Lecture 1

The Canonical Gaussian Measure on $\mathbb{R}^n$

1. Introduction

The main goal of this course is to study "Gaussian measures." The simplest example of a Gaussian measure is the *canonical Gaussian measure* $\mathbb{P}_n$ on $\mathbb{R}^n$ where $n \geqslant 1$ is an arbitrary integer. The measure $\mathbb{P}_n$ is one of the main objects of study in this course, and is defined as

$$\mathbb{P}_n(A) := \int_A \gamma_n(x)\,\mathrm{d}x \qquad \text{for all Borel sets } A \subseteq \mathbb{R}^n,$$

where

$$\gamma_n(x) := \frac{\mathrm{e}^{-\|x\|^2/2}}{(2\pi)^{n/2}} \qquad [x \in \mathbb{R}^n]. \tag{1.1}$$

gamma_n

The function γ_1 describes the famous "bell curve." In dimensions $n \geqslant 1$, the curve of γ_n looks like a suitable "rotation" of the curve of γ_1.

We frequently drop the subscript n from $\mathbb{P}_n$ when it is clear which dimension we are in. We also choose and fix the integer $n \geqslant 1$, tacitly, consider the probability space $(\Omega\,,\mathcal{F}\,,\mathbb{P})$ where we have dropped the subscript n from $\mathbb{P}_n$ [as we will some times], and set

$$\Omega := \mathbb{R}^n, \quad \text{and} \quad \mathcal{F} := \mathcal{B}(\mathbb{R}^n).$$

Throughout we designate by Z the random vector,

$$Z_j(x) := x_j \qquad \text{for all } x \in \mathbb{R}^n \text{ and } 1 \leqslant j \leqslant n. \tag{1.2}$$

Z

Then $Z := (Z_1, \ldots, Z_n)$ is a random vector of n i.i.d. standard normal random variables on our probability space. In particular,

$$\mathbb{P}_n(A) = \mathbb{P}\{Z \in A\} \qquad \text{for all Borel sets } A \subseteq \mathbb{R}^n.$$

One of the elementary, though important, properties of the measure $\mathbb{P}_n$ is that its "tails" are vanishingly small.

lem:tails **Lemma 1.1.** *As $t \to \infty$,*

$$\mathbb{P}\left\{x \in \mathbb{R}^n : \|x\| > t\right\} = \frac{2 + o(1)}{2^{n/2}\Gamma(n/2)} t^{n-2} e^{-t^2/2}.$$

Proof. Since

$$S_n := \|Z\|^2 = \sum_{i=1}^n Z_i^2 \tag{1.3}$$

S_n

has a χ_n^2 distribution,

$$\mathbb{P}\{x \in \mathbb{R}^n : \|x\| > t\} = \mathbb{P}\{S_n > t^2\} = \frac{1}{2^{n/2}\Gamma(n/2)} \int_{t^2}^{\infty} x^{(n-2)/2} e^{-x/2}\, dx,$$

for all $t \geqslant 0$. Now apply l'Hôpital's rule of calculus. $\square$

The following large-deviations estimate is one of the consequences of Lemma 1.1:

$$\lim_{t\to\infty} \frac{1}{t^2} \log \mathbb{P}\left\{x \in \mathbb{R}^n : \|x\| > t\right\} = -\tfrac{1}{2}. \tag{1.4}$$

eq:LD

Of course, this is a weaker statement than Lemma 1.1. But it has the advantage of being "dimension independent." Dimension independence properties play a prominent role in the analysis of Gaussian measures. Here, for example, (1.4) teaches us that the tails of $\mathbb{P}_n$ behave roughly as do the tails of $\mathbb{P}_1$ for all $n \geqslant 1$.

Still, many of the more interesting properties of $\mathbb{P}_n$ are radically different from those of $\mathbb{P}_1$ when n is large. In low dimensions—say $1 \leqslant n \leqslant 3$—one can visualize the probability density function γ_n from (1.1). Based on that, or other methods, one knows that in low dimensions most of the mass of $\mathbb{P}_n$ lies near the origin. For example, an inspection of the standard normal table reveals that more than half of the total mass of $\mathbb{P}_1$ is within one unit of the origin; in fact, $\mathbb{P}_1[-1, 1] \approx 0.68268$.

In higher dimensions, however, the structure of $\mathbb{P}_n$ can be quite different. Recall the random variable S_n from (1.3), and apply Khintchine's weak law of large numbers XXX to see that S_n/n converges in probability to one, as $n \to \infty$.[1] This is another way to say that for every

[1] In the present setting, it does not make sense to discuss almost-sure convergence since the underlying probability space is $(\mathbb{R}^n, \mathscr{B}(\mathbb{R}^n), \mathbb{P}_n)$.

$\varepsilon > 0$,

$$\lim_{n\to\infty} \mathbb{P}\left\{x \in \mathbb{R}^n : (1-\varepsilon)n^{1/2} \leqslant \|x\| \leqslant (1+\varepsilon)n^{1/2}\right\} = 1. \tag{1.5}$$

ex:CoM

The proof is short and can reproduced right here: Let $\mathbb{E}$ denote the expectation operator for $\mathbb{P} := \mathbb{P}_n$, Var := the variance, etc. Since $\mathbb{E}\, S_n = \operatorname{Var} S_n = n$, Chebyshev's inequality yields $\mathbb{P}\{|S_n - \mathbb{E}\, S_n| > \varepsilon n\} \leqslant \varepsilon^{-2}n^{-1}$. Equivalently,

$$\mathbb{P}\left\{x \in \mathbb{R}^n : (1-\varepsilon)n^{1/2} \leqslant \|x\| \leqslant (1+\varepsilon)n^{1/2}\right\} \geqslant 1 - \frac{1}{n\varepsilon^2}. \tag{1.6}$$

WLLN

Thus we see that, when n is large, the measure $\mathbb{P}_n$ concentrates much of its total mass near the boundary of the centered ball of radius $n^{1/2}$, very far from the origin. A more careful examination shows that, in fact, very little of the total mass of $\mathbb{P}_n$ is elsewhere when n is large. The following theorem makes this statement more precise. Theorem 1.2 is a simple example of a remarkable property of Gaussian measures that is known commonly as *concentration of measure* XXX. We will discuss this topic in more detail in due time.

