B8.6 High-Dimensional Probability

BY ZHONGMIN QIAN

December 30, 2025

Contents

1 Introduction **2**

2 Measures, integration and probability distributions **4**
2.1 Measures and Lebesgue's integration 4

3 General concentration inequalities **11**
3.1 One-dimensional distributions 11
3.2 The Cramér theorem 13

4 Gaussian distributions **19**
4.1 High-dimensional normal distributions 19
4.1.1 Sudakov-Fernique's inequality 23
4.2 Heat kernel 25
4.3 Geometric properties of normal distributions 27

5 The Ornstein-Uhlenbeck semi-group **30**
5.1 The Mehler formula 30
5.2 Entropy and the logarithmic Sobolev inequality 38
5.3 Poincaré inequality 40
5.4 The concentration inequality 41
5.5 Estimates of exponential type 44
5.6 Gaussian isoperimetric inequality 48

6 Brunn-Minkowski's inequality, Isoperimetric inequality **55**
6.1 Prékopa-Leindler's inequality 55
6.2 Brunn-Minkowski's theorem 61

7 Appendix **62**
7.1 Analysis of Lebesgue's measure 62
7.1.1 Density property 63
7.1.2 Generalized derivatives, Sobolev spaces 65
7.1.3 Lipschitz functions 66
7.2 Inverting a square matrix 67

1 Introduction

[*UNDER CONSTRUCTION* !] In data science and in many applications such as quantum field theories, we have to handle datasets with a large number of attributes, and often labels and attributes demonstrating a dataset are not independent. It is convenient to represent datasets with D many attributes as vectors in the Euclidean space of D dimensions, where D though is very large. In many applications, D is larger than the size of the sample data. Often datasets in applications are located in a lower dimensional sub-manifolds, so there is a question of reducing dimensions in datasets. This course does not address this kind of questions, nor to address anything about learning from data or about regenerating datasets. Rather, we attempt to develop an array of mathematical tools to address the question of describing the distributions of datasets. The main tool we shall develop in this course is the Ornstein-Uhlenbeck (OU) diffusion process, although we shall only study this model from a deterministic dynamic point-view. We however would like to point out that this OU process plays a crucial role in the recent year AI revolution, namely *the regenerative diffusion model* in this new phase of AI technology.

Prerequisite: It is essential that you have good computational skills from (1) *Prelims Calculus*, (2) A2.1 *Metric Spaces*, (3) First half of A8 *Probability*, and (4) A4 *Integration*.

Main tools: We shall introduce a few new concepts on the way, but no one of them is particularly new, and they are introduced mainly for convenience. We shall mainly use the computational tools developed in elementary calculus such as finding derivatives using various rules, finding some simple integrals, a little bit algebra for helping organizing your computations and etc. *A4 Integration* is required to backup and to justify your computations. You shall enjoy the powerful techniques developed in this course, and you shall appreciate the results established in this course like the isoperimetric inequalities both for Gaussian measures and for the Lebesgue measures. You shall be able to appreciate the main method developed in this course, i.e. the method of stochastic quantization in its simplest form.

About this course: This is not a course about data science, it is a course which is quite useful for understanding datasets. It is a probability course with strong flavor of analysis. While I hope in near future these tools shall be used widely in data science.

The standard one dimensional normal distribution, even in high-dimensional probability, remains to play an important role as in elementary Probability Theory. The Gaussian distribution function

$$\Phi(x) = \int_{-\infty}^{x} \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}} \mathrm{d}x \quad \text{for } x \in \mathbb{R}$$

whose probability density function (PDF) is its derivative: $\Phi'(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}}$. Clearly its second derivative $\Phi''(x) = -x\Phi'(x)$. Φ is strictly increasing on $(-\infty, \infty)$ taking values in $(0,1)$, whose inverse function $\Phi^{-1} : (0,1) \mapsto (-\infty, \infty)$ is also strictly increasing. A fundamental fact about normal distribution is that the tail probability

$$1 - \Phi(r) = \int_{r}^{\infty} \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}} \mathrm{d}x$$

decays to zero in a speed like $e^{-r^2/2}$ as $r \to \infty$.

In fact we have more precise quantitative decay estimates.

Exercise. For $r > 0$ we have

$$\left(r+\frac{1}{r}\right)^{-1}\Phi'(r) \leq 1-\Phi(r) < \frac{1}{r}\Phi'(r).$$

[*Hint*: Observe that

$$\begin{aligned}\int_r^\infty \frac{1}{\sqrt{2\pi}}e^{-\frac{x^2}{2}}\mathrm{d}x &< \int_r^\infty \frac{x}{r}\frac{1}{\sqrt{2\pi}}e^{-\frac{x^2}{2}}\mathrm{d}x \\ &= \int_r^\infty \left(1+\frac{1}{x^2}\right)\frac{1}{\sqrt{2\pi}}e^{-\frac{x^2}{2}}\mathrm{d}x \\ &< \left(1+\frac{1}{r^2}\right)\int_r^\infty \frac{1}{\sqrt{2\pi}}e^{-\frac{x^2}{2}}\mathrm{d}x.\end{aligned}$$

You may read page 4 in H. P. McKean: *Stochastic Integrals*. Academic Press New York and London (1969), or any other books on probability.]

Therefore we conclude that

$$1-\Phi(r) = \int_r^\infty \frac{1}{\sqrt{2\pi}}e^{-\frac{x^2}{2}}\mathrm{d}x \leq \min\left\{\frac{1}{2}, \frac{1}{\sqrt{2\pi}}\frac{1}{r}e^{-\frac{r^2}{2}}\right\}$$

for any $r > 0$.

Suppose X has a normal distribution with mean zero and variance σ^2, then for every $r > 0$

$$\begin{aligned}\mathbb{P}[X > r] &= \int_r^\infty \frac{1}{\sqrt{2\pi\sigma^2}}e^{-\frac{x^2}{2\sigma^2}}\mathrm{d}x \\ &= \sigma\int_{r/\sigma}^\infty \frac{1}{\sqrt{2\pi}}e^{-\frac{x^2}{2}}\mathrm{d}x \\ &\leq \min\left\{\frac{\sigma}{2}, \frac{\sigma^2}{\sqrt{2\pi}}\frac{1}{r}e^{-\frac{r^2}{2\sigma^2}}\right\} \\ &\sim \exp\left(-\frac{r^2}{2\sigma^2}\right)\end{aligned}$$

which maybe called the Gaussian decay rate. We shall later on prove that

$$\mathbb{P}[X > r] \leq \exp\left(-\frac{r^2}{2\sigma^2}\right)$$

for every $r > 0$.

In this course, we shall develop an array of mathematical tools for establishing effective tail estimates for high-dimensional probability distributions. In contrast with the traditional probability theory and classical stochastic analysis, where the concepts such as independence, martingale property, Markov property, play dominated roles, in High-Dimensional Probability, we seek for tools which can be used for handling distributions of random fields which do not possess these properties. These tools shall be particularly useful for the study of distributions of datasets with large numbers of attributes with complex (dependent) structures.

Let us collect several notions, notations and a few elementary facts which shall be used in this course.

Suppose (X,d) is a metric space, then the topology on X defined by the metric d is the collection of all open subsets, that is all subset U which have the following property: for every $x \in U$, there is a positive number r (depending on x in general though) such that the open ball centered at x with radius r, $B_x(r)$ is a subset of U. A metric space is *separable* if it has a countable dense subset. A metric space is *complete* if every Cauchy sequence has a limit. A complete and separable metric space is called a *Polish space*.

The σ-algebra generated by open subsets, i.e. the smallest σ-algebra on X, containing all open subsets (and therefore all closed subsets as well) is called the Borel σ-algebra, denoted by $\mathscr{B}(X)$. By saying a measure on a metric space, we mean a measure on the Borel σ-algebra on a metric space, unless otherwise specified. In particular, any continuous function on a metric space is measurable (with respect to the Borel σ-algebra), cf. A4 Integration.

Most distributions one has to deal with in applications are probability measures on sample spaces with additional space structures, such as linear structures you studied in Linear Algebras. The most convenient way to introduce a distance on a vector space X is through a norm. We recall that a function $x \mapsto \|x\|$ from a vector space $X \mapsto [0,\infty)$ if $\|x\| = 0$ only for $x = 0$, $\|\lambda x\| = |\lambda| \, \|x\|$ for every scalar λ and $x \in X$, and the triangle inequality holds: $\|x+y\| \leq \|x\| + \|y\|$ for any $x, y \in X$. The topology (i.e. the collection of open sets) on X is defined by the induced distance $d(x,y) = \|x-y\|$ (for $x,y \in X$). In this way we call $(X, \|\cdot\|)$ is a normed (linear, or vector) space, that is, a vector space equipped with a norm. Such normed space is called a *Banach space* if it is complete as a metric space (cf. A2.1 Metric Spaces).

A scalar (or inner) product on X is a mapping $\langle \cdot, \cdot \rangle$ from the product space $X \times X$ to $\mathbb{C}$,which sends an ordered pair (x,y) to a number $\langle x,y \rangle$ which satisfies the following properties: $\langle x,y \rangle = \overline{\langle y,x \rangle}$ for every pair $x,y \in X$, $\langle x,x \rangle \geq 0$ for every x and $= 0$ only for $x = 0$, the mapping $x \mapsto \langle x,y \rangle$ is linear (in x) for every y, and $y \mapsto \langle x,y \rangle$ is conjugate linear (in y) for every x, i.e. $\langle x, y_1 + y_2 \rangle = \langle x,y_1 \rangle + \langle x,y_2 \rangle$ and $\langle x, \lambda y \rangle = \bar{\lambda} \langle x,y \rangle$ for any number λ, and $x,y \in X$. $\|x\| = \sqrt{\langle x,x \rangle}$ for $x \in X$ defines a norm on X, the norm $\|\cdot\|$ induced by the scalar product. A Banach space whose norm is induced by a scalar product is called a *Hilbert space*.

2 Measures, integration and probability distributions

In this section we give a quick review about the foundation of probability theory.

2.1 Measures and Lebesgue's integration

We shall not develop Lebesgue's theory of integration in detail, which the reader may learn from a standard textbook such as Halmos [9]. We shall however introduce the notations, notions and the fundamental results sufficient enough so that the reader may follow the main content of the book without need to refer to more theoretical approach of the measure theory.

A *measurable space* $(E, \mathscr{F})$ consists of two objects, a space (simply a non-empty set) E, and a σ-algebra $\mathscr{F}$ on E. By a σ-algebra on E we mean a collection $\mathscr{F}$ of some subsets of E which satisfies the following properties: the empty set ï¿œ and the whole space E belong to the collection $\mathscr{F}$, if $A, B \in \mathscr{F}$, then $A \setminus B \in \mathscr{F}$, and if $A_i \in \mathscr{F}$ (where $i = 1, 2, \ldots$), then $\bigcup_{i=1}^{\infty} A_i \in \mathscr{F}$.

Clearly, the collection which contains only the empty set and the whole space is a σ-algebra, which is called the trivial σ-algebra on any space. On the other hand, the totality of all subsets of E is a σ-algebra, which shall be the default choice of a σ-algebra when E is a finite or countable space, unless otherwise specified.

If $\mathscr{C}$ is a non-empty collection of some subsets of E, then

$$\sigma(\mathscr{C}) = \bigcap \{\mathscr{F} : \mathscr{C} \subset \mathscr{F} \text{ and } \mathscr{F} \text{ is } \sigma\text{-algebra on } E\}$$

is indeed a σ-algebra, which is the smallest σ-algebra containing $\mathscr{C}$, called the σ-algebra generated by $\mathscr{C}$.

Example 2.1. *Let S be a metric space. Then the Borel σ-algebra on S, denoted by $\mathscr{B}(S)$, is the σ-algebra generated by the collection of all open (hence closed) subsets of S. Unless it is said otherwise, the default σ-algebra on a metric space is the Borel σ-algebra $\mathscr{B}(S)$.*

Definition 2.2. *Suppose $(E_1,\mathscr{F}_1)$ and $(E_2,\mathscr{F}_2)$ are two measurable spaces, and $F : E_1 \mapsto E_2$ is a mapping. Then F is called a measurable mapping (or called a measurable function) if $F^{-1}(\mathscr{F}_2) \subset \mathscr{F}_1$. That is, for every $A \in \mathscr{F}_2$, the pre-image $F^{-1}(A) \in \mathscr{F}_1$.*

We shall add several comments about this definition.

Remark 2.3. *1) The concept of measurable mappings (functions) between two spaces depends on their carried σ-algebras.*

2) $F^{-1}(\mathscr{F}_2)$, which is collection of all $F^{-1}(A)$ (where A runs through $\mathscr{F}_2$), is itself a σ-algebra on E_1, called the pull-back σ-algebra of $\mathscr{F}_2$ by the mapping F. $F^{-1}(\mathscr{F}_2)$ is the smallest σ-algebra $\mathscr{F}$ on E_1 such that F is measurable (with respect to the σ-algebra $\mathscr{F}$ on E_1 and $\mathscr{F}_2$ on E_2, and therefore $F^{-1}(\mathscr{F}_2)$ is also called the σ-algebra on E_1 generated by the mapping F.

3) A measurable mapping $F : E_1 \mapsto E_2$, where $(E_i,\mathscr{F}_i)$ (where $i = 1,2$) are measurable spaces, is also called an E_2-valued random variable.

4) Let $(E,\mathscr{F})$ be a measurable space, and let S be a metric space. Then a mapping $F : E \mapsto S$ is called an S-valued random variable if F is measurable with respect to the σ-algebra $\mathscr{F}$ and the Borel σ-algebra $\mathscr{B}(S)$, i.e. $F^{-1}(A) \in \mathscr{F}$ for every Borel measurable subset $A \subset S$.

5) It is convenient to introduce two symbols ∞ and $-\infty$ in $\mathbb{R}$ of real numbers, with the convention that $-\infty < a < \infty$ for any real number a, $0 \cdot \infty = 0$, $a \cdot \infty = \infty$ if $a > 0$, and $\infty \cdot \infty = \infty$. Let $[-\infty,\infty] = \mathbb{R} \cup \{-\infty,\infty\}$. Then the Borel σ-algebra $\mathscr{B}([-\infty,\infty])$ is the σ-algebra generated by $\{-\infty\}$, $\{\infty\}$ and $\mathscr{B}(\mathbb{R})$. A $[-\infty,\infty]$ -valued measurable function on a measurable space $(E,\mathscr{F})$ (where the generalized real line $[-\infty,\infty]$ carries the Borel σ-algebra) is called a (generalized) real random variable on $(E,\mathscr{F})$.

Proposition 2.4. *Let $(E,\mathscr{F})$ be a measurable space, and let $f, f_n : E \mapsto [-\infty,\infty]$ be (generalized) real functions (for $n = 1,2,\ldots$).*

1) f is measurable if and only if $f^{-1}(-\infty)$, $f^{-1}(\infty)$ and $\{f < a\}$ are measurable for every number a.

2) If f is measurable, then $f^+ = \max\{f,0\}$ and $f^- = \max\{-f,0\}$ are non-negative and measurable.

3) Suppose f_n are measurable (for $n = 1,2,\ldots$), then $\sup_n f_n$, $\inf_n f_n$, $\limsup f_n$ and $\liminf f_n$ are measurable.

Let $(E,\mathscr{F})$ be a measurable space. An *outer measure* μ on $(E,\mathscr{F})$ is a function defined on $\mathscr{F}$ taking values in $[0,\infty]$ such that $\mu(ï¿œ) = 0$, and

$$\mu\left(\bigcup_{i=1}^{\infty} A_i\right) \leq \sum_{i=1}^{\infty} \mu(A_i)$$

for any $A_i \in \mathscr{F}$ (where $i = 1,2,\ldots$). An outer measure μ is called a *measure* if in addition μ is countably additive:

$$\mu\left(\bigcup_{i=1}^{\infty} A_i\right) = \sum_{i=1}^{\infty} \mu(A_i)$$

for any disjoint measurable A_i (where $i = 1,2,\ldots$).

If μ is an outer measure on $(E,\mathscr{F})$, then $A \in \mathscr{F}$ is called μ-measurable if

$$\mu(B) = \mu(B \cap A) + \mu(B \cap A^c) \quad \text{for any } B \in \mathscr{F}.$$

According to a theorem of Caratheodory's, the collection $\mathscr{M}_\mu$ of all μ-measurable subsets is a sub σ-algebra of $\mathscr{F}$, and μ restricted on $\mathscr{M}_\mu$ is a measure.

If μ is a measure $(E,\mathscr{F})$, then the triple $(E,\mathscr{F},\mu)$ is called a measure space. μ is called a finite measure if $\mu(E) < \infty$, and it is σ-finite if there is a sequence $A_n \in \mathscr{F}$ (for $n = 1,2,\ldots$) such that $\bigcup_{i=1}^{\infty} A_i = E$ and $\mu(A_n) < \infty$ for every $n = 1,2,\ldots$.

A measure μ with total mass 1, that is, $\mu(E) = 1$, is called a probability, a probability distribution, a probability measure, or simply a distribution on $(E,\mathscr{F})$.

Let us now work with a σ-finite measure space $(E,\mathscr{F},\mu)$. The integration theory over this measure space can be constructed as the following. A non-negative function $\varphi : E \mapsto [0,\infty]$ is called $\mathscr{F}$-simple if $\varphi = \sum_{i=1}^{k} c_i 1_{A_i}$ for some positive integer k, some $A_i \in \mathscr{F}$ and some $c_i \in [0,\infty]$. For such a simple function, its integral

$$\int_E \varphi \mathrm{d}\mu = \sum_{i=1}^{k} c_i \mu(A_i)$$

which may be infinity though. If $f : E \mapsto [0,\infty]$ is measurable, then its integral

$$\int_E f \mathrm{d}\mu = \sup\left\{\int_E \varphi \mathrm{d}\mu : \varphi \ \mathscr{F}\text{-simple, and } \varphi \leq f\right\}$$

where $\sup I = \infty$ if I is not bounded from above. An non-negative, $\mathscr{F}$-measurable function is integrable (with respect to the measure μ) if its integral $\int_E f \mathrm{d}\mu < \infty$. For a general (i.e. not necessary non-negative) $\mathscr{F}$-measurable function f, then $f = f^+ - f^-$ and $|f| = f^+ + f^-$, where both f^+ and f^- are non-negative and $\mathscr{F}$-measurable. If $\int_E f^{\pm} \mathrm{d}\mu < \infty$, then f is called integrable (w.r.t. the measure μ) and its integral

$$\int_E f \mathrm{d}\mu = \int_E f^+ \mathrm{d}\mu - \int_E f^- \mathrm{d}\mu.$$

The totality of all $\mathscr{F}$-measurable and μ-integrable functions is denoted by $L^1(E,\mathscr{F},\mu)$. For simplicity, if f is measurable, and if f is non-negative or integrable, then its integral $\int_E f \mathrm{d}\mu$ is also denoted by $\int_E f(x)\mu(\mathrm{d}x)$, $\int f \mathrm{d}\mu$, or by $\mu(f)$ if no confusion arises.

Suppose $p > 0$ is a constant, and f is $\mathscr{F}$-measurable, then $|f|^p$ is $\mathscr{F}$-measurable, and define $\|f\|_p = \left(\int_E |f|^p \mathrm{d}\mu\right)^{1/p}$ (which may be infinity). If $p = \infty$, then

$$\|f\|_\infty = \inf\{C \geq 0 : |f| \leq C \; \mu \text{ almost surely } \}$$

For $p \in (0, \infty]$, $L^p(E, \mathscr{F}, \mu)$ denotes the totality of all $\mathscr{F}$-measurable functions f such that $\|f\|_p < \infty$.

For discussions involving functions in $L^p(E, \mathscr{F}, \mu)$, two measurable functions f and g are identified as the same element (in any L^p-space) as long as $\{f \neq g\}$ has μ-measure zero.

Theorem 2.5. *Let $p \in [1, \infty]$, and $d_p(f, g) = \|f - g\|_p$ for any two measurable functions on $(E, \mathscr{F}, \mu)$.*
1) $L^p(E, \mathscr{F}, \mu)$ is a linear space, and d_p is a metric on $L^p(E, \mathscr{F}, \mu)$.
2) The metric space $L^p(E, \mathscr{F}, \mu)$ equipped with the metric d_p is complete.

$L^p(E, \mathscr{F}, \mu)$ is called the L^p-space over the measure space $(E, \mathscr{F}, \mu)$. It is a very important fact that for every $p \geq 1$, $f \mapsto \|f\|_p$ is a norm on the L^p-space. In particular if $p \geq 1$, then

$$\|f + g\|_p \leq \|f\|_p + \|g\|_p \qquad \text{for any } f, g \in L^p(E, \mathscr{F}, \mu),$$

which is called the *Minkowski inequality*. This inequality can be proved by using the convexity of the power function x^p on $(0, \infty)$ if $p \geq 1$. The detail of the proof is left as an exercise (see Problem Sheet 1).

Let us now review several important results in the integration theory.

Recall that a real function ρ defined on an interval (a, b) (not necessary bounded) is convex if

$$\rho(\lambda s + (1 - \lambda)t) \leq \lambda \rho(s) + (1 - \lambda)\rho(t) \tag{2.1}$$

for any $s, t \in (a, b)$ and $\lambda \in [0, 1]$. A function ρ is concave if $-\rho$ is convex.

Theorem 2.6. *(*Jensen's inequality*) Let $(E, \mathscr{F}, \mu)$ be a finite measurable space. If ρ is convex on (a, b) and f is measurable and takes values in (a, b), then*

$$\rho\left(\frac{1}{\mu(E)} \int_E f d\mu\right) \leq \frac{1}{\mu(E)} \int_E \rho(f) d\mu \tag{2.2}$$

as long as both f and $\rho(f)$ are integrable.

Theorem 2.7. *(*The Hölder inequality*) If f and g are two measurable functions on a σ-finite measure space $(E, \mathscr{F}, \mu)$, then*

$$\int_E |fg| d\mu \leq \left(\int_E |f|^p d\mu\right)^{\frac{1}{p}} \left(\int_E |g|^q d\mu\right)^{\frac{1}{q}} \tag{2.3}$$

if $p > 1$ and $\frac{1}{p} + \frac{1}{q} = 1$. In particular if $f \in L^p(E, \mu)$ and $g \in L^q(E, \mu)$ then $fg \in L^1(E, \mu)$. The case where $p = q = 2$ is called the Cauchy-Schwartz inequality.

Proof. If one of the integral on the right-hand side vanishes, then f or g equals zero almost surely, which forces that $fg = 0$ almost surely too, thus both sides of the inequality are zero. The inequality is trivial in this case. Thus let us assume both integrals on the right-hand side are greater than zero (but may be ∞). For this case, if one of the integral on the right-hand side is ∞, the the right-hand side is infinity, so the inequality is surely true and of course is also trivial. Therefore we may assume that

$$0 < \|f\|_p = \left(\int_E |f|^p \mathrm{d}\mu\right)^{\frac{1}{p}} < \infty$$

and

$$0 < \|g\|_q = \left(\int_E |g|^q \mathrm{d}\mu\right)^{\frac{1}{q}} < \infty.$$

For this case, by replacing f by $f/\|f\|_p$ and $g/\|g\|_q$, we may further assume that $\|f\|_p = \|g\|_q = 1$. Now we use the elementary inequality

$$st \leq \frac{1}{p}s^p + \frac{1}{q}t^q$$

for any non-negative s,t [This inequality follows by inspecting the function $\varphi(x) = x - \frac{1}{p}x^p - \frac{1}{q}$ (for $x \geq 0$) and showing the maximum $\varphi(1) \leq 0$]. □

The Hölder inequality may be stated as the following convenient form

$$\int_E |f|^\alpha |g|^{1-\alpha} \mathrm{d}\mu \leq \left(\int_E |f| \mathrm{d}\mu\right)^\alpha \left(\int_E |g| \mathrm{d}\mu\right)^{1-\alpha} \tag{2.4}$$

where $\alpha \in (0,1)$ is a constant, f,g are μ-integrable.

A special case for probabilities is worthy of mention.

Corollary 2.8. *Let $(\Omega, \mathscr{F}, \mathbb{P})$ be a probability space. Then*

$$(\mathbb{E}|X|)^p \leq \mathbb{E}(|X|^p)$$

for every $p \geq 1$, X is p-th integrable. Equivalently

$$\mathbb{E}(|X|^\alpha) \leq (\mathbb{E}|X|)^\alpha$$

for every constant $\alpha \in (0,1)$, and X is integrable.

Exercise 2.9. *Suppose $X > 0$ and Y are two measurable functions on a σ-finite measure space $(E, \mathscr{F}, \mu)$. Then*

$$\mu\left(\frac{Y^2}{X}\right) \geq \frac{(\mu(|Y|))^2}{\mu(X)}. \tag{2.5}$$

Here we use also $\mu(f)$ to denote the integral $\int_E f d\mu$.

Proof. In fact by Cauchy-Schwartz inequality

$$\mu(|Y|) = \mu\left(\sqrt{X}\frac{|Y|}{\sqrt{X}}\right) \leq \sqrt{\mu(X)}\sqrt{\mu\left(\frac{Y^2}{X}\right)}$$

which yields (2.5). □

It should be understood that the main task in probability theory (i.e. statistical mechanics) is to give a good description of the distribution of a random variable. For a real random variable X, we are interested in its distribution function $F_X(t) = \mathbb{P}[X \le t]$, which is a reason we are so interested in tail estimates such as $\mathbb{P}[X \ge t]$.

Theorem 2.10. *Let $\rho : (0,\infty) \mapsto [0,\infty)$ be right-continuous and increasing with its right-hand limit at 0: $\rho(0+) = 0$. Let m_ρ denote the Lebesgue–Stieltjes measure associated with ρ (cf. A4* Integration*), i.e. m_ρ is the unique measure on $([0,\infty), \mathscr{B}([0,\infty)))$ such that $m_\rho((s,t]) = \rho(t) - \rho(s)$ for any $t > s \ge 0$, and $m_\rho(\{0\}) = 0$.*

Let X and Y be two non-negative *measurable functions on a σ-finite measure space $(E, \mathscr{F}, \mu)$.*

1) It holds that

$$\int_E \rho(X) d\mu = \int_0^\infty \mu[X \ge \lambda] m_\rho(d\lambda). \tag{2.6}$$

2) Suppose that there is a constant $C > 0$ such that $\mu[X \ge \lambda] \le C\mu[Y \ge \lambda]$ for all $\lambda > 0$. Then $\int_E \rho(X) d\mu \le C \int_E \rho(Y) d\mu$.

Proof. The proof follows from the construction of m_ρ and the Fubini theorem (cf. A4 *Integration*). Indeed

$$\begin{aligned}
\int_E \rho(X(\omega))\mu(\mathrm{d}\omega) &= \int_E (\rho(X(\omega)) - \rho(0+))\mu(\mathrm{d}\omega) = \int_E m_\rho((0,X(\omega)])\mu(\mathrm{d}\omega) \\
&= \int_E \left[\int_{(0,X(\omega)]} m_\rho(\mathrm{d}\lambda)\right]\mu(\mathrm{d}\omega) = \int_E \left[\int_0^\infty 1_{[\lambda \le X(\omega)]} m_\rho(\mathrm{d}\lambda)\right]\mu(\mathrm{d}\omega) \\
&= \int_{E\times(0,\infty)} 1_{[X(\omega)\ge\lambda]} m_\rho(\mathrm{d}\lambda)\mu(\mathrm{d}\omega) \\
&= \int_{(0,\infty)} \mu(\{X \ge \lambda\}) m_\rho(\mathrm{d}\lambda)
\end{aligned}$$

where we have used the fact that $m_\rho((s,t]) = \rho(t) - \rho(s)$ for any $t \ge s \ge 0$ by definition. □

Theorem 2.11. *If f is a non-negative, Borel measurable function on $\mathbb{R}^D$, then*

$$\int_{\mathbb{R}^D} f(x) dx = \int_0^\infty Leb(\{f > t\}) dt \tag{2.7}$$

where Leb denotes the Lebesgue measure on $\mathbb{R}^D$.

Proof. We may observe that, if ρ is increasing, continuous and $\rho(0+) = 0$ in Lemma 2.10, then $\mu[X \ge \lambda]$ can be replaced by $\mu[X > \lambda]$. In fact

$$\int_{\mathbb{R}^D} f(x)\mathrm{d}x = \int_0^\infty \mathrm{Leb}(\{f \ge t\})\mathrm{d}t.$$

Since $t \mapsto \mathrm{Leb}(\{f > t\})$ is decreasing so that

$$\{t \ge 0 : \mathrm{Leb}(\{f > t\}) \ne \mathrm{Leb}(\{f \ge t\})\}$$

is at most countable, and therefore is a null subset with respect to the Lebesgue measure. Therefore (2.7) follows immediately. □

Theorem 2.12. *Suppose X and A are two non-negative random variables on a probability space, and suppose*

$$\mathbb{P}[X \geq \lambda] \leq \frac{1}{\lambda}\mathbb{E}[A : X \geq \lambda] \quad \text{for any } \lambda > 0.$$

Then, for any $p > 1$

$$\mathbb{E}[X^p] \leq \left(\frac{p}{p-1}\right)^p \mathbb{E}[A^p]. \tag{2.8}$$

Proof. We can assume that X is bounded, otherwise we use $\min\{X,n\}$ (for $n = 1,2,\ldots$) instead and take limit as $n \to \infty$. Let $\rho(t) = t^p$ for $t > 0$. Then, by (2.6) [with $\rho(t) = t^p$ for $t > 0$]

$$\begin{aligned}\mathbb{E}[X^p] &= \int_0^\infty \mathbb{P}[X \geq \lambda]\, m_\rho(\mathrm{d}\lambda) \leq \int_0^\infty \frac{1}{\lambda}\mathbb{E}[A : X \geq \lambda]\,\rho'(\lambda)\mathrm{d}\lambda \\ &\leq p\int_0^\infty \mathbb{E}[A : X \geq \lambda]\,\lambda^{p-2}\mathrm{d}\lambda.\end{aligned}$$

Using Fubini's theorem for the last integration, we obtain that

$$\mathbb{E}[X^p] \leq p\mathbb{E}\left[A\int_0^X \lambda^{p-2}\mathrm{d}\lambda\right] = \frac{p}{p-1}\mathbb{E}\left[AX^{p-1}\right].$$

Apply Hölder's inequality to obtain that

$$\mathbb{E}[X^p] \leq \frac{p}{p-1}\left(\mathbb{E}[A^p]\right)^{\frac{1}{p}}\left(\mathbb{E}[X^p]\right)^{\frac{1}{q}}$$

where $\frac{1}{p} + \frac{1}{q} = 1$. Rearranging the inequality to complete the proof. □

The previous results, which though are very useful, can be stated in terms of Riemann integrals (if one is happy with Riemann integrals rather than abstract integration), the usefulness of Lebesgue's integration however lies in its powerful capability of handling orders of taking various limits as stated in the following fundamental theorem below, which is the core part of Lebesgue's integration.

