



INSTITUTO DE MATEMÁTICA

Universidade Federal do Rio de Janeiro



UFRJ

On Log-Sobolev Inequalities and their Applications

Patrick Oliveira Santos

Rio de Janeiro, Brasil

August 25, 2020

On Log-Sobolev Inequalities and their Applications

Patrick Oliveira Santos

Dissertação de mestrado apresentada ao Programa de Pós-graduação em Matemática do Instituto de Matemática da Universidade Federal do Rio de Janeiro, como parte dos requisitos necessários à obtenção do título de Mestre em Matemática.

Universidade Federal do Rio de Janeiro

Instituto de Matemática

Programa de Pós-Graduação em Matemática

Supervisor: César Javier Niche Mazzeo

Rio de Janeiro, Brasil

August 25, 2020

CIP - Catalogação na Publicação

0058o

Oliveira Santos, Patrick
On Log-Sobolev Inequalities and their
Applications / Patrick Oliveira Santos. -- Rio de
Janeiro, 2020.
221 f.

Orientador: César Javier Niche Mazzeo.
Dissertação (mestrado) - Universidade Federal do
Rio de Janeiro, Instituto de Matemática, Programa
de Pós-Graduação em Matemática, 2020.

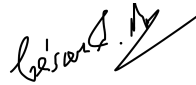
1. Concentração de Medida. 2. Desigualdades Log
Sobolev. 3. Entropia. 4. Teoria da Informação. I.
Niche Mazzeo, César Javier, orient. II. Título.

Patrick Oliveira Santos

On Log-Sobolev Inequalities and their Applications

Dissertação de mestrado apresentada ao Programa de Pós-graduação em Matemática do Instituto de Matemática da Universidade Federal do Rio de Janeiro, como parte dos requisitos necessários à obtenção do título de Mestre em Matemática.

Trabalho aprovado por



Prof. César Javier Niche Mazzeo
Doutor - IM/UFRJ, Presidente



Prof. Guilherme Ost de Aguiar
Doutor - IM/UFRJ



Prof. Olivier Guédon
Doutor - Université Gustave Eiffel



Dr. Pierre Youssef
Doutor - Université Paris Diderot



Prof. Roberto Imbuzeiro Moraes
Felinto de Oliveira
Doutor - IMPA

Rio de Janeiro, Brasil
Julho de 2020

*This work is dedicated to my sister:
she is my everything.*

Acknowledgements

The content of this part is written in Portuguese, my native language.

Queria começar agradecendo especialmente a minha família. Aos meus pais Ulisses Souza dos Santos e Janaina dos Santos Oliveira Santos: vocês são fonte de inspiração e amor constante em minha vida. Se hoje cheguei até onde cheguei, boa parte se deve ao empenho de vocês em me manter nesse caminho. Também estendo esses agradecimentos aos meus avós Ulisses Barnabé dos Santos e Neyde Souza dos Santos. Se eu me mantive firme durante essa morada no Rio de Janeiro, muito se deve ao carinho de vocês. Quero também agradecer aos meus primos Danielle Menezes, Milleny Menezes e Roberto Pereira por sempre estarem dispostos a me ajudar, seja consertando o carro, seja com algum problema de internet, seja com remédios. Por fim, Bruna.... não chegou sua hora ainda.

Para além da minha família, quero agradecer aos meus amigos que conheci no CELD ou numa vizinhança do CELD. Aqui cito nomes como Larissa Corrêa, Paulo Victor, Anderson Santos, Jessica Assis, Pietro Bollo, Jonathan Assis, Camila Bilbao, Lucas Baldomero, Erika Beatriz, entre muitos outros. A presença de vocês sempre foi motivo de muita alegria. Cito, em especial, a Beatriz Fraga e a Milenna Mesquita, que presente é tê-las comigo toda semana. Amo vocês!

Mas não só de CELD vive o homem, então aqui agradeço também a família Oficina que me acolheu. Cito alguns nomes, mas os agradecimentos vão a toda família: Marcelo Manga, Adriana Pimentel, Ricardo Leite, Andrea Moraes, Leandro Turano, Tami Lima, Juliana Oliveira, Tuca Alves, Allan Valdivia... entre muitos outros. Sempre gostei de música, mas aprender e expandir os horizontes do conhecimento técnico e moral na música todo sábado (e demais dias) são alegrias que guardo comigo.

Mas também não posso deixar de citar aos amigos da faculdade, aqueles que fizeram o Bob. São muitos nomes, então vamos lá: Bruno Lima, Cynthia Herkenhoff, Karolayne Dessabato, Ricardo Turano, Gabriel Picanço, Ivani Ivanova, Leonardo Gama,

Marcelo Carneiro, Maria Luiza Avlis, Matheus Fontoura, Rodrigo Lima, Thiago Holleben, Ramiz Oliveira, Israel Tuponi, Thayz Ferreira, Giovanna Rieken e Artur Souza. Um agradecimento especial ao Iago Leal, Pedro Aragão, Alexandre Moreira e Jéssica Richard Nascimento (Sirina) por serem maravilhosas pessoas que sempre me ajudaram, mas, acima de tudo, por rirem de minhas piadas. Além dos citados, um especial agradecimento a Gabriela Lewenfus, que querendo ou não, estava comigo durante boa parte da minha graduação e sempre foi fonte de inspiração, além de motivação para melhorar minhas piadas, digo, estudos.

Aqui, também quero agradecer aos amigos que atravessaram a flecha do tempo e estão comigo desde muito tempo. A liga... Da Justica! Matheus Ribeiro e Julio Rama. Palavras não descreveriam o que vocês são pra mim e tudo que eu queria agradecer. Amo vocês e espero que nossa amizade continue até o fim, até porque não há como continuar depois do fim e só acabará no fim mesmo...

Para muito além do Rio de Janeiro, quero agradecer aos seres que comigo estão desde Boa Vista. Vocês são parte da minha infância, mas muito mais do que isso, são pessoas que me fizeram ser quem sou. Aqui cito: Íngrid Suéllen, Beatriz Belo, Tháilla Jasminie, Emily Pinheiro, Amanda Ikuta, Halaine Pessoa, Tulio Marroquim, Gabriel Fin, entre muitos outros. Aos companheiros de jogos: Bruno Silva Zardo e Cheyenne Oliveira. Que haja muito jogos entre a gente ainda, diversão e companherismo! Entretanto, tenho que enfatizar dois nomes. Primeiramente: Natália Araújo Carim. Você é luz. Saiba sempre disso. Amo você. E também Arthur Philipe (que se fala Cleyton) Barbosa Almeida. Você será eternamente meu melhor amigo, fonte de motivação e piadas.

Quero agradecer também as instituições que investiram em mim e na minha jornada. Anteriormente a universidade, cito: Instituto Batista de Roraima, Acadêmico News, Sistema Elite de Ensino. Em nível acadêmico, meu agradecimento é à UFRJ, que proporcionou muitos conhecimentos e experiências. Em parapeço, agradeço a CAPES e a FAPERJ por investirem em mim por meio de bolsas de estudo.

Além disso, agradeço também aos meus professores que marcaram minha jornada e me proporcionaram inúmeros conhecimentos. Cito aqui, dentre muitos, os seguintes nomes: Sérgio Lima, Katrin Gelfert e Leandro Pimentel. Com certeza vocês poderão achar em minha dissertação um pouco do que eu aprendi com suas aulas!

Meus sinceros agradecimentos aos membros da banca professores Guilherme Ost, Roberto Imbuzeiro, Pierre Youssef e Olivier Guédon. Obrigado por disporem de tempo para ler e comentar essa dissertação.

Agradeço ao meu orientador César Niche por sempre ter estado presente nesses momentos de dificuldade. Seja pelas reuniões sobre as mais diversas entropias, seja por toda ajuda com documentos incansáveis, o senhor se fez um ponto de suporte em meio a

todos os problemas que enfrentei durante a graduação e mestrado. Além disso, vale citar nossas reuniões durante a pandemia, que sem dúvida me iluminaram o caminho e me ajudaram a manter a disciplina em tempos de crise. Obrigado por tudo.

Por fim, cito minha irmã Bruna Oliveira Santos, cuja importância seja imensurável. Entre ciclos e ciclos de 30 dias, você se fez igualmente uma inspiração e um ponto de suporte. Nossas personalidades Yin e Yang serão, para sempre, um marco de irmandade e companherismo, além de sermos irmãos ótimos, não é? Agradeço por todas as trocas que tivemos, seja por me mandar fazer coisas que eu não quero (por que burocracia existe?), seja por pedir para assistirmos filmes ruins (às vezes bons). Mana, você é realmente um espetáculo de pessoa, com um coração que não cabe em ti e eu agradeço por me acolher do jeito que sou, embora esporadicamente não nos entendamos. Não sei me expressar às vezes, mas quero que fique claro que não consigo imaginar minha vida antes de você (até porque você é mais velha...) e sem você. Obrigado por tudo. Te amo.

Resumo

O presente trabalho é dedicado ao estudo e entendimento das desigualdades Log-Sobolev no Cubo de Hamming e no Espaço de Gauss. Algumas ferramentas serão estudadas, como semigrupos de operadores e desigualdades em Teoria da Informação, que nos permitirão obter os corolários desejados. Com isso, iremos abordar algumas aplicações importantes, tais como o fenômeno de concentração de medida em ambos os espaços, as complexidades de Rademacher e Gauss e suas consequências e a conexão entre as desigualdades Log-Sobolev com Teoria da Informação.

Palavras-chave: Desigualdades Log-Sobolev, Concentração de Medida, Entropia, Teoria da Informação.

Abstract

The present work is dedicated to study and understanding the Log-Sobolev Inequalities in the Hamming Cube and Gauss Space. Some tools are going to be studied, such as semigroup of operators and inequalities in Information Theory, that will allow us to obtain the desired corollaries. Thereby, we will address some important applications, such as the concentration of measure phenomenon in both spaces, the Rademacher and Gauss Complexities and their consequences and the connection between the Log-Sobolev Inequalities and Information Theory.

Keywords: Log-Sobolev Inequalities, Concentration of Measure, Entropy, Information Theory.

List of Figures

Figure 1	– A convex function in the interval $[0, 2]$, in blue, and a straight line connecting the points $(1, 1)$ and $(2, 8)$, in red.	58
Figure 2	– The function e^x and the tangent at $x = 0$, namely, $y = x + 1$	59
Figure 3	– Representative Veen Diagram with two circles: the first one is concerning $H(X)$ and the second $H(Y)$. Notice that $H(X, Y) = H(X) + H(Y X)$, $H(X, Y) \leq H(X) + H(Y)$ and equality holds if and only if $I(X, Y) = 0$, that is, they are independent.	93
Figure 4	– Diagram representing the transmission. A message W is encoded in $f(W) \in \mathcal{X}^n$. The channel transforms this input into a noisy sign $Y \in \mathcal{Y}^n$ and the decode g guesses the best candidate $\hat{W} = g(Y)$ for the original message.	107
Figure 5	– Given the input $x \in \{0, 1\}$, the output is x with probability $1 - p$ and $1 - x$ with probability p	109
Figure 6	– The graph of the function $c(p)$	173
Figure 7	– The graph of the function $r(p)$	176

List of Tables

Table 1	– Table of a code for the set \mathcal{X} .	89
Table 2	– Table of a code for the set \mathcal{X}^2 .	89

Contents

1	INTRODUCTION	23
2	ON DICE AND COINS	27
2.1	Introduction	27
2.2	Probability Spaces	28
2.2.1	Carathéodory's Theorem	30
2.2.2	Borel, Lebesgue and Kolmogorov	31
2.3	Random Variables and Random Vectors	32
2.3.1	Random Variables and Distribution	33
2.3.2	Independence	35
2.4	Integral and Expected Value	37
2.4.1	The Three Steps	37
2.4.2	Radon-Nikodym Theorem	41
2.4.3	Product Measure and Fubini's Theorem	44
2.5	Computing Integrals	46
2.5.1	Riemann Integral	46
2.5.2	Change of Variables	48
2.5.3	The Weak Derivative	49
2.6	The Fourier Transform and Moments	51
2.6.1	The Convolution Rule	51
2.6.2	Moments	52
2.6.3	The Generating Function	53
2.6.4	The Fourier Transform and The Characteristic Function	55
2.7	Inequalities in Probability	57
2.7.1	Convex Function and Jensen Inequality	57
2.7.2	Markov's Inequality	63
2.7.3	Chernoff's Inequality	64
2.7.4	Inequalities in Hilbert Space	66
2.8	Conditional Expectation	66
2.9	Notions of Convergence and Laws of Large Numbers	71
2.9.1	Weak Law and Convergence in Probability	71
2.9.2	Almost Surely Convergence and Strong Law	72
2.9.3	Convergence in Distribution and Central Limit Theorem	73
2.10	Markov Chains	74
2.10.1	Discrete Time and Countable State Space	74

2.10.2	Continuous Time and Countable State Space	76
2.10.3	Uncountable State Space	78
	3 INFORMATION AND ITS MYSTERIES	81
3.1	Introduction	81
3.2	Shannon Entropy	82
3.3	Compression and Codes	88
3.4	Differential Entropy and Information	94
3.4.1	Differential Entropy of Shannon	94
3.4.2	Maximum Entropy	95
3.4.3	Exponential Entropy of Shannon	98
3.4.4	Fisher Information according to a parameter	99
3.4.5	Fisher Information	102
3.4.6	Fisher Matrix	104
3.4.7	Fisher and Kullback-Leibler Divergence	106
3.5	Channel	106
3.5.1	Discrete Channel	107
3.5.2	Continuous Channel	110
3.6	Inequalities in Information Theory	111
3.6.1	Fisher Information Inequality	112
3.6.2	Exponential Entropy Inequality of Shannon	116
	4 WE WON'T GO INTO PDES!	121
4.1	Introduction	121
4.2	Semigroups and Generators	122
4.2.1	Semigroups	122
4.2.2	Heat Semigroup and DeBrujn's Identity	126
4.2.3	Ornstein-Uhlenbeck Semigroup	129
4.2.4	Discrete and Binary Semigroups	130
4.3	Functional Entropy	132
4.3.1	Convexity and duality formulas	135
4.3.2	Evolution of Entropy	141
4.3.3	Tensorization	143
4.4	Poincaré's Inequality	145
4.4.1	Spectral Gap Inequality	149
4.4.2	Tensorization	151
4.4.3	Perturbation	152
4.4.4	Concentration	154
4.5	Log-Sobolev Inequality	156

4.5.1	Tensorization and Perturbation	158
4.5.2	Concentration and the Herbst Method	161
4.5.3	Equivalent Definitions	165
	5 BETTER START WITH 2 THAN MANY!	169
5.1	Introduction	169
5.2	Definitions and Properties	170
5.3	Main Theorem	170
5.4	Application I: Concentration in the Hamming Cube	172
5.5	Application II: Rademacher Complexity	177
5.6	Application IV: Supervised Classification Problem	181
5.7	Application III: Concentration in Graphs	184
	6 OPEN THE WAY FOR GAUSS!	189
6.1	Introduction	189
6.2	Definitions	191
6.3	Main Theorem	191
6.4	Application I: Concentration in Gaussian Spaces	196
6.5	Application II: Gaussian Complexity	198
6.6	Application III: The Crámer-Rao Inequality	203
6.7	Application IV: The Uncertainty Principle	207
	Bibliography	211

Introduction

Our main goals in this Dissertation are to prove the Rademacher Log-Sobolev Inequality and the Gaussian Log-Sobolev Inequality and to use them to obtain results in Functional Analysis, Probability Theory and Information Theory. The statement of the Rademacher Log-Sobolev Inequality is the following: let $H_n = \{-1, 1\}^n$ and μ be the uniform measure in H_n , then for all $f : H_n \rightarrow \mathbb{R}$, we have

$$\int_{H_n} f^2 \log(f^2) d\mu - \left(\int_{H_n} f^2 d\mu \right) \log \left(\int_{H_n} f^2 d\mu \right) \leq 2 \int_{H_n} \|\nabla f\|^2 d\mu,$$

where ∇f is the discrete gradient. The Gaussian Log-Sobolev Inequality states that for all $f \in C^1(\mathbb{R}^n)$, we have

$$\int_{\mathbb{R}^n} f^2 \log(f^2) d\mu - \left(\int_{\mathbb{R}^n} f^2 d\mu \right) \log \left(\int_{\mathbb{R}^n} f^2 d\mu \right) \leq 2 \int_{\mathbb{R}^n} \|\nabla f\|^2 d\mu,$$

where μ is the standard Gaussian measure in \mathbb{R}^n . There are important applications of these results in many different fields, such as Compressed Sensitive, High Dimensional Probability and Statistic Theory, Convex Geometry, Functional Analysis and Information Theory. We will apply the Rademacher Log-Sobolev Inequality and the Gaussian Log-Sobolev Inequality to study Concentration in the Hamming Cube, Graphs and Gaussian Spaces, Rademacher and Gaussian Complexities, the Crámer-Rao Inequality and the Uncertainty Principle, among other results.

In order to achieve this goal, in the first chapters we introduce the main definitions and results we need. In Chapter 2, we provide the basic definitions and results from Probability Theory. Amongst these results, we emphasize the importance of Chernoff's Inequality, which says that for an exponentially integrable random variable, we have

$$\mathbb{P}(X - \mathbb{E}[X] \geq t) \leq e^{-t} \mathbb{E}[e^{(X - \mathbb{E}[X])}].$$

This inequality is important in order to prove concentration for Lipschitz functions.

In Chapter 3 we introduce the basic ideas of Information Theory, such as Shannon Entropy

$$H(X) = - \sum_{x \in \mathcal{X}} p(x) \log p(x),$$

and describe some applications such as in *Coding and Compressing* and the Theory of Channels. Furthermore, we will also derive, in Section 3.6, some useful inequalities, such as the Fisher Information Inequality, which says that for all X, Y with smooth densities f, g , we have

$$\frac{1}{J(X+Y)} \geq \frac{1}{J(X)} + \frac{1}{J(Y)}.$$

This is the first inequality we need to prove the Gaussian Log-Sobolev Inequality by means of Information Theory.

In Chapter 4 we will introduce the ideas from Functional Analysis, in particular, Semigroup Theory. There, we define the Ornstein-Uhlenbeck Semigroup P_t , defined as

$$P_t f(x) = \int_{\mathbb{R}^n} f(e^{-t}x + \sqrt{1-e^{-2t}}y) \frac{1}{(2\pi)^{n/2}} \exp\left(-\frac{\|y\|^2}{2}\right) dy.$$

Moreover, we will define the Functional Entropy

$$\text{Ent}(X) = \mathbb{E}[X \log X] - \mathbb{E}[X] \log \mathbb{E}[X],$$

and we will introduce the two main functional inequalities we will associate to Concentration of Measure phenomena, namely the Poincaré's Inequality and the Log-Sobolev Inequality.

Our first main result lies in Chapter 5, where we study the Rademacher Log-Sobolev Inequality in H_n . In Section 5.3 we prove this theorem, namely, for every $f : H_n \rightarrow \mathbb{R}$, we have

$$\text{Ent}(f^2) \leq 2\mathcal{E}(f),$$

with μ uniform in H_n . We will then use this result in some applications in the following sections. In Section 5.4, we will prove the Concentration of Measure in the Hamming Cube, that is, for all 1-Lipschitz function $f : H_n \rightarrow \mathbb{R}$ we have

$$\mathbb{P}(|f(X) - \mathbb{E}[f(X)]| \geq t) \leq 2 \exp(-nt^2).$$

After that, in Section 5.5, we will introduce the Rademacher Complexity of a set $V \subset \mathbb{R}^n$, namely,

$$r(V) := \mathbb{E}[\sup_{t \in V} \langle t, X \rangle],$$

where X is uniformly distributed in H_n . Furthermore, we will also prove some of its basic properties, in particular, the following bound:

$$r(V) \leq 2\sqrt{\sigma^2 \log(|V|)},$$

where $\sigma^2 = \sup_{t \in T} \|t\|^2$ and $|V|$ denotes the cardinality of V . In Section 5.6, we use Rademacher Complexity to study the Supervised Classification Problem. Finally, in Section 5.7, we will introduce the Log-Sobolev Inequality for graphs (V, E) equipped with a probability measure μ , namely

$$\text{Ent}(f^2) \leq c\mathcal{E}(f),$$

where \mathcal{E} is the energy is associated with the graph:

$$\mathcal{E}(f) = \frac{1}{4} \sum_{x \in V} \sum_{y \in E_x} (f(x) - f(y))^2 \mu(x),$$

and E_x is the set of all y such that $(x, y) \in E$.

Finally, Chapter 6 is dedicated to the Gaussian Log-Sobolev Inequality. We will prove the main result in Section 6.3, namely, for $X \sim \mathcal{N}(0, \text{Id})$ and for all $f \in C^2(\mathbb{R}^n)$ we have

$$\text{Ent}(f^2(X)) \leq 2\mathbb{E}[\|\nabla f(X)\|^2].$$

However, the first nontrivial result we will obtain is that it is equivalent with an inequality in Information Theory, say, for all random vector X with finite second moment and density $f \in C^2(\mathbb{R}^n)$, we have

$$N(X)J(X) \geq n,$$

where $N(X)$ and $J(X)$ are the exponential entropy of Shannon and the Fisher Information, respectively. The first application appears in Section 6.4, where we will prove Gaussian Concentration, namely, for all $f : \mathbb{R}^n \rightarrow \mathbb{R}$ 1-Lipschitz and $X \sim \mathcal{N}(0, \text{Id})$, we have

$$\mathbb{P}\left(|f(X) - \mathbb{E}[f(X)]| \geq t\right) \leq 2\exp\left(-t^2/2\right).$$

As a consequence of this result we will prove, for example, that

$$\mathbb{P}\left(\left|\|X\| - \mathbb{E}[\|X\|]\right| \geq t\right) \leq 2\exp(-t^2/2).$$

In Section 6.5, we will define the Gaussian Complexity of a set $V \subset \mathbb{R}^n$ as

$$w(V) = \mathbb{E}[\sup_{t \in V} \langle g, t \rangle],$$

where $g \sim \mathcal{N}(0, \text{Id})$. We will prove some of its basic properties and we will state a very powerful bound, known as the M^* Bound. Using this, we will prove one of the simplest

theorems in Compressed Sensing, namely, if we want to recover a signal $x \in T \subset \mathbb{R}^n$ according to a random measurement $Ax = y \in \mathbb{R}^m$, where $A_{ij} \sim \mathcal{N}(0, 1)$ independent, then any solution of $Az = y$ and $z \in T$ satisfies

$$\mathbb{E}[z - x] \leq \frac{Cw(T)}{\sqrt{m}}.$$

In Section 6.6, we will explore the Crámer-Rao Inequality, which says that, for any random variable X we have

$$\sigma^2(X) \geq \frac{1}{J(X)},$$

where $\sigma^2(X)$ is the variance of X . We will give some examples of applications of this inequality, which is proved through the Gaussian Log-Sobolev Inequality. Finally, in Section 6.7, we will prove that for associated random variables X and Y , which means that their densities are the Fourier Transform of one another, we have

$$16\pi^2\sigma^2(X)\sigma^2(Y) \geq 1.$$

On Dice and Coins

2.1 Introduction

The idea behind this chapter is to make this dissertation as self-contained as possible and to provide a quick reference for the readers that are not familiar with some results in Measure Theory. This is done in Sections 2.2 to 2.5. Readers that have a working knowledge of Measure Theory should skip this chapter and proceed to the following ones. However, in Sections 2.6 to 2.10 we gather and briefly discuss some results on Probability Theory that are essential for the rest of this text.

First, we will start with the definition of a Probability Space in Section 2.2. Moreover, we will state the Caratheodory's Extension Theorem. Lastly, we will introduce the Lebesgue Measure and Kolmogorov's Extension Theorem.

After this, in Section 2.3, we will give the standard definition of a Random Variable and Random Element. We will introduce the idea of Distribution and the Push Forward Measure. Finally, we will define independent random variables.

In Section 2.4, the concept of Integral and Expectation will be introduced. We will define these quantities and state some properties and theorems, such as the Monotone and Dominated Convergence Theorems. Also in this section we will define the Product Measure and the Fubini's Theorem.

Moving on, we will give some methods to compute integrals in the Section 2.5. We will describe three methods: first, using the Riemann Integral, then the Change of Variables Theorem and finally the Weak Derivative Property.

Section 2.6 is about the Characteristic Function and the Generating Function. We will explore these definitions and why they are important. We also will state some of their properties.

After this section, we study some inequalities in Probability Theory, in Section 2.7.

Here we will explore some of the most fundamental inequalities concerning measures and probability, such as Markov's Inequality and Chernoff's Inequality. These inequalities play a role in future sections since their bounds induce the Subgaussian Property, which we will mention later.

The Section 2.8 is about Conditional Expectation in a more general framework, say, Conditional Expectation with respect to a σ -field. We will state some of its properties, in particular, the Projection Property.

After that, we will see in Section 2.9 some notions of convergence of random variables. Moreover, we will define and prove the main theorems about convergence, such as Theorems 2.9.1, 2.9.2 and 2.9.5.

Finally, in Section 2.10 we will begin to introduce the idea of a *Markov chain* and *semigroups*. However, we will explain these concepts in a more general framework only in Chapter 4.

2.2 Probability Spaces

To introduce some notations and definitions from Measure Theory, let us define *semialgebra*, *algebra* and σ -*algebra*.

Definition 2.2.1. Let V be a set. A **semialgebra** \mathcal{S} is a nonempty collection of subsets of V such that

1. If $A, B \in \mathcal{S}$, then $A \cap B \in \mathcal{S}$; and
2. If $A \in \mathcal{S}$, then A^c , its complement, is a disjoint union of elements in \mathcal{S} .

We can extend this concept to an *algebra*.

Definition 2.2.2. Let V be a set. A **algebra** \mathcal{A} is a collection of subsets of V such that

1. The set V is in \mathcal{A} ;
2. If $A, B \in \mathcal{A}$, then $A \cap B \in \mathcal{A}$; and
3. If $A \in \mathcal{A}$, then $A^c \in \mathcal{A}$.

Properties 2 and 3 from Definition 2.2.2 can be used to prove the following lemma.

Lemma 2.2.1. Let \mathcal{A} be an algebra. If $A_1, \dots, A_n \in \mathcal{A}$, then

1. It is closed under finite intersections: $\bigcap_{i=1}^n A_i \in \mathcal{A}$; and

2. It is closed under finite unions: $\bigcup_{i=1}^n A_i \in \mathcal{A}$.

However, we cannot extend it to countable unions or intersections. Thereby we define a σ -algebra \mathcal{F} as those algebras \mathcal{F} which satisfies Lemma 2.2.1, but with countable union and intersection instead.

Definition 2.2.3. Let V be a set. A σ -algebra \mathcal{F} is a collection of subsets of V such that

1. The set V is in \mathcal{F} ;
2. If $A \in \mathcal{F}$, then $A^c \in \mathcal{F}$; and
3. If $(A_i)_{i \in \mathbb{N}} \subset \mathcal{F}$, then

$$\bigcup_{i \in \mathbb{N}} A_i \in \mathcal{F}.$$

It will be convenient to define the *smallest semialgebra* (resp. *algebra* and σ -algebra) generated by a collection \mathcal{B} of subsets of V .

Definition 2.2.4. Let V be a set and \mathcal{B} a collection of subsets of V . The **semialgebra generated by \mathcal{B}** (resp. **algebra** and σ -algebra) is the smallest **semialgebra generated by \mathcal{B}** (resp. **algebra** and σ -algebra) which contains \mathcal{B} . We denote by $\sigma(\mathcal{B})$ the smallest σ -algebra generated by \mathcal{B} . In this case, we say that \mathcal{B} **generates** $\sigma(\mathcal{B})$.

This is well-defined, since there is at least one semialgebra generated by \mathcal{B} (resp. algebra and σ -algebra) which contains \mathcal{B} , namely $\mathcal{P}(V)$, the collection of all subsets of V .

Definition 2.2.5. If \mathcal{F} is a σ -algebra, then a set $A \in \mathcal{F}$ is called a **measurable set**.

Summarizing, we obtain the definition of a *measurable space*.

Definition 2.2.6. Let V be a nonempty set and \mathcal{F} be a σ -algebra of V , then the pair (V, \mathcal{F}) is called a **measurable space**.

We can now define a measure in (V, \mathcal{F}) .

Definition 2.2.7. Let \mathcal{F} be a σ -algebra. Then a measure $\mu : \mathcal{F} \rightarrow [0, \infty]$ is a function such that

1. The empty set has measure zero: $\mu(\emptyset) = 0$; and
2. If $(A_n)_{n \in \mathbb{N}} \in \mathcal{F}$ is a countable disjoint collection, then

$$\mu\left(\bigcup_{i \in \mathbb{N}} A_i\right) = \sum_{i \in \mathbb{N}} \mu(A_i).$$

Definition 2.2.8. Let μ be a measure in (V, \mathcal{F}) . If $\mu(V) = 1$, then μ is called a **probability measure**, or just **probability**. Also, if there is a countable collection $(A_i)_{i \in \mathbb{N}}$ such that

1. The collection covers V : $V = \bigcup_{i \in \mathbb{N}} A_i$; and
2. For all $i \in \mathbb{N}$, we have that $\mu(A_i) < \infty$,

then the measure μ is called a **σ -finite measure**,

For this reason, we have the following definition.

Definition 2.2.9. Let V be a nonempty set, \mathcal{F} be a σ -algebra and \mathbb{P} be a probability in V , then the triple $(V, \mathcal{F}, \mathbb{P})$ is called a **probability space**.

Let us just state some properties of probabilities.

Lemma 2.2.2. *Let $(V, \mathcal{F}, \mathbb{P})$ be a Probability Space, then the following are true.*

1. If $A, B \in \mathcal{F}$ and $A \subseteq B$, then we have $\mathbb{P}(A) \leq \mathbb{P}(B)$;
2. If $A_i \in \mathcal{F}$ for all $i \in \mathbb{N}$, $A_i \subseteq A_{i+1}$ and $A := \bigcup_{i \in \mathbb{N}} A_i$, then

$$\mathbb{P}(A) = \lim_{i \rightarrow \infty} \mathbb{P}(A_i);$$

3. If $A \in \mathcal{F}$, then $\mathbb{P}(A^c) = 1 - \mathbb{P}(A)$; and
4. If $(A_i)_{i \in \mathbb{N}} \subset \mathcal{F}$, then

$$\mathbb{P}\left(\bigcup_{i \in \mathbb{N}} A_i\right) \leq \sum_{i=1}^{\infty} \mathbb{P}(A_i).$$

We will not prove this lemma. The proof can be found in [Durrett \(2019\)](#), [Shiryaev \(2016\)](#) or [Folland \(2013\)](#).

2.2.1 Carathéodory's Theorem

Let \mathcal{A} be a semialgebra and $\mu : \mathcal{A} \rightarrow \mathbb{R}_+$ such that

$$\mu\left(\bigcup_{i \leq n} A_i\right) = \sum_{i \leq n} \mu(A_i),$$

whenever $(A_i)_{i=1}^n \subset \mathcal{A}$ is a disjoint collection such that $\bigcup_{i \leq n} A_i \in \mathcal{A}$ and $\mu(\emptyset) = 0$. We should expect that we can extend μ to an unique measure $\hat{\mu}$ in the algebra generated by \mathcal{A} and

$$\hat{\mu}(A) = \mu(A),$$

for all $A \in \mathcal{A}$. Carathéodory's Theorem provides sufficient conditions for this to hold.

Theorem 2.2.1 (Carathéodory's Theorem). *Let \mathcal{B} be a semialgebra and $\mu : \mathcal{B} \rightarrow [0, \infty]$ such that*

1. *The empty set has measure zero: $\mu(\emptyset) = 0$;*
2. *If $(A_i)_{i=1}^n \subset \mathcal{B}$ is a disjoint finite collection of sets and $\bigcup_{i=1}^n A_i \in \mathcal{B}$, then*

$$\mu\left(\bigcup_{i=1}^n A_i\right) = \sum_{i=1}^n \mu(A_i); \text{ and}$$

3. *If $(A_i)_{i \in \mathbb{N}} \subset \mathcal{B}$ is a disjoint countable collection and $\bigcup_{i \in \mathbb{N}} A_i \in \mathcal{B}$, then*

$$\mu\left(\bigcup_{i \in \mathbb{N}} A_i\right) \leq \sum_{i \in \mathbb{N}} \mu(A_i).$$

Then μ has an unique extension $\hat{\mu}$ to the algebra \mathcal{A} generated by \mathcal{B} , in the sense that $\hat{\mu}$ restricted to \mathcal{B} is equal to μ and

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

whenever A_i is a finite disjoint family in the algebra \mathcal{A} . Also, if $\hat{\mu}$ is σ -finite, then $\hat{\mu}$ has an unique extension to $\sigma(\mathcal{A})$.

Proof. The reader can find two different proofs: using π - λ Theorem (see [Durrett \(2019\)](#)) or using outer measures (see [Folland \(2013\)](#)). \square

2.2.2 Borel, Lebesgue and Kolmogorov

In this section, we define the most frequent examples of measurable sets and measures. First, let us define the Borel σ -algebra.

Definition 2.2.10. A **Topological Space** is (V, τ) where τ is a collection of subsets of V such that

1. The empty set and V itself are in τ : $\emptyset \in \tau$ and $V \in \tau$;
2. If $A, B \in \tau$, then $A \cap B \in \tau$; and
3. For all collection of index $\lambda \in \Lambda$ and $(A_\lambda)_{\lambda \in \Lambda} \subseteq \tau$, we have

$$\bigcup_{\lambda \in \Lambda} A_\lambda \in \tau.$$

A set $A \in \tau$ is called **open**.

Definition 2.2.11. Let (V, τ) be a topological space. The smallest σ -algebra generated by τ , and denoted by $\mathcal{B}(V)$, is called the **Borel σ -algebra**.

Now we can define the *Lebesgue Measure* in \mathbb{R}^n .

Definition 2.2.12. Let $(\mathbb{R}^n, \mathcal{B}(\mathbb{R}^n))$, where the topology is generated by the Euclidean norm, then Theorem 2.2.1 implies that there is an unique σ -finite measure λ , called the **Lebesgue Measure**, such that for all $a_i \leq b_i$ we have

$$\lambda\left(\prod_{i=1}^n (a_i, b_i]\right) = \prod_{i=1}^n (b_i - a_i).$$

Remark 2.2.1. We will interchangeably denote the Lebesgue measure by λ , λ_n or dx when the context is clear. This measure is invariant by translation and rotation. (for a proof, see [Folland \(2013\)](#)).

The last theorem we need to state in this section is *Kolmogorov's Extension Theorem*. First, let us define the space $(\mathbb{R}^\infty, \mathcal{B}(\mathbb{R}^\infty))$.

Definition 2.2.13. Let \mathbb{R}^∞ be the space of all sequences $(x_n)_{n \in \mathbb{N}} \subset \mathbb{R}$. A cylinder A in \mathbb{R}^∞ is a subset such that there are an n and sets $B_1, \dots, B_n \in \mathcal{B}(\mathbb{R})$ and

$$A = \prod_{i=1}^n B_i \times \prod_{i=n+1}^{\infty} \mathbb{R}.$$

Hence $\mathcal{B}(\mathbb{R}^\infty)$ is defined as the smallest σ -algebra that contains all the cylinders.

Now we can state Kolmogorov's Extension Theorem.

Theorem 2.2.2 (Kolmogorov's Extension Theorem). *Let \mathbb{P}_n be probabilities in $(\mathbb{R}^n, \mathcal{B}(\mathbb{R}^n))$, for $n \in \mathbb{N}$, possessing the consistency property, which means that for all $n \in \mathbb{N}$ and all $B \in \mathcal{B}(\mathbb{R}^n)$, we have*

$$\mathbb{P}_{n+1}(B \times \mathbb{R}) = \mathbb{P}_n(B).$$

Then there is an unique probability \mathbb{P} in $(\mathbb{R}^\infty, \mathcal{B}(\mathbb{R}^\infty))$ such that

$$\mathbb{P}(B_1 \times \dots \times B_n \times \mathbb{R} \times \mathbb{R} \times \dots) = \mathbb{P}_n(B_1 \times \dots \times B_n),$$

for all $n \in \mathbb{N}$ and all $B_i \in \mathcal{B}(\mathbb{R})$.

Proof. For a proof, we recommend [Shiryaev \(2016\)](#). □

2.3 Random Variables and Random Vectors

Definition 2.3.1. Let (V_1, \mathcal{F}_1) and (V_2, \mathcal{F}_2) be two Measurable Spaces. We say that a function $f : V_1 \rightarrow V_2$ is **measurable** if $f^{-1}(A) \in \mathcal{F}_1$ whenever $A \in \mathcal{F}_2$.

This definition is similar to the definition of continuity in topological spaces, say, a function $f : (V_1, \tau_1) \rightarrow (V_2, \tau_2)$ is continuous if $f^{-1}(A) \in \tau_1$ whenever $A \in \tau_2$. In fact, we have the following theorem linking these two definitions.

Theorem 2.3.1. *Let (V_1, τ_1) and (V_2, τ_2) be two Topological Spaces and let $\mathcal{B}(V_1)$, $\mathcal{B}(V_2)$ be the respective Borel σ -algebra. If $f : V_1 \rightarrow V_2$ is continuous, then it is measurable.*

We define now *null sets* and *almost surely* properties.

Definition 2.3.2. A **null set** A is a set such that $\mu(A) = 0$.

Definition 2.3.3. We say that a property P holds **for almost all points** $x \in V$, or **almost surely**, if there is a null set A such that P holds for all $x \in A^c$. We denote this by μ -a.s., or, when the measure μ is implicit, just by *a.s.*

For instance, we say that $f = g$ μ -a.s. if there is a null set A such that $f(x) = g(x)$ for all $x \in A^c$.

2.3.1 Random Variables and Distribution

Now we are able to define a *random variable* and a *random vector*.

Definition 2.3.4. Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a Probability Space and $(\mathbb{R}^n, \mathcal{B}(\mathbb{R}^n))$ be measurable spaces. A **random variable** X is a measurable function $X : \Omega \rightarrow \mathbb{R}$ and a **random vector** Y is a measurable function $Y : \Omega \rightarrow \mathbb{R}^n$.

Remark 2.3.1. We can also define a **random element** by measurable functions $X : \Omega \rightarrow S$, where (S, \mathcal{S}) is a measurable space (see [Durrett \(2019\)](#)).

For every random vector $X : \Omega \rightarrow \mathbb{R}^n$, there is a probability in $(\mathbb{R}^n, \mathcal{B}(\mathbb{R}^n))$ associated with it, namely, the *push forward measure*.

Definition 2.3.5. Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a Probability Space and $X : \Omega \rightarrow \mathbb{R}^n$ be a Random Vector, then the **push forward measure**, or the **distribution of X** , is the measure \mathbb{P}_X defined in \mathbb{R}^n , so that

$$\mathbb{P}_X(A) := \mathbb{P}(X \in A) := \mathbb{P}\{\omega \in \Omega : X(\omega) \in A\}.$$

Remark 2.3.2. We will denote $X \sim \mathbb{P}_X$ to say that X has distribution \mathbb{P}_X .

Definition 2.3.6. Let $(\Omega_i, \mathcal{F}_i, \mathbb{P}_i)$, for $i = 1, 2$, be two Probability Spaces and $X_i : (\Omega_i, \mathcal{F}_i) \rightarrow (S, \mathcal{S})$ be two random elements. We say that X_1 is **equal in distribution** to X_2 , and denote by $X_1 \stackrel{d}{=} X_2$, if

$$\mathbb{P}_{X_1} = \mathbb{P}_{X_2}.$$

Remark 2.3.3. Note that this does not imply that $X_1 - X_2 \stackrel{d}{=} 0$, since this is not well-defined.

In some books, the distribution of a random variable or vector X in \mathbb{R}^n is defined in terms of rectangles:

$$\mathbb{F}_X(a) := \mathbb{P}(\{X \leq a\}) = \mathbb{P}\left(\prod_{i=1}^n \{X_i \leq a_i\}\right),$$

for $a \in \mathbb{R}^n$. These two definitions are equivalent because of the following lemma.

Lemma 2.3.1. *Let $(\mathbb{R}^n, \mathcal{B}(\mathbb{R}^n))$. Let \mathcal{R} be the set of rectangles R of the form*

$$R = \prod_{i=1}^n \{x \in \mathbb{R} : x \leq a_i\},$$

for some $a \in \mathbb{R}^n$. Then

$$\mathcal{B}(\mathbb{R}^n) = \sigma(\mathcal{R}).$$

There are several properties that define a distribution \mathbb{F}_X , which we state below.

First, let (\mathbb{R}^n, \preceq) where \preceq is the partial order generated by the canonical positive cone \mathbb{R}_+^n in \mathbb{R}^n . For $I = (a, b] \subset \mathbb{R}$ and $g : \mathbb{R}^n \rightarrow \mathbb{R}$, let $\Delta_I g : \mathbb{R}^{n-1} \rightarrow \mathbb{R}$ such that

$$\Delta_I g(x_1, \dots, x_n) := g(x_1, \dots, x_{n-1}, b) - g(x_1, \dots, x_{n-1}, a).$$

Then we have the following theorem.

Theorem 2.3.2. *Let $\mathbb{F} : \mathbb{R}^n \rightarrow \mathbb{R}$ be a function. Then \mathbb{F} is a distribution of a probability measure μ , in the sense that $\mathbb{F}(x) = \mu((-\infty, x])$ if and only if the following properties hold.*

1. *For all i and all $(x_1, \dots, x_{i-1}, x_{i+1}, \dots, x_n) \in \mathbb{R}^{n-1}$ we have that*

$$\lim_{x_i \rightarrow -\infty} F(x) = 0;$$

2. *If $x_i \rightarrow \infty$ for all i , then*

$$\lim_{\forall i \ x_i \rightarrow \infty} F(x) \rightarrow 1;$$

3. *F is an increasing function: if $x \succeq y$, then $F(x) \geq F(y)$; and*

4. *For any $a \preceq b$, we have that*

$$\Delta_{I_n} \Delta_{I_{n-1}} \cdots \Delta_1 F(x_1, \dots, x_n) \geq 0,$$

where $I_k = (a_k, b_k]$.

Given a distribution \mathbb{F} , is there a random vector X such that $\mathbb{F} = \mathbb{F}_X$?

Theorem 2.3.3. *Let μ be a probability in \mathbb{R}^n , then there is a probability space $(\Omega, \mathcal{F}, \mathbb{P})$ and a random vector $X : \Omega \rightarrow \mathbb{R}^n$ such that $\mathbb{P}_X = \mu$.*

Proof. Let $(\Omega, \mathcal{F}, \mathbb{P}) = (\mathbb{R}^n, \mathcal{B}(\mathbb{R}^n), \mu)$. Then $X(\omega) = \omega$ is the desired random variable. \square

This theorem states that we can always take our Probability Space as $(\mathbb{R}^n, \mathcal{F}, \mathbb{P})$.

2.3.2 Independence

We are able now to define one of the most important concepts in Probability: *independence*.

Definition 2.3.7. Let X and Y be two random vectors in \mathbb{R}^n and \mathbb{R}^m , respectively. We say that X and Y are **independent** if, for all $A \in \mathcal{B}(\mathbb{R}^n)$ and $B \in \mathcal{B}(\mathbb{R}^m)$ we have

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

Remark 2.3.4. This is equivalent to say that the distribution $\mathbb{P}_{X,Y}$ of (X, Y) is a product measure of \mathbb{P}_X and \mathbb{P}_Y (see Section 2.4.3). Therefore, we can always assume the existence of such random variables.

Because of Kolmogorov's Theorem 2.2.2, we can always assume the existence of countably many independent random variables. Therefore, from now on, we will assume the existence of independent and identically distributed (i.i.d.) random variables (X_1, X_2, \dots) with distribution μ .

Definition 2.3.8. Let (S, \mathcal{S}, μ) be a probability space. We say that a random element $X \in S$ is a **sample** according to μ if $X \sim \mu$.

Definition 2.3.9. Let X and Y be random vectors in \mathbb{R}^n and \mathbb{R}^m , respectively. If $\mathbb{P}(Y \in B) > 0$, then the **conditional probability of X given $Y \in B$** is defined as the distribution

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

Remark 2.3.5. In fact, we can define the conditional probability without random variables, namely,

$$\mathbb{P}(A|B) := \frac{\mathbb{P}(A \cap B)}{\mathbb{P}(B)},$$

if $\mathbb{P}(B) > 0$.

Let us give some examples of all these concepts. For now on we will denote random variables or vector by the initials “r.v.” and all the functions will be measurable.

Example 2.3.1. Let X be a random variable. If X takes values in some discrete set $\mathcal{X} \subset \mathbb{R}^n$, then it is known as a **discrete r.v.** Hence we just need to look at the probability of

$$X = x,$$

for $x \in \mathcal{X}$, that is, the distribution of X is entirely determined by the values

$$p(x) := \mathbb{P}(X = x), \quad x \in \mathcal{X}.$$

In the case $\mathcal{X} = \{0, 1\}$ and $\mathbb{P}(X = 1) = p$, we say that $X \sim \text{Ber}(p)$, a **Bernoulli r.v.**

Example 2.3.2. By a **Rademacher r.v.** X we mean that X takes two values, $+1$ and -1 , with probability p and $1 - p$, respectively, that is

$$\mathbb{P}(X = 1) = p, \quad \mathbb{P}(X = -1) = 1 - p.$$

We will denote this random variable as $X \sim \text{Rad}(p)$.

Example 2.3.3. A r.v. X taking values **uniformly** in a compact set $V \subset \mathbb{R}^n$ is such that its distribution is

$$\mathbb{P}_X(A) = \mathbb{P}(X \in A) = \frac{\lambda(A \cap V)}{\lambda(V)}.$$

We will denote this r.v. as $X \sim \text{Unif}(V)$.

Example 2.3.4. We say that a random variable X has an **absolutely continuous distribution** if there is a measurable function $f : \mathbb{R}^n \rightarrow \mathbb{R}$, namely the density of X , such that

$$\mathbb{P}(X \in A) := \int_A f(x) \, dx.$$

Remark 2.3.6. Even though we have not yet defined the integral, this is an important example to introduce here.

The main example of an absolutely continuous distribution is the *Gaussian*. Let A be a $n \times n$ matrix and denote $|A|$ the absolute value of its determinant.

Example 2.3.5. Let $\mu \in \mathbb{R}^n$ and Σ be a $n \times n$ positive definite matrix, that is, for all $x \in \mathbb{R}^n \setminus \{0\}$, we have

$$x^T \Sigma x = \sum_{i,j=1}^n \Sigma_{ij} x_i x_j > 0.$$

Then the function $f : \mathbb{R}^n \rightarrow \mathbb{R}_+$, defined as

$$f(x) := \frac{|\Sigma|^{-1/2}}{(2\pi)^{n/2}} \exp\left(-\frac{\langle(x-\mu), \Sigma^{-1}(x-\mu)\rangle}{2}\right),$$

is a density, that is,

$$\int_{\mathbb{R}^n} f(x) dx = 1$$

The **Gaussian Measure** γ is the measure defined by f , that is,

$$\gamma(A) := \int_A f d\lambda.$$

If X has distribution γ , then we denote it by $X \sim \mathcal{N}(\mu, \Sigma)$. In the case $\mu = 0$ and $\Sigma = \text{Id}$, we say that X is a **standard Gaussian**.

2.4 Integral and Expected Value

In this section, we will define the Integral and Expected Value of a general measurable function and random variable, and prove some of their properties.

2.4.1 The Three Steps

The classical approach to introduce the Integral is to define it in three steps: first, we consider the integral of simple function. Then, we define it to positive ones. Finally, we extend it to general functions.

Consequently, let us define a *simple function*.

Definition 2.4.1. Let f be a measurable function in (V, \mathcal{F}, μ) . We say that f is a **simple measurable function** if there is an $n \in \mathbb{N}$, disjoint sets $A_1, \dots, A_n \in \mathcal{F}$ with $\mu(A_i) < \infty$ and $c_1, \dots, c_n \in \mathbb{R}$ such that for all $x \in V$ we have

$$f(x) = \sum_{k=1}^n c_k \mathbf{1}_{A_k}(x) \quad \mu - \text{a.s.}$$

where $\mathbf{1}_A(x) = 1$ if $x \in A$ and 0 otherwise, that is, $\mathbf{1}_A$ is the **indicator function** of A .

Remark 2.4.1. We say that $(A_i, c_i)_{i=1}^n$ is a **representation** of f . There is no unique representation of a function, since we can always split one set A_i into $B \subset A_i$ and $B^c \cap A_i \subset A_i$.

Now we can define the integral of simple functions.

Definition 2.4.2. Let f be a simple function with representation $(A_i, c_i)_{i=1, \dots, n}$. Then the **integral** of f , defined as $\int f d\mu$, is

$$\int f d\mu := \sum_{k=1}^n c_k \mu(A_k).$$

Remark 2.4.2. We will equivalently write $\int f d\mathbb{F}$, with \mathbb{F} the distribution generated by μ . Also, when the parameter x of integration is not evident, we will write $d\mu(dx)$ or $d\mathbb{F}(x)$.

It can be shown this integral is well-defined in the sense that it does not depend on the particular representation of f .

Let us now state some of its properties.

Lemma 2.4.1. *Let f and g be simple functions, then*

1. *If $f \geq 0$ μ -a.s., then*

$$\int f d\mu \geq 0;$$

2. *If $a \in \mathbb{R}$, then*

$$\int (f + ag) d\mu = \int f d\mu + a \int g d\mu;$$

3. *For a simple function f , it is true that*

$$|\int f d\mu| \leq \int |f| d\mu; \text{ and}$$

4. *If $f \geq 0$ a.s. and*

$$\int f d\mu = 0,$$

then $f = 0$ a.s.

We can now define the integral for a *positive measurable functions*.

Definition 2.4.3. Let f be a positive measurable function, which means that $f \geq 0$ μ -a.s., then its **integral** $\int f$ is defined as

$$\int f d\mu := \sup \left\{ \int g d\mu : g \text{ is simple and } 0 \leq g \leq f \right\}.$$

Using this, we can state the *Monotone Convergence Theorem*.

Theorem 2.4.1. *Let $(f_n)_{n \in \mathbb{N}}$, f be positive functions such that $f_n \leq f_{n+1}$ for all $n \in \mathbb{N}$ and*

$$f(x) = \lim_{n \rightarrow \infty} f_n(x),$$

then

$$\int f d\mu = \lim_{n \rightarrow \infty} \int f_n d\mu.$$

Proof. For a proof of this theorem, we recommend [Folland \(2013\)](#). \square

Now, we define the integral for a *general function* as follows.

Definition 2.4.4. Let f be a measurable function and let $f^+(x) := f(x) \vee 0 = \max\{f(x), 0\}$ and $f^-(x) := (-f(x)) \vee 0 = \max\{-f(x), 0\}$. Then f^+ and f^- are positive functions and the **integral** of f is defined as

$$\int f \, d\mu = \int f^+ \, d\mu - \int f^- \, d\mu,$$

when the difference is not $\infty - \infty$.

Definition 2.4.5. We say that a function f is **integrable** if

$$\int |f| \, d\mu < \infty.$$

Now we can state the *Dominated Convergence Theorem*.

Theorem 2.4.2. Let $(f_n)_{n \in \mathbb{N}}$ be a sequence of functions with $f_n \rightarrow f$ almost surely. Suppose $|f_n| \leq g$ for an integrable function g and all n , then

$$\int f \, d\mu = \lim_{n \rightarrow \infty} \int f_n \, d\mu.$$

Definition 2.4.6. Let E be a measurable set and f be an integrable function, then the **integral** of f in E is defined as

$$\int_E f \, d\mu := \int f \mathbf{1}_E \, d\mu,$$

and it is clear that

$$\int_V f \, d\mu = \int f \, d\mu.$$

Now, we have a simple lemma.

Lemma 2.4.2. Let \mathcal{L} be the space of all measurable functions. Then

$$R := \{(f, g) \in \mathcal{L}^2 : f = g \text{ a.s.}\}$$

is an equivalence relation.

Now we are able to define the L^p Spaces for $p \in [1, \infty]$. Let \sim be the equivalence relation in Lemma 2.4.2 and $[f]$ be the class of equivalence of f , then the L^p is defined as the following.

Definition 2.4.7. Let $p \in [1, \infty)$. The $L^p(V, \mathcal{F}, \mu)$ **Space** in (V, \mathcal{F}, μ) is the quotient space of all measurable functions f such that

$$\int_V |f|^p d\mu < \infty,$$

that is,

$$L^p(\mu) := \{[f] : \int_V |f|^p d\mu < \infty\}.$$

In the case $p = \infty$, we need to define the essential supremum of f .

Definition 2.4.8. Let f be a measurable function. Then its **essential supremum** is defined as

$$\text{ess sup } f := \inf\{t : \mu(f > t) = 0\},$$

and if there is no $t \in \mathbb{R}$ such that $\mu(f > t) = 0$, then $\text{ess sup } f := \infty$. A function f such that $\text{ess sup } |f| < \infty$ is called **bounded**.

Note that, for all $t \in \mathbb{R}$ such that $\mu(f > t) = 0$ we have $f \leq t$ almost surely, hence almost surely we have

$$f \leq \text{ess sup } f,$$

thereby the name.

The L^∞ is the quotient space of all limited measurable functions.

Definition 2.4.9. Let (V, \mathcal{F}, μ) . The $L^\infty(V, \mathcal{F}, \mu)$ **Space** is defined as

$$L^\infty := \{[f] : \text{ess sup } |f| < \infty\}.$$

Remark 2.4.3. When the space (V, \mathcal{F}, μ) is clear from the context, we will denote $L^p(V, \mathcal{F}, \mu)$ by just L^p or $L^p(\mu)$. Moreover, since Lemma 2.4.1, the integral $\int |f|^p d\mu$ does not depend on which element $g \in [f]$ we take to compute the integral.

Theorem 2.4.3. Let $\|f\|_p := \left(\int |f|^p d\mu \right)^{1/p}$ when $p \in [1, \infty)$ or $\|f\|_\infty := \text{ess sup } |f|$ when $p = \infty$, then $\|\cdot\|_p$ is a norm and $(L^p, \|\cdot\|_p)$ is a Banach Space.

Remark 2.4.4. In the case of L^2 , it is a Hilbert Space with inner product $\langle f, g \rangle := \int f g d\mu$.

We can generalize these ideas to define the integral of complex-valued as

$$\int f d\mu := \int \text{Re}(f) d\mu + i \int \text{Im}(f) d\mu.$$

The L^2 case will have the inner product $\langle f, g \rangle = \int f \bar{g} d\mu$ and all absolute values are now $|f| = \sqrt{f \bar{f}}$.

Finally, the expected value of a random variable is just its integral.

Definition 2.4.10. Let X be a random variable in $(\Omega, \mathcal{F}, \mathbb{P})$, then the **expected value** $\mathbb{E}X$, or **mean** of X , is defined as

$$\mathbb{E}(X) := \int X d\mathbb{P} = \int_{\Omega} X d\mathbb{P}.$$

Remark 2.4.5. We can extend this concept to vectors with some abuse of notation: if $X = (X_1, \dots, X_n) \in \mathbb{R}^n$ is a random vector, then

$$\mathbb{E}(X) := (\mathbb{E}[X_1], \dots, \mathbb{E}[X_n]) \in \mathbb{R}^n.$$

We call this the **expected value of the vector X** . Also, we say that $X \in L^p(\Omega, \mathcal{F}, \mathbb{P})$ if

$$\mathbb{E}[\|X\|_p^p] = \mathbb{E}\left[\sum_{i=1}^n |X_i|^p\right] < \infty.$$

Therefore, all the properties we stated for the integral are also true for the expected value of X .

The last theorem in this section refers to sufficient conditions for taking limits under the integral.

Theorem 2.4.4. Let (V, \mathcal{F}, μ) be a measurable space. Let $[a, b]$ be a finite interval in \mathbb{R} and $f : V \times [a, b] \rightarrow \mathbb{C}$ be a measurable function with the property that, for all $t \in [a, b]$, $f(\cdot, t)$ is integrable. Let $F(t) := \int f(\cdot, t) d\mu$. Then we have the following:

1. Suppose $|f(\cdot, t)| \leq g(\cdot)$ for all t and g is integrable. Take $t_0 \in [a, b]$. If $f(x, \cdot)$ is continuous at t_0 , for all x , then

$$\lim_{t \rightarrow t_0} F(t) = F(t_0); \text{ and}$$

2. Suppose $f(x, \cdot) \in C^1([a, b])$, for all x , that $\left|\frac{df(\cdot, t)}{dt}\right| \leq g(\cdot)$ for all t and g is integrable. If $t_0 \in [a, b]$, then

$$F'(t_0) = \int \partial_t f(\cdot, t_0) d\mu.$$

Proof. For a proof, see [Folland \(2013\)](#). □

2.4.2 Radon-Nikodym Theorem

As we saw in the Examples [2.3.4](#) and [2.3.5](#), we need to be able to compute $\int_A f dx$ or prove that this defines a measure in \mathbb{R}^n .

Lemma 2.4.3. Let f be an integrable function in \mathbb{R}^n with respect to the Lebesgue Measure, then

$$\mu(A) := \int_A |f| dx$$

defines a finite measure in \mathbb{R}^n , and all null sets A with respect to the Lebesgue Measure are also null with respect to μ .