th:CoM:n **Theorem 1.2.** *For every $\varepsilon > 0$,*

$$\mathbb{P}\left\{x \in \mathbb{R}^n : (1-\varepsilon)n^{1/2} \leqslant \|x\| \leqslant (1+\varepsilon)n^{1/2}\right\} \geqslant 1 - 2\mathrm{e}^{-n\varepsilon}. \tag{1.7}$$

CoM:n

Theorem 1.2 does not merely improve the crude bound (1.6). Rather, it describes an entirely new phenomenon in high dimensions. To wit, suppose we are working in dimension $n = 500$. When $\varepsilon = 0$, the left-hand side of (1.7) is equal to 0. But if we increase ε slightly, say to $\varepsilon = 0.01$, then the left-hand side of (1.7) increases to a probability $\geqslant 0.986$ (!). By comparison, (1.6) reports a silly bound in this case.

Proof. From now on, we let $\mathbb{E} := \mathbb{E}_n$ denote the expectation operator for $\mathbb{P} := \mathbb{P}_n$. That is,

$$\mathbb{E}(f) := \mathbb{E}_n(f) := \int f\, \mathrm{d}\mathbb{P}_n = \int f\, \mathrm{d}\mathbb{P} \qquad \text{for all } f \in L^1(\mathbb{P}_n).$$

Since $S_n := \sum_{i=1}^n Z_i^2$ has a χ_n^2 distribution,

$$\mathbb{E}\, \mathrm{e}^{\lambda S_n} = (1-2\lambda)^{-n/2} \quad \text{for } -\infty < \lambda < 1/2, \tag{1.8}$$

mgf:chi2

and $\mathbb{E}\exp(\lambda S_n) = \infty$ when $\lambda \geqslant 1/2$.

We use the preceding as follows: For all $t > 0$ and $\lambda \in (0\,, 1/2)$,

$$\mathbb{P}\left\{x : \|x\| > n^{1/2}t\right\} = \mathbb{P}\{S_n > nt^2\} = \mathbb{P}\left\{\mathrm{e}^{\lambda S_n} > \mathrm{e}^{\lambda n t^2}\right\} \leqslant (1-2\lambda)^{-n/2}\mathrm{e}^{-\lambda n t^2},$$

thanks to Chebyshev's inequality. Since the left-hand side is independent of $\lambda \in (0\,,1/2)$, we may optimize the right-hand side over $\lambda \in (0\,,1/2)$ to find that

$$\begin{aligned}\mathbb{P}\left\{x:\, \|x\| > n^{1/2}t\right\} &\leqslant \exp\left\{-n \sup_{0<\lambda<1/2}\left[\lambda t^2 + \frac{1}{2}\log(1-2\lambda)\right]\right\}\\ &= \exp\left\{-\frac{n}{2}\left[t^2 - 1 - 2\log t\right]\right\}.\end{aligned}$$

Taylor expansion yields $\log t < t-1-\frac{1}{2}(t-1)^2$ when $t>1$. This and the previous inequality together yield

$$\mathbb{P}\left\{x:\, \|x\| > n^{1/2}t\right\} \leqslant \mathrm{e}^{-nt(t-1)} < \mathrm{e}^{-n|t-1|}. \tag{1.9}$$

BooBooBound

When $t=1$, the same inequality holds vacuously. Therefore, it suffices to consider the case that $t<1$.

In that case, we argue similarly and write

$$\begin{aligned}\mathbb{P}\left\{x:\, \|x\| < n^{1/2}t\right\} &= \mathbb{P}\left\{\mathrm{e}^{-\lambda S_n} > \mathrm{e}^{-\lambda n t^2}\right\} \qquad [\text{for all } \lambda>0]\\ &\leqslant \exp\left\{-n\sup_{\lambda>0}\left[-\lambda t^2 + \frac{1}{2}\log(1+2\lambda)\right]\right\}\\ &= \exp\left\{-\frac{n}{2}\left[1-t^2-2\log t\right]\right\}.\end{aligned}$$

Since $t<1$, a Taylor expansion yields $-2\log t > 2(1-t)+(1-t)^2$, whence

$$\mathbb{P}\left\{x:\, \|x\| < n^{1/2}t\right\} \leqslant \exp\left\{-\frac{n}{2}\left[1-t^2-2\log t\right]\right\} \leqslant \mathrm{e}^{-n(1-t)}.$$

This estimate and (1.9) together complete the proof. □

The preceding discussion shows that when n is large, $\mathbb{P}\{\|Z\| \approx n^{1/2}\}$ is extremely close to one. One can see from another appeal to the weak law of large numbers that for all $\varepsilon>0$,

$$\mathbb{P}\left\{(1-\varepsilon)\mu_p n^{1/p} \leqslant \|Z\|_p \leqslant (1+\varepsilon)\mu_p n^{1/p}\right\} \to 1 \qquad \text{as } n\to\infty,$$

for all $p\in[1\,,\infty)$, where $\mu_p := \mathbb{E}(|Z_1|^p)$ and $\|\cdot\|_p$ [temporarily] denotes the ℓ^p-norm on $\mathbb{R}^n$; that is,

$$\|x\|_p := \begin{cases}\left(\sum_{i=1}^n |x_i|^p\right)^{1/p} & \text{if } 1\leqslant p<\infty,\\ \max_{1\leqslant i\leqslant n}|x_i| & \text{if } p=\infty,\end{cases}$$

for all $x\in\mathbb{R}^n$. These results suggest that the n-dimensional Gauss space $(\mathbb{R}^n, \mathscr{B}(\mathbb{R}^n), \mathbb{P}_n)$ has unexpected geometry when $n \gg 1$.

Interestingly enough, the case $p=\infty$ is different still. The following might help us anticipate which ℓ^∞-balls might carry most of the measure $\mathbb{P}_n$ when $n\gg 1$.

pr:max **Proposition 1.3.** *Let* $M_n := \max_{1\leqslant i\leqslant n}|Z_i|$ *or* $:= \max_{1\leqslant i\leqslant n} Z_i$. *Then,*

$$\lim_{n\to\infty}\frac{1}{\sqrt{2\log n}}\,\mathbb{E}(M_n) = 1.$$

It is possible to compute $\mathbb{E}(M_n)$ directly, but the preceding bound turns out to be more informative for our ends, and good enough for our present needs.

