GAUSSIAN DISTRIBUTIONS WITH APPLICATIONS

MAT 477, FALL 2016

TABLE OF CONTENTS

1 Gaussian distributions on $\mathbb{R}$

Let us start with some basic definitions in a language that assumes minimal or no background in probability. The function

$$p(x) = \frac{1}{\sqrt{2\pi}} e^{-x^2/2} \tag{1}$$

is called the *density function*, or simply density, of the *standard Gaussian distribution*. (This notation p is not standard and we will at times use different notation.) This means that, for any measurable set A on $\mathbb{R}$, we define its standard Gaussian measure by Lebesgue's integral

$$\gamma(A) = \int_A \frac{1}{\sqrt{2\pi}} e^{-x^2/2}\, dx. \tag{2}$$

For example, if A is an interval $[a,b]$, then

$$\gamma([a,b]) = \int_a^b \frac{1}{\sqrt{2\pi}} e^{-x^2/2}\, dx. \tag{3}$$

Notice that, in the case when A is the whole real line $\mathbb{R}$, we can use the following trick, using polar coordinates, to compute

$$\Big(\frac{1}{\sqrt{2\pi}} \int_{\mathbb{R}} e^{-x^2/2}\, dx\Big)^2 = \frac{1}{2\pi} \iint_{\mathbb{R}^2} e^{-(x^2+y^2)/2}\, dxdy = \frac{1}{2\pi} \int_0^{2\pi}\!\!\int_0^{\infty} e^{-r^2/2} r\, dr d\theta = 1,$$

so the measure of the whole line $\gamma(\mathbb{R})$ is equal to 1. By the countable additivity of the Lebesgue integral, if $A = \cup_{\ell \geq 1} A_\ell$ is a disjoint union of countably many sets A_ℓ for $\ell \geq 1$, then

$$\gamma(A) = \int_A \frac{1}{\sqrt{2\pi}} e^{-x^2/2}\, dx = \sum_{\ell \geq 1} \int_{A_\ell} \frac{1}{\sqrt{2\pi}} e^{-x^2/2}\, dx = \sum_{\ell \geq 1} \gamma(A_\ell).$$

If you imagine an experiment whose outcome is a random real number, you can interpret measure $\gamma(A)$ as the *probability* that the outcome will belong to the set A. Remember, that this is just an interpretation and, even when we introduce different notation to better match this interpretation, at the end of the day, we are always interested in various properties of the above measure γ and its multivariate analogues below.

In a basic undergraduate probability class, you would denote the random outcome of this imaginary experiment by, say, g, call it a *random variable*, and use the notation $\mathbb{P}(g \in A)$ to denote the "probability that the random variable g takes value in the set A". When

$$\mathbb{P}(g \in A) = \gamma(A) = \int_A \frac{1}{\sqrt{2\pi}} e^{-x^2/2}\, dx, \tag{4}$$

we would call the standard Gaussian measure γ the *distribution* of the random variable g. Of course, there are many other distributions. The standard Gaussian distribution is also often called the standard normal distribution and is denoted by $N(0,1)$. An abbreviation $g \sim N(0,1)$ means that

g has the standard Gaussian distribution, i.e. (4) holds.

In a graduate probability class, such vagueness as "outcome of a random experiment" is not allowed, so "random variable" has precise definition, but this is not crucial to us.

Given a (measurable) function $f\colon \mathbb{R} \to \mathbb{R}$, we can plug in g into f and define the *expectation* of $f(g)$, denoted by $\mathbb{E}f(g)$, as the average of f weighted by the density $p(x)$,

$$\mathbb{E}f(g) := \int_{\mathbb{R}} f(x)p(x)\,dx. \tag{5}$$

Informally, we will often call the expectation $\mathbb{E}f(g)$ the average of $f(g)$. We will assume that this is defined only when $f(x)$ is absolutely integrable, as is usual in the Lebesgue integration, i.e.

$$\mathbb{E}|f(g)| = \int_{\mathbb{R}} |f(x)|p(x)\,dx < \infty.$$

Notice the basic connection between expectations and probability (or integrals and measure),

$$\mathbb{P}(g \in A) = \mathbb{E}\,\mathrm{I}(g \in A), \tag{6}$$

where $\mathrm{I}(x \in A)$ is the indicator of the set A,

$$\mathrm{I}(x \in A) = \begin{cases} 1 & \text{if } x \in A \\ 0 & \text{if } x \notin A. \end{cases}$$

Of course, this connection is just another notation for

$$\gamma(A) = \int_A p(x)\,dx = \int_{\mathbb{R}} \mathrm{I}(x \in A)p(x)\,dx.$$

When $f(x) = x^k$ for integer $k \geq 1$, the average

$$\mathbb{E}g^k = \int_{\mathbb{R}} x^k p(x)\,dx \tag{7}$$

is called the k^{th} *moment* of g. The first moment is also called the *mean*. If g has the standard Gaussian distribution $N(0,1)$ then, by symmetry, $\mathbb{E}g = 0$, and, by integration by parts,

$$\mathbb{E}g^2 = \frac{1}{\sqrt{2\pi}} \int_{\mathbb{R}} x^2 e^{-x^2/2}\,dx = \frac{1}{\sqrt{2\pi}} \int_{\mathbb{R}} x\,d(-e^{-x^2/2}) = \frac{1}{\sqrt{2\pi}} \int_{\mathbb{R}} e^{-x^2/2}\,dx = 1.$$

The quantity $\mathrm{Var}(g) = \mathbb{E}(g - \mathbb{E}g)^2 = \mathbb{E}g^2 - (\mathbb{E}g)^2$ is called the *variance* of a random variable g, so for standard Gaussian $\mathrm{Var}(g) = 1$.

Given two parameters $\mu \in \mathbb{R}$ and $\sigma > 0$, and let us denote $X := \sigma g + \mu$. It is easy to see by

the change of variables that

$$\mathbb{P}(X \in A) = \mathbb{P}\big(\sigma g + \mu \in A\big) = \mathbb{P}\big(g \in \frac{A-\mu}{\sigma}\big) = \int_{\frac{A-\mu}{\sigma}} \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}}\, dx = \int_A \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}}\, dx.$$

The integrand inside the last integral,

$$p_{\mu,\sigma}(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}}, \tag{8}$$

is called the *density of the Gaussian distribution* $N(\mu, \sigma^2)$ *with the mean* μ *and variance* σ^2, because $\mu = \mathbb{E}X$ and $\sigma^2 = \mathrm{Var}(X)$. When $\mu = 0$, Gaussian distribution $N(0, \sigma^2)$ is the distribution of the linear transformation σg of a standard Gaussian random variable g.

Notice how the probabilistic notation $\mathbb{P}(\sigma g + \mu \in A), \mathbb{E}f(\sigma g + \mu)$ conveniently allows you to think of these quantities either in terms of the random variable g and express, for example,

$$\mathbb{E}f(\sigma g + \mu) = \int_{\mathbb{R}} f(\sigma x + \mu) p(x)\, dx,$$

or in terms of the random variable $X = \sigma g + \mu$ and express this as

$$\mathbb{E}f(X) = \int_{\mathbb{R}} f(x) p_{\mu,\sigma}(x)\, dx.$$

In other words, the notation $\mathbb{P}$ and $\mathbb{E}$ conveniently hides the change of density, depending on what you choose to view as your random variable.

Let us record a couple of basic properties of probabilities and expectations, which will be useful to us. Suppose that the function f in (5) is nonnegative, and consider any $t > 0$. Then, integrating separately over two sets $\{g \mid f(g) \geq t\}$ and $\{g \mid f(g) < t\}$, we can write

$$\begin{aligned}\mathbb{E}f(g) &= \mathbb{E}f(g)\,\mathrm{I}(f(g) \geq t) + \mathbb{E}f(g)\,\mathrm{I}(f(g) < t) \\ &\geq \mathbb{E}f(g)\,\mathrm{I}(f(g) \geq t) \geq t\mathbb{E}\,\mathrm{I}(f(g) \geq t) = t\mathbb{P}(f(g) \geq t).\end{aligned}$$

This simple computation gives what is known as *Chebyshev's inequality,*

$$\mathbb{P}(f(g) \geq t) \leq \frac{\mathbb{E}f(g)}{t}. \tag{9}$$

If you are familiar with general distributions and expectations, the same calculation gives

$$\mathbb{P}(X \geq t) \leq \frac{\mathbb{E}X}{t} \tag{10}$$

for any nonnegative random variable X and $t > 0$. In particular, for any random variable X, any

$t \in \mathbb{R}$ and $\lambda \geq 0$, this implies *Markov's inequality*,

$$\mathbb{P}(X \geq t) = \mathbb{P}(e^{\lambda X} \geq e^{\lambda t}) \leq e^{-\lambda t}\mathbb{E}e^{\lambda X}. \tag{11}$$

Now, let $f\colon \mathbb{R} \to \mathbb{R}$ be a convex function. By convexity, $f(x) - f(a) \geq f'(a)(x-a)$. If f is not differentiable at a, one can replace $f'(a)$ by the slope of any line below f passing through the point $(a, f(a))$. For $x = X$ and $a = \mathbb{E}X$, $f(X) - f(\mathbb{E}X) \geq f'(\mathbb{E}X)(X - \mathbb{E}X)$, and taking expectations (integrating) on both sides, we get

$$\mathbb{E}f(X) - f(\mathbb{E}X) \geq f'(\mathbb{E}X)(\mathbb{E}X - \mathbb{E}X) = 0.$$

This means that for any convex function f, we have

$$\mathbb{E}f(X) \geq f(\mathbb{E}X), \tag{12}$$

which is called *Jensen's inequality.*

2 Standard Gaussian distributions on $\mathbb{R}^n$

Let us now consider $x = (x_1, \ldots, x_n) \in \mathbb{R}^n$ and denote its Euclidean norm by

$$\|x\| = \left(x_1^2 + \ldots + x_n^2\right)^{1/2}.$$

The function

$$p_n(x) = \frac{1}{(\sqrt{2\pi})^n} e^{-\|x\|^2/2} \tag{13}$$

is called the *density function*, or simply density, of the *standard Gaussian distribution* on $\mathbb{R}^n$. This means that, for any measurable set A on $\mathbb{R}^n$, we define its standard Gaussian measure by

$$\gamma_n(A) = \int_A \frac{1}{(\sqrt{2\pi})^n} e^{-\|x\|^2/2}\, dx. \tag{14}$$

If p is the standard Gaussian density on $\mathbb{R}$ in the previous section then, obviously,

$$p_n(x) = p(x_1) \cdots p(x_n).$$

As before, we can imagine an outcome of some random experiment $g = (g_1, \ldots, g_n)$ such that

$$\mathbb{P}(g \in A) = \gamma_n(A) = \int_A \frac{1}{(\sqrt{2\pi})^n} e^{-\|x\|^2/2}\, dx. \tag{15}$$

We would call the standard Gaussian measure γ_n the *distribution* of the random vector g. In some sense, this vector g or measure γ_n will be our most basic objects of study.

Let us notice that, if we take the set A to be a rectangle $A = \prod_{\ell \leq n} A_\ell$ with sides A_ℓ then, by

Fubini's theorem,

$$\mathbb{P}(g \in A) = \mathbb{P}(g_1 \in A_1, \ldots, g_n \in A_n) = \int_A p(x_1) \cdots p(x_n)\, dx_1 \ldots dx_n = \prod_{\ell \leq n} \int_{A_\ell} p(x)\, dx.$$

In particular, if we take all sides to be $\mathbb{R}$ except one side A_ℓ, this gives that

$$\mathbb{P}(g_\ell \in A_\ell) = \int_{A_\ell} p(x)\, dx.$$

This means that each coordinate of the standard Gaussian random vector $g \in \mathbb{R}^n$ is a standard Gaussian random variable. Again, this is just a probabilistic way of saying that, if you want to measure along only one coordinate without any constraints on the other coordinates, after integrating out all the other coordinates, you get

$$\gamma_n(\mathbb{R} \times \ldots \times A_\ell \times \cdots \times \mathbb{R}) = \gamma(A_\ell),$$

i.e. the standard Gaussian measure on the real line. The measure on one coordinate when you don't put any constraints on the other coordinates is called the *marginal* on this coordinate, so the marginals of the standard Gaussian measure on $\mathbb{R}^n$ are standard Gaussian on $\mathbb{R}$.