Proof. As

$$\int_{\emptyset} |f| \, dx = \int |f| \mathbf{1}_{\emptyset} \, dx = 0,$$

since $\mathbf{1}_{\emptyset} \equiv 0$, we have that $\mu(\emptyset) = 0$. Also, if $(A_i)_{i \in \mathbb{N}}$ is a countable family of disjoint sets and $A = \bigcup_{i=1}^{\infty} A_i$, then

$$\mathbf{1}_A(x) = \sum_{i=1}^{\infty} \mathbf{1}_{A_i}(x),$$

for almost all $x \in \mathbb{R}^n$. Then

$$|f| \sum_{k=1}^n \mathbf{1}_{A_k} \nearrow |f| \mathbf{1}_A.$$

Moreover, we also have that

$$\sum_{i=1}^{\infty} \mu(A_i) = \lim_{n \rightarrow \infty} \sum_{k=1}^n \mu(A_k) = \lim_{n \rightarrow \infty} \int |f| \sum_{k=1}^n \mathbf{1}_{A_k} \, dx,$$

and

$$\mu(A) = \int |f| \mathbf{1}_A \, dx.$$

Consequently the Monotone Convergence Theorem implies that

$$\lim_{n \rightarrow \infty} \int |f| \sum_{k=1}^n \mathbf{1}_{A_k} \, dx = \int |f| \mathbf{1}_A \, dx,$$

therefore, we have

$$\sum_{i=1}^{\infty} \mu(A_i) = \mu(A),$$

hence μ is a measure. It is finite since

$$\mu(\mathbb{R}^n) = \int |f| \mathbf{1}_{\mathbb{R}^n} \, dx = \int |f| \, dx < \infty.$$

We now prove the second part of the statement. Let A be a null Lebesgue set. Let us prove that $\int_A |f| \, dx$ is 0 for simple functions f . Then we will prove it also is zero for positive functions, and finally for general functions.

First, if $|f| := \sum_{i=1}^n c_i \mathbf{1}_{A_i}$, then

$$|f| \mathbf{1}_A = \sum_{i=1}^n c_i \mathbf{1}_{A_i \cap A},$$

which clearly implies that $\int |f| \mathbf{1}_A \, dx = 0$, since

$$\mu(A_i \cap A) \leq \mu(A) = 0.$$

To prove it for positive functions, let g_n be a sequence of simple functions with

$$0 \leq g_n \leq |f|\mathbf{1}_A,$$

and $g_n \rightarrow |f|\mathbf{1}_A$, which can be done considering the family $g_n = |f|\mathbf{1}_A \wedge n$. Notice that

$$0 \leq g_n \mathbf{1}_A \leq g_n \leq |f|\mathbf{1}_A.$$

Let $x \in A$, then $g_n(x) = g_n(x)\mathbf{1}_A(x)$, hence

$$g_n(x)\mathbf{1}_A(x) \rightarrow |f|(x)\mathbf{1}_A(x).$$

If $x \notin A$, then

$$g_n(x)\mathbf{1}_A = 0 = |f|(x)\mathbf{1}_A(x),$$

hence $g_n \mathbf{1}_A \rightarrow |f|\mathbf{1}_A$. By the Dominated Convergence Theorem, we have that

$$0 = \int g_n \mathbf{1}_A \, dx \rightarrow \int |f|\mathbf{1}_A \, dx,$$

The general case follows by taking negative and positive parts. Hence, $\mu(A) = 0$. \square

Definition 2.4.11. Let μ and ν be two σ -finite measures in the same Measurable Space (V, \mathcal{F}) . We say that μ is **absolutely continuous with respect** ν , and denote by $\mu \ll \nu$, if all ν -null sets are μ -null sets.

The name *absolutely continuous* can be justified because of the following lemma.

Lemma 2.4.4. Let μ and ν be two σ -finite measures in the same Measurable Space. Then $\mu \ll \nu$ if and only if for all $\varepsilon > 0$, there is a $\delta > 0$ such that for all A with $\nu(A) < \delta$ we have that $\mu(A) < \varepsilon$.

Lemma 2.4.3 states that all measures μ such that

$$\mu(A) = \int_A |f| \, d\nu$$

are absolutely continuous with respect to ν . In the next theorem we state the converse.

Theorem 2.4.5 (Radon-Nikodym Theorem). Let $\mu \ll \nu$ be two σ -finite measures in (V, \mathcal{F}) then there is an almost surely unique positive function f such that $\mu(A) = \int_A f \, d\nu$ for all measurable sets A .

Definition 2.4.12. Let $\mu \ll \nu$ and f be the unique non-negative function such that for all A we have

$$\mu(A) = \int_A f \, d\nu.$$

Then f is called the **Radon-Nikodym derivative** $\frac{d\mu}{d\nu}$.

We state below some properties of Radon-Nikodym derivative.

Lemma 2.4.5. *We have that*

1. *If $\mu_i \ll \nu$, for $i = 1, 2$, then $\mu_1 + \mu_2 \ll \nu$ and*

$$\frac{d(\mu_1 + \mu_2)}{d\nu} = \frac{d\mu_1}{d\nu} + \frac{d\mu_2}{d\nu};$$

2. *If $\mu \ll \nu$ and f is integrable with respect μ , then*

$$\int f d\mu = \int f \frac{d\mu}{d\nu} d\nu; \text{ and}$$

3. *If $\mu \ll \nu$ and $\nu \ll \mu$, then*

$$\frac{d\nu}{d\mu} = \left(\frac{d\mu}{d\nu} \right)^{-1}.$$

2.4.3 Product Measure and Fubini's Theorem

The first step to define the product measure in a product space is to define the product σ -algebra.

Lemma 2.4.6. *Let (V_1, \mathcal{F}_1) and (V_2, \mathcal{F}_2) be two measurables spaces. Let*

$$\mathcal{B} := \{A \times B : A \in \mathcal{F}_1, B \in \mathcal{F}_2\},$$

then \mathcal{B} is a semialgebra.

The σ -algebra generated by this semialgebra is the *product σ -algebra*.

Definition 2.4.13. Let (V_1, \mathcal{F}_1) and (V_2, \mathcal{F}_2) be two measurables spaces. Let

$$\mathcal{B} := \{A \times B : A \in \mathcal{F}_1, B \in \mathcal{F}_2\}$$

be the family of **cylinders sets**. Then $\sigma(\mathcal{B})$ is called the **product σ -algebra** and denoted by the symbol $\mathcal{F}_1 \times \mathcal{F}_2$.

As a consequence of the following theorem, we can define the product measure.

Theorem 2.4.6. *Let μ_1 and μ_2 be σ -finite measures in (V_1, \mathcal{F}_1) and (V_2, \mathcal{F}_2) , respectively. Then there is an unique measure μ defined in $(V_1 \times V_2, \mathcal{F}_1 \times \mathcal{F}_2)$ such that*

$$\mu(A \times B) = \mu_1(A)\mu_2(B),$$

*for all $(A \times B) \in \mathcal{B}$. This measure is called the **product measure** of μ_1 and μ_2 , and denoted by $\mu_1 \times \mu_2$.*

Proof. Let \mathcal{B} the family of cylinders sets and $\mu : \mathcal{B} \rightarrow \mathbb{R}_+$ be such that

$$\mu(A \times B) := \mu_1(A)\mu_2(B).$$

Then $\mu(\emptyset) = 0$.

Let $(A_i \times B_i)_{i \in \mathbb{N}}$ be a countable disjoint collection of sets in \mathcal{B} and suppose there is $A \times B \in \mathcal{B}$ such that

$$A \times B = \bigcup_{i=1}^{\infty} (A_i \times B_i).$$

Then

$$\mathbf{1}_A(x)\mathbf{1}_B(y) = \mathbf{1}_{A \times B}(x, y) = \sum_{i=1}^{\infty} \mathbf{1}_{A_i \times B_i}(x, y) = \sum_{i=1}^{\infty} \mathbf{1}_{A_i}(x)\mathbf{1}_{B_i}(y).$$

Applying the Monotone Convergence Theorem to the variable x and integrating we obtain

$$\mu_1(A)\mathbf{1}_B(y) = \sum_{i=1}^{\infty} \mu_1(A_i)\mathbf{1}_{B_i}(y).$$

By the same reason we have

$$\mu(A \times B) = \mu_1(A)\mu_2(B) = \sum_{i=1}^{\infty} \mu_1(A_i)\mu_2(B_i).$$

Finally, because μ_1 and μ_2 are σ -finite, we can take $(A_i)_{i \in \mathbb{N}} \subset \mathcal{F}_1$ and $(B_j)_{j \in \mathbb{N}} \subset \mathcal{F}_2$ such that $\mu_1(A_i) < \infty$ and $\mu_2(B_j) < \infty$ and

$$V_1 \times V_2 = \bigcup_{i,j=1}^{\infty} (A_i \times B_j),$$

hence we have that μ are σ -finite. Because of this, the Carathéodory's Theorem 2.2.1 implies that μ has an unique extension $\mu_1 \times \mu_2$ to $\mathcal{F}_1 \times \mathcal{F}_2$ and this measure is σ -finite. \square

As we have defined the product measure $\mu_1 \times \mu_2$, we now want to compute integrals with respect to $\mu_1 \times \mu_2$. Fortunately, we can compute it easily using simpler integrals. This is the content of *Fubini's Theorem*.

Theorem 2.4.7 (Fubini's Theorem). *Let $(V_1 \times V_2, \mathcal{F}_1 \times \mathcal{F}_2, \mu_1 \times \mu_2)$ be a product space. Let also $f : V_1 \times V_2 \rightarrow \mathbb{R}$ be a measurable function. If either $f \geq 0$ or f is integrable with respect $\mu_1 \times \mu_2$, then*

$$g(x) := \int_{V_2} f(x, y) \, d\mu_2(dy)$$

exists almost surely, it is integrable and

$$\int_{V_1} g(x) \, d\mu_1(dx) = \int_{V_1 \times V_2} f(x, y) \, d\mu_1 \times \mu_2(dx, dy).$$

Remark 2.4.6. Fubini's Theorem states that

$$\int_{V_1 \times V_2} f(x, y) \, d\mu_1 \times \mu_2(dx, dy) = \int_{V_1} \left(\int_{V_2} f(x, y) \, d\mu_2(dy) \right) d\mu_1(dx).$$

But the symmetry in V_1 and V_2 also implies that

$$\int_{V_1 \times V_2} f(x, y) \, d\mu_1 \times \mu_2(dx, dy) = \int_{V_2} \left(\int_{V_1} f(x, y) \, d\mu_1(dx) \right) d\mu_2(dy).$$

Corollary 2.4.1. *Let $h(x) = \prod_{i=1}^n f_i(x_i)$ be an integrable or positive function in $(\mathbb{R}^n, \mathcal{B}(\mathbb{R}^n), \mu)$, where $\mu = \prod_{i=1}^n \mu_i$, then*

$$\int_{\mathbb{R}^n} h \, d\mu = \prod_{k=1}^n \int_{\mathbb{R}} f_k \, d\mu_k.$$

We need this corollary to compute expected value of independent random variables.

2.5 Computing Integrals

In this section we present some methods to compute integrals. The first one is the relation between Riemann Integral, which we can compute, and the Lebesgue Integral. The second is the Change of Variables Formula, which is related to the Change of Variables in the Riemann Integral. Later, we present the relation between weak derivative and Riemann integration by parts.

2.5.1 Riemann Integral

Let A be a compact set in \mathbb{R}^n and $f : A \rightarrow \mathbb{R}$ be a bounded function. Let us denote the Riemann integral of f over A by $I(f, A)$, to not cause any confusion with the Lebesgue Integral $\int_A f \, d\lambda$. We know, by an important result in Analysis, that $I(f, A)$ exists if and only if the set of discontinuities of f is a Lebesgue null set. Also, we know that

$$I(f, A) = \sup \left\{ \sum_{i=1}^n f(x_i^*) \lambda(P_i) \right\}$$

where the supremum is over $n \in \mathbb{N}$ and all finite families $P = (P_i)_{i=1}^n$ of disjoint rectangles sets such that $\bigcup_{i=1}^n P_i \subset A$ and $x_i^* \in P_i$ is any point of P_i .

Theorem 2.5.1. *Let $f : A \rightarrow \mathbb{R}$ be a bounded function and $A \subset \mathbb{R}^n$ a compact set. If $I(f, A)$ exists, then f is measurable, integrable and also*

$$I(f, A) = \int_A f \, d\lambda.$$

Proof. For a proof see [Folland \(2013\)](#). □

We can extend this result as follows.

Corollary 2.5.1. *Let $A_k \subset \mathbb{R}^n$ be compact sets so that $A_k \subseteq A_{k+1}$ for all $k \in \mathbb{N}$ and*

$$\bigcup_{k=1}^{\infty} A_k = \mathbb{R}^n.$$

Let $f : \mathbb{R}^n \rightarrow \mathbb{R}$ be Riemann integrable in A_k for all k , and Lebesgue integral or positive. Suppose also there is $c \in \mathbb{R}$ such that

$$\lim_{k \rightarrow \infty} I(f, A_k) = c,$$

then

$$\int_{\mathbb{R}^n} f \, d\lambda = c.$$

Proof. By Theorem 2.5.1, we have that

$$I(f, A_k) = \int_{A_k} f \, d\lambda.$$

If f is integrable, then $|f \mathbf{1}_{A_k}| \leq |f|$ and $f \mathbf{1}_{A_k} \rightarrow f$, when $k \rightarrow \infty$. Therefore, by the Dominated Convergence Theorem we have that

$$I(f, A_k) \rightarrow \int_{\mathbb{R}^n} f \, d\lambda.$$

If $f \geq 0$, then $f \mathbf{1}_{A_k} \nearrow f$. Consequently, the Monotone Convergence Theorem implies

$$I(f, A_k) \nearrow \int_{\mathbb{R}^n} f \, d\lambda,$$

and the corollary is proved. □

A simple application of this result allows us to prove that the Gaussian density in Example 2.3.5 is in fact a density.

Corollary 2.5.2. *Let f be a Gaussian density with parameters $\mu \in \mathbb{R}^n$ and positive definite matrix Σ (see Example 2.3.5), then*

$$\int_{\mathbb{R}^n} f \, d\lambda = 1.$$

Proof. We know by elementary calculus that the Improper Riemann Integral of f exists and it is equal to 1. Since f is positive, Corollary 2.5.1 leads to the result. □

Therefore, we can use all the Riemann Integral tools to compute integrals.

2.5.2 Change of Variables

We will next introduce the *Change of Variables Formula*. This is a well-known way to generalize the substitution rule in the Riemann Integral, but for the Lebesgue Integral instead. Let us state the theorem.

Theorem 2.5.2. *Let $(\Omega, \mathcal{F}, \mathbb{P})$, be a Probability Space and (V, \mathcal{B}) a Measurable Space. Suppose $X : \Omega \rightarrow V$ is a random element with distribution $\mathbb{P}_X(A) := \mathbb{P}(X \in A)$, with $A \in \mathcal{B}$ and $f : V \rightarrow \mathbb{R}$ is measurable. If $f \geq 0$ or it is integrable with respect to \mathbb{P}_X , then*

$$\mathbb{E}[f(X)] = \int_V f(x) d\mathbb{P}_X(dx).$$

Remark 2.5.1. We can also change the condition that f is integrable with respect to \mathbb{P}_X to $f(X)$ is integrable with respect to \mathbb{P} . Also, we can assume that the original space is not a probability space, but a σ -finite space and the random variable X is a measurable function g .

Remark 2.5.2. The interpretation of this theorem is as follow. In order to compute

$$\mathbb{E}[f(X)] := \int_{\Omega} f(X) d\mathbb{P},$$

we do not need to compute the right-hand side, but instead we just have to compute it by using the equality

$$\mathbb{E}[f(X)] = \int_V f(x) d\mathbb{P}_X(dx).$$

Proof. See [Durrett \(2019\)](#). □

We have the following corollary.

Corollary 2.5.3. *Let X, Y be two random vectors in \mathbb{R}^n and \mathbb{R}^m , respectively, with joint distribution $\mathbb{F}_{X,Y}$. Let also $f : \mathbb{R}^n \rightarrow \mathbb{R}$ and $g : \mathbb{R}^m \rightarrow \mathbb{R}$. Suppose $\mathbb{E}|f(X)g(Y)| < \infty$ or both functions are positive, then the following are true.*

1. *We have that*

$$\mathbb{E}[f(X)g(Y)] = \int_{\mathbb{R}^n \times \mathbb{R}^m} f(x)g(y) d\mathbb{F}_{X,Y}(x, y); \text{ and}$$

2. *If X, Y are independent, then*

$$\mathbb{E}[f(X)g(Y)] = \mathbb{E}[f(X)]\mathbb{E}[g(Y)].$$

If X has a density, it is in fact easier to use the following corollary.

Corollary 2.5.4. *Suppose X is a random vector with density f . Let $H : \mathbb{R}^n \rightarrow \mathbb{R}$ so that Hf is Lebesgue integrable or $H \geq 0$, then*

$$\mathbb{E}[H(X)] = \int_{\mathbb{R}^n} H(x)f(x) \, d\lambda(dx).$$

In general, we will use this formula to compute integrals and expected values. For instance, we have the following corollary.

Corollary 2.5.5. *Let X be a random vector with density f and let $H : \mathbb{R}^n \rightarrow \mathbb{R}^n$ be an invertible affine transformation, say, $H(x) = b + Tx$, where T is an invertible matrix and $b \in \mathbb{R}^n$. Then the density of $H(X)$ is equal to $g(x) = \frac{1}{|T|}f(H^{-1}x)$.*

The final example of this subsection is about the Gaussian r.v.

Example 2.5.1. Let $X \sim \mathcal{N}(0, \text{Id})$, $\mu \in \mathbb{R}^n$ and Σ be a positive definite matrix. As Σ has a square root, we have that

$$\mu + \Sigma^{1/2}X \sim \mathcal{N}(\mu, \Sigma).$$

This result is a special case of Corollary 2.5.5. Moreover, $Y \sim \mathcal{N}(\mu, \Sigma)$ if and only if there is a $X \sim \mathcal{N}(0, \text{Id})$ so that $Y = \mu + \Sigma^{1/2}X$, therefore, all Gaussian random variables or vectors are generated by $\mathcal{N}(0, 1)$ and $\mathcal{N}(0, \text{Id})$, respectively.

2.5.3 The Weak Derivative

In this subsection we extend the Integration by Parts idea for the Lebesgue Integral. Let us first recall the method.

Theorem 2.5.3. *Let $f, g \in C^1([a, b])$. Then*

$$I(f'g, [a, b]) = fg \Big|_a^b - I(fg', [a, b]),$$

*and this formula is called **integration by parts**.*

Before generalizing, suppose that $g \in C_c^1(\mathbb{R})$, that is, $\{x : g(x) \neq 0\}$ is compact, then

$$I(f'g, (-\infty, \infty)) = -I(f, g', (-\infty, \infty)).$$

Theorem 2.5.3 can easily be extended to \mathbb{R}^n . Let $\alpha \in \mathbb{N}^n$, which means that α is a **multi-index**,

$$|\alpha| := \sum_{i=1}^n \alpha_i,$$

and

$$\partial^\alpha g = \frac{\partial^{|\alpha|} g}{\partial_1^{\alpha_1} \dots \partial_n^{\alpha_n}}.$$

We thus have the following corollary.

Corollary 2.5.6. *Let $f \in C^\alpha(\mathbb{R}^n)$ and $g \in C_c^\alpha(\mathbb{R}^n)$, then*

$$I(\partial^\alpha f g, (-\infty, \infty)^n) = (-1)^{|\alpha|} I(f \partial^\alpha g, (-\infty, \infty)^n).$$

This last result will inspire us to define the *weak derivative*. Let $C_c^\infty(\mathbb{R}^n)$ be the space of all compact supported C^∞ functions ψ endowed with the uniform topology, that is, $\psi_n \rightarrow \psi$ if for all multi-index α and all compact $K \subset \mathbb{R}^n$, we have $\partial^\alpha \psi_n \rightarrow \partial^\alpha \psi$ in the uniform metric $d(f, g) = \sup_{x \in \mathbb{R}^n} |f(x) - g(x)|$. Let also $L_{loc}(\mathbb{R}^n)$ be the space of functions f such that $\int_K |f| d\lambda < \infty$ for all compact sets K . We define the weak derivative as follows.

Definition 2.5.1. Let $f \in L^1 = L^1(\mathbb{R}^n, \mathcal{B}(\mathbb{R}^n), \lambda)$. The **weak derivative of order α** of f is any $g \in L^1$ such that

$$\int_{\mathbb{R}^n} g \psi d\lambda = (-1)^{|\alpha|} \int_{\mathbb{R}^n} f \partial^\alpha \psi d\lambda,$$

for all $\psi \in C_c^\infty(\mathbb{R}^n)$. We denote the weak derivative by $\partial^\alpha f$.

Remark 2.5.3. The weak derivative is unique almost surely, hence the single notation $\partial^\alpha f$. For more results and properties of the weak derivative, see [Folland \(2013\)](#).

The weak derivative has the following properties.

Lemma 2.5.1. *Let $f, g : \mathbb{R}^n \rightarrow \mathbb{R}$ and suppose they have weak derivatives $\partial^\alpha f$ and $\partial^\alpha g$ for a multi-index α . Then the following are true.*

1. *For $a \in \mathbb{R}$, we have $\partial^\alpha(f + ag) = \partial^\alpha f + a\partial^\alpha g$; and*
2. *If $f \in C^\alpha(\mathbb{R}^n)$, then the weak derivative agrees with the true derivative of f .*

Our final corollary is the following.

Corollary 2.5.7. *Let $f, \psi \in C^1(\mathbb{R})$ such that $\psi, \psi' \in L^1(\lambda)$ and $f\psi', f'\psi \in L^1(\lambda)$, then*

$$\int_{\mathbb{R}} f \psi' dx = - \int_{\mathbb{R}} f' \psi dx.$$

This result is going to be important when proving the main theorem in the last chapter (see Theorem [6.3.1](#)).

2.6 The Fourier Transform and Moments

In this section, we will explore four related topics. The first one is the convolution rule for distributions. The second is the definition of moments of a random variable. Finally we have the Generating Function and the Fourier Transform.

2.6.1 The Convolution Rule

Given two independent random variables X and Y , what is the distribution of $X + Y$? The convolution rule gives the desired formula.

Definition 2.6.1. Let \mathbb{F}_i , for $i = 1, 2$, be two distributions, then we define the **convolution function** $\mathbb{F}_1 * \mathbb{F}_2$ by

$$\mathbb{F}_1 * \mathbb{F}_2(x) := \int_{\mathbb{R}} \mathbb{F}_1(x - y) d\mathbb{F}_2(dy).$$

Lemma 2.6.1. Let \mathbb{F}_i , $i = 1, 2, 3$, be three distributions. Then the following are true.

1. The convolution $\mathbb{F}_1 * \mathbb{F}_2$ is a distribution;
2. It is commutative: $\mathbb{F}_1 * \mathbb{F}_2 = \mathbb{F}_2 * \mathbb{F}_1$; and
3. It is distributive: $(\mathbb{F}_1 * \mathbb{F}_2) * \mathbb{F}_3 = \mathbb{F}_1 * (\mathbb{F}_2 * \mathbb{F}_3)$.

The importance of the convolution in Probability Theory lies in the following theorem.

Theorem 2.6.1. Let X and Y be two independent random variables with distribution \mathbb{F}_X and \mathbb{F}_Y , then the distribution of the sum $X + Y$ is $\mathbb{F}_X * \mathbb{F}_Y$.

Proof. Let $\mathbb{P}_{X,Y} = \mathbb{P}_X \times \mathbb{P}_Y$ be their joint distribution. Then Theorem 2.5.2 implies that

$$\mathbb{F}_{X+Y}(z) := \mathbb{P}(X + Y \leq z) = \int_{\mathbb{R}} \mathbf{1}_{(-\infty, z]}(x + y) d(\mathbb{P}_X(dx) \times \mathbb{P}_Y(dy)).$$

Since $\mathbf{1}_{(-\infty, z]}$ is a positive function, we can apply Fubini's Theorem, hence

$$\int_{\mathbb{R}} \mathbf{1}_{(-\infty, z]}(x + y) d(\mathbb{P}_X(dx) \times \mathbb{P}_Y(dy)) = \int_{\mathbb{R}} \left(\int_{\mathbb{R}} \mathbf{1}_{(-\infty, z]}(x + y) d\mathbb{P}_X(dx) \right) d\mathbb{P}_Y(dy).$$

The inner integral is equal to $\mathbb{F}_X(z - y)$, therefore we have

$$\mathbb{F}(z) = \int_{\mathbb{R}} \mathbb{F}_X(z - y) d\mathbb{F}_Y(y) = \mathbb{F}_X * \mathbb{F}_Y(z).$$

□

In particular, we have the following corollary.

Corollary 2.6.1. Let $X_i \sim \mathcal{N}(0, \Sigma_i)$, $i = 1, 2$, be two independent Gaussian r.v. and $\Sigma_1 = \sigma \Sigma_2$, then the convolution of their densities is also a Gaussian density, with parameters 0 and $\Sigma_1 + \Sigma_2$.

2.6.2 Moments

To estimate some properties of X , we need to compute some quantities as $\mathbb{E}[X]$ or $\mathbb{E}[X^2]$. These are called the *moments* of the random variable X . Precisely, we have the following definition.

Definition 2.6.2. Let X be a random variable, then $\mathbb{E}[X^n]$ is called the **n -th moment** of X , provided it exists.

By a change of variables, we can use the distribution of X to compute the moments.

Lemma 2.6.2. Let X be a random variable and \mathbb{F} be its distribution, then

$$\mathbb{E}[X^n] = \int_{\mathbb{R}} x^n d\mathbb{F}.$$

In particular, if X has a density $f : \mathbb{R} \rightarrow \mathbb{R}_+$, then

$$\mathbb{E}[X^n] = \int_{\mathbb{R}} x^n f(x) dx.$$

We now define the covariance matrix and variance of a random variable or random vector.

Definition 2.6.3. Let $X = (X_1, \dots, X_n)$ be a random vector, then its **covariance matrix** is defined as

$$\Sigma(X)_{ij} := \mathbb{E}[(X_i - \mathbb{E}X_i)(X_j - \mathbb{E}X_j)],$$

provided the quantity on the right hand side exists.

Remark 2.6.1. For the case $n = 1$, we define the **variance** as

$$\text{Var}(X) := \mathbb{E}(X - \mathbb{E}X)^2 = \mathbb{E}[X^2] - \mathbb{E}^2[X].$$

The meaning of the variance will be clear from Corollary 2.7.2.

Example 2.6.1. For $X = (X_1, \dots, X_n) \sim \mathcal{N}(\mu, \Sigma)$, we have that $\mathbb{E}[X_i] = \mu_i$ and $\Sigma(X) = \Sigma$. If $\mu = 0$ and $\Sigma = \text{Id}$, then $\mathbb{E}[\|X\|^2] = n$.

Example 2.6.2. Let $X \sim \text{Rad}(p)$, then $\mathbb{E}[X] = 2p - 1$ and $\mathbb{E}[X^2] = 1$, hence

$$\text{Var}(X) = 1 - (2p - 1)^2 = 4p(1 - p).$$

Example 2.6.3. Let $X \sim \text{Unif}([-1, 1])$, then $\mathbb{E}[X] = 0$, and

$$\text{Var}(X) = \mathbb{E}[X^2] = \int_{-1}^1 x^2 dx = 2/3.$$

Definition 2.6.4. Let X, Y be two random variables. We say that X and Y are **uncorrelated** if the covariance of X, Y , defined as

$$\text{cov}(X, Y) := \mathbb{E}(X - \mathbb{E}X)(Y - \mathbb{E}Y)$$

is equal to 0.

It is easy to see that independency is a sufficient condition for uncorrelatedness, but it is not necessary. Let us state some properties of these quantities.

Lemma 2.6.3. *Let $(X_i)_{i=1}^n \subset L^2(\Omega, \mathcal{F}, \mathbb{P})$. Then*

1. *The variance of the sum is*

$$\text{Var}(X_1 + \dots + X_n) = \sum_{i=1}^n \text{Var}(X_i) + \sum_{i < j} \text{cov}(X_i, X_j);$$

2. *The variance of an affine transformation is*

$$\text{Var}(aX_1 + b) = a^2 \text{Var}(X_1);$$

3. *If $X = (X_1, \dots, X_n)$, then $\Sigma(X)$ is positive semidefinite matrix;*

4. *The covariance operator $\text{cov} : L^2(\Omega, \mathcal{F}, \mathbb{P}) \times L^2(\Omega, \mathcal{F}, \mathbb{P}) \rightarrow \mathbb{R}$ is a bilinear operator; and*

5. *If $\mathbb{E}[X_1] = \mathbb{E}[X_2] = 0$, then $\text{cov}(X_1, X_2)$ is the classic inner product in $L^2(\Omega, \mathcal{F}, \mathbb{P})$.*

2.6.3 The Generating Function

In this section we present one important tool in Probability Theory.

Definition 2.6.5. Let X be a random variable. Then the **generating function** of X is

$$\varphi_X(t) := \mathbb{E}[e^{tX}],$$

provided the right hand side exists.

The next theorem provides the relation between the generating function and the moments of a random variable.

Theorem 2.6.2. *Let X be a positive random variable and suppose that, for some $t_0 > 0$, $\varphi_X(t_0)$ exists, then X has all its moments well-defined and*

$$\varphi_X^{(k)}(0) = \mathbb{E}[X^k].$$

Now let us state some its properties.

Lemma 2.6.4. *Let X, Y be random variables and suppose $\varphi_X(t)$ and $\varphi_Y(t)$ exists for all $t \in \mathbb{R}$, then*

1. *The generating function of $aX + b$ is*

$$\varphi_{aX+b}(t) = e^{tb}\varphi_X(at),$$

for all $a, b \in \mathbb{R}$ and $t \in \mathbb{R}$;

2. *If X has density f , then*

$$\varphi_X(t) = \int_{\mathbb{R}} f(x)e^{tx} dx;$$

3. *If X is symmetric, in the sense that $X \stackrel{d}{=} -X$, then $\varphi_X(t) = \varphi_X(-t)$;*

4. *If X, Y are independent, then $\varphi_{X+Y}(t) = \varphi_X(t)\varphi_Y(t)$; and*

5. *φ_X uniquely defines X .*

Let us now give two examples.

Example 2.6.4. It is easy to see that if $X \sim \text{Rad}(p)$, then $\psi_X(t) = pe^t + (1-p)e^{-t}$. In case $p = 1/2$, the real inequality

$$\frac{e^t + e^{-t}}{2} \leq e^{t^2/2},$$

implies that $\psi_{\lambda X}(t) \leq e^{\lambda^2 t^2/2}$.

Example 2.6.5. Let $X \sim \mathcal{N}(0, 1)$, then

$$\psi_X(t) = \int_{\mathbb{R}} (2\pi)^{-1/2} e^{-x^2/2} e^{tx} dx = e^{t^2/2} \int_{\mathbb{R}} (2\pi)^{-1/2} e^{-(x-t)^2/2} dx,$$

hence $\psi_X(t) = e^{t^2/2}$ and $\psi_{\lambda X}(t) = e^{\lambda^2 t^2/2}$.

Remark 2.6.2. Note that both Generating Function of the Rademacher and the Gaussian are bounded by the same function $e^{\lambda^2 t^2/2}$. This bound defines a *Subgaussian Random Variable* (see Definition 2.7.2). We will explore these two cases of Subgaussian Random Variables in Chapters 5 and 6.

2.6.4 The Fourier Transform and The Characteristic Function

Let $L^2(\mathbb{R}, \mathcal{B}(\mathbb{R}), \lambda)$ be the Hilbert Space of all complex-valued functions $f : \mathbb{R} \rightarrow \mathbb{C}$ with norm

$$\|f\|_{L^2}^2 := \int_{\mathbb{R}} |f|^2 dx < \infty,$$

and inner product

$$\langle f, g \rangle := \int_{\mathbb{R}} f \bar{g} dx.$$

We can define the *Fourier Transform* in this space as follows.

Definition 2.6.6. Let $\mathcal{F} : L^2(\mathbb{R}, \mathcal{B}(\mathbb{R}), \lambda) \rightarrow L^2(\mathbb{R}, \mathcal{B}(\mathbb{R}), \lambda)$ be such that

$$[\mathcal{F}(f)](t) = \hat{f}(t) := \int_{\mathbb{R}} f(x) e^{-2\pi i x t} dx,$$

then \mathcal{F} is called the **Fourier transform**.

Remark 2.6.3. We also define the Fourier Transform for *signed measures* μ . By a **signed measure** we mean a function $\mu : \mathcal{B}(\mathbb{R}) \rightarrow \mathbb{R}$ such that

1. The empty set has measure zero: $\mu(\emptyset) = 0$; and
2. Let $(A_i)_{i=1}^{\infty} \subset \mathcal{B}(\mathbb{R})$ be a countable family of disjoint sets, then

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

where the series converges absolutely.

Let μ be a signed measure with

$$|\mu| := |\mu(\mathbb{R})| < \infty,$$

then

$$[\mathcal{F}(\mu)](t) := \int_{\mathbb{R}} e^{-2i\pi t x} d\mu(dx).$$

Note that the Fourier Transform of f is the Fourier Transform of the signed measure defined by the Radon-Nikodym derivative $\frac{d\mu}{d\lambda} = f$.

And we can also define the *Characteristic Function* of a random variable X .

Definition 2.6.7. Let X be a r.v, then its **characteristic function** is the function $\psi : \mathbb{R} \rightarrow \mathbb{C}$ such that

$$\psi_X(t) := \mathbb{E}[e^{itX}].$$

Notice that if $X \sim \mu$, then

$$[\mathcal{F}(\mu)]\left(\frac{-t}{2\pi}\right) = \psi_X(t).$$

The Characteristic Function of a Random Variable X is, rather than a function of X , a function of \mathbb{F}_X .

Let us now state some properties of the Fourier Transform.

Lemma 2.6.5. *Let $f, g \in L^2(\mathbb{R}, \mathcal{B}(\mathbb{R}), \lambda)$ and $a \in \mathbb{R}$. Then*

1. *It is well-defined:*

$$|[\mathcal{F}f](x)| \leq \int_{\mathbb{R}} |f| < \infty,$$

for almost all $x \in \mathbb{R}$;

2. *It is linear: $\mathcal{F}(f + ag) = \mathcal{F}f + a\mathcal{F}f$;*

3. *It has the delay property: if $g(x) := f(x - x_0)$, then*

$$[\mathcal{F}g](t) = e^{-2\pi i t x_0} [\mathcal{F}]f(t);$$

4. *It has the rotation property: if $g(x) = e^{2\pi i x t_0} f(x)$, then*

$$[\mathcal{F}g](t) = [\mathcal{F}f](t - t_0);$$

5. *The convolution property holds: if $h = g * f$, then $h \in L^2(\mathbb{R}, \mathcal{B}(\mathbb{R}), \lambda)$ and $\hat{h} = \hat{g}\hat{f}$;*

6. *It is bijective;*

7. *It is an isometry:*

$$\langle f, g \rangle = \langle \hat{f}, \hat{g} \rangle;$$

8. *It obeys **Parseval's Identity**: $\|f\|_{L^2} = \|\hat{f}\|_{L^2}$;*

9. *Its second and fourth power follows the rule: $[\mathcal{F}\hat{f}](t) = f(-t)$, and $\mathcal{F}^{(4)} = \text{Id}$; and*

10. *If $f \in C^n(\mathbb{R})$, then*

$$[\mathcal{F}(f^{(k)})](t) = (2\pi i t)^k \hat{f}(t),$$

for all $k \leq n$.

Proof. The reader can find a proof of this lemma in [Brémaud \(2014\)](#). □

We can prove some of these properties for the characteristic function.

Lemma 2.6.6. *Let X, Y be two random variables and $a, b \in \mathbb{R}$, then*

1. If $Y = aX + b$, then $\psi_Y(t) = e^{bit}\psi_X(at)$;
2. The characteristic function is always bounded: $\psi_X(t) \leq 1$;
3. The characteristic function is uniformly continuous;
4. If X is integrable, then $\psi_X \in C^1(\mathbb{R})$ and

$$\psi_X^{(n)}(t) = \mathbb{E}[iX e^{itX}]; \text{ and}$$

5. If $X \in L^p(\Omega, \mathcal{F}, \mathbb{P})$, then $\psi_X \in C^p(\mathbb{R})$ and

$$\psi_X^{(k)}(t) = \mathbb{E}[(iX)^k e^{itX}],$$

for all $k \leq p$.

The following lemma proves the uniqueness of the characteristic function and using it we can prove a decomposition of ψ_{X+Y} when X and Y are independent.

Lemma 2.6.7. *If X and Y have the same characteristic function, then $\mathbb{F}_X = \mathbb{F}_Y$.*

Proof. For a proof, see [Shiryaev \(2016\)](#). □

Corollary 2.6.2. *Let X, Y be two random variables. Then they are independent if and only if $\psi_{X+Y} = \psi_X \psi_Y$.*

Proof. It is easy to see that $\psi_{X+Y} = \psi_X \psi_Y$ if X, Y are independent. The other direction is a consequence of the Uniqueness of the Characteristic Function. □

2.7 Inequalities in Probability

This is the most important section in this chapter, where we state some useful inequalities such as Jensen, Markov and Chernoff's Inequalities.

2.7.1 Convex Function and Jensen Inequality

Definition 2.7.1. Let $f : \mathbb{R}^n \rightarrow \mathbb{R}$. We say that f is **convex** if, for every $t \in [0, 1]$ and $x, y \in \mathbb{R}^n$ we have

$$f(tx + (1-t)y) \leq tf(x) + (1-t)f(y).$$

We also say that f is **strictly convex** if the strict inequality holds whenever $t \in (0, 1)$ and $x \neq y$.

Remark 2.7.1. In some cases, we will consider a small domain $A \subset \mathbb{R}^n$. In this case, we need to assume that A is a **convex set**, that is, if $x, y \in A$ and $\lambda \in [0, 1]$, then

$$\lambda x + (1 - \lambda)y \in A.$$

From a geometrical point of view, a function f is convex if the line segment from $(x, f(x))$ to $(y, f(y))$ lies above the graph of $f(t)$ in $t \in [x, y]$ (see example below).

Example 2.7.1. Let $f : [0, \infty) \rightarrow \mathbb{R}$ be such that

$$f(x) := x^3,$$

then f is convex. The graph of f is shown in Figure 1, as well as the line between two fixed points of the graph.

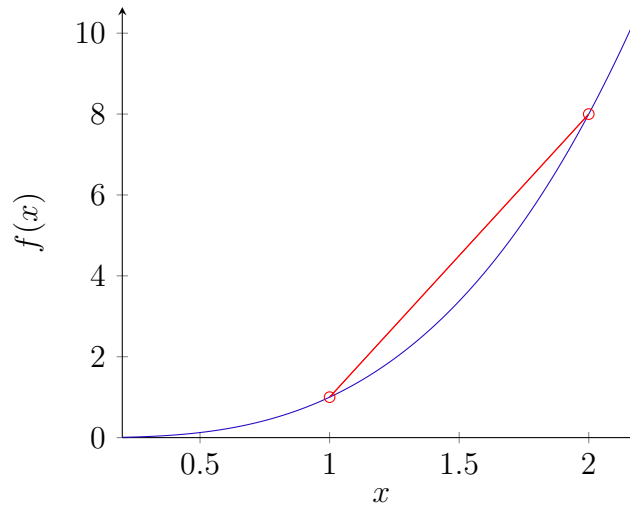


Figure 1 – A convex function in the interval $[0, 2]$, in blue, and a straight line connecting the points $(1, 1)$ and $(2, 8)$, in red.

For $f \in C^1(\mathbb{R}^n)$ we have the following.

Lemma 2.7.1. *If $f \in C^1(\mathbb{R}^n)$, then f is convex if and only if for every $x, y \in \mathbb{R}^n$ we have*

$$f(y) \geq f(x) + \langle \nabla f(x), x - y \rangle.$$

The strict inequality holds for $y \neq x$ whenever f is strictly convex.

Proof. For a proof, we recommend [Boyd and Vandenberghe \(2004\)](#). □

This lemma states that the graph of the function lies above the tangent plane in $(x, f(x))$. See the example below.

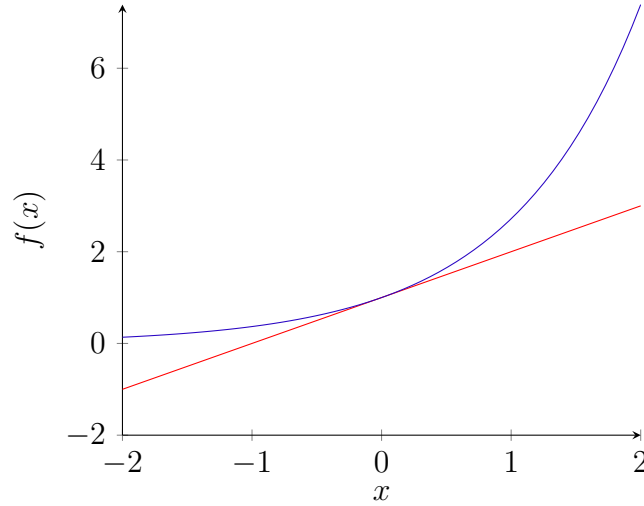


Figure 2 – The function e^x and the tangent at $x = 0$, namely, $y = x + 1$.

Example 2.7.2. Let $\exp : \mathbb{R} \rightarrow \mathbb{R}_+$ be the exponential e^x , then it is convex and its graph is shown in Figure 2, as well as its tangent at $x = 0$.

We can also generalize the gradient condition using the notion of *subgradient*.

Lemma 2.7.2. Let $f : \mathbb{R}^n \rightarrow \mathbb{R}$ be a convex function, then for every $x \in \mathbb{R}^n$ there is at least one **subgradient** at the point x , that is, one vector $g \in \mathbb{R}^n$ such that

$$f(y) \geq f(x) + \langle g, y - x \rangle.$$

For a strictly convex function, there is at least of subgradient such that the strict inequality holds whenever $y \neq x$.

Proof. This lemma is a consequence of the Support Hyperplane Theorem which we will not prove here. See [Boyd and Vandenberghe \(2004\)](#). \square

Remark 2.7.2. The set of all subgradients in x is denoted by $\partial f(x)$.

If f is twice differentiable, then we have a second order condition for convexity.

Lemma 2.7.3. If $f \in C^2(\mathbb{R}^n)$, then f is convex if and only if for all $x \in \mathbb{R}^n$ we have

$$\text{Hess}\{f\}(x) \succeq 0,$$

where \succeq is the partial order in the cone of positive semidefinite matrices. If f is strictly convex, then $\text{Hess}\{f\}(x)$ is positive definite.

We can now prove *Jensen's Inequality*.

Theorem 2.7.1 (Jensen's Inequality). *Let X be a random vector and $f : \mathbb{R}^n \rightarrow \mathbb{R}$ a convex function. If $f(X)$ is integrable and X is integrable, then*

$$\mathbb{E}[f(X)] \geq f(\mathbb{E}[X]).$$

Also, if f is strictly convex, then equality holds if and only if X is constant almost surely.

Proof. Let $x, y \in \mathbb{R}^n$ and $\omega \in \Omega$. Let $g \in \partial f(x)$, then

$$f(y) \geq f(x) + \langle g, y - x \rangle.$$

Now fix $y = X(\omega)$ and $x = \mathbb{E}[X]$, then we obtain

$$f(X(\omega)) \geq f(\mathbb{E}[X]) + \langle g, X(\omega) - \mathbb{E}[X] \rangle.$$

Since $\mathbb{E}[X - \mathbb{E}(X)] = 0$, we obtain after taking expected value:

$$\mathbb{E}[f(X)] \geq f(\mathbb{E}[X]).$$

For a strictly convex property, let g be a subgradient such that the equality in the Lemma 2.7.2 holds if and only if $y = x$. Again we obtain

$$f(X(\omega)) \geq f(\mathbb{E}[X]) + \langle g, X(\omega) - \mathbb{E}[X] \rangle.$$

From the properties of the expected value, if $X \geq Y$ and $\mathbb{E}[X] = \mathbb{E}[Y]$, then $X = Y$ almost surely, therefore we obtain that equality in Jensen's Inequality holds if and only if

$$f(X) = f(\mathbb{E}[X]) + \langle g, X - \mathbb{E}[X] \rangle \text{ a.s.}$$

Finally, by the subgradient property of g , we have that $X = \mathbb{E}[X]$ a.s. □

This theorem is an useful tool to prove a lot of results. We will point out a few of them.

Corollary 2.7.1. *Let $t \in [1, \infty]$ and $X \in L^t := L^t(\Omega, \mathcal{F}, \mathbb{P})$, then $X \in L^s$ for all $s \leq t$ and*

$$\|X\|_{L^s} \leq \|X\|_{L^t}.$$

Remark 2.7.3. Corollary 2.7.1 holds even when X^p is not integrable, and in this case we have the trivial inequality

$$\|X\|_{L^k} \leq \infty.$$

Proof. First, let $0 < s < t$, then

$$|x|^s \leq 1 + |x|^t,$$

since if $|x| > 1$, then $|x|^s \leq |x|^t$, otherwise $|x|^s \leq 1$. Therefore, if X^p is integrable, so is X^k for all $k \leq p$. Now, take $\psi(x) = |x|^{t/s}$, then

$$\psi''(x) = \frac{t}{s} \frac{t-s}{s} |x|^{t/s-2} \geq 0,$$

hence ψ is convex. Because $\psi(|X|^s) = |X|^t$ and $|X|^s$ is integrable, we can apply Jensen's Inequality for $|X|^s$ and obtain

$$\mathbb{E}[\psi(|X|^s)] \geq \psi(\mathbb{E}|X|^s),$$

that is,

$$\mathbb{E}|X|^t \geq |\mathbb{E}|X|^s|^{t/s},$$

hence $\|X\|_{L^s} \leq \|X\|_{L^t}$. □

Remark 2.7.4. This inequality is equivalent to Holder's Inequality in the case of finite Measure Space (M, \mathcal{B}, μ) , say, if $f \in L^p(M, \mathcal{B}, \mu)$, $g \in L^q(M, \mathcal{B}, \mu)$ and p and q are **conjugate exponents**, that is

$$\frac{1}{p} + \frac{1}{q} = 1,$$

then $fg \in L^1(M, \mathcal{B}, \mu)$ and

$$\|fg\|_{L^1} \leq \|f\|_{L^p} \|g\|_{L^q}.$$

Let us state this in the case of a probability space.

Theorem 2.7.2. *Let $(\Omega, \mathcal{B}, \mathbb{P})$ be a probability space and $L^p := L^p(\Omega, \mathcal{B}, \mathbb{P})$. Then Holder's Inequality and Corollary 2.7.1 are equivalent.*

Proof. Let $X \in L^s$ and $r \leq s$. Notice that the constant function $\mathbf{1} \in L^q$, for all $q \in [1, \infty]$. Now let $t \geq 0$ be the conjugate exponent of s/r , then $X^r \in L^{s/r}$ and Holder's Inequality implies that

$$\|X^r \cdot \mathbf{1}\|_{L^1} \leq \|X^r\|_{L^{s/r}} \|\mathbf{1}\|_{L^t},$$

that is,

$$\int_{\Omega} |X|^r d\mathbb{P} \leq \left(\int_{\Omega} |X|^s d\mathbb{P} \right)^{r/s}.$$

Taking the r -root leads to Corollary 2.7.1.

Suppose now q and p are conjugate exponents and $X \in L^p$ and $Y \in L^q$. Since Y^q is integrable, we can define a probability measure \mathbb{Q} such that

$$\frac{d\mathbb{Q}}{d\mathbb{P}} = \frac{|Y|^q}{\int_{\Omega} |Y|^q d\mathbb{P}}.$$

Now let $Z = XY^{1-q}$. Using Corollary 2.7.1 to Z and the measure \mathbb{Q} we obtain

$$\|Z\|_{L^1(\Omega, \mathcal{B}, \mathbb{Q})} \leq \|Z\|_{L^p(\Omega, \mathcal{B}, \mathbb{Q})}. \quad (2.1)$$

Moreover, we have that

$$\int_{\Omega} |XY| d\mathbb{P} = \int_{\Omega} |Y|^q d\mathbb{P} \int_{\Omega} \left(|X| |Y|^{1-q} \frac{|Y|^q}{\int_{\Omega} |Y|^q d\mathbb{P}} \right) d\mathbb{P},$$

that is,

$$\int_{\Omega} |XY| d\mathbb{P} = \int_{\Omega} |Y|^q d\mathbb{P} \int_{\Omega} |Z| d\mathbb{Q}.$$

Using Inequality 2.1, we obtain

$$\int_{\Omega} |XY| d\mathbb{P} \leq \|Z\|_{L^p(\Omega, \mathcal{B}, \mathbb{Q})} \int_{\Omega} |Y|^q d\mathbb{P}.$$

Now, the $L^p(\Omega, \mathcal{B}, \mathbb{Q})$ norm of Z can be compute as

$$\|Z\|_{L^p} = \left(\int_{\Omega} \frac{|X|^p |Y|^{p(1-q)} |Y|^q}{\int_{\Omega} |Y|^q d\mathbb{P}} d\mathbb{P} \right)^{1/p}.$$

Since $p(1-q) + q = 0$, we have

$$\|Z\|_{L^p} = \frac{\|X\|_{L^p}}{\left(\int_{\Omega} |Y|^q d\mathbb{P} \right)^{1/p}}.$$

Using

$$\frac{1}{q} = 1 - \frac{1}{p},$$

we also obtain

$$\|XY\|_{L^1} \leq \|Y\|_{L^q} \|X\|_{L^p},$$

which is Holder's Inequality.

□

We also have the following example which we will use later.

Example 2.7.3. Let $\phi : A \subseteq \mathbb{R} \rightarrow \mathbb{R}$ be a strictly convex function, where A is a convex set (typically $A = \mathbb{R}$ or $A = \mathbb{R}_+$). Then the ϕ -**Entropy** is defined as

$$\text{Ent}_\phi(X) := \mathbb{E}[\phi(X)] - \phi(\mathbb{E}[X]).$$

Jensen's Inequality states that $\text{Ent}_\phi(X) \geq 0$ and $\text{Ent}_\phi(X) = 0$ if and only if X is constant almost surely. This is an useful quantity to describe how concentrated a random variable is around its mean (see Herbst's Method 4.5.6).

Two particular cases are when $\phi : \mathbb{R} \rightarrow \mathbb{R}$ such that $\phi(x) = x^2$, then

$$\text{Ent}_\phi(X) = \text{Var}(X),$$

and $\phi(x) : \mathbb{R}_+ \rightarrow \mathbb{R}$ such that $\phi(x) = x \log x$, which later we will prove of its properties, see Section 4.3.

2.7.2 Markov's Inequality

Theorem 2.7.3 (Markov's Inequality). *Let $\psi : \mathbb{R} \rightarrow \mathbb{R}$ such that $\psi \geq 0$. Let $A \in \mathcal{B}(\mathbb{R})$ and let*

$$p_A := \min\{\psi(x) : x \in A\}.$$

Then, for a random variable X such that $\psi(X)$ is integrable we have

$$p_A \mathbb{P}(X \in A) \leq \mathbb{E}[\psi(X)].$$

Remark 2.7.5. The classical statement of Markov's Inequality is the following: let X be a positive integrable r.v., then for all $\lambda > 0$ we have

$$\mathbb{P}(X \geq \lambda) \leq \frac{\mathbb{E}[X]}{\lambda}.$$

Proof. By definition of p_A , we have

$$p_A \mathbf{1}_A(X) \leq \psi(X) \mathbf{1}_A(X) \leq \psi(X).$$

Taking expectation leads to the result. □

This result is simple nevertheless powerful because of its consequences.

Corollary 2.7.2 (Chebyshev's Inequality). *Let X be a square integrable random variable, then for all $\lambda > 0$ we have*

$$\mathbb{P}(|X - \mathbb{E}[X]| \geq \lambda) \leq \frac{\text{Var}(X)}{\lambda^2}.$$

Proof. Let $\psi(x) = (x - \mathbb{E}[X])^2$ and $A = \{x : |x - \mathbb{E}[X]| \geq \lambda\}$, then $p_A = \lambda^2$ and hence

$$\lambda^2 \mathbb{P}(|X - \mathbb{E}[X]| \geq \lambda) \leq \mathbb{E}(X - \mathbb{E}[X])^2,$$

which is the result. \square

Corollary 2.7.3. *Let $X \in L^p(\Omega, \mathcal{F}, \mathbb{P})$, then*

$$\mathbb{P}(|X| \geq t) \leq \frac{\mathbb{E}|X|^p}{t^p}.$$

Proof. Let $\psi(x) = x^p$ and $A = \{x : |x| \geq t\}$, then $p_A = t^p$ and the result follows from Markov's Inequality. \square

2.7.3 Chernoff's Inequality

Chernoff's Inequality, which we will derive below, is just a particular case of Markov's Inequalities, nevertheless it is important to state it separately because it is closely related to our study.

Theorem 2.7.4 (Chernoff's Inequality). *Let X be a random variable and suppose the generating function $\varphi_X(\lambda)$ exists for some $\lambda_0 \in \mathbb{R}_+$, then*

$$\mathbb{P}(X \geq t) \leq e^{-\lambda_0 t} \varphi_X(\lambda_0),$$

for all $t \in \mathbb{R}$.

Remark 2.7.6. A useful expression for Chernoff's inequality can be obtained through the **Log-Generating Function**

$$\phi_X(t) := \log \varphi_X(t),$$

and Chernoff's Inequality is rewritten as

$$\mathbb{P}(X \geq t) \leq e^{-\lambda t + \phi_X(\lambda)}.$$

Proof. Let $\psi(x) = e^{\lambda x}$ and $A = \{x : x \geq t\}$, then $p_A = e^{\lambda t}$, hence Markov's Inequality implies

$$e^{\lambda t} \mathbb{P}(X \in A) \leq \mathbb{E}[e^{\lambda X}],$$

which is the desired inequality. \square

Hence, in order to control how the tail of a random variable goes to zero we need to control its Generating Function. For instance, we have the following corollary.

Corollary 2.7.4. *Let X be a random variable with $\phi_X(\lambda) \leq \frac{\sigma^2 \lambda^2}{2}$, for some $\sigma > 0$ and every $\lambda > 0$, then for all $t \geq 0$ we have*

$$\mathbb{P}(X \geq t) \leq e^{-t^2/(2\sigma^2)}.$$

Proof. Using Chernoff's Inequality, we have

$$\mathbb{P}(X \geq t) \leq e^{-\lambda t + \frac{\sigma^2 \lambda^2}{2}}.$$

The value $\lambda^* = t/\sigma^2$ gives the desired result. \square

Such variables are called *Subgaussian Random Variables* with parameter σ .

Definition 2.7.2. Let X be a r.v. It is called **Subgaussian** with parameter σ if

$$\phi_X(\lambda) \leq \frac{\sigma^2 \lambda^2}{2},$$

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

Example 2.7.4. As we saw in Example 2.6.4, $X \sim \text{Rad}(1/2)$ is Subgaussian with parameter $\sigma = 1$.

Example 2.7.5. $X \sim \mathcal{N}(0, 1)$ is Subgaussian with parameter $\sigma = 1$, as we saw in Example 2.6.5.

Remark 2.7.7. For $X \sim \text{Rad}(1/2)$ or $X \sim \mathcal{N}(0, 1)$ we have $X \stackrel{d}{=} -X$ hence we can also get a bound to $\mathbb{P}(X \leq -t)$. These bounds combined yields

$$\mathbb{P}(|X| \geq t) \leq 2e^{-t^2/2}.$$

We also have the following corollary for sum of independent Subgaussian r.v.

Corollary 2.7.5 (Hoeffding's Inequality). *Let $X = (X_1, \dots, X_n)$ be a random vector with independent subgaussian coordinates with parameters $\sigma_1, \dots, \sigma_n$, respectively and let $v \in \mathbb{R}^n$. Take*

$$\sigma^2 = \sum_{i=1}^n v_i^2 \sigma_i^2,$$

then $\langle X, v \rangle$ is subgaussian with parameter σ , therefore

$$\mathbb{P}\left(\sum_{i=1}^n v_i X_i \geq t\right) \leq e^{-t^2/(2\sigma^2)}.$$

Proof. Since X_i are independent, we have

$$\mathbb{E}[e^{\lambda \langle X, v \rangle}] = \prod_{i=1}^n \mathbb{E}[e^{\lambda v_i X_i}].$$

By the subgaussian property, we obtain

$$\mathbb{E}[e^{\lambda \langle X, v \rangle}] \leq \exp\left(\sum_{i=1}^n \lambda^2 v_i^2 \sigma_i^2 / 2\right) = \exp(\lambda^2 \sigma^2 / 2),$$

and the result follows from Chernoff's Inequality. \square

2.7.4 Inequalities in Hilbert Space

We saw in Remark 2.4.4 that $L^2(\Omega, \mathcal{F}, \mathbb{P})$ is a Hilbert space with $\langle X, Y \rangle = \mathbb{E}[XY]$, hence we can get Cauchy-Schwarz' inequality and Hölder's inequality.

Theorem 2.7.5 (Cauchy-Schwarz' Inequality). *Let X and Y be two random variables with finite second moments, then*

$$[\text{Cov}(X, Y)]^2 \leq \text{Var}(X)\text{Var}(Y).$$

Theorem 2.7.6 (Hölder's Inequality). *Let $L^p := L^p(\Omega, \mathcal{F}, \mathbb{P})$, $X \in L^p$, $Y \in L^q$ and $\frac{1}{p} + \frac{1}{q} = 1$. Then XY is integrable and*

$$\mathbb{E}|XY| \leq \|X\|_{L^p} \|Y\|_{L^q}.$$

We can also get some useful bounds in the Variance using the distance in the Hilbert Space.