Theorem 2.13. *Let $(E,\mathscr{F},\mu)$ be a σ-finite measure space. Let f_n be measurable on E (where $n = 1,2,\ldots$).*

*1) (*Fatou's lemma*) If f_n are non-negative, then*

$$\int_E \liminf_{n\to\infty} f_n d\mu \leq \liminf_{n\to\infty}\int_E f_n d\mu.$$

*2) (*Monotone Convergence Theorem, MCT*) If f_n is an increasing sequence of non-negative and measurable functions, then*

$$\int_E \lim_{n\to\infty} f_n d\mu = \lim_{n\to\infty}\int_E f_n d\mu.$$

*3) (*Dominated Convergence Theorem*) Suppose $f_n \to f$ almost surely and $|f_n| \leq g$ (for every n) for some integrable function g, then all f_n and f are integrable and*

$$\int_E f d\mu = \lim_{n\to\infty}\int_E f_n d\mu.$$

Item 3), the Dominated Convergence Theorem, is the theoretical foundation for justifying our differentiation and taking limits under integration, though very often one should carefully check the control condition required in this theorem, such details though are often omitted.

3 General concentration inequalities

Let us begin with a very general concentration principle of high-dimensional distributions, which is not quantitative as we wish and therefore it has a very limited value.

Lemma 3.1. *Let (E,ρ) be a Polish space, and $\mathbb{P}$ be any probability measure on $(E,\mathscr{B}(E))$. Then for every $\varepsilon>0$ there is a compact subset $K\subset E$, such that $\mathbb{P}[E\setminus K]<\varepsilon$.*

Proof. Since E is separable, for every $\delta>0$, E can be covered by countable many balls with radius δ. Therefore, for every n, there is a sequence of *closed* balls $B_i^{(n)}$ of radius $\frac{1}{2^n}$ (where $i=1,2,\ldots$) such that $\cup_i B_i^{(n)}=E$ for each n. By construction

$$\lim_{k\to\infty}\mathbb{P}\left(\bigcup_i^k B_i^{(n)}\right)=\mathbb{P}\left(\bigcup_i B_i^{(n)}\right)=\mathbb{P}(E)=1.$$

Hence for each n, there is k_n such that

$$\mathbb{P}\left(\bigcup_i^{k_n} B_i^{(n)}\right)>1-\frac{\varepsilon}{2^n}.$$

Let $K=\cap_{n=1}^{\infty}\cup_i^{k_n}B_i^{(n)}$. K is totally bounded by definition and is also closed. Since E is complete, therefore K is compact. Since

$$\mathbb{P}(K^c)\le\sum_{n=1}^{\infty}\mathbb{P}\left[\left(\bigcup_i^{k_n}B_i^{(n)}\right)^c\right]<\sum_{n=1}^{\infty}\frac{\varepsilon}{2^n}=\varepsilon$$

and therefore $\mathbb{P}(K)>1-\varepsilon$. □

3.1 One-dimensional distributions

The most familiar estimates are perhaps those derived from the Markov inequality. Recall that if X is a real and integrable random variable on a probability space $(\Omega,\mathscr{F},\mathbb{P})$, then for every $\lambda>0$ we have

$$\mathbb{P}[X\ge\lambda]=\mathbb{E}\left[1_{\{X\ge\lambda\}}\right]\le\mathbb{E}\left[\frac{X}{\lambda}1_{\{X\ge\lambda\}}\right]=\frac{1}{\lambda}\mathbb{E}\left[X1_{\{X\ge\lambda\}}\right]$$

In particular, if X is non-negative

$$\mathbb{P}[X\ge\lambda]\le\frac{1}{\lambda}\mathbb{E}[X]\qquad\text{for }\lambda>0\tag{3.1}$$

which is called the Markov inequality.

There are variations of the Markov inequality. Suppose $\phi:\mathbb{R}\to(0,\infty)$ is increasing, then

$$\begin{aligned}\mathbb{P}[X\ge\lambda]=\mathbb{P}[\phi(X)\ge\phi(\lambda)]&\le\mathbb{E}\left[\frac{\phi(X)}{\phi(\lambda)}1_{\{X\ge\lambda\}}\right]\\&=\frac{1}{\phi(\lambda)}\mathbb{E}[\phi(X):X\ge\lambda]\end{aligned}$$

which of course yields that

$$\mathbb{P}[X \geq \lambda] \leq \frac{1}{\phi(\lambda)}\mathbb{E}[\phi(X) : X \geq \lambda] \tag{3.2}$$

for any λ and increasing, positive function ϕ. In particular

$$\mathbb{P}[|X-\mu| \geq \lambda] \leq \frac{1}{\lambda^p}\mathbb{E}[|X-\mu|^p] \quad \text{for } \lambda > 0 \tag{3.3}$$

for any μ and $p \geq 0$. The inequality reduces to the Chebyshev inequality where $\mu = \mathbb{E}[X]$ and $p = 2$. Similarly if $\psi : \mathbb{R} \to (0,\infty)$ is decreasing, then

$$\mathbb{P}[X \leq \lambda] = \mathbb{P}[\psi(X) \geq \psi(\lambda)] \leq \mathbb{E}\left[\frac{\psi(X)}{\psi(\lambda)} 1_{\{X \leq \lambda\}}\right].$$

Therefore

$$\mathbb{P}[X \leq \lambda] = \mathbb{P}[\psi(X) \geq \psi(\lambda)] \leq \mathbb{E}\left[\frac{\psi(X)}{\psi(\lambda)}\right]$$

for any λ and any positive and decreasing function ψ.

Proposition 3.2. *(*Chernoff's inequality*) Suppose* $\mathbb{E}\left[e^{\lambda X}\right]$ *exists for all* λ*, then*

$$\mathbb{P}[X \geq t] \leq e^{-I_X^+(t)} \quad \textit{for every } t \in \mathbb{R}, \tag{3.4}$$

where

$$I_X^+(t) = \sup_{\lambda \geq 0}\left\{\lambda t - \ln \mathbb{E}\left[e^{\lambda X}\right]\right\}. \tag{3.5}$$

Proof. $\phi(x) = \mathrm{e}^{\lambda x}$ (where $\lambda \geq 0$) is increasing, therefore

$$\mathbb{P}[X \geq t] \leq \frac{1}{\mathrm{e}^{\lambda t}}\mathbb{E}\left[\mathrm{e}^{\lambda X}\right] = \mathrm{e}^{-\left(\lambda t - \ln \mathbb{E}\left[\mathrm{e}^{\lambda X}\right]\right)}$$

for every t and $\lambda \geq 0$. However the left-hand side is independent of $\lambda \geq 0$, therefore

$$\mathbb{P}[X \geq t] \leq \mathrm{e}^{-\sup_{\lambda \geq 0}\left(\lambda t - \ln \mathbb{E}\left[\mathrm{e}^{\lambda X}\right]\right)}$$

which completes the proof. □

The function I_X^+ (which takes non-negative values, but maybe infinity) is called the Cramér transform of (the distribution of) X. We will revisit this function later on.

Example. Let X has a normal distribution $N(0,\sigma^2)$. Then

$$\begin{aligned}
\mathbb{E}\left[\mathrm{e}^{\lambda X}\right] &= \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{x^2}{2\sigma^2} + \lambda x\right) \mathrm{d}x \\
&= \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x-\sigma^2\lambda)^2}{2\sigma^2} + \frac{\sigma^2\lambda^2}{2}\right) \mathrm{d}x \\
&= \exp\left(\frac{\sigma^2\lambda^2}{2}\right)
\end{aligned}$$

so that

$$\mathbb{P}[X \ge t] \le \mathrm{e}^{-\sup_{\lambda\ge 0}\left(\lambda t - \frac{\sigma^2\lambda^2}{2}\right)}$$

where the sup is achieved at $\lambda = \frac{t}{\sigma^2}$, and therefore

$$\mathbb{P}[X \ge t] \le \exp\left(-\frac{t}{2\sigma^2}\right).$$

3.2 The Cramér theorem

Let $X_1, X_2, \ldots$ be an independent identically distributed sequence of (real) random variables on a probability space $(\Omega, \mathscr{F}, \mathbb{P})$, with a common distribution μ which is a probability measure on $(\mathbb{R}, \mathscr{B}(\mathbb{R}))$. Assume that X_1 is integrable, and let $a = \mathbb{E}[X_1] = \int_{\mathbb{R}} x\mu(\mathrm{d}x)$. Then the strong law of large numbers says

$$\frac{1}{n}\sum_{i=1}^{n} X_i \to a \quad \text{almost surely.}$$

That is to say, the distribution of the average $\frac{1}{n}\sum_{i=1}^{n} X_i$ is concentrated about the mean value a, and tends to Dirac's delta measure δ_a at a as $n \to \infty$. This result is at the core of probability, statistics and AI technology. In this section, we give more precise information about the concentration of the distribution μ_n of $\frac{1}{n}\sum_{i=1}^{n} X_i$.

The distribution μ_n of $\frac{1}{n}\sum_{i=1}^{n} X_i$ (for $n = 1, 2, \ldots$) is a probability measure on $(\mathbb{R}, \mathscr{B}(\mathbb{R}))$, by definition

$$\mu_n(A) = \mathbb{P}\left[\frac{1}{n}\sum_{i=1}^{n} X_i \in A\right] \qquad \text{for } A \in \mathscr{B}(\mathbb{R}).$$

Let us assume that the exponential moment of $X = X_1$ is finite, that is, $\mathbb{E}(e^{\lambda X}) < \infty$ for every λ. For simplicity, let $\psi_X(\lambda) = \ln \mathbb{E}(e^{\lambda X})$. The Legendre transform of ψ_X is defined by

$$I_X(x) = \sup_{\lambda \in \mathbb{R}} \{\lambda x - \psi_X(\lambda)\} \qquad \text{for } x \in \mathbb{R}.$$

I_X takes values in $[0, \infty]$.

Now we are in a position to state the first example of large deviation principle.

Theorem 3.3. *(H. Cramèr) Suppose* $\mathbb{E}(e^{\lambda X}) < \infty$ *for every* λ*, then* $\frac{1}{n}\sum_{i=1}^{n} X_i$ *(for* $n = 1, 2, \ldots$*) satisfies the* large deviation principle *(LDP) with the rate function* I_X*, in the sense that*

$$\limsup_{n\to\infty} \frac{1}{n} \log \mathbb{P}\left[\frac{1}{n}\sum_{i=1}^{n} X_i \in F\right] \le -\inf_{x\in F} I_X(x) \tag{3.6}$$

for every closed subset $F \subset \mathbb{R}$*, and*

$$\liminf_{n\to\infty} \frac{1}{n} \log \mathbb{P}\left[\frac{1}{n}\sum_{i=1}^{n} X_i \in G\right] \ge -\inf_{x\in G} I_X(x) \tag{3.7}$$

for every open subset $G \subset \mathbb{R}$*.*

We divide the proof of this theorem into several steps.

Lemma 3.4. *1) The function $\lambda \mapsto \mathbb{E}(e^{\lambda X})$ is smooth and log-convex, that is $\lambda \to \psi_X(\lambda)$ is convex.*

2) I_X is a convex, and $K_c = \{x : I_X(x) \le c\}$ is compact for every c.

3) $I_X(a) = 0$ where , and $I_X \uparrow$ on (a,∞) and $I_X \downarrow$ on $(-\infty,a)$.

4) We have

$$\inf_{(x,y]} I_X = I_X(y) \qquad \text{if } x < y \le a$$

and

$$\inf_{[x,y)} I_X = I_X(x) \qquad \text{if } a \le x < y\,.$$

Proof. 1) We only need to show that $\log\mathbb{E}(e^{\lambda X})$ is convex. For every $\alpha \in (0,1)$

$$\begin{aligned}\mathbb{E}(e^{(\alpha\lambda_1+(1-\alpha)\lambda_2)X}) &= \int e^{\alpha\lambda_1 x} e^{(1-\alpha)\lambda_2 x}\mu(\mathrm{d}x)\\ &\le \left(\int e^{\lambda_1 x}\mu(\mathrm{d}x)\right)^{\alpha}\left(\int e^{\lambda_2 x}\mu(\mathrm{d}x)\right)^{1-\alpha}\end{aligned}$$

(μ is the distribution of $X = X_1$), where the inequality follows from Hölder inequality with $p = \frac{1}{\alpha}$. Therefore $\lambda \mapsto \log\mathbb{E}(e^{\lambda X})$ is convex.

2) I_X is non-negative, and is convex as it is the supremum of the linear functions. In particular I_X is continuous on $\{x : I_X(x) < \infty\}$. We show that for every $c > 0$

$$K_c = \{x \in \mathbb{R} : I_X(x) \le c\}$$

is compact. Since I_X is continuous on $\{I_X < \infty\}$, so K_c is closed, thus we only need to show that K_c is bounded. If $x \in K_c$ then

$$\pm x - \psi_X(\pm 1) \le c$$

which implies that

$$|x| \le c + |\psi_X(1)| + |\psi_X(-1)|$$

for every $x \in K_c$. Hence K_c is bounded.

3) Since $-\ln x$ is convex on $(0,\infty)$, by Jensen's inequality

$$\begin{aligned}\log\mathbb{E}(e^{\lambda X}) &= \log\int e^{\lambda x}\mu(\mathrm{d}x)\\ &\ge \lambda\int x\mu(\mathrm{d}x) = \lambda a\end{aligned}$$

which implies that

$$\lambda a - \psi_X(\lambda) \le 0 \quad \text{for all } \lambda$$

Therefore we must have $I_X(a) = 0$ so a is the global minimum of I_X. The other claims then follows immediately as I_μ is convex. □

Lemma 3.5. *1) We have*

$$x\lambda - \psi_X(\lambda) \le (x-a)\lambda \tag{3.8}$$

for any x and λ. Here we recall that $\psi_X(\lambda) = \ln\mathbb{E}(e^{\lambda X})$.

2) We have

$$I_X(x) = \sup_{\lambda \geq 0} \{\lambda x - \psi_X(\lambda)\} \quad \textit{for} \quad x \geq a \tag{3.9}$$

and

$$I_X(x) = \sup_{\lambda \leq 0} \{\lambda x - \psi_X(\lambda)\} \quad \textit{for} \quad x \leq a\,. \tag{3.10}$$

Proof. By the proof of 3) in the previous lemma, (3.8) follows from Jensen's inequality. In particular, $\lambda x - \psi_X(\lambda) \leq 0$ for any x and λ such that $(x-a)\lambda \leq 0$. Therefore

$$I_X(x) = \sup_{\lambda:(x-a)\lambda \geq 0} \{\lambda x - \psi_X(\lambda)\}$$

for any x, which implies (3.9, 3.10) immediately. □

Lemma 3.6. *Let μ be the distribution of $X = X_1$ and $a = \mathbb{E}X$. Then*

$$\mu([x,\infty)) \leq \exp(-I_X(x)) = \exp\left(-\inf_{[x,\infty)} I_X\right) \quad \textit{for } x \geq a$$

and

$$\mu((-\infty,x]) \leq \exp(-I_X(x)) = \exp\left(-\inf_{(-\infty,x]} I_X\right) \quad \textit{for } x \leq a\,.$$

Proof. Indeed we have already proven the first inequality: if $\lambda \geq 0$ and $x \geq a$

$$\mu([x,\infty)) = \int_{z\geq x} \mu(\mathrm{d}z) \leq \int_{z\geq x} \frac{e^{\lambda z}}{e^{\lambda x}} \mu(\mathrm{d}z) \leq \int_{\mathbb{R}} \frac{e^{\lambda z}}{e^{\lambda x}} \mu(\mathrm{d}z) = \mathrm{e}^{-(\lambda x - \psi_X(\lambda))}$$

which yields that

$$\mu([x,\infty)) \leq \exp\left\{-\sup_{\lambda\geq 0}(\lambda x - \psi_X(\lambda))\right\} = \exp\{-I_X(x)\}\,.$$

Similarly we may prove the case where $x \leq a$. □

After having established the elementary facts we are now in a position to prove the LDP bounds.

Proof of upper bound (3.6). If $F = \emptyset$ or $a \in F$ then $\inf I_\mu = 0$ so that $\inf_F I_\mu = 0$ the bound is trivial in this case. Therefore we assume that $a \notin F$. If $F \subset [a,\infty)$, then $F \subset [y,\infty)$ where $y = \inf\{z : z \in F\}$. Hence

$$\inf_F I_X = I_X(y) = \sup_{\lambda\geq 0}\{\lambda y - \psi_X(\lambda)\}\,. \tag{3.11}$$

For every $\lambda > 0$

$$\begin{aligned}
\mathbb{P}\left[\frac{1}{n}\sum_{i=1}^n X_i \in F\right] &\leq \mathbb{P}\left[\frac{1}{n}\sum_{i=1}^n X_i \geq y\right] \leq \int_{\{\frac{1}{n}\sum_{i=1}^n X_i \geq y\}} \frac{e^{\frac{1}{n}\lambda\sum_{i=1}^n X_i}}{e^{\lambda y}} \mathrm{d}\mathbb{P} \\
&\leq \int_\Omega \frac{e^{\frac{1}{n}\sum_{i=1}^n \lambda X_i}}{e^{\lambda y}} \mathrm{d}\mathbb{P} = \int_\Omega \frac{\prod_{i=1}^n e^{\frac{\lambda}{n}X_i}}{e^{\lambda y}} \mathrm{d}\mathbb{P} \\
&= e^{-\lambda y} \prod_{i=1}^n \int_\Omega e^{\frac{\lambda}{n}X_i} \mathrm{d}\mathbb{P} = e^{-\lambda y}\left(\mathbb{E}\left(e^{\frac{\lambda}{n}X}\right)\right)^n.
\end{aligned}$$

Taking log both sides to obtain that

$$\frac{1}{n}\ln\mathbb{P}\left[\frac{1}{n}\sum_{i=1}^{n}X_i\in F\right]\leq-\left\{\frac{\lambda}{n}y-\ln M_\mu\left(\frac{\lambda}{n}\right)\right\}$$

for every $\lambda\geq 0$. It thus follows that

$$\begin{aligned}\frac{1}{n}\ln\mathbb{P}\left[\frac{1}{n}\sum_{i=1}^{n}X_i\in F\right]&\leq-\sup_{\lambda\geq 0}\left\{\lambda y-\psi_X(\lambda)\right\}=-I_X(y)\\&=-\inf_F I_X=-I_X(\min F).\end{aligned}$$

We thus have proven the upper bound for the case that $F\subset[a,\infty)$.

Similarly we may show that

$$\frac{1}{n}\ln\mathbb{P}\left[\frac{1}{n}\sum_{i=1}^{n}X_i\in F\right]\leq-\inf_F I_X=-I_\mu(\max F)\quad\text{if}\ \ F\subset(-\infty,a]\,.$$

Finally for an arbitrary closed set F in $\mathbb{R}$, let $F_1=F\cap(-\infty,a]$ and $F_2=F\cap[a,\infty)$. Then

$$\frac{1}{n}\ln\mathbb{P}\left[\frac{1}{n}\sum_{i=1}^{n}X_i\in F\right]\leq\frac{1}{n}\ln\left(\mathbb{P}\left[\frac{1}{n}\sum_{i=1}^{n}X_i\in F_1\right]+\mathbb{P}\left[\frac{1}{n}\sum_{i=1}^{n}X_i\in F_2\right]\right)$$

so that

$$\begin{aligned}\limsup_{n\to\infty}\frac{1}{n}\ln\mathbb{P}\left[\frac{1}{n}\sum_{i=1}^{n}X_i\in F\right]&\leq\max_{k=1,2}\left\{\limsup_{n\to\infty}\frac{1}{n}\ln\mathbb{P}\left[\frac{1}{n}\sum_{i=1}^{n}X_i\in F_k\right]\right\}\\&\leq\max\left\{-I_X(\max F_1)\,;\ -I_X(\min F_2)\right\}\\&=-\min\left\{I_X(\max F_1)\,;\ I_X(\min F_2)\right\}\\&\leq-\inf_F I_X\end{aligned}$$

which is the upper bound for large deviations.

Proof of lower bound (3.7) Let G be an open subset of $\mathbb{R}$. We are going to show that for every $x\in G$,

$$\liminf_{n\to\infty}\frac{1}{n}\ln\mathbb{P}\left[\frac{1}{n}\sum_{i=1}^{n}X_i\in G\right]\geq-I_X(x)\,.$$

Obviously we only need to prove the previous inequality for those $x\in G$ such that $I_X(x)<\infty$.

We consider two cases.

Firstly let us consider the case that the supremum $I_X(x)$ of $\sup_\lambda(\lambda x-\psi_X(\lambda))$ is not achievable. Then $x\neq a$ (as $I_X(a)=0$ which is achieved when $\lambda=0$). Without loss of generality, let us assume that $x>a$. Then we may choose a sequence of $\lambda_n>0$ such that $\lambda_n\to\infty$ and $\lambda_n x-\psi_X(\lambda_n)\to I_X(x)$ as $n\to\infty$.

By Lebesgue's dominated convergence theorem

$$\lim_{n\to\infty}\int_{(-\infty,x)} e^{\lambda_n(z-x)}\mu(\mathrm{d}z)=0$$

and therefore

$$\begin{aligned}\lim_{n\to\infty}\int_{[x,\infty)} e^{\lambda_n(z-x)}\mu(\mathrm{d}z) &= \lim_{n\to\infty}\int_{\mathbb{R}} e^{\lambda_n(z-x)}\mu(\mathrm{d}z)\\ &= \lim_{n\to\infty} e^{-\left\{\lambda_n x-\log\int_{\mathbb{R}}\exp(\lambda_n z)\mu(\mathrm{d}z)\right\}}\\ &= \exp\left(-I_X(x)\right)<\infty. \end{aligned} \tag{3.12}$$

On the other hand, for any $\delta>0$ we have

$$\int_{[x+\delta,\infty)} e^{\lambda_n(z-x)}\mu(\mathrm{d}z)\geq e^{\delta\lambda_n}\mu([x+\delta,\infty))$$

so that

$$\begin{aligned}\mu([x+\delta,\infty)) &\leq e^{-\delta\lambda_n}\int_{[x+\delta,\infty)} e^{\lambda_n(z-x)}\mu(\mathrm{d}z)\\ &\leq e^{-\delta\lambda_n}\int_{\mathbb{R}} e^{\lambda_n(z-x)}\mu(\mathrm{d}z)\\ &\leq e^{-\delta\lambda_n}e^{-\left\{\lambda_n x-\log\int_{\mathbb{R}} e^{\lambda_n z}\mu(\mathrm{d}z)\right\}}. \end{aligned}$$

Letting $n\to\infty$ we conclude that

$$\mu([x+\delta,\infty))\leq e^{-\lim_{n\to\infty}\left\{\lambda_n x-\log\int_{\mathbb{R}} e^{\lambda_n z}\mu(dz)\right\}}\lim_{n\to\infty} e^{-\delta\lambda_n}=0$$

for every $\delta>0$. Therefore $\mu((x,\infty))=0$. Hence by (3.12)

$$\lim_{n\to\infty}\int_{[x,\infty)} e^{\lambda_n(z-x)}\mu(\mathrm{d}z)=\mu(\{x\})=\exp\left(-I_X(x)\right).$$

Now

$$\begin{aligned}\mathbb{P}\left[\frac{1}{n}\sum_{i=1}^n X_i\in G\right]\geq \mathbb{P}\left[\frac{1}{n}\sum_{i=1}^n X_i=x\right] &\geq \mathbb{P}\left[X_i=x \text{ for all } i=1,\ldots,n\right]\\ &= \left(\mathbb{P}\left[X_1=x\right]\right)^n\end{aligned}$$

and therefore

$$\liminf_{n\to\infty}\frac{1}{n}\ln\mathbb{P}\left[\frac{1}{n}\sum_{i=1}^n X_i\in G\right]\geq \ln\mathbb{P}\left[X_1=x\right]=\ln\mu(\{x\})=-I_X(x).$$

Similarly one may handle the case that $x<a$.

Next we consider the case that $x \in G$ and there is λ_0 such that $I_X(x) = \lambda_0 x - \psi_X(\lambda_0)$. Then $(x-a)\lambda_0 \geq 0$ (see (3.8)), and λ_0 is a critical point of the function $\lambda \mapsto \lambda x - \psi_X(\lambda)$, so its partial derivative w.r.t. λ at λ_0 vanishes. Hence

$$x = \frac{\int_{\mathbb{R}} z e^{\lambda_0 z} \mu(dz)}{\int_{\mathbb{R}} e^{\lambda_0 z} \mu(dz)} . \tag{3.13}$$

Without losing generality, assume that $x \geq a$ so that $\lambda_0 \geq 0$. Choose $\delta > 0$ such that $(x-\delta, x+\delta) \subset G$. Then

$$\begin{aligned}
\mathbb{P}\left[\frac{1}{n}\sum_{i=1}^n X_i \in G\right] &\geq \mathbb{P}\left[\left|\frac{1}{n}\sum_{i=1}^n X_i - x\right| < \delta\right] \\
&\geq \mathbb{E}\left\{\frac{e^{\lambda_0 \sum_{i=1}^n X_i}}{e^{n\lambda_0(x+\delta)}} : \left|\frac{1}{n}\sum_{i=1}^n X_i - x\right| < \delta\right\} \\
&= e^{-n\lambda_0(x+\delta)} \mathbb{E}\left\{e^{\lambda_0 \sum_{i=1}^n X_i}; \left|\frac{1}{n}\sum_{i=1}^n X_i - x\right| < \delta\right\} \\
&= e^{-n\lambda_0(x+\delta)} \int_{\mathbb{R}^n} e^{\lambda_0 \sum_{i=1}^n z_i} 1_{\{|\frac{1}{n}\sum_{i=1}^n z_i - x| < \delta\}} \mu(dz_1)\cdots\mu(dz_n)
\end{aligned}$$

Define a new probability measure ν on $\mathbb{R}$ by

$$\nu(dz) = \frac{e^{\lambda_0 z}}{\int_{\mathbb{R}} e^{\lambda_0 z} \mu(dz)} \mu(dz)$$

which is a probability measure on $(\mathbb{R}, \mathscr{B}(\mathbb{R}))$. Then the previous inequality may be written as

$$\begin{aligned}
\mathbb{P}\left[\frac{1}{n}\sum_{i=1}^n X_i \in G\right] &\geq e^{-n\lambda_0(x+\delta)} \left(\int_{\mathbb{R}} e^{\lambda_0 z} \mu(dz)\right)^n \int_{\mathbb{R}^n} 1_{\{|\frac{1}{n}\sum_{i=1}^n z_i - x| < \delta\}} \nu(dz_1)\cdots\nu(dz_n) \\
&= e^{-n\lambda(x+\delta)} \left(\int_{\mathbb{R}} e^{\lambda_0 z} \mu(dz)\right)^n \mathbb{P}\left\{\left|\frac{1}{n}\sum_{i=1}^n Y_i - x\right| < \delta\right\}
\end{aligned}$$

where Y_i are i.i.d distribution ν, so that its mean (see equation (3.13))

$$\begin{aligned}
\mathbb{E}[Y_i] &= \int_{\mathbb{R}} z_i \nu(dz_i) = \int_{\mathbb{R}} \frac{z_i e^{\lambda_0 z_i}}{\int_{\mathbb{R}} e^{\lambda_0 z} \mu(dz)} \mu(dz_i) \\
&= \frac{1}{\int_{\mathbb{R}} e^{\lambda_0 z} \mu(dz)} \int_{\mathbb{R}} z_i e^{\lambda_0 z_i} \mu(dz_i) \\
&= x.
\end{aligned}$$

By the strong law of large numbers

$$\mathbb{P}\left\{\left|\frac{1}{n}\sum_{i=1}^n Y_i - x\right| < \delta\right\} \to 1 \text{ as } n \to \infty$$

and therefore the previous estimate yields that

$$\begin{aligned}\frac{1}{n}\ln\mathbb{P}\left[\frac{1}{n}\sum_{i=1}^{n}X_i\in G\right] &\geq -\lambda_0(x+\delta)+\psi_X(\lambda_0) \\ &\quad +\frac{1}{n}\log\mathbb{P}\left\{\left|\frac{1}{n}\sum_{i=1}^{n}Y_i-x\right|<\delta\right\} \\ &\to -\lambda_0(x+\delta)+\psi_X(\lambda_0) \quad \text{as } n\to\infty .\end{aligned}$$

Therefore

$$\begin{aligned}\liminf_{n\to\infty}\frac{1}{n}\ln\mathbb{P}\left[\frac{1}{n}\sum_{i=1}^{n}X_i\in G\right] &\geq -(\lambda_0 x-\psi_X(\lambda_0))-\delta\lambda_0 \\ &= -I_X(x)-\delta\lambda_0 \qquad \forall\delta>0.\end{aligned}$$

By letting $\delta\downarrow 0$ we obtain

$$\liminf_{n\to\infty}\frac{1}{n}\ln\mathbb{P}\left[\frac{1}{n}\sum_{i=1}^{n}X_i\in G\right]\geq -I_X(x) \quad \text{for every } x\in G.$$

Thus we have completed the proof of Cramér's theorem.

The proof is complete.

4 Gaussian distributions

Unfortunately it is a rather challenging problem for describing the distributions of general high-dimensional datasets. Here we give a detailed study of a class of random datasets with high-dimensional Gaussian distributions. The approach we have adapted is a primary version called *stochastic quantization*.

