SOME TOPICS IN PROBABILITY THEORY, COMBINATORICS AND INFORMATION THEORY

by

Jiange Li

A dissertation submitted to the Faculty of the University of Delaware in partial fulfillment of the requirements for the degree of Doctor of Philosophy in Mathematics

Spring 2016

SOME TOPICS IN PROBABILITY THEORY, COMBINATORICS AND INFORMATION THEORY

by

Jiange Li

Approved: _____

Louis Rossi, Ph.D.
Chair of the Department of Mathematical Sciences

Approved: _____

George H. Watson, Ph.D.
Dean of the College of Arts & Sciences

Approved: _____

Ann L. Ardis, Ph.D.
Senior Vice Provost for Graduate and Professional Education

I certify that I have read this dissertation and that in my opinion it meets the academic and professional standard required by the University as a dissertation for the degree of Doctor of Philosophy.

Signed: _____

Mokshay Madiman, Ph.D.
Professor in charge of dissertation

I certify that I have read this dissertation and that in my opinion it meets the academic and professional standard required by the University as a dissertation for the degree of Doctor of Philosophy.

Signed: _____

Yuk J. Leung, Ph.D.
Member of dissertation committee

I certify that I have read this dissertation and that in my opinion it meets the academic and professional standard required by the University as a dissertation for the degree of Doctor of Philosophy.

Signed: _____

Nayantara Bhatnagar, Ph.D.
Member of dissertation committee

I certify that I have read this dissertation and that in my opinion it meets the academic and professional standard required by the University as a dissertation for the degree of Doctor of Philosophy.

Signed: _____

Brian Rider, Ph.D.
Member of dissertation committee

ACKNOWLEDGEMENTS

As regards my Ph.D., above all, I would like to thank my advisors, Wenbo Li (deceased) and Mokshay Madiman, for advising me in these years. Without their support, none of the accomplishments in this thesis would be possible. Wenbo was a great probabilist, mentor, and friend. I enjoyed very much the stimulating discussions with him. I am grateful to Mokshay's kindness and support after Wenbo's death. I really benefit a lot from Mokshay's open-mindedness and diverse research interests. This will definitely be one of the big fortunes in my academic career. Also I would like to thank Y. J. Leung and Nayantara Bhatnagar for their help on my research.

I benefited immeasurably from many courses I took. Particularly, I learnt significantly from Sebastian Cioabă, Felix Lazebnik, and Qing Xiang on spectral graph theory and extremal combinatorics. I would like to thank my fellow graduate students Weiqiang (Ricky) Li for getting me interested in discrete mathematics, and Lei Chen for wonderful discussions and big help on numerical simulations. I am also very grateful to the department staff who have been immensely helpful throughout. Also I thank my thesis committee members for reading the draft and making useful suggestions.

Even more influential on my development as a person were my parents, my teachers at school and college, and my relatives and friends. Particularly, I would like to thank Prof. Mingzhu Liu from Harbin Institute of Technology and my long-time friend Chenxi Guo for wonderful discussions far beyond mathematics. It would be silly to even try to adequately acknowledge in words my deep appreciation for my father and my mother for their unconditional support in my life. My sisters, my aunts and uncles were quite as important in shaping me. Almost none of them is a mathematician, but they have that high respect for knowledge even without full comprehension. This thesis is dedicated to all of them.

TABLE OF CONTENTS

LIST OF FIGURES

ABSTRACT

This dissertation explores three topics at the intersection of probability theory, combinatorics and information theory. The first part focuses on studying small ball inequalities for sums and differences of independent, identically distributed random variables taking values in very general sets. Depending on the setting (abelian or non-abelian groups, or vector spaces, or Banach spaces) we provide a collection of inequalities relating different small ball probabilities that are sharp in many cases of interest. We show that underlying these distribution-free probabilistic inequalities are inequalities of extremal combinatorial nature, related among other things to classical packing problems such as the kissing number problem. As regards applications, we develop various moment inequalities.

The second part is devoted to exploring a formal parallel relation between entropy inequalities in information theory and sumset estimates in additive combinatorics. Our work is closely related to the study of more-sum-than-difference sets in additive combinatorics. Various information theoretical inequalities are obtained, such as the entropy analogue of Freiman-Pigarev inequality. We also present applications of our results in the construction of polar codes with significantly improved error probability compared to the canonical construction.

Concentration of measure principle is one of the cornerstones in geometric functional analysis and probability theory, and it is widely used in many other areas. In the third part, we study the concentration property of information content, which is one of the central interests in information theory, and it has great relevance with various other areas such as probability theory, statistics and statistical physics. Sharp exponential deviation estimates for the information content as well as a sharp bound on the varentropy for convex probability measures are obtained on Euclidean spaces.

Chapter 1

INTRODUCTION

This dissertation explores three topics: small ball inequalities in probability theory, information theoretical inequalities analogous to sumset estimates in additive combinatorics, and concentration properties of information content for convex probability measures. In this chapter, we give basic introduction of these topics and demonstrate some motivations for our study. Detailed results will be discussed in the following three chapters.

1.1 Small ball inequalities

Both the theory of large deviations (or tail probabilities) and the theory of small deviations (or small ball probabilities) study the occurrence of rare events. The theory of large deviation studies the asymptotic behavior of the probability that a random variable is far away from its mean, i.e.

$$\mathbb{P}(\|X - \mathbb{E}X\| > t)$$

as $t \to \infty$. While the small ball deviation theory seeks to control the probability that a random variable is very small, i.e.

$$\mathbb{P}(\|X\| < \epsilon)$$

as $\epsilon \to 0$. The theory of large deviations goes back the study of Cramér about actuarial "ruin problems", and it culminates in Varadhan's landmark paper [160]. The large deviation principle has been well developed in the last few decades, see e.g. Donsker and Varadhan [39, 40, 41] for Markov processes, Ledoux and Talagrand [96], Ledoux [94] and Bogachev [19] for Gaussian measures, Varadhan [161], Dembo and Zeitouni

[34] for the general theory. Small ball probabilities have been extensively studied in the setting of Gaussian processes and associated Banach or Hilbert spaces. It has been found that the small ball estimate has close connections with various approximation quantities of compact sets and operators [134, 90, 104], and has a variety of applications in studies of fractal properties of random sets [169], rates of convergence in Strassen's law of the iterated logarithm [156, 91]. A nice exposition of the state of the art in the theory of small deviations can be found in Li and Shao [105].

Motivation and goal. Given the ubiquity of sums of independent, identically distributed (i.i.d.) random variables in probability theory, it is natural to look for ways to estimate the probability that their sum lies in a given measurable set. If the measurable set is selected to be a normed ball, we are actually studying small ball probabilities, although the normed ball is not necessary to be small. In general, this can be a rather complex calculation, and is often intractable. The *raison d'etre* of the first part is the fact that it is often much easier to estimate the probability that a symmetric random variable lies in a symmetric set; so if we can find a way to relate the desired probability to a probability of this type, then in many circumstances our task is significantly simplified.

The most general setting in which we can talk about sums (and symmetry) is that of group-valued random variables, where the group operation represents summation. To state our problem more precisely, consider i.i.d. random variables X, Y taking values in a (possibly non-abelian) topological group with group operation "$+$" and the Borel σ-algebra generated by all open sets; then our problem is to find good bounds on $\mathbb{P}(X+Y \in F)$ for arbitrary measurable set F in terms of $\mathbb{P}(X-Y \in K)$ for symmetric measurable sets K. Since the distribution of $X-Y$ is always symmetric, this would provide a reduction of the form mentioned earlier. We also study a related problem, namely that of estimating $\mathbb{P}(X-Y \in F)$ for arbitrary measurable F in terms of $\mathbb{P}(X-Y \in K)$ for symmetric measurable sets K.

It might seem that the problem stated is somewhat abstruse; however, it is

closely related to a number of influential streams of recent research. To highlight these connections, we discuss the problem from various perspectives.

Symmetrization. Symmetrization is one of the most basic and powerful meta-techniques that arises in many different guises in different parts of mathematics. Instances include Steiner symmetrization in convex geometry and the study of isoperimetric phenomena, Rademacher symmetrization in empirical process theory, use of rearrangements in functional inequalities and the study of partial differential equations, and others too numerous to mention. One goal of this part is to develop a symmetrization technique for estimating small ball probabilities of sums and differences of i.i.d. random variables. We call these small ball probabilities even though there may be no "ball" under consideration (for instance, no norm in the general group settings that we will consider), because when considered in the context of finite dimensional vector spaces, these are related to inequalities for the probability of lying in a ball with respect to some norm.

