R}$ with an absolutely continuous spectral measure $d\mu(x) = h(x)dx$ (then h is called the *spectral density* of the process). Find the spectral measures of the following processes:

1. $Y_t = X_{\alpha t}$ where $\alpha > 0$.
2. $Y_m = X_{\delta m}$ for $m \in \mathbb{Z}$, where $\delta > 0$.
3. $Y_m = (-1)^m X_{\delta m}$ for $m \in \mathbb{Z}$, where $\delta > 0$.

Exercise 16. Let $p_n(t) = Z_0 + Z_1 t + \ldots + Z_n t^n$ be the *Kac polynomial*, where Z_k are i.i.d. standard Gaussians. Consider the rescaled process around 1, namely $q_n(t) = p_n(1 + \frac{t}{n})$, $-\infty < t < \infty$. Show that q_n converges to a stationary Gaussian process as $n \to \infty$ (in the sense of convergence of finite dimensional distributions) and find the limiting process and its spectral measure.

On general groups

Let G be an Abelian topological group that is locally compact. Then there is a unique Haar measure μ (meaning $\mu(g.A) = \mu(A.g) = \mu(A)$ for all Borel sets A and all $g \in G$). Now consider the regular action of G on $L^2(G,\mu)$ defined by $\tau_g f(h) = f(g+h)$, for $g,h \in G$. Each

τ_g is a unitary transformation on $L^2(G)$. Further, τ_g and τ_h commute for any $g,h \in G$. Consequently, they have a simultaneous spectral decomposition.

For simplicity, assume that G is finite with $n = |G|$, and then μ is the counting measure on G. Then there is an orthogonal basis of common eigenfunctions $\chi_1, \ldots, \chi_n$. That is, $\tau_g \chi_k = \lambda_k(g)\chi_k$ for all $g \in G$ and $1 \le k \le n$. To understand what the eigenfunctions and eigenvalues are, consider one of these and write it as $\tau_g \chi = \lambda(g)\chi$.

Then $\chi(g+h) = \lambda(g)\chi(h)$ for all g,h. This shows that χ cannot vanish, and we may normalize it so that $\chi(0) = 1$, where 0 is the identity of G. Then we see that $\lambda(g) = \chi(g)$ (set $h = 0$) and consequently, $\chi(g+h) = \chi(g)\chi(h)$. Thus, χ is a homomorphism from G to the multiplicative group $\mathbb{C} \setminus \{0\}$. It is easy to see that it must map into S^1. Thus, the common eigenfunctions are just homomorphisms from G to S^1. to complete or delete

2.3 Gaussian free field and related processes

Let $G = (V,E)$ be a finite connected graph. What we wish to find are jointly Gaussian random variables X_v, $v \in V$, having joint density proportional to $\exp\{-\frac{1}{2}Q(\mathbf{x})\}$, where $Q(\mathbf{x}) = \sum_{u \sim v}(x_u - x_v)^2$. But this is not integrable, since Q is not strictly positive definite, as $Q(\mathbf{1}) = 0$. Alternately, consider the simplest case of a graph with a single edge, in which case we have $e^{-\frac{1}{2}(x_1 - x_2)^2}$, whose integral over $\mathbb{R}^2$ is infinite.

To overcome this problem, one must impose a constraint to a lower dimensional space of $\mathbb{R}^V$. Indeed, $Q(\mathbf{x}) = 0$ if and only if $x_u = x_v$ for all $u \sim v$, which happens if and only if $\mathbf{x}$ is constant (because G is connected). Hence, if we restrict to any subspace $W \subseteq \mathbb{R}^V$ that contains no constant functions (we refer to vectors in $\mathbb{R}^V$ as functions on V) other than the zero function, then Q is strictly positive definite on W, and then the Gaussian distribution with density $\exp\{-\frac{1}{2}Q(\mathbf{x})\}$ on this subspace does exist. Two examples:

1. Fix a non-empty subset $B \subseteq V$ and set $W = \{\mathbf{x} \in \mathbb{R}^V : x_v = 0 \text{ for } v \in B\}$. The corresponding Gaussian field is called the GFF with Dirichlet boundary condition on B.

2. Let $W := \mathbf{1}^{\perp} = \{\mathbf{x} \in \mathbb{R}^V : \sum_v x_v = 0\}$. The corresponding GFF is called the GFF with zero average.

Note that we have not specified the covariance of the GFF directly. To find that, it is necessary to write $Q(\mathbf{x}) = \mathbf{x}^t \Sigma^{-1} \mathbf{x}$ for $\mathbf{x} \in W$. For this purpose, define the Laplacian matrix

L of the graph, which is the $V \times V$ matrix such that

$$L(u,v) = \begin{cases} d_u & \text{if } u = v, \\ -1 & \text{if } u \sim v \\ 0 & \text{otherwise} \end{cases}$$

where d_v is the degree of v. Then, it is easy to check that $\langle L\mathbf{x}, \mathbf{x}\rangle = Q(\mathbf{x})$. Since $L\mathbf{1} = 0$, it has no inverse. But on W it has an inverse, and that is the covariance of the GFF. We shall analyse it in more detail in greater generality soon, but here is the summary: The covariance of X_u and X_v is $\mathcal{G}(u,v)$ where $\mathcal{G}(u,v)$ is the expected number of visits to v by a simple random walk on G started at u and killed when it hits the set B.

On trees it is easy to understand the GFF by hand. Here is the simplest example of the line graph.

Example 17. Let G be the subgraph of $\mathbb{Z}$ induced by the vertices $V = \{0, 1, \ldots, n\}$. Then $Q(\mathbf{x}) = \sum_{j=1}^{n} (x_j - x_{j-1})^2$. Two cases of the GFF with Dirichlet boundary condition.

1. $B = \{0\}$. $(X_1, \ldots, X_n)$ have density proportional to $\exp\{-\frac{1}{2}\sum_{j=1}^{n}(X_j - X_{j-1})^2\}$. Hence we may represent them as $X_j = Z_1 + \ldots + Z_j$ where Z_j are i.i.d. $N(0,1)$. Thus X is jsut random walk with $N(0,1)$ steps and $\mathbf{E}[X_j X_k] = j \wedge k$.

2. $B = \{0, n\}$. The way to construct this is $X_j = Z_1 + \ldots + Z_j - \frac{j}{n}(Z_1 + \ldots + Z_n)$ where Z_i are i.i.d. $N(0,1)$. The covariance is $\mathbf{E}[X_j X_k] = j(n-k)/n$. This is random walk bridge with Gaussian steps. One can also describe is the the Gaussian random walk conditioned to be at 0 at time n.

Exercise 18. Let T be a finite rooted tree and let $B = \{\text{root}\}$ be a singleton. In this case, describe the GFF on T (if the question sounds vague, assume that you have a supply of i.i.d. $N(0,1)$ random variables and use them to construct the GFF on T). Either from your construction or by finding the Green's function, give an expression for $\mathbf{E}[X_u X_v]$ for two distinct vertices u, v.

Extension to general reversible Markov chains: Let P be a transition matrix for a Markov chain on a finite state space V. We shall assume that the chain is reversible with respect to some $\pi : V \mapsto (0, \infty)$, hence this chain can be thought of as random walk on the graph

$G = (V, E, (c_e)_{e \in E})$, where the edges are $\{u, v\}$ with $P_{u,v} > 0$ and the conductances are $C_{u,v} = \pi(u)P_{u,v}$. The assumption of reversibility ensures that the graph is undirected so are the conductances. The random walk transitions are $P_{u,v} = C_{u,v}/\sum_{w \in V} C_{u,w}$ and the reversible measure is $\pi(u) = \sum_{w \sim u} C_{u,w}$.

For simplicity, we work with the case $\pi(u) = 1$, so $P_{u,v} = P_{v,u}$. The general case will be left as exercise. Then the matrix P is symmetric and has a spectral decomposition. All its eigenvalues are in $[-1, 1]$. One of the eigenvalues is 1. Hence for $0 < \theta < 1$, the matrix $(I - \theta P)^{-1}$ is positive definite. Let us assume that P is aperiodic, so that by the Perron-Frobenius theorem, all other eigenvalues are in $(-1, 1)$ and we can write the power series expansion

$$(I - \theta P)^{-1} = I + \theta P + \theta^2 P^2 + \theta^3 P^3 + \dots$$

We would like to set $\theta = 1$, but the series on the right does not converge and the inverse on the left does not exist. By fixing a non-empty set $B \subseteq V$ and imposing Dirichlet conditions there, we are changing the Markov chain so that the states in B become absorbing states, and the transition probabilities from vertices outside B remain unchanged. Then the chain is transient and the above series converges. The centered Gaussian process indexed by V and having covariance $(I - P)^{-1}$ is the generalized notion of GFF.

Agreement with the original definition: If all the conductances are 1, the chain becomes SRW on G. Further, if π has to be uniform, then the degrees must agree (recall that $\pi(u) = \sum_{w \sim u} C_{u,w}$), let us say d is the common degree. Then $I - P = \frac{1}{d} L$, where L is the Laplacian matrix of G, defined by $L(i,i) = \deg(i)$ and $L(i,j) = -1$ for $i \sim j$ (other entries are zero). Thus, up to a scaling by $\sqrt{d}$, this agrees with the earlier definition of a Gaussian free field on an unweighted graph.

Green's function: For a transient Markov chain (ona. countable state space), we define its Green's function

$$G(u,v) = \sum_{t=0}^{\infty} P^t(u,v) \;=\; \mathbf{E}_u \left[\sum_{t=0}^{\infty} \mathbf{1}_{X_t = v} \right],$$

the expected number of visits to v by a chain that started at u. This is the covariance kernel of the Gaussian free field for a transient chain, and what we did above in case of random walk on a finite graph (which is certainly recurrent) is to convert it to a transient chain by introducing one or more absorbing states and then taking the Green's function of the resulting transient chain as the covariance function of our Gaussian process.

Continuum GFF: In the continuum setting, there is a well-known definition of Green's function, which happens to be $G(x,y) = c_2 \log|x-y|^{-2}$ in $\mathbb{R}^2$ and $G(x,y) = c_d|x-y|^{-d+2}$ in $\mathbb{R}^d$ for $d \geq 3$. If Ω is a reasonable subset of $\mathbb{R}^d$ (e.g., bounded open set with smooth boundary), then there is a notion of Green's function G_Ω (it corresponds to the Laplacian with Dirichlet boundary condition or equivalently to Brownian motion that is killed at the boundary of Ω), but that also has singularity (of the kind $\log(1/|x-y|)$ in 2-dimensions and $1/|x-y|^{d-2}$ in higher dimensions) and hence there is no Gaussian process whose covariance function is $G(x,y)$ (the variance would have to be infinite!). But nevertheless, it is an object one can make sense of in various limiting senses or as a random distribution. Working in the discrete setting avoids these technicalities, but also loses out on some of the symmetries such as scaling (there are richer symmetries such as conformal invariance in 2-dimensions). However, features of the continuum GFF can often be seen in some asymptotic sense in the discrete GFF.

2.3.1 Continuum GFF

Suppose X is a centered Gaussian process on an open set $U \subseteq \mathbb{R}^2$ with covariance kernel K. Assume that it so happens that $t \mapsto X_t(\omega)$ is almost surely continuous. Then we can create new fields on more exotic index sets as follows.

- For $f \in C_c(U)$, define $Y(f) := \int_U f(t)X(t)dt$. This is a Gaussian process on $C_c(U)$ with covariance $\mathbb{K}(f,g) = \int_{U\times U} f(t)g(s)K(t,s)dtds$.
- For any curve $\gamma : [0,1] \mapsto U$, define $Z(\gamma) = \int_\gamma X d\gamma := \int_0^1 X(\gamma(u))\gamma'(u)du$. Then Z is a Gaussian field on the space of all curves in U with covariance

$$\mathbb{K}(\gamma,\eta) = \int_0^1 \int_0^1 K(\gamma(u),\eta(v))\gamma'(u)\eta'(v)\, du\, dv.$$

- Let $\mathcal{M}$ be the space of finite Borel measures on U and define $W(\mu) = \int_U X(t)d\mu(t)$. This is a Gaussian field on $\mathcal{M}$ with covariance kernel

$$\mathbb{K}(\mu,\nu) = \int_U \int_U K(t,s)d\mu(t)d\nu(s).$$

In some sense, the original field can be got from Y by taking a sequence $f_n \in C_c(U)$ that approaches δ_t. For example, if $\varphi \in C_c(\mathbb{R}^2)$ whose integral is 1, we can set $f_n(t) = \frac{1}{\varepsilon^2}\varphi((t_0 +$

$t)/\varepsilon)$ (for small enough ε) and it is easy to see that $Y(f_n) \to X(t_0)$ as $n \to \infty$. Similarly one can recover X from Z (take curves that stay close to t_0) or from W (let $\mu_n \to \delta_{t_0}$).

The continuum GFF is a peculiar case where we have the approximating fields Y, Z, W, but the original field X does not exist! We imagine it as existing as a limit of the others.

GFF on a bounded domain having Green's function: Let U be a bounded domain that has a Green's function G_U. By definition, the (Dirichlet) Green's function is the unique function on $U \times U \mapsto$ such that (1) $G_U(\cdot, y)$ is harmonic on $U \setminus \{y\}$ for any $y \in U$, (2) $G_U(x,y) \to 0$ as $x \to \partial U$, (3) $G_U(x,y) = -\log|x-y| + O(1)$ as $x \to y$.

The existence of a Green's function requires mild regularity on the domain U. A sufficient condition is that for every boundary point, there is a continuous curve whose image intersects U in that boundary point alone. But for our purposes one may just stick to the unit disk $\mathbb{D}$, in which case the Green's function has the explicit formula $G_{\mathbb{D}}(z,w) = \log \frac{|z-w|}{|1-\overline{w}z|}$. As in this example, in general, the Green's function is symmetric in the co-ordinates. Further, the harmonicity and behaviour near y may be summarised as saying $\Delta G_U(\cdot, y) = -2\pi\delta_y(\cdot)$ in the weak sense that

$$\int_U \Delta\varphi(x)\, G(x,y)\, dx = -2\pi\varphi(y), \text{ for any } \varphi \in C_c^2(U).$$

Informally, Green's function is the inverse of the Laplacian (with Dirichlet boundary conditions).

With this preparation, the GFF on U is supposed to be the Gaussian process on U with covariance kernel G_U. For this to make sense, G_U must be positive semi-definite. We see that this poses some problems.

It is a non-trivial fact that the $-\Delta$ has a discrete spectrum $0 < \lambda_1 \leq \lambda_2 \leq \ldots$ and the corresponding eigenfunctions $\{\varphi_n\}$ (properly chosen) form an orthonormal basis for $L^2(U)$. Thus, $-\Delta\varphi_n = \lambda_n\varphi_n$. Then, the Green's function has the representation

$$G_U(x,y) = \sum_{n=1}^{\infty} \frac{\varphi_n(x)\varphi_n(y)}{\lambda_n}.$$

This shows that G_U is formally positive semi-definite. The fact is, $G_U(x,x) = \infty$ (recall the singularity at $y = x$), hence this is only formally true. Hence we must take an indirect route[2].

[2]This is taken from the beautiful notes Introduction to the Gaussian Free Field and Liouville Quantum Gravity of Nathanaël Berestycki.

Let $\mathcal{M}_+$ be the set of all compactly supported positive Borel measures on U such that $\int\int G_U(x,y)d\mu(x)d\mu(y) < \infty$. Then let $\mathcal{M} = \{\mu_+ - \mu_- : \mu_\pm \in \mathcal{M}_+\}$ be the set of signed measures having a similar property. Observe that if a measure has atoms, it is not in $\mathcal{M}$ whereas if it has continuous density, then it is definitely in $\mathcal{M}$ (continuity is more than enough to integrate out the logarithmic singularity). Uniform measures on open subsets and even length measures on smooth curves are in $\mathcal{M}_+$.

Define (why is it finite?)

$$\mathbb{K}(\mu,\nu) = \int\int G_U(x,y)d\mu(x)d\nu(y).$$

By the formal expansion of G_U, it follows that (need justification)

$$\mathbb{K}(\mu,\nu) = \sum_{n\geq 1} \frac{1}{\lambda_n} L_n(\mu)L_n(\nu), \qquad L_n(\mu) := \int \varphi_n d\mu,$$

from which the positive semi-definiteness of $\mathbb{K}$ is clear.

Definition 19. The Gaussian free field on U is defined to be the centered Gaussian process on $\mathcal{M}$ with covariance kernel $\mathbb{K}$.

Of course, is a Gaussian process X on U with covariance G_U were to exist, then we would have got the above process by integrating: $Y(\mu) = \int_U X(z)d\mu(z)$. The whole point is that X does not exist but Y does. By considering $Y(\mu_n)$ where $\mu_n \to \delta_z$, we imagine that we are probing X.

Exercise 20. Consider GFF F on a domain $U \subseteq \mathbb{R}^2$. Suppose $\mathbb{D}(x,R) \subseteq U$. Let $\mu_{x,t}$ denote the uniform probability measure on $\mathbb{D}(x,e^{-t})$. Define $X(t) = F(\mu_{x,t})$ for $t \geq -\log R$. Find the covariance of the process X.

2.4 Processes on a tree and on its boundary

Let $\mathcal{T}$ be a rooted tree in which all vertices having finite degrees. Let V denote its vertex set, and let 0 denote the root. We write $|u|$ for the distance of u from the root. Let $\mathcal{T}_n = \{v \in V : |v| = n\}$ denote the n-th *generation* of vertices. We write $u \rightsquigarrow v$ to mean that v is a child of u (i.e., v is adjacent to u and $|v| = |u|+1$) or that u is the parent of v. We write $u \mapsto v$ to mean that v is a descendent of u or that u is an ancestor of v. This means that the path from the root to v passes through u. For two vertices u,v, the vertex $u \wedge v$ is

the farthest vertex from the root that is a common ancestor of both u and v. It is called the last common ancestor of u and v.

Let $w : V \mapsto \mathbb{R}_+$ be given. Let Z_v, $v \in V$, be i.i.d. $N(0,1)$ random variables. Define

$$X_v = \sum_{u \mapsto v} w(u) Z_u.$$

Then, X is a centered Gaussian process on V with covariance $\mathbf{E}[X_u X_v] := \sum_{y \mapsto u \wedge v} w_y^2$. There are different choices of trees and weights for which this process is interesting. A common sort of interest is in defining $X^{(n)}$ as the restriction of X to $\mathcal{T}_n$ and studying its asymptotics as $n \to \infty$. For example, when $\mathcal{T}$ is the regular 4-ary tree (where every vertex including the root has exactly four children) and $w_v = 1$ for all v, then the processes $X^{(n)}$ may be considered a crude version of the Gaussian free field on the unit square in the plane. This we explain in one dimension first.

Example 21. Let $\mathcal{T}$ be a regular binary tree (all vertices have two children) and let $w_v = 1$ for all v. Consider the process $X^{(n)} = (X_v)_{|v|=n}$ on the n-th generation vertices. Then, $\mathbf{E}[X_u X_v] = |u \wedge v|$.

One may code the vertices of $\mathcal{T}$ by binary strings, starting with the root vertex (0), its children (00) and (01), the children of (00) are (000) and (001), and so on, with the n-th generation vertices coded by binary strings of length n. These binary strings may in turn by identified with dyadic rational numbers in $[0,1]$. For example, (01011) corresponds to $\frac{1}{2} + \frac{1}{8} + \frac{1}{16}$. Define the distance d_λ on $\mathcal{T}_n$ by $d_\lambda(u,v) = \lambda^{-\frac{1}{2}|u \wedge v|}$.

With these notations, we see that $\mathbf{E}[X_u X_v] = \frac{2}{\log \lambda} \log \frac{1}{d_\lambda(u,v)}$. Although the distance d_λ on dyadic rationals is not quite the same as the Euclidean distance, it may be considered a reasonable substitute and then X may be considered a substitute or a toy version of a *logarithmically correlated field*. Then one may study the processes $X^{(n)}$ as $n \to \infty$. For example, one can look at[3] $M_n = \max_{v \in \mathcal{T}_n} X_v$.

Similarly, if $\mathcal{T}$ is a 4-ary tree, then using the dyadic decomposition of the unit square into sub-squares, one can see that $X^{(n)}$ are logarithmically correlated. Since the Gaussian free field on the unit square has (in a formal sense) covariance kernel $G(x,y)$, and the Green's function $G(x,y) \sim c \log(1/|x-y|)$, the processes $X^{(n)}$ may be thought of as crude approximants to the GFF on the unit square.

[3] See for example, the lecture notes of Ofer Zeitouni on Gaussian processes where this is studied at length ("branching random walk").

Processes on the boundary of the tree: To the tree $\mathcal{T}$, one can associate its boundary $\partial\mathcal{T}$, which is a metric space whose elements are infinite simple paths from the root:

$$\partial\mathcal{T} := \{\xi = (0 = u_0, u_1, u_2, \ldots) : u_k \rightsquigarrow u_{k+1} \text{ for } k \geq 0\}.$$

To define a metric on $\mathcal{T}$, fix $\lambda > 1$ and define $d_\lambda(\xi,\eta) = \lambda^{-\frac{1}{2}|\xi\wedge\eta|}$, where $\xi\wedge\eta$ is the largest n for which the first n co-ordinates of ξ and η agree (we call it the last common ancestor of ξ and η).

Exercise 22. Check that d_λ is a metric and that $\partial\mathcal{T}$ is compact under this metric space. [In fact, d_λ is an *ultrametric*, i.e., $d_\lambda(\xi,\eta) \leq d_\lambda(\xi,\zeta) \vee d_\lambda(\zeta,\eta)$ for all $\xi,\eta,\zeta \in \partial\mathcal{T}$.]

Let $w_v = w_n$ for all v with $|v| = n$. If $\sum_v w_v^2 < \infty$, then we can define the Gaussian process X on $\partial\mathcal{T}$ by $X_\xi = \sum_{k\geq 0} w_k Z_{v_k}$ if $\xi = (v_0, v_1, \ldots)$. Here Z_v, $v \in V$, are i.i.d. $N(0,1)$ variables. Then X is a Gaussian process on $\partial\mathcal{T}$ with covariance kernel $K(\xi,\eta) = \sum_{k\leq|\xi\wedge\eta|} w_k^2$. Alternately, we may write $\mathbf{E}[|X_\xi - X_\eta|^2] = 2\sum_{k>|\xi\wedge\eta|} w_k^2$.

For the particular choice of weights $w_k = \lambda^{-k/2}$, we see that $\mathbf{E}[|X_\xi - X_\eta|^2] = \frac{2\lambda}{\lambda-1}\lambda^{-\frac{1}{2}|\xi\wedge\eta|}$, which is the same as d_λ up to a constant factor. That is, d_λ is essentially the pull back of the $L^2(\mathbf{P})$ metric under the map $\xi \mapsto X_\xi$ from $\partial\mathcal{T}$ to $L^2(\mathbf{P})$. This pull-back metric will play much role in the study of the Gaussian process later.

Chapter 3

Gaussian isoperimetric inequality and concentration

Always γ_m denotes the standard Gaussian measure on $\mathbb{R}^m$, or on any vector space with a given inner product (for example, if W is a k-dimensional subspace of $\mathbb{R}^m$, we use γ_k for the Gaussian measure on W with the inherited inner product). For any set $A \subseteq \mathbb{R}^d$ and $\varepsilon > 0$, let $A^\varepsilon = \{y \in \mathbb{R}^d : |y - x| \leq \varepsilon \text{ for some } x \in A\}$.

Theorem 1 (Borell, Tsirelson-Ibragimov-Sudakov (1970s))**.** *Let A be any Borel subset of $\mathbb{R}^m$ with $\gamma_m(A) > 0$ and let H be a half-space in $\mathbb{R}^m$ with $\gamma_m(H) = \gamma_m(A)$. Then $\gamma_m(A^\varepsilon) \geq \gamma_m(H^\varepsilon)$ for all $\varepsilon > 0$. If A is a closed set with $\gamma_m(A) > 0$, then equality holds for some $\varepsilon > 0$ if and only if A is a half-space.*

We present two proof of this theorem. The first one, which is the original proof of Borrell and of Sudakov–Tsirelson, uses isoperimetric inequality on spheres and the fact that Gaussian measure arises as the limit of uniform measures on high dimensional spheres. The second one, due to Ehrardt, is a self-contained proof using the idea of symmetrization which is also one approach to proving isoperimetric inequality on spheres and in Euclidean space[1].

[1]Our proof is cobbled together from the paper of Ehrhard, *Symétrisation de l'espace de Gauss* and the appendix to the paper of Figiel, Lindenstrauss and Milman, *The dimension of almost spherical sections of convex bodies*. The symmetrization idea is from Ehrhard. But the rest of the details needed to complete the proof seems most cleanly presented in the paper of Figiel, Lindenstrass and Milman, albeit for the isoperimetric inequality on the sphere. These details appear to go through for the Gaussian case with minimal modification. If there are gaps or mistakes, please let me know.

First, the one-dimensional case as an exercise. We give a solution later, since the symmetrization proof of Theorem 1 works by induction on m.

Exercise 2. For any closed set $A \subseteq \mathbb{R}$ and any $\varepsilon > 0$, we have $\Phi^{-1}(\gamma_1(A^\varepsilon)) \geq \Phi^{-1}(\gamma_1(A)) + \varepsilon$. [**Hint:** Try proving it for one interval and then a finite union of intervals. From there to closed sets may be omitted.]

3.1 Proof of GIE via isoperimetric inequality on spheres

Let σ_{n-1} denote the uniform probability measure on $\mathbb{S}^{n-1}$. Endow it with the spherical metric d and let $A_\varepsilon := \{y \in \mathbb{S}^{n-1} : d(y,A) \leq \varepsilon\}$ denote the ε-enlargement of A. The isoperimetric inequality on the sphere says the following.

Lemma 3. *Let $A \in \mathcal{B}(\mathbb{S}^{n-1})$ and let $B = \{x \in \mathbb{S}^{n-1} : x_1 > t\}$ (for some $t \in (-1,1)$) be a* spherical cap *such that $\sigma_{n-1}(A) = \sigma_{n-1}(B)$. Then, $\sigma_{n-1}(A_\varepsilon) \geq \sigma_{n-1}(B_\varepsilon)$ for any $\varepsilon > 0$.*

Let $\Pi_{n,d} : \mathbb{S}^{n-1} \mapsto \mathbb{R}^d$ be defined by $\Pi_{n,d}(x) = \sqrt{n}(x_1, \ldots, x_d)$. Let $\mu_{n,d} = \sigma_{n-1} \circ \Pi_{n,d}^{-1}$ denote the push-forward probability measure on $\mathbb{R}^d$. We showed earlier that $\mu_{n,d} \xrightarrow{d} \gamma_d$. This means that $\mu_{n,d}(A) \to \gamma_d(A)$ for $A \subseteq \mathbb{R}^d$ with $\gamma_d(\partial A) = 0$. A stronger statement is true.

Exercise 4. For any $A \subseteq \mathbb{R}^d$ measurable, $\mu_{n,d}(A) \to \gamma_d(A)$.

Proof of the GIE

Fix $A \in \mathcal{B}_{\mathbb{R}^d}$ (or even measurable?) with $\Phi(\alpha) = \gamma_d(A) \in (0,1)$. Fix any $\beta < \alpha$ and let $B = \{x \in \mathbb{R}^d : x_1 \leq \beta\}$. Then B is a a half-space with probability $\Phi(\beta)$ under γ_d and $B^\varepsilon = \{x : x_1 \leq \beta + \varepsilon\}$ has probability $\Phi(\beta + \varepsilon)$ under γ_d.

For $n \geq d$, let $A_n, B_n, A_{n,\varepsilon}, B_{n,\varepsilon}$ denote the inverse images under $\Pi_{n,d}$ of $A, B, A^\varepsilon, B^\varepsilon$ respectively. These are subsets of $\mathbb{S}^{n-1}$ and $B_n, B_{n,\varepsilon}$ are spherical caps. Further, σ_{n-1} measure of these sets converge to γ_d measure of the corresponding sets in $\mathbb{R}^d$ (i.e., $\sigma_{n-1}(A_n) \to \gamma_d(A)$ etc.). One last observation is that $A_{n,\varepsilon}$ contains $(A_n)^{\varepsilon/\sqrt{n}}$ (this last is an enlargement in $\mathbb{S}^{n-1}$ with respect to the spherical metric in which $u, v \in \mathbb{S}^{n-1}$ have distance $\cos^{-1}(\langle u, v \rangle)$).

Since $\beta < \alpha$, for large enough n, we have $\sigma_{n-1}(A_n) > \sigma_{n-1}(B_n)$. By the isoperimetric inequality and the observation above, we see that $\sigma_{n-1}(A_{n,\varepsilon}) \geq \sigma_{n-1}((B_n)^{\varepsilon/\sqrt{n}})$. However, it is easy to see that $\sigma_{n-1}((B_n)^{\varepsilon/\sqrt{n}}) \to \gamma_d(B^\varepsilon)$ (direct calculation). Putting all this together, we see that $\liminf \sigma_{n-1}(A_n) \geq \gamma_d(B^\varepsilon)$ and hence $\gamma_d(A^\varepsilon) \geq \gamma_d(B^\varepsilon)$. As this is true for all $\beta < \alpha$,

letting $\beta \uparrow \alpha$, we see that $\gamma_d(A^\varepsilon) \geq \gamma_d(H^\varepsilon)$ where $H = \{x \in \mathbb{R}^d : x_1 \leq \alpha\}$. This completes the proof.