4.1 High-dimensional normal distributions

Let $X=(X_1,\cdots,X_D)$ be a (random) data set of D dimensions. Suppose X has a normal distribution, hence its distribution can be determined by its mean vector $\mu=(\mu_i)$ and its co-variance matrix $\Sigma=(\sigma_{ij})$, where $\mu_i=\mathbb{E}[X_i]$ and $\sigma_{ij}=\mathbb{E}\left[(X_i-\mu_i)(X_j-\mu_j)\right]$ (for $i,j=1,\cdots,D$). More precisely, the law of X is a probability measure on $\mathbb{R}^D$ with a probability density function (pdf) $G_\Sigma(x-\mu)$ with respect to the Lebesgue measure on $\mathbb{R}^D$, where

$$G_\Sigma(x)=\frac{1}{(2\pi)^{D/2}\sqrt{\det\Sigma}}\exp\left(-\frac{1}{2}x\cdot\Sigma^{-1}x\right) \qquad \text{for } x\in\mathbb{R}^D,$$

which is a central Gaussian density with co-variance matrix Σ. Here Σ^{-1} denotes the inverse of Σ. We will write $\Sigma^{-1}=(\sigma^{ij})$, so that $\sum_l\sigma^{il}\sigma_{lj}=\delta_{ij}$ for any $i,j\leq D$. $\Sigma=(\sigma_{ij})$ defines a scalar product on $\mathbb{R}^D$: $\langle x,y\rangle_{\Sigma^{-1}}=x\cdot\Sigma^{-1}y$ for $x,y\in\mathbb{R}^D$ and its a Hilbert norm $\|x\|_{\Sigma^{-1}}=\sqrt{x\cdot\Sigma^{-1}x}$. The Gaussian density

$$G_\Sigma(x)=\frac{1}{(2\pi)^{D/2}\sqrt{\det\Sigma}}\exp\left(-\frac{1}{2}\|x\|^2_{\Sigma^{-1}}\right) \qquad \text{for } x\in\mathbb{R}^D. \tag{4.1}$$

By means of change of variables we may see that $\int_{\mathbb{R}^D} G_\Sigma(x)\mathrm{d}x = 1$.

Lemma 4.1. *The norm distance*

$$\|x-y\|_{\Sigma^{-1}} = \sup\left\{f(x)-f(y) : f \in C^1 \text{ such that } \nabla f \cdot \Sigma \nabla f \leq 1\right\}.$$

Note that, since Σ is a constant *matrix, therefore the right-hand side is translation invariant.*

The proof is left as an exercise.

Remark 4.2. *A centered Gaussian random variable $X = (X_1, \cdots, X_D)$ is symmetric, that is, X and $-X$ have the same distribution.*

The distribution of a centered Gaussian random is parameterized by the co-variance matrix Σ, which is positive definite and symmetry, so that $|\sigma_{ij}| \leq \sigma_i \sigma_j$ where $\sigma_i^2 = \sigma_{ii}$ is the variance of X_i, where $i, j = 1, \cdots, D$. Since G_Σ is positive, it is a good idea to look at its logarithm

$$\ln G_\Sigma(x) = -\frac{D}{2}\ln(2\pi) - \frac{1}{2}\ln\det\Sigma - \frac{1}{2}x\cdot\Sigma^{-1}x.$$

To calculate its derivatives with respect to variables σ_{ij} (for $i < j$) and $\sigma_{ii} = \sigma_i^2$ (for $i = 1, \cdots, D$), we shall calculate its differential with respect to Σ.

Lemma 4.3. *Let $\Sigma(\varepsilon)$ (for $\varepsilon > 0$ but small enough so that $\Sigma(\varepsilon)$ remains positive definite) be a variation such that $\Sigma(0) = \Sigma$ and $\frac{d}{d\varepsilon}\big|_{\varepsilon=0}\Sigma(\varepsilon) = A$, where A is a symmetric matrix. Then*

$$\left.\frac{d}{d\varepsilon}\right|_{\varepsilon=0} \ln G_{\Sigma(\varepsilon)}(x) = -\frac{1}{2}tr(\Sigma^{-1}A) + \frac{1}{2}x\cdot\Sigma^{-1}A\Sigma^{-1}x \quad \text{for } x \in \mathbb{R}^D.$$

Proof. Clearly we have

$$\left.\frac{d}{d\varepsilon}\right|_{\varepsilon=0} \ln G_{\Sigma(\varepsilon)}(x) = -\frac{1}{2}\left.\frac{d}{d\varepsilon}\right|_{\varepsilon=0}\ln\det\Sigma(\varepsilon) - \frac{1}{2}x\cdot\left.\frac{d}{d\varepsilon}\right|_{\varepsilon=0}\Sigma(\varepsilon)^{-1}x. \tag{4.2}$$

Now observe that

$$\begin{aligned}\left.\frac{d}{d\varepsilon}\right|_{\varepsilon=0}\ln\det\Sigma(\varepsilon) &= \sum_{i=1}^{D}\frac{\frac{d}{d\varepsilon}\big|_{\varepsilon=0}\lambda_i(\varepsilon)}{\lambda_i} = \mathrm{tr}\left(\Sigma^{-1}\left.\frac{d}{d\varepsilon}\right|_{\varepsilon=0}\Sigma(\varepsilon)\right) \\ &= \mathrm{tr}\left(\Sigma^{-1}A\right),\end{aligned}$$

(which is called Jacobi's formula), and

$$0 = \left.\frac{d}{d\varepsilon}\right|_{\varepsilon=0}\left(\Sigma\Sigma^{-1}\right) = \Sigma\left.\frac{d}{d\varepsilon}\right|_{\varepsilon=0}\Sigma^{-1} + A\Sigma^{-1}$$

which yields that

$$\left.\frac{d}{d\varepsilon}\right|_{\varepsilon=0}\Sigma^{-1} = -\Sigma^{-1}A\Sigma^{-1}.$$

Using these equations in (4.2) we prove the lemma. □

Corollary 4.4. *Let $\Sigma = (\sigma_{ij})$ be symmetric and positive. Then*

$$\frac{\partial}{\partial \sigma_{ii}} G_\Sigma = \frac{1}{2} \frac{\partial^2}{\partial x_i^2} G_\Sigma \quad \text{for } i = 1, \ldots, D \tag{4.3}$$

and

$$\frac{\partial}{\partial \sigma_{ij}} G_\Sigma = \frac{\partial^2}{\partial x_j \partial x_i} G_\Sigma \quad \text{for } i \neq j. \tag{4.4}$$

Proof. Set $A = (a_{kl})$ where $a_{ii} = 1$ otherwise $a_{kl} = 0$ (i.e. $a_{kl} = \delta_{ki}\delta_{li}$) in Lemma 4.3. Then

$$\mathrm{tr}(\Sigma^{-1}A) = \sigma^{kl} a_{lk} = \sigma^{kl} \delta_{li}\delta_{ki} = \sigma^{ii}$$

and

$$x \cdot \Sigma^{-1} A \Sigma^{-1} x = x_k \sigma^{kb} a_{bc} \sigma^{cl} x_l = x_k \sigma^{ki} \sigma^{il} x_l = \left(\sum_{k=1}^{D} \sigma^{ki} x_k \right)^2$$

hence

$$\frac{\partial}{\partial \sigma_{ii}} \ln G_\Sigma(x) = \frac{1}{2} \left(\sum_{k=1}^{D} \sigma^{ki} x_k \right)^2 - \frac{1}{2} \sigma^{ii}.$$

Similarly, if $i \neq j$, we set in Lemma 4.3, $A = (a_{kl})$ where $a_{ij} = a_{ji} = 1$ (for $i \neq j$) and otherwise $a_{kl} = 0$. That is, $a_{kl} = \delta_{ki}\delta_{lj} + \delta_{li}\delta_{kj}$, we deduce that

$$\frac{\partial}{\partial \sigma_{ij}} \ln G_\Sigma(x) = \sum_{k=1}^{D} \sigma^{ki} x_k \sum_{k=1}^{D} \sigma^{lj} x_l - \sigma^{ij}.$$

On the other hand, we may differentiate G_Σ in the space variables $x = (x_1, \cdots, x_D)$ to obtain

$$\frac{\partial}{\partial x_i} G_\Sigma(x) = -G_\Sigma(x) \sum_{l=1}^{D} \sigma^{il} x_l$$

and

$$\frac{\partial^2}{\partial x_j \partial x_i} G_\Sigma(x) = G_\Sigma(x) \left(\sum_{k=1}^{D} \sigma^{jk} x_k \sum_{l=1}^{D} \sigma^{il} x_l - \sigma^{ij} \right).$$

Comparing the previous equations our corollary follows immediately. □

Remark 4.5. *Jacob's formula holds for any matrix valued function:*

$$\frac{d}{d\varepsilon} \det \Gamma(\varepsilon) = tr \left(adj(\Gamma(\varepsilon)) \frac{d}{d\varepsilon} \Gamma(\varepsilon) \right)$$

where $adj(\Gamma(\varepsilon))$ denotes the adjugate matrix of $\Gamma(\varepsilon)$. If $\Gamma(\varepsilon)^{-1}$ exists, then

$$\Gamma(\varepsilon)^{-1} = \frac{1}{\det \Gamma(\varepsilon)} adj(\Gamma(\varepsilon))$$

that we have learned from linear algebra, so that for this case

$$\frac{d}{d\varepsilon} \det \Gamma(\varepsilon) = \det \Gamma(\varepsilon) tr \left(\Gamma(\varepsilon)^{-1} \frac{d}{d\varepsilon} \Gamma(\varepsilon) \right)$$

which is Jacobi's formula for differentiation of determinants.

Theorem 4.6. *(Joag-Dev, Pelman and Pitt 1983) Let $f : \mathbb{R}^D \mapsto \mathbb{R}$ be a C^2-function whose derivatives are at most polynomial growth. Let*

$$h(\sigma_{ij}) = \int_{\mathbb{R}^D} f(x) G_\Sigma(x) dx$$

where $\Sigma = (\sigma_{ij})$ is symmetric and positive definite (so h is considered as a function of σ_{ij} for $i \leq j$). Suppose that $k < l$ is a pair, such that $\frac{\partial^2}{\partial x_k \partial x_l} f \geq 0$ on $\mathbb{R}^D$. Then h is increasing in the variable σ_{kl}.

Proof. By an inspection, we are justified for differentiating σ_{kl} under the integration, to obtain that

$$\frac{\partial}{\partial \sigma_{kl}} h = \int_{\mathbb{R}^D} f(x) \frac{\partial}{\partial \sigma_{kl}} G_\Sigma(x) \mathrm{d}x = \int_{\mathbb{R}^D} f(x) \frac{\partial^2}{\partial x_k \partial x_l} G_\Sigma(x) \mathrm{d}x,$$

where the second equality follows from (4.4). Integration by parts twice, we then deduce that

$$\frac{\partial}{\partial \sigma_{kl}} h = \int_{\mathbb{R}^D} G_\Sigma(x) \frac{\partial^2}{\partial x_k \partial x_l} f(x) \mathrm{d}x \geq 0$$

and the conclusion follows immediately. □

Theorem 4.7. (Slepian's Inequality) *If $X = (X_1, \cdots, X_D)$ and $Y = (Y_1, \cdots, Y_D)$ are two centered Gaussian vectors. Suppose that $\mathbb{E}X_i^2 = \mathbb{E}Y_i^2$ and $\mathbb{E}|X_i - X_j|^2 \leq \mathbb{E}|Y_i - Y_j|^2$ for any $i, j = 1, \ldots, D$. Then*

$$\mathbb{P}\left[\sup_i X_i \geq t\right] \leq \mathbb{P}\left[\sup_i Y_i \geq t\right]$$

for all t, and

$$\mathbb{E}\left[\sup_i X_i\right] \leq \mathbb{E}\left[\sup_i Y_i\right].$$

Proof. The assumptions imply that the variances $\mathbb{E}\left[X_i X_j\right] \geq \mathbb{E}\left[Y_i Y_j\right]$ for any i, j. Let $t > 0$. Since $1_{(-\infty,t]}$ is non-negative, and decreasing, we may choose a sequence of functions h_n which are C^1, decreasing, non-negative, such that h_n and their derivatives are uniformly bounded, and $h^{(n)} \to 1_{(-\infty,t]}$ as $n \to \infty$. Let $f_n(x_1, \cdots, x_D) = h_n(x_1) \cdots h_n(x_D)$. Then

$$\frac{\partial^2 f_n}{\partial x_i \partial x_j}(x) = h_n'(x_i) h_n'(x_j) \prod_{k \neq i,j} h_n(x_k) \geq 0$$

for any $i \neq j$. Since $\mathbb{E}X_i^2 = \mathbb{E}Y_i^2$ for every i, by Theorem 4.6, we have

$$\mathbb{E}\left[f_n(X_1, \cdots, X_D)\right] \geq \mathbb{E}\left[f_n(Y_1, \cdots, Y_D)\right].$$

Letting $n \to \infty$, we obtain that

$$\mathbb{P}\left[\sup_i X_i \leq t\right] \geq \mathbb{P}\left[\sup_i X_i \leq t\right]$$

which is equivalent to the first inequality. To show the second inequality, we observe that

$$\begin{aligned}
\mathbb{E}\left[\sup_i X_i\right] &= \mathbb{E}\left[\left(\sup_i X_i\right)^+\right] - \mathbb{E}\left[\left(\sup_i X_i\right)^-\right] \\
&= \int_0^\infty \mathbb{P}\left[\left(\sup_i X_i\right)^+ > t\right] \mathrm{d}t - \int_0^\infty \mathbb{P}\left[\left(\sup_i X_i\right)^- > t\right] \mathrm{d}t \\
&= \int_0^\infty \mathbb{P}\left[\sup_i X_i > t\right] \mathrm{d}t - \int_0^\infty \mathbb{P}\left[-\sup_i X_i > t\right] \mathrm{d}t \\
&= \int_0^\infty \mathbb{P}\left[\sup_i X_i > t\right] \mathrm{d}t - \int_0^\infty \mathbb{P}\left[\sup_i X_i < -t\right] \mathrm{d}t \\
&= \int_0^\infty \mathbb{P}\left[\sup_i X_i > t\right] \mathrm{d}t - \int_{-\infty}^0 \mathbb{P}\left[\sup_i X_i < t\right] \mathrm{d}t \\
&\leq \int_0^\infty \mathbb{P}\left[\sup_i Y_i > t\right] \mathrm{d}t - \int_{-\infty}^0 \mathbb{P}\left[\sup_i Y_i < t\right] \mathrm{d}t \\
&= \mathbb{E}\left[\sup_i Y_i\right]
\end{aligned}$$

which completes the proof. □

4.1.1 Sudakov-Fernique's inequality

Let $X = (X_1, \cdots, X_D)$ and $Y = (Y_1, \cdots, Y_D)$ be two independent D-dimensional centered Gaussian random vectors, whose co-variance matrices are $\Sigma = (\sigma_{ij})$ and $\tilde{\Sigma} = (\tilde{\sigma}_{ij})$ respectively. Let $X(t) = X\sin t + Y\cos t$ where $t \in (-\infty, \infty)$ is a real parameter, and let $\Sigma(t)$ denote the co-variance matrix for every t.

Lemma 4.8. *The following facts hold true.*

1) For every t, $X(t)$ and $\frac{d}{dt}X(t)$ are centered Gaussian with the co-variance matrices $\Sigma(t) = (\sigma_{ij}(t))$ and $\Sigma(t+\frac{\pi}{2})$ respectively, where

$$\sigma_{ij}(t) = \sigma_{ij}\sin^2 t + \tilde{\sigma}_{ij}\cos^2 t$$

for every t and $i, j = 1, \ldots, D$.

2) The co-variance between $X(t)$ and $\frac{d}{dt}X(t)$ are given by

$$\mathbb{E}\left(X_i(t)\frac{d}{dt}X_j(t)\right) = \frac{1}{2}\sin(2t)\left(\sigma_{ij} - \tilde{\sigma}_{ij}\right)$$

for every t, and $i, j = 1, \ldots, D$.

The proof follows from direct computation.

Lemma 4.9. *Consider the probability density $G_{\Sigma(t)}(x)$. Then*

$$\frac{\partial}{\partial t}G_{\Sigma(t)} = \sin(2t)\Delta_\Sigma G_{\Sigma(t)}(x) - \sin(2t)\Delta_{\tilde{\Sigma}} G_{\Sigma(t)}(x)$$

for all t.

Proof. By chain rule we have

$$\begin{aligned}
\frac{\partial}{\partial t}G_{\Sigma(t)} &= \sum_{i<j}\frac{\partial}{\partial\sigma_{ij}}G_{\Sigma(t)}\frac{d}{dt}\sigma_{ij}(t)+\sum_i\frac{\partial}{\partial\sigma_{ii}}G_{\Sigma(t)}\frac{d}{dt}\sigma_{ii}(t)\\
&= \sum_{i<j}\frac{\partial^2}{\partial x_i\partial x_j}G_{\Sigma(t)}\frac{d}{dt}\sigma_{ij}(t)+\frac{1}{2}\sum_i\frac{\partial^2}{\partial x_i\partial x_i}G_{\Sigma(t)}\frac{d}{dt}\sigma_{ii}(t)\\
&= \frac{1}{2}\sum_{i,j=1}^{D}\frac{\partial^2}{\partial x_i\partial x_j}G_{\Sigma(t)}\frac{d}{dt}\sigma_{ij}(t)\\
&= \frac{1}{2}\sin(2t)\sum_{i,j=1}^{D}\left(\sigma_{ij}-\tilde{\sigma}_{ij}\right)\frac{\partial^2}{\partial x_i\partial x_j}G_{\Sigma(t)}\\
&= \frac{1}{2}\sin(2t)\,(\Delta_\Sigma-\Delta_{\tilde{\Sigma}})\,G_{\Sigma(t)}\\
&= \frac{1}{2}\sin(2t)\left(\Delta_\Sigma G_{\Sigma(t)}-\Delta_{\tilde{\Sigma}}G_{\Sigma(t)}\right)
\end{aligned}$$

which completes the proof. □

Let h be a C^1 function on $\mathbb{R}^D$ with bounded derivatives, and let us consider the function

$$\phi(t)=\int_{\mathbb{R}^D}h(x)G_{\Sigma(t)}(x)\mathrm{d}x \quad \text{for } t\in\mathbb{R}.$$

Then

$$\begin{aligned}
\frac{d}{dt}\phi(t) &= \int_{\mathbb{R}^D}h(x)\frac{\partial}{\partial t}G_{\Sigma(t)}(x)\mathrm{d}x\\
&= \frac{1}{2}\sin(2t)\left(\int_{\mathbb{R}^D}h(x)\Delta_\Sigma G_{\Sigma(t)}(x)\mathrm{d}x-\int_{\mathbb{R}^D}h(x)\Delta_{\tilde{\Sigma}}G_{\Sigma(t)}(x)\mathrm{d}x\right)\\
&= \frac{1}{2}\sin(2t)\left(-\int_{\mathbb{R}^D}\nabla h(x)\cdot\Sigma\nabla G_{\Sigma(t)}(x)\mathrm{d}x+\int_{\mathbb{R}^D}\nabla h(x)\cdot\tilde{\Sigma}\nabla G_{\Sigma(t)}(x)\mathrm{d}x\right)
\end{aligned}$$

where the second equality follows from integration by parts. On the other hand

$$\nabla G_{\Sigma(t)}(x)=-G_{\Sigma(t)}(x)\Sigma(t)^{-1}x$$

so by substituting this into the previous equation for the differentiation of ϕ, we deduce that

$$\frac{d}{dt}\phi(t)=\sin(2t)\int_{\mathbb{R}^D}\nabla h(x)\cdot\left((\Sigma-\tilde{\Sigma})\Sigma(t)^{-1}x\right)G_{\Sigma(t)}(x)\mathrm{d}x$$

Since

$$\Sigma(t)=\sin^2 t\Sigma+\cos^2 t\tilde{\Sigma}=\sin^2 t(\Sigma-\tilde{\Sigma})+\tilde{\Sigma}$$

so that

$$(\Sigma-\tilde{\Sigma})\Sigma(t)^{-1}=\frac{1}{\sin^2 t}\left(I-\tilde{\Sigma}\Sigma(t)^{-1}\right)$$

Lemma 4.10. *Let* $h_D(x_1,\cdots,x_D) = \max\{x_1,\cdots,x_D\}$*. Then*

$$\begin{aligned} h_D(x_1,\cdots,x_D) &= \max\{x_i, h_{D-1}(x_1,\cdots,\hat{x}_i,\cdots,x_D)\} \\ &= \frac{1}{2}\left(x_i + h_{D-1}(x_1,\cdots,\hat{x}_i,\cdots,x_D) + |x_i - h_{D-1}(x_1,\cdots,\hat{x}_i,\cdots,x_D)|\right) \end{aligned}$$

and therefore

$$\frac{\partial}{\partial x_i} h_D = \frac{1}{2} + \frac{1}{2}\frac{x_i - h_{D-1}(x_1,\cdots,\hat{x}_i,\cdots,x_D)}{|x_i - h_{D-1}(x_1,\cdots,\hat{x}_i,\cdots,x_D)|}$$

4.2 Heat kernel

The heat kernel on $\mathbb{R}^D$ equipped with the *metric* Σ is defined by

$$p_\Sigma(t,x,y) = \frac{1}{(4\pi t)^{D/2}\sqrt{\det\Sigma}} \exp\left(-\frac{1}{4t}(y-x)\cdot\Sigma^{-1}(y-x)\right) \tag{4.5}$$

for $t > 0$, $x, y \in \mathbb{R}^D$. By definition, $G_\Sigma(x) = p_\Sigma(\frac{1}{2}, 0, x)$ and $p_\Sigma(t,x,y) = G_{2t\Sigma}(y-x)$.

For every pair $t > 0$ and $x \in \mathbb{R}^D$, $P_\Sigma(t,x,\mathrm{d}y) = p_\Sigma(t,x,y)\mathrm{d}y$ is a probability measure on $\mathbb{R}^D$ (with the Borel σ-algebra $\mathscr{B}(\mathbb{R}^D)$), that is,

$$A \mapsto P_\Sigma(t,x,A) = \int_A p_\Sigma(t,x,y)\mathrm{d}y$$

for $A \in \mathscr{B}(\mathbb{R}^D)$ defines clearly a probability on $(\mathbb{R}^D, \mathscr{B}(\mathbb{R}^D))$. The mapping $P_\Sigma : (t,x,A) \mapsto P_\Sigma(t,x,A)$, for $t > 0$, $x \in \mathbb{R}^D$ and $A \in \mathscr{B}(\mathbb{R}^D)$ is a *transition probability kernel* from $(\mathbb{R}^D, \mathscr{B}(\mathbb{R}^D))$ to $(0,\infty) \times \mathbb{R}^D$ with its Borel σ-algebra in the following sense:

1) for every pair $t > 0$ and $x \in \mathbb{R}^D$, $P_\Sigma(t,x,\cdot)$ is a probability measure on $\mathscr{B}(\mathbb{R}^D)$,

2) For every $A \in \mathscr{B}(\mathbb{R}^D)$, the function $(t,x) \mapsto P_\Sigma(t,x,A)$ is Borel measurable on $(0,\infty) \times \mathbb{R}^D$.

Indeed, the function $(t,x) \mapsto P_\Sigma(t,x,A)$ smooth in $t > 0$ and $x \in \mathbb{R}^D$ for this example.

Proposition 4.11. *For every* $x \in \mathbb{R}^D$*, the probability measures* $P_\Sigma(t,x,dy)$ *converge* weakly*, as* $t \downarrow 0$*, to Dirac measure* $\delta_x(dy)$*. That is*

$$\lim_{t\downarrow 0} \int_{\mathbb{R}^D} p_\Sigma(t,x,y) f(y) dy = f(x) \quad \text{for any } x \in \mathbb{R}^D$$

for every bounded and continuous function f*.*

Proof. Since Σ is positive definite and symmetry, so that there is a square root $\Sigma^{\frac{1}{2}}$ of Σ, a symmetric positive definite matrix such that $\Sigma^{\frac{1}{2}}\Sigma^{\frac{1}{2}} = \Sigma$. Making change of variable: $y = \sqrt{2t}\Sigma^{\frac{1}{2}} z + x$, whose Jacobi is $\det\Sigma^{\frac{1}{2}} = (2t)^{\frac{D}{2}}\sqrt{\det\Sigma}$. Therefore

$$\begin{aligned} \int_{\mathbb{R}^D} p_\Sigma(t,x,y) f(y)\mathrm{d}y &= \int_{\mathbb{R}^D} \frac{1}{(2\pi)^{D/2}} \exp\left(-\frac{1}{2}|z|^2\right) f\left(\sqrt{2t}\Sigma^{\frac{1}{2}} z + x\right) \mathrm{d}z \\ &\to \int_{\mathbb{R}^D} \frac{1}{(2\pi)^{D/2}} \exp\left(-\frac{1}{2}|z|^2\right) f(x)\,\mathrm{d}z = f(x) \end{aligned}$$

as $t \downarrow 0$, where the limit taking under integration is justified by Lebesgue's dominated convergence theorem [cf. A4: Integration]. □

In view of this lemma, we may define for each $t > 0$ an *operator* P_t which maps a function f to another function $P_t f$, by the following formula:

$$P_t f(x) = \int_{\mathbb{R}^D} f(y) p_\Sigma(t,x,y) \mathrm{d}y = \int_{\mathbb{R}^D} f(y) P_\Sigma(t,x,\mathrm{d}y) \quad \text{for } x \in \mathbb{R}^D$$

as long as the right-hand side is well defined. For example, for any f which is non-negative and is measurable, for f in $L^p(\mathbb{R}^D)$ for any $p \geq 1$, for f which is bounded and measurable, i.e. $f \in L^\infty(\mathbb{R}^D)$.

Remark 4.12. *If f is measurable and non-negative, then $P_t f$ is also non-negative. Therefore the operator P_t preserves the positivity.*

Remark 4.13. *If f is bounded and measurable, then, according to the theorem of taking derivatives under integration (cf. A4 Integration), the function $u(t,x) \equiv P_t f(x)$ is smooth in both variables $t > 0$ and $x \in \mathbb{R}^D$.*

Remark 4.14. *Suppose X is a random variable in $\mathbb{R}^D$ with a normal distribution $N(m,\Sigma)$, then with the definitions above, $\mathbb{E}[f(X)] = P_2 f(m)$.*

By a slightly complicated but completely elementary computation, we prove the following lemma.

Proposition 4.15. *The heat kernel $\{p_\Sigma(t,x,y) : t > 0\}$ possesses the following properties.*

1) $p_\Sigma(t,x,y)$ is positive, smooth for $t > 0$, x,y in $\mathbb{R}^D$, and $p_\Sigma(t,x,y) = p_\Sigma(t,x,y)$ for any $t > 0$ and x,y.

2) The following equality holds:

$$p_\Sigma(s,x,z) p_\Sigma(t,z,y) = p_\Sigma(s+t,x,y) p_\Sigma\left(\frac{2st}{t+s}, \frac{t}{t+s}x + \frac{s}{t+s}y, z\right) \tag{4.6}$$

for any $s > 0, t > 0$ and $x,y,z \in \mathbb{R}^D$.

3) Chapman-Kolmogorov's equality holds:

$$\int_{\mathbb{R}^D} p_\Sigma(s,x,z) p_\Sigma(t,z,y) dz = p_\Sigma(s+t,x,y) \tag{4.7}$$

for any $s > 0, t > 0$ and $x,y \in \mathbb{R}^D$.

Proof. 1) is obvious by the expression (4.5). Clearly (4.7) follows by integrating (4.6) and the fact

$$\int_{\mathbb{R}^D} p_\Sigma\left(\frac{2st}{t+s}, a, z\right) \mathrm{d}z = 1$$

for every $a \in \mathbb{R}^D$. To show 2) we use the polar identity for the scalar product $\langle x,y \rangle_{\Sigma^{-1}}$ which yields that

$$\left\| \frac{z-x}{\sqrt{2s}} \right\|_{\Sigma^{-1}}^2 + \left\| \frac{y-z}{\sqrt{2t}} \right\|_{\Sigma^{-1}}^2 = \left\| \frac{z-a}{\sqrt{\frac{2st}{t+s}}} \right\|_{\Sigma^{-1}}^2 + \left\| \frac{y-x}{\sqrt{2t+2s}} \right\|_{\Sigma^{-1}}^2$$

where $a = \frac{t}{t+s}x + \frac{s}{t+s}y$, and the equality (4.6) follows immediately. □

Proposition 4.16. *The family of operators P_t for $t > 0$ together with $P_0 = I$ the identity operator forms a semi-group of linear operators, denoted by $(P_t)_{t\geq 0}$, in the following sense.*

1) For each $t \geq 0$, P_t is linear: $P_t(f+g) = P_t f + P_t g$ and $P_t(cf) = cP_t f$ for any constant c, for any measurable function f, g which are bounded, or non-negative.

2) For any $s,t \geq 0$, it holds that $P_{t+s}f = P_t(P_s f)$ for any measurable function f which is bounded or non-negative.

3) For each $t > 0$, P_t is self-adjoint, and P_t is a contraction in $L^p(\mathbb{R}^d)$ for every $p \geq 1$.

The first item follows from the definition of P_t and the second item shows that $P_{t+s} = P_t \circ P_s$ (often shall write $P_t P_s$ for simplicity), called the semi-group property. The family $(P_t)_{t\geq 0}$ is the heat semi-group on $\mathbb{R}^D$ with the metric Σ. 3) follows from the symmetry that $p_\Sigma(t,x,y) = p_\Sigma(t,y,x)$. Indeed

$$\begin{aligned} \int fP_t g &= \int\int f(x)g(y)p_\Sigma(t,x,y)\mathrm{d}y \\ &= \int\int f(x)g(y)p_\Sigma(t,y,x)\mathrm{d}y \\ &= \int gP_t f \end{aligned}$$

for any $f,g \in L^2(\mathbb{R}^D)$.

Proposition 4.17. *The Lebesgue measure is the invariant measure of $(P_t)_{t>0}$, that is,*

$$\int_{\mathbb{R}^D} P_t f(x)dx = \int_{\mathbb{R}^D} f(x)dx \quad \text{for all } t > 0$$

for any $f \in L^1(\mathbb{R}^D)$.

