Printed: 5 March 2025 at 18:54

Chapter 6

Multivariate normal

Gaussian::Gaussian

SECTION 6.1 *describes some wondrous properties of the (multivariate) normal distribution. By way of preview, it presents two examples of what can be achieved by using the masterly trick of studying relationships between different normals by integrating along a path.*

SECTION 6.2 *describes the path method. It also draws your attention to a simple integration-by-parts trick used by Charles Stein to prove amazing facts about normal approximation.*

SECTION 6.3 *presents three different path methods for slightly different versions of the* LIPSCHITZ *concentration inequality.*

*SECTION 6.4 *describes the gaussian isoperimetric inequality.*

SECTION 6.5 *derives some comparison inequalities (initially due to Slepian, with a generalization due to Gordon) by means of the path method.*

*SECTION 6.6 *uses the* GORDON *inequality to derive a special case of a result by Dvoretzky regarding cross sections of convex bodies in high dimensional Euclidean spaces.*

SECTION 6.7 *presents Chatterjee's proof of Fernique's inequality.*

6.1 Introduction

Gaussian::S:intro

The normal distribution has many amazing properties. This Chapter presents a few results that give a glimpse of the heights achieved by the those who have worked on the theory of gaussian processes in the last half century or so. Even for those readers who are primarily interested in nongaussian processes, it helps to see what is possible in the cleanest case before plunging into more general theory; often the gaussian results provide the benchmarks against which the nongaussian analogs are judged. In any case, the gaussian results provide an example of beautiful probability theory, which is worthy of attention for aesthetic reasons.

A little bit of notation is needed before I can describe the main focus of the Chapter. As in Section 3.3, the symbol Φ denotes the $N(0,1)$ distribution function with $\bar{\Phi} = 1 - \Phi$, and ϕ denotes the standard normal density. The symbol γ_n denotes the $N(0, I_n)$, the probability measure on $\mathcal{B}(\mathbb{R}^n)$ with

density $\phi_n(x) = \prod_j \phi(x_j) = (2\pi)^{-n/2} \exp(-|x|^2/2)$ with respect to Lebesgue measure on $\mathbb{R}^n$. More generally, if $\mu \in \mathbb{R}^n$ and V is an $n \times n$ positive semi-definite matrix then $N(\mu, V)$ denotes the probability measure on $\mathcal{B}(\mathbb{R}^n)$ with Fourier transform $\exp(it'\mu - t'Vt/2)$. Such a measure is often called a ***multivariate normal*** (MVN for short) distribution. It has mean μ and variance matrix V. If μ is zero the distribution is said to be a ***centered*** MVN. Only if V is nonsingular does the $N(\mu, V)$ have a density with respect to Lebesgue measure. Sometimes it helps to add a subscript, $N_n(\dots)$, if there is any ambiguity about the dimension of the Euclidean space where the measure lives. For example, if $X \sim N_n(\mu, V)$ and F is an $m \times n$ constant matrix then FX has a $N_m(F\mu, FVF')$ distribution. As a special case for m equal to n, if $FF' = V$ and $Z \sim N_n(0, I_n)$ then $\mu + FZ \sim N_n(\mu, V)$. Finally, the general fact that independence implies zero correlation can be run in the other direction for the MVN: if

$$\begin{pmatrix} X \\ Y \end{pmatrix} \sim N(\mu, V) \qquad \text{with } V = \begin{pmatrix} V_x & C_{x,y} \\ C'_{x,y} & V_y \end{pmatrix},$$

then X and Y are independent if and only if $C_{x,y} = 0$.

All the assertions in the previous paragraph are easy to prove using Fourier transforms.

Now back to the main focus of the Chapter. The unifying concept is that of a ***path argument*** (described in Section 6.2) to establish two very useful properties of the MVN:

(i) subgaussian concentration of $f(Z)$ if $Z \sim N(0, I_n)$ and $f : \mathbb{R}^n \to \mathbb{R}$ is LIPSCHITZ (in Section 6.3);

(ii) comparison inequalities relating probabilistic bounds for two centered MVN distributions based on simple comparisons between their variance matrices (in Sections 6.5 and 6.7).

Here is a preview.

6.1.1 Subgaussian concentration

Gaussian::Lip.subg

Remember that a function $f : \mathbb{R}^n \to \mathbb{R}$ is said to be LIPSCHITZ if there exists a finite constant κ for which

$$|f(y) - f(z)| \le \kappa |y - z| \qquad \text{for all } y, z \text{ in } \mathbb{R}^n.$$

Here $|\cdot|$ denotes the usual Euclidean (ℓ^2) distance, $|w|^2 := \sum_i w_i^2$. The smallest κ for which the inequality holds is called the LIPSCHITZ constant, which I denote by $\|f\|_{Lip}$.

The following concentration result is proved in Section 6.3.3. A much simpler argument, which produces a slightly larger subgaussian scale parameter, appears in in Section 6.3.1. Of course I suggest that you start with the easier derivation.

Gaussian::Lipschitz.subg <1> **Theorem.** *If $Z \sim \gamma_n = N(0, I_n)$ and $f : \mathbb{R}^n \to \mathbb{R}$ is a* LIPSCHITZ *function with $\|f\|_{Lip} \le \kappa$ then $f(Z) - \gamma_n f$ has a subgaussian distribution,*

$$\mathbb{P}e^{\lambda(f(Z)-\gamma_n f)} \le e^{\lambda^2\kappa^2/2} \qquad \text{for each } \lambda \in \mathbb{R},$$

□ *which implies $\mathbb{P}\{|f(Z) - \gamma_n f| \ge t\kappa\} \le 2e^{-t^2/2}$ for each $t \ge 0$.*

Remark. Notice that n does not appear explicitly in the upper bound, although it might be hiding inside the constant κ.

The bound is sharp, in the following sense: If u is a unit vector then the linear function $f(x) = u'x$ is LIPSCHITZ with $\|f\|_{Lip} = 1$ and the function $f(Z)$ has a $N(0,1)$ distribution if $Z \sim \gamma_n$.

Gaussian::Borell.subg <2> **Example.** Let M equal $\sup_i Y_i$ for a gaussian process $\{Y_i : i \in \mathbb{N}\}$. If both $m := \mathbb{P}M$ and $\sigma^2 := \sup_{i\in\mathbb{N}} \operatorname{var}(Y_i)$ are finite then Theorem <1> can be used to show that

\E@ max.conc <3> $$\mathbb{P}\{|M - m| \ge \sigma x\} \le 2\exp(-x^2/2) \qquad \text{for all } x \ge 0.$$

In special cases—such as the maximum of independent $N(0,1)$-distributed variables, as discussed in Section 3.3—there exist much tighter bounds. However, inequality <3> has the great virtues of both being impervious to the effects of possible dependence between the Y_i and of not depending (except through m and σ) on the size of the index set.

All the real work is carried out for the case of a finite index set. Suppose $W := (Y_1, \ldots, Y_n) \sim N_n(\mu, V)$. We may write W as $\mu + LZ$ with $Z \sim \gamma_n$ and an $n \times n$ constant matrix L for which $LL' = V$. If ℓ_i denotes the ith row of L then $Y_i = \mu_i + \ell_i Z$ and $\sigma^2 \ge \operatorname{var}(Y_i) = \operatorname{var}(\mu_i + \ell_i Z) = \ell_i I_n \ell_i' = |\ell_i|^2$.

Define a real-valued function $f_n(z) := \max_{i\le n} (\mu_i + \ell_i z)$ on $\mathbb{R}^n$. It is LIPSCHITZ with $\|f_n\|_{Lip} \le \sigma$: for $z_1, z_2 \in \mathbb{R}^n$,

$$\begin{aligned} |f_n(z_1) - f_n(z_2)| &= |\max_{i\le n}(\mu_i + \ell_i z_1) - \max_{i\le n}(\mu_i + \ell_i z_2)| \\ &\le \max_i |(\mu_i + \ell_i z_1) - (\mu_i + \ell_i z_2)| \\ &\le \max_i |\ell_i|\,|z_1 - z_2| \qquad \text{by CAUCHY-SCHWARZ} \\ &\le \sigma|z_1 - z_2|. \end{aligned}$$

Define $M_n := \max_{i\le n} Y_i = \max_{i\le n}(\mu_i + \ell_i Z) = f_n(Z)$. Theorem <1> gives $\mathbb{P}\exp(\lambda(M_n - \mathbb{P}M_n)) \le \exp(\lambda^2\sigma^2/2)$ for all real λ, which implies $\mathbb{P}\{M_n \ge \mathbb{P}M_n + \sigma x\} \le e^{-x^2/2}$. Put another way,

$$\mathbb{P}\{M_n > m + \sigma(x - \epsilon)\} \le e^{-(x-\epsilon)^2/2} \qquad \text{for } 0 < \epsilon < x \text{ and each } n.$$

As n tends to ∞ the events $\{M_n > r\}$ increase to $\{M > r\}$ for each r. Consequently, $\mathbb{P}\{M \ge m + \sigma x\} \le \mathbb{P}\{M > m + \sigma(x - \epsilon)\} \le e^{-(x-\epsilon)^2/2}$ for $0 < \epsilon < x$. Let ϵ decrease to zero to obtain a one-sided analog of the asserted inequality. A similar argument applied to $-f_n$ followed by a passage to the limit leads to a similar bound for the lower tail of $M - \mathbb{P}M$. □

Remark. You might be puzzled by the sneaky use of a strict inequality near the end of the argument. The trick is made necessary by the annoying fact that if $M(\omega) > r$ then $M_n(\omega) > r$ for all n large enough, but if $M(\omega) \ge r$ then we might have $M_n(\omega) < r$ for all n.

6.1.2 Comparison inequalities

Gaussian::comparison

The following inequality is a special case of a result proved in Section 6.7.

Gaussian::Fernique0 <4> **Theorem.** *(Fernique, 1975, Section 2.1) Suppose both $X = (X_1, \ldots, X_n)$ and $Y = (Y_1, \ldots, Y_n)$ have centered* MVN *distributions. If $\mathbb{P}|X_i - X_j|^2 \le \mathbb{P}|Y_i - Y_j|^2$ for all i, j then $\mathbb{P}\max_i X_i \le \mathbb{P}\max_i Y_i$.* □

Gaussian::Sudakov <5> **Example.** (SUDAKOV "minoration") If $Y := (Y_1, Y_2, \ldots, Y_n)$ has a centered MVN distribution with $\mathbb{P}|Y_j - Y_k|^2 \ge \delta^2$ for all distinct $j \ne$ and k, then $\mathbb{P}\max_{i \le n} Y_i \ge C_{\text{sud}}\delta\sqrt{\log_2 n}$ with C_{sud} a universal (positive) constant.

I leave to you the trival case where $n = 1$.

The asserted inequality follows directly from Theorem <4>. Consider first the case where $n = 2^k$, for some $k \in \mathbb{N}$, so that the index set can be identified with $\mathcal{S} := \{-1, +1\}^k$. Construct $\{X_\mathfrak{s} : \mathfrak{s} \in \mathcal{S}\}$ from a set $Z_1, \ldots, Z_k$ of independendent $N(0,1)$'s:

$$X_\mathfrak{s} := \tfrac{1}{2}\delta k^{-1/2} \sum_{j=1}^k \mathfrak{s}_j Z_j.$$

For $\mathfrak{s} \ne \mathfrak{s}'$,

$$\mathbb{P}|X_\mathfrak{s} - X_{\mathfrak{s}'}|^2 = \tfrac{1}{4}\delta^2 k^{-1} \sum_{j=1}^k (\mathfrak{s}_j - \mathfrak{s}'_j)^2 \le \delta^2 \le \mathbb{P}|Y_\mathfrak{s} - Y_{\mathfrak{s}'}|^2.$$

The X process is more tractable than Y. Its expected maximum can be calculated exactly.

$$\begin{aligned}\mathbb{P}\max_\mathfrak{s} X_\mathfrak{s} &= \tfrac{1}{2}\delta k^{-1/2}\mathbb{P}\left(\max_\mathfrak{s} \sum_{j=1}^k \mathfrak{s}_j Z_j\right) \\ &= \tfrac{1}{2}\delta k^{-1/2}\mathbb{P}\sum_{j=1}^k |Z_j| = \tfrac{1}{2}\delta k^{1/2}\mathbb{P}|Z_1| = \tfrac{1}{2}\delta k^{1/2}\sqrt{2/\pi}.\end{aligned}$$

For a general $n \ge 2$ define $k = \lfloor \log_2 n \rfloor$

$$\mathbb{P}\max_{i \le n} Y_i \ge \mathbb{P}\max_{i \le 2^k} Y_i \ge \tfrac{1}{2}\delta k^{1/2}\sqrt{2/\pi}.$$

The choice $C_{\text{sud}} = (4\pi)^{-1/2}$ suffices. □

Remark. The lower bound for $\mathbb{P}\max_i Y_i$ is sharp within a constant, in the following sense. Suppose $\mathbb{P}|Y_j - Y_k|^2 \le (A\delta)^2$ for all j, k, with A a positive constant. First note that

$$\mathbb{P}\max_i Y_i = \mathbb{P}Y_1 + \mathbb{P}\max_i(Y_i - Y_1) = \mathbb{P}\max_i(Y_i - Y_1).$$

Then argue as in Section 3.3. For each $\lambda > 0$,

$$\begin{aligned}\exp\left(\lambda\mathbb{P}\max_i Y_i\right) &= \mathbb{P}\exp\left(\lambda\max_i(Y_i - Y_1)\right) \qquad \text{by Jensen's inequality} \\ &\le \mathbb{P}\sum_i \exp\left(\lambda(Y_i - Y_1)\right) \le n e^{\lambda^2 A^2 \delta^2/2}.\end{aligned}$$

Take logs of both sides, divide through by λ, then choose $\lambda = \sqrt{2\log n}/(A\delta)$ to minimize, leaving $\mathbb{P}\max_i Y_i \le A\delta\sqrt{2\log n}$.