Concentration function. The notion of the concentration function was introduced by Lévy, as a means of describing in a flexible way the spread or concentration of a real-valued random variable that may not have finite moments. For a real-valued random variable X with distribution P_X, the Lévy's concentration function is given by $Q(X, s) = \sup_{x \in \mathbb{R}} P_X([x, x + s]))$ for $s > 0$. While there was already much attention paid to concentration functions in classical probability theory (see, e.g., [99, 37, 88, 137, 48, 49, 84, 70, 74, 123]), their study received renewed attention in recent years [35, 36, 139, 57, 45, 128] because of the relevance of arithmetic structure to the concentration function of linear combinations of i.i.d. random variables, as well as applications to random matrix theory. While we do not directly address the literature on concentration functions in our note, our results may be seen as providing bounds on multidimensional or non-Euclidean analogs of concentration functions in general spaces. Indeed, a natural way to define the concentration function in a general setting,

say an abelian group G, would be to set

$$Q(X, F) = \sup_{x \in G} P_X(x + F),$$

where the set-valued parameter F plays the role of the parameter s in the definition $Q(X, s)$ of the concentration function for real-valued random variables. Since the constants that appear in our results are covering/packing numbers $N(F, K)$ that are invariant with respect to translations of F, our results directly imply concentration function bounds; for instance, Theorem 2.3.1 implies that for F an arbitrary measurable subset of an abelian topological group G and K a measurable subset of G containing the identity in its interior,

$$Q(X + Y, F) \leq N(F, K) \cdot Q(X - Y, K).$$

Packing problems/Extremal combinatorics. In 1995, Alon and Yuster [4] showed that for any two i.i.d. real-valued random variables X, Y,

$$\mathbb{P}(|X - Y| \leq b) < (2\lceil b/a \rceil - 1) \cdot \mathbb{P}(|X - Y| \leq a), \tag{1.1}$$

thus answering a question raised by Peres and Margulis. They also observed that the optimal constants in such estimates are closely related to the kissing number problem, which is a long-standing problem in geometry; indeed, the kissing number in $\mathbb{R}^3$ was a subject of discussion between Isaac Newton and David Gregory in 1694. A similar probabilistic inequality proved by Katona [76] is closely related to Turán-type theorems for triangle-free graphs. It turns out that behind the main results of this paper, which among other things generalize significantly the inequality (1.1) of [4], are actually statements from extremal combinatorics, which we prove en route to proving our main results. This strengthens the link between extremal combinatorial phenomena and probabilistic inequalities, in a much more general setting than that of [4], in analogy with similar links developed by Katona in a series of papers (see, e.g., [75, 83]).

Moment inequalities. Probability bounds are of course closely related to moment inequalities, and our results in particular can be used to develop a number of moment

inequalities for functions of sums and differences of random variables under various assumptions on the distribution and/or the function. Such inequalities are of intrinsic interest since they serve as tools in a variety of areas.

Random walks. For $0 < a < 2b$, the following sharp symmetrization inequality for i.i.d. real-valued random variables X, Y is proved in [38]:

$$\mathbb{P}(|X+Y| \leq b) < \lceil 2b/a \rceil \cdot \mathbb{P}(|X-Y| \leq a). \tag{1.2}$$

For $a \geq 2b$, the estimate still holds with " $\leq$ " in the middle. This generalizes the early work of Schultze and von Weizsächer [148], which considered the special case $a = b$ and used it as a key ingredient in studying the level crossing probabilities for random walks on the real line. The first part contains those of [38], and although we do not investigate this direction further here; it is likely that our results would be useful in the study of random walks on groups.

1.2 Information theoretical inequalities

Classical information theoretical inequalities have been driven to solve communication theoretical problems. Nowadays information theory is no longer restricted to the domain of communication theory. Information theoretical inequalities also play important roles in many other areas, such as probability theory, convex geometry and combinatorics. The second part of this dissertation studies certain entropy inequalities analogous to some sumset estimates in additive combinatorics.

Introduction to entropy. Entropy made its first appearance in the middle of the 19th century in the context of thermodynamics. It was introduced by Clausius in 1865 as a macroscopic description of a thermodynamic system. Later, Boltzmann in 1877 developed a statistical mechanical interpretation of entropy as a measure of uncertainty or disorderedness of a system. It is proportional to the natural logarithm of the number of possible microscopic states, which gives rise to the observed macroscopic state of the system. Boltzmann entropy is now regarded as one of the cornerstones of statistical mechanics. A statistical concept of entropy called Shannon entropy was

introduced by Claude Shannon in his seminal paper [146] to study the communication and transmission of information. According to the folklore Tribus and MacIrvine [159] [1] , the term entropy was suggested to Shannon by von Neumann for both its fuzziness and resemblance with Boltzmann entropy. In nowadays, entropy is a fundamental concept in many disciplines, such as probability theory, information theory, statistical mechanics, dynamical systems and computer science, etc.

Let X be a random variable taking values in a finite set A with probability mass function $p(x)$ for $x \in A$. Its *Shannon entropy* $H(X)$ is defined as

$$H(X) = -\sum_{x \in A} p(x) \log p(x) = -\mathbb{E} \log p(X). \tag{1.3}$$

Here we adopt the usual abuse of notation: we write $H(X)$ even though the entropy is a functional depending only on the distribution of X and not on the value of X. Shannon entropy measures the uncertainty of a distribution or the average missing information from a random source. One basic fact implied by the concavity of $-x \log x$ is that

$$0 \leq H(X) \leq \log |A|, \tag{1.4}$$

where $|A|$ denotes the cardinality of A. Equality in the lower bound holds if only if X is deterministic, in which there is no uncertainty. Equality in the upper bound holds if only if X is uniform on A, in which case we have the largest uncertainty. For any discrete random variables X, Y, we have the following sub-additive property

$$H(X \pm Y) \leq H(X) + H(Y). \tag{1.5}$$

[1] When John von Neumann asked him how he was getting on with his information theory, Shannon replied: "The theory was in excellent shape. My greatest concern was what to call it. I thought of calling it 'information', but the word was overly used, so I decided to call it 'uncertainty'." John von Neumann told him, "You should call it entropy, for two reasons. In the first place your uncertainty function has been used in statistical mechanics under that name, so it already has a name. In the second place, and more important, nobody knows what entropy really is, so in a debate you will always have the advantage".

If X, Y are independent, we have

$$\max\{H(X), H(Y)\} \leq H(X \pm Y). \tag{1.6}$$

The continuous analogous of Shannon entropy is called differential entropy, which is defined for continuous random variables. For a real-valued continuous random variable X with density function $f(x)$, its *differential entropy* $h(X)$ is defined as

$$h(X) = -\int_{\mathbb{R}} f(x) \log f(x) dx = -\mathbb{E} \log f(X). \tag{1.7}$$

Unlike Shannon entropy, differential entropy could be negative even $-\infty$. The lower bound (1.6) does generalize for differential entropy, but the sub-additive property (1.5) in general fails for differential entropy. Shannon entropy and differential entropy have a lot of properties in common. But there is indeed a lot of properties, which hold for Shannon entropy, fails for differential entropy. That is one of the reasons why differential entropy is relatively harder to handle.

Information theoretical interpretation. Let X be a discrete random variables taking values in a finite set A with probability mass function $p(x)$ for $x \in A$. Let $X_1, \cdots, X_n$ be a sequence of independent copies of X. Without confusion, we use $p(x_1, \cdots, x_n)$ to denote the joint probability mass function of this sequence. By the weak law of large number, we have

$$-\frac{\log p(X_1, \cdots, X_n)}{n} \xrightarrow{p} H(X). \tag{1.8}$$

For small $\epsilon > 0$, we introduce the *typical set* $T_{n,\epsilon}(X)$ defined by

$$T_{n,\epsilon}(X) = \{(x_1, \cdots, x_n) \in A^n : e^{-n(H(X)+\epsilon)} \leq p(x_1, \cdots, x_n) \leq e^{-n(H(X)-\epsilon)}\}, \tag{1.9}$$

where we denote by A^n the n-th Cartesian product of A. For large n, from (1.8) we can see that

$$\mathbb{P}(T_{n,\epsilon}(X)) = 1 - o_n(1). \tag{1.10}$$

Estimate of the $o(1)$ term can follow from Hoeffding's inequality. Using the definition of $T_{n,\epsilon}(X)$, we have

$$|T_{n,\epsilon}(X)| = e^{n(H(X)+o(1))}. \tag{1.11}$$

The concentration property (1.10) is a special case of *Asymptotic Equipartition Property* (AEP) for i.i.d. random variables. For a discrete-time stationary ergodic process, AEP is known as *Shannon-McMillan-Breiman theorem*. We refer to [9] for a definitive version of this theorem, as well as for a nice account of the history of it. The notion of typical set and AEP plays an important role in coding theory [33].

Then we can see that after a long run the sequence $X_1, \cdots, X_n$ is roughly uniformly distributed on a set with probability close to 1 and approximately $e^{nH(X)}$ elements. This gives the traditional information theoretical interpretation of Shannon entropy $H(X)$ as the measure of the logarithm of the effective support of a large sample of X. It is analogous to the definition of entropy in thermodynamics. In the continuous setting, the typical set $T_{n,\epsilon}(X)$ is defined in a similar way, with the probability mass function $p(x_1, \cdots, x_n)$ replaced by the joint density function $f(x_1, \cdots, x_n)$. The concentration property (1.10) still holds. The quantity $|T_{n,\epsilon}(X)|$ in equation (1.11) will be interpreted as the volume of $T_{n,\epsilon}(X)$ and the Shannon entropy $H(X)$ will be replaced by the differential entropy $h(X)$.