Proof. For all $t > 0$,

$$\mathbb{P}\left\{M_n > t\right\} \leqslant 1 \wedge n\mathbb{P}\{|Z_1| > t\} \leqslant 1 \wedge 2n\mathrm{e}^{-t^2/2}.$$

Now choose and fix some $\varepsilon \in (0\,,1)$ and write, using this bound,

$$\begin{aligned}
\mathbb{E}\left(\max_{1\leqslant i\leqslant n}|Z_i|\right) &= \int_0^\infty \mathbb{P}\left\{\max_{1\leqslant i\leqslant n}|Z_i| > t\right\}\mathrm{d}t = \int_0^{2(1+\varepsilon)\log n} + \int_{2(1+\varepsilon)\log n}^\infty\\
&\leqslant \sqrt{2(1+\varepsilon)\log n} + 2n\int_{\sqrt{2(1+\varepsilon)\log n}}^\infty \mathrm{e}^{-t^2/2}\,\mathrm{d}t\\
&= \sqrt{2(1+\varepsilon)\log n} + o\left(n^{-\varepsilon}\right).
\end{aligned}$$

Thus, we have $\mathbb{E}(M_n) \leqslant (1+o(1))\sqrt{2\log n}$, which implies half of the proposition. For the other half, we use Lemma 1.1 to see that,

$$\begin{aligned}
\mathbb{P}\left\{\max_{1\leqslant i\leqslant n} Z_i \leqslant \sqrt{2(1-\varepsilon)\log n}\right\} &= \left(1 - \frac{1}{\sqrt{2\pi}}\int_{\sqrt{2(1-\varepsilon)\log n}}^\infty \mathrm{e}^{-t^2/2}\,\mathrm{d}t\right)^n\\
&= \left(1 - \frac{1+o(1)}{\sqrt{2\pi\log n}}\,n^{-1+\varepsilon}\right)^n = o(1),
\end{aligned}$$

for example because $1 - x \leqslant \exp(-x)$ for all $x \in \mathbb{R}$. Therefore,

$$\begin{aligned}
\mathbb{E}\left(\max_{1\leqslant i\leqslant n} Z_i;\ \max_{1\leqslant i\leqslant n} Z_i \geqslant 0\right) &\geqslant \mathbb{E}\left(\max_{1\leqslant i\leqslant n} Z_i;\ \max_{1\leqslant i\leqslant n} Z_i > \sqrt{2(1-\varepsilon)\log n}\right)\\
&\geqslant (1+o(1))\sqrt{2(1-\varepsilon)\log n}.
\end{aligned}$$

On the other hand,

$$\begin{aligned}
\left|\mathbb{E}\left(\max_{1\leqslant i\leqslant n} Z_i;\ \max_{1\leqslant i\leqslant n} Z_i < 0\right)\right| &\leqslant \sum_{i=1}^n \mathbb{E}\left(|Z_i|;\ \max_{1\leqslant i\leqslant n} Z_i < 0\right)\\
&\leqslant n\sqrt{\mathbb{P}\left\{\max_{1\leqslant i\leqslant n} Z_i < 0\right\}} = n2^{-n/2} = o(1),
\end{aligned}$$

as $n \to \infty$, thanks to the Cauchy–Schwarz inequality. The preceding two displays together imply that $\mathbb{E}\,M_n \geqslant (1+o(1))\sqrt{2(1-\varepsilon)\log n}$ as $n\to\infty$ for every $\varepsilon \in (0\,,1)$, and prove the remaining half of the proposition. $\square$

2. The Projective CLT

The characteristic function of the random vector Z is

$$\widehat{\mathbb{P}}_n(w) := \mathbb{E}\, e^{iw\cdot Z} = e^{-\|w\|^2/2} \qquad [w \in \mathbb{R}^n].$$

If M denotes any $n \times n$ orthogonal matrix, then

$$\mathbb{E}\, e^{iw\cdot MZ} = e^{-\|Mw\|^2/2} = e^{-\|w\|^2/2}.$$

Therefore, the distribution of MZ is $\mathbb{P}_n$ as well. In particular, the normalized random vectors $Z/\|Z\|$ and $MZ/\|MZ\|$ have the same distribution. Since the uniform distribution on the n-sphere $\mathbb{S}^{n-1} := \{x \in \mathbb{R}^n : \|x\| = 1\}$ is the unique probability measure on $\mathbb{S}^{n-1}$ that is invariant under orthogonal transformations, it follows that $Z/\|Z\|$ is distributed uniformly on $\mathbb{S}^{n-1}$. By the weak law of large numbers, $\|Z\|/\sqrt{n}$ converges to 1 in probability as $n \to \infty$. Therefore, for all fixed $k \geqslant 1$, the random vector $\sqrt{n}(Z_1 \ldots, Z_k)/\|Z\|$ converges weakly to a k-vector of i.i.d. $\mathrm{N}(0\,,1)$ random variables as $n \to \infty$.

In other words, we have proved the following.

pr:ProjCLT **Proposition 2.1.** *Choose and fix an integer $k \geqslant 1$ and a bounded and continuous function $f : \mathbb{R}^k \to \mathbb{R}$. Let μ_n denote the uniform measure on $\sqrt{n}\mathbb{S}^{n-1}$. Then,*

$$\lim_{n\to\infty} \int_{\sqrt{n}\mathbb{S}^{n-1}} f(x_1\,,\ldots,x_k)\,\mu_n(dx_1\cdots dx_n) = \int_{\mathbb{R}^k} f\, d\mathbb{P}_k.$$

In other words, this means roughly that the conditional distribution of Z, given that $(1-\varepsilon)n^{1/2} \leqslant \|Z\| \leqslant (1+\varepsilon)n^{1/2}$, is close to the uniform distribution on $\sqrt{n}\mathbb{S}^{n-1}$ as $n \to \infty$ and $\varepsilon \to 0$. It is in fact possible to make this statement precise using a Bayes-type argument. But we will not do so here. Instead we end this section by noting the following: Since the probability of the conditioning event $\{(1-\varepsilon)n^{1/2} \leqslant \|Z\| \leqslant (1+\varepsilon)n^{1/2}\}$ is almost one—see (1.5)—we can see that Proposition 2.1 is a way of stating that the canonical Gaussian measure on $\mathbb{R}^n$ is very close to the uniform distribution on $\sqrt{n}\mathbb{S}^{n-1}$ when $n \gg 1$.