3.2 Proof of GIE by symmetrization

Notation: For a unit vector $\mathbf{u} \in \mathbb{R}^m$ and $t \in \mathbb{R}$, define the closed half-space $H_{\mathbf{u}}(t) := \{\mathbf{x} : \langle \mathbf{x}, \mathbf{u} \rangle \leq t\}$. For a closed subset $A \subseteq \mathbb{R}^m$, define

- $M(A) := \{B \subseteq \mathbb{R}^m : \text{B is closed, } \gamma_m(A) = \gamma_m(B), \gamma_m(A^\varepsilon) \geq \gamma_m(B^\varepsilon) \text{ for all } \varepsilon > 0\}$.
- $r(A) := \inf\{t \in \mathbb{R} : A \subseteq H_{\mathbf{u}}(t) \text{ for some unit vector } \mathbf{u}\}$.

The set $M[A]$ is the collection of all closed sets that are at least as good as A from the isoperimetry point of view. The quantity $r(A)$ will be of use in proofs. We now collect some basic facts about $M[A]$ and $r(A)$.

Lemma 5. *Let $\mathcal{C}$ be the set of all closed subsets of $\mathbb{R}^m$ endowed with the Hausdorff metric d.*

1. *The function $A \to r(A)$ is continuous.*
2. *The function $A \to \gamma_m(A)$ is upper semicontinuous.*
3. *If A is a closed subset of $\mathbb{R}^m$ with $\gamma_m(A) > 0$, then $r(\cdot)$ attains its minimum on $M(A)$.*

The main idea in proving Theorem 1 is a symmetrization procedure due to Antoine Ehrhard (analogous to Steiner's symmetrization for the classical isoperimetric inequality in Euclidean space) that takes a set and produces another that is better in the isoperimetric sense.

Ehrhard's symmetrization: Let ℓ be a one-dimensional affine subspace in $\mathbb{R}^m$ and let $\mathbf{u} \in \ell^\perp$ be a unit vector. For any $A \subseteq \mathbb{R}^m$, define its symmetrization w.r.t. $(\ell, \mathbf{u})$ as the subset $B = S_{\ell,\mathbf{u}}[A]$ such that

1. for any $t \in \ell$, the section $B \cap (t + \ell^\perp)$ is a half-space in $t + \ell^\perp$ whose boundary is orthogonal to $\mathbf{u}$,
2. $\gamma_{m-1}(B \cap (t + \ell^\perp)) = \gamma_{m-1}(A \cap (t + \ell^\perp))$.

Here is a more explicit description of B. For each $t \in \mathbb{R}$, find the unique $a = a_t \in \mathbb{R} \cup \{\pm\infty\}$ such that $\gamma_{m-1}(H_{\mathbf{u}}(a) \cap (t + \ell^{\perp})) = \gamma_{m-1}(A \cap (t + \ell^{\perp}))$ and set $B = \bigcup_{t \in \ell}(H_{\mathbf{u}}(a_t) \cap (t + \ell^{\perp}))$.

As $S_{\ell,\mathbf{u}}[A]$ is defined by an uncountable union of sections, it is not obvious that it is measurable, even for a nice set A. The following lemma shows that any symmetrization transforms closed sets to closed sets, in particular measurable.

Lemma 6. *Let A be a closed set. Then $S_{\ell,\mathbf{v}}[A]$ is also closed.*

The following two lemmas show why symmetrization improves a set and that the only sets that cannot be improved by further symmetrizations are half-spaces. They justify the use of symmetrization as a tool and their proofs form the heart of the proof of Theorem 1.

Lemma 7. *Let A be closed and non-empty in $\mathbb{R}^m$. Then $S_{(\ell,\mathbf{v})}[A] \in M(A)$ for any symmetrization $(\ell, \mathbf{v})$.*

Lemma 8. *Let A be a non-empty closed subset of $\mathbb{R}^m$. Then there exist a finite sequence of symmetrizations under which A transforms to a set B with $r(B) < r(A)$.*

Now we prove the main theorem assuming all the lemmas stated so far.

Proof of Theorem 1. Let A be any closed set with $\gamma_m(A) > 0$. By the third part of Lemma 5, there is some $B \in M[A]$ with $r(B) \leq r(X)$ for all $X \in M[A]$. If B is not a half-space, then by Lemma 8 we could get apply a finite number of symmetrizations to get a set C with $r(C) < r(B)$. Lemma 7 implies that $C \in M[B]$. But since $M[B] \subseteq M[A]$ this contradicts the minimality of $r(B)$. Thus, B must be a half-space. This proves the isoperimetric inequality for closed sets A. Recall that (2) is an equivalent form of the inequality and thus it has been proved now for closed sets.

If A is any Borel set, by regularity of γ_m, for any $\delta > 0$ there exists a compact sets $K \subseteq A$ with $\gamma_m(K) \geq \gamma_m(A) - \delta$. Then

$$\begin{aligned}
\Phi^{-1}(\gamma_m(A^{\varepsilon})) &\geq \Phi^{-1}(\gamma_m(K^{\varepsilon})) \ \text{(because } K \subseteq A\text{)} \\
&\geq \Phi^{-1}(\gamma_m(K)) + \varepsilon \ \text{(by the proved inequality (2) for closed sets)} \\
&\geq \Phi^{-1}(\gamma_m(A) - \delta) + \varepsilon. \ \text{(because } \Phi^{-1} \text{ is increasing)}
\end{aligned}$$

Let $\delta \downarrow 0$ to get $\Phi^{-1}(\gamma_m(A^{\varepsilon})) \geq \Phi^{-1}(\gamma_m(A)) + \varepsilon$. ∎

3.2.1 Proofs of lemmas used to prove the Gaussian isoperimetric inequality

Proof of Lemma 5. 1. Suppose $d(A,B) < \delta$ for some $A, B \in \mathcal{C}$. If a half-space H contains A, then H^δ contains B. Therefore $r(B) \le r(A) + \delta$. Reversing the roles of A and B we see that $A \to r(A)$ is in fact a Lipschitz function on $\mathcal{C}$.

2. If $d(A,B) < \delta$ then $A^\delta \supseteq B$ and hence $\gamma_m(A^\delta) \ge \gamma(B)$. Hence, if $d(A_k, A) \to 0$, then $\gamma_m(A^\delta) \ge \limsup \gamma_m(A_k)$ as $k \to \infty$. This holds for any δ and $\gamma_m(A^\delta) \to \gamma_m(A)$ as $\delta \to 0$. Therefore $\gamma_m(A^\delta) \ge \limsup_{k\to\infty} \gamma_m(A_k)$ showing that γ_m is u.s.c. on $\mathcal{C}$.

3. Let $r = \inf\{r(X) : X \in A\}$. Since $\Phi(r(X)) \ge \gamma_m(A) > 0$ for all $X \in M[A]$, it follows that $r > -\infty$. If $r = +\infty$, then we may take $B = A$. Thus we assume that r is finite.

   Let $B_k \in M[A]$ with $r_k := r(B_k) \downarrow r$. Then $B_k \subseteq H_{\mathbf{u}_k}(r_k + 1/k)$ for some unit vectors $\mathbf{u}_k$. By passing to a subsequence we may assume that $\mathbf{u}_k \to \mathbf{u}$ for some unit vector $\mathbf{u}$. Since $\gamma_m(B_k) = \gamma_m(A) > 0$, there is a finite number R_0 such that $B(0,R_0)$ has a non-empty intersection with B_k for all k. By Lemma 10, we can pass to a further subsequence and assume that $B_k \cap K \to B \cap K$ in Hausdorff metric for every compact set K. Here B is a closed set.

   By the second part, $\gamma_m(B \cap K) \ge \limsup \gamma_m(B_k \cap K) \ge \limsup \gamma_m(B_k) - \gamma_m(K^c)$. Since $B_k \in M[A]$, by taking arbitrarily large K we get $\gamma_m(B) \ge \gamma_m(A)$.

   Now fix K. For any $\delta > 0$ we have $B \cap K \subseteq (B_k \cap K)^\delta$ for large enough k and hence $\gamma_m((B\cap K)^\varepsilon) \le \liminf \gamma_m((B_k \cap K)^{\delta+\varepsilon}) \le \gamma_m(A^{\varepsilon+\delta})$ since each $B_k \in M[A]$. Now let $\delta \downarrow 0$ to get $\gamma_m((B\cap K)^\varepsilon) \le \gamma_m(A^\varepsilon)$ for all $\varepsilon > 0$. Then let K increase to $\mathbb{R}^m$ and conclude that $\gamma_m(B^\varepsilon) \le \gamma_m(A^\varepsilon)$. Thus, $B \in M[A]$.

   We claim that $B \subseteq H_{\mathbf{u}}(r)$. For if not, then for some small enough $\delta > 0$ and large enough compact set K we must have $(B \cap K) \cap \partial H_{\mathbf{u}}(r+\delta) \ne \emptyset$. But for large enough k we have $B_k \cap K \subseteq H_{\mathbf{u}}(r + \delta/3)$ and $B \cap K \subseteq (B_k \cap K)^{\delta/3}$ which implies that $B \cap K \subseteq H_{\mathbf{u}}(r + 2\delta/3)$, a contradiction.

   Putting everything together, we have found a set $B \in M[A]$ and $B \subseteq H_{\mathbf{u}}(r)$. Thus $r(B) = r$ and the proof is complete. ■

Proof of Lemma 6. Fix ℓ and $\mathbf{v}$ and write points of $\mathbb{R}^m$ as $(t, \mathbf{x})$ with $t \in \ell$ and $\mathbf{x} \in \ell^\perp$. For any set A, let $A_t = A \cap (t + \ell^\perp)$ for $t \in \ell$.

Suppose $t_k \to t$. If $(t_k, \mathbf{x}_k) \in A$ and $(t_k, \mathbf{x}_k) \to (t, \mathbf{x})$, then $(t, \mathbf{x}) \in A$ as A is closed. Therefore, $A_t \supseteq \limsup A_{t_k}$, in particular $\gamma_{m-1}(A_t) \geq \limsup \gamma_{m-1}(A_{t_k})$. This implies $a_t \geq \limsup a_{t_k}$.

Now let $B := S_{\ell,\mathbf{v}}[A]$ and suppose that $(t_k, \mathbf{y}_k) \in B$ and $(t_k, \mathbf{y}_k) \to (t, \mathbf{y})$. By definition of symmetrization, $\langle \mathbf{y}_k, \mathbf{v} \rangle \leq a_{t_k}$ and hence $\langle \mathbf{y}, \mathbf{v} \rangle \leq \limsup a_{t_k} \leq a_t$ which implies that $(t, \mathbf{y}) \in B$. Thus B is closed. ■

Proof of Lemma 7. Fix $\ell, \mathbf{v}$ and let $B = S_{(\ell,\mathbf{v})}[A]$. We need to prove two things.

$$(a)\ \ \gamma_m(B) = \gamma_m(A) \qquad \text{and} \qquad (b)\ \ \gamma_m(B^\varepsilon) \leq \gamma_m(A^\varepsilon) \text{ for each } \varepsilon > 0.$$

The first assertion is easy. Use Fubini's theorem to see that

$$\gamma_m(A) = \int_{\mathbb{R}} \gamma_{m-1}[(t\mathbf{u} + \ell^\perp) \cap A] d\gamma_1(t) = \int_{\mathbb{R}} \gamma_{m-1}[(t\mathbf{u} + \ell^\perp) \cap B] d\gamma_1(t) = \gamma_m(B).$$

The proof of (b) is non-trivial and it is the key step in the entire proof of Theorem 1. By Fubini's theorem, it suffices to show that $\gamma_{m-1}[(B^\varepsilon)_t] \leq \gamma_{m-1}[(A^\varepsilon)_t]$ for all $t \in \ell$, where $A_t := A \cap (t + \ell^\perp)$ is the t-section of A.

Without loss of generality let $\ell = \mathbb{R}e_1$ and $\mathbf{v} = e_2$. For each $s \in \mathbb{R}$, then $B_s = \{(s, u_2, \ldots, u_n) : u_2 \leq a_s\}$ where $\Phi(a_s) = \gamma_{m-1}(A_s)$. Let π denote the orthogonal projection from $\mathbb{R}^m$ onto $\ell^\perp = \text{span}\{e_2, \ldots, e_n\}$.

Fix $t \in \mathbb{R}$. Then $(t, \mathbf{x}) \in B^\varepsilon$ if and only if there exists s with $|s - t| \leq \varepsilon$ and $\mathbf{y} \in B_s$ with $|\mathbf{y} - \mathbf{x}| \leq \delta_s := \sqrt{\varepsilon^2 - (s-t)^2}$. This means

$$\pi[(B^\varepsilon)_t] = \bigcup_{s:|s-t|<\varepsilon} (\pi[B_s])^{\delta_s}, \qquad \pi[(A^\varepsilon)_t] = \bigcup_{s:|s-t|<\varepsilon} (\pi[A_s])^{\delta_s}. \tag{1}$$

In $\mathbb{R}^{n-1}$, $\pi(B_s)$ is a half-space with the same γ_{m-1} measure as $\pi(A_s)$ (by definition of symmetrization). Therefore, inductively assuming the the Gaussian isoperimetric inequality for lower dimensions (the ground case $m = 1$ is checked in Exercise 2), we get $\gamma_{m-1}[(\pi[B_s])^{\delta_s}] \leq \gamma_{m-1}[(\pi[A_s])^{\delta_s}]$ for each s. Therefore, using the second set-identity in (1) we get $\gamma_{m-1}[(\pi[B_s])^{\delta_s}] \leq \gamma_{m-1}[\pi[(A^\varepsilon)_t]$ for each $s \in [t - \varepsilon, t + \varepsilon]$.

Now note that $(\pi[B_s])^{\delta_s} = \{(u_2, \ldots u_n) : u_2 \leq a_s + \delta_s\}$ are all half-spaces. For any two of them, one contains the other. Hence, their union is an increasing union of a countable number of them. Therefore,

$$\gamma_{m-1}[\pi((B^\varepsilon)_t)] = \sup_{s:|s-t|\leq\varepsilon} \gamma_{m-1}[(\pi[B_s])^{\delta_s}] \leq \gamma_{m-1}[\pi[(A^\varepsilon)_t].$$

Equivalently $\gamma_{m-1}[(B^\varepsilon)_t] \leq \gamma_{m-1}[(A^\varepsilon)_t]$. By Fubini's theorem, this proves (b). ■

Proof of Lemma 8. Since A is closed, the infimum in the definition of $r(A)$ is a minimum. Let $\mathbf{v}$ be a unit vector such that $H_{\mathbf{v}}(r) \supseteq A$ with $r = r(A)$. Without loss of generality we assume $\mathbf{v} = e_n$. Let $W = re_n + e_n^{\perp}$, the boundary of the half-space $H := H_{e_n}(r)$.

First pick any line ℓ_0 inside W and let $A' = S_{\ell_0, e_n}[A]$. Since A is closed and not the whole half-space, A' is a closed proper subset of H. Further, if $\mathbf{x} \in A'$ and $\mathbf{y} \in H$ has $y_i = x_i$ for $i \leq n-1$ and $y_n < x_n$, then $\mathbf{y} \in A'$ too. Therefore, it is clear that there is a point $p \in W$ and $\delta > 0$ such that $A' \cap Q_p(2\delta) = \emptyset$ where $Q_p(2\delta) = p + (-2\delta, 2\delta)^n$.

Now let $\ell_i = p + re_n + \mathbb{R}e_i$ for $i = 1, 2, \ldots, n-1$. These are lines inside W, passing through p and parallel to the co-ordinate directions.

Let $A'' = S_{\ell_1, p_1}[A']$. For each $t \in [-\delta, \delta]$ the section $(t + \ell_1^{\perp}) \cap A'$ is a subset of $[(t + \ell_1^{\perp}) \cap (H \setminus Q_p(\delta))]$. Therefore, there is some $\delta' > 0$ such that $A'' \cap ([-\delta, \delta] \times \mathbb{R}^{m-1})$ is contained in $H_{\mathbf{v}}(r - \delta')$.

Now symmetrize w.r.t. $(\ell_2, \mathbf{v})$ and let $A''' = S_{\ell, \mathbf{v}}[A'']$. For each $t \in \ell_2$, the section $A'' \cap (t + \ell_2^{\perp})$ is a subset of $H_{\mathbf{v}}(r - \delta')$. Therefore, there is some $\delta'' > 0$ such that $A''' \subseteq H_{\mathbf{v}}(r - \delta'')$. Thus in (at most) three symmetrizations we arrive at a set A''' with $r(A''') < r(A)$. ■

Solution to Exercise 2. For $p \in (0,1)$ define $Q_p = \Phi^{-1}(1-p)$, the $(1-p)$-quantile. For $x \leq Q_p$, define $b_p(x)$ by the equation $\gamma_1[x, b_p(x)] = p$. Let α_p denote the unique x such that $b_p(x) = -x$. Differentiating $p = \int_x^{b_p(x)} \varphi(t)dt$, we get $\varphi(b_p(x))b_p'(x) - \varphi(x) = 0$.

Fix $p \in (0,1)$, $\varepsilon > 0$ and define $h(x) = \gamma_1[x - \varepsilon, b_p(x) + \varepsilon]$ and observe that

$$\begin{aligned} h'(x) &= \varphi(b_p(x) + \varepsilon)b_p'(x) - \varphi(x - \varepsilon) \\ &= \varphi(x)\left\{\frac{\varphi(b_p(x) + \varepsilon)}{\varphi(b_p(x))} - \frac{\varphi(x - \varepsilon)}{\varphi(x)}\right\} \\ &= \varphi(x)\left\{\frac{\varphi(b_p(x) + \varepsilon)}{\varphi(b_p(x))} - \frac{\varphi(-x + \varepsilon)}{\varphi(-x)}\right\}. \end{aligned}$$

Note that $\varphi(u + \varepsilon)/\varphi(u) = e^{-u\varepsilon - \frac{1}{2}\varepsilon^2}$ is decreasing in x. Hence, when $x > \alpha_p(x)$ (which is equivalent to $b_p(x) > -x$), we have $h'(x) < 0$. Thus $h(\alpha_p) > h(x) > h(Q_p)$ for all $x \in (\alpha_p, Q_p)$.
Case of one closed interval: If A is an interval with $\gamma_1(x) = p$, then it is of the form $[x, b_p(x)]$ for some x. We may also assume that $x \geq \alpha_p$ (otherwise replace A by $-A$). Thus, by the above deduction, $\gamma(A^{\varepsilon})$ is minimized when $x = Q_p$.
Case of multiple closed intervals: We write A as $I_1 \sqcup I_2 \ldots \sqcup I_k$ with $I_j = [x_j, b_p(x_j)]$ with $b_p(x_{i-1}) < x_i$ for all i. There are two reductions which improve our set in isoperimetric setting.

1. Suppose that I_k and I_{k-1} differ by less than 2ε, i.e., $b_p(x_{k-1}) + \varepsilon > x_k - \varepsilon$. In this case, if we move the interval $[x_k, b_p(x_k)]$ to the left (i.e., decrease x_k), then $\gamma_1(A)$ stays the same but $\gamma_1(A^\varepsilon)$ decreases till x_k hits $b_p(x_{k-1})$. This results in a set with $(k-1)$ intervals and better isoperimetric profile.

2. Suppose that I_{k-1} and I_k are separated by at least 2ε. Without loss of generality $b_p(x_k) > -x_k$ (otherwise, replace I_k by $-I_k$, which would be even further to the right than I_k and the separation condition continues to hold). Then, by the earlier deduction, as x_k increases, $\gamma_1(A)$ stays the same but $\gamma_1(A^\varepsilon)$ decreases, till $x_k = Q_{p_k}$.

Repeatedly applying these two reductions, we can reduce A to the interval $[Q_p, \infty)$.
Case of an arbitrary closed set: Let A be closed with $\gamma_1(A) = p$. For any small $\eta > 0$, the set A^η is the closure of an open set, and hence it is a union of countably many disjoint closed intervals. At the cost of losing an η probability, we drop all but finitely many intervals. This gives us a set B with the property that $B \subseteq A^\eta$ and $p' := \gamma_1(B) \geq p - \eta$. By the already proved inequality, $\gamma_1(B^\varepsilon) \geq \gamma_1[Q_{p'} - \varepsilon, \infty)$. Of course $B^\varepsilon \subseteq A^{\eta+\varepsilon}$ and therefore $\gamma_1(A^\varepsilon) \geq \gamma_1[Q_{p'} - \varepsilon, \infty)$. Letting $\eta \downarrow 0$ and noticing that $p' \to p$, we get $\gamma_1(A^\varepsilon) \geq \gamma_1[Q_p - \varepsilon, \infty)$. ■

3.2.2 Appendix: Hausdroff metric

Let (X, d) be a metric space and let $\mathcal{C}_X$ denote the set of all non-empty closed subsets of X. The Hausdorff distance between two closed sets A, B is defined by $d_H(A,B) = \inf\{r > 0 : A^r \supseteq B \text{ and } B^r \supseteq A\}$ where $A^r = \{\mathbf{x} : d(\mathbf{x}, A) \leq r\}$. The value $+\infty$ is allowed and $(\mathcal{C}, d)$ is a metric space (if you are not comfortable with a metric that takes infinite values, just use $d_H(A,B) \wedge 1$ which is a finite metric).

Exercise 9. Let (X, d) be a compact metric space. Then $(\mathcal{C}_X, d_H)$ is a compact metric space.

We shall work in $\mathbb{R}^m$ which is not compact.

Lemma 10. *Let A_k be a sequence of closed non-empty sets in $\mathbb{R}^m$. Assume that $A_k \cap B(0, R_0) \neq \emptyset$ for all k for some R_0. Then, there exists a subsequence k_j and a non-empty closed set X such that $A_{k_j} \cap K \to X \cap K$ in Hausdorff metric for every non-empty compact $K \subseteq \mathbb{R}^m$.*

Proof. For each $j > R_0$, use Exercise 9 to see that $A_k \cap B(\bar{0}, j)$ has a subsequence that converges in Hausdorff metric to some set $X_j \subseteq B(\bar{0}, j)$. Set $X = \cup_j X_j$. Then it is easy to see that X is closed and the conclusions hold (check!). ■

3.2.3 Appendix: Gap in the proof!

In lecture we realized that there is a gap in the proof that we gave for the isoperimetric inequality. It is in the proof of Lemma 8. The given proof is correct in dimensions 3 and higher but not in dimension 2 as there is only one line contained in the boundary of a half space in $\mathbb{R}^2$! We fix this below[2].

Lemma 11. *Let $\mathbf{v}_k = (\cos\theta_k, \sin\theta_k)$ with $\theta_0 = 0$ and $\theta_k = \frac{\pi+\theta_{k-1}}{2}$ for $k \geq 1$. Let $\ell_k = \mathbf{v}_k^{\perp}$ and let $S_k := S_{\ell_k, -\mathbf{v}_k}$. Given a closed set $A \subseteq \mathbb{R}^2$, define $A_0 = S_0[A]$ and $A_k = S_k[A_{k-1}]$ for $k \geq 1$.*

1. *If $\mathbf{x} \in A_k$ then $\mathbf{x} + t\mathbf{v}_0 + s\mathbf{v}_k \in A_k$ for all $t, s \geq 0$.*

2. *Let $H = \{(x,y) : y \geq \Phi^{-1}(\gamma_2(A))\}$. Then A_k converges to H on compacta in Hausdorff metric i.e., $A_k \cap K \to H \cap K$ in Hausdorff metric for every compact set K.*

Proof. 1. By definition of symmetrization, it is clear that if $\mathbf{x} \in A_k$ then $\mathbf{x} + t\mathbf{v}_k \in A_k$ for $t > 0$. It remains to prove for $k \geq 1$ that if $\mathbf{x} \in A_k$ then $\mathbf{x} + t\mathbf{v}_0 \in A_k$. The case $k = 0$ is trivial.

Consider $k = 1$. By the $\gamma_1(A \cap (t\mathbf{v}_0 + \ell_1^{\perp}))$ is increasing in t (because of the case $k = 0$), which shows that if $\mathbf{x} \in A_1$ then $\mathbf{x} + t\mathbf{v}_0 \in A_1$. This completes the proof for $k = 1$.

Fix $k \geq 2$ and let π denote the projection onto $\ell_k^{\perp}$ and let $A_{k-1,t} = A_{k-1} \cap (t\mathbf{v}_k^{\perp} + \ell_k^{\perp})]$ and $A_{k,t} = A_k \cap (t\mathbf{v}_k^{\perp} + \ell_k^{\perp})]$ so that $A_{k,t}$ is a half-line with $\gamma_1(A_{k,t}) = \gamma_1(A_{k-1,t})$. Observe that ℓ_k is the angle bisector of $\mathbf{v}_k$ and $\mathbf{v}_0$. Therefore, inductively assuming the lemma for $k-1$, we see that $\pi[A_{k-1,t+\varepsilon}] \supseteq \pi[A_{k-1,t}]^{\varepsilon}$ (the ε-enlargement in $\ell^{\perp} = \mathbb{R}$). Consequently, by the one-dimensional isoperimetric inequality we deduce that $\pi[A_{k,t+\varepsilon}] \supseteq \pi[A_{k,t}]^{\varepsilon}$. Draw a picture to see that this precisely implies that if $\mathbf{x} \in A_k$ then $\mathbf{x} + t\mathbf{v}_0 \in A_k$ for $t \geq 0$.

2. If $\gamma_2(A) = 0$ then A_k is empty for all k and the statement is valid. Hence assume $\gamma_2(A) > 0$. By properties of symmetrization, for every k we have $\gamma_2(A_k) = \gamma_2(A)$ and $\gamma_2(A_k^{\varepsilon}) \leq \gamma_2(A^{\varepsilon})$ for all $\varepsilon > 0$. Also define the cone $C_k = \{s\mathbf{v}_0 + t\mathbf{v}_k : s,t \geq 0\}$ and $C_\infty = \{(x,y) : y \geq 0\}s$.

Let R be large enough such that $\gamma_2(B_0(R)^c) < \gamma_2(A)$. Then there exists $\mathbf{x}_k \in A_k \cap B_0(R)$. Having fixed $\varepsilon > 0$ and $R > 0$, it is clear that for large enough k and every $\mathbf{x} \in B_0(R)$

[2]Proof is taken from Bogachev's book, chapter 4

we have $((\mathbf{x}+\mathcal{C}_k)\cap B_0(R))^\varepsilon \supseteq (\mathbf{x}+\mathcal{C}_\infty)\cap B_0(R)$. Since $\mathcal{C}_k \subseteq \mathcal{C}_\infty$ we obviously have $((\mathbf{x}+\mathcal{C}_\infty)\cap B_0(R))^\varepsilon \supseteq (\mathbf{x}+\mathcal{C}_k)\cap B_0(R)$.

■

3.3 Some consequences of the Gaussian isoperimetric inequality

Equivalent formulations: We present two equivalent ways of writing the Gaussian isoperimetric inequality. The first one, without explicit reference to half-spaces is

$$\Phi^{-1}(\gamma_m(A^\varepsilon)) \geq \Phi^{-1}(\gamma_m(A)) + \varepsilon \quad \text{for all Borel sets } A \text{ and any } \varepsilon > 0. \tag{2}$$

Exercise 12. Deduce (2) and Theorem 1 from each other.

Here is yet another formulation[3].

Proposition 13. *Let $f : \mathbb{R}^n \to \mathbb{R}$ be a Lip(κ) function. Then there exists a Lip(κ) function $g : \mathbb{R} \to \mathbb{R}$ such that $\gamma_n \circ f^{-1} = \gamma_1 \circ g^{-1}$. In other words, the distribution of the random variable f on the probability space $(\mathbb{R}^n, \mathcal{B}_{\mathbb{R}^n}, \gamma_n)$ is the same as the distribution of the random variable g on $(\mathbb{R}, \mathcal{B}_{\mathbb{R}}, \gamma_1)$.*

Exercise 14. Deduce Proposition 13 and Theorem 1 from each other.

Proposition 13 shows the dimension-free nature of isoperimetric inequality. In other words, the isoperimetric inequality will hold for standard Gaussian measures in infinite dimensions, once we make sense of such an object! This would not have been the case if Proposition 13 only asserted that g is Lip($\kappa \log n$), for example.

Log-concave densities: What made the proof of isoperimetric inequality in one dimension click? Looking back, we see that the key point was that $\varphi(u+\varepsilon)/\varphi(u)$ is decreasing in u, for any fixed $\varepsilon > 0$. Any other density f satisfying this will also satisfy the isoperimetric inequality (perhaps we need symmetry?). This condition is equivalent to $\log f(u+\varepsilon) - \log f(u)$ being decreasing in u. Assuming smoothness for simplicity, this happens if and only if $(\log f)'(u)$ is decreasing in u, which in turn is equivalent to $(\log f)''(u)$ being negative. In other words, equivalent to $\log f$ being a concave function.

[3]Taken from Boris Tsirelson's lecture notes available on his home page.

Any density (in any dimension) for which $\log f$ is concave, is called *log-concave*. Examples in one dimension are symmetric exponential density $\frac{1}{2}e^{-|x|}$, uniform density on an interval, and of course the Gaussian. Examples in higher dimensions are uniform measures on compact convex sets and the densities $\exp\{-|\mathbf{x}|^p\}$ for $p \geq 1$. One can get many more from these few, since log-concave densities are closed under convolutions and under linear transformations (eg., marginals). Log-concave densities are a very important class of densities that share many properties of Gaussian measures, in particular, concentration properties.