Remark 4.18. *Let us recall, for a given $p \geq 1$, that $L^p(\mathbb{R}^D)$ denotes the normed space of all p-th integrable functions (identified up to almost surely) with respect to the Lebesgue measure on $\mathbb{R}^D$ whose norm $\|\cdot\|_p$ defined by $\|f\|_p = \left(\int_{\mathbb{R}^D} |f(x)|^p dx\right)^{\frac{1}{p}}$. $L^p(\mathbb{R}^D)$ is complete and separable, so that $L^p(\mathbb{R}^D)$ is a Banach space. Similarly $L^\infty(\mathbb{R}^D)$ is a separable Banach space to. As a matter of fact, for every $p \geq 1$, P_t can be extended to be a linear operator from $L^p(\mathbb{R}^D)$ to $L^p(\mathbb{R}^D)$ such that $P_{t+s} = P_t \circ P_s$ for any $s,t > 0$. Every P_t is a contraction on $L^p(\mathbb{R}^D)$, i.e. $\|P_t f\|_p \leq \|f\|_p$ for every $f \in L^p(\mathbb{R}^D)$. Moreover $P_t f \mapsto f$ in $L^p(\mathbb{R}^D)$ as $t \downarrow 0$.*

4.3 Geometric properties of normal distributions

In this part we study the geometric aspects of the heat kernel $p_\Sigma(t,x,y)$. Firstly we observe that

$$\ln p_\Sigma(t,x,y) = -\frac{D}{2}\ln(4\pi t) - \frac{1}{2}\ln\det\Sigma - \frac{1}{4t}(y-x)\cdot\Sigma^{-1}(y-x)$$

which allows us to work out the derivatives of p_Σ with respect to all variables $t > 0$, x (equivalently y too) and $\Sigma = (\sigma_{ij})$. In fact

$$\frac{\partial}{\partial t}\ln p_\Sigma(t,x,y) = -\frac{D}{2t} + \frac{1}{4t^2}(y-x)\cdot\Sigma^{-1}(y-x), \tag{4.8}$$

$$\frac{\partial}{\partial x_i}\ln p_\Sigma(t,x,y) = \frac{1}{2t}\sum_{l=1}^{D}\sigma^{il}(y^l - x^l). \tag{4.9}$$

We therefore have proved the following important fact.

Theorem 4.19. *Let $\Sigma = (\sigma_{ij})$ be a positive definite and symmetric $D\times D$ matrix, and $\Delta_\Sigma = \sum_{i,j=1}^{D}\sigma_{ij}\frac{\partial^2}{\partial x_j \partial x_i}$ a differential operator of second order in $\mathbb{R}^D$. Then $p_\Sigma(t,x,y)$ is the fundamental solution to the heat operator $\frac{\partial}{\partial t} - \Delta_\Sigma$ in the following sense:*

$$\left(\frac{\partial}{\partial t} - \Delta_\Sigma\right) p_\Sigma(t,x,y) = 0 \quad \text{for } t>0, x,y\in\mathbb{R}^D$$

(where Δ_Σ either acts on the variable x or y with the other variables being fixed), and $p_\Sigma(t,x,y)dy \to \delta_x$ weakly as $t\downarrow 0$ for each x.

Proof. First we have the time derivative of p_Σ is given by

$$\frac{\partial}{\partial t}p_\Sigma(t,x,y) = \left(-\frac{D}{2t} + \frac{1}{4t^2}(y-x)\cdot\Sigma^{-1}(y-x)\right)p_\Sigma(t,x,y).$$

While the space derivative of $p_\Sigma(t,x,y)$ can be calculated as the following:

$$\frac{\partial^2}{\partial x_i \partial x_j}\ln p_\Sigma(t,x,y) = -\frac{1}{2t}\sigma^{ij}$$

which reflects the fact that $\ln p_\Sigma(t,x,y)$ is a quadratic polynomial of x,y. Therefore

$$\begin{aligned}
\frac{\partial^2}{\partial x_i \partial x_j}p_\Sigma(t,x,y) &= \frac{\partial}{\partial x_j}\left(p_\Sigma(t,x,y)\frac{\partial}{\partial x_i}\ln p_\Sigma(t,x,y)\right)\\
&= \frac{\partial}{\partial x_j}p_\Sigma(t,x,y)\frac{\partial}{\partial x_i}\ln p_\Sigma(t,x,y) + p_\Sigma(t,x,y)\frac{\partial^2}{\partial x_j \partial x_i}\ln p_\Sigma(t,x,y)\\
&= \left(\frac{\partial}{\partial x_j}\ln p_\Sigma(t,x,y)\frac{\partial}{\partial x_i}\ln p_\Sigma(t,x,y) + \frac{\partial^2}{\partial x_j \partial x_i}\ln p_\Sigma(t,x,y)\right)p_\Sigma(t,x,y)\\
&= \left(\frac{1}{4t^2}\sum_{k,l=1}^{D}\sigma^{ik}\sigma^{jl}(y_l - x_l)(y_k - x_k) - \frac{1}{2t}\sigma^{ij}\right)p_\Sigma(t,x,y),
\end{aligned}$$

and therefore

$$\begin{aligned}
\Delta_\Sigma p_\Sigma(t,x,y) &= \left(\frac{1}{4t^2}\sum_{k,l=1}^{D}\sum_{i,j=1}^{D}\sigma_{ij}\sigma^{ik}\sigma^{jl}(y_l - x_l)(y_k - x_k) - \frac{1}{2t}\sum_{i,j=1}^{D}\sigma_{ij}\sigma^{ij}\right)p_\Sigma(t,x,y)\\
&= \left(\frac{1}{4t^2}(y-x)\cdot\Sigma^{-1}(y-x) - \frac{D}{2t}\right)p_\Sigma(t,x,y)\\
&= \frac{\partial}{\partial t}p_\Sigma(t,x,y).
\end{aligned}$$

This completes the proof. □

Corollary 4.20. *Suppose that f is a bounded measurable function on $\mathbb{R}^D$. Let $u(t,x) = P_t f(x)$ (for $t > 0$ and $x \in \mathbb{R}^D$). Then u is smooth on $(0,\infty)\times\mathbb{R}^D$, and u solves the heat equation*

$$\left(\frac{\partial}{\partial t} - \Delta_\Sigma\right) u(t,x) = 0 \quad \text{in } (0,\infty)\times\mathbb{R}^D. \tag{4.10}$$

If in addition f is continuous, then $u(t,x) \to f(x)$ as $t \downarrow 0$ for every $x \in \mathbb{R}^D$.

Proof. Since $u(t,x) = \int_{\mathbb{R}^D} f(y) p_\Sigma(t,x,y)\mathrm{d}y$, all conclusions follow by using the theorem of differentiation under intergrals. □

The heat equation (4.10) may be written as $\frac{\partial}{\partial t} P_t f = \Delta_\Sigma(P_t f)$ for every bounded (or non-negative) measurable function f, so by abusing notation, the last equation may be written as $\frac{\partial}{\partial t} P_t = \Delta_\Sigma P_t$ for every $t > 0$. In this sense, we say Δ_Σ is the infinitesimal generator of the heat semi-group $(P_t)_{t\geq 0}$, and formally write as $P_t = \mathrm{e}^{\Delta_\Sigma t}$ for $t > 0$.

Remark 4.21. *The heat semigroup P_t (hence its heat kernel $p_\Sigma(t,x,y)$) is uniquely determined by the second-order differential operator Δ_Σ, and equivalently determined by the quadratic form:*

$$\begin{aligned}\int_{\mathbb{R}^D} -\psi(x)\Delta_\Sigma\varphi(x)dx &= \int_{\mathbb{R}^D} -\psi(x)\sigma_{ij}\frac{\partial^2}{\partial x_j \partial x_i}\varphi(x)dx \\ &= \int_{\mathbb{R}^D} \sigma_{ij}\frac{\partial\varphi}{\partial x_i}\frac{\partial\psi}{\partial x_j}dx\end{aligned}$$

for any φ,ψ belonging to $W^{2,1}(\mathbb{R}^D)$.

Proposition 4.22. *It holds that*

$$\|\nabla \ln p_\Sigma(t,x,y)\|_\Sigma^2 - \frac{\partial}{\partial t}\ln p_\Sigma(t,x,y) = \frac{D}{2t} \tag{4.11}$$

for every $t > 0$, $x,y \in \mathbb{R}^D$, where $\|a\|^2 = a\cdot\Sigma a$ [Note that it is not $\|a\|^2_{\Sigma^{-1}}$].

Proof. The verification is completely elementary. In fact

$$\frac{\partial}{\partial t}\ln p_\Sigma(t,x,y) = -\frac{D}{2t} + \frac{1}{4t^2}(y-x)\cdot\Sigma^{-1}(y-x), \tag{4.12}$$

and

$$\sum_{i,j}\sigma_{ij}\frac{\partial}{\partial x_i}\ln p_\Sigma(t,x,y)\frac{\partial}{\partial x_j}\ln p_\Sigma(t,x,y) = \frac{1}{4t^2}(y-x)\cdot\Sigma^{-1}(y-x) \tag{4.13}$$

which completes the proof. □

Exercise. [Hard] Suppose $u(x,t) = P_t\varphi$ where φ is a positive continuous function. Let $f(x,t) = \ln u(x,t)$, $X = \nabla \ln f\cdot\Sigma\nabla\ln f$ and $Y = \frac{\partial}{\partial t}\ln f$.

(1) Work out $\left(\frac{\partial}{\partial t} - \Delta_\Sigma\right)X$ and $\left(\frac{\partial}{\partial t} - \Delta_\Sigma\right)Y$.

(2) Show that

$$X(x,t) - Y(x,t) \leq \frac{D}{2t}$$

for all x and $t > 0$.

[*Hint*: you may look at the paper by D. Bakry and Z. Qian: Harnack inequalities on a manifold with positive or negative Ricci curvature, in *Revista Matemática Iberoamericana* (1999) Volume: 15, Issue: 1, page 143-179.]

5 The Ornstein-Uhlenbeck semi-group

In the previous section we have studied a few properties of Gaussian measures on $\mathbb{R}^D$. In particular we demonstrate that the Lebesgue measure is the invariant measure of heat semi–group $P_t = e^{t\Delta_\Sigma}$ (for $t \geq 0$) defined via the heat kernel $p_\Sigma(t,x,y)$. In this section we introduce a dynamical system whose invariant measure is the Gaussian measure $G_\Sigma(x)\mathrm{d}x$. More precisely, we construct a semi-group Q_t (for $t > 0$) in analogs with the heat semigroup, such that $G_\Sigma(x)\mathrm{d}x$ is the invariant measure of $(Q_t)_{t>0}$.

For simplicity we use $\gamma(\mathrm{d}x)$ denote the Gaussian measure $G_\Sigma(x)\mathrm{d}x$ on the Borel σ-algebra $\mathscr{B}(\mathbb{R}^D)$, if no confusion may arise. Let $L^p(\gamma)$ (for every $p \in [1,\infty]$) denote the L^p-space over the measure space $(\mathbb{R}^D, \mathscr{B}(\mathbb{R}^D), \gamma)$.

5.1 The Mehler formula

The simplest way to construct the Orenstein-Uhlenbeck semigroup Q_t is to apply the Mehler formula. For every $t > 0$ define linear operator $Q_t : f \mapsto Q_t f$ by setting

$$Q_t f(x) = \int_{\mathbb{R}^D} f\left(e^{-t}x + \sqrt{1-e^{-2t}}y\right) G_\Sigma(y)\mathrm{d}y \tag{5.1}$$

for every $t > 0$ and $x \in \mathbb{R}^D$, where f is a Borel measurable function as long as the integral on the right-hand is defined – for example f is bounded or f is non-negative. Clearly $Q_t 1 = 1$ for every $t > 0$, and $Q_t f \geq 0$ as long as f is non-negative.

Making a change of variable one can rewrite the above formula as the following

$$\begin{aligned} Q_t f(x) &= \int_{\mathbb{R}^D} f(y) \frac{\exp\left(-\frac{1}{2\left(1-e^{-2t}\right)}(y-e^{-t}x)\cdot\Sigma^{-1}(y-e^{-t}x)\right)}{(2\pi(1-e^{-2t}))^{\frac{D}{2}}\sqrt{\det\Sigma}}\mathrm{d}y \\ &= \int_{\mathbb{R}^D} f(y) q_\Sigma(t,x,y) G_\Sigma(\mathrm{d}y) \end{aligned} \tag{5.2}$$

where

$$q_\Sigma(t,x,y) = \frac{1}{G_\Sigma(y)} \frac{\exp\left(-\frac{1}{2\left(1-e^{-2t}\right)}(y-e^{-t}x)\cdot\Sigma^{-1}(y-e^{-t}x)\right)}{(2\pi(1-e^{-2t}))^{\frac{D}{2}}\sqrt{\det\Sigma}} \tag{5.3}$$

is called the transition probability density function of the OU semi-group.

Recall that the heat kernel associated a positive definite and symmetric Σ is given by

$$p_\Sigma(t,x,y) = \frac{1}{(4\pi t)^{D/2}\sqrt{\det\Sigma}} \exp\left(-\frac{1}{4t}(y-x)\cdot\Sigma^{-1}(y-x)\right)$$

so that

$$q_\Sigma(t,x,y) = p_\Sigma\left(\frac{1-e^{-2t}}{2}, e^{-t}x, y\right)\frac{1}{G_\Sigma(y)} \tag{5.4}$$

for every $t > 0$ and $x, y \in \mathbb{R}^D$. Here the Gaussian density $G_\Sigma(y)$ is inserted in the definition of the probability density kernel q_Σ, since we expect that the Gaussian measure $G_\Sigma(y)\mathrm{d}y$ is the invariant measure for Q_t.

Lemma 5.1. *Suppose f is continuous and is of at most polynomial growth, then*

$$\lim_{t\downarrow 0} Q_t f(x) = f(x) \quad \text{and} \quad \lim_{t\to\infty} Q_t f(x) = \int_{\mathbb{R}^D} f(y) G_\Sigma(dy) \tag{5.5}$$

for every $x \in \mathbb{R}^D$.

This follows immediately from the Mehler formula (5.1).

Lemma 5.2. *The transition probability function of the Orenstein-Uhlenbeck semi-group is given by*

$$q_\Sigma(t,x,y) = \frac{1}{(1-e^{-2t})^{\frac{D}{2}}} \exp\left(-\frac{1}{2}\frac{y\cdot\Sigma^{-1}y + x\cdot\Sigma^{-1}x - 2e^t x\cdot\Sigma^{-1}y}{e^{2t}-1}\right) \tag{5.6}$$

for every $t > 0$ and for any x, y. In particular q is symmetric: $q_\Sigma(t,x,y) = q_\Sigma(t,y,x)$.

Proof. By (5.3) the transition probability density function

$$q_\Sigma(t,x,y) = \frac{1}{(1-e^{-2t})^{\frac{D}{2}}} \exp\left(-I(t,x,y)\right)$$

where

$$I(t,x,y) = \frac{1}{2(1-e^{-2t})}(y - e^{-t}x)\cdot\Sigma^{-1}(y - e^{-t}x) - \frac{1}{2} y\cdot\Sigma^{-1} y.$$

Collecting the quadratic terms of y together we have

$$I(t,x,y) = \frac{1}{2}\frac{e^{-2t}}{1-e^{-2t}}\left(y\cdot\Sigma^{-1}y + x\cdot\Sigma^{-1}x - 2e^t x\cdot\Sigma^{-1}y\right)$$

ant the conclusion follows immediately. □

Lemma 5.3. *We have*

$$q_\Sigma(s,x,y) q_\Sigma(t,y,z) = q_\Sigma(s+t,x,z) q_\Sigma(T(s,t), c_{s,t}(x,z), y)$$

where $T = T(s,t)$ and $c_{s,t}(x,z)$ are given by

$$\frac{1}{e^{2T}-1} = \frac{1}{e^{2s}-1} + \frac{1}{e^{2t}-1}$$

and

$$c_{s,t}(x,z) = \frac{e^T}{e^{2(t+s)}-1}\left((e^{2t}-1)e^s x + (e^{2s}-1)e^t z\right)$$

for $s, t > 0$ and $x, y, z \in \mathbb{R}^D$.

Therefore the Chapman-Kolmogorov equality holds

$$\int_{\mathbb{R}^D} q_\Sigma(s,x,y) q_\Sigma(t,y,z) G_\Sigma(y) dy = q_\Sigma(s+t,x,z)$$

for any $s, t > 0$ and $x, z \in \mathbb{R}^D$.

Proof. Let $a(t) = \left(1 - e^{-2t}\right)^{\frac{D}{2}}$ and

$$I(t,x,y) = \frac{y \cdot \Sigma^{-1} y + x \cdot \Sigma^{-1} x - 2e^t x \cdot \Sigma^{-1} y}{e^{2t} - 1}.$$

Then $q_\Sigma(s,x,y) = a(s)^{-1} \exp\left(-\frac{1}{2} I(s,x,y)\right)$, and

$$\frac{q_\Sigma(s,x,y) q_\Sigma(t,y,z)}{q_\Sigma(s+t,x,z)} = \frac{a(s+t)}{a(s)a(t)} \exp\left(-\frac{1}{2}\left(I(s,x,y) + I(t,y,z) - I(s+t,x,z)\right)\right).$$

Let us calculate $J = I(s,x,y) + I(t,y,z) - I(s+t,x,z)$. By definition $T = T(s,t) > 0$ is given b

$$\frac{1}{e^{2T} - 1} = \frac{1}{e^{2s} - 1} + \frac{1}{e^{2t} - 1} = \frac{e^{2t} + e^{2s} - 2}{(e^{2s} - 1)(e^{2t} - 1)}.$$

Hence

$$e^T = \sqrt{1 + \frac{(e^{2s} - 1)(e^{2t} - 1)}{e^{2t} + e^{2s} - 2}} = \sqrt{\frac{e^{2(t+s)} - 1}{e^{2t} + e^{2s} - 2}}$$

and

$$\frac{a(s+t)}{a(s)a(t)} = \left(\frac{e^{2(t+s)} - 1}{(e^{2s} - 1)(e^{2t} - 1)}\right)^{\frac{D}{2}} = \left(\frac{e^{2T}}{e^{2T} - 1}\right)^{\frac{D}{2}} = \frac{1}{a(T)}.$$

Moreover, one can verify that

$$J = \frac{1}{e^{2T} - 1}\left(y \cdot \Sigma^{-1} y - 2e^T c \cdot \Sigma^{-1} y + c \cdot \Sigma^{-1} c\right)$$

and therefore

$$\frac{q_\Sigma(s,x,y) q_\Sigma(t,y,z)}{q_\Sigma(s+t,x,z)} = \frac{1}{a(T)} \exp\left(-\frac{1}{2} I(T,c,y)\right)$$

which completes the proof. □

In what follows we will work with a fixed symmetric, positive definite $D \times D$ matrix Σ, and we will use $\gamma(\mathrm{d}x)$ to denote the Gaussian measure $G_\Sigma(x)\mathrm{d}x$ on $(\mathbb{R}^D, \mathscr{B}(\mathbb{R}^D))$. Let $L^p(\gamma)$ denote the L^p-space over the probability space $(\mathbb{R}^D, \mathscr{B}(\mathbb{R}^D), \gamma)$.

Proposition 5.4. *The OU semi-group $(Q_t)_{t \geq 0}$ possesses the following properties.*

1) For every $t > 0$, Q_t is symmetric:

$$\int_{\mathbb{R}^D} f(x) Q_t g(x) \gamma(dx) = \int_{\mathbb{R}^D} g(x) Q_t f(x) \gamma(dx)$$

for any f and g belonging to $L^2(\gamma)$. In particular, γ is an invariant measure of Q_t. That is

$$\int_{\mathbb{R}^D} Q_t f(x) \gamma(dx) = \int_{\mathbb{R}^D} f(x) \gamma(dx)$$

2) $(Q_t)_{t \geq 0}$ is a semi-group: $Q_s Q_t = Q_{t+s}$ for any $s,t \geq 0$, where $Q_0 = I$ is the identity operator.

3) For every $t > 0$, Q_t is a contraction on $L^p(\gamma)$, in the sense that $\|Q_t f\|_{L^p(\gamma)} \leq \|f\|_{L^p(\gamma)}$ for every $p \geq 1$ and $f \in L^p(\gamma)$.

Proof. 1) follows from the fact that $q_\Sigma(t,x,y) = q_\Sigma(t,y,x)$:

$$\int g(x)Q_t f(x)\gamma(\mathrm{d}x) = \int\int g(x)f(y)q_\Sigma(t,y,x)\gamma(\mathrm{d}y)\gamma(\mathrm{d}x)$$
$$= \int f(y)Q_t g(y)\gamma(\mathrm{d}y).$$

2) follows from Lemma 5.3

$$\begin{aligned} Q_s Q_t f(x) &= \int\int q_\Sigma(s,x,y)q_\Sigma(t,y,z)f(z)\gamma(\mathrm{d}z)\gamma(\mathrm{d}y) \\ &= \int q_\Sigma(s+t,x,z)f(z)\left(\int q_\Sigma(T(s,t),c_{s,t}(x,z),y)\gamma(\mathrm{d}y)\right)\gamma(\mathrm{d}z) \\ &= Q_{t+s}f(x) \end{aligned}$$

which proves the semi-group property.

We only need to prove 3) for bounded and continuous function f. Then, by using Hölder's inequality,

$$\begin{aligned} \|Q_t f\|^p_{L^p(\gamma)} &= \int \left|\int f(y)q_\Sigma(t,x,y)\gamma(dy)\right|^p \gamma(dx) \\ &\le \int\int |f(y)|^p q_\Sigma(t,x,y)\gamma(dy)\gamma(dx) \\ &= \int\int |f(y)|^p q_\Sigma(t,y,x)\gamma(dy)\gamma(dx) \\ &= \int |f(y)|^p \gamma(dy) \end{aligned}$$

where the inequality follows from the Hölder's inequalty to f and constant function 1 with probability measure $m(dy) = q_\Sigma(t,x,y)\gamma(dy)$ for each x, and the last equality follows from Fubini's theorem by integrating the variable x first to give 1. ☐

Using the fact that the space $C_b(\mathbb{R}^D)$ of bounded and continuous functions is dense in $L^p(\mathbb{R}^D)$ for every $p \ge 1$, the following proposition follows immediately.

Proposition 5.5. *Suppose that* $f \in L^p(\gamma)$,

$$\lim_{t\to\infty}\left\|Q_t f - \int_{\mathbb{R}^D} f d\gamma\right\|_{L^p(\gamma)} = 0$$

and

$$\lim_{t\downarrow 0}\|Q_t f - f\|_{L^p(\gamma)} = 0.$$

Remark 5.6. *Let* $t,s > 0$. *Consider two linear mappings* $T,S : \mathbb{R}^D \times \mathbb{R}^D \to \mathbb{R}^D$ *defined by*

$$T(x,y) = e^{-t}x + \sqrt{1-e^{-2t}}y$$

and

$$S(y,z) = e^{-s}\frac{\sqrt{1-e^{-2t}}}{\sqrt{1-e^{-2(t+s)}}}y + \frac{\sqrt{1-e^{-2s}}}{\sqrt{1-e^{-2(t+s)}}}z$$

for $x, y, z \in \mathbb{R}^D$. Then

$$\int_{\mathbb{R}^D \times \mathbb{R}^D} f \circ T(x,y) \gamma(dx)\gamma(dy) = \int_{\mathbb{R}^D} f(x)\gamma(dx)$$

and similarly

$$\int_{\mathbb{R}^D \times \mathbb{R}^D} f \circ S(y,z) \gamma(dy)\gamma(dz) = \int_{\mathbb{R}^D} f(x)\gamma(dx)$$

for any Borel measurable function f. The proof is left as an exercise.

We next establish the most remarkable property of the OU semi-group $(Q_t)_{t>0}$.

Proposition 5.7. *1) For every $t > 0$ it holds that*

$$\frac{\partial}{\partial x^i} Q_t f = e^{-t} Q_t \left(\frac{\partial f}{\partial x^i} \right)$$

for any C^1 function f whose partial derivatives $\frac{\partial f}{\partial x^i}$ are γ-integrable, where $i = 1, \ldots, D$.

Proof. Suppose f is differentiable with a compact support, then we may differentiate $Q_t f(x)$ under integration to obtain

$$\begin{aligned} \frac{\partial Q_t f}{\partial x^i}(x) &= \int_{\mathbb{R}^D} \frac{\partial}{\partial x^i} f\left(e^{-t}x + \sqrt{1-e^{-2t}}y\right) \gamma(\mathrm{d}y) \\ &= \int_{\mathbb{R}^D} e^{-t} \frac{\partial f}{\partial x^i}\left(e^{-t}x + \sqrt{1-e^{-2t}}y\right) \gamma(\mathrm{d}y) \\ &= e^{-t} Q_t \left(\frac{\partial f}{\partial x^i} \right) \end{aligned}$$

which completes the proof. □

Theorem 5.8. *(*Domination inequality*) The following domination inequality holds*

$$\sqrt{\nabla Q_t f \cdot \Sigma \nabla Q_t f} \le e^{-t} Q_t \left(\sqrt{\nabla f \cdot \Sigma \nabla f} \right) \tag{5.7}$$

for every C^1 function f and $t \ge 0$. The domination inequality implies the following weak domination inequality

$$\nabla Q_t f \cdot \Sigma \nabla Q_t f \le e^{-2t} Q_t \left(\nabla f \cdot \Sigma \nabla f \right)$$

for every C^1 function f and $t \ge 0$.

Proof. The proof relies on the Cauchy-Schwartz inequality $|a \cdot \Sigma b| \le \sqrt{a \cdot \Sigma a}\sqrt{b \cdot \Sigma b}$ for any $a, b \in \mathbb{R}^D$ (its proof is left as an exercise). By an approximation procedure, we may prove the domination inequality for C^1-function f with bounded derivatives. For simplicity, use f_i to denote the partial derivative $\frac{\partial}{\partial x_i} f$. By the Mehler formula

$$\frac{\partial}{\partial x_i} Q_t f(x) = e^{-t} \int_{\mathbb{R}^D} f_i \left(e^{-t}x + \sqrt{1-e^{-2t}}y\right) \gamma(\mathrm{d}y)$$

for $i=1,\dots,D$, and Fubini's theorem, we have

$$\begin{aligned}
\nabla Q_t f\cdot\Sigma\nabla Q_t f &= e^{-2t}\int_{\mathbb{R}^D}\int_{\mathbb{R}^D}\nabla f(e^{-t}x+\sqrt{1-e^{-2t}}y)\cdot\Sigma\nabla f(e^{-t}x+\sqrt{1-e^{-2t}}z)\gamma(\mathrm{d}y)\gamma(\mathrm{d}z)\\
&\le e^{-2t}\int\int_{\mathbb{R}^D}\sqrt{\nabla f\cdot\Sigma\nabla f|_{e^{-t}x+\sqrt{1-e^{-2t}}y}}\sqrt{\nabla f\cdot\Sigma\nabla f|_{e^{-t}x+\sqrt{1-e^{-2t}}z}}\gamma(\mathrm{d}y)\gamma(\mathrm{d}z)\\
&= e^{-2t}\left(\int_{\mathbb{R}^D}\sqrt{\nabla f(e^{-t}x+\sqrt{1-e^{-2t}}y)\cdot\Sigma\nabla f(e^{-t}x+\sqrt{1-e^{-2t}}y)}\gamma(\mathrm{d}y)\right)^2\\
&= e^{-2t}\left(Q_t\left(\sqrt{\nabla f\cdot\Sigma\nabla f}\right)\right)^2
\end{aligned}$$

which yields (5.7). □

We next goal is to identify the infinitesimal generator of Q_t, which is the elliptic differential operator $L=\Delta_\Sigma - x\cdot\nabla$.

Proposition 5.9. *The infinitesimal generator of the Ornstein-Uhlenbeck semi-group $(Q_t)_{t\geq 0}$ is $L=\Delta_\Sigma - x\cdot\nabla$, in the following sense. If f is continuous with at most polynomial growth, then $u(t,x)=Q_tf(x)$ belongs to $C^{1,2}((0,\infty)\times\mathbb{R}^D)$ and solves the following initial value problem of the parabolic equation:*

$$\left(L-\frac{\partial}{\partial t}\right)u(t,x)=0,\quad \lim_{t\downarrow 0}u(t,x)=f(x).$$

Therefore $\frac{\partial}{\partial t}Q_t=LQ_t$ for $t\geq 0$. This fact may be denoted as formally $Q_t=e^{tL}$.