Proposition 2.1 has other uses as well.

3. Anderson's Shifted-Ball Inequality

One of the deep properties of $\mathbb{P}_n$ is that it is "unimodal," in a certain sense that we will describe soon. That result hinges on a theorem of T. W. Anderson XXX in convex analysis. Anderson's theorem has many deep applications in probability theory, as well as multivariate statistics, which originally was one of the main motivations for Anderson's work.

We will see some applications later on. For now we contend ourselves with a statement and proof.

In order to state Anderson's theorem efficiently, let us recall a few basic notions from [undergraduate] real analysis.

Recall that a set $E \subset \mathbb{R}^n$ is *convex* if $\lambda x + (1-\lambda)y \in E$ for all $x, y \in E$ and $\lambda \in [0\,,1]$. You should check that E is convex if and only if

$$E = \lambda E + (1-\lambda)E \qquad \text{for all } \lambda \in [0\,,1],$$

where for all $\alpha, \beta \in \mathbb{R}$ and $A, B \subseteq \mathbb{R}^n$,

$$\alpha A + \beta B := \{\alpha x + \beta y : x \in A\,, y \in B\}.$$

pr:convex:meas **Proposition 3.1.** *Every convex set $E \subset \mathbb{R}^n$ is Lebesgue measurable.*

Remark 3.2. Suppose $n \geqslant 2$ and $E = B(0\,,1) \cup F$, where $B(0\,,1)$ is the usual notation for the Euclidean ball of radius one about $0 \in \mathbb{R}^n$, and $F \subset \partial B(0\,,1)$. Then E is always convex. But E is not Borel measurable unless F is. Still, E is always Lebesgue measurable, in this case because F is Lebesgue null in $\mathbb{R}^n$.

Proof. We will prove *Claim. Every bounded convex set is measurable.*

This does the job since whenever E is convex and $n \geqslant 1$, $E \cap B(0\,,n)$ is a bounded convex set, which is measurable according to the Claim. Therefore, $E = \cup_{n=1}^{\infty} E \cap B(0\,,n)$ is also measurable.

Since ∂E is closed, it is measurable. We will prove that $|\partial E| = 0$. This shows that the difference between E and the open set E^0 is null, whence E is Lebesgue measurable. There are many proofs of this fact. Here is an elegant one, due to Lang XXX.

Define

$$\mathscr{M} := \left\{ B \in \mathscr{B}(\mathbb{R}^n) : \ |B \cap \partial E| \leqslant \left(1 - 3^{-n}\right) |B| \right\}.$$

Then $\mathscr{M}$ is clearly a monotone class, and closed under disjoint unions. We plan to prove that every upright *rectangle*, that is every nonempty set of of the form $\prod_{i=1}^n (a_i\,, b_i]$, is in $\mathscr{M}$. If so, then the monotone class theorem implies that $\mathscr{M} = \mathscr{B}(\mathbb{R}^n)$. That would show, in turn, that $|\partial E| = |\partial E \cap \partial E| \leqslant (1 - 3^{-n})|\partial E|$, which proves the claim.

Choose and fix a *rectangle* $B := \prod_{i=1}^n (a_i\,, b_i]$, where $a_i < b_i$ for all $1 \leqslant i \leqslant n$. Subdivide each 1-D interval $(a_i\,, b_i]$ into 3 equal-sized parts: $(a_i\,, a_i + r_i]$, $(a_i + r_i\,, a_i + 2r_i]$, and $(a_i + 2r_i\,, a_i + 3r_i]$ where $r_i := (b_i - a_i)/3$. We can write B as a disjoint union of 3^n equal-sized rectangles, each of which has the form $\prod_{i=1}^n (a_i + c_i r_i\,, a_i + (1 + c_i] r_i]$ where $c_i \in \{0\,,1\,,2\}$. Call these rectangles $B_1, \ldots, B_{3^n}$. Direct inspection shows that there must exist an integer $1 \leqslant \ell \leqslant 3^n$ such that $\partial E \cap B_\ell = \varnothing$. For otherwise the middle rectangle $\prod_{i=1}^n (a_i + r_i\,, a_i + 2r_i]$ would have to lie entirely in E^0 and intersect ∂E at the same time; this would contradict the existence of a supporting hyperplane at every point of ∂E. Let us fix the integer ℓ alluded to here.

Since the B_j's are translates of one another they have the same measure. Therefore,

$$|B \cap \partial E| \leqslant \sum_{\substack{1 \leqslant j \leqslant 3^n \\ j \neq \ell}} |B_j| = |B| - |B_\ell| = \left(1 - 3^{-n}\right)|B|.$$

This proves that every rectangle $B \in \mathscr{M}$, and completes the proof. □

We will also recall two standard definitions.

Definition 3.3. A set $E \in \mathbb{R}^n$ is *symmetric* if $E = -E$.

Definition 3.4. If $f : \mathbb{R}^n \to \mathbb{R}$ is a measurable function, then its *level set at level* $r \in \mathbb{R}$ is defined as $f^{-1}[r\,,\infty) := \{x \in \mathbb{R}^n : f(x) \geqslant r\} := \{f \geqslant r\}$.

The main result of this section is Anderson's inequality XXX, which I mention next.

th:Anderson

Theorem 3.5 (Anderson's inequality). *Let $f \in L^1(\mathbb{R}^n)$ be a non-negative symmetric function that has convex level sets. Then,*

$$\int_E f(x - \lambda y)\,\mathrm{d}x \geqslant \int_E f(x - y)\,\mathrm{d}x,$$

for all symmetric convex sets $E \subset \mathbb{R}^n$, every $y \in \mathbb{R}^n$, and all $\lambda \in [0\,,1]$.