Formal parallel relation. The link between random variables and typical sets suggests a formal parallel relation between entropy inequalities in information theory and sumset estimates in additive combinatorics (and convex geometry): replace sets by random variables, and replace the logarithm of cardinality (volume) of each set by the entropy of the corresponding random variable. First identified by Ruzsa [143], this connection has been studied extensively in the last few years. Useful tools in additive combinatorics have been developed in the entropy setting, such as Plünnecke-Ruzsa inequalities by Madiman, Marcus and Tetali [110], and Freiman-Green-Ruzsa and Balog-Szemerédi-Gowers theorems by Tao [158]. Much more work has also recently emerged on related topics, such as efforts towards an entropy version of the Cauchy-Davenport

inequality [64, 71, 164, 166], an entropy analogue of the doubling-difference inequality [109], extensions from discrete groups to locally compact abelian groups [89, 108], and applications of additive combinatorics in information theory [92, 107, 29, 50, 167].

In the following we demonstrate this relation by some examples. For two finite subsets A, B of an additive group, the sumset $A+B$ and difference set $A-B$ are defined by

$$A+B := \{a+b : a \in A, b \in B\},$$

and

$$A-B := \{a-b : a \in A, b \in B\}.$$

Inequalities (1.5) and (1.6) are exactly entropy analogs of the following trivial bounds

$$\max\{|A|, |B|\} \leq |A \pm B| \leq |A||B|.$$

This is, of course, an analogy but not a proof. Another typical sumset estimate in additive combinatorics is Ruzsa's triangle inequality [142], which says that

$$|A-C| \leq \frac{|A-B||B-C|}{|B|}. \tag{1.12}$$

Its entropy analog [158] asserts that for independent discrete random varialbes X, Y, Z, we have

$$H(X-Z) \leq H(X-Y) + H(Y-Z) - H(Z). \tag{1.13}$$

Its ifferential entropy analog is proved in [89]. The famous *entropy power inequality* (EPI) [146, 150] asserts that for independent continuous random variables X, Y in $\mathbb{R}^n$, we have

$$e^{\frac{2}{n}h(X+Y)} \geq e^{\frac{2}{n}h(X)} + e^{\frac{2}{n}h(Y)}. \tag{1.14}$$

It has a strong formal resemblance to the well-known Brunn-Minkowski inequality in convex geometry. It says that for non-empty compact Borel subsets $A, B \subset \mathbb{R}^n$, we have

$$|A+B|^{1/n} \geq |A|^{1/n} + |B|^{1/n}. \tag{1.15}$$

Here we denote by $|\cdot|$ the Lebesgure measure. We refer to [153] for the derivation of EPI from Brunn-Minkowski inequality for restricted sumsets.

Our motivation. In an abelian group, since addition is commutative while subtraction is not, two generic elements generate one sum but two differences. Likely motivated by this observation, J. H. Conway had posed the following conjecture (contained in H. T. Croft's Research Problems, 1967):

> "Let $A = \{a_1, a_2, \ldots, a_N\}$ be a finite set of integers, and define $A + A = \{a_i + a_j : 1 \leq i, j \leq N\}$ and $A - A = \{a_i - a_j : 1 \leq i, j \leq N\}$. Prove that $A - A$ always has more members than $A + A$, unless A is symmetric about 0."

However, that is not always the case. In 1969, Marica [113] showed that the conjecture is false by exhibiting the set $A = \{1, 2, 3, 5, 8, 9, 13, 15, 16\}$, for which $A + A$ has 30 elements and $A - A$ has 29 elements. (Conway himself is also said to have found the counter example $\{0, 2, 3, 4, 7, 11, 12, 14\}$ in the 1960's, thus disproving his own conjecture– some history about more-sum-than-difference (MSTD) sets is discussed in [127, 126].) Subsequently, Stein [152] showed that one can construct sets A for which the ratio $|A - A|/|A + A|$ is as close to 0 or as large as we please; apart from his own proof, he observed that such constructions also follow by adapting arguments in an earlier work of Piccard [132] that focused on the Lebesgue measure of $A + A$ and $A - A$ for subsets A of $\mathbb{R}$. A stream of recent papers aims to quantify how rare or frequent MSTD sets are (see, e.g., [114, 69] for work on the integers, and [171] for finite abelian groups more generally), or try to provide denser constructions of infinite families of MSTD sets (see, e.g., [118, 170]); however these are not directions we will explore in this part.

Since convolutions of uniforms are always distributed on the sumset of the supports, but are typically not uniform distributions, it is not immediately obvious from the Conway and Marica constructions whether there exist i.i.d. random variables X and Y such that $H(X+Y) > H(X-Y)$. The purpose of the second part is to explore this sum-difference problem for entropy. A natural related question to ask is for some

description of the coefficient λ that maximizes $H(X+\lambda Y)$ for i.i.d. random variables X, Y taking values in cyclic groups; restricting the choice of coefficients to $\{+1, -1\}$ would correspond to the sum-difference question. This question is motivated by applications to the class of polar codes, which is a very promising class of codes that has attracted much recent attention in information and coding theory.

1.3 Concentration of information content

Information content is one of the central interests in information and coding theory. It also has important relevance with other areas, such as probability theory, statistical physics and statistics. The third part of the thesis devotes to the study of concentration properties of information content for convex probability measures.

Concentration of measure principle. The concentration of measure phenomenon roughly says if a function depends in a reasonably continuous way on a large number of small variables, then it is almost always close to its expected value. This idea goes back to the work of Lévy [100] on the spherical isoperimetric problem. But its full strength was first realized by Milman in his revolutionary proof [119] of Dvoretzky's theorem [42]. Specially the proof is a milestone in the local theory of Banach spaces. This concentration principle is responsible for many counterintuitive phenomenons in high dimensional spaces. One simple example is that the volume of the unit ball goes to 0 when the dimension increases to infinity. It also leads to new understandings of some traditional probabilistic conditions, such as independence and martingale. This principle is one of the cornerstones in geometric functional analysis and probability theory, and it is widely used in many other areas. This concentration phenomenon has been extensively studied in the last several decades by Milman [119, 120], Gromov and Milman [60, 61], Milman and Shechtman [122], Maurey [116], Pisier [134], Shechtman [144], Talagrand [154, 155, 157], Ledoux [95], and others.

Information content. Let $\mathbb{X} = (X_1, X_2, \cdots)$ be a stochastic process with each X_i taking values in $\mathbb{R}$. Suppose that the joint distribution of $X^n = (X_1, \cdots, X_n)$ has

a density f with respect to the Lebesgue measure on $\mathbb{R}^n$. We are interested in the random variable

$$\widetilde{h}(X^n) = -\log f(X^n), \tag{1.16}$$

which may be thought of as the (random) *information content* of X^n. In the discrete case, the quantity $\widetilde{h}(X^n)$ (using f for the probability mass function) is essentially the number of bits needed to represent X^n by a coding scheme that minimizes the average code length [146]. In the continuous case, one may still call $\widetilde{h}(X^n)$ the information content even though this coding interpretation no longer holds. The quantity $\widetilde{h}(X^n)$ is of central interest in information theory and also naturally arises in several other areas such as probability theory, statistical physics and statistics. The average value of information content is known more commonly as the (differential) entropy defined by

$$h(X^n) = -\int_{\mathbb{R}^n} f(x)\log f(x)dx = \mathbb{E}\widetilde{h}(X^n). \tag{1.17}$$

Background. A typical problem is to study the deviation estimate of information content from the entropy, either through the *varentropy*, which is defined as the variance of $\widetilde{h}(X^n)$, or through deviation inequalities for the random variable $\widetilde{h}(X^n)$. The *entropy rate* of the stochastic process $\mathbb{X}$ is defined by

$$h(\mathbb{X}) = \lim_{n\to\infty} \frac{h(X^n)}{n}, \tag{1.18}$$

when the limit exists. The question of whether the information content per coordinate $\frac{\widetilde{h}(X^n)}{n}$ converges to the entropy rate has been extensively studied. In the discrete case, the affirmative answer goes back to Shannon [146], McMillan [117] and Breiman [25] for stationary ergodic process. The theorem particularly implies the existence of a set of roughly $e^{nh(\mathbb{X})}$ typical sequences of length n all having roughly equal probability (a fact that plays a fundamental role in compressing discrete data from ergodic sources). McMillan [117] called this the asymptotic equipartition property (AEP). Extensions

to more general (including continuous) settings were obtained independently in [9] and [131].