We can rewrite the above Fubini theorem as

$$\mathbb{P}(g_1 \in A_1, \ldots, g_n \in A_n) = \mathbb{P}(g_1 \in A_1) \cdots \mathbb{P}(g_n \in A_n). \tag{16}$$

In probability, this property is called *independence*, or that the random variables $g_1 \ldots, g_n$ are *independent.* This name comes from the fact that for, let's say two random vectors X and Y,

$$\mathbb{P}(X \in A | Y \in B) := \frac{\mathbb{P}(X \in A, Y \in B)}{\mathbb{P}(Y \in B)}$$

is called the *conditional probability* that X is in A given that Y is in B. This proportion of outcomes $X \in A$ inside the set $Y \in B$ represents the "probability to observe X from A given that you already observed Y from B". If the information above Y does not affect the chances of X, we should have

$$\mathbb{P}(X \in A | Y \in B) = \mathbb{P}(X \in A),$$

which, by the above definition, is equivalent to

$$\mathbb{P}(X \in A, Y \in B) = \mathbb{P}(X \in A)\mathbb{P}(Y \in B).$$

The random variables $g_1 \ldots, g_n$ are called independent if information about any subset of them does not affect the chances of the others, which is exactly (16).

A slight modification of the standard Gaussian distribution γ_n on $\mathbb{R}^n$ would be to scale each

coordinate by a constant and consider a random vector $(\sigma_1 g_1, \ldots, \sigma_n g_n)$, whose density is

$$\prod_{\ell \le n} p_{0,\sigma_\ell}(x_\ell) = \prod_{\ell \le n} \frac{1}{\sqrt{2\pi}\sigma_\ell} e^{-\frac{x_\ell^2}{2\sigma_\ell^2}}. \tag{17}$$

Under this measure, the coordinates are independent with Gaussian distributions $N(0, \sigma_\ell^2)$.

3 General Gaussian distributions on $\mathbb{R}^n$: non-degenerate case

In probabilistic language, general Gaussian distributions are defined as the distributions of linear maps of the standard Gaussian random vector $(g_1, \ldots, g_n)$. The maps could be between spaces of different dimensions $\mathbb{R}^n$ and $\mathbb{R}^k$ but, for simplicity, we will first use only linear maps on $\mathbb{R}^n$ corresponding to square matrices and in this section we will consider only non-degenerate case.

Consider a $n \times n$ matrix A such that $\det(A) \neq 0$, and define a random vector $X = Ag$, where $g = (g_1, \ldots, g_n)$ is a standard Gaussian random vector (of course, when we use linear algebra notation, such as Ag, we think of g as a column vector). For any (measurable) set Ω in $\mathbb{R}^n$, we can write the probability that X belong to this set as

$$\mathbb{P}(X \in \Omega) = \mathbb{P}(Ag \in \Omega) = \mathbb{P}(g \in A^{-1}\Omega) = \int_{A^{-1}\Omega} \frac{1}{(\sqrt{2\pi})^n} \exp\Big(-\frac{1}{2}\|x\|^2\Big)\, dx.$$

Let us now make the change of variables $y = Ax$ or $x = A^{-1}y$. Then

$$\mathbb{P}(X \in \Omega) = \int_{\Omega} \frac{1}{(\sqrt{2\pi})^n} \exp\Big(-\frac{1}{2}\|A^{-1}y\|^2\Big) \frac{1}{|\det(A)|}\, dy.$$

First of all,

$$\|A^{-1}y\|^2 = (A^{-1}y)^T (A^{-1}y) = y^T (A^T)^{-1} A^{-1} y = y^T (AA^T)^{-1} y = y^T C^{-1} y,$$

where in the last step we introduced the notation

$$C = AA^T. \tag{18}$$

We will see in a second that C is what is called the *covariance matrix* of X. It is obvious that this matrix is symmetric and positive semidefinite. Since

$$\det(C) = \det(AA^T) = \det(A)\det(A^T) = \det(A)^2,$$

we have $|\det(A)| = \sqrt{\det(C)}$. Therefore, we can rewrite the above probability as

$$\mathbb{P}(X \in \Omega) = \int_{\Omega} \frac{1}{(\sqrt{2\pi})^k} \frac{1}{\sqrt{\det(C)}} \exp\Big(-\frac{1}{2} y^T C^{-1} y\Big)\, dy.$$

This means that the integrand

$$p_C(x) := \frac{1}{(\sqrt{2\pi})^n} \frac{1}{\sqrt{\det(C)}} \exp\Big(-\frac{1}{2} x^T C^{-1} x\Big) \tag{19}$$

is the density of X. The distribution with this density, denoted $N(0,C)$, is called the *Gaussian distribution on* $\mathbb{R}^n$ *with the covariance* C. Notice that it does not depend on the choice of the matrix A but only on the matrix C.

- The *covariance matrix* of a random vector X is the matrix C with elements

$$c_{i,j} = \mathbb{E} X_i X_j,$$

when these expectations exist. When $X = Ag$ as above, using that

$$\mathbb{E} g_k^2 = 1 \text{ and } \mathbb{E} g_k g_{k'} = 0 \text{ for } k \neq k'$$

(integrate with respect to density $p_n(x)$ and use Fubini's theorem), we get by linearity of integral (expectation),

$$\mathbb{E} X_i X_j = \mathbb{E} \sum_{k=1}^n a_{i,k} g_k \sum_{k=1}^n a_{j,k} g_k = \sum_{k,k'=1}^n a_{i,k} a_{j,k'} \mathbb{E} g_k g_{k'} = \sum_{k=1}^n a_{i,k} a_{j,k}.$$

This is the scalar product of the i^{th} and j^{th} rows of matrix A or, in other words, of the i^{th} row of matrix A and j^{th} column of matrix A^T, which is exactly the (i,j) element $c_{i,j}$ of $C = AA^T$. So C is the covariance matrix of X,

$$\mathrm{Cov}(X) = C,$$

as we mentioned above.

- Notice that, in the case when $X = Qg$ for some orthogonal matrix Q, the covariance $C = QQ^T$ is the identity matrix, and the density of X in (19) is standard Gaussian on $\mathbb{R}^n$,

$$p_Q(x) = \frac{1}{(\sqrt{2\pi})^n} e^{-\|x\|^2/2}. \tag{20}$$

This is basically because the standard Gaussian density is a function of $\|x\|$ and is, therefore, rotationally invariant.

- If we take another invertible matrix B, and consider random vector $Y = BX = BAg$, it will have a density function (19), only with the covariance

$$\mathrm{Cov}(Y) = (BA)(BA)^T = B(AA^T)B^T = BCB^T = B\,\mathrm{Cov}(X)B^T. \tag{21}$$

It is easy to check that this relationship, $\mathrm{Cov}(Y) = B\,\mathrm{Cov}(X)B^T$, between the covariances of X and $Y = BX$ is always true, not only in the Gaussian case.

- The distribution $N(0,C)$ is completely determined by the covariance matrix C. Suppose that

the covariance matrix is of the block-diagonal form

$$C = \begin{bmatrix} C_1 & 0 \\ 0 & C_2 \end{bmatrix},$$

where C_1 is $n_1 \times n_1$ and C_2 is $n_2 \times n_2$, where $n = n_1 + n_2$. If we write $X = (U,V)$, where U consists of the first n_1 coordinates and V consists of the last n_2 coordinates, then the covariances $\mathbb{E}X_iX_j$ are equal to 0 if $i \le n_1$ and $j \ge n_1 + 1$. In other words, the coordinates of U and V are *uncorrelated*. In this case, if we represent $x \in \mathbb{R}^n$ and $x = (u,v)$ for $u \in \mathbb{R}^{n_1}$ and $v \in \mathbb{R}^{n_2}$, we can write

$$C^{-1} = \begin{bmatrix} C_1^{-1} & 0 \\ 0 & C_2^{-1} \end{bmatrix}, \; x^T C^{-1} x = u^T C_1^{-1} u + v^T C_2^{-1} v$$

and, $\det(C) = \det(C_1)\det(C_2)$. Therefore, the density $p_C(x)$ in (19) can be rewritten as

$$p_C(x) = p_{C_1}(u) p_{C_2}(v) \tag{22}$$

where

$$p_{C_1}(u) = \frac{1}{(\sqrt{2\pi})^{n_1}} \frac{1}{\sqrt{\det(C_1)}} \exp\Big(-\frac{1}{2} u^T C_1^{-1} u\Big)$$

and

$$p_{C_2}(v) = \frac{1}{(\sqrt{2\pi})^{n_2}} \frac{1}{\sqrt{\det(C_2)}} \exp\Big(-\frac{1}{2} v^T C_2^{-1} v\Big)$$

are Gaussian densities on $\mathbb{R}^{n_1}$ and $\mathbb{R}^{n_2}$. Given two sets $A_1 \subseteq \mathbb{R}^{n_1}$ and $A_2 \subseteq \mathbb{R}^{n_2}$, we can calculate the probability that $X = (U,V)$ belongs to the rectangle $A_1 \times A_2$ as

$$\begin{aligned} \mathbb{P}(U \in A_1, V \in A_2) = \mathbb{P}(X \in A_1 \times A_2) &= \int_{A_1 \times A_2} p_C(x)\, dx = \iint_{A_1 \times A_2} p_{C_1}(u) p_{C_2}\, du dv \\ &= \int_{A_1} p_{C_1}(u)\, du \int_{A_2} p_{C_2}(v)\, dv = \mathbb{P}(U \in A_1)\mathbb{P}(V \in A_2). \end{aligned}$$

(Why does the last equality hold?) This means that uncorrelated coordinates of a Gaussian vector are independent in the probabilistic sense described above. Analytically, it just means that the density (if it exists) decouples into a product of two densities on the corresponding subspaces. What we showed is that for Gaussian distributions, *uncorrelated means independent.*

4 General Gaussian distributions on $\mathbb{R}^n$

Let us now consider a standard Gaussian vector $g = (g_1, \dots, g_m)$ on $\mathbb{R}^m$ and let A be $n \times m$ matrix. Consider $X = Ag$, which is now a random vector in $\mathbb{R}^n$. We will see that the distribution of this vector depends only on the covariance matrix of X,

$$\mathrm{Cov}(X) = \mathbb{E}Ag(Ag)^T = A(\mathbb{E}gg^t)A^T = AA^T =: C,$$

and does not really depend on the matrix A or dimension m, except through this covariance matrix. As before, the distribution of X is denoted $N(0,C)$ and is called the *Gaussian distribution on* $\mathbb{R}^n$ *with the covariance* C.

To prove the above statement, we will use the singular value decomposition of the matrix A. The matrix $C = AA^T$ is, obviously, symmetric and positive-semidefinite. Let

$$C = QDQ^T$$

be its eigenvalue decomposition for some $n \times n$ orthogonal matrix Q and diagonal matrix D that has eigenvalues $d_1, \ldots, d_n$ of C on the diagonal, $D = \mathrm{diag}(d_1, \ldots, d_n)$. Singular value decomposition tells us that A can be written as

$$A = QD^{1/2}R,$$

where $D = \mathrm{diag}(d_1^{1/2}, \ldots, d_n^{1/2})$ and R is some $m \times m$ orthogonal matrix on $\mathbb{R}^m$. We can rewrite X, using this representation, as

$$X = Ag = QD^{1/2}Rg. \tag{23}$$

Recall that the random vector Rg on $\mathbb{R}^m$ has the standard Gaussian distribution, just like g, because R is orthogonal. From the point of view of calculating probabilities, we can redefine

$$X = QD^{1/2}g, \tag{24}$$

which simply amounts to making the change of variables $Rg \to g$, which does not affect the density. Let us denote columns of the matrix Q by $q_1, \ldots, q_n$,

$$Q = \begin{bmatrix} q_1 & q_2 & \cdots & q_n \end{bmatrix},$$

and let us suppose that the eigenvalues of C are arranged in the decreasing order, $d_1 \geq d_2 \geq \ldots \geq d_n$. Some of them can be zero, so let us suppose that the first r are non-zero. In this case, we can rewrite

$$X = QD^{1/2}g = g_1 d_1^{1/2} q_1 + \ldots + g_r d_r^{1/2} q_r. \tag{25}$$

Notice that this formula only involves the first r coordinates of the standard Gaussian random vector $g = (g_1, \ldots, g_m)$, and their distribution is standard Gaussian on $\mathbb{R}^r$, so the dependence on the dimension m disappeared. The dependence on the original matrix A also disappeared, since this representation depends only on the eigenvalues and eigenvectors of the covariance matrix C. In other words, any linear map of the standard Gaussian random vector on a space of arbitrary dimension has distribution that depends only on the resulting covariance matrix C. For example, we can always choose A to be a square matrix if we like, for example, $A = QD^{1/2}$. Let us collect several properties of these distributions that will be useful to us.

- Any linear map $Y = BX$ of $X \sim N(0,C)$ is Gaussian $N(0, BCB^T)$. This is because we can think of Y as $Y = BAg$ and calculate

$$\mathrm{Cov}(Y) = BA(BA)^T = B(AA^T)B^T = BCB^T,$$

just like in the previous section.