6.2 The path method

Gaussian::S:paths

The path method usually involves the construction of a smooth multivariate gaussian process $\{W_t : 0 \le t \le 1\}$ with W_0 and W_1 suggested by the particular problem at hand. For example, to attack Theorem <4> the random vector W_0 could be chosen to have the same distribution as X and the random vector W_1 could be chosen to have the same distribution as Y. For f a suitably smooth function on $\mathbb{R}^n$ we have a representation

\E@ path.rep <6>

$$\begin{aligned} f(W_1) - f(W_0) &= \int_0^1 df(W_t)/dt\, dt \\ &= \int_0^1 \langle \dot{f}(W_t), \dot{W}_t \rangle\, dt = \sum\nolimits_{i \le n} \int_0^1 \dot{f}_i(W_t) W_{t,i}\, dt, \end{aligned}$$

where $\dot{W}_t$ denotes the derivative of W_t with respect to t, with ith component $\dot{W}_{t,i}$, and $\dot{f} = \nabla f$ denotes the derivative of the map $x \mapsto f(x)$, with ith component $\dot{f}_i(x) := \partial f(x_1, \ldots, x_n)/\partial x_i$. To learn something about $\mathbb{P}f(W_1) - \mathbb{P}f(W_0)$ we need to control $\mathbb{P}\dot{f}_i(W_t)\dot{W}_{t,i}$ at each point of the path.

Remark. As I have been using the prime symbol $'$ to denote transpose of a vector or matrix I need something different to indicate derivatives. For a real-valued function $D = D(x_1, \ldots, x_n)$ on $\mathbb{R}^n$ the derivative will be denoted by $\dot{D}$ (or sometimes ∇D), with components $\dot{D}_i = \partial D(x)/\partial x_i$. If D also depends on another real argument, t, then $\dot{D}_t$ will denote $\partial D(x,t)/\partial t$. Similarly $\ddot{D}_{i,j}$ will denote the partial derivative $\partial^2 D(x)/\partial x_i \partial x_j$. The function $\sum_i \ddot{D}_{i,i}$ is called the Laplacian. It is often denoted by ΔD, a notation that I need to avoid because I want to reserve Δ to denote an increment.

In general, the choice of a good path is a rather mysterious business, although the examples I know all seem to share a few tricks. Talagrand (2003, Section 1.3) suggested that the choice of starting point of the path is the real subtlety:

> To study a difficult situation one can compare it to a simpler one, by finding a path between them and controlling derivatives along this path. This is an old idea. In practice we are given the difficult situation, and the key to the effectiveness of the method is to find the correct simple situation to which it should be compared. This can be done only after the problem is well understood. To insist upon the fact that the choice of the path is the real issue, we call this method the smart path method. (More precisely, the real issue is in the choice of the "easy end of the path". Once this has been chosen, the choice of the path itself will be rather canonical, except for its "orientation". We make the convention that the "smart path" moves from the "easy end" to the "hard end") The smart path method, under various forms, will be the main tool throughout the book.

Often the path is constructed in a rather rigid way, starting from independent random vectors $W_0 \sim N(0, V_0)$ and $W_1 \sim N(0, V_1)$. For smooth deterministic functions a_t and b_t with $a_0 = 1 = b_1$ and $a_1 = b_0 = 0$ we define $W_t = a_t W_0 + b_t W_1$. That gives $\dot{W}_t = \dot{a}_t W_0 + \dot{b}_t W_1$, which has a $N\left(0, (\dot{a}_t)^2 V_0 + (\dot{b}_t)^2 V_1\right)$ distribution. More importantly, the joint distribution of W_t and $\dot{W}_t$ is also MVN, with

\E@ covWdotW <7>
$$\operatorname{cov}(\dot{W}_t, W_t) = \dot{a}_t a_t V_0 + \dot{b}_t b_t V_1.$$

It often seems to help if we choose a_t and b_t so that $a_t^2 + b_t^2 = 1$ for each t, a constraint that implies

\E@ dot.unitvector <8>
$$0 = \partial(a_t^2 + b_t^2)/\partial t = 2\left(\dot{a}_t a_t + \dot{b}_t b_t\right)$$

so that

\E@ WdotW.cov <9>
$$\operatorname{cov}(\dot{W}_t, W_t) = \dot{b}_t b_t \left(V_1 - V_0\right).$$

This simplification is most convenient when dealing with theorems that impose assumptions on the difference $V_1 - V_0$. See Theorem <33> for an example.

Strangely enough, the argument in Section 6.3.1 involves a path construction with $V_1 = V_0$, so that $\operatorname{cov}(\dot{W}_t, W_t) = 0$, implying that $\dot{W}_t$ and W_t are independent. In that case, the random variables $g(W_0)$ and $g(W_1)$ have the same distribution, so it would be pointless to take expected values in <6>. Instead, as you'll see, a little JENSEN trick gives a bound on the expected value of $\exp\left(\lambda(g(W_1) - g(W_0)\right)$.

If $V_1 \neq V_0$ the expected value $\mathbb{P}\dot{g}_i(W_t)\dot{W}_{t,i}$ can still be simplified by using an integration-by-parts trick that was often used by Charles Stein to perform miracles. See the Notes.

Gaussian::i-b-p <10> **Theorem.** *[i-b-p trick] Suppose $(Z, X_1, \ldots, X_m)$ has a* MVN *distribution with $\mathbb{P}Z = 0$. (No assumptions are made about the means or covariances for the X_i's.) Suppose also that a continuously differentiable function $G : \mathbb{R}^m \to \mathbb{R}$ has partial derivatives $\dot{G}_j(x_1, \ldots, x_m) := \partial G(x_1, \ldots, x_m)/\partial x_j$ for which $\mathbb{P}|\dot{G}_j(X_1, \ldots, X_m)| < \infty$ for each i. Then*

$$\mathbb{P}ZG(X_1, \ldots, X_m) = \sum\nolimits_{j \le m} \tau_j \mathbb{P}\dot{G}_j(X_1, \ldots, X_m) \qquad \text{where } \tau_j := \mathbb{P}(ZX_j). \qquad \square$$

For the proof see Problems [2] (the one dimensional case) and [3], which involves many appeals to the one-dimensional case.

For future reference, here is a result that summarizes these ideas. Rather than imposing explicit assumptions on the f in <6>, I'll stick with the vague term 'suitably regular' to suggest any requirement that would ensure integrability of various random variables and justify taking derivatives insides integrals with respect to $\mathbb{P}$ and integrations-by-parts. Twice continuous differentiability with second partial derivatives $\ddot{f}_{i,j}(x)$ that grow no faster than $\exp(C|x|)$ would suffice in the next Lemma.

Gaussian::general.path <11>

Lemma. *Suppose $W_t = a_t W_0 + b_t W_1$ for $0 \le t \le 1$, where W_0 and W_1 are independent random vectors with $W_\alpha \sim N_n(0, V_\alpha)$ and a_t, b_t are chosen so that $a_t^2 + b_t^2 = 1$ and $0 = b_0 < b_t \uparrow 1$ as $t \uparrow 1$. If $G(t) := \mathbb{P}g(W_t)$ for a suitably regular real-valued function g on $\mathbb{R}^n$ then*

$$\dot{G}(t) = \sum_{i \le n} \mathbb{P}\dot{W}_{t,i} \dot{g}_i(W_t) = \dot{b}_t b_t \sum_{i,j \in [\![n]\!]} (V_1[i,j] - V_0[i,j])\, \mathbb{P}\ddot{g}_{i,j}(W_t).$$

□

Remark. The assumption on b_t ensures that $\dot{b}_t b_t = d(b_t^2/2)/dt > 0$, a convenience if we want to show H is an increasing function of t.

The choice $a_t = \sqrt{1-t}$ and $b_t = \sqrt{t}$, or its reparametrization $a_t = \cos(\pi t/2)$ and $b_t = \sin(\pi t/2)$, works well for all of the applications in this Chapter. We could also change the index set to some other subinterval, such as $a_t = \cos(t)$ and $b_t = \sin(t)$ with $0 \le t \le \pi/2$.

6.3 Concentration of LIPSCHITZ functionals

Gaussian::S:Lipschitz

This Section contains three different path arguments, presented in order of increasing subtlety, that show $f(Z) - \gamma_n f$ has a subgaussian distribution if $Z \sim \gamma_n = N(0, I_n)$ and f is LIPSCHITZ. The easiest argument (in Section 6.3.1) yields a subgaussian bound with scale parameter $\pi \|f\|_{Lip}/2$. Section 6.3.2 improves the scale parameter to $\sqrt{2}\|f\|_{Lip}$ by using a slightly better path. Finally, Section 6.3.3 uses a path constructed from a BROWNIAN motion to obtain the scale factor $\|f\|_{Lip}$. A completely rigorous argument in the third case uses the ITÔ formula from stochastic calculus. On the assumption that my readers might not be completely familiar with that formula, I also include a heuristic argument that gives some insight into what is really going on.

The three different proofs helped me when I was first trying to understand what made one path smarter than another.

An analog of Theorem <1>, with centering at the median rather than at the expected value, can also be derived easily from a very deep result known as the ***gaussian isoperimetric inequality***, which is discussed briefly in Section 6.4.

Remark. Unfortunately, most proofs that I know of for that inequality are quite involved. It is worth knowing at least a little about isoperimetry because it lurks behind several interesting facts about the normal distribution. It would, however, be discouraging if the only path to Theorem <1> ran through isoperimetry.

Roughly speaking, when proving mgf results about a generic LIPSCHITZ function f there is no harm in pretending that it is infinitely differentiable, with derivative $\dot{f}(x)$ (sometimes denoted by ∇f) whose euclidean length is everywhere bounded above by $\|f\|_{Lip}$. Such a simplification is justified by Problem [1], which shows how to use convolution smoothing to construct a

family $\{f_\sigma : \sigma > 0\}$ of infinitely differentiable functions with $\|f_\sigma\|_{Lip} \le \|f\|_{Lip}$ and $\sup_x |f_\sigma(x) - f(x)| \to 0$ as $\sigma \to \infty$ and

\E@ Lip.grad <12>
$$|\dot{f}_\sigma(x)|_2 = \sqrt{\sum\nolimits_i (\partial f_\sigma(x)/\partial x_i)^2} \le \|f\|_{Lip} \qquad \text{for all } x.$$

Moreover, $\mathbb{P}\exp(\lambda f_\sigma(W)) \to \mathbb{P}\exp(\lambda f(W))$ as $\sigma \to 0$ for each real λ if $\mathbb{P}\{|W| \ge t\}$ decreases fast enough. Subgaussianity suffices.

6.3.1 The PISIER-MAUREY method

Gaussian::Lip-PisierMaurey

The method in this subsection comes from Theorem 2.2 of Pisier (1986, page 176), who commented that "The proof below is a simplification, due to Maurey, of my original proof which used an expansion in Hermite polynomials". Pisier was primarily interested in a concentration bound for the norm of a BANACH-valued gaussian process X. He also noted that his method would "apply to more general functions of X than the norm of X (and even vector valued functions) provided a suitable bound is known for the gradients of the functions".

The proof starts with a symmetrization trick. (Compare with Chapter 13.) For notational reasons, which should soon be apparent, let me rewrite the mgf for $f(Z) - \mathbb{P}f(Z)$ as an iterated integral, then invoke the JENSEN inequality for the integral appearing within the exponent:

\E@ symm.MGF <13>
$$\begin{aligned}\mathbb{P}e^{\lambda(f(X)-\mathbb{P}f(X))} &= \gamma_n^x \exp\left(\lambda(f(x) - \gamma_n^y f(y))\right) \\ &\le \gamma_n^x \gamma_n^y \exp\left(\lambda(f(x) - f(y))\right) = \mathbb{P}e^{\lambda(f(X)-f(Y))},\end{aligned}$$

where X and Y are independent random vectors that both have distribution γ_n.

Remark. The argument could also be written as a conditional JENSEN inequality using the fact that $\mathbb{P}(f(Y) \mid X) = \gamma_n f$. I prefer FUBINI to conditioning when I am trying to deal carefully with symmetrization arguments involving independence.

Notice that $\operatorname{var}(f(X) - f(Y)) = 2\operatorname{var}(f(X))$. This symmetrization approach inevitably leads to at least a doubling of the squared scaling factor.

The distribution of $|X - Y|$ depends on n. If we were to bound the difference $f(X) - f(Y)$ in the exponent by $\kappa|X - Y|$ we would introduce an explicit dependence on n in the upper bound. Instead we need to exploit cancellations due to independence along a one-dimensional path from $W_0 = Y$ to $W_1 = X$. For $0 \le t \le 1$ define $a_t = \cos(t\pi/2)$ and $b_t = \sin(t\pi/2)$ and $W_t = a_t W_0 + b_t W_1$. The derivative with respect to t along the path equals

$$\dot{W}_t = \frac{\partial W_t}{\partial t} = \dot{a}_t W_0 + \dot{b}_t W_1 = \frac{\pi}{2}\left(-b_t W_0 + a_t W_1\right) \sim N(0, (\pi/2)^2 I_n).$$

As explained in Sections 6.2, this choice ensures that $\operatorname{cov}(\dot{W}_t, W_t) = 0$, making the random vectors $(2/\pi)\dot{W}_t$ and W_t independent, each with distribution γ_n.

Moreover

$$f(X) - f(Y) = f(W_1) - f(W_0) = \int_0^1 \frac{\partial f(W_t)}{\partial t}\, dt = \int_0^1 \langle \dot{W}_t, \dot{f}(W_t)\rangle\, dt.$$

By the JENSEN inequality for LEBESGUE measure on $[0,1]$,

$$\mathbb{P}\exp\left(\lambda\left(f(X) - f(Y)\right)\right) \le \mathbb{P}\int_0^1 \exp\left(\lambda\langle \dot{W}_t, \dot{f}(W_t)\rangle\right)\, dt.$$

For each fixed s in $\mathbb{R}^n$ we have $\mathbb{P}\exp(\langle \dot{W}_t, s\rangle) = \exp(\pi^2|s|_2^2/8)$. Independence of $\dot{W}_t$ and $\dot{f}(W_t)$ lets us treat $\lambda\dot{f}(W_t)$ like a constant s (more conditioning? or Fubini?) to deduce that

$$\begin{aligned}\mathbb{P}\exp\left(\lambda\left(f(X) - f(Y)\right)\right) &\le \int_0^1 \mathbb{P}\exp\left(\lambda^2\pi^2|\dot{f}(W_t)|_2^2/8\right)\, dt \\ &\le \exp\left(\lambda^2\kappa^2\pi^2/8\right) \qquad \text{by } <12>.\end{aligned}$$

We have a subgaussian bound with scale parameter $\pi\kappa/2$. Derive the tail bound as in Section 3.3.

Remark. The argument does not work with the choice $a_t = \sqrt{1-t}$ and $b_t = \sqrt{t}$, because then $\dot{W}_t \sim N(0, \sigma_t^2 I_n)$ with $1/\sigma_t^2 = 4t(1-t)$. The average over the $\dot{W}_t$ distribution then leaves a σ_t^2 in the exponent. The trigonometric parametrization, which gives a constant variance matrix for $\dot{W}_t$, seems better in this case.