Remark 3.6. It follows immediately from Theorem 3.5 that the map $\lambda \mapsto \int_E f(x - \lambda y)\,\mathrm{d}x$ is non increasing for $\lambda \in [0\,,1]$.

The proof will take up the remainder of this chapter. For now, let us remark briefly on how the Anderson inequality might be used to analyse the Gaussian measure $\mathbb{P}_n$.

Recall γ_n from (1.1), and note that for every $r > 0$, the level set

$$\gamma_n^{-1}[r\,,\infty) = \left\{x \in \mathbb{R}^n : \|x\| \leqslant \sqrt{2\log r + n\log(2\pi)}\right\}$$

is a closed ball, whence convex and symmetric. Therefore, we can apply Anderson's inequality with $\lambda = 0$ to see the following immediate corollary.

co:Anderson

Corollary 3.7. *For all symmetric convex sets $E \subset \mathbb{R}^n$, $0 \leqslant \lambda \leqslant 1$, and $y \in \mathbb{R}^n$, $\mathbb{P}_n(E + \lambda y) \geqslant \mathbb{P}_n(E + y)$. In particular,*

$$\mathbb{P}_n(E + y) \leqslant \mathbb{P}_n(E).$$

It is important to emphasize the remarkable fact that Corollary 3.7 is a "dimension-free theorem." Here is a typical consequence: $\mathbb{P}\{\|Z-a\| \leqslant r\}$ is maximized at $a = 0$ for all $r > 0$. For this reason, Corollary 3.7 is sometimes referred to as a "shifted-ball inequality."

One can easily generalize the preceding example with a little extra effort. Let us first note that if M is an $n \times n$ positive-semidefinite matrix, then $E := \{x \in \mathbb{R}^n : x \cdot Mx \leqslant r\}$ is a symmetric convex set for every real number $r > 0$. Equivalently, E is the event—in our probability space $(\mathbb{R}^n, \mathcal{B}(\mathbb{R}^n), \mathbb{P}_n)$—that $Z \cdot MZ \leqslant r$. Therefore, Anderson's shifted-ball inequality implies that

$$\mathbb{P}_n\{(Z-\mu)\cdot M(Z-\mu) \leqslant r\} \leqslant \mathbb{P}_n\{Z \cdot MZ \leqslant r\} \quad \forall r > 0 \text{ and } \mu \in \mathbb{R}^n.$$

This inequality has applications in multivariate statistics, particularly to the analysis of "Hotelling's T^2 statistic." We will see other interesting examples later on.

§1. Part 1. The Brunn–Minkowski Inequality. The Brunn–Minkowski inequality XXX is a classical result from convex analysis, and has profound connections to several other areas of research. In this subsection we state and prove the Brunn–Minkowski inequality.

It is easy to see that if A and B are compact, then so is $A+B$, since the latter is clearly bounded and closed. In particular, $A + B$ is measurable.

The Brunn–Minkowski inequality relates the Lebesgue measure of the Minkowski sum $A + B$ to those of A and B. Let $|\cdots|$ denote the Lebesgue measure on $\mathbb{R}^n$. Then we have the following.

Theorem 3.8 (The Brunn–Minkowski Inequality). *For all compact sets $A, B \subset \mathbb{R}^n$,*

$$|A+B|^{1/n} \geqslant |A|^{1/n} + |B|^{1/n}.$$

We can replace A by λA and B by $(1-\lambda)B$, where $0 \leqslant \lambda \leqslant 1$, and recast the Brunn–Minkowski inequality in the following equivalent form:

$$|\lambda A + (1-\lambda)B|^{1/n} \geqslant \lambda |A|^{1/n} + (1-\lambda)|B|^{1/n},$$

for all compact sets $A, B \subset \mathbb{R}^n$ and $\lambda \in [0,1]$. This formulation suggests the existence of deeper connections to convex analysis because if A and B are convex sets, then so is $\lambda A + (1-\lambda)B$ for all $\lambda \in [0,1]$.

Proof. The proof is elementary but tricky. In order to clarify the underlying ideas, we will divide it up into 3 small steps.

Step 1. Say that $K \subset \mathbb{R}^n$ is a *rectangle* when K has the form,

$$K = [x_1, x_1 + k_1] \times \cdots \times [x_n, x_n + k_n],$$

for some $x := (x_1, \ldots, x_n) \in \mathbb{R}^n$ and $k_1, \ldots, k_n > 0$. We refer to the point x as the *lower corner* of K, and $k := (k_1, \ldots, k_n)$ as the *length* of K.

In this first step we verify the theorem in the case that A and B are rectangles with respective lengths a and b. In this case, we can see that $A + B$ is an rectangle of sidelength $a + b$. The Brunn–Minkowski inequality, in this case, follows from the following application of Jensen's inequality [the *arithmetic–geometric mean inequality*]:

$$\begin{aligned}\left(\prod_{i=1}^n \frac{a_i}{a_i + b_i}\right)^{1/n} + \left(\prod_{i=1}^n \frac{b_i}{a_i + b_i}\right)^{1/n} &\leqslant \frac{1}{n}\sum_{i=1}^n \left(\frac{a_i}{a_i + b_i}\right) + \frac{1}{n}\sum_{i=1}^n \left(\frac{b_i}{a_i + b_i}\right)\\ &= 1.\end{aligned}$$

Step 2. Now we consider the case that A and B are interior-disjoint [or "ID"] finite unions of rectangles.

For every compact set K let us write $K^+ := \{x \in K : x_1 \geqslant 0\}$ and $K^- := \{x \in K : x_1 \leqslant 0\}$.