Gaussian Brunn-Minkowski inequality: In Euclidean space, we deduced the isoperimetric inequality from the Brunn-Minkowski inequality. Is there an analogue for the Gaussian measure? Ehrhard initiated this study and proved the inequality below for convex sets, again using his symmetrization procedure. The convexity assumption was relaxed by Latala and completely removed by Borell.

Result 15 (Ehrhard, Latala, Borell)**.** If $A, B \subseteq \mathbb{R}^n$ (Borel sets), and $\alpha \in [0,1]$, then $\Phi^{-1}(\gamma_n(\alpha A + (1-\alpha)B)) \geq \alpha\Phi^{-1}(\gamma_n(A)) + (1-\alpha)\Phi^{-1}(\gamma_n(B))$.

We shall not use this and hence not give a proof[4].

Concentration inequalities: The isoperimetric inequality implies concentration inequalities for various functions of Gaussian random variables. This is its primary importance in probability. It is possible to deduce some of these concentration bounds directly without using the isoperimetric inequality, albeit with poorer constants, but the isoperimetric inequality yields the sharpest general bounds.

Theorem 16. *Let $f : \mathbb{R}^n \to \mathbb{R}$ be a Lip(κ) function. Let M_f be a median of f, defined by $\gamma_n\{f \geq M_f\} \geq \frac{1}{2}$ and $\gamma_n\{f \leq M_f\} \geq \frac{1}{2}$. Then, for every $t > 0$, we have*

$$\gamma_n\left\{f - M_f \geq t\right\} \leq \bar{\Phi}\left(\frac{t}{\kappa}\right) \leq e^{-\frac{t^2}{2\kappa^2}}, \tag{3}$$

$$\gamma_n\left\{|f - M_f| \geq t\right\} \leq 2\bar{\Phi}\left(\frac{t}{\kappa}\right) \leq 2e^{-\frac{t^2}{2\kappa^2}}. \tag{4}$$

[4]Potential presentation topic! See Borell's paper *The Ehrhard inequality*. Another potential topic is a very different proof of the Gaussian isoperimteric inequality by Bobkov, see *An isoperimetric inequality on the discrete cube, and an elementary proof of the isoperimetric inequality in Gauss space.*

Proof. If $A = \{f \le M_f\}$ then $A^t \subseteq \{f \le M_f + \kappa t\}$. But $\Phi^{-1}(\gamma_n(A)) \ge 0$ and hence by (2) we get $\Phi^{-1}(\gamma_n(A^t)) \ge t$. Hence $\gamma_n\{f \ge M_f + \kappa t\} \le \bar{\Phi}(t)$ which shows the first claim. The second follows by adding the same estimate for $\gamma_n\{f \le M_f - t\}$. ■

Remark 17. Since $\bar{\Phi}(t)$ is strictly smaller than $\frac{1}{2}$ for every $t > 0$, it follows that the median is unique! Incidentally, we have been writing statements in terms of measures, but one can equivalently state them in terms of random variables. If $X_1, \ldots, X_n$ are i.i.d. $N(0,1)$ random variables on some probability space, and $V = f(X_1, \ldots, X_n)$ for a Lip(κ) function f, then

$$\mathbf{P}\{|V - \text{Med}[V]| \ge t\} \le 2e^{-t^2/2\kappa^2}.$$

The random variable is concentrated around its median. Incidentally, inequalities of this type, with perhaps not the optimal constants on the right, can be obtained by easier methods (see the end of this section). Often that suffices in applications but we decided to go through the isoperimetric inequality for its natural appeal, in addition to sharpness of constants.

Example 18. Some examples of Lipschitz functions of interest are $\max_i x_i$, $\|\mathbf{x}\|_p$ (or any norm, for that matter), $d(\mathbf{x}, A)$ for a fixed closed set A. A smooth function is Lipschitz if and only if its gradient is bounded.

What about functions of correlated Gaussians? Here is a simple exercise.

Exercise 19. Suppose $X \sim N_n(\mu, \Sigma)$ and let $f : \mathbb{R}^n \to \mathbb{R}$ is a Lip(κ) function. Let $V = f(X)$. Then $\mathbf{P}\{|V - \text{Med}[V]| \ge t\} \le 2e^{-t^2/2\lambda_1\kappa^2}$ with λ_1 being the maximal eigenvalue of Σ.

Concentration inequalities of the type given by Theorem 16 are desirable to have for many other probability measures too. Deduce the following from Theorem 16.

Exercise 20. Let V_n be the uniform probability measure on $[0,1]^n$. If $f : [0,1]^n \to \mathbb{R}$ is Lip(κ), show that

$$\gamma_n\left\{f - M_f \ge t\right\} \le e^{-ct^2/\kappa^2},$$
$$\gamma_n\left\{|f - M_f| \ge t\right\} \le 2e^{-ct^2/\kappa^2}.$$

Here c is a numerical constant (find it!).

For general product measures, for example uniform measure on the discrete cube $\{0,1\}^n$, getting a similar concentration inequality is hard. This is the famous *Talagrand's inequality*, proved by Talagrand and now a cornerstone in probability.

Concentration about the mean: Usually mean is easier to compute than median and concentration inequalities are often expressed around the mean. Here is a simple way to get a (sub-optimal) concentration inequality around the mean for the same setting as above. Let $f : \mathbb{R}^n \to \mathbb{R}$ be a Lip(κ) function and let M_f be its median under γ_n and let $E_f = \int f(\mathbf{x}) d\gamma_n(\mathbf{x})$ be its expectation.

Using the bound in Theorem 16 we get

$$\mathbf{E}[(f - M_f)_+] = \int_0^\infty \gamma_n\{f > M_f + t\}dt \leq \int_0^\infty \bar{\Phi}(t/\kappa)dt = \frac{\kappa}{\sqrt{2\pi}}.$$

The same bound holds for $\mathbf{E}[(f - M_f)_-]$ and we get $\mathbf{E}[|f - M_f|] \leq \sqrt{\frac{2}{\pi}}\kappa < \kappa$. In particular, $|E_f - M_f| < \kappa$. Therefore, for $t \geq 2$, we get

$$\gamma_n\{f - E_f > t\kappa\} \leq \gamma_n\left\{f - M_f > \frac{t}{2}\kappa\right\} \leq \bar{\Phi}(t/2),$$

by another application of Theorem 16. For $t \leq 2$, we use the trivial bound $\gamma_n\{f - E_f > t\kappa\} \leq 1$. Putting all this together and using the same for deviations below E_f we arrive at the following result.

Theorem 21. *Let $f : \mathbb{R}^n \to \mathbb{R}$ be a Lip(κ) function. Let $E_f = \int f d\gamma_n$. Then, for every $t > 0$, we have (with $C = 1/\bar{\Phi}(1)$)*

$$\gamma_n\left\{f - E_f \geq t\right\} \leq C\bar{\Phi}\left(\frac{t}{2\kappa}\right) \leq Ce^{-\frac{t^2}{8\kappa^2}}, \tag{5}$$

$$\gamma_n\left\{|f - E_f| \geq t\right\} \leq C\bar{\Phi}\left(\frac{t}{2\kappa}\right) \leq Ce^{-\frac{t^2}{8\kappa^2}}. \tag{6}$$

Weaker forms of concentration by easier methods: As we remarked earlier, weaker forms of concentration inequalities can be obtained by easier methods some of which we mention here[5].

Let $f : \mathbb{R}^n \to \mathbb{R}$ be a Lip(1) function and let $X \sim \gamma_n$. We look for a number A_f such that $f(X)$ is well-concentrated about A_f. The crudest bound is as follows. Let Y be an independent copy of X on the same probability space, and use $\mathbf{E}[|f(X) - f(Y)|] \leq \mathbf{E}[\|X - Y\|] \asymp \sqrt{n}$. Observing that $\min_a \mathbf{E}[|f(X) - a|] \leq \mathbf{E}[|f(X) - f(Y)|]$, we get a number A_f such

[5]For a spectacular presentation leading up from simpler inequalities up to the Borell-TIS inequality, see the lecture notes of Boris Tsirelson http://www.tau.ac.il/~tsirel/Courses/Gaussian/lect2.pdf. Here I have taken a couple of points from those notes.

that $\mathbf{E}[|f(X) - A_f|] \lesssim \sqrt{n}$. By Markov's inequality this gives weak bounds like $\mathbf{P}\{|f(X) - A_f| \geq t\} \lesssim \frac{\sqrt{n}}{t}$. This compares poorly with the bound in (16).

To improve this, we introduce a technique that will be used many times later. Interpolate between X and Y by setting $Z(\theta) = (\cos\theta)X + (\sin\theta)Y$ for $0 \leq \theta \leq \frac{\pi}{2}$ so that $Z(0) = X$ and $Z(\pi/2) = Y$. The key property of this interpolation is that for any θ, the random vectors $Z_\theta = (\cos\theta)X + (\sin\theta)Y$ and $\dot{Z}_\theta = -(\sin\theta)X + (\cos\theta)Y$ are independent and have γ_n distribution.

Now assume that f is smooth, then the Lipschitz condition is equivalent to $|\nabla f(\mathbf{x})| \leq \kappa$ for all $\mathbf{x} \in \mathbb{R}^n$. It is easy to approximate Lipschitz functions uniformly by smooth Lipschitz functions and thus extend the bounds obtained below to all Lipschitz functions, a step we shall not elaborate on. Then, write $f(X) - f(Y)$ as the integral of $\frac{d}{d\theta} f(Z_\theta) = \langle \nabla f(Z_\theta), \dot{Z}_\theta \rangle$ to get

$$\begin{aligned}
\mathbf{E}[|f(X) - f(Y)|] &\leq \int_0^{\pi/2} \mathbf{E}[\langle \nabla f(Z_\theta), \dot{Z}_\theta \rangle] d\theta \\
&= \sqrt{\frac{2}{\pi}} \int_0^{\pi/2} \mathbf{E}[\|\nabla f(Z_\theta)\|] d\theta \\
&\leq \sqrt{\frac{\pi}{2}}.
\end{aligned}$$

From this we get some number A_f such that $\mathbf{P}\{|f(X) - A_f| \geq t\} \lesssim \frac{1}{t}$. This does not decay fast in t, but is free of n, already a remarkable improvement over the crude bound.

By bounding $\mathbf{E}[G(|f(X) - f(Y)|)]$ for some convex increasing function G we can get better bounds along the same lines.

Exercise 22. For $b > 0$ and $x \in \mathbb{R}^n$ define $G_b(x) = (|x| - b)_+$. Use the convexity of G to obtain the bound $\mathbf{E}[|G_b(X) - G_b(Y)| \geq t] \leq \mathbf{E}[G(\frac{\pi}{2}X)]$.

What concentration of $f(X)$ does this yield?

Chapter 4

Comparison inequalities

The study of the maximum (or supremum) of a collection of Gaussian random variables is of fundamental importance. In such cases, certain comparison inequalities are helpful in reducing the problem at hand to the same problem for a simpler correlation matrix. In another direction, if two covariance matrices are close, the distributions of the corresponding Gaussian vectors are close. Quantitative statements to this effect are useful. In this chapter we study such general results and illustrate their uses with applications.

4.1 Preparatory lemmas

We start with a lemma of this kind and from which we derive two important results - Slepian's inequality and the Sudakov-Fernique inequality[1]. When the function does not depend on the covariance matrix, this lemma is due to Kahane, but the same proof applies in this more general situation.

Let $\mathcal{P}_n$ be the space of $n \times n$ symmetric, positive semi-definite matrices (local notation) and let $F : \mathbb{R}^n \times \mathcal{P}_n \mapsto \mathbb{R}$ be a smooth function. We write $F(z, \Sigma)$ with $z \in \mathbb{R}^n$ and $\Sigma \in \mathcal{P}_n$ and write $\partial_i f$ for $\frac{\partial f}{\partial z_i}$ and $\overline{\partial}_{(i,j)} f$ for $\frac{\partial f}{\partial \sigma_{i,j}}$.

[1]The presentation here is cooked up from Ledoux-Talagrand (the book titled *Probability on Banach spaces*) and from Sourav Chatterjee's paper on Sudakov-Fernique inequality. Chatterjee's proof can be used to prove Kahane's inequality too, and consequently Slepian's, and that is the way we present it here. I have not seen Lemma 1 written anywhere but the idea that one can allow the function to depend on the random vector and the covariance matrix of the random vector was inspired by one special case which appears as a key lemma in the theory of *Gaussian multiplicative chaos*.

Lemma 1. *Let X and Y be $n \times 1$ mutivariate Gaussian vectors with equal means, i.e., $\mathbf{E}[X_i] = \mathbf{E}[Y_i]$ for all i. Let $F : \mathbb{R}^n \times \mathcal{P}_n \to \mathbb{R}$ be any C^2 function all of whose partial derivatives up to second order have subgaussian growth in the first variable. Assume that (a) $(\partial_i\partial_i + 2\overline{\partial}_{(i,i)})F$ has the same sign as $\sigma^Y_{i,i} - \sigma^X_{i,i}$ for all i, and (b) $(\partial_j\partial_i + \overline{\partial}_{(i,j)})F$ has the same sign as $\sigma^Y_{i,j} - \sigma^X_{i,j}$ for all $i \neq j$. Then $\mathbf{E}[F(X, \Sigma_X)] \leq \mathbf{E}[F(Y, \Sigma_Y)]$.*

To be clear about the assumption, what we mean is that pointwise

$$(\sigma^Y_{i,i} - \sigma^X_{i,i}) \times (\partial_i\partial_i + 2\overline{\partial}_{(i,i)})F \geq 0, \qquad (\sigma^Y_{i,i} - \sigma^X_{i,i}) \times (\partial_i\partial_j + \overline{\partial}_{(i,j)})F \geq 0.$$

Note that no condition is imposed in cases where $\sigma^X_{i,j} = \sigma^Y_{i,j}$.

When the function depends only on the Gaussian variables and not on the covariance matrix, we get the following corollary.

Corollary 2 (Kahane)**.** *Let $f : \mathbb{R}^n \mapsto \mathbb{R}$ be a C^2 function whose first two partial derivatives have sub-Gaussian growth. If X, Y are as in the statement of the theorem and $(\sigma^Y_{i,j} - \sigma^X_{i,j})\partial_i\partial_j f(\mathbf{x}) \geq 0$ for all i, j, and for all $\mathbf{x} \in \mathbb{R}^n$, then $\mathbf{E}[f(X)] \leq \mathbf{E}[f(Y)]$.*

The key idea in the proof is one that is widely useful. Instead of comparing two Gaussians, it is better to interpolate between them smoothly and use the power of Calculus to prove a differential inequality which can then be integrated. In all examples of this chapter, the interpolation is a straight line between the two covariance matrices. In principle nothing precludes consideration of other curves.

Proof of Lemma 1. First assume that both X and Y are centered. Without loss of generality, assume that X and Y are defined on the same probability space and independent of each other.

Interpolate between them by setting $Z(\theta) = (\cos\theta)X + (\sin\theta)Y$ for $0 \leq \theta \leq \frac{\pi}{2}$ so that $Z(0) = X$ and $Z(\pi/2) = Y$. Let $\Sigma(\theta) = (\cos^2\theta)\Sigma_X + (\sin^2\theta)\Sigma_Y$ denote the covariance matrix of $Z(\theta)$. Let $W(\theta) = (Z(\theta), \Sigma(\theta))$. Then, differentiating under the expectation,

$$\begin{aligned}\frac{d}{d\theta}\mathbf{E}[F(W(\theta))] = &-\sin\theta \sum_{i=1}^n \mathbf{E}[X_i\partial_i F(W(\theta))] + \cos\theta \sum_{i=1}^n \mathbf{E}[Y_i\partial_i F(W(\theta))] \\ &+ 2\sin\theta\cos\theta \sum_{i\leq j}(\sigma^Y_{i,j} - \sigma^X_{i,j})\mathbf{E}[\overline{\partial}_{(i,j)}F(W(\theta))]\end{aligned}$$

The integration by parts formula for Gaussians says that $\mathbf{E}[X_i g(X)] = \sum_{j=1}^n \sigma^X_{i,j}\mathbf{E}[\partial_j g(X)]$. Use the independence of X and Y and apply this formula by first conditioning on Y (or vice

versa) to get

$$\mathbf{E}[X_i\partial_i F(W(\theta))] = \cos\theta \sum_{j=1}^{n} \sigma^X_{i,j}\mathbf{E}[\partial_j\partial_i F(W(\theta))],$$

$$\mathbf{E}[Y_i\partial_i F(W(\theta))] = \sin\theta \sum_{j=1}^{n} \sigma^Y_{i,j}\mathbf{E}[\partial_j\partial_i F(W(\theta))].$$

Therefore,

$$\begin{aligned}\frac{d}{d\theta}\mathbf{E}[F(Z(\theta),\Sigma(\theta))] &= \sin\theta\cos\theta\sum_{i=1}^{n}\sum_{j=1}^{n}(\sigma^Y_{i,j}-\sigma^X_{i,j})\mathbf{E}[\partial_j\partial_i F(W(\theta))]\\ &\quad + 2\sin\theta\cos\theta\sum_{i\le j}(\sigma^Y_{i,j}-\sigma^X_{i,j})\mathbf{E}[\overline{\partial}_{(i,j)}F(W(\theta))]\\ &= \sin\theta\cos\theta\sum_{i=1}^{n}(\sigma^Y_{i,i}-\sigma^X_{i,i})\mathbf{E}[(\partial_i\partial_i + 2\overline{\partial}_{(i,i)})F(W(\theta))]\\ &\quad + 2\sin\theta\cos\theta\sum_{i<j}(\sigma^Y_{i,j}-\sigma^X_{i,j})\mathbf{E}[(\partial_j\partial_i + \overline{\partial}_{(i,j)})F(W(\theta))]. \qquad (1)\end{aligned}$$

The assumptions were made so that each summand is positive. Therefore, $\theta \mapsto \mathbf{E}[F(W(\theta)]$ is increasing on $[0, \frac{1}{2}\pi]$, hence $\mathbf{E}[F(W(0))] \le \mathbf{E}[F(W(1))]$ which is exactly what we wanted to prove.

It remains to consider the case when the means are not zero. Let $\mu_i = \mathbf{E}[X_i] = \mathbf{E}[Y_i]$ and set $\hat{X}_i = X_i - \mu_i$ and $\hat{Y}_i = Y_i - \mu_i$ and let $g(x_1,\dots,x_n) = F(x_1+\mu_1,\dots,x_n+\mu_n)$. Then $F(X) = g(\hat{X})$ and $F(Y) = g(\hat{Y})$ while $\partial_i\partial_j g(x) = \partial_i\partial_j F(x+\mu)$. Thus, the already proved statement for centered variables implies the one for non-centered variables. ■

Remark 3. The interpolation is often alternately written as $\sqrt{t}X + \sqrt{1-t}Y$, $0 \le t \le 1$ or as $e^{-t}X + \sqrt{1-e^{-2t}}\,Y$, $0 \le t \le \infty$. There is no difference in substance.

A speculative (possibly naive) question: In Corollary 2, we compare $\mathbf{E}[F(X)]$ and $\mathbf{E}[F(Y)]$ for some $F : \mathbb{R}^n \mapsto \mathbb{R}$. In Lemma 1 this is extended this to a function $F : \mathbb{R}^n \times \mathcal{P}_n \mapsto \mathbb{R}$ and compare $\mathbf{E}[F(X,A)]$ to $\mathbf{E}[F(Y,B)]$ where A, B are the covariance matrices of X and Y. Can we go on to generalize to compare quantities like $\mathbf{E}[F(X,A,\Gamma)]$ where F takes three arguments, $X \in \mathbb{R}^n$, $A \in \mathcal{P}_n$ and $\Gamma \in T_A(\mathcal{P}_n)$ is a tangent vector to $\mathcal{P}_n$ at the point A? And so on, to higher derivatives, getting a hierarchy of more and more general inequalities?

4.2 Slepian's and Gordon's inequalities

We write X^* for $\max_i X_i$. The general intuition is that when X_is are positively correlated, then they tend to stick together, and X^* is unlikely to be large. To illustrate this, let Z_1, Z_2 be i.i.d. standard Gaussians and consider $U = (Z_1, Z_1)$, $V = (Z_1, Z_2)$, and $W = (Z_1, -Z_1)$, in which the correlations are $1, 0, -1$, respectively. Then $U^* = Z_1$, $V^* = Z_1 \vee Z_2$ and $W^* = |Z_1|$. Thus,

$$\mathbf{P}\{U^* < t\} = \Phi(t), \qquad \mathbf{P}\{V^* < t\} = \Phi(t)^2, \qquad \mathbf{P}\{W^* < t\} = 2\Phi(t) - 1.$$

Clearly these are in decreasing order, showing that[2] $U^* \prec V^* \prec W^*$. Extending this intuition, Slepian showed the following general stochatic comparison inequality.

Lemma 4 (Slepian's inequality)**.** *Let X and Y be $n \times 1$ mutivariate Gaussian vectors with equal means, and equal variance, i.e., $\mathbf{E}[X_i] = \mathbf{E}[Y_i]$ and $\mathbf{E}[X_i^2] = \mathbf{E}[Y_i^2]$ for all i. Assume that $\mathbf{E}[X_i X_j] \geq \mathbf{E}[Y_i Y_j]$ for all i, j. Then,*

1. *For any real $t_1, \ldots, t_n$, we have $\mathbf{P}\{X_i < t_i \text{ for all } i\} \geq \mathbf{P}\{Y_i < t_i \text{ for all } i\}$.*
2. *$X^* \prec Y^*$, i.e., $\mathbf{P}\{X^* > t\} \leq \mathbf{P}\{Y^* > t\}$ for all t.*

We would like to say that the first conclusion follows from Corollary 2 by taking $f(x_1, \ldots, x_n) = \prod_{i=1}^n \mathbf{1}_{x_i < t_i}$. The only wrinkle is that f is not smooth. Approximating the indicator with smooth decreasing functions, this can be converted to a rigorous proof.

Proof. To elaborate, let $\psi_i : \mathbb{R} \mapsto [0,1]$ be smooth decreasing functions with $\psi_i(t) = 1$ for $t \leq t_i$ and $\psi_i(t) = 0$ for $t \geq t_i + \varepsilon$. Let $f_\varepsilon(x_1, \ldots, x_n) = \prod_{i=1}^n \psi_i(x_i)$. Then

$$\partial_i \partial_j f(x) = \psi_i'(x_i)\psi_j'(x_j) \prod_{k \neq i,j} \psi_k(x_k) \geq 0.$$

Corollary 2 applies to show that $\mathbf{E}[f_\varepsilon(X)] \leq \mathbf{E}[f_\varepsilon(Y)]$. Let $\varepsilon \downarrow 0$ and apply monotone convergence theorem to get the first conclusion.

Taking $t_i = t$, we immediately get the second conclusion from the first. ■

Remark 5. The second statement is not less general than the first. Indeed, applying the second statement to $(X_1/t_1, \ldots, X_n/t_n)$ and $(Y_1/t_1, \ldots, Y_n/t_n)$, one gets the first.

Here is another inequality which specializes to Slepian's inequality when $m = 1$.

[2] We say that U is stochastically dominated by V and write $U \prec V$ if $\mathbf{P}\{U > t\} \leq \mathbf{P}\{V > t\}$ for all $t \in \mathbb{R}$.

Lemma 6 (Gordon's inequality)**.** *Let $X_{i,j}$ and $Y_{i,j}$ be $m \times n$ arrays of joint Gaussians with equal means. Assume that (1) $Cov(X_{i,j}, X_{i,\ell}) \geq Cov(Y_{i,j}, Y_{i,\ell})$, (2) $Cov(X_{i,j}, X_{k,\ell}) \leq Cov(Y_{i,j}, Y_{k,\ell})$ if $i \neq k$, (3) $Var(X_{i,j}) = Var(Y_{i,j})$. Then*

1. *For any real $t_{i,j}$ we have* $\mathbf{P}\left\{\bigcap_i \bigcup_j \{X_{i,j} < t_{i,j}\}\right\} \geq \mathbf{P}\left\{\bigcap_i \bigcup_j \{Y_{i,j} < t_{i,j}\}\right\}$,
2. $\min_i \max_j X_{i,j} \prec \min_i \max_j Y_{i,j}$.

Exercise 7. Deduce Gordon't inequality from Lemma 1 (or Corollary 2).

4.3 Sudakov-Fernique inequality

Studying the maximum of a Gaussian process is a very important problem. Slepian's (or Gordon's) inequality helps to control the maximum of our process by that of a simpler process. For example, if $X_1, \ldots, X_n$ are standard normal variables with positive correlation between any pair of them, then $\max X_i$ is stochastically smaller than the maximum of n independent standard normals (which is easy). However, the equality of variances condition of Slepian's inequality is restrictive, and the conclusion is much stronger than what one needs in many situations. Here is a more applicable substitute.

Theorem 8 (Sudakov-Fernique inequality)**.** *Let X and Y be $n \times 1$ Gaussian vectors satisfying $\mathbf{E}[X_i] = \mathbf{E}[Y_i]$ for all i and $\mathbf{E}[(X_i - X_j)^2] \leq \mathbf{E}[(Y_i - Y_j)^2]$ for all $i \neq j$. Then, $\mathbf{E}[X^*] \leq \mathbf{E}[Y^*]$.*

Remark 9. Assume that the processes are centered. If the two processes had the same variances, then the condition $\mathbf{E}[(X_i - X_j)^2] \leq \mathbf{E}[(Y_i - Y_j)^2]$ would be the same as $\mathrm{Cov}(X_i, X_j) \geq \mathrm{Cov}(Y_i, Y_j)$. In that case, Slepian's inequality would apply and we would get the much stronger conclusion of $X^* \prec Y^*$. The point here is that we relax the assumption of equal variances and settle for the weaker conclusion which only compares expectations of the maxima.

For non-centered processes one may wonder whether it would not be more appropriate to compare $\mathrm{Var}(X_i - X_j)$ with $\mathrm{Var}(Y_i - Y_j)$ in the assumption. But since $\mathbf{E}[(X_i - X_j)^2] = \mathrm{Var}(X_i - X_j) + (\mathbf{E}[X_i] - \mathbf{E}[X_j])^2$, and the means are assumed to be equal, that would be the same condition!

Proof. As in the proof of Lemma 1, we interpolate between X and Y using $Z(\theta) = \cos\theta\, X + \sin\theta\, Y$ for $0 \leq \theta \leq \frac{\pi}{2}$. Since our function now depends on the random variables only (i.e.,

$\overline{\partial}_{(}i,j)F = 0$ for all i,j), (1) becomes

$$\frac{d}{d\theta}\mathbf{E}[F(Z(\theta))] = \sin\theta\cos\theta \sum_{i,j=1}^{n} (\sigma^Y_{i,j} - \sigma^X_{i,j})\mathbf{E}[\partial_j\partial_i F(W(\theta))].$$

Now we specialize to the function $f_\beta(x) = \frac{1}{\beta}\log\sum_{i=1}^n e^{\beta x_i}$ where $\beta > 0$ is fixed (again, there is no dependence on Σ). Let $p_i(x) = \frac{e^{\beta x_i}}{\sum_{i=1}^n e^{\beta x_i}}$, so that $(p_1(x),\ldots,p_n(x))$ is a probability vector for each $x \in \mathbb{R}^n$. Observe that

$$\partial_i f(x) = p_i(x), \qquad \partial_i\partial_j f(x) = \beta p_i(x)\delta_{i,j} - \beta p_i(x)p_j(x).$$

Thus, (1) gives (writing $P_i = p_i(Z_\theta)$ for simplicity of notation)

$$\begin{aligned}\frac{1}{\beta(\cos\theta)(\sin\theta)}\,\frac{d}{d\theta}\mathbf{E}[f_\beta(Z_\theta)] &= \sum_{i,j=1}^n (\sigma^Y_{ij} - \sigma^X_{ij})\mathbf{E}\left[P_i\delta_{i,j} - P_iP_j\right]\\ &= \sum_{i=1}^n (\sigma^Y_{ii} - \sigma^X_{ii})\mathbf{E}[P_i] - \sum_{i,j=1}^n (\sigma^Y_{ij} - \sigma^X_{ij})\mathbf{E}[P_iP_j]\end{aligned}$$

Since $\sum_i p_i(x) = 1$ for any x, we can write $P_i = \sum_j P_iP_j$ and hence

$$\begin{aligned}\frac{1}{\beta(\cos\theta)(\sin\theta)}\,\frac{d}{d\theta}\mathbf{E}[f_\beta(Z_\theta)] &= \sum_{i,j=1}^n (\sigma^Y_{ii} - \sigma^X_{ii})\mathbf{E}[P_iP_j] - \sum_{i,j=1}^n (\sigma^Y_{ij} - \sigma^X_{ij})\mathbf{E}[P_iP_j]\\ &= \sum_{i<j}\mathbf{E}[P_iP_j]\left(\sigma^Y_{ii} - \sigma^X_{ii} + \sigma^Y_{jj} - \sigma^X_{jj} - 2\sigma^Y_{ij} + 2\sigma^X_{ij}\right)\\ &= \sum_{i<j}\mathbf{E}[P_iP_j]\left(\gamma^Y_{ij} - \gamma^X_{ij}\right) \qquad (2)\end{aligned}$$

where $\gamma^X_{ij} = \sigma^X_{ii} + \sigma^X_{jj} - 2\sigma^X_{ij} = \mathbf{E}[(X_i - \mu_i - X_j + \mu_j)^2]$. Of course, the latter is equal to $\mathbf{E}[(X_i - X_j)^2] - (\mu_i - \mu_j)^2$. Since the μ_i are the same for X as for Y we get $\gamma^X_{ij} \le \gamma^Y_{ij}$. Clearly $p_i(x) \ge 0$ too and hence $\mathbf{E}[P_iP_j] \ge 0$. Therefore, $\frac{d}{d\theta}\mathbf{E}[f_\beta(Z_\theta)] \ge 0$ and we get $\mathbf{E}[f_\beta(X)] \le \mathbf{E}[f_\beta(Y)]$. Letting $\beta \uparrow \infty$ we get $\mathbf{E}[X^*] \le \mathbf{E}[Y^*]$. ■

Remark 10. This proof contains another useful idea - to express $\max_i x_i$ in terms of $f_\beta(x)$. The advantage is that f_β is smooth while the maximum is not. And for large β, the two are close because $\max_i x_i \le f_\beta(x) \le \max_i x_i + \frac{\log n}{\beta}$.