For processes that are not asymptotically mean stationary, the entropy rate typically does not exist; so there is no convergence question of $\frac{\widetilde{h}(X^n)}{n}$. With a global restriction on the joint distribution of the process, namely log-concavity, but without assuming an asymptotic framework (i.e., for a density on $\mathbb{R}^n$ for fixed n), [14] proved that $\frac{\widetilde{h}(X^n)}{n}$ is highly concentrated around the entropy rate. It demonstrates that high-dimensional log-concave measures are in a sense close to uniform distributions on the annulus between two nested convex sets. The argument of [14] is non-trivial and depends on the rather heavy machinery of the so-called Lovász-Simonovits localization technique; however, optimal concentration bounds were recently obtained in [53] using a much simpler approach.

Our goal. The purpose of this part is to extend the concentration property of the information content from log-concave measures to the more general class of "convex measures". The class of convex measures (which we define more carefully in Section 4.3) includes all probability distributions with quasiconcave densities, i.e., densities such that the value at a convex combination of two points is at least the minimum of the values at the two points. In particular, these include all log-concave densities (such as Gaussians and exponentials) as well as Pareto distributions (only some of whose moments are finite) and the Cauchy distribution (whose mean does not exist). Perhaps most importantly from the information theory point of view, we expect our results to have implications for the study of fundamental limits of finite-blocklength performance in contexts involving convexity, e.g., for additive noise channels where the noise is drawn from a convex measure. Let us note that a very special case of our results, namely information concentration for Gaussians (which can be proved by explicit computations), is a key ingredient in the results of Cover and Pombra [32] on feedback capacities of Gaussian channels. This is because bounds on concentration of information content are often useful in obtaining concentration inequalities for the

Pinsker information density (which plays a key role in finite-blocklength analysis of communication channels); this is laid out in the log-concave case in [16]. Developing this direction and the resulting applications to communications, however, requires additional work and we do not attempt it in this part.

Our study is also closely related to many interesting problems in convex geometry and probability theory. For example, Corollary 4.3.4 was used in [15] (for log-concave measures) to give an information-theoretic formulation of Bourgain's famous hyperplane conjecture [23]. A weaker form of Corollary 4.3.5 and Corollary 4.3.4 are key ingredients used by [18] to obtain a reverse entropy power inequality for convex measures; and the log-concave case of our main result also has applications in random matrix theory [112].

Chapter 2

SMALL BALL INEQUALITIES

In this chapter, we will develop a symmetrization technique for estimating small ball probabilities for sums and differences of i.i.d. random variables. Let X, Y be i.i.d. random variables taking values in certain measurable space, and let F, K be two measurable subsets. More precisely we are looking for the smallest possible constants c_+, c_- such that

$$\mathbb{P}(X \pm Y \in F) \leq c_{\pm} \cdot \mathbb{P}(X - Y \in K) \tag{2.1}$$

hold for all i.i.d. random variables. As mentioned before, this problem is related to several other influential research streams. Depending on the space where X, Y take values, various small ball inequalities are obtained. Estimates for real-valued random variables are provided in Section 2.1. Extensions are made in Section 2.2 and Section 2.3 for random variables taking values in separable Banach spaces and general topological groups, respectively. We will discuss the tightness problem for various estimates in Section 2.4. Regarding applications, various moment inequalities are obtained in Section 2.5. Estimates in Section 2.1 and Section 2.2 can be found in [38], and results in the other three sections are from [101].

2.1 Small ball inequalities for real-valued random variables

In this section, we assume that X, Y are real-valued i.i.d. random variables. Margulis raised the problem of determining the smallest possible constant c such that the following inequality holds for any real-valued i.i.d. random variables X, Y,

$$\mathbb{P}(|X - Y| \leq 2) \leq c \cdot \mathbb{P}(|X - Y| \leq 1). \tag{2.2}$$

Since X, Y have the same distribution, their difference $X - Y$ is symmetric and has positive probability concentrating around 0, this provides a reason for the possible existence of such a distribution-free constant. Many researchers observed that the optimal constant should satisfy $3 \leq c \leq 5$. The lower bound follows by considering the example that X, Y are independent and uniformly distributed on $\{2, 4, \cdots, 2n\}$. In this case, we have $\mathbb{P}(|X - Y| \leq 1) = 1/n$ and $\mathbb{P}(|X - Y| \leq 2) = 3/n - 2/n^2$. The optimal constant $c = 3$ was obtained by Alon and Yuster [4], and independently by Kotlov [1]. Using a combinatorial approach, Alon and Yuster [4] actually proved the following general result.

Theorem 2.1.1 (Alon and Yuster [4]). *Let a, b be two positive numbers. For any real-valued i.i.d. random variables X, Y, we have*

$$\mathbb{P}(|X - Y| \leq b) < (2\lceil b/a \rceil - 1) \cdot \mathbb{P}(|X - Y| \leq a). \tag{2.3}$$

Moreover the constant $2\lceil b/a \rceil - 1$ cannot be improved.

During the study of level crossing probabilities for random walks with symmetric independent increments, Schultze and von Weizsächer [148] obtained a similar sharp inequality: for any real-valued i.i.d. random variables X, Y, we have

$$\mathbb{P}(|X + Y| \leq 1) < 2 \cdot \mathbb{P}(|X - Y| \leq 1). \tag{2.4}$$

This estimate plays an important role in removing the symmetry assumption. Their proof depends the following key lemma, which shows how to derive two-variable integral inequalities from one-variable integral estimate. We state the lemma in a form suitable for our purpose.

Lemma 2.1.1 (Schultze and von Weizsächer [148]). *Let $(\mathcal{X}, \mathcal{B})$ be a measurable space and $f : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ be a $\mathcal{B} \otimes \mathcal{B}$ measurable bounded symmetric function. Let $\mathcal{P}$ be the set of all probability measures on $\mathcal{B}$. Then the following statements are equivalent:*

[1] The authors of [4] learnt from Y. Peres that A. Kotlov also obtained this result without strict inequality. But we can not find the reference.

- *For all* $\mu \in \mathcal{P}$,
$$\int_{\mathcal{X}\times\mathcal{X}} f(x,y)d\mu(x)d\mu(y) > 0.$$
- *For all* $\mu \in \mathcal{P}$,
$$\mu\left(\left\{x \in \mathcal{X} : \int_{\mathcal{X}} f(x,y)d\mu(y) > 0\right\}\right) > 0.$$

Using this lemma, we can obtain the following extension of (2.4).

Theorem 2.1.2. *Let* a, b *be positive real numbers such that* $0 < a < 2b$. *For any real-valued i.i.d. random variables* X, Y, *we have*

$$\mathbb{P}(|X+Y| \le b) < \lceil 2b/a \rceil \cdot \mathbb{P}(|X-Y| \le a). \tag{2.5}$$

Moreover the constant $\lceil 2b/a \rceil$ *can not be improved. When* $a \ge 2b$, *the inequality is still tight with* " $\le$ " *in the middle.*

Remark. When $a \ge 2b$, equality can indeed happen. To see that we can take X, Y to be independent random variables with the same distribution $\mathbb{P}(X = 0) = 1$. In this case, it is easy to see $\mathbb{P}(|X+Y| \le b) = \mathbb{P}(|X-Y| \le a) = 1$.

Proof. Upper bound: Without loss of generality, we can assume $a = 1$. Then we can rewrite the inequality (2.5) as

$$\int_{\mathbb{R}\times\mathbb{R}} \varphi(x,y)d\mu(x)d\mu(y) > 0,$$

where

$$\varphi(x,y) = \lceil 2b \rceil \cdot 1_{\{(x,y):|x-y|\le 1\}} - 1_{\{(x,y):|x+y|\le b\}},$$

and μ is the probability measure induced by X. It is easy to see that $\varphi(x,y)$ is bounded and symmetric with respect to x, y. By Lemma 2.1.1, it is equivalent to prove

$$\mu\left(\left\{x \in \mathbb{R} : \int_{\mathbb{R}} \varphi(x,y)d\mu(y) > 0\right\}\right) > 0$$

for all $\mu \in \mathcal{P}$. We define $\mu_r(x) := \mu([x-r, x+r])$. Then the above statement can be rewritten as

$$\mu(\{x \in \mathbb{R} : \mu_b(-x) < \lceil 2b \rceil \cdot \mu_1(x)\}) > 0.$$

Suppose that the claim is not true, then there is some μ such that $\mu(S) = 1$, where

$$S = \{x \in \mathbb{R} : \mu_b(-x) \geq \lceil 2b \rceil \cdot \mu_1(x)\}. \tag{2.6}$$

Define $\alpha = \sup_{x \in S} \mu_1(x)$, which is positive. For $\epsilon > 0$ small, we will show that there exists a sequence of disjoint intervals $\{I_k\}$ such that $\mu(I_k) > \alpha - \lceil 2b \rceil^{2k} \epsilon$. For M large enough, we have

$$\mu\left(\cup_{k=0}^{M} I_k\right) > \sum_{k=0}^{M} (\alpha - \lceil 2b \rceil^{2k} \epsilon) > 1,$$

which is impossible. So the claim must be true. Firstly, we can pick $x_0 \in S$ such that $\mu_1(x_0) > \alpha - \epsilon$, and define

$$I_0 = [x_0 - 1, x_0 + 1]. \tag{2.7}$$

Since $x_0 \in S$, we have $\mu_b(-x_0) > \lceil 2b \rceil (\alpha - \epsilon)$. Without loss of generality, we assume $x_0 \geq 0$. It is easy to see that $[-x_0 - b, -x_0 + b]$ can be divided into $\lceil 2b \rceil$ disjoint intervals of the form

$$[-x_0 + b - 1, -x_0 + b], [-x_0 + b - 2, -x_0 + b - 1), \cdots, [-x_0 - b, -x_0 + b + 1 - \lceil 2b \rceil).$$