Proof. According to Lemma 5.2 the transition probability density function

$$q_\Sigma(t,x,y)=\frac{1}{(1-e^{-2t})^{\frac{D}{2}}}\exp\left(-\frac{1}{2}\frac{y\cdot\Sigma^{-1}y+x\cdot\Sigma^{-1}x-2e^tx\cdot\Sigma^{-1}y}{e^{2t}-1}\right)$$

so that

$$\ln q_\Sigma(t,x,y)=-\frac{D}{2}\ln\left(1-e^{-2t}\right)-\frac{1}{2}\frac{y\cdot\Sigma^{-1}y+x\cdot\Sigma^{-1}x-2e^tx\cdot\Sigma^{-1}y}{e^{2t}-1}.$$

Thus

$$\begin{aligned}
\frac{\partial}{\partial t}\ln q_\Sigma(t,x,y) &= -D\frac{e^{-2t}}{1-e^{-2t}}+\frac{e^tx\cdot\Sigma^{-1}y}{e^{2t}-1}\\
&\quad+\frac{e^{2t}}{(e^{2t}-1)^2}\left(y\cdot\Sigma^{-1}y+x\cdot\Sigma^{-1}x-2e^tx\cdot\Sigma^{-1}y\right)
\end{aligned}$$

$$\frac{\partial}{\partial x_i}\ln q_\Sigma(t,x,y)=-\sigma^{ik}\frac{x_k-e^ty_k}{e^{2t}-1}$$

and

$$\begin{aligned}
\frac{\partial^2}{\partial x_j\partial x_i}q_\Sigma(t,x,y) &= \frac{\partial}{\partial x_j}\left(-\sigma^{ik}\frac{x_k-e^ty_k}{e^{2t}-1}q_\Sigma(t,x,y)\right)\\
&= \sigma^{ik}\sigma^{jl}\frac{x_k-e^ty_k}{e^{2t}-1}\frac{x_l-e^ty_l}{e^{2t}-1}q_\Sigma(t,x,y)-\sigma^{ij}\frac{1}{e^{2t}-1}q_\Sigma(t,x,y).
\end{aligned}$$

Hence

$$\begin{aligned} x\cdot\nabla q_\Sigma(t,x,y) &= -x_i\sigma^{ik}\frac{x_k-e^t y_k}{e^{2t}-1}q_\Sigma(t,x,y) \\ &= -\frac{x\cdot\Sigma^{-1}(x-e^t y)}{e^{2t}-1}q_\Sigma(t,x,y) \end{aligned}$$

and

$$\Delta_\Sigma q_\Sigma(t,x,y) = \left(\frac{(x-e^t y)\cdot\Sigma^{-1}(x-e^t y)}{(e^{2t}-1)^2} - \frac{D}{e^{2t}-1}\right)q_\Sigma(t,x,y).$$

Therefore

$$\begin{aligned} \left(\frac{\partial}{\partial t}-\Delta_\Sigma\right)q_\Sigma(t,x,y) &= \frac{x\cdot\Sigma^{-1}(x-e^t y)}{e^{2t}-1}q_\Sigma(t,x,y) \\ &= -x\cdot\nabla q_\Sigma(t,x,y). \end{aligned}$$

which implies that

$$\left(\frac{\partial}{\partial t}-\Delta_\Sigma + x\cdot\nabla\right)q_\Sigma(t,x,y)=0.$$

Suppose f is continuous with at most polynomial growth, then

$$\begin{aligned} \frac{\partial}{\partial t}u(x,t) &= \int_{\mathbb{R}^D} f(y)\frac{\partial}{\partial t}q_\Sigma(t,x,y)\gamma(\mathrm{d}y) \\ &= \int_{\mathbb{R}^D} f(y)\left(\Delta_\Sigma q_\Sigma(t,x,y)-x\cdot\nabla q_\Sigma(t,x,y)\right)\gamma(\mathrm{d}y) \\ &= \Delta_\Sigma\int_{\mathbb{R}^D} f(y)q_\Sigma(t,x,y)\gamma(\mathrm{d}y) - x\cdot\nabla\int_{\mathbb{R}^D} f(y)q_\Sigma(t,x,y)\gamma(\mathrm{d}y) \\ &= \left(\Delta_\Sigma - x\cdot\nabla\right)u(x,t) \end{aligned}$$

which completes the proof. □

Since Q_t is symmetric on $L^2(\gamma)$, so we expect its infinitesimal generator $L=\Delta_\Sigma - x\cdot\nabla$ is also symmetric on $L^2(\gamma)$, which is the context of the following lemma.

Lemma 5.10. *(Integration by parts) The differential operator $L=\Delta_\Sigma - x\cdot\nabla$ is symmetric on $L^2(\gamma)$, in the sense that*

$$\begin{aligned} \int_{\mathbb{R}^D}\psi(x)L\varphi(x)\gamma(dx) &= \int_{\mathbb{R}^D}\varphi(x)L\psi(x)\gamma(dx) \\ &= -\int_{\mathbb{R}^D}\nabla\varphi\cdot\Sigma\nabla\psi\gamma(dx) \end{aligned} \tag{5.8}$$

for any C^2-functions φ, ψ, whose first and second derivatives belong to $L^2(\gamma)$.

Proof. By using the identity

$$\frac{\partial}{\partial x^j}\ln G_\Sigma(x) = -\sum_{l=1}^{D}\sigma^{jl}x^l \tag{5.9}$$

we obtain that

$$\int_{\mathbb{R}^D} \sum_{i,j} \sigma_{ij} \frac{\partial \varphi(x)}{\partial x^i} \frac{\partial \psi(x)}{\partial x^j} \gamma(\mathrm{d}x) = - \int_{\mathbb{R}^D} \sum_{i,j} \sigma_{ij} \frac{\partial}{\partial x^j} \left(\frac{\partial \varphi}{\partial x^i} G_\Sigma \right) \psi \mathrm{d}x$$
$$= - \int_{\mathbb{R}^D} \left(\Delta_\Sigma \varphi + \sum_{i,j} \sigma_{ij} \frac{\partial \ln G_\Sigma}{\partial x^j} \frac{\partial \varphi}{\partial x^i} \right) \psi G_\Sigma \mathrm{d}x$$
$$= - \int_{\mathbb{R}^D} \left(\Delta_\Sigma \varphi - \sum_i x^i \frac{\partial \varphi}{\partial x^i} \right) \psi \gamma(\mathrm{d}x),$$

which implies (5.8) as $\Sigma = (\sigma_{ij})$ is symmetric. □

Remark 5.11. [Not examinable] *You may wonder where the Mehler formula comes from. Let us give its derivation. Recall that we wish to define a Markov semi-group Q_t whose invariant measure is the Gaussian measure $\gamma(dx)$. From the theory of diffusion processes [to be learned in SDE course, C8.1], we first identify the infinitesimal generator L of Q_t, which must satisfy the equality:*

$$\int_{\mathbb{R}^D} -\psi L \varphi d\gamma = \int_{\mathbb{R}^D} \nabla \varphi \cdot \Sigma \nabla \psi d\gamma.$$

Now integration by parts gives

$$\int_{\mathbb{R}^D} \nabla \varphi \cdot \Sigma \nabla \psi d\gamma = \int_{\mathbb{R}^D} G_\Sigma \Sigma \nabla \varphi \cdot \nabla \psi dx = - \int_{\mathbb{R}^D} \psi div \left(G_\Sigma \Sigma \nabla \varphi \right) dx$$

which gives that

$$L\varphi = \frac{1}{G_\Sigma} div \left(G_\Sigma \Sigma \nabla \varphi \right) = \Delta_\Sigma \varphi - x \cdot \nabla \varphi.$$

This is exactly the generator we have already seen. The diffusion process, whose transition probability function gives the semi-group Q_t, can be constructed as the solution to the following stochastic differential equation

$$dX_t = \sqrt{2} \Sigma^{\frac{1}{2}} dB_t - X_t dt, \quad X_0 = x$$

which can be solved explicitly

$$X_t = e^{-t} x + e^{-t} \int_0^t \sqrt{2} \Sigma^{\frac{1}{2}} e^s dB_s$$

which implies that the distribution of X_t has a normal distribution with a mean $e^{-t}x$ and covariance matrix $\left(1 - e^{-2t}\right) \Sigma$. Therefore

$$Q_t f(x) = \mathbb{E}\left[f(X_t) | X_0 = x \right]$$
$$= \int_{\mathbb{R}^D} f(y) dN \left(e^{-t}x, \left(1 - e^{-2t}\right) \Sigma \right)$$
$$= \int_{\mathbb{R}^D} f(y) \frac{\exp \left(-\frac{1}{2\left(1 - e^{-2t}\right)} (y - e^{-t}x) \cdot \Sigma^{-1} (y - e^{-t}x) \right)}{(2\pi(1 - e^{-2t}))^{\frac{D}{2}} \sqrt{\det \Sigma}} dy$$

which leads to the Mehler formula.

5.2 Entropy and the logarithmic Sobolev inequality

Recall that $\gamma(\mathrm{d}x)$ is the central Gaussian measure with Gaussian density $G_\Sigma(x)$ on $\mathscr{B}(\mathbb{R}^D)$. The entropy functional Ent (associated with the measure $\gamma(dx)$) is defined by

$$\mathrm{Ent}(h)=\int_{\mathbb{R}^D} h\ln h\mathrm{d}\gamma-\left(\int_{\mathbb{R}^D} h\mathrm{d}\gamma\right)\ln\left(\int_{\mathbb{R}^D} h\mathrm{d}\gamma\right) \tag{5.10}$$

for every non-negative $h\in L^1(\gamma)$, where $s\ln s$ is assigned to be $0=\lim_{s\downarrow 0}s\ln s$ at $s=0$. Since $s\mapsto s\ln s$ is convex on $(0,\infty)$, according to the Jensen inequality, $\mathrm{Ent}(h)\geq 0$ for every non-negative $h\in L^1(\gamma)$.

Theorem 5.12. (L. Gross) *For every $f\in W^{2,1}(\gamma)$, that is, both f and its derivative belong to $L^2(\gamma)$, it holds that*

$$Ent(f^2)\leq 2\int_{\mathbb{R}^D}(\nabla f\cdot\Sigma\nabla f)\,d\gamma. \tag{5.11}$$

Proof. By approximation property, we may assume that $f\in C^2$. Since $|\nabla|f||=|\nabla f|$ almost surely (with respect to the Lebesgue measure),

$$\int_{\mathbb{R}^D}(\nabla f\cdot\Sigma\nabla f)\,\mathrm{d}\gamma=\int_{\mathbb{R}^D}(\nabla|f|\cdot\Sigma\nabla|f|)\,\mathrm{d}\gamma.$$

Thus we may assume that f is non-negative. By replace f by $f+\varepsilon$ for any constant $\varepsilon>0$ then send $\varepsilon\downarrow 0$, we can further assume that f is bounded by a positive constant.

Let $\psi(s)=s\ln s$ and consider one variable function

$$F(t)=\int_{\mathbb{R}^D}\psi\left(Q_t(f^2)\right)\mathrm{d}\gamma=\int_{\mathbb{R}^D}Q_t(f^2)\ln Q_t(f^2)\mathrm{d}\gamma$$

for $t\in(0,\infty)$. Then $\lim_{t\downarrow 0}F(t)=\int f^2\ln f^2\mathrm{d}\gamma$,

$$\lim_{t\to\infty}F(t)=\left(\int_{\mathbb{R}^D}f^2\mathrm{d}\gamma\right)\ln\left(\int_{\mathbb{R}^D}f^2\mathrm{d}\gamma\right)$$

and therefore

$$\mathrm{Ent}(f^2)=\lim_{t\downarrow 0}F(t)-\lim_{t\to\infty}F(t)=-\int_0^\infty\frac{d}{dt}F(t)\mathrm{d}t. \tag{5.12}$$

On the other hand

$$\begin{aligned}-\frac{d}{dt}F(t)&=-\int_{\mathbb{R}^D}\psi'\left(Q_t(f^2)\right)\frac{\partial}{\partial t}Q_t(f^2)\mathrm{d}\gamma\\&=-\int_{\mathbb{R}^D}\psi'\left(Q_t(f^2)\right)LQ_t(f^2)\mathrm{d}\gamma\\&=\int_{\mathbb{R}^D}\nabla\psi'\left(Q_t(f^2)\right)\cdot\Sigma\nabla Q_t(f^2)\mathrm{d}\gamma\\&=\int_{\mathbb{R}^D}\psi''\left(Q_t(f^2)\right)\nabla Q_t(f^2)\cdot\Sigma\nabla Q_t(f^2)\mathrm{d}\gamma\end{aligned}$$

where the third equality follows from Lemma 5.10. Since $\psi'(s) = \ln s + 1$ and $\psi''(s) = \frac{1}{s}$, we deduce that

$$-\frac{d}{dt}F(t) = \int_{\mathbb{R}^D} \frac{1}{Q_t(f^2)} \nabla Q_t(f^2) \cdot \Sigma \nabla Q_t(f^2) \mathrm{d}\gamma \quad \text{for } t > 0. \tag{5.13}$$

By the domination inequality

$$\begin{aligned}
\sqrt{\nabla Q_t(f^2) \cdot \Sigma \nabla Q_t(f^2)} &\leq e^{-t} Q_t\left(\sqrt{\nabla f^2 \cdot \Sigma \nabla f^2}\right) \\
&= 2e^{-t} Q_t\left(|f|\sqrt{\nabla f \cdot \Sigma \nabla f}\right) \\
&\leq 2e^{-t}\sqrt{Q_t(f^2)}\sqrt{Q_t\left(\nabla f \cdot \Sigma \nabla f\right)}
\end{aligned}$$

where the last inequality follows from Cauchy-Schwartz inequality. Rearrange the previous inequality we deduce that

$$\frac{1}{Q_t(f^2)} \nabla Q_t(f^2) \cdot \Sigma \nabla Q_t(f^2) \leq 4e^{-2t} Q_t\left(\nabla f \cdot \Sigma \nabla f\right).$$

Together with (5.13)

$$-\frac{d}{dt}F(t) \leq 4e^{-2t} \int_{\mathbb{R}^D} Q_t\left(\nabla f \cdot \Sigma \nabla f\right) \mathrm{d}\gamma = 4e^{-2t} \int_{\mathbb{R}^D} \nabla f \cdot \Sigma \nabla f \mathrm{d}\gamma$$

and, by integrating the inequality over $(0, \infty)$ to obtain that

$$\mathrm{Ent}(f^2) \leq \int_0^\infty 4e^{-2t} \mathrm{d}t \int_{\mathbb{R}^D} \nabla f \cdot \Sigma \nabla f \mathrm{d}\gamma = 2\int_{\mathbb{R}^D} \nabla f \cdot \Sigma \nabla f \mathrm{d}\gamma$$

and therefore the proof is complete. □

Remark 5.13. *If $f \in C^2$, then the logarithmic Sobolev inequality may be written as*

$$Ent(f^2) \leq -2\int_{\mathbb{R}^D} fLf d\gamma.$$

Exercise 1. In this exercise we are going to prove the *hyper-contractivity* of the Ornstein-Uhlenbeck semi-group. Let $\gamma(\mathrm{d}x) = G_\Sigma(x)\mathrm{d}x$, and let $q : (0, \infty) \to [1, \infty)$ be differentiable, to be chosen later. Let f be a positive, bounded and continuous function on $\mathbb{R}^D$. Consider two functions on $(0, \infty)$: $F(t) = \int (Q_t f)^{q(t)} \mathrm{d}\gamma$ and $G(t) = \|Q_t f\|_{L^{q(t)}(\gamma)}$. Then $G(t) = F(t)^{\frac{1}{q(t)}}$ and $\ln G(t) = \frac{1}{q(t)} \ln F(t)$. Therefore

$$G'(t) = G(t)\frac{1}{q(t)}\left(-\frac{q'(t)}{q(t)}\ln F(t) + \frac{F'(t)}{F(t)}\right)$$

and

$$\begin{aligned}
F'(t) &= \int_{\mathbb{R}^D} \frac{d}{dt}\left(Q_t f\right)^{q(t)} \mathrm{d}\gamma \\
&= q'(t)\int_{\mathbb{R}^D} \left(Q_t f\right)^{q(t)} \ln Q_t f \mathrm{d}\gamma + q(t)\int_{\mathbb{R}^D} \left(Q_t f\right)^{q(t)-1} \frac{d}{dt} Q_t f \mathrm{d}\gamma \\
&= \frac{q'(t)}{q(t)}\int_{\mathbb{R}^D} \left(Q_t f\right)^{q(t)} \ln\left(Q_t f\right)^{q(t)} \mathrm{d}\gamma + q(t)\int_{\mathbb{R}^D} \left(Q_t f\right)^{q(t)-1} LQ_t f \mathrm{d}\gamma \\
&= \frac{q'(t)}{q(t)}\left[\mathrm{Ent}\left(\left(Q_t f\right)^{q(t)}\right) + F(t)\ln F(t)\right] + q(t)\int_{\mathbb{R}^D} \left(Q_t f\right)^{q(t)-1} LQ_t f \mathrm{d}\gamma.
\end{aligned}$$

Let us now choose function q which increasing, i.e. $q'(t) \geq 0$. Applying the logarithmic Sobolev inequality

$$\mathrm{Ent}\left((Q_t f)^{q(t)}\right) \leq 2\int_{\mathbb{R}^D} \nabla (Q_t f)^{\frac{q(t)}{2}} \cdot \Sigma \nabla (Q_t f)^{\frac{q(t)}{2}} \,\mathrm{d}\gamma$$

in the previous equality, one deduces that

$$\begin{aligned}
F'(t) &\leq \frac{q'(t)}{q(t)} F(t) \ln F(t) + 2\frac{q'(t)}{q(t)} \int_{\mathbb{R}^D} \nabla (Q_t f)^{\frac{q(t)}{2}} \cdot \Sigma \nabla (Q_t f)^{\frac{q(t)}{2}} \,\mathrm{d}\gamma \\
&\quad - q(t) \int_{\mathbb{R}^D} \nabla (Q_t f)^{q(t)-1} \cdot \Sigma \nabla Q_t f \mathrm{d}\gamma \\
&= \frac{q'(t)}{q(t)} F(t) \ln F(t) + q(t)\left(\frac{1}{2} q'(t) - (q(t)-1)\right) \int_{\mathbb{R}^D} (Q_t f)^{q(t)-2} \nabla Q f_t \cdot \Sigma \nabla Q_t f \mathrm{d}\gamma.
\end{aligned}$$

The best choice of q for the previous inequality is given as solutions to

$$\frac{1}{2} q'(t) - (q(t)-1)) = 0. \tag{5.14}$$

Suppose $q(t) \geq 1$ is a solution of (5.14). Then

$$F'(t) \leq \frac{q'(t)}{q(t)} F(t) \ln F(t)$$

and

$$\begin{aligned}
G'(t) &= G(t) \frac{1}{q(t)} \left(-\frac{q'(t)}{q(t)} \ln F(t) + \frac{F'(t)}{F(t)}\right) \\
&\leq G(t) \frac{1}{q(t)} \left(-\frac{q'(t)}{q(t)} \ln F(t) + \frac{q'(t)}{q(t)} \ln F(t)\right) \\
&= 0.
\end{aligned}$$

Therefore $t \to G(t)$ is decreasing, so that $G(t) \leq G(0)$. The solution to (5.14) with $q(0) = p$ for a given $p \geq 1$ is $q(t) = 1 + (p-1)e^{2t}$. Therefore

$$\|Q_t f\|_{L^{q(t)}(\gamma)} \leq \|f\|_{L^p(\gamma)} \qquad \text{for every } t \geq 0 \text{ and } f \in L^p(\gamma)$$

where $q(t) = 1 + (p-1)e^{2t}$. This is called the hypercontractivity of the Ornstein-Uhlenbeck semi-group $(Q_t)_{t \geq 0}$.

5.3 Poincaré inequality

The variance of f (with respect to the Gaussian measure $\gamma(\mathrm{d}x) = G_\Sigma(x)\mathrm{d}x$)

$$\begin{aligned}
\mathrm{var}(f) &= \int_{\mathbb{R}^D} \left(f - \int_{\mathbb{R}^D} f \mathrm{d}\gamma\right)^2 \mathrm{d}\gamma \\
&= \int_{\mathbb{R}^D} f^2 \mathrm{d}\gamma - \left(\int_{\mathbb{R}^D} f \mathrm{d}\gamma\right)^2.
\end{aligned}$$

The following inequality is called the Poincaré inequality.

Theorem 5.14. *Let $\gamma(dx) = G_\Sigma(x)dx$ be the Gaussian measure. Then*

$$\int_{\mathbb{R}^D} \left(f - \int_{\mathbb{R}^D} f d\gamma \right)^2 d\gamma \le \int_{\mathbb{R}^D} (\nabla f \cdot \Sigma \nabla f)\, d\gamma$$

for any C^1-function f such that $|\nabla f|^2$ is γ-integrable.

Proof. Let $F(t) = \int_{\mathbb{R}^D} (Q_t f)^2 \mathrm{d}\gamma$. Then $\lim_{t\to 0} F(t) = \int_{\mathbb{R}^D} f^2 \mathrm{d}\gamma$ and

$$\lim_{t\to\infty} F(t) = \int_{\mathbb{R}^D} \left(\int_{\mathbb{R}^D} f \mathrm{d}\gamma \right)^2 \mathrm{d}\gamma = \left(\int_{\mathbb{R}^D} f \mathrm{d}\gamma \right)^2 .$$

Therefore

$$\mathrm{var}(f) = -\int_0^\infty \frac{d}{dt} F(t) \mathrm{d}t.$$

Next calculate the derivative

$$\begin{aligned} -\frac{d}{dt} F(t) &= -\int_{\mathbb{R}^D} \frac{d}{dt} (Q_t f)^2 \mathrm{d}\gamma \\ &= -2 \int_{\mathbb{R}^D} Q_t f \frac{d}{dt} Q_t f \mathrm{d}\gamma \\ &= -2 \int_{\mathbb{R}^D} Q_t f L Q_t f \mathrm{d}\gamma \\ &= 2 \int_{\mathbb{R}^D} \nabla Q_t f \cdot \Sigma \nabla Q_t f \mathrm{d}\gamma. \end{aligned}$$

Using the weak domination inequality we thus deduce that

$$-\frac{d}{dt} F(t) \le 2e^{-2t} \int_{\mathbb{R}^D} Q_t (\nabla f \cdot \Sigma \nabla f) \mathrm{d}\gamma = 2e^{-2t} \int_{\mathbb{R}^D} \nabla f \cdot \Sigma \nabla f \mathrm{d}\gamma.$$

Integrating the previous inequality over $(0,\infty)$ to get that

$$\mathrm{var}(f) \le \int_0^\infty 2e^{-2t} \mathrm{d}t \int_{\mathbb{R}^D} (\nabla f \cdot \Sigma \nabla f) \mathrm{d}\gamma = \int_{\mathbb{R}^D} (\nabla f \cdot \Sigma \nabla f) \mathrm{d}\gamma.$$

Thus we have completed the proof. □

5.4 The concentration inequality

In this section we prove the major concentration inequality for Gaussian measure $\gamma(\mathrm{d}x) = G_\Sigma(x)\mathrm{d}x$.

If g is a function on $\mathbb{R}^D$, we shall use $\|g\|_\infty$ to denote the supremum norm of g, that is, $\|g\|_\infty = \sup_{x\in\mathbb{R}^D} |g(x)|$.

Theorem 5.15. *Let $\gamma(dx) = G_\Sigma(x)dx$ be a centered Gaussian measure on $(\mathbb{R}^D, \mathscr{B}(\mathbb{R}^D))$, and let f be a C^1-function with bounded derivatives. Then*

$$\int_{\mathbb{R}^D} \exp\left[\lambda \left(f - \int_{\mathbb{R}^D} f d\gamma \right) \right] \le \exp\left(\frac{\lambda^2}{2} \|\nabla f \cdot \Sigma \nabla f\|_\infty \right) \tag{5.15}$$

for every $\lambda \in \mathbb{R}$, where

$$\|\nabla f \cdot \Sigma \nabla f\|_\infty = \sup_{\mathbb{R}^D} (\nabla f \cdot \Sigma \nabla f)$$

is the supremum norm of $\nabla f \cdot \Sigma \nabla f$ over $\mathbb{R}^D$.

Proof. By considering $f(x) - \int_{\mathbb{R}^D} f \mathrm{d}\gamma$ instead, without losing generality we may assume that $\int_{\mathbb{R}^D} f \mathrm{d}\gamma = 0$. Let $\psi(s) = e^{\lambda s}$. Then $\psi' = \lambda \psi$ and $\psi'' = \lambda^2 \psi$. Consider

$$F(t) = \int_{\mathbb{R}^D} \psi(Q_t f) \mathrm{d}\gamma = \int_{\mathbb{R}^D} \exp(\lambda Q_t f) \mathrm{d}\gamma \quad \text{for } t \geq 0$$

Then

$$\lim_{t\to\infty} F(t) = \int_{\mathbb{R}^D} \exp\left(\lambda \int f \mathrm{d}\gamma\right) \mathrm{d}\gamma = 1$$

and therefore

$$F(t) - 1 = -\int_t^\infty \frac{d}{dt} F(t) \mathrm{d}t \quad \text{for } t \geq 0.$$

As before we differentiate under integration, and use the equation that $\frac{d}{dt} Q_t f = L Q_t f$, to obtain that

$$\begin{aligned}
-\frac{d}{dt} F(t) &= -\int_{\mathbb{R}^D} \frac{d}{dt} \psi(Q_t f) \mathrm{d}\gamma = -\int_{\mathbb{R}^D} \psi'(Q_t f) \frac{d}{dt} Q_t f \mathrm{d}\gamma \\
&= -\int_{\mathbb{R}^D} \psi'(Q_t f) L Q_t f \mathrm{d}\gamma.
\end{aligned}$$

Next perform integration in the last integral, to get that

$$\begin{aligned}
-\frac{d}{dt} F(t) &= \int_{\mathbb{R}^D} \nabla \psi'(Q_t f) \cdot \Sigma \nabla Q_t f \mathrm{d}\gamma \\
&= \int_{\mathbb{R}^D} \psi''(Q_t f) \nabla Q_t f \cdot \Sigma \nabla Q_t f \mathrm{d}\gamma \\
&= \lambda^2 \int_{\mathbb{R}^D} \psi(Q_t f) \nabla Q_t f \cdot \Sigma \nabla Q_t f \mathrm{d}\gamma
\end{aligned}$$

Since ψ is positive, we may use the weak domination inequality

$$\nabla Q_t f \cdot \Sigma Q_t f \leq e^{-2t} Q_t (\nabla f \cdot \Sigma \nabla f) \leq e^{-2t} \|\nabla f \cdot \Sigma \nabla f\|_\infty$$

we thus conclude that

$$\begin{aligned}
-\frac{d}{dt} F(t) &\leq \lambda^2 e^{-2t} \|\nabla f \cdot \Sigma \nabla f\|_\infty \int_{\mathbb{R}^D} \psi(Q_t f) \mathrm{d}\gamma \\
&= \lambda^2 e^{-2t} \|\nabla f \cdot \Sigma \nabla f\|_\infty F(t),
\end{aligned}$$

i.e.

$$-\frac{1}{F(t)} \frac{d}{dt} F(t) \leq \lambda^2 e^{-2t} \|\nabla f \cdot \Sigma \nabla f\|_\infty$$

for $t > 0$. Integrating the inequality over $[t, \infty)$ to obtain that

$$\begin{aligned}
\ln F(t) - \ln F(\infty) &= -\int_t^\infty \frac{1}{F(t)} \frac{d}{dt} F(t) dt \\
&\leq \lambda^2 \int_t^\infty e^{-2t} dt \|\nabla f \cdot \Sigma \nabla f\|_\infty = \frac{\lambda^2}{2} \|\nabla f \cdot \Sigma \nabla f\|_\infty
\end{aligned}$$

Letting $t \downarrow 0$ we conclude that

$$\int_{\mathbb{R}^D} \exp\left(\lambda\left(f - \int_{\mathbb{R}^D} f \mathrm{d}\gamma\right)\right) \mathrm{d}\gamma \leq \exp\left(\frac{\lambda^2}{2} \|\nabla f \cdot \Sigma f\|_\infty\right).$$

The second inequality follows from Markov inequality. □

We next prove the well-known Borell's inequality for family of Gaussian random variables.

Corollary 5.16. *Let $Y = (Y_1, \cdots, Y_D)$ be an $\mathbb{R}^D$-valued random variable with the standard normal distribution $N(0, I)$ (where I is the identity matrix), and let $f : \mathbb{R}^D \mapsto \mathbb{R}$ be Lipschitz continuous.*

(a) We have

$$\mathbb{E}\left(e^{\lambda(f(Y) - \mathbb{E}f(Y))}\right) \leq \exp\left(\frac{\lambda^2}{2} \|f\|_{Lip}^2\right) \tag{5.16}$$

for any $\lambda \in \mathbb{R}$.

(b) The following Gaussian estimate holds:

$$\mathbb{P}(|f(Y) - \mathbb{E}f(Y)| > r) \leq 2\exp\left(-\frac{r^2}{2\|f\|_{Lip}^2}\right) \tag{5.17}$$

for every $r > 0$.

Proof. Let f_ε be constructed in Lemma 7.12 for every $\varepsilon > 0$. By Theorem 5.15,

$$\mathbb{E}\left(\exp\left[\lambda\left(f_\varepsilon(Y) - \mathbb{E}f_\varepsilon(Y)\right)\right]\right) \leq \exp\left(\frac{\lambda^2}{2} \|\nabla f_\varepsilon\|_\infty^2\right) \leq \exp\left(\frac{\lambda^2}{2} (\|f\|_{\text{Lip}} + \varepsilon)^2\right)$$

for every $\varepsilon > 0$. Letting $\varepsilon \downarrow 0$ we obtain (5.16). The Gaussian estimate (5.17) follows from (5.16) as we have seen in Section 1. □

Theorem 5.17. (Borell's inequality). *Let $X = (X_1, \cdots, X_D)$ be a random variable with central Gaussian distribution with co-variance matrix $\Sigma = (\sigma_{ij})$. Then*

$$\mathbb{P}\left[\left|\sup_{i=1,\ldots,D} X^i - \mathbb{E} \sup_{i=1,\ldots,D} X^i\right| > r\right] \leq 2\exp\left(-\frac{r^2}{2\sup_i \sigma_{ii}}\right) \tag{5.18}$$

for every $r > 0$.

Proof. Let $Y = (Y_1, \cdots, Y_D)$ be a random variable in $\mathbb{R}^D$ with the standard normal distribution $N(0, I)$, as in the previous corollary. Then $Z = \Sigma^{\frac{1}{2}} Y$ has the same distribution as that of X, where $\Sigma^{\frac{1}{2}} = (\rho_{ij})$ is a positive square root of Σ. Let $f(x) = \max_{i=1,\cdots,D}\left(\sum_{k=1}^D \rho_{ik} x_k\right)$. For given x, y, there are i and j such that

$$f(x) = \sum_{k=1}^D \rho_{ik} x_k \quad \text{and } f(y) = \sum_{k=1}^D \rho_{jk} y_k$$

(where i, j depend on x, y of course), so that

$$f(x) - f(y) = \sum_{k=1}^D \rho_{ik} x_k - \sum_{k=1}^D \rho_{jk} y_k \leq \sum_{k=1}^D \rho_{ik} x_k - \sum_{k=1}^D \rho_{ik} y_k$$

and similarly

$$f(y)-f(x)\le\sum_{k=1}^{D}\rho_{jk}y_k-\sum_{k=1}^{D}\rho_{jk}x_k,$$

which implies that

$$\begin{aligned}|f(x)-f(y)| &\le \max_i\left|\sum_{k=1}^{D}\rho_{ik}(y_k-x_k)\right| \\ &\le \max_{i=1,\cdots,D}\sqrt{\sum_{k=1}\rho_{ik}^2}|x-y| \\ &= \max_{i=1,\cdots,D}\sqrt{\sigma_{ii}}|x-y|.\end{aligned}$$

Thus f is Lipschitz continuous with Lipschitz constant less than $\max_i\sqrt{\sigma_{ii}}$. Therefore, according to (5.17)

$$\begin{aligned}\mathbb{P}\left[\left|\sup_{i=1,\ldots,D}X^i-\mathbb{E}\sup_{i=1,\ldots,D}X^i\right|>r\right] &= \mathbb{P}\left(|f(Z)-\mathbb{E}f(Z)|>r\right) \\ &\le 2\exp\left(-\frac{r^2}{2\sup_i\sigma_{ii}}\right).\end{aligned}$$

□

Remark 5.18. *(a) As long as $\mathbb{E}\sup_i X_i$ is finite (in this case the family of centered Gaussian random variables (X_i) is called bounded), then the Borell's inequality is still valid in exactly the same form, by letting $D\to\infty$. That is, if $(X_t)_{t\in\Lambda}$ is a family of centered Gaussian random variables, where Λ is any countable set, such that $\mathbb{E}\sup_{t\in\Lambda}X_t<\infty$, then*

$$\mathbb{P}\left[\left|\sup_{t\in\Lambda}X_t-\mathbb{E}\sup_{t\in\Lambda}X_t\right|>r\right]\le 2\exp\left(-\frac{r^2}{2\sup_{t\in\Lambda}\sigma_{tt}}\right) \tag{5.19}$$

for every $r>0$, where $\sigma_{tt}=var(X_t)$.