Now we apply a so-called "Hadwiger–Ohmann cut": Notice that if we translate A and/or B, then we do not alter $|A + B|$, $|A|$, or $|B|$. Therefore, after we translate the sets suitably, we can always ensure that: (a) A^+ and B^+ are rectangles; (b) A^- and B^- are ID unions of rectangles; and (c)

$$\frac{|A^+|}{|A|} = \frac{|B^+|}{|B|}.$$

With this choice in mind, we find that

$$|A + B| \geqslant |A^+ + B^+| + |A^- + B^-| \geqslant \left(|A^+|^{1/n} + |B^+|^{1/n}\right)^n + |A^- + B^-|,$$

thanks to Step 1 and the fact that $A^+ + B^+$ is disjoint from $A^- + B^-$. Now,

$$\left(|A^+|^{1/n} + |B^+|^{1/n}\right)^n = |A^+| \left(1 + \frac{|B^+|^{1/n}}{|A^+|^{1/n}}\right)^n = |A^+| \left(1 + \frac{|B|^{1/n}}{|A|^{1/n}}\right)^n,$$

whence

$$|A + B| \geqslant |A^+| \left(1 + \frac{|B|^{1/n}}{|A|^{1/n}}\right)^n + |A^- + B^-|.$$

Now split up, after possibly also translating, A^- into $A^{-,\pm}$ and B^- into $B^{-,\pm}$ such that: (i) $A^{-,\pm}$ are interior disjoint; (ii) $B^{-,\pm}$ are interior disjoint; and (iii) $|A^{-,+}|/|A^-| = |B^{-,+}|/|B^-|$. Thus, we can apply the preceding to A^- and B^- in place of A and B in order to see that

$$\begin{aligned}|A + B| &\geqslant |A^+| \left(1 + \frac{|B|^{1/n}}{|A|^{1/n}}\right)^n + |A^{-,+}| \left(1 + \frac{|B|^{1/n}}{|A|^{1/n}}\right)^n + |A^{-,-} + B^{-,-}|\\ &= \left(|A^+| + |A^{-,-}|\right) \left(1 + \frac{|B|^{1/n}}{|A|^{1/n}}\right)^n + |A^{-,-} + B^{-,-}|.\end{aligned}$$

And now continue to split and translate $A^{-,-}$ and $B^{-,-}$, etc. In this way we obtain a countable sequence $A_0 := A^+$, $A_1 := A^{-,+}$, ..., $B_0 := B^+$, $B_1 := B^{-,+}, \ldots$ of ID rectangles such that: (i) $\cup_{j=0}^{\infty} B_j = B$ [after translation]; (ii) $\cup_{j=0}^{\infty} A_j = A$ [after translation]; and (iii)

$$|A+B| \geqslant \sum_{j=0}^{\infty} |A_j| \left(1 + \frac{|B|^{1/n}}{|A|^{1/n}}\right)^n = |A| \left(1 + \frac{|B|^{1/n}}{|A|^{1/n}}\right)^n = \left(|A|^{1/n} + |B|^{1/n}\right)^n.$$

This proves the result in the case that A and B are ID unions of rectangles.

Step 3. Every compact set can be written as an countable union of ID rectangles. In other words, we can find $A^1, A^2, \ldots$ and $B^1, B^2, \ldots$ such that: (i) Every A^j and B^k is a finite union of ID rectangles; (ii) $A^j \subseteq A^{j+1}$ and $B^k \subseteq B^{k+1}$ for all $j, k \geqslant 1$; and (iii) $A = \cup_{j=1}^{\infty} A^j$ and $B = \cup_{j=1}^{\infty} B^j$. By the previous step, $|A+B|^{1/n} \geqslant |A^m + B^m|^{1/n} \geqslant |A^m|^{1/n} + |B^m|^{1/n}$ for all $m \geqslant 1$. Let $m \uparrow \infty$ and appeal to the inner continuity of Lebesgue measure in order to find deduce the theorem in its full generality. $\square$

§2. Part 2. Change of Variables. In the second part of the proof we develop an elementary fact from integration theory.

Let $A \subseteq \mathbb{R}^n$ be a Borel set, and $g : A \to \mathbb{R}_+$ a Borel-measurable function.

Definition 3.9. The *distribution function* of g is the function $\bar{G} : [0\,,\infty) \to \mathbb{R}_+$, defined as

$$\bar{G}(r) := \left|g^{-1}[r\,,\infty)\right| := |\{x \in A : g(x) \geqslant r\}| := |\{g \geqslant r\}| \qquad \text{for all } r \geqslant 0.$$

This is standard notation in classical analysis, and should not be mistaken with the closely-related definition of *cumulative distribution functions* in probability and statistics. In any case, the following should be familiar to you.

pr:ChangeofVar **Proposition 3.10** (Change of Variables Formula). *For every Borel measurable function $F : \mathbb{R}_+ \to \mathbb{R}_+$,*

$$\int_A F(g(x))\,\mathrm{d}x = -\int_0^\infty F(r)\,\mathrm{d}\bar{G}(r).$$

If, in addition, A is compact and F is absolutely continuous, then

$$\int_A F'(g(x))\,\mathrm{d}x = \int_0^\infty F'(r)\bar{G}(r)\,\mathrm{d}r.$$

Proof. First consider the case that $F = \mathbb{1}_{[a,b)}$ for some $b \geqslant a > 0$. In that case,

$$\int_0^\infty F(r)\,\mathrm{d}\bar{G}(r) = \bar{G}(b-) - \bar{G}(a) = -\left|\{x \in A : a \leqslant g(x) < b\}\right|$$
$$= -\int_A F(g(x))\,\mathrm{d}x.$$

This proves our formula when F is a simple function. By linearity, it holds also when F is an elementary function. The general form of the first assertion of the proposition follows from this and Lebesgue's dominated convergence theorem. The second follows from the first and integration by parts for Stieltjes integrals. □

§3. Part 3. The Proof of Anderson's Inequality. Let us define a new number $\alpha \in [0\,,1]$ by $\alpha := (1+\lambda)/2$. The number α is chosen so that

$$\alpha y + (1-\alpha)(-y) = \lambda y.$$

Since E is convex, we have $E = \alpha E + (1-\alpha)E$. Therefore, the preceding display implies that

$$(E + \lambda y) \supseteq \alpha(E + y) + (1-\alpha)(E - y).$$

And because the intersection of two convex sets is a convex set, we may infer that

$$(E+\lambda y)\cap f^{-1}[r\,,\infty) \supseteq \alpha\left[(E+y)\cap f^{-1}[r\,,\infty)\right] + (1-\alpha)\left[(E-y)\cap f^{-1}[r\,,\infty)\right].$$