6.3.2 The smart path method

Gaussian::Lip-Talagrand

This refinement of the path method from Section 6.3.1 comes from Talagrand (2011, Section 1.3). The method this time is similar to the one in 6.3.1 , except that now the path argument is carried out directly using the function

$$g(w) := g(x,y) := \exp(\lambda f(x) - \lambda f(y)) \qquad \text{for } w = (x,y) \in \mathbb{R}^{2n}.$$

Notice that

$$\begin{aligned}\dot{g}(w) &= \left(\lambda\dot{f}(x)g(w), -\lambda\dot{f}(y)g(w)\right) \\ &= \left(\dot{f}_1(x), \ldots, \dot{f}_n(x), -\dot{f}_1(y), \ldots, -\dot{f}_n(y)\right)\lambda g(w).\end{aligned}$$

To avoid some notational confusion, I'll use Greek letters (α, β) when referring to elements of $[\![2n]\!] = \{1, 2, \ldots, 2n\}$ and Roman letters $(i, j, \ldots)$ when referring to elements of $[\![n]\!] = \{1, 2, \ldots, n\}$.

The smarter path is defined using three independent random vectors X, Y an Z, each distributed $N(0, I_n)$, and $a_t = \sqrt{1-t}$ and $b_t = \sqrt{t}$:

$$\begin{aligned}W_t &= (X_t, Y_t) \qquad \text{for } 0 \le t \le 1, \\ X_t &= (X_{t,1}, \ldots, X_{t,n}) = a_t Z + b_t X, \\ Y_t &= (Y_{t,1}, \ldots, Y_{t,n}) = a_t Z + b_t Y.\end{aligned}$$

Remark. You might find it informative to check that the argument still works with $a_t = \cos(t)$ and $b_t = \sin(t)$ for $0 \le t \le \pi/2$.

For some calculations it helps to think of the $\{X_t\}$ and $\{Y_t\}$ as two n-dimensional gaussian processes, each with the same distribution as the $\{W_t\}$ process from Section 6.3.1. In particular,

$$\operatorname{cov}(\dot{X}_t, X_t) = \left(\dot{a}_t a_t + \dot{b}_t b_t\right) I_n = 0 = \operatorname{cov}(\dot{Y}_t, Y_t).$$

The processes inherit a dependence from the Z that they share. In particular, $W_0 = (Z, Z)$ and $W_1 = (X, Y)$, so that

\E@ Tal-var-diff <14>

$$\Delta V := V_1 - V_0 := \operatorname{var}(W_1) - \operatorname{var}(W_0) = I_{2n} - \begin{pmatrix} I_n & I_n \\ I_n & I_n \end{pmatrix} = \begin{pmatrix} 0 & -I_n \\ -I_n & 0 \end{pmatrix}.$$

Consequently

$$\Delta V[\alpha, \beta] = \begin{cases} -1 & \text{if } \alpha = \beta - n = i \in [\![n]\!] \text{ or } \alpha - n = \beta = i \in [\![n]\!] \\ 0 & \text{otherwise} \end{cases}.$$

The sparsity of ΔV greatly simplifies the application of Lemma <11> to the function $G(t) := \mathbb{P} g(W_t)$:

$$\begin{aligned} \dot{G}(t) &= \dot{b}_t b_t \sum\nolimits_{\alpha,\beta \in [\![2n]\!]} \Delta V[\alpha, \beta] \mathbb{P} \ddot{g}_{\alpha,\beta}(W_t) \\ &= \tfrac{1}{2} \sum\nolimits_{i \in [\![n]\!]} \mathbb{P} 2\lambda^2 \dot{f}_i(X_t) \dot{f}_i(Y_t) g(W_t) \\ &\le \lambda^2 \mathbb{P} |\dot{f}(X_t)| \, |\dot{f}(Y_t)| g(W_t) \qquad \text{by CAUCHY-SCHWARZ} \\ &\le \lambda^2 \kappa^2 G(t) \qquad \text{by inequality <12>.} \end{aligned}$$

Remark. It helped me to write the the αth component of $\dot{g}$ as

$$\dot{g}_\alpha(w) = \begin{cases} \lambda \dot{f}_i(x) g(w) & \text{if } \alpha = i \in [\![n]\!] \\ -\lambda \dot{f}_i(y) g(w) & \text{if } \alpha - n = i \in [\![n]\!] \end{cases}.$$

The partial derivative with respect to w_β, with $|\alpha - \beta| = n$, then acts only on the $g(w)$ factor.

That is, $d(\log(G(t))/dt \le \lambda^2 \kappa^2$ and

$$\log G(1) \le \log G(0) + \int_0^1 \lambda^2 \kappa^2 \, dt = \lambda^2 \kappa^2.$$

Take exponentials, using the fact that $G(1) = \mathbb{P} g(X, Y)$ and $G(0) = \mathbb{P} g(Z, Z) = 1$, to deduce that $\mathbb{P} g(X, Y) \le \exp\left(\lambda^2 \kappa^2\right)$. Finally, again invoke the JENSEN inequality <13> to conclude that

$$\mathbb{P} \exp\left(\lambda f(X) - \lambda \gamma_n f\right) \le \mathbb{P} \exp\left(\lambda f(X) - \lambda f(Y)\right) = G(1) \le e^{\lambda^2 (2\kappa^2)/2}.$$

The centered random variable $f(X) - \gamma_n f$ is subgaussian with scale parameter $\sqrt{2}\kappa$.

6.3.3 The stochastic calculus method

Gaussian::Lip-Ito

The PISIER-MAUREY argument showed that $f(Z) - \gamma_n f$ is subgaussian with scale factor approximately 1.57κ Talagrand's argument brought the scale factor down to approximately 1.4κ. In this Section, with a lot more technical effort the scale factor is reduced to κ, the best possible. For many purposes, where the size of various constants is not a great concern, the PISIER-MAUREY bound should suffice. If you find yourself in that situation then you might want to skip this Section. However, the stochastic calculus approach (apparently due to Maurey) has proved itself useful in other problems. You might find it helpful, therefore, to at least glance at the heuristic argument just to get a feel for how this approach works.

This very smart proof creates a different sort of path, from $\gamma_n f$ to $f(Z)$, using a stochastic integral with respect to an n-dimensional BROWNIAN motion, $B_t = (B_{t,1}, \ldots, B_{t,n})$ for $0 \le t \le 1$. That is, the $\{B_{t,i} : 0 \le t \le 1\}$, for $i = 1, \ldots, n$, are independent BROWNIAN motions on $[0, 1]$.

Remark. I have no satisfying intuitive explanation for why a wiggly Brownian motion path does better than the more rigid paths from the previous two Sections. A miracle, perhaps.

Write $\mathcal{F}_t$ for $\sigma\{B_s : 0 \le s \le t\}$, a sigma-field that corresponds to what we learn by watching $\{B_s : 0 \le s \le t\}$. Given $\mathcal{F}_t$, the process $\{B_s : t \le s \le 1\}$ is BROWNIAN motion started from B_t. In particular, the conditional distribution of $B_1 - B_t$ is $N_n(0, (1-t)I_n)$, so that

\E@ M.def <15>
$$M_t := \mathbb{P}_{\mathcal{F}_t} f(B_1) = F(B_t, t) \qquad \text{where } F(x,t) := \gamma_n^z f(x + z\sqrt{1-t}\,).$$

The process $\{(M_t, \mathcal{F}_t) : 0 \le t \le 1\}$ is a martingale.

Notice the reappearance of our old friend $a_t = \sqrt{1-t}$. The function F satisfies the boundary conditions $F(x,1) = f(x)$ and $F(0,0) = \gamma_n f$. If we again assume f to be very smooth then the derivative $\dot{F} = \partial F(x,t)/\partial x$, with components $\dot{F}_i(x,t)$, is controlled by the LIPSCHITZ property:

\E@ grad2 <16>
$$\sqrt{\sum\nolimits_i \dot{F}_i(x,t)^2} = |\dot{F}(x,t)| \le \gamma_n^z |\dot{f}(x + a_t z)| \le \kappa.$$

The key to the whole subgaussian proof is a stochastic integral representation of M in a form that allows us to exploit <16>. The rigorous argument uses Itô's formula to express M as a stochastic integral:

\E@ Ito1 <17>
$$M_t - \gamma_n f = \int_0^t \langle \dot{F}(B_s, s), dB_s \rangle = \sum\nolimits_i \int_0^t \dot{F}_i(B_s, s)\, dB_{s,i}.$$

Instead of starting with the ITÔ formula, I'll first present a non-rigorous calculation that gives a rough idea of what is going on. If you are comfortable with stochastic calculus you can skip the next few pages of heuristics by jumping straight to equation <24>.

The trick with stochastic integration is to carry TAYLOR expansions of functions of B_t out to second order, using the fact that for $\delta > 0$

$$\Delta B_t := B_{t+\delta} - B_t = (\Delta B_{t,1}, \ldots, \Delta B_{t,n})$$

has a $N(0, \delta I_n)$ distribution independent of $\mathcal{F}_t$. In particular,

\E@ DelB <18>
$$\mathbb{P}_{\mathcal{F}_t} \Delta B_{t,i} = 0 \quad \text{AND} \quad \mathbb{P}_{\mathcal{F}_t} \Delta B_{t,i} B_{t,j} = \begin{cases} \delta & \text{if } i = j \\ 0 & \text{otherwise} \end{cases}.$$

Accordingly, if $0 \le t < 1$ and δ is positive and small enough,

\E@ M.incr1 <19>
$$\begin{aligned} \Delta M_t := M_{t+\delta} - M_t \approx \dot{F}_t(B_t, t)\delta + \sum\nolimits_{i \in [\![n]\!]} \dot{F}_i(B_t, t) \Delta B_{t,i} \\ + \tfrac{1}{2} \sum\nolimits_{i,j \in [\![n]\!]} \ddot{F}_{i,j}(B_t, t) \Delta B_{t,i} \Delta B_{t,j}. \end{aligned}$$

The approximation sign ($\approx$) is intended to suggest that terms of order smaller than δ (in some probabilistic sense) have been omitted. A rigorous argument would need to include error terms. As before, $\dot{F}_t$ denotes $\partial F(x,t)/\partial t$, and $\dot{F}_i$ denotes $\partial F(x,t)/\partial x_i$, and so on.

We can get a good approximation for $M_1 - M_0$ by summing up the increments ΔM_t over a fine grid $\mathbb{G}(m) = \{\ell\delta : \ell = 0, 1, \ldots, m-1\}$, where $\delta = 1/m$ for a large positive integer m.

The martingale property for M_t and <18> give

$$0 = \mathbb{P}_{\mathcal{F}_t} \Delta M_t \approx \dot{F}_t(B_t, t)\delta + \tfrac{1}{2} \sum\nolimits_i \ddot{F}_{i,i}(B_t, t)\delta,$$

which suggests (correctly) that $\dot{F}_t + \frac{1}{2}\sum_i \ddot{F}_{i,i} = 0$. That is,

\E@ backward.heat <20>
$$\partial F(x,t)/\partial t + \tfrac{1}{2} \sum\nolimits_i \partial^2 F(x,t)/\partial x_i^2 = 0 \qquad \text{for } x \in \mathbb{R}^n \text{ and } 0 < t < 1,$$

an equality that is easy to verify by means of an appeal to the i-b-t trick: see Problems [2] and [5].

With the help from <20>, the approximation <19> simplifies to

\E@ M.incr2 <21>
$$\Delta M_t \approx \sum\nolimits_i \dot{F}_i(B_t, t) \Delta B_{t,i} + R_1(t,\delta) + R_2(t,\delta)$$

where

$$\begin{aligned} R_1(t,\delta) &= \tfrac{1}{2} \sum\nolimits_i \ddot{F}_{i,i}(B_t, t) \left(\Delta B_{t,i}^2 - \delta \right), \\ R_2(t,\delta) &= \tfrac{1}{2} \sum\nolimits_{i,j} \{i \ne j\} \ddot{F}_{i,j}(B_t, t) \Delta B_{t,i} \Delta B_{t,j}. \end{aligned}$$

Both $\mathbb{P}_{\mathcal{F}_t} R_1$ and $\mathbb{P}_{\mathcal{F}_t} R_2$ are zero. If we could keep B_t in a bounded region then the contributions from the $\ddot{F}_{i,j}$'s would stay bounded and a conditioning argument could be used to show that the contributions from R_1 and R_2 would be small in an $\mathcal{L}^2$ sense if m were large.

Such a boundedness effect could be achieved by borrowing a stopping time trick that is often used in proofs of Itô's formula (see Karatzas and Shreve, 1988, page 149, for example): replace B_t by the stopped process $Z_t = B_{t\wedge\tau(r)}$ where $\tau(r) := \inf\{t \in \mathbb{R}^+ : |B_t| > r\}$. Unfortunately, such a substitution would complicate the heuristics based on <18> but, as you'll soon see, it is easy to accommodate in a rigorous stochastic calculus argument.

In the heuristic spirit I'll just ignore the R_1 and R_2 contributions, simplifying <21> to $\Delta M_t \approx \sum_i \dot{F}_i(B_t,t)\Delta B_{t,i}$. Another conditioning arguments (which kills off cross-product terms) then gives

$$\mathbb{P}_{\mathcal{F}_t}(\Delta M_t)^2 \approx \sum\nolimits_i \mathbb{P}_{\mathcal{F}_t}\dot{F}_i(B_t,t)^2\delta = \mathbb{P}_{\mathcal{F}_t}|\dot{F}(B_t,t)|^2\delta \le \kappa^2\delta \qquad \text{by <16>.}$$

Remark. In stochastic calculus jargon, this approximation corresponds to a DOOB-MEYER decomposition of the submartingale M_t^2.

Further approximation then leads to

$$\mathbb{P}_{\mathcal{F}_t}e^{\lambda\Delta M_t} \approx 1 + \lambda\mathbb{P}_{\mathcal{F}_t}\Delta M_t + \tfrac{1}{2}\lambda^2\mathbb{P}_{\mathcal{F}_t}(\Delta M_t)^2 \lessapprox \exp(\lambda^2\kappa^2\delta/2).$$

Successive conditioning steps finally produce (approximately) the conclusion

\E@ peel <22>
$$\mathbb{P}e^{\lambda M_1 - \lambda\gamma_n f} \approx \mathbb{P}\prod\nolimits_{t\in\mathbb{G}(m)} e^{\lambda\Delta M_t} \le \exp\left(m\lambda^2\kappa^2\delta/2\right) = e^{\lambda^2\kappa^2/2},$$

which should give the desired subgaussian bound in the limit.