(b) It remains to control the quantity $\mathbb{E}\sup_{t\in\Lambda}X_t$. This can be done by using the technique of metric entropy, a topic we left for your own study. The reader may refer to the small book by R. J. Adler [1].

5.5 Estimates of exponential type

In this section we introduce another idea for deriving typical Gaussian type exponential decay estimates, which is in a matter of transport distributions, an idea which is quite useful. It yields interesting results, though it does not lead to better results as we have developed so far.

Lemma 5.19. *Let $X=(X^i)_{i=1,\cdots,D}$ and $Y=(Y^i)_{i=1,\cdots,D}$ be two independent random variables with the same distribution $\gamma(dx)=G_\Sigma(x)dx$, where Σ is symmetric, positive definite. Let $X(t)=X\sin t+Y\cos t$ and $\frac{d}{dt}X(t)=X\cos t-Y\sin t$ for $t\in\mathbb{R}$. Then for every t, $X(t)$ and $\frac{d}{dt}X(t)$ have independent, and have the same distribution $\gamma(dx)$.*

Proof. For each t we have

$$\begin{aligned}\mathbb{E}\left[X(t)^iX(t)^j\right] &= \mathbb{E}\left[\left(X^i\sin t + Y^i\cos t\right)\left(X^j\sin t + Y^j\cos t\right)\right]\\ &= \sin^2 t\mathbb{E}\left[X^iX^j\right] + \cos^2 t\mathbb{E}\left[Y^iY^j\right]\\ &= \sigma_{ij}\end{aligned}$$

hence $X(t)$ has distribution γ as well. Let $Z(t) = \frac{d}{dt}X(t)$. Then

$$\begin{aligned}\mathbb{E}\left[X(t)^iZ(t)^j\right] &= \mathbb{E}\left[\left(X^i\sin t + Y^i\cos t\right)\left(X^j\cos t - Y^j\sin t\right)\right]\\ &= \sin t\cos t\left(\mathbb{E}\left[X^iX^j\right] - \mathbb{E}\left[Y^iY^j\right]\right)\\ &\quad + \cos^2 t\mathbb{E}\left[Y^iX^j\right] - \sin^2 t\mathbb{E}\left[Y^jX^i\right]\\ &= 0\end{aligned}$$

which implies X and Z are independent. □

Let begin with the following general Gaussian estimate.

Theorem 5.20. *Let $f : \mathbb{R}^D \mapsto \mathbb{R}^n$ be a C^1-function, and $\Psi : \mathbb{R}^n \mapsto \mathbb{R}$ be a convex function. Then*

$$\int_{\mathbb{R}^D}\int_{\mathbb{R}^D}\Psi(f(x)-f(y))\gamma(dy)\gamma(dx) \le \int_{\mathbb{R}^D}\int_{\mathbb{R}^D}\Psi\left(\frac{\pi}{2}\nabla f(x)\cdot y\right)\gamma(dx)\gamma(dy) \tag{5.20}$$

and

$$\int_{\mathbb{R}^D}\Psi\left(f(x)-\int_{\mathbb{R}^D} f d\gamma\right)\gamma(dx) \le \int_{\mathbb{R}^D}\int_{\mathbb{R}^D}\Psi\left(\frac{\pi}{2}\nabla f(x)\cdot y\right)\gamma(dx)\gamma(dy) \tag{5.21}$$

where $f = (f_1,\cdots,f_n)$ and $\nabla f(x)\cdot y = (\nabla f_1(x)\cdot y,\cdots,\nabla f_n(x)\cdot y)$ for any $x,y \in \mathbb{R}^D$.

Proof. By considering $f^i - \int_{\mathbb{R}^D} f^i \mathrm{d}\gamma$ instead, without losing generality, we assume that $\int_{\mathbb{R}^D} f^i \mathrm{d}\gamma = 0$ for $i = 1,\dots,n$. Let X and Y be independent random variables with the same distribution γ, and $X(t) = X\sin t + Y\cos t$. Then

$$\begin{aligned}f(X) - f(Y) &= \int_0^{\frac{\pi}{2}}\frac{d}{dt}f(X(t))\mathrm{d}t\\ &= \int_0^{\frac{\pi}{2}}\nabla f(X(t))\cdot\frac{d}{dt}X(t)\mathrm{d}t\end{aligned}$$

and therefore

$$\Psi\left(f(X)-f(Y)\right) = \Psi\left(\int_0^{\frac{\pi}{2}}\nabla f(X(t))\cdot\frac{d}{dt}X(t)\mathrm{d}t\right)$$

Since Ψ is convex, applying Jensen's inequality (with respect to the $\frac{2}{\pi}\mathrm{d}t$ on $[0,\frac{\pi}{2}]$), to obtain

$$\Psi\left(f(X)-f(Y)\right) \le \frac{2}{\pi}\int_0^{\frac{\pi}{2}}\Psi\left(\frac{\pi}{2}\nabla f(X(t))\cdot\frac{d}{dt}X(t)\right)\mathrm{d}t.$$

Taking expectation both sides of the inequality to deduce that

$$\mathbb{E}\left[\Psi\left(f(X)-f(Y)\right)\right] \le \frac{2}{\pi}\int_0^{\frac{\pi}{2}}\mathbb{E}\left[\Psi\left(\frac{\pi}{2}\nabla f(X(t))\cdot\frac{d}{dt}X(t)\right)\right]\mathrm{d}t. \tag{5.22}$$

By Lemma 5.19, both (X,Y), and $(X(t),\frac{d}{dt}X(t))$ (for every t) has the same distribution $\gamma\otimes\gamma$, so that

$$\mathbb{E}\left[\Psi\left(f(X)-f(Y)\right)\right]=\int_{\mathbb{R}^D}\int_{\mathbb{R}^D}\Psi(f(x)-f(y))\gamma(\mathrm{d}y)\gamma(\mathrm{d}x)$$

and

$$\mathbb{E}\left[\Psi\left(\frac{\pi}{2}\nabla f(X(t))\cdot\frac{d}{dt}X(t)\right)\right]=\int_{\mathbb{R}^D}\int_{\mathbb{R}^D}\Psi\left(\frac{\pi}{2}\nabla f(x)\cdot y\right)\gamma(\mathrm{d}x)\gamma(\mathrm{d}y)$$

for every t, so the first inequality follows.

To prove the second inequality, we use Jensen's inequality again, to deduce that

$$\int_{\mathbb{R}^D}\Psi\left(f(x)-f(y)\right)\gamma(\mathrm{d}y)\geq\Psi\left(f(x)-\int_{\mathbb{R}^D}f\mathrm{d}\gamma\right)$$

for every x. Integrating out the variable x, we then deduce that

$$\int_{\mathbb{R}^D}\int_{\mathbb{R}^D}\Psi\left(f(x)-f(y)\right)\gamma(\mathrm{d}y)\gamma(\mathrm{d}x)\geq\int_{\mathbb{R}^D}\Psi\left(f(x)-\int_{\mathbb{R}^D}f\mathrm{d}\gamma\right)\gamma(\mathrm{d}x).$$

Therefore the second inequality follows from the first inequality. □

Corollary 5.21. *Let $\gamma(dx)=G_\Sigma(x)dx$. Suppose $f:\mathbb{R}^D\mapsto\mathbb{R}$ is a C^1-function, and $p\geq 1$. Then*

$$\int_{\mathbb{R}^D}\left|f-\int_{\mathbb{R}^D}fd\gamma\right|^p d\gamma\leq C_p\int_{\mathbb{R}^D}|\nabla f|^p d\gamma \tag{5.23}$$

where

$$C_p=\left(\frac{\pi}{2}\right)^p\int_{\mathbb{R}^D}|y|^p\gamma(dy),\quad |y|=\sqrt{\sum_{i=1}^D(y_i)^2}.$$

Proof. We apply Theorem 5.20 to convex function $\Psi(x)=|x|^p$. Then

$$\begin{aligned}\int_{\mathbb{R}^D}\int_{\mathbb{R}^D}\Psi\left(\frac{\pi}{2}\nabla f(x)\cdot y\right)\gamma(\mathrm{d}x)\gamma(\mathrm{d}y)&=\left(\frac{\pi}{2}\right)^p\int_{\mathbb{R}^D}\int_{\mathbb{R}^D}|\nabla f(x)\cdot y|^p\gamma(\mathrm{d}x)\gamma(\mathrm{d}y)\\&\leq C_p\int_{\mathbb{R}^D}|\nabla f|^p\mathrm{d}\gamma\end{aligned}$$

which yields the conclusion. □

If $p=2$, then estimate (5.23) becomes a variation of the Poincaré inequality:

$$\int_{\mathbb{R}^D}\left|f-\int_{\mathbb{R}^D}f\mathrm{d}\gamma\right|^2\mathrm{d}\gamma\leq C_2\int_{\mathbb{R}^D}|\nabla f|^2\mathrm{d}\gamma$$

where

$$C_2=\left(\frac{\pi}{2}\right)^2\sum_{i=1}^D\int_{\mathbb{R}^D}(y^i)^2\gamma(\mathrm{d}y)=\left(\frac{\pi}{2}\right)^2\mathrm{tr}\Sigma$$

while the variance $\mathrm{var}(f)$ is dominated by the quadratic form $\int\nabla f\cdot\nabla f\mathrm{d}\gamma$, instead of $\int\nabla f\cdot\Sigma\nabla f\mathrm{d}\gamma$.

Corollary 5.22. *Suppose f is Lipschitz continuous from $\mathbb{R}^D \mapsto \mathbb{R}$ with Lipschitz constant C. Then*

$$\int_{\mathbb{R}^D} \exp\left(\alpha \left| f - \int_{\mathbb{R}^D} f d\gamma \right|^2\right) d\gamma \leq \int_{\mathbb{R}^D} \exp\left(\frac{\pi}{2}\alpha C\lambda |y|^2\right) G_I(y) dy$$

where λ is the largest eigenvalue of Σ. The right hand-side is finite as long as $\alpha < \frac{2}{\pi^2 C^2 \lambda^2}$.

Proof. Let $\Psi(t) = \exp(\alpha t^2)$ where $\alpha \geq 0$ is a constant. Then

$$\Psi''(t) = 2\alpha \exp(\alpha t^2) + (2\alpha t)^2 \exp(\alpha t^2) \geq 0$$

so Ψ is convex. We apply (5.21) with Ψ. Then

$$\begin{aligned}\Psi\left(\frac{\pi}{2}\alpha f'(x) y\right) &= \exp\left(\frac{\pi}{2}\alpha \left(\sum_{i=1}^{D} \frac{\partial f(x)}{\partial x^i} y^i\right)^2\right) \\ &\leq \exp\left(\frac{\pi}{2}\alpha |\nabla f|^2 |y|^2\right)\end{aligned}$$

and therefore, according to (5.21),

$$\int_{\mathbb{R}^D} \exp\left(\alpha \left| f - \int_{\mathbb{R}^D} f \mathrm{d}\gamma \right|\right) \mathrm{d}\gamma \leq \int_{\mathbb{R}^D} \exp\left(\frac{\pi}{2}\alpha C |y|^2\right) \gamma(\mathrm{d}y).$$

For the integral on the right-hand side we make a change of variable $\Sigma^{\frac{1}{2}} z = y$, so that

$$\begin{aligned}\int_{\mathbb{R}^D} \exp\left(\frac{\pi}{2}\alpha C |y|^2\right) \gamma(\mathrm{d}y) &= \int_{\mathbb{R}^D} \exp\left(\frac{\pi}{2}\alpha C y \cdot \Sigma y\right) G_I(y) \mathrm{d}y \\ &\leq \int_{\mathbb{R}^D} \exp\left(\frac{\pi}{2}\alpha \lambda_D C |y|^2\right) G_I(y) \mathrm{d}y\end{aligned}$$

where now $G_I(y)$ is the standard Gaussian density on $\mathbb{R}^D$ and λ_D is the largest eigenvalue. By a standard computation we have

$$\int_{\mathbb{R}^D} \exp\left(\frac{\pi}{2}\alpha \lambda_D C |y|^2\right) G_I(y) \mathrm{d}y \leq \frac{1}{\sqrt{1 - \frac{\alpha}{2}\pi^2 C^2 \lambda_D^2}}$$

which completes the proof. □

Corollary 5.23. *If f is C^1, then*

$$\int_{\mathbb{R}^D} \exp\left(f(x) - \int_{\mathbb{R}^D} f d\gamma\right) \gamma(dx) \leq \int_{\mathbb{R}^D} \exp\left(\frac{\pi^2}{8} \nabla f \cdot \Sigma \nabla f\right) d\gamma \tag{5.24}$$

Proof. Let us apply (5.21) with $\Psi(t) = e^t$ which is convex, to obtain that

$$\begin{aligned}\int_{\mathbb{R}^D} \exp\left(f(x) - \int_{\mathbb{R}^D} f \mathrm{d}\gamma\right) \gamma(\mathrm{d}x) &\leq \int_{\mathbb{R}^D}\int_{\mathbb{R}^D} \exp \frac{\pi}{2}\left(\sum_{i=1}^{D} \frac{\partial f(x)}{\partial x^i} y^i\right) \gamma(\mathrm{d}x)\gamma(\mathrm{d}y) \\ &\leq \int_{\mathbb{R}^D}\int_{\mathbb{R}^D} \exp\left(\frac{\pi}{2}\sum_{i=1}^{D} \frac{\partial f(x)}{\partial x^i} y^i\right) \gamma(\mathrm{d}x)\gamma(\mathrm{d}y).\end{aligned}$$

For every x (but fixed), $Y=(Y^i)$ has a distribution γ. Then $Z=\frac{\pi}{2}\sum_{i=1}^{D}\frac{\partial f(x)}{\partial x^i}Y^i$ is Gaussian random variable whose variance is

$$\mathrm{var}(Z)=\frac{\pi^2}{4}\nabla f\cdot\Sigma\nabla f$$

and therefore

$$\int_{\mathbb{R}^D}\exp\left(\frac{\pi}{2}\sum_{i=1}^{D}\frac{\partial f(x)}{\partial x^i}y^i\right)\gamma(\mathrm{d}y)=\exp\left(\frac{\pi^2}{8}\nabla f\cdot\Sigma\nabla f\right).$$

Hence (5.24) follows immediately. □

5.6 Gaussian isoperimetric inequality

In this section we derive Lévy-Gromov's isoperimetric function for centered Gaussian measure $\gamma(\mathrm{d}x)=G_\Sigma(x)\mathrm{d}x$, following the approach put forward by D. Bakry and M. Ledoux [3] via the Ornstein-Uhlenbeck semigroup $(Q_t)_{t\geq 0}$, whose invariant measure is $\gamma(\mathrm{d}x)$. B-L [3] aims to give a general version of Lévy-Gromov's isoperimetric inequality (for metric-measure spaces with positive curvature) by using Bakry-Emery's Γ_2 formulation (Ricci curvature) and the idea of quantization. While the most useful case remains the isoperimetric inequality (independent of dimensions) for Gaussian measures, which is going to be presented in this part.

Let us now introduce the *isoperimetric function* for Gaussian measure. Suppose ξ is a real random variable with a standard normal distribution $N(0,1)$, then

$$\Phi(r)=\mathbb{P}[\xi\leq r]=\int_{-\infty}^{r}\frac{1}{\sqrt{2\pi}}\exp\left(-\frac{x^2}{2}\right)\mathrm{d}x \tag{5.25}$$

which strictly increasing, whose inverse $\Phi^{-1}:(0,1)\mapsto(-\infty,\infty)$ is also increasing. The isopermetric function is defined to be $\mathscr{U}=\Phi'\circ\Phi^{-1}$ on $(0,1)$, where the derivative Φ' is nothing but just the 1-D standard Gaussian density, i.e. $\Phi'(x)=\frac{1}{\sqrt{2\pi}}\exp\left(-\frac{x^2}{2}\right)$. Naturally we extend the definition of $\mathscr{U}$ to $[0,1]$ by setting

$$\mathscr{U}(0)=0 \text{ and } \mathscr{U}(1)=0$$

so that $\mathscr{U}$ is differentiable (of any degree) on $(0,1)$ and is continuous on $[0,1]$. By chain rule and use the fact that $\Phi''(x)=-x\Phi'$, we have

$$\mathscr{U}'=\Phi''\circ\Phi^{-1}\frac{1}{\Phi'\circ\Phi^{-1}}=-\Phi^{-1} \tag{5.26}$$

and

$$\mathscr{U}''=-\frac{1}{\Phi'\circ\Phi^{-1}}=-\frac{1}{\mathscr{U}}. \tag{5.27}$$

In particular $\mathscr{U}''<0$ on $(0,1)$. Therefore $x\to\mathscr{U}(x)$ is (strictly) concave on $(0,1)$, symmetric again the vertical line $x=\frac{1}{2}$ at which it attains its maximum $\frac{1}{\sqrt{2\pi}}$. Moreover

$$\lim_{x\downarrow 0}\frac{\mathscr{U}(x)}{\sqrt{2\ln\frac{1}{x}}}=1. \tag{5.28}$$

Let us begin with several facts we shall use.

Recall that $L=\Delta_\Sigma-x\cdot\nabla$ is the infinitesimal generator of the Ornstein-Uhlenbeck semi-group $(Q_t)_{t>0}$, in the sense that $\frac{d}{dt}Q_t=LQ_t$ for $t>0$.

Lemma 5.24. *Let Ψ be a C^2-function on $\mathbb{R}$. Then*

$$L(\Psi(f)) = \Psi'(f)Lf + \Psi''(f)\nabla f \cdot \Sigma \nabla f \tag{5.29}$$

for any C^2-function f on $\mathbb{R}^D$.

Proof. The equality may be called a chain rule for L, which follows immediately from the rules of computing derivatives. Let f_i and f_{ij} denote the partial derivatives $\frac{\partial}{\partial x_i} f$ and $\frac{\partial^2}{\partial x_i \partial x_j} f$ respectively for simplicity. Then

$$\begin{aligned} L(\Psi(f)) &= \sum_{i,j=1}^{D} \sigma_{ij} \Psi(f)_{ij} - \sum_{i=1}^{D} x_i \Psi(f)_i \\ &= \sum_{i,j=1}^{D} \sigma_{ij} \left(\Psi'(f) f_i\right)_j - \Psi'(f) \sum_{i=1}^{D} x_i f_i \\ &= \Psi'(f) \sum_{i,j=1}^{D} \sigma_{ij} f_{ij} + \Psi''(f) \sum_{i,j=1}^{D} \sigma_{ij} f_j f_i - \Psi'(f) \sum_{i=1}^{D} x_i f_i \\ &= \Psi'(f) Lf + \Psi''(f) \nabla f \cdot \Sigma \nabla f \end{aligned}$$

which completes the proof. □

Lemma 5.25. *Let $f : \mathbb{R}^D \mapsto [0,1]$ be a C^2-function whose derivatives have at most polynomial growth. Let $t > 0$ be fixed but arbitrary, and consider $G(s) = Q_s(\mathscr{U}(Q_{t-s} f))$, that is,*

$$G(s)(x,t) = \int_{\mathbb{R}^D} q_\Sigma(s,x,y) \mathscr{U}(Q_{t-s} f(y)) \gamma(dy) \tag{5.30}$$

for $s \in (0,t)$ and $x \in \mathbb{R}^D$. [The argument (x,t) is suppressed if no confusion may arise]. Then

$$\frac{\partial}{\partial s} G(s) = Q_s\left(\mathscr{U}''(Q_{t-s} f) \nabla(Q_{t-s} f) \cdot \Sigma \nabla(Q_{t-s} f)\right) \tag{5.31}$$

for every $s \in (0,t)$.

Proof. For simplicity we suppress the argument x in $G(s)(x)$ which is fixed though arbitrary. By differentiating in s under integration (which is allowed under our assumptions on f), we obtain

$$\begin{aligned} \frac{\partial}{\partial s} G(s) &= \int_{\mathbb{R}^D} \mathscr{U}(Q_{t-s} f(y)) \frac{\partial}{\partial s} q_\Sigma(s,x,y) \gamma(\mathrm{d}y) \\ &\quad - \int_{\mathbb{R}^D} q_\Sigma(s,x,y) \mathscr{U}'(Q_{t-s} f(y)) \frac{\partial}{\partial s} Q_{t-s} f(y) \gamma(\mathrm{d}y) \\ &= \int_{\mathbb{R}^D} \mathscr{U}(Q_{t-s} f(y)) L q_\Sigma(s,x,y) \gamma(\mathrm{d}y) \\ &\quad - \int_{\mathbb{R}^D} q_\Sigma(s,x,y) \mathscr{U}'(Q_{t-s} f(y)) \frac{\partial}{\partial s} Q_{t-s} f(y) \gamma(\mathrm{d}y) \end{aligned}$$

where we have used the fundamental equation that

$$\frac{\partial}{\partial s} q_\Sigma(s,x,y) = L q_\Sigma(s,x,y)$$

where L operates on the variable y, while x is fixed. Next for the first term we use the symmetry of L, so that

$$\begin{aligned} J_1 &= \int_{\mathbb{R}^D} \mathscr{U}(Q_{t-s}f(y)) L q_\Sigma(s,x,y)\gamma(\mathrm{d}y) \\ &= \int_{\mathbb{R}^D} q_\Sigma(s,x,y) L\mathscr{U}(Q_{t-s}f(y))\gamma(\mathrm{d}y) \\ &= \int_{\mathbb{R}^D} q_\Sigma(s,x,y)\mathscr{U}'(Q_{t-s}f(y))L(Q_{t-s}f)(y))\gamma(\mathrm{d}y) \\ &+ \int_{\mathbb{R}^D} q_\Sigma(s,x,y)\mathscr{U}''(Q_{t-s}f(y))\nabla(Q_{t-s}f)\cdot\Sigma\nabla(Q_{t-s}f)(y))\gamma(\mathrm{d}y) \end{aligned}$$

where the second equality from the chain rule for L. Substituting J_1 into the previous equation for $G'(s)$, and using the fundamental equation

$$\frac{\partial}{\partial r} Q_r f = L(Q_r f)$$

(with $r = t - s > 0$), we obtain that

$$G'(s) = \int_{\mathbb{R}^D} \mathscr{U}''(Q_{t-s}f(y))\left(\nabla(Q_{t-s}f)\cdot\Sigma\nabla(Q_{t-s}f)\right)(y) q_\Sigma(s,\cdot,y)\gamma(\mathrm{d}y) \tag{5.32}$$

for every $s \in (0,t)$, which is equivalent to (5.31). □

Lemma 5.26. *Under the same assumptions as in Lemma 5.25. Let*

$$F(s) = \left(Q_s(\mathscr{U}(Q_{t-s}f))\right)^2 \quad \text{for } s \in (0,t).$$

Then

$$F'(s) = 2Q_s(\mathscr{U}(Q_{t-s}f))\, Q_s\left(\mathscr{U}''(Q_{t-s}f)\nabla(Q_{t-s}f)\cdot\Sigma\nabla(Q_{t-s}f)\right)$$

for $s \in (0,t)$.

Proof. This follows from the previous lemma. Indeed $F = G^2$, so that

$$\begin{aligned} F'(s) &= 2G(s)G'(s) \\ &= 2Q_s(\mathscr{U}(Q_{t-s}f))\, Q_s\left(\mathscr{U}''(Q_{t-s}f)\nabla(Q_{t-s}f)\cdot\Sigma\nabla(Q_{t-s}f)\right) \end{aligned}$$

for every $s \in (0,t)$. □

Lemma 5.27. *Suppose that f is a C^1 function with values in $[0,1]$, and suppose both f and its partial derivatives are γ-integrable. Then*

$$\frac{\sqrt{\nabla(Q_t f)\cdot\Sigma\nabla(Q_t f)}}{\mathscr{U}(Q_t f)} \le \frac{1}{\sqrt{e^{2t}-1}} \quad \text{for every } t > 0. \tag{5.33}$$

Proof. We only need to show this for any C^2-function f taking values in $[0,1]$. Let $t > 0$ and let $F(s) = \left(Q_s(\mathscr{U}(Q_{t-s}f))\right)^2$ for $s \in (0,t)$. Then $F(t) = (Q_t(\mathscr{U}(f)))^2$, $F(0) = (\mathscr{U}(Q_t f))^2$, and

$$\begin{aligned} F(t) - F(0) &= \int_0^t \frac{d}{ds} F(s)\mathrm{d}s \\ &= 2\int_0^t Q_s(\mathscr{U}(Q_{t-s}f))\, Q_s\left(\mathscr{U}''(Q_{t-s}f)\nabla(Q_{t-s}f)\cdot\Sigma\nabla(Q_{t-s}f)\right)\mathrm{d}s. \end{aligned} \tag{5.34}$$

Using the differential equation that $\mathscr{U}'' = -\frac{1}{\mathscr{U}}$ in the previous equality, we obtain that

$$\begin{aligned} F(t) - F(0) &= -2\int_0^t Q_s\left(\mathscr{U}\left(Q_{t-s}f\right)\right) Q_s\left(\frac{\nabla(Q_{t-s}f)\cdot\Sigma\nabla(Q_{t-s}f)}{\mathscr{U}\left(Q_{t-s}f\right)}\right)\mathrm{d}s \\ &\le -2\int_0^t \left(Q_s\left(\sqrt{\nabla(Q_{t-s}f)\cdot\Sigma\nabla(Q_{t-s}f)}\right)\right)^2 \mathrm{d}s \end{aligned} \tag{5.35}$$

where the second inequality follows from the Cauchy-Schwartz inequality:

$$Q_s\left(\sqrt{\nabla(Q_{t-s}f)\cdot\Sigma\nabla(Q_{t-s}f)}\right) \le \sqrt{Q_s\left(\mathscr{U}\left(Q_{t-s}f\right)\right)}\sqrt{Q_s\left(\frac{\nabla(Q_{t-s}f)\cdot\Sigma\nabla(Q_{t-s}f)}{\mathscr{U}\left(Q_{t-s}f\right)}\right)}$$

which implies that

$$Q_s\left(\mathscr{U}\left(Q_{t-s}f\right)\right) Q_s\left(\frac{\nabla(Q_{t-s}f)\cdot\Sigma\nabla(Q_{t-s}f)}{\mathscr{U}\left(Q_{t-s}f\right)}\right) \ge \left(Q_s\left(\sqrt{\nabla(Q_{t-s}f)\cdot\Sigma\nabla(Q_{t-s}f)}\right)\right)^2.$$

By the domination inequality (cf. Theorem 5.8):

$$\begin{aligned} \sqrt{\nabla(Q_tf)\cdot\Sigma\nabla(Q_tf)} &= \sqrt{\nabla(Q_s(Q_{t-s}f))\cdot\Sigma\nabla(Q_s(Q_{t-s}f))} \\ &\le e^{-s}Q_s\left(\sqrt{\nabla(Q_{t-s}f)\cdot\Sigma\nabla(Q_{t-s}f)}\right) \end{aligned}$$

for every $s \in (0,t)$. Rearrange the inequality to obtain that that

$$\left(Q_s\left(\sqrt{\nabla(Q_{t-s}f)\cdot\Sigma\nabla(Q_{t-s}f)}\right)\right)^2 \ge e^{2s}\nabla(Q_tf)\cdot\Sigma\nabla(Q_tf) \tag{5.36}$$

for any $s \in (0,t)$. Substituting this into (5.35) we thus get that

$$\begin{aligned} F(t) - F(0) &\le -2\int_0^t e^{2s}\nabla(Q_tf)\cdot\Sigma\nabla(Q_tf)\mathrm{d}s \\ &= -(e^{2t}-1)\nabla(Q_tf)\cdot\Sigma(\nabla Q_tf) \end{aligned}$$

which yields that

$$\nabla(Q_tf)\cdot\Sigma\nabla(Q_tf) \le \frac{1}{e^{2t}-1}\left[(\mathscr{U}(Q_tf))^2 - (Q_t(\mathscr{U}(f)))^2\right]$$

and therefore

$$\frac{\sqrt{\nabla(Q_tf)\cdot\Sigma\nabla(Q_tf)}}{\mathscr{U}(Q_tf)} \le \frac{1}{\sqrt{e^{2t}-1}}\sqrt{1-\left(\frac{Q_t(\mathscr{U}(f))}{\mathscr{U}(Q_tf)}\right)^2}$$

for every $t > 0$. This completes the proof. □

Exercise. Let ψ be an increasing C^1 function on $[0,\infty)$, and f is a C^1 function on $\mathbb{R}^D$ taking values in $[0,1]$. Prove that

$$\psi(Q_t(\mathscr{U}(f))) - \psi(\mathscr{U}(Q_tf)) \le -(\nabla(Q_tf)\cdot\Sigma\nabla(Q_tf))\int_0^t e^{2s}\frac{\psi'(Q_s(\mathscr{U}(Q_{t-s}f)))}{Q_s(\mathscr{U}(Q_{t-s}f))}\mathrm{d}s$$

for any $t > 0$.

[*Hint*: For any $t > 0$ be any but fixed. Consider $\varphi(s) = \psi(Q_s(\mathscr{U}(Q_{t-s}f)))$ for $s \in [0,t]$. Then

$$\psi(Q_t(\mathscr{U}(f))) - \psi(\mathscr{U}(Q_t f)) = \int_0^t \frac{d}{ds}\varphi(s)\mathrm{d}s.$$

Compute $\varphi(s)$ and use Theorem 5.8 as in the proof of the previous lemma.]