Lemma 2.7.4. *Let $X \in L^2 := L^2(\Omega, \mathcal{F}, \mathbb{P})$, then $\mathbb{E}[X]$ is the closest constant element in L^2 to X , that is,*

$$\text{Var}(X) = \mathbb{E}(X - \mathbb{E}[X])^2 = \min_{c \in \mathbb{R}} \mathbb{E}(X - c)^2.$$

Proof. Let us compute $\mathbb{E}(X - c)^2$. We have that

$$\begin{aligned} \mathbb{E}(X - c)^2 &= \mathbb{E}(X - \mathbb{E}[X] + \mathbb{E}[X] - c)^2 \\ &= \mathbb{E}(X - \mathbb{E}[X])^2 + (\mathbb{E}[X] - c)^2 + 2\mathbb{E}(X - \mathbb{E}[X])(\mathbb{E}[X] - c), \end{aligned}$$

that is,

$$\mathbb{E}(X - c)^2 = \text{Var}(X) + (\mathbb{E}[X] - c)^2.$$

Therefore, we have that $\mathbb{E}(X - c)^2 \geq \text{Var}(X)$ and equality holds if and only if $c = \mathbb{E}[X]$. \square

2.8 Conditional Expectation

Let $L^2 = L^2(\Omega, \mathcal{F}, \mathbb{P})$ throughout the section. To define the conditional expectation in a more general framework, let us define a *sub σ -algebra*.

Definition 2.8.1. Let (Ω, \mathcal{F}) be a measurable space. Then $\mathcal{G} \subseteq \mathcal{F}$ is a **sub σ -algebra** of \mathcal{F} if it is a σ -algebra itself.

We now can define what is a \mathcal{G} -measurable r.v.

Definition 2.8.2. Let $\mathcal{G} \subseteq \mathcal{F}$ be a sub σ -algebra of \mathcal{F} . We say that a random variable is **\mathcal{G} -measurable** if

$$X^{-1}(A) \in \mathcal{G},$$

for all $A \in \mathcal{B}(\mathbb{R})$. We denote it by $X \in \mathcal{G}$.

Given a r.v X , we can always construct a sub σ -algebra, called *the σ -algebra generated by X* .

Definition 2.8.3. Let X be a r.v. Then the set

$$\sigma(X) := \{X^{-1}(A) : A \in \mathcal{B}(\mathbb{R})\}$$

is called the **σ -algebra generated by X** .

Now we can define the Conditional Expectation.

Definition 2.8.4. Let $X \in L^2$ and $\mathcal{G} \subseteq \mathcal{F}$ be a sub σ -algebra. The **conditional expected value of X given \mathcal{G}** is any random variable Y such that

1. Y is \mathcal{G} -measurable; and
2. The integral of Y over a \mathcal{G} measurable sets agrees with the the integral of X over the same set:

$$\int_A Y \, d\mathbb{P} = \int_A X \, d\mathbb{P},$$

for all $A \in \mathcal{G}$.

We will denote any such r.v. by $\mathbb{E}[X|\mathcal{G}]$.

Lemma 2.8.1. *Let $X \in L^2$ and $\mathcal{G} \subseteq \mathcal{F}$. Then there is an unique (almost surely) Y such that*

1. Y is \mathcal{G} -measurable; and
2. For all $A \in \mathcal{G}$ we have

$$\int_A Y \, d\mathbb{P} = \int_A X \, d\mathbb{P}.$$

Proof. Let us prove first uniqueness. Let Y_1 and Y_2 such as in the definition. Take $A_n = \{Y_2 > Y_1 + 1/n\}$, then $A_n \in \mathcal{G}$ and

$$\int_{A_n} Y_1 \, d\mathbb{P} = \int_{A_n} X \, d\mathbb{P} = \int_{A_n} Y_2 \, d\mathbb{P},$$

hence

$$0 \geq 1/n\mathbb{P}(A_n),$$

that is, $\mathbb{P}(A_n) = 0$ for all n . Since

$$\mathbb{P}(Y_2 > Y_1) = \mathbb{P}\left(\bigcup_{i=1}^n A_n\right) = 0,$$

we have that $Y_2 \leq Y_1$ almost surely. Likewise, we also obtain $Y_1 \leq Y_2$ almost surely hence $Y_1 = Y_2$ almost surely.

To prove the existence, suppose $X \geq 0$ and let

$$\nu(A) := \int_A X \, d\mathbb{P},$$

for $A \in \mathcal{G}$. Then ν is a measure in the space (Ω, \mathcal{G}) . We also have that

$$\nu \ll \mathbb{P}|_{\mathcal{G}},$$

therefore, by Radon-Nikodym Theorem, there is a \mathcal{G} -measurable function $f \geq 0$ such that

$$\int_A X \, d\mathbb{P} = \nu(A) = \int_A f \, d\mathbb{P},$$

and hence $\mathbb{E}[X|\mathcal{G}] = f$. The general case follows from $X = X^+ - X^-$. \square

The conditional expectation has all the properties of the expectation, but now they hold almost surely, since $\mathbb{E}[X|\mathcal{G}]$ is a random variable.

Lemma 2.8.2. *Let $X, Y \in L^2$ and $\mathcal{G} \subset \mathcal{F}$ be a sub σ -algebra. Then*

1. $\mathbb{E}[(X + aY)|\mathcal{G}] = \mathbb{E}[X|\mathcal{G}] + a\mathbb{E}[Y|\mathcal{G}]$ a.s.;
2. If $X \leq Y$ then $\mathbb{E}[X|\mathcal{G}] \leq \mathbb{E}[Y|\mathcal{G}]$ a.s.;
3. If $X \leq Y$ and $\mathbb{E}[X|\mathcal{G}] = \mathbb{E}[Y|\mathcal{G}]$, then $X = Y$ a.s.;
4. If X, Y are independent, then $\mathbb{E}[X|Y] := \mathbb{E}[X|\sigma(Y)] = \mathbb{E}[X]$ a.s.;
5. If ψ is convex, then $\mathbb{E}[\psi(X)|\mathcal{G}] \geq \psi(\mathbb{E}[X|\mathcal{G}])$ a.s.;
6. If $0 \leq X_n \nearrow X$, then $\mathbb{E}[X_n|\mathcal{G}] \nearrow \mathbb{E}[X|\mathcal{G}]$ a.s.;
7. If $X_n \rightarrow X$ and $|X_n| \leq Y \in L^1$, then $\mathbb{E}[X_n|\mathcal{G}] \rightarrow \mathbb{E}[X|\mathcal{G}]$ a.s.;
8. If $X \in \mathcal{G}$, then $\mathbb{E}[X|\mathcal{G}] = X$ a.s.;
9. If $\mathcal{G}_1 \subseteq \mathcal{G}_2$, then $\mathbb{E}(\mathbb{E}(X|\mathcal{G}_1)|\mathcal{G}_2) = \mathbb{E}(\mathbb{E}(X|\mathcal{G}_2)|\mathcal{G}_1) = \mathbb{E}(X|\mathcal{G}_1)$ a.s (the Tower Property);

10. The operator $\mathbb{E}[\cdot|\mathcal{G}] : L^2 \rightarrow L^2$ is well-defined and it is a projection, that is, $\mathbb{E}(\mathbb{E}[X|\mathcal{G}]) = \mathbb{E}[X]$ and

$$\mathbb{E}[\mathbb{E}(X|\mathcal{G})|\mathcal{G}] = \mathbb{E}[X|\mathcal{G}];$$

11. If $Y \in \mathcal{G}$, then $\mathbb{E}[XY|\mathcal{G}] = Y\mathbb{E}[X|\mathcal{G}]$; and

12. $\mathbb{E}[X|\mathcal{G}]$ minimizes the distance between X and the space of all \mathcal{G} -r.v.

Proof. This lemma can be found in [Durrett \(2019\)](#). □

We can also define the conditional expectation given $Y = y$.

Definition 2.8.5. Let $X, Y \in L^2$, then the **conditional expectation** $\mathbb{E}[X|Y = y]$, as a function of $y \in \mathbb{R}$, is any measurable function $m : \mathbb{R} \rightarrow \mathbb{R}$ such that

$$\int_{Y^{-1}(B)} X d\mathbb{P} = \int_B m(y) d\mathbb{P}_Y(dy).$$

It can be shown that this is well-defined and it is unique. It has almost all the properties shown in Lemma 2.8.2, but instead of a.s., it is \mathbb{P}_Y -a.s. Before prove some other properties of this conditional expectation, we need the following lemma.

Lemma 2.8.3. Let X, Y be independent random variables and $f \in L^2(\mathbb{R}^2, \mathcal{B}(\mathbb{R}^2), \mathbb{P}_X \times \mathbb{P}_Y)$. If $m : \mathbb{R} \rightarrow \mathbb{R}$ is such that

$$m(y) := \mathbb{E}[f(X, y)],$$

then $m(Y) = \mathbb{E}[f(X, Y)|Y]$ almost surely.

Proof. Since $m(Y) \in \sigma(Y)$ already, we just have to prove that for all $A \in \mathcal{B}(\mathbb{R})$, we have that

$$\int_{Y^{-1}(A)} f(X, Y) d\mathbb{P} = \int_{Y^{-1}(A)} m(Y) d\mathbb{P}.$$

First, notice that Theorem 2.5.2 implies that

$$m(y) = \int_{\mathbb{R}} f(x, y) d\mathbb{P}_X(dx).$$

Now let $A \in \mathcal{B}(\mathbb{R})$. Then we have that

$$\int_{Y^{-1}(A)} m(Y) d\mathbb{P} = \int_A m(y) d\mathbb{P}_Y(dy) = \int_A \left(\int_{\mathbb{R}} f(x, y) d\mathbb{P}_X(dx) \right) d\mathbb{P}_Y(dy).$$

Using Fubini's Theorem 2.4.7 and the fact X and Y are independent, we obtain

$$\int_{Y^{-1}(A)} m(Y) d\mathbb{P} = \int_{\mathbb{R} \times A} f(x, y) d\mathbb{P}_{X,Y}(dx, dy).$$

Using the indicator function of $\mathbb{R} \times A$, we obtain

$$\int_{Y^{-1}(A)} m(Y) \, d\mathbb{P} = \int_{\mathbb{R}^2} \mathbf{1}_{\mathbb{R} \times A}(x, y) f(x, y) \, d\mathbb{P}_{X,Y}(dx, dy)$$

Now we can use Theorem 2.5.2 to change back the variables, hence

$$\int_{Y^{-1}(A)} m(Y) \, d\mathbb{P} = \int_{\Omega} \mathbf{1}_{\mathbb{R} \times A}(X, Y) f(X, Y) \, d\mathbb{P}.$$

Since

$$\mathbf{1}_{\mathbb{R} \times A}(X, Y) = \mathbf{1}_{Y^{-1}(A)},$$

we finally obtain

$$\int_{Y^{-1}(A)} m(Y) \, d\mathbb{P} = \int_{Y^{-1}(A)} f(X, Y) \, d\mathbb{P}.$$

□

Now we can state the following properties.

Lemma 2.8.4. *Let $X, Y \in L^2$ and $m(y) = \mathbb{E}[X|Y = y]$, then*

1. $m(Y) = \mathbb{E}[X|Y]$ \mathbb{P}_Y -a.s.; and
2. If $f \in L^2(\mathbb{R}^2, \mathcal{B}(\mathbb{R}^2), \mathbb{P}_X \times \mathbb{P}_Y)$ and X and Y are independent, then

$$\mathbb{E}[f(X, Y)|Y = y] = \mathbb{E}[f(X, y)] \quad \mathbb{P}_Y - \text{a.s.}$$

Finally, we can define the *conditional probability to a σ -algebra*.

Definition 2.8.6. Let $\mathcal{G} \subseteq \mathcal{F}$ be a sub σ algebra and $A \in \mathcal{F}$. We define the **conditional probability** as

$$\mathbb{P}(A|\mathcal{G}) := \mathbb{E}[\mathbf{1}_A|\mathcal{G}].$$

In undergraduate Probability courses, the **conditional probability of A given a set B** is defined as

$$\mathbb{P}(A|B) := \frac{\mathbb{P}(A \cap B)}{\mathbb{P}(B)},$$

for $\mathbb{P}(B) > 0$.

We can recover this definition using the following lemma.

Lemma 2.8.5. *Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space and $B \in \mathcal{F}$ be such that $\mathbb{P}(B) \in (0, 1)$. Let also \mathcal{G} be the following sub σ -algebra:*

$$\mathcal{G} := \{\emptyset, B, B^c, \Omega\}.$$

Then, for $\omega \in B$, we have

$$\mathbb{P}(A|\mathcal{G})(\omega) = \mathbb{P}(A|B),$$

and for $\omega \in B^c$ we obtain

$$\mathbb{P}(A|\mathcal{G})(\omega) = \mathbb{P}(A|B^c).$$

Proof. The property of the conditional expectation means that

$$\int_B \mathbb{E}[\mathbf{1}_A|\mathcal{G}] d\mathbb{P} = \mathbb{P}(A \cap B).$$

Since $\mathbb{E}[\mathbf{1}_A|\mathcal{G}] \in \mathcal{G}$, it is constant in B , then

$$\mathbb{E}[\mathbf{1}_A|\mathcal{G}](\omega) = \mathbb{P}(A \cap B)/\mathbb{P}(B) = \mathbb{P}(A|B),$$

for all $\omega \in B$. Likewise, we have

$$\mathbb{E}[\mathbf{1}_A|\mathcal{G}](\omega) = \mathbb{P}(A|B^c),$$

for all $\omega \in B^c$. □

2.9 Notions of Convergence and Laws of Large Numbers

In this section we define different kinds of convergence in our space and state the Weak and Strong Law of Large Numbers, as well as the Central Limit Theorem.

2.9.1 Weak Law and Convergence in Probability

Definition 2.9.1. Let $(X_n)_{n=1}^\infty$ and X be random elements in a metric space (S, d) , then we say that X_n **converges in probability** to X and denote $X_n \xrightarrow{\mathbb{P}} X$ if for all $\varepsilon > 0$ we have

$$\mathbb{P}(d(X_n, X) > \varepsilon) \rightarrow 0.$$

We also have *convergence in L^p* .

Definition 2.9.2. Let $(X_n)_{n \in \mathbb{N}} \subset L^p(\Omega, \mathcal{F}, \mathbb{P})$. We say that X_n **converges in L^p** to X if

$$\mathbb{E}|X_n - X|^p \rightarrow 0.$$

Remark 2.9.1. Because of Markov's Inequality 2.7.3, if $X_n \xrightarrow{L^p} X$ for some p , then

$$\mathbb{P}(|X_n - X| > \varepsilon) \leq \frac{\mathbb{E}|X_n - X|^p}{\varepsilon^p} \rightarrow 0,$$

then $X_n \xrightarrow{\mathbb{P}} X$.

We are now able to state and prove the *Weak Law of Large Numbers*.

Theorem 2.9.1. *Let $(X_n)_{n=1}^\infty$ be independent r.v. such that $\text{Var}(X_i) \leq C$, for all $i \in \mathbb{N}$, then*

$$\frac{1}{n} \sum_{i=1}^n (X_i - \mathbb{E}[X_i]) \xrightarrow{\mathbb{P}} 0.$$

Proof. Notice that

$$\mathbb{P}\left(\frac{1}{n} \sum_{i=1}^n (X_i - \mathbb{E}X_i) \geq \varepsilon\right) \leq \sum_{i=1}^n \frac{\text{Var}(X_i)}{n^2 \varepsilon^2},$$

by Markov's Inequality. Since $\text{Var}(X_i) \leq C$, the right-hand side converges to 0 and the theorem is proved. \square

2.9.2 Almost Surely Convergence and Strong Law

Definition 2.9.3. Let X and $(X_n)_{n=1}^\infty$ be random elements in a metric space (S, d) . We say that X_n **converges to X almost surely** and denote $X_n \rightarrow X$ a.s. if there is a null set A such that there is pointwise convergence in its complement.

Example 2.9.1. Let $(\Omega, \mathcal{F}, \mathbb{P}) = ([0, 1], \mathcal{B}(\mathbb{R}), \lambda)$, then $X_n = n\mathbf{1}_{[0, 1/n]} \rightarrow 0$ almost surely. However, notice that

$$\mathbb{E}|X_n|^p = n^{p-1},$$

hence X_n does not converges to 0 in L^p .

It can be shown the following lemma.

Lemma 2.9.1. *Let X and $(X_n)_{n=1}^\infty$ be random elements in a metric space (S, d) . Then $X_n \rightarrow X$ a.s. if and only if*

$$\sum_{i=1}^{\infty} \mathbb{P}(d(X_n, X) > \varepsilon) < \infty,$$

for all $\varepsilon > 0$.

We can now state the *Strong Law of Large Numbers*.

Theorem 2.9.2. Let $(X_i)_{i=1}^\infty$ be integrable i.i.d random variables, then

$$\frac{1}{n} \sum_{i=1}^n X_i \rightarrow \mathbb{E}[X_1] \text{ a.s.}$$

Proof. See [Durrett \(2019\)](#). □

2.9.3 Convergence in Distribution and Central Limit Theorem

Definition 2.9.4. Let $(\mu_n)_{n \in \mathbb{N}}$ and μ be probability measures in a Polish Space (M, d) , that is, a complete separable metric space (M, d) . We say that μ_n **converges weakly** to μ and denote $\mu \xrightarrow{w} \mu$ if, for all $f \in C_b(M)$, the space of all continuous and bounded real-valued functions, we have that

$$\int_M f d\mu_n \rightarrow \int_M f d\mu.$$

It can be shown this definition is equivalent to many others.

Theorem 2.9.3 (Portmanteau's Theorem). *Let $(\mu_n)_{n \in \mathbb{N}}$ and μ be probability measures in a Polish Space (M, d) , then all affirmations below are equivalent.*

1. $\mu_n \xrightarrow{w} \mu$;
2. For all closed sets F we have

$$\limsup_{n \rightarrow \infty} \mu_n(F) \leq \mu(F);$$

3. For all open sets A we have

$$\liminf_{n \rightarrow \infty} \mu_n(A) \geq \mu(A); \text{ and}$$

4. For all sets B such that $\mu(\partial B) = 0$ we have

$$\lim_{n \rightarrow \infty} \mu_n(B) = \mu(B).$$

Proof. For a proof see [Billingsley \(2013\)](#). □

Example 2.9.2. If δ_x is the probability measure concentrated at $x \in M$, then $\delta_{x_n} \xrightarrow{w} \delta_x$ if and only if $x_n \rightarrow x$.

Definition 2.9.5. We say that a family $(X_n)_{n \in \mathbb{N}}$ **converges in distribution** to X and denote $X_n \xrightarrow{d} X$ if their distribution converges weakly to the distribution of X .

We have the following theorem concerning convergence in distribution.

Theorem 2.9.4 (Paul-Lévy's Theorem). *Let $(X_n)_{n \in \mathbb{N}}$ and X be random variables with characteristic function $(\phi_n)_{n \in \mathbb{N}}$ and ϕ . Then X_n converges in distribution to X if and only if*

$$\phi_n(t) \rightarrow \phi(t),$$

for all $t \in \mathbb{R}$. In fact, if there exists a function $\psi(t)$ continuous at $t = 0$ and $\psi_n(t) \rightarrow \psi(t)$ for all $t \in \mathbb{R}$, then ψ is a characteristic function of a random variable X and $X_n \xrightarrow{d} X$.

Proof. For a proof, see [Shiryaev \(2016\)](#). □

A consequence of Paul-Lévy's Theorem is the Central Limit Theorem.

Theorem 2.9.5. *Let $(X_n)_{n \in \mathbb{N}}$ be square integrable i.i.d r.v. such that $\sigma^2 = \text{Var}(X_1)$. Then*

$$\frac{1}{\sqrt{n}} \sum_{k=1}^n (X_k - \mathbb{E}[X_k]) \xrightarrow{d} \mathcal{N}(0, \sigma^2).$$

Proof. For a proof, see [Shiryaev \(2016\)](#). □

2.10 Markov Chains

In this section, we define Markov Chain and give some simple examples.

2.10.1 Discrete Time and Countable State Space

Definition 2.10.1. A sequence of r.v. $(X_n)_{n \in \mathbb{N}}$ taking values in a discrete set E is a **discrete Markov Chain** if for all $n \in \mathbb{N}$ and all $x_1, \dots, x_{n+1} \in E$, we have

$$\mathbb{P}(X_{n+1} = x_{n+1} | X_1 = x_1, \dots, X_n = x_n) = \mathbb{P}(X_{n+1} = x_{n+1} | X_n = x_n).$$

The family $(\mathbf{P}_n)_{n \in \mathbb{N}}$ of matrices, defined as

$$\mathbf{P}_n(i, j) := \mathbb{P}(X_{n+1} = j | X_n = i),$$

is called the **transition matrices**. Equivalently, we denote the elements $\mathbf{P}_n(i, j)$ as $p_n(i, j)$ or use

$$p_n(B_n, B_{n+1}) := \mathbb{P}(X_{n+1} \in B_{n+1} | X_n \in B_n),$$

for Borel sets B_n and B_{n+1} .

Remark 2.10.1. Informally, we describe the Markov property as follows: given the *present* X_n , the *future* X_{n+1} does not depend on the *past* X_1, \dots, X_{n-1} .

Also, we define a *homogeneous Markov Chain*.

Definition 2.10.2. Let $(X_n)_{n \in \mathbb{N}}$ be a discrete Markov chain taking values in E . It is **homogeneous** if \mathbf{P}_n does not depend on $n \in \mathbb{N}$. In this case, we describe the Markov Chain through the initial distribution μ and the transition matrix $\mathbb{P} := \mathbb{P}_1$.

For short, we will denote a discrete homogeneous Markov Chain as (DHMC) or $(X_1 \rightarrow X_2 \rightarrow \dots)$. Given any law μ of X_1 , we can construct a Markov Chain using the previous definition: we just have to use Kolmogorov Extension Theorem to the distributions

$$\mu_n(B_1 \times B_2 \times \dots \times B_n) = \mu(B_1)p(B_2, B_1)p(B_3, B_2)\dots p(B_n, B_{n-1}),$$

that is,

$$\mathbb{P}(X_1 \in B_1, \dots, X_n \in B_n) = \mathbb{P}(X_1 \in B_1)\mathbb{P}(X_2 \in B_2|X_1 \in B_1)\dots\mathbb{P}(X_n \in B_n|X_{n-1} \in B_{n-1}).$$

It is easy to see that $X_1 \rightarrow \dots \rightarrow X_n$ if and only if $X_n \rightarrow \dots \rightarrow X_1$, since

$$\mathbb{P}(X_1 = x_1|X_n = x_n, \dots, X_2 = x_2) = \frac{\mathbb{P}(X_n = x_n, \dots, X_1 = x_1)}{\mathbb{P}(X_n = x_n, \dots, X_2 = x_2)},$$

and

$$\mathbb{P}(X_n = x_n, \dots, X_1 = x_1) = \mathbb{P}(X_n = x_n|X_{n-1} = x_{n-1})\dots\mathbb{P}(X_2 = x_2|X_1 = x_1)\mathbb{P}(X_1 = x_1),$$

therefore,

$$\begin{aligned} \mathbb{P}(X_1 = x_1|X_n = x_n, \dots, X_2 = x_2) &= \frac{\mathbb{P}(X_2 = x_2|X_1 = x_1)\mathbb{P}(X_1 = x_1)}{\mathbb{P}(X_2 = x_2)} \\ &= \mathbb{P}(X_1 = x_1|X_2 = x_2), \end{aligned}$$

that is, X_1 only depends on X_2 , not on X_3, \dots, X_n , which consists the Markovian property.

Example 2.10.1. Let $E = \{1, \dots, n\}$. Suppose X_1 is uniform in E . Now, consider the transition matrix \mathbf{P} given by the elements

$$p(i, j) = \begin{cases} 0, & \text{if } i = j \\ \frac{1}{n-1}, & \text{if } i \neq j. \end{cases}$$

Then the family $(X_n)_{n \in \mathbb{N}}$, given by

$$\mathbb{P}(X_{n+1} = j|X_n = i) := p(i, j),$$

is a Markov chain. Informally, given the present X_n , the future X_{n+1} is uniform in $E \setminus \{X_n\}$.

For any initial distribution, we can derive the distribution in time $n \in \mathbb{N}$ by a matrix product.

Lemma 2.10.1. Let μ be the initial distribution of a DHMC with transition matrix \mathbf{P} , then the distribution μ_n at time n is given by

$$\mu_n(j) = \sum_{i \in E} \mu(i) \mathbf{P}^n(i, j),$$

or

$$\mu_n = \mu \mathbf{P}^n,$$

when we write $\mu_n = (\mu_n(i))_{i \in E}$.

Definition 2.10.3. We say that a distribution μ is **stationary** if $\mu = \mu \mathbf{P}$.

This means that, at any given time n , we have the same initial distribution μ . Notice that, if

$$\mu(i) \mathbb{P}(X_2 = j | X_1 = i) = \mu(j) \mathbb{P}(X_1 = i | X_2 = j),$$

for all $(i, j) \in E^2$, then the process $X_n \rightarrow \dots \rightarrow X_1$ is Markovian and also stationary for the distribution μ . To see this, notice that if X_2 has distribution μ , then

$$\mathbb{P}(X_1 = i) = \sum_{j \in E} \mathbb{P}(X_2 = j) \mathbb{P}(X_1 = i | X_2 = j) = \sum_{j \in E} \mu(j) \mathbb{P}(X_2 = j | X_1 = i) = \mu(i),$$

that is, if $X_n \sim \mu$, then $X_i \sim \mu$ for all $i \leq n$.

Definition 2.10.4. We say that a Markov Chain with stationary distribution μ is **reversible** if for all n , $X_n \rightarrow X_{n-1} \rightarrow \dots \rightarrow X_1$ with initial distribution $X_n \sim \mu$ has the same joint law of $X_1 \rightarrow \dots \rightarrow X_n$ with initial distribution $X_1 \sim \mu$.

2.10.2 Continuous Time and Countable State Space

Definition 2.10.5. Let $(X_t)_{t \geq 0}$ be r.v. taking values in a discrete set E . We say that (X_t) is a **Markov chain** if for all $t_1 < \dots < t_n < t$, all $s \geq 0$ and all $i_1, \dots, i_{n+2} \in E$ we have

$$\mathbb{P}(X_{t+s} = i_{n+2} | X_t = i_{n+1}, X_{t_n} = x_n, \dots, X_{t_1} = i_1) = \mathbb{P}(X_{t+s} = i_{n+2} | X_t = i_{n+1}).$$

It is homogeneous if the right-hand side is independent of t and we define the **transition probabilities** as

$$p_s(i, j) := \mathbb{P}(X_{t+s} = j | X_t = i),$$

and \mathbf{P}_s is the matrix $(p_s(i, j))_{i, j \in E}$.

It is easy to see that $(\mathbf{P}_s)_{s \geq 0}$ defines a *semigroup* of matrices.

Definition 2.10.6. Let $(\mathbf{P}_s)_{s \geq 0}$ be quadratic matrices. Then they are a **semigroup** if

1. $\mathbf{P}_{t+s} = \mathbf{P}_t \mathbf{P}_s$;
2. $\mathbf{P}_0 = \text{Id}$.

The distribution μ_t of X_t at any given time is related to the distribution μ of X_0 through the following formula.

$$\mu_t(j) = \sum_{i \in E} \mu(i) \mathbf{P}_t(i, j),$$

or, the matrical form, $\mu_t = \mu \mathbf{P}_t$. We define the stationary distribution and reversibility as before.

It is usually to assume that the semigroup is *right-continuous at $t = 0$* , which we state below.

Definition 2.10.7. Let $(X_t)_{t \geq 0}$ be a continuous Markov process with discrete state space E and transition semigroup $(\mathbf{P}_t)_{t \geq 0}$. We say that the semigroup is right-continuous at $t = 0$ if

$$\lim_{t \rightarrow 0+} \mathbf{P}_t = \text{Id},$$

where the convergence is the convergence of each entry. This means that the process $(X_t)_{t \geq 0}$ is right-continuous in the sense that if $X_0 = i$, then for small $s \geq 0$, with high probability $X_s = i$.

For notation, we need to define a *stochastic process*.

Definition 2.10.8. A **stochastic process** is a family $(X_t)_{t \in T}$ of random variables, indexed by some set T .

Now we can define a *discrete Markov Chain*.

Definition 2.10.9. A **discrete homogeneous Markov Chain** $(X_t)_{t \geq 0}$ in E is a stochastic process such that there is a right-continuous semigroup $(\mathbf{P}_t)_{t \geq 0}$ in E that it is stochastic, that is

$$\sum_{j \in E} \mathbf{P}_t(i, j) = 1,$$

for all $i \in E$, all $t \geq 0$, and $\mathbf{P}_t(i, j) \geq 0$ for all $(i, j) \in E$ and $t \geq 0$. It also has to satisfy the Markov property:

$$\mathbb{P}(X_{t+s} = i_{n+2} | X_t = i_{n+1}, X_{t_n} = x_n, \dots, X_{t_1} = i_1) = \mathbf{P}_s(i_{n+1}, i_{n+2}),$$

for all $0 \leq s, t_1 < \dots < t_n < t$, and all $(i_k)_{k=1}^{n+2} \subseteq E$.

In fact, given a stochastic right-continuous semigroup $(\mathbf{P}_t)_{t \geq 0}$ in E , we can always define a discrete homonegenous Markov Chain $(X_t)_{t \geq 0}$ with these transitions probabilities.

Our last definition is the generator of the semigroup.

Lemma 2.10.2. *For a stochastic right-continuous semigroup \mathbf{P}_t in E , there exist a matrix A such that*

$$A = \lim_{h \rightarrow 0_+} \frac{\mathbf{P}_h - \text{Id}}{h},$$

where the convergence is for each entry. The matrix A is known as the **infinitesimal generator** of the semigroup $(\mathbf{P}_t)_{t \geq 0}$.

Given $(\mathbf{P}_t)_{t \geq 0}$ and its infinitesimal generator A , they satisfy the following differential equation:

$$\frac{d}{dt} \mathbf{P}_t = A \mathbf{P}_t = \mathbf{P}_t A.$$

For a first example, we have the *Poisson Process*.

Example 2.10.2. Let $(\tau_n)_{n \in \mathbb{N}}$ be i.i.d exponential r.v with parameter $\lambda > 0$, that is, $\mathbb{P}(\tau_n \geq x) = e^{-\lambda x}$ for all $x \geq 0$ and let $(N_t)_{t \geq 0} \subseteq \mathbb{N}$ be the stochastic process such that

$$N_t = n \Leftrightarrow \sum_{i=1}^n \tau_i \leq t < \sum_{i=1}^{n+1} \tau_i,$$

then $(N_t)_{t \geq 0}$ is a discrete homogeneous Markov Chain, called the **Poisson Process**. In fact, if $N(a, b] := N_b - N_a$ for $b \geq a$, then N_a and $N(a, b]$ are independent, and the transition probabilities are given by

$$\mathbb{P}(N_t = n + k | N_0 = n) = \frac{(\lambda t)^k e^{-\lambda t}}{k!}.$$

For more information about discrete continuous Markov Chains, see [Brémaud \(2013\)](#).

2.10.3 Uncountable State Space

For the uncountable state space case, we have to assume some technicalities. First, let (S, \mathcal{S}) be a measurable space where (S, τ) is a topological space and \mathcal{S} is the Borel σ -algebra.

Definition 2.10.10. Let $T = [0, \infty)$ and $(X_t)_{t \in T}$ be a stochastic process in (S, \mathcal{S}) . Suppose that given $X_t = x$, the law of X_{t+s} for $s \geq 0$ is given by the transition probabilities $p_t(x, \cdot)$, that is,

$$\mathbb{P}(X_{t+s} \in B | X_t = x) = p_s(x, B),$$

independently of the past for all $x \in S$ and all $B \in \mathcal{S}$ fixed, or, equivalently

$$\mathbb{P}(X_{t+s} \in B | \sigma(X_r : r < t), X_t = x) = p_s(x, B).$$

Then $(X_t)_{t \in T}$ is called a **homogeneous Markov process**.

We will assume that the transition probabilities satisfies some regularities conditions, namely,

1. $p_t(\cdot, B)$ is measurable for fixed B and t ;
2. $p_t(x, \cdot)$ is a probability measure for t and x fixed;
3. $p_0(x, B) = \delta_x(B)$;
4. For fixed t , if $x_n \rightarrow x$, then $p_t(x_n, \cdot) \xrightarrow{w} p_t(x, \cdot)$;
5. For every neighborhood $U(x)$ of x , we have $p_t(x, U(x)) \rightarrow 1$ for $t \searrow 0$; and
6. The Chapman-Kolmogorov equation are satisfied:

$$p_{s+t}(x, B) = \int_S p_s(y, B) dp_t(x, dy).$$

(see [Itô \(2013\)](#) for more).

If $C_b(S)$ is the space of all bounded continuous real-valued functions in S and it is endowed with the supremum norm, that is,

$$d(f, g) := \sup_{x \in S} |f(x) - g(x)|,$$

then for all $f \in C_b(S)$ we can set

$$[P_t f](x) := \int_S f(y) dp_t(x, dy),$$

and conditions (1) – (6) can be rewritten as

1. $P_t : C_b(S) \rightarrow C_b(S)$ is linear;
2. $P_0 = \text{Id}$;
3. $[P_t f](x) \rightarrow f(x)$ for $x \in S$ and $t \searrow 0$;
4. $P_{t+s} = P_t \circ P_s$ (semigroup property); and
5. $P_t 1 = 1$ and $P_t \geq 0$, that is, $P_t f \geq 0$ for $f \geq 0$.

If $(P_t)_{t \geq 0}$ satisfies (1) – (5), then we can find a Markov Process $(X_t)_{t \geq 0}$ with this semigroup $(P_t)_{t \geq 0}$. (see [Guionnet and Zegarliński \(2003\)](#)).

Example 2.10.3. Let $T = [0, \infty)$ and $(X_t)_{t \in T} \in \mathbb{R}^n$ be the Markov Process defined by the Radon-Nikodym derivative

$$\frac{dp_t(y, dx)}{dx} = \frac{1}{(2\pi t)^{n/2}} e^{-\frac{\|x-y\|^2}{2t}},$$

then $(X_t)_{t \in T}$ is called the **Brownian motion** in \mathbb{R}^n .

Definition 2.10.11. Let $\mathcal{D}(L) := \{f \in C_b(S) : \lim_{t \rightarrow 0+} \frac{P_t f - f}{t} \text{ exists}\}$, then the **infinitesimal generator** of P_t is the operator L such that

$$Lf = \lim_{t \rightarrow 0+} \frac{P_t f - f}{t},$$

for $f \in \mathcal{D}$.

Example 2.10.4. It can be shown that if $(P_t)_{t \in T}$ is the semigroup of the Brownian Motion $(X_t)_{t \in T}$, then the infinitesimal generator L has domain

$$\mathcal{D}(L) \subseteq C^2(\mathbb{R}^n),$$

and

$$Lf = \frac{1}{2} \Delta f.$$

Finally, the following theorem provides conditions for a linear operator in $C_b(S)$ to generate a semigroup.

Theorem 2.10.1 (Hille-Yoshida's Theorem). *Let $L : C_b(S) \rightarrow C_b(S)$ be a linear operator. Then L is the infinitesimal generator of a Markov Semigroup $(P_t)_{t \geq 0}$ if and only if*

1. $\mathcal{D}(L)$ is dense in $C_b(S)$;
2. The constant function 1 is in $\mathcal{D}(L)$ and $L1 = 0$;
3. L is closed, that is, for all $f_n \in \mathcal{D}(L)$ such that $f_n \rightarrow f$ and Lf_n converges, then

$$Lf_n \rightarrow Lf;$$

and

4. If $\lambda > 0$, then $(\lambda - L)$ is invertible, $(\lambda - L)^{-1}f \geq 0$ whenever $f \geq 0$ and

$$\sup_{\|f\| \leq 1} \|(\lambda - L)^{-1}f\| \leq \frac{1}{\lambda}.$$

Proof. See [Guionnet and Zegarlinski \(2003\)](#). □

Information and Its Mysteries

3.1 Introduction

In this chapter, we will introduce the basic ideas from Information Theory. Our goal is to provide enough background to understand the theorems coming from this area, as well as their applications.

In Section 3.2 we define the *Shannon entropy*. It measures the uncertainty of a random variable according to the formula:

$$H(X) = \sum_{i=1}^n \mathbb{P}(X = x_i) \log \mathbb{P}(X = x_i),$$

and we can see that X is constant if and only if $H(X) = 0$.

Moreover in this first section, we will define others entropies, say, *joint entropy*, *conditional entropy* and two related quantities: *Kullback-Leibler divergence* and *mutual information*. All these quantities are related to the Shannon Entropy and they can be used to prove several results. For instance, we can prove that two random variables X, Y are independent if and only if their conditional entropy is 0.

In Section 3.3 we will introduce the idea of *codes*. This goes back to Shannon and the problem of *compressing* a message.

The problem can be described as follows: let X be a discrete random variable in some discrete set \mathcal{X} and an alphabet \mathcal{D} of D symbols. Suppose we want to *compress* X , that is, for all $x \in \mathcal{X}$, we want to associate to it a string of letters in the alphabet \mathcal{D} . For instance, suppose $\mathcal{X} = \{0, 1\}$ and $\mathcal{D} = \{a, b\}$, then one way to compress it is the following Table 1.

Now suppose we want to compress X using the smallest numbers of letters in the alphabet per element of \mathcal{X} , according to the probability of each element. One of the ways

of doing it is to minimize the expected value of the *length*:

$$\mathbb{E}[l(X)] = \sum_{x \in \mathcal{X}} l(x)p(x).$$

Moreover, we want our code to be efficient in decoding, that is, given a sequence of letters that we know it came from compressing a sample X_1, \dots, X_n of X , we want to recover uniquely the sample.

This problem has a solution and it is directly associated with the Shannon Entropy, as we will see in this section.

After codes and compression, we will extend the discrete entropy to the continuous case in Section 3.4. The extension is almost obvious through the *differential entropy*:

$$H(X) = - \int_{\mathbb{R}} f(x) \log f(x) dx,$$

nevertheless some properties of the discrete entropy are lost in the continuous case. For instance, there are random variables with a negative differential entropy. Although this seemingly incompatibility, we will find a confluence of ideas between them. For instance, we can see an analogy between Corollary 3.2.1 and Theorem 3.4.1.

We will also see other definitions using densities and random variables with absolutely continuous distributions, namely, the *Fisher information* and *exponential entropy*.

In Section 3.5, we will see the second real application of Shannon's ideas, namely, the study of *channels*. By a channel we mean a triple $(\mathcal{X}, p(x|y), \mathcal{Y})$, where \mathcal{X} is the *input*, \mathcal{Y} is the *output* and $p(y|x)$ is the transition family of probabilities, that is, the probability of the output $y \in \mathcal{Y}$, given input $x \in \mathcal{X}$. The problem consists in sending a message x through an imperfect channel, where errors can happen. For the receiver to decode the message, it is necessary to send the message x with redundancy, but can we quantify this in an efficient way? The answer is yes, and it is express in Theorems 3.5.1 and 3.5.2.

Finally, in Section 3.6, we will prove a couple of useful inequalities in Information Theory that will be necessary to prove our main theorem in Chapter 6, namely, the *Fisher Information Inequality* and *Exponential Entropy Inequality*.

3.2 Shannon Entropy

Definition 3.2.1. Let X be a discrete r.v. taking values in \mathcal{X} and $p(x) = \mathbb{P}(X = x)$, then the **Shannon entropy** of X is defined as

$$H(X) := - \sum_{x \in \mathcal{X}} p(x) \log p(x),$$

with the convention $0 \log 0 = 0$.

Remark 3.2.1. If $|\mathcal{X}| < \infty$ we have that $H(X) < \infty$, but in the countable case it is necessary to assume that $\sum_{x \in \mathcal{X}} p(x) \log p(x)$ converges.

It is easy to see that $H(X) = \mathbb{E}[-\log p(X)]$. Also, since $0 \leq p(x) \leq 1$, we have $p(x) \log p(x) \leq 0$, then

$$H(X) \geq 0,$$

and equality only holds if $p(x) \log p(x) = 0$ for all $x \in \mathcal{X}$, that is, there is only one $x_0 \in \mathcal{X}$ that $X = x_0$ almost surely.

Now, let \mathbb{R}_+^∞ be the space of all sequences $(x_n)_{n \in \mathbb{N}}$ such that $x_n \geq 0$ for all $n \in \mathbb{N}$. Also, by \mathbb{R}_+^n , we mean the canonical positive cone in \mathbb{R}^n , that is,

$$\mathbb{R}_+^n = \{x \in \mathbb{R}^n : x_i \geq 0, \forall i \leq n\}.$$

By the definition of the Shannon Entropy, $H(X)$ only depends on the probabilities of $x \in \mathcal{X}$, therefore we can consider H defined in

$$A := \left(\bigcup_{n \in \mathbb{N}} \{p \in \mathbb{R}_+^n : \sum_{i=1}^n p_i = 1\} \right) \cup \{p \in \mathbb{R}_+^\infty : \sum_{i \in \mathbb{N}} p_i = 1, \sum_{i \in \mathbb{N}} p_i \log 1/p_i < \infty\}.$$

Because of this, if $f : \mathcal{X} \rightarrow \mathcal{Y}$ is an injective function, then $H(X) = H(f(X))$.

We can define the *joint Shannon Entropy* in a similar manner.

Definition 3.2.2. Let X, Y be two discrete random variables taking values in \mathcal{X} and \mathcal{Y} , respectively, and $p(x, y) = \mathbb{P}(X = x, Y = y)$, then their **joint entropy** is

$$H(X, Y) := - \sum_{(x, y) \in \mathcal{X} \times \mathcal{Y}} p(x, y) \log p(x, y).$$

We prove now an elementary lemma.

Lemma 3.2.1. *If X, Y are independent discrete random variables, then*

$$H(X, Y) = H(X) + H(Y).$$

Proof. Just notice that $\mathbb{P}(X = x, Y = y) = \mathbb{P}(X = x)\mathbb{P}(Y = y)$ and replace this in the expression of the joint Shannon Entropy. \square

Later we will prove that in fact we have that

$$H(X, Y) \leq H(X) + H(Y),$$

with equality if only if X, Y are independent.

Notice that, if we compute $H(X, Y) - H(Y)$, we have

$$\begin{aligned} H(X, Y) - H(Y) &= - \sum_{(x,y) \in \mathcal{X} \times \mathcal{Y}} \mathbb{P}(X = x, Y = y) \log \mathbb{P}(X = x, Y = y) \\ &\quad + \sum_{y \in \mathcal{Y}} \mathbb{P}(Y = y) \log \mathbb{P}(Y = y). \end{aligned}$$

Using that $\mathbb{P}(X = x|Y = y) = \mathbb{P}(X = x, Y = y)/\mathbb{P}(Y = y)$, we obtain that

$$H(X, Y) - H(Y) = - \sum_{(x,y) \in \mathcal{X} \times \mathcal{Y}} \mathbb{P}(X = x, Y = y) \log \mathbb{P}(X = x|Y = y).$$

Therefore, we have the following defining of the *conditional Shannon Entropy*.

Definition 3.2.3. Let X, Y are two discrete random variables taking values in \mathcal{X} and \mathcal{Y} , respectively, $p(x, y) = \mathbb{P}(X = x, Y = y)$ and $p(x|y) = \mathbb{P}(X = x|Y = y)$, then the **conditional entropy of X given Y** is

$$H(X|Y) := - \sum_{(x,y) \in \mathcal{X} \times \mathcal{Y}} p(x, y) \log p(x|y).$$

By definition, we have that $H(X, Y) = H(Y) + H(X|Y)$. Likewise, $H(X, Y) = H(X) + H(Y|X)$.

To prove some properties of these quantities, we need the following useful definition.

Definition 3.2.4. Let p and q two discrete probability measures in \mathcal{X} . Suppose $p \ll q$, then the **Kullback-Leibler Divergence** is defined as

$$\mathcal{D}(p||q) = \sum_{x \in \mathcal{X}} p(x) \log \left(\frac{p(x)}{q(x)} \right),$$

with the convention $0 \log \frac{0}{0} = 0$.

Remark 3.2.2. If $\text{supp}(q) = \{x \in \mathcal{X} : q(x) \neq 0\}$, then

$$\mathcal{D}(p||q) = \sum_{x \in \text{supp}(q)} p(x) \log \left(\frac{p(x)}{q(x)} \right).$$

Remark 3.2.3. In the literature, $\mathcal{D}(p||q)$ are also called Relative Entropy or Kullback-Leibler Distance, even though is not a true distance, since it is not symmetric.

Let us prove the first reason why $\mathcal{D}(p||q)$ is called a distance.

Lemma 3.2.2. Let $p \ll q$, then $\mathcal{D}(p||q) \geq 0$ and equality holds if and only if $p = q$.

Proof. Notice that

$$\mathcal{D}(p||q) = - \sum_{x \in \text{supp}(p)} p(x) \log \left(\frac{q(x)}{p(x)} \right).$$

Now, the elementary inequality $\log x \leq x - 1$, that holds for all $x > 0$ with equality if and only if $x = 1$, implies

$$\mathcal{D}(p||q) \geq - \sum_{x \in \text{supp}(p)} p(x) \left(\frac{q(x)}{p(x)} - 1 \right) = \sum_{x \in \text{supp}(p)} (p(x) - q(x)).$$

Finally, we have that

$$\sum_{x \in \text{supp}(p)} q(x) \leq 1,$$

then

$$\mathcal{D}(p||q) \geq 0.$$

The equality holds if and only if $\frac{q(x)}{p(x)} = 1$, for all $x \in \mathcal{X}$, that is, $p = q$. \square

There are some applications of this result. The first one shows that the uniform distribution maximizes the entropy.

Corollary 3.2.1. *If X takes values in a finite set \mathcal{X} , then $H(X) \leq \log |\mathcal{X}|$ with equality if and only if X is uniform in \mathcal{X} .*

Proof. Let $p(x) = \mathbb{P}(X = x)$ and q be the uniform distribution in \mathcal{X} . It is easy to see that $p \ll q$, then

$$0 \leq \mathcal{D}(p||q).$$

Using that

$$q(x) = \frac{1}{|\mathcal{X}|},$$

for all $x \in \mathcal{X}$, we can open up the expression of $\mathcal{D}(p||q)$ and obtain

$$\mathcal{D}(p||q) = \log |\mathcal{X}| - H(X).$$

Then $H(X) \leq \log |\mathcal{X}|$ with equality if and only if $p = q$, that is, p is uniform in \mathcal{X} . \square

Corollary 3.2.2. *Let X, Y be two discrete random variables, then*

$$H(X, Y) \leq H(Y) + H(X),$$

with equality if and only if X and Y are independent.

Proof. Let

$$\begin{aligned} p(x, y) &:= \mathbb{P}(X = x, Y = y); \\ p_1(x) &:= \mathbb{P}(X = x); \text{ and} \\ p_2(y) &:= \mathbb{P}(Y = y). \end{aligned}$$

It is easy to see that $p \ll p_1 \times p_2$, then

$$\mathcal{D}(p||p_1 \times p_2) \geq 0$$

Now, the left-hand side is equal to

$$\sum_{(x,y) \in \mathcal{X} \times \mathcal{Y}} p(x, y) \log \left(\frac{p(x, y)}{p_1(x)p_2(y)} \right),$$

that is,

$$\mathcal{D}(p||p_1 \times p_2) = -H(X, Y) - \sum_{(x,y) \in \mathcal{X} \times \mathcal{Y}} p(x, y) \log p_1(x) - \sum_{(x,y) \in \mathcal{X} \times \mathcal{Y}} p(x, y) \log p_2(y).$$

Finally, we see that

$$- \sum_{(x,y) \in \mathcal{X} \times \mathcal{Y}} p(x, y) \log p_1(x) = H(X),$$

and likewise

$$- \sum_{(x,y) \in \mathcal{X} \times \mathcal{Y}} p(x, y) \log p_2(y) = H(Y),$$

then

$$0 \leq \mathcal{D}(p||p_1 \times p_2) = H(X) + H(Y) - H(X, Y),$$

with equality if and only if $p = p_1 \times p_2$, that is, X, Y are independent. □

The expression $\mathcal{D}(p||p_1 \times p_2)$ has a special name.

Definition 3.2.5. Let X, Y be two discrete random variables and $p(x, y)$, $p_1(x)$ and $p_2(y)$ representing their joint distribution, the distribution of X and the distribution of Y , respectively. Then the **mutual information** of X and Y is defined as

$$I(X; Y) := \mathcal{D}(p||p_1 \times p_2).$$

We have the final corollary concerning the conditional entropy.

Corollary 3.2.3. *Let X, Y be two discrete random variables, then $0 \leq H(X|Y) \leq H(X)$. The first equality holds if and only if $X = f(Y)$ for some measurable function f and the second holds if and only if X and Y are independent. We also have $H(g(X)) \leq H(X)$ with equality if and only if g is injective.*

Proof. For the second inequality, we have that

$$H(X, Y) = H(Y) + H(X|Y) \leq H(X) + H(Y),$$

then $H(X|Y) \leq H(X)$, and equality holds if and only if X, Y are independent.

For the first, since $p(x, y) \geq 0$ and $0 \leq p(x|y) \leq 1$, we have that $H(X|Y) \geq 0$ and equality holds if and only if

$$p(x, y) \log p(x|y) = 0,$$

for all $(x, y) \in \mathcal{X} \times \mathcal{Y}$. That is, fixed y , there is only one x_0 that $p(x_0|y) = 1$ and for all other we have $p(x|y) = 0$. If we set $f(y) := x_0$, then $X = f(Y)$.

For the second part, notice that

$$H(X, g(X)) = H(X) + H(g(X)|X) = H(g(X)) + H(X|g(X)).$$

Since $H(g(X)|X) = 0$ and $H(X|g(X)) \geq 0$, we have the result with equality if and only if $H(X|g(X)) = 0$, that is, if g is injective. \square

We summarize now all the results proved in this section.

Corollary 3.2.4. *Let X, Y be two discrete random variables in \mathcal{X} and \mathcal{Y} , respectively. Then*

1. *The entropy is bounded:*

$$0 \leq H(X) \leq \log |\mathcal{X}|.$$

The first equality holds if and only if X is constant and the second holds if and only if X is uniform;

2. *The joint entropy follows the chain rule and it is bounded:*

$$H(X, Y) = H(X) + H(Y|X) \leq H(X) + H(Y),$$

and equality holds if and only if X and Y are independent;

3. *The mutual information can be decomposed as*

$$0 \leq I(X, Y) = H(X) + H(Y) - H(X, Y) = H(X) - H(X|Y),$$

and $I(X, Y) = 0$ if and only if X, Y are independent;

4. The conditional entropy is bounded:

$$0 \leq H(X|Y) \leq H(X).$$

The first equality holds if and only if $X = f(Y)$ for some f and the second holds if and only if X and Y are independent; and

5. The relative entropy is always nonnegative: if p and q are two probabilities measures in some common finite space $(\mathcal{X}, \mathcal{F})$, then $\mathcal{D}(p||q) \geq 0$ and equality holds if and only if $p = q$.

In the next section, we will explore an interpretation of these equalities and inequalities through the notion of *Information*.

3.3 Compression and Codes

The goal of this section is to describe a code and the role of entropy in compression. We will follow [Cover and Thomas \(2012\)](#). To begin with, we define a *code*.

Definition 3.3.1. Let X be a discrete random variable taking values in \mathcal{X} and \mathbf{D} a finite nonempty set, known as the **alphabet**. A **code** C for X is an injective map $C : \mathcal{X} \rightarrow \mathcal{D}$, where

$$\mathcal{D} = \bigcup_{n=1}^{\infty} \mathbf{D}^n,$$

The value $C(x)$ is known as the **codeword** of $x \in \mathcal{X}$. Moreover, if $|\mathbf{D}| = D$, we say that the alphabet is D -ary.

Remark 3.3.1. Instead of denoting $C(x) = (a_1, \dots, a_k) \in \mathbf{D}^k$, for some $k \in \mathbb{N}$, we will simply concatenate the letters a_1, \dots, a_k , that is, $C(x) = a_1 \dots a_k$.

Associated with a code, we can quantify the *length* of each codeword.

Definition 3.3.2. Let X be a discrete r.v. taking values in \mathcal{X} , \mathbf{D} be a D -ary alphabet and $C : \mathcal{X} \rightarrow \mathcal{D}$ a code. For each $x \in \mathcal{X}$, let $l(x) \in \mathbb{N}$ be the number such that $C(x) \in \mathbf{D}^{l(x)}$. Then $l(x)$ is the **length** of the codeword $C(x)$. The **expected length** is

$$L(C) := \mathbb{E}[l(X)].$$

Take a look at the following simple example.

Example 3.3.1. Let X be a r.v. taking values in $\mathcal{X} = \{0, 1\}$ with probability

$$\mathbb{P}(X = 1) = p = 1 - \mathbb{P}(X = 0),$$

\mathcal{X}	Code C
0	aab
1	ba

Table 1 – Table of a code for the set \mathcal{X} .

and the alphabet $\mathbf{D} = \{a, b\}$, then a code C for X can be described by the values $C(0) = aab$ and $C(1) = ba$, as shown in Table 1. Furthermore, we notice that $l(0) = 3$ and $l(1) = 2$, then the expected length is

$$L(C) = 2p + 3(1 - p) = 3 - p.$$

Notice that in Example 3.3.1, we code just one element of \mathcal{X} at time, that is, the domain of C is \mathcal{X} . However, we can extend C to a code C^* where its domain is bigger than \mathcal{X} .

Definition 3.3.3. Let C be a code for X and

$$\mathbf{X} = \bigcup_{n \in \mathbb{N}} \mathcal{X}^n.$$

The **extension of the code** C is a code $C^* : \mathbf{X} \rightarrow \mathcal{D}$ such that

$$C^*(x_1 x_2 \dots x_n) = C(x_1) C(x_2) \dots C(x_n),$$

where again $x_1 \dots x_n$ indicates (x_1, \dots, x_n) and $C(x_1) C(x_2) \dots C(x_n)$ indicates the concatenation of the corresponding codewords.

To illustrate the idea, we can extend our Example 3.3.1 to the concatenation of two elements.

Example 3.3.2. Let X, \mathbf{D} and C be as in Example 3.3.1. The extension to two elements is shown in Table 2.

\mathcal{X}^2	Code C
00	aabaab
01	aabba
10	baaab
11	baba

Table 2 – Table of a code for the set \mathcal{X}^2 .

A code is sometimes called a **Compression** of the random variable X . The main idea of Shannon is to find a code C such that we can always recover the concatenation $x_1 \dots x_n$ from the value $C(x_1 \dots x_n)$ (that is, it is injective) and it minimizes the expected length. Among all the codes, the *instantaneous* ones are the most interesting.

Definition 3.3.4. Let C be a code for X . The code C is **instantaneous** if no codeword is a prefix of any other codewords, that is, if $x, y \in \mathcal{X}$ and $l(x) < l(y)$, then there is no element $a_1 \dots a_n \in \mathbf{D}$ such that $C(y) = C(x)a_1 \dots a_n$.

The idea behind an instantaneous code is that we can recover a concatenation $x_1 \dots x_n$ “reading” the extension $C(x_1 \dots x_n)$, but rather than look at all the extension, we can already identify any particular x_i without the future codewords. For instance, we have the following example.

Example 3.3.3. Let X taking values in $\{0, 1, 2\}$, $\mathbf{D} = \{0, 1\}$ and the code C such that $C(0) = 0$, $C(1) = 11$ and $C(2) = 10$, then the code C is instantaneous and, for instance, if we have that $C(x_1 x_2 x_3) = 10011$, we can identify $x_1 = 2$ without the reference of x_2 and x_3 , as well as for $x_2 = 0$ and $x_3 = 1$.

On the other hand, the following example from [Cover and Thomas \(2012\)](#) is not instantaneous.

Example 3.3.4. Let X be a r.v. taking values in $\{0, 1, 2, 3\}$ and the code C such that $C(0) = 10$, $C(1) = 00$, $C(2) = 11$ and $C(3) = 110$. Then C is not instantaneous because $C(2)$ is a prefix of $C(3)$. Note that if $C(x_1 \dots x_n) = 110$, we could not identify that $n = 1$ and $x_1 = 3$ if we just read the string “11”, which is the codeword for 2.

For instantaneous codes, we have the following inequality.

Theorem 3.3.1 (Kraft’s Inequality). *Let C an instantaneous D -ary code with codeword lengths l_1, \dots, l_n , then we have*

$$\sum_{k=1}^n D^{-l_k} \leq 1. \quad (3.1)$$

Conversely, if a code satisfies this inequality, then there is an instantaneous code with the same codeword lengths.

Proof. For a proof, we recommend [Cover and Thomas \(2012\)](#). □

Notice that, as a consequence of the Kraft’s Inequality, we have that the family of all instantaneous code is countable.

Lemma 3.3.1. *If I is the set of all instantaneous codes $C : \mathcal{X} \rightarrow \mathbf{D}$, then I is countable*

Proof. Since $I \subseteq \mathbf{D}^{\mathcal{X}}$, we have that I is at most countable. Now, if $n = |\mathcal{X}|$ and $l_i \geq \log_D n$ for all i , then

$$\sum_{k=1}^n D^{-l_k} \leq \sum_{k=1}^n \frac{1}{n} = 1,$$

then Kraft's Inequality implies that there is an instantaneous code with these lengths. Since the set of all (l_1, \dots, l_n) such that $l_i \geq \log_D n$ for all i is countable, we also have that I is countable. \square

Therefore, if we want to minimize the expected length $\sum_{x \in \mathcal{X}} p(x)l(x)$ among all instantaneous codes, we have to solve the following optimization problem.

$$\begin{aligned} L = \min \quad & \sum_{x \in \mathcal{X}} p(x)l(x) \\ \text{s.t.} \quad & \sum_{x \in \mathcal{X}} D^{-l(x)} \leq 1 \\ & l(x) \in \mathbb{N}, \forall x \in \mathcal{X}. \end{aligned}$$

The relaxation problem is

$$\begin{aligned} \hat{L} = \min \quad & \sum_{x \in \mathcal{X}} p(x)l(x) \\ \text{s.t.} \quad & \sum_{x \in \mathcal{X}} D^{-l(x)} \leq 1 \\ & l(x) \in \mathbb{R}_+, \forall x \in \mathcal{X}. \end{aligned}$$

The relaxation problem is solved by Lagrange Multipliers and its minimum is achieved for $l(x) = -\log_D p(x)$, therefore,

$$\hat{L} = H_D(X) := - \sum_{x \in \mathcal{X}} p(x) \log_D p(x),$$

that is, the Shannon Entropy with logarithm in base D .