End of heuristics.

To turn the last few pages of approximations into a rigorous argument one would need to engage in exactly the sort of calculations that lead to the ITÔ formula. Instead of providing those details I'll just refer you to the places in that excellent book by C&W = Chung and Williams (2014) where the stochastic calculus is derived.

Start again with the martingale $M_t = F(B_t,t)$ for $0 \le t \le 1$. By C&W Theorem 5.10 (the multidimensional version of the ITÔ formula), the crude approximation <19> is replaced by

\E@ multiIto <23>
$$F(B_t,t) - F(B_0,0) = \int_0^t \dot{F}_t(B_s,s)ds + \sum\nolimits_{i\in[\![n]\!]}\int_0^t \dot{F}_i(B_s,s)\,dB_{s,i} + \tfrac{1}{2}\int_0^t \sum\nolimits_{i,j\in[\![n]\!]}\int_0^t \ddot{F}_{i,j}(B_s,s)d[B_{\cdot,i},B_{\cdot,j}]_s \quad .$$

The quadratic variation-covariation process of B is a deterministic matrix valued-process with (i,j)th component

$$[B_{\cdot,i},B_{\cdot,j}]_t = \begin{cases} t & \text{for } i = j \\ 0 & \text{for } i \ne j \end{cases} .$$

All the terms with $i \ne j$ in the final double sum in <23> vanish, leaving only the $i = j$ terms:

$$\int_0^t \ddot{F}_{i,i}(B_s,s)d[B_{\cdot,i},B_{\cdot,i}]_s = \int_0^t \ddot{F}_{i,i}(B_s,s)\,ds.$$

Thus <23> becomes

$$F(B_t,t) - F(B_0,0) = \int_0^t \dot{F}_t(B_s,s) + \tfrac{1}{2}\sum\nolimits_i \ddot{F}_{i,i}(B_s,s)\,ds + \sum\nolimits_i \int_0^t \dot{F}_i(B_s,s)\,dB_{s,i}.$$

\E@ Ito2 <24>

Once again the identity <20> kills the $\int_0^t \ldots ds$ contribution, leaving

$$M_t - \gamma_n f = \sum\nolimits_i N_{t,i} \qquad \text{where } N_{t,i} := \int_0^t \dot{F}_i(B_s,s)\,dB_{s,i}.$$

\E@ Ito3 <25>

Remark. In fact we don't even need <20> because $M_t - \sum_i N_{t,i}$ is a continuous local martingale whose sample paths are locally of bounded variation, which forces it to be (for almost all paths) equal to $M_0 = \gamma_n f$. See C&W Corollary 4.5.

The quadratic variation $[M,M]$ process, which is usually abbreviated to $[M]$, corresponds to the limiting form of the heuristic $\sum_{t\in\mathbb{G}(m)} (\Delta M_t)^2$. It can be derived from the variation-covariation processes of the BROWNIAN motions: by C&W Theorem 5.7,

$$\begin{aligned}
[M]_t &= \sum\nolimits_{i,j} [N_{\cdot,i}, N_{\cdot,j}]_t \\
&= \sum\nolimits_{i,j} \int_0^t \dot{F}_i(B_s,s) F_j(B_s,s)\, d[B_{\cdot,i}, B_{\cdot,j}]_s \\
&= \sum\nolimits_i \int_0^t \dot{F}_i(B_s,s)^2 ds = \int_0^t |\dot{F}(B_s,s)|^2 \le \kappa^2 t.
\end{aligned}$$

\E@ [M] <26>

Finally, to bound the MGF $\mathbb{P}e^{\lambda(M_t - M_0)}$ we can use another stochastic calculus trick to replace the peeling argument <22>. By C&W Theorem 6.2, the process $Z_t := \exp(\lambda M_t - \frac{1}{2}\lambda^2 [M]_t)$ is a local martingale, that is, for some sequence of bounded stopping times τ_j that increases pointwise to ∞, the process $Z(t\wedge\tau_j)$ is a martingale. For each j,

$$\begin{aligned}
e^{\lambda M_0} = \mathbb{P}Z_0 = \mathbb{P}Z(1\wedge\tau_j) &= \mathbb{P}\exp\left(\lambda M_{1\wedge\tau_j} - \tfrac{1}{2}\lambda^2 [M]_{1\wedge\tau_j}\right) \\
&\ge \mathbb{P}\exp\left(\lambda M_{1\wedge\tau_j} - \tfrac{1}{2}\lambda^2\kappa^2\right).
\end{aligned}$$

That is,

$$\mathbb{P}\exp\left(\lambda M_{1\wedge\tau_j}\right) \le \exp\left(\lambda\gamma_n f + \tfrac{1}{2}\lambda^2\kappa^2\right) \qquad \text{for each } j.$$

Complete the argument by letting j tend to infinity. An appeal to the FATOU lemma as $j\to\infty$ then gives

$$\mathbb{P}e^{\lambda f(B_1)} = \mathbb{P}e^{\lambda M_1} \le \liminf_j \mathbb{P}e^{\lambda M_{1\wedge\tau_j}} \le \exp\left(\lambda\gamma_n f + \tfrac{1}{2}\lambda^2\kappa^2\right),$$

the desired subgaussian bound.

*6.4 The gaussian isoperimetric inequality

Gaussian::S:iso

For each subset A of $\mathbb{R}^n$ define $A^\delta := \{z \in \mathbb{R}^n : d(a, A) \le \delta\}$, where $d(z, A) := \inf\{|z - y| : y \in A\}$.

Gaussian::gisop <27> **Gaussian isoperimetric inequality.** *If A is a* BOREL *subset of $\mathbb{R}^n$ with $\gamma_n A = \Phi(\alpha)$ then $\gamma_n A^\delta \ge \Phi(\alpha + \delta)$ for each $\delta \ge 0$. The lower bound is achieved when A is any closed halfspace with gaussian measure $\Phi(\alpha)$.*

It is the reduction from an n-dimensional problem, with n arbitrarily large, to a one-dimensional calculation for the lower bound that makes the isoperimetric inequality so powerful. The next Example, which derives concentration of LIPSCHITZ functionals around a median, illustrates.

Recall that a median of a (real valued) random variable X is any constant m for which $\mathbb{P}\{X \ge m\} \ge 1/2$ and $\mathbb{P}\{X \le m\} \ge 1/2$. Such an m always exists, but it need not be unique.

Gaussian::gauss.conc <28> **Example.** Suppose f is a LIPSCHITZ function on $\mathbb{R}^n$ with $\|f\|_{Lip} \le \kappa$. Under γ_n, the random variable $f(z)$ has at least one median, a number M for which

$$\gamma_n\{f(z) \le M\} \ge \tfrac{1}{2} \qquad \text{AND} \qquad \gamma_n\{f(z) \ge M\} \ge \tfrac{1}{2}.$$

Define $A = \{z \in \mathbb{R}^n : f(z) \le M\}$ so that $\gamma_n A \ge 1/2 = \Phi(0)$. If $d(x, A) < u$ then there exist a point $z \in A$ with $d(z, x) < u$. From the LIPSCHITZ property and the fact that $f(x) \le M$ we then get

$$f(x) < f(z) + \kappa u < M + \kappa u.$$

Conversely, if $f(x) \ge M + \kappa u$ then $d(x, A) \ge u$. It follows that

$$\gamma_n\{f(x) \ge M + \kappa u\} \le \gamma_n\{d(x, A) \ge u\} \le \bar{\Phi}(0 + u) \le \tfrac{1}{2}\exp(-u^2/2).$$

An analogous argument for deviations from the set $\{z : f(z) \ge M\}$ gives the companion lower bound. Together the two bounds give a concentration property for f,

\E@ med.conc <29> $$\gamma_n\{z : |f(z) - M| \ge \kappa y\} \le 2\bar{\Phi}(u) \le \exp(-y^2/2),$$

□ where M is a median for f under γ_n.

Remark. Inequality <29> also gives concentration around the mean $\mu = \mathbb{P}f(Z)$ because it also implies that the mean and the median are close to each other:

$$|\mu - M| \le \gamma_n|f - M| \le \kappa \int_0^\infty \gamma_n\{|f(z) - M| \ge \kappa y\}\, dy = C\kappa,$$

with $C = 1/(2\sqrt{2\pi})$. Thus

$$\gamma_n\{|f - \mu| \ge \kappa(C + y)\} \le \gamma_n\{|f - M| \ge \kappa y\} \le \exp(-y^2/2).$$

6.5 Tail bound comparisons

Gaussian::S:TailComparison

One of the earliest comparison inequalities for the MVN is due to Slepian (1962, page 498).

Gaussian::Slepian <30> **Slepian's comparison inequality.** *For some finite index set B suppose $X = (X_b : b \in B)$ and $Y = (Y_b : b \in B)$ both have centered* MVN *distributions. Suppose also that, for all $(b, b') \in B \times B$,*

\E@ Slep <31> $$\Delta(b,b') := \mathbb{P}(Y_b Y_{b'}) - \mathbb{P}(X_b X_{b'}) \le 0 \qquad \text{with equality when } b = b'.$$

□ *Then $\mathbb{P} \cup_b \{X_b \ge r_b\} \le \mathbb{P} \cup_b \{Y_b \ge r_b\}$ for each r in $\mathbb{R}^B$.*

Remarks.

(i) Of course the prime on b' does not denote a derivative. I just need some way of distinguishing between two elements of B. I'll also write $X(b)$ for X_b, and so on, if the subscripts get too messy.

(ii) If you are wondering whether the equality constraint in <31> could be omitted, consider the case where $B = \{1, 2\}$ with $Y \sim N_2(0, I_2)$ and $X_1 = Y_1$ and $X_2 = 100Y_2$. Note that

$$\mathbb{P}\left(\{X_1 \ge 5\} \cup \{X_2 \ge 5\}\right) \approx 1/2,$$
$$\mathbb{P}\left(\{Y_1 \ge 5\} \cup \{Y_2 \ge 5\}\right) \approx 0,$$

which clearly would violate the conclusion of the Theorem.

(iii) I often have trouble remembering which way the inequalities should go. It helps me to rewrite the covariance assumptions as $\mathbb{P}X_b^2 = \mathbb{P}Y_b^2$ and $\mathbb{P}|X_b - X_{b'}|^2 \le \mathbb{P}|Y_b - Y_{b'}|^2$ for all b, b'. The Y components are more 'spread out' than the X components, which makes $\max_b Y_b$ stochastically larger than $\max_b X_b$, that is,

$$\mathbb{P}\{\max_b X_b \ge r\} \le \mathbb{P}\{\max_b Y_b \ge r\} \qquad \text{for each real } r.$$

This property ensures (Marshall et al., 2011, Chapter 11) existence of a pair of random variables on some new probability space, M_X with the same distribution as $\max_b X_b$ and M_Y with the same distribution as $\max_b Y_b$, such that $M_X \le M_Y$ almost surely. For each increasing function h we then have $h(M_X) \le h(M_Y)$ almost surely and hence

\E@ Slepian.max <32> $$\mathbb{P}h(\max_b X_b) = \mathbb{P}h(M_X) \le \mathbb{P}h(M_Y) = \mathbb{P}h(\max_b Y_b),$$

provided the expected values are well defined (no $\infty - \infty$ difficulties). In particular $\mathbb{P}\max_b Y_b \le \mathbb{P}\max_b Y_b$. Compare with the version of the Fernique inequality stated in Theorem <4>, which does not require the equality $\mathbb{P}X_b^2 = \mathbb{P}Y_b^2$.

Gordon (1985) extended the SLEPIAN inequality to doubly indexed arrays with a more complicated comparison between tail events. If you are wondering why anyone would be interested in this ∩∪ stuff take a look at Section 6.6

Gaussian::Gordon <33>

Theorem. *Suppose A and B are finite sets and $X = (X_{a,b} : (a,b) \in A \times B)$ and $Y = (Y_{a,b} : (a,b) \in A \times B)$ both have centered* MVN *distributions. Suppose also that, for all $i = (a,b)$ and $j = (a',b')$ in $A \times B$,*

$$\Delta(i,j) := \mathbb{P}Y_iY_j - \mathbb{P}X_iX_j \qquad \begin{cases} = 0 & \text{if } i = j \\ \le 0 & \text{if } a = a' \text{ and } b \ne b' \\ \ge 0 & \text{if } a \ne a' \end{cases}.$$

Then

$$\mathbb{P} \cap_{a\in A} \cup_{b\in B} \{X_{a,b} \ge r_{a,b}\} \le \mathbb{P} \cap_{a\in A} \cup_{b\in B} \{Y_{a,b} \ge r_{a,b}\}$$

for each r in $\mathbb{R}^{A\times B}$. □

Gaussian::Gordon.minmax <34>

Corollary. *Under the conditions of the Theorem, by taking $r_{a,b} = r$ for all a, b we get $\mathbb{P}\{\min_a \max_b X_{a,b} \ge r\} \le \mathbb{P}\{\min_a \max_b Y_{a,b} \ge r\}$ for each real r. That is, $\min_a \max_b Y_{a,b}$ is stochastically larger than $\min_a \max_b X_{a,b}$, implying*

$$\mathbb{P}h\left(\min_a \max_b X_{a,b}\right) \le \mathbb{P}h\left(\min_a \max_b Y_{a,b}\right)$$

for each increasing function h, provided there are no $\infty - \infty$ difficulties.

Notice that the GORDON inequality reduces to the SLEPIAN inequality if the set A is taken to be a singleton.

Remark. The covariance assumptions of Theorem <33> imply, for all distinct $i = (a,b)$ and $j = (a',b')$ in $A \times B$,

$$\begin{aligned} &\mathbb{P}|Y_i - Y_j|^2 - \mathbb{P}|X_i - X_j|^2 \\ &= \left(\mathbb{P}Y_i^2 + \mathbb{P}Y_j^2 - \mathbb{P}X_i^2 - \mathbb{P}X_j^2\right) - 2\left(\mathbb{P}Y_iY_j - \mathbb{P}X_iX_j\right) \\ &\begin{cases} \ge 0 & \text{if } a = a' \text{ and } b \ne b' \\ \le 0 & \text{if } a \ne a' \end{cases}. \end{aligned} \qquad \text{\E@ better.GF} \quad <35>$$

Gordon (1985, Theorem 1.4) actually established an analog of the FERNIQUE inequality,

$$\mathbb{P} \min_{a\in A} \max_{b\in B} X_{a,b} \le \mathbb{P} \min_{a\in A} \max_{b\in B} Y_{a,b}, \qquad \text{\E@ Gordon.Fernique} \quad <36>$$

under assumption <35>. That is, he derived the special case of Corollary <34> where h is the identity map without requiring that $\mathbb{P}X_{a,b}^2 = \mathbb{P}Y_{a,b}^2$ for all (a,b) in $A \times B$. Unfortunately, the proof is more complicated than the proof of Theorem <33>, just as the original proof of the FERNIQUE inequality is more complicated than the proof of Theorem <30>.