Due to $\mu(S) = 1$, the interval above with positive measure must have non-empty intersection with S. So it can be covered by $[y - 1, y + 1]$ for some $y \in S$. Then we can see that every interval above has measure at most α, which implies

$$\mu\left([-x_0 - b, -x_0 + b + 1 - \lceil 2b \rceil)\right) > \lceil 2b \rceil (\alpha - \epsilon) - (\lceil 2b \rceil - 1)\alpha = \alpha - \lceil 2b \rceil \epsilon.$$

For any $x_1 \in [-x_0 - b, -x_0 + b + 1 - \lceil 2b \rceil) \cap S$, we have $\mu_1(x_1) > \alpha - \lceil 2b \rceil \epsilon$ and

$$\mu_b(-x_1) > \lceil 2b \rceil (\alpha - \lceil 2b \rceil \epsilon). \tag{2.8}$$

When $b > 1/2$, we always have

$$-x_1 + b > x_0 + \lceil 2b \rceil - 1 > x_0 + 1. \tag{2.9}$$

For $1/2 < b \leq 1$, we can see

$$x_0 \geq -x_1 - b > x_0 + \lceil 2b \rceil - 2b - 1 \geq x_0 - 1. \tag{2.10}$$

Combining (2.8), (2.9) and (2.10), we have

$$\mu((x_0+1,-x_1+b])\geq \mu_b(-x_1)-\mu_1(x_0)>\alpha-\lceil 2b\rceil^2\epsilon.$$

For $b>1$, we have

$$-x_1-b-1+\lceil 2b\rceil > x_0+2(\lceil 2b\rceil-b-1)\geq x_0+1.$$

In this case, we also have

$$\mu((-x_1-b-1+\lceil 2b\rceil,-x_1+b])\geq \alpha-\lceil 2b\rceil^2\epsilon.$$

Hence, we can define

$$I_1=\begin{cases}(x_0+1,-x_1+b] & 1/2<b\leq 1,\\ (-x_1-b-1+\lceil 2b\rceil,-x_1+b] & b>1.\end{cases}$$

Apparently, we have $I_0\cap I_1=\emptyset$. Proceeding recursively we can construct a sequence of disjoint intervals $\{I_k\}$ with properties as we mentioned before. So, the claim is true.

Lower bound: The following example shows that our estimate in Theorem 2.1.2 is sharp. Let X,Y be independent random variables with the same distribution $\mathbb{P}(X=x_i)=(2n)^{-1}$, where

$$x_i=\begin{cases}i(1+\epsilon)a & i=1,2,\cdots,n,\\ i(1+\epsilon)a-r & i=0,-1,\cdots,-n+1,\end{cases}\tag{2.11}$$

with $\epsilon>0$ small and $0<r\leq a(1+\epsilon)/2$. It is easy to see

$$\mathbb{P}(|X-Y|\leq a)=\mathbb{P}(X=Y)=(2n)^{-1},\tag{2.12}$$

and

$$\mathbb{P}(|X+Y|\leq 1)=(2n)^{-1}\Big(\sum_{i\in I_1}+\sum_{i\in I_2}+\sum_{i\in I_3}\Big)\mathbb{P}(-x_i-1\leq X\leq -x_i+1),\tag{2.13}$$

where $\{I_1,I_2,I_3\}$ is a partition of the index set $\{i:-n+1\leq i\leq n\}$. The sets I_1,I_2 are defined by

$$I_1=\{i:-x_0+1\leq x_i\leq -x_{-n+1}-1\},$$
$$I_2=\{i:-x_n+1\leq x_i\leq -x_1-1\}.$$

Elementary calculations show that

$$|I_1| = \lfloor n - 1 - (1-r)(1+\epsilon)^{-1}a^{-1} \rfloor - \lceil (1+r)(1+\epsilon)^{-1}a^{-1} \rceil + 1, \tag{2.14}$$

$$|I_2| = \lfloor n - (1+r)(1+\epsilon)^{-1}a^{-1} \rfloor - \lceil 1 + (1-r)(1+\epsilon)^{-1}a^{-1} \rceil + 1. \tag{2.15}$$

For any $i \in I_1 \cup I_2$, we have

$$\begin{aligned} &\mathbb{P}(-x_i - 1 \le X \le -x_i + 1) = (2n)^{-1} \cdot |\{k : -x_i - 1 \le x_k \le -x_i + 1\}| \\ = \;& (2n)^{-1} \cdot \left(1 + \lfloor (1-r)(1+\epsilon)^{-1}a^{-1} \rfloor + \lfloor (1+r)(1+\epsilon)^{-1}a^{-1} \rfloor\right). \end{aligned} \tag{2.16}$$

For any $i \in I_3$, we can see

$$\mathbb{P}(-1 - x_i \le X \le 1 - x_i) = O(n^{-1}). \tag{2.17}$$

Combining (2.12)-(2.17), we have

$$\lim_{n\to\infty} \frac{\mathbb{P}(|X+Y| \le 1)}{\mathbb{P}(|X-Y| \le a)} = 1 + \lfloor (1-r)(1+\epsilon)^{-1}a^{-1} \rfloor + \lfloor (1+r)(1+\epsilon)^{-1}a^{-1} \rfloor. \tag{2.18}$$

For all $a > 0$, we will see that there are always appropriate ϵ, r such that the right hand side of (2.18) can achieve $\lceil 2/a \rceil$.

1. When $k < 1/a \le k + 1/2$, for some non-negative integer k, and $r > 0$ small, we have
$$k < (1-r)a^{-1} < k+1, \quad k < (1+r)a^{-1} < k+1.$$
For $\epsilon > 0$ small, we have
$$1 + \lfloor (1-r)(1+\epsilon)^{-1}a^{-1} \rfloor + \lfloor (1+r)(1+\epsilon)^{-1}a^{-1} \rfloor = 2k + 1 = \lceil 2/a \rceil.$$

2. When $k + 1/2 < 1/a \le k + 1$, and $r = a/2$, we have
$$k < (1-r)a^{-1} < k+1 < (1+r)a^{-1} < k+2.$$
Then we can choose $\epsilon > 0$ small such that
$$1 + \lfloor (1-r)(1+\epsilon)^{-1}a^{-1} \rfloor + \lfloor (1+r)(1+\epsilon)^{-1}a^{-1} \rfloor = 2k + 2 = \lceil 2/a \rceil.$$

Now we finish the proof.

■

2.2 Small ball inequalities in Banach spaces

In this section, we will extend small ball inequalities for real-valued random variables to rather general settings, such as for random variables taking values in Banach spaces. Moreover we will see the geometric meanings of the optimal constants in these estimates. We denote by $\tau(n,r)$ the maximal number of points that can be placed in a closed ball of radius r such that the ball is centering at one of these points and all mutual distances exceed 1. Then the extension of Theorem 2.1.1 for random variables taking values in high dimensional Euclidean spaces $\mathbb{R}^n$ can be stated in the following way.

Theorem 2.2.1 (Alon and Yuster [4]). *Let $r > 0$ be a positive number. For any i.i.d. random variables X, Y taking values in $\mathbb{R}^n$, we have*

$$\mathbb{P}(\|X - Y\| \leq r) \leq \tau(n,r) \cdot \mathbb{P}(\|X - Y\| \leq 1). \tag{2.19}$$

In the note [38], we gave an unified treatment for the small ball problem (2.1) in the more general Banach space setting. Before stating our results, let us specify some notations that will be used.

Let $(\mathcal{X}, \|\cdot\|)$ be a separable Banach space with the norm $\|\cdot\|$, and let $F, K \subseteq \mathcal{X}$ be two measurable subsets. Their Minkowski sum $F + K$ is defined as

$$F + K = \{x + y : x \in F,\ y \in K\}.$$

For $\rho > 0$, we denote by $N(F, K, \rho)$ the *ρ-covering number* of F by K, which is defined to be

$$N(F, K, \rho) = \inf\{|A| : A \subseteq \mathcal{X}, F \subseteq A + \rho K\}, \tag{2.20}$$

where $\rho K = \{\rho y : y \in K\}$. It measures the minimal number of points needed to cover the set F under the translation of the dilated set ρK. If $\rho = 1$, then it is corresponding to the standard definition of covering number. The *diameter* $d(K)$ of K is defined in the usual way

$$d(K) = \sup_{x,y \in K} \|x - y\|. \tag{2.21}$$

Another notation to be used is the *inner radius* $r(K)$ of K, which is defined by

$$r(K) = \sup\{r \geq 0 : B(r) \subseteq K\}, \tag{2.22}$$

where $B(r)$ is the closed ball of radius r centered at the origin. With all these notations introduced, we can state one of our main results as follows.