• One simple consequence of this is the following *stability property* of real-valued Gaussian distributions. Let us take two independence standard Gaussian random variables g_1 and g_2, which can be viewed as coordinates of the standard Gaussian random vector (g_1, g_2) on $\mathbb{R}^2$. The maps

$$X_1 = \sigma_1 g_1,\ X_2 = \sigma_2 g_2,\ X = X_1 + X_2 = \sigma_1 g_1 + \sigma_2 g_2$$

have Gaussian distributions $N(0, \sigma_1^2)$, $N(0, \sigma_2^2)$ and $N(0, \sigma_1^2 + \sigma_2^2)$ on $\mathbb{R}$. In other words, if random variables $X_1 \sim N(0, \sigma_1^2)$, $X_2 \sim N(0, \sigma_2^2)$, and X_1, X_2 are independent, then their sum

$$X_1 + X_2 \sim N(0, \sigma_1^2 + \sigma_2^2).$$

More generally, if Gaussian random variables $X_i \sim N(0, \sigma_i^2)$ for $i \leq n$ are independent, then

$$X_1 + \ldots + X_n \sim N(0, \sigma_1^2 + \ldots + \sigma_n^2). \tag{26}$$

Next, we consider a multidimensional analogue of this stability property.

• Let us consider two independent Gaussian random vectors $X \sim N(0, C_1)$ and $Y \sim N(0, C_2)$ on $\mathbb{R}^n$. Independence means that we can view $X = Ag$ and $Y = Bg'$ as linear maps of independent standard Gaussian random vectors g and g' on $\mathbb{R}^n$. If you like, you can view g and g' as the first n and the last n coordinates of the standard Gaussian random vector (g, g') on $\mathbb{R}^{2n}$. Then the sum

$$X + Y \sim N(0, C_1 + C_2)$$

is again a Gaussian vector on $\mathbb{R}^n$, because it is a linear map of (g, g'). Its covariance can be easily computed and it is equal to $C_1 + C_2$.

• As in the previous section, suppose that the covariance matrix of $X \sim N(0, C)$ is of the block-diagonal form

$$C = \begin{bmatrix} C_1 & 0 \\ 0 & C_2 \end{bmatrix},$$

where C_1 is $n_1 \times n_1$ and C_2 is $n_2 \times n_2$, where $n = n_1 + n_2$. Consider any $n_1 \times n_1$ matrix A and $n_2 \times n_2$ matrix B such that $C_1 = AA^T$ and $C_2 = BB^T$. Let g be standard Gaussian random vector on $\mathbb{R}^{n_1}$ and g' to be standard Gaussian on $\mathbb{R}^{n_2}$. If we define $U = Ag$ and $V = Bg'$ then the vector (U, V) is Gaussian on $\mathbb{R}^n$, because it is a linear map of the standard Gaussian vector (g, g') on $\mathbb{R}^{2n}$. On the other hand, using that $\mathbb{E} g (g')^T = 0$, it is easy to check that the covariance of (U, V) is equal to C. This means that X and (U, V) have the same distribution, which means that the first n_1 coordinates of X are independent of the last n_2 coordinates.

5 Gaussian integration by parts

Let g be a Gaussian random variable with variance $\mathbb{E}g^2 = \sigma^2$. Let us denote its density function by

$$\varphi(x) = \frac{1}{\sqrt{2\pi v}} \exp\Big(-\frac{x^2}{2\sigma^2}\Big). \tag{27}$$

Notice that $x\varphi(x) = -\sigma^2\varphi'(x)$. Therefore, given a continuously differentiable function $F : \mathbb{R} \to \mathbb{R}$, we can integrate by parts,

$$\begin{aligned}\mathbb{E}gF(g) &= \int xF(x)\varphi(x)dx = -\sigma^2 F(x)\varphi(x)\Big|_{-\infty}^{+\infty} + \sigma^2\int F'(x)\varphi(x)dx \\ &= \sigma^2\int F'(x)\varphi(x)dx = \sigma^2\mathbb{E}F'(g),\end{aligned}$$

if the limits $\lim_{x\to\pm\infty} F(x)\varphi(x) = 0$ and the integrals on both sides are finite. Therefore,

$$\mathbb{E}gF(g) = \mathbb{E}g^2\mathbb{E}F'(g). \tag{28}$$

This computation can be generalized to Gaussian vectors as follows.

Let $g = (g_\ell)_{1\le\ell\le n}$ be a Gaussian random vector. Given a continuously differentiable function

$$F = F\big((x_\ell)_{1\le\ell\le n}\big) : \mathbb{R}^n \to \mathbb{R}$$

that satisfies some mild growth conditions (discussed below), let us show how one can integrate $\mathbb{E}g_1F(g)$ by parts. If $\sigma^2 = \mathbb{E}g_1^2 > 0$ then the Gaussian vector $\hat{g} = (\hat{g}_\ell)_{1\le\ell\le n}$ defined by

$$\hat{g}_\ell = g_\ell - \lambda_\ell g_1 \text{ where } \lambda_\ell = \frac{\mathbb{E}g_1g_\ell}{\sigma^2}, \tag{29}$$

is independent of g_1, since the covariance

$$\mathbb{E}g_1\hat{g}_\ell = \mathbb{E}g_1g_\ell - \lambda_\ell\sigma^2 = 0.$$

If we denote $\lambda = (\lambda_\ell)_{1\le\ell\le n}$ then we can write $g = \hat{g} + g_1\lambda$. If $\mathbb{E}_1$ denotes the expectation in g_1 only for a fixed $\hat{g}$, then using (28) implies that

$$\mathbb{E}_1g_1F(g) = \mathbb{E}_1g_1F(\hat{g}+g_1\lambda) = \sigma^2\mathbb{E}_1\frac{\partial F}{\partial t}(\hat{g}+t\lambda)\Big|_{t=g_1} \tag{30}$$

if, for all $\hat{g}$, $\lim_{t\to\pm\infty} F(\hat{g}+t\lambda)\varphi(t) = 0$ and both sides of (30) are finite, which can be ensured by some mild growth conditions on F and its partial derivatives (see below). If we assume that

$$g_1F(\hat{g}+g_1\lambda) \text{ and } \frac{\partial F}{\partial x}(\hat{g}+t\lambda)\Big|_{t=g_1} \tag{31}$$

are absolutely integrable then, integrating (30) in $\hat{g}$, by Fubini's theorem,

$$\mathbb{E}g_1F(g) = \sigma^2\mathbb{E}\frac{\partial F}{\partial t}(\hat{g}+t\lambda)\Big|_{t=g_1}. \tag{32}$$

Finally, if we compute the derivative,

$$\frac{\partial F}{\partial t}(\hat{g}+t\lambda)\Big|_{t=g_1} = \sum_{\ell\le n}\lambda_\ell\frac{\partial F}{\partial x_\ell}(\hat{g}+g_1\lambda) = \sum_{\ell\le n}\lambda_\ell\frac{\partial F}{\partial x_\ell}(g), \tag{33}$$

the equation (32) can be rewritten as

$$\mathbb{E}g_1F(g) = \sum_{\ell\le n}\mathbb{E}(g_1g_\ell)\mathbb{E}\frac{\partial F}{\partial x_\ell}(g). \tag{34}$$

This formula is called the *Gaussian integration by parts* formula.

The conditions that were used in the derivation of this formula can usually be easily verified in applications. For example, it is sufficient to have at most exponential growth of F and its partial derivatives. If we assume that, for some constants $c_1, c_2 > 0$,

$$|F(x)| \le c_1e^{c_2\|x\|} \text{ and, for } \ell\le n, \text{ either } \left|\frac{\partial F}{\partial x_\ell}(x)\right| \le c_1e^{c_2\|x\|} \text{ or } \mathbb{E}(g_1g_\ell) = 0, \tag{35}$$

where $\|x\|$ is the Euclidean norm of $x\in\mathbb{R}^n$, then it is easy to see that all the assumptions above are satisfied. When $\mathbb{E}(g_1g_\ell) = 0$, the partial derivative $\frac{\partial F}{\partial x_\ell}$ does not appear in (33), so we do not need to make any assumptions on its growth.

6 Gaussian interpolation

We will now explain a certain canonical Gaussian interpolation technique, and in the following sections show several applications. Pedagogically, it is better to understand this calculation and carry it out in each application from the beginning, instead of using the general formula. We will not do this here, and simply use the general formula.

Consider two Gaussian random vectors $X = (X_i)_{i\le n}$ and $Y = (Y_i)_{i\le n}$ with the covariances

$$a_{i,j} = \mathbb{E}X_iX_j \text{ and } b_{i,j} = \mathbb{E}Y_iY_j. \tag{36}$$

If we suppose that X and Y are independent then, for $0\le t\le 1$,

$$Z(t) = \sqrt{t}X + \sqrt{1-t}Y \tag{37}$$

is a Gaussian random vector $Z(t) = (Z_i(t))_{i\le n}$ with the covariance

$$\mathbb{E}Z_i(t)Z_j(t) = ta_{i,j} + (1-t)b_{i,j},$$

which is a linear interpolation between the covariances of X and Y.

Take a twice continuously differentiable function

$$F = F\big((x_\ell)_{1\le\ell\le n}\big)\colon \mathbb{R}^n \to \mathbb{R},$$

whose derivatives satisfy some growth conditions determined below and, for $0 \le t \le 1$, let us consider its average along the above interpolation,

$$f(t) = \mathbb{E}F(Z(t)) = \mathbb{E}F(\sqrt{t}X + \sqrt{1-t}Y). \tag{38}$$

The end points of this interpolation are

$$f(0) = \mathbb{E}F(Y) \text{ and } f(1) = \mathbb{E}F(X),$$

which can often be compared in some useful way by analyzing the derivative of $f(t)$ for $0 < t < 1$. When the first partial derivatives have at most exponential growth as in (35), it is easy to check that one can interchange the derivative and integral to write

$$f'(t) = \mathbb{E}\frac{d}{dt}F(Z(t)) = \mathbb{E}\sum_{i\le n}\frac{\partial F}{\partial x_i}(Z(t))Z_i'(t) = \sum_{i\le n}\mathbb{E}\frac{\partial F}{\partial x_i}(Z(t))Z_i'(t).$$

Let us apply Gaussian integration by parts to each of the terms on the right hand side. Since the covariance

$$\mathbb{E}Z_i'(t)Z_j(t) = \mathbb{E}\Big(\frac{1}{2\sqrt{t}}X_i - \frac{1}{2\sqrt{1-t}}Y_i\Big)(\sqrt{t}X_j + \sqrt{1-t}Y_j) = \frac{1}{2}(a_{i,j} - b_{i,j}),$$

the Gaussian integration by parts formula (34) implies that

$$\mathbb{E}\frac{\partial F}{\partial x_i}(Z(t))Z_i'(t) = \frac{1}{2}\sum_{j\le n}(a_{i,j} - b_{i,j})\mathbb{E}\frac{\partial^2 F}{\partial x_i \partial x_j}(Z(t)),$$

if we assume that, for all $1 \le i, j \le n$,

$$\Big|\frac{\partial F}{\partial x_i}(x)\Big| \le c_1 e^{c_2|x|} \text{ and, either } \Big|\frac{\partial^2 F}{\partial x_i \partial x_j}(x)\Big| \le c_1 e^{c_2|x|} \text{ or } a_{i,j} = b_{i,j}. \tag{39}$$

Adding up over $i \le n$,

$$f'(t) = \frac{d}{dt}\mathbb{E}F(Z(t)) = \frac{1}{2}\sum_{i,j\le n}(a_{i,j} - b_{i,j})\mathbb{E}\frac{\partial^2 F}{\partial x_i \partial x_j}(Z(t)). \tag{40}$$

This *Gaussian interpolation formula* is very useful, as we will see in the applications below. Notice

that, if g is standard Gaussian random variable, and $f(t,x) = \mathbb{E}F(x+\sqrt{t}g)$ then

$$\frac{\partial f}{\partial t} = \frac{1}{2}\mathbb{E}F''(x+\sqrt{t}g) = \frac{1}{2}\frac{\partial^2 f}{\partial x^2},$$

(assuming some growth condition on F) so f solves the heat equation with the boundary condition $f(0,x) = F(x)$.

7 Gaussian concentration

Let us consider a Lipschitz function $F = F\big((x_\ell)_{1\le\ell\le n}\big) : \mathbb{R}^n \to \mathbb{R}$ such that, for some $L > 0$,

$$|F(x) - F(y)| \le L\|x-y\| \ \text{ for all } x, y \in \mathbb{R}^n. \tag{41}$$

The smallest such L is called the Lipschitz seminorm $\|F\|_{\mathrm{L}ip}$ of F.