See Section 6.6 for a Lemma and an Example that show the extra costs incurred by the extra assumption that $\mathbb{P}X_{a,b}^2 = \mathbb{P}Y_{a,b}^2$ for all (a,b).

Kahane (1986) provided a most elegant derivation of both the SLEPIAN and GORDON inequalities, by a single theorem whose proof can be slightly modified to become a path argument with an appeal to i-b-p trick. The theorem involves a twice continuously differentiable function f for which various differentiations inside expectations are legitimate. Once again, rather than hardcoding the necessary regularity requirements into the Theorem I will merely refer to f as having 'suitably integrable partial derivatives'. For example, it would suffice that f be derived from a LIPSCHITZ function by means of smoothing operations like those in Problem [1]. For the derivation of the SLEPIAN and GORDON inequalities, the f is used to approximate sets that are products of intervals of the form $[r, \infty)$.

Gaussian::Kahane <37>

Theorem. *(Kahane) Suppose $X = (X_i : i \in I)$ and $Y = (Y_i : i \in I)$ both have centered* MVN *distributions, for some finite index set I. Suppose also that*

(i) a function $f : \mathbb{R}^I \to \mathbb{R}$ is twice continuously differentiable and has 'suitably integrable partial derivatives';

(ii) the set $I \times I$ can be partitioned into three disjoint subsets $\mathcal{Z}$, $\mathcal{P}$, and $\mathcal{N}$ for which

$$\Delta(i,j) := \mathbb{P}Y_iY_j - \mathbb{P}X_iX_j \qquad \begin{cases} = 0 & \text{if } (i,j) \in \mathcal{Z} \\ \geq 0 & \text{if } (i,j) \in \mathcal{P} \\ \leq 0 & \text{if } (i,j) \in \mathcal{N} \end{cases}$$

$$\ddot{f}_{i,j}(w) := \frac{\partial^2 f}{\partial w_i \partial w_j} \qquad \begin{cases} \geq 0 & \text{if } (i,j) \in \mathcal{P} \\ \leq 0 & \text{if } (i,j) \in \mathcal{N} \end{cases} \qquad \text{for all } w \in \mathbb{R}^I.$$

□ *Then $\mathbb{P}f(X) \leq \mathbb{P}f(Y)$.*

Proof. Let W_0 and W_1 be independent random vectors, with W_0 having the same distribution as X and W_1 having the same distribution as Y. As before, define $W_t = a_tW_0 + b_tW_1$, this time (just for the sake of variety) with $a_t = \cos(t)$ and $b_t = \sin(t)$ for $0 \leq t \leq \pi/2$. Define $H(t) := \mathbb{P}f(W_t)$. Note that $\dot{a}_t = -b_t$ and $\dot{b}_t = a_t$, so that

$$\mathbb{P}\left(\dot{W}_tW_t'\right) = \dot{a}_ta_t\text{var}(X) + \dot{b}_tb_t\text{var}(Y) = a_tb_t\left(\text{var}(Y) - \text{var}(X)\right).$$

Or, in coordinate form, $\tau_t(i,j) := \mathbb{P}\dot{W}_{t,i}W_{t,j} = a_tb_t\Delta(i,j)$. By assumption and the fact that $a_tb_t \geq 0$, the product $\tau_t(i,j)\ddot{f}_{i,j}(w)$ is nonegative for every w in $\mathbb{R}^I$.

As before, we now have

$$\begin{aligned} \dot{H}(t) &= \mathbb{P}\langle \dot{W}_t, \dot{f}(W_t)\rangle \\ &= \sum\nolimits_{i\in I} \mathbb{P}\dot{W}_{t,i}\dot{f}_i(W_t) \\ &= \sum\nolimits_{i\in I, j\in I} \tau_t(i,j)\mathbb{P}\ddot{f}_{i,j}(W_t) \qquad \text{by Theorem <10>} \\ &\geq 0. \end{aligned}$$

□ It follows that $\mathbb{P}f(X) = F(0) \leq F(1) = \mathbb{P}f(Y)$, as asserted.

The remainder of this Section is devoted to explaining how Kahane's result covers both Theorems <30> (SLEPIAN) and <33> (GORDON). Even though the first Theorem is a special case of the second, I'll prove both results because the first proof gives some insight into the second.

The SLEPIAN inequality involves two centered MVN distributions, represented by random vectors $X = (X_b : b \in B)$ and $Y = (Y_b : b \in B)$. The Theorem gives conditions under which is that $\mathbb{P}g(X - r) \le \mathbb{P}g(Y - r)$ for each r in $\mathbb{R}^B$, where g is the indicator function of $\{z \in \mathbb{R}^B : z_b \ge 0\}$. In de Finetti notation, g can be reexpressed as

$$g(z) = 1 - \prod\nolimits_{b\in B}\{z_b < 0\},$$

a convenient representation of a union as a complement of an intersection.

To employ Kahane's result we need to approximate the indicator function of $(-\infty, 0)$ by a smooth, decreasing function ψ_σ that equals 1 on $(-\infty, -\sigma)$ and 0 on $[0, \infty)$. (In fact, existence of a bounded, continuous derivative $\dot\psi_\sigma$ will suffice, as far as smoothness is concerned.) At some stage the positive parameter σ is sent to zero, which ensures that

$$f(z) := 1 - \prod\nolimits_{b\in B}\psi_\sigma(z_b)$$

converges pointwise to $g(z)$. If we can show that $\mathbb{P}f(X - r) \le \mathbb{P}f(Y - r)$ for each r in $\mathbb{R}^B$ then an appeal to dominated convergence will give the coreesponding inequality with f replaced by g.

The function f has partial derivatives, for distinct b and b' in B,

$$\dot f_b(z) := \frac{\partial f(z)}{\partial z_b} = -\dot\psi_\sigma(z_b)\prod\nolimits_{\beta\in B\setminus\{b\}}\psi_\sigma(z_\beta),$$
$$\ddot f_{b,b'}(z) := \frac{\partial^2 f(z)}{\partial z_b \partial z_{b'}} = -\dot\psi_\sigma(z_b)\dot\psi_\sigma(z_{b'})\prod\nolimits_{\beta\in B\setminus\{b,b'\}}\psi_\sigma(z_\beta).$$

The function $\dot f_b$ is everywhere ≥ 0 because ψ_σ is decreasing and non-negative. The function $\ddot f_{b,b'}$ is everywhere ≤ 0.

The signs of the partial derivatives are unchanged if $f(z)$ is replaced by $f(z - r)$.

The assumptions of Theorem <37> hold when applied to $f(\cdot - r)$ for a fixed r in $\mathbb{R}^B$ with $I = B$ and $\mathcal{Z} = \{(b, b') \in B \times B : b = b'\}$ and $\mathcal{N} = I\setminus\mathcal{N}$. It follows that $\mathbb{P}f(X - r) \le \mathbb{P}f(Y - r)$, which implies the SLEPIAN inequality in the limit as $\sigma \to 0$.

For the GORDON inequality, $I = A \times B$ and

$$\mathcal{Z} = \{(a, b, a', b') \in I \times I : a = a', b = b'\},$$
$$\mathcal{N} = \{(a, b, a', b') \in I \times I : a = a', b \ne b'\},$$
$$\mathcal{P} = \{(a, b, a', b') \in I \times I : a \ne a'\}.$$

Think of elements of $\mathbb{R}^I$ as $|A|$ by $|B|$ matrices. If $Z \in \mathbb{R}^I$ define $Z_a := (Z[a, b] : b \in B)$, the Ath row of Z. The function g is replaced by

$$G(Z) = \prod\nolimits_{a\in A} g(Z_a) = \cap_{a\in A}\cup_{b\in B}\{Z[a, b] \ge 0\},$$

which is approximated by the smooth function

$$F(Z) := \prod_{a\in A} f(Z_a).$$

If $i = (a, b)$ and $j = (a', b')$ then

$$\dot{F}_i(Z) := \frac{\partial F(Z)}{\partial Z[a,b]} = \dot{f}_b(Z_a) \prod_{\alpha\in A\setminus\{a\}} f(Z_\alpha),$$

which is everywhere ≥ 0, and

$$\ddot{F}_{i,j}(Z) = \begin{cases} \ddot{f}_{b,b'}(Z_a) \prod_{\alpha\in A\setminus\{a\}} f(Z_\alpha) \leq 0 & \text{if } a = a' \\ \dot{f}_b(Z_a)\dot{f}_{b'}(Z_{a'}) \prod_{\alpha\in A\setminus\{a,a'\}} f(Z_\alpha) \geq 0 & \text{if } a \neq a' \end{cases}.$$

That is, the partial derivative $\ddot{F}_{i,j}$ is everywhere ≥ 0 if $(i, j) \in \mathcal{P}$ and is everywhere ≤ 0 if $(i, j) \in \mathcal{N}$, which agrees with the sign of $\Delta(i, j)$. The partial derivatives for $F(Z - R)$ follow the same pattern, for each $R = (r_{a,b})$ in $\mathbb{R}^{A\times B}$. Theorem <37> gives $\mathbb{P}F(X - R) \leq \mathbb{P}F(Y - R)$. The limit as σ tends to zero then gives the inequality asserted by Theorem <33>.

*6.6 An application of the GORDON inequality

Gaussian::S:Gordon

Both inequalities let us control complicated gaussian processes by means of simpler gaussian processes, as illustrated by the following result of Gordon (1985, Theorem 2.1).

Gaussian::Gthm2.1 <38>

Lemma. *Let $G = [g_{i,j}]$ be an $N \times k$ matrix of independent standard normals, and let $Z \sim N(0, I_N)$ be independent of $W \sim N(0, I_k)$ If $\mathbb{A}$ is a compact subset of $\mathbb{R}^k$ and $\mathbb{B}$ is a compact subset of $\mathbb{R}^N$, with $0 \in \mathbb{B}$, then*

$$\begin{aligned} &\alpha_{\min}\mathbb{P}\sup_{b\in\mathbb{B}}\langle b, Z\rangle - \beta_{\max}\mathbb{P}\sup_{a\in\mathbb{A}}|\langle a, W\rangle| - \alpha_{\max}\beta_{\max} \\ &\qquad \leq \mathbb{P}\inf_{a\in\mathbb{A}}\sup_{b\in\mathbb{B}}\langle b, Ga\rangle \leq \mathbb{P}\sup_{a\in\mathbb{A}}\sup_{b\in\mathbb{B}}\langle b, Ga\rangle \\ &\leq \alpha_{\max}\mathbb{P}\sup_{b\in\mathbb{B}}\langle b, Z\rangle + \beta_{\max}\mathbb{P}\sup_{a\in\mathbb{A}}|\langle a, W\rangle| + \alpha_{\max}\beta_{\max} \end{aligned}$$

where $\alpha_{\min} := \inf_{a\in\mathbb{A}}|a|_2$ and $\alpha_{\max} := \inf_{a\in\mathbb{A}}|a|_2$ and $\beta_{\max} = \sup_{b\in\mathbb{B}}|b|_2$.

Proof. It suffices to prove an analogous set of inequalities with $\mathbb{A}$ replaced by a finite subset A and $\mathbb{B}$ replaced by a finite subset B. (Then invoke dominated convergence as A expands up to a countable dense subset of $\mathbb{A}$ and B expands up to a countable dense subset of $\mathbb{B}$.)

Define two gaussian processes indexed by $\mathbb{A} \times \mathbb{B}$:

$$\begin{aligned} X_{a,b} &= |a|_2|b|_2\xi + \langle b, Ga\rangle = |a|_2|b|_2\xi + \sum_{i,j} b_i g_{i,j} a_j, \\ Y_{a,b} &= |a|_2\langle b, Z\rangle + |b|_2\langle a, W\rangle, \end{aligned}$$

where $\xi \sim N(0, 1)$ is independent of G. The role of ξ is to ensure that $\mathbb{P}X_{a,b}^2 = \mathbb{P}Y_{a,b}^2$. It comes at the slight $\pm\alpha_{\max}\beta_{\max}$ cost because

$$|X_{a,b} - \langle b, Ga\rangle| \leq \alpha_{\max}\beta_{\max}|\xi| \qquad \text{for all } (a, b) \in \mathbb{A} \times \mathbb{B},$$

which implies

$$|\inf_{a\in\mathbb{A}}\sup_{b\in\mathbb{B}} X_{a,b} - \inf_{a\in\mathbb{A}}\sup_{b\in\mathbb{B}}\langle a, Gb\rangle| \le \alpha_{\max}\beta_{\max}|\xi|,$$
$$|\sup_{a\in\mathbb{A},b\in\mathbb{B}} X_{a,b} - \sup_{a\in\mathbb{A},b\in\mathbb{B}}\langle a, Gb\rangle| \le \alpha_{\max}\beta_{\max}|\xi|.$$

The GORDON inequality does most of the work. It is easy to check, for (a, b) and (a^*, b^*) in $A \times B$, that

$$\mathbb{P}X_{a,b}X_{a^*,b^*} = |a|_2\,|a^*|_2\;|b|_2\,|b^*|_2 + \langle a, a^*\rangle\langle b, b^*\rangle,$$
$$\mathbb{P}Y_{a,b}Y_{a^*,b^*} = |a|_2\,|a^*|_2\;\langle b, b^*\rangle + |b|_2\,|b^*|_2\;\langle a, a^*\rangle,$$

which gives

$$\mathbb{P}X_{a,b}X_{a^*,b^*} - \mathbb{P}Y_{a,b}Y_{a^*,b^*} = \left(|b|_2\,|b^*|_2 - \langle b, b^*\rangle\right)\left(|a|_2\,|a^*|_2 - \langle a, a^*\rangle\right)$$
$$\begin{cases} = 0 & \text{if } a = a^* \\ \ge 0 & \text{if } a \ne a^* \end{cases}.$$

\E@ Gord <39>

The inequality

$$\mathbb{P}\min_{a\in A}\max_{b\in B} Y_{a,b} \le \mathbb{P}\min_{a\in A}\max_{b\in B} X_{a,b}$$

follows from Corollary <34> with the roles of X and Y reversed. The inequality

$$\mathbb{P}\max_{a\in A,b\in B} X_{a,b} \le \mathbb{P}\max_{a\in A,b\in B} Y_{a,b}$$

follows from the SLEPIAN inequality via <32>.