Since $\mathbb{N} \subset \mathbb{R}_+$, we have that $\hat{L} \leq L$ and equality is achieved if and only if $-\log_D p(x) \in \mathbb{N}$ for all x . We can summarize this in the following theorem.

Theorem 3.3.2. *For all instantaneous D -ary codes we have that $H_D(X) \leq \mathbb{E}[l(X)]$ and equality is achieved when the codeword lengths $l(x)$ satisfy*

$$l(x) = -\log_D p(x) \in \mathbb{N},$$

for all $x \in \mathcal{X}$. Therefore, $H_D(X)$ is known as the limit of compression of a random variable X .

Before we provide an interpretation for this result, let us just mention an example of a code.

Example 3.3.5. Let $l(x) = \lceil \log_D 1/p(x) \rceil$, where $\lceil x \rceil$ represents the smallest integer greater than x , then

$$\sum_{x \in \mathcal{X}} D^{-l(x)} \leq \sum_{x \in \mathcal{X}} D^{-\log_D p(x)} = 1,$$

that is, $l(x)$ satisfies the Kraft's Inequality, hence there is an instantaneous code with length $l(x)$.

For this code, we have

$$H_D(X) \leq \mathbb{E}[l(X)] = \sum_{x \in \mathcal{X}} p(x)l(x) \leq \sum_{x \in \mathcal{X}} p(x)(\log_D 1/p(x) + 1),$$

hence,

$$H_D(X) \leq \mathbb{E}[l(X)] \leq H_D(X) + 1.$$

This is known as the **Shannon's Code**.

In fact, if we encode (X_1, \dots, X_n) , an i.i.d sample of X , by the Shannon's Code, we have that the expected length *per one symbol* is

$$L_n = \frac{1}{n} \mathbb{E}[l(X_1, \dots, X_n)] \leq \frac{1}{n} (H(X_1, \dots, X_n) + 1),$$

hence,

$$nH_D(X) = H_D(X_1, \dots, X_n) \leq \mathbb{E}[l(X_1, \dots, X_n)] \leq H_D(X_1, \dots, X_n) + 1 = nH_D(X) + 1,$$

that is,

$$H_D(X) \leq L_n \leq H_D(X) + \frac{1}{n}.$$

Therefore, the expected length per symbol can get arbitrarily close to the Entropy.

Entropy is, therefore, **Information**. In fact, $H(X)$ is known as *self-information*, since

$$I(X, X) = H(X) - H(X|X) = H(X).$$

$H(X|Y)$ is the information X still has given the knowledge of Y .

Theorem 3.3.2 expresses that we can not compress more than the Information contained in the random variable X if we want to recover it exactly, that is, if we want an instantaneous code for X .

In the previous section, we derived some properties of Entropy of a random variable X , and now we can provide an interpretation for it using the compression of X .

First, we have that $H(X, Y) \leq H(X) + H(Y)$. That means that, if we want to compress X and Y simultaneously, we do not need to compress independently, but we can use the Information of one to get Information of the other, that is, we can compress efficiently their joint distribution of X, Y if they are correlated. In fact, we can first compress X and then use the information of X to compress Y , and it is precisely $H(Y|X)$, that is, $H(X, Y) = H(X) + H(Y|X)$. However, in the independent case, knowing X gives nothing about Y , therefore, we will have to compress independently X and Y , that is

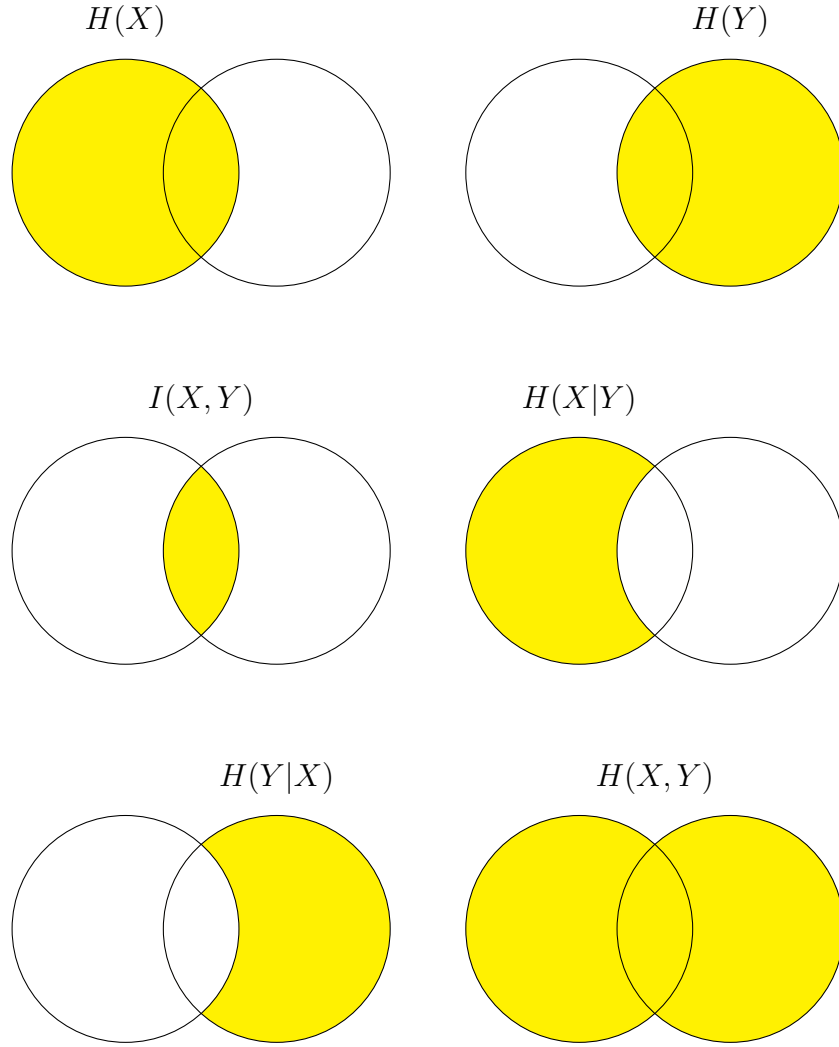


Figure 3 – Representative Veen Diagram with two circles: the first one is concerning $H(X)$ and the second $H(Y)$. Notice that $H(X, Y) = H(X) + H(Y|X)$, $H(X, Y) \leq H(X) + H(Y)$ and equality holds if and only if $I(X, Y) = 0$, that is, they are independent.

why $H(X, Y) = H(X) + H(Y)$ in this case. Symbolically, we can express the relations between these quantities through a Veen Diagram, as in Figure 3.

Futhermore, if $H(X) = 0$ and therefore X is constant, we can get codes with expected length arbitrarily close to zero. We just need to compress more symbols with just one codeword, for instance, we can compress X_1, \dots, X_n with $C(X_1 \dots X_n) = 1$, and, per symbol, we have

$$L_n = \frac{1}{n} \rightarrow 0.$$

Finally, the Kullback-Leibler divergence drives the error when we encode a random variable X with the wrong distribution q .

Theorem 3.3.3. Let X be a finite random variable with distribution $p(x)$, for $x \in \mathcal{X}$. Suppose we guess a wrong distribution $q(x)$ for $x \in \mathcal{X}$ and use Shannon's Code according to this distribution, that is, $l(x) = \lceil \log 1/q(x) \rceil$, then

$$H_D(X) + \mathcal{D}(q||p) \leq \mathbb{E}l(X) \leq H_D(X) + \mathcal{D}(q||p) + 1.$$

Proof. You can find the proof in [Cover and Thomas \(2012\)](#). □

3.4 Differential Entropy and Information

3.4.1 Differential Entropy of Shannon

Definition 3.4.1. Let X be a continuous random vector in \mathbb{R}^n with density f , then the **differential entropy of Shannon** is defined as

$$H(X) := - \int_{\mathbb{R}^n} f \log f \, dx,$$

if the integral exists.

Remark 3.4.1. We will denote the differential entropy by the same letter H as the discrete case and we hope no confusion will be made.

Example 3.4.1. Let $X \sim \text{Unif}([0, a])$, then $f(x) = \frac{1}{a} \mathbf{1}_{[0, a]}$, hence

$$H(X) = - \int_0^a \frac{1}{a} \log \frac{1}{a} \, dx = \log a.$$

If we set $0 < a < 1$, then $H(X) < 0$.

In this example we see that the differential entropy lacks the positivity property of the discrete case. It turns out that $H(X) \in (-\infty, \infty]$ and we will see later that $H(X) \rightarrow -\infty$ corresponds to the case where X is constant, that is, the differential entropy renormalize the constants r.v. to $-\infty$, instead of 0 in the discrete case.

We also have a definition for the *Kullback-Leibler Divergence*.

Definition 3.4.2. Let μ and ν two absolutely continuous measures with respect Lebesgue in \mathbb{R}^n and $\mu \ll \nu$. Then we define the **Kullback-Leibler Divergence** as

$$\mathcal{D}(\mu||\nu) := \int_{\mathbb{R}^n} f \log \frac{f}{g} \, dx,$$

where f, g are the densities of μ and ν with respect to the Lebesgue measure, respectively.

The Kullback-Leibler Divergence preserves $\mathcal{D}(\mu||\nu) \geq 0$ and equality holds if and only if $\mu = \nu$ (this is a consequence of Lemma [4.3.1](#)).

We can define $H(X|Y)$, $H(X, Y)$ similarly as the discrete case, but they will lack the positivity property as well. However, since $I(X, Y)$ is defined in terms of \mathcal{D} , $I(X, Y) \geq 0$ and equality holds if and only if X and Y are independent.

It turns out that the differential entropy has other properties. The first ones are the *dilation* and *translation* property.

Lemma 3.4.1. *Let $a > 0$ and $b \in \mathbb{R}^n$. If X is a random vector in \mathbb{R}^n with finite entropy, then*

$$H(aX + b) = H(X) + n \log a.$$

Proof. It is easy to see that $H(aX + b) = H(aX)$, since we just translate the density and thus the integral does not change. Hence we can consider $b = 0$.

Let $f(x)$ be the density of X , then $\frac{1}{a^n}f(x/a)$ is the density of aX , therefore

$$H(aX) = - \int_{\mathbb{R}^n} \frac{1}{a^n} f(x/a) \log \left(\frac{1}{a^n} f(x/a) \right) dx.$$

Changing variable to $y = x/a$, we have that $a^n dy = dx$, then

$$H(aX) = - \int_{\mathbb{R}^n} f(y) \log \frac{f(y)}{a^n} dy = H(X) + n \log a,$$

and the lemma is proved. □

We can see in this case that if $a \searrow 0$, then $aX \rightarrow 0$ in probability and

$$H(aX) \searrow -\infty.$$

Moreover, we can generalize the dilation property to an affine invertible transformation $T(X) = AX + b$.

Corollary 3.4.1. *Let A be an invertible matrix $n \times n$ and $b \in \mathbb{R}^n$, then*

$$H(AX + b) = H(X) + \log |A|,$$

where $|A|$ is the absolute value of the determinant of A .

3.4.2 Maximum Entropy

We will explore here the problem of maximizing entropy for families of random variables, first for the case of random variables with the same mean and variance, and then for the case where the random variables satisfy a general condition $\mathbb{E}[W(X)] = c$, where W will be defined later in this subsection and c is a constant.

Corollary 3.4.2. *Let X be a r.v. with finite variance $\sigma^2 = \text{Var}(X)$ and $\mu = \mathbb{E}[X]$, then $H(X) \leq H(Y)$ for $Y \sim \mathcal{N}(\mu, \sigma^2)$ and equality only holds if and only if $X \stackrel{d}{=} Y$.*

Proof. Because of Lemma 3.4.1, we can assume $\sigma^2 = 1$ and $\mu = 0$. Let f be the density of X and $\gamma(x) := \frac{1}{\sqrt{2\pi}}e^{-x^2/2}$ the density of the Gaussian. Then

$$\begin{aligned} 0 \leq \mathcal{D}(f dx || \gamma dx) &= \int_{\mathbb{R}} f \log \frac{f}{\gamma} dx \\ &= -H(X) - \int_{\mathbb{R}} f \log \gamma dx, \end{aligned}$$

but

$$\begin{aligned} - \int_{\mathbb{R}} f \log \gamma dx &= \int_{\mathbb{R}} f \log[\sqrt{2\pi}] dx + \frac{1}{2} \int_{\mathbb{R}} x^2 f dx \\ &= \log \sqrt{2\pi} + 1/2, \end{aligned}$$

because the variance of X is 1. Since

$$\int_{\mathbb{R}} \gamma dx = 1,$$

and

$$\int_{\mathbb{R}} x^2 \gamma dx = 1,$$

we also have that

$$- \int_{\mathbb{R}} f \log \gamma dx = - \int_{\mathbb{R}} \gamma \log \gamma dx,$$

hence

$$0 \leq -H(X) - \int_{\mathbb{R}} \gamma \log \gamma dx = -H(X) + H(Y),$$

and the theorem is proved. □

As a corollary, we can generalize it to \mathbb{R}^n .

Corollary 3.4.3. *Let X be a random vector in \mathbb{R}^n with mean $\mu \in \mathbb{R}^n$ and positive definite covariance matrix Σ , then $H(X) \leq H(Y)$ where $Y \sim \mathcal{N}(\mu, \Sigma)$ with equality whenever $X \stackrel{d}{=} Y$.*

In the previous proof, we indirectly computed the entropy of the Gaussian.

Lemma 3.4.2. *Let $X \sim \mathcal{N}(0, 1)$, then $H(X) = \log \sqrt{2\pi} + 1/2 = \frac{1}{2} \log 2\pi e$. Also, if $Y \sim \mathcal{N}(0, \sigma^2)$, then $Y \stackrel{d}{=} \sigma X$ and $H(Y) = H(X) + \log \sigma = \frac{1}{2} \log 2\pi e \sigma^2$. Finally, if $X \in \mathbb{R}^n$ and $X \sim \mathcal{N}(\mu, \Sigma)$, then $H(X) = \frac{n}{2} \log(2\pi e |\Sigma|^{1/n})$.*

Now we can compute the renormalized constant of the differential entropy.

Corollary 3.4.4. *Let X be a random variable with $\text{Var}(X) = \sigma^2$ and $\mu = \mathbb{E}[X]$. If $\sigma \rightarrow 0$ then $X \xrightarrow{\mathbb{P}} \mu$ and $H(X) \rightarrow -\infty$.*

Proof. The first one is a consequence of the Weak Law of Large Numbers. The second part is a consequence of $H(X) \leq \frac{1}{2} \log(2\pi e \sigma^2)$ when $\sigma \rightarrow 0$. \square

Using the same relative entropy argument, we can also maximize the entropy on compact sets.

Theorem 3.4.1. *Let X be a random vector distributed on a compact set $K \subset \mathbb{R}^n$, with density f , that is $f(K^c) \equiv 0$ and*

$$\int_K f \, dx = 1,$$

then $H(X) \leq \log \lambda(K)$, where $\lambda(K)$ is the Lebesgue measure of K and equality only holds if X is uniform in K .

Finally, we can generalize Theorem 3.4.2 for a large class of densities in the form $f(x) = \frac{1}{Z} e^{-W(x)}$ where W satisfies some conditions.

Definition 3.4.3. Let $W : \mathbb{R}^n \rightarrow \mathbb{R}$ be a strongly convex function, that is, the Hessian of W is positive definite matrix and

$$\text{Hess } W(x) \succeq c \text{Id},$$

for some $c > 0$ and all $x \in \mathbb{R}^n$, where \succeq is the partial order induced by the cone of positive semidefinite matrices in $n \times n$. Let

$$Z := \int_{\mathbb{R}^n} e^{-W(x)} \, dx,$$

which is finite, and

$$f(x) := \frac{1}{Z} e^{-W(x)}.$$

Set μ the probability measure in \mathbb{R}^n such that

$$\frac{d\mu}{dx} = f,$$

then μ is known as the **Boltzmann Measure** associated with the **potential** W .

Remark 3.4.2. The strong convexity condition is not necessary in this definition: we just have to assume that Z is well-defined. However, we use this definition because we will need it as a hypothesis in Theorem 4.4.6.

Now we can prove that the Boltzmann measures maximizes the entropy under some conditions.

Theorem 3.4.2. Let $W : \mathbb{R}^n \rightarrow \mathbb{R}$ be a strongly convex function, μ the Boltzmann measure associated with W and $Y \sim \mu$. Hence, for all X such that $\mathbb{E}[W(X)] = \mathbb{E}[W(Y)]$ we have that $H(X) \leq H(Y)$ with equality whenever $X \stackrel{d}{=} Y$.

Remark 3.4.3. When $W(x) := \|x\|^2$, we recover Corollary 3.4.2, since

$$\mathbb{E}[W(X)] = \mathbb{E}[X^2] = \text{Var}(X),$$

for $\mathbb{E}[X] = 0$ and $Y \sim \mathcal{N}(0, \text{Var}(X))$.

Proof. We will use the same procedure as in the Gaussian case. Let g, f be the densities of X and Y , respectively. Then

$$\begin{aligned} 0 \leq \mathcal{D}(gdx||\mu) &= \int_{\mathbb{R}^n} g \log g/f \, dx \\ &= -H(X) - \int_{\mathbb{R}^n} g \log f \, dx. \end{aligned}$$

To compute the last term in the right-hand side, notice that

$$\begin{aligned} - \int_{\mathbb{R}^n} g \log f \, dx &= - \int_{\mathbb{R}^n} g \log \left(\frac{1}{Z} e^{-W} \right) \, dx \\ &= \int_{\mathbb{R}^n} g \log Z \, dx + \int_{\mathbb{R}^n} gW \, dx. \end{aligned}$$

Since

$$\int_{\mathbb{R}^n} g \log Z \, dx = \log Z = \int_{\mathbb{R}^n} f \log Z \, dx,$$

because g and f are densities, and

$$\int_{\mathbb{R}^n} gW \, dx = \mathbb{E}[W(X)],$$

by hypothesis we have that

$$\int_{\mathbb{R}^n} gW \, dx = \mathbb{E}[W(Y)] = \int_{\mathbb{R}^n} fW \, dx,$$

then

$$- \int_{\mathbb{R}^n} g \log f \, dx = - \int_{\mathbb{R}^n} f \log f \, dx = H(Y),$$

hence we obtain the result. □

3.4.3 Exponential Entropy of Shannon

Definition 3.4.4. Let X be a random vector in \mathbb{R}^n with finite differential entropy $H(X)$, then the **exponential entropy** of X is defined as

$$N(X) = \frac{1}{2\pi e} e^{\frac{2}{n} H(X)}.$$

Therefore, the exponential entropy renormalizes back to 0 the constants vectors. The constants $\frac{1}{2\pi e}$ and $2/n$ also normalize the standard Gaussian case.

Lemma 3.4.3. *Let $X \sim \mathcal{N}(0, \text{Id})$, then $N(X) = 1$.*

Proof. We have already seen that $H(X) = \frac{n}{2} \log(2\pi e)$, hence the result. \square

The maximum entropy and the dilation property can be rewritten for the exponential entropy.

Corollary 3.4.5. *Let X be a centered random vector with positive definite covariance matrix Σ and $Y \sim \mathcal{N}(0, \Sigma)$, then $N(X) \leq |\Sigma|^{1/n} = N(Y)$ and equality holds whenever $X \stackrel{d}{=} Y$.*

Corollary 3.4.6. *Let X be a random vector in \mathbb{R}^n with finite differential entropy $H(X)$. Let A be a nonsingular matrix, then*

$$N(AX) = |A|^{2/n} N(X).$$

In particular, $N(aX) = a^2 N(X)$ for $a \in \mathbb{R}$.

Later in Section 3.6, we will explore an inequality relating the exponential entropy of the sum $N(X + Y)$ and the single ones $N(X)$ and $N(Y)$. In the Gaussian case, we have the following example.

Example 3.4.2. Let $X \sim \mathcal{N}(0, \Sigma_1)$ and $Y \sim \mathcal{N}(0, \Sigma_2)$ and $\Sigma_1 = r\Sigma_2$ for some $r \in \mathbb{R}_+$. Then $X + Y \sim \mathcal{N}(0, \Sigma_1 + \Sigma_2)$ and

$$N(X + Y) = |(1 + r)\Sigma_2|^{1/n} = (1 + r)|\Sigma_2|^{1/n} = N(Y) + N(X),$$

and hence the exponential entropy of the sum is equal to the sum of the exponential entropies. We will explore this equality in Subsection 3.6.2.

3.4.4 Fisher Information according to a parameter

Let $f : A \subset \mathbb{R}^n \times \mathbb{R}^d \rightarrow \mathbb{R}$ be a function and $(x, y) \in \mathbb{R}^n \times \mathbb{R}^d$. We will denote the gradient in the x variable as

$$\nabla_x f(x, y) := \left(\frac{\partial}{\partial x_i} f(x, y) \right)_{i=1}^n,$$

and likewise for $y \in \mathbb{R}^d$.

We also need to define the *classes of parameters* and *densities*.

Definition 3.4.5. A **class of parameters**, or **family of parameters**, is a subset $\Theta \subseteq \mathbb{R}^d$, for some $d \in \mathbb{N}$. An element $\theta \in \Theta$ is called **parameter**.

Definition 3.4.6. Let $\Theta \subseteq \mathbb{R}^d$ be a class of parameters. By a **class of densities** according to Θ , we mean a function $f : \mathbb{R}^n \times \Theta \rightarrow \mathbb{R}$ such that for each $\theta \in \Theta$, the function $f(\cdot, \theta)$ is a density. We denote $f(x, \theta)$ simply by $f(x|\theta)$.

Remark 3.4.4. We will also denote a class of densities according to Θ as $\{f(x|\theta) : \theta \in \Theta\}$.

Now we can establish the first important definition of this subsection.

Definition 3.4.7. Let $\Theta \subseteq \mathbb{R}^d$ be a class of parameters and $\mathcal{F} := \{f(x|\theta) : \theta \in \Theta\}$ be a family of densities in \mathbb{R}^n according to Θ such that, for each $x \in \mathbb{R}^n$, $f(x|\cdot) \in C^1(\Theta)$. Given $\theta \in \Theta$, let X be a sample of the distribution with density $f(\cdot|\theta)$. The **score function** is defined as

$$V := \nabla_{\theta} \log f(X|\theta),$$

or, in the discrete case,

$$V := \nabla_{\theta} \log p(X|\theta),$$

where $p(x|\theta) = \mathbb{P}(X = x|\theta)$.

Remark 3.4.5. Notice that V depends on X , \mathcal{F} and the particular $\theta \in \Theta$ we took. However, the dependence on X will disappear as soon as we start to take expected values.

The next lemma gives sufficient conditions to avoid computing the expected value of V .

Lemma 3.4.4. Let $\|\cdot\|$ be any norm in \mathbb{R}^d and suppose $\|\nabla_{\theta} f(\cdot|\theta)\| \leq g(\cdot)$ for all $\theta \in \Theta$ and $g \in L^1(\mathrm{d}x)$, then $\mathbb{E}[V] = 0$.

Proof. Set $n = 1$. The general case $n \in \mathbb{N}$ is a corollary of the case $n = 1$ and the Fubini's Theorem. Notice first that

$$\mathbb{E}[V] = \int_{\mathbb{R}} f(x|\theta) \frac{\partial}{\partial \theta} \log f(x|\theta) \mathrm{d}x = \int_{\mathbb{R}} \frac{\partial}{\partial \theta} f(x|\theta) \mathrm{d}x.$$

But $|\partial_{\theta} f(x|\theta)| \leq g(x)$ is a sufficient condition to change the order in the derivative (see Theorem 2.4.4). Hence

$$\mathbb{E}[V] = \partial_{\theta} \int_{\mathbb{R}} f(x|\theta) \mathrm{d}x = \partial_{\theta} 1 = 0,$$

and the lemma is proved. □

Example 3.4.3. Let $\Theta = [0, 1]$ and $\mathcal{X} = \{0, 1\}$. Let X be Bernoulli with parameter $\theta \in \Theta$, that is, $p(1|\theta) = \theta$ and $p(0|\theta) = 1 - \theta$, hence we have that

$$V = \begin{cases} \theta^{-1} & \text{if } X = 1; \\ -(1 - \theta)^{-1} & \text{if } X = 0. \end{cases}$$

Therefore we obtain

$$\mathbb{E}[V] = \frac{\theta}{\theta} - \frac{1 - \theta}{1 - \theta} = 0.$$

Example 3.4.4. Let $\sigma^2 > 0$ and $\Theta = \mathbb{R}^n$. Given $\theta \in \mathbb{R}^n$, set $X \sim \mathcal{N}(\theta, \sigma^2 \text{Id})$, then

$$f(x|\theta) = \frac{1}{(2\pi\sigma^2)^{n/2}} e^{-\frac{\|x-\theta\|^2}{2\sigma^2}}.$$

Its log-derivative is

$$\nabla \log f(x|\theta) = \frac{1}{\sigma^2}(x - \theta).$$

Therefore, we have

$$\mathbb{E}[V] = \frac{1}{\sigma^2} \int_{\mathbb{R}} (x - \theta) f(x|\theta) dx = 0,$$

since

$$\int_{\mathbb{R}} x f(x|\theta) dx = \theta.$$

This is a case where the conditions in Lemma 3.4.4 fail, although we still have $\mathbb{E}[V] = 0$. Indeed, the maximum of $|\partial f(x|\theta)|$ is constant in x , therefore not Lebesgue integrable.

Throughout this dissertation, we will only consider cases where $\mathbb{E}[V] = 0$.

The variance of the Score Function is known as the *Fisher Information*.

Definition 3.4.8. Let $\mathcal{F} := \{f(x|\theta) : \theta \in \Theta\}$ be a family of densities in \mathbb{R}^n according to Θ . Set $V = \nabla \log f(X|\theta)$ and assume $\mathbb{E}[V] = 0$, for X with density $f(x|\theta)$. We define the **Fisher Information** of the family \mathcal{F} as the covariance matrix of V , that is,

$$J(\theta) := \left(\mathbb{E}[V_i V_j] \right)_{i,j=1}^d$$

In the real case, we have explicitly

$$J(\theta) = \mathbb{E} \left(\partial_{\theta} \log f(X|\theta) \right)^2.$$

If we consider $(X_i)_{i=1}^n$ be i.i.d according to the densities $f(x|\theta)$, their joint score function is

$$V = \partial_{\theta} \log f(X_1, \dots, X_n|\theta) = \sum_{i=1}^n V_i,$$

hence

$$J(\theta) = nJ_1(\theta),$$

and we recover the property that, in the independent case, the joint Information is the sum of the individual Informations.

Example 3.4.5. Let $\Theta = \mathbb{R}$, $\sigma^2 > 0$ and $X_i \sim \mathcal{N}(\theta, \sigma^2)$ i.i.d for $i \leq n$, hence

$$J(\theta) = \frac{n}{\sigma^2},$$

Indeed, if f is the density of $\mathcal{N}(\theta, \sigma^2)$, then

$$\log f(x) = -\frac{1}{2} \log(2\pi\sigma^2) - \frac{(x - \theta)^2}{2\sigma^2},$$

and

$$\partial_\theta \log f(x) = \frac{(x - \theta)}{\sigma^2},$$

hence

$$J_1(\theta) = \int_{\mathbb{R}} \frac{(x - \theta)^2}{\sigma^4} f(x) dx,$$

and we recognize the second moment of the Gaussian r.v., that is,

$$J_1(\theta) = \frac{\sigma^2}{\sigma^4} = \frac{1}{\sigma^2}.$$

3.4.5 Fisher Information

If we fix a random vector X in \mathbb{R}^n with density f with respect to the Lebesgue measure and consider the family of densities $\{f(x - \theta) : \theta \in \Theta\}$, for $\theta \in \mathbb{R}^n$, then we can define the *Fisher Information* of X .

Definition 3.4.9. Let X be a random vector in \mathbb{R}^n with density f . Let $\Theta = \mathbb{R}^n$, hence the **Fisher Information** of X is defined as the Fisher Information of the family

$$\mathcal{F} := \{f(x - \theta) : \theta \in \Theta\},$$

that is,

$$J(X) := J(\theta).$$

Lemma 3.4.5. We have that $J(X)$ depends only on f .

Proof. Notice that

$$J(X) = J(\theta) = \int_{\mathbb{R}^n} \left\| \nabla_\theta \log f(x - \theta) \right\|^2 f(x - \theta) dx,$$

but for all $x, \theta \in \mathbb{R}^n$ we have that

$$\left\| \nabla_\theta \log f(x - \theta) \right\|^2 = \left(\frac{\left\| \nabla_x f(x - \theta) \right\|}{f(x - \theta)} \right)^2,$$

hence

$$J(X) = \int_{\mathbb{R}^n} \frac{\left\| \nabla f(y) \right\|^2}{f(y)} dy,$$

by change of variables $y = x - \theta$. □

By the chain rule, we can equivalently define $J(X)$.

Lemma 3.4.6. *Let X be a random vector in \mathbb{R}^n with finite Fisher Information $J(X)$. Then*

$$J(X) = \int_{\mathbb{R}^n} \frac{\|\nabla f(x)\|^2}{f(x)} dx = 4 \int_{\mathbb{R}^n} \|\nabla \sqrt{f}\|^2 dx = \int_{\mathbb{R}^n} \langle \nabla f, \nabla \log f \rangle dx,$$

or, in terms of expected value,

$$J(X) = \mathbb{E} \left(\frac{\|\nabla f(X)\|^2}{f(X)} \right).$$

As we have seen in Example 3.4.5, if $X \sim \mathcal{N}(0, \sigma^2)$, then $J(X) = 1/\sigma^2$ and we have that $J(X)N(X) = 1$, a relation we will explore in the final Chapter 6. Therefore, if X and Y are independent Gaussian random variables, we have

$$\frac{1}{J(X+Y)} = N(X+Y) = N(X) + N(Y) = \frac{1}{J(X)} + \frac{1}{J(Y)}.$$

We will also explore this equality in the final section of this chapter. (see Subsection 3.6.1).

Example 3.4.6. For the standard Gaussian $X \sim \mathcal{N}(0, \text{Id})$, $J(X) = n$. Indeed, the property that the joint Fisher Information of i.i.d random variables is n times the Fisher Information of the first gives the result.

We also have the dilation property.

Lemma 3.4.7. *Let X with finite Fisher Information, then $J(aX + b) = a^{-2}J(X)$, for $a > 0$ and $b \in \mathbb{R}^n$.*

Proof. The translation does not affect the Fisher Information, then we can consider $b = 0$. The density of aX is equal to $\frac{1}{a^n}f(x/a)$, where f is the density of X , hence

$$J(aX) = 4 \int_{\mathbb{R}^n} \left\| \nabla_x \sqrt{\frac{1}{a^n}f(x/a)} \right\|^2 dx = \frac{4}{a^n} \int_{\mathbb{R}^n} \frac{1}{a^2} \left\| \nabla_{x/a} \sqrt{f(x/a)} \right\|^2 dx.$$

By the change of variables to $x = ay$, we have

$$J(aX) = 4 \int_{\mathbb{R}^n} \frac{1}{a^2} \left\| \nabla_y \sqrt{f(y)} \right\|^2 dy = \frac{1}{a^2} J(X).$$

□

The formula for the Fisher Information of AX is a little more complicated and we will explore a more general concept in the next subsection.

3.4.6 Fisher Matrix

We can give a *matrix definition* of the Fisher Information. Let ∇f be a vector $n \times 1$, then we have the following.

Definition 3.4.10. Let X be a random vector in \mathbb{R}^n with density f with respect to the Lebesgue measure. Then the **Fisher Matrix** is defined as

$$\mathbb{J}(X) := \int_{\mathbb{R}^n} \nabla f \cdot (\nabla f)^T \frac{1}{f} dx.$$

Remark 3.4.6. This definition is consistent in the following sense:

$$\text{tr}(\mathbb{J}(X)) = J(X),$$

and we have that \mathbb{J} is the covariance matrix Σ of $\nabla \log f(X)$, since

$$\Sigma = \mathbb{E}[\nabla \log f(X) \nabla \log f(X)^T] = \int_{\mathbb{R}^n} \nabla \log f(x) \nabla \log f(x)^T f(x) dx,$$

where the last one is the Fisher Matrix, by the chain rule. Hence $\mathbb{J}(X)$ is a positive semidefinite matrix and is singular if and only if X lies in a lower dimension subspace.

Example 3.4.7. $\mathbb{J}(X) = \text{Id}$, for $X \sim \mathcal{N}(0, \text{Id})$. Indeed, let $f(x) = (2\pi)^{n/2} \exp(-\|x\|^2/2)$ be the density of the standard Gaussian, then

$$\partial_i \log f(x) \partial_j \log f(x) = x_i x_j,$$

hence

$$\mathbb{J}(X)_{ij} = \int_{\mathbb{R}^n} x_i x_j f(x) dx = \text{Id}_{ij}.$$

For the Fisher Matrix, the dilation property is the following.

Lemma 3.4.8. Let X be a random vector in \mathbb{R}^n with finite Fisher Matrix $\mathbb{J}(X)$ and A be an $n \times n$ invertible matrix. If $Y = AX$, then $\mathbb{J}(Y) = A^{-1} \mathbb{J}(X) (A^{-1})^T$.

Proof. The density of Y is

$$g(y) = |A|^{-1} f(A^{-1}y).$$

Therefore, the Fisher Matrix of Y is

$$\mathbb{J}(Y) = |A|^{-1} \int_{\mathbb{R}^n} \nabla f(A^{-1}y) \nabla f(A^{-1}y)^T \frac{1}{f(A^{-1}y)} dy.$$

Let $A^{-1} = B$ and $u = By$, then

$$\partial_{y_i} f(By) = \sum_{j=1}^n \partial_{u_j} f(u) \partial_{y_i} u_j = \sum_{j=1}^n \partial_{u_j} f(u) B_{ji},$$

then $\nabla_y f(By) = B^T \nabla_u f(u)$, therefore

$$\mathbb{J}(Y) = |A|^{-2} \int_{\mathbb{R}^n} B^T \nabla_u f(u) \nabla_f(u) B \frac{1}{f(u)} dy.$$

Finally, we have $dy = |A| du$, hence

$$\mathbb{J}(Y) = \int_{\mathbb{R}^n} B^T \nabla_u f(u) \nabla_f(u) B \frac{1}{f(u)} du.$$

And the result follows taking B^T and B out of the integral. \square

Corollary 3.4.7. *Let $A = \mathbb{J}(X)^{1/2}$, the square root of the Fisher Matrix, then $\mathbb{J}(AX) = \text{Id}$.*

Therefore, we have the following dilation property for the Fisher Information.

Corollary 3.4.8. *Let A be a nonsingular $n \times n$ matrix and X be a random vector in \mathbb{R}^n with finite Fisher Matrix, then*

$$J(AX) = \text{tr}(A^{-1} \mathbb{J}(X) (A^{-1})^T).$$

For the Guassian case $X \sim \mathcal{N}(0, \Sigma)$, we have

$$J(X) = \sum_{i=1}^n \frac{1}{\sigma^2(X_i)},$$

where $\sigma^2(X_i) = \Sigma_{ii}$, for $i = 1, \dots, n$.

Proof. The first is immediate, since $\text{tr}(\mathbb{J}(X)) = J(X)$. For the second, we have that $X \stackrel{d}{=} \Sigma^{1/2} Y$, where $Y \sim \mathcal{N}(0, \text{Id})$, hence

$$\mathbb{J}(X) = \Sigma^{-1/2} \Sigma^{-1/2} = \Sigma^{-1}.$$

The eigenvalues of Σ^{-1} are the reciprocal of the eigenvalues of Σ , hence the result. \square

Again, if $X \sim \mathcal{N}(0, \Sigma_1)$, $Y \sim \mathcal{N}(0, \Sigma_2)$, $\Sigma_1 = \sigma \Sigma_2$, for some $\sigma > 0$, and X, Y are independent, then

$$\mathbb{J}(X + Y)^{-1} = (\Sigma_1 + \Sigma_2) = \mathbb{J}(X)^{-1} + \mathbb{J}(Y)^{-1}.$$

Likewise, we can define the parametric version of the Fisher Matrix.

Definition 3.4.11. Let $\mathcal{F} := \{f(x|\theta) : \theta \in \Theta\}$ be a parametric family of densities in \mathbb{R}^n . Hence, the **Fisher Matrix** according to this family is

$$\mathbb{J}(\theta) := \int_{\mathbb{R}^n} \nabla_{\theta} f(x|\theta) \nabla_{\theta} f(x|\theta)^T \frac{1}{f(x|\theta)} dx.$$

Hence we have the following lemma.

Lemma 3.4.9. *Let X be a random vector in \mathbb{R}^n with density f . Let $\{f(x - \theta) : \theta \in \mathbb{R}^n\}$ be a parametric family, hence*

$$\mathbb{J}(X) = \mathbb{J}(\theta).$$

3.4.7 Fisher and Kullback-Leibler Divergence

There is a wonderful relation between the Fisher Matrix and the Kullback-Leibler Divergence. Let $\{f(x|\theta) : \theta \in \Theta\}$ be a parametric family of densities and Θ be an open set in \mathbb{R}^n and, for $\theta \in \Theta$, consider μ_θ the probability measure with density $f(\cdot|\theta)$. For a fixed $\theta \in \Theta$, consider $g : \Theta \rightarrow \mathbb{R}_+$ the function $g(\alpha) := \mathcal{D}(\mu_\theta || \mu_\alpha)$. Let us compute the Hessian of g .

The first partial derivative is

$$\partial_i g(\alpha) = - \int_{\mathbb{R}^n} f(x|\theta) \frac{\partial_i f(x|\alpha)}{f(x|\alpha)} dx, \quad (3.2)$$

if we can differentiate under the integral (for sufficient conditions, see Theorem 2.4.4). Notice that, for $\alpha = \theta$, we've already known that θ is a global minimum of the relative entropy, hence

$$\partial_i g(\theta) = 0.$$

The second partial derivative is

$$\partial_j \partial_i g(\alpha) = - \int_{\mathbb{R}^n} f(x|\theta) \left[\frac{f(x|\alpha) \partial_j \partial_i f(x|\alpha) - \partial_i f(x|\alpha) \partial_j f(x|\alpha)}{[f(x|\alpha)]^2} \right] dx, \quad (3.3)$$

where again we differentiate under the integral sign. For $\alpha = \theta$, we have

$$\partial_j \partial_i g(\theta) = - \int_{\mathbb{R}^n} \partial_j \partial_i f(x|\theta) dx + \int_{\mathbb{R}^n} \frac{\partial_i f(x|\theta) \partial_j f(x|\theta)}{f(x|\theta)} dx.$$

The first integral is zero by exchanging integral with derivatives. The second is the ij -entry of the Fisher Matrix $\mathbb{J}(\theta)$. Hence we have the following theorem.

Theorem 3.4.3. *Let $\Theta \subseteq \mathbb{R}^n$ be an open set, $\{f(x|\theta) : \theta \in \Theta\}$ be a parametric family of densities and $\{\mu_\theta : \theta \in \Theta\}$ be the corresponding family of measures. If the Equations 3.2 and 3.3 hold, then for $\theta \in \Theta$ fixed, the Hessian of $\mathcal{D}(\mu_\theta || \mu_\alpha)$ in $\alpha = \theta$ is precisely $\mathbb{J}(\theta)$ and we have the second order Taylor Expansion:*

$$\mathcal{D}(\mu_\theta || \mu_\alpha) \sim \frac{1}{2}(\alpha - \theta)^T \mathbb{J}(\theta)(\alpha - \theta).$$

3.5 Channel

In this section, we define the notion of Channel and the principal theorem derived by Shannon (1948) about the Channel Capacity and efficient codes.

3.5.1 Discrete Channel

Definition 3.5.1. A **discrete channel** is a triple $(\mathcal{X}, \mathcal{Y}, p(y|x))$ where \mathcal{X} and \mathcal{Y} are two finites sets and $p(y|x)$ is a family of conditional distributions in \mathcal{Y} given $x \in \mathcal{X}$.

Definition 3.5.2. The **extension of discrete memoryless without feedback channel** $(\mathcal{X}, p(y|x), \mathcal{Y})$ is the channel $(\mathcal{X}^n, p(y^n, x^n), \mathcal{Y}^n)$ where

$$p(y^n|x^n) = \prod_{k=1}^n p(y_k|x_k).$$

We will only consider memoryless without feedback channels and denote them by just *DMC*.

Associated to a channel, we can define its capacity.

Definition 3.5.3. Let $(\mathcal{X}, \mathcal{Y}, p(y|x))$ be a channel, then its **capacity** is

$$C := \max_p I(X, Y),$$

where the maximum is over all the probabilities distribution in \mathcal{X} , X has distribution p and Y is the induced distribution of X in \mathcal{Y} by the family $p(y|x)$, that is,

$$\mathbb{P}(Y = y) = \sum_{x \in \mathcal{X}} p(x)p(y|x),$$

and $I(X, Y)$ is the mutual information.

Now that we have defined a channel, we need to know what is a code for it.

Definition 3.5.4. An (M, n) **code for the channel** $(\mathcal{X}, p(y|x), \mathcal{Y})$ is a triple (S, f, g) , where S is a set with $|S| = M$; $f : S \rightarrow \mathcal{X}^n$ and $g : \mathcal{Y}^n \rightarrow S$ are two functions.

The set S is called the set of **messages**, the function f is the **encode function** and g is the **decode function**.

We can see a code as the diagram in Figure 4.

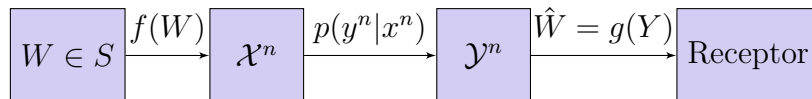


Figure 4 – Diagram representing the transmission. A message W is encoded in $f(W) \in \mathcal{X}^n$. The channel transforms this input into a noisy sign $Y \in \mathcal{Y}^n$ and the decode g guesses the best candidate $\hat{W} = g(Y)$ for the original message.

To a code (M, n) , we can associated its rate of transmission

Definition 3.5.5. Let (M, n) be a code of a DMC, then its **rate** is defined as

$$R := \frac{\log M}{n}$$

The error of transmission is defined as below.

Definition 3.5.6. To a code (M, n) , the **error** associated with the transmission of the message $i \in S$ is

$$\lambda_i := \mathbb{P}(g(Y^n) \neq i | x^n(i)) = \sum_{y^n \in \mathcal{Y}^n} p(y^n | x^n(i)) \mathbf{1}_{\{i\}}(y^n).$$

The error of the code is defined as $\lambda^{(n)} := \max_{i \in S} \lambda_i$.

Using these two last definitions, we can define *achievable rates*.

Definition 3.5.7. We say that a rate R is **achievable** if there is a sequence $(\lceil 2^{nR} \rceil, n)$ of codes such that $\lambda^{(n)} \rightarrow 0$ and it **achieves** error $\varepsilon \in (0, 1)$ if there is a sequence $(\lceil 2^{nR} \rceil, n)$ of codes such that

$$\limsup_{n \rightarrow \infty} \lambda^{(n)} \leq \varepsilon.$$

We denote \mathcal{R} all the achievable rates and $\mathcal{R}(\varepsilon)$ all rates which achieve error ε .

Note the following relation between achievable rates.

$$\mathcal{R} = \bigcap_{\varepsilon \in \mathbb{Q} \cap (0, 1)} \mathcal{R}(\varepsilon).$$

For simplicity, we denote a code $(\lceil 2^{nR} \rceil, n)$ just by $(2^{nR}, n)$.

Definition 3.5.8. Given $R_1, R_2 \in \mathcal{R}$ or in $\mathcal{R}(\varepsilon)$, we say that R_1 is more **efficient** than R_2 if $R_1 > R_2$.

The central example of DMC in this section will be the Binary Channel, perhaps the simplest channel.

Example 3.5.1. Let $\mathcal{X} = \mathcal{Y} = \{0, 1\}$ and

$$p(1|1) = 1 - p = p(0|0),$$

with $p \in (0, 1/2)$. We can represent this channel as in Figure 5. Its capacity is $C = 1 - H(p)$, where $H(p)$ is the entropy of the distribution $(p, 1 - p)$. Indeed, the mutual information can be decompose as

$$I(X, Y) = H(Y) - H(Y|X) \leq 1 - H(p),$$

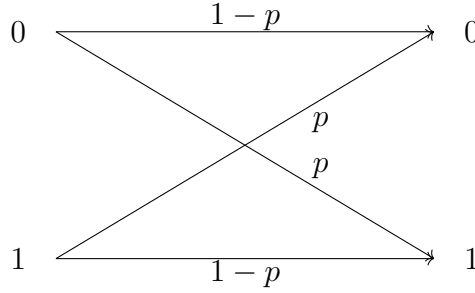


Figure 5 – Given the input $x \in \{0, 1\}$, the output is x with probability $1 - p$ and $1 - x$ with probability p .

with equality if and only if Y is uniform. It can be reached if X is uniform, since

$$\mathbb{P}(Y = 1) = \mathbb{P}(X = 1)(1 - p) + \mathbb{P}(X = 0)p = 1/2.$$

Suppose we want to transmit two messages, say $S = \{0, 1\}$ and $M = 2$. Given an $n \in \mathbb{N}$, we can encode this message by just sending it n times, that is, $f(x) = xx \dots x$ n times. The decode function g guesses x if the numbers of x in the output is greater than $1 - x$. This is known as the **repetition code**. Because of the symmetry of this channel, the error is $\lambda^{(n)} = \max_{i \in \{0, 1\}} \lambda_i = \lambda_1$. Let $X_1, \dots, X_n \sim \text{Ber}(p)$ i.i.d, then the error can be measured by

$$\lambda^{(n)} = \mathbb{P}\left(\frac{1}{n} \sum_{i=1}^n X_i > 1/2\right).$$

Since $p < 1/2$, the Weak Law of Large Numbers says that

$$\lambda^{(n)} \rightarrow 0.$$

However, this code is not efficient, since

$$R = \frac{\log M}{n} = \frac{\log 2}{n} \rightarrow 0.$$

In Chapter 5, we will derive a nonasymptotic version of this code (see Example 5.4.1).

We only consider $p < 1/2$ since we just have to exchange labels in the case $p > 1/2$ and the case $p = 1/2$ has capacity 0.

By this example, we would imagine that to get $\lambda^{(n)} \rightarrow 0$, the rate necessarily converges to 0. This is not true and it was proved in Shannon (1948).

Theorem 3.5.1 (Shannon's Theorem). *Let C be the capacity of a DMC. Then all rates $R < C$ are achievable. Conversely, if a rate R is achievable, then $R \leq C$.*

Proof. For the proof, see Cover and Thomas (2012). □

3.5.2 Continuous Channel

Definition 3.5.9. The **Gaussian channel** is a time-discrete channel with output Y_i at instant i and input X_i such that

$$Y_i = X_i + Z_i,$$

where $Z_i \sim \mathcal{N}(0, N)$ is independent of X_i for all i and $(X_i)_{i=1}^n$ are assumed to be independent.

If we assume no other conditions in this channel, we can recover the signal X_i with arbitrarily small probability. For instance, if we set the input space \mathcal{X} as a well-separated set, that is, all points are distant to each other, say,

$$d(x, \tilde{x}) \gg N, \forall x, y \in \mathcal{X},$$

then we can recover it with small probability. Therefore, we will impose a condition on the input space.

Definition 3.5.10. Let $(Y_i, X_i)_{i=1}^n$ be a Gaussian Channel. The **power constraint** in the input (X_1, \dots, X_n) is

$$\frac{1}{n} \sum_{i=1}^n \mathbb{E}[X_i] \leq P,$$

for a power $P > 0$.

It is worth to mention that, even if X is discrete, $Y = X + Z$ is continuous, since, for this case, we have

$$\mathbb{P}(Y \leq y) = \sum_{x \in \mathcal{X}} \mathbb{P}(X = x) \mathbb{P}(Z \leq y - x).$$

But

$$\mathbb{P}(Z \leq y - x) = \int_{-\infty}^{y-x} f(z) dz = \int_{-\infty}^y f(z - x) dz,$$

where f is the density of Z . Hence the density of Y is

$$g(y) := \sum_{x \in \mathcal{X}} \mathbb{P}(X = x) f(y - x).$$

Let us compute the capacity of this channel $C = \max I(X, Y)$, where the maximum is over all continuous distribution X such that $\mathbb{E}[X^2] \leq P$.

Lemma 3.5.1. *We have that*

$$C = \frac{1}{2} \log \left(1 + P/N \right).$$

Proof. The mutual information is equal to

$$I(X, Y) = H(Y) - H(Y|X) = H(Y) - H(X + Z|X).$$

Since X is constant given X and the entropy is invariant under translation, we have

$$I(X, Y) = H(Y) - H(Z|X) = H(Y) - H(Z).$$

The power constraint means that

$$\mathbb{E}[Y^2] = \mathbb{E}(X + Z)^2 = \mathbb{E}[X^2] + \mathbb{E}[Z^2] \leq P + N,$$

hence we obtain that $H(Y) \leq \frac{1}{2} \log 2\pi e(P + N)$ by Corollary 3.4.2, and then

$$I(X, Y) \leq \frac{1}{2} \log 2\pi e(P + N) - \frac{1}{2} \log 2\pi eN = \frac{1}{2} \log \left(1 + P/N\right).$$

The equality holds if $X \sim \mathcal{N}(0, P)$. □

Likewise, we define a *code*.

Definition 3.5.11. An (M, n) **code** for the Gaussian Channel with Power Constraint P is a triple (S, f, g) such that $|S| = M$, $f : S \rightarrow \mathbb{R}^n$ and $g : \mathbb{R}^n \rightarrow S$ such that

$$\frac{1}{n} \sum_{i=1}^n (f(w))_i^2 \leq P,$$

for all $w \in S$.

Similarly to the discrete case, the function f is the encode function and g is the decode function. Moreover, we also define **error**, **rate**, **efficiency** and **achievable rates** similarly in the discrete case. Finally, we also have the Shannon Theorem in this case.

Theorem 3.5.2 (Shannon's Theorem). *If $R < C$, then there is a sequence $(2^{nR}, n)$ of codes such that $\lambda^{(n)} \rightarrow 0$. Conversely, if $\lambda^{(n)} \rightarrow 0$, then $R \leq C$.*

Proof. For the proof, see Cover and Thomas (2012). □

3.6 Inequalities in Information Theory

In this section we will prove two inequalities that will be important in Chapter 6. They are Fisher Information Inequality, presented in Theorem 3.6.4, and Exponential Entropy Inequality, presented in Theorem 3.6.5.

3.6.1 Fisher Information Inequality

The main tool to prove Fisher Information Inequality is the following *Fisher Matrix Inequality*.

Theorem 3.6.1. *Let $X, Y \in \mathbb{R}^n$ be independent random vectors and $Z = X + Y$. Let f, g and $f * g$ be their densities with respect Lebesgue and assume they have finite Fisher matrices $\mathbb{J}(X), \mathbb{J}(Y)$ and $\mathbb{J}(Z)$. Assume the integrability condition holds for all $y, z \in \mathbb{R}^n$ and all i :*

$$|g(y)\partial_{z_i}f(z-y)| \leq h(z),$$

where $h \in L^1(dx)$. Then, for all $n \times n$ matrices A , we have

$$A\mathbb{J}(X)A^T + (\text{Id} - A)\mathbb{J}(Y)(\text{Id} - A)^T - \mathbb{J}(Z) \succeq 0.$$

Remark 3.6.1. The integrability condition is just a sufficient condition to exchange integral and derivative (see Theorem 2.4.4).

In order to prove this theorem, we need a *matrix-valued Jensen's Inequality* whose proof we omit. We can find further informations on matrix-valued functions and convexity in [Boyd and Vandenberghe \(2004\)](#).

Theorem 3.6.2. *Let $S \subseteq \mathbb{R}^n$ be a convex set and (M_n, \preceq) be the set of all $n \times n$ matrices with the partial order defined by the positive semidefinite cone. Let $f : S \rightarrow M_n$ be a convex function, that is, $f(\lambda x + (1 - \lambda)y) \preceq \lambda f(x) + (1 - \lambda)f(y)$, for all $x, y \in S$ and $\lambda \in [0, 1]$. Then, for all integrable random vectors $X \in \mathbb{R}^n$, we have*

$$\mathbb{E}[f(X)] \succeq f(\mathbb{E}[X]).$$

We just need one example of convex matrix-valued function.

Corollary 3.6.1. *The function $f : \mathbb{R}^n \rightarrow M_n$, given by $f(u) = uu^T$, is convex.*

Proof. See [Boyd and Vandenberghe \(2004\)](#). □

Now we can prove Theorem 3.6.1

Proof. The proof we present was first shown by [Dembo \(1990\)](#). Set

1. $S_X = \nabla \log f(X)$;
2. $S_Y = \nabla \log g(Y)$;
3. $S_Z = \nabla \log f * g(Z)$; and

$$4. S_z = \nabla \log f * g(z),$$

then $\mathbb{J}(X)$ is the covariance matrix of S_X , $\mathbb{J}(Y)$ is the covariance matrix of S_Y and $\mathbb{J}(Z)$ is the covariance matrix of S_Z . The density of X given $Z = z$ is equal to

$$h(x|z) = \frac{f(x)g(z-x)}{f * g(z)},$$

hence the convolution rule and the integrability condition imply that

$$\begin{aligned} \mathbb{E}[S_X|Z = z] &= \mathbb{E}\left[\frac{\nabla f(X)}{f(X)} \middle| Z = z\right] = \frac{\int_{\mathbb{R}^n} \nabla f(x)g(z-x) dx}{f * g(z)} \\ &= \frac{\int_{\mathbb{R}^n} \nabla_z f(z-y)g(y) dy}{f * g(z)} \\ &= \frac{\nabla(f * g(z))}{f * g(z)} = S_z. \end{aligned}$$

Likewise, we have that

$$\mathbb{E}[S_Y|Z = z] = S_z.$$

Let $U = AS_X + (\text{Id} - A)S_Y$, then

$$\mathbb{E}[U|Z] = \mathbb{E}\left([AS_X + (\text{Id} - A)S_Y]|Z\right) = AS_Z + (\text{Id} - A)S_Z = S_Z.$$

Conditional Matrix-Valued Jensen's Inequality applied to the function $u \rightarrow uu^T$ implies that

$$\mathbb{E}[UU^T|Z] - \mathbb{E}[U|Z](\mathbb{E}U|Z)^T \succeq 0. \quad (3.4)$$

The expression $\mathbb{E}[UU^T|Z]$ can be rewritten as

$$\begin{aligned} \mathbb{E}[UU^T|Z] &= \mathbb{E}\left([AS_X + (\text{Id} - A)S_Y][AS_X + (\text{Id} - A)S_Y]^T|Z\right) \\ &= A[\mathbb{E}(S_X S_X^T)]A^T + (\text{Id} - A)\mathbb{E}(S_Y S_Y)(\text{Id} - A)^T \\ &\quad + A\mathbb{E}[S_X S_Y^T](\text{Id} - A)^T + (\text{Id} - A)\mathbb{E}[S_Y S_X^T]A^T. \end{aligned}$$

The last line is equal to 0, since X and Y are independent. Finally, using the fact that the Fisher Matrix is the covariance matrix of S , we conclude

$$A\mathbb{J}(X)A^T + (\text{Id} - A)\mathbb{J}(Y)(\text{Id} - A)^T - \mathbb{J}(X + Y) \succeq 0.$$

□

We can use Theorem 3.6.1 to prove the following corollary concerning the *convexity* of Fisher Matrix. Set $A = \lambda \text{Id}$ and rescale $X \rightarrow \sqrt{\lambda}X$ and $Y \rightarrow \sqrt{1-\lambda}Y$, then the following is true.

Corollary 3.6.2. *For all independent r.v. X, Y in \mathbb{R}^n and $\lambda \in [0, 1]$, we have that*

$$\mathbb{J}(\sqrt{\lambda}X + \sqrt{1-\lambda}Y) \preceq \lambda\mathbb{J}(X) + (1-\lambda)\mathbb{J}(Y).$$

Taking the trace in the above corollary, we have *Blachman-Stam's Inequality*.

Theorem 3.6.3 (Blachman-Stam's Inequality). *For all X, Y independent random vectors and $\lambda \in [0, 1]$, we have*

$$J(\sqrt{\lambda}X + \sqrt{1-\lambda}Y) \leq \lambda J(X) + (1-\lambda)J(Y).$$

Theorem 3.6.3 can be rewritten equivalently in several forms and all of them can be called *Fisher Information Inequality*.

Theorem 3.6.4. *Let X, Y be two independent r.v with Fisher Information $J(X)$ and $J(Y)$. Then the following inequalities are true and equivalent.*

1. *Let $Z = X + Y$, then*

$$\frac{1}{J(Z)} \geq \frac{1}{J(X)} + \frac{1}{J(Y)};$$

2. *Let $\lambda \in [0, 1]$, then we have that*

$$\frac{1}{J(\sqrt{\lambda}X + \sqrt{1-\lambda}Y)} \geq \frac{\lambda}{J(X)} + \frac{1-\lambda}{J(Y)};$$

3. *Let X_0 and Y_0 be two independent Gaussian r.v with proportional covariance matrices, $J(X_0) = J(X)$ and $J(Y_0) = J(Y)$, then*

$$J(X + Y) \leq J(X_0 + Y_0); \text{ and}$$

4. *Let $\lambda \in [0, 1]$, then $J(\sqrt{\lambda}X + \sqrt{1-\lambda}Y) \leq \lambda J(X) + (1-\lambda)J(Y)$.*

The equality in all above happens only if X and Y are independent Gaussian with proportional covariance matrices (which may depend on λ).