Theorem 1 *If $g = (g_i)_{i\le n}$ is a standard Gaussian vector on $\mathbb{R}^n$ then, for any $t \ge 0$,*

$$\mathbb{P}\big(F(g) - \mathbb{E}F(g) \ge t\big) \le \exp\Big(-\frac{t^2}{4\|F\|^2_{\mathrm{L}ip}}\Big). \tag{42}$$

The constant 4 in the exponent is not optimal and can be replaced by 2. Notice that applying this result to $-F$ gives

$$\mathbb{P}\big(F(g) - \mathbb{E}F(g) \le -t\big) \le \exp\Big(-\frac{t^2}{4\|F\|^2_{\mathrm{L}ip}}\Big) \tag{43}$$

and, combining two inequalities (using the union bound), we get

$$\mathbb{P}\big(|F(g) - \mathbb{E}F(g)| \ge t\big) \le 2\exp\Big(-\frac{t^2}{4\|F\|^2_{\mathrm{L}ip}}\Big). \tag{44}$$

This shows that the probability that $F(g)$ deviates from its average $\mathbb{E}F(g)$ by more than $t \ge 0$ decreases exponentially fast in t^2, and this statement is independent of the dimension n. We will see many applications of this inequality, but first let us prove it using the Gaussian interpolation technique.

Proof. First, let us suppose that F is differentiable and its gradient is bounded by L, $\|\nabla F\| \le L$. Take any $\lambda \ge 0$. The Gaussian interpolation we would like to consider is of the form

$$f(t) = \mathbb{E}\exp\lambda\big(F(\sqrt{t}g^1 + \sqrt{1-t}g) - F(\sqrt{t}g^2 + \sqrt{1-t}g)\big), \tag{45}$$

where $g = (g_i)_{i\le n}$, $g^1 = (g^1_i)_{i\le n}$ and $g^2 = (g^2_i)_{i\le n}$ are three independent standard Gaussian vectors on $\mathbb{R}^n$. To calculate the derivative $f'(t)$, one can repeat the computation in the last section using the Gaussian integration by parts. Alternatively, we can use the general formula (40) above, but we

first have to set up this interpolation in the form (38). Let us consider a function $G\colon \mathbb{R}^{2n} \to \mathbb{R}$ given by the formula

$$G(x_1,\dots,x_n,x_{n+1},\dots,x_{2n}) = \exp\lambda\big(F(x_1,\dots,x_n) - F(x_{n+1},\dots,x_{2n})\big),$$

and let us define Gaussian random vectors X and Y on $\mathbb{R}^{2n}$ by $X = (g^1, g^2), Y = (g, g)$. Then the above interpolation can be rewritten in the form (38) as

$$f(t) = \mathbb{E}G(\sqrt{t}X + \sqrt{1-t}Y).$$

Notice that the covariance matrices of X and Y are equal to

$$A = \mathrm{Cov}(X) = \begin{bmatrix} I_n & 0 \\ 0 & I_n \end{bmatrix},\ B = \mathrm{Cov}(Y) = \begin{bmatrix} I_n & I_n \\ I_n & I_n \end{bmatrix},$$

where I_n is $n \times n$ identity matrix. Therefore,

$$A - B = \begin{bmatrix} 0 & -I_n \\ -I_n & 0 \end{bmatrix},$$

and the difference $a_{i,j} - b_{i,j}$ is non-zero only when $1 \le i \le n, j = i+n$ or $1 \le j \le n, i = j+n$, in which case $a_{i,j} - b_{i,j} = -1$. If we denote by $F_i = \frac{\partial F}{\partial x_i}$ the partial derivative in the ith coordinate of $F(x_1,\dots,x_n)$ for $1 \le i \le n$, then it is easy to see that

$$\frac{\partial^2 G}{\partial x_i \partial x_{i+n}} = \lambda^2 G F_i(x_1,\dots,x_n)F_i(x_{n+1},\dots,x_{2n}).$$

By assumption, $\|\nabla F\| \le L$, so the partial derivatives F_i are all bounded, and F grows at most linearly, $|F(x)| \le c_1 + c_2\|x\|$. So the growth conditions in (39) (for G) are satisfied, and the formula (40) gives

$$f'(t) = \lambda^2 \mathbb{E}G(\sqrt{t}X + \sqrt{1-t}Y)\sum_{i=1}^{n} F_i(\sqrt{t}g^1 + \sqrt{1-t}g)F_i(\sqrt{t}g^2 + \sqrt{1-t}g).$$

By the Cauchy-Schwarz inequality,

$$\sum_{i=1}^{n} F_i(\sqrt{t}g^1 + \sqrt{1-t}g)F_i(\sqrt{t}g^2 + \sqrt{1-t}g) \le \|\nabla F(\sqrt{t}g^1 + \sqrt{1-t}g)\|\|\nabla F(\sqrt{t}g^2 + \sqrt{1-t}g)\| \le L^2,$$

which means that (since G is positive)

$$f'(t) \le \lambda^2 L^2 \mathbb{E}G(\sqrt{t}X + \sqrt{1-t}Y) = \lambda^2 L^2 f(t).$$

This implies that the derivative

$$\big(f(t)e^{-\lambda^2 L^2 t}\big)' = e^{-\lambda^2 L^2 t}\big(f'(t) - \lambda^2 L^2 f(t)\big) \le 0,$$

so $f(t)e^{-\lambda^2 L^2 t}$ is decreasing and, therefore, $f(1) \le e^{\lambda^2 L^2} f(0)$. Recalling the definition (45), we see that $f(0) = 1$ and, therefore, we proved that

$$f(1) = \mathbb{E}\exp\lambda\big(F(g^1) - F(g^2)\big) \le e^{\lambda^2 L^2}.$$

The Gaussian interpolation and the assumption on the gradient, $\|\nabla F\| \le L$, have played their roles. What remains is to apply Jensen's and Markov's inequalities (12) and (11).

Since g^1 and g^2 are independent, integrating in g^2 first (let us denote this integral $\mathbb{E}_2$) and using that exp is convex,

$$\exp\lambda\big(F(g^1) - \mathbb{E}_2 F(g^2)\big) \le \mathbb{E}_2 \exp\lambda\big(F(g^1) - F(g^2)\big).$$

Integrating this in g^1, and using that g^1, g^2 have the same distribution as g, we get

$$\mathbb{E}\exp\lambda\big(F(g) - \mathbb{E}F(g)\big) \le \mathbb{E}\exp\lambda\big(F(g^1) - F(g^2)\big) \le e^{\lambda^2 L^2}.$$

Using Markov's inequality,

$$\mathbb{P}\big(F(g) - \mathbb{E}F(g) \ge t\big) \le e^{-\lambda t}\exp\lambda\big(F(g) - \mathbb{E}F(g)\big) \le e^{-\lambda t + \lambda^2 L^2}.$$

This inequality holds for any $\lambda \ge 0$, and minimizing the right hand side over λ, in other words, setting $\lambda = t/2L^2$, finishes the proof of (42) in the case when F is differentiable and its gradient is bounded by L, $\|\nabla F\| \le L$.

Finally, in the general case when we do not assume differentiability and only assume (41), one can use the standard smoothing technique. Namely, for $\varepsilon > 0$, we define

$$F_\varepsilon(x) = \mathbb{E}F(x + \varepsilon g), \tag{46}$$

where g is a standard Gaussian vector on $\mathbb{R}^n$. This function is also Lipschitz,

$$|F_\varepsilon(x) - F_\varepsilon(y)| \le \mathbb{E}|F(x+\varepsilon g) - F(y+\varepsilon g)| \le L\|x-y\|,$$

but it is now differentiable (even smooth), because

$$\mathbb{E}F(x+\varepsilon g) = \frac{1}{(\sqrt{2\pi})^n}\int_{\mathbb{R}^n} F(x+\varepsilon y)e^{-\|y\|^2/2}\,dy = \frac{1}{(\varepsilon\sqrt{2\pi})^n}\int_{\mathbb{R}^n} F(y)e^{-\|y-x\|^2/2\varepsilon^2}\,dy.$$

The case proved above shows that

$$\mathbb{P}\big(F_\varepsilon(g) - \mathbb{E}F_\varepsilon(g) \geq t\big) \leq \exp\Big(-\frac{t^2}{4\|F\|_{\mathrm{L}ip}^2}\Big). \tag{47}$$

Since F_ε approximates F uniformly for small $\varepsilon > 0$,

$$|F(x) - F_\varepsilon(x)| \leq \mathbb{E}|F(x) - F(x+\varepsilon g)| \leq L\varepsilon\mathbb{E}\|g\|,$$

letting $\varepsilon \downarrow 0$ finishes the proof of the general case. □

8 Examples of concentration inequalities

Example 1. Our first example will be a supremum of linear functionals in $\mathbb{R}^n$. Let us consider a bounded set A in $\mathbb{R}^n$, and consider the function

$$F(x) = \sup_{a\in A}(a,x) = \sup_{a\in A}\big(a_1x_1 + \ldots + a_nx_n\big).$$

Since

$$\begin{aligned}|F(x) - F(y)| &= \Big|\sup_{a\in A}\big(a_1x_1 + \ldots + a_nx_n\big) - \sup_{a\in A}\big(a_1y_1 + \ldots + a_ny_n\big)\Big| \\ &\leq \sup_{a\in A}\Big|\big(a_1(x_1-y_1) + \ldots + a_n(x_n-y_n)\big)\Big| \leq \sup_{a\in A}\|a\|\|x-y\|,\end{aligned}$$

the Lipschitz constant of F is bounded by $\|F\|_{\mathrm{L}ip} \leq \sup_{a\in A}\|a\|$. Therefore, if g is a standard Gaussian vector in $\mathbb{R}^n$ then

$$\mathbb{P}\Big(\big|\sup_{a\in A}(a,g) - \mathbb{E}\sup_{a\in A}(a,g)\big| \geq t\Big) \leq 2\exp\Big(-\frac{t^2}{4\sup_{a\in A}\|a\|^2}\Big). \tag{48}$$

Let us give a couple of special cases of this inequality.

Example 2. Let X by a Gaussian random vector in $\mathbb{R}^n$ with arbitrary covariance matrix. We know that X is equal in distribution to a linear transformation Cg of the standard Gaussian random vector g on $\mathbb{R}^n$. If $C_i = (c_{i,j})_{j\leq n}$ is the ith row of C then $X_i = C_ig$ and

$$\mathbb{E}X_i^2 = \mathbb{E}(C_ig)^2 = \sum_{j=1}^n c_{i,j}^2 = \|C_i\|^2.$$

If we apply the Example 1 with the set A given by the collection of rows $C_1,\ldots,C_n$, we get

$$\mathbb{P}\Big(\big|\max_{i\leq n}X_i - \mathbb{E}\max_{i\leq n}X_i\big| \geq t\Big) \leq 2\exp\Big(-\frac{t^2}{4\max_{i\leq n}\mathbb{E}X_i^2}\Big). \tag{49}$$

A related example is

$$\mathbb{P}\Big(\Big|\log\sum_{i=1}^n e^{X_i} - \mathbb{E}\log\sum_{i=1}^n e^{X_i}\Big| \geq t\Big) \leq 2\exp\Big(-\frac{t^2}{4\max_{i\leq n}\mathbb{E}X_i^2}\Big), \tag{50}$$

which follows by the same argument.

Example 3. *(Concentration of finite dimensional norms of Gaussian vectors)* Let us consider n-dimensional normed vector space E with the norm $\|\cdot\|_E$ and let $b_1,\ldots,b_n$ be any basis of E. If g is a standard Gaussian vector in $\mathbb{R}^n$, then

$$X = g_1 b_1 + \ldots + g_n b_n \tag{51}$$

is one way to produce a random vector in E. The norm of a vector $x \in E$ can be written as the supremum over the unit ball of the dual space E^* of linear functionals

$$\|x\|_E = \sup\Big\{\zeta(x) \mid \zeta \in E^*, \|\zeta\|_{E^*} \leq 1\Big\}, \tag{52}$$

and, in particular, the norm of the random vector X in (51) can be written as

$$\|X\|_E = \sup\Big\{g_1\zeta(b_1) + \ldots + g_n\zeta(b_n) \mid \zeta \in E^*, \|\zeta\|_{E^*} \leq 1\Big\}. \tag{53}$$

This functional is of the same form as in the Example 1, with the set A in $\mathbb{R}^n$ given by

$$A = \Big\{(\zeta(b_1),\ldots,\zeta(b_n) \mid \zeta \in E^*, \|\zeta\|_{E^*} \leq 1\Big\}.$$

The Lipschitz norm of this function is bounded by $\sup_{a\in A}\|a\|$, denoted by

$$\sigma(X) := \sup\Big\{\big(\zeta(b_1)^2 + \ldots + \zeta(b_n)^2\big)^{1/2} \mid \zeta \in E^*, \|\zeta\|_{E^*} \leq 1\Big\}. \tag{54}$$

This shows that the norm of a random vector X with Gaussian coordinates satisfies the following concentration inequality,

$$\mathbb{P}\Big(\big|\|X\|_E - \mathbb{E}\|X\|_E\big| \geq t\Big) \leq 2\exp\Big(-\frac{t^2}{4\sigma(X)^2}\Big). \tag{55}$$

Another common way to write this, replacing t by $t\mathbb{E}\|X\|_E$,

$$\mathbb{P}\Big(\big|\|X\|_E - \mathbb{E}\|X\|_E\big| \geq t\mathbb{E}\|X\|_E\Big) \leq 2\exp\Big(-\frac{t^2}{4}\Big(\frac{\mathbb{E}\|X\|_E}{\sigma(X)}\Big)^2\Big). \tag{56}$$

The quantity $d(X) := \big(\frac{\mathbb{E}\|X\|_E}{\sigma(X)}\big)^2$ is called the *concentration dimension* of X.