If Sudakov-Fernique inequality is considered a modification of Slepian's inequality, the analogous modification of Gordon's inequality is the following. We leave it as exercise as we may not use it in the course.

Exercise 11. (optional) Let $X_{i,j}$ and $Y_{i,j}$ be $n \times m$ arrays of joint Gaussians with equal means. Assume that

1. $\mathbf{E}[|X_{i,j} - X_{i,\ell}|^2] \geq \mathbf{E}[|Y_{i,j} - Y_{i,\ell}|^2]$,
2. $\mathbf{E}[|X_{i,j} - X_{k,\ell}|^2] \leq \mathbf{E}[|Y_{i,j} - Y_{k,\ell}|^2]$ if $i \neq k$.

Then $\mathbf{E}[\min_i \max_j X_{i,j}] \leq \mathbf{E}[\min_i \max_j Y_{i,j}]$.

4.4 Positive association

Definition 12. A random vector $X = (X_1, \ldots, X_n)$ is said to have *positive association* if any two bounded increasing functions of X are positively correlated. That is, if $f, g : \mathbb{R}^n \mapsto \mathbb{R}$ are bounded and non-decreasing in each co-ordinate, then $\mathbf{E}[f(X)g(X)] \geq \mathbf{E}[f(X)]\mathbf{E}[g(X)]$.

Some remarks.

1. If positive association holds, two decreasing functions are also positively correlated. An increasing function of X is negatively correlated with a decreasing function of X.
2. The boundedness condition is only to ensure that the expectations exist. If f, g are increasing in each co-ordinate and all expectations in the definition exist, then let $f_N := (f \wedge N) \vee (-N)$ and $g_N = (g \wedge N) \vee (-N)$. These are bounded functions that are increasing in each co-ordinate, hence $\mathbf{E}[f_N(X)g_N(X)] \geq \mathbf{E}[f_N(X)]\mathbf{E}[g_N(X)]$. By DCT (since $|f_N| \leq |f|$ and $|g_N| \leq |g|$), we see that $\mathbf{E}[f_N(X)]$, $\mathbf{E}[g_N(X)]$, $\mathbf{E}[f_N(X)g_N(X)]$ converge to $\mathbf{E}[f(X)]$, $\mathbf{E}[g(X)]$, $\mathbf{E}[f(X)g(X)]$ respectively as $N \to \infty$, and the inequality extends to f, g.
3. The definition naturally extends to infinite collections of random variables.

Positive association is a very stringent requirement. When it holds, it is a powerful tool. In percolation and certain models of statistical mechanics, this is often known as *FKG inequality* as it was proved by Fortouin, Kastelyn and Ginibre in those settings. One basic well-known example of positive association is the following.

Result 13 (Harris's inequality)**.** Let X_k be independent real valued random variables. Then X is positively associated.

We skip the well-known proof here[3]. Instead, we come to the main result of interest.

Theorem 14 (Loren Pitt (1982)). *Let X be a centered Gaussian vector. Then X is positively associated if and only if all correlations are positive (i.e., $\mathbf{E}[X_iX_j] \geq 0$).*

One direction is obvious (and has nothing to do with Gaussians): Positive association implies positive correlation by applying to the increasing functions $X \mapsto X_k$ and $X \mapsto X_j$. We prove the other direction now.

Proof that positively correlated Gaussians are positively associated. Let X, Y be i.i.d. copies and as in earlier proofs, define the interpolation $Z(\theta) = \cos\theta\, X + \sin\theta\, Y$ for $0 \leq \theta \leq \frac{\pi}{2}$. Let $f, g : \mathbb{R}^n \mapsto \mathbb{R}$ be smooth functions such that $\partial_k f, \partial_k g \geq 0$ for all k. Then $\mathbf{E}[f(X)g(Z(\theta))]$ is equal to $\mathbf{E}[f(X)g(X)]$ when $\theta = 0$ and equal to $\mathbf{E}[f(X)]\mathbf{E}[g(X)]$ when $\theta = \frac{\pi}{2}$. Therefore, it suffices to show that $\theta \mapsto \mathbf{E}[f(X)g(Z(\theta))]$ is decreasing. To this end, consider

$$\frac{d}{d\theta}\mathbf{E}[f(X)g(Z(\theta)] = \sum_{k=1}^{n} \mathbf{E}[f(X)\partial_k g(Z_\theta)(-\sin\theta\, X_k + \cos\theta\, Y_k)]$$

$$= \cos\theta\sin\theta \sum_{k=1}^{n}\sum_{j=1}^{n} (\sigma^Y_{k,j} - \sigma^Y_{k,j})\mathbf{E}[f(X)\partial_j\partial_k g(Z(\theta)] - \cos\theta \sum_{k=1}^{n} \sigma^X_{k,j}\mathbf{E}[\partial_j f(X)\partial_k g(Z(\theta))].$$

This calculation is almost the same as in the proof of Lemma 1, except that when we use integration by parts on $\mathbf{E}[X_k f(X)g(Z(\theta))]$, the partial derivative with respect to X_k can fall on $f(X)$ or on $g(Z(\theta))$ whereas the corresponding partial derivative with respect to Y_k in $\mathbf{E}[Y_k f(X)g(Z(\theta))]$ falls only on $g(Z(\theta))$.

Now, X, Y have the same distribution, hence $\sigma^X_k = \sigma^Y_k$ and the first summand vanishes. In the second summand, $\partial_j f, \partial_k g, \sigma_{k,j}, \cos\theta$ are all positive. Hence the derivative of $\mathbf{E}[f(X)g(Z(\theta))]$ is decreasing in θ. This completes the proof for smooth f, g.

For general bounded functions, a standard approximation argument via smooth functions must be used, but we skip the details here. ■

4.5 Negative association

Negative association is a stronger form of negative correlation just as positive association is a stronger form of positive correlation.

[3] See section 5.8 of Probability on trees and networks by Lyons and Peres, for example.

Definition 15. Let $X = (X_1, \ldots, X_n)$ be a random vector. We say that it is negatively associated if for any disjoint sets $A, B \subseteq [n]$ and any increasing bounded $f : \mathbb{R}^A \mapsto \mathbb{R}$ and increasing bounded $g : \mathbb{R}^B \mapsto \mathbb{R}$, we have $\text{Cov}(f(X_A), g(X_B)) \leq 0$.

Unlike in the case of positive association, disjointness of the "supports" of the two functions is essential. For example, no random variable would be negatively correlated with itself. Independent random variables are trivially negatively associated. An important and intuitive non-trivial example of a negatively associated random vector is this[4].

Example 16. Consider a box with n coupons carrying labels (not necessarily distinct) $u_1, \ldots, u_n$. Sample uniformly at random from this box, *without replacement*, and note the labels to get random variables $X_1, \ldots, X_n$ (in other words, pick a permutation $\pi \in \mathcal{S}_n$ uniformly at random and set $X_k = u_{\pi(k)}$ for all k). Then X_is are negatively associated.

Theorem 17 (Joag-Dev, Proschan)**.** *A Gaussian vector X is negatively associated if and only if $Cov(X_i, X_j) \leq 0$ for all $i \neq j$.*

For $j \neq k$, the functions $f(X) = X_j$ and $g(X) = X_k$ depend on disjoint subsets of variables and are increasing, hence negative correlation is necessary for negative association. It is the other direction that needs proof. We shall prove the following more precise and stronger statement. What it says is that in the language of Theorem 17 and the definition of negative association, to get the conclusion that $\mathbf{E}[f(X_A)g(X_B)] \leq 0$, it suffices to assume that $\text{Cov}(X_i, X_j) \leq 0$ for $i \in A$ and $j \in B$.

Lemma 18. *Let $X_{m\times 1}, W_{n\times 1}$ be jointly Gaussian vectors and let $f : \mathbb{R}^m \mapsto \mathbb{R}$ and $g : \mathbb{R}^n \mapsto \mathbb{R}$ be increasing in each co-ordinate. If $Cov(X_i, W_j) \leq 0$ for all i, j, then $Cov(f(X), g(W)) \leq 0$.*

Proof. By subtracting the means, we may assume that X and W are centered. We also assume at first that f, g are smooth. Without loss of generality, we assume that there is a random vector Y that is independent of (X, W) and has the same distribution as X. As always, interpolate between them with $Z(\theta) = \cos\theta\, X + \sin\theta\, Y$, $0 \leq \theta \leq \frac{\pi}{2}$. We show that $\mathbf{E}[f(Z(\theta))g(W)]$ is increasing by computing its derivative

$$\frac{d}{d\theta}\mathbf{E}[f(Z(\theta))g(W)] = \sum_{k=1}^{m} \mathbf{E}[g(W)\partial_k f(Z(\theta))(-\sin\theta\, X_k + \cos\theta\, Y_k)].$$

[4]For proof of this and more, see the paper *Negative Association of Random Variables with Applications* by Kumar Joag-Dev and Frank Proschan (Annals of Statstics, **11**, No. 1, 286–295, (1983)).

As in the proof of Lemma 1, we condition on Y (or (X,W)) and use the Gaussian integration by parts formula. We get

$$\begin{aligned}
\mathbf{E}[X_k g(W)\partial_k f(Z_\theta)] &= \cos\theta \sum_{j=1}^{m} \mathbf{E}[X_k X_j]\mathbf{E}[g(W)\partial_j\partial_k f(Z_\theta)] + \sum_{j=1}^{n} \mathbf{E}[X_k W_j]\mathbf{E}[\partial_j g(W)\partial_k f(Z(\theta))],\\
\mathbf{E}[Y_k g(W)\partial_k f(Z_\theta)] &= \sin\theta \sum_{j=1}^{m} \mathbf{E}[Y_k Y_j]\mathbf{E}[g(W)\partial_j\partial_k f(Z_\theta)] + \sum_{j=1}^{n} \mathbf{E}[Y_k W_j]\mathbf{E}[\partial_j g(W)\partial_k f(Z(\theta))]\\
&= \sin\theta \sum_{j=1}^{m} \mathbf{E}[Y_k Y_j]\mathbf{E}[g(W)\partial_j\partial_k f(Z_\theta)]
\end{aligned}$$

since $\mathbf{E}[Y_k W_j] = 0$ for all k, j. As X and Y have the same distribution, $\mathbf{E}[X_k X_j] = \mathbf{E}[Y_k Y_j]$ for all k, j. Plugging back into the earlier equation, we get

$$\frac{d}{d\theta}\mathbf{E}[g(W)f(Z_\theta)] = -\sin\theta \sum_{k=1}^{m}\sum_{j=1}^{n} \mathbf{E}[X_k W_j]\mathbf{E}[\partial_j g(W)\partial_k f(Z(\theta))].$$

As f, g are increasing in each co-ordinate, $\partial_j g \geq 0$ and $\partial_j f \geq 0$, hence the second expectation is positive. The first expectation is negative, and $\sin\theta \geq 0$, hence $\mathbf{E}[g(W)f(Z_\theta)]$ is increasing in θ. At $\theta = 0$ this is $\mathbf{E}[f(X)g(W)]$ whereas at $\theta = \frac{\pi}{2}$ this is $\mathbf{E}[g(W)]\mathbf{E}[f(X)]$. Hence, the comparison at these two points shows that $\mathrm{Cov}(f(X), g(W)) \leq 0$.

The boring step of approximating general increasing functions by smooth functions is omitted as exercise. ■

Here is a use of negative association.

Exercise 19. Suppose $X = (X_1, \ldots, X_n)$ is negatively associated. Assume that $\mathbf{E}[X_i] = 0$ and that each X_i is bounded. Then, Hoeffding's inequality holds for $S = X_1 + \ldots + X_n$.

The intuition behind negative correlations extends to the following situation where the functions are not dependent on disjoint sets of variables. How do you deal with it?

Exercise 20. Let $f : \mathbb{R} \mapsto \mathbb{R}_+$ be an increasing function. Let X be a Gaussian vector with negative correlation. Define the random probability vector P with co-ordinates $P_k = f(X_k)/\sum_{j=1}^{n} f(X_j)$. Show that P_1 and P_2 are negatively correlated. Is P negatively associated? (I have not checked the last statement myself).

4.6 Piterbarg's identity

Piterbarg wrote an identity for the difference in the probabilities of a set under two Gaussian distributions. We state the simplest special case of this.

Proposition 21 (Piterbarg). *Let X, Y be centered Gaussian random vectors in $\mathbb{R}^n$ with covariance matrices Σ^X and Σ^Y. Assume that $\sigma^X_{i,i} = \sigma^Y_{i,i}$ for all i. Let $A = (a_1^-, a_1^+] \times \ldots \times (a_n^-, a_n^+]$. Let $Z(\theta) = \cos\theta\ X + \sin\theta\ Y$ and let $\varphi^\theta_{i,j}$ denote the density of $(Z_i(\theta), Z_j(\theta))$. Then, $\mathbf{P}\{Y \in A\} - \mathbf{P}\{X \in A\}$ is equal to*

$$\sum_{i<j}(\sigma^Y_{i,j} - \sigma^X_{i,j}) \sum_{\mu,\lambda\in\{+,-\}} \mu\lambda \int_0^{\pi/2} \varphi^\theta_{i,j}(a_i^\mu, a_j^\lambda)\mathbf{P}\{Z(\theta) \in \overline{A} \mid Z_i(\theta) = a_i^\mu, Z_j(\theta) = a_j^\lambda\}\ \sin\theta\cos\theta\ d\theta$$

Some remarks.

1. In the interpolation, the covariance matrices of $Z(\theta)$ lie on the line segment connecting Σ^X with Σ^Y. As $\det(t\Sigma^X + (1-t)\Sigma^Y)$ is a polynomial in t, there are at most n exceptional values of θ where the covariance matrix of $Z(\theta)$ is singular. At all other θ, the density $\psi^\theta_{i,j}$ exists for all $i < j$. In Piterbarg's statement, which is written in greater generality, he assumes that $\sigma^X_{i,i} = 1 = \sigma^Y_{i,i}$ and $|\sigma^X_{i,j}| < 1$ for $i \neq j$. I am not seeing the need for this in the proof below.

2. I cannot explain the meaning of the inequality. One point is that the terms on the right hand side comes from cases when $Z(\theta) \in \overline{A}$ but only barely - two of the coordinates are at the boundaries of their intervals. Also see the corollary below.

3. The expression on the right appears very complicated. The way to use it is to get bounds on the integrals, which then shows that if Σ^X and Σ^Y are entry-wise close, then $\mathbf{P}\{X \in A\}$ and $\mathbf{P}\{Y \in A\}$ are also close. For example, it could be that $\sigma^X_{i,j}$ is small for all $i \neq j$, and we wish to see if we can replace X_i by independent Gaussians having the same variances.

If we let $a_i^- \to -\infty$ for all i, then we get the following limiting formulation (since the densities $\varphi^\theta_{i,j}(a_i^\mu, a_j^\lambda)$ goes to zero if one of λ or μ is '$-$'.

Corollary 22. *Let X, Y be as in Proposition 21 and let $A = (-\infty, a_1] \times \ldots \times (-\infty, a_n]$. Then, $\mathbf{P}\{X \in A\} - \mathbf{P}\{Y \in A\}$ is equal to*

$$\sum_{i<j}(\sigma^Y_{i,j} - \sigma^X_{i,j}) \int_0^{\pi/2} \varphi^\theta_{i,j}(a_i, a_j)\mathbf{P}\{Z(\theta) \in \overline{A} \mid Z_i(\theta) = a_i, Z_j(\theta) = a_j\}\ \sin\theta\cos\theta d\theta.$$

In fact, the proposition can be deduced from this corollary by inclusion-exclusion formula.

Proof. Fix $\varepsilon > 0$ and let ψ_i be a smooth function such that $\mathbf{1}_{(a_i^-, a_i^+]} \le \psi_i \le \mathbf{1}_{(a_i^- - \varepsilon, a_i^+ + \varepsilon]}$. Let $f_\varepsilon(x) = \prod_{i=1}^n \psi_i(x_i)$. From (1),

$$\frac{d}{d\theta}\mathbf{E}[f_\varepsilon(Z(\theta))] = \sin\theta\cos\theta \sum_{i<j}(\sigma^Y_{i,j} - \sigma^X_{i,j})\mathbf{E}[\partial_i\partial_j f_\varepsilon(Z(\theta))]$$

and hence $\mathbf{E}[f_\varepsilon(Y)] - \mathbf{E}[f_\varepsilon(X)]$ is the integral of the right side quantity over $\theta \in [0, 2\pi]$.

Now, $\partial_i\partial_j f_\varepsilon(x) = \psi_i'(x_i)\psi_j'(x_j) \prod_{k\neq i,j} \psi_k(x_k)$. Observe that ψ_k' vanishes outside the intervals $I_k^- := (a_k^- - \varepsilon, a_k^-)$ and $I_k^+ := (a_k^+, a_k^+ + \varepsilon)$. Thus,

$$\begin{aligned}\mathbf{E}[\partial_i\partial_j f_\varepsilon(Z(\theta))] &= \sum_{\mu,\lambda\in\{+,-\}} \mathbf{E}\left[\mathbf{1}_{I_i^\mu}(Z_i(\theta))\mathbf{1}_{I_j^\lambda}(Z_j(\theta))\psi_i'(Z_i(\theta))\psi_j'(Z_j(\theta))\prod_{k\neq i,j}\psi_k(Z_k(\theta))\right] \\ &= \sum_{\mu,\lambda\in\{+,-\}} \int_{I_i^\mu}\int_{I_j^\lambda} \psi_i'(u)\psi_j'(v)\mathbf{E}\left[\prod_{k\neq i,j}\psi_k(Z_k(\theta)) \;\middle|\; Z_i(\theta) = u,\ Z_j(\theta) = v\right]\varphi^\theta_{i,j}(u,v)\,du dv.\end{aligned}$$

We could choose ψ_i so that ψ_i' has constant sign on each I_i^μ (positive if $\mu = -$ and negative if $\mu = +$). Since the integral of ψ_i' over such an interval is ± 1, the (μ,λ) term above can be considered an average over $I_i^\mu \times I_j^\lambda$ of the function (up to sign)

$$(u,v) \mapsto \mathbf{E}\left[\prod_{k\neq i,j}\psi_k(Z_k(\theta)) \;\middle|\; Z_i(\theta) = u,\ Z_j(\theta) = v\right]\varphi^\theta_{i,j}(u,v).$$

This is a continuous function of (u,v) (except at the exceptional θ values referred to earlier), hence as $\varepsilon \downarrow 0$, the integral over $I_i^\mu \times I_j^\lambda$ converges to

$$\mu\lambda\mathbf{E}\left[\prod_{k\neq i,j}\mathbf{1}_{I_k^\varepsilon}(Z_k(\theta)) \;\middle|\; Z_i(\theta) = a_i^\mu,\ Z_j(\theta) = a_j^\lambda\right]\varphi^\theta_{i,j}(a_i^\mu, a_j^\lambda).$$

It is obvious that $\mathbf{E}[f_\varepsilon(Y)] - \mathbf{E}[f_\varepsilon(X)]$ is converges to $\mathbf{P}\{Y \in A\} - \mathbf{P}\{X \in A\}$. Putting everything together, we get the claim in the statement of the proposition. ∎

Remark 23. If $\mathcal{G}$ is the algebra generated by the collection of left-open, right-closed rectangles of the form considered in the proposition above. Any element of $\mathcal{G}$ can be written as a finite, pairwise disjoint union of such rectangles. Hence the result of the proposition can be added for the rectangles.

4.7 Going beyond finite vectors

Now that we have stated all the general comparison inequalities that we want, it is worth stating that while the essence of things is already in finite dimensional Gaussian vectors, all these comparison inequalities also apply to general Gaussian processes. Let us indicate how.

Assume that $X = (X_I)_{i\in I}$ and $Y = (Y_i)_{i\in I}$ are Gaussian processes. If they have sufficient regularity so that $X^* = \sup_i X_i$ and $Y^* = \sup_i Y_i$ are measurable, then it shall also be the case that $X^* = \sup\{\mathbf{E}[X_F^*] : F \subseteq I \text{ is finite}\}$ where $X_F = (X_i)_{i\in F}$. Similar expression for Y^*.

Now suppose the means of X_i and Y_i agree for all i, and $\mathbf{E}[(X_i - X_j)^2] \leq \mathbf{E}[(Y_i - Y_j)^2]$ for all i, j. Then by Sudakov-Fernique inequality, $\mathbf{E}[X_F^*] \leq \mathbf{E}[Y_F^*]$ for all finite $F \subseteq I$. Clearly that implies that $\mathbf{E}[X^*] \leq \mathbf{E}[Y^*]$.

Similar considerations apply to Slepian's inequality, Gordon's inequality, positive association, etc. We shall use these inequalities without further comment in future. The only subtlety one should be aware of (in 'real examples' that never arises!) is that without any conditions on the process, X^* need not be a random variable (back to the poorness of the cylinder sigma algebra).

4.8 Application: Mean width of convex bodies

Consider the problem of maximizing or minimizing $\mathbf{E}[X^*]$ among all Gaussian vectors $X \sim N_n(0,\Sigma)$ for which $\sigma_{k,k} = 1$ for all k. Equating the variances to 1 provides the right normalization so that the comparison makes sense. In this section, let us refer to these as the admissible Gaussians.

The minimization question: Let $Y \sim N_n(0, J_n)$ where J_n is the all ones matrix. In other words, $Y_1 = \ldots = Y_n \sim N(0,1)$. Clearly, for any admissible Gaussian X, we have $\mathbf{E}[X_i X_j] \leq \mathbf{E}[Y_i Y_j]$ for any i, j. And the means and variances of X_i agree with those of Y_i, for each i. Hence, Slepian's inequality applies and we see that $\mathbf{E}[X^*] \geq \mathbf{E}[Y^*]$. In fact, there is even a stochastic comparison. Thus, Y is the solution to the minimization problem.

The maximization problem: By the same logic, to push the expectation of X^* as high as possible, we should make the covariances as low as possible. However, we cannot make all cross-covariances equal to -1, as the resulting matrix is not p.s.d. That leaves us with

a dilemma - is it better to make some of them very negative or make all of them equally (but less) negative? It is an unproved conjecture that the latter is better.

Conjecture 24 (Gritzmann–Klee, as reformulated by Kabluchko–Litvak–Zaporozhets)**.** *For any admissible Gaussian,* $\mathbf{E}[X^*] \le \mathbf{E}[Y^*]$ *where* $Y \sim N_n(0, \frac{1}{n-1}(nI_n - J_n))$*. That is,* $\mathbf{E}[Y_iY_j] = -\frac{1}{n-1}$ *for all* $i \ne j$*.*

The original formulation of Gritzmann and Klee concerned the *mean width of convex bodies* among all convex bodies. If K is a convex body (a compact convex set with non-empty interior) in $\mathbb{R}^n$, its mean width is defined as the expected length of the projection of a convex body in a uniformly chosen random direction. The conjecture was that among all convex hulls of n points on $\mathbb{S}^{n-1}$, the one that maximizes the mean width is the regular simplex centered at the origin.

If $Z \sim \gamma_n$, then one can show that the mean width is equal to $\sqrt{2\pi}\mathbf{E}[\max_{u\in K}\langle Z,u\rangle]$ (Sudakov's formula). Hence the conjecture about mean-width can be transformed to the above conjecture about Gaussians.

Exercise 25. Prove Sudakov's formula and derive the equivalence of the two conjectures.

While the conjecture is open, here is an exercise (actually a theorem of Kabluchko–Litvak–Zaporozhets) that suggests why the dilemma stated earlier is resolved in this way.

Proposition 26. *Assume* $n = 2m$ *is even and let* X, Y *be centered Gaussian vectors with unit variances and (1)* $\mathbf{E}[Y_iY_j] = -\frac{1}{2n-1}$ *for all* $i \ne j$*, (2)* $\mathbf{E}[X_{2i-1}X_{2i}] = -1$ *for* $1 \le i \le m$ *and all other cross-covariances are zero.Then* $\mathbf{E}[X^*] \le \mathbf{E}[Y^*]$*.*

Proof. Let X, Y be independently constructed on a common probability space and introduce the usual interpolation $Z(\theta) = \cos\theta\ X + \sin\theta\ Y$, $0 \le \theta \le \frac{\pi}{2}$. From (2) in the proof of the Sudakov-Fernique inequality, we have

$$\frac{d}{d\theta}\mathbf{E}[f_\beta(Z(\theta))] = \sin\theta\ \cos\theta \sum_{i<j} \mathbf{E}[P_iP_j]\left(\gamma_{ij}^Y - \gamma_{ij}^X\right)$$

where $f_\beta(x) = \frac{1}{\beta}\log\sum_{k=1}^n e^{\beta x_k}$ and $P_i = e^{\beta Z_i(\theta)}/\sum_{k=1}^n e^{\beta Z_i(\theta)}$ and

$$\begin{aligned}\gamma_{i,j}^Y - \gamma_{i,j}^X &= \mathbf{E}[(Y_i - Y_j)^2] - \mathbf{E}[(X_i - X_j)^2] \\ &= \begin{cases} 2 - \frac{2}{n-1} & \text{if } \{i,j\} = \{2k, 2k-1\} \text{ for some } k \le m, \\ -\frac{2}{n-1} & \text{otherwise.} \end{cases}\end{aligned}$$

Therefore,

$$\frac{1}{\sin\theta\,\cos\theta}\frac{d}{d\theta}\mathbf{E}[f_\beta(Z(\theta))] = -\frac{2}{n-1}\sum_{i<j}\mathbf{E}[P_iP_j] + 2\sum_{k=1}^{m}\mathbf{E}[P_{2k-1}P_{2k}]$$
$$= -\frac{1}{n-1}(1-\sum_{i=1}^{n}\mathbf{E}[P_i^2]) + 2\sum_{k=1}^{m}\mathbf{E}[P_{2k-1}P_{2k}]$$

since $\sum_i P_i = 1$. By symmetry, it is clear that at any θ, the variables P_i are identically distributed and the variables (P_{2k-1}, P_{2k}) are identically distributed. Therefore,

$$\frac{1}{\sin\theta\,\cos\theta}\frac{d}{d\theta}\mathbf{E}[f_\beta(Z(\theta))] = \frac{1}{n-1}\left\{-1+n\mathbf{E}[P_1^2]+n(n-1)\mathbf{E}[P_1P_2]\right\}.$$

Now we wish to claim that $\mathbf{E}[P_1P_2] \le \mathbf{E}[P_iP_j]$ for any $i \neq j$ (the idea being that P_1, P_2 are more negatively correlated than P_i, P_j). Granting this, the right hand side of the above expression is bounded by

$$-1+\sum_{i=1}^{n}\mathbf{E}[P_i^2]+2\sum_{i<j}\mathbf{E}[P_iP_j] = -1+\mathbf{E}[(\sum_i P_i)^2] \;=\; 0.$$

It remains to show that[5] $\mathbf{E}[P_1P_2] \le \mathbf{E}[P_1P_3]$. ■

4.9 Application: Kahane's convexity inequality

Here is the only application I know so far, where the extra power of Lemma 1 over Corollary 2 is required. This lemma is crucially used in the theory of Gaussian multiplicative chaos.

Lemma 27 (Kahane's convexity inequality)**.** *Let X and Y be centered Gaussian vectors in $\mathbb{R}^n$. Assume that $\mathbf{E}[X_iX_j] \le \mathbf{E}[Y_iY_j]$ for all i, j. Then, for any $p \in \mathbb{R}^n_+$ and any convex $f : \mathbb{R}_+ \mapsto \mathbb{R}$ that grows slowly enough, we have*

$$\mathbf{E}\left[f\left(\sum_{k=1}^{n} p_k e^{X_k - \frac{1}{2}\mathbf{E}[X_k^2]}\right)\right] \le \mathbf{E}\left[f\left(\sum_{k=1}^{n} p_k e^{Y_k - \frac{1}{2}\mathbf{E}[Y_k^2]}\right)\right]$$

[5] In their paper, Kabluchko-Litvak-Zaporozhets invoke negative association (not clear how, but believable) to say that $\mathbf{E}[P_1P_2] \le \mathbf{E}[P_1]\mathbf{E}[P_2]$. But then, the desired derivative is negative if and only if $\mathbf{E}[P_1^2] \le \frac{1}{n^2}$, whereas $\mathbf{E}[P_1^2] \ge \frac{1}{n^2}$, by Cauchy-Schwarz inequality.