Proof. We will not prove the equality condition in those equations. The reader can find the proof in [Blachman \(1965\)](#).

Let $\lambda \in [0, 1]$, $\lambda' = 1/2$ and X_0, Y_0 as in (3).

(1) \Rightarrow (2). Applying (1) to $X' = \sqrt{\lambda}X$ and $Y' = \sqrt{1-\lambda}Y$ leads the result in (2).

(2) \Rightarrow ((3) and (1)). Take λ' in (2) we have

$$\begin{aligned} \frac{1}{2(J(X+Y))} &= \frac{1}{J\left(\frac{1}{\sqrt{2}}(X+Y)\right)} \\ &\geq \frac{1}{2J(X)} + \frac{1}{2J(Y)} \\ &= \frac{1}{2J(X_0)} + \frac{1}{2J(Y_0)} \\ &= \frac{1}{2J(X_0+Y_0)}. \end{aligned}$$

(3) \Rightarrow (2). Take $X' = \sqrt{\lambda}X$ and $Y' = \sqrt{1-\lambda}Y$, then, by (3), we have

$$\begin{aligned} \frac{1}{J(X'+Y')} &\geq \frac{1}{J(\sqrt{\lambda}X_0 + \sqrt{1-\lambda}Y_0)} \\ &= \frac{\lambda}{J(X_0)} + \frac{1-\lambda}{J(Y_0)} \\ &= \frac{\lambda}{J(X)} + \frac{1-\lambda}{J(Y)}. \end{aligned}$$

So we have already proved that (1) \Leftrightarrow (2) \Leftrightarrow (3).

(2) \Rightarrow (4). Let $u : \mathbb{R}_+ \rightarrow \mathbb{R}_+$ so that $u(x) = 1/x$. Then u is decreasing function and convex. Therefore, because of (2) we have

$$J(\sqrt{\lambda}X + \sqrt{1-\lambda}Y) = u\left(\frac{1}{J(\sqrt{\lambda}X + \sqrt{1-\lambda}Y)}\right) \leq u\left(\frac{\lambda}{J(X)} + \frac{1-\lambda}{J(Y)}\right),$$

as u is decreasing. Now, by convexity,

$$\begin{aligned} u\left(\frac{\lambda}{J(X)} + \frac{1-\lambda}{J(Y)}\right) &\leq \lambda u\left(1/J(X)\right) + (1-\lambda)u\left(1/J(Y)\right) \\ &= \lambda J(X) + (1-\lambda)J(Y). \end{aligned}$$

(4) \Rightarrow (1). As (4) is true to all X, Y and $\lambda \in (0, 1)$, take $X' = \frac{X}{\sqrt{\lambda}}$ and $Y' = \frac{Y}{\sqrt{1-\lambda}}$, then

$$\begin{aligned} J(X+Y) &= J(\sqrt{\lambda}X' + \sqrt{1-\lambda}Y') \\ &\leq \lambda J(X') + (1-\lambda)J(Y') \\ &= \lambda^2 J(X) + (1-\lambda)^2 J(Y). \end{aligned}$$

Minimizing this quadratic form in λ we have that $\lambda^* = \frac{J(Y)}{J(X)+J(Y)} \in (0, 1)$, so by replacing this value in $\lambda^2 J(X) + (1-\lambda)^2 J(Y)$ we finally have

$$J(X+Y) \leq \frac{J(X)J(Y)}{J(X)+J(Y)} \Rightarrow \frac{1}{J(X+Y)} \geq \frac{1}{J(X)} + \frac{1}{J(Y)},$$

which is what we wanted to prove. \square

3.6.2 Exponential Entropy Inequality of Shannon

The last inequality we prove is *Exponential Entropy Inequality of Shannon*.

Theorem 3.6.5. *Let X and Y be independent random vectors with densities $f, g \in C^2(\mathbb{R}^n)$. If $N(X + Y)$, $N(X)$ and $N(Y)$ exist, then*

$$N(X + Y) \geq N(X) + N(Y).$$

There are three key points in proving this inequality. First, we need some equivalent inequalities; then, we need Fisher Information Inequality. Finally we need an identity concerning the Shannon Entropy and the Fisher Information. Let first state several equivalences of Theorem 3.6.5.

Theorem 3.6.6. *Let X and Y be independent random vectors and suppose $N(X)$, $N(Y)$ and $N(X + Y)$ exist. Then the following are equivalent.*

1. *We have that*

$$N(X + Y) \geq N(X) + N(Y);$$

2. *Let X_0 and Y_0 be two independent Gaussian vectors with proportional covariance matrices and $H(X_0) = H(X)$ and $H(Y_0) = H(Y)$, then*

$$H(X + Y) \geq H(X_0 + Y_0); \text{ and}$$

3. *For $\lambda \in [0, 1]$, we have that*

$$H(\sqrt{\lambda}X + \sqrt{1-\lambda}Y) \geq \lambda H(X) + (1-\lambda)H(Y).$$

Remark 3.6.2. Item 3 in this theorem is called **Shannon-Stam Inequality**.

Proof. Let $\lambda \in [0, 1]$ and X_0 and Y_0 as in (2).

(1) \Rightarrow (2). We have $N(X_0 + Y_0) = N(X_0) + N(Y_0) = N(X) + N(Y)$, by the exponential entropy of the Gaussian. Hence

$$N(X + Y) \geq N(X) + N(Y) = N(X_0 + Y_0),$$

therefore $H(X + Y) \geq H(X_0 + Y_0)$.

(2) \Rightarrow (1). Because of (2), we have

$$N(X + Y) \geq N(X_0 + Y_0) = N(X_0) + N(Y_0) = N(X) + N(Y).$$

(1) \Rightarrow (3). Set $u(x) := \frac{n}{2}(\log x + \log 2\pi e)$. Then u is an increasing function, concave and $u(N(X)) = H(X)$. Therefore, applying (1) for $X' = \sqrt{\lambda}X$ and $Y' = \sqrt{1-\lambda}Y$ we have

$$N(\sqrt{\lambda}X + \sqrt{1-\lambda}Y) \geq \lambda N(X) + (1-\lambda)N(Y).$$

This implies

$$\begin{aligned} H(\sqrt{\lambda}X + \sqrt{1-\lambda}Y) &= u\left(N(\sqrt{\lambda}X + \sqrt{1-\lambda}Y)\right) \\ &\geq u\left(\lambda N(X) + (1-\lambda)N(Y)\right). \end{aligned}$$

By the concavity of u , we finally obtain

$$\begin{aligned} u\left(\lambda N(X) + (1-\lambda)N(Y)\right) &\geq \lambda u(N(X)) + (1-\lambda)u(N(Y)) \\ &= \lambda H(X) + (1-\lambda)H(Y). \end{aligned}$$

(3) \Rightarrow (1). The dilation property of H gives that

$$H\left(\frac{X}{\sqrt{\lambda}}\right) = H(X) - \frac{n}{2} \log \lambda,$$

and likewise for Y , for $\lambda \in (0, 1)$. Hence, applying (3) for $X' = \frac{X}{\sqrt{\lambda}}$ and $Y' = \frac{Y}{\sqrt{1-\lambda}}$ gives

$$H(X + Y) \geq \lambda H(X) + (1-\lambda)H(Y) - \frac{n\lambda}{2} \log \lambda - \frac{n(1-\lambda)}{2} \log(1-\lambda),$$

that is,

$$H(X + Y) \geq \lambda(H(X) - H(Y)) + H(Y) + \frac{n}{2}H(\lambda) =: \phi(\lambda), \quad (3.5)$$

where $H(\lambda) = -\lambda \log \lambda - (1-\lambda) \log(1-\lambda)$ is the discrete entropy of Shannon. It is well-known that $H(\lambda)$ is a concave function of λ , therefore ϕ is concave, hence there is only one maximum and it happens when $\phi'(\lambda^*) = 0$, hence

$$\phi'(\lambda^*) = H(X) - H(Y) - \frac{n}{2} \left(\log \lambda^* - \log(1-\lambda^*) \right) = 0,$$

and hence we obtain

$$\frac{\lambda^*}{1-\lambda^*} = \exp \left[\frac{2}{n} \left(H(X) - H(Y) \right) \right].$$

We also know, by the definition the Exponential Entropy of Shannon, that

$$\frac{N(X)}{N(Y)} = \exp \left[\frac{2}{n} \left(H(X) - H(Y) \right) \right],$$

which leads to

$$\lambda^* = \frac{N(X)}{N(X) + N(Y)};$$

$$1 - \lambda^* = \frac{N(Y)}{N(X) + N(Y)}.$$

The value $H(\lambda^*)$ is given by

$$\begin{aligned} H(\lambda^*) &= - \left[\frac{N(X)}{N(X) + N(Y)} \log \left(\frac{N(X)}{N(X) + N(Y)} \right) + \frac{N(Y)}{N(X) + N(Y)} \log \left(\frac{N(Y)}{N(X) + N(Y)} \right) \right] \\ &= - \left[\frac{N(X)}{N(X) + N(Y)} \log N(X) + \frac{N(Y)}{N(X) + N(Y)} \log N(Y) - \log(N(X) + N(Y)) \right]. \end{aligned}$$

Therefore

$$\begin{aligned} \frac{n}{2} H(\lambda^*) &= - \frac{n}{2} \left(\frac{N(X)}{N(X) + N(Y)} \log N(X) + \frac{N(Y)}{N(X) + N(Y)} \log N(Y) \right) \\ &\quad + \frac{n}{2} \log(N(X) + N(Y)). \end{aligned}$$

We also have

$$\begin{aligned} \lambda^* H(X) + (1 - \lambda^*) H(Y) &= \frac{N(X) H(X) + N(Y) H(Y)}{N(X) + N(Y)} \\ &= \frac{n}{2} \frac{N(X) [2H(X)/n] + N(Y) [2H(Y)/n]}{N(X) + N(Y)}. \end{aligned}$$

Replacing $\frac{2}{n} H(X) = \log N(X) + \log 2\pi e$, we have

$$\begin{aligned} \lambda^* H(X) + (1 - \lambda^*) H(Y) &= \frac{n}{2} \frac{N(X) \log N(X) + \log(2\pi e) N(X)}{N(X) + N(Y)} \\ &\quad + \frac{n}{2} \frac{N(Y) \log N(Y) + \log(2\pi e) N(Y)}{N(X) + N(Y)}. \end{aligned}$$

Rearranging we obtain

$$\begin{aligned} \lambda^* H(X) + (1 - \lambda^*) H(Y) &= \frac{n}{2} \left(\frac{N(X)}{N(X) + N(Y)} \log N(X) + \frac{N(Y)}{N(X) + N(Y)} \log N(Y) \right) \\ &\quad + \frac{n}{2} \log(2\pi e). \end{aligned}$$

Hence

$$\begin{aligned} \phi(\lambda^*) &= \frac{n}{2} \left[\log(2\pi e) + \log \left(\frac{N(X)}{N(X) + N(Y)} + \frac{N(Y)}{N(X) + N(Y)} \right) \right] \\ &= \frac{n}{2} \log \left(\exp[2H(X)/n] + \exp[2H(Y)/n] \right). \end{aligned}$$

Replacing these values in Inequality 3.5 we have

$$H(X + Y) \geq \frac{n}{2} \log \left(\exp[2H(X)/n] + \exp[2H(Y)/n] \right).$$

Finally, applying the function $g(x) := \frac{1}{2\pi e} \exp(x)$, we obtain

$$N(X + Y) \geq N(X) + N(Y),$$

and the theorem is proved. \square

Let us now state an important identity, which we will only prove in Chapter 4 (see Theorem 4.2.3).

Theorem 3.6.7 (DeBruijn's Identity). *Let X be a random variable with density $f \in C^2(\mathbb{R})$ and $Z \sim \mathcal{N}(0, \text{Id})$ independent of X . Suppose $J(X + \sqrt{u}Z)$ is finite for some u , then*

$$\frac{d}{dt} H(X + \sqrt{t}Z) \Big|_{t=u} = \frac{1}{2} J(X + \sqrt{u}Z).$$

Now we can prove Shannon-Stam's Inequality based on Blachman-Stam's Inequality.

Proof. Let $\lambda \in [0, 1]$ fixed, $t \in [0, 1]$, X and Y independent random variable with finite Fisher Information $J(X)$ and $J(Y)$. Let X_0 and Y_0 be two independent standard Gaussian r.v with X_0 is independent of X , and Y_0 is independent of Y . Moreover, let

$$\begin{aligned} X_t &= \sqrt{t}X + \sqrt{1-t}X_0; \\ Y_t &= \sqrt{t}Y + \sqrt{1-t}Y_0; \text{ and} \\ V_t &= \sqrt{\lambda}X_t + \sqrt{1-\lambda}Y_t. \end{aligned}$$

Finally, let

$$\phi(t) := H(V_t) - \lambda H(X_t) - (1-\lambda)H(Y_t).$$

Because $X_1 = X$ and $Y_1 = Y$, we want to prove that $\phi(1) \geq 0$. Notice that, for $t = 0$, we have that V_0 is a standard gaussian vector, therefore $\phi(0) = 0$. Also, we have the following decomposition:

$$V_t = \sqrt{\lambda}\sqrt{t}X + \sqrt{\lambda}\sqrt{1-t}X_0 + \sqrt{1-\lambda}\sqrt{t}Y + \sqrt{1-\lambda}\sqrt{1-t}Y_0.$$

Collecting the terms with \sqrt{t} and with $\sqrt{1-t}$, we have

$$\sqrt{t}V_1 + \sqrt{1-t}V_0 = \sqrt{t}(\sqrt{\lambda}X + \sqrt{1-\lambda}Y) + \sqrt{1-t}(\sqrt{\lambda}X_0 + \sqrt{1-\lambda}Y_0),$$

with V_1 independent of V_0 . Let $r(t) = \frac{1-t}{t}$, then

$$\begin{aligned} V_t &= \sqrt{t}(V_1 + \sqrt{r(t)}V_0); \\ X_t &= \sqrt{t}(X + \sqrt{r(t)}X_0); \text{ and} \\ Y_t &= \sqrt{t}(Y + \sqrt{r(t)}Y_0). \end{aligned}$$

Because of the dilation property, we have that

$$\phi(t) = H(V_1 + \sqrt{r(t)}V_0) - \lambda H(X + \sqrt{r(t)}X_0) - (1 - \lambda)H(Y + \sqrt{r(t)}Y_0).$$

By differentiation and DeBruijn's Identity 3.6.7 we have

$$\phi'(t) = \frac{r'(t)}{2} \left(J(V_1 + \sqrt{r(t)}V_0) - \lambda J(X + \sqrt{r(t)}X_0) - (1 - \lambda)J(Y + \sqrt{r(t)}Y_0) \right).$$

Finally, notice that

$$V_1 + \sqrt{r(t)}V_0 = \sqrt{\lambda}(X + \sqrt{r(t)}X_0) + \sqrt{1 - \lambda}(Y + \sqrt{r(t)}Y_0),$$

therefore the Blanchman-Stam's Inequality 3.6.3 implies that

$$J(V_1 + \sqrt{r(t)}V_0) - \lambda J(X + \sqrt{r(t)}X_0) - (1 - \lambda)J(Y + \sqrt{r(t)}Y_0) \leq 0.$$

Replacing this in the expression of $\phi'(t)$ and using the fact that $r'(t) = -1/t^2 \leq 0$ we get $\phi'(t) \geq 0$, hence $\phi(1) \geq \phi(0) \geq 0$. \square

We won't go into PDEs!

4.1 Introduction

In this chapter, we will introduce the main ideas from Semigroup Theory and inequalities in Functional Analysis. They will lead us to the study of two examples of Concentration of Measure: the Binary Case, which we will prove in Chapter 5, and the Gaussian case, in Chapter 6.

In Section 4.2, we will define a semigroup of operators, its generators and some related quantities, such as the Energy and the Carré du Champ operator. We will also study some examples, such as the Heat Semigroup and the Ornstein-Uhlenbeck Semigroup. We will use the former to prove the DeBruijn's Identity (see Theorem 4.2.3) and the latter to prove the Poincaré's Inequality for the Gaussian measure (see Theorem 4.4.3).

In Section 4.3, we will define the Functional Entropy $\text{Ent}(X)$. This is a quantity that measures how concentrated the random variable X is, in the sense that $\text{Ent}(X) = 0$ if and only if X is constant. Moreover, we will study some of its properties, such as the tensorization rule and convexity.

In Section 4.4 we study the main object in this chapter. Poincaré's Inequality is a functional inequality, relating two quantities: the energy $\mathcal{E}(f)$, and the variance $\text{Var}(f)$. If a probability measure μ satisfies Poincaré's Inequality with constant c , then for all f smooth enough, we have

$$\text{Var}_\mu(f) \leq c\mathcal{E}(f).$$

At first sight, this does not seem impressive. However, if X has distribution μ , this inequality says that

$$\mathbb{P}(f(X) - \mathbb{E}[f(X)] \geq t) \leq 2\exp(-t/c),$$

for all f 1-Lipschitz. That is, $f(X)$ is *exponentially close* to its mean.

We will also prove some properties of such inequalities, such as the tensorization rule, the perturbation rule and the relation with the spectral gap of the generator of the semigroup.

Finally, in Section 4.5, we will introduce another functional inequality known as Log-Sobolev Inequality. This is stronger than the Poincaré's Inequality, in the sense that the former implies the latter. We say that a probability measure μ satisfies Log-Sobolev Inequality with constant c , then for all suitable f , we have

$$\text{Ent}_\mu(f^2) \leq c\mathcal{E}(f).$$

In some cases, this inequality will take the following form:

$$\text{Ent}_\mu(f^2) \leq c\mathbb{E}_\mu[\|\nabla f\|^2].$$

While Poincaré's Inequality gives an exponential concentration, Log-Sobolev Inequality provides a Gaussian concentration: let f be 1-Lipschitz and X with distribution μ , then

$$\mathbb{P}(|f(X) - \mathbb{E}[f(X)]| \geq t) \leq 2\exp(-t^2/c).$$

We will also prove its basics properties, such as tensorization and perturbation rules, and others equivalent definitions the reader may find in the literature.

4.2 Semigroups and Generators

4.2.1 Semigroups

Definition 4.2.1. A family of linear operators $(P_t)_{t \geq 0} : \mathcal{B} \rightarrow \mathcal{B}$ on a Banach Space $(\mathcal{B}, \|\cdot\|)$ is a **semigroup** if

1. $P_0 = \text{Id}$;
2. For all $f \in \mathcal{B}$, the map $t \rightarrow P_t f$ is continuous; and
3. for all $t, s \geq 0$, we have $P_{t+s} = P_t \circ P_s$.

In our work, the space \mathcal{B} will be the space $C_b(M)$ of real-valued bounded continuous functions f of some Polish Space (M, d) endowed with the uniform norm $\|f\| = \sup_{x \in M} |f(x)|$. In this case there is a partial order: $f \geq 0$ if and only if $f(x) \geq 0$ for all $x \in M$. Then we can define a *Markovian Semigroup*.

Definition 4.2.2. A semigroup $(P_t)_{t \geq 0}$ is **Markovian** if

4. For the constant function $f \equiv 1$ we have $P_t f = f$ for all $t \geq 0$; and

5. $P_t f \geq 0$, whenever $f \geq 0$ (preserves positivity).

A nice property of Markovian Semigroups is the *Cauchy-Schwarz' Inequality*.

Lemma 4.2.1. *Let $(P_t)_{t \geq 0}$ be a Markovian semigroup, then $[P_t(fg)]^2 \leq P_t(f)^2 P_t(g)^2$. In particular, $[P_t f]^2 \leq (P_t f^2)$ when $g = 1$.*

Proof. For a proof, see [Guionnet and Zegarlinksi \(2003\)](#) or [van Handel \(2014\)](#). □

We also have that the semigroup is contractive.

Lemma 4.2.2. *Let $(P_t)_{t \geq 0}$ be a Markovian semigroup, then the semigroup is contractive, that is, $\|P_t f\| \leq \|f\|$.*

Proof. Let $r = \|f\| = \sup_{x \in M} |f(x)|$, then $g := -f + r$ is positive. Hence $P_f g \geq 0$. Linearity and positivity imply

$$0 \leq r - P_t f,$$

hence $P_f f \leq r$. Taking sup in x gives the result. □

Similarly to Subsection 2.10.3, we can define the generator of a Markovian semigroup.

Definition 4.2.3. Let $(P_t)_{t \geq 0}$ be a Markovian semigroup and let

$$\mathcal{D}(\mathcal{L}) := \{f \in \mathcal{B} : \exists \lim_{t \rightarrow 0_+} \frac{P_t f - f}{t}\},$$

where the limit is taken with respect to the norm $\|\cdot\|$ on the Banach Space. Then we can define the operator $\mathcal{L} : \mathcal{D}(\mathcal{L}) \rightarrow \mathcal{B}$ as

$$\mathcal{L}f = \lim_{t \rightarrow 0_+} \frac{P_t f - f}{t},$$

and it is called the **generator** of the Markovian semigroup.

It can be proved that the generator uniquely defines the semigroup (see the book [Pazy \(2012\)](#)) and we also have the Hille-Yoshida's Theorem.

Theorem 4.2.1 (Hille-Yoshida's Theorem). *A linear operator \mathcal{L} is the generator of a Markovian semigroup if and only if*

1. $1 \in \mathcal{D}(\mathcal{L})$ and $\mathcal{L}1 = 0$;
2. $\mathcal{D}(\mathcal{L})$ is dense in \mathcal{B} ;

3. \mathcal{L} is closed and preserves positivity; and
4. For all $\lambda > 0$, $(\lambda Id - \mathcal{L})$ is invertible and

$$\sup_{\|f\| \leq 1} \|(\lambda Id - \mathcal{L})^{-1} f\| \leq \frac{1}{\lambda}.$$

Since we are considering $\mathcal{B} = C_b(M)$, for any probability measure μ in M we have

$$\mu(f) := \int_M f \, d\mu < \infty,$$

We can define invariant measures by the following property.

Definition 4.2.4. Let μ be a probability measure in M . Then μ is an **invariant measure** of $(P_t)_{t \geq 0}$ if $\mu(P_t f) = \mu(f)$, for all $t \geq 0$.

It can be proved that μ is invariant if and only if $\mu(\mathcal{L}f) = 0$ for all $f \in \mathcal{D}(\mathcal{L})$. Also, we can extend the semigroup to $L^p(\mu) := \{f : M \rightarrow \mathbb{R} : \mu(|f|^p) < \infty\}$ for all $p \in [1, \infty)$ (see [Guionnet and Zegarliński \(2003\)](#)).

Definition 4.2.5. We say that a Markovian semigroup is **ergodic** in $\mathcal{B}(M)$ if $P_t f$ converges to $\mu(f)$ in the uniform norm.

There are other definitions that could lead to ergodicity.

Definition 4.2.6. Let $(P_t)_{t \geq 0}$ be a semigroup, μ an invariant probability measure and $L^2 := L^2(\mu)$. We say that $P_t : L^2 \rightarrow L^2$ is **ergodic** if

$$\lim_{t \rightarrow \infty} \frac{1}{t} \int_0^t \langle P_s f, g \rangle \, ds = \mu(f)\mu(g).$$

We also say P_t is **weak-mixing** if

$$\lim_{t \rightarrow \infty} \frac{1}{t} \int_0^t |\langle P_s f, g \rangle - \mu(f)\mu(g)| \, ds = 0.$$

Finally, it is **strong-mixing** if

$$\lim_{t \rightarrow \infty} \langle P_t f, g \rangle = \mu(f)\mu(g),$$

for all $f, g \in L^2$.

Naturally, we have that ergodicity in $\mathcal{B}(M)$ implies strong-mixing, which implies weak-mixing and this implies ergodicity.

When a semigroup is ergodic? Von Neumann Ergodic Theorem for semigroup (see [Krengel \(2011\)](#), [Newman \(2015\)](#) and [Bakry \(1997\)](#)) states that

$$\lim_{t \rightarrow \infty} \frac{1}{t} \int_0^t P_s f \, ds = f^*$$

in L^2 where f^* is the projection of f onto the space of invariant functions $P_t g = g$ for all $t \geq 0$. It is easy to see that

$$\{f : P_t f = f, \forall t \geq 0\} \subset \{f : \mathcal{L}f = 0\},$$

therefore, if $\mathcal{L}f = 0$ only when f is constant, we recover the ergodicity of the semigroup.

For more about ergodicity, see [Walters \(2000\)](#) and [Krengel \(2011\)](#).

We can also define reversible measures.

Definition 4.2.7. Let μ be an invariant probability of $(P_t)_{t \geq 0}$. We say that μ is **reversible** if $\mu(f P_t g) = \mu(g P_t f)$, for all $f, g \in C_b(M)$ and all $t \geq 0$.

This means that P_t is a self-adjoint operator in $L^2(\mu)$, since $\mu(fg)$ is the standard inner product in $L^2(\mu)$. We can likewise define it equivalently using \mathcal{L} and it will be a self-adjoint operator as well.

Let J be the set of all reversible probabilities of $(P_t)_{t \geq 0}$.

Definition 4.2.8. Let $\mu \in J$. The **Dirichlet form** associated with \mathcal{L} is

$$\mathcal{E}(f, g) := \mu(f(-\mathcal{L})g).$$

The Dirichlet form is thereby the quadratic form associated with the self-adjoint operator \mathcal{L} . In the case $f = g$, we define the *energy*.

Definition 4.2.9. The **energy** of $f \in L^2(d\mu)$ is

$$\mathcal{E}(f) := \mu(f(-\mathcal{L})f).$$

Definition 4.2.10. The **Carré du Champ operator** is

$$\Gamma(f, g) := \frac{1}{2}[\mathcal{L}fg - f\mathcal{L}g - g\mathcal{L}f].$$

In the case $g = f$, we have $\Gamma(f, f) = \frac{1}{2}(\mathcal{L}f^2 - 2f\mathcal{L}f)$.

Note that $\Gamma(f, g)$ is always symmetric. Moreover, notice that

$$\begin{aligned} \lim_{t \rightarrow 0_+} \frac{P_t f^2 - (P_t f)^2}{2t} &= \lim_{t \rightarrow 0_+} \frac{(P_t f^2 - f^2) - (P_t f)^2 + f^2}{2t} \\ &= \lim_{t \rightarrow 0_+} \frac{(P_t f^2 - f^2)}{2t} + \lim_{t \rightarrow 0_+} \frac{(f - P_t f)(f + P_t f)}{2t}. \end{aligned}$$

The first term is equal to $\mathcal{L}(f^2/2)$ and the second is equal to $-f\mathcal{L}f$, hence the following lemma.

Lemma 4.2.3. *We have that*

$$\Gamma(f, f) = \lim_{t \rightarrow 0_+} \frac{P_t f^2 - (P_t f)^2}{2t},$$

therefore, $\Gamma(f, f) \geq 0$.

Also, we have that

$$\mu(\Gamma(f, f)) = \frac{1}{2} \mu(\mathcal{L} f^2) + \mu(f(-\mathcal{L})f).$$

Since $\mu \in J$, the first term is 0 and then we obtain the following lemma.

Lemma 4.2.4. *We have that $\mathcal{E}(f, f) = \mu(\Gamma(f, f))$. Hence $-\mathcal{L}$ is a positive semidefinite operator in $\mathcal{D}(\mathcal{L}) \cap L^2(M, \mathcal{B}(M), \mu)$.*

Because of this, the semigroup is also contractive in $L^2(\mu)$.

Lemma 4.2.5. *Let $(P_t)_{t \geq 0}$ be a Markovian Semigroup in $C_b(M)$ with reversible invariant measure $\mu \in J$ and \mathcal{L} , then $\|P_t f\|_{L^2(\mu)} \leq \|f\|_{L^2(\mu)}$.*

Proof. Let $f \in \mathcal{D}(\mathcal{L})$ and $u(t) = \|P_t f\|_{L^2(\mu)}^2$. Then

$$\frac{d}{dt} u(t) = 2 \int_M P_t f \mathcal{L} P_t f \, d\mu = -2\mathcal{E}(P_t f) \leq 0,$$

hence $u(t) \leq u(0)$. For general $f \in L^2(\mu)$, we can take $f_n \in \mathcal{D}(\mathcal{L})$ with $f_n \rightarrow f$ in $L^2(\mu)$, since $\mathcal{D}(\mathcal{L})$ is dense in $L^2(\mu)$ and use the result for f_n . \square

Lastly, given a Markovian semigroup $(P_t)_{t \geq 0}$, there is a unique Markov Chain with such transition semigroup.

Theorem 4.2.2. *Let $(P_t)_{t \geq 0}$ be a Markovian semigroup in $C_b(M)$, then there is a unique $(X_t)_{t \geq 0}$ Markov Chain with transition semigroup $(P_t)_{t \geq 0}$.*

Proof. The proof can be found in [Guionnet and Zegarlinski \(2003\)](#). \square

4.2.2 Heat Semigroup and DeBrujin's Identity

Let us briefly describe some important semigroups.

Definition 4.2.11. Let $M = \mathbb{R}^n$ and μ_t be the centered Gaussian measure in \mathbb{R}^n with covariance matrix

$$\Sigma := t \text{Id},$$

then we can define the **heat semigroup** $(P_t)_{t \geq 0} : C_b(M) \rightarrow C_b(M)$ as

$$P_t f(x) := \int_{\mathbb{R}^n} f(x - y) \, d\mu_t(dy) = \int_{\mathbb{R}^n} f(y) \frac{1}{(2\pi t)^{n/2}} \exp\left(-\frac{\|x - y\|^2}{2t}\right) dy.$$

It is easy to see that $P_t f = f * \gamma_t$, where γ_t is the density of μ_t . Therefore, $P_t f$ is the density of $X + \sqrt{t}Y$, where X, Y are independent, X has density f and $Y \sim \mathcal{N}(0, \text{Id})$.

Furthermore, let $M = S^1 = \{x \in \mathbb{R}^2 : \|x\|_2^2 = 1\}$ and μ the uniform probability measure in M . For $f \in L^2(\mu)$, let $f = \sum_{n \in \mathbb{Z}} c_n e^{inx}$ its Fourier series. Then we can also define the Heat Semigroup $P_t f$ by

$$P_t f(x) = \sum_{n \in \mathbb{Z}} e^{-n^2 t} c_n e^{inx}.$$

The generator of this semigroup satisfies

$$\mathcal{L}f = \frac{1}{2}f'',$$

and $\mathcal{D}(\mathcal{L}) = C_b^2(M)$. Therefore, the heat semigroup satisfies the *Heat Equation*:

$$\partial_t P_t f = \frac{1}{2} \Delta P_t f,$$

thereby it has neither invariant nor reversible measures. Indeed, suppose there is such invariant probability measure μ . Let $f \in C_b^2(\mathbb{R}^n) \cap L^1(dx)$, then

$$P_t f(x) \rightarrow 0,$$

for all $x \in \mathbb{R}^n$, by the Dominated Convergence Theorem. Also,

$$|P_t f(x)| \leq C,$$

for some constant C , which depends on f . Then, the Dominated Convergence Theorem applied to μ implies that

$$\mu(f) = \mu(P_t f) \rightarrow 0,$$

hence, for all $f \in L^1(dx) \cap C_b^2(\mathbb{R}^n)$, we have that $\mu(f) = 0$, thereby the contradiction, since there are functions $f \in L^1(dx) \cap C_b^2(\mathbb{R}^n)$ such that $\mu(f) \neq 0$.

However, the connection between the Heat Equation and the Convolution Formula provides a powerful identity between the Shannon Entropy and the Fisher Information, shown in the following theorem.

Theorem 4.2.3 (DeBruijn's Identity). *If X is a random vector with density $f \in C^2(\mathbb{R}^n)$, $Z \sim \mathcal{N}(0, \text{Id})$ is independent of X and $J(X + \sqrt{t}Z)$ exists for $t \in \mathbb{R}^+$, then*

$$\partial_u H(X + \sqrt{u}Z) \Big|_{u=t} = \frac{1}{2} J(X + \sqrt{t}Z).$$

Proof. Let $(P_t)_{t \geq 0}$ be the Heat semigroup, therefore $X + \sqrt{t}Z$ has density $P_t f$. Hence

$$\begin{aligned} J(X + \sqrt{t}Z) &= 4 \int_{\mathbb{R}^n} \|\nabla \sqrt{P_t f}\|^2 dx \\ &= \int_{\mathbb{R}^n} \frac{\|\nabla P_t f\|^2}{P_t f} dx \\ &= \sum_{i=1}^n \int_{\mathbb{R}^n} \frac{(\partial_i P_t f)^2}{P_t f} dx. \end{aligned}$$

Using integration by parts and Corollary 2.5.7, we can see that

$$\begin{aligned} \int_{\mathbb{R}^n} \frac{(\partial_i P_t f)^2}{P_t f} dx &= \int_{\mathbb{R}^n} (\partial_i P_t f)(\partial_i \log(P_t f)) dx \\ &= - \int_{\mathbb{R}^n} [\partial_{ii} P_t f] \log P_t f dx, \end{aligned}$$

which means

$$J(X + \sqrt{t}Z) = - \int_{\mathbb{R}^n} \log[P_t f] \Delta P_t f dx.$$

Let $u(t) := f(y)h_t(x - y)$, where h_t is the density of Z , then

$$\begin{aligned} \log u(t) &= \log f(y) + \log h_t(x - y) \\ &= \log f(y) + \frac{n}{2} \log(2\pi) + \frac{n}{2} \log t - \frac{\|x - y\|^2}{2t}, \end{aligned}$$

then

$$\partial_t u(t) = u(t) \partial_t \log u(t) = u(t) \left(\frac{n}{2t} + \frac{\|x - y\|^2}{2t^2} \right).$$

Because $e^{-\|x-y\|^2/(2t)}$ converges faster than $1/t^k$ diverges for all k when $t \rightarrow 0_+$, we have that $|\partial_t u| \leq C|f(y)|$, which is integrable, therefore we can apply the Dominated Convergence Theorem and get

$$\begin{aligned} \frac{1}{2} \int_{\mathbb{R}^n} \Delta P_t f dx &= \int_{\mathbb{R}^n} \partial_t P_t f dx \\ &= \partial_t \int_{\mathbb{R}^n} P_t f dx \\ &= \partial_t \int_{\mathbb{R}^n} f(x) dx \\ &= 0. \end{aligned}$$

Therefore we have

$$\begin{aligned} \frac{1}{2} J(X + \sqrt{t}Z) &= -\frac{1}{2} \int_{\mathbb{R}^n} [\Delta P_t f] \log P_t f dx - \frac{1}{2} \int_{\mathbb{R}^n} \Delta P_t f dx \\ &= - \int_{\mathbb{R}^n} \partial_t \left(P_t f \log P_t f \right) dx. \end{aligned}$$

Using again the Dominated Convergence Theorem, we can exchange the order of integration and differentiation and then

$$\frac{1}{2} J(X + \sqrt{t}Z) = \partial_t \left(- \int_{\mathbb{R}^n} P_t f \log P_t f dx \right) = \partial_t H(X + \sqrt{t}Z).$$

□

This relation provides a connection between Fisher Information and the Shannon Entropy. In fact, we have the following corollary.

Corollary 4.2.1. *Let X be a random vector in \mathbb{R}^n with a density $f \in C^2(\mathbb{R}^n)$. If $J(X)$ and $N(X)$ are finite, then $J(X)N(X) \geq n$.*

Remark 4.2.1. This inequality also can be called Stam's Inequality (see the book [Raginsky et al. \(2013\)](#)). Moreover, we will show this is equivalent to the Gaussian Log-Sobolev Inequality (see Section 6.3).

Proof. Take $Z \sim \mathcal{N}(0, \text{Id})$ independent of X . Shannon Exponential Entropy Inequality 3.6.5 implies that

$$N(X + \sqrt{t}Z) \geq N(X) + N(\sqrt{t}Z) = N(X) + t.$$

Both sides of this inequality are continuously differentiable and equal to $N(X)$ when $t = 0$, therefore, we must have that $N'(X + \sqrt{t}Z) \geq (N(X) + t)' = 1$ when $t = 0$. Using the chain rule, we have

$$1 \leq N'(X + \sqrt{t}Z) \Big|_{t=0} = N(X + \sqrt{t}Z) \Big|_{t=0} \frac{2}{n} H'(X + \sqrt{t}Z) \Big|_{t=0} = \frac{1}{n} N(X) J(X),$$

and this is the desired inequality. \square

We will derive some useful equivalences of this inequality in Chapter 6.

4.2.3 Ornstein-Uhlenbeck Semigroup

Definition 4.2.12. Let $M = \mathbb{R}^n$, then we define the **Ornstein-Uhlenbeck semigroup** as

$$P_t f(x) = \int_{\mathbb{R}^n} f(e^{-t}x + \sqrt{1 - e^{-2t}}y) \frac{1}{(2\pi)^{n/2}} \exp\left(-\frac{\|y\|^2}{2}\right) dy.$$

Notice that $P_t f(x) = \mathbb{E}[f(e^{-t}x + \sqrt{1 - e^{-2t}}Y)]$, where $Y \sim \mathcal{N}(0, \text{Id})$. It is not difficult to prove that indeed $(P_t)_{t \geq 0}$ is a Markovian semigroup.

First of all, let μ be the distribution of a standard Gaussian. Then

$$\mu(P_t f) = \mathbb{E}[f(e^{-t}X + \sqrt{1 - e^{-2t}}Y)],$$

where X, Y are independent standard Gaussians. Let $Z = e^{-t}X + \sqrt{1 - e^{-2t}}Y$, then Z is a centered Gaussian vector. Also, since

$$\mathbb{E}[Z_i Z_j] = \mathbb{E}[e^{-2t} X_i X_j] + \mathbb{E}[(1 - e^{-2t}) Y_i Y_j] = \delta_{ij},$$

we also have that $Z \sim \mathcal{N}(0, \text{Id})$. Then

$$\mu(P_t f) = \mathbb{E}[f(Z)] = \mu(f),$$

hence μ is invariant. Also, it is not hard to see that, for $f \in C^2(\mathbb{R}^n)$, the generator of P_t is

$$\mathcal{L}f := \Delta f - \langle x, \nabla f \rangle.$$

If $U(x) := \|x\|^2/2$, it can be rewritten in terms of U :

$$\mathcal{L}f = \Delta f - \langle \nabla U, \nabla f \rangle,$$

and the Gaussian measure μ can be rewritten as well:

$$\frac{d\mu}{dx} = \frac{1}{Z} e^{-U},$$

where Z is a normalization factor. Thereby, the semigroup $(P_t)_{t \geq 0}$ satisfies the following PDE:

$$\partial_t P_t f = \Delta P_t f - \langle \nabla U, \nabla P_t f \rangle.$$

Using the Dominated Convergence Theorem in the definition of $P_t f$, we obtain

$$P_t f \rightarrow \mu(f) \text{ a.s.},$$

and in fact we have $L^2(\mu)$ -convergence for $f \in C_b(\mathbb{R}^n)$. Hence the semigroup is strong-mixing. Using Poincaré's Inequality 4.4, we will later prove its rate of convergence.

The Dirichlet form and the Carré du Champ do not depend on U :

$$\begin{aligned} \mathcal{E}(f, g) &= \mathbb{E} \langle \nabla f(X), \nabla g(X) \rangle = \mu(\langle \nabla f, \nabla g \rangle); \\ \Gamma(f, g) &= \langle \nabla f, \nabla g \rangle, \end{aligned}$$

and hence μ is reversible and $Lf = 0$ if and only if f is constant almost surely.

Remark 4.2.2. This semigroup is a particular case of semigroups defined by the generators

$$\mathcal{L}f := \Delta f - \langle \nabla U, \nabla f \rangle,$$

for some strongly convex function $U : \mathbb{R}^n \rightarrow \mathbb{R}$. The reversible probability measure is the Boltzmann Measure associated with U and all of them are strong-mixing (see [Guionnet and Zegarlinski \(2003\)](#)).

4.2.4 Discrete and Binary Semigroups

Definition 4.2.13. Let $(X_i)_{i \in \mathbb{N}}$ be an i.i.d. sequence in \mathbb{R}^n with distribution μ . Let $(N_t)_{t \geq 0}$ be a Poisson Process independent of $(X_i)_{i \in \mathbb{N}}$. Then the Process $(X_t)_{t \geq 0}$, defined as

$$X_t = X_{N_t},$$

is a Markov Process, known as the **discrete process**.

It is easy to see that μ is an invariant measure for $(X_t)_{t \geq 0}$, that the transition semigroup is defined as

$$P_t f(x) = e^{-t} f(x) + (1 - e^{-t}) \mu(f),$$

and the Dirichlet form is

$$\mathcal{E}(f, g) = \text{cov}_\mu(f, g) = \int_{\mathbb{R}^n} (f - \mu(f))(g - \mu(g)) d\mu.$$

By the formula of P_t , we can see that it is strong-mixing and $\mathcal{L}f = 0$ if and only if f is constant μ -a.s.

Also, if $X_1 \sim \text{Rad}(1/2)$ and $X_{n+1} := -X_n$ recursively, then $(X_i)_{i \in \mathbb{N}}$ are dependent, but identical distributed and $X_t := X_{N_t}$ defines another Markov Process.

Definition 4.2.14. Let $X_1 \sim \text{Rad}(1/2)$ and $X_{n+1} := -X_n$ recursively and $(N_t)_{t \geq 0}$ an independent Poisson Process, then $(X_{N_t})_{t \geq 0}$ is called the **binary process**.

Notice that the transition semigroup is thereby

$$P_t f(x) = \mathbb{E}[f(X_t)],$$

given $X_0 = x$. Since X_t only takes two values and X_t changes at each jump between them, we have

$$P_t f(x) = \mathbb{P}(N_t \text{ is even}) f(x) + \mathbb{P}(N_t \text{ is odd}) f(-x).$$

These probabilities can be compute directly as

1. $\mathbb{P}(N_t \text{ is even}) = \frac{1}{2}(1 + e^{-t})$; and
2. $\mathbb{P}(N_t \text{ is odd}) = \frac{1}{2}(1 - e^{-t})$.

Hence

$$P_t f(x) = \frac{1 + e^{-t}}{2} f(x) + \frac{1 - e^{-t}}{2} f(-x).$$

Taking expected value with respect the Rademacher distribution μ , we obtain that $\mu(P_t f) = \mu(f)$, hence μ is invariant. We also have the generator

$$P_t f(x) - f(x) = \frac{1 - e^{-t}}{2} f(-x) + \frac{e^{-t} - 1}{2} f(x).$$

Dividing by t and letting $t \rightarrow 0$ we obtain

$$\mathcal{L}f = -\frac{1}{2}f(x) + \frac{1}{2}f(-x).$$

The Dirichlet Form is given by

$$\mathcal{E}(f, g) = \mu(f(-\mathcal{L})g) = \mu\left(f(x)\left[\frac{1}{2}g(x) - \frac{1}{2}g(-x)\right]\right).$$

Hence

$$\mathcal{E}(f, g) = \frac{1}{4}\left(f(x)[g(x) - g(-x)] + f(-x)[g(-x) - g(x)]\right) = \frac{1}{4}\left(f(x) - f(-x)\right)\left(g(x) - g(-x)\right).$$

Thereby μ is also reversible. Finally, if

$$\nabla f(x) := \frac{f(x) - f(-x)}{2},$$

then

$$\mathcal{E}(f, g) = \mathbb{E}[\nabla f \nabla g].$$

Letting $t \rightarrow \infty$ in $P_t f$, we see that $P_t f(x) \rightarrow \frac{1}{2}f(x) + \frac{1}{2}f(-x) = \mu(f)$, hence we also have the strong mixing property here.

4.3 Functional Entropy

Definition 4.3.1. Let X be a positive integrable random variable and $\psi(x) := x \log x$, for $x \geq 0$, then the **functional entropy** of X is defined as

$$\text{Ent}(X) = \mathbb{E}[\psi(X)] - \psi(\mathbb{E}[X]),$$

with the convention $0 \log 0 = 0$. In case $\mathbb{E}[X] = 1$, we have $\text{Ent}(X) = \mathbb{E}[X \log X]$.

Remark 4.3.1. We can also extend this concept to an arbitrary probability space (M, \mathcal{F}, μ) . We will denote $\text{Ent}_\mu(f)$ as the functional entropy in this space, that is:

$$\text{Ent}_\mu(f) = \mu(f \log f) - \mu(f) \log \mu(f),$$

for a positive measurable function.

Jensen's Inequality implies the following simple lemma.

Lemma 4.3.1. *We have that*

$$\text{Ent}(X) \geq 0,$$

and equality holds if and only if X is constant almost surely.

Lemma 4.3.1 is the first reason for the name *entropy*. In this section we will see that there are differences between the functional entropy and Shannon's entropy.

A simple example is the following.

Example 4.3.1. Let $X = \mathbf{1}_A$, for some A and $p := \mathbb{P}(A)$, then $\text{Ent}(X) = -p \log p$. Moreover, we have

$$\text{Ent}(\mathbf{1}_A) + \text{Ent}(\mathbf{1}_{A^c}) = H(p).$$

We also have the dilation property.

Lemma 4.3.2. *Let X be a random variable with finite entropy $\text{Ent}(X)$ and $a > 0$. Then*

$$\text{Ent}(aX) = a\text{Ent}(X).$$

Proof. It can be shown directly:

$$\text{Ent}(aX) = a\mathbb{E}[X \log X] + a\mathbb{E}[X \log a] - a\mathbb{E}[X \log \mathbb{E}X] - a\mathbb{E}[X \log a],$$

and the result follows canceling the second and last terms. \square

This already shows a difference between the entropies, namely,

$$\lim_{a \rightarrow \infty} \frac{\text{Ent}(aX)}{a} = \text{Ent}(X),$$

but

$$\lim_{a \rightarrow \infty} \frac{H(aX)}{a} = 0,$$

since $H(aX) = H(X)$ in the discrete case and $H(aX) = H(X) + \log a$ in the continuous case.

A generalization of Lemma 4.3.2 can be computed for independent random variables.

Lemma 4.3.3. *Let $X, Y \geq 0$ be independent, then*

$$\text{Ent}(XY) = \mathbb{E}[X]\text{Ent}(Y) + \mathbb{E}[Y]\text{Ent}(X).$$

In particular, if $\mathbb{E}[X] = \mathbb{E}[Y] = 1$, then

$$\text{Ent}(XY) = \text{Ent}(Y) + \text{Ent}(X).$$

We can bound from above the functional entropy by the covariance of certain random variables.

Lemma 4.3.4. *Let X be a random variable, then $\text{Ent}(e^X) \leq \text{cov}(X, e^X)$.*

Proof. The concavity of \log implies that $\log \mathbb{E}[e^X] \geq \mathbb{E}[X]$, then

$$\text{Ent}(e^X) \leq \mathbb{E}[Xe^X] - \mathbb{E}[X]\mathbb{E}[e^X] = \text{cov}(X, e^X),$$

hence the result. \square

The final definition of this section is the relative entropy with respect the Functional Entropy.

Definition 4.3.2. Let \mathbb{Q}, \mathbb{P} be two probabilities in the space measurable space (Ω, \mathcal{F}) . Then we define the **relative entropy** as

$$\mathcal{D}(\mathbb{Q}||\mathbb{P}) := \begin{cases} \text{Ent}_{\mathbb{P}}\left(\frac{d\mathbb{Q}}{d\mathbb{P}}\right), & \text{if } \mathbb{Q} \ll \mathbb{P}; \\ +\infty, & \text{otherwise.} \end{cases}$$

The relative entropy is a generalization of the Kullback-Leibler divergence, so we use the same notation. To see that, if \mathbb{Q}, \mathbb{P} have density f, g with respect to the Lebesgue measure in \mathbb{R}^n , then $\frac{d\mathbb{Q}}{d\mathbb{P}} = \frac{f}{g}$, hence

$$\mathcal{D}(\mathbb{Q}||\mathbb{P}) = \int_{\mathbb{R}^n} \frac{f}{g} \log \frac{f}{g} g dx = \int_{\mathbb{R}^n} f \log \frac{f}{g} dx,$$

Lemma 4.3.5. Let X be discrete in \mathcal{X} and $p(x) = \mathbb{P}(X = x)$ and let μ be the counting measure in \mathcal{X} , then

$$\text{Ent}_{\mu}(p) = -H(X).$$

In the continuous case, if f is the density of X , then $\text{Ent}_{dx}(f) = -H(X)$. Moreover, if U is uniform in \mathcal{X} , then

$$\text{Ent}(p(U)) = \frac{1}{|\mathcal{X}|} \left(\log |\mathcal{X}| - H(X) \right).$$

In the continuous case, if X has compact support K with density f and U is uniform in K , then

$$\text{Ent}(f(U)) = \frac{1}{\lambda(K)} \left(\log \lambda(K) - H(X) \right),$$

where $\lambda(K)$ is the Lebesgue measure of K .

Proof. To show the first part of the lemma we just notice that

$$\mu(p) = \sum_{x \in \mathcal{X}} p(x) = 1,$$

hence the second factor of $\text{Ent}_{\mu}(p)$ is zero. The continuous case also follows from

$$\lambda(f) = \int_{\mathbb{R}^n} f dx = 1.$$

The second part follows since

$$\mathbb{E}[p(U)] = \sum_{x \in \mathcal{X}} \frac{1}{|\mathcal{X}|} p(x) = \frac{1}{|\mathcal{X}|},$$

then

$$\text{Ent}(p(U)) = \frac{1}{|\mathcal{X}|} \sum_{x \in \mathcal{X}} p(x) \log p(X) - \frac{1}{|\mathcal{X}|} \log \frac{1}{|\mathcal{X}|},$$

hence

$$\text{Ent}(p(U)) = \frac{1}{|\mathcal{X}|} \left(\log |\mathcal{X}| - H(X) \right).$$

The continuous case follows similarly. \square

The second part of the theorem provides another proof for $H(X) \leq \log |\mathcal{X}|$ in the discrete case and equality if and only if X is uniform, since in this case $p(U)$ is constant. The continuous case also provides that $H(X) \leq \log \lambda(K)$ if equality if and only if X is uniform in K .

4.3.1 Convexity and duality formulas

In this section we explore more properties of the Functional Entropy. The first we will prove is the *Variational Formula*.

Theorem 4.3.1. *Let $X \geq 0$, then*

$$\text{Ent}(X) = \sup\{\mathbb{E}[XZ] : \mathbb{E}[e^Z] = 1\}.$$

Proof. Take Z such that $\mathbb{E}[e^Z] = 1$, then we can define a probability measure such that $d\mathbb{Q} = e^Z d\mathbb{P}$. Therefore

$$0 \leq \text{Ent}_{\mathbb{Q}}(e^{-Z}X) = \mathbb{E}[X \log(e^{-Z}X)] - \mathbb{E}[X] \log \mathbb{E}[X] = \text{Ent}(X) - \mathbb{E}[XZ],$$

hence $\text{Ent}(X) \geq \mathbb{E}[XZ]$. Equality occurs when $Z = \log(X/\mathbb{E}[X])$. \square

Remark 4.3.2. In fact, we have that

$$\text{Ent}(X) = \sup_{f: \mathbb{E}[e^{f(X)}] = 1} \mathbb{E}[Xf(X)].$$

Indeed, if $Z = f(X)$, we see that

$$\sup\{\mathbb{E}[Xf(X)] : \mathbb{E}[e^{f(X)}] = 1\} \leq \sup\{\mathbb{E}[XZ] : \mathbb{E}[e^Z] = 1\},$$

and equality holds for $f = \frac{\log x}{\log(\mathbb{E}[X])}$.

Let \mathbf{F} be the space of all positive random variables. Then the first impressive corollary is that $\text{Ent}(Z)$ is the supremum of affine functions of Z , hence it is convex in \mathbf{F} .

Corollary 4.3.1. *The function $\text{Ent} : \mathbf{F} \rightarrow \mathbb{R}_+$ is convex. Since $\text{Ent}(aX) = a\text{Ent}(X)$, for $a \geq 0$, it is also subadditive.*

A consequence of this is an upper bound for the Functional entropy.

Corollary 4.3.2. *Let X be supported on a compact set $K \subset [0, \infty)$, that is,*

$$X(\Omega) \subseteq K,$$

and $c = \|X\|_\infty$. Then $\text{Ent}(X) \leq cH(X)$. In particular, if $\lambda(K)$ is the Lebesgue measure of K , then $\text{Ent}(X) \leq c \log \lambda(K)$.

Proof. Let $X_n := \sum_{k=1}^n c_{k,n} \mathbf{1}_{E_{k,n}}$ be a sequence of simple functions which increase to X . Then, by the Dominated Convergence Theorem, we have

1. $\text{Ent}(X_n) \rightarrow \text{Ent}(X)$; and
2. $H(X_n) \rightarrow H(X)$.

But the subadditive property implies that

$$\text{Ent}(X_n) \leq \sum_{k=1}^n c_{n,k} \text{Ent}(\mathbf{1}_{E_{k,n}}) \leq cH(X_n),$$

and the desired result follows by taking the limit in n . □

The subadditive property also gives an upper bound on the entropy of $X + a$.

Corollary 4.3.3. *Let $f : \mathbb{R}_+ \rightarrow \mathbb{R}_+$ be defined as*

$$f(x) := \text{Ent}(X + x),$$

then f is nonincreasing.

Proof. Let $x \leq y$, then

$$f(y) = \text{Ent}(X + y) = \text{Ent}(X + x + (y - x)) \leq \text{Ent}(X + x) + \text{Ent}(y - x).$$

Since $y - x$ is constant, $\text{Ent}(y - x) = 0$, hence the result. □

We are now able to prove the Dualities Formula for the Relative Entropy.

Theorem 4.3.2. *Let \mathbb{Q}, \mathbb{P} be two probability measures and $\mathbb{E}_{\mathbb{Q}}, \mathbb{E}$ denotes the expected value with respect these measures, then*

$$\mathcal{D}(\mathbb{Q}||\mathbb{P}) = \sup_{Z: \mathbb{E}[e^Z] < \infty} \{\mathbb{E}_{\mathbb{Q}}[Z] - \log \mathbb{E}[e^Z]\}$$

Remark 4.3.3. In Optimization Theory, the **Fenchel conjugate function** of $f : V \rightarrow \mathbb{R}$ is

$$f^*(y) := \sup_x \{\langle x, y \rangle - f(x)\},$$

where $y \in V^*$, the dual space of V , and $\langle y, x \rangle := y(x)$. Thereby, noticing that if

$$\langle \mathbb{Q}, Z \rangle := \mathbb{E}_{\mathbb{Q}} Z,$$

then this theorem says that

$$\mathcal{D}(\cdot || \mathbb{P}) = (Z \rightarrow \log \mathbb{E}[e^Z])^*,$$

the Fenchel conjugate function of the log-Generating Function.

Proof. Suppose $\mathbb{Q} \ll \mathbb{P}$ and set $X = \frac{d\mathbb{Q}}{d\mathbb{P}}$, then

$$\mathcal{D}(\mathbb{Q} || \mathbb{P}) = \text{Ent}(X) = \sup_{Z: \mathbb{E}[e^Z]=1} \mathbb{E}[XZ].$$

Now, we can change variables to $Y = Z + a$, then $a = \log(\mathbb{E}[e^Y])$ and

$$\mathbb{E}[XZ] = \mathbb{E}[X(Y - a)] = \mathbb{E}_{\mathbb{Q}} Y - \log \mathbb{E}[e^Y],$$

hence

$$\mathcal{D}(\mathbb{Q} || \mathbb{P}) = \sup_{Y: \mathbb{E}[e^Y] < \infty} \{\mathbb{E}_{\mathbb{Q}}[Y] - \log \mathbb{E}[e^Y]\}.$$

If \mathbb{Q} is not absolutely continuous with respect \mathbb{P} , there is a set A such that $\mathbb{P}(A) = 0$, but $\mathbb{Q}(A) = a > 0$. Let $Y = n\mathbf{1}_A$, then

$$\log \mathbb{E}[e^Y] = 1,$$

since $Y = 0$ \mathbb{P} -a.s, but $\mathbb{E}_{\mathbb{Q}}[Y] = na$, hence the supremum is also infinity and the theorem is proved. \square

If a function $f : \mathbb{R}^n \rightarrow \mathbb{R}$ is convex and closed, that is, if $\{(x, t) : f(x) \leq t\}$ is a closed set, then it is a well-known fact that $(f^*)^* = f$ (see [Boyd and Vandenberghe \(2004\)](#)). The next theorem is the second duality formula and states that this also holds for $\mathcal{D}(\mathbb{Q} || \mathbb{P})$.