Example 4. *(Moment comparison for norms of Gaussian vectors)* One useful consequence of the concentration inequality in Example 3 is that the L^p-norms for $p \geq 1$ of the norm $\|X\|_E$ of the

Gaussian vector X in (51) are comparable to each other,

$$\big(\mathbb{E}\|X\|_E\big)^p \le \mathbb{E}\|X\|_E^p \le c_p\big(\mathbb{E}\|X\|_E\big)^p, \tag{57}$$

where c_p is some constant that depends only on p. The first inequality is just Jensen's inequality, so we only need to show the second one.

The following simple observation will be useful to us here and in the next section. If $p \ge 1$, and a random variable $Z \ge 0$ is nonnegative then we can write the p^{th} moment of Z as

$$\mathbb{E}Z^p = p\int_0^\infty x^{p-1}\mathbb{P}(Z \ge x)\,dx. \tag{58}$$

Suppose that f is the density function of Z on $\mathbb{R}^+$. If, for $x \ge 0$, we write $x^p = p\int_0^\infty s^{p-1}\,\mathrm{I}(s \le x)\,ds$ then switching the order of integration (Fubini's theorem),

$$\begin{aligned}\mathbb{E}Z^p = \int_0^\infty x^p f(x)\,dx &= p\int_0^\infty\int_0^\infty s^{p-1}\,\mathrm{I}(s \le x) f(x)\,ds dx \\ &= p\int_0^\infty\int_0^\infty s^{p-1}\,\mathrm{I}(s \le x) f(x)\,dx ds = p\int_0^\infty s^{p-1}\mathbb{P}(Z \ge s)\,ds.\end{aligned}$$

If X does not have density, just write the distribution $d\mathbb{P}(x)$ instead of $f(x)dx$.

If we denote $x^+ = \max(0,x)$ the positive part of x, then we can bound

$$\|X\|_E \le \mathbb{E}\|X\|_E + (\|X\|_E - \mathbb{E}\|X\|_E)^+.$$

Let us denote $Z = (\|X\|_E - \mathbb{E}\|X\|_E)^+$. Using the inequality $(a+b)^p \le 2^{p-1}(a^p + b^p)$, we get

$$\mathbb{E}\|X\|_E^p \le 2^{p-1}(\mathbb{E}\|X\|_E)^p + 2^{p-1}\mathbb{E}Z^p.$$

For simplicity of notation, let us write σ instead of $\sigma(X)$. By (55), we know that, $\mathbb{P}(Z \ge x) \le e^{-x^2/4\sigma^2}$ for $x \ge 0$ and, using (58),

$$\mathbb{E}Z^p \le p\int_0^\infty x^{p-1}e^{-x^2/4\sigma^2}\,dx = \sigma^p p\int_0^\infty t^{p-1}e^{-t^2/4}\,dt = a_p\sigma^p,$$

where we made the change of variables $t = \sigma x$ and then denoted by a_p the constant $p\int_0^\infty t^{p-1}e^{-t^2/4}\,dt$. Using the concentration inequality in the previous example, we proved that

$$\mathbb{E}\|X\|_E^p \le 2^{p-1}(\mathbb{E}\|X\|_E)^p + 2^{p-1}a_p\sigma(X)^p. \tag{59}$$

It remains to understand how to bound $\sigma(X)$ in (54). For $\zeta \in E^*$, the random variable

$$\zeta(X) = g_1\zeta(b_1) + \ldots + g_n\zeta(b_n)$$

is Gaussian with the variance

$$\mathbb{E}\zeta(X)^2 = \zeta(b_1)^2 + \ldots + \zeta(b_n)^2.$$

If Y is an arbitrary (centred) Gaussian random variable $\sim N(0, v^2)$ then

$$\mathbb{E}|Y| = \frac{1}{\sqrt{2\pi}v}\int_{-\infty}^{\infty}|x|e^{-x^2/2v^2}\,dx = \frac{v}{\sqrt{2\pi}}\int_{-\infty}^{\infty}|y|e^{-y^2/2}\,dy = v\sqrt{2/\pi},$$

which implies that $v^2 = \mathbb{E}Y^2 = \frac{\pi}{2}(\mathbb{E}|Y|)^2$. Using this for $Y = \zeta(X)$, we get

$$\zeta(b_1)^2 + \ldots + \zeta(b_n)^2 = \mathbb{E}\zeta(X)^2 = \frac{\pi}{2}(\mathbb{E}|\zeta(X)|)^2.$$

Taking supremum over $\zeta \in E^*$ such that $\|\zeta\|_{E^*} \le 1$, we get that

$$\sigma(X)^2 = \frac{\pi}{2}\big(\sup_{\zeta}\mathbb{E}|\zeta(X)|\big)^2 \le \frac{\pi}{2}\big(\mathbb{E}\sup_{\zeta}|\zeta(X)|\big)^2 = \frac{\pi}{2}\big(\mathbb{E}\|X\|_E\big)^2.$$

This means that $\sigma(X) \le \sqrt{\pi/2}\mathbb{E}\|X\|_E$, and plugging this into (59), finally proves (57) with the constant $c_p = 2^{p-1}(1 + a_p(\pi/2)^{p/2})$.

9 Gaussian comparison

The first two results we will prove here are generally called *Slepian's inequality*.

Theorem 2 *Let $X = (X_i)_{i\le n}$ and $Y = (Y_i)_{i\le n}$ be two Gaussian vectors in $\mathbb{R}^n$ such that*

1. $\mathbb{E}X_i^2 = \mathbb{E}Y_i^2$ *for all* $i \le n$,
2. $\mathbb{E}X_iX_j \le \mathbb{E}Y_iY_j$ *for all* $i, j \le n$.

Then, for any choice of parameters $(\lambda_i)_{i\le n}$,

$$\mathbb{P}\Big(\bigcap_{i=1}^{n}\{X_i \le \lambda_i\}\Big) \le \mathbb{P}\Big(\bigcap_{i=1}^{n}\{Y_i \le \lambda_i\}\Big) \tag{60}$$

and

$$\mathbb{E}\max_i X_i \ge \mathbb{E}\max_i Y_i. \tag{61}$$

What this means is that, if the coordinates of X have the same variance, but are less correlated, then they are less likely to stay below given thresholds (λ_i) and, therefore, their maximum is bigger on average. After we prove this result, we will give another proof of the second statement (61) under less restrictive assumptions.

Proof. Let us rewrite the indicator

$$\mathrm{I}\Big(\bigcap_{i=1}^n \{x_i \le \lambda_i\}\Big) = \prod_{i=1}^n \mathrm{I}\big(x_i \le \lambda_i\big).$$

Let us approximate each indicator $\mathrm{I}(x_i \le \lambda_i)$ by a smooth nonnegative decreasing function $\varphi_i(x_i)$. Define

$$\varphi(x) = \prod_{i=1}^n \varphi_i(x_i).$$

If we consider the interpolation $f(t) = \mathbb{E}\varphi(\sqrt{t}X + \sqrt{1-t}Y)$ and use the Gaussian interpolation formula (40), we will now check that the assumptions of the theorem about the covariances imply that $f'(t) \le 0$. Indeed, for $j \ne i$,

$$\frac{\partial^2}{\partial x_i \partial x_j} \prod_{\ell=1}^n \varphi_\ell(x_\ell) \ge 0,$$

because the derivatives applied to the factors i and j in the product will both be negative, since all functions φ_ℓ are decreasing. On the other hand, by assumption, the difference of the covariances $\mathbb{E}X_iX_j - \mathbb{E}Y_iY_j \le 0$ is negative in this case, so the corresponding term in (40) will be negative. The derivatives $\partial^2/\partial x_i^2$ are not important because the variances are equal and $\mathbb{E}X_i^2 - \mathbb{E}Y_i^2 = 0$. This proves that $f'(t) \le 0$ and, therefore,

$$f(1) = \mathbb{E}\varphi(X) \le f(0) = \mathbb{E}\varphi(Y).$$

Now, letting $\varphi_{i,j}$'s converge to the corresponding indicators, proves that

$$\mathbb{E}\mathrm{I}\Big(\bigcap_{i=1}^n \{X_i \le \lambda_i\}\Big) \le \mathbb{E}\mathrm{I}\Big(\bigcap_{i=1}^n \{Y_i \le \lambda_i\}\Big),$$

which is the same as (60). Let us show how this implies (61).

Notice that, if we take all $\lambda_i = \lambda$ then (60) can be rewritten as

$$\mathbb{P}\Big(\max_{i=1} X_i \le \lambda\Big) \le \mathbb{P}\Big(\max_{i=1} Y_i \le \lambda\Big). \tag{62}$$

Then the inequality (61) for the averages is an immediate consequence of integration by parts, as follows. Now we jsut need to use (58) in the previous section with $p = 1$: if $Z \ge 0$ then

$$\mathbb{E}Z = \int_0^\infty \mathbb{P}(Z \ge x)\, dx.$$

Let $X = \max_{i=1} X_i$ and $Y = \max_{i=1} Y_i$, and decompose X and Y into positive and negative parts,

$$X = X^+ - X^-,\ Y = Y^+ - Y^-.$$

If we take $\lambda \geq 0$, and apply the inequality (62) with λ and $-\lambda$, we get

$$\mathbb{P}\big(X^+ \geq \lambda\big) \geq \mathbb{P}\big(Y^+ \geq \lambda\big),\ \mathbb{P}\big(X^- \geq \lambda\big) \leq \mathbb{P}\big(Y^- \geq \lambda\big).$$

This implies that $\mathbb{E}X^+ \geq \mathbb{E}Y^+$, $\mathbb{E}X^- \leq \mathbb{E}Y^-$, and subtracting two inequalities we get (61). □

Next, we will prove (61) under less restrictive assumptions. If we write

$$\mathbb{E}(X_i - X_j)^2 = \mathbb{E}X_i^2 + \mathbb{E}X_j^2 - 2\mathbb{E}X_iX_j,\ \mathbb{E}(Y_i - Y_j)^2 = \mathbb{E}Y_i^2 + \mathbb{E}Y_j^2 - 2\mathbb{E}Y_iY_j$$

then, under the first assumption that all variances are equal, $\mathbb{E}X_i^2 = \mathbb{E}Y_i^2$, the second assumption that $\mathbb{E}X_iX_j \leq \mathbb{E}Y_iY_j$ for all $i, j \leq n$ is equivalent to

$$\mathbb{E}(X_i - X_j)^2 \geq \mathbb{E}(Y_i - Y_j)^2 \text{ for all } i, j \leq n.$$

We will now show that with this reformulation of the second assumption, we can drop the first assumption, which is often convenient in applications.