Theorem 2.2.2. *Let $F, K \subseteq \mathcal{X}$ be two measurable subsets. Suppose that K is symmetric with non-empty interior around the origin. For i.i.d. random variables X, Y taking values in $\mathcal{X}$, we have*

$$\mathbb{P}(X + Y \in F) \leq N(F, K, \rho_K) \cdot \mathbb{P}(X - Y \in K), \tag{2.23}$$

where $\rho_K = r(K)/d(K)$. If F is also symmetric, we have

$$\mathbb{P}(X - Y \in F) \leq [N(F\backslash K, K, \rho_K) + 1] \cdot \mathbb{P}(X - Y \in K), \tag{2.24}$$

where $F\backslash K$ is the set consisting of all elements in F but not in K.

Remark. Before showing the proof, let us demonstrate how this general result can give us the estimates for real-valued random variables. We can take $F = [-b, b]$ and $K = [-a, a]$. In this case, we have $r(K) = a$ and $d(K) = 2a$, which implies $\rho_K = 1/2$. Then $N(F, K, \rho_K)$ is the number of translations of $[-a/2, a/2]$ needed to cover the interval $[-b, b]$. Using elementary geometric argument, it is not hard to see that $N(F, K, \rho_K) = \lceil 2b/a \rceil$. Similarly we have $N(F\backslash K, K, \rho_K) + 1 = 2\lceil b/a \rceil - 1$. That gives us the slightly weaker versions of Theorem 2.1.1 and Theorem 2.1.2 without the strict inequality.

Proof. We use $\mathcal{P}$ to denote the set of all probability measures on $\mathcal{X}$. Without confusion, we let $\rho := \rho_K$ and $N := N(F, K, \rho)$. Apparently, the theorem is true for $N = \infty$. In the following, we always assume N is finite. By Lemma 2.1.1, in order to prove (2.23), we only need to show that for any constant $C > N$,

$$\mathbb{P}(X + Y \in F) < C \cdot \mathbb{P}(X - Y \in K) \tag{2.25}$$

for all i.i.d. random variables X, Y. The inequality above can be rewritten as

$$\int_{\mathcal{X}\times\mathcal{X}} \varphi(x,y)d\mu(x)d\mu(y) > 0, \tag{2.26}$$

where

$$\varphi(x,y) = C \cdot 1_{\{(x,y):x-y\in K\}} - 1_{\{(x,y):x+y\in F\}}, \quad x, y \in \mathcal{X},$$

and $\mu \in \mathcal{P}$ is induced by X. Since K is symmetric, we can see $\varphi(x,y)$ is symmetric and bounded. By Lemma 1, it is equivalent to prove

$$\mu\left(\left\{x \in \mathcal{X} : \int_{\mathcal{X}} \varphi(x,y)d\mu(y) > 0\right\}\right) > 0 \tag{2.27}$$

for all $\mu \in \mathcal{P}$. Assume otherwise, then there exists some $\mu \in \mathcal{P}$ such that $\mu(S) = 1$, where

$$\begin{aligned} S &= \left\{x \in \mathcal{X} : \int_{\mathcal{X}} \varphi(x,y)d\mu(y) \le 0\right\} \\ &= \{x \in \mathcal{X} : \mu(-x+F) \ge C \cdot \mu(x-K)\}. \end{aligned} \tag{2.28}$$

Let's define

$$\alpha = \sup_{x\in S} \mu(x-K). \tag{2.29}$$

Since $r(K) > 0$ and $\mathcal{X}$ is separable, there exists a countable subset $S' \subseteq S$ such that $S \subseteq S' - K := \cup_{x\in S'}(x-K)$, which implies $\alpha > 0$. For $\epsilon > 0$ small, we can pick $x^* \in S$ such that

$$\mu(x^* - K) > \alpha - \epsilon. \tag{2.30}$$

By the definition of N, there exists a subset $\{x_i\}_{i=1}^N \subseteq \mathcal{X}$ such that

$$F \subseteq \cup_{i=1}^N (x_i + \rho K).$$

So we have

$$-x^* + F \subseteq \cup_{i=1}^N (x_i - x^* + \rho K) = \cup_{i=1}^N (x_i - x^* - \rho K). \tag{2.31}$$

From (2.30), (2.28) and (2.31), we have

$$C \cdot (\alpha - \epsilon) < C \cdot \mu(x^* - K) \leq \mu(-x^* + F) \leq N \cdot \sup_{x \in \mathcal{X}} \mu(x - \rho K). \tag{2.32}$$

Since $\mu(S) = 1$, for any set $x - \rho K$ with positive measure, there is

$$x_0 \in (x - \rho K) \cap S. \tag{2.33}$$

Next we will show

$$x - \rho K \subseteq B(x_0, r(K)) \subseteq x_0 - K. \tag{2.34}$$

By (2.33), there exists $y_0 \in K$ such that $x_0 = x - \rho y_0$. For any $y \in K$,

$$\|x - \rho y - x_0\| = \rho \|y_0 - y\| \leq \rho \cdot d(K) = r(K),$$

which implies the first part of (2.34). The second part follows from the assumption on K and the definition of $r(K)$. Combining (2.32), (2.33) and (2.34), we have

$$C \cdot (\alpha - \epsilon) < N \cdot \sup_{x \in \mathcal{X}} \mu(x - \rho K) \leq N \cdot \sup_{x \in S} \mu(x - K).$$

Taking $\epsilon = \alpha \cdot (1 - N/C)$, we have

$$N \cdot \alpha = C \cdot (\alpha - \epsilon) < N \cdot \sup_{x \in S} \mu(x - K), \tag{2.35}$$

which contradicts the definition of α in (2.29). So we proved (2.23).

To prove (2.24), we only need to make a slight modification of the previous proof. Similar to (2.25), we need to prove that for any $C > N := N(F \backslash K, K, \rho) + 1$,

$$\mathbb{P}(X - Y \in F) < C \cdot \mathbb{P}(X - Y \in K). \tag{2.36}$$

Instead of (2.28), we redefine

$$S = \{x \in \mathcal{X} : \mu(x - F) \geq C \cdot \mu(x - K)\}, \tag{2.37}$$

and α is defined in the same way as in (2.29). For $\epsilon > 0$ small, we can pick $x^* \in S$ such that

$$\mu(x^* - K) > \alpha - \epsilon. \tag{2.38}$$

By the definition of N, there exists a subset $\{x_i\}_{i=1}^{N-1} \subseteq \mathcal{X}$ such that

$$F\backslash K \subseteq \cup_{i=1}^{N-1}(x_i + \rho K).$$

Hence

$$x^* - F \subseteq (x^* - K) \cup \left(\cup_{i=1}^{N-1}(x^* - x_i - \rho K)\right). \tag{2.39}$$

From (2.38), (2.37) and (2.39), we have

$$\begin{aligned} C \cdot (\alpha - \epsilon) &< C \cdot \mu(x^* - K) \leq \mu(x^* - F) && (2.40)\\ &\leq \mu(x^* - K) + (N-1) \cdot \sup_{x \in \mathcal{X}} \mu(x - \rho K). && (2.41) \end{aligned}$$

Combining (2.40), (2.41), (2.33) and (2.34), we have

$$C \cdot (\alpha - \epsilon) < \mu(x^* - K) + (N-1) \cdot \sup_{x \in \mathcal{X}} \mu(x - \rho K) \leq N \cdot \sup_{x \in S} \mu(x - K).$$

Taking $\epsilon = \alpha \cdot (1 - N/C)$, we get (2.35) again, which is in contradiction to the definition of α. So we proved (2.24). ■

Let us assign two norms $\|\cdot\|_1$ and $\|\cdot\|_2$ on the measurable vector space $\mathcal{X}$. We denote by $B_1(r)$ and $B_2(r)$ the closed balls centered at the origin with radius r under the gauges $\|\cdot\|_1$ and $\|\cdot\|_2$, respectively. Then the following result is an immediate consequence of Theorem 2.2.2.

Corollary 2.2.1. *Let $a, b > 0$ be positive numbers. For any i.i.d. random variables X, Y taking values in $\mathcal{X}$, we have*

$$\mathbb{P}(\|X + Y\|_1 \leq b) \leq N(B_1(b), B_2(a), 1/2) \cdot \mathbb{P}(\|X - Y\|_2 \leq a),$$

and

$$\mathbb{P}(\|X - Y\|_1 \leq b) \leq [N(B_1(b)\backslash B_2(a), B_2(a), 1/2) + 1] \cdot \mathbb{P}(\|X - Y\|_2 \leq a).$$

We denote by $\|\cdot\|_\infty$ the l^∞ norm in $\mathbb{R}^n$. Then Theorem 2.1.1 in conjugation with Theorem 2.1.2 imply the following sharp estimates.