Theorem 4.3.3. *Let Z be a random variable, then*

$$\log \mathbb{E}[e^Z] = \sup_{\mathbb{Q}: \mathbb{Q} \ll \mathbb{P}} \{\mathbb{E}_{\mathbb{Q}}[Z] - \mathcal{D}(\mathbb{Q} || \mathbb{P})\}.$$

Proof. Suppose $\mathbb{E}[e^Z] < \infty$. Let $X = \frac{d\mathbb{Q}}{d\mathbb{P}}$ and $U = Z - \log \mathbb{E}[e^Z]$, thereby

$$\mathbb{E}[e^U] = \mathbb{E}[e^Z] / \mathbb{E}[e^Z] = 1,$$

then

$$\mathcal{D}(\mathbb{Q}||\mathbb{P}) = \text{Ent}(X) \geq \mathbb{E}[XU] = \mathbb{E}[XZ] - \log \mathbb{E}[e^Z],$$

hence

$$\log \mathbb{E}[e^Z] \geq \mathbb{E}[XZ] - \mathcal{D}(\mathbb{Q}||\mathbb{P}) = \mathbb{E}_{\mathbb{Q}}[Z] - \mathcal{D}(\mathbb{Q}||\mathbb{P}),$$

thus $\log \mathbb{E}[e^Z] \geq \sup_{\mathbb{Q} \ll \mathbb{P}} \{\mathbb{E}_{\mathbb{Q}}[Z] - \mathcal{D}(\mathbb{Q}||\mathbb{P})\}$. Setting

$$\frac{d\mathbb{Q}}{d\mathbb{P}} := \frac{e^Z}{\mathbb{E}[e^Z]},$$

we obtain

$$\mathbb{E}_{\mathbb{Q}}[Z] - \mathcal{D}(\mathbb{Q}||\mathbb{P}) = \frac{\mathbb{E}[Ze^Z]}{\mathbb{E}[e^Z]} - \mathbb{E}\left(\frac{e^Z}{\mathbb{E}[e^Z]} \log \frac{e^Z}{\mathbb{E}[e^Z]}\right),$$

which leads to

$$\mathbb{E}_{\mathbb{Q}}[Z] - \mathcal{D}(\mathbb{Q}||\mathbb{P}) = \log \mathbb{E}[e^Z],$$

then

$$\log \mathbb{E}[e^Z] = \sup_{\mathbb{Q} \ll \mathbb{P}} \{\mathbb{E}_{\mathbb{Q}}[Z] - \mathcal{D}(\mathbb{Q}||\mathbb{P})\}.$$

If $\mathbb{E}[e^Z] = \infty$, let $Z_n = Z \mathbf{1}_{|Z| \leq n}$ and

$$\frac{d\mathbb{Q}_n}{d\mathbb{P}} := \frac{e^{Z_n}}{\mathbb{E}[e^{Z_n}]},$$

then

$$\mathbb{E}_{\mathbb{Q}_n}[Z_n] - \mathcal{D}(\mathbb{Q}_n||\mathbb{P}) = \log \mathbb{E}[e^{Z_n}],$$

by the previous argument. Since $|\mathbb{E}_{\mathbb{Q}_n}[Z_n] - \mathbb{E}_{\mathbb{Q}_n}[Z]| \rightarrow 0$ and $\log \mathbb{E}[e^{Z_n}] \rightarrow \infty$, we obtain

$$\mathbb{E}_{\mathbb{Q}_n}[Z] - \mathcal{D}(\mathbb{Q}_n||\mathbb{P}) \rightarrow \infty,$$

hence the desired result in both cases. \square

This result is impressive, since Relative Entropy and log-Generating function do not seem to be related at first sight, but they are the Fenchel conjugate of each other.

Also, we can actually recover the log-Generating function as a corollary of this.

Corollary 4.3.4. *Let X be a r.v. and $\psi(\lambda) := \log \mathbb{E}[e^{\lambda X}]$, then*

$$\psi(\lambda) = \sup_{\mathbb{Q} : \mathbb{Q} \ll \mathbb{P}} \{ \lambda \mathbb{E}_{\mathbb{Q}}[X] - \mathcal{D}(\mathbb{Q} || \mathbb{P}) \}.$$

When proving the Duality Formula, we used

$$\frac{d\mathbb{Q}}{d\mathbb{P}} = \frac{e^Z}{\mathbb{E}[e^Z]}.$$

Since this is a common and well-explored technique, let us define this kind of measures.

Definition 4.3.3. Let μ be a probability in M , and $f : M \rightarrow \mathbb{R}$ an exponentially integrable function, that is,

$$\mu(f) = \int_M e^f d\mu < \infty.$$

Then we can define the f -**tilting** as the measure μ_f such that

$$\frac{d\mu_f}{d\mu} = \frac{e^f}{\mu(e^f)}.$$

In the case $f(x) = \lambda x$ and $X \sim \mu$, we denote $\mathbb{Q}_\lambda := \mu_f$ and

$$\frac{d\mathbb{Q}_\lambda}{d\mathbb{P}} = \frac{e^{\lambda X}}{\mathbb{E}[e^{\lambda X}]}$$

This kind of technique helps to simplify expression, as well as to verify properties of μ (see Theorem 4.5.6, for instance).

The final result in this section is another duality formula, due to [Donsker and Varadhan \(1975\)](#).

Theorem 4.3.4. *Let X be a centered r.v. and $\psi_X(\lambda) = \log \mathbb{E}[e^{\lambda X}]$. Suppose*

$$\psi_{|X|}(\lambda) < \infty,$$

for all $\lambda \in \mathbb{R}$, then its Fenchel conjugate $\psi^(t) := \sup_{\lambda \in \mathbb{R}} \{ \lambda t - \psi(\lambda) \}$ is equal to:*

$$g(t) := \inf \{ \mathcal{D}(\mathbb{Q} || \mathbb{P}) : \mathbb{Q} \ll \mathbb{P}, \mathbb{E}_{\mathbb{Q}}[X] \geq t \}.$$

Proof. First of all, it is easy to show that $\psi^*(t) \leq g(t)$, for all $t \geq 0$. It is a direct consequence of the duality formula. Also, the fact that $t > 0$ imposes a condition on the supremum in λ , say, by Jensen's Inequality

$$\lambda t - \psi(\lambda) \leq \lambda t < 0,$$

for $\lambda \leq 0$, then we can take the supremum over $\lambda \geq 0$.

Suppose X is constant, that is, $X = 0$ almost surely (we supposed X centered). Then $\psi^*(0) = 0$ and $\psi^*(t) = \infty$ for all $t > 0$. On the other hand, if X is constant almost surely with respect to \mathbb{P} and $\mathbb{Q} \ll \mathbb{P}$, then X is also constant almost surely with respect to \mathbb{Q} , then

$$\mathbb{E}_{\mathbb{Q}}X = 0.$$

If $t = 0$, then we can set $\mathbb{Q} = \mathbb{P}$ and get $g(0) = 0$, and for $t > 0$ the optimization problem in g is infeasible, hence $g(t) = \infty$. Therefore the constant X case is proved. Thus we can consider X not constant almost surely, that is, $\text{var}(X) > 0$.

Let $s = \text{ess sup } X := \inf\{c > 0 : \mathbb{P}(X > c) = 0\}$. The first case we will analyze will be $t < s$. The condition $\psi|_X(\lambda) < \infty$ implies that we can differentiate under the integral sign and all derivatives of $\psi_X(\lambda)$ are finite. Therefore we can easily deduce that

1. $\psi'(\lambda) = \frac{\mathbb{E}[Xe^{\lambda X}]}{\mathbb{E}[e^{\lambda X}]}$; and
2. $\psi''(\lambda) = \text{var}_{\mathbb{Q}_{\lambda}}(X) > 0$,

for all $\lambda \geq 0$, where \mathbb{Q} is the λ -tilting. Hence ψ' is strictly increasing function. It is well-known fact that

$$\psi'(\lambda) = \frac{\mathbb{E}[Xe^{\lambda X}]}{\mathbb{E}[e^{\lambda X}]}$$

is a smooth approximation of s , that is, $\psi'(\lambda) \nearrow s$ and it is a bijective increasing map from $(0, \infty)$ to $(0, s)$. By Intermediate Value Theorem, we can find a λ_0 such that

$$\psi'(\lambda_0) = t,$$

then $\psi^*(t) = t\lambda_0 - \psi(\lambda_0)$. Setting \mathbb{Q}_{λ_0} , the tilting measure, we ensure that

$$\mathbb{E}_{\mathbb{Q}_{\lambda_0}}[X] = t,$$

and we can easily see that $\mathcal{D}(\mathbb{Q}_{\lambda_0} || \mathbb{P}) = \psi^*(\lambda_0)$. Hence the equality. If $s = \infty$, then the theorem ends here.

Thus suppose $s < \infty$. If $t > s$, then we can note that both $\psi^*(t)$ and $g(t)$ are infinity and it is also proved the equality.

Finally the case $t = s$. Let $a = \mathbb{P}(X = s)$, and then we have two situations.

(*Situation I: $a = 0$*). We know that $X \leq s$ \mathbb{Q} -a.s for all $\mathbb{Q} \ll \mathbb{P}$ and $\mathbb{Q}(X = s) = 0$. If $\mathbb{E}_{\mathbb{Q}}[X] \geq s$, we obtain that $X = s$ \mathbb{Q} -a.s., which contradicts $\mathbb{Q}(X = s) = 0$, therefore the optimization problem in g is infeasible, hence $g(s) = \infty$.

Also, we have that $\psi^*(s) = \lim_{\lambda} \{\lambda s - \psi(\lambda)\}$. It can also be written as

$$\psi^*(s) = \lim_{\lambda \rightarrow \infty} \log \left(\mathbb{E}[e^{\lambda(s-X)}] \right).$$

Since $a = 0$, we have that $\lambda(s - X) \nearrow \infty$ almost surely and the Monotone Convergence Theorem implies that $\mathbb{E}[e^{\lambda(s-X)}] \rightarrow \infty$, hence $\psi^*(s)$ is also infinity.

(*Situation II: $a > 0$*). Let \mathbb{Q} be such that

$$\frac{d\mathbb{Q}}{d\mathbb{P}} = \frac{\mathbf{1}_{X=s}}{a},$$

there is, $\mathbb{Q}(A) = \frac{\mathbb{P}(A \cap \{X=s\})}{\mathbb{P}(X=s)}$, the conditional probability. We clearly have

$$\mathbb{E}_{\mathbb{Q}}[X] = s,$$

and we can compute the relative entropy and find $\mathcal{D}(\mathbb{Q}||\mathbb{P}) = -\log a$. Hence

$$g(s) \leq -\log a.$$

Then we just need to show that $\psi^*(s) = -\log a$. Let \mathbb{F} be the distribution of X . We can decompose \mathbb{F} in a part \mathbb{G} and the other will be a Dirac measure in s , that is,

$$\mathbb{F} = C\mathbb{G} + a\delta_s,$$

where C is a normalization constant and \mathbb{G} is supported in $(-\infty, s)$. Therefore, we have

$$\psi^*(s) = \log \left[\lim_{\lambda \rightarrow \infty} \left(\frac{e^{\lambda s}}{\int_{\mathbb{R}} e^{\lambda x} d\mathbb{G} + a e^{\lambda s}} \right) \right] = \log \left[\lim_{\lambda \rightarrow \infty} \frac{1}{a + C \int_{\mathbb{R}} e^{\lambda(x-s)} d\mathbb{G}} \right].$$

Since \mathbb{G} is supported in $(-\infty, s)$ and $e^{\lambda(x-s)} \leq 1$, for all $\lambda \geq 0$ and $x \in (-\infty, s)$, we can apply the Dominated Convergence Theorem and get

$$C \int_{\mathbb{R}} e^{\lambda(x-s)} d\mathbb{G} \rightarrow 0,$$

hence $\psi^*(s) = -\log a$. Thereby the theorem is proved. \square

To know more about the Duality Formulas, see the books [van Handel \(2014\)](#) and [Boucheron et al. \(2013\)](#). Moreover, they are useful in Large Deviation Theory. More information about Large Deviations can be found in [Durrett \(2019\)](#), [Dupuis and Ellis \(2011\)](#), [Den Hollander \(2008\)](#) and [Mörters \(2008\)](#).

4.3.2 Evolution of Entropy

In this section, we study the evolution of Entropy under the action of a semigroup.

Theorem 4.3.5. *Let $(P_t)_{t \geq 0}$ be a Markovian Semigroup in $C_b(M)$ and μ an invariant measure. Then*

$$\frac{d}{dt} \text{Ent}_\mu(P_t f) = -\mathcal{E}(\log P_t f, P_t f).$$

Proof. Let $f \in \mathcal{D}(\mathcal{L})$ and assume, without loss of generality, that $\mu(f) = 1$. Then the entropy is

$$\text{Ent}_\mu(P_t) = \int_M P_t f \log P_t f \, d\mu.$$

The differential of $P_t f \log P_t f$ is

$$\frac{d}{dt} P_t f \log P_t f = [\mathcal{L} P_t f] \log P_t f + \mathcal{L} P_t f.$$

Since

$$\lim_{t \rightarrow 0+} \frac{P_t f \log P_t f - f \log f}{t} = \frac{d}{dt} P_t f \log P_t f,$$

we can differentiate under the integral, that is,

$$\frac{d}{dt} \text{Ent}_\mu(P_t f) = \mu\left([P_t \mathcal{L} f] \log P_t f\right) + \mu(P_t \mathcal{L} f).$$

The second term is zero and the first is $-\mathcal{E}(\log P_t f, P_t f)$. □

This theorem implies that the evolution of $\text{Ent}_\mu(P_t f)$ is controlled by the Dirichlet Form. Let us apply this to the special cases we mentioned in the last section.

Example 4.3.2. The Dirichlet Form of the Ornstein-Uhlenbeck Semigroup is

$$\mathcal{E}(f, g) = \mu(\langle \nabla f, \nabla g \rangle),$$

where μ is the standard Gaussian measure. Hence

$$\mathcal{E}(\log f, f) = \mu\left(\|\nabla f\|^2 / f\right).$$

For $f \geq 0$, we obtain $\mathcal{E}(\log f, f) \geq 0$, hence $\text{Ent}_\mu(P_t f)$ is nonincreasing in time. We will see in Theorems 4.5.1 and 6.3.1 that we actually have an *exponential entropy rate of convergence*, that is,

$$\text{Ent}_\mu(P_t f) \leq e^{-2t} \text{Ent}_\mu(f).$$

Example 4.3.3. The Dirichlet Form of the Discrete Semigroup is

$$\mathcal{E}(f, g) = \text{cov}_\mu(f, g) = \mu(fg) - \mu(f)\mu(g).$$

Hence

$$\mathcal{E}(\log f, f) = \mu(f \log f) - \mu(\log f)\mu(f).$$

Since $\text{Ent}_\mu(f) \leq \text{cov}(f, \log f)$, we obtain

$$\mathcal{E}(\log f, f) \geq \text{Ent}_\mu(f) \geq 0,$$

hence $\text{Ent}_\mu(P_t f)$ is nonincreasing in time. In fact, notice that

$$\frac{d}{dt} \text{Ent}_\mu(P_t f) = -\mathcal{E}(\log P_t f, P_t f) = -\text{cov}(P_t f, \log P_t f) \leq -\text{Ent}_\mu(P_t f)$$

Let $u(t) = \text{Ent}_\mu(P_t f)$, then we have

$$u'(t) + u(t) \leq 0.$$

Multiplying by e^t , we obtain

$$(u(t)e^t)' \leq 0,$$

therefore the function $u(t)e^t$ is nonincreasing, that is,

$$\text{Ent}_\mu(f) = u(0) \geq e^t \text{Ent}_\mu(P_t f),$$

hence $\text{Ent}_\mu(P_t f) \leq e^{-t} \text{Ent}_\mu(f)$ and we obtain the *exponential entropy ergodicity*.

The binary case has Dirichlet form

$$\mathcal{E}(\log f, f) = \frac{1}{4} \left(\log f(x) - \log f(-x) \right) \left(f(x) - f(-x) \right).$$

Now, fix $a \in \mathbb{R}_+$ and the function $f(y) := (\log y - \log a)(y - a)$, for $y > 0$, then

$$f'(y) = \frac{y-a}{y} + (\log y - \log a),$$

hence f achieves its minimum value in $y = a$, that is, $f(y) \geq f(a) = 0$. Therefore, we also obtain here that $\text{Ent}_\mu(P_t f)$ is nonincreasing in time.

4.3.3 Tensorization

In this section we study the Tensorization Property of Entropy. In order to do so, we first define Conditional Entropy.

Definition 4.3.4. Let X be a positive random variable and $\mathcal{G} \subseteq \mathcal{F}$ be a sub- σ algebra of \mathcal{F} . Let $\psi(x) = x \log x$, then we define the **conditional entropy** of X to \mathcal{G} as

$$\text{Ent}(X|\mathcal{G}) = \mathbb{E}[\psi(X)|\mathcal{G}] - \psi(\mathbb{E}[X|\mathcal{G}]).$$

Again, $\text{Ent}(X|\mathcal{G}) = 0$ if and only if $\mathbb{E}[X|\mathcal{G}]$ is constant almost surely with respect to \mathcal{G} and we have the following variational formula.

Corollary 4.3.5. *The conditional entropy satisfies the variational formula below:*

$$\text{Ent}(X|\mathcal{G}) = \sup_{Z: \mathbb{E}[e^Z] = 1} \{\mathbb{E}[XZ|\mathcal{G}]\}.$$

We can now state and prove the *Tensorization Rule*.

Theorem 4.3.6. *Let X_1, \dots, X_n be independent r.v. and $f : \mathbb{R}^n \rightarrow \mathbb{R}_+$ be a positive function. Let $Z = f(X_1, \dots, X_n)$ and*

$$\text{Ent}^i(Z) := \text{Ent}(Z|(X_1, \dots, X_{i-1}, X_{i+1}, \dots, X_n))$$

be the entropy conditioned to all X_k but X_i . Then

$$\text{Ent}(Z) \leq \mathbb{E} \left(\sum_{k=1}^n \text{Ent}^k(Z) \right).$$

Proof. Let $U_1 = \log \mathbb{E}[Z|X_1] - \log \mathbb{E}[Z]$ and

$$U_k := \log \mathbb{E}[Z|(X_k, \dots, X_1)] - \log \mathbb{E}[Z|(X_{k-1}, \dots, X_1)],$$

for $2 \leq k \leq n$. Therefore, U_k is such that

$$\text{Ent}(Z) = \mathbb{E}[Z(\log Z - \log \mathbb{E}Z)] = \sum_{k=1}^n \mathbb{E}[ZU_k].$$

Now, let $k \geq 2$ be fixed (the case $k = 1$ is similar) and define, for simplicity

$$Z_k := \mathbb{E}[Z|(X_k, \dots, X_1)].$$

Take $\mathcal{G}_k := \sigma(X_1, \dots, X_{k-1}, X_{k+1}, \dots, X_n)$. Then we have that

$$e^{U_k} = \frac{Z_k}{Z_{k-1}} \in \sigma(X_1, \dots, X_k),$$

since it only depends on the first k coordinates. Now, we can take conditional and then

$$\mathbb{E}[e^{U_k}|\mathcal{G}_k] = \mathbb{E} \left(\frac{Z_k}{Z_{k-1}} \middle| (X_1, \dots, X_{k-1}) \right).$$

Because Z_{k-1} is $\sigma(X_1, \dots, X_{k-1})$ -measurable, we can take it out of the conditional expectation and hence

$$\mathbb{E}[e^{U_k}|\mathcal{G}_k] = \frac{1}{Z_{k-1}} \mathbb{E}[Z_k|(X_1, \dots, X_{k-1})].$$

The Tower Property of conditional expectation implies the Martingale Property, namely,

$$\mathbb{E}[Z_k|(X_1, \dots, X_{k-1})] = Z_{k-1},$$

then

$$\mathbb{E}[e^{U_k}|\mathcal{G}_k] = 1.$$

The projection property implies that $\mathbb{E}[e^{U_k}] = 1$ and the Variational Formula for the Conditional Entropy ensures that

$$\mathbb{E}[ZU_k|\mathcal{G}_k] \leq \text{Ent}(Z|\mathcal{G}_k) = \text{Ent}^k(Z).$$

Taking expectation leads to $\mathbb{E}[ZU_k] \leq \mathbb{E}[\text{Ent}(Z|\mathcal{G}_k)]$. The sum of $\mathbb{E}[ZU_k]$ gives the desired result. \square

This theorem implies that in order of bound the Entropy of a random function $f(X_1, \dots, X_n)$, we can bound the individual Entropy of $f(X_1, \dots, X_n)$ with all but one coordinate fixed and then sum them up. Although it does not seem to lead to interesting results, it is easier to control the fluctuations of single variable functions and in Section 4.5 we will derive some powerful tools and consequences using only tensorization and the Herbst's Method 4.5.6.

4.4 Poincaré's Inequality

In this section, we will describe the first important functional inequality that will help us to prove an exponential concentration of measure.

The *Poincaré's Inequality* reads as follow.

Definition 4.4.1. Let $(P_t)_{t \geq 0}$ be a Markovian Semigroup in $C_b(M)$ and μ an invariant probability measure. Let also $\text{Var}_\mu(f) := \mu[f - \mu(f)]^2$ and $\mathcal{E}(f)$ be the energy. If

$$\text{Var}_\mu(f) \leq c\mathcal{E}(f),$$

for all $f \in \mathcal{D}(\mathcal{L})$ and for some $c > 0$, then we say that (\mathcal{L}, μ) satisfies **Poincaré's Inequality** with constant c .

Although this does not seem too strong, in fact we have the following equivalence.

Theorem 4.4.1. Let $(P_t)_{t \geq 0}$ be a Markovian Semigroup in $C_b(M)$ and μ an invariant probability measure. Then the following are equivalent.

1. We have that

$$\text{Var}_\mu(f) \leq c\mathcal{E}(f),$$

for all $f \in \mathcal{D}(\mathcal{L})$; and

2. If $f \in L^2(\mu)$, then

$$\|P_t f - \mu(f)\|_{L^2(\mu)} \leq e^{-t/c} \|f - \mu(f)\|_{L^2(\mu)}.$$

Thus we can see that Poincaré's Inequality is equivalent with an exponential ergodic rate of L^2 -convergence.

Proof. (1) \Rightarrow (2). Note first that

$$\frac{d}{dt} \text{Var}_\mu(P_t f) = -2\mathcal{E}(P_t f)$$

by a direct computation. Then, applying (1) to $P_t f$ we obtain

$$\text{Var}_\mu(P_t f) \leq c\mathcal{E}(P_t f) = -\frac{c}{2} \frac{d}{dt} \text{Var}_\mu(P_t f).$$

Let $u(t) = \text{Var}_\mu(P_t f)$, then the inequality above means

$$u'(t) + \frac{2}{c}u(t) \leq 0.$$

Multiplying by $e^{2t/c}$, we have

$$[e^{2t/c}u(t)]' \leq 0,$$

hence $e^{2t/c}u(t) \leq u(0) = \text{Var}_\mu(f)$. Replacing $\text{Var}_\mu(P_t) = \|P_t f - \mu(f)\|_{L^2(\mu)}^2$ gives the result.

(2) \Rightarrow (1). Since

$$2\mathcal{E}(f) = 2\mathcal{E}(P_t f) \Big|_{t=0} = -\frac{d}{dt} \text{Var}_\mu(P_t f) \Big|_{t=0},$$

we obtain

$$2\mathcal{E}(f) = \lim_{t \rightarrow 0} \frac{\text{Var}_\mu(f) - \text{Var}_\mu(P_t f)}{t}.$$

By (2), we have that $\text{Var}_\mu(P_t f) \leq e^{-2t/c} \text{Var}_\mu(f)$, then

$$2\mathcal{E}(f) \geq \lim_{t \rightarrow 0} \frac{1 - e^{-2t/c}}{t} \text{Var}_\mu(f) = \frac{2}{c} \text{Var}_\mu(f),$$

and it is proved. □

In fact, we can prove more if we assume reversibility. Since our goal is not Poincaré's Inequality, let us just state the result.

Theorem 4.4.2. *If μ is an invariant and reversible measure, then (1) and (2) are also equivalent to*

3. $\mathcal{E}(P_t f) \leq e^{-2t/c} \mathcal{E}(f)$, for all $f \in \mathcal{D}(\mathcal{L})$ and $t \geq 0$.

Proof. The proof can be found in [van Handel \(2014\)](#). \square

Item 3 is perhaps the easiest way to prove Poincaré's Inequality in general cases. The first example of its use is to prove *Gaussian Poincaré's Inequality*.

Theorem 4.4.3. *Let γ be the standard Gaussian measure in \mathbb{R}^n . Then, for all $f \in C_b^1(\mathbb{R}^n)$ we have*

$$\text{Var}_\gamma(f) \leq \mathbb{E}_\mu \|\nabla f\|^2.$$

Proof. Let $(P_t)_{t \geq 0}$ be the Ornstein-Uhlenbeck Semigroup with the reversible and invariant measure γ , which is the standard Gaussian measure in \mathbb{R}^n . We have seen that

$$\mathcal{E}(f) = \mu(\|\nabla f\|^2),$$

and a direct computation in the definition of $P_t f$ gives

$$\nabla P_t(x) = e^{-t} P_t \nabla f(x),$$

in the sense that $P_t \nabla f(x)$ is the vector with coordinates $P_t \partial_i f(x)$. Hence

$$\mathcal{E}(P_t f) = e^{-2t} \mu(\|P_t \nabla f\|^2).$$

Since P_t is contractive in $L^2(\mu)$, we have that

$$\mathcal{E}(P_t f) \leq e^{-2t} \mu(\|\nabla f\|^2) = e^{-2t} \mathcal{E}(f).$$

Therefore we have Gaussian Poincaré's Inequality with constant $c = 1$. \square

This shows the power of the Ornstein-Uhlenbeck Semigroup, since it is easy to manipulate.

Example 4.4.1. Let us remember the definition of the binary semigroup P_t .

$$\begin{aligned} P_t f(x) &= \frac{1 + e^{-t}}{2} f(x) + \frac{1 - e^{-t}}{2} f(-x). \\ \mathcal{E}(f) &= \frac{1}{4} \left(f(1) - f(-1) \right)^2 \end{aligned}$$

Therefore, we have

$$\mathcal{E}(P_t f) = \frac{1}{4} \left(e^{-t} f(1) - e^{-t} f(-1) \right)^2 = e^{-2t} \mathcal{E}(f),$$

and this leads to a Poincaré's Inequality with constant $c = 1$. Even though we did not define the Semigroup in $\{-1, 1\}^n$, we could get the same result using

$$\mathcal{L}(f) := \mathbb{E} \|\nabla f(X)\|^2,$$

where

$$\nabla_i f(x_1, \dots, x_n) = \frac{1}{2}(f(x_1, \dots, x_n) - f(x_1, \dots, x_{i-1}, -x_i, x_{i+1}, \dots, x_n)),$$

and X is uniform in $H_n = \{-1, 1\}^n$. The $\text{Var}(f)$ is the variance of $f(X)$ and we have Poincaré's Inequality below.

Theorem 4.4.4. *If $X \in H_n$ is uniform and $f : H_n \rightarrow \mathbb{R}$, then*

$$\text{Var}(f(X)) \leq \mathbb{E} \|\nabla f(X)\|^2.$$

Our final Poincaré's Inequality is for Boltzmann measures. The proof will be based on Caffarelli's Contraction Theorem and Brenier Optimal Transport function. (see [Chafai and Lehec \(2018\)](#), [Fathi et al. \(2019\)](#), [Kim and Milman \(2012\)](#) and [Caffarelli \(2000\)](#)).

Theorem 4.4.5 (Caffarelli's Contraction Theorem). *Let $V : \mathbb{R}^n \rightarrow \mathbb{R}$ be a strongly convex function with $\text{Hess}(V) \succeq p \text{Id}$, for some $p > 0$ and let μ the Boltzmann measure associated with V . Let also γ_p be the Gaussian measure in \mathbb{R}^n with density proportional to $e^{-p\|x\|^2/2}$. Then there is a convex function $\phi \in C^1(\mathbb{R}^n)$ such that $T = \nabla \phi$ is 1-Lipschitz and μ is the push forward of the Gaussian measure γ_p , that is, $\mu(A) = \gamma_p(T^{-1}A)$, for all Borel sets A .*

The map T is called the **Brenier Optimal Transport function**. It is a map that minimizes

$$\inf_{T: T_*\gamma_p = \mu} \int_{\mathbb{R}^n} \|T(x) - x\|^2 d\gamma_p,$$

where $T_*\gamma_p$ denotes the push forward measure (see the books [Ané et al. \(2000\)](#) and [van Handel \(2014\)](#)).

Now we can prove Poincaré's Inequality for Boltzmann Measures.

Theorem 4.4.6. *Let $V : \mathbb{R}^n \rightarrow \mathbb{R}$ be a strongly convex function such that $\text{Hess}(V) \succeq p \text{Id}$, for some $p > 0$ and let μ be the Boltzmann measure associated with V . Then for all $f \in C^1(\mathbb{R}^n)$, we have*

$$\text{Var}_\mu(f) \leq \frac{1}{p} \mathbb{E}_\mu \|\nabla f\|^2.$$

Proof. Let T be Brenier map between γ_p and μ . Since $T_*\gamma_p = \mu$, we have

$$\text{Var}_\mu(f) = \text{Var}_{\gamma_p}(f \circ T).$$

By Gaussian Poincaré's Inequality (note that the variance is no longer equal to 1), we have

$$\text{Var}_\mu(f) = \text{Var}_{\gamma_p}(f \circ T) \leq \frac{1}{p} \int_{\mathbb{R}^n} \|\nabla f \circ T\|^2 d\gamma_p.$$

The chain rule and the fact T is 1-Lipschitz imply

$$\mathrm{Var}_\mu(f) \leq \frac{1}{p} \int_{\mathbb{R}^n} \|\nabla f\|^2 \circ T \, d\gamma_p.$$

Replacing back $T_*\gamma_p = \mu$, we obtain Poincaré's Inequality for μ . \square

4.4.1 Spectral Gap Inequality

Poincaré's Inequality is also known as *Spectral Gap Inequality*. To explain this name, let us recall some definitions in Functional Analysis. For simplicity, let

$$L^2(\mu) := L^2(M, \mathcal{B}(M), \mu),$$

where (M, d) is a Polish Space.

Definition 4.4.2. Let $\mathcal{L} : L^2(\mu) \rightarrow L^2(\mu)$ be a linear operator. Then the set $\rho(\mathcal{L}) \subset \mathbb{C}$, defined as

$$\rho(\mathcal{L}) := \{\lambda \in \mathbb{C} : (\lambda \mathrm{Id} - \mathcal{L})^{-1} \text{ is a continuous operator}\},$$

is known as the **resolvent set** of \mathcal{L} .

Definition 4.4.3. The set $\sigma(\mathcal{L}) = \mathbb{C} \setminus \rho(\mathcal{L})$ is called the **spectrum** of \mathcal{L}

Definition 4.4.4. A pair $(\lambda, f) \in \mathbb{C} \times L^2(\mu)$, where $f \neq 0$, which satisfies the equation

$$\mathcal{L}f = \lambda f,$$

is known as the **eigenvalue** and associated **eigenfunction**.

In Linear Algebra, the spectrum is the set of eigenvalues. However, for infinite dimensional operators, the spectrum can be strictly larger than the set of eigenvalues. That is one of the biggest differences between finite and infinite dimensional operators.

Suppose now we have a generator of a Markovian Semigroup \mathcal{L} with reversible invariant measure μ . Then \mathcal{L} is a **self-adjoint** operator, that is,

$$\langle f, \mathcal{L}g \rangle := \mu(f\mathcal{L}g) = \langle g, \mathcal{L}f \rangle$$

Moreover, our operator is **nonpositive**, that is, $\langle f, \mathcal{L}f \rangle \leq 0$. Therefore, we can restrict σ_p to the set of real λ such that $\lambda f = \mathcal{L}f$. Finally, $0 \in \sigma_p$, since $\mathcal{L}1 = 0$.

With this in mind, we can define the *spectral gap*.

Definition 4.4.5. The **spectral gap** of a Markovian Generator is $|\lambda_1|$, where

$$\lambda_1 := \sup\{\lambda : \lambda f = \mathcal{L}f, \text{ for some nonconstant } f\}.$$

Therefore, the spectral gap is the difference between the trivial eigenvalue 0 and the first nontrivial eigenvalue λ_1 .

Now suppose \mathcal{L} has spectral gap $m > 0$, and take $f = f_0 + f_1$, where f_0 and f_1 are the orthogonal projection of f onto the constant space and its perpendicular. Then

$$\mathcal{E}(f) = -\langle (f_0 + f_1)\mathcal{L}(f_0 + f_1) \rangle.$$

Since $\mathcal{L}f_0 = 0$ and \mathcal{L} is self-adjoint, we have

$$\mathcal{E}(f) = -\langle f_0 + f_1, \mathcal{L}f_1 \rangle = -\langle f_1, \mathcal{L}f_1 \rangle.$$

Hence we can assume, for our purposes, that $f \perp 1$, that is, $\mu(f) = 0$. In finite dimensional analysis, we have the formula (see [Lax \(2007\)](#)):

$$m = \inf_{f \perp 1, f \neq 0} \frac{\langle f, (-\mathcal{L})f \rangle}{\|f\|^2}.$$

However, in infinite dimensional spaces, we require compactness of the operator to argue this (see the book [Conway \(2010\)](#)). Let

$$B := \{f : \|f\| = 1\},$$

the unit sphere in $L^2(\mu)$. Then we have the following definition.

Definition 4.4.6. We say that an operator $L : L^2(\mu) \rightarrow L^2(\mu)$ is compact if the closure of $L(B)$ is compact.

Using this definition, we have the following theorem.

Theorem 4.4.7. Let T be a compact positive self-adjoint operator in $L^2(\mu)$ and

$$\lambda := \inf\{\sigma : \sigma \text{ is eigenvalue}\}.$$

Then

$$\lambda = \inf_{f \in B} \langle f, Tf \rangle.$$

Suppose then $(-\mathcal{L})$ is compact, hence

$$m = \inf_{f \perp 1} \frac{\langle f, (-\mathcal{L})f \rangle}{\|f\|^2},$$

and it follows that

$$\mathcal{E}(f) = \langle f, (-\mathcal{L})f \rangle \geq m\|f\|^2 = m\text{Var}(f),$$

that is, (\mathcal{L}, μ) satisfies Poincaré's Inequality with constant $c = 1/m$.

On the other hand, let C be the infimum of all $c > 0$ such that (\mathcal{L}, μ) satisfies Poincaré's Inequality with constant c , that is,

$$C = \inf_{f \perp 1} \frac{\text{Var}(f)}{\langle f, (-\mathcal{L})f \rangle},$$

then the spectral gap of $-\mathcal{L}$ is $m = 1/C$.

Therefore, we can find the Poincaré's Inequality constant just taking the inverse of the Spectral Gap constant. That is why some references call it the *spectral gap inequality*.

Remark 4.4.1. Of course, we could redefine the spectral gap in terms of $\inf_{f \in B} \langle f, (-\mathcal{L})f \rangle$ and get the same result without the assumption that $(-\mathcal{L})$ is compact.

Because of Theorem 4.4.1, we can see that the spectral gap m controls the ergodicity of P_t . This is expressed in the following last lemma.

Lemma 4.4.1. *Let $(P_t)_{t \geq 0}$ be a Markovian Semigroup, with invariant probability measure μ and generator \mathcal{L} . Let m be the spectral gap constant of \mathcal{L} , then*

$$\|P_t f - \mu(f)\|_{L^2(\mu)} \leq e^{-mt} \|f - \mu(f)\|_{L^2(\mu)}.$$

4.4.2 Tensorization

In this subsection, we will prove an useful formula which helps to prove Poincaré's Inequality in the case of product measure. We will restrict ourselves to the case where the energy is

$$\mathcal{E}(f) = \mathbb{E}_\mu \|\nabla f\|^2.$$

For the general energy proof, we can see it in [Guionnet and Zegarlinski \(2003\)](#). Let us state the theorem.

Theorem 4.4.8. *Let μ_1 and μ_2 be two measures in $(\Omega_1, \mathcal{F}_2)$ and $(\Omega_2, \mathcal{F}_2)$, respectively. Suppose μ_1 and μ_2 satisfy a Poincaré's Inequality with constant c_1 and c_2 , respectively, then the product measure $\mu_1 \times \mu_2$ satisfies a Poincaré's Inequality with constant $c = \max\{c_1, c_2\}$, that is, for all $f \in C^1(\Omega_1 \times \Omega_2)$, we have*

$$\text{Var}_{\mu_1 \times \mu_2}(f) \leq c \mathbb{E}_{\mu_1 \times \mu_2} \|\nabla f\|^2.$$

Remark 4.4.2. As a particular case, we can see that if μ satisfies Poincaré's Inequality with constant c , then $\mu \times \dots \times \mu$ also satisfies Poincaré's Inequality with the same constant, for all finite dimensional product of μ .

Proof. Let $f(x, y) \in L^2(\mu)$, where $\mu = \mu_1 \times \mu_2$. Then Fubini's Theorem states that

$$\text{Var}_\mu(f) = \int_{\Omega_2} \int_{\Omega_1} (f - \mu(f))^2 d\mu_1 d\mu_2.$$

The inside integral can be computed as

$$\int_{\Omega_1} (f - \mu(f))^2 d\mu_1 = \mu_1(f^2) - 2\mu_1(f)\mu(f) + [\mu(f)]^2.$$

Now, if we note that

$$\mu_1(f - \mu_1(f))^2 = \mu_1(f^2) - 2[\mu_1(f)]^2 + [\mu_1(f)]^2 = \mu_1(f^2) - [\mu_1(f)]^2,$$

then

$$\int_{\Omega_1} (f - \mu(f))^2 d\mu_1 = \mu_1(f - \mu_1(f))^2 + [\mu_1(f)]^2 - 2\mu_1(f)\mu(f) + [\mu(f)]^2.$$

The second term is $(\mu_1(f) - \mu(f))^2$, hence

$$\text{Var}_\mu(f) = \mu_2[\mu_1(f - \mu_1(f))^2] + \mu_2[(\mu_1(f) - \mu(f))^2].$$

For a fixed y , Poincaré's Inequality in μ_1 states

$$\mu_1(f - \mu_1(f))^2 \leq c_1 \mathbb{E}_{\mu_1} \|\nabla_x f(x, y)\|^2.$$

Also, since $\mu(f) = \mu_2(\mu_1(f))$, Poincaré's Inequality for μ_2 says that

$$\mu_2[(\mu_1(f) - \mu(f))^2] \leq c_2 \mathbb{E}_{\mu_2} \|\nabla \mu_1(f)\|^2.$$

Hence

$$\mu_1(f - \mu_1(f))^2 \leq c\mu_2\left(\mu_1[\|\nabla_x f(x, \cdot)\|^2] + \|\nabla \mu_1(f)\|^2\right).$$

Finally, we must prove that

$$\mu_2[\|\nabla \mu_1(f)\|^2] \leq \mu_2(\mu_1[\|\nabla_y f\|^2]), \quad (4.1)$$

but this is a direct consequence of Jensen's Inequality. Therefore the theorem is proved. \square

Remark 4.4.3. It is worth to mention that the proof in the general case is the same, except for Inequality 4.1, where we have to prove the convexity of the Carre Du Champ Operator.

The Tensorization Theorem is an important tool to prove Poincaré's Inequality for product measures, since it is sufficient to prove it for each of the marginal measures. Because of this, a tensorization lemma is very useful in many situations. For instance, we will also see the tensorization theorem in the Log-Sobolev Inequality (see Theorem 4.5.3).

4.4.3 Perturbation

Suppose (μ, Γ) satisfies Poincaré's Inequality, in the sense that

$$\text{Var}_\mu(f) \leq c\mu(\Gamma(f)),$$

for all $f \in \mathcal{D}(\mathcal{L})$. Is it true that all small perturbations of μ satisfy Poincaré's Inequality with the same Carré Du Champ Operator?

Definition 4.4.7. Let μ and ν be two measures in the same measurable space. We say that ν is a **bounded perturbation** of μ if $\nu \ll \mu$ and there are two constants $\varepsilon, \delta > 0$ such that

$$\varepsilon \leq \frac{d\nu}{d\mu} \leq \delta.$$

Remark 4.4.4. We could define it equivalently as all the ν in the form $\frac{1}{Z}e^{-V}d\mu$, where Z is a normalization factor and V is a bounded function.

The *Perturbation Theorem* reads as follow.

Theorem 4.4.9. Suppose (μ, Γ) satisfies a Poincaré's Inequality with constant c . Let ν be a bounded perturbation of μ with constant (ε, δ) , then (ν, Γ) satisfies Poincaré's Inequality with constant $\frac{\delta c}{\varepsilon}$.

Proof. Since $\nu(f)$ minimizes $\mathbb{E}_\nu(f - c)^2$, we have

$$\text{Var}_\nu(f) \leq \mathbb{E}_\nu(f - \mu(f))^2.$$

Because of the upper bound on the Radon-Nikodym derivative, we obtain

$$\mathbb{E}_\nu(f - \mu(f))^2 \leq \delta \mathbb{E}_\mu(f - \mu(f))^2.$$

Poincaré's Inequality for μ implies that

$$\delta \mathbb{E}_\mu(f - \mu(f))^2 \leq \delta c \mu(\Gamma(f)).$$

Now applying the lower bound in the derivative, we finally have

$$\text{Var}_\nu(f) \leq \frac{c\delta}{\varepsilon} \mu(\Gamma(f)).$$

□

This Theorem, combined with the Tensorization Property, will be an important tool for proving Poincaré's Inequality for general measures. For instance, we have the following corollary.

Corollary 4.4.1. Suppose that a function V is sandwiched between two quadratic functions:

$$x^2 + a \leq V(x) \leq x^2 + b,$$

for all x and $a \leq b$. Then the measure $d\nu = \frac{1}{Z}e^{-V}dx$ satisfies Poincaré's Inequality:

$$\text{Var}_\nu(f) \leq e^{2(b-a)} \mathbb{E}_\nu[\|\nabla f\|^2].$$

Proof. To prove it, we just notice that ν is a bounded perturbation of Gaussian measure and use the Gaussian Poincaré's Inequality. \square

We can also prove Poincaré's Inequality for a perturbation of the binary case.

Corollary 4.4.2. *Let $\mu\{1\} = \mu\{-1\} = 1/2$ be the Rademacher measure. Let $p \in (0, 1)$, then $\nu\{1\} = p = 1 - \nu\{-1\}$ is a bounded perturbation of μ :*

$$\frac{d\nu}{d\mu}(1) = 2p = 2 - \frac{d\nu}{d\mu}(-1),$$

that is,

$$2 \min\{p, 1 - p\} \leq \frac{d\nu}{d\mu} \leq 2 \max\{p, 1 - p\}.$$

Hence, Poincaré's Inequality for the assymmetric Rademacher is

$$\text{Var}_\nu(f) \leq \frac{\max\{p, 1 - p\}}{\min\{p, 1 - p\}} \mathbb{E}_\nu \|\nabla f\|^2,$$

where ∇f is the discrete gradient.

Remark 4.4.5. The constant $\frac{\max\{p, 1-p\}}{\min\{p, 1-p\}}$ is not optimal. Indeed, we can use Theorems 5.3.2 and 4.5.5 to get a shaper constant.

4.4.4 Concentration

In this last subsection, we will present a method to derive concentration of measure if Poincaré's Inequality is satisfied. Again, we will restrict our study to the energy $\mathcal{E}(f) = \mathbb{E}_\mu[\|\nabla f\|^2]$. In Chapter 5 we will see a stronger concentration result for the binary case in which the energy does not have this form.

The general procedure is as follows. Let $f \in C^1(\mathbb{R}^n)$ be a 1-Lipschitz function, that is, $\|\nabla f\| \leq 1$. Suppose, without loss of generality, that $\mathbb{E}_\mu[f] = 0$. Let $\lambda > 0$ and

$$\text{Var}_\mu(e^{\lambda f/2}) = \mathbb{E}_\mu[e^{\lambda f}] - \mathbb{E}^2[e^{\lambda f/2}].$$

If (μ, \mathcal{L}) satisfies Poincaré's Inequality with constant c^2 , then

$$\text{Var}_\mu(e^{\lambda f/2}) \leq c^2 \mathbb{E}_\mu[\|\nabla e^{\lambda f/2}\|^2] \leq \frac{c^2 \lambda^2}{4} \mathbb{E}_\mu[e^{\lambda f}].$$

If we set $\psi(\lambda) = \mathbb{E}_\mu[e^{\lambda f}]$, then we obtain inequality below:

$$(1 - c^2 \lambda^2 / 4) \psi(\lambda) \leq [\psi^2(\lambda/2)].$$

Just to simplify, let $g(x) := \psi(2x/c)$. This inequality applied to $\lambda = 2x/c$ yields to

$$(1 - x^2)g(x) \leq g^2(x/2).$$

Since we are assuming $\mathbb{E}_\mu[f] = 0$, L'Hôpital Rule leads to

$$\lim_{x \rightarrow 0} \frac{g(x) - 1}{x} = \lim_{x \rightarrow 0} \mathbb{E}_\mu e^{2xf/c} = \lim_{x \rightarrow 0} \frac{2}{c} \mathbb{E}_\mu f e^{2xf/c} = 0.$$

Therefore, we can apply it to the following lemma.

Lemma 4.4.2. *Let $g : (0, 1) \rightarrow (0, \infty)$ be a function such that*

$$\lim_{x \rightarrow 0} (g(x) - 1)/x = 0,$$

and

$$(1 - x^2)g(x) \leq g^2(x/2),$$

for all $x \in (0, 1)$. Then we have

$$g(x) \leq \left(1 - x^2\right)^{-2}.$$

Proof. The proof of this lemma can be found in [Boucheron et al. \(2013\)](#). □

Thereby we have that

$$\psi(2x/c) = g(x) \leq \left(1 - x^2\right)^{-2}.$$

Hence, for every $\lambda \in (0, 2/c)$, we obtain

$$\psi(\lambda) \leq \left(1 - \frac{\lambda^2 c^2}{4}\right)^{-2}.$$

The value $\lambda_0 = 1/c$ yields

$$\psi(\lambda_0) \leq \left(1 - \frac{1}{4}\right)^{-2} = \frac{16}{9} \leq 2.$$

Thus, Markov's Inequality implies that

$$\mathbb{P}(f(X) \geq t) = \mathbb{P}(e^{\lambda_0 f(X)} \geq e^{\lambda_0 t}) \leq e^{-\lambda_0 t} \psi(\lambda_0),$$

therefore

$$\mathbb{P}(f(X) \geq t) \leq 2e^{-t/c}.$$

This shows the *subexponential* behavior of measures satisfying Poincaré's Inequality. We summarize it in the following theorem.

Theorem 4.4.10. *Let (M, d) be a Polish Space. Let μ be a probability in M satisfying Poincaré's Inequality*

$$\text{Var}_\mu(f) \leq c^2 \mathbb{E}_\mu[\|\nabla f\|^2],$$

for all $f \in C^1(M)$. Then, for all 1-Lipschitz functions f , its tail behaves like the exponential distribution:

$$\mathbb{P}\left(|f(X) - \mathbb{E}[f(X)]| \geq t\right) \leq 4e^{-t/c},$$

for all $t > 0$.

4.5 Log-Sobolev Inequality

The basic ingredient to prove a *subgaussian* concentration is the so called Log-Sobolev Inequality. This is a functional inequality concerning the functional entropy and the energy. It is expressed in the definition below.

Definition 4.5.1. Let \mathcal{L} be a generator of a Markovian Semigroup and μ its invariant probability measure. We say that the pair (μ, \mathcal{L}) satisfies **Log-Sobolev Inequality** for some constant $c > 0$ if

$$\text{Ent}_\mu(f^2) \leq c\mathcal{E}(f),$$

where $\text{Ent}_\mu(f^2) = \mu(f^2 \log f^2) - \mu(f^2) \log \mu(f^2)$ is the functional entropy and $\mathcal{E}(f)$ is the energy.

Remark 4.5.1. We can also say that (μ, Γ) satisfies the Log-Sobolev Inequality where Γ is the Carre du Champ operator.

In Subsection 4.5.3 we will prove several equivalent definitions of Log-Sobolev Inequality under some hypothesis such as reversibility and chain rule of the Carre du Champ.

We will postpone the proof of the main Log-Sobolev Inequality in the Binary Case and the Gaussian Case to the Chapters 5 and 6, respectively. For now, we will prove some useful properties. The first one is the *entropy ergodicity*.

Theorem 4.5.1. If (μ, \mathcal{L}) satisfies Log-Sobolev Inequality with constant c and μ is a reversible measure of P_t , then for all $f \geq 0$ such that $\mu(f) = 1$, we have

$$\text{Ent}_\mu(P_t f) \leq e^{-4t/c} \text{Ent}_\mu(f).$$

To prove this theorem, the following two lemmas are necessary.

Lemma 4.5.1. Let $g \in \mathcal{D}(\mathcal{L})$ be positive and $p_t(x, dy)$ be the family of transition probabilities of P_t , then

$$\lim_{t \rightarrow 0} \frac{1}{2t} \int_M \int_M [g(x) - g(y)][\log g(x) - \log g(y)] dp_t(x, dy) d\mu(dx) = \mu[\log g(-\mathcal{L})g].$$

Proof. Let

$$L := \int_M [g(x) - g(y)][\log g(x) - \log g(y)] dp_t(x, dy),$$

and

$$R := g(x) \log g(x) - \log g(x) P_t g(x) - g(x) P_t [\log g](x) + P_t [g \log g](x).$$

Moreover, notice that for every function h we have

$$\int_M h(y) \, dp_t(x, dy) = P_t h(x),$$

by definition of P_t . Hence

$$L = R.$$

Therefore, the left-hand side in the statement of the theorem is equal to

$$I := \lim_{t \rightarrow 0} \frac{\mu(g \log g) - \mu(\log g P_t g)}{t},$$

by the definition of the invariant and reversible measure μ . This is precisely the definition of the derivative of $\mu(\log g(-P_t)g)$, therefore we obtain

$$I = -\frac{d}{dt} \mu(\log g P_t g) = \mu[\log g(-\mathcal{L})g],$$

and the lemma is proved. \square

Furthermore, we need an elementary inequality in \mathbb{R} .

Lemma 4.5.2. *For all $x, y \geq 0$, we have $[\log x - \log y](x - y) \geq 4(x^{1/2} - y^{1/2})^2$.*

Remark 4.5.2. Note that the case $x < y$ leads to the equivalently inequality

$$[\log a - \log b](a + b) \leq 2(a - b).$$

As a corollary, we have the following inequality.

Corollary 4.5.1. *For all positive $g \in \mathcal{D}(\mathcal{L})$, we have*

$$\mathcal{E}(\log g, g) \geq 4\mathcal{E}(g^{1/2}).$$

Proof. Lemmas 4.5.2 and 4.5.1 imply that

$$\mathcal{E}(\log g, g) \geq 4 \lim_{t \rightarrow 0} \frac{1}{2t} \int \int [g^{1/2}(x) - g^{1/2}(y)]^2 dp_t(x, dy) \mu(dx).$$

Using the same argument as in Lemma 4.5.1, we can see that

$$\lim_{t \rightarrow 0} \frac{1}{2t} \int \int [g^{1/2}(x) - g^{1/2}(y)]^2 dp_t(x, dy) \mu(dx) = \mathcal{E}(g^{1/2}),$$

and the corollary is proved. \square

Now we can prove Theorem 4.5.1.

Proof. We have already seen in previous sections that

$$\frac{d}{dt} \text{Ent}_\mu(P_t f) = -\mathcal{E}(\log f, f),$$

then Corollary 4.5.1 implies

$$\frac{d}{dt} \text{Ent}_\mu(P_t f) \leq -4\mathcal{E}(f^{1/2}).$$

Using Log-Sobolev Inequality for $f^{1/2}$, we obtain that

$$\frac{d}{dt} \text{Ent}_\mu(P_t f) \leq -4c \text{Ent}_\mu(f).$$

Solving this differential inequality gives the result. \square

In fact, Lemma 4.5.1 implies the following theorem.

Theorem 4.5.2. *Let μ be an invariant reversible measure of the Markov Semigroup generated by \mathcal{L} . If it satisfies Log-Sobolev Inequality with constant c , then*

$$\text{Ent}_\mu(f) \leq \frac{c}{4} \mathcal{E}(\log f, f),$$

for all positive $f \in \mathcal{D}(\mathcal{L})$.

Remark 4.5.3. This is known as *Modified Log-Sobolev Inequality*. As we will see in Subsection 4.5.3, Modified Log-Sobolev Inequality and Log-Sobolev Inequality are equivalent under some conditions. Furthermore, Modified Log-Sobolev Inequality is equivalent to the exponential entropy ergodicity.

4.5.1 Tensorization and Perturbation

As we saw in Poincaré's Inequality, a tensorization formula can provide a way of verifying if a product measure satisfies a certain inequality. Likewise, Log-Sobolev Inequality also satisfies a tensorization formula.

Theorem 4.5.3. *Suppose (μ_i, \mathcal{L}_i) satisfy Log-Sobolev Inequality with constant c_i , for $i = 1, 2$. Let*

$$\mathcal{D}(\mathcal{L}) := \{f : \Omega_1 \times \Omega_2 \rightarrow \mathbb{R} : f(x, \cdot) \in \mathcal{D}(\mathcal{L}_2), f(\cdot, y) \in \mathcal{D}(\mathcal{L}_1), \forall (x, y) \in \Omega_1 \times \Omega_2\}.$$

For $f \in \mathcal{D}(\mathcal{L})$, define the generator \mathcal{L} as the following. Let f_1 and f_2 be the marginals of f , that is, for fixed y , $f_y(x) = f(x, y)$ and for fixed x , $f_x(y) = f(x, y)$, then

$$\mathcal{L}f(x, y) := \mathcal{L}_1 f_y(x) + \mathcal{L}_2 f_x(y).$$

Then $(\mu_1 \times \mu_2, \mathcal{L})$ satisfies Log-Sobolev Inequality with constant $c = \max\{c_1, c_2\}$.

Proof. Let $X_i \sim \mu_i$ be independent and, for $f \in \mathcal{D}(\mathcal{L})$, let $Z = f(X_1, X_2)$. Then $\mu_1 \times \mu_2$ is the distribution of (X_1, X_2) and

$$\text{Ent}_{\mu_1 \times \mu_2}(f^2) = \text{Ent}(Z^2).$$

The tensorization rule for the entropy gives

$$\text{Ent}(Z^2) \leq \mathbb{E} \left(\text{Ent}^{(1)}(Z^2) + \text{Ent}^{(2)}(Z^2) \right).$$

Now let

$$\begin{aligned} m_1(x) &:= \text{Ent}[f^2(x, Y)] = \text{Ent}[f_x^2(Y)]; \\ m_2(y) &:= \text{Ent}[f^2(X, y)] = \text{Ent}[f_y^2(X)], \end{aligned}$$

then $m_1(X) = \text{Ent}[Z^2|X]$ and $m_2(Y) = \text{Ent}[Z^2|Y]$, therefore

$$\text{Ent}(Z^2) \leq \mathbb{E}(m_1(X) + m_2(Y)) = \mu_1(m_1) + \mu_2(m_2).$$

Since $f_x \in \mathcal{D}(\mathcal{L}_2)$ and likewise for f_y , Log-Sobolev Inequality for each measure gives

$$\begin{aligned} m_1(x) &\leq c_2 \mu_2[f_x(-\mathcal{L}_2)f_x]; \\ m_2(y) &\leq c_1 \mu_1[f_y(-\mathcal{L}_1)f_y], \end{aligned}$$

then

$$\begin{aligned} \mu_1(m_1) &\leq c_2 \mu_1 \left(\mu_2[f_x(-\mathcal{L}_2)f_x] \right); \\ \mu_2(m_2) &\leq c_1 \mu_2 \left(\mu_1[f_y(-\mathcal{L}_1)f_y] \right). \end{aligned}$$

Because the Carre Du Champ Operator is a positive operator, we can apply Fubini's Theorem, hence

$$\mu_1(m_1) + \mu_2(m_2) \leq c \mu_1 \times \mu_2 \left(f_x(-\mathcal{L}_2)f_x + f_y(-\mathcal{L}_1)f_y \right).$$

Finally, just notice that the integrand is

$$f_x(-\mathcal{L}_2)f_x + f_y(-\mathcal{L}_1)f_y = f(-\mathcal{L})f.$$

□

Furthermore, Log-Sobolev Inequality is also stable under perturbations.

Theorem 4.5.4. *Suppose that (μ, \mathcal{L}) satisfies Log-Sobolev Inequality with constant c . If $\nu \ll \mu$ is a bounded perturbation of μ with constants (ε, δ) then (ν, \mathcal{L}) satisfies Log-Sobolev Inequality with constant $c\delta/\varepsilon$.*

Proof. Let $f \in \mathcal{D}(\mathcal{L})$ and $h = \frac{d\nu}{d\mu}$. Suppose we proved the following second variational formula of the entropy:

$$\text{Ent}(Z) = \inf_{t>0} \{\mathbb{E}(Z \log Z - Z \log t - Z + t)\}. \quad (4.2)$$

Then

$$\text{Ent}_\nu(f) = \inf_{t>0} \{\nu(f \log f - f \log t - f + t)\} \leq \delta \inf_{t>0} \{\mu(f \log f - f \log t - f + t)\} = \delta \text{Ent}_\mu(f).$$

Using Log-Sobolev Inequality for μ , we obtain

$$\text{Ent}_\nu(f) \leq c\delta c\mu(\Gamma(f)) \leq \frac{c\delta}{\varepsilon} \nu(\Gamma(f)),$$

and the theorem is proved.

To prove Identity 4.2, note that if $\mathbb{E}[Z] \neq 0$, then

$$\mathbb{E}(Z \log Z - Z \log t - Z + t) = \mathbb{E}[Z \log Z - Z \log \mathbb{E}[Z] + Z \log \mathbb{E}[Z] - Z \log t - Z + t].$$

The first two terms are just the entropy, then

$$\mathbb{E}(Z \log Z - Z \log t - Z + t) = \text{Ent}(Z) - \mathbb{E}[Z \log(t/\mathbb{E}[Z])] - \mathbb{E}[Z] + t.$$

Using $\log x \leq x - 1$ for all $x \in \mathbb{R}_+$, we obtain for $x = t/\mathbb{E}[Z]$

$$\mathbb{E}(Z \log Z - Z \log t - Z + t) \geq \text{Ent}(Z) - \mathbb{E}[Z] \left(\frac{t}{\mathbb{E}[Z]} - 1 \right) - \mathbb{E}[Z] + t.$$

Therefore, we have

$$\mathbb{E}(Z \log Z - Z \log t - Z + t) \geq \text{Ent}(Z).$$

Hence

$$\inf_{t>0} \mathbb{E}(Z \log Z - Z \log t - Z + t) \geq \text{Ent}(Z).$$

If $t = \mathbb{E}[Z]$ we obtain the equality.