Theorem 3 *Let $X = (X_i)_{i\leq n}$ and $Y = (Y_i)_{i\leq n}$ be two Gaussian vectors in $\mathbb{R}^n$ such that*

$$\mathbb{E}(X_i - X_j)^2 \geq \mathbb{E}(Y_i - Y_j)^2 \textit{ for all } i, j \leq n.$$

Then, the inequality (61) holds, i.e.

$$\mathbb{E}\max_i X_i \geq \mathbb{E}\max_i Y_i. \tag{63}$$

Proof. We will apply the same interpolation as in the above proof, but to a special smooth approximation of the maximum function $\max_{i\leq n} x_i$ given by

$$F(x) = \frac{1}{\beta}\log\sum_{i\leq n} e^{\beta x_i},$$

for positive parameter $\beta > 0$. The reason we can view this as a smooth approximation of the maximum is because

$$\max_{i\leq n} x_i \leq F(x) = \frac{1}{\beta}\log\sum_{i\leq n} e^{\beta x_i} \leq \max_{i\leq n} x_i + \frac{\log n}{\beta}.$$

Indeed, if you keep only the largest term in the sum $\sum_{i\leq n} e^{\beta x_i}$ you get the lower bound, and if you replace each term in this sum by the largest one, you get the upper bound. Now, you can make the second term $\log n/\beta$ becomes as small as you like by taking β large. If we prove that

$$\mathbb{E}F(X) \geq \mathbb{E}F(Y),$$

we recover (63) by letting β go to infinity. To use the Gaussian interpolation (40), let us compute the derivatives of F. First of all,

$$p_i(x) := \frac{\partial F}{\partial x_i} = \frac{e^{\beta x_i}}{\sum_{j\le n} e^{\beta x_j}}.$$

Differentiating this, we see that

$$\frac{\partial^2 F}{\partial x_i^2} = \beta(p_i(x) - p_i(x)^2),\ \frac{\partial^2 F}{\partial x_i \partial x_j} = -\beta p_i(x)p_j(x) \text{ if } i \neq j,$$

Therefore, the Gaussian interpolation formula (40) gives (to simplify notation, we write p_i instead of $p_i(Z(t))$)

$$\begin{aligned} f'(t) &= \frac{d}{dt}\mathbb{E}F(Z(t)) = \frac{1}{2}\sum_{i,j\le n}(\mathbb{E}X_iX_j - \mathbb{E}Y_iY_j)\mathbb{E}\frac{\partial^2 F}{\partial x_i \partial x_j}(Z(t)) \\ &= \frac{\beta}{2}\sum_{i\le n}(\mathbb{E}X_i^2 - \mathbb{E}Y_i^2)\mathbb{E}(p_i - p_i^2) - \frac{\beta}{2}\sum_{i\neq j}(\mathbb{E}X_iX_j - \mathbb{E}Y_iY_j)\mathbb{E}p_ip_j. \end{aligned}$$

However, we chose F in such a way that,

$$\sum_{i\le n} p_i(x) = \sum_{i\le n}\frac{e^{\beta x_i}}{\sum_{j\le n} e^{\beta x_j}} = 1.$$

If we multiply both sides by $p_i(x)$ and subtract $p_i^2(x)$, we get

$$p_i(x) - p_i^2(x) = \sum_{j\neq i} p_i(x)p_j(x).$$

This means that

$$\sum_{i\le n}(\mathbb{E}X_i^2 - \mathbb{E}Y_i^2)\mathbb{E}(p_i - p_i^2) = \sum_{i\le n}(\mathbb{E}X_i^2 - \mathbb{E}Y_i^2)\mathbb{E}\sum_{j\neq i} p_ip_j = \sum_{i\neq j}(\mathbb{E}X_i^2 - \mathbb{E}Y_i^2)\mathbb{E}p_ip_j.$$

Switching indices i and j, this can also be written as $\sum_{i\neq j}(\mathbb{E}X_j^2 - \mathbb{E}Y_j^2)\mathbb{E}p_ip_j$, and taking average of the two, we get

$$\sum_{i\le n}(\mathbb{E}X_i^2 - \mathbb{E}Y_i^2)\mathbb{E}(p_i - p_i^2) = \frac{1}{2}\sum_{i\neq j}(\mathbb{E}X_i^2 + \mathbb{E}X_j^2 - \mathbb{E}Y_i^2 - \mathbb{E}Y_j^2)\mathbb{E}p_ip_j.$$

Plugging it into the above expression for the derivative $f'(t)$ and collecting the terms, we get

$$f'(t) = \frac{\beta}{4}\sum_{i\neq j}(\mathbb{E}(X_i - X_j)^2 - \mathbb{E}(Y_i - Y_j)^2)\mathbb{E}p_ip_j \ge 0,$$

since, by the assumption of the theorem, each term is positive. This implies that $f(1) \geq f(0)$ or, in other words, $\mathbb{E}F(X) \geq \mathbb{E}F(Y)$. This finishes the proof. □

Example. Let us consider a Gaussian process

$$X(t) = (t,g) = t_1 g_1 + \ldots + t_n g_n, \tag{64}$$

where $g = (g_1, \ldots, g_n)$ is a standard Gaussian random vector on $\mathbb{R}^n$, and $t = (t_1, \ldots, t_n) \in T$ for some bounded subset $T \subseteq \mathbb{R}^n$. Let us consider a 1-Lipschitz function $f\colon \mathbb{R}^n \to \mathbb{R}^n$, i.e.

$$\|f(t) - f(t')\| \leq \|t - t'\|,$$

and, if we write $f = (f_1, \ldots, f_n)$, consider a process

$$Y(t) = (f(t), g) = f_1(t) g_1 + \ldots + f_n(t) g_n. \tag{65}$$

Obviously,

$$\mathbb{E}(Y(t) - Y(t'))^2 = \|f(t) - f(t')\|^2 \leq \|t - t'\|^2 = \mathbb{E}(X(t) - X(t'))^2,$$

so Theorem 3 implies that

$$\mathbb{E} \sup_{t \in T} Y(t) = \mathbb{E} \sup_{t \in T} (f(t), g) \leq \mathbb{E} \sup_{t \in T} X(t) = \mathbb{E} \sup_{t \in T} (t, g). \tag{66}$$

A particular example would be to take a 1-Lipschitz function $\sigma\colon \mathbb{R} \to \mathbb{R}$, $|\sigma(x) - \sigma(y)| \leq |x - y|$, and set $f(t) = (\sigma(t_1), \ldots, \sigma(t_n))$. This example will appear in Project 8. □

Next, we will prove a more general minimax analogue of the first theorem above known as *Gordon's comparison inequality*.

Theorem 4 *Let $(X_{i,j})$ and $(Y_{i,j})$ be two Gaussian vectors indexed by $i \leq n$, $j \leq m$ such that*

1. $\mathbb{E}X_{i,j}^2 = \mathbb{E}Y_{i,j}^2$ *for all* i, j,
2. $\mathbb{E}X_{i,j}X_{i,\ell} \leq \mathbb{E}Y_{i,j}Y_{i,\ell}$ *for all* i, j, ℓ,
3. $\mathbb{E}X_{i,j}X_{k,\ell} \geq \mathbb{E}Y_{i,j}Y_{k,\ell}$ *for all* i, j, k, ℓ *such that* $i \neq k$.

Then, for any choice of parameters $(\lambda_{i,j})$,

$$\mathbb{P}\Big(\bigcup_{i=1}^{n} \bigcap_{j=1}^{m} \{X_{i,j} \leq \lambda_{i,j}\}\Big) \leq \mathbb{P}\Big(\bigcup_{i=1}^{n} \bigcap_{j=1}^{m} \{Y_{i,j} \leq \lambda_{i,j}\}\Big) \tag{67}$$

and

$$\mathbb{E} \min_i \max_j X_{i,j} \geq \mathbb{E} \min_i \max_j Y_{i,j}. \tag{68}$$

This reduces to Slepian's inequality when the index i takes only one value, so the condition 3 is empty.

Proof. Let us rewrite the indicator

$$\mathrm{I}\Big(\bigcup_{i=1}^{n}\bigcap_{j=1}^{m}\{x_{i,j}\le\lambda_{i,j}\}\Big)=1-\prod_{i=1}^{n}\Big(1-\prod_{j=1}^{m}\mathrm{I}\big(x_{i,j}\le\lambda_{i,j}\big)\Big).$$

Let us approximate each indicator $\mathrm{I}(x_{i,j}\le\lambda_{i,j})$ by a smooth nonnegative decreasing function $\varphi_{i,j}(x_{i,j})$. Define

$$\varphi(x)=1-\prod_{i=1}^{n}\Big(1-\prod_{j=1}^{m}\varphi_{i,j}(x_{i,j})\Big).$$

If we consider the interpolation $f(t)=\mathbb{E}\varphi(\sqrt{t}X+\sqrt{1-t}Y)$ and use the Gaussian interpolation formula (40), it is easy to check that the conditions of the Theorem on the covariance imply that $f'(t)\le 0$. Indeed, for $j\neq\ell$,

$$\frac{\partial^2\varphi}{\partial x_{i,j}\partial x_{i,\ell}}=\prod_{k\neq i}^{n}\Big(1-\prod_{p=1}^{m}\varphi_{k,p}(x_{k,p})\Big)\frac{\partial^2}{\partial x_{i,j}\partial x_{i,\ell}}\prod_{p=1}^{m}\varphi_{i,p}(x_{i,p})\ge 0,$$

because the derivatives applied to two factors in the last product will both be negative (since all functions $\varphi_{i,j}$ are decreasing). On the other hand, by assumption, the difference of the covariances $\mathbb{E}X_{i,j}X_{i,\ell}-\mathbb{E}Y_{i,j}Y_{i,\ell}\le 0$ is negative in this case, so the corresponding term in (40) will be negative. Similarly, for $i\neq k$,

$$\frac{\partial^2\varphi}{\partial x_{i,j}\partial x_{k,\ell}}=-\prod_{k'\neq i,k}^{n}\Big(1-\prod_{p=1}^{m}\varphi_{k',p}(x_{k',p})\Big)\frac{\partial}{\partial x_{i,j}}\prod_{p=1}^{m}\varphi_{i,p}(x_{i,p})\frac{\partial}{\partial x_{k,\ell}}\prod_{p=1}^{m}\varphi_{k,p}(x_{k,p})\le 0.$$

By assumption, the difference of the covariances $\mathbb{E}X_{i,j}X_{k,\ell}-\mathbb{E}Y_{i,j}Y_{k,\ell}\ge 0$ is positive in this case, so the corresponding term in (40) will again be negative. This proves that $f'(t)\le 0$ and, therefore,

$$f(1)=\mathbb{E}\varphi(X)\le f(0)=\mathbb{E}\varphi(Y).$$

Now, letting $\varphi_{i,j}$'s converge to the corresponding indicators, proves that

$$\mathbb{E}\mathrm{I}\Big(\bigcup_{i=1}^{n}\bigcap_{j=1}^{m}\{X_{i,j}\le\lambda_{i,j}\}\Big)\le\mathbb{E}\mathrm{I}\Big(\bigcup_{i=1}^{n}\bigcap_{j=1}^{m}\{Y_{i,j}\le\lambda_{i,j}\}\Big),$$

which is the same as (67). If we take all $\lambda_{i,j}=\lambda$, this can be rewritten as

$$\mathbb{P}\Big(\min_i\max_j X_{i,j}\le\lambda\Big)\le\mathbb{P}\Big(\min_i\max_j Y_{i,j}\le\lambda\Big).$$

and (68) follows by integration by parts as before. □

For the last statement of the previous theorem, one can also remove the assumption about the equalities of variances.

Theorem 5 *Let $(X_{i,j})$ and $(Y_{i,j})$ be two Gaussian vectors indexed by $i \le n$, $j \le m$ such that*

1. $\mathbb{E}(X_{i,j} - X_{i,\ell})^2 \ge \mathbb{E}(Y_{i,j} - Y_{i,\ell})^2$ *for all* i, j, ℓ,
2. $\mathbb{E}(X_{i,j} - X_{k,\ell})^2 \le \mathbb{E}(Y_{i,j} - Y_{k,\ell})^3$ *for all* i, j, k, ℓ *such that* $i \neq k$.

Then,

$$\mathbb{E} \min_i \max_j X_{i,j} \ge \mathbb{E} \min_i \max_j Y_{i,j}. \tag{69}$$

To prove this inequality, one can use that

$$F(x) = \frac{1}{\beta} \log \sum_i \frac{1}{\sum_j e^{\beta X_{i,j}}} \to - \min_i \max_j X_{i,j} \text{ as } \beta \to \infty,$$

and, as in the proof of Theorem 3, show that $\mathbb{E}F(X) \le \mathbb{E}F(Y)$. The calculation is a bit more involved but straightforward. Letting $\beta \to \infty$ proves (69).

10 Comparison of bilinear and linear forms

Let us consider the following random bilinear form

$$X(t,u) = \sum_{i=1}^n \sum_{j=1}^m g_{i,j} t_i u_j, \tag{70}$$

where $(g_{i,j})_{i \le n, j \le m}$ are i.i.d. standard Gaussian random variables and parameters

$$t = (t_1, \ldots, t_n) \in \mathbb{R}^n \text{ and } u = (u_1, \ldots, u_m) \in \mathbb{R}^m.$$

In various applications, one is interested in quantities of the form

$$\max_{t \in T} \max_{u \in U} X(t,u), \ \min_{t \in T} \max_{u \in U} X(t,u) \text{ or } \min_{u \in U} \max_{t \in T} X(t,u)$$

for some sets of parameters $T \subseteq \mathbb{R}^n$ and $U \subseteq \mathbb{R}^m$, which can be difficult to calculate or estimate directly. It turns out that one can use comparison inequalities from the previous section to relate these quantities to similar quantities for another, linear or 'almost' linear form. We will give several examples below.