Now we apply the Brunn–Minkowski inequality in order to see that

$$\left|(E+\lambda y)\cap f^{-1}[r\,,\infty)\right|^{1/n}$$
$$\geqslant \alpha\left|(E+y)\cap f^{-1}[r\,,\infty)\right|^{1/n} + (1-\alpha)\left|(E-y)\cap f^{-1}[r\,,\infty)\right|^{1/n}.$$

Since E is symmetric, $E - y = -(E + y)$. Because of this identity and the fact that f has symmetric level sets, it follows that

$$(E-y)\cap f^{-1}[r\,,\infty) = -\left[(E+r)\cap f^{-1}[r\,,\infty)\right].$$

Therefore,

$$\left|(E+y)\cap f^{-1}[r\,,\infty)\right|^{1/n} = \left|(E-y)\cap f^{-1}[r\,,\infty)\right|^{1/n},$$

whence

$$\bar{H}_\lambda(r) := \left|(E+\lambda y)\cap f^{-1}[r\,,\infty)\right| \geqslant \left|(E+y)\cap f^{-1}[r\,,\infty)\right| := \bar{H}_1(r).$$

Now two applications of the change of variables formula [Proposition 3.10] yield the following:

$$\begin{aligned}\int_E f(x-\lambda y)\,\mathrm{d}x - \int_E f(x-y)\,\mathrm{d}x &= \int_{E+\lambda y} f(x)\,\mathrm{d}x - \int_{E+y} f(x)\,\mathrm{d}x\\ &= -\int_0^\infty r\mathrm{d}\bar{H}_\lambda(r) + \int_0^\infty r\,\mathrm{d}\bar{H}_1(r)\\ &= \int_0^\infty \left[\bar{H}_\lambda(r) - \bar{H}_1(r)\right]\mathrm{d}r \geqslant 0.\end{aligned}$$

This completes the proof of Anderson's inequality. □

4. Gaussian Random Vectors

Let $(\Omega\,,\mathcal{F}\,,\mathbb{Q})$ be a probability space, and recall the following.

Definition 4.1. A random n-vector $X = (X_1\,,\ldots,X_n)$ in $(\Omega\,,\mathcal{F}\,,\mathbb{Q})$ is *Gaussian* if $a\cdot X$ has a normal distribution for every non-random n-vector a.

General theory ensures that we can always assume that $\Omega = \mathbb{R}^n$, $\mathcal{F} = \mathcal{B}(\mathbb{R}^n)$, and $\mathbb{Q} = \mathbb{P}_n$, which we will do from now on without further mention in order to save on the typography.

If X is a Gaussian random vector in $\mathbb{R}^n$ then $a\cdot X$ has moments of all orders. Let μ and Q respectively denote the mean vector and the covariance matrix of X.[2] Then it is easy to see that $a\cdot X$ must have a $\mathrm{N}(a\cdot\mu\,,a\cdot Qa)$ distribution. In particular, the characteristic function of X is given by

$$\mathbb{E}\left[\mathrm{e}^{ia\cdot X}\right] = \exp\left(ia\cdot\mu - \tfrac{1}{2}a\cdot Qa\right) \qquad \text{for all } a\in\mathbb{R}^n.$$

Definition 4.2. Let X be a Gaussian random vector in $\mathbb{R}^n$ with mean μ and covariance matrix Q. The distribution of X is then called a *multivariate normal distribution on* $\mathbb{R}^n$ and is denoted by $\mathrm{N}_n(\mu\,,Q)$.

When Q is non singular we can invert the Fourier transform to find that the probability density function of X is

$$p_X(x) = \frac{1}{(2\pi)^{n/2}\,|\det Q|^{1/2}}\exp\left\{-\tfrac{1}{2}(x-\mu)\cdot Q^{-1}(x-\mu)\right\} \qquad [x\in\mathbb{R}^n].$$

When Q is singular, the distribution of X is singular with respect to the Lebesgue measure on $\mathbb{R}^n$, and hence does not have a density.

[2] This means that $\mu_i = \mathbb{E}(X_i)$ and $Q_{i,j} = \mathrm{Cov}(X_i\,,X_j)$ respectively denote the expectation and covariance operators with respect to $\mathbb{P}$.

Example 4.3. Suppose $n = 2$ and W has a $\mathrm{N}(0\,,1)$ distribution on the line [which you might recall is denoted by $\mathbb{P}_1$]. Then, the distribution of $X = (W\,,W)$ is concentrated on the diagonal $\{(x\,,x) : x \in \mathbb{R}\}$ of $\mathbb{R}^2$. Since the diagonal has zero Lebesgue measure, it follows that the distribution of X is singular with respect to the Lebesgue measure on $\mathbb{R}^2$.

The following are a series of simple, though useful, facts from elementary probability theory.

lem:G1 **Lemma 4.4.** *Suppose X has a $\mathrm{N}_n(\mu\,,Q)$ distribution. Then for all $b \in \mathbb{R}^m$ and every $m \times n$ matrices A, the distribution of $AX + b$ is $\mathrm{N}_m(A\mu + b\,,A \cdot QA)$.*

lem:G2 **Lemma 4.5.** *Suppose X has a $\mathrm{N}_n(\mu\,,Q)$ distribution. Choose and fix an integer $1 \leqslant K \leqslant n$, and suppose in addition that $I_1, \ldots, I_K$ are K disjoint subsets of $\{1\,,\ldots,n\}$ such that*

$$\mathrm{Cov}(X_i\,,X_j) = 0 \qquad \textit{whenever } i \textit{ and } j \textit{ lie in distinct } I_\ell\textit{'s.}$$

Then, $(X_i)_{i\in I_1}, \ldots, (X_i)_{i\in I_K}$ are independent, each having a multivariate normal distribution.

lem:G3 **Lemma 4.6.** *Suppose X has a $\mathrm{N}_n(\mu\,,Q)$ distribution, where Q is a symmetric, non-singular, $n \times n$ covariance matrix. Then $Q^{-1/2}X$ has the same distribution as Z.*