The case $\mathbb{E}[Z] = 0$ is trivial, since $Z = 0$ almost surely, and

$$\inf_{t>0} \mathbb{E}(Z \log Z - Z \log t - Z + t) = \inf_{t>0} t = 0.$$

□

Again, combined with a Tensorization Formula, this theorem can help us to verify if certain Log-Sobolev Inequality is satisfied.

The last result, which we will not prove, is that Log-Sobolev Inequality implies Poincaré's Inequality.

Theorem 4.5.5. *Suppose (μ, Γ) satisfies Log-Sobolev Inequality*

$$\text{Ent}_\mu(f^2) \leq 2c\mu(\Gamma(f, f)),$$

for all $f \in \mathcal{D}(\mathcal{L})$. Then (μ, Γ) satisfies Poincaré's Inequality

$$\text{Var}_\mu(f) \leq c\mu(\Gamma(f, f)).$$

The proof can be found in several books in the subject (see [van Handel \(2014\)](#), [Guionnet and Zegarlinski \(2003\)](#), and [Royer \(2007\)](#)).

For more about Log-Sobolev Inequality, see [Ané et al. \(2000\)](#) and [Royer \(2007\)](#).

4.5.2 Concentration and the Herbst Method

Log-Sobolev Inequality is a powerful tool to prove several properties for measures and semigroups. For instance, we can prove that it is equivalent to hypercontractive of the Markov Semigroup $(P_t)_{t \geq 0}$ (see [Guionnet and Zegarlinski \(2003\)](#)). In this subsection, we will show another very important fact about it: the subgaussian concentration of measure.

To begin with, we will show a general argument that is frequently used to prove *subgaussian concentration*. It is known as Herbst's Method. Then we will verify that we can apply this method to measures satisfying Log-Sobolev Inequality.

The method is described in the following theorem.

Theorem 4.5.6 (Herbst's Method). *Let X be an integrable random variable satisfying $\psi(\lambda) := \log \mathbb{E}[e^{\lambda(X - \mathbb{E}[X])}] < \infty$ and*

$$\text{Ent}(e^{\lambda X}) \leq \frac{\sigma^2 \lambda^2}{2} \mathbb{E}[e^{\lambda X}],$$

for all $\lambda \geq 0$, then, for all positive λ , we have

$$\psi(\lambda) \leq \frac{\sigma^2 \lambda^2}{2},$$

and therefore

$$\mathbb{P}(X - \mathbb{E}[X] \geq t) \leq \exp(-t^2/(2\sigma^2)),$$

by Corollary 2.7.4.

Remark 4.5.4. We could write equivalently the property

$$\text{Ent}(e^{\lambda X}) \leq \frac{\sigma^2 \lambda^2}{2} \mathbb{E}[e^{\lambda X}]$$

with the λ -tilting measures. Let $X \sim \mu$ and $\frac{d\mu_\lambda}{d\mu} := \frac{e^{\lambda X}}{\mathbb{E}[e^{\lambda X}]}$, then the above inequality is equivalent to

$$\mathcal{D}(\mu_\lambda || \mu) \leq \frac{\sigma^2 \lambda^2}{2}.$$

Proof. Since X is integrable, we have that

$$\lim_{\lambda \rightarrow 0} \frac{\psi(\lambda)}{\lambda} = \lim_{\lambda \rightarrow 0} \frac{\log \mathbb{E}[e^{\lambda(X - \mathbb{E}[X])}]}{\lambda}.$$

Using L'Hôpital Rule, we obtain

$$\lim_{\lambda \rightarrow 0} \frac{\psi(\lambda)}{\lambda} = \lim_{\lambda \rightarrow 0} \frac{\mathbb{E}[(X - \mathbb{E}[X])e^{\lambda(X - \mathbb{E}[X])}]}{\mathbb{E}[e^{\lambda(X - \mathbb{E}[X])}]} = 0.$$

Then, by the Fundamental Theorem of Calculus, we obtain

$$\frac{\psi(\lambda)}{\lambda} = \int_0^\lambda \frac{d}{du} \frac{\psi(u)}{u} du.$$

Now let us bound $\frac{d}{du} \left(\frac{\psi(u)}{u} \right)$. First note that

$$\frac{\psi(u)}{u} = \frac{\log \mathbb{E}[e^{uX}]}{u} - \mathbb{E}[X],$$

hence

$$\begin{aligned} \frac{d}{du} \left(\frac{\psi(u)}{u} \right) &= \frac{1}{u^2} \left(\frac{\mathbb{E}[uX e^{uX}]}{\mathbb{E}[e^{uX}]} - \log \mathbb{E}[e^{uX}] \right) \\ &= \frac{1}{u^2 \mathbb{E}[e^{uX}]} \left(\text{Ent}(e^{uX}) \right) \\ &\leq \sigma^2/2. \end{aligned}$$

Then

$$\frac{\psi(\lambda)}{\lambda} = \int_0^\lambda \frac{d}{du} \left(\frac{\psi(u)}{u} \right) du \leq \frac{\lambda \sigma^2}{2},$$

and the theorem is proved. \square

This method shows that, in order to prove the subgaussian property, we just have to control $\mathcal{D}(\mu_\lambda || \mu)$. In the next theorem we show that up to a constant factor, both properties are equivalent.

Theorem 4.5.7. *If $\psi(\lambda) \leq \frac{\sigma^2 \lambda^2}{8}$ for all $\lambda \geq 0$, then*

$$\text{Ent}(e^{\lambda X}) \leq \frac{\sigma^2 \lambda^2}{2} \mathbb{E}[e^{\lambda X}],$$

for all $\lambda \geq 0$.

Proof. Let $Z = \frac{e^{\lambda X}}{\mathbb{E}[e^{\lambda X}]}$ and $Y = e^{\lambda(X - \mathbb{E}[X])}$. Then we have to prove that

$$\text{Ent}(Z) \leq \frac{\sigma^2 \lambda^2}{2}.$$

By Jensen's Inequality, we have that $\mathbb{E}[Y] \geq e^{\mathbb{E}[\lambda(X - \mathbb{E}[X])]} = 1$, then

$$\text{Ent}(Y) = \mathbb{E}[Y \log Y] - \mathbb{E}[Y] \log \mathbb{E}[Y] \leq \mathbb{E}[Y \log Y].$$

Since $\log y \leq y$ for all $y \geq 0$, we obtain

$$\text{Ent}(Y) \leq \mathbb{E}[Y^2] = \mathbb{E}[e^{2\lambda(X - \mathbb{E}[X])}].$$

Since $\psi(\lambda) \leq \frac{\sigma^2 \lambda^2}{8}$, we have

$$\text{Ent}(Y) \leq \frac{\sigma^2 \lambda^2}{2}.$$

To complete the proof, note that

$$\text{Ent}(Y) = \text{Ent}(e^{-\lambda \mathbb{E}[X]} e^{\lambda X}) = \frac{1}{e^{\lambda \mathbb{E}[X]}} \text{Ent}(e^{\lambda X}).$$

Multiplying by $1 = \frac{\mathbb{E}[e^{\lambda X}]}{\mathbb{E}[e^{\lambda X}]}$ and using $\mathbb{E}[e^{\lambda X}] \geq e^{\lambda \mathbb{E}[X]}$, we get that

$$\text{Ent}(Y) = \frac{\mathbb{E}[e^{\lambda X}]}{e^{\lambda \mathbb{E}[X]}} \text{Ent}(Z) \geq \text{Ent}(Z).$$

Hence

$$\text{Ent}(Z) \leq \text{Ent}(Y) \leq \frac{\sigma^2 \lambda^2}{2}.$$

□

The first application we will give of Herbst's Method is that Log-Sobolev Inequality leads to concentration.

Corollary 4.5.2. *Let (M, d) a Polish Space and μ a probability in the measurable space $(M, \mathcal{B}(M))$ satisfying Log-Sobolev Inequality below.*

$$\text{Ent}_\mu(f^2) \leq c\mu(\|\nabla f\|^2),$$

for all $f \in C^1(M)$. Then for all 1-Lipschitz function $f : M \rightarrow \mathbb{R}$ we have

$$\mathbb{P}(|f(X) - \mathbb{E}[f(X)]| \geq t) \leq 2 \exp\left(-t^2/c\right),$$

for all $t \geq 0$ and $X \sim \mu$.

Remark 4.5.5. When the energy is not the integral of $\|\nabla f\|^2$, we can still prove concentration using Herbst's Method if the Carre du Champ satisfies the chain rule, which we will define in the next Subsection 4.5.3. Furthermore, if it does not, a direct computation can still help us to prove concentration, such as the Binary Case (see Chapter 5).

Proof. Let $f \in C^1(M)$ be a 1-Lipschitz function, that is, $\|\nabla f\| \leq 1$ and take $g = e^{\lambda f/2}$, for $\lambda > 0$. Then Log-Sobolev Inequality for g implies

$$\text{Ent}_\mu(e^{\lambda f}) = \text{Ent}_\mu(g^2) \leq c\mu(\|\nabla e^{\lambda f/2}\|^2).$$

The chain rule of ∇ leads to

$$\text{Ent}_\mu(e^{\lambda f}) \leq \frac{c\lambda^2}{4}\mu(\|\nabla f\|^2 e^{\lambda f}) \leq \frac{c\lambda^2}{4}\mu(e^{\lambda f}).$$

Applying Herbst's Method leads to the desired result for $\mathbb{P}(f(X) - \mathbb{E}[f(X)] \geq t)$. To bound

$$\mathbb{P}(f(X) - \mathbb{E}[f(X)] \leq -t) \leq \exp\left(-t^2/c\right),$$

we just have to use the same argument to $-f$. If f is not continuously differentiable, we can use a standard approximation and get the same result. \square

Moreover, the tensorization yields the same result for the product space.

Corollary 4.5.3. *Let $(\Omega_i, d_i)_{i=1}^n$ be Polish Spaces and $(\mu_i)_{i=1}^n$ be probability measures in $(\Omega_i, \mathcal{B}(\Omega_i))_{i=1}^n$ such that*

$$\text{Ent}_{\mu_i}(f_i^2) \leq c_i \mu_i(\|\nabla f_i\|^2),$$

for all $f_i \in C^1(\Omega_i)$ and $i \leq n$. Let $g \in C^1(\Omega_1 \times \dots \times \Omega_n)$ in the product space such that

$$\|\nabla_{x_i} g\| \leq a_i,$$

for all $i \leq n$. If $X_i \sim \mu_i$ are independent, then $g(X_1, \dots, X_n)$ is subgaussian and

$$\mathbb{P}\left(g(X_1, \dots, X_n) - \mathbb{E}[g(X_1, \dots, X_n)] \geq t\right) \leq \exp\left(-\frac{t^2}{\sum_{i=1}^n a_i^2 c_i}\right),$$

for all $t \geq 0$.

Remark 4.5.6. Note that this result is stronger than the tensorization of Log-Sobolev Inequality, since

$$\sum_{i=1}^n a_i^2 c_i \leq \max_{i \leq n} \{c_i\} \sum_{i=1}^n a_i^2.$$

Proof. Using the tensorization of the entropy and the chain rule, we obtain for $\mu = \prod \mu_i$

$$\text{Ent}_\mu(e^{\lambda f}) \leq \frac{\lambda^2}{4} \sum_{i=1}^n c_i a_i^2 \mathbb{E}[e^{\lambda f}],$$

hence the result. \square

4.5.3 Equivalent Definitions

Most of the classical literature of Log-Sobolev Inequality uses the version we showed with $\mathcal{E}(f)$ in the right-side of the inequality (see for instance [Ané et al. \(2000\)](#), [Guionnet and Zegarlinski \(2003\)](#) or [Ledoux \(1999\)](#) and even a nonclassical approach in [Boucheron et al. \(2013\)](#)). However, some modern sources in Log-Sobolev Inequality use different definitions (see [van Handel \(2014\)](#) or [Raginsky et al. \(2013\)](#)). In this subsection we will prove the equivalence of such definitions under some conditions in the generator \mathcal{L} or directly in the Carre du Champ operator Γ .

Let us first define the four forms of Log-Sobolev Inequality. The first we have already defined and we will call it the **classical LSI**. The second is the following.

Definition 4.5.2. We say that (μ, Γ) satisfies a **exponential LSI** if

$$\text{Ent}_\mu(e^f) \leq \frac{c}{4} \mu[\Gamma(f, f)e^f],$$

for all $f \in \mathcal{D}(\mathcal{L})$.

The next one we have also already presented.

Definition 4.5.3. We say that (μ, Γ) satisfies a **modified LSI** if

$$\text{Ent}_\mu(f) \leq \frac{c}{4} \mathcal{E}(\log f, f),$$

for all positive $f \in \mathcal{D}(\mathcal{L})$.

And the last is a symmetric version of the modified LSI.

Definition 4.5.4. We say that (μ, Γ) satisfies a **symmetric modified LSI** if

$$\text{Ent}_\mu(f) \leq \frac{c}{4} \mathcal{E}(f, \log f),$$

for all positive $f \in \mathcal{D}(\mathcal{L})$

There are two conditions for equivalence of these definitions. The first one is reversibility of the measure. The second is known as the chain rule of the Carre du Champ operator.

Definition 4.5.5. We say that Γ satisfies the **chain rule** if

$$\Gamma(f, gh) = \Gamma(f, g)h + \Gamma(f, h)g,$$

for all $f, g, h \in \mathcal{D}(\mathcal{L})$.

The chain rule is important when we have $\Gamma(f, u \circ g)$ and we want to express it in term of $\Gamma(f, g)$.

Lemma 4.5.3. *Suppose Γ satisfies the chain rule in $C_b(M)$. Then for all $u \in C^\infty(M)$ we have*

$$\Gamma(f, u \circ g) = \Gamma(f, g)u' \circ g.$$

Now we can prove the equivalences.

Theorem 4.5.8. *For a pair (μ, Γ) we have the following.*

1. *If the measure μ is reversible, then the symmetric modified LSI is equivalent to the modified LSI;*
2. *If Γ satisfies the chain rule, then the classical LSI is equivalent to the Exponential LSI; and*
3. *If Γ satisfies the chain rule and μ is reversible, then they all are equivalent.*

Proof. The first item is trivial since

$$\begin{aligned} \mathcal{E}(\log f, f) &= \mu[\log f(-\mathcal{L})f] = \mu[f(-\mathcal{L})\log f] \\ &= \mathcal{E}(f, \log f). \end{aligned}$$

To prove the second item, note that the classical LSI implies

$$\text{Ent}_\mu(e^f) \leq c\mu(\Gamma(e^f/2, e^f/2)).$$

Now, if $u(x) = e^{x/2}$, then

$$\Gamma(e^{f/2}, u(f)) = \Gamma(e^{f/2}, f)\frac{1}{2}e^{f/2}.$$

By symmetry of Γ , we have

$$\Gamma(e^{f/2}, f)\frac{1}{2}e^{f/2} = \Gamma(f, f)\frac{1}{4}e^f,$$

hence

$$\text{Ent}_\mu(e^f) \leq \frac{c}{4}\mu(\Gamma(f, f)e^f),$$

which is the Exponential LSI. Likewise, if we apply the Exponential LSI to

$$g = 2 \log f,$$

we obtain

$$\text{Ent}_\mu(f^2) \leq \frac{c}{4}\mu(\Gamma(2 \log f, 2 \log f)f^2) = c\mu(\Gamma(f, f)),$$

which is the classical LSI.

Now, to prove (3), we just have to prove the equivalence between Exponential LSI and the Modified LSI. Take $g = \log f$ in the Exponential LSI, then

$$\text{Ent}_\mu(f) \leq \frac{c}{4} \mu(\Gamma(\log f, \log f)f) = \frac{c}{4} \mu(\Gamma(\log f, f)).$$

But

$$\mu(\Gamma(\log f, f)) = \frac{1}{2} \mu(\mathcal{L}f \log f - \log f \mathcal{L}f - f \mathcal{L} \log f).$$

By reversibility and $\mu(\mathcal{L}f \log f) = 0$, we obtain

$$\mu(\Gamma(\log f, f)) = \mu(\log f(-\mathcal{L})f) = \mathcal{E}(\log f, f).$$

Replacing it in the Exponential LSI gives the first implication. To prove that the Modified LSI also implies the Exponential LSI, we just have to set $g = e^f$ in the Modified LSI and use the same arguments above. \square

It is important to note that each version of the LSI is useful for different applications. For instance, in order to prove concentration, the Exponential LSI is the most indicated since it already gives the concentration in terms of $\|\Gamma(f)\|_\infty$ (see [Raginsky et al. \(2013\)](#)). However, the Modified Log-Sobolev provides some easier ways to prove the LSI because it is equivalent to the exponential entropy ergodicity (see [van Handel \(2014\)](#)). Finally, the most classical literature in hypercontractive of semigroups and properties of some semigroups are best understood in terms of the classical LSI (see [Guionnet and Zegarliński \(2003\)](#), [Gross and Stroock \(1993\)](#) and [Ané et al. \(2000\)](#)) and it can be easy to prove in examples where the chain rule fails (see [Boucheron et al. \(2013\)](#)).

Better start with 2 than many!

5.1 Introduction

In this section, we will prove the simplest version of Log-Sobolev Inequality, where our space will be the Hamming Cube $H_n = \{-1, 1\}^n$ with the uniform measure \mathbb{P}_n . Although the proof is just a matter of manipulation and calculus, we will derive some impressive consequences from this result. We will also extend it to the asymmetric case where

$$\mu_n = \prod_{i=1}^n \mu_i,$$

and $\mu_i(\{1\}) = p_i \in (0, 1)$.

The first application of this Log-Sobolev Inequality is the standard result on concentration of Rademacher random variables and Lipschitz functions in H_n . It says that

$$\mu_n\left(\{x \in H_n : |f(x) - \mu(f)| \geq t\}\right) \leq 2 \exp\left(-nt^2\right),$$

whenever $f : H_n \rightarrow \mathbb{R}$ is 1-Lipschitz according to the discrete distance in H_n (see Section 5.2 for the definitions).

In Section 5.5 we will define Rademacher Complexity and stress its importance. Basically, this is a quantity that measures the size of a set $T \subset \mathbb{R}^n$. We will apply this concept in a concrete example in Section 5.6, where we will study the *supervised classification problem*.

Finally, we will study the concentration phenomenon in graphs in Section 5.7. The main idea is, given a graph (V, E) and a probability measure in V , we want to prove concentration of 1-Lipschitz function $f : V \rightarrow \mathbb{R}$, where the distance in V is the *graph distance*.

5.2 Definitions and Properties

Let $H_n = \{-1, 1\}^n$ be the Hamming Cube, endowed with the normalized Hamming distance d , that is,

$$d(x, y) := \frac{1}{2n} \sum_{i=1}^n |x_i - y_i| = \frac{\#\{i : x_i \neq y_i\}}{n}.$$

Let \mathbb{P}_n be the uniform measure in H_n , which is equivalent to say that \mathbb{P}_n is the law of n independent Rademacher random variables X_1, \dots, X_n with parameter $1/2$. We will denote $X := (X_1, \dots, X_n)$.

Let Ent be the **Functional Entropy** for $f : H_n \rightarrow \mathbb{R}$ defined as

$$\text{Ent}(f^2) = \mathbb{E}[f^2(X) \log f^2(X)] - \mathbb{E}[f^2(X)] \log[\mathbb{E}f^2(X)].$$

Notice that, using Theorem 4.5.3, the entropy functional satisfies

$$\text{Ent}(f^2) = \text{Ent}(f^2(X)) \leq \mathbb{E}\left(\sum_{i=1}^n \text{Ent}^{(i)}[f^2(X)]\right).$$

Moreover, let the energy function \mathcal{E} be defined as above. Let

$$\nabla_i f(x) = \frac{1}{2}(f(x) - f(\bar{x}_i)),$$

where $\bar{x}_i = (x_1, \dots, x_{i-1}, -x_i, x_{i+1}, \dots, x_n)$ is the vector x with flipped i -th coordinate. Take $\nabla f(x) = \left(\nabla_i f(x)\right)_{i=1}^n$ and finally the energy is

$$\begin{aligned} \mathcal{E}(f) &:= \int \|\nabla f(x)\|^2 d\mathbb{P}_n \\ &= \mathbb{E}[\|\nabla f(X)\|^2]. \end{aligned}$$

5.3 Main Theorem

Now we can state and prove the main theorem of this chapter, the *Rademacher Log-Sobolev Inequality*.

Theorem 5.3.1. *Let $f : H_n \rightarrow \mathbb{R}$, then*

$$\text{Ent}(f^2) \leq 2\mathcal{E}(f).$$

Proof. Because of the tensorization rule, it suffices to prove that

$$\text{Ent}^{(i)}(f^2(X)) \leq \frac{1}{2} \mathbb{E}^{(i)} \left(f(X) - f(\bar{X}_i) \right)^2,$$

since

$$\mathbb{E}\left(\frac{1}{2}\sum_{i=1}^n \mathbb{E}^{(i)}\left(f(X) - f(\overline{X}_i)\right)^2\right) = 2\mathcal{E}(f).$$

Now, fixed all X_j but X_i , we have that $f(X)$ only takes two values with the same probability, say a and b . In fact, these two values are precisely $f(X)$ and $f(\overline{X}_i)$, then

$$\text{Ent}^{(i)}(f^2(X)) = \frac{a^2}{2} \log a^2 + \frac{b^2}{2} \log b^2 - \frac{a^2 + b^2}{2} \log \frac{a^2 + b^2}{2},$$

and

$$\frac{1}{2}\mathbb{E}^{(i)}\left(f(X) - f(\overline{X}_i)\right)^2 = \frac{1}{2}(a - b)^2.$$

Therefore, we just have to prove that, for all $a, b \in \mathbb{R}$, the following inequality is true

$$\frac{a^2}{2} \log a^2 + \frac{b^2}{2} \log b^2 - \frac{a^2 + b^2}{2} \log \frac{a^2 + b^2}{2} \leq \frac{(a - b)^2}{2}.$$

Because symmetry, we can assume $a \geq b$ and, for fixed b , define the function $h : [b, \infty) \rightarrow \mathbb{R}$ such that

$$h(a) := \frac{a^2}{2} \log a^2 + \frac{b^2}{2} \log b^2 - \frac{a^2 + b^2}{2} \log \frac{a^2 + b^2}{2} - \frac{(a - b)^2}{2}.$$

It is easy to see that $h(b) = 0$. Differentiating, we have

$$h'(a) = a \log a^2 + a - a \log \frac{a^2 + b^2}{2} - a - (a - b),$$

hence

$$h'(a) = a \log \frac{2a^2}{a^2 + b^2} - (a - b).$$

Then we also have $h'(b) = 0$. Differentiating again, we obtain

$$\begin{aligned} h''(a) &= \log \frac{2a^2}{a^2 + b^2} + a \left(\frac{2}{a} - \frac{2a}{a^2 + b^2} \right) - 1 \\ &= \log \frac{2a^2}{a^2 + b^2} + \left(\frac{2b^2}{a^2 + b^2} \right) - 1 \\ &= \log \frac{2a^2}{a^2 + b^2} - \frac{2a^2}{a^2 + b^2} + 1 \\ &\leq 0, \end{aligned}$$

because $\log x \leq x - 1$, for all $x \in \mathbb{R}_+$, then h' is a decreasing function, hence $h'(a) \leq 0$ for all $a \geq b$ and then

$$h(a) \leq h(b) = 0,$$

so the theorem is proved. □

We can extend this theorem to product measures

$$\mu_n := \prod_{i=1}^n \mu,$$

such that

$$\mu(\{1\}) = p = 1 - \mu(\{-1\}).$$

This is done in the following theorem.

Theorem 5.3.2. *Let the metric space (H_n, d) be equipped with the probability measure*

$$\mu_n = \prod_{i=1}^n \mu,$$

where $\mu(\{1\}) = p \in (0, 1)$. Hence, for all $f : H_n \rightarrow \mathbb{R}$, we have

$$\text{Ent}_{\mu_n}(f^2) \leq c(p)\mathcal{E}(f),$$

where

$$\mathcal{E}(f) = \mu_n[\|\nabla f\|^2],$$

and

$$c(p) = \frac{4p(1-p)}{1-2p} \log\left(\frac{1-p}{p}\right).$$

Remark 5.3.1. Note that Theorem 5.3.2 generalizes Theorem 5.3.1, since

$$\lim_{p \rightarrow 1/2} c(p) = 2.$$

Proof. A proof of this theorem using other techniques can be found in [Ané et al. \(2000\)](#) and [Bobkov et al. \(2006\)](#). □

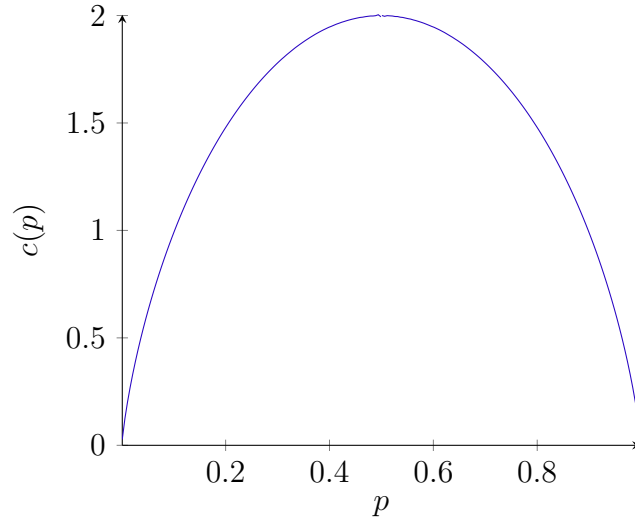
The graph of $c(p)$ is shown in Figure 6.

5.4 Application I: Concentration in the Hamming Cube

The first application of Theorem 5.3.1 is the so called *Concentration in the Hamming Cube*. The standard method (see Corollary 4.5.2) fails since the Carré du Champ Operator does not satisfy the chain rule. However, we can still modify its argument to prove concentration.

Corollary 5.4.1. *Let $f : H_n \rightarrow \mathbb{R}$ be a 1-Lipschitz function, then*

$$\mathbb{P}\left(|f(X) - \mathbb{E}[f(X)]| \geq t\right) \leq 2 \exp\left(-nt^2\right).$$

Figure 6 – The graph of the function $c(p)$.

Proof. Let $\lambda > 0$ and $g(x) = e^{\lambda f(x)/2}$. First, notice that

$$\mathbb{P}\left(f(X) - f(\bar{X}_i) \geq 0\right) = 1/2,$$

by symmetry. Therefore, we have

$$\frac{1}{2}\mathbb{E}\left([g(X) - g(\bar{X}_i)]^2\right) = \mathbb{E}\left(\left[(g(X) - g(\bar{X}_i))_+\right]^2\right),$$

where $(x)_+ := x$ if $x \geq 0$ and 0 otherwise. Indeed, we always have

$$[g(X) - g(\bar{X}_i)]^2 = \left[(g(X) - g(\bar{X}_i))_+\right]^2 + \left[(g(\bar{X}_i) - g(X))_+\right]^2.$$

Now, the convexity of the function e^x implies that

$$0 \leq (g(x) - g(\bar{x}_i))_+ \leq \frac{\lambda(f(x) - f(\bar{x}_i))_+}{2} e^{\lambda f(x)/2},$$

hence Theorem 5.3.1 applied to g yields

$$\begin{aligned} \text{Ent}(e^{\lambda f}) &\leq 2\mathcal{E}(e^{\lambda f/2}) \\ &= \mathbb{E}\left(\frac{1}{2} \sum_{i=1}^n \mathbb{E}^{(i)} \left(e^{\lambda f(X)/2} - e^{\lambda f(\bar{X}_i)/2} \right)^2\right) \\ &\leq \frac{\lambda^2}{4} \mathbb{E}\left(e^{\lambda f(X)} \sum_{i=1}^n \left[(f(X) - f(\bar{X}_i))_+\right]^2\right). \end{aligned}$$

Now, the assumption that f is 1-Lipschitz gives that

$$(f(X) - f(\bar{X}_i))_+ \leq d(x, \bar{x}_i) = \frac{1}{n},$$

hence

$$\text{Ent}(e^{\lambda f}) \leq \frac{\lambda^2}{4n} \mathbb{E}[e^{\lambda f}].$$

Finally, Herbst's Method 4.5.6 yields

$$\mathbb{P}\left(|f(X) - \mathbb{E}[f(X)]| \geq t\right) \leq \exp\left(-nt^2\right),$$

and the theorem is proved. \square

Remark 5.4.1. Notice that this result depends on the metric d we chose at the beginning. If we set $d_0 := nd$, that is,

$$d_0(x, y) = \#\{i : x_i \neq y_i\},$$

then Corollary 5.4.1 reads as follows.

Corollary 5.4.2. *Let $f : H_n \rightarrow \mathbb{R}$ be a 1-Lipschitz function according to the metric d_0 , then*

$$\mathbb{P}\left(|f(X) - \mathbb{E}[f(X)]| \geq t\right) \leq \exp\left(-t^2/n\right).$$

In fact, we can get a stronger bound using the marginal Lipschitz constants.

Corollary 5.4.3. *Let $f : H_n \rightarrow \mathbb{R}$ be a function such that, for all $x \in H_n$ and all $i \leq n$, we have*

$$|f(x) - f(\bar{x}_i)| \leq L_i d(x, \bar{x}_i) = \frac{L_i}{n},$$

then

$$\mathbb{P}\left(|f(X) - \mathbb{E}[f(X)]| \geq t\right) \leq 2 \exp\left(-\frac{n^2 t^2}{\sum_{i=1}^n L_i^2}\right).$$

For a first example, we can state a quantitative Law of Large Numbers.

Example 5.4.1. Let $X_1, \dots, X_n \sim \text{Rad}(p)$, then

$$\mathbb{P}\left(\left|\frac{1}{n} \sum_{i=1}^n X_i - (2p - 1)\right| \geq t\right) \leq 2 \exp\left(-\frac{nt^2}{2c(p)}\right),$$

where $c(p)$ is defined as in Theorem 5.3.2. Indeed, we have that the function $f : H_n \rightarrow \mathbb{R}$, defined as

$$f(x) = \frac{1}{n} \sum_{i=1}^n x_i,$$

satisfies

$$|f(x) - f(\bar{x}_i)| = 2d(x, \bar{x}_i), \quad \forall x \in H_n,$$

then the result follows by the Herbst's Method and Theorem 5.3.2.

Using this, we can show some other properties of the *Repetition Code* (see 3.5.1).

Example 5.4.2. Let $m \in H_k$ be a message we want to transmit through a Binary Channel as Example 3.5.1. Let $f : H_k \rightarrow H_{nk}$ be the encode function, such that

$$f(x) = x \dots x,$$

n times. The decode function $g : H_{nk} \rightarrow H_k$ guesses individually each entry as follows: given $y \in H_{nk}$, let $x_i \in H_1$ such that x_i is the most common number in the vector $(y_{i+lk})_{l=0}^{n-1}$, then

$$g(y) = x.$$

Moreover, let $Y \in H_{nk}$ be the random variable corresponding to the channel output of $f(m)$ and

$$\lambda^{(n)} = \mathbb{P}(g(Y) \neq m),$$

then we have the following corollary.

Corollary 5.4.4. *We have that*

$$\lambda^{(n)} \leq \exp \left(\log k - \frac{n(1-2p)^2}{2c(p)} \right).$$

Proof. By a simple union bound argument, we can bound the error λ_i in each entry m_i and then

$$\lambda^{(n)} \leq \sum_{i=1}^k \lambda_i.$$

Also, using the fact that all random variables are i.i.d, we have that λ_i is constant, then

$$\lambda^{(n)} \leq k\lambda_1.$$

Now, it is easy to see that

$$(g(Y))_i = m_i \operatorname{sign} \left(\sum_{l=1}^n X_l \right),$$

where $X_l \sim \operatorname{Rad}(1-p)$ are independent random variables. Hence, we have

$$\lambda_1 = \mathbb{P} \left(\sum_{l=1}^n X_l > 0 \right).$$

Now we can use Example 5.4.1 with $t = 1 - 2p > 0$ and get

$$\begin{aligned} \lambda_1 &= \mathbb{P} \left(\frac{1}{n} \sum_{l=1}^n X_l > 0 \right) \\ &\leq \mathbb{P} \left(\frac{1}{n} \sum_{l=1}^n X_l > t + (2p - 1) \right) \\ &\leq \exp \left(- \frac{nt^2}{2c(1-p)} \right). \end{aligned}$$

Since $c(1-p) = c(p)$, we get

$$\lambda_1 \leq \exp\left(-\frac{n(1-2p)^2}{2c(p)}\right),$$

and it is proved. \square

Therefore, to guarantee an error $\lambda^{(n)} \leq \varepsilon$, we can bound

$$\log k - \frac{n(1-2p)^2}{2c(p)} \leq \log \varepsilon,$$

that is,

$$n \geq \frac{c(p)}{(1-2p)^2} \log(k/\varepsilon) =: r(p) \log(k/\varepsilon).$$

The graph of $r(p)$ is shown in Figure 7. Notice that

$$\lim_{p \rightarrow 1/2} r(p) = \infty.$$

Thereby in order to recover the message m , we need to transmit it $n \sim O(\log k)$ times. However, this code is not efficient, since

$$R = \frac{\log_2 2^k}{nk} = \frac{1}{n} \rightarrow 0.$$

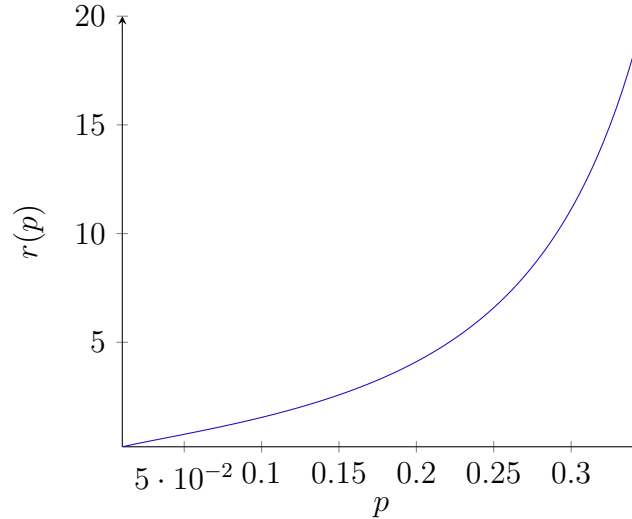


Figure 7 – The graph of the function $r(p)$.

We now explore concentration in matrices that have only $+1$ or -1 as entries. For notation, let $M_n(H_1) := H_{n \times n}$ be the set of all $n \times n$ matrices with only -1 and $+1$ in each entrance.

Corollary 5.4.5. *Let $n \in \mathbb{N}$ and $M \in M_n(H_1)$ be a random matrix such that each entry is Rademacher with parameter $1/2$. Let $\|\cdot\| : M_n(H_1) \rightarrow \mathbb{R}_+$ be the operator norm, then*

$$\mathbb{P}\left(\left|\|M\| - \mathbb{E}[\|M\|]\right| \geq t\right) \leq 2 \exp\left(-\frac{t^2}{4n}\right).$$

Proof. It is well-known that

$$\|A\| = \sup_{x,y \in \mathbb{S}^{n-1}} \langle x, Ay \rangle,$$

where \mathbb{S}^{n-1} is the unit sphere and $A \in M_n(\mathbb{R})$. Let \bar{A}^{ij} be the matrix where the ij entry of A is flipped. Then

$$\left|\|A\| - \|\bar{A}^{ij}\|\right| \leq \|A - \bar{A}^{ij}\|.$$

Since

$$\begin{aligned} \|A - \bar{A}^{ij}\| &= \sup_{x,y \in \mathbb{S}^{n-1}} \langle x, (A - \bar{A}^{ij})y \rangle \\ &= \sup_{x,y \in \mathbb{S}^{n-1}} \sum_{m,n=1}^n (A_{mn} - \bar{A}_{mn}^{ij}) x_m y_n \\ &\leq 2 \sup_{x,y \in \mathbb{S}^{n-1}} x_i y_j \\ &= 2. \end{aligned}$$

Hence the operator $\|\cdot\|$ satisfies

$$\left|\|A\| - \|\bar{A}^{ij}\|\right| \leq L_{ij} d(A, \bar{A}_{ij}),$$

where $L_{ij} = 2n^2$. Therefore Corollary 5.4.3 implies that

$$\begin{aligned} \mathbb{P}\left(\left|\|M\| - \mathbb{E}[\|M\|]\right| \geq t\right) &\leq 2 \exp\left(-\frac{n^4 t^2}{\sum_{i,j=1}^n (2n^2)^2}\right) \\ &= 2 \exp\left(-\frac{t^2}{4n}\right), \end{aligned}$$

and the corollary is proved. □

5.5 Application II: Rademacher Complexity

Corollary 5.4.5 tell us that in order to control $\|M\|$, we need to control

$$\mathbb{E}[\|M\|] = \mathbb{E}\left(\sup_{x,y \in \mathbb{S}^{n-1}} \langle x, My \rangle\right).$$

This works as a motivation to the following definition.

Definition 5.5.1. Let $T \subseteq \mathbb{R}^n$ and $X \in H_n$ be a Rademacher r.v., that is, a random vector distributed according the uniform measure in H_n . Then the **Rademacher complexity** of T is defined as

$$r(T) := \mathbb{E}[\sup_{t \in T} \langle t, X \rangle].$$

Remark 5.5.1. In some books, the Rademacher Complexity is defined as

$$\tilde{r}(T) = \mathbb{E}[\sup_{t \in T} |\langle t, X \rangle|].$$

These definitions are equivalent, in the sense that

$$r(T) \leq \tilde{r}(T) \leq 2r(T).$$

Let us first state some of its properties. Recall that the **Minkowski Sum** of two sets is

$$A + B := \{a + b : a \in A, b \in B\}.$$

Lemma 5.5.1. Let $T, S \subseteq \mathbb{R}^n$ and $a \in \mathbb{R}$. Then

1. We have that $r(T + S) = r(T) + r(S)$;
2. We also have that $r(aT) = |a|r(T)$;
3. In particular, we obtain $r(T - T) = 2r(T)$;
4. The relation between the Rademacher Complexity and the Diameter (in the Euclidean norm) is the following:

$$\frac{1}{\sqrt{8\pi \log n}} \text{diam}(T) \leq r(T) \leq \frac{\sqrt{n}}{2} \text{diam}(T); \text{ and}$$

5. If $\text{conv}(T)$ denotes the convex hull of T , then

$$r(\text{conv}(T)) = r(T).$$

Proof. (1) Notice that

$$\begin{aligned} r(T + S) &= \mathbb{E}[\sup_{(x,y) \in T \times S} \langle X, x + y \rangle] \\ &= \mathbb{E}[\sup_{x \in T} \langle X, x \rangle] + \mathbb{E}[\sup_{y \in S} \langle X, y \rangle]. \end{aligned}$$

- (2) For $a \geq 0$, we have trivially $r(aT) = ar(T)$. Now, note that

$$\begin{aligned} r(-T) &= \mathbb{E}[\sup_{x \in T} \langle X, -x \rangle] \\ &= \mathbb{E}[\sup_{x \in T} \langle -X, x \rangle] \\ &= r(T), \end{aligned}$$

since X is symmetric. Hence we get, for $a < 0$,

$$r(aT) = |a|r(-T) = |a|r(T).$$

(4) We will just prove the upper bound. Using (3), we get

$$\begin{aligned} r(T) &= \frac{1}{2} \mathbb{E} \left[\sup_{x, y \in T} \langle X, x - y \rangle \right] \\ &\leq \frac{1}{2} \mathbb{E} [\|X\| \sup_{x, y \in T} \|x - y\|] \\ &= \frac{\sqrt{n}}{2} \text{diam}(T), \end{aligned}$$

hence the upper bound.

(5) It is clear that

$$r(\text{conv}(T)) \geq r(T).$$

Now, for any point $t \in \text{conv}(T)$, there is an $n = n(t) \in \mathbb{N}$, $t_1, \dots, t_n \in T$ and $p(t) \in \mathbb{R}_+^n$ such that $\sum_{i=1}^n p_i(t) = 1$ and

$$t = \sum_{i=1}^n p_i(t) t_i,$$

hence

$$\begin{aligned} \sup_{t \in \text{conv}(T)} \langle X, t \rangle &= \sup_{t \in \text{conv}(T)} \sum_{i=1}^n p_i(t) \langle X, t_i \rangle \\ &\leq \sup_{t \in \text{conv}(T)} \sum_{i=1}^n p_i(t) \sup_{s \in T} \langle X, s \rangle \\ &= \sup_{s \in T} \langle X, s \rangle, \end{aligned}$$

hence we have the other direction. □

Lemma 5.5.2. *Let $T \subset \mathbb{R}^n$ be a finite set and $\sigma^2 = \sup_{t \in T} \|t\|^2$, then*

$$r(T) \leq 2\sqrt{\sigma^2 \log |T|}.$$

Proof. Using the concavity of $\log x$, for all $\lambda > 0$ we have

$$\begin{aligned} r(T) &= \mathbb{E} \left[\frac{1}{\lambda} \log e^{\lambda \sup_{t \in T} \langle X, t \rangle} \right] \\ &\leq \frac{1}{\lambda} \log \mathbb{E} [e^{\lambda \sup_{t \in T} \langle X, t \rangle}] \\ &\leq \frac{1}{\lambda} \log \left(\sum_{t \in T} \mathbb{E} [e^{\lambda \langle X, t \rangle}] \right). \end{aligned}$$

Now we need to control the generating function of $\langle t, X \rangle$. Notice that

$$\langle t, x \rangle - \langle t, \bar{x}_i \rangle \leq 2|t_i|,$$

hence the proof of Corollary 5.4.1 shows that

$$\begin{aligned} \text{Ent}(e^{\lambda \langle t, X \rangle}) &\leq \frac{\lambda^2}{4} \mathbb{E} \left(e^{\lambda \langle t, X \rangle} \sum_{i=1}^n (\langle t, X \rangle - \langle t, \bar{X}_i \rangle)_+^2 \right) \\ &\leq \lambda^2 \sigma^2 \mathbb{E}[e^{\lambda \langle t, X \rangle}]. \end{aligned}$$

Therefore, Herbst's Method 4.5.6 implies that

$$\mathbb{E}[e^{\lambda \langle t, X \rangle}] \leq e^{\lambda^2 \sigma^2},$$

for all $t \in T$. Hence

$$r(T) \leq \frac{\log |T| + \lambda^2 \sigma^2}{\lambda}.$$

Using $\lambda = \sqrt{\log |T|}/\sigma$, we finally obtain

$$r(T) \leq 2\sqrt{\sigma^2 \log |T|},$$

and the lemma is proved. \square

Now we can state the final theorem of this section, gathering these results.

Theorem 5.5.1. *Let $T \subset \mathbb{R}^n$ be a finite set, X be a Rademacher random vector in \mathbb{R}^n and*

$$Z := \sup_{t \in T} \langle t, X \rangle.$$

Let $\sigma^2 = \sup_{t \in T} \|t\|^2$. Then $\mathbb{E}[Z] \leq 2\sqrt{\sigma^2 \log |T|}$ and

$$\mathbb{P}(Z \geq 2\sqrt{\sigma^2 \log |T|} + u) \leq \exp\left(-\frac{u^2}{4\sigma^2}\right).$$

Proof. We have already proved the first part of the theorem. For the second, notice that

$$\mathbb{P}(Z \geq s) \leq \sum_{t \in T} \mathbb{P}(\langle t, X \rangle \geq s).$$

Herbst's Method 4.5.6 implies that

$$\mathbb{P}(\langle t, X \rangle \geq s) \leq \exp\left(-\frac{s^2}{4\sigma^2}\right).$$

Hence

$$\mathbb{P}(Z \geq s) \leq \exp\left(\log |T| - \frac{s^2}{4\sigma^2}\right)$$

Using $s = u + 2\sqrt{\sigma^2 \log |T|}$ yields the result. \square

Let us give an example of $r(T)$. For $p \in [1, \infty]$, let

$$\|x\|_p = \left(\sum_{i=1}^n |x_i|^p \right)^{1/p},$$

and $\|x\|_\infty = \sup_{i \leq n} |x_i|$ be the l_p norm. Then we have the following example.

Example 5.5.1. Let $T = B_p^n := \{x \in \mathbb{R}^n : \|x\|_p \leq 1\}$, then

$$r(T) = n^{1/q},$$

since

$$\sup_{t \in T} \langle t, x \rangle = \|x\|_q,$$

where q is the conjugate exponent of p and $\|x\|_q = n^{1/q}$ for $x \in H_n$.

We remark that we can extend Theorem 5.5.1 to Totally Bounded sets. For a proof, see [Vershynin \(2017\)](#), [van Handel \(2014\)](#) and [Boucheron et al. \(2013\)](#).

5.6 Application IV: Supervised Classification Problem

In this section, we explore what we consider one of the most important applications of Rademacher Complexity. Our goal here is to predict the classification of a random object $X \in \mathcal{X}$ in two different classes. The problem can be modeled as follows. Suppose $(X, Y) \in \mathcal{X} \times \{-1, 1\}$ is drawn according to a probability measure \mathbb{P} . Then we want to find a function $h : \mathcal{X} \rightarrow \{-1, 1\}$ such that it minimizes the error

$$L(h) := \mathbb{P}(h(X) \neq Y).$$

If \mathbb{P} is known, then we can find the best minimizer exactly.

Lemma 5.6.1. *We have that*

$$\min_h \mathbb{P}(h(X) \neq Y) = \mathbb{P}(h^*(H) \neq Y),$$

where

$$h^*(x) = \text{sign}(\mathbb{E}[Y|X = x]).$$

However, in most cases \mathbb{P} is unknown and we only have a sample $(X_i, Y_i)_{i=1}^n \sim \mathbb{P}$. Moreover, we want to find the best h in a finite class \mathcal{F} of functions $f : \mathcal{X} \rightarrow \{-1, 1\}$. In this case, we will minimize the empirical error

$$L_n(h) := \frac{1}{n} \sum_{k=1}^n \mathbf{1}_{h(X_i) \neq Y_i}.$$

Let $h^*, h_n : \mathcal{X} \rightarrow \{-1, 1\}$ be such that

$$L(h^*) = \min_{h \in \mathcal{F}} L(h), \quad (5.1)$$

and

$$L_n(h_n) = \min_{h \in \mathcal{F}} L_n(h). \quad (5.2)$$

Hence h^* is also unknown. But we can still quantify the error. Let us proceed to do this.

First, let us define the *shattering coefficient*.

Definition 5.6.1. Let \mathcal{F} be a class of functions $f : \mathcal{X} \rightarrow \{-1, 1\}$, then the **shattering coefficient** is defined as

$$S_n(\mathcal{F}) = \max_{x_1, \dots, x_n \in \mathcal{X}} |\{(h(x_1), \dots, h(x_n)) : h \in \mathcal{F}\}|.$$

Then we have the following theorem.

Theorem 5.6.1. Let $(X_i, Y_i)_{i=1}^n \subset \mathcal{X} \times \{-1, 1\}$ be an independent sample according to the probability \mathbb{P} , \mathcal{F} be a class of functions $f : \mathcal{X} \times \{-1, 1\}$ and h^* and h_n be as in Equalities 5.1 and 5.2, then

$$\mathbb{E}[|L_n(h_n) - L(h_n)|] \leq 8\sqrt{\frac{\log(S_n(\mathcal{F}))}{n}},$$

and

$$\mathbb{E}[L(h_n) - L(h^*)] \leq 16\sqrt{\frac{\log(S_n(\mathcal{F}))}{n}}.$$

Remark 5.6.1. We could also state a high dimensional version of this result, but it will not provide much improvement than the bound on the expected value. Moreover, our constants are not optimal, but they capture the right order of fluctuation. See [Giraud \(2014\)](#) for more information.

Proof. Let

$$\Delta_n = \sup_{h \in \mathcal{F}} |L_n(h) - L(h)|,$$

then

$$|L_n(h_n) - L(h_n)| \leq \Delta_n.$$

By the definition of h_n , we have $L_n(h_n) \leq L_n(h^*)$, hence we also have that

$$\begin{aligned} L(h_n) - L(h^*) &= L(h_n) - L_n(h_n) + L_n(h_n) - L(h^*) \\ &\leq L(h_n) - L_n(h_n) + L_n(h^*) - L(h^*) \\ &\leq 2\Delta_n. \end{aligned}$$

Therefore, we just need to estimate $\mathbb{E}[\Delta_n]$ from above. Notice first that

$$\mathbb{E}[L_n(h) - L(h)] = 0.$$

Now we need a simple symmetrization lemma

Lemma 5.6.2. *We have that*

$$\mathbb{E}[\Delta_n] \leq 2\mathbb{E}\left(\mathbb{E}_\varepsilon\left[\sup_{h \in \mathcal{F}} \left|\frac{1}{n} \sum_{k=1}^n \varepsilon_k \mathbf{1}_{Y_i \neq h(X_i)}\right|\right]\right),$$

where ε is uniform in H_n and independent of $(X_i, Y_i)_{i=1}^n$.

This lemma can be found in [Vershynin \(2017\)](#) and [Giraud \(2014\)](#).

Therefore, we can simply bound $\mathbb{E}[\Delta_n]$ by

$$\mathbb{E}[\Delta_n] \leq \frac{2}{n} \sup_{y \in H_n} \sup_{x \in \mathcal{X}^n} \mathbb{E}\left[\sup_{h \in \mathcal{H}} \sum_{k=1}^n \varepsilon_k \mathbf{1}_{y_i \neq h(x_i)}\right].$$

Now, for $(x, y) \in \mathcal{X}^n \times \{-1, 1\}^n$, we can consider the set

$$T_n(x, y) := \{(\mathbf{1}_{y_i \neq h(x_i)}, \dots, \mathbf{1}_{y_n \neq h(x_n)}) : h \in \mathcal{F}\},$$

hence

$$\mathbb{E}[\Delta_n] \leq \frac{2}{n} \sup_{y \in H_n} \sup_{x \in \mathcal{X}^n} \mathbb{E}\left[\sup_{t \in T_n(x, y)} |\langle \varepsilon, t \rangle|\right].$$

Now, using the Rademacher Complexity and Remark 5.5.1, we have that

$$\mathbb{E}[\Delta_n] \leq \frac{4}{n} \sup_{y \in H_n} \sup_{x \in \mathcal{X}^n} r(T_n(x, y)).$$

Corollary 5.5.1 implies that

$$r(T_n(x, y)) \leq 2\sqrt{n \log(|T_n(x, y)|)},$$

since

$$\sup_{t \in T_n(x, y)} \|t\|^2 \leq n.$$

Finally, there is a bijection between H_n and $\{0, 1\}^n$, hence

$$|T_n(x, y)| \leq S_n(\mathcal{F}),$$

for all $(x, y) \in \mathcal{X}^n \times \{-1, 1\}^n$, therefore

$$r(T_n(x, y)) \leq 2\sqrt{n \log(S_n(\mathcal{F}))},$$

and the result is proved. \square

Therefore, in order to control the error, we need to control the Rademacher Complexity of the sets $T_n(x, y)$, or simply the shattering coefficient. The latter motivates the definition of the *VC dimension* of \mathcal{F} , which we are not going to introduce here. For more details, see [Giraud \(2014\)](#) and [Vershynin \(2017\)](#).

5.7 Application III: Concentration in Graphs

In this section, we will explore the connection between the Rademacher Log-Sobolev Inequality and Log-Sobolev Inequality in graphs. First, let us define a graph and the graph distance.

Definition 5.7.1. By a **connected and undirected graph** (V, E) , we mean a finite set V , called the **vertices** and a collection $E \subset V \times V$, called the **edges**, such that

1. For all $x \in V$, $(x, x) \notin E$;
2. If $(x, y) \in E$, then $(y, x) \in E$;
3. For all $x, y \in V$ such that $x \neq y$, there are an $n \in \mathbb{N}$ and a sequence $(x_i)_{i=1}^n \subset V$ such that $x_1 := x$, $(x_n, y) \in E$ and

$$(x_i, x_{i+1}) \in E,$$

for all $i \leq n - 1$ if $n > 1$.

Remark 5.7.1. We will simply call a connected and undirected graph by a graph.

Moreover, we have the following definition.

Definition 5.7.2. For $x \in V$, we denote E_x the set of its neighbors, that is

$$E_x := \{y \in V : (x, y) \in E\},$$

and $d(x) := |E_x|$, the **degree** of x .

The graph distance is associated with the Item 3 in Definition 5.7.1.

Definition 5.7.3. Let (V, E) be a graph. Then the distance $d(x, y)$ between two distinct points $x, y \in V$ is defined as the smallest n satisfying Item 3. If $x = y$, we define $d(x, x) := 0$.

Using this distance, we can define an 1-Lipschitz function $f : V \rightarrow \mathbb{R}$.

Definition 5.7.4. A function $f : V \rightarrow \mathbb{R}$ is 1-Lipschitz in the graph (V, E) if for all $x, y \in V$, we have

$$|f(x) - f(y)| \leq d(x, y).$$

Now we can formulate the problem. Let μ be a probability measure in V . We want to find the best constant σ^2 , such that

$$\mathbb{E}_\mu[e^{\lambda(f - \mathbb{E}[f])}] \leq e^{\sigma^2 \lambda^2 / 2}, \quad (5.3)$$

for all $\lambda \geq 0$ and 1-Lipschitz function f in (V, E) . Equivalently, we want to find the best constant σ^2 such that

$$\mathcal{D}(\mu_\lambda || \mu) \leq \frac{\sigma^2 \lambda^2}{2},$$

where

$$\frac{d\mu_\lambda}{d\mu} = \frac{e^{\lambda f}}{\mathbb{E}_\mu[e^{\lambda f}]}.$$

To begin with, we proved this result to the Hamming Cube in Corollary 5.4.2.

Example 5.7.1. Let $V = H_n$, $(x, y) \in E$ if and only if $d(x, y) = 1/n$ in the Hamming distance and let $d_0 = nd$ be the graph distance. Let also $\mu(\{1\}) = p$ and

$$\mu_n = \prod_{i=1}^n \mu,$$

Then for all 1-Lipschitz functions $f : V \rightarrow \mathbb{R}$ in the graph distance we have that

$$\text{Ent}_\mu(e^{\lambda f}) \leq \frac{nc(p)\lambda^2}{8} \mathbb{E}_\mu[e^{\lambda f}],$$

hence Herbst's Method implies that the best constant σ^2 is bounded:

$$\sigma^2 \leq \frac{nc(p)}{4}.$$

Let us define now the energy on a graph (V, E) .

Definition 5.7.5. Let μ be a probability measure on a graph (V, E) , then the **energy** associated with μ is

$$\mathcal{E}(f) := \frac{1}{4} \sum_{x \in V} \sum_{y \in E_x} [f(x) - f(y)]^2 \mu(x),$$

where $\mu(x) = \mu(\{x\})$.

Remark 5.7.2. Note that this generalizes the energy on the Hamming Cube.

We could also define the *generator* \mathcal{L} of this energy.

Definition 5.7.6. Let μ be a probability measure on a graph (V, E) and

$$\mathcal{F} := \{f : V \rightarrow \mathbb{R}\}.$$

Then the generator $\mathcal{L} : \mathcal{F} \rightarrow \mathcal{F}$ is defined as

$$\mathcal{L}f(x) := \frac{1}{2} \sum_{y \in E_x} \mu(y)(f(y) - f(x)).$$

Indeed, we have the following lemma.

Lemma 5.7.1. *We have that*

$$\langle f, -\mathcal{L}f \rangle = \mathcal{E}(f).$$

Proof.

$$\begin{aligned} \langle f, -\mathcal{L}f \rangle &= \frac{1}{2} \sum_{x \in V} \sum_{y \in E_x} f(x)[f(x) - f(y)]\mu(x)\mu(y) \\ &= \frac{1}{2} \sum_{x, y \in V} [f(x) - f(y)]^2 q(x, y), \end{aligned}$$

where $q(x, y) := \mu(x)\mu(y)$ if $(x, y) \in E$ and 0 otherwise. By symmetry, we have that

$$\begin{aligned} \langle f, -\mathcal{L}f \rangle &= \frac{1}{4} \sum_{x \in V} \sum_{y \in E_x} [f(x) - f(y)]^2 q(x, y) \\ &= \mathcal{E}(f), \end{aligned}$$

and the result is proved. □

We can now define the Log-Sobolev Inequality on the graph.

Definition 5.7.7. We say that a probability μ on a graph (V, E) satisfies a Log-Sobolev Inequality with constant c if

$$\text{Ent}_\mu(f^2) \leq c\mathcal{E}(f),$$

for all $f : V \rightarrow \mathbb{R}$.

Using this, we can simply bound the σ^2 constant in Inequality 5.3 using the following lemma.

Lemma 5.7.2. *Suppose (V, E, μ) satisfies a Log-Sobolev Inequality with constant c . Then, the best constant σ^2 in Inequality 5.3 is bounded:*

$$\sigma^2 \leq \frac{c \max_{x \in E} d(x)}{4}.$$

Proof. The Log-Sobolev Inequality implies that

$$\text{Ent}(e^{\lambda f}) \leq \frac{c\lambda^2}{8} \mathbb{E} \left(e^{\lambda f(X)} \sum_{u \in E_X} (f(X) - f(u))_+^2 \right).$$

Using that f is 1-Lipschitz and that $d(X) \leq \max_{x \in E} d(x)$ a.s., we obtain the result by Herbst's Method. □

Moreover, there is also a tensorization lemma for graphs. First, let us define the product of graphs.