Example 1. Let us first consider the following random process

$$Y(t,u) = \|u\| \sum_{i=1}^n h_i t_i + \|t\| \sum_{j=1}^m g_j u_j, \tag{71}$$

where $h_1,\dots,h_n,g_1,\dots,g_m$ are i.i.d. standard Gaussian random variables. This is not quite a linear form because of the factors $\|t\|$ and $\|u\|$, but the dependence on t and u is simpler and one can often analyze more easily the quantities

$$\max_{t\in T}\max_{u\in U} Y(t,u),\ \min_{t\in T}\max_{u\in U} Y(t,u)\ \text{ or }\ \min_{u\in U}\max_{t\in T} Y(t,u).$$

Let us take any two pairs of parameters (t,u) and (t',u') and compute

$$\begin{aligned}\mathbb{E}\big(X(t,u)-X(t',u')\big)^2 &= \mathbb{E}\Big(\sum_{i=1}^n\sum_{j=1}^m g_{i,j}(t_iu_j-t'_iu'_j)\Big)^2\\ &= \sum_{i=1}^n\sum_{j=1}^m (t_iu_j-t'_iu'_j)^2 = \|t\|^2\|u\|^2+\|t'\|^2\|u'\|^2-2(t,t')(u,u'). \qquad (72)\end{aligned}$$

Similarly, one can compute

$$\begin{aligned}\mathbb{E}\big(Y(t,u)-Y(t',u')\big)^2 &= \mathbb{E}\Big(\sum_{i=1}^n h_i(\|u\|t_i-\|u'\|t'_i)\Big)^2+\mathbb{E}\Big(\sum_{j=1}^m g_j(\|t\|u_j-\|t'\|u'_j)\Big)^2\\ &= 2\|t\|^2\|u\|^2+2\|t'\|^2\|u'\|^2-2\|u\|\|u'\|(t,t')-2\|t\|\|t'\|(u,u'). \qquad (73)\end{aligned}$$

Subtracting and rearranging the terms, it is easy to see that

$$\begin{aligned}&\mathbb{E}\big(Y(t,u)-Y(t',u')\big)^2-\mathbb{E}\big(X(t,u)-X(t',u')\big)^2 =\\ &=\Big(\|t\|\|u\|-\|t'\|\|u'\|\Big)^2+2\Big(\|t\|\|t'\|-(t,t')\Big)\Big(\|u\|\|u'\|-(u,u')\Big). \qquad (74)\end{aligned}$$

By the Cauchy-Schwarz inequality, $(t,t')\le\|t\|\|t'\|$ and $(u,u')\le\|u\|\|u'\|$, so

$$\mathbb{E}\big(Y(t,u)-Y(t',u')\big)^2\ge\mathbb{E}\big(X(t,u)-X(t',u')\big)^2. \qquad (75)$$

By Theorem 3 in the previous section,

$$\mathbb{E}\max_{t\in T}\max_{u\in U} X(t,u)\le\mathbb{E}\max_{t\in T}\max_{u\in U} Y(t,u). \qquad (76)$$

This inequality, for example, will be used in the Project 11. □

In order to apply the results for the minimax $\min_{t\in T}\max_{u\in U}$, we need the reverse inequality for $t=t'$. Because of (75), we can only hope to get the equality

$$\mathbb{E}\big(Y(t,u)-Y(t,u')\big)^2=\mathbb{E}\big(X(t,u)-X(t,u')\big)^2. \qquad (77)$$

There are a couple of ways this can be achieved, as we will see in the next examples.

Example 2. One way to do this is to modify the definition of the bilinear form slightly and consider

$$X^+(t,u) = \sum_{i=1}^{n}\sum_{j=1}^{m} g_{i,j}t_i u_j + z\|t\|\|u\|, \tag{78}$$

where z is a standard Gaussian random variable independent of all $g_{i,j}$. Including this additional term will add $(\|t\|\|u\| - \|t'\|\|u'\|)^2$ to (72) and, as a result, (74) will become

$$\begin{aligned}&\mathbb{E}\big(Y(t,u) - Y(t',u')\big)^2 - \mathbb{E}\big(X^+(t,u) - X^+(t',u')\big)^2 = \\ &= 2\Big(\|t\|\|t'\| - (t,t')\Big)\Big(\|u\|\|u'\| - (u,u')\Big).\end{aligned} \tag{79}$$

This is equal to zero when $t = t'$ so, by Theorem 5,

$$\mathbb{E}\min_{t\in T}\max_{u\in U} X^+(t,u) \geq \mathbb{E}\min_{t\in T}\max_{u\in U} Y(t,u). \tag{80}$$

Notice that the inequality is reversed in this case and together with (76), we can write

$$\mathbb{E}\min_{t\in T}\max_{u\in U} Y(t,u) \leq \mathbb{E}\min_{t\in T}\max_{u\in U} X^+(t,u) \leq \mathbb{E}\max_{t\in T}\max_{u\in U} X^+(t,u) \leq \mathbb{E}\max_{t\in T}\max_{u\in U} Y(t,u). \tag{81}$$

Moreover, if we take $t' = 0$ in (79), we get that

$$\mathbb{E}X^+(t,u)^2 = \mathbb{E}Y(t,u)^2 \text{ for all } (t,u),$$

so we are in a position to apply Theorem 4 to compare the probabilities as in (67),

$$\mathbb{P}\Big(\bigcup_{t\in T}\bigcap_{u\in U}\big\{X^+(t,u) \leq \lambda(t,u)\big\}\Big) \leq \mathbb{P}\Big(\bigcup_{t\in T}\bigcap_{u\in U}\big\{Y(t,u) \leq \lambda(t,u)\big\}\Big), \tag{82}$$

for arbitrary function $\lambda(t,u)$. Notice that (79) is also equal to zero if $u = u'$ so, by the same logic, we can switch the role of the parameters t and u,

$$\mathbb{E}\min_{u\in U}\max_{t\in T} Y(t,u) \leq \mathbb{E}\min_{u\in U}\max_{t\in T} X^+(t,u) \leq \mathbb{E}\max_{u\in U}\max_{t\in T} X^+(t,u) \leq \mathbb{E}\max_{u\in U}\max_{t\in T} Y(t,u) \tag{83}$$

and

$$\mathbb{P}\Big(\bigcup_{u\in U}\bigcap_{t\in T}\big\{X^+(t,u) \leq \lambda(t,u)\big\}\Big) \leq \mathbb{P}\Big(\bigcup_{u\in U}\bigcap_{t\in T}\big\{Y(t,u) \leq \lambda(t,u)\big\}\Big). \tag{84}$$

This example will be used for example, in Project 6. □

Example 3. There is another special case that is very useful, when

$$\|t\| = a \text{ for all } t \in T,\ \|u\| \leq b \text{ for all } u \in U, \tag{85}$$

for some constants $a, b > 0$. In other words, T is a subset of the sphere or radius a in $\mathbb{R}^n$ and U is a subset of the ball of radius b in $\mathbb{R}^m$. In this case, we will modify the definition (71) slightly and

consider a proper random linear form

$$Y^+(t,u) = b\sum_{i=1}^{n} h_i t_i + a\sum_{j=1}^{m} g_j u_j. \tag{86}$$

As before, by a direct calculation, one can check that

$$\mathbb{E}\big(Y^+(t,u) - Y^+(t',u')\big)^2 - \mathbb{E}\big(X(t,u) - X(t',u')\big)^2 = 2\big(a^2 - (t,t')\big)\big(b^2 - (u,u')\big). \tag{87}$$

By the Cauchy-Schwarz inequality, this difference is nonnegative. Moreover, it is equal to zero when $t = t'$ by the assumption that $\|t\| = a$. This implies, by Theorem 3 and Theorem 5,

$$\mathbb{E}\min_{t\in T}\max_{u\in U} Y^+(t,u) \le \mathbb{E}\min_{t\in T}\max_{u\in U} X(t,u) \le \mathbb{E}\max_{t\in T}\max_{u\in U} X(t,u) \le \mathbb{E}\max_{t\in T}\max_{u\in U} Y^+(t,u). \tag{88}$$

This comparison is used in Projects 3 and 4. □

Example 4. In the setting of the Example 1, let us suppose that

$$\|t\| = a \ \text{ for all } \ t \in T, \tag{89}$$

i.e. T is a subset of the sphere or radius a in $\mathbb{R}^n$. If we take $u = u'$ in (74), we get the equality

$$\mathbb{E}\big(Y(t,u) - Y(t',u)\big)^2 = \mathbb{E}\big(X(t,u) - X(t',u)\big)^2. \tag{90}$$

As in (83), this implies

$$\mathbb{E}\min_{u\in U}\max_{t\in T} Y(t,u) \le \mathbb{E}\min_{u\in U}\max_{t\in T} X(t,u) \le \mathbb{E}\max_{u\in U}\max_{t\in T} X(t,u) \le \mathbb{E}\max_{u\in U}\max_{t\in T} Y(t,u). \tag{91}$$

This example is not used in any projects, but we mention it here just in case. □

Example 5. Let us take $T = S^{n-1}$ and $U = S^{m-1}$ to be unit spheres in $\mathbb{R}^n$ and $\mathbb{R}^m$. We will use Example 3 with $a = b = 1$ and the role of t and u reversed in (88). First of all,

$$\min_{\|u\|=1}\max_{\|t\|=1}\Big(\sum_{i=1}^{n} h_i t_i + \sum_{j=1}^{m} g_j u_j\Big) = \max_{\|t\|=1}(h,t) - \max_{\|u\|=1}(g,u) = \|h\| - \|g\|,$$

so

$$\mathbb{E}\min_{u\in U}\max_{t\in T} Y^+(t,u) = \mathbb{E}\|h\| - \mathbb{E}\|g\|.$$

This can be computed explicitly, and it is well-known that

$$\frac{n}{\sqrt{n+1}} \le \mathbb{E}\|h\| = \frac{\sqrt{2}\Gamma(\frac{n+1}{2})}{\Gamma(\frac{n}{2})} \le \sqrt{n} \tag{92}$$

(we do not reproduce this here). The same holds for $\mathbb{E}\|g\|$ with n replaced by m. On the other hand,

$$\min_{u\in U}\max_{t\in T} X(t,u) = \min_{\|u\|=1}\Big(\sum_{i=1}^{n}\Big(\sum_{j=1}^{m} g_{i,j}u_j\Big)^2\Big)^{1/2},$$

which we can rewrite as follows. Let us introduce the notation $g_i = (g_{i,1},\dots,g_{i,m})^T \in \mathbb{R}^m$ and

$$Z = \frac{1}{n}\sum_{i=1}^{n} g_i g_i^T. \tag{93}$$

Vectors g_i are i.i.d. standard Gaussian in $\mathbb{R}^m$ and $m\times m$ matrix Z is their sample covariance matrix. We can then rewrite

$$\sum_{j=1}^{n}\Big(\sum_{j=1}^{m} g_{i,j}u_j\Big)^2 = \sum_{j=1}^{n}(g_i,u)^2 = \sum_{j=1}^{n} u^T g_i g_i^T u = \sum_{j=1}^{n}(g_i g_i^T u,u) = n(Zu,u).$$

Since Z is symmetric and positive semi-definite,

$$\min_{\|u\|=1}(Zu,u) = \lambda_{\min}(Z) \ge 0,$$

where $\lambda_{\min}(Z)$ is the smallest eigenvalue of Z. The first inequality in (88) shows that

$$\mathbb{E}\|h\| - \mathbb{E}\|g\| \le \mathbb{E}\sqrt{n\lambda_{\min}(Z)}, \tag{94}$$

and using (92), we can write

$$\sqrt{\frac{n}{n+1}} - \sqrt{\frac{m}{n}} \le \mathbb{E}\sqrt{\lambda_{\min}(Z)}. \tag{95}$$

When $n = (1+\varepsilon)m$, i.e. the sample size n is relatively bigger than the dimension m of our space, the left hand side is separated away from zero, so the smallest eigenvalue $\lambda_{\min}(Z)$ is separated away from zero, at least on average. To obtain similar statement in probability, we can use Example 2. The only difference will be that

$$\min_{u\in U}\max_{t\in T} X^+(t,u) = \sqrt{n\lambda_{\min}(Z)} + z,$$

and using (84) with $\lambda(t,u)\equiv\lambda$, we get

$$\mathbb{P}\Big(\sqrt{n\lambda_{\min}(Z)} + z \le \lambda\Big) \le \mathbb{P}\Big(\|h\| - \|g\| \le \lambda\Big). \tag{96}$$

If we take $\lambda = t\sqrt{n}$ and flip the inequality, we get

$$\mathbb{P}\Big(\sqrt{\lambda_{\min}(Z)} + \frac{z}{\sqrt{n}} \ge t\Big) \ge \mathbb{P}\Big(\frac{\|h\| - \|g\|}{\sqrt{n}} \ge t\Big). \tag{97}$$

By the law of large number, $(\|h\| - \|g\|)/\sqrt{n} \approx 1 - \sqrt{m/n}$, so the second probability will be close to one if n is such that $1 - \sqrt{m/n} > t$, or $n > m/(1-t)^2$. Since $z/\sqrt{n}$ is negligible for large n, this shows that with probability close to one, $\lambda_{\min}(Z) \geq t^2$. This shows that when $n = (1+\varepsilon)m$, $\lambda_{\min}(Z)$ is separated away from zero with high probability, and not only on average.