Proof. Define $F(x,A) = f\left(\sum_{k=1}^n p_k e^{x_k - \frac{1}{2}a_{k,k}}\right)$. As $\sigma^Y_{i,j} \geq \sigma^X_{i,j}$ for all i,j, to apply Lemma 1 to obtain the desired comparison, we only need to check that for any $i \neq j$,

$$(\partial_i\partial_j + \overline{\partial}_{(i,j)})F \geq 0 \qquad \text{and} \qquad (\partial_i^2 + 2\overline{\partial}_{(i,i)})F \geq 0.$$

Denoting $w = \sum_{k=1}^n p_k e^{x_k - \frac{1}{2}a_{k,k}}$, direct computation gives

$$\begin{aligned}
\partial_j F(x) &= f'(w) p_j e^{x_j - \frac{1}{2}a_{j,j}} \\
\partial_i\partial_j F(x) &= f''(w) p_i p_j e^{x_i + x_j - \frac{1}{2}a_{i,i} - \frac{1}{2}a_{j,j}} + \delta_{i,j} f'(w) p_j e^{x_j - \frac{1}{2}a_{j,j}} \\
\overline{\partial}_{(i,j)} F(x) &= -\delta_{i,j}\frac{1}{2} f'(w) p_j e^{x_j - \frac{1}{2}a_{j,j}}.
\end{aligned}$$

As $f'' \geq 0$, the first summand in $\partial_i\partial_j F(x)$ is non-negative for any i,j. That first term is equal to $(\partial_i\partial_j + \overline{\partial}_{(i,j)})F$ if $i \neq j$, and equal to $(\partial_i^2 + 2\overline{\partial}_{(i,i)})F$ if $i = j$. Hence the conditions of Lemma 1 are satisfied and we get the claimed inequality. ■

4.10 Application: Eigenvalues of random matrices

Here we present[6] few applications of the basic results on Gaussian processes, namely concentration of measure and comparison theorems.

Extreme singular values of a rectangular Gaussian matrix: Let $A_{m,n} = (a_{i,j})_{i \leq m, j \leq n}$ be a matrix whose entries are i.i.d. $N(0,1)$. We assume $m \leq n$ and denote the singular values of A by $s_1 \leq s_2 \leq \ldots \leq s_m$ (by definition s_i^2 are the eigenvalues of AA^t). The following result gives bounds for the smallest and largest singular values.

Theorem 28 (Gordon). *With A as above, $\mathbf{E}[s_1] \geq \sqrt{n} - \sqrt{m}$ and $\mathbf{E}[s_n] \leq \sqrt{n} + \sqrt{m}$.*

Proof. For $(u,v) \in T := S^{m-1} \times S^{n-1}$ define $X(u,v) = u^t A v = \sum_{i=1}^m \sum_{j=1}^n a_{i,j} u_i v_j$. It has zero mean and $\mathbf{E}[|X_{u,v} - X_{u',v'}|^2] = 2 - 2\langle u,u'\rangle\langle v,v'\rangle$ (check!).

Consider a different Gaussian process on the same index set defined by $Y(u,v) = \sum_{i=1}^m u_i \xi_i + \sum_{j=1}^n v_i \eta_i$ where ξ_i, η_i are i.i.d. $N(0,1)$. Then $\mathbf{E}[|Y_{u,v} - Y_{u',v'}|^2] = |u-u'|^2 + |v-v'|^2 = 4 - 2\langle u,u'\rangle - 2\langle v,v'\rangle$. Both X and Y are continuous on T and hence the comparison theorems are applicable.

[6] This material is taken from the paper *Local operator theory, random matrices and Banach spaces*, by Davidson and Szarek. Roman Vershynin has several lecture notes that cover this and much more.

Thus,

$$\mathbf{E}[|Y_{u,v} - Y_{u',v'}|^2] - \mathbf{E}[|X_{u,v} - X_{u',v'}|^2] = 2(1 - \langle u, u' \rangle)(1 - \langle v, v' \rangle)$$

which is non-negative for all $(u,v),(u',v') \in T$. Therefore, by the Sudakov-Fernique inequality we get $\mathbf{E}[X^*] \leq \mathbf{E}[Y^*]$. Clearly $Y^* \leq \|\xi\| + \|\eta\|$ and $\mathbf{E}[\|\xi\|] \leq \sqrt{\mathbf{E}[\|\xi\|^2]} = \sqrt{m}$ and $\mathbf{E}[\|\eta\|] \leq \sqrt{\mathbf{E}[\|\eta\|^2]} = \sqrt{n}$. But X^* is precisely s_m. Therefore $\mathbf{E}[s_m] \leq \sqrt{n} + \sqrt{m}$.

Next observe that $s_1 = \min_u \max_v X_{u,v}$. We have already seen that

$$\mathbf{E}[|Y_{u,v} - Y_{u',v'}|^2] \geq \mathbf{E}[|X_{u,v} - X_{u',v'}|^2] \text{ for all } u, v, u', v',$$
$$\mathbf{E}[|Y_{u,v} - Y_{u,v'}|^2] = \mathbf{E}[|X_{u,v} - X_{u,v'}|^2] \text{ for all } u, v, v'.$$

For the second, observe that $\langle u, u' \rangle = 1$ when $u = u'$. Gordon's inequality applies to give $\mathbf{E}[s_1] \geq \mathbf{E}[\min_u \max_v Y_{u,v}]$. As the last step in the proof, observe that picking $v = \eta/\|\eta\|$ and $u = -\xi/\|\xi\|$ achieves the $\min_u \max_v Y_{u,v}$ and gives $\mathbf{E}[\min_u \max_v Y_{u,v}] = \mathbf{E}[\|\eta\|] - \mathbf{E}[\|\xi\|]$. Since $\|\eta\|^2 \sim \chi^2_{n-1}$,

$$\mathbf{E}[\|\eta\|] = \frac{1}{2^{n/2}\Gamma(n/2)} \int_0^\infty \sqrt{x} e^{-x} x^{\frac{n}{2}-1} dx = \frac{\sqrt{2}\Gamma(\frac{n+1}{2})}{\Gamma(\frac{n}{2})}$$

and similarly $\mathbf{E}[\|\xi\|] = \frac{\sqrt{2}\Gamma(\frac{m+1}{2})}{\Gamma(\frac{m}{2})}$. Thus the theorem is proved if we show that $\mathbf{E}[\|\eta\|] - \mathbf{E}[\|\xi\|] \geq \sqrt{n} - \sqrt{m}$. Deduce this from Exercise 29. ■

Exercise 29. Show that $\nu \to \frac{\sqrt{2}\Gamma(\frac{\nu+1}{2})}{\Gamma(\frac{\nu}{2})} - \sqrt{\nu}$ is increasing for $\nu \geq 1$.

Location of individual singular values of a Gaussian matrix: Let $A_{m,n}$ be a real symmetric matrix such that $a_{i,j}$, $i \leq j$ are i.i.d. $N(0,1)$ (it is okay to allow the diagonals to have variance 2 to make it exactly a GOE matrix). Let $\lambda_{n,1} < \ldots < \lambda_{n,n}$ be the eigenvalues of $A_n/\sqrt{n}$ (normalized so that the empirical distribution of eigenvalues converges to the semicircle distribution as n tends to infinity)

Theorem 30. *There exist deterministic numbers $t_{n,k}$ such that $\mathbf{P}\{|\lambda_{n,k} - t_{n,k}| \geq u\} \leq Ce^{-cnu^2}$ for all $k \leq n$.*

Proof. Recall the min-max representation

$$\lambda_{n,n-k+1} = \frac{1}{\sqrt{n}} \min_{\mathbf{v}_1,\ldots,\mathbf{v}_{k-1}} \max_{\mathbf{u}:\mathbf{u}\perp\mathbf{v}_j} \mathbf{u}^t A \mathbf{u}.$$

From this (since $A \to \mathbf{u}^t A\mathbf{u}$ is linear for each $\mathbf{u}$) it follows that the function $(a_{i,j})_{i\le j\le n} \to \lambda_{n,n-k+1}$ is $\mathrm{Lip}(2/\sqrt{n})$. By the Gaussian concentration inequality, if $t_{n,n-k+1}$ is a median of $\lambda_{n,n-k+1}$ then $\mathbf{P}\{|\lambda_{n,n-k+1} - t_{n,n-k+1}| \ge u\} \le 2\bar{\Phi}(u/2\sqrt{n}) \le 2e^{-nu^2/8}$. ■

Remark 31. The well-known Wigner's semicircle law says that the histogram of eigenvalues is close to the semi-circle density $\frac{1}{2\pi}\sqrt{4-x^2}$. This does not imply a quantitative estimate for the location of individual eigenvalues. In contrast, the above theorem shows that each eigenvalue is concentrated in a window of length essentially $1/\sqrt{n}$. However the actual facts (proved by harder methods specific to the problem) are that eigenvalues are concentrated in even smaller windows (of length $1/n$ if k is away from 1 and n and of length $n^{-2/3}$ if k is close to 1 or n).

4.11 Application: Persistence probability

Let $X = (X_n)_{n\in\mathbb{Z}}$ be a stationary Gaussian process. Define the *persistence probability*

$$H_X(n) = \mathbf{P}\{X_1 > 0, \dots, X_n > 0\}.$$

This is the probability that the process persists above level zero. It is a quantity that has been studied considerably. Here we get a lower bound for positively correlated processes.

Claim 32. *Assume that* $\mathbf{E}[X_n] = 0$ *and* $\mathbf{E}[X_n X_m] \ge 0$ *for all* m,n. *Then* $H_X(n) \ge 2^{-n}$.

Proof. Fix $m,n \ge 0$ and let X,Y be two i.i.d. copies of the process. Observe that $U = (X_1,\dots,X_{m+n})$ and $V = (X_1,\dots,X_m,Y_{m+1},\dots,Y_{m+n})$ have equal means and variances, and $\mathbf{E}[U_iU_j] \ge \mathbf{E}[V_iV_j]$ for all i,j. Therefore, by Slepian's inequality, we see that $\mathbf{P}\{U^* < 0\} \ge \mathbf{P}\{V^* < 0\}$, which is the same as $H_X(m+n) \ge H_X(m)H_X(n)$. In particular, $H_X(n) \ge H_X(1)^n = 2^{-n}$. ■

In fact, by Fekete's lemma, $\frac{1}{n}\log H_X(n)$ exists and is equal to $\kappa := \sup_n \frac{1}{n}H_X(n)$. Consequently, if one can calculate $H_X(p)$ for a specific p, then one can get the better bound $\frac{1}{p}\log H_X(p)$ for κ.

The positivity of correlations used here is necessary (in general; we do not mean that it is always needed).

Example 33. Let Z_n be i.i.d. standard Gaussians and let $X_n = Z_n - Z_{n-1}$. Then X is a stationary Gaussian process and $H_X(n) = \mathbf{P}\{Z_0 < Z_1 < \dots < Z_n\} = \frac{1}{n!}$, which decays faster than exponentially. In this example, $\mathbf{E}[X_nX_{n-1}] = -1$, not positive.

4.12 Application: Mixing

Let $X = (X_n)_{n\in\mathbb{Z}}$ be a centered stationary Gaussian process with $\kappa(n) = \mathbf{E}[X_m, X_{m+n}]$. Without loss of generality, assume that $\kappa(0) = 1$. If the covariance decays fast, we expect the process at far off times to be approximately independent. To make a precise definition, introduce the sigma-algebras $\mathcal{F}_t = \sigma\{X_k : k \le t\}$ and $\overline{\mathcal{F}}_t = \sigma\{X_k : k \ge t\}$. These are generated by the corresponding algebras $\mathcal{A}_t$ and $\overline{\mathcal{A}}_t$ that are generated by left-open, right closed cylinders. More precisely, $\mathcal{A}_t$ consists of finite unions of disjoint cylinder sets of the form $\cap_{i=1}^{p}\{a_j < X_{t_j} \le b_j\}$ where $p \ge 1$, $t_j \le t$, $a_j, b_j \in \mathbb{R}$. The algebra $\overline{\mathcal{A}}_t$ is similar, except that we require $t_j \ge t$.

The mixing coefficient is defined as

$$\begin{aligned}\alpha(T) &:= \sup\{|\mathbf{P}(A\cap B) - \mathbf{P}(A)\mathbf{P}(B) : A \in \mathcal{F}_0,\, B \in \overline{\mathcal{F}}_T\}\\ &= \sup\{|\mathbf{P}(A\cap B) - \mathbf{P}(A)\mathbf{P}(B) : A \in \mathcal{A}_0,\, B \in \overline{\mathcal{A}}_T\}.\end{aligned}$$

The equality comes from the general fact that if $\mathbf{P}$ is a probability measure on a sigma-algebra $\mathcal{F}$ generated by an algebra $\mathcal{A}$, then for any $A \in \mathcal{F}$ and $\varepsilon > 0$, there is a $B \in \mathcal{A}$ such that $\mathbf{P}(A\Delta B) < \varepsilon$.

A sufficient condition for mixing in stationary Gaussian sequences: We need to bound

$$\mathbf{P}\{X_{-i} \le a_i, 1 \le i \le m,\, X_{T+j} \le b_j, 1 \le j \le n\} - \mathbf{P}\{X_{-i} \le a_i, 1 \le i \le m\}\mathbf{P}\{X_{T+j} \le b_j, 1 \le j \le n\}$$

We apply Piterbarg's identity to $U = ((X_{-i})_{i\le m}, (X_{T+i})_{i\le n})$ and $V = ((X_{-i})_{i\le m}, (Y_{T+i})_{i\le n})$, where Y is an independent copy of X. The left hand side of the identity is exactly the difference in probabilities that we want. On the right, many terms vanish and we are left with (here $\varphi_{i,j}$ is the density of (X_{-i}, X_{T+j}))

$$\sum_{i=1}^{m}\sum_{j=1}^{n}(\kappa(T+j+i)-1)\int_0^{\pi/2}\varphi_{i,j}^{\theta}(a_i,b_j)\mathbf{P}\{\star\big|Z_i(\theta) = a_i, Z_{T+j}(\theta) = b_j\}\,\sin\theta\cos\theta\,d\theta$$

which can be bounded by

$$C\sum_{i,j\ge 1}|\kappa(T+i+j)|$$

where $C = \sup_{i,j\ge1}\|\varphi_{i,j}^{\theta}\|_{\sup}$. Observe that $\varphi_{i,j}^{\theta}$ is bounded by $\frac{1}{2\pi\sqrt{1-\kappa(T+j+i)^2}}$. Hence, if $\kappa(t) \to 0$ as $t \to \infty$, then for large enough t, this can be bounded by 2. Thus,

$$\alpha(T) \le C\sum_{i,j\ge1}|\kappa(T+i+j)| \le C\sum_{\ell\ge1}\ell\,|\kappa(T+\ell)|.$$

If $\sum_j j|\kappa(j)| < \infty$, then this quantity goes to zero as $T \to \infty$. Hence we arrive at

Proposition 34. *Let X be a a centered stationary Gaussian process with correlation function $\kappa(t) = \mathbf{E}[X_0 X_t]$. If $\sum_{j\geq 1} j|\kappa(j)| < \infty$, then $\alpha(T) \to 0$ as $T \to \infty$.*

4.13 Application: LLN for maximum

For the i.i.d. sequence Z_n, we have seen that with $M_n = \max\{Z_1, \ldots, Z_n\}$, then,

$$\frac{M_n}{\sqrt{2\log n}} \xrightarrow{P} 1.$$

In fact we proved the stronger statement $M_n - \sqrt{2\log n} \xrightarrow{P} 0$. To what extend these asymptotics extend to approximately independent Gaussian sequences? Berman found pleasantly simple and mild conditions.

Let $X_0, X_1, \ldots$ be jointly Gaussian random variables with zero means and unit variances. Let $K(m,n) = \mathbf{E}[X_m X_n]$ and let $r(n) = \sup_m K(m, m+n)$. If X is a stationary sequence, then $K(m,n) = r(n-m)$ for all m,n. Let $M_n^X = \max\{X_0, \ldots, X_{n-1}\}$.

Theorem 35 (Berman)**.** *If $r(n) \to 0$ as $n \to \infty$, then $\frac{M_n^X}{\sqrt{2\log n}} \xrightarrow{P} 1$.*

One side of this is trivial. If $M_n^X > (1+\varepsilon)\sqrt{2\log n}$, then $X_i > (1+\varepsilon)\sqrt{2\log n}$ for some $0 \leq i \leq n-1$. Union bound shows that

$$\mathbf{P}\{M_n^X > (1+\varepsilon)\sqrt{2\log n}\} \leq n e^{-(1+\varepsilon)^2 \log n} \leq \frac{1}{n^{2\varepsilon}}. \tag{3}$$

What remains is to show that $\mathbf{P}\{M_n^X > (1-\varepsilon)\sqrt{2\log n}\} \to 1$ for any $\varepsilon > 0$. We shall need the following claim.

Exercise 36. Let $Y_1, \ldots, Y_n$ be jointly Gaussian with $\mathbf{E}[Y_i] = 0$, $\mathbf{E}[Y_i^2] = 1$ and $\mathbf{E}[Y_i Y_j] = \rho > 0$ for all $i \neq j$. Show that for any $\varepsilon > 0$, there is a sequence $\delta_n(\varepsilon) \to 0$ such that

$$\mathbf{P}\{M_n^Y \geq \sqrt{2(1-\rho-\varepsilon)\log n}\} \geq 1 - \delta_n(\varepsilon).$$

Proof of Berman's theorem. Fix $\rho > 0$ and find m such that Since $|r(k)| < \rho$ for all $k \geq m$. Consider the vector $(X_0, X_m, X_{2m}, \ldots, X_{(\ell-1)m})$ and compare it to the vector Y in the exercise above. The means and variances are the same, and Y_is are more positively correlated, hence by Slepian's inequality,

$$\mathbf{P}\{\max_{0\leq j\leq \ell-1} X_{jm} \geq \sqrt{2(1-2\rho)\log \ell}\} \geq 1 - \delta_\ell$$

for some $\delta_\ell \to 0$. Consequently, taking $\ell_n = \lfloor n/m \rfloor$ as a function of n, we have

$$\mathbf{P}\{M_n^X \geq \sqrt{2(1-2\rho)\log n}\} \geq 1 - \delta_{\ell_n}.$$

since $M_n^X \geq \max\{X_{jm} : 0 \leq j \leq \ell_n - 1\}$. As $\delta_{\ell_n} \to 0$ and ρ is arbitrary, together with (3) this completes the proof. ■

Remark 37. Berman also shows that under the stronger assumption (say for stationary Gaussian sequence) that $nr(n) \to 0$, the stronger conclusion $M_n - \sqrt{2\log n} \xrightarrow{P} 0$ holds.

Remark 38. What happens in continuous time? If $X = (X_t)_{t\in\mathbb{R}}$ is a stationary Gaussian process, then so is $X_a := (X_{an})_{n\in\mathbb{Z}}$, for any $a > 0$. Consequently $M^X(T) = \sup\{X_t : 0 \leq t \leq T\}$ is at least as large as $M^{X_a}(T)$. Hence, under the condition $\mathrm{Cov}(X_0, X_t) \to 0$ as $t \to \infty$, the lower bound of $\sqrt{2\log T}$ remains valid in continuous time.

Without regularity of paths, one can have $M^X(T) = \infty$ a.s. for all $T > 0$. Hence, to get an upper bound, some regularity is required. After that, the upper bound of $\sqrt{2\log T}$ can be deduced.

Chapter 5

Boundedness and continuity of a Gaussian process

5.1 The questions

Let $X = (X_t)_{t\in T}$ be a centered Gaussian process on a probability space $(\Omega, \mathcal{F}, \mathbf{P})$. Two important questions that we shall address in this chapter: Can the Gaussian process be constructed so that for *a.e.* $\omega \in \Omega$, the sample paths are (1) bounded? (2) continuous? For the second question to make sense, T must be assumed to have a topology. The continuity question is natural, but why boundedness? It is really the study of $X^* := \sup_{t\in T} X_t$ and $\|X\| = \sup_{t\in T} |X_t|$, which we have studied in specific situations earlier. Even the continuity question, which is about the study of the (uniform) *modulus of continuity*

$$\omega_X(\varepsilon) := \sup\{X_t - X_s : t, s \in T,\ d(t,s) \le \varepsilon\}$$

as $\varepsilon \downarrow 0$, is about the supremum of the Gaussian process $X_t - X_s$ on the set $\{(t,s) : d(t,s) \le \varepsilon\}$.

Irrespective of whether T comes with a metric or not, a big role will be played by the metric $\tau(s,t) = \|X_s - X_t\|_{L^2(\mathbf{P})}$ that comes by pulling back the metric under the curve $t \mapsto X_t$ in the Hilbert space $L^2(\mathbf{P})$. Observe that τ may be a pseudo-metric, but this will not cause any problems. Eg., If $X_t = Z$ for all t, then $K(t,s) = 1$ for all t, s and $\tau(t,s) = 0$ for all t, s.

We make a few preliminary observations.

1. As a standing convention, assume that X^* is measurable so that quantities such as $\mathbf{P}\{X^* > u\}$ and $\mathbf{E}[X^*]$ make sense. Alternately, simply assume that T is countable, in which case X^* is indeed measurable. In all cases of interest (e.g., $T = \mathbb{R}$), one can

work with a countable dense set and then extend the conclusions to the full space in the end.

2. Firstly, bounds on tail of X^* and bounds on the expected value of X^* imply each other.

   - By Markov's inequality, $\mathbf{P}\{X^* \geq u\} \leq \mathbf{E}[(X^*)_+]/u$.
   - We have $\mathbf{E}[(X^*)_+] = \int_0^\infty \mathbf{P}\{X^* > u\}du$. The lower side never poses a problem since $X^* \geq X_{t_0}$ for any $t_0 \in T$ and hence $\mathbf{E}[(X^*)_-] \leq \mathbf{E}[(X_{t_0})_-] = \sqrt{K(t_0,t_0)}/\sqrt{2\pi}$.

3. Obviously $(X^*)_+ \leq \|X\|$. For a bound the other way, observe that for any $t_0 \in T$,

   $$\|X\| \leq |X_{t_0}| + \sup_{t\in T}(X_t - X_{t_0}) + \sup_{t\in T}(X_{t_0} - X_t).$$

   Hence it suffices to study X^*, for example, $\mathbf{E}[\|X\|] \leq 2\mathbf{E}[X^*] + \mathbf{E}[|X_{t_0}|]$ which is $2\mathbf{E}[X^*] + \sqrt{K(t_0,t_0)}\sqrt{\frac{2}{\pi}}$.

4. For boundedness, a necessary condition is that $\sigma_T^2 := \sup_t \mathbf{E}[X_t^2]$ be finite. Indeed, for any fixed $t \in T$

   $$\mathbf{P}\{X^* \geq \sigma_t\} \geq \mathbf{P}\{X_t \geq \sigma_t\} = \overline{\Phi}(1)$$

   is a fixed constant. Hence by taking t such that σ_t approaches σ_T, we see that $\mathbf{P}\{X^* \geq \sigma_T\} \geq \overline{\Phi}(1)$. In particular, if $\sigma_T = \infty$, then $X^* = \infty$ with positive probability (you can make that probability equal to $1/2$, or $\mathbf{P}\{\|X\| = \infty\} = 1$).

5. For continuity, a necessary condition for continuity of X (when T is a metric space) is that the mean function and covariance kernel be continuous. Other ways to say this are that the curve $t \mapsto X_t$ in $L^2(\mathbf{P})$ must be continuous or that d must be stronger than τ.

   To see this, fix $t, s \in T$ and sequences $t_n \to t$ and $s_n \to s$. From almost sure convergence follows the convergence in distribution of (X_{t_n}, X_{s_n}) to (X_t, X_s), and as everything is Gaussian, $\mathbf{E}[X_{t_n}X_{s_n}] \to \mathbf{E}[X_tX_s]$. That is, $K(t_n,s_n) \to K(t,s)$, showing that $K : T \times T \mapsto \mathbb{R}$ is continuous. Similarly, the mean function has to be continuous. Thus, continuity of the mean and covariance kernel are necessary for the continuity of sample paths.

   That this condition is not sufficient is perhaps surprising at first, but that is the precise topic of the chapter - to work out additional conditions that ensure continuity.

6. When d is stronger than τ (i.e., if the covariance is continuous), then $\mathrm{id} : (T,d) \mapsto (T,\tau)$ is continuous. Thus if X has continuous sample paths w.r.t. τ, then it has continuous sample paths w.r.t. d.

   Now suppose (T,d) is compact. Then $\mathrm{id} : (T,d) \mapsto (T,\tau)$ is actually a homeomorphism (being continuous, it maps compact sets to compact sets, and hence open sets to open sets, proving the continuity in the reverse direction).

   Thus, for compact metric spaces, continuity of the Gaussian process in the original metric is equivalent to its continuity in the τ-metric. The restriction to compact spaces is not a restriction - our spaces will be σ-compact, and in any case one does not expect processes to be bounded or uniformly continuous on non-compact spaces.

7. We claim that whenever $\mathbf{E}[X^*]$ is finite, X^* has Gaussian tails above its mean (with exponent given by the maximal variance).

   Consider the case of finite $T = \{1,2,\ldots,n\}$. The mapping $(x_1,\ldots,x_n) \mapsto x^* = \max_i x_i$ is a Lipschitz function with Lipschitz constant 1. If $X \sim N_n(0,\Sigma)$, we represent it as $X = BZ$ where $Z \sim \gamma_n$ and $BB^t = \Sigma$. For any $z,w \in \mathbb{R}^n$,

$$\begin{aligned} |(Bz)^* - (Bw)^*| &= \left|\max_{i\le n} \sum_{j=1}^n b_{i,j}(z_j - w_j)\right| \\ &\le \|z-w\|_2 \max_{i\le n} \sqrt{\sum_{j=1}^n b_{i,j}^2} \;=\; \left(\max_i \sigma_{i,i}\right) \|z-w\|_2. \end{aligned}$$

   Thus, the mapping is Lipschitz with constant σ_T. By the Gaussian isoperimetric inequality (see the form in Theorem 21), it follows that

$$\mathbf{P}\{X^* \ge \mathbf{E}[X^*] + \sigma_T x\} \le Ce^{-cx^2}$$

   for universal constants C,c. In all cases of interest, it will be the case that X^* is approximable by finite subsets of T, and hence (why?) we have the above concentration bound for the upper tail of X^*.

With all these comments, let us not forget that we have not actually calculated the maximum except in the i.i.d. case and a weakly dependent case (Berman's theorem). But we have already built essential tools, such as the isoperimetric inequality that gave us the concentration results above and comparison inequalities that will come in handy below.

5.2 A lower bound for the expectation of the maximum

For a metric space T, let $N(\varepsilon)$ denote the smallest size of an ε-net for T in the metric τ (i.e., the smallest m for which there are points $t_1,\dots,t_m \in T$ such that $B(t_1,\varepsilon)\cup\dots\cup B(t_m,\varepsilon) = T$). A closely related quantity if $N'(\varepsilon)$, defined as the maximum size of an ε-separated set (i.e., the maximum m for which there are points $t_1,\dots,t_m$ such that $\tau(t_i,t_j) \geq \varepsilon$ for all $i \neq j$). It is also called $\varepsilon/2$-packing number, since it is the maximum number of pairwise disjoint $\varepsilon/2$-radius balls that one can pack into T. The two quantities N and N' are comparable, and can be used interchangeably.

Claim 1. $N(\varepsilon) \leq N'(\varepsilon) \leq N(\varepsilon/2)$.

Proof. Let $t_1,\dots,t_m$ be a maximal-cardinality ε-separated set in T so that $m = N'(\varepsilon)$. The the balls $B(t_j,\varepsilon/2)$ are pairwise disjoint, showing that any $\varepsilon/2$-net of T must contain a point in each of these balls. Hence $N(\varepsilon/2) \geq m$. Further, $B(\mathbf{t}_j,\varepsilon)$, $1 \leq j \leq m$, cover T, or else we could add one more point to $\{t_1,\dots,t_m\}$ maintaining ε-separation. Hence $N(\varepsilon) \leq N'(\varepsilon)$. ■

The quantity $\log N(\varepsilon)$ is called the *metric entropy* of (T,τ). If one uses logarithm to base 2, the metric entropy can be interpreted as the number of bits needed to identify any point of T to an accuracy of ε. Because of the following inequalities, the same interpretation applies to $\log N'(\varepsilon)$ and the same. The metric entropy function measures the size of the metric space.

Theorem 2 (Fernique/Sudakov minoration)**.** *Let X be a centered Gaussian process on T. Let τ be the associated metric on T. Then, if X^* is measurable, then $\mathbf{E}[X^*] \geq \kappa\varepsilon\sqrt{\log N(\varepsilon)}$ for any $\varepsilon > 0$ (where κ is a universal cosntant). In particular, if T is not totally bounded in the metric τ, then X must be unbounded w.p.1.*

Proof. Let $t_1,\dots,t_N$ be a minimal ε-net for T. Then $\tau(t_i,t_j) \geq \frac{1}{2}\varepsilon$ Then $\mathbf{E}[|X(t_i)-X(t_j)|^2] \geq \frac{1}{2}\varepsilon^2\mathbf{E}[|Z_i - Z_j|^2]$, where Z_i are i.i.d. standard Gaussians. By Sudakov-Fernique inequality, $X^* \geq \max_{i\leq N} X(t_i)$ has greater expectation than $\max_{i\leq N}\varepsilon Z_i$. But we know that $\mathbf{E}[\max_{i\leq n} Z_i] \sim \sqrt{2\log n}$ as $n\to\infty$, hence $\mathbf{E}[\max_{i\leq n} Z_i] \geq \kappa\sqrt{\log n}$ for some $\kappa > 0$. Consequently, $\mathbf{E}[X^*] \geq \kappa\varepsilon\sqrt{\log N(\varepsilon)}$.

To say that T is not totally bounded is the same as saying that $N(\varepsilon) = \infty$ for some $\varepsilon > 0$, hence the second statement in the theorem. ■

This theorem shows the relevance of the metric entropy to the problems of boundedness of a Gaussian process. The following exercise shows the same for the continuity question.