The calculations now involve only the Y process. Write $\mathcal{Y}_{\mathbb{A}}$ for $\sup_{a\in\mathbb{A}}|\langle a, W\rangle|$ and $\mathcal{Y}_B$ for $\max_{b\in B}\langle b, Z\rangle$. Then, from the definition of $Y_{a,b}$,

$$|a|_2\mathcal{Y}_B - \beta_{\max}\mathcal{Y}_{\mathbb{A}} \le \max_{b\in B} Y_{a,b} \le |a|_2\mathcal{Y}_B + \beta_{\max}\mathcal{Y}_{\mathbb{A}},$$

□ Take the maximum or minimum over $a \in A$ then take expectations.

Gaussian::Dvor <40>

Example. A compact, convex subset K of $\mathbb{R}^N$ with non-empty interior is called a ***convex body***. It is called symmetric if $-x \in K$ for each $x \in K$. A famous result of Dvoretzky (1961) asserts that low-dimensional cross sections of symmetric, convex bodies look like balls. More precisely, for each $\epsilon > 0$ there is a small $\eta(\epsilon) > 0$ such that: if K is a symmetric, convex body in $\mathbb{R}^N$ and k is a positive integer with $k \le \eta(\epsilon)\log N$ then there exists a k-dimensional subspace $\mathcal{H}$ of $\mathbb{R}^N$ and a closed ball B in $\mathcal{H}$ for which

\E@ Dvor.statement <41>

$$B \subset K \cap \mathcal{H} \subset (1+\epsilon)B.$$

In fact it also doesn't much matter if we require B to be a ball or an ellipsoid. See Problem [6].

In this Example I'll derive Dvoretzky's theorem by applying the GORDON inequality, but only for the special case where $K = [-1, +1]^N$.

Remark. The tricks for this special case can be extended to cover more general convex bodies. See Gordon (1985, 1988, 2005), for example.

The set K is the unit ball for the ℓ_∞^N norm, $|x|_\infty := \max_j |x_j|$. It is an easy exercise to prove that $|x|_\infty$ is equal to $\sup\{\langle b, x\rangle : b \in K^o\}$, where $K^o = \{b \in \mathbb{R}^N : |b|_1 \le 1\}$ is the unit ball for the ℓ_1^N norm. The little superscript 'o' means that the convex body K^o is the polar of K, that is, $K^o = \{b \in \mathbb{R}^N : \langle x, b\rangle \le 1 \text{ for all } x \in K\ \}$.

The subspace $\mathcal{H}$ will be defined as $\operatorname{span}\{z_1, \ldots, z_k\}$, where $\mathfrak{z} = (z_1, \ldots, z_k)$ is a carefully chosen realization of $\mathfrak{Z} = (Z_1, \ldots, Z_k)$, for independent $N(0, I_N)$-distributed random vectors Z_j.

If $a = ru$ with $r \ge 0$ and $|u|_2 = 1$ then $|a|_\infty = r|u|_\infty$. Thus

$$|a|_2 h(\mathfrak{Z}) \le \Big|\sum\nolimits_j a_j Z_j\Big|_\infty \le |a|_2 H(\mathfrak{Z}) \qquad \text{for each } a \text{ in } \mathbb{R}^k,$$

where

$$h(\mathfrak{Z}) := \inf_{u\in\mathbb{A}}\Big|\sum\nolimits_j u_j Z_j\Big|_\infty = \inf_{u\in\mathcal{U}} \sup_{b\in K^o}\langle b, \sum\nolimits_j u_j Z_j\rangle$$
$$H(\mathfrak{Z}) := \sup_{u\in\mathbb{A}}\Big|\sum\nolimits_j u_j Z_j\Big|_\infty = \sup_{u\in\mathcal{U}} \sup_{b\in K^o}\langle b, \sum\nolimits_j u_j Z_j\rangle.$$

We are now in the setting described by Lemma <38> for $\mathbb{B} = K^o$, with $\alpha_{\max} = \alpha_{\min} = 1 = \beta_{\max}$. The Lemma gives

\E@ Gordon.bounds <42>
$$\mathbb{P}|Z|_\infty - \mathbb{P}|W|_2 - 1 \le \mathbb{P}h(\mathfrak{Z}) \le \mathbb{P}H(\mathfrak{Z}) \le \mathbb{P}|Z|_\infty + \mathbb{P}|W|_2 - 1,$$

where $Z \sim N(0, I_N)$ and $W \sim N(0, I_k)$.

The quantity $\mathbb{P}|W|_2$ is bounded above by $\sqrt{\mathbb{P}|W|_2^2} = \sqrt{k}$. As was explained in Section MGF:sharp.normal , $\max_{j\le N} Z[j]$ concentrates very tightly around a value only slightly smaller than $\sqrt{2\log N}$. A cruder form of the argument (Problem [8]) shows that $\mathbb{P}|Z_0|_\infty = \mathbb{P}\max_{j\le N}|Z[j]| \ge (1 - o(1))\sqrt{\log N}$, with $\mathbb{P}|Z_0|_\infty \ge 0.65\sqrt{\log N}$ for $N \ge 2$ suggested numerically. It is safe to assert for $N \ge 2$ that, for some positive constant c,

$$c\sqrt{\log N} - \sqrt{k} - 1 \le \mathbb{P}h(\mathfrak{Z}) \le \mathbb{P}H(\mathfrak{Z}) \le c\sqrt{\log N} + \sqrt{k} + 1,$$

and hence, if $\sqrt{k} + 1 \le c\epsilon\sqrt{\log(N)}/3$ and $0 < \epsilon < 1$,

$$\frac{\mathbb{P}H(\mathfrak{Z})}{\mathbb{P}h(\mathfrak{Z})} \le \frac{c\sqrt{\log N} + \sqrt{k} + 1}{c\sqrt{\log N} - \sqrt{k} - 1} \le \frac{1 + \epsilon/3}{1 - \epsilon/3} \le 1 + \epsilon.$$

From Problem [7] there must exist some realization $\mathfrak{z} = (z_1, \ldots, z_k)$ of $\mathfrak{Z}$ for which $0 < h(\mathfrak{z}) \le H(\mathfrak{z}) \le (1 + \epsilon)h(\mathfrak{z})$. If we write R for $H(\mathfrak{z})$ then

$$|a|_2 R/(1 + \epsilon) \le |\sum\nolimits_i a_i z_i|_\infty \le |a|_2 R \qquad \text{for each } a \text{ in } \mathbb{R}^k.$$

Let $\mathcal{H}$ be the k-dimensional subspace of $\mathbb{R}^N$ spanned by $z_1, \ldots, z_k$ and define $T : \mathbb{R}^k \to \mathcal{H}$ by $Te_i = z_i$, where $\{e_1, \ldots, e_k\}$ is the usual orthonormal basis for $\mathbb{R}^k$. Define $B = \{a \in \mathbb{R}^k : |a|_2 \le 1/R\}$. The ellipsoid $D := TB$ has the properties asserted by Dvoretzky's theorem.

□
$$D \subset \mathcal{H} \cap K_\infty = \Big\{\sum\nolimits_i a_i z_i : |\sum\nolimits_i a_i z_i|_\infty \le 1\Big\} \subset (1 + \epsilon)D.$$

6.7 A generalized FERNIQUE inequality

Gaussian::S:Fernique

The main topic of this Section is the proof of a stronger version of Theorem <4> from Section 6.1.2: for random vectors $X = (X_1, \ldots, X_n)$ and $Y = (Y_1, \ldots, Y_n)$ with centered MVN distributions, if

\E@ Fernique.assump <43>
$$\mathbb{P}|X_i - X_j|^2 \le \mathbb{P}|Y_i - Y_j|^2 \qquad \text{for all } i, j.$$

then

\E@ Fernique1 <44>
$$\mathbb{P} \max_i X_i \le \mathbb{P} \max_i Y_i.$$

Fernique (1975, Section 2.1) deduced <44> from a stronger inequality for the ranges $\max(Y_i - Y_j)$ and $\max(X_i - X_j)$. He pointed out that <43> is weaker than the assumption made for the SLEPIAN inequality, in that it does not require $\mathbb{P}X_i^2 = \mathbb{P}Y_i^2$ for each i.

Fernique's proof combined a path argument, FOURIER inversion, and a decomposition of $\mathbb{R}^n$ into polyhedral subsets corresponding to the regions where the max is achieved by different coordinates. Subsequently, Chatterjee (2005) gave a much simpler path proof based on a smooth approximation to the maximum function. Chatterjee's method also covered a result of Vitale (2000), who had employed an ingenious limit argument to show that <44> also implies result,

\E@ Fernique2 <45>
$$\mathbb{P} \max_i (X_i + \mu_i) \le \mathbb{P} \max_i (Y_i + \mu_i) \qquad \text{for all } \mu \in \mathbb{R}^n.$$

That is, Vitale effectively removed the assumption that the variables are centered to zero expected values.

Remark. See Pollard (2001, Section 12.3) for a a slightly flawed exposition of Fernique's original method. As a friend pointed out to me, with my usual interpretation of $\{i \ne j\}$ as the indicator function of the set $\{i \in [\![n]\!] : i \ne j\}$ the quantity $\max\{i \ne j\} x_i$ on page 276 would correspond to the function $L_j(x)^+$, not the intended $L_j(x) := \max\{x_i : i \in [\![n]\!] \backslash \{j\}\}$.

Chatterjee's smooth approximation to $m(w) := \max_{i \le n} w_i$ is sometimes called a ***soft maximum***. The relevant facts are summarized in the next Lemma.

Gaussian::soft.max <46>
Lemma. *For w in $\mathbb{R}^n$ and $\lambda > 0$ define $S(w, \lambda) = \sum_{i \le n} \exp(\lambda w_i)$ and $\mathfrak{M}(w, \lambda) := \lambda^{-1} \log S(w, \lambda)$. Then*

(i) $\mathfrak{M}(w, \lambda) = m(w) + R(w, \lambda)$ where $0 \le R(w, \lambda) \le \lambda^{-1} \log n \to 0$ as $\lambda \to \infty$.

(ii) The function $w \mapsto \mathfrak{M}(w, \lambda)$ is infinitely differentiable with

$$p_i(w, \lambda) := \dot{\mathfrak{M}}_i(w, \lambda) := \frac{\partial \mathfrak{M}(w, \lambda)}{\partial w_i} = e^{\lambda w_i} / S(w, \lambda)$$

and

$$\ddot{\mathfrak{M}}_{i,j}(w,\lambda) := \frac{\partial^2 \mathfrak{M}(w,\lambda)}{\partial w_i \partial w_j} = \begin{cases} \lambda p_i(w,\lambda) - \lambda p_i(w,\lambda)^2 & \text{if } j = i \\ -\lambda p_i(w,\lambda) p_j(w,\lambda) & \text{if } j \neq i \end{cases}.$$

Proof. For (i) write $S(w,\lambda)$ as $e^{\lambda m w)} \sum_{i \le n} e^{\lambda(w_i - m(w))}$ then note that $\lambda(w_i - m(w)) \le 0$ for all i with equality at any w_i that achieves the maximum, $m(w)$.

□ The equalities in (ii) are just calculus exercises.

The proof of <45> also depends on a simple real variable fact: for any real numbers $p_1, \ldots, p_n$ with $\sum_i p_i = 1$ and real numbers $z_1, \ldots, z_n$ we have the identity

\E@ identity <47>

$$\begin{aligned} \sum\nolimits_{i,j} z_i z_j \Big(\{i = j\} p_i - p_i p_j \Big) &= \sum\nolimits_i p_i z_i^2 - \Big(\sum\nolimits_i p_i z_i \Big)^2 \\ &= \tfrac{1}{2} \sum\nolimits_{i,j} p_i p_j (z_i - z_j)^2 \end{aligned}$$

Define $G(t) := G(t,\mu,\lambda) = \mathbb{P} g(W_t)$ for $g(w) := g(w,\mu,\lambda) := \mathfrak{M}(w+\mu,\lambda)$, with both μ and λ held fixed. Write $p_i(w) = p_i(w,\mu,\lambda)$ for $\dot{g}_i(w,\mu,\lambda) = \dot{\mathfrak{M}}_i(w+\mu,\lambda)$ and let $\widetilde{X}, \widetilde{Y}, W_0, W_1$ be independent random vectors, with both $\widetilde{X}$ and W_0 distributed like X and both $\widetilde{Y}$ and W_1 distributed like Y. Then Lemma <11> and <47> give

$$\begin{aligned} \dot{G}(t) &= \dot{b}_t b_t \sum\nolimits_{i,j \in [\![n]\!]} \mathbb{P} \left(\widetilde{Y}_i \widetilde{Y}_j - \widetilde{X}_i \widetilde{X}_j \right) \mathbb{P} \ddot{\mathfrak{M}}_{i,j}(W_t + \mu, \lambda) \\ &= \lambda \dot{b}_t b_t \mathbb{P} \sum\nolimits_{i,j \in [\![n]\!]} \left(\widetilde{Y}_i \widetilde{Y}_j - \widetilde{X}_i \widetilde{X}_j \right) \Big(\{i = j\} p_i(W_t) - p_i(W_t) p_j(W_t) \Big) \\ &= \tfrac{1}{2} \lambda \dot{b}_t b_t \mathbb{P} \sum\nolimits_{i,j \in [\![n]\!]} \left((\widetilde{Y}_i - \widetilde{Y}_j)^2 - (\widetilde{X}_i - \widetilde{X}_j)^2 \right) p_i(W_t) p_j(W_t) \\ &= \tfrac{1}{2} \lambda \dot{b}_t b_t \sum\nolimits_{i,j \in [\![n]\!]} D_{ij} \mathbb{P} p_i(W_t) p_j(W_t) \end{aligned}$$

where $D_{i,j} := \mathbb{P}|Y_i - Y_j|^2 - \mathbb{P}|X_i - X_j|^2 \ge 0$ for all (i,j).