We are now in a position to prove the isoperimetric inequality for Gaussian measures.

Theorem 5.28. (Isoperimetric inequality for Gaussian measures) *Let $f : \mathbb{R}^D \mapsto [0,1]$ be C^1-function and $|\nabla f|$ is γ-integrable. Then*

$$\mathscr{U}\left(\int_{\mathbb{R}^D} f d\gamma\right) - \int_{\mathbb{R}^D} \mathscr{U}(f) d\gamma \leq \int_{\mathbb{R}^D} \sqrt{\nabla f \cdot \Sigma \nabla f} d\gamma. \tag{5.37}$$

Proof. Let us apply the approach we have tested in the previous sections. Consider

$$F(t) = \int_{\mathbb{R}^D} \mathscr{U}(Q_t f)\,\mathrm{d}\gamma.$$

Then $F(\infty) = \mathscr{U}\left(\int_{\mathbb{R}^D} f \mathrm{d}\gamma\right)$ and $F(0) = \int_{\mathbb{R}^D} \mathscr{U}(f)\,\mathrm{d}\gamma$, and

$$\mathscr{U}\left(\int_{\mathbb{R}^D} f \mathrm{d}\gamma\right) - \int_{\mathbb{R}^D} \mathscr{U}(f)\,\mathrm{d}\gamma = \int_0^\infty \frac{d}{dt}F(t)\mathrm{d}t.$$

Next we compute the derivative: differentiating under integration gives

$$\begin{aligned}\frac{d}{dt}F(t) &= \int_{\mathbb{R}^D} \frac{d}{dt}\mathscr{U}(Q_t f)\,\mathrm{d}\gamma \\ &= \int_{\mathbb{R}^D} \mathscr{U}'(Q_t f)\frac{d}{dt}Q_t f \mathrm{d}\gamma.\end{aligned}$$

Using the equation $\frac{d}{dt}Q_t f = LQ_t f$ and performing integration by parts we obtain

$$\begin{aligned}\frac{d}{dt}F(t) &= \int_{\mathbb{R}^D} \mathscr{U}'(Q_t f) LQ_t f \mathrm{d}\gamma \\ &= -\int_{\mathbb{R}^D} \nabla(\mathscr{U}'(Q_t f)) \cdot \Sigma \nabla(Q_t f)\mathrm{d}\gamma \\ &= -\int_{\mathbb{R}^D} \mathscr{U}''(Q_t f)\nabla(Q_t f) \cdot \Sigma \nabla(Q_t f)\mathrm{d}\gamma.\end{aligned}$$

Since $\mathscr{U}'' = -\frac{1}{\mathscr{U}}$, we therefore have

$$\frac{d}{dt}F(t) = \int_{\mathbb{R}^D} \frac{\nabla(Q_t f) \cdot \Sigma \nabla(Q_t f)}{\mathscr{U}(Q_t f)}\mathrm{d}\gamma$$

for every $t > 0$. Finally we apply the estimate we have proven in Lemma 5.27

$$\frac{\sqrt{\nabla(Q_t f) \cdot \Sigma \nabla(Q_t f)}}{\mathscr{U}(Q_t f)} \leq \frac{1}{\sqrt{e^{2t} - 1}}$$

and deduce that

$$\begin{aligned}\frac{d}{dt}F(t) &\leq \frac{1}{\sqrt{e^{2t}-1}}\int_{\mathbb{R}^D}\sqrt{\nabla(Q_tf)\cdot\Sigma\nabla(Q_tf)}\mathrm{d}\gamma \\ &\leq \frac{1}{\sqrt{e^{2t}-1}}\int_{\mathbb{R}^D}e^{-t}Q_t(\sqrt{\nabla f\cdot\Sigma\nabla f})\mathrm{d}\gamma \\ &= \frac{e^{-t}}{\sqrt{e^{2t}-1}}\int_{\mathbb{R}^D}\sqrt{\nabla f\cdot\Sigma\nabla f}\mathrm{d}\gamma\end{aligned}$$

Integrating both sides of the previous inequality on $(0,\infty)$ we therefor obtain that

$$\begin{aligned}\mathscr{U}\left(\int_{\mathbb{R}^D}f\mathrm{d}\gamma\right)-\int_{\mathbb{R}^D}\mathscr{U}(f)\mathrm{d}\gamma &\leq \int_0^\infty\frac{e^{-t}}{\sqrt{e^{2t}-1}}\mathrm{d}t\int_{\mathbb{R}^D}\sqrt{\nabla f\cdot\Sigma\nabla f}\mathrm{d}\gamma \\ &= \int_{\mathbb{R}^D}\sqrt{\nabla f\cdot\Sigma\nabla f}\mathrm{d}\gamma\end{aligned}$$

which completes the proof. □

If $A\in\mathbb{R}^D$ be a closed subset with a C^1-boundary, then

$$\gamma_S(\partial A)=\liminf_{\varepsilon\downarrow 0}\frac{\gamma(A_\varepsilon)-\gamma(A)}{\varepsilon}$$

where $A_\varepsilon=\left\{x\in\mathbb{R}^D : d(x,A)<\varepsilon\right\}$, is called the Minkowski outer content of the boundary of A. Here the distance d is the metric associated with Σ, i.e.

$$d(x,y)=\sup_{f\in C^1}\left\{|f(x)-f(y)| : \nabla f\cdot\Sigma\nabla f\leq 1\right\}.$$

Indeed $d(x,y)=\sqrt{(x-y)\cdot\Sigma^{-1}(x-y)}$ for any $x,y\in\mathbb{R}^D$. Note that if $\varepsilon\mapsto\gamma(A_\varepsilon)$ is differentiable (from right), then

$$\gamma_S(\partial A)=\left.\frac{d}{d\varepsilon}\right|_{\varepsilon=0+}\gamma(A_\varepsilon).$$

Corollary 5.29. *Let $\gamma(dx)=G_\Sigma(x)dx$ be a central Gaussian measure with co-variance matrix Σ. Then*

$$\mathscr{U}(\gamma(A))\leq\gamma_S(\partial A)$$

for any closed subset $A\subset\mathbb{R}^D$ with a C^1-boundary.

Proof. Choose C^1-functions f_n valued in $[0,1]$which tends to 1_A. Then

$$\mathscr{U}\left(\int_{\mathbb{R}^D}f_n\mathrm{d}\gamma\right)-\int_{\mathbb{R}^D}\mathscr{U}(f_n)\mathrm{d}\gamma\leq\int_{\mathbb{R}^D}\sqrt{\nabla f_n\cdot\Sigma\nabla f_n}\mathrm{d}\gamma$$

for every n. Since $\mathscr{U}(0)=\mathscr{U}(1)=0$ so that

$$\mathscr{U}\left(\int_{\mathbb{R}^D}f_n\mathrm{d}\gamma\right)\to\mathscr{U}(\gamma(A)),\quad\int_{\mathbb{R}^D}\mathscr{U}(f_n)\mathrm{d}\gamma\to 0$$

and

$$\int_{\mathbb{R}^D}\sqrt{\nabla f_n\cdot\Sigma\nabla f_n}\mathrm{d}\gamma\to\gamma_S(\partial A)$$

which thus yields the isopermetric inequality. □

Theorem 5.30. *Suppose $\gamma(dx) = G_\Sigma(x)dx$ is a Gaussian measure on $\mathbb{R}^D$, and $A \subset \mathbb{R}^D$ be Borel measurable with C^1-boundary. Then*

$$\gamma(A_t) \geq \Phi\left(\Phi^{-1}(\gamma(A)) + t\right) \quad \text{for } t \geq 0, \tag{5.38}$$

where $A_\varepsilon = \{x \in \mathbb{R}^D : d(x,A) \leq \varepsilon\}$ for every $\varepsilon > 0$, and the distance d is the metric associated with Σ, i.e.

$$d(x,y) = \sup_{f \in C^1} \{|f(x) - f(y)| : \nabla f \cdot \Sigma \nabla f \leq 1\}.$$

It is a fact that $d(x,y) = \sqrt{(x-y)\cdot \Sigma^{-1}(x-y)}$ for any $x, y \in \mathbb{R}^D$.

Proof. The isoperimetric inequality may be written as

$$\frac{d}{dr}\gamma(A_r) \geq \mathscr{U}(\gamma(A_r))$$

for $r \geq 0$, i.e.

$$\frac{1}{\mathscr{U}(\gamma(A_r))}\frac{d}{dr}\gamma(A_r) \geq 1 \quad \text{for } r \geq 0.$$

Integrating the inequality over $[0,t]$ (for $t > 0$) to obtain that

$$\int_0^t \frac{1}{\mathscr{U}(\gamma(A_r))}\frac{d}{dr}\gamma(A_r)\mathrm{d}r = \int_{\gamma(A)}^{\gamma(A_t)} \frac{1}{\mathscr{U}(s)} ds \geq t$$

On the other hand

$$\int_{\gamma(A)}^{\gamma(A_t)} \frac{1}{\mathscr{U}(s)} ds = \int_{\gamma(A)}^{\gamma(A_t)} \frac{1}{\Phi' \circ \Phi^{-1}(s)} ds = \int_{\gamma(A)}^{\gamma(A_t)} \frac{d}{ds}\Phi^{-1}(s) ds$$
$$= \Phi^{-1}(\gamma(A_t)) - \Phi^{-1}(\gamma(A))$$

and therefore

$$\Phi^{-1}(\gamma(A_t)) - \Phi^{-1}(\gamma(A)) \geq t$$

which yield the inequality (5.38). □

As a consequence we deduce the following concentration estimate.

Theorem 5.31. *Let $\gamma(dx) = G_\Sigma(x)dx$ be a centered Gaussian measure on $\mathbb{R}^D$. Let $f : \mathbb{R}^D \to \mathbb{R}$ be a function such that $\nabla f \cdot \Sigma \nabla f \leq 1$. Let $m \in \mathbb{R}^D$ such that $\gamma(\{f \leq m\}) \geq \frac{1}{2}$. Then*

$$\gamma(\{f > m + r\}) \leq \int_r^\infty \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}} dx \tag{5.39}$$

for any $r \geq 0$.

Proof. Let $A = \{f \leq m\}$. Then $\gamma(A) \geq \frac{1}{2} = \Phi(0)$ which implies that $\Phi^{-1}(\gamma(A)) \geq 0$. Also the condition that $\nabla f \cdot \Sigma \nabla f \leq 1$ implies that $A_r \subset \{f \leq m + r\}$, and therefore, (5.38) yields that

$$\gamma(\{f \leq m + r\}) \geq \Phi(r) = \int_{-\infty}^r \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}} \mathrm{d}x$$

and the conclusion follows immediately. □

By an approximation procedure, we therefore have the following.

Proposition 5.32. *Let $X = (X_1, \cdots, X_D)$ be a D-dimensional random vector on $(\Omega, \mathscr{F}, \mathbb{P})$ with the standard normal distribution $N(0,I)$ on $\mathbb{R}^D$, $f : \mathbb{R}^D \mapsto \mathbb{R}$ is Lipschitz such that $\|f\|_{Lip} \leq 1$, and let m be a number such that $\mathbb{P}[f(X) \leq m] \geq \frac{1}{2}$. Then*

$$\mathbb{P}[f(X) > m+r] \leq \int_r^\infty \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}} dx$$

for every $r > 0$.

Theorem 5.33. *Let $Y = (Y_1, \cdots, Y_D)$ be a D-dimensional Gaussian random vector on $(\Omega, \mathscr{F}, \mathbb{P})$ with mean zero and co-variance matrix $\Sigma = (\sigma_{ij})$, and let m be a number such that $\mathbb{P}[\sup_i Y_i \leq m] \geq \frac{1}{2}$. Then*

$$\mathbb{P}\left[\sup_{i=1,\cdots,D} Y_i > m+r\right] \leq \int_{\frac{r}{\sup_{i=1,\cdots,D}\sqrt{\sigma_{ii}}}}^\infty \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}} dx \tag{5.40}$$

for every $r > 0$, where $\sigma_{ii} = \mathbb{E}(Y_i^2)$ is the variance of Y_i for $i = 1, \ldots, D$.

Proof. As in the proof of Theorem 5.17, Y and $\Sigma^{\frac{1}{2}} X$ have the same distribution $N(0, \Sigma)$ (where X has the standard normal distribution $N(0,I)$). Apply Proposition 5.32 with

$$f(x) = \frac{1}{\sup_i \sqrt{\sigma_{ii}}} \sup_i \sum_j \rho_{ij} x_j$$

where $\Sigma^{\frac{1}{2}} = (\rho_{ij})$ is a square root of Σ. Then $\|f\|_{\text{Lip}} \leq 1$ (see the proof of the Borell inequality, Theorem 5.17), and the concentration inequality (5.40) follows immediately. □

This theorem implies Borell's inequality we have proved.

6 Brunn-Minkowski's inequality, Isoperimetric inequality

In this part we demonstrate some special features of datasets lying in convex domains. The main tool is the isoperimetric inequality for the Lebesgue measure on $\mathbb{R}^D$.

As in the previous sections, if $A \subset \mathbb{R}^D$ is a Borel measurable subset, then $|A|$ denotes the Lebesgue measure of A. If A is a box with sides parallel to axises, and if the length of the side parallel to x^i-axis is α_i, then $|A| = \prod_{i=1}^D \alpha_i$. If A and B are two Borel measurable sets of $\mathbb{R}^D$, then $A + B = \{a + b : a \in A, b \in B\}$ and $\lambda A = \{\lambda x : x \in A\}$ are Borel measurable too. In particular, if $a \in \mathbb{R}^D$, then $a + A = \{a\} + A$ is measurable and $|a + A| = |A|$, i.e. the Lebesgue measure is translation invariant.

6.1 Prékopa-Leindler's inequality

Let us begin with a lemma which is the Brunn-Minkowski inequality on $\mathbb{R}$.

Lemma 6.1. *Let A, B be two Borel measurable subsets of $\mathbb{R}$. Then*

$$|A+B| \geq |A|+|B| \tag{6.1}$$

and

$$|\lambda A+(1-\lambda)B| \geq \lambda|A|+(1-\lambda)|B| \tag{6.2}$$

for every $\lambda \in (0,1)$.

Proof. The second inequality follows from the first as $|\lambda A| = \lambda|A|$. Let us prove the first inequality for non-empty compact subsets A and B. Choose a and b such that $\tilde{A} = \{a\}+A \subset \mathbb{R}_-$, $\tilde{B} = \{b\}+B \subset \mathbb{R}_+$ and $\tilde{A} \cap \tilde{B} = \{0\}$. Then $\tilde{A} \cup \tilde{B} \subset \tilde{A}+\tilde{B} = a+b+A+B$. Therefore

$$|A+B| = |\tilde{A}+\tilde{B}| \geq |\tilde{A} \cup \tilde{B}| = |\tilde{A}|+|\tilde{B}| = |A|+|B|$$

and the proof is complete. □

Lemma 6.2. *Let a, b are two positive numbers. Then*

$$\lambda a+(1-\lambda)b \geq a^{\lambda} b^{1-\lambda} \tag{6.3}$$

for any $\lambda \in (0,1)$.

Proof. This follows from Jensen's inequality. Since $x \mapsto \ln x$ is concave (i.e. $-\ln x$ is convex) on $(0,\infty)$, therefore

$$\ln(\lambda a+(1-\lambda)b) \geq \lambda \ln a+(1-\lambda)\ln b$$

and the inequality follows immediately. □

Lemma 6.3. *Let f and g be two non-negative, continuous functions on $\mathbb{R}$, and let $\lambda \in (0,1)$ be a constant. Then*

$$\int_{\mathbb{R}} h(x)dx \geq \lambda \int_{\mathbb{R}} f(x)dx+(1-\lambda)\int_{\mathbb{R}} g(x)dx \tag{6.4}$$

where h is defined by

$$h(x) = \sup_{y\in\mathbb{R}} f\left(\frac{x-y}{\lambda}\right)^{\lambda} g\left(\frac{y}{1-\lambda}\right)^{1-\lambda}$$

for $x \in \mathbb{R}$.

Proof. To prove (6.4), we consider

$$A(t) = \{x \in \mathbb{R} : f(x) > t\}, \quad B(t) = \{x \in \mathbb{R} : g(x) > t\}, \quad C(t) = \{x \in \mathbb{R} : h(x) > t\}$$

for every $t > 0$. By definition of $h_\lambda(f,g)$, we have

$$\lambda A(t)+(1-\lambda)B(t) \subset C(t) \tag{6.5}$$

for any $t \geq 0$, and therefore

$$\begin{aligned} |C(t)| &\geq |\lambda A(t)+(1-\lambda)B(t)| \\ &\geq \lambda|A(t)|+(1-\lambda)|B(t)|, \end{aligned}$$

where the second inequality follows from Lemma 6.1. Integrating the previous inequality in $t \in (0,\infty)$ and using the dis-integration formula (2.7) we have

$$\int_{\mathbb{R}} h(x)\mathrm{d}x = \int_0^\infty |C(t)|\mathrm{d}t \geq \lambda \int_0^\infty |A(t)|\mathrm{d}t + (1-\lambda)\int_0^\infty |B(t)|\mathrm{d}t$$
$$= \lambda \int_{\mathbb{R}} f(x)\mathrm{d}x + (1-\lambda)\int_{\mathbb{R}} g(x)\mathrm{d}x$$

which completes the proof of (6.4). □

Theorem 6.4. (Prékopa-Leindler Inequality) *Let f and g be two non-negative Borel measurable functions on $\mathbb{R}^D$ and $\lambda \in (0,1)$. Then*

$$\int_{\mathbb{R}^D} h(x)dx \geq \left(\int_{\mathbb{R}^D} f(x)dx\right)^\lambda \left(\int_{\mathbb{R}^D} g(x)dx\right)^{1-\lambda} \tag{6.6}$$

where $h = h_\lambda(f,g)$ defined by

$$h_\lambda(f,g)(x) = \sup_{y\in\mathbb{R}^D} f\left(\frac{x-y}{\lambda}\right)^\lambda g\left(\frac{y}{1-\lambda}\right)^{1-\lambda} \quad \textit{for } x \in \mathbb{R}^D. \tag{6.7}$$

Proof. [*The proof is not examinable.*] For simplicity we use h to denote $h_\lambda(f,g)$ if no confusion may arise, and by a simple approximation procedure, we may assume that f and g are continuous. Without losing generality we shall assume that

$$\int_{\mathbb{R}^D} f(x)\mathrm{d}x > 0 \quad \text{and} \quad \int_{\mathbb{R}^D} g(x)\mathrm{d}x > 0,$$

as otherwise the inequality is trivial.

Let us prove (6.6) by using induction argument on the dimension D.

If $D = 1$, then (6.6) follows from (6.4) and (6.3). Indeed

$$\int_{\mathbb{R}} h(x)\mathrm{d}x \geq \lambda \int_{\mathbb{R}} f(x)\mathrm{d}x + (1-\lambda)\int_{\mathbb{R}} g(x)\mathrm{d}x$$
$$\geq \left(\int_{\mathbb{R}} f(x)\mathrm{d}x\right)^\lambda \left(\int_{\mathbb{R}} g(x)\mathrm{d}x\right)^{1-\lambda}.$$

Now assume that $D > 2$ and let $\lambda \in (0,1)$. Suppose that (6.6) holds for any non-negative functions f,g on $\mathbb{R}^{D-1}$.

Let $f(x), g(x)$ be two non-negative, continuous functions on $\mathbb{R}^D$ (where $x \in \mathbb{R}^D$). Write $x = (x, x_D)$ where $x \in \mathbb{R}^{D-1}$ and define

$$f_0(x) = \int_{-\infty}^\infty f(x,s)\mathrm{d}s, \quad g_0(x) = \int_{-\infty}^\infty g(x,s)\mathrm{d}s.$$

By assumptions

$$h_\lambda(f,g)(x,x_D) \geq \sup_{s\in\mathbb{R}} f\left(\frac{x-y}{\lambda}, \frac{x_D - s}{\lambda}\right)^\lambda g\left(\frac{y}{1-\lambda}, \frac{s}{1-\lambda}\right)^{1-\lambda}$$

for every $y \in \mathbb{R}^{D-1}$. For any $x, y \in \mathbb{R}^{D-1}$ fixed but arbitrary, we apply Lemma 6.3, (6.4), with one dimensional functions $s \mapsto f\left(\frac{x-y}{\lambda}, s\right)^\lambda$ and $s \mapsto g\left(\frac{y}{1-\lambda}, s\right)$, to obtain that

$$\begin{aligned}\int_{-\infty}^{\infty} h_\lambda(f,g)(x,s)\mathrm{d}s &\geq \lambda \int_{-\infty}^{\infty} f\left(\frac{x-y}{\lambda}, s\right)\mathrm{d}s + (1-\lambda)\int_{-\infty}^{\infty} g\left(\frac{y}{1-\lambda}, s\right)\mathrm{d}s \\ &\geq \left(\int_{-\infty}^{\infty} f\left(\frac{x-y}{\lambda}, s\right)\mathrm{d}s\right)^\lambda \left(\int_{-\infty}^{\infty} g\left(\frac{y}{1-\lambda}, s\right)\mathrm{d}s\right)^{1-\lambda}\end{aligned}$$

where the second inequality follows from (6.3). Since $y \in \mathbb{R}^{D-1}$ is arbitrary, so that

$$\begin{aligned}\int_{-\infty}^{\infty} h_\lambda(f,g)(x,s)\mathrm{d}s &\geq \sup_{y\in\mathbb{R}^{D-1}} \left(\int_{-\infty}^{\infty} f\left(\frac{x-y}{\lambda}, s\right)\mathrm{d}s\right)^\lambda \left(\int_{-\infty}^{\infty} g\left(\frac{y}{1-\lambda}, s\right)\mathrm{d}s\right)^{1-\lambda} \\ &= h_\lambda(f_0, g_0)(x)\end{aligned} \tag{6.8}$$

for every $x \in \mathbb{R}^{D-1}$. Using induction assumption with f_0 and g_0 which are non-negative functions on $\mathbb{R}^{D-1}$, we thus obtain that

$$\int_{\mathbb{R}^{D-1}} h_\lambda(f_0,g_0)(x)\mathrm{d}x \geq \left(\int_{\mathbb{R}^{D-1}} f_0(x)\mathrm{d}x\right)^\lambda \left(\int_{\mathbb{R}^{D-1}} g_0(x)\mathrm{d}x\right)^{1-\lambda}.$$

On the other hand, by (6.8) and Fubini's theorem

$$\begin{aligned}\int_{\mathbb{R}^D} h_\lambda(f,g)(x)\mathrm{d}x &= \int_{\mathbb{R}^{D-1}}\int_{-\infty}^{\infty} h_\lambda(f,g)(x,s)\mathrm{d}s \\ &\geq \int_{\mathbb{R}^{D-1}} h_\lambda(f_0,g_0)(x)\mathrm{d}x \\ &\geq \left(\int_{\mathbb{R}^{D-1}} f_0(x)\mathrm{d}x\right)^\lambda \left(\int_{\mathbb{R}^{D-1}} g_0(x)\mathrm{d}x\right)^{1-\lambda} \\ &= \left(\int_{\mathbb{R}^D} f(x)\mathrm{d}x\right)^\lambda \left(\int_{\mathbb{R}^{D-1}} g(x)\mathrm{d}x\right)^{1-\lambda}\end{aligned}$$

and therefore (6.6) holds for any non-negative, continuous functions f and g. The proof is complete. □

Theorem 6.4 is formulated by H. Brascamp and E. H. Lieb [6] (this paper has an unusual long title as if the JFA journal printed its Abstract as the title !) The original P-L inequality follows of course from the above version immediately.

Theorem 6.5. (Pékopa-Leindler Inequality) *Let f, g and h be non-negative measurable functions on $\mathbb{R}^D$ and $\lambda \in (0,1)$. Suppose*

$$h(\lambda x + (1-\lambda)y) \geq f(x)^\lambda g(y)^{1-\lambda} \quad \text{for any } x, y \in \mathbb{R}^D. \tag{6.9}$$

Then

$$\int_{\mathbb{R}^D} h(x)dx \geq \left(\int_{\mathbb{R}^D} f(x)dx\right)^\lambda \left(\int_{\mathbb{R}^D} g(x)dx\right)^{1-\lambda}. \tag{6.10}$$

Proof. Under assumption, $h(x) \geq h_\lambda(f,g)(x)$ for every x, and therefore the P-L inequality follows immediately from (6.6). □

Definition 6.6. *Let f be a non-negative function on $\mathbb{R}^D$. Then f is* log-concave *(i.e. logarithmically concave) if*

$$f(\lambda x + (1-\lambda)y) \geq f(x)^\lambda f(y)^{1-\lambda}$$

for any $\lambda \in [0,1]$ and $x, y \in \mathbb{R}^D$.

By definition, f is log-concave if and only if $-\ln f$ is convex on $\{f > 0\}$.

Exercise. Let ρ be log-concave on $\mathbb{R}^D = \mathbb{R}^{D_1} \times \mathbb{R}^{D_2}$ (where $D_1 + D_2 = D$). Let

$$\rho_1(x_1) = \int_{\mathbb{R}^{D_2}} \rho(x_1, x_2) \mathrm{d}x_2$$

where $x_i \in \mathbb{R}^{D_i}$ $(i = 1,2)$. Show that ρ_1 is log-concave too. [*Hint: Use Theorem* 6.4].

Theorem 6.7. *If ρ is non-negative and log-concave on $\mathbb{R}^D$, then*

$$\int_{\lambda A + (1-\lambda)B} \rho(x) dx \geq \left(\int_A \rho(x) dx \right)^\lambda \left(\int_B \rho(x) dx \right)^{1-\lambda}$$

for any Borel measurable subsets $A, B \subset \mathbb{R}^D$ and for any $\lambda \in (0,1)$.

Proof. We shall apply Theorem 6.4 to $f = 1_A \rho$ and $g = 1_B \rho$. Since ρ is log-concave, for every $\lambda \in (0,1)$,

$$\rho\left(\frac{x-y}{\lambda}\right)^\lambda \rho\left(\frac{y}{1-\lambda}\right)^{1-\lambda} \leq \rho(x)$$

for any x and y. If $\frac{x-y}{\lambda} \in A$ and $\frac{y}{1-\lambda} \in B$, than $x \in \lambda A + (1-\lambda)B$, which implies that $h_\lambda(f,g) \leq 1_{\lambda A + (1-\lambda)B} \rho$. Therefore according to (6.6) we have

$$\begin{aligned}\int_{\mathbb{R}^D} 1_{\lambda A + (1-\lambda)B} \rho(x) \mathrm{d}x &\geq \int_{\mathbb{R}^D} h_\lambda(f,g) \mathrm{d}x \\ &\geq \left(\int_{\mathbb{R}^D} 1_A \rho(x) \mathrm{d}x \right)^\lambda \left(\int_{\mathbb{R}^D} 1_B \rho(x) \mathrm{d}x \right)^{1-\lambda}\end{aligned}$$

which yields (6.11). □

Lemma 6.8. *Let Σ be a symmetric, positive definite $D \times D$-matrix. Then the central Gaussian kernel $G_\Sigma(x)$ is log-concave.*

Proof. Recall that

$$\ln G_\Sigma(x) = -\frac{1}{2} \ln\left((2\pi)^D \det \Sigma \right) - \frac{1}{2} x \cdot \Sigma^{-1} x.$$

Hence we only need to show that $x \mapsto x \cdot \Sigma^{-1} x$ is convex. Let $x, y \in \mathbb{R}^D$ be any two points. Consider

$$\varphi(\lambda) = (\lambda x + (1-\lambda)y) \cdot \Sigma^{-1} (\lambda x + (1-\lambda)y)$$

for $\lambda \in [0,1]$. Then

$$\varphi'(\lambda) = 2(x-y)\cdot \Sigma^{-1}(\lambda x + (1-\lambda)y)$$

and

$$\varphi''(\lambda) = 2(x-y)\cdot \Sigma^{-1}(x-y) \geq 0$$

as Σ^{-1} is symmetric, positive definite. Hence φ is convex on $[0,1]$, and therefore

$$\varphi(\lambda) = \varphi(\lambda 1 + (1-\lambda)0) \leq \lambda\varphi(1) + (1-\lambda)\varphi(0)$$

for any $\lambda \in (0,1)$. That is

$$-(\lambda x + (1-\lambda)y)\cdot \Sigma^{-1}(\lambda x + (1-\lambda)y) \geq -\lambda x\cdot \Sigma^{-1}x - (1-\lambda)y\cdot \Sigma^{-1}y$$

which in turn yields that $\ln G_\Sigma$ is concave. □

As a consequence, we have the following result for Gaussian distributions.

Theorem 6.9. (Geometric form of the isoperimetric inequality for Gaussian measure) *Let $\gamma(dx) = G_\Sigma(x)dx$ be a centered Gaussian measure on $\mathscr{B}(\mathbb{R}^D)$ with co-variance matrix Σ. Then*

$$\gamma(\lambda A + (1-\lambda)B) \geq \gamma(A)^\lambda \gamma(B)^{1-\lambda} \tag{6.11}$$

for any Borel measurable subsets $A, B \subset \mathbb{R}^D$ and for any $\lambda \in (0,1)$.

This follows from the fact that $x \mapsto G_\Sigma(x)$ is log-concave, Lemma 6.8.

Exercise. Let $\gamma(\mathrm{d}x)$ be the centered Gaussian measure $G_\Sigma(x)\mathrm{d}x$. Let A be a symmetric convex subset of $\mathbb{R}^D$ and $a \in \mathbb{R}^D$.

(a) Prove that

$$\gamma(A+a) \leq \gamma(A+ta)$$

for any $t \in [0,1]$, and $t \mapsto \gamma(A+ta)$ is non-increasing on $[0,\infty)$.

[*Hint*: You may assume that $\Sigma = I$, otherwise consider $\Sigma^{-\frac{1}{2}}A$ and $\Sigma^{-\frac{1}{2}}a$ instead. Apply Theorem 6.9 to $\lambda = \frac{1}{2}(t+1)$, use the fact that $\gamma(A+a) = \gamma(A-a)$, and the fact that

$$A + ta = \lambda(A+a) + (1-\lambda)(A-a)$$

in (6.11).]