However, although entropy bounds are very powerful and sufficient for most purposes, we shall see that the precise geometric feature of the metric space (T,τ) that controls the boundedness/continuity is a different one.

Exercise 3. With the notation as in Fernique's theorem, show that $\mathbf{E}[\omega_X(\varepsilon)] \geq \kappa\varepsilon\sqrt{\log N(\varepsilon)}$.

5.3 The generic chaining upper bound

As we have remarked earlier, the lower tail of X^* poses no difficulty, it is only the upper tail that we need to control. Hence we can either write our bounds for $\mathbf{E}[(X^*)_+]$ and $\mathbf{P}\{X^* \geq u\}$, or define $Y(t) = X(t) - X(t_0)$ for some fixed $t_0 \in T$ and study Y^* (the convenience being that $Y^* \geq 0$ necessarily). The approaches are equivalent since $X^* = Y^* - X(t_0)$. Here is the basic lemma by the method of *generic chaining*[1].

Setting: Let $X = (X_t)_{t\in T}$ be a stochastic process with zero mean random variables indexed by a metric space (T,τ) such that $\mathbf{P}\{|X_t - X_s| \geq u\tau(t,s)\} \leq 2\exp\{-\frac{1}{2}u^2\}$ for all $u > 0$ and for all $t,s \in T$. The case of interest for us is that of a centered Gaussian process with $\tau(s,t) := \sqrt{\mathbf{E}[|X_t - X_s|^2]}$.

Lemma 4 (The generic chaining bound)**.** *Let T be finite or countable. Fix $t_0 \in T$ and numbers $u_k \geq 1$. Choose any finite subsets $T_k \subseteq T$ with $T_0 = \{t_0\}$ and such that each $t \in T$ is contained in T_k for all large k. Then for any $x > 0$ we have*

$$\mathbf{P}\{X^* - X_{t_0} \geq A\} \leq Q.$$

where $A = 2\sup_{t\in T}\sum_{k=1}^{\infty} u_k\tau(t,T_k)$ *and* $Q = 2\sum_{k=1}^{\infty} |T_k|\cdot|T_{k-1}|e^{-u_k^2/2}$.

Proof. First take $x = 0$. Let $\pi_k(t)$ be any point of T_k closest to t, i.e., $\tau(t,\pi_k(t)) = \tau(t,T_k)$. Then $X_t - X_{t_0} = \sum_{k=1}^{\infty} X_{\pi_k(t)} - X_{\pi_{k-1}(t)}$ (the sum is finite by the assumption that $\pi_k(t) = t$ for large enough k). If $X_t - X_{t_0} > 2\sum_k u_k\tau(t,T_k)$, then there must be at least one k such that $X_{\pi_k(t)} - X_{\pi_{k-1}(t)} > u_k(\tau(t,T_k) + \tau(t,T_{k-1}))$. For any $t,s \in T$ and $u \geq 1$,

$$\mathbf{P}\{|X_t - X_s| \geq u\,\tau(t,s)\} \leq 2e^{-\frac{1}{2}u^2}.$$

[1]This topic is beautifully explained in the book *Upper and lower bounds for stochastic processes* by Talagrand, who was after all the discoverer of many of these things. The older book *Generic chaining* by Talagrand is also sufficient for our purposes (although strongly discouraged by Talagrand!).

Hence, as $\tau(\pi_k(t),\pi_{k-1}(t)) \le \tau(t,T_k)+\tau(t,T_{k-1})$, it follows that

$$\mathbf{P}\{|X_{\pi_k(t)} - X_{\pi_{k-1}(t)}| > u_k(\tau(t,T_k)+\tau(t,T_{k-1}))\} \le 2e^{-\frac{1}{2}u_k^2}.$$

There are $|T_k|$ possibilities for $\pi_k(t)$, hence the union bound gives

$$\mathbf{P}\{\sup_t |X_t - X_{t_0}| \ge A\} \le 2\sum_{k\ge 1} |T_k| \times |T_{k-1}|\ e^{-\frac{1}{2}u_k^2}.$$

The right side is exactly Q. ■

What is the right choice of u_ks and T_ks? That is what we investigate next.

5.4 Dudley's integral

For any choice of u_ks, to reduce the bound A, it is important that $\tau(t,T_k)$ be small. Hence it seems natural to take T_k to be a minimal ε_k-net for some sequence $\varepsilon_k \downarrow 0$. Then the kth summand is like $N(\varepsilon_k)N(\varepsilon_{k-1})e^{-\frac{1}{2}u_k^2}$. But then to get a finite Q, we must take $u_k \gtrsim C\sqrt{\log N(\varepsilon_k)}$.

Let us make the following choices: Let $\varepsilon_k = 2^{-k}$, let T_k be a minimal cardinality ε_k-net and $u_k = u(1+4\sqrt{\log(N_k+k)})$ where $u \ge 1$ and $N_k = N(\varepsilon_k)$. Then,

$$\begin{aligned} Q &\le \sum_{k\ge 0} N_k^2 e^{-\frac{1}{2}u^2(1+16\log(N_k+k))} \\ &\le e^{-\frac{1}{2}u^2} \sum_{k\ge 0} \frac{N_k^2}{(N_k+k)^8} \\ &\le e^{-\frac{1}{2}u^2}. \end{aligned}$$

In the last line we used the fact that $(N_k+k)^8 \ge 8N_k^2k^6$ to bound the series in the previous line by $\sum_k (8k^6)^{-1}$, which is safely bounded by 1. Further, since $\sup_t \tau(t,T_k) \le \varepsilon_k$,

$$\begin{aligned} A &\le 2\sum_{k=1}^{\infty} u_k \sup_{t\in T} \tau(t,T_k) \\ &\le 2u\sum_{k=1}^{\infty} \varepsilon_k(1+4\sqrt{\log(N_k+k)}). \end{aligned}$$

Now write $\sqrt{\log(x+y)} \le \sqrt{\log x}+\sqrt{\log y}$ to bound the right hand side by

$$2u\left(\sum_{k=1}^{\infty} 2^{-k}(1+4\sqrt{\log k}) \ +\ 4\sum_{k=1}^{\infty} 2^{-k}\sqrt{\log N_k}\right) \le Cu(1+J'(T))$$

where C is a constant (it can be easily bounded by 10) and $J'(T) = \sum_{k=1}^{\infty} 2^{-k}\sqrt{\log N_k}$. By the fact that $\varepsilon \mapsto N(\varepsilon)$ is decreasing, we see that

$$2^{-k-1}\sqrt{\log N_k} \le \int_{2^{-k-1}}^{2^{-k}} \sqrt{\log N(\varepsilon)}d\varepsilon \le 2^{-k-1}\sqrt{\log N_{k+1}}.$$

and hence $J'(T) \asymp J(T)$ where the *Dudley integral* $J(T)$ is defined as

$$J(T) := \int_0^{\infty} \sqrt{\log N(\varepsilon)}d\varepsilon.$$

The conclusion is summarized in the theorem below.

Theorem 5 (Dudley's integral for boundedness)**.** *Let T be countable and fix $t_0 \in T$. With the notations above, for any $u \ge 1$, we have*

$$\mathbf{P}\{X^* - X_{t_0} \ge Cu(1 + J(T))\} \le e^{-\frac{1}{2}u^2}.$$

Further, $\mathbf{E}[(X^*)_+] \le CJ(T)$. Here C is a pure number. In particular, the finiteness of the Dudley integral is a sufficient condition for the almost sure boundedness of the Gaussian process.

The extra additive term 1 is irrelevant and may be removed to write $\mathbf{P}\{X^* - X_{t_0} \ge CuJ(T)\} \le e^{-\frac{1}{2}u^2}$. We leave it as an exercise.

Compare the upper bound from Dudley integral to the lower bound due to Fernique that we have seen before:

$$\mathbf{E}[X^*] \ge \kappa \sup_{\varepsilon} \varepsilon\sqrt{\log N(\varepsilon)}.$$

Dudley's upper bound and Fernique's lower bound are almost enough to answer the boundedness question decisively, but not quite. The ambiguity remains when the Dudley integral is finite but Fernique's lower bound is finite.

- If $N(\varepsilon) \le \exp\{-\varepsilon^{-c}\}$ for $c < 2$, then the Dudley integral is finite, showing boundedness.
- If $N(\varepsilon) \ge \exp\{-\varepsilon^{-c}\}$ for some $c > 2$, then the lower bound of Fernique is inifinite, showing that the process is not bounded.
- If $N(\varepsilon) \asymp \exp\{\varepsilon^{-2}\}$, then the lower bound is finite and the upper bound is infinite, and hence the boundedness question remains unanswered.

As it happens, both Dudley's and Fernique's bounds are loose, and the right quantity that determines boundedness is a different one, given in the next section. However, we shall also see later that for stationary Gaussian processes, convergence of Dudley's integral is necessary for boundedness.

5.5 Talagrand's γ_2-functional

We define two fundamental quantities.

Definition 6. For a metric space (T,τ), define the following by taking infima over all choices of the sets $\{T_k\}$ subject to the condition $|T_k| = 2^{2^k}$.

1. *Talagrand's γ_2-functional:* $\gamma_2(T) := \inf_{\{T_k\}} \sup_t \sum_{k=0}^{\infty} 2^{k/2}\tau(t,T_k)$.

2. *Dudley's integral:* $\mathcal{J}'(T) := \inf_{\{T_k\}} \sum_{k=0}^{\infty} 2^{k/2} \sup_t \tau(t,T_k)$.

Observe that the choice $u_k = 2^{k/2}$ is consistent with the need that $u_k \gtrsim \sqrt{\log |T_k|}$. The name Dudley integral is justified by the following exercise.

Exercise 7. Show that $\mathcal{J}'(T)$ is (up to constants) the same as $J(T)$ defined earlier.

The difference between $\mathcal{D}(T)$ and $\gamma_2(T)$ is that in the former, the supremum is taken inside the sum. The extra flexibility of γ_2 comes from the fact that the sequence of sets $\{T_k\}$ an be chosen adapted to the point t. It gives both an upper and lower bound, and thus settles the problem of finding the right condition for boundedness of a Gaussian process!

Theorem 8 (Talagrand)**.** $\mathbf{E}[\sup_t X_t - X_{t_0}] \asymp \gamma_2(T)$.

We have already proved the upper bound $\mathbf{E}[X^* - X_{t_0}] \lesssim \gamma_2(T)$ (the first genereic chaining bound!). The lower bound was conjectured by Fernique (in a different form) and proved by Talagrand. We shall not prove this theorem here.

5.6 Dudley integral criterion for continuity

Recall that the *modulus of continuity* of a function $f : X \mapsto \mathbb{R}$ on a metric space (X,d) is defined as $\omega_f(\delta) := \sup\{|f(x) - f(y)| : x,y \in T,\ d(x,y) < \delta\}$. The uniform continuity of f is equivalent to $\omega_f(\delta) \downarrow 0$ as $\delta \downarrow 0$. If $\omega_f(\delta) \le C_f \delta^\alpha$ for some $\alpha \in (0,1]$ and $C_f < \infty$, we say that f is Hölder(α). When $\alpha = 1$, we say that f is Lipschitz with Lipschitz constant C_f.

We now want to investigate the modulus of continuity of a Gaussian process X on a set T. We assume that T is countable as before. Let K be the covariance kernel and let τ be the associated metric. The analysis is parallel to the generic chaining bound for the supremum, and we constantly refer to the reasoning there for details.

As in the generic chaining bound, let $\{t_0\} = T_0 \subseteq T_1 \subseteq T_2 \subseteq \dots$ that increase to T, and let $\pi_k(t)$ be the "projection" of t to T_k as defined there. For $t, s \in T$, write the telescoping sum as before:

$$X_t - X_{t_0} = \sum_{k=1}^{\infty} X_{\pi_k(t)} - X_{\pi_{k-1}(t)}, \qquad X_s - X_{t_0} = \sum_{k=1}^{\infty} X_{\pi_k(s)} - X_{\pi_{k-1}(s)}.$$

Now suppose $\pi_k(t) = \pi_k(s)$ for $k \le m(t,s)$. The idea being that if t and s are close, then $m(t,s)$ is large (one must choose T_ks reasonably, of course). Then,

$$\begin{aligned} |X_t - X_s| &\le \sum_{k > m(t,s)} |X_{\pi_k(t)} - X_{\pi_{k-1}(t)}| + \sum_{k > m(t,s)} |X_{\pi_k(s)} - X_{\pi_{k-1}(s)}| \\ &\le 4 \sum_{k \ge m(t,s)} \max_{u \in T_k, v \in T_{k-1}} |X_u - X_v|. \end{aligned}$$

Hence if we write $A_m = 4 \sup_{t \in T} \sum_{k>m} u_k \tau(t, T_k)$ and $Q_m = 2 \sum_{k>m} |T_k| \cdot |T_{k-1}| e^{-u_k^2/2}$ (for some choice of u_ks), then we see that for fixed $\delta > 0$, writing $m(\delta) = \min\{m(t,s) : \tau(t,s) \le \delta\}$,

$$\mathbf{P}\{\omega_X(\delta) > A_{m(\delta)}\} \le Q_{m(\delta)}.$$

Now let us make the choice of u_ks and T_ks, as in the Dudley integral bound earlier. That is, fix $\varepsilon_k = 2^{-k}$ and let T_k be a maximal cardinality ε_k-separated set in T (earlier we chose it to be a minimal cardinality ε_k-net) and set $u_k = 1 + 5\sqrt{\log(N_k + k)}$. Then, with similar analysis as before,

$$\begin{aligned} Q_m &\le \sum_{k \ge m} \frac{N_k^2}{(N_k + k)^5} \\ &\le \sum_{k \ge m} \frac{1}{1 + k^3} \\ &\le \frac{1}{m^2} \end{aligned}$$

and

$$A_m \le C \sum_{k \ge m} 2^{-k} \sqrt{\log N_k}$$

where $N_k = N(2^{-k})$.

For this choice of T_ks, if $\tau(t,s) \le 2^{-r-2}$, then

$$\tau(\pi_k(t), \pi_k(s)) \le \tau(\pi_k(t), t) + \tau(\pi_k(s), s) + \tau(s,t) \le 32^{-r},$$

for $k \le r$ and hence $\pi_k(t) = \pi_k(s)$ for $k \le r$ (otherwise $\tau(u,v) \ge 2^{-r}$ for $u,v \in T_r$). Thus, $m(\delta) \ge c\log(1/\delta)$. Putting this together with the previous bounds, we arrive at

$$\mathbf{P}\left\{\omega_X(2^{-r}) \ge C\int_0^{2^{-r}} \sqrt{\log N(\varepsilon)}d\varepsilon\right\} \le \frac{1}{r^2}.$$

As this is summable, we see that with probability one, $\omega_X(2^{-r}) \le C\int_0^{2^{-r}} \sqrt{\log N(\varepsilon)}d\varepsilon$ for all large enough r, or equivalently

$$\omega_X(\delta) \le K\int_0^{\delta} \sqrt{\log N(\varepsilon)}d\varepsilon$$

for all $\delta \in (0,1]$, for some random constant K. This shows that whenever the Dudley integral is finite, the process X is uniformly continuous on T. We summarize the conclusions.

Theorem 9. *Let T be a countable set and let X be a centered Gaussian process on T with the associated pseudo-metric $\tau(t,s) := \|X_t - X_s\|_{L^2}$. If the Dudley integral $J(T) = \int_0^{\infty} \sqrt{\log N(\varepsilon)}d\varepsilon$ converges, then the process X is uniformly continuous on T and has (w.r.t. τ) the modulus of continuity $\omega_X(\delta) \le K\int_0^{\delta} \sqrt{\log N(\varepsilon)}d\varepsilon$ for all δ for a random variable K that is finite almost surely.*

In the remaining sections, we apply these general theorems based on Dudley's bound in many examples of interest.

5.7 Example: Independent Gaussians

Let $X_k \sim N(0,\sigma_k^2)$ be independent. Assume that σ_k^2 decreases to 0. Then $\tau(m,n) = \sqrt{\sigma_n^2 + \sigma_m^2}$. Note that for $m < n$ we have $\sigma_m \le \tau(m,n) \le \sigma_m\sqrt{2}$. For simplicity let us pretend that $\tau(m,n) = \sigma_{m\wedge n}$ (we leave it as an exercise to make appropriate modifications).

If $0 < \varepsilon < \sigma_1$, then there is a unique n such that $\sigma_n \le \varepsilon < \sigma_{n-1}$. Then $\{1,2,\ldots,n\}$ is an ε-net whence $N(\varepsilon) \le n$. Since $\tau(i,j) > \sigma_{n-1} > \varepsilon$ for $i,j \le n-1$, it is clear that $N(\varepsilon) \ge n-1$. Thus the Dudley integral is (as always ignoring constant factors) $J = \sum_{k=2}^{\infty}(\sigma_{k-1} - \sigma_k)\sqrt{\log k}$.

On the other hand, we may write $X_n = \sigma_n\xi_n$ where ξ_n are i.i.d. $N(0,1)$ variables. Recall that $\limsup_{n\to\infty} \frac{\xi_n}{\sqrt{2\log n}} = 1$ a.s. (if not clear, provide a proof!). Thus, $\sup_n X_n < \infty$ a.s. whenever $\limsup \sigma_n\sqrt{\log n} < \infty$.

Thus, by choosing, for example, $\sigma_n = \frac{1}{\sqrt{\log n \log\log n}}$, we see that the Dudley integral may diverge but the supremum is finite.

Exercise 10. Compute γ_2 or at lease verify that it is finite for this choice of σ_ns.

5.8 Example: Brownian motion

Suppose X is a Gaussian process on $[0,1]$ satisfying $\mathbf{E}[|X_t - X_s|^2] \leq C|t-s|^\alpha$ for some $\alpha > 0$ and constant C. We restrict to a coutnable dense subset such as dyadic rationals to apply the theorem above.

Then $\tau(t,s) \leq C|t-s|^\alpha$ and $N(\varepsilon) \leq C\varepsilon^{-1/\alpha}$ and hence the Dudley integral converges. Thus X is almost surely uniformly continuous on dyadic rationals, and hence extends continuously to $[0,1]$ almost surely. Further, the modulus of continuity of X on $[0,1]$ is the same as on the dense subset. It is

$$\omega_X(\delta) \leq \frac{K}{\sqrt{\alpha}} \int_0^\delta \sqrt{\log \frac{1}{\varepsilon}} d\varepsilon \;\leq\; K'\delta\sqrt{\log\frac{1}{\delta}}.$$

Observe that this is with respect to the metric τ. If $|t-s| \leq \delta$, then $\tau(t,s) \leq C\delta^\alpha$ and hence with respect to the Euclidean metric $\omega_X(\delta) \leq K''\delta^\alpha\sqrt{\log\frac{1}{\delta}}$. In particular, the sample paths of X are almost surely Hölder(β) for any $\beta < \alpha$.

In the special case of Brownian motion, $K(t,s) = t \wedge s$ and hence $\tau(t,s) = \sqrt{|t-s|}$. This corresponds to the case $\alpha = \frac{1}{2}$. Therefore, the modulus of continuity is $O(\sqrt{\delta \log \frac{1}{\delta}})$ and the paths are Hölder($\frac{1}{2} - \varepsilon$) for any $\varepsilon > 0$.

5.9 Example: Processes on the boundary of spherically symmetric trees

Let $\mathcal{T}$ be a rooted locally finite tree. Fix $\lambda > 1$ and recall the Gaussian process on the boundary $\partial\mathcal{T}$ defined by [2].

$$X_\xi := \sum_{k=0}^{\infty} Z_{v_k} \lambda^{-k/2}$$

for $\xi = (v_0, v_1, \ldots) \in \partial\mathcal{T}$. We have fixed $\lambda > 1$ and chosen Z_v i.i.d. standard Gaussians. We have seen that the associated metric τ is (up to some constant multiple that we ignore) given by $\tau(\xi,\eta) = \lambda^{-|\xi\wedge\eta|}$.

[2]The lower bounds here are all due to Fernique. Our presentation is essentially from Kahane's book *Some random series of functions*. He does not mention trees but what we call spherically symmetric trees are referred to as generalized Cantor sets there. But the essence is the same. We have simplified some proofs, perhaps correctly.

The goal: We wish to say that for a class of such processes, convergence of the Dudley integral is necessary for the boundedness of the process.

With this goal in mind, we need an upper bound for the Dudley integral and a lower bound for the expectation of the supremum. We restrict to a class of trees known as *spherically symmetric trees*. By definition, this means that there are numbers $m_k \geq 1$ such that all vertices in the kth generation have exactly m_k children.

Upper bound for the Dudley integral: Given $\varepsilon > 0$, choose m such that $\lambda^{-n} \leq \varepsilon < \lambda^{-n+1}$. If we take a collection of $|\mathcal{T}_n|$ paths from the root, one passing through each vertex in the n-th generation, then that gives an ε-net for $\partial\mathcal{T}$ (since any path from the root has to pass through one of these vertices). Therefore, $N(\varepsilon) \leq |\mathcal{T}_n|$. Hence the Dudley integral

$$J(\partial\mathcal{T}) \leq C_\lambda \sum_{k\geq 0} \lambda^{-k/2} \sqrt{\log |\mathcal{T}_k|}.$$

Under the assumption of spherical symmetry, $|\mathcal{T}_k| = m_0 m_1 \dots m_{k-1}$. Therefore, using $\sqrt{\log(x+y)} \leq \sqrt{\log x} + \sqrt{\log y}$, we see that

$$\begin{aligned} J(\partial\mathcal{T}) &\leq C_\lambda \sum_{k\geq 0} \lambda^{-k/2} \sum_{j=0}^{k-1} \sqrt{\log m_j} \\ &= C_\lambda \sum_{j=0}^{\infty} \sqrt{\log m_j} \sum_{k=j+1}^{\infty} \lambda^{-k/2} \\ &\leq C'_\lambda \sum_{j=0}^{\infty} \lambda^{-j/2} \sqrt{\log m_j}. \end{aligned} \tag{1}$$

Lower bound for the expected supremum: One way to get a lower bound for X^* is to use a greedy algorithm. We define $u_0 = 0$ (the root) and having chosen $u_0, \dots, u_k$, we choose u_{k+1} as the child of u_k for which Z_u is maximized. That is, $Z_{u_{k+1}} \geq Z_v$ for $u_k \rightsquigarrow v$. Let the resulting path be $\xi = (u_0, u_1, \dots) \in \partial\mathcal{T}$. Then

$$X^* \geq X_\xi = \sum_{k\geq 0} \lambda^{-k/2} M(u_k)$$

where $M(u) = \max\{Z_v : u \rightsquigarrow v\}$.

Recall that the expectation of the maximum of n i.i.d. standard Gaussians is at least $\kappa\sqrt{\log n}$ for some $\kappa > 0$. Therefore, if $\mathcal{F}_k = \sigma\{Z_v : |v| \leq k\}$, then $\mathbf{E}[M(u_k) \mid \mathcal{F}_k] \geq \kappa\sqrt{\log m_k}$

and the same lower bound must hold for the unconditional expectation. Consequently,

$$\mathbf{E}[X^*] \geq \kappa \sum_{k\geq 0} \lambda^{-k/2} \sqrt{\log m_k}. \tag{2}$$

By comparing (2) with (1),and the general upper bound $\mathbf{E}[X^*] \lesssim J$, we arrive at the following theorem.

Theorem 11. *Let $\mathcal{T}$ be a locally finite, infinite, rooted, spherically symmetric tree. Let Z_v be i.i.d. standard Gaussian variables indexed by the vertices of $\mathcal{T}$ and let $\lambda > 1$. Let X be the Gaussian process on $\partial\mathcal{T}$ defined by $X(\xi) = \sum_{k\geq 0} Z_{\xi_k} \lambda^{-k/2}$. Then, $\mathbf{E}[X^*] \asymp J(\partial\mathcal{T})$. In particular, the process is bounded if and only if the Dudley integral is finite.*

Exercise 12. Under the same conditions, is it true that the process is continuous on $\partial\mathcal{T}$ if and only if the Dudley integral is finite?

Exercise 13. Exact homogeneity is not needed. Let m_k be as above and suppose each vertex in generation k has between $m_k^{0.01}$ and m_k^{100} children. The tree need no longer be spherically symmetric, but show Dudley integral is a lower bound for the expected supremum.

5.10 Example: Stationary processes

If G is a group and X is a centered Gaussian process indexed by G, then recall that X is said to be *left-stationary* if $(X_{hg})_g \stackrel{d}{=} (X_g)_{g\in G}$ for any $h \in G$. For Gaussian processes, this just means that $\tau(hg, hg') = \tau(g, g')$ for all g, g', h.

Theorem 14 (Fernique)**.** *Let G be a locally compact group and let X be a centered stationary Gaussian process on G. For any compact $K \subseteq G$, let $J(K)$ denote the Dudley integral of (K, τ_X). Then, for any $g_0 \in K$, we have $\mathbf{E}[\sup_{g\in K} X_g - X_{g_0}] \geq \kappa J(K)$ for some constant $\kappa > 0$ (which may depend on the group).*

Proof. Let $\tau = \tau_X$. Without loss of generality, assume that $\mathrm{dia}_\tau(K) = 1$. Let $S_0 = \{1\}$ and for $k \geq 1$ let S_k be a maximal 2^{-k}-separated set in $B_\tau(1, 2^{-k+1})$ (open ball centered at the identity). If $g_i \in S_i$ for all $i \geq 1$, then

$$\tau(g_1 \dots g_m, g_1 \dots g_{m-1}) = \tau(1, g_m) \leq 2^{-m+1}$$

as S_m forms a 2^{-m}-net for $B_\tau(1,2^{-m+1})$. From this, just as in binary expansion of numbers, it follows that all finite products of the form $g_1 g_2 \dots g_m$ where $m \geq 1$ and $g_i \in S_i$, are distinct. The set of all such elements are dense in $B_\tau(1,1)$ and hence in K. To see this, start with $g \in B_\tau(1,1)$, find $g_1 \in S_1$ so that $\tau(g,g_1) \leq 2^{-1}$. Then $g_1^{-1}g \in B_\tau(1,2^{-1})$ and hence within 2^{-2} distance of some $g_2 \in S_2$, which means that $g_2^{-1}g_1^{-1}g \in B_\tau(1,2^{-2})$, and so on, showing that $g_1 \dots g_m$ converges to g.

Now let $\mathcal{T}$ be the tree whose vertex set is the collection of all such finite products $g_1 \dots g_m$ with $m \geq 1$ and $g_i \in S_i$, where $g_1 \dots g_{m+1}$ is a child of $g_1 \dots g_m$. Then $\mathcal{T}$ is a spherically symmetric tree in which all vertices in the kth generation have $m_{k+1} = |S_{k+1}|$ children. From the earlier discussion, it is clear that the each element of $\xi = (u_0, u_1, \dots) \in \mathcal{T}$ may be mapped to the group element of $g_\xi = \lim u_n$. It is possible that $g_\xi = g_\eta$ for distinct ξ, η. In fact, in terms of the metric $\tau_4(\xi,\eta) = 4^{-|\xi \wedge \eta|}$, we have

$$\tilde{\tau}(\xi,\eta) := \tau(g_\xi, g_\eta) \leq 3\tau_4(\xi,\eta) \quad \text{for all } \xi,\eta \in \partial\mathcal{T},$$

but no such inequality holds in the other direction. However, the reverse inequality can hold on appropriate subsets of $\partial\mathcal{T}$. For example, consider the subsets $\partial_0\mathcal{T}$ and $\partial_1\mathcal{T}$ defined by

$$\partial_r\mathcal{T} = \{\xi = (u_0, u_1, \dots) \in \partial\mathcal{T} : u_{k+1} \text{ is the first child of } u_k \text{ if } k \neq r (\text{mod } 2)\}$$

If $\xi = (u_0, u_1, \dots)$ and $\eta = (v_0, v_1, \dots)$ are in $\partial_r\mathcal{T}$, then $u_j \neq v_j$ for $|\xi \wedge \eta| + 1 \leq j \leq |\xi \wedge \eta| + 9$, which implies that

$$\begin{aligned} \tau(\xi,\eta) &\geq 4^{-|\xi\wedge\eta|}\left(1 + \frac{1}{4} - 3\sum_{j=2}^{\infty} 4^{-j}\right) \\ &\geq \tau_4(\xi,\eta) \end{aligned} \tag{3}$$

as the sum in the brackets on the previous line is equal to 1. This is the reverse inequality we wanted.

Now we have two processes on $\partial\mathcal{T}$: The process X (or more precisely the process $\xi \mapsto X(g_\xi)$ from the Gaussian process on G and the process $Y_\xi = \sum_k 4^{-k/2} Z_{u_k}$ for $\xi = (1 = u_0, u_1, u_2, \dots)$, where Z_v are i.i.d. standard Gaussians. The metrics on $\partial\mathcal{T}$ associated to X and Y are $\tilde{\tau}$ and τ_4, respectively. Therefore, it is clear that

$$J_{\tilde{\tau}}(\partial\mathcal{T}) \leq 3J_{\tau_4}(\partial\mathcal{T}).$$

In particular, if $J_{\tilde{\tau}}(\partial\mathcal{T})$ is infinite, then so is $J_{\tau_4}(\partial\mathcal{T})$. We also know that $\mathbf{E}[Y^*] \asymp J_{\tau_4}(\mathcal{T})$.