The function G is increasing, which implies

$$G(0) = \mathbb{P}\mathfrak{M}(X + \mu, \lambda) \le \mathbb{P}\mathfrak{M}(Y + \mu, \lambda) = G(1).$$

The desired inequality <45> then emerges as the limit when λ tends to ∞.

Remark. It seemed equally natural to me that Gordon's extension of Fernique's inequality, inequality <36>, should be derivable by replacing min max by $\mathfrak{m}\mathfrak{M}$, where $\mathfrak{m}$ is a 'soft minimum' function. I tried this approach but did not succeed. Maybe some clever reader can make this idea work, perhaps by using a diferent form of smooth approximation to min max.

6.8 Problems

Gaussian::S:problems

Gaussian::P:smooth.Lip

[1] Suppose f is a real valued function on $\mathbb{R}^n$ with $\|f\|_{Lip} \le \kappa$. Let ψ be an infinitely differentiable, nonnegative function on $\mathbb{R}^n$ with compact support and $\int \psi(z)\,dz = 1$. For each $\sigma > 0$, define

$$f_\sigma(x) := \int_{\mathbb{R}^n} f(x+\sigma z)\psi(z)\,dz = \sigma^{-n}\int f(w)\psi\left((w-x)/\sigma\right)\,dw.$$

(i) Show that

$$\sup_x |f_\sigma(x) - f(x)| \le \sup_x \int |f(x+\sigma z) - f(x)|\psi(z)\,dz \le \kappa\sigma \int |z|\psi(z)\,dz.$$

Deduce that f_σ converges uniformly to f as σ tends to zero.

(ii) Show that f_σ is infinitely differentiable.

(iii) Use the inequality

$$|f_\sigma(x) - f_\sigma(y)| \le \int |f(x+\sigma z) - f(y+\sigma z)|\psi(z)\,dz \le \kappa|x-y|$$

to show that $\|f_\sigma\|_{Lip} \le \kappa$.

(iv) Show that all the partial derivatives of f_σ are also LIPSCHITZ functions (with LIPSCHITZ norms depending on σ and ψ).

(v) From (iii), we have $|f_\sigma(x+tu) - f_\sigma(x)| \le t\kappa$ for each unit vector u and each $t > 0$. Use the fact that $\left(f_\sigma(x+tu) - f_\sigma(x)\right)/t \to \langle u, \nabla f_\sigma\rangle$ as $t \searrow 0$ to show that $\sum_i (\partial f_\sigma/\partial x_i)^2 = |\nabla f_\sigma|^2 \le \kappa^2$.

(vi) Suppose W has a subgaussian distribution. Show that $\mathbb{P}e^{\lambda f_\sigma(W)} \to \mathbb{P}e^{\lambda f(W)}$ for each real λ as $\sigma \to 0$. Hint: dominated convergence using $|f_\sigma(x)| \le |f(0)| + \kappa|x| + 1$ for small enough σ.

Gaussian::P:Stein

[2] Suppose Z is $N(0,1)$ distributed and F is an absolutely continuous, real-valued function on the real line with almost sure (LEBESGUE) derivative f. That is f is integrable at least on each bounded interval and $F(b) - F(a) = \int_a^b f(t)\,dt$ for $-\infty < a < b < \infty$. If $\mathbb{P}|f(Z)| < \infty$, show that $\mathbb{P}|ZF(Z)| < \infty$ and $\mathbb{P}ZF(Z) = \mathbb{P}f(Z)$ by these steps. The argument is essentially just a very careful integration-by-parts, making sure there are no hidden $\infty - \infty$ cancellations.

(i) Explain why we may assume $F(0) = 0$ and $f \ge 0$ without loss of generality. Note that these assumptions imply that $xF(x) \ge 0$ for all x. Hint: Split into $f^\pm$ contributions.

(ii) Let ϕ denote the $N(0,1)$ density. Use the FUBINI theorem to show that

$$\mathbb{P}\{Z > 0\}ZF(Z) = \int_0^\infty \int_0^\infty z\phi(z)\{0 \le t \le z\}f(t)\,dt\,dz = \int_0^\infty f(t)\phi(t)\,dt.$$

(iii) Argue similarly for $\mathbb{P}\{Z < 0\}ZF(Z)$.

Gaussian::P:gauss.ibp

[3] Suppose $(Z, X_1, \ldots, X_m)$ has a MVN distribution with $Z \sim N(0, \sigma^2)$ (but no assumptions about the means or covariances for the X_i's). Suppose also that $G : \mathbb{R}^m \to \mathbb{R}$ is continuously differentiable, with partial derivatives $\dot{G}_i(x_1, \ldots, x_m) := \partial G(x_1, \ldots, x_m)/\partial x_i$. If $\mathbb{P}|ZG(X_1, \ldots, X_m)| < \infty$ and $\mathbb{P}|\dot{G}_i(X_1, \ldots, X_m)| < \infty$ for each i, show that

$$\mathbb{P}ZG(X_1, \ldots, X_m) = \sum\nolimits_{i \le m} \tau_i \mathbb{P}\dot{G}_i(X_1, \ldots, X_m) \qquad \text{where } \tau_i := \mathbb{P}ZX_i.$$

Follow these steps.

(i) Without loss of generality suppose σ equals 1. (Equivalently, divide both sides of the asserted representation by σ.)

(ii) Show that Z is independent of the random variables $Y_i := X_i - \tau_i Z$ for $i \in [\![m]\!]$. [Check that $\operatorname{cov}(Z, Y_i) = 0$.] For fixed y_i values, use Problem [2] to show that

$$\begin{aligned} &\mathbb{P}\left(ZG(X) \mid Y_i = y_i \text{ for } i \in [\![m]\!]\right) \\ &= \gamma_n^z z G(y_1 + \tau_1 z, \ldots, y_m + \tau_m z) \\ &= \sum\nolimits_i \tau_i \gamma_n^z \dot{G}_i(y_1 + \tau_1 z, \ldots, y_m + \tau_m z). \end{aligned}$$

(iii) Average out over the y_i's with respect to the joint distribution of $Y_1, \ldots, Y_m$.

Gaussian::P:Isserlis

[4] (cf. Isserlis 1918; or google "Wick"; or look at Janson 1997, Theorem 1.28) Suppose $X \sim N_n(0, V)$. For each subset A of $[\![n]\!]$ define $M_A := \mathbb{P}\prod_{i \in A} X_i$.

(i) Use Problem [3] to show that $M_{[\![n]\!]} = \sum_{j \ge 2} V_{1,j} M_{[\![n]\!] \setminus \{1,j\}}$.

(ii) Deduce that $M_{[\![n]\!]}$ is either zero (for n odd) or it can be written as a sum of products of off-diagonal elements of V.

(iii) Calculate the representations for $n = 4$ and $n = 5$.

Gaussian::P:pde

[5] For suitably regular f (such as the f_σ from Problem [1]) and $a_t = \sqrt{1-t}$, show that the function $F(x,t) = \gamma_n^z f(x + z a_t)$ (as in equation <15>) satisfies

$$\begin{aligned} \partial F(x,t)/\partial t &= \dot{a}_t \sum\nolimits_i \gamma_n^z z_i \dot{f}_i(x + z a_t) \qquad \text{for } 0 < t < 1 \\ &= a_t \dot{a}_t \sum\nolimits_i \gamma_n^z \ddot{f}_{i,i}(x + z a_t) \qquad \text{by Problem [2]} \\ &= -\tfrac{1}{2} \sum\nolimits_i \partial^2 F(x,t)/\partial x_i^2. \end{aligned}$$

Remark. If the "−" sign in the final line were removed the equation would become $\partial F(x,t)/\partial t = \frac{1}{2}\sum_i \partial^2 F(x,t)/\partial x_i^2$, the heat equation. The negative sign comes from the $1 - t$ in the definition of a_t; time is running backwards. Maybe the partial differential equation could be called the 'backwards heat equation'.

Gaussian::P:ellipsoid

[6] A subset D of $\mathbb{R}^N$ is called an ellipsoid in a k-dimensional subspace $\mathcal{H}$ of $\mathbb{R}^N$ if there is a linear map T that is a bijection from $\mathbb{R}^k$ onto $\mathcal{H}$ of $\mathbb{R}^N$ such that such that $D = h + TB$, where $B = \{x \in \mathbb{R}^k : |x|_2 \le 1\}$ and $h \in \mathcal{H}$.

(i) Suppose T has the singular value decomposition $Tv_i = \lambda_i u_i$ for $1 \le i \le k$, for singular values $\lambda_1 \ge \lambda_2 \ge \cdots \ge \lambda_k > 0$ with $\{v_i : i \le k\}$ an onb of $\mathbb{R}^k$ and $\{u_i : i \le k\}$ an onb of $\mathcal{H}$. Show that

$$TB = \Big\{ \sum\nolimits_{i\le k} t_i u_i : \sum\nolimits_{i\le k} (t_i/\lambda_i)^2 \le 1 \Big\}.$$

In particular, TB is a closed ball in $\mathcal{H}$ if $\lambda_i = \lambda_1$ for all i.

(ii) Show that there is a subspace $\mathcal{L}_0$ of $\mathbb{R}^k$ of dimension approximately $k/2$ for which $D \cap (T\mathcal{L}_0)$ is a closed ball. Argue as follows. Suppose k is even, $k = 2\ell$. Choose any value λ with $\lambda_\ell \ge \lambda \ge \lambda_{\ell+1}$. For $i \le \ell$ choose positive numbers c_i and s_i for which $c_i^2 + s_i^2 = 1$ and $c_i^2 \lambda_i + s_i^2 \lambda_{k-i+1} = \lambda^2$. Define $V_i = c_i v_i + s_i v_{k-i+1}$ and $U_i = (c_i \lambda_i u_i + s_i \lambda_{k-i+1} u_{k-i+1})/\lambda$. Show that V_i and U_i are unit vectors with $TV_i = \lambda U_i$ for $i \le \ell$. Define $\mathcal{L}_0 = \operatorname{span}\{V_1, \ldots, V_\ell\}$.

If k is odd, $k = 2\ell + 1$, argue similarly with $\lambda = \lambda_{\ell+1}$ and $V_{\ell+1} = v_{\ell+1}$.

Gaussian::P:select

[7] Suppose h and H are random variables defined on a probability space $(\Omega, \mathcal{F}, \mathbb{P})$ for which $0 < h \le H$ almost surely and $\mathbb{P}H \le (1+\epsilon)\mathbb{P}h < \infty$ for some positive ϵ. Argue as follows to show that there exists an ω in Ω for which $1 \le H(\omega)/h(\omega) \le 1+\epsilon$. Define a new probability measure $\mathbb{Q}$ by $d\mathbb{Q}/d\mathbb{P} = h/\mathbb{P}h$. Show that $\mathbb{Q}(H/h) \le 1 + \epsilon$, so that $\mathbb{Q}\{\omega : H(\omega)/h(\omega) > 1 + \epsilon\} < 1$.

Gaussian::P:PP.max

[8] (repeated from Chapter 3) Suppose $\xi_1, \ldots, \xi_N$ are independent random variables, each $N(0,1)$ distributed. Define $M_n = \max_{j\le N} |\xi_j|$ and $x_n = \sqrt{\log n}$.

(i) Show that

$$\mathbb{P}M_n = \int_0^\infty \mathbb{P}\{M_n > t\}\, dt \ge x_n \mathbb{P}\{M_n > x_n\}.$$

(ii) Show that $\mathbb{P}\{M_n > t\} = 1 - (1 - 2\bar{\Phi}(t))^n \ge 1 - \exp\big(-2n\bar{\Phi}(t)\big)\big)$.

(iii) Use Laplace's bound from Section MGF:sharp.normal , $\bar{\Phi}(t) \ge (t^{-1} - t^{-3})\phi(t)$ for $t > 1$, to show that $\mathbb{P}\{M_n > x_n\} \to 1$ as $n \to \infty$.

(iv) Deduce that $\mathbb{P}M_n \ge x_n(1 - o(1))$.

Gaussian::P:bdd.sup

[9] Suppose $\{X_i : i \in \mathbb{N}\}$ is a centered Gaussian process with $X_i \sim N(0, \sigma_i^2)$. Define $M_n = \max_{i\le n} X_i$ and $M_\infty = \sup_{i\in\mathbb{N}} X_i$. Show that $M_\infty < \infty$ almost surely if and only if $\mathbb{P}M_\infty < \infty$. Argue as follows for the nontrivial implication. Suppose $\mathbb{P}\{M_\infty < \infty\} = 1$.

(i) Show that there exists an $R \in \mathbb{R}$ for which $\mathbb{P}\{M_n > R\} < 1/4$ for all large enough n.

(ii) Show that $1/4 > \mathbb{P}\{X_n > R\} = \bar{\Phi}(R/\sigma_n)$ if n is large enough. Deduce that $\sigma^2 = \sup_{i\in\mathbb{N}} \sigma_i^2$ is finite.

(iii) Write m_n for $\mathbb{P}M_n$. Use the concentration inequality from Section 6.1 to show that $\mathbb{P}\{|M_n - m_n| \ge \sigma r\} \le 2\exp(-r^2/2)$ for each n and each $r \ge 0$. Deduce that there exists an r for which $\mathbb{P}\{M_n > m_n - \sigma r\} \ge 3/4$ for each n.

(iv) From (i) and (iii) deduce that $m_n \le R + \sigma r$ for all n large enough.

(v) Show that $\mathbb{P}M_\infty = \lim_{n\to\infty} m_n \le R + \sigma r$. Hint: $0 \le M_n + |X_1| \uparrow M_\infty + |X_1|$.

(vi) Extend the argument to a two-sided equivalence: $\mathbb{P}\{\sup_i |X_i| < \infty\} = 1$ iff $\mathbb{P}\sup_i |X_i| < \infty$. Hint: $\sup_i |X_i| = \max(\sup_i X_i, \sup_i(-X_i))$.

Gaussian::P:embed.hilbert

[10] Let $\{X_t : t \in T\}$ be a centered GAUSSIAN process, defined on a probability space $(\Omega, \mathcal{F}, \mathbb{P})$. Equip T with the semi-metric $d(s,t) := \|X_s - X_t\|_2 := \big(\mathbb{P}|X_s - X_t|^2\big)^{1/2}$.

(i) Show that $X_s = X_t$ ae[$\mathbb{P}$] if $d(s,t) = 0$. Explain why this fact might cause difficulties with sample-path continuity if the inequality could not be strengthened to $X_s(\omega) = X_t(\omega)$ for all ω.