We can frequently use one, or more, of these results to study the general Gaussian distribution on $\mathbb{R}^n$ via the canonical Gaussian measure $\mathbb{P}_n$. Here is a typical example.

th:Anderson:Gauss **Theorem 4.7** (Anderson's Shifted-Ball Inequality). *If X has a $\mathrm{N}_n(0\,,Q)$ distribution and Q is positive definite, then for all convex symmetric sets $F \subset \mathbb{R}^n$ and $a \in \mathbb{R}^n$,*

$$\mathbb{P}\{X \in a + F\} \leqslant \mathbb{P}\{X \in F\}.$$

Proof. Since $Q^{-1/2}X$ has the same distribution as Z,

$$\mathbb{P}\{X \in a + F\} = \mathbb{P}\left\{Z \in Q^{-1/2}a + Q^{-1/2}F\right\}.$$

Now $Q^{-1/2}F$ is symmetric and convex because F is. Apply Anderson's shifted-ball inequality for $\mathbb{P}_n$ [Corollary 3.7] to see that

$$\mathbb{P}\left\{Z \in Q^{-1/2}a + Q^{-1/2}F\right\} \leqslant \mathbb{P}\left\{Z \in Q^{-1/2}F\right\}.$$

This proves the theorem. $\square$

The following comparison theorem is one of the noteworthy corollaries of the preceding theorem.

th:Anderson:Gauss:2 **Corollary 4.8.** *Suppose X and Y are respectively distributed as $\mathrm{N}_n(0\,,Q_X)$ and $\mathrm{N}_n(0\,,Q_Y)$, where $Q_X - Q_Y$ is positive semidefinite. Then,*

$$\mathbb{P}\{X \in F\} \leqslant \mathbb{P}\{Y \in F\},$$

for all symmetric, closed convex sets $F \subset \mathbb{R}^n$.

Proof. First consider the case that Q_X, Q_Y, and $Q_X - Q_Y$ are positive definite. Let W be independent of Y and have a $\mathrm{N}_n(0\,,Q_X - Q_Y)$ distribution. The distribution of W has a probability density p_W, and $W + Y$ is distributed as X, whence

$$\mathbb{P}\{X \in F\} = \mathbb{P}\{W + Y \in F\} = \int_{\mathbb{R}^n} \mathbb{P}\{Y \in -a + F\} p_W(a)\,\mathrm{d}a \leqslant \mathbb{P}\{Y \in F\},$$

thanks to Theorem 4.7. This proves the theorem in the case that $Q_X - Q_Y$ is positive definite. If Q_Y is positive definite and $Q_X - Q_Y$ is positive semidefinite, then we define for all $0 < \delta < \varepsilon < 1$,

$$X(\varepsilon) := X + \varepsilon U, \quad Y(\delta) := Y + \delta U,$$

where U is independent of $(X\,,Y)$ and has the same distribution $\mathrm{N}_n(0\,,I)$ as Z. The respective distributions of $X(\varepsilon)$ and $Y(\delta)$ are $\mathrm{N}_n(0\,,Q_{X(\varepsilon)})$ and $\mathrm{N}_n(0\,,Q_{Y(\delta)})$, where $Q_{X(\varepsilon)} := Q_X + \varepsilon I$ and $Q_{Y(\delta)} := Q_Y + \delta I$. Since $Q_{X(\varepsilon)}$, $Q_{Y(\delta)}$, and $Q_{X(\varepsilon)} - Q_{Y(\delta)}$ are positive definite, the portion of the theorem that has been proved so far implies that $\mathbb{P}\{X(\varepsilon) \in F\} \leqslant \mathbb{P}\{Y(\delta) \in F\}$, for all symmetric convex sets $F \subset \mathbb{R}^n$. Let ε and δ tend down to zero, all the while ensuring that $\delta < \varepsilon$. Since $F = \bar{F}$, this proves the result. □

Example 4.9 (Comparison of Moments)**.** For every $1 \leqslant p \leqslant \infty$, the ℓ^p-norm of $x \in \mathbb{R}^n$ is

$$\|x\|_p := \begin{cases} \left(\sum_{i=1}^n |x_i|^p\right)^{1/p} & \text{if } p < \infty, \\ \max_{1 \leqslant i \leqslant n} |x_i| & \text{if } p = \infty. \end{cases}$$

It is easy to see that all centered ℓ^p-balls of the form $\{x \in \mathbb{R}^n :\ \|x\|_p \leqslant t\}$ are convex and symmetric. Therefore, it follows immediately from Corollary 4.8 that if $Q_X - Q_Y$ is positive semidefinite, then

$$\mathbb{P}\left\{\|X\|_p > t\right\} \geqslant \mathbb{P}\left\{\|Y\|_p > t\right\} \qquad \text{for all } t > 0 \text{ and } 1 \leqslant p \leqslant \infty.$$

Multiply both sides by rt^{r-1} and integrate both sides $[\mathrm{d}t]$ from $t = 0$ to $t = \infty$ in order to see that

$$\mathbb{E}\left[\|X\|_p^r\right] \geqslant \mathbb{E}\left[\|Y\|_p^r\right] \qquad \text{for } r > 0 \text{ and } 1 \leqslant p \leqslant \infty.$$

These are examples of moment comparison, and can sometimes be useful in estimating expectation functionals of X in terms of expectation functionals of a Gaussian random vector Y with a simpler covariance matrix than that of X. Similarly, $\mathbb{P}\{\|X + a\|_p > t\} \geqslant \mathbb{P}\{\|X\|_p > t\}$ for all $a \in \mathbb{R}^n$, $t > 0$, and $1 \leqslant p \leqslant \infty$ by Theorem 4.7. Therefore,

$$\mathbb{E}\left(\|X\|_p^r\right) = \inf_{a \in \mathbb{R}^n} \mathbb{E}\left(\|X + a\|_p^r\right) \qquad \text{for all } 1 \leqslant p \leqslant \infty \text{ and } r > 0.$$

This is a nontrivial generalization of the familiar fact that when $n = 1$, $\mathrm{Var}(X) = \inf_{a \in \mathbb{R}} \mathbb{E}(|X - a|^2)$.