The Sudakov-Fernique inequality gives $\mathbf{E}[X^*] \le 3\mathbf{E}[Y^*]$, but we need an inequality in the other direction (something that gives a lower bound for $\mathbf{E}[X^*]$). For that purpose, consider the subtrees $\partial_r \mathcal{T}$, $r = 0, 1$. Indeed, from the inequality (3), we see that $\mathbf{E}[X^*_{\partial_r \mathcal{T}}] \le \mathbf{E}[Y^*_{\partial_r \mathcal{T}}]$ for $r = 0, 1$.

Now, we leave it as an exercise to check that $\mathbf{E}[Y^*_{\partial \mathcal{T}}] \asymp J(\partial_r \mathcal{T})$ for $r = 0, 1$. And also that $J_{\tau_4}(\partial \mathcal{T}) \le J_{\tau_4}(\partial_0 \mathcal{T}) + J_{\tau_4}(\partial_1 \mathcal{T})$.

Combining all these, we arrive at $\mathbf{E}[X^*_K] \asymp J(\tau_K)$. ■

Exercise 15. Let X be a stationary Gaussian process on a locally compact group G. If X is continuous *a.s.*, then is it true that the Dudley integral must converge?

5.11 Example: GFF on a domain

Recall that the Gaussian free field on a domain U having Green's function G was defined to be a Gaussian process on the space $\mathcal{M}$ of signed-measures having finite logarithmic energy (i.e., $\iint G(x,y)d|\mu|(x)d|\mu|(y)$ is finite).

Question: Is this process continuous on $\mathcal{M}$? Bounded on compact subsets of $\mathcal{M}$?

5.12 A lemma of Garsia, Rodemich and Rumsey

Around the same time as Dudley's criterion, a different approach was found to generalize Dudley's criterion. The relationship between the two, their relative strengths etc.,?. The key to this approach is a real-variable lemma that has no randomness in it.

Lemma 16 (Garsia, Rodemich, Rumsey)**.** *Let $f \in [0,1] \mapsto \mathbb{R}$ be a continuous function. Let $\Psi, p : \mathbb{R}_+ \mapsto \mathbb{R}_+$ be strictly increasing continuous functions such that $\Psi(0) = 0$, $p(0) = 0$, $\psi(x) \to \infty$ as $x \to \infty$. Suppose*

$$I := \int_0^1 \int_0^1 \Psi\left(\frac{|f(t) - f(s)|}{p(t-s)}\right) dt\, ds < \infty.$$

Then, $\omega_f(\delta) \le 8 \int_0^\delta \Psi^{-1}(4I/u^2)\, dp(u)$ for any $\delta > 0$.

For now, we skip the proof[3]. Let us see how this is useful to the question of continuity of Gaussian processes.

[3] It is explained in good detail in several lecture notes of S. R. S. Varadhan, this one for example.

Remark 17. One may worry that the hypothesis of the lemma already has continuity in it. It can be rephrased for a measurable function f satisfying $I < \infty$ that

$$|f(x) - f(y)| \leq 8 \int_0^{\delta} \Psi^{-1}(4I/u^2)\, dp(u) \quad \text{for } a.e.\ x, y \in [0,1].$$

Then of course, one may consider the restriction of f to this set of full measure and extend it continuously to $[0,1]$ preserving the same modulus of continuity.

Let $X = (X_t)_{t\in[0,1]}$ be a stochastic process (not necessarily Gaussian) on $[0,1]$. Assume, like in the Kolmogorov-Centsov theorem, that $\mathbf{E}[|X_t - X_s|^b] \leq C|t-s|^{1+a}$ for all $t, s \in [0,1]$, for some $a, b > 0$ and $C < \infty$. Then let $p(u) = u^c$ for some $c > 0$ and let $\Psi(u) = u^{\beta}$. These functions are as in the Lemma above, and further,

$$\begin{aligned}
\mathbf{E}[I] &= \int_0^1 \int_0^1 \mathbf{E}\left[\frac{|X_t - X_s|^b}{|t-s|^{cb}}\right] dt\, ds \\
&\leq C \int_0^1 \int_0^1 |t-s|^{1+a-cb}\, dt\, ds \\
&< \infty \qquad \text{if } 1 + a - cb > -1.
\end{aligned}$$

Thus for $c < \frac{a+2}{b}$, almost every sample path satisfies the conditions of the lemma, and we conclude that the modulus of continuity is

$$\omega_X(\delta) \leq 8 \int_0^{\delta} (4I/u^2)^{1/b}\, cu^{c-1} du = CI^{\frac{1}{b}} \delta^{c-\frac{2}{b}}.$$

Thus, we get Hölder continuity with any exponent less that $\frac{a}{b}$. Note that the constant I is random, as one would expect, but it is finite almost surely, as its expectation is finite. This may be the point to note in applying the GRR lemma to random processes. To show that $I < \infty$ w.p.1., one can work on its expectation, and that only requires knowledge of the bivariate distributions (X_t, X_s).

Exercise 18. Assume that the process X is Gaussian. Choose Ψ and p differently to prove that $\omega_X(\delta) \leq C\delta^{\frac{a}{b}} \sqrt{\log \frac{1}{\delta}}$ for all $\delta > 0$, for a random finite constant C.

5.13 Remark on sigma-algebras when T is not countable

Recall that the main lemmas using chaining were stated for countable index sets. What about in general? For definiteness, let us take $T = [0,1]$, the general situation being similar.

As defined, a stochastic process is a random variable X on some $(\Omega, \mathcal{F}, \mathbf{P})$, taking values in $(\mathbb{R}^T, \mathcal{C}_T)$, where $\mathcal{C}_T := \otimes_{t\in T}\mathcal{B}_{\mathbb{R}}$ is the cylinder sigma-algebra. Any event measurable with respect to $\mathcal{C}_T$ depends only on countably many co-ordinates. Hence, natural subsets of $\mathbb{R}^T$ such as $C(T)$, $C^1(T)$, $C^\infty(T)$, etc., are not measurable. The question: "Does $X(\omega)$ belong to $C(T)$ for almost all ω?" does not make literal sense. There are two ways in which we make sense of this and similar questions.

First, let use say that Y (on a possibly different probability space) is a version of X if $Y \stackrel{d}{=} X$. As $\mathcal{C}_T$ is generated by finite dimensional cylinder sets, this is the same as saying that the finite dimensional distributions agree: $(Y(t_1),\ldots,Y(t_k)) \stackrel{d}{=} (X(t_1),\ldots,X(t_k))$ for all $k \geq 1$ and $t_1,\ldots,t_k \in T$. The question, "Is X continuous a.s." is then interpreted as "Is there a version of X that has continuous sample paths a.s.". This distinction is necessary because even if one version is continuous, other versions need not be. For example, when $T = [0,1]$ and X is a continuous stochastic process, if we pick $U \sim \text{uniform}[0,1]$ independent of X, and set $Y_t = X_t \mathbf{1}_{t\neq U}$, then then Y is not continuous (in general). Similar considerations apply to questions such as "Does X have smooth sample paths?". Since there is no reason to use versions that are not the best possible in sample path behaviour, one simply says "X is continuous/smooth", implying that one is considering a version that is continuous/smooth.

A second way, which is just as good for everything of interest, is to consider a countable dense set $D \subseteq T$, such as the set of dyadic rationals. For the restricted stochastic process $X_D = (X_t)_{t\in D}$, the cylinder sigma-algebra has all the events one wants. For example, $\{X \text{ is uniformly continuous on } D\}$ is an event in $\mathcal{C}_D$, as it can be written as

$$\bigcap_{n\geq 1}\bigcup_{k\geq 1}\bigcap_{t,s\in D:|t-s|\leq\frac{1}{k}}\left\{X_t - X_s \leq \frac{1}{n}\right\}.$$

Similarly, $\{X_D \text{ extends to } T \text{ as a smooth function}\}$ is also an event in $\mathcal{C}_D$ (why?). Now, the original question "Is X continuous on T a.s.?" may be interpreted as "Is X_D uniformly continuous on D?". Similarly, "Does X have smooth sample paths?" is interpreted as "Does the event $\{X_D \text{ extends to } T \text{ as a smooth function}\}$ have probability 1?". The only point one should check in this approach is that the answers do not depend on the chosen countable dense set, which we leave as an exercise.

5.14 Smooth Gaussian processes

So far we have been worried about whether the process is good enough to be continuous or bounded. What about asking for more? It turns out to be easier, in a way.

Proposition 19. *Let X be a Gaussian process on a open subset T of $\mathbb{R}^d$ with covariance function K. The sample paths of X are $C^\infty(T)$ almost surely if and only if K is $C^\infty(T\times T)$.*

One can also write counterparts of this theorem for $C^{(p)}$ paths, which is roughly (but not exactly) equivalent to the covariance being $C^{(2p)}$.

Proof. For simplicity of notation we assume $d=1$. And we assume the process is centered as always.

Suppose that X has smooth sample paths almost surely. Let $X_h(t) = (X(t+h)-X(t))/h$ for $h>0$. For each h, the process X_h is a Gaussian process, in fact, X and X_h are jointly Gaussian and

$$\mathbf{E}[X_h(t)X(s)] = \frac{1}{h}(K(t+h,s)-K(t,s)).$$

As $X_h(t)\to X'(t)$ *a.s.* as $h\to 0$, and for Gaussians this implies convergence of the parameters, the left hand side converges to $\mathbf{E}[X'(t)X(s)]$. This shows that $K(\cdot,s)$ is differentiable and has derivative $\partial_1 K(t,s) = \mathbf{E}[X'(t)X(s)]$. Now play the same game to see that

$$\frac{\partial_1 K(t,s+h)-\partial_1 K(t,s)}{h} = E[X'(t)X_h(s)] \to \mathbf{E}[X'(t)X'(s)]$$

to see that $\partial_1\partial_2 K(t,s) = \mathbf{E}[X'(t)X'(s)]$. More generally, inductively one can see that $\partial_1^p\partial_2^q K(t,s) = \mathbf{E}[X^{(p)}(t)X^{(q)}(s)]$ for any $p,q\geq 1$.

Conversely, suppose K is smooth. Without loss of generality, let us assume that $X(0)=0$ *a.s.* (otherwise we work with the process $X(t)-X(0)$). From the calculations above, we see that

$$\mathbf{E}[X_h(t)X_h(s)] = \frac{K(t+h,s+h)-K(t+h,s)-K(t,s+h)+K(t,s)}{h^2} \to \partial_1\partial_2 K(t,s).$$

This shows that $\partial_1\partial_2 K(t,s)$ is a limit of covariance functions and hence positive semi-definite. Let Y be a Gaussian process with covariance $\partial_1\partial_2 K$. As the covariance is smooth, it is clear from the Dudley integral that Y is continuous (i.e., we may choose Y to be continuous) a.s. Define $\hat{X}(t) = \int_0^t Y(s)ds$. Since we assumed that $X(0)=0$, it follows that

$K(t,0) = 0 = K(0,s)$ for all t,s, from which it is easy to see that (this is for the second line below)

$$\begin{aligned}\mathbf{E}[\hat{X}(t)\hat{X}(s)] &= \int_0^t \int_0^s \partial_1\partial_2 K(u,v)\, dv\, du \qquad \text{(by linearity)}\\ &= K(t,s).\end{aligned}$$

Thus, $\hat{X}$ is a version of X. But $\hat{X}$ is C^1 with derivative equal to Y. Continuing this way, we see that for any p, the process X has a version that is $C^{(p)}$.

To conclude that X has an infinitely smooth version, fix a countable dense subset $D \subseteq \mathbb{R}$ as in the previous section. Let $\mathcal{A}_p$ be the even that $X_D = (X_t)_{t\in D}$ extends as a $C^{(p)}$ function to all of $\mathbb{R}$. By the proof so far, $\mathbf{P}\{\mathcal{A}_p\} = 1$ for all $p \geq 1$, hence $\mathbf{P}\{\cap_p \mathcal{A}_p\} = 1$. That means that there is an infinitely smooth version. ■

Another way to prove this proposition is to express X as a random series (which is possible as we shall see later). The following exercise is a simpler illustration of this idea.

Exercise 20. Show that X has a real-analytic version on $\mathbb{R}^d$ if and only if K is real-analytic on $\mathbb{R}^d \times \mathbb{R}^d$.

Stationary processes

From Fourier analysis, we know that if μ is a finite measure on $\mathbb{R}$, then $\hat{\mu}$ is infinitely smooth if and only if μ has moments of all orders. Therefore, a stationary Gaussian process has smooth sample paths if and only if its spectral measure has moments of all orders. Similarly, real-analyticity of the sample paths (and hence of the covariance function) corresponds to existence of moment generation function (in a neighbourhood of the origin) of the spectral measure.

Chapter 6

Stationary Gaussian processes

Let $T = \mathbb{R}^d$ or $\mathbb{Z}^d$. For some parts of this chapter we stick to one or the other, or $d = 1$, while other parts may be generalizable to other groups.

6.1 Parameterizing stationary Gaussian processes

Let T be $\mathbb{Z}^d$ or $\mathbb{R}^d$ and let $X = (X_t)_{t\in T}$ be a stationary Gaussian process on T. Its distribution is determined by the covariance function, which is of the form $\mathbf{E}[X_t X_s] = K(t-s)$, where $K : T \mapsto \mathbb{R}$ is a positive semi-definite function. Assuming continuity, we have seem that such a function is necessarily the Fourier transform of a finite symmetric measure μ on $\hat{T}$ (where $\hat{T} = [-\pi,\pi]^d$ if $T = \mathbb{Z}^d$ and $\hat{T} = \mathbb{R}^d$ if $T = \mathbb{R}^d$), which is known as its spectral measure. Sometimes it gets annoying and restrictive to impose symmetry on the spectral measure, hence some authors consider complex-valued Gaussian processes (meaning $X + iY$, where X, Y are jointly Gaussian processes), in which case, the spectral measure is any finite measure on $\hat{T}$. But we stick to real-valued processes.

In this sense, stationary Gaussian processes on T are parameterized by either the class of positive semi-definite functions or by the class of finite symmetric measures. There are other parametrizations that are possible and sometimes convenient. We introduce one of them now[1].

Case $T = \mathbb{Z}$: Let μ be any finite measure on $S^1 = [-\pi,\pi]$. One can apply Gram-Schmidt pro-

[1]The parts about Verblunsky coefficients and OPUC are taken from Barry Simon's book *Orthogonal polynomials on the unit circle Part 1: Classical theory* and Bingham's paper *Szegö's theorem and its probabilistic descendents*.

cedure to $e_k(z) = z^k$, $k \geq 0$, in $L^2(S^1, \mu)$, to get an orthonormal basis $\{\varphi_0(z), \varphi_1(z), \varphi_2(z), \ldots\}$. Then φ_k is a polynomial of degree k, and we write $\varphi_n(z) = \kappa_n \Phi_n$, where Φ_n is a monic polynomial and $\kappa_n^{-2} = \|\Phi_n\|^2_{L^2(\mu)}$. If μ is not finitely supported, these polynomial sequences are infinite. Let $P^*(z) = z^n \overline{P(1/\bar{z})}$. If $P(z) = a_0 + a_1 z + \ldots + a_n z^n$, then $P^*(z) = \bar{a}_n + \bar{a}_{n-1} z + \ldots + \bar{a}_1 z^{n-1} + \bar{a}_0 z^n$. Observe that $\Phi_0^*, \ldots, \Phi_n^*$ also span the space of degree n polynomials. The key point is that Φ_k is orthogonal to $e_0, \ldots, e_{k-1}$ while Φ_k^* is orthogonal to $e_1, \ldots, e_k$.

The polynomial $e_1 \Phi_k - \Phi_{k+1}$ has degree k and hence ($\bar{e}_1 = e_{-1}$ on $\mathbb{S}^1$)

$$\langle e_1 \Phi_k - \Phi_{k+1}, e_j \rangle = \langle \Phi_k, e_{j-1} \rangle - \langle \Phi_{k+1}, e_j \rangle \; = \; 0$$

for $1 \leq j \leq k-1$. Therefore, we must have $e_1 \Phi_k - \Phi_{k+1} = \overline{\alpha}_k \Phi_k^*$ for some $\alpha_k \in \mathbb{C}$. By the orthogonality,

$$\begin{aligned} \|e_1 \Phi_k\|^2 &= \|\Phi_{k+1}\|^2 + |\alpha|^2 \|\Phi_k^*\|^2 \\ &= \|\Phi_{k+1}\|^2 + |\alpha|^2 \|\Phi_k\|^2 \end{aligned}$$

showing that $|\alpha_k| < 1$ and $\kappa_{n+1}^{-2} = \kappa_n^{-2}(1 - |\alpha_n|^2)$. Hence, $\kappa_n^{-2} = (1 - |\alpha_0|^2) \ldots (1 - |\alpha_{n-1}|^2)$ for all n (the right side is interpreted as 1 when $n = 0$). The numbers α_k are called *Verblunsky coefficients*. We summarize the *Szegö recursions*

$$\begin{aligned} \Phi_{n+1}(z) &= z \Phi_n(z) - \overline{\alpha}_n \Phi_n^*(z), \\ \Phi_{n+1}^*(z) &= -\alpha_n z \Phi_n(z) + \Phi_n^*(z). \end{aligned}$$

The second equation can be got by applying the $*$ operation to the first equation. It is a fact that we do not justify here, that the correspondence between measures μ on $\mathbb{S}^1$ with infinite support and sequences $\alpha = (\alpha_n)_{n \geq 0} \in \mathbb{D}^{\mathbb{N}}$, is in fact a bijection. The measure μ is symmetric if and only if α_n is real for each n (why?).

Remark 1. It is often useful to write this in matrix form

$$\begin{bmatrix} \Phi_{n+1}(z) \\ \Phi_{n+1}^*(z) \end{bmatrix} = \begin{bmatrix} z & -\overline{\alpha}_n \\ -\alpha_n z & 1 \end{bmatrix} \begin{bmatrix} \Phi_n(z) \\ \Phi_n^*(z) \end{bmatrix}.$$

This suggests that the linear fractional transformation $z \mapsto \frac{z - \overline{\alpha}_n}{1 - \alpha_n z}$ is lurking in the background.

The upshot of all this is that centered Gaussian processes (on $\mathbb{Z}$ may also be parameterized by $I^{\mathbb{N}}$ where $I = (-1, 1)$(strictly speaking, by $I^{\mathbb{N}} \sqcup \{-1, 1\} \sqcup I \times \{-1, 1\} \sqcup I^2 \times \{-1, 1\} \sqcup \ldots$,

if you allow spectral measures with finite support) via the Verblunsky coefficients of the spectral measure. This is in some ways better than the original parameterization by the correlation function $K = (K(n))_{n\geq 0}$, because the correlation coefficients satisfy complicated positivity inequalities, while the Verblunsky coefficients are *unrestricted*! On the other hand, the probabilistic interpretation of the Verblunsky coefficients is less obvious.

Exercise 2. Let $P_{[j,k]}$ denote the projection operator in $L^2(\mu)$ onto $\text{span}\{e_q : j \leq q \leq k\}$. Then $\alpha_n = \langle (I - P_{[0,n]})e_n, (I - P_{[-n,-1]})e_0 \rangle$. By the isomorphism, this is the same as the correlation between $X_n - \mathbf{E}[X_n \,\big|\, X_0, \ldots, X_{n-1}]$ and $X_0 - \mathbf{E}[X_0 | X_{-1}, \ldots, X_{-n}]$. In Statistics, it is known as the partial autocorrelation function of n.

Remark 3. If μ was supported on n points, then $\Phi_n = 0$, which shows that $|\alpha_{n-1}|^2 = 1$. Thus, in the finite support case, we have $(\alpha_0, \ldots, \alpha_n) \in \mathbb{D}^{n-1} \times \mathbb{S}^1$. This is in fact a one-one parameterization of measures supported on exactly n points.

6.2 The one sided prediction problem

Consider a stationary Gaussian process on $\mathbb{Z}$ or $\mathbb{R}$. The one-dimensionality is important here (there may be analogues in higher dimension). For definiteness, let us consider $X = (X_n)_{n\in\mathbb{Z}}$, a centered Gaussian process with spectral measure μ on $[-\pi, \pi]$. Let $\mathcal{F}_t = \sigma\{X_s : s \leq t\}$. The best predictor in L^2-sense for X_0 given $\mathcal{F}_{-1}$ is $\mathbf{E}[X_0 \,\big|\, \mathcal{F}_{-1}]$ and the prediction error is

$$\mathbf{E}\left[|X_0 - \mathbf{E}[X_0 \,\big|\, \mathcal{F}_{-1}]|^2\right] = \text{Var}(X_0 \,\big|\, \mathcal{F}_{-1}).$$

On the right one should say $\mathbf{E}[\text{Var}(X_0 \,\big|\, \mathcal{F}_{-1})]$ in general, but we know that for Gaussians, the conditional variance does not depend on the conditioned values of the random variables.

Here is a characterisation of the situations in which the prediction is perfect!

Theorem 4. *[Szegö]Let $d\mu(x) = w(x)\frac{dx}{2\pi} + d\mu_s(x)$ where μ_s is singular to Lebesgue measure and w is the density of the absolutely continuous part. Then, the prediction error is zero if and only if $\int_{[-\pi,\pi]} \log w(x)dx > -\infty$. Further,*

1. *The prediction error is precisely $\exp\left\{\frac{1}{2\pi}\int_{[-\pi,\pi]} \log w(x)dx\right\}$.*
2. *In case where the prediction error is zero, $Var(X_n \,\big|\, \mathcal{F}_m) = 0$ for any $m < n$.*

First we make some preparations. Let $\Sigma_n = (\hat{\mu}(j-k))_{1\le j,k\le n}$ and let $v_n^t = (\hat{\mu}(1),\ldots,\hat{\mu}(n))$. Let $\sigma_n^2 = \mathrm{Var}(X_0 \mid X_{-1},\ldots,X_{-n})$ which is the same as $\|(I-P_{[-n,-1]})e_0\|^2_{L^2(\mu)}$ by the isomorphism theorem. Clearly σ_n^2 is decreasing, its limit σ^2 exists. By the interpretation as norms of projections onto decreasing subspaces, it is also clear that $\sigma^2 = \mathrm{Var}(X_0 \mid \mathcal{F}_{-1})$ which is the same as $\|(I-P_{(-\infty,-1]})e_0\|^2_{L^2(\mu)}$. Therefore

$$\sigma_n^2 = \min\left\{\int |U|^2 d\mu : U = e_0 + c_1 e_{-1} + \ldots + c_n e_{-n},\ c_k \in \mathbb{R}\right\}. \tag{1}$$

For U as above, Ue_n is a monic polynomial, and has the same $L^2(\mu)$ norm, hence the above infimum may also be taken over monic polynomials of degree n in one complex variable.

This is the form that we shall use, but it is worth noting that there are several other forms that one could write. For instance, by the formulas for conditional distribution of Gaussians,

$$\sigma_n^2 \;=\; \frac{\det(\Sigma_{n+1})}{\det(\Sigma_n)} \;=\; \hat{\mu}(0) - v_n^t \Sigma_n^{-1} v. \tag{2}$$

Yet another formula is in terms of the Verblunsky coefficients:

$$\sigma_n^2 = \prod_{k=0}^{n-1}(1-|\alpha_k|^2). \tag{3}$$

Therefore, the following quantities are all equal.

1. $\mathrm{Var}(X_0 \mid \mathcal{F}_{-1})$.
2. $\inf\{\|U\|_{L^2(\mu)} : U \text{ is a monic polynomial in one variable}\}$.
3. $\exp\left\{\frac{1}{2\pi}\int_{-[\pi,\pi]} \log w(x)dx\right\}$.
4. $\lim_{n\to\infty} \frac{\det(\Sigma_{n+1})}{\det(\Sigma_n)}$ and $\lim_{n\to\infty} \det(\Sigma_n)^{1/n}$.
5. $\prod_{n=0}^{\infty}(1-|\alpha_n|^2)$.

Even if one forgets the probabilistic interpretation, the equality of the other quantities is non-trivial and interesting! For example, the equality of the third and fourth expressions above gives asymptotics of Toeplitz determinants (a Toeplitz matrix is one whose (i,j)

entry depends only on $j-i$). There is a more refined theorem on this asymptotics due to Ibragimov and Golinskii which we do not go into here[2].

First half of the proof of Theorem 4. For any U as in (1),

$$\begin{aligned}\int |U|^2 d\mu &\geq \int |U(e^{ix})|^2 w(x)\frac{dx}{2\pi} \\ &\geq \exp\left\{\int \log(|U(e^{ix})|^2 w(x))\frac{dx}{2\pi}\right\} \\ &= \exp\left\{\int \log|U(e^{ix})|^2\frac{dx}{2\pi}\right\}\exp\left\{\int \log w(x)\frac{dx}{2\pi}\right\}.\end{aligned}$$

But $\overline{U}(z) = c_n z^n + \ldots + c_1 z + 1$ is holomorphic, hence $\log|U|^2$ is subharmonic[3], and hence

$$\int \log|U(x)|^2\frac{dx}{2\pi}. \;\geq\; \log|U(0)|^2 \;=\; 0.$$

Thus, we have shown that for all n,

$$\sigma_n^2 \geq \exp\left\{\int \log w(x)\frac{dx}{2\pi}\right\}$$

and hence $\sigma^2 \geq \exp\left\{\int \log w(x)\frac{dx}{2\pi}\right\}$. ■

It remains to prove that $\sigma^2 \leq \exp\left\{\int \log w(x)\frac{dx}{2\pi}\right\}$. Looking back to where the inequality appeared in the lower bound, there are three places:(1) When we replaced μ by its absolutely continuous part $w(x)dx$. (2) When we applied Jensen's inequality to $|U|^2 w$. (3) When we used subharmonicity of $\log|U|^2$. According to Remark 6 below, $|U|^2$ can never have roots in the unit disk, hence $\log|U|^2$ is actually harmonic, showing that the third circumstance was in fact spurious. In addition, this allows us to rewrite (1) with the infimum over non-negative trigonometric polynomials and by approximation, over continuous functions satisfying $\int \log h(x)dx = 0$.

$$\sigma^2 = \inf\left\{\int h d\mu : h \in C(\mathbb{S}^1),\ h \geq 0,\ \int \log h(x)dx = 0\right\}$$

[2]A comprehensive reference for the Szegö theorems are Barry Simon's two books *Orthogonal polynomials on the unit circle*, Part-1 and Part-2. In particular Part 1 contains at least five proof of Szegö's theorem and the stronger form mentioned here. However, our presentation of the proof is taken from the older book of Grenander and Szegö *Toeplitz forms and their applications*.

[3]If subharmonicity is not familiar, here is the point: Write $e_n(z)U(z) = (z-\zeta_1)\ldots(z-\zeta_n)$ to see that $\int \log|U|^2\frac{dx}{2\pi} = 2\sum_{k=1}^n \int_{-\pi}^{\pi}\log|e^{ix}-\zeta_k|\frac{dx}{2\pi}$. We leave it as an exercise to show that $\int \log|e^{ix}-\zeta|\frac{dx}{2\pi} = \mathbf{1}_{|\zeta|\geq 1}\log|\zeta|$. Then it follows that $\int \log|U(x)|^2\frac{dx}{2\pi} = 2\sum_{k:|\zeta_k|\geq 1}\log|\zeta_k|$ which is no less than $\log|U(0)|^2 = 2\sum_k \log|\zeta_k|$.

We need to deal with the first two.

For the second one, we can achieve equality in Jensen's if we can choose U so that $|U|^2 w$ is a constant. That is not possible, hence we must proceed by approximation. We do this first assuming that μ is absolutely continuous and in the end take care of the singular part if it exists.

Proof of the second half of Theorem 4 for absolutely continuous μ. First assume that $w(x) \geq \delta > 0$ for a.e. x. Then $1/w$ is integrable, hence by Lemma 5 and the density of continuous functions in $L^1(\mathbb{S}^1)$, there exist U_n having no roots inside the disk such that $|U_n|^2 \to \frac{1}{w}$ in $L^1(\mathbb{S}^1)$. ■

To give a very quick idea of what works, consider the restricted case where $d\mu(x) = w(x)\frac{dx}{2\pi}$ where w is continuous on $\mathbb{S}^1$ and $w \geq \delta$ for some $\delta > 0$. In this case, $\frac{1}{\sqrt{w}}$ is a continuous function and can be approximated uniformly by trigonometric polynomials U_n. Then $|U_n|^2 \to \frac{1}{w}$ and $wU_n^2 \to 1$ and $\log(w|U_n|^2) \to 0$, all uniformly on $\mathbb{S}^1$. By Lemma 5, w e can also arrange it so that U_n has no roots inside the unit disk, in which case $\int \log |U_n|^2 = 0$. Consequently,

First suppose that $\frac{1}{w} \in L^1$, i.e., $h := \frac{1}{\sqrt{w}} \in L^2$. Then the partial sums of the Fourier series of h (which are just $h * D_n$, where D_n is the Dirichlet kernel) converge to h in L^2. Hence, $h * K_n$ also converge to h in L^2, where K_n is the Fejer kernel. Recall that $K_n = \frac{1}{n+1}(D_0 + \ldots + D_n)$, and that K_n is in fact a probability density on $\mathbb{S}^1$.

To avoid problems of taking reciprocals, fix $\delta > 0$ and let $w_\delta = w + \delta$, which is bounded below. By the density of trigonometric polynomials in L^1, we can find some U such that $\|U - \frac{1}{\sqrt{w_\delta}}\|_{L^1} \leq \varepsilon$. Here L^1 is with respect to uniform measure on $\mathbb{S}^1$. Then, $\|U^2 w_\delta - 1\|_{L^1} \leq$.