Corollary 2.2.2. *Let X, Y be i.i.d. random vectors in $\mathbb{R}^n$ with independent entries. For $0 < a < 2b$, we have*

$$\mathbb{P}(\|X+Y\|_\infty \leq b) < (\lceil 2b/a \rceil)^n \, \mathbb{P}(\|X-Y\|_\infty \leq a).$$

For $a \geq 2b$, the above inequality still holds with "$\leq$" in the middle. For all $a, b > 0$, we have

$$\mathbb{P}(\|X-Y\|_\infty \leq b) < (2\lceil b/a \rceil - 1)^n \, \mathbb{P}(\|X-Y\|_\infty \leq a).$$

Proof. Let $X = (X_1, \cdots, X_n)$ and $Y = (Y_1, \cdots, Y_n)$. Since Y is an independent copy of X and their entries are independent, we have

$$\begin{aligned} \mathbb{P}(\|X+Y\|_\infty \leq b) &= \prod_{i=1}^n \mathbb{P}(|X_i + Y_i| \leq b) \\ &\leq (\lceil 2b/a \rceil)^d \prod_{i=1}^n \mathbb{P}(|X_i - Y_i| \leq a) \\ &= (\lceil 2b/a \rceil)^d \, \mathbb{P}(\|X-Y\|_\infty \leq a). \end{aligned}$$

The second estimate can be proved in a similar way. So we omit the proof. ■

2.3 Small ball inequalities in groups

The most general setting in which we can talk about sums (and symmetry) is that of group-valued random variables, where the group operation represents summation. In this section, we will explore small ball inequalities for random variables taking values in a topological group from a combinatorial point of view. The reason is two-fold: firstly, it seems to be impossible to generalize the analytical technique developed in [38] to the group setting because it relies essentially on the availability of a dilation operation on the space, and secondly (and perhaps more importantly), it is reasonable to expect a deterministic phenomenon behind these estimates since they are independent of the probability distributions imposes on our random variables.

2.3.1 Combinatorial perspective on distribution-free inequalities

Firstly we demonstrate a combinatorial approach, that enables us to prove distribution-free probabilistic inequalities by considering their combinatorial analogs.

This idea originated from Katona's proof certain probabilistic inequalities [76] using results from extremal graph theory.

Let X be a random variable taking values in certain measurable space, and let F, K be two measurable subsets of the k-fold product space. Given a sequence $X_1, \cdots, X_m$ of independent copies of X, the random variable $T_m(X, F)$ is defined as

$$T_m(X, F) = |\{(i_1, \cdots, i_k) : i_1 \neq \cdots \neq i_k, (X_{i_1}, \cdots, X_{i_k}) \in F\}|. \tag{2.42}$$

Similarly we can define $T_m(X, K)$. For a deterministic sequence $x_1, \cdots, x_m$, we define

$$T_m(F) = |\{(i_1, \cdots, i_k) : i_1 \neq \cdots \neq i_k, (x_{i_1}, \cdots, x_{i_k}) \in F\}|. \tag{2.43}$$

The quantity $T_m(K)$ is defined similarly.

Proposition 2.3.1. *Suppose that there is a function $h_k(m) = o(m^k)$ and an absolute constant $C(F, K)$ such that the inequality*

$$T_m(F) \leq h_k(m) + C(F, K) \cdot T_m(K) \tag{2.44}$$

holds for all deterministic sequences $x_1, \cdots, x_m$. Then the following inequality

$$\mathbb{P}((X_1, \cdots, X_k) \in F) \leq C(F, K) \cdot \mathbb{P}((X_1, \cdots, X_k) \in K). \tag{2.45}$$

holds for any i.i.d. random variables $X_1, \cdots, X_k$.

Proof. The assumption (2.44) for any deterministic sequences implies that

$$T_m(X, F) \leq h_k(m) + C(F, K) \cdot T_m(X, K).$$

In particular, we have

$$\mathbb{E}(T_m(X, F)) \leq h_k(m) + C(F, K) \cdot \mathbb{E}(T_m(X, K)). \tag{2.46}$$

Notice that $T_m(X, F)$ can be written as the summation of Bernoulli random variables with the same distribution

$$T_m(X, F) = \sum 1_{\{(X_{i_1}, \cdots, X_{i_k}) \in F\}}, \tag{2.47}$$

where the summation is taken over all ordered k-tuples $(i_1, \cdots, i_k)$ with distinct coordinates. Therefore we have

$$\mathbb{E}(T_m(X, F)) = (m)_k \cdot \mathbb{P}((X_1, \cdots, X_k) \in F). \tag{2.48}$$

where the notation $(m)_k$ stands for the product $m(m-1)\cdots(m-k+1)$. Similarly we have

$$\mathbb{E}(T_m(X, K)) = (m)_k \cdot \mathbb{P}((X_1, \cdots, X_k) \in K). \tag{2.49}$$

Combining (2.46), (2.48) and (2.49), we have

$$\mathbb{P}((X_1, \cdots, X_k) \in F) \leq \frac{h(m)}{(m)_k} + C(F, K) \cdot \mathbb{P}((X_1, \cdots, X_k) \in K).$$

As m is large, the quantities m^k and $(m)_k$ are of the same magnitude for fixed k. Since $h_k(m) = o(m^k)$, the proposition follows by taking the limit $m \to \infty$. ■

Although the proof is very simple, let us demonstrate the heuristic idea behind Proposition 2.3.1. We will see that the assumption (2.44) is not artificial and it has to be true if the inequality (2.45) holds for all distributions. Using the representation (2.47), it is not hard to show that

$$\mathbb{E}\left(\frac{T_m(X, F)}{m^k} - \mathbb{P}((X_1, \cdots, X_k) \in F)\right)^2 \longrightarrow 0, \text{ as } m \to \infty.$$

In particularly, we have

$$\frac{T_m(X, F)}{(m)_k} \xrightarrow{a.s.} \mathbb{P}((X_1, \cdots, X_k) \in F), \text{ as } m \to \infty.$$

We have similar convergence for $T_m(X, K)$. Then the inequality (2.45) for a fixed random variable X will imply

$$T_m(X, F) \leq o(m^k) + C(F, K) \cdot T_m(X, K), \ a.s.$$

Therefore, for almost all realizations of $X_1, \cdots, X_m$, i.e. deterministic sequences $x_1, \cdots, x_m$, we will have

$$T_m(F) \leq o(m^k) + C(F, K) \cdot T_m(K).$$

We should notice that such sequences depend on the support of X and the $o(m^k)$ term may depend on the sequences. However, if the inequality (2.45) holds for all distributions, it will be reasonable to expect that (2.44) holds for all deterministic sequences.

2.3.2 Abelian groups

Let G be a topological group equipped with the Borel σ-algebra generated by all open sets. Let X, Y be i.i.d. random variables taking values in G. A subset of G is said to be symmetric if it contains the group inverse of each element of this set. In this section, we assume that G is abelian with the identity 0.

Before showing our results, let us introduce some notations to be used. For two subsets $F, K \subseteq G$, their Minkowski sum $F + K$ is defined as

$$F + K = \{x + y : x \in F, y \in K\}. \tag{2.50}$$

Similarly we can define the difference set $F - K$. The *generalized entropy number* $N(F, K)$ is defined to be the maximal number of elements we can select from F such that the difference of any two distinct elements does not belong to K. To state more precisely, it is defined by

$$N(F, K) = \sup\{|S| : S \subseteq F, (S - S) \cap K \subseteq \{0\}\}. \tag{2.51}$$

Let $T = \{x_1, x_2, \cdots, x_m\}$ be a multi-set (or sequence) of G, i.e. the elements of T are selected from G and are not necessary distinct. For any $s \in \mathbb{R}$, the quantities $T_+(F, s)$ and $T_-(K, s)$ are defined as

$$T_+(F, s) = ms + |\{(i, j) : i \neq j, x_i + x_j \in F\}|, \tag{2.52}$$

and

$$T_-(K, s) = ms + |\{(i, j) : i \neq j, x_i - x_j \in K\}|. \tag{2.53}$$

The relation between these two quantities is given in the following lemma, which is similar in spirit to Lemma 2.1 and Lemma 3.2 in [4].