We can reformulate these results for general Gaussian vectors on $\mathbb{R}^m$ with the covariance C. If $X_i = Ag_i$ for some matrix A such that $C = AA^T$, then $(X_i)_{i \leq n}$ are i.i.d. $N(0,C)$ and

$$Z_C = \frac{1}{n}\sum_{i=1}^{n} X_i X_i^T = AZA^T \tag{98}$$

is their sample covariance matrix, where Z was defined in (93). Since

$$\begin{aligned}\lambda_{\min}(Z_C) &= \inf_{\|u\|=1}(Z_C u, u) = \inf_{\|u\|=1}(AZA^T u, u) = \inf_{\|u\|=1}(ZA^T u, A^T u) \\ &\geq \lambda_{\min}(Z)\inf_{\|u\|=1}(A^T u, A^T u) = \lambda_{\min}(Z)\inf_{\|u\|=1}(Cu,u) = \lambda_{\min}(Z)\lambda_{\min}(C),\end{aligned}$$

the statements we obtained above for the sample covariance matrix Z can be transferred to similar statements for Z_C, only now scaled by $\lambda_{\min}(C)$. □

11 The central limit theorem on $\mathbb{R}$

Let us begin with the definition of convergence in distribution for real-valued random variables. Let X and X_n for $n \geq 1$ be some random variables on $\mathbb{R}$, and let $\mathbb{P}$ and $\mathbb{P}_n$ for $n \geq 1$ be their distributions correspondingly. In other words, we know how to calculate probabilities $\mathbb{P}(A) = \mathbb{P}(X \in A)$ for any measurable set in $\mathbb{R}$, and the same for X_n. This means that we also know how to calculate expectations

$$\mathbb{E}f(X) = \int_{\mathbb{R}} f(x)\,d\mathbb{P}(x),\ \mathbb{E}f(X_n) = \int_{\mathbb{R}} f(x)\,d\mathbb{P}_n(x).$$

Let us denoted the set of all continuous and bounded functions on $\mathbb{R}$ by

$$C_b = \big\{f\colon \mathbb{R} \to \mathbb{R} - \text{continuous and bounded}\big\}.$$

We say that $X_n \to X$ *in distribution*, or $\mathbb{P}_n \to \mathbb{P}$ *weakly* if

$$\lim_{n\to\infty} \mathbb{E}f(X_n) = \lim_{n\to\infty}\int_{\mathbb{R}} f\,d\mathbb{P}_n = \mathbb{E}f(X) = \int_{\mathbb{R}} f\,d\mathbb{P} \ \text{ for all } \ f \in C_b. \tag{99}$$

This is often denoted by $\mathbb{P}_n \xrightarrow{d} \mathbb{P}$ or $\mathbb{P}_n \Longrightarrow \mathbb{P}$.

The *cumulative distribution functions* of X and X_n are defined by

$$F(t) = \mathbb{P}\big((-\infty,t]\big) = \mathbb{P}(X \leq t),\ F(t) = \mathbb{P}_n\big((-\infty,t]\big) = \mathbb{P}(X_n \leq t).$$

We will now show that the convergence of probability measures can be expressed in terms of their

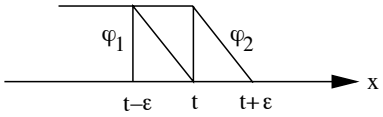
cumulative distribution functions.

Theorem 6 *$\mathbb{P}_n \xrightarrow{d} \mathbb{P}$ if and only if $F_n(t) \to F(t)$ for any point of continuity t of F.*

Proof. "$\Longrightarrow$" Suppose that (99) holds. Let us approximate the indicator $\mathrm{I}(x \le t)$ by continuous functions so that

$$\mathrm{I}(x \le t-\varepsilon) \le \varphi_1(x) \le \mathrm{I}(x \le t) \le \varphi_2(x) \le \mathrm{I}(x \le t+\varepsilon),$$

as in the figure below.



Obviously, $\varphi_1, \varphi_2 \in C_b$. Then, using (99) for φ_1 and φ_2,

$$F(t-\varepsilon) \le \int \varphi_1\, dF = \lim_{n\to\infty} \int \varphi_1\, dF_n \le \lim_{n\to\infty} F_n(t) \le \lim_{n\to\infty} \int \varphi_2\, dF_n = \int \varphi_2\, dF \le F(t+\varepsilon).$$

Therefore, for any $\varepsilon > 0$, we can write

$$F(t-\varepsilon) \le \lim_{n\to\infty} F_n(t) \le F(t+\varepsilon).$$

More carefully, we should write lim inf and lim sup but, since t is a point of continuity of F, letting $\varepsilon \downarrow 0$ proves that the limit $\lim_{n\to\infty} F_n(t)$ exists and is equal to $F(t)$.

"$\Longleftarrow$" Let $PC(F)$ be the set of all points of continuity of F. Since F is monotone, the set $PC(F)$ is dense in $\mathbb{R}$. Take M large enough such that both $M, -M \in PC(F)$ and $\mathbb{P}((-M,M]^c) \le \varepsilon$. Clearly, for large enough $n \ge 1$ we have $\mathbb{P}_n((-M,M]^c) \le 2\varepsilon$. For any $k > 1$, consider a sequence of points $-M = x_1^k \le x_2^k \le \ldots \le x_k^k = M$ such that all $x_i \in PC(F)$ and $\max_i |x_{i+1}^k - x_i^k| \to 0$ as $k \to \infty$. Given a function $f \in C_b$, consider an approximating function

$$f_k(x) = \sum_{1<i\le k} f(x_i^k)\, \mathrm{I}\big(x \in (x_{i-1}^k, x_i^k]\big) + 0 \cdot \mathrm{I}\big(x \notin (-M,M]\big).$$

Since f in continuous, we get that

$$\delta_k(M) := \sup_{|x|\le M} \big|f_k(x) - f(x)\big| \to 0,\ k \to \infty.$$

Since all $x_i^k \in PC(F)$, by assumption, we can write

$$\int f_k\, dF_n = \sum_{1<i\le k} f_k(x_i^k)\big(F_n(x_i^k) - F_n(x_{i-1}^k)\big) \xrightarrow{n\to\infty} \sum_{1<i\le k} f_k(x_i^k)\big(F(x_i^k) - F(x_{i-1}^k)\big) = \int f_k\, dF.$$

On the other hand,

$$\left|\int f\,dF - \int f_k\,dF\right| \le \|f\|_\infty \mathbb{P}\big((-M,M]^c\big) + \delta_k(M) \le \|f\|_\infty \varepsilon + \delta_k(M)$$

and, similarly, for large enough $n \ge 1$,

$$\left|\int f\,dF_n - \int f_k\,dF_n\right| \le \|f\|_\infty \mathbb{P}_n\big((-M,M]^c\big) + \delta_k(M) \le \|f\|_\infty 2\varepsilon + \delta_k(M).$$

Letting $n \to \infty$, then $k \to \infty$ and, finally, $\varepsilon \downarrow 0$ (or $M \uparrow \infty$), proves that $\int f\,dF_n \to \int f\,dF$. □

Notice that in the above proof, we could have used smooth approximations of indicators, let's say, with the third bounded derivative, instead of simply continuous approximations. This means that, in order to check convergence in distribution, it is enough to check (99) for bounded functions with the third bounded derivative.

Let us now consider a sequence X_n for $n \ge 1$ of independent random variables, which all have the same distribution. Such random variables are called *independent identically distributed*, or simply i.i.d.. We will suppose they have finite (absolute) third moment,

$$\mathbb{E}|X_1|^3 < \infty,$$

and denote their mean and variance by

$$\mu = \mathbb{E}X_1 \text{ and } \sigma^2 = \mathrm{Var}(X_1) < \infty.$$

We will denote their sum by $S_n = X_1 + \ldots + X_n$, and introduce a notation

$$Z_n := \frac{S_n - n\mu}{\sqrt{n\sigma^2}} = \frac{1}{\sqrt{n}} \sum_{i=1}^n \frac{X_i - \mu}{\sigma}. \tag{100}$$

The following result is known as the *Central Limit Theorem*. We will prove it as a consequence of the stability property of the normal distribution proved above.

Theorem 7 *The distribution of Z_n converges weakly to the standard Gaussian distribution $N(0,1)$.*

Remark. One can easily modify the proof to get rid of the unnecessary assumption $\mathbb{E}|X_1|^3 < \infty$. The condition $\sigma^2 = \mathrm{Var}(X_1) < \infty$ suffices.

Proof. First of all, notice that the random variables $(X_i - \mu)/\sigma$ have mean 0 and variance 1 so, by changing names, it is enough to prove the result for

$$Z_n = \frac{1}{\sqrt{n}} \sum_{i=1}^n X_i$$

under the assumption that $\mathbb{E}X_1 = 0, \mathbb{E}X_1^2 = 1$. Let $(g_i)_{i \ge 1}$ be independent standard normal random

variables. Then, by the stability property,

$$Z = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} g_i$$

also has standard normal distribution $N(0,1)$. If, for $1 \leq m \leq n+1$, we define

$$T_m = \frac{1}{\sqrt{n}} \big(g_1 + \ldots + g_{m-1} + X_m + \ldots + X_n \big)$$

then, for any bounded function $f : \mathbb{R} \to \mathbb{R}$, we can write

$$\big| \mathbb{E} f(Z_n) - \mathbb{E} f(Z) \big| = \Big| \sum_{m=1}^{n} \big(\mathbb{E} f(T_m) - \mathbb{E} f(T_{m+1}) \big) \Big| \leq \sum_{m=1}^{n} \big| \mathbb{E} f(T_m) - \mathbb{E} f(T_{m+1}) \big|.$$

If we introduce the notation

$$S_m = \frac{1}{\sqrt{n}} \big(g_1 + \ldots + g_{m-1} + X_{m+1} + \ldots + X_n \big)$$

then $T_m = S_m + X_m/\sqrt{n}$ and $T_{m+1} = S_m + g_m/\sqrt{n}$. By the comment above, we can suppose that $f \in C_b$ and has uniformly bounded third derivative. Then, by Taylor's formula,

$$\Big| f(T_m) - f(S_m) - \frac{f'(S_m) X_m}{\sqrt{n}} - \frac{f''(S_m) X_m^2}{2n} \Big| \leq \frac{\|f'''\|_\infty |X_m|^3}{6 n^{3/2}}$$

and

$$\Big| f(T_{m+1}) - f(S_m) - \frac{f'(S_m) g_m}{\sqrt{n}} - \frac{f''(S_m) g_m^2}{2n} \Big| \leq \frac{\|f'''\|_\infty |g_m|^3}{6 n^{3/2}}.$$

Notice that S_m is independent of X_m and g_m and, therefore,

$$\mathbb{E} f'(S_m) X_m = \mathbb{E} f'(S_m) \mathbb{E} X_m = 0 = \mathbb{E} f'(S_m) \mathbb{E} g_m = \mathbb{E} f'(S_m) g_m$$

and, similarly,

$$\mathbb{E} f''(S_m) X_m^2 = \mathbb{E} f''(S_m) \mathbb{E} X_m^2 = \mathbb{E} f''(S_m) = \mathbb{E} f''(S_m) \mathbb{E} g_m^2 = \mathbb{E} f''(S_m) g_m^2.$$

As a result, taking expectations and subtracting the above inequalities, we get

$$\big| \mathbb{E} f(T_m) - \mathbb{E} f(T_{m+1}) \big| \leq \frac{\|f'''\|_\infty (\mathbb{E}|X_1|^3 + \mathbb{E}|g_1|^3)}{6 n^{3/2}}.$$

Adding up over $1 \leq m \leq n$, we have shown that

$$\big| \mathbb{E} f(Z_n) - \mathbb{E} f(Z) \big| \leq \frac{\|f'''\|_\infty (\mathbb{E}|X_1|^3 + \mathbb{E}|g_1|^3)}{6 \sqrt{n}},$$

which implies that $\lim_{n\to\infty} \mathbb{E}f(Z_n) = \mathbb{E}f(Z)$. This finishes the proof. □