Trigonometric polynomials and approximation

By the Stone-Weierstrass theorem, trigonometric polynomials (which are, by definition, finite linear combinations of e_n, $n \in \mathbb{Z}$) are dense in $C(\mathbb{S}^1)$ (with the sup-norm metric). Hence they are are also dense in $L^p(\mathbb{S}^1)$ for $1 \leq p < \infty$. Here is a more refined approximation statement that we used above.

Lemma 5 (Fejér-Riesz)**.** *Suppose $f \in C(\mathbb{S}^1)$ and $f \geq \delta$. Then there are trigonometric polynomials T_n such that (a) $T_n \geq \delta$ and (b) T_n has no roots inside the open unit disk and (c) $T_n \to f$ uniformly on $\mathbb{S}^1$.*

In this statement and elsewhere, we freely extend trigonometric polynomials holomorphically outside $\mathbb{S}^1$, since they are all of the form $P(z)/z^n$ for a polynomial P.

Proof. One can add δ at the end to f and T_n, hence we assume that $\delta = 0$.

Suppose $f \in C(\mathbb{S}^1)$ and $f \geq 0$. Take trigonometric polynomials T_n converging uniformly to $\sqrt{f}$. Then $\{T_n\}$ and f are uniformly bounded in absolute value by some number M. Hence $\|f - T_n^2\|_{\sup} \leq 2\sqrt{M}\|\sqrt{f} - T_n\|_{\sup}$ also goes to zero. Of course T_n^2 is a non-negative trigonometric polynomial.

To satisfy the condition about roots, suppose T_n has roots $\alpha_1, \ldots, \alpha_k$ in the unit disk. Let

$$\tilde{T}_n(z) = T_n(z) \prod_{j=1}^{k} \frac{1 - \overline{\alpha}_j z}{z - \alpha_j}.$$

Then $\tilde{T}_n$ has no roots inside the disk. As $|\tilde{T}_n| = |T_n|$ on $\mathbb{S}^1$ it follows that $|\tilde{T}_n|^2$ is a trigonometric polynomial (even if T_n is not) and that $|\tilde{T}_n|^2 \to f$ uniformly on $\mathbb{S}^1$. ■

Remark 6. The proof in fact shows that $|T|^2$ cannot have zeros in the open unit disk. This is because it is equal to $|\tilde{T}|^2$ on $\mathbb{S}^1$, and hence everywhere. As we show below, every non-negative trigonometric polynomial is of the form $|T|^2$ for a trigonometric polynomial T, hence no non-negative trigonometric polynomial has zeros in the open unit disk.

Let us justify the claim that any non-negative trigonometric polynomial S are of the form $|T|^2$ for some trigonometric polynomial T (we are allowing complex coefficients here). To see this, write $S(z) = P(z)/z^n$ for some polynomial P. Since $z = 1/\bar{z}$ on $\mathbb{S}^1$ and S is real-valued there, $S(z) = \overline{S}(1/\bar{z})$ for all $z \in \mathbb{C} \setminus \{0\}$ and consequently $z^{2n}\overline{P}(1/\bar{z}) = P(z)$. Thus, if ζ is a root of P, then so is $1/\overline{\zeta}$ (what about $\zeta = 0$?). The roots that are on the unit circle must occur with even multiplicity (or else on one side of $\mathbb{S}^1$ of such a root, S must take negative values). Conclude that $S = |T|^2$ for some trigonometric polynomial T.

6.3 Ergodicity and mixing: Statements of the results

Everywhere T will denote $\mathbb{R}^d$ or $\mathbb{Z}^d$. When writing proofs we often stick to $d = 1$ for simplicity of notation. Let $\mathbb{Q}$ be a probability measure on $(\mathbb{R}^T, \mathcal{C})$, where $\mathcal{C}$ denotes the cylinder sigma-algebra on $\mathbb{R}^T$. The group T acts on $\mathbb{R}^T$ by translations: $\tau_s\omega(\cdot) = \omega(\cdot + s)$. We say that $\mathbb{Q}$ is stationary (for shifts) if $\mathbb{Q} \circ \tau_s^{-1} = \mathbb{Q}$ for all $s \in T$. We also say that shifts are *measure preserving* on $(\mathbb{R}^T, \mathcal{C}, \mathbb{Q})$.

If $X = (X_t)_{t\in T}$ is a stationary stochastic process on some $(\Omega, \mathcal{F}, \mathbf{P})$, then its distribution $\mathbb{Q} = \mathbf{P} \circ X^{-1}$ on $(\mathbb{R}^T, \mathcal{C})$ is stationary in the above sense. Conversely, for any stationary $\mathbb{Q}$, the projections $\Pi = (\Pi_t)_{t\in T}$ form a stationary stochastic process in the sense defined earlier. In short, we may assume that the probability space on which our stochastic process is defined is $(\mathbb{R}^T, \mathcal{C}, \mathbb{Q})$.

Below T will denote $\mathbb{R}^d$ or $\mathbb{Z}^d$. What is special about these index sets is that they are (abelian) groups. If $X = (X_t)_{t\in T}$ is a stochastic process, then its translate X^τ is another stochastic process on T defined as $X^\tau(t) = X(t-\tau)$. The process X is called *stationary* (or translation invariant) if $X^\tau \stackrel{d}{=} X$ for all $\tau \in T$.

Ergodicity: An event $A \in \mathcal{C}$ is said to be invariant if $\tau_s^{-1}(A) = A$ for all $s \in T$. The measure $\mathbb{Q}$ is said to be ergodic if any invariant event A has $\mathbb{Q}(A) = 0$ or $\mathbb{Q}(A) = 1$. Observe that the set of all invariant events $\mathbb{I}$ forms a sigma-algebra. Hence ergodicity is the statement that $\mathbb{Q}$ is trivial on $\mathbb{I}$.

Mixing: $\mathbb{Q}$ is said to be *mixing* if $\mathbb{Q}(A \cap \tau_s^{-1}B) \to \mathbb{Q}(A)\mathbb{Q}(B)$ for all $A, B \in \mathcal{C}$.

As always, we say that a stationary process is ergodic or mixing, if its distribution $\mathbb{Q}$ on $(\mathbb{R}^T, \mathcal{C})$ is ergodic or mixing, respectively. The two main theorems that we wish to prove are as follows.

Theorem 7 (Maruyama, Grenander, Fomin)**.** *A stationary Gaussian process on $\mathbb{R}^d$ or $\mathbb{Z}^d$ is ergodic if and only if its spectral measure has no atoms.*

Theorem 8. *A stationary Gaussian process on $\mathbb{R}^d$ or $\mathbb{Z}^d$ with covariance $K(t-s)$ is mixing if and only if $K(t) \to 0$ as $|t| \to \infty$.*

There are multiple notions of mixing. Here are a few others, in increasing order of strength.

1. The Cesaro average of $|Q(A \cap \tau_s^{-1}B) - \mathbb{Q}(A)\mathbb{Q}(B)|$ converges to 0 as $s \to \infty$, for any $A, B \in \mathcal{C}$.
2. $|Q(A \cap \tau_s^{-1}B) - \mathbb{Q}(A)\mathbb{Q}(B)|$ converges to 0 as $s \to \infty$, for any $A, B \in \mathcal{C}$. This was what we called mixing above.
3. $\sup_{B\in\mathcal{C}} |\mathbb{Q}(A \cap \tau_s^{-1}B) - \mathbb{Q}(A)\mathbb{Q}(B)| \to 0$ for all $A \in \mathcal{C}$ as $s \to \infty$.
4. $\sup_{A,B\in\mathcal{C}} |\mathbb{Q}(A \cap \tau_s^{-1}B) - \mathbb{Q}(A)\mathbb{Q}(B)| \to 0$ as $s \to \infty$.

5. $\sup_{A\in\mathcal{C}} |\mathbb{Q}(\tau_s^{-1}B \,\big|\, A) - \mathbb{Q}(B)| \to 0$ as $s\to\infty$.

What for Gaussian processes?

6.4 Proof of the mixing theorem

Here we prove Theorem 8.

Since $K(t) = \mathbf{E}[X_0X_t]$, it is clear that it must go to zero to have mixing. But since the definition of mixing was for events and not unbounded random variables, let us elaborate on this point. assume $K(0) = 1$ without loss of generality and observe that

$$\mathbf{P}\{X_0 > 0, X_t > 0\} = \frac{1}{4} - \frac{1}{2\pi}\arccos K(t).$$

For mixing to hold, this must converge to $\mathbf{P}\{X_0 > 0\}^2 = \frac{1}{4}$, as $t\to\infty$. That happens if and only if $K(t)\to 0$ as $t\to\infty$.

To prove the other way implication, assume that $K(t)\to 0$. Consider any two cylinder sets $A = \Pi^{-1}_{t_1,\ldots,t_n}(C)$ and $B = \Pi^{-1}_{s_1,\ldots,s_m}(D)$ where $C\in\mathcal{B}_{\mathbb{R}^n}$ and $D\in\mathcal{B}_{\mathbb{R}^m}$. We may and shall assume that $(X(t_1),\ldots,X(t_n))$ and $(X(s_1),\ldots,X(s_m))$ have non-singular covariance matrices. For, if say $X(t_1),\ldots,X(t_n)$ satisfy some linear relationship among themselves, we can choose a subset $\{X(t_{i_1}),\ldots,X(t_{i_p})\}$ of them that forms a basis for $\text{span}\{X(t_1),\ldots,X(t_n)\}$ and express A as $\Pi^{-1}_{t_{i_1},\ldots,t_{i_p}}(C')$ for some $C'\in\mathcal{B}_{\mathbb{R}^p}$.

Then $\mathbb{Q}(A\cap\tau_s^{-1}B)$ is the probability that the vector $(X_{t_1},\ldots,X_{t_n},X_{s_1+s},\ldots,X_{s_m+s})$ belongs to $C\times D$. As the covariance matrix is

$$\begin{bmatrix} (K(t_i-t_j))_{i,j\le n} & (K(t_i-s_j-s))_{i\le n,j\le m} \\ (K(t_j-s_i-s))_{i\le n,j\le m} & (K(s_i-s_j))_{i,j\le m}\end{bmatrix}$$

we see that as $s\to\infty$, it converges to block-diagonal form. By our assumption, the two diagonal blocks are non-singular, hence for large s the above matrix is also non-singular, and the Gaussian density of the above covariance converges to that of $(X_{t_1},\ldots,X_{t_n},Y_{s_1},\ldots,Y_{s_m})$, where Y is an independent copy of X. From the convergence of densities, it follows that for *any* Borel C,D, as $s\to\infty$,

$$\begin{aligned}\mathbf{P}\{(X_{t_1},\ldots,X_{t_n},X_{s_1+s},\ldots,X_{s_m+s})\in C\times D\} &\to \mathbf{P}\{(X_{t_1},\ldots,X_{t_n},X_{s_1+s},\ldots,X_{s_m+s})\in C\times D\}\\ &= \mathbf{P}\{(X_{t_1},\ldots,X_{t_n})\in C\}\times\mathbf{P}\{(X_{s_1},\ldots,X_{s_m})\in D\}.\end{aligned}$$

which is the same as $\mathbb{Q}(A \cap \tau_s^{-1} B) \to \mathbb{Q}(A)\mathbb{Q}(B)$. This completes the proof for cylinder sets.

If $A, B \in \mathcal{C}$, then for any $\varepsilon > 0$, there are cylinder sets A_1, B_1 such that $\mathbb{Q}(A \Delta A_1) < \varepsilon$ and $\mathbb{Q}(B \Delta B_1) < \varepsilon$. Hence

$$\begin{aligned}|\mathbb{Q}(A \cap \tau_s^{-1} B) - \mathbb{Q}(A)\mathbb{Q}(B)| &\leq |\mathbb{Q}(A_1 \cap \tau_s^{-1} B_1) - \mathbb{Q}(A_1)\mathbb{Q}(B_1)| \\ &\quad + |\mathbb{Q}(A \cap \tau_s^{-1} B) - \mathbb{Q}(A \cap \tau_s^{-1} B)| + |\mathbb{Q}(A)\mathbb{Q}(B) - \mathbb{Q}(A_1)\mathbb{Q}(B_1)|.\end{aligned}$$

The second and third summands are bounded by 2ε and the first goes to zero as $s \to \infty$. As ε is arbitrary, this shows that $\mathbb{Q}(A \cap \tau_s^{-1} B) \to \mathbb{Q}(A)\mathbb{Q}(B)$. ■

6.5 Proof of the ergodicity theorem

Now we prove Theorem 7.

First suppose the spectral measure μ has an atom. Then there is some $a \in T$ such that $\mu\{a\} = \mu\{-a\} = p > 0$. Consider the isomorphism between $\mathbb{H} = \overline{\text{span}}\{\Pi_t : t \in T\}$ (a closed subspace of $L^2(\mathbb{R}^T, \mathcal{C}, \mathbb{Q})$) with $L^2(\mu)$ given by $X_t \leftrightarrow e_t$, where $e_t : T \mapsto \mathbb{C}$ is given by $e_t(x) = e^{i\langle t, x\rangle}$. The definition of the spectral measure is that $\mathbf{E}[X_t X_s] = \hat{\mu}(t-s) = \langle e_t, e_s \rangle_{L^2(\mu)}$, hence this is an isomorphism. Now let $\xi \in \mathbb{H}$ correspond to $\mathbf{1}_a$. Then $\mathbf{E}[\xi^2] = \mu\{a\} > 0$ and $\mathbf{E}[\xi] = 0$ (all random variables in $\mathbb{H}$ have zero mean, as we assume that X is centered). Hence ξ is not a constant. Further, we claim that ξ is invariant.

To see this, observe that in the case $T = \mathbb{Z}$, we have

$$\frac{1}{2T} \sum_{k=-T}^{T} e_k(x-a) \to \mathbf{1}_a(x)$$

pointwise and in $L^2(\mu)$, therefore, ξ is the limit in $\mathbb{H}$ of

$$\frac{1}{2T} \sum_{k=-T}^{T} e^{-ika} \Pi_k.$$

But then $\tau_m \xi$ is the limit in $\mathbb{H}$ of

$$\frac{1}{2T} \sum_{k=-T}^{T} e^{-ika} \Pi_{k+m} = e^{ima} \frac{1}{2T} \sum_{k=-T}^{T} e^{-i(k+m)} \Pi_{k+m}$$

which differs from

Chapter 7

Karhunen Loeve expansion

Reproducing kernel Hilbert spaces: A Hilbert space (we take the scalar field to be reals) H of functions on a set T is called a reproducing kernel Hilbert space (rkhs) if for each t, the functional $\Lambda_t(f) = f(t)$ defines a bounded linear functional Λ_t on H. Then, there must exist an element $K_t \in H$ such that $\Lambda_t(f) = \langle f, K_t \rangle$. We shall write $K(s,t)$ for $K_t(s)$ and consequently $K(\cdot,t)$ for $K_t(\cdot)$. Some observations.

1. By symmetry of inner product $\langle K_s, K_t \rangle = \langle K_t, K_s \rangle$. Therefore $K(s,t) = K(t,s)$.

2. For any $f \in H$ we have $\langle f, K(\cdot,t) \rangle = f(t)$, in particular,

Let X be a centered Gaussian process on T with covariance kernel K. We define two Hilbert spaces.

1. For each $t,s \in T$, define the inner product between the functions $K(\cdot,t)$ and $K(\cdot,s)$ to be $K(t,s)$. This is a valid definition of a (pseudo) inner product because K is p.s.d. In other words, on $\text{span}\{K(\cdot,t) : t \in T\}$ we get a well-defined inner product by setting

$$\langle \sum_{i=1}^{n} a_i K(\cdot,t_i), \sum_{j=1}^{m} b_j K(\cdot,s_j) \rangle := \sum_{i,j} a_i b_j K(t_i,s_j).$$

2.

Chapter 8

Zeros of smooth Gaussian processes

Let X be a smooth Gaussian process on a subset of $\mathbb{R}^n$. The random set $X^{-1}\{u\}$ is called the *level set of level u*. We are interested in studying the properties of this random set. In particular, in measuring its size. More generally, one may be interested in simultaneous zeros of several jointly Gaussian processes $X_1, \ldots, X_m$. Before going into that, one must know what kind of a set it is.

By dimension considerations, one would expect that generically, $X_1^{-1}\{u_1\} \cap \ldots \cap X_m^{-1}\{u_m\}$ must be a manifold of dimension $n-m$ if $n \geq m$. When $n = m$ the zero-dimensional set is perhaps a discrete set. When $n < m$, we expect the set to be empty.

First we show that these expectations are indeed correct. Once that is done, one may measure the size of the set by the appropriate Hausdoff measure. Of course, one may choose other measures, such as counting the number of connected components of the set.

8.1 Bulinskaya's lemma

Let us show that generically $n+1$ functions on $\mathbb{R}^n$ will not have a common zero. Here and in general, we do not aim for optimal conditions under which theorems hold, but only for sufficiently powerful theorems that cover examples of interest to us (which are usually very "nice").

Lemma 1 (Bulinskaya)**.** *Let U be an open set in $\mathbb{R}^n$ and let $g : U \mapsto \mathbb{R}^{n+1}$ be a random function. Assume that (A) $g \in C^1(U)$ a.s. and (B) The vector $g(x)$ has a density on $\mathbb{R}^{n+1}$ that is bounded uniformly over x in compact subsets of U. Then $g^{-1}\{0\}$ is almost surely empty.*

Proof. It suffices to assume that $g = (g_0, g_1, \ldots, g_n)$ is defined on a neighbourhood of $I_{n+1} =$

$[0,1]^{n+1}$ and show that $g^{-1}\{0\} \cap I_{n+1} = \emptyset$ a.s. If not, there is a point $x \in I_{n+1}$ such that $g_j(x) = 0$ for all $0 \le j \le n$.

Fix $q \ge 1$ and let $D_q = \{k2^{-q} : 1 \le k \le 2^q\}$. Then there is a point $t \in D_q^{n+1}$ such that $\|x-t\|_\infty \le 2^{-q}$. Writing $\|Dg\|_\infty := \sup\{\|\nabla g_j(y)\| : 0 \le j \le n,\ y \in I_{n+1}\}$, it follows that $|g_j(t)| \le \|Dg\|_\infty 2^{-q}$ for $0 \le j \le n$. This implies that for any $M < \infty$,

$$\begin{aligned}
\mathbf{P}\{g^{-1}\{0\} \cap I_{n+1} \neq \emptyset\} &\le \sum_{t \in D_q^{n+1}} \mathbf{P}\{\|g(t)\|_\infty \le M2^{-q}\} + \mathbf{P}\{\|Dg\|_\infty > M\} \\
&\le 2^{qn} C_0 (2M2^{-q})^{n+1} + \mathbf{P}\{\|Dg\|_\infty \ge M\} \\
&\le C_0 2^{-q+n+1} + \mathbf{P}\{\|Dg\|_\infty \ge M\}.
\end{aligned}$$

As $q \to \infty$, the first term goes to zero. As $M \to \infty$, the second term goes to zero since $\|Dg\|_\infty$ is a finite random variable. Thus, $\mathbf{P}\{g^{-1}\{0\} \cap I_{n+1} \neq \emptyset\} = 0$. ■

As a corollary, we deduce that under mild (but not optimal) conditions, zero sets of random functions are smooth manifolds.

Corollary 2. *Let U be an open subset of $\mathbb{R}^n$ and let $f : U \mapsto \mathbb{R}^m$ be a random function. Assume that $m \le n$ and that $f \in C^1(U)$ a.s. and also that $(f(x), \nabla f(x))$ has a density $\rho(x)$ on $\mathbb{R}^{n+1}$ and that ρ is uniformly bounded on compact subsets of U.*

1. *If $m = n$, then $f^{-1}\{0\}$ is almost surely discrete.*

2. *If $m < n$, then $f^{-1}\{0\}$ is almost surely an $n-m$ dimensional manifold.*

Proof. Let $g_0 = f$ and $g_j = \partial_j f$ for $1 \le j \le n$. The assumptions of Bulinskaya's lemma are satisfied and hence there is almost surely no point where $f(x) = 0$ and $\nabla f(x) = 0$.

If $m = n$ and $f^{-1}\{0\} \ni x^{(j)} \to x \in U$, then $f(x) = 0$, then $\nabla f(x) = 0$, contradicting the conclusion above. Hence the zero set of f can have no accumulation points in U.

If $m < n$, then since $\nabla f(x) \neq 0$ for any $x \in f^{-1}\{0\}$ (almost surely), it follows from the implicit function theorem that all points of $f^{-1}\{0\}$ are regular and hence $f^{-1}\{0\}$ is a manifold of dimension $n-m$. ■

Remark 3. These conclusions can be drawn in greater generality. For example, in the book of Cramer and Leadbetter, they show this (or rather, some equivalent reformulations) even without assuming smoothness of f.

8.2 Kac-Rice formula in one dimension

Let $f : U \mapsto \mathbb{R}$ be a smooth random function on an open interval $U \subseteq \mathbb{R}$. By Corollary 2, the level set $Z_f := f^{-1}\{0\}$ is almost surely discrete. We wish to calculate $\mathbf{E}[|Z_f \cap I|]$ for $I \subseteq U$.

Lemma 4 (Kac-Rice)**.** *Let $f : U \mapsto \mathbb{R}$ be a random function that is almost surely C^1. Assume that $(f(x), f'(x))$ has a density $\rho(x;u,v)$ on $\mathbb{R}^2$ that is continuous in $(x;u,v)$ that is bounded over $(x;u,v) \in I \times \mathbb{R} \times \mathbb{R}$ for any compact $I \subseteq U$. Then,*

$$\mathbf{E}[|Z_f \cap I|] = \int_I \int_{\mathbb{R}} |v| \rho(x;0,v)\, dv\, dx.$$

Let X be a smooth centered Gaussian process on $\mathbb{R}^d$ with covariance K and spectral measure μ. The random closed set $Z := \{t \in \mathbb{R}^d : X(t) = 0\}$ is called the nodal set of X. One can also consider other level lines $Z_a := \{t \in \mathbb{R}^d : X(t) = a\}$ for $a \in \mathbb{R}$. Of interest is to study the statistical properties of Z (or Z_a).

Here is a quick discussion.

- In $d = 1$, under mild conditions on K, the random sets Z_a are discrete w.p.1. That is, almost surely, there are only finitely many zeros of X in any bounded interval, and in fact these zeros are simple.
- In $d = 2$, again under mild conditions on K, the random sets Z_a consist of simple closed loops and simple bi-infinite paths (the bi-infinite paths, if they exist, converge to infinity on both sides)
- In $d = 1$, the primary object of study are the random variables $Z(I) := \#(Z \cap I)$, the number of zeros in I. More generally one can consider the *linear statistics* $Z[\varphi] := \sum_{t \in Z} \varphi(t)$ for a compactly supported test function φ. When $\varphi = \mathbf{1}_I$ we recover the particle counts $Z(I)$.
- There exist formulas for the joint intensities of Z. That is, there exist function $p_k : \mathbb{R}^k \to \mathbb{R}$ such that the kth moment of $Z(I)$ can be expressed in terms of $\int_{I^r} p_r(\cdot)$ for $1 \le r \le k$. The formula for p_k is given by

  Kac-Rice formula: Under some conditions on the process X, we have

$$p_k(t_1, \ldots, t_k) = \mathbf{E}\left[\prod_{j=1}^{k} |X'(t_j)| \; ; \; X(t_1) = \ldots = X(t_k) = 0 \right].$$

- In $d = 2$, the counting function of zeros can be generalized in two different directions.

  1. $\ell_Z(A)$, the length of the nodal set intersected with a bounded set $A \subseteq \mathbb{R}^2$.
  2. $N_Z(A)$, the number of components of the nodal set contained entirely inside A.

  While ℓ_Z is a local quantity, N_Z is a not (and hence much harder to study). In particular, there exist formulas (again called Kac-Rice formulas) that give the joint intensities of Z in terms of the process X, namely

$$p_k(t_1,\ldots,t_k) := \mathbf{E}\left[\prod_{j=1}^{k} |\nabla X'(t_j)| \; ; \; X(t_1) = \ldots = X(t_k) = 0\right].$$

  And moments of $\ell_Z(A)$ can be expressed in terms of $\int_{A^r} p_r(t_1,\ldots,t_r)$ for $1 \leq r \leq k$. Although the formulas get complicated quickly, the mean is always easy, the variance is sometimes calculable and even asymptotic normality (as the region gets larger and larger) can be proved sometimes.

- No such formulas can be expected for $N_Z(A)$ as it is not a local quantity! Nevertheless the following result is known (if you include some closely related results, it is one of the main advances in recent times in the subject).

  Theorem 5 (Malevich, Nazarov-Sodin)**.** *Let $Q_T = [-T,T]^2$. Under some mild conditions on the process X, we have $\mathbf{E}[N_Z(Q_T)] \sim CT^2$ where C is a constant that depends on the covariance kernel only. Further, $T^{-2}N_Z(Q_T) \xrightarrow{P} C$ as $T \to \infty$.*

  In contrast to ℓ_Z, it is an entirely open problem to say anything about $\mathrm{Var}(N_Z(Q_T))$, even in special Gaussian processes (eg., the random plane wave)!

Chapter 9

Hypercontractivity - lecture by Nanda Kishore Reddy, Kartick Adhikari and Tulasi Ram Reddy

For a Gaussian Hilbert space $H \subseteq L^2(\Omega, \mathcal{F}, \mathbf{P})$ we define $H^{:n:} = \mathcal{P}_n \cap \mathcal{P}_{n-1}^{\perp}$ where $\mathcal{P}_n$ is the closed span of the products $\xi_1 \dots \xi_n$ with $\xi_i \in H$.

Theorem 1 (Wiener's chaos decomposition)**.** *Let $\mathcal{F}_H$ be the sigma-algebra generated by H. Then $L^2(\Omega, \mathcal{F}_H, \mathbf{P}) = H^{:0:} \oplus H^{:1:} \oplus H^{:2:} \oplus \dots$ (direct sum, by definition includes closure).*

Here is the prime example.

Example 2. Let $H = \mathbb{R}\xi$ be a one dimensional GHS where $\xi \sim N(0,1)$. Then $H^{:n:} = \mathbb{R}h_n(\xi)$ where h_n is the nth (monic) Hermite polynomial. $L^2(\Omega, \mathcal{F}_H, \mathbf{P})$ is naturally isomorphic to $L^2(\mathbb{R}, \gamma_1)$ (under that isomorphism ξ maps to the identity function). The theorem above is just the assertion that $h_n(\cdot)/\sqrt{n!}$ is an ONB for $L^2(\mathbb{R}, \gamma_1)$.

Let $\pi_n : L^2(\Omega, \mathcal{F}_H, \mathbf{P}) \to H^{:n:}$ be the orthogonal projection. Let $A : H \to H$ be a linear operator. We define $A : H^{:n:} \to H^{:n:}$ by setting

$$A^{:n:}(\pi_n[\xi_1 \dots \xi_n]) = \pi_n[(A\xi_1) \dots (A\xi_n)].$$

For this definition to make sense, prove the following exercise.

Exercise 3. If $\xi_i, \eta_i \in H$ and $\pi_n[\xi_1 \dots \xi_n] = \pi_n[\eta_1 \dots \eta_n]$ then show that $(A\xi_1) \dots (A\xi_n) - (A\eta_1) \dots (A\eta_n) \in \mathcal{P}_{n-1}$. Conclude that $A^{:n:}$ is well defined and $\|A^{:n:}\| \leq \|A\|^n$.

Definition 4. If $A : H \to H$ and $\|A\| \leq 1$, then define the operator $\Gamma[A]$ from $L^2(\Omega, \mathcal{F}_H, \mathbf{P})$ to itself by $\Gamma[A](X) = \sum_{n \geq 0} A^{:n:}(\pi_n[X])$.

Exercise 5. Check that $\Gamma[A]$ does map $L^2(\mathbf{P})$ into itself and $\|\Gamma[A]\| \leq 1$.

Example 6. If $A = rI$ with $r \leq 1$ then $\Gamma[rI](X) = \sum_n r^n \pi_n[X]$.

Observe that H and hence (by Hölder's inequality) $\mathcal{P}_n$ and $H^{:n:}$ are contained in $L^p(\Omega, \mathbf{P})$ for all $p \geq 1$. In H all variables are centered Gaussian, hence $\|\xi\|_p \leq c_p \|\xi\|_2^{p/2}$ for a universal constant c_p. How the different norms compare on $H^{:n:}$ is less clear. The following results are asserted.

Theorem 7. *Let $A \leq 1$ and let $\Gamma[A]$ be defined as above,*

1. *If $X \in \bigcap_p L^p$ then $\Gamma[A](X) \in \bigcap_p L^p$ and further $\|\Gamma[A]X\|_p \leq \|X\|_p$.*
2. *By the density of $\bigcap L^p$ in any L^p, it follows that $\Gamma[A]$ extends as a contraction to all of L^p.*
3. *$\Gamma[AB] = \Gamma[A]\Gamma[B]$.*

Theorem 8 (The Hypercontractivity theorem)**.** *If $\|A\| \leq r \leq 1$, then $\|\Gamma[A]X\|_q \leq \|X\|_p$ for every $1 \leq p \leq q \leq (p-1)r^{-2} + 1$.*

Example 9. Consider the basic case $H = \mathbb{R}\xi$ with $A = rI$. The hypercontractivity theorem says in this case that

$$\|\sum_n a_n r^n h_n\|_{L^q(\gamma_1)} \leq \|\sum_n a_n h_n\|_{L^p(\gamma_1)}$$

whenever $r \leq \sqrt{\frac{q-1}{p-1}}$ and any coefficients a_n.