(ii) Replace T by a subset T_0 for each equivalence class $[t] := \{s \in T : d(s,t) = 0\}$ contains at most one point of T_0. Show that $\{X_t : t \in T_0\}$ can be identified with a subset of the HILBERT space $\mathbb{H} := L^2(\Omega, \mathcal{F}, \mathbb{P})$ via the map $X_t \leftrightarrow [X_t]$ in such a way that

$$\operatorname{cov}(X_s, X_t) = \langle [X_s], [X_t] \rangle$$

where $[X_t]$ denotes the equivalence class of X_t in $L^2(\Omega, \mathcal{F}, \mathbb{P})$. Remark: This representation explains why some authors regard all centered GAUSSIAN processes as subsets of an isonormal process indexed by a HILBERT space.

Gaussian::P:iso.normal

[11] Let $\mathbb{H}$ be a HILBERT space with inner product $\langle \cdot, \cdot \rangle$, and an orthonormal basis $\{e_\alpha : \alpha \in I\}$. Let $\{\eta_\alpha : \alpha \in I\}$ be a set of independent random variables, each distributed $N(0,1)$, defined on a probability space $(\Omega, \mathcal{F}, \mathbb{P})$. (Kolmogorov's existence theorem would provide such a space.)

(i) For h in $\mathbb{H}$ define $I(h) := \{\alpha \in I : \langle h, e_\alpha \rangle \ne 0\}$, a countable set for which $h = \sum_{\alpha \in I(h)} e_\alpha \langle h, e_\alpha \rangle$, a series that converges in norm. Show that the sum $Z_h := \sum_{\alpha \in I(h)} \eta_\alpha \langle h, e_\alpha \rangle$ converges in the $\mathcal{L}^2(\Omega, \mathcal{F}, \mathbb{P})$ sense. Hint: For each finite subset A of I, show that

$$\mathbb{P}\Big|\sum\nolimits_{\alpha \in A} \eta_\alpha \langle h, e_\alpha \rangle\Big|^2 = \sum\nolimits_{\alpha \in A} |\langle h, e_\alpha \rangle|^2.$$

Note: The limit is defined only up to a P-equivalence. Perhaps it would be more elegant to work with limits in $L^2(\Omega, \mathcal{F}, \mathbb{P})$.

(ii) Show that $\{Z_h : h \in \mathbb{H}\}$ is a centered GAUSSIAN process with $\operatorname{cov}(Z_g, Z_h) = \langle g, h \rangle$ for $g, h \in \mathbb{H}$. Remark: This process is sometimes called the ***isonormal process indexed by*** $\mathbb{H}$. See Dudley (2014, §2.2.1) for an insightful discussion.

6.9 Notes

Gaussian::S:Notes

I know little about the early history of the path idea, except that it appears in the work of Plackett (1954, page 353), Chover (1962), and Slepian (1962, page 483).

The most stunning fact about γ_n—the so-called isoperimetric inequality—was established independently by Borell (1975) and Sudakov and Tsirel'son

(1978), a translation from a 1974 paper in Russian. See the concise and informative book by Ledoux (2001) for more about concentration inequalities. For an exposition of a proof due to Ehrhard (1983a,b) see Pollard (2001, Section 12.5). The inequality can be rewritten more compactly as

$$\gamma_k(A^\delta) \geq \Phi\left(\Phi^{-1}(\gamma_n A) + \delta\right),$$

slightly disguising the fact that equality is achieved by halfspaces but leading towards a functional form of the inequality that was developed by Bobkov (1996, 1997). See Boucheron, Lugosi, and Massart (2000, pages 290–298, 303–307) for an insightful discussion of Bobkov's method. For elegant reformulations of the functional approach see Ledoux (1998) and Barthe and Maurey (2000).

The lower bound in Example <5> was proved by Sudakov (1973) using a different method.

The integration by parts formula, $\mathbb{P}ZF(Z) = \mathbb{P}\dot{F}(Z)$ if $Z \sim N(0,1)$, (see Problem [2]) actually characterizes the $N(0,1)$ distribution. This fact lies at the heart of Charles Stein's astounding theory of normal approximation, as exposited in his Lecture notes (Stein, 1986). See page 21 of that volume for a proof of the characterization. Undoubtedly the same trick has been used for many other purposes. See Problem [4], for example.

A huge part of the modern theory of Gaussian process (and stochastic processes in general) grew from the ideas of Xavier Fernique. For example, see Fernique (1975, 1983) or search for his name at the French digital mathematics library, http://www.numdam.org/. See also Ledoux and Talagrand (1991, page 87–88) and Dudley (2014, Chapter 2 Notes) for more about the history of comparison methods and where credit is due.

My introduction to the stochastic integration method described in Section 6.3.1 came from Adler (1990, Section 2.1), who attributed the method to Maurey and Pisier, citing lectures by Pisier (1986). In those lectures Pisier (page 180) had provided a sketch of Maurey's argument, prefaced by the comment "B. Maurey found a proof of theorem 2.1 with the best constant ... His proof uses stochastic integrals and apparently does not extend to the setting of theorem 2.2." Only later did I notice that Ledoux (2001, page 45) gave credit for the stochastic calculus proof to Cirel'son, Ibragimov, and Sudakov (1976), with the comment that their paper "was unfortunately ignored for a long time". That 1976 paper had attributed the concentration result (via the isoperimetric inequality) to the 1974 Russian version of Sudakov and Tsirel'son (1978), with the remark (page 25) that the 1974 proof was "not a purely probabilistic one ... But sometimes the following assertion, provable in a purely probabilistic way, can replace it". They then gave a stochastic calculus argument that started in the same way as the proof in Section 6.3.1, with the identification of the martingale M, but then they invoked a time-change argument (Chung and Williams, 2014, Section 9.3) to represent M as $\gamma_n f + W([M]_t)$ for a new Brownian motion W. In my notation, the rest

of their proof used the fact (inequality <26>) that $[M]_1 \le \kappa^2$ to deduce that

$$\begin{aligned}\mathbb{P}\{M_1 \ge \gamma_n f + \kappa x\} &\le \mathbb{P}\{\sup_{0\le s\le \kappa^2} W_s \ge \kappa x\} \\ &= 2\mathbb{P}\{W_{\kappa^2}/\kappa \ge x\} \qquad \text{reflection principle} \\ &= 2\bar{\Phi}(x).\end{aligned}$$

That is, they actually established a result sharper than the $e^{-x^2/2}$ bound.

See Davidson and Szarek (2001, Section II) and Vershynin (2018) for discussion of ways that concentration and comparison inequalities enter the theory of random matrices.

References

Adler90gauss

Adler, R. J. (1990). *An Introduction to Continuity, Extrema, and Related Topics for General Gaussian Processes*, Volume 12 of *Lecture Notes–Monograph series*. Hayward, CA: Institute of Mathematical Statistics.

BartheMaurey2000AIHP

Barthe, F. and B. Maurey (2000). Some remarks on isoperimetry of Gaussian type. *Annales de l'Institut Henri Poincaré, Probability and Statistics 36*(4), 419–434.

Bobkov1996JFA

Bobkov, S. (1996). A functional form of the isoperimetric inequality for the Gaussian measure. *Journal of Functional Analysis 135*, 39–49.

Bobkov1997AnnProb

Bobkov, S. G. (1997). An isoperimetric inequality on the discrete cube, and an elementary proof of the isoperimetric inequality in Gauss space. *Annals of Probability 25*(1), 206–214.

Borell:75

Borell, C. (1975). The Brunn-Minkowski inequality in Gauss space. *Inventiones Mathematicae 30*, 207–216.

cheronLugosimassart2000ras

Boucheron, S., G. Lugosi, and P. Massart (2000). A sharp concentration inequality with applications. *Random Structures and Algorithms 16*, 277–292.

Chatterjee2005Fernique

Chatterjee, S. (2005). An error bound in the Sudakov-Fernique inequality. Technical report, arXiv:math/0510424.

Chover1962Duke

Chover, J. (1962, March). Certain convexity conditions on matrices with applications to Gaussian processes. *Duke Mathematical Journal 29*(1), 141–150.

ChungWilliams2014book

Chung, K. L. and R. J. Williams (2014). *Introduction to Stochastic Integration* (Second ed.). Boston: Birkhäuser.

lIbragimovSudakov1976norms

Cirel'son, B., I. Ibragimov, and V. Sudakov (1976). Norms of Gaussian sample functions. In *Proceedings of the Third Japan-USSR Symposium on Probability Theory*, Volume 550 of *Springer Lecture Notes in Mathematics*, pp. 20–41. Springer.

DavidsonSzarek2001local Davidson, K. R. and S. J. Szarek (2001). Local operator theory, random matrices and Banach spaces. *Handbook of the geometry of Banach spaces 1*(317-366), 131. (Available in preprint form at http://case.edu/artsci/math/szarek/publications.html).

Dudley2014UCLT Dudley, R. M. (2014). *Uniform Central Limit Theorems* (2nd ed.), Volume 142 of *Cambridge studies in advanced mathematics*. Cambridge University Press. (First edition, 1999).

Dvoretzky1960theorem Dvoretzky, A. (1961). Some results on convex bodies and Banach spaces. In *Proceedings of the International Symposium on linear spaces (held at Hebrew University of Jerusalem, July 5-12, 1960)*, pp. 123–160. Academic Press; New York, Pergamon Press.

Ehrhard83MathScand Ehrhard, A. (1983a). Symétrisation dans l'espace de Gauss. *Mathematica Scandinavica 53*, 281–301.

Ehrhard83slnm Ehrhard, A. (1983b). Un principe de symétrisation dans les espaces de Gauss. *Springer Lecture Notes in Mathematics 990*, 92–101.

Fernique75StFlour Fernique, X. (1975). Regularité des trajectoires des fonctions aléatoires gaussiennes. *Springer Lecture Notes in Mathematics 480*, 1–97. École d'Été de Probabilités de Saint-Flour IV, 1974.

Fernique83StFlour Fernique, X. (1983). Regularité de fonctions aléatoires non gaussiennes. *Springer Lecture Notes in Mathematics 976*, 1–74. École d'Été de Probabilités de St-Flour XI, 1981.

Gordon1985IJM Gordon, Y. (1985). Some inequalities for Gaussian processes and applications. *Israel Journal of Mathematics 50*(4), 265–289.

Gordon1988AnnProb Gordon, Y. (1988). Gaussian processes and almost spherical sections of convex bodies. *The Annals of Probability 16*(1), 180–188.

Gordon1910SLNM Gordon, Y. (2004-2005). A note on an observation of G. Schechtman. *Springer Lecture Notes in Mathematics 1910*, 127–132. Geometric Aspects of Functional Analysis, Israel Seminar.

Isserlis1918Biom Isserlis, L. (1918). On a formula for the product-moment coefficient of any order of a normal frequency distribution in any number of variables. *Biometrika 12*(1/2), 134–139.

Janson1997Gaussian Janson, S. (1997). *Gaussian Hilbert Spaces*. Cambridge Tracts in Mathematics. Cambridge University Press.

Kahane1986IsraelJM Kahane, J.-P. (1986). Une inégalité du type de Slepian et Gordon sur les processus gaussiens. *Israel Journal of Mathematics 55*(1), 109–110.

KaratzasShreve88book Karatzas, I. and S. E. Shreve (1988). *Brownian Motion and Stochastic Calculus* (First ed.). New York: Springer-Verlag.

Ledoux1998PP Ledoux, M. (1998). A short proof of the Gaussian isoperimetric inequality. In *High Dimensional Probability*, Volume 43 of *Progress in Probability*, pp. 229–232. Springer.

Ledoux01conc Ledoux, M. (2001). *The Concentration of Measure Phenomenon*, Volume 89 of *Mathematical Surveys and Monographs*. American Mathematical Society.

LedouxTalagrand91book Ledoux, M. and M. Talagrand (1991). *Probability in Banach Spaces: Isoperimetry and Processes*. New York: Springer.

MarshallOlkinArnold2011 Marshall, A. W., I. Olkin, and B. C. Arnold (2011). *Inequalities: Theory of Majorization and its Applications* (Second ed.), Volume 143 of *Series in Statistics*. Springer.

Pisier1986SLNM1206 Pisier, G. (1986). Probabilistic methods in the geometry of Banach spaces. In G. Letta and M. Pratelli (Eds.), *Probability and Analysis*, Volume 1206 of *Lecture Notes in Mathematics*, pp. 167–241. Springer Berlin Heidelberg.

Plackett1954Biometrika Plackett, R. L. (1954). A reduction formula for normal multivariate integrals. *Biometrika 41*(3/4), 351–360.

PollardUGMTP Pollard, D. (2001). *A User's Guide to Measure Theoretic Probability*. Cambridge University Press.

Slepian62 Slepian, D. (1962). The one-sided barrier problem for Gaussian noise. *Bell System Technical Journal 41*, 463–501.

Stein86ims Stein, C. (1986). *Approximate Computation of Expectations*, Volume 7 of *Lecture Notes–Monograph series*. Institute of Mathematical Statistics.

Sudakov1973minoration Sudakov, V. (1973). A remark on the criterion of continuity of Gaussian sample function. In *Proceedings of the Second Japan-USSR Symposium on Probability Theory, Kyoto, August 2–9, 1972*, Volume 330 of *Springer Lecture Notes in Mathematics*, pp. 444–454.

SudakovTsirelson1978JSM Sudakov, V. N. and B. S. Tsirel'son (1978). Extremal properties of half-spaces for spherically invariant measures. *Journal of Soviet Mathematics 9*, 419–434. (Translated from Zapiski Nauchnykh Seminarov Leningradskogo Otdeleniya Matematicheskogo instituta im. V. A. Steklova AN SSSR, Vo!. 41, pp. 14-24, 1974.).

Talagrand03spin Talagrand, M. (2003). *Spin Glasses: a Challenge to Mathematicians*, Volume 46 of *Ergbnisse der Mathematik und ihrer Grenzgebiete*. New York: Springer-Verlag.

Talagrand2011spin1 Talagrand, M. (2011). *Mean Field Models for Spin Glasses: Volume I: Basic Examples*, Volume 54 of *Ergebnisse der Mathematik und ihrer Grenzgebiete*. Springer Science & Business Media.

Vershynin2020HDP Vershynin, R. (2018). *High-Dimensional Probability: An Introduction with Applications in Data Science*. Cambridge University Press.

Vitale2000PAMS

Vitale, R. A. (2000). Some comparisons for Gaussian processes. *Proceedings of the American Mathematical Society 128*(10), 3043–3046.