(b) Suppose f is convex and $f(x) = f(-x)$ for every x. Show that

$$\int_{\mathbb{R}^D} f(x)\gamma(dx) \leq \int_{\mathbb{R}^D} f(x+a)\gamma(\mathrm{d}x)$$

for any $a \in \mathbb{R}^D$, and conclude that $t \to \int_{\mathbb{R}^D} f(x+ta)\gamma(\mathrm{d}x)$ is non-decreasing.

[*Hint*: Apply (a) to level sets $\{f \leq c\}$ for every c.]

(c) Prove that

$$\int_{\mathbb{R}^D} |x|^p\gamma(dx) \leq \int_{\mathbb{R}^D} |x+a|^p\gamma(\mathrm{d}x)$$

for any $a \in \mathbb{R}^D$ and $p \geq 1$.

6.2 Brunn-Minkowski's theorem

This is a deep result about the Lebesgue measure. Let begin with a weak version which is independent of the dimension D.

Theorem 6.10. *Suppose A, B are two Borel measurable subsets of $\mathbb{R}^D$ and $\lambda \in (0,1)$. Then*

$$|\lambda A + (1-\lambda)B| \geq |A|^{\lambda}|B|^{1-\lambda}. \tag{6.12}$$

Proof. It follows immediately from the Prékopa-Leindler inequality. Indeed, if $f = 1_A$ and $g = 1_B$, then $h_\lambda(f,g) = 1_{\lambda A+(1-\lambda)B}$. Hence (6.6) gives (6.12). □

In fact this weak version, in which the dimension seems missing, is equivalent to the Brunn-Minkowski inequality, and the dimension may be recovered from the scaling property: $|\lambda A| = \lambda^D |A|$ for $A \in \mathscr{B}(\mathbb{R}^D)$.

Theorem 6.11. *Let A and B be two bounded Borel measurable subsets of $\mathbb{R}^D$. Then*

$$|A+B|^{\frac{1}{D}} \geq |A|^{\frac{1}{D}} + |B|^{\frac{1}{D}}. \tag{6.13}$$

Proof. We may assume that $|A| > 0$ and $|B| > 0$. Let $\tilde{A} = |A|^{-1/D}A$ and $\tilde{B} = |B|^{-1/D}B$. Then $|\tilde{A}| = |\tilde{B}| = 1$, and therefore by (6.12) we deduce that

$$|\lambda\tilde{A} + (1-\lambda)\tilde{B}| \geq 1 \quad \forall \lambda \in (0,1).$$

Set

$$\lambda = \frac{|A|^{1/D}}{|A|^{1/D} + |B|^{1/D}}$$

so that

$$1-\lambda = \frac{|B|^{1/D}}{|A|^{1/D} + |B|^{1/D}}.$$

The previous inequality may be written as

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

which yields (6.13). The proof is complete. □

We are now in a position to prove the well-known isoperimetric inequality. To this end we shall define the area measure. Suppose $\Omega \subset \mathbb{R}^D$ with a C^1 boundary $\partial\Omega$. Then the area of $\partial\Omega$ is given by

$$A(\partial\Omega) = \liminf_{\varepsilon \downarrow 0} \frac{|\Omega + \varepsilon B_1| - |\Omega|}{\varepsilon}$$

where B_1 is the unit ball in $\mathbb{R}^D$ with center 0.

Theorem 6.12. (The isoperimetric inequality) *Let $\Omega \subset \mathbb{R}^D$ be a relatively compact region with a C^1 boundary $\partial\Omega$. Then*

$$\frac{A(\partial\Omega)}{|\Omega|^{1-\frac{1}{D}}} \geq \frac{A(S^{D-1})}{|B_1|^{1-\frac{1}{D}}}$$

where S^{D-1} is the unit sphere in D-dimensional space $\mathbb{R}^D$. In particular if $|\Omega| = |B_1|$, then the area of S^{D-1} is smaller than that of $\partial\Omega$, which gives the name of the isopreimetric inequality when $D = 2$.

Proof. For every $\varepsilon > 0$, by the Brunn-Minkowski inequality, we have

$$|\Omega + \varepsilon B_1| \geq \left(|\Omega|^{\frac{1}{D}} + |\varepsilon B_1|^{\frac{1}{D}}\right)^D = \left(|\Omega|^{\frac{1}{D}} + \varepsilon |B_1|^{\frac{1}{D}}\right)^D$$

so that

$$\begin{aligned} A(\partial\Omega) &= \liminf_{\varepsilon\downarrow 0} \frac{|\Omega + \varepsilon B_1| - |\Omega|}{\varepsilon} \\ &\geq \lim_{\varepsilon\to 0} \frac{\left(|\Omega|^{\frac{1}{D}} + \varepsilon |B_1|^{\frac{1}{D}}\right)^D - |\Omega|}{\varepsilon} \\ &= D|\Omega|^{1-\frac{1}{D}}|B_1|^{\frac{1}{D}} \\ &= \frac{A(S^{D-1})}{|B_1|^{1-\frac{1}{D}}}|\Omega|^{1-\frac{1}{D}} \end{aligned}$$

and the proof is complete. □

By an elementary computation, we know that the area of the Euclidean unit sphere S^{D-1} in $\mathbb{R}^D$ equals $\frac{2\pi^{D/2}}{\Gamma(D/2)}$, where $\Gamma(1/2) = \sqrt{\pi}$, and therefore the volume of the unit ball B_1 in $\mathbb{R}^D$ is $\frac{1}{D}A(S^{D-1}) = \frac{1}{D}\frac{2\pi^{D/2}}{\Gamma(D/2)}$. If $D = 2$, then the isoperimetric inequality becomes

$$\frac{A(\partial\Omega)}{\sqrt{|\Omega|}} \geq 2\sqrt{\pi}$$

so that

$$L^2 - 4\pi A \geq 0$$

where L and A are the length of the perimeter and the area of a region $\Omega \subset \mathbb{R}^2$.

7 Appendix

In this part we collect several facts about properties of matrices, which are useful in dealing with high-dimensional datasets.

7.1 Analysis of Lebesgue's measure

[*This part brings together a few useful facts in Analysis, which can be considered as a general background or general knowledge. These facts can be obtained by using what you learned in Prelim Analysis and Lebesgue's Integration Theory (A4 Paper). The lecturer shall not present this part in lectures, rather you may refer back when we need them through the course.*]

The Lebesgue measure Leb on the Euclidean space $\mathbb{R}^D$ is the unique measure on $(\mathbb{R}^D, \mathscr{B}(\mathbb{R}^D))$ such that

$$\text{Leb}\left((a_1, b_1] \times \cdots \times (a_D, b_D]\right) = (b_1 - a_1)\cdots(b_D - a_D)$$

for any $a_i \leq b_i$ ($i = 1, \ldots, D$), where $\mathscr{B}(\mathbb{R}^D)$ is the Borel σ-algebra on $\mathbb{R}^D$, the smallest σ-algebra containing all open (hence as well closed) subsets. It is the D-fold product measure of one dimensional Lebesgue measure. The integral of a Borel measurable function f against the Lebesgue measure may be written as $\int_{\mathbb{R}^D} f(x)\mathrm{d}x$. In applications to datasets, the dimension D is rather large, and therefore it is not practical to evaluate an integral such as $\int_{\mathbb{R}^D} f(x)\mathrm{d}x$ unless for very simple functions. Therefore the density properties of nice functions in L^p-spaces are very important, which we shall review now.

If $R \subset \mathbb{R}^D$ is a Borel subset, then the Lebesgue restricted on $\mathscr{B}(R)$ is a called the Lebesgue measure on $(R, \mathscr{B}(R))$, and we shall use $L^p(R)$ to denote the L^p-space $L^p(R, \mathscr{B}(R), \text{Leb})$ for simplicity, and the L^p-norm of a function f on R may be denoted by $\|f\|_p$ or $\|f\|_{L^p(R)}$ if no confusion may arise.

7.1.1 Density property

Let $\Omega \subset \mathbb{R}^D$ be an open subset.

A Borel measurable function f is locally L^p-integrable on Ω, denoted by $f \in L^p_{\text{loc}}(\Omega)$, if for every $x \in \Omega$, there is a ball $B(x,r)$ centered at x with $r > 0$ such that $B(x,r) \subset \Omega$ and $\int_{B(x,r)} |f|^p(y)\mathrm{d}y < \infty$. Clearly $f \in L^p_{\text{loc}}(\Omega)$ if and only if f is Borel measurable and $\int_K |f|^p(y)\mathrm{d}y < \infty$ for every compact subset $K \subset \Omega$.

If m is an integer, then $C^m(\Omega)$ denote the linear space of all functions with continuous partial derivatives up to m-order, and $C^\infty(\Omega) = \bigcap_{m\geq 1} C^m(\Omega)$. Recall that if f is a function on Ω, then the closure of $\{f \neq 0\}$ is called the support of f, denoted by $\text{supp}(f)$. A function $\varphi \in C^\infty(\Omega)$ is called a *test function* on Ω, if its support $\text{supp}(\varphi)$ is a compact subset of Ω, i.e. $\text{supp}(\varphi) \subset \Omega$ and $\text{supp}(f)$ is compact. The linear space of all test functions on Ω shall be denoted by $C^\infty_C(\Omega)$.

Example 7.1. *The function $\varphi(x) = \exp(1/(|x|^2 - 1))$ for $|x| < 1$ and $\varphi(x) = 0$ for $|x| \geq 1$ belongs to $C^\infty_C(\mathbb{R}^d)$, whose support $\text{supp}(\varphi)$ is the closed unit ball at 0. φ is non-negative.*

Definition 7.2. *A non-negative function $\alpha \in C^\infty_C(\mathbb{R}^D)$ with $\int_{\mathbb{R}^D} \alpha(x)dx = 1$ is called a smoothing function on $\mathbb{R}^D$.*

Given a smoothing function α on $\mathbb{R}^D$, with a compact support $\text{supp}(\alpha)$ inside the closed unit ball centered at 0, define $\alpha_\varepsilon(x) = \varepsilon^{-D}\alpha(x/\varepsilon)$ for $x \in \mathbb{R}^D$, for every $\varepsilon > 0$. Then α_ε is a smoothing function too, and $\text{supp}(\alpha_\varepsilon) \subset \{x : |x| \leq \varepsilon\}$ for every $\varepsilon > 0$. If f is local integrable, then

$$f_\varepsilon(x) = \int_{\mathbb{R}^D} f(x-y)\alpha_\varepsilon(y)\mathrm{d}y = \int_{\mathbb{R}^D} f(y)\alpha_\varepsilon(x-y)\mathrm{d}y$$

(for every $x \in \mathbb{R}^D$) is well defined for every $\varepsilon > 0$, which is called the convolution f and α_ε, denoted by $f_\varepsilon = f \star \alpha_\varepsilon$. Then, by using differentiation under integral, justified by Theorem 2.13, we have the following simple facts:

1) $f_\varepsilon \in C^\infty(\mathbb{R}^D)$ and $\frac{\partial}{\partial x^\beta} f(x) = \int_{\mathbb{R}^D} f(y) \frac{\partial}{\partial x^\beta} \alpha_\varepsilon(x-y)\mathrm{d}y$ for every $\varepsilon > 0$ for any any indices $\beta = (\beta_1, \cdots, \beta_D)$.

2) If f has a compact support, then $\text{suup}(f_\varepsilon) \subset \text{supp}(f)_\varepsilon$ for every $\varepsilon > 0$, where $A_\varepsilon = \{x : d(x,A) \leq \varepsilon\}$, where $d(x,A) = \inf\{|x-y| : y \in A\}$.

3) If f is continuous on $\mathbb{R}^D$, then $f_\varepsilon \to f$ as $\varepsilon \downarrow 0$ uniformly on every compact subset (hence on any bounded subset) $K \subset \mathbb{R}^D$.

4) If $f \geq 0$ then $f_\varepsilon \geq 0$ for every $\varepsilon > 0$. Similarly, if $f \leq C$ for some constant, then $f_\varepsilon \leq C$ for every $\varepsilon > 0$. Hence if $f \in L^\infty(\mathbb{R}^D)$, then $f_\varepsilon \in C_b(\mathbb{R}^D)$ the space of all bounded and continuous functions on $\mathbb{R}^D$, and $\|f_\varepsilon\|_\infty \leq \|f\|_\infty$ for each $\varepsilon > 0$.

We show that the function space $C_C^\infty(\mathbb{R}^D)$ is dense in any $L^p(\mathbb{R}^D)$ space, and therefor $C_C^\infty(\Omega)$ is dense in $L^p(\Omega)$ for every $p \in [1,\infty]$. To this end we begin with the following fact.

Lemma 7.3. *Any continuous function on a closed subset of $\mathbb{R}^D$ can be extend to be a continuous function on $\mathbb{R}^D$.*

Lemma 7.4. *(*Lusian Theorem*) If f is a Borel measurable on $\mathbb{R}^d$, then for every $\delta > 0$, there is a closed subset F such that 1) Leb$(F^c) < \delta$, and 2) f is continuous on F.*

Proof. For every $k = 1,2,\cdots$ and every integer $n \in \mathbb{Z}$, set

$$E_{n,k} = \left\{ \frac{n}{k} \le f < \frac{n+1}{k} \right\}.$$

Then for every k, $E_{n,k}$ are disjoint and $\bigcup_n E_{n,k} = \mathbb{R}^D$, and therefore there is an positive integer n_k such that

$$\mathrm{Leb}\left[\left(\bigcup_{|n| \le n_k} E_{n,k} \right)^c \right] < \frac{\delta}{2^{k+1}}.$$

We then for each $n = 0, \pm 1, \cdots, \pm n_k$, choose a closed subset $F_{n,k} \subset E_{n,k}$ such that

$$\sum_{|n| \le n_k} \mathrm{Leb}(E_{n,k} \setminus F_{n,k}) < \frac{\delta}{2^{k+1}}$$

and therefore

$$\mathrm{Leb}\left[\left(\bigcup_{|n| \le n_k} F_{n,k} \right)^c \right] < \frac{\delta}{2^k}$$

for every $k = 1,2,\cdots$. Let $F_k = \bigcup_{|n| \le n_k} F_{n,k}$ which is closed, where $F_{n,k}$ are disjoint closed subset. Define f_k on F_k by $f_k(x) = \frac{n}{k}$ if $x \in F_{n,k}$. Then f_k are continuous, and $|f(x) - f_k(x)| \le \frac{1}{k}$ for every $x \in F_k$. f_k is continuous, so is continuous on $F = \bigcap_{k=1}^\infty F_k$ and $f_k \to f$ uniformly on F. Therefore f is continuous on F. □

Corollary 7.5. *If f is Borel measurable on $\mathbb{R}^D$, then for every $\delta > 0$, there is a continuous function g on $\mathbb{R}^D$ such that Leb$\{f \neq g\} < \delta$.*

Theorem 7.6. *Let $p \in [1,\infty)$. If $f \in L^p(\mathbb{R}^D)$, then $f_\varepsilon \in L^p(\mathbb{R}^D)$ for every $\varepsilon > 0$, and $f_\varepsilon \to f$ in $L^p(\mathbb{R}^D)$. Therefore $C_C^\infty(\mathbb{R}^D)$ is dense in $L^p(\mathbb{R}^D)$ for every $p \in [1,\infty)$.*

Proof. First note that for every $\varepsilon > 0$, $\alpha_\varepsilon(y)\mathrm{d}y$ is a probability measure, so that, using the Fubini theorem

$$\begin{aligned} \|f_\varepsilon\|_p^p &= \int_{\mathbb{R}^D} \left| \int_{\mathbb{R}^D} f(x-y)\alpha_\varepsilon(y)\mathrm{d}y \right|^p \mathrm{d}x \\ &\le \int_{\mathbb{R}^D} \int_{\mathbb{R}^D} |f(x-y)|^p \,\mathrm{d}x\alpha_\varepsilon(y)\mathrm{d}y \\ &= \int_{\mathbb{R}^D} \int_{\mathbb{R}^D} |f(x)|^p \,\mathrm{d}x\alpha_\varepsilon(y)\mathrm{d}y = \|f\|_p^p \end{aligned}$$

which implies that $f_\varepsilon \in L^p$.

We may assume that f is bounded. For every $\delta > 0$, according to the previous corollary there is a bounded continuous function g such that $\mathrm{Leb}\{f \neq g\} < \frac{\delta}{(2\|f\|_\infty+1)^p}$. Then

$$\begin{aligned}
\|f_\varepsilon - g_\varepsilon\|_p^p &= \int_{\mathbb{R}^D} \left| \int_{\mathbb{R}^D} (f(x)-g(x))\, \alpha_\varepsilon(x-y)\mathrm{d}y \right|^p \mathrm{d}x \\
&= \int_{\mathbb{R}^D}\int_{\mathbb{R}^D} |f(x)-g(x)|^p \alpha_\varepsilon(x-y)\mathrm{d}y\mathrm{d}x \\
&= \int_{\mathbb{R}^D} |f(x)-g(x)|^p \mathrm{d}x \leq \delta.
\end{aligned}$$

Therefore we may assume that f is continuous with a compact support, so that $f_\varepsilon \in C_C^\infty(\mathbb{R}^D)$ for all $\varepsilon > 0$, and

$$\begin{aligned}
\|f_\varepsilon - f\|_p^p &= \int_{\mathbb{R}^D} \left| \int_{\mathbb{R}^D} (f(x-y)-f(x))\, \alpha_\varepsilon(y)\mathrm{d}y \right|^p \mathrm{d}x \\
&= \int_{\mathbb{R}^D} \left| \int_{\mathbb{R}^D} (f(x-\varepsilon y)-f(x))\, \alpha(y)\mathrm{d}y \right|^p \mathrm{d}x \\
&\leq \int_{\mathbb{R}^D} \left| \int_{\mathbb{R}^D} (f(x-\varepsilon y)-f(x)) \right|^p \mathrm{d}x\alpha(y)\mathrm{d}y.
\end{aligned}$$

Since

$$\left| \int_{\mathbb{R}^D} (f(x-\varepsilon y)-f(x)) \right|^p \mathrm{d}x \leq 2^p \|f\|_p^p$$

and

$$\left| \int_{\mathbb{R}^D} (f(x-\varepsilon y)-f(x)) \right|^p \mathrm{d}x \to 0 \quad \text{as } \varepsilon \downarrow 0.$$

Therefore by Dominated Convergence Theorem,

$$\int_{\mathbb{R}^D} \left| \int_{\mathbb{R}^D} (f(x-\varepsilon y)-f(x)) \right|^p \mathrm{d}x\alpha(y)\mathrm{d}y \to 0 \quad \text{as } \varepsilon \downarrow 0$$

which yields that $\|f_\varepsilon - f\|_p \to 0$ as $\varepsilon \downarrow 0$. The proof is complete. □

7.1.2 Generalized derivatives, Sobolev spaces

Definition 7.7. *Let $\Omega \subset \mathbb{R}^D$ be an open set, and $f \in L^1_{loc}(\Omega)$. We say a locally integrable function f_i is the generalized partial derivative with respect to x^i (for $i = 1, \cdots, D$), if*

$$\int_\Omega f_i(x)\varphi(x)dx = -\int_\Omega f(x)\frac{\partial}{\partial x^i}\varphi(x)dx, \quad \text{for any } \varphi \in C_C^\infty(\Omega).$$

In this case f_i is denoted by $\frac{\partial}{\partial x^i}f$ (called the generalized derivative of f), if no confusion may arise, and we say the generalized derivative $\frac{\partial}{\partial x^i}f$ is locally integrable. If in addition, $\frac{\partial}{\partial x^i}f$ is p-th integrable (where $p \geq 1$), then we say the generalized derivative $\frac{\partial}{\partial x^i}f \in L^p(\Omega)$. The generalized gradient of f is defined to be naturally as $(\frac{\partial}{\partial x^1}f, \cdots, \frac{\partial}{\partial x^D}f)$, denoted by ∇f, and we say the generalized gradient ∇f is locally integrable (resp. belongs to the L^p-space).

Remark 7.8. *1) Generalized derivatives of a locally integrable function f always exist as generalized functions (i.e. distributions). The proper treatment of this approach requires certain preparation and therefore we do not give a general definition of generalized functions in this book, the reader may refer to K. Yosida: Functional Analysis.*

2) The locally integrable function f_i in the definition, if exists, then it is unique up to almost surely. We often say $\frac{\partial}{\partial x^i} f = f_i$ in the sense of distribution in this case.

This definition can be generalized to higher order generalized derivatives, which we shall not discussed further, the reader may refer to standard textbooks such as .

Definition 7.9. *Let f be a locally integrable function on an open subset $\Omega \subset \mathbb{R}^D$. Then we say $f \in H^1(\Omega)$ (some authors use $W^{1,2}(\Omega)$ instead), if both f and its generalized gradient ∇f belong to the $L^2(\Omega)$, that is, $f \in L^2(\Omega)$ and the generalized derivatives $\frac{\partial}{\partial x^i} f \in L^2(\Omega)$. For $f \in H^1(\Omega)$, its Sobolev norm is defined to be*

$$\|f\|_{H^1(\Omega)} = \sqrt{\int_\Omega |f|^2(x)dx + \sum_{i=1}^D \left|\frac{\partial}{\partial x^i} f(x)\right|^2 dx.}$$

Theorem 7.10. *1) $H^1(\Omega)$ is a complete metric space under the distanced defined by the norm $\|\cdot\|_{H^1(\Omega)}$.*

2) $C^\infty(\Omega)$ is dense in $H^1(\Omega)$ under the $\|\cdot\|_{H^1(\Omega)}$-norm distance.

The proof is left as exercise, see

Definition 7.11. *Define $H_0^1(\Omega)$ to be the closure of $C_C^\infty(\Omega)$ under the the $\|\cdot\|_{H^1(\Omega)}$-norm distance.*

These results can be generalized to measures which are absolutely continuous with respect to the Lebesgue measure. Suppose $\mu(\mathrm{d}x)$ is a σ-finite measure on $(\mathbb{R}^D, \mathscr{B}(\mathbb{R}^D))$ and $\mu(\mathrm{d}x)$ is absolutely continuous with the Lebesgue measure on $\mathbb{R}^d$. That is, there is a non-negative Borel measurable, locally integarable function $\rho(x)$ such that $\mu(\mathrm{d}x) = \rho(x)\mathrm{d}x$. For simplicity the L^p-space over $(\mathbb{R}^D, \mathscr{B}(\mathbb{R}^D), \mu)$ is denoted by $L^p(\mu)$. Similarly we shall use $H^1(\mu)$ to denote the space of all locally integrable functions f, such that both f and its generalized gradient ∇f belong to $L^2(\mu)$, equipped with the norm

$$\|f\|_{H^1(\mu)} = \sqrt{\|f\|^2_{L^2(\mu)} + \|\nabla f\|^2_{L^2(\mu)}}.$$

Then $H^1(\mu)$ is a Banach space, and $C_C^\infty(\mathbb{R}^D)$ is dense in $H^1(\mu)$.

7.1.3 Lipschitz functions

Finally we shall recall several elementary facts about Lipschitz functions. Recall that a function f on $\mathbb{R}^D$ is Lipschitz, if $|f(x) - f(y)| \le C|x-y|$ for every $x, y \in \mathbb{R}^D$, where $C \ge 0$ is a constant. The least C is called the Lipschitz norm of f, denoted by $\|f\|_{\mathrm{Lip}}$. That is

$$\|f\|_{\mathrm{Lip}} = \sup_{x \ne y} \frac{|f(x) - f(y)|}{|x-y|}.$$

Let α be a non-negative smoothing function and $\alpha_\varepsilon(x) = \varepsilon^{-D}\alpha(x/\varepsilon)$ and $f_\varepsilon = f \star \alpha_\varepsilon$ for every $\varepsilon > 0$.

Lemma 7.12. *Let $f : \mathbb{R}^D \mapsto \mathbb{R}$ be Lipschitz continuous (with respect to the standard metric on $\mathbb{R}^D$ and $\mathbb{R}$. Then $f_\varepsilon \to f$ as $\varepsilon \downarrow 0$ uniformly on any bounded subset, and $\|\nabla f_\varepsilon\|_\infty \le \|f\|_{Lip}$ for every $\varepsilon > 0$.*

Proof. $f_\varepsilon \in C^\infty(\mathbb{R}^D)$ for every $\varepsilon > 0$. Since f is continuous, so that $f_\varepsilon \to f$ uniformly on any bounded subset of $\mathbb{R}^D$. Since

$$f_\varepsilon(x+ha) - f_\varepsilon(x) = \int_{\mathbb{R}^D} (f(x+ha-y) - f(x-y))\alpha_\varepsilon(y)\mathrm{d}y$$

for every x and $h \neq 0$, and f is Lipschitz continuous, so that

$$\begin{aligned} \left| \frac{f_\varepsilon(x+ha) - f_\varepsilon(x)}{h} \right| &\le \|f\|_{\mathrm{Lip}} \int_{\mathbb{R}^D} |a|\alpha_\varepsilon(y)\mathrm{d}y \\ &= \|f\|_{\mathrm{Lip}} |a| \end{aligned}$$

for every $\varepsilon > 0$ and $h \neq 0$. Letting $h \to 0$, we then obtain that

$$|\nabla f_\varepsilon(x) \cdot a| \le \|f\|_{\mathrm{Lip}} |a|$$

for every $a \in \mathbb{R}^D$, which yields that $|\nabla f_\varepsilon(x)| \le \|f\|_{\mathrm{Lip}}$ for every x and for every $\varepsilon > 0$. □

7.2 Inverting a square matrix

Let $A = (a_{ij})$ be an $n \times n$ square matrix. Then its determinant

$$|A| = \det A = \sum_{\sigma \in S_n} (-1)^\sigma a_{1\sigma_1} \cdots a_{n\sigma_n}$$

where σ runs over the permutation group S_n of $\{1, \cdots, n\}$, and also $\sigma = 0$ or 1 according to the parity of the arrangement $\sigma = \{\sigma_1, \ldots, \sigma_n\}$.

For every pair (i, j), $\Lambda_{ij} = (-1)^{i+j}$ times the determinant of the $(n-1) \times (n-1)$-square matrix with the i-th row, j-th column delated. Then

$$\det A = \sum_{i=1}^n a_{ij}\Lambda_{ij} = \sum_{j=1}^n a_{ij}\Lambda_{ij}$$

(for every j, resp. for every i). It is known that A is invertible if and only if $\det A \neq 0$. In this case the inverse of A, denoted by A^{-1}, is given by

$$A^{-1} = \frac{1}{\det A}(\Lambda_{ij})^T,$$

where T labels the transport.

Suppose we write a square matrix A in blocks:

$$A = \begin{pmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{pmatrix}$$

where A_{11} and A_{22} are square matrices (but not necessary having the same rank).

1) Suppose A_{11} is invertible, then

$$\begin{pmatrix} I & 0 \\ -A_{21}A_{11}^{-1} & I \end{pmatrix} \begin{pmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{pmatrix} \begin{pmatrix} I & -A_{11}^{-1}A_{12} \\ 0 & I \end{pmatrix} = \begin{pmatrix} A_{11} & 0 \\ 0 & A_{22} - A_{21}A_{11}^{-1}A_{12} \end{pmatrix}.$$

2) Suppose both A and A_{11} are invertible, then

$$A^{-1} = \begin{pmatrix} A_{11}^{-1}\left(I + A_{12}B^{-1}A_{21}A_{11}^{-1}\right) & -A_{11}^{-1}A_{12}B^{-1} \\ -B^{-1}A_{21}A_{11}^{-1} & B^{-1} \end{pmatrix}$$

where $B = A_{22} - A_{21}A_{11}^{-1}A_{12}$.

3) If A_{11} is invertible, then

$$\det \begin{pmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{pmatrix} = \det A_{11} \det\left(A_{22} - A_{21}A_{11}^{-1}A_{12}\right)$$

and, similarly, if A_{22} is invertible,

$$\det \begin{pmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{pmatrix} = \det A_{22} \det\left(A_{11} - A_{12}A_{22}^{-1}A_{21}\right).$$

Lemma 7.13. *Suppose A and B are two square matrices, then the non-zero eigenvalues of AB and BA are the same with the same multiplicity. In particular, $tr(AB) = tr(BA)$.*

References

[1] R. J. Adler: *An introduction to continuity , extrema, and related topics for general Gaussian processes*. Inst. of Math. stat. Lect. Notes, Monograph Ser., 12. Inst. of Math. Stat., Hayward, CA, 1990.

[2] Azencott, R. (1980). Grandes déviations et applications, *Ecoles s'Été de Probabilités de Saint-Flou*r VIII-1978. Lecture Notes in Math. 774, Springer, pp. 1-176 (1980). English translation is available in "*Large Deviations at Saint-Flour*", R. Azencott, M. I. Freidlin and S. R. S. Varadhan, Springer (2013).

[3] D. Bakry and M. Ledoux: Lévy-Gromov's isoperimetric inequality for an infinite dimensional diffusion generator. *Invent. math.* 123, 259-281 (1996).

[4] V. I. Bogachev: *Gaussian Measures*. Mathematical Surveys and Monographs Volume 62. American Math. Society (1998).

[5] S. Boucheron, G. Lugosi and P. Massart: *Concentration Inequalities - A nonasymptotic theory of independence*. Oxford University Press (2013).

[6] H. Brascamp and E. H. Lieb: On extensions of the Brunn-Minkowski and Prékopa-Leindler theorems, including inequalities for log concave functions, and with an application to the diffusion equation. *J. Funct. Anal.* 22 (1976), 366-389.

[7] Cramér, H. (1938). Sur un nouveau théorème-limite de la théorie des probabilités, *Actualités Scientifiques et Industrielles* 736 (1938), 5-23, *Colloque consacré à la théorie des probabilités*, Vol. 3, Hermann, Paris.

[8] Ethier S. N. and T. G. Kurtz (1986). *Markov Processes – Characterization and Convergence*. John Wiley & Sons (1986).

[9] P. R. Halmos: Measure Theory. Springer (1974). Originally published by Litton Educational Publishing, Inc. (1950).

[10] Stroock, D. W and Varadhan, S. R. S. (1979). *Multidimensional Diffusion Processes*, Springer (1979, 1997).

[11] Yosida, K: *Functional Analysis*. 6th Edition. Springer-Verlag.