Definition 5.7.8. Let (V_i, E_i, d_i) , for $i = 1, 2$, be two graphs. Then the **graph product** $(V_1 \times V_2, E_1 \times E_2)$ is a graph such that

$$((x_1, y_1), (x_2, y_2)) \in E_1 \times E_2,$$

if $(x_1, x_2) \in E_1$ or $(y_1, y_2) \in E_2$, but not simultaneously. The graph distance $d_1 \times d_2$ on the product is trivially

$$d_1 \times d_2((x_1, y_1), (x_2, y_2)) = d_1(x_1, x_2) + d_2(y_1, y_2).$$

Then we have the Tensorization Formula.

Lemma 5.7.3. Suppose (V_i, E_i, d_i, μ_i) satisfy a Log-Sobolev Inequality with constant c_i , for $i = 1, 2$. Then $(V_1 \times V_2, E_1 \times E_2, d_1 \times d_2, \mu_1 \times \mu_2)$ also satisfies a Log-Sobolev Inequality with constant $c = \max\{c_1, c_2\}$.

Our next and final example will be the complete graph.

Definition 5.7.9. We say that (V, E) is **complete** if $(x, y) \in E$ whenever $x \neq y$.

Thus we have the following theorem.

Theorem 5.7.1. Let μ be a probability measure in a complete graph (V, E) and suppose that

$$p := \min_{x \in V} \mu(\{x\}) > 0,$$

then (V, E, μ) satisfies a Log-Sobolev Inequality with constant

$$c = c(p).$$

Proof. The proof is based on [Diaconis et al. \(1996\)](#). Note first that

$$\frac{1}{c} = \inf \left\{ \frac{\mathcal{E}(g)}{\text{Ent}(g^2)} : \text{Ent}(g^2) \neq 0 \right\}.$$

Since

$$(|x| - |y|)^2 \leq (x - y)^2,$$

we have that $\mathcal{E}(|g|) \leq \mathcal{E}(g)$, and then we can consider the infimum to be restricted to nonnegative functions g . Let $f : V \rightarrow \mathbb{R}$ be any minimizer, therefore

$$\text{Ent}(f^2) = c\mathcal{E}(f).$$

Let g be any function and $r : \mathbb{R}_+ \rightarrow \mathbb{R}_+$ such that

$$r(\lambda) := \text{Ent}[(f + \lambda g)^2] - c\mathcal{E}(f + \lambda g).$$

By definition of f , we have that

$$r'(0) = 0.$$

Computing the derivative, we obtain that

$$\begin{aligned} r'(0) &= 2\mu\left(gf \log \left[\frac{f^2}{\mu(f^2)}\right] - cg(-\mathcal{L}f)\right) \\ &= 2\langle g, f \log \left[\frac{f^2}{\mu(f^2)}\right] + c\mathcal{L}f \rangle. \end{aligned}$$

Since this must be 0 for all g , we obtain that

$$f \log f^2 - f \log \mu(f^2) + c\mathcal{L}f = 0. \quad (5.4)$$

In our case, the generator satisfies

$$\begin{aligned} \mathcal{L}f(x) &= -\frac{1}{2} \sum_{y \in E_x} \mu(y)[f(x) - f(y)] \\ &= -\frac{1}{2} \sum_{y \in V} \mu(y)[f(x) - f(y)] \\ &\quad - \frac{f(x)}{2} + \frac{1}{2} \mathbb{E}_\mu[f], \end{aligned}$$

then Equation 5.4 can be rewritten as

$$2f \log f = f \log \mu(f^2) + \frac{cf}{2} - \frac{c}{2} \mathbb{E}_\mu(f). \quad (5.5)$$

The right side is a linear function of f and the left is the function $t \mapsto t \log t$ applied to f . Now, since the latter is convex, it can only intersect a line in at most two points. Therefore Equation 5.5 means that f takes at most two values and we recover the binary case. Let $a, b \in \mathbb{R}$ be these two values and

$$s := \mu(\{x \in V : f(x) = a\}).$$

Then, by Theorem 5.3.2, we have

$$c = \sup_s c(s).$$

Notice that, by symmetry, we can assume that $s \in [p, 1/2]$. Finally, we can easily check that the supremum is achieved at $s = p$. \square

For further information on Log-Sobolev Inequalities in graphs, see [Bobkov et al. \(2006\)](#), [Diaconis et al. \(1996\)](#), and [Bobkov and Tetali \(2006\)](#).

Open the way for Gauss!

6.1 Introduction

In the previous chapter we discussed the discrete Log-Sobolev Inequality scenario and some applications, such as concentration in the Hamming Cube. The proof of this result was based on nothing but calculus and some inequalities in \mathbb{R} . However, the Gaussian Log-Sobolev Inequality (GLS) will require some previous results from Information Theory.

Before we introduce the definitions to guide us to the main theorem, let us take a moment to analyze what we want. The Log-Sobolev Inequality states that the entropy of the square of a smooth enough function is controlled by the square of the L^2 norm of its gradient. In another words, if a function does not vary too much, then the entropy functional is close to zero. As we did for the Rademacher case in the previous chapter, here we can derive concentration in Gaussian Spaces and applications.

There are many proofs of this result, which we briefly describe now. The first one uses the Central Limit Theorem and the Discrete Log-Sobolev Inequality. For the proof, see [Boucheron et al. \(2013\)](#) and [Ané et al. \(2000\)](#). In a second proof, we check that the Ornstein-Uhlenbeck semigroup satisfies the Modified Log-Sobolev Inequality, see [van Handel \(2014\)](#). Log-Sobolev Inequality can also be proved for the large class of Boltzmann measures

$$d\mu_W(x) := \frac{1}{Z} e^{-W(x)} dx,$$

where dx is the Lebesgue measure, $W : \mathbb{R}^n \rightarrow \mathbb{R}$ is a strongly convex function with $\text{Hess}(W) \geq c\text{Id}$ and

$$Z := \int e^{-W(x)} dx,$$

which is finite. This proof is based on the Bakry-Émery Criterion, see [Ledoux \(1999\)](#). Note that the Gaussian is a special case when $W(x) = \|x\|^2/2$. There is also a proof based on the Herbst's Method. As we have seen, we use LSI and Herbst's Method to prove

concentration. *Herbst's Inverse Method* uses this idea to obtain LSI from a reverse version of Herbst's argument. There are some conditions, known as Wang's Conditions, that allow us to do that, see [Ané et al. \(2000\)](#).

Finally, there is a proof based on Information Theory Inequalities, such as Exponential Entropy Inequality and Blanchman-Stam Inequality. This is the proof we show in Section 6.3. Even though it is not the most direct one, its connections with Information Theory are clearer and help us find formulas and expressions that are more suitable for some applications.

In Section 6.4, we will use the standard Gaussian Log-Sobolev Inequality to prove concentration in Gaussian Spaces, or concentration of functions of Gaussian vectors. The standard method is Herbst's Method 4.5.6, as we have already shown in Chapter 4.

In Section 6.5 we will extend the Gaussian Concentration to the *Gaussian Complexity*. The main idea is to bound the quantity

$$r(T) = \mathbb{E}[\sup_{t \in T} \langle t, g \rangle],$$

where $g \sim \mathcal{N}(0, \text{Id})$ and $T \subset \mathbb{R}^n$. This is an important quantity in many other applications (see [Vershynin \(2017\)](#)).

In Section 6.6, we will use the equivalent form of Gaussian Log-Sobolev Inequality, namely, for any X random vector with density $f \in C^1(\mathbb{R}^n)$ and finite second moment, we have

$$N(X)J(X) \geq n,$$

where N and J are the exponential entropy and Fisher Information, respectively. We will prove, in particular, the *Crámer-Rao Bound*, namely,

$$|\Sigma(X)| \leq |\mathbb{J}(X)|^{-1},$$

where $\Sigma(X)$ is the covariance matrix of X and $\mathbb{J}(X)$ is the Fisher matrix.

Finally, in Section 6.7, we will use again the Version $N(X)J(X) \geq n$ to prove an *Uncertainty Principle*. This is an important result relating the variance of two associated densities f and g , that is,

$$\begin{aligned} \mathcal{F}(g) &= f; \text{ and} \\ \mathcal{F}(f) &= g, \end{aligned}$$

where \mathcal{F} is the Fourier Transform.

6.2 Definitions

Let us recall some definitions from Chapters 2, 3 and 4. The Gaussian measure γ in \mathbb{R}^n is the probability measure such that

$$\frac{d\gamma}{dx} = \frac{1}{(2\pi)^{n/2}} e^{-\frac{\|x\|^2}{2}}.$$

Given a function $f \geq 0$, we defined its entropy as

$$\begin{aligned} \text{Ent}_\gamma(f) &= \int_{\mathbb{R}^n} f \log(f) d\gamma - \left(\int_{\mathbb{R}^n} f d\gamma \right) \log \left(\int_{\mathbb{R}^n} f d\gamma \right) \\ &= \text{Ent}(f(X)), \end{aligned}$$

where $X \sim \mathcal{N}(0, \text{Id})$. Moreover, we defined its energy as

$$\begin{aligned} \mathcal{E}(f) &= \int_{\mathbb{R}^n} \|\nabla f\|^2 dx \\ &= \mathbb{E}[\|\nabla f(X)\|^2]. \end{aligned}$$

Now, for a random vector X with density $f \in C^1(\mathbb{R}^n)$, we defined the Fisher Information as

$$J(X) = 4 \int_{\mathbb{R}^n} \|\nabla \sqrt{f}\|^2 dx,$$

and the exponential entropy as

$$\begin{aligned} N(X) &= \frac{1}{2\pi e} e^{\frac{2}{n} H(X)} \\ &= \frac{1}{2\pi e} e^{-\frac{2}{n} \int_{\mathbb{R}^n} f \log(f) dx}. \end{aligned}$$

Finally, we need to introduce the space of all *finite second moment* functions, when we see a function as a density of some random vector X .

Definition 6.2.1. Let $(\mathbb{R}^n, \mathcal{B}(\mathbb{R}^n), \mu)$ be a probability space. Then $M^2(\mu)$ is the space of all functions f such that $\|x\|^2 f$ is integrable, that is,

$$M^2(\mu) := \{f : \mathbb{R}^n \rightarrow \mathbb{R} : \int_{\mathbb{R}^n} \|x\|^2 |f| d\mu < \infty\}.$$

6.3 Main Theorem

We are now able to state and prove the main theorem, the *Gaussian Log-Sobolev Inequality*.

Theorem 6.3.1. Let $f \in C^2(\mathbb{R}^n) \cap L^2(\gamma)$ and $f^2 \in M^2(\gamma)$. Then

$$\text{Ent}_\gamma(f^2) \leq 2\mathbb{E}_\gamma[\|\nabla f\|^2].$$

Remark 6.3.1. Even though the assumption $f^2 \in M^2(\gamma)$ is not required, we need it for our proof. For a proof in which this condition is not needed, see [van Handel \(2014\)](#), where he uses properties of the Ornstein-Uhlenbeck Semigroup. Moreover, by a standard approximation, the condition $f \in C^2(\mathbb{R}^n)$ can be weakened to $f \in C^1(\mathbb{R}^n)$.

Before prove it, we need some equivalent statements. The first one is the following.

Lemma 6.3.1. *The following are equivalent.*

1. For all $f \in C^2(\mathbb{R}^n) \cap L^2(\gamma)$ and $f^2 \in M^2(\gamma)$ we have that

$$\text{Ent}_\gamma(f^2) \leq 2\mathbb{E}_\gamma[\|\nabla f\|^2]; \text{ and}$$

2. For all $f \in C^2(\mathbb{R}^n) \cap L^2(dx)$, $f^2 \in M^2(dx)$ and $\|f\|_{L^2(dx)} = 1$ we have that

$$\text{Ent}_{dx}(f^2) \leq \frac{n}{2} \log \left(\frac{2}{e\pi n} \int_{\mathbb{R}^n} \|\nabla f\|^2 dx \right).$$

Remark 6.3.2. Item 2 is known as Lebesgue Log-Sobolev Inequality.

Proof. (\Rightarrow) Let us prove Item 2 for a function p satisfying the hypothesis. We can always assume

$$\int_{\mathbb{R}^n} \|\nabla p\|^2 dx < \infty,$$

otherwise the theorem is trivial. Let us set two auxiliar functions:

$$\begin{aligned} g(x) &:= \lambda^{n/2} p(\lambda x); \\ f(x) &= g(x)(2\pi)^{n/4} e^{\|x\|^2/4}. \end{aligned}$$

It is easy to see that

$$\|p\|_{L^2(dx)} = \|g\|_{L^2(dx)} = \|f\|_{L^2(\gamma)} = 1,$$

and

$$\int_{\mathbb{R}^n} \|\nabla g\|^2 dx, \int_{\mathbb{R}^n} \|\nabla f\|^2 d\gamma < \infty.$$

The following equality holds $f^2(x)h(x) = g^2(x)$, where h is the Gaussian density, that is,

$$h(x) = \frac{1}{(2\pi)^{n/2}} e^{-\|x\|^2/2}.$$

Item 1 applied to f leads to

$$\int_{\mathbb{R}^n} f^2 \log[f^2] h(x) dx \leq 2 \int_{\mathbb{R}^n} \|\nabla f\|^2 h(x) dx. \quad (6.1)$$

Also, we can relate the gradients of f and g as follows. First, note that

$$f(x) = g(x)(2\pi)^{n/4}e^{\|x\|^2/4},$$

then

$$\nabla f(x) = \nabla g(x)(2\pi)^{n/4}e^{\|x\|^2/4} + \frac{1}{2}xg(2\pi)^{n/4}e^{\|x\|^2/4},$$

therefore

$$\|\nabla f(x)\|^2 h(x) = \left(\|\nabla g(x)\|^2 + \frac{1}{4}\|x\|^2 g^2(x) + g\langle \nabla g(x), x \rangle \right).$$

Using this in Inequality 6.1 we have that

$$\int_{\mathbb{R}^n} g^2 \log(g^2(2\pi)^{n/2}e^{\|x\|^2/2}) dx \leq 2 \int_{\mathbb{R}^n} \left(\|\nabla g(x)\|^2 + \frac{1}{4}\|x\|^2 g^2(x) + g\langle \nabla g(x), x \rangle \right) dx.$$

To simplify, let us call L the left side and R the right hand side, then

$$L = \text{Ent}_\lambda(g^2) + \frac{n}{2} \log(2\pi) + \frac{1}{2} \int_{\mathbb{R}^n} g^2 \|x\|^2 dx,$$

and

$$R = 2 \int_{\mathbb{R}^n} \|\nabla g(x)\|^2 dx + \frac{1}{2} \int_{\mathbb{R}^n} g^2(x) \|x\|^2 dx + 2 \int_{\mathbb{R}^n} g(x) \langle \nabla g(x), x \rangle dx.$$

Cancelling terms (and finite by assumption), we obtain

$$\text{Ent}_\lambda(g^2) + \frac{n}{2} \log(2\pi) \leq 2 \int_{\mathbb{R}^n} \|\nabla g(x)\|^2 dx + 2 \int_{\mathbb{R}^n} g(x) \langle \nabla g(x), x \rangle dx.$$

Using Corollary 2.5.7, we obtain

$$2 \int_{\mathbb{R}^n} g(x) \langle \nabla g(x), x \rangle dx = - \sum_{i=1}^n \int_{\mathbb{R}^n} g^2(x) dx = -n,$$

hence Item 1 is equivalent to

$$\text{Ent}_{dx}(g^2) \leq 2 \int_{\mathbb{R}^n} \|\nabla g\|^2 dx - n - \frac{n}{2} \log(2\pi), \quad (6.2)$$

with $\|g\|_{L^2(dx)} = 1$. As $g(x) = \lambda^{n/2}p(\lambda x)$, the left side of Inequality 6.2 becomes

$$\int_{\mathbb{R}^n} g^2(x) \log[g^2(x)] dx = \int_{\mathbb{R}^n} \lambda^n p^2(\lambda x) n \log \lambda dx + \int_{\mathbb{R}^n} \lambda^n p^2(\lambda x) \log p(\lambda x) dx.$$

Set $u = \lambda x$, we have $du = \lambda^n dx$, therefore

$$\int_{\mathbb{R}^n} g^2(x) \log[g^2(x)] dx = n \log(\lambda) \int_{\mathbb{R}^n} p^2(u) du + \int_{\mathbb{R}^n} p^2(u) \log[p^2(u)] du.$$

On the other hand, the first term on the right side in Inequality 6.2 is

$$2 \int_{\mathbb{R}^n} \|\nabla g(x)\|^2 dx = 2 \int_{\mathbb{R}^n} \lambda^n \|\nabla_x p(\lambda x)\|^2 dx,$$

where ∇_x represents the gradient with respect to x . Setting again $u = \lambda x$, we have

$$\begin{aligned} 2 \int_{\mathbb{R}^n} \|\nabla g\|^2 dx &= 2 \int_{\mathbb{R}^n} \|\nabla g(x)\|^2 dx \\ &= 2\lambda^2 \int_{\mathbb{R}^n} \lambda^n \|\nabla_u p(u)\|^2 du. \end{aligned}$$

Summarizing, Item 1 is equivalent to the following with $\lambda > 0$:

$$\text{Ent}_{dx}(p^2) \leq 2\lambda^2 \int_{\mathbb{R}^n} \|\nabla p^2(x)\|^2 dx - n \log(\sqrt{2\pi}\lambda e) =: 2\lambda^2 a - n \log(\lambda b), \quad (6.3)$$

where

$$a = \int \|\nabla p^2(x)\|^2 dx,$$

and $b = e\sqrt{2\pi}$. Minimizing with respect to λ , we have the optimal value $\lambda^* = \sqrt{\frac{n}{4a}}$. Therefore we finally have

$$\begin{aligned} \text{Ent}_{dx}(p^2) &\leq \frac{n}{2} - n \log[b(n(4a)^{-1})^{1/2}] \\ &= \frac{n}{2} \log \left(\frac{2}{e\pi n} \int_{\mathbb{R}^n} \|\nabla p(x)\|^2 dx \right), \end{aligned}$$

which is Item 2.

(\Leftarrow) As seen in the proof above, the right side of the conclusion in Item 2 is the minimum with respect $\lambda > 0$ of $2\lambda^2 a - n \log(\lambda b)$, where a and b are the same in Inequality 6.3. Hence

$$\text{Ent}_{dx}(f^2) \leq \frac{n}{2} \log \left(\frac{2}{e\pi n} \int \|\nabla f(x)\|^2 dx \right) \leq 2\lambda^2 a - n \log(\lambda b).$$

Therefore we can invert the substitutions in the previous proof. \square

Now we can prove Lebesgue Log-Sobolev Inequality.

Lemma 6.3.2. *For all $f \in C^2(\mathbb{R}^n) \cap L^2(dx)$, $f^2 \in M^2(dx)$ and $\|f\|_{L^2(dx)} = 1$ we have that*

$$\text{Ent}_{dx}(f^2) \leq \frac{n}{2} \log \left(\frac{2}{e\pi n} \int_{\mathbb{R}^n} \|\nabla f\|^2 dx \right).$$

Proof. Let f be such as in lemma above and

$$\int_{\mathbb{R}^n} \|\nabla f\|^2 dx < \infty.$$

Let X be a random vector with density f^2 , which is by assumption a density. Then the fact that $f^2 \in M^2(dx)$ and Corollary 3.4.2 imply that

$$H(X) < \infty.$$

On the other hand, we also have

$$J(X) = 4 \int_{\mathbb{R}^n} \|\nabla f\|^2 dx < \infty,$$

hence Corollary 4.2.1 implies that

$$N(X)J(X) \geq n,$$

that is,

$$\frac{2}{\pi e} \exp\left(\frac{-2\text{Ent}_{dx}(f^2)}{n}\right) \int_{\mathbb{R}^n} \|\nabla f\|^2 dx \geq n,$$

or, equivalently,

$$\text{Ent}_{dx}(f^2) \leq \frac{n}{2} \log\left(\frac{2}{n\pi e} \int_{\mathbb{R}^n} \|\nabla f\|^2 dx\right),$$

and the theorem is proved. \square

Remark 6.3.3. Notice that we used Corollary 4.2.1 to prove Lebesgue Log-Sobolev Inequality and by Lemma 6.3.1 we also have Gaussian Log-Sobolev Inequality. However, it is easy to see that Corollary 4.2.1 and Lebesgue Log-Sobolev Inequality are equivalent. Indeed, we just have to invert the above proof. Therefore Corollary 4.2.1, Theorem 6.3.1 and Lemma 6.3.2 are equivalent.

Using Fisher Information Matrix, we can state another equivalence.

Lemma 6.3.3. *The following are equivalent.*

1. Let $X \in L^2(\Omega, \mathcal{F}, \mathbb{P})$ be a random vector with $J(X) < \infty$ and density $f \in C^2(\mathbb{R}^n)$, then $N(X)J(X) \geq n$; and
2. Let $X \in L^2(\Omega, \mathcal{F}, \mathbb{P})$ be a random vector and $J(X) < \infty$ and density $f \in C^2(\mathbb{R}^n)$, then $N(X)|\mathbb{J}(X)|^{1/n} \geq 1$.

Remark 6.3.4. Item 2 is known as Strong Log-Sobolev Inequality.

Proof. (\Rightarrow) Let $Y = \mathbb{J}(X)^{1/2}X$. Then Corollary 3.4.6 and Lemma 3.4.8 take the form

$$N(Y) = |\mathbb{J}(X)|^{1/n} N(X),$$

and

$$\mathbb{J}(Y) = \text{Id}.$$

Therefore $J(Y) = n$. Hence Inequality $J(Y)N(Y) \geq n$ is rewritten as

$$|\mathbb{J}(X)|^{1/n} N(X) \geq 1.$$

(\Leftarrow) Since $|\mathbb{J}(X)|^{1/n}$ and $J(X)/n$ are the geometric and arithmetic means of the eigenvalues of $\mathbb{J}(X)$, the mean inequality states that

$$J(X)/n \geq |\mathbb{J}(X)|^{1/n},$$

hence

$$\frac{1}{n} N(X) J(X) \geq N(X) |\mathbb{J}(X)|^{1/n} \geq 1,$$

and then we have the result. \square

Summarizing, we have the following theorem.

Theorem 6.3.2. *All sentences below are equivalent.*

1. Let $f \in C^2(\mathbb{R}^n) \cap L^2(d\gamma)$ and $f^2 \in M^2(d\gamma)$, then

$$\text{Ent}_\gamma(f^2) \leq 2\mathbb{E}_\gamma[\|\nabla f\|^2];$$

2. Let $f \in C^2(\mathbb{R}^n) \cap L^2(dx)$, $f^2 \in M^2(dx)$ and $\|f\|_{L^2(dx)} = 1$, then

$$\text{Ent}_{dx}(f^2) \leq \frac{n}{2} \log \left(\frac{2}{e\pi n} \int_{\mathbb{R}^n} \|\nabla f\|^2 dx \right);$$

3. Let $X \in L^2(\Omega, \mathcal{F}, \mathbb{P})$ be a random vector with $J(X) < \infty$ and density $f \in C^2(\mathbb{R}^n)$, then $N(X)J(X) \geq n$; and

4. Let $X \in L^2(\Omega, \mathcal{F}, \mathbb{P})$ be a random vector and $J(X) < \infty$ and density $f \in C^2(\mathbb{R}^n)$, then $N(X)|\mathbb{J}(X)|^{1/n} \geq 1$.

In the next sections, we will explore the applications of Theorem 6.3.1. Without trying to run out all the applications, we will only mention four results which we consider relevant: Concentration in Gaussian Spaces; Gaussian Process and Gaussian Complexity; The Crámer-Rao Bound and the Uncertainty Principle.

6.4 Application I: Concentration in Gaussian Spaces

Let X be a standard Gaussian random vector, then Theorem 6.3.1 is equivalent to saying that, for every $f \in C^1(\mathbb{R}^n)$ we have

$$\text{Ent}(f^2(X)) \leq 2\mathbb{E}[\|\nabla f(X)\|^2],$$

according to Remark 6.3.1.

Now, let $g \in C^1(\mathbb{R}^n)$ be a Lipschitz function with $\|\nabla g\|^2 \leq L^2$ and let

$$f = \exp(\lambda g/2),$$

then by the chain rule we have

$$\begin{aligned} \text{Ent}(e^{\lambda g(X)}) &\leq \frac{\lambda^2}{2} \mathbb{E}[\|\nabla g(X)\|^2 e^{\lambda g(X)}] \\ &\leq \frac{\lambda^2 L^2}{2} \mathbb{E}[e^{\lambda g(X)}]. \end{aligned}$$

By Herbst's Method 4.5.6, we have that

$$\mathbb{P}\left(|g(X) - \mathbb{E}[g(X)]| \geq t\right) \leq 2 \exp\left(-t^2/(2L^2)\right).$$

Notice that this bound does not depend on n . Therefore, we have the following theorem.

Theorem 6.4.1. *Let $X \in \mathbb{R}^n$ be a random Gaussian vector. Let $f \in C^1(\mathbb{R}^n)$ be a Lipschitz function with constant L , then*

$$\mathbb{P}\left(|f(X) - \mathbb{E}[f(X)]| \geq t\right) \leq 2 \exp\left(-\frac{t^2}{2L^2}\right).$$

In fact, we can weaken the hypothesis required to functions which are Lipschitz on each coordinate.

Corollary 6.4.1. *Let $X \in \mathbb{R}^n$ be a random Gaussian vector. Let $f \in C^1(\mathbb{R}^n)$ be a Lipschitz function such that*

$$|f(x_1, \dots, x_{i-1}, u, \dots, x_{n-1}) - f(x_1, \dots, x_{i-1}, v, \dots, x_{n-1})| \leq L_i |u - v|,$$

for all $u, v \in \mathbb{R}$, all $x \in \mathbb{R}^{n-1}$ and all i , then

$$\mathbb{P}\left(|f(X) - \mathbb{E}[f(X)]| \geq t\right) \leq 2 \exp\left(-\frac{t^2}{2 \sum_{k=1}^n L_k^2}\right).$$

We will give two examples to illustrate this impressive result.

Example 6.4.1. Let X be a Gaussian vector and $f(x) = \|x\|$, then f satisfies Theorem 6.4.1 with constant $L = 1$, and

$$\mathbb{P}\left(\left|\|X\| - \mathbb{E}[\|X\|]\right| \geq t\right) \leq 2 \exp\left(-t^2/2\right).$$

Notice that $\mathbb{E}[\|X\|] \leq \left(\mathbb{E}[\|X\|^2]\right)^{1/2} = \sqrt{n}$, therefore $\mathbb{E}[\|X\|]$ is at most of order \sqrt{n} , so by replacing $t = \sqrt{n}u$, we have

$$\mathbb{P}\left(\left|\frac{\|X\|}{\sqrt{n}} - \frac{\mathbb{E}[\|X\|]}{\sqrt{n}}\right| \geq u\right) \leq 2 \exp\left(-\frac{nu^2}{2}\right).$$

This last result captures the right order of fluctuation of $\|X\|$.

Example 6.4.2. Let $x \in \mathbb{R}^n$ be an unknown vector. Let A be a deterministic $m \times n$ matrix and let

$$y = Ax + \varepsilon,$$

where ε is a standard Gaussian vector in \mathbb{R}^m . Suppose we want to estimate x from the output y . The quadratic regression gives the best solution as $x^* = \operatorname{argmin}_{z \in \mathbb{R}^n} \|y - Az\|$. How good is this solution? We now give a simple bound on the probability of error $\|x - x^*\|$.

Notice first that the real solution is $x = (A^T A)^{-1} A^T (y - \varepsilon)$ and $x^* = (A^T A)^{-1} A^T y$, then

$$\|x - x^*\| = \|(A^T A)^{-1} A^T \varepsilon\| \leq k_A \|\varepsilon\|,$$

where k_A is the condition number of A , which satisfies

$$\|(A^T A)^{-1} A^T\| \leq \|A^T\| \|(A^T A)^{-1}\| \leq \frac{\sigma_{\max}(A)}{\sigma_{\min}(A)} := k_A.$$

Therefore, we can apply the previous example to bound $\|x - x^*\|$, which gives

$$\mathbb{P}\left(\left|\frac{\|x - x^*\|}{k_A \sqrt{m}} - \frac{\mathbb{E}[\|x - x^*\|]}{k_A \sqrt{m}}\right| \geq u\right) \leq 2 \exp\left(-mt^2/2\right).$$

Although this result gives a bad fluctuation $\|x - x^*\| \sim k_A \sqrt{m}$, this order does not depend on n , which is impressive. However, we can strengthen this result using a *random matrix* instead of a deterministic. We will do this in the next section, particularly in Corollary 6.5.1.

6.5 Application II: Gaussian Complexity

Let $V \subset \mathbb{R}^n$. We have already defined a way to measure the complexity of V in Section 5.5. In this section we will give another way to define precisely this quantity, but we will use Gaussian random vectors, instead of Rademachers.

Definition 6.5.1. Let $V \subset \mathbb{R}^n$ and $g \sim \mathcal{N}(0, \text{Id})$, then the **Gaussian Complexity** of V is defined as

$$w(V) = \mathbb{E}[\sup_{v \in V} \langle v, g \rangle]$$

This is precisely the same definition as the Rademacher Complexity, but with a Gaussian vector. Likewise, we get the same properties, but with a different constant in the relation with the diameter.

Lemma 6.5.1. *Let $T, S \subseteq \mathbb{R}^n$, $a \in \mathbb{R}$. Then the following are true.*

1. *We have: $w(T + S) = w(T) + w(S)$;*
2. *We also have $w(aT) = |a|w(T)$;*
3. *In particular, we have $w(T - T) = 2w(T)$;*
4. *The relation between the Rademacher Complexity and the Diameter (in the Euclidean norm) is the following:*

$$\frac{1}{\sqrt{2\pi}}\text{diam}(T) \leq w(T) \leq \frac{\sqrt{n}}{2}\text{diam}(T); \text{ and}$$

5. *If $\text{conv}(T)$ denotes the convex hull of T , then*

$$w(\text{conv}(T)) = w(T).$$

Proof. Let us just prove that

$$\frac{1}{\sqrt{2\pi}}\text{diam}(T) \leq w(T).$$

First, note that

$$\begin{aligned} w(T) &= \frac{1}{2} \mathbb{E} \left[\sup_{x, y \in T} \langle g, x - y \rangle \right] \\ &\geq \frac{1}{2} \sup_{x, y \in T} \mathbb{E} [|\langle g, x - y \rangle|]. \end{aligned}$$

Now, we can easily check that

$$\mathbb{E}[|X|] = \sqrt{\frac{2}{\pi}},$$

for $X \sim \mathcal{N}(0, 1)$, hence

$$\begin{aligned} w(T) &\geq \frac{1}{2} \sqrt{\frac{2}{\pi}} \sup_{x, y \in T} \|x - y\| \\ &= \frac{1}{\sqrt{2\pi}} \text{diam}(T), \end{aligned}$$

and it is proved. □

Also, we can bound the quantity $w(V)$ in a similar manner.

Theorem 6.5.1. *Let V be a finite set in \mathbb{R}^n and let g be a standard Gaussian random vector in \mathbb{R}^n . Let $L = \sup_{v \in V} \|v\|$, then*

$$w(V) \leq \sqrt{2L^2 \log |V|},$$

and

$$\mathbb{P} \left(\sup_{v \in V} g_v \geq \sqrt{2L^2 \log |V|} + u \right) \leq e^{-u^2/(2L^2)}.$$

Remark 6.5.1. Even though we bound just

$$\mathbb{P}\left(\sup_{v \in V} g_v \geq \sqrt{2L^2 \log |V|} + u\right) \leq e^{-u^2/(2L^2)},$$

we can get a lower bound as well. Indeed, note that

$$\left| \sup_{v \in V} \langle v, x \rangle - \sup_{v \in V} \langle v, y \rangle \right| \leq L \|x - y\|,$$

hence Corollary 6.4.1 implies that

$$\mathbb{P}(|\sup_{v \in V} \langle v, X \rangle - w(V)| \geq t) \leq 2 \exp\left(-\frac{t^2}{2L^2}\right).$$

Proof. Let $g_v := \langle v, g \rangle$. Then $(g_v)_{v \in V}$ becomes a Stochastic Process. Notice that the function $f(x) = \langle v, x \rangle$ is a Lipschitz function with constant $L_v = \|v\| \leq L$. Then Corollary 6.4.1 implies that

$$\psi_v(\lambda) \leq \frac{\lambda^2 L^2}{2},$$

by Herbst's Method 4.5.6 and $\psi_v(\lambda) := \log \mathbb{E}[e^{\lambda g_v}]$.

Let $\lambda > 0$, then, by Jensen's Inequality to the function $\frac{1}{\lambda} \log x$ the Gaussian Complexity satisfies

$$\begin{aligned} w(V) &= \mathbb{E}\left(\frac{1}{\lambda} \log \exp\left(\lambda \sup_{v \in V} g_v\right)\right) \\ &\leq \frac{1}{\lambda} \log \mathbb{E}\left(\exp\left(\lambda \sup_{v \in V} g_v\right)\right) \\ &\leq \frac{1}{\lambda} \log \sum_{v \in V} \mathbb{E}\left(\exp(\lambda g_v)\right) \\ &\leq \frac{\log |V| + \lambda^2 L^2 / 2}{\lambda}. \end{aligned}$$

Minimizing it over $\lambda > 0$ we get the bound

$$w(V) \leq \sqrt{2L^2 \log |V|}.$$

The bound in the probability of $\sup_{v \in V} g_v$ follows by Chernoff's Inequality 2.7.4, since

$$\begin{aligned} \mathbb{P}\left(\sup_{v \in V} g_v \geq t\right) &= \mathbb{P}\left(\bigcup_{v \in V} g_v \geq t\right) \\ &\leq \sum_{v \in V} \mathbb{P}(g_v \geq t) \\ &\leq \sum_{v \in V} e^{-t^2/(2L^2)} \\ &= e^{\log |V| - t^2/(2L^2)}. \end{aligned}$$

Taking $t = \sqrt{2L^2 \log |V|} + u$, we have

$$\mathbb{P}\left(\sup_{v \in V} g_v \geq \sqrt{2L^2 \log |V|} + u\right) \leq e^{-u^2/(2L^2)},$$

which gives a Gaussian bound to this random variable. □

A simple reason why we want to bound the Gaussian Complexity can be illustrated in the following lemma.

Lemma 6.5.2. *Let $p \in [1, \infty]$ and $X \sim \mathcal{N}(0, \text{Id})$ in \mathbb{R}^n . Let also $q \in [1, \infty]$ be the conjugate exponent of p and $B_q^n = \{x \in \mathbb{R}^n : \|x\|_q \leq 1\}$, then*

$$\mathbb{E}[\|X\|_p] = w(B_q^n).$$

Proof. It is a direct consequence of the duality formula between the norms:

$$\|x\|_p = \sup_{y \in B_q^n} \langle x, y \rangle.$$

□

Finally, we can relate both complexity in the following lemma.

Lemma 6.5.3. *Let $T \subset \mathbb{R}^n$ be a bounded set. Then*

$$\sqrt{\frac{2}{\pi}} r(T) \leq w(T) \leq \lceil 2\sqrt{\log n} \rceil r(T).$$

We will give a simple application of the Gaussian Complexity based on the M^* Bound. Let us first introduce the *Grassmanian* $(G_{n,m}, \mathcal{B}(G_{n,m}), \mu)$.

Definition 6.5.2. Let $G_{n,m}$ be the space of all m -dimensional subspaces of \mathbb{R}^n endowed with the projection distance, that is,

$$d(E, F) := \|P_E - P_F\|,$$

where P_E is the projection onto E and $\|\cdot\|$ is the operator norm. Then $(G_{n,m}, \mathcal{B}(G_{n,m}))$ is the **Grassmanian**.

We can endow the Grasmannian with an uniform measure. This is done in the following theorem.

Theorem 6.5.2. *Let $O(n)$ be the space of all ortogonal operators in \mathbb{R}^n endowed with the Haar measure γ and $E \in G_{n,m}$. Then there is an unique measure μ in the Grassmanian $G_{n,m}$ such that*

$$\mu(A) = \gamma(\{U \in O(n) : O(E) \in A\}),$$

for all $A \in \mathcal{B}(G_{n,m})$.

Now we can state the M^* Bound.

Theorem 6.5.3. Let $K \subset \mathbb{R}^n$ be a bounded subset. Fix $m < n$ and let E be a random element $(G_{n,n-m}, \mathcal{B}(G_{n,n-m}))$, distributed according to the uniform measure in $G_{n,n-m}$. Then

$$\mathbb{E}[\text{diam}(K \cap E)] \leq \frac{Cw(K)}{\sqrt{m}},$$

for some universal constant C .

Proof. We recommend the elegant proof by Vershynnin in [Pfander \(2015\)](#). \square

Remark 6.5.2. We can also prove a concentration inequality, as shown in [Pfander \(2015\)](#):

$$\mathbb{P}\left(\text{diam}(K \cap E) \geq C\frac{w(K)}{\sqrt{m}} + Ct\right) \leq 2\exp\left(-\frac{mt^2}{2\text{diam}^2(K)}\right).$$

Remark 6.5.3. Let $m = 1$ and K be a convex body with barycenter at the origin, then $G_{n,n-1}$ is the space of all hyperplanes in \mathbb{R}^n . The theorem says that, in mean, the size of $K \cap E$ is of order $w(K)$. In other words, almost all the volume of K lies in the set

$$\tilde{K} = K \cap (Cw(K)B_2^n),$$

which is known as the *bulk* of K . The set $K \setminus \tilde{K}$ is the *outliers* of K , carrying *exponential* less volume (see [Ball et al. \(1997\)](#)).

Now we can state and prove one of the main and simple theorems in *Compressive Sensing Theory*. Suppose we want to recover a signal $x \in \mathbb{R}^n$ from a random measurement Ax , which means that A is a random $m \times n$ matrix. We will also suppose that x lies in some set K . The most naive idea is to take any $y \in K$ which satisfies $Ay = Ax$. Can we quantify the error doing this? The following corollary expresses it.

Corollary 6.5.1. Let $x \in K$ and A be a random $m \times n$ matrix such that $A_{ij} \sim \mathcal{N}(0, 1)$ are independent, for all i, j . Set $y = Ax$ and z any solution of the system

$$\begin{cases} Az = y; \\ z \in K. \end{cases}$$

Then we have

$$\mathbb{E}[\|z - x\|] \leq \frac{Cw(K)}{\sqrt{m}}.$$

Proof. Note that

$$\mathbb{E}[\|z - x\|] \leq \mathbb{E}[\text{diam}(K \cap E)],$$

where $E = \ker(A)$. Since the columns of A are standard Gaussian vectors, we can see that E is distributed according to the uniform measure in $G_{n,n-m}$, hence we can apply Theorem 6.5.3 and get

$$\mathbb{E}[\text{diam}(K \cap E)] \leq \frac{Cw(K)}{\sqrt{m}},$$

and the corollary is proved. \square

We can transform this idea into an optimization problem. Suppose K is a symmetric convex body, which means that K is closed, convex, origin symmetric and has nonempty interior. We can define a norm using K .

Definition 6.5.3. Let K be a symmetric convex body. Then the function $\|\cdot\|_K : \mathbb{R}^n \rightarrow \mathbb{R}$, defined as

$$\|x\|_K := \inf\{\lambda > 0 : \lambda^{-1}x \in K\},$$

is known as the **Minkowski functional** of K .

Lemma 6.5.4. *The Minkowski functional is a norm and the unit ball is K :*

$$\{x : \|x\|_K \leq 1\} = K.$$

Now we can transform Corollary 6.5.1 into an optimization problem.

Corollary 6.5.2. *Let K be a symmetric convex body, $x \in K$ and A be a random $m \times n$ matrix such that $A_{ij} \sim \mathcal{N}(0, 1)$ are independent, for all i, j . Consider $Ax = y$ and x^* the solution of the following optimization problem:*

$$\begin{aligned} p^* = \min \quad & \|z\|_K \\ \text{s.t.} \quad & Az = y. \end{aligned}$$

Then $p^* \leq 1$ and

$$\mathbb{E}[\|x^* - x\|] \leq \frac{Cw(K)}{\sqrt{m}}.$$

Proof. Since $x \in K$ and x is a feasible point, we have that $p^* \leq 1$. Now, Since x^* achieves the minimum, we have

$$\|x^*\|_K \leq \|x\|_K \leq 1,$$

hence $x^* \in K$. Therefore the result follows by Corollary 6.5.1. \square

For more applications of Gaussian complexity, see [Vershynin \(2017\)](#) and [Pfander \(2015\)](#).

The next application is about a Statistic Inequality.

6.6 Application III: The Crámer-Rao Inequality

Let X be a centered random variable with variance equals to σ^2 . Suppose we want to estimate a parameter θ from the observation of $Y = X + \theta$. What is the best estimator

$T(Y)$ of θ ? How can we quantify the error? In this section we will try to explore this idea and prove the Crámer-Rao lower bound on the error through Theorem 6.3.1, using the version

$$N(X)J(X) \geq n.$$

In our case, we will consider only linear estimation, which means that, given a sample Y_1, \dots, Y_N , the estimator $T(Y_1, \dots, Y_N)$ is a linear function of Y_1, \dots, Y_N .

Definition 6.6.1. An unbiased estimator T satisfies

$$\mathbb{E}[T(Y_1, \dots, Y_N)] = \theta.$$

Therefore, for an unbiased estimator T , we have

$$T(y_1, \dots, y_N) = \langle v, y \rangle,$$

where $v_1 + \dots + v_N = 1$. For the error, we will consider the variance error

$$e_T := \text{Var}[T(Y_1, \dots, Y_N)].$$

The main theorem of this section is the following.

Theorem 6.6.1. Let X be a random variable with Fisher Information $J(X)$. Let θ be in \mathbb{R} and take a linear unbiased estimator of θ with sample Y_1, \dots, Y_N according to the law of $X + \theta$. Then, the error of estimation is bounded from below:

$$e_T \geq \frac{1}{NJ(X)}.$$

Proof. Because $\text{Var}[Y] = \text{Var}[X] = \sigma^2$, we have

$$e_T = \sigma^2 \sum_{i=1}^n v_i^2 = \sigma^2 \|v\|^2.$$

Now, because Corollary 3.4.2, we have $\sigma^2 \geq N(X)$, therefore

$$e_T \geq N(X) \|v\|^2 \geq \frac{\|v\|^2}{J(X)},$$

where the last inequality is due to Corollary 4.2.1 with $n = 1$. Therefore, the error is bounded away from 0. To minimize over v , notice that

$$\min_{v_1 + \dots + v_N = 1} \|v\|^2 = 1/N,$$

therefore

$$e_T \geq \frac{1}{NJ(X)},$$

and the theorem is proved. □

Remark 6.6.1. In fact, this bound works not only for linear, but for all unbiased estimator. For see this (take $N = 1$ for simplicity), notice that

$$J(X)\text{Var}[T(X + \theta)] = \mathbb{E}\left(\frac{f'(X)}{f(X)}\right)^2 \mathbb{E}\left(T(X + \theta) - \theta\right)^2,$$

where f is the density of X . Applying Cauchy-Schwartz Inequality, we have

$$J(X)\text{Var}[T(Y)] \geq \left(\mathbb{E}\left(\frac{f'(X)}{f(X)}\right)\left(T(X + \theta) - \theta\right)\right)^2.$$

We've already seen in the proof of Theorem 3.6.4 that $\mathbb{E}\left(\frac{f'(X)}{f(X)}\right) = 0$. Finally, notice that

$$1 = \frac{d}{d\theta}\theta = \frac{d}{d\theta}\mathbb{E}[T(X + \theta)] = \frac{d}{d\theta} \int f(x)T(x + \theta) dx.$$

Changing the order and since $\frac{d}{d\theta}T(x + \theta) = \frac{d}{dx}T(x + \theta)$, we have

$$1 = \int_{\mathbb{R}} f(x)T'(x + \theta) dx.$$

Therefore, by the weak derivative property we have

$$1 = \left(\int_{\mathbb{R}} f'(x)T(x + \theta) dx\right)^2,$$

hence

$$J(X)\text{Var}[T(Y)] \geq \left(\int_{\mathbb{R}} f'(x)T(x + \theta) dx\right)^2 = 1,$$

hence the nonlinear case is proved.

Let us look to the Example 6.4.2 with this perspective.

Example 6.6.1. Let $Y = \theta + X$, where X is a standard Gaussian noise. Suppose we want to estimate θ from an independent sample $Y_1, \dots, Y_N \sim Y$. We can look this as a problem of recovering a signal $x \in \mathbb{R}^N$ such that $Y = \text{Id } x + Z$, where Z is a standard Gaussian vector and we know, by the prior information, that $x = (\theta, \dots, \theta)$. The best unbiased estimation is

$$T(Y_1, \dots, Y_N) = \frac{1}{N} \sum_{i=1}^N Y_i,$$

since

$$\text{Var}[T(Y_1, \dots, Y_N)] = \frac{1}{N} \text{Var}[Y_i] = \frac{1}{N} = \frac{1}{NJ(X)},$$

and $J(X) = 1$, by Example 3.4.6. Similarly, according to the perspective of the recovery problem, we have that

$$\theta^* = \text{argmin}_{x=(\theta, \dots, \theta)} \|Y - \text{Id } x\|^2,$$

therefore

$$2 \sum_{i=1}^n (Y_i - \theta^*) = 0,$$

hence

$$\theta^* := T(Y_1, \dots, Y_N) = \frac{1}{N} \sum_{i=1}^n Y_i$$

is also the best solution of the recovery problem.

We can get a higher dimensional version of this result.

Corollary 6.6.1. *Let $Y = \theta + X$, where X is a standard Gaussian noise and a sample $Y_1, \dots, Y_N \sim Y$. Let also*

$$T(Y_1, \dots, Y_N) = \frac{1}{N} \sum_{i=1}^N Y_i,$$

hence

$$\mathbb{P}\left(|T(Y_1, \dots, Y_N) - \theta| \geq t\right) \leq 2 \exp\left(-Nt^2/2\right).$$

Proof. Notice that $T(Y_1, \dots, Y_N)$ is a Gaussian random variable with mean θ and variance $1/N$, which means that $T(Y_1, \dots, Y_N)$ is equal in distribution to

$$f(X) := \theta + \frac{1}{\sqrt{N}}X$$

where X is a standard Gaussian r.v. Since f is a Lipschitz function with constant $1/\sqrt{N}$, we have the desired result applying Theorem 6.4.1. \square

Because of this, to guarantee an error ε , with probability, say, at least 0.99, we just have to take

$$2 \exp\left(-N\varepsilon^2/2\right) \leq 0.01,$$

that is, $N \geq \frac{10}{\varepsilon^2}$ is enough.

Notice that the Crámer-Rao Bound is about the inequality

$$\sigma^2(X) \geq \frac{1}{J(X)},$$

for all random variables with variance σ^2 , therefore, we can strengthen this inequality using the Strong Log-Sobolev Inequality in Theorem 6.3.2 and get the following theorem.

Theorem 6.6.2 (Crámer-Rao's Inequality). *Let X be a random vector in \mathbb{R}^n with covariance matrix $\Sigma(X)$ and Fisher Matrix $\mathbb{J}(X)$, then*

$$|\Sigma(X)| \geq \frac{1}{|\mathbb{J}(X)|}.$$

In fact, we can obtain a Matrix Inequality $\Sigma(X) \succeq \mathbb{J}^{-1}(X)$, where \succeq is the partial order defined by the cone of Positive Semidefinite Matrices. Let us just state this, without proving it.

Theorem 6.6.3. *Let X be a random vector with covariance matrix $\Sigma(X)$ and Fisher Matrix $\mathbb{J}(X)$ and let \succeq be the partial order defined by the Positive Semidefinite Cone of Matrices $n \times n$. Then*

$$\Sigma(X) \succeq \mathbb{J}^{-1}(X).$$

6.7 Application IV: The Uncertainty Principle

Perhaps the most surprising result comes from the relation between the Uncertainty Principle and Corollary 4.2.1.

First, some notation: let

$$L^2 := \{f : \mathbb{R} \rightarrow \mathbb{C} : \|f\|^2 = \int_{\mathbb{R}} |f(x)|^2 dx < \infty\},$$

be the space of square integrable complex value functions. Let $\mathcal{F} : L^2 \rightarrow L^2$ be the Fourier Transform:

$$(\mathcal{F}(f))(u) := \int_{\mathbb{R}} f(x) e^{-2\pi i x u} dx.$$

And let, for simplify the notation, $\hat{f} = \mathcal{F}(f)$.

Definition 6.7.1. Let $\psi \in L^2$, $\|\psi\| = 1$, and X be a random variable with density $|\psi|^2$. Let $\phi = \hat{\psi}$, then $|\phi|^2$ is also a density, say, of Y . We say that X and Y are **associated random variables** and ψ and ϕ are **associated densities**.

The following theorem was proved recently in Dembo (1990) and before in Stam (1959).

Theorem 6.7.1. *Let X and Y be associated random variables with finite variances $\text{Var}(X)$ and $\text{Var}(Y)$. Then*

$$16\pi^2 \text{Var}(X) \text{Var}(Y) \geq 1.$$

Remark 6.7.1. This is known as an **Uncertainty Principle**.

Proof. From Corollary 4.2.1 we obtain

$$\text{Var}(X)J(X) \geq N(X)J(X) \geq 1. \quad (6.4)$$

Let us now compute $J(X)$. Let $u_0 \in \mathbb{R}$, then $|e^{-2\pi i x u_0}| = 1$ for all $x \in \mathbb{R}$, hence

$$J(X) = 4 \int_{\mathbb{R}} \left(\frac{d}{dx} |\psi(x) e^{-2\pi i x u_0}| \right)^2 dx. \quad (6.5)$$

To compute the derivative, notice that for a complex-valued function $f(x) = r e^{i\theta}$ we have

$$f'(x) = r'(x) e^{i\theta} + r i \theta'(x) e^{i\theta} = r' e^{i\theta} + i \theta' f,$$

and

$$\overline{f}'(x) = r' e^{-i\theta} - r i \theta' e^{-i\theta} = r' e^{-i\theta} - i \theta' \overline{f}.$$

Multiplying these equalities, we obtain

$$f'(x) \overline{f}'(x) = (r')^2 + (\theta')^2 |f|^2.$$

Therefore

$$\left(\frac{d}{dx} |f| \right)^2 = \frac{df}{dx} \frac{d\overline{f}}{dx} - |f|^2 \left(\arg f \right)' \leq \frac{df}{dx} \frac{d\overline{f}}{dx}. \quad (6.6)$$

Let $f(x) = \psi(x) e^{-2\pi i x u_0}$, then

$$\left(\frac{d}{dx} |\psi(x) e^{-2\pi i x u_0}| \right)^2 \leq \frac{d\psi(x) e^{-2\pi i x u_0}}{dx} \frac{d\overline{\psi(x) e^{-2\pi i x u_0}}}{dx}.$$

Now, let us take a look at the following expression:

$$\begin{aligned} A &= \int_{\mathbb{R}} (u - u_0)^2 |\phi(u)|^2 du \\ &= \int_{\mathbb{R}} v^2 |\phi(v + u_0)|^2 dv. \end{aligned}$$

The fourth property stated at Lemma 2.6.5 implies that

$$\phi(v + u_0) = \left(\mathcal{F}(e^{-2\pi i x u_0} \psi(x)) \right)(v).$$

Since $f(x) = \psi(x) e^{-2\pi i x u_0}$, we have that

$$A = \int_{\mathbb{R}} \left| v \mathcal{F}[f](v) \right|^2 dv.$$

Because of the tenth property at Lemma 2.6.5, we have that

$$v \mathcal{F}[f](v) = \frac{1}{2\pi i} \mathcal{F}[f'](v),$$

then

$$A = \frac{1}{4\pi^2} \int_{\mathbb{R}} |\mathcal{F}[f'](v)|^2 dv.$$

Since \mathcal{F} is an isometry, we obtain

$$A = \frac{1}{4\pi^2} \int_{\mathbb{R}} |f'(x)|^2 dx = \frac{1}{4\pi^2} \int_{\mathbb{R}} \frac{df}{dx} \frac{d\bar{f}}{dx} dx.$$

Replacing this in Inequality 6.6 and using it in Equation 6.5, we obtain

$$J(X) \leq 16\pi^2 A.$$

Of course, setting $u_0 = \mathbb{E}[Y]$ gives that $A = \text{Var}(Y)$, then

$$J(X) \leq 16\pi^2 \text{Var}(Y).$$

Finally, replacing this bound in Inequality 6.4 gives the uncertainty principle. □

Bibliography

- C. Ané, S. Blachère, D. Chafaï, P. Fougères, I. Gentil, F. Malrieu, C. Roberto, and G. Scheffer. *Sur les inégalités de Sobolev logarithmiques*. Panoramas et Synthèses, 2000.
- D. Bakry. On sobolev and logarithmic sobolev inequalities for markov semigroups. *New trends in stochastic analysis (Charingworth, 1994)*, pages 43–75, 1997.
- K. Ball et al. *An elementary introduction to modern convex geometry*, 1997.
- P. Billingsley. *Convergence of Probability Measures*. John Wiley & Sons, 2013.
- N. Blachman. The convolution inequality for entropy powers. *IEEE Transactions on Information Theory*, 11(2):267–271, 1965.
- S. G. Bobkov and P. Tetali. Modified logarithmic sobolev inequalities in discrete settings. *Journal of Theoretical Probability*, 19(2):289–336, 2006.
- S. G. Bobkov, C. Houdré, and P. Tetali. The subgaussian constant and concentration inequalities. *Israel Journal of Mathematics*, 156(1):255–283, 2006.
- S. Boucheron, G. Lugosi, and P. Massart. *Concentration inequalities: A nonasymptotic theory of independence*. Oxford University Press, 2013.
- S. Boyd and L. Vandenberghe. *Convex Optimization*. Cambridge University Press, 2004.
- P. Brémaud. *Markov chains: Gibbs Fields, Monte Carlo Simulation, and Queues*. Springer, 2013.
- P. Brémaud. *Fourier Analysis and Stochastic Processes*. Springer, 2014.
- L. A. Caffarelli. Monotonicity properties of optimal transportation and the fkg and related inequalities. *Communications in Mathematical Physics*, 214(3):547–563, 2000.

- D. Chafai and J. Lehec. On poincare and logarithmic sobolev inequalities for a class of singular gibbs measures, 2018.
- J. B. Conway. *A Course in Functional Analysis*. Springer, 2010.
- T. M. Cover and J. A. Thomas. *Elements of Information Theory*. John Wiley & Sons, 2012.
- A. Dembo. Information inequalities and uncertainty principles. *Dept. Statistics, Stanford Univ., Stanford, CA, Tech. Rep*, 75, 1990.
- F. Den Hollander. *Large Deviations*. American Mathematical Soc., 2008.
- P. Diaconis, L. Saloff-Coste, et al. Logarithmic sobolev inequalities for finite markov chains. *The Annals of Applied Probability*, 6(3):695–750, 1996.
- M. D. Donsker and S. S. Varadhan. Asymptotic evaluation of certain markov process expectations for large time, i. *Communications on Pure and Applied Mathematics*, 28(1):1–47, 1975.
- P. Dupuis and R. S. Ellis. *A Weak Convergence Approach to the Theory of Large Deviations*. John Wiley & Sons, 2011.
- R. Durrett. *Probability: Theory and Examples*. Cambridge University Press, 2019.
- M. Fathi, N. Gozlan, and M. Prodhomme. A proof of the caffarelli contraction theorem via entropic regularization, 2019.
- G. B. Folland. *Real Analysis: Modern Techniques and Their Applications*. John Wiley & Sons, 2013.
- C. Giraud. *Introduction to High-Dimensional Statistics*. CRC Press, 2014.
- E. F. M. F. L. Gross and C. K. M. R. D. WStroock. *Dirichlet Forms*. Springer, 1993.
- A. Guionnet and B. Zegarlinski. Lectures on logarithmic sobolev inequalities. In *Séminaire de probabilités XXXVI*, pages 1–134. Springer, 2003.
- K. Itô. *Stochastic Processes: lectures given at Aarhus University*. Springer, 2013.
- Y.-H. Kim and E. Milman. A generalization of caffarelli’s contraction theorem via (reverse) heat flow. *Mathematische Annalen*, 354(3):827–862, 2012.
- U. Krengel. *Ergodic Theorems*. Walter de Gruyter, 2011.
- P. D. Lax. *Linear Algebra and its Applications*. 2007. John Wiley & Sons, 2007.

- M. Ledoux. Concentration of measure and logarithmic sobolev inequalities. In *Seminaire de probabilites XXXIII*, pages 120–216. Springer, 1999.
- P. Mörters. Large deviation theory and applications, 2008.
- J. Newman. Ergodic theory for semigroups of markov kernels, 2015.
- A. Pazy. *Semigroups of Linear Operators and Applications to Partial Differential Equations*. Springer, 2012.
- G. E. Pfander. *Sampling Theory, a Renaissance: Compressive Sensing and Other Developments*. Springer, 2015.
- M. Raginsky, I. Sason, et al. Concentration of measure inequalities in information theory, communications, and coding. *Foundations and Trends® in Communications and Information Theory*, 10(1-2):1–246, 2013.
- G. Royer. *An Initiation to Logarithmic Sobolev Inequalities*. American Mathematical Soc., 2007.
- C. E. Shannon. A mathematical theory of communication. *Bell system technical journal*, 27(3):379–423, 1948.
- A. N. Shiryaev. *Probability-1*. Springer, 2016.
- A. J. Stam. Some inequalities satisfied by the quantities of information of fisher and shannon. *Information and Control*, 2(2):101–112, 1959.
- R. van Handel. Probability in high dimension, 2014.
- R. Vershynin. *High-Dimensional Probability*. Cambridge University Press, 2017.
- P. Walters. *An Introduction to Ergodic Theory*. Springer, 2000.