Lemma 2.3.1. *Suppose K is a symmetric set with $0 \in K$. For $s \geq 2$ and any multi-set T, we have*

$$T_+(F, s) \leq N(F, K) \cdot T_-(K, 2s). \tag{2.54}$$

Proof. If $N(F, K) = \infty$, the above statement is obviously true. So, we will assume that $N(F, K)$ is finite and prove the lemma by induction on the cardinality of T. When counting the cardinality of a multiset, every element counts even for two elements with the same value. For the base case $|T| = 1$, we have $T_+(F, s) = s$ and $T_-(K, 2s) = 2s$. Since $N(F, K) \geq 1$, it is clear that the lemma has to be true. We assume that the lemma holds for any multi-set T with cardinality $|T| \leq m - 1$. Let t be some non-negative integer such that

$$\max_{x \in T} |(x + K) \cap T| = t + 1.$$

Here we use $(x + K) \cap T$ to denote the multi-set consisting of elements of T which lie in $x + K$. We will use similar notations without further clarification. Let $x^* \in T$ be an element that can achieve the above maximum and we set $T^* = T \backslash \{x^*\}$, where '\' is the standard set subtraction notation. (We only throw x^* away but not other elements with the same value). Since K is a symmetric set containing 0, we have

$$T_-(K, 2s) = T^*_-(K, 2s) + 2s + 2t. \tag{2.55}$$

We also have

$$T_+(F, s) \leq T^*_+(F, s) + s + 2|(-x^* + F) \cap T|. \tag{2.56}$$

The definition in (2.51) implies that we can select at most $N(F, K)$ elements from $(-x^* + F) \cap T$, say $\{y_1, y_2, \cdots, y_k\}$ with $k \leq N(F, K)$, such that their mutual differences are not in K. Therefore we have

$$(-x^* + F) \cap T \subseteq \cup_i (y_i + K) \cap T. \tag{2.57}$$

Combining with (2.56), we have

$$T_+(F, s) \leq T^*_+(F, s) + s + 2(t + 1)N(F, K). \tag{2.58}$$

By the induction assumption, the lemma holds for T^*. Combining (2.55) and (2.58), it is not hard to check that the lemma holds when

$$s \geq \frac{2N(F,K)}{2N(F,K)-1},$$

which is implied by the assumption $s \geq 2$. ■

By Proposition 2.3.1, we have the following result.

Theorem 2.3.1. *Let $F, K \subseteq G$ be measurable subsets. Suppose that K is symmetric and contains the identity of G in its interior. For any G-valued i.i.d. random variables X, Y, we have*

$$\mathbb{P}(X+Y \in F) \leq N(F,K) \cdot \mathbb{P}(X-Y \in K). \tag{2.59}$$

The same argument can be used to study the comparison between $\mathbb{P}(X-Y \in F)$ and $\mathbb{P}(X-Y \in K)$. Let $T_-(F,s)$ be defined the same as (2.53) with K replaced by F. Similar to Lemma 2.3.1, we have

Lemma 2.3.2. *Suppose K is a symmetric set with $0 \in K$. For $s \geq 2$ and any multi-set T, we have*

$$T_-(F,s) \leq (1+N(F\backslash K, K)) \cdot T_-(K,2s). \tag{2.60}$$

Proof. We only need to make a slight modification of the proof of Lemma 2.3.1. Let x^* be chosen in the same way as before and we set $T^* = T\backslash\{x^*\}$. It is clear that the equation (2.55) still holds. Instead of (2.56), we have

$$T_-(F,s) \leq T_-^*(F,s) + s + |((x^*+F)\cap T)\backslash\{x^*\}| + |((x^*-F)\cap T)\backslash\{x^*\}|. \tag{2.61}$$

Notice the following set-inclusion relations

$$(x^*+F) \subseteq (x^*+K) \cup (x^*+F\backslash K),$$

and

$$(x^*-F) \subseteq (x^*+K) \cup (x^*-F\backslash K).$$

We use the symmetry assumption of K in the second inclusion relation. Applying the covering argument (2.57) again to $x^* + F\backslash K$ and $x^* - F\backslash K$, we have

$$T_-(F,s) \leq T_-^*(F,s) + s + 2t + 2(t+1)N(F\backslash K, K). \tag{2.62}$$

Combining (2.55) and (2.62), for $s \geq 2$, we will have

$$T_-(F,s) \leq (1 + N(F\backslash K, K)) \cdot T_-(K, 2s).$$

So we complete the proof of the lemma. ■

By Proposition 2.3.1, we have the following result.

Theorem 2.3.2. *Let $F, K \subseteq G$ be two measurable subsets. Suppose that K is symmetric and contains the identity of G in its interior. For any G-valued i.i.d. random variables X, Y, we have*

$$\mathbb{P}(X - Y \in F) \leq (1 + N(F\backslash K, K)) \cdot \mathbb{P}(X - Y \in K). \tag{2.63}$$

2.3.3 Non-abelian groups

In this section, we let G be a general group with the identity e. We will show that Theorem 2.3.1 and Theorem 2.3.2 still hold for certain measurable sets F, K in this general setting. Similar to the sumset $F + K$ in the abelian case, we define the product set $F \cdot K$ in this non-abelian setting as

$$F \cdot K = \{xy : x \in F, y \in K\}. \tag{2.64}$$

The *generalized entropy number* $N(F,K)$ is redefined as

$$N(F,K) = \sup\{|S| : S \subseteq F, (S \cdot S^{-1}) \cap K \subseteq \{e\}\}, \tag{2.65}$$

where S^{-1} is the set of all inverses of the elements of S. For $s \in \mathbb{R}$ and a multi-set $T = \{x_1, \cdots, x_m\}$, the quantities $T_+(F,s)$ and $T_-(K,s)$ are redefined as

$$T_+(F,s) = ms + |\{(i,j) : i \neq j, x_i x_j \in F\}|, \tag{2.66}$$

and

$$T_-(K,s) = ms + |\{(i,j) : i \neq j, x_i x_j^{-1} \in K\}|. \tag{2.67}$$

Similar to Lemma 2.3.1, we have the following result.

Lemma 2.3.3. *Suppose K is a normal subgroup of G. For $s \geq 2$ and any multi-set T, we have*

$$T_+(F,s) \leq N(F,K) \cdot T_-(K,2s). \tag{2.68}$$

Proof. The lemma can be proved with a slight modification of the proof of Lemma 2.3.1. In order to see how the assumption of K is used, we write the proof again. Let t be some non-negative integer such that

$$\max_{x \in T} |(xK) \cap T| = t+1,$$

where xK is the set of the products of x and the elements of K. Let x^* be an element such that the maximum can achieved and $T^* = T \backslash \{x^*\}$. By the definition of $T_-(K,2s)$, we have

$$T_-(K,2s) = T_-^*(K,2s) + 2s + |((Kx^*) \cap T) \backslash \{x^*\}| + |((x^* K^{-1}) \cap T) \backslash \{x^*\}|$$

Since K is a normal subgroup, the estimate of $T_-(K,2s)$ in (2.55) still holds. Similar to (2.56), we have

$$T_+(F,s) \leq T_+^*(F,s) + s + |((x^*)^{-1}F) \cap T| + |(F(x^*)^{-1}) \cap T|.$$

Let $\alpha_1, \alpha_2 \in F$ be any two elements, and $u_1 = (x^*)^{-1}\alpha_1$, $u_2 = (x^*)^{-1}\alpha_2$. Since K is a normal subgroup, we can see that $u_1 u_2^{-1} \in K$ if only if $\alpha_1 \alpha_2^{-1} \in K$. (The assumption that K is a normal subgroup is important here). By the definition of $N(F,K)$ in (2.65), we can select at most $N(F,K)$ elements from $((x^*)^{-1}F)) \cap T$, say $\{y_1, \cdots, y_k\}$, such that $y_i y_j^{-1} \notin K$ for any $y_i \neq y_j$. Then we have the following covering relation

$$((x^*)^{-1}F) \cap T \subseteq \cup_i (y_i K) \cap T,$$

which implies

$$|((x^*)^{-1}F) \cap T| \leq (t+1)N(F,K).$$

Similarly we have the same estimate for $|(F(x^*)^{-1}) \cap T|$. Then we can see that the estimate of $T_+(F,s)$ is exactly the same as (2.58). So we proved the lemma. ■

Using Proposition 2.3.1, we have the following result.

Theorem 2.3.3. *Let $F, K \subseteq G$ be measurable subsets. Suppose that K is a normal subgroup and contains the identity in its interior. For any G-valued i.i.d. random variables X, Y, we have*

$$\mathbb{P}(XY \in F) \leq N(F,K) \cdot \mathbb{P}(XY^{-1} \in K). \tag{2.69}$$

The following lemma can be proved in the same way as done for Lemma 2.3.2. Thus we omit its proof.

Lemma 2.3.4. *Suppose $K \subseteq G$ is a normal subgroup. For $s \geq 2$ and any multi-set T, we have*

$$T_-(F,s) \leq (1 + N(F \backslash K, K)) \cdot T_-(K, 2s). \tag{2.70}$$

Using Proposition 2.3.1, we have the following result.

Theorem 2.3.4. *Let $F, K \subseteq G$ be two measurable subsets. Suppose that K is a normal subgroup and contains the identity in its interior. For any G-valued i.i.d. random variables X, Y, we have*

$$\mathbb{P}(XY^{-1} \in F) \leq (1 + N(F \backslash K, K)) \cdot \mathbb{P}(XY^{-1} \in K). \tag{2.71}$$

2.3.4 Topological vector spaces

Let V be a topological vector space over a field $\mathbb{F}$ with the Borel σ-algebra generated by all open sets. Let $F, K \subseteq V$ be measurable subsets and let $a, b \in \mathbb{F}$. Then we can consider the comparison between $\mathbb{P}(aX + bY \in F)$ and $\mathbb{P}(X - Y \in K)$ for i.i.d. random variables X, Y taking values in V.

Theorem 2.3.5. *Let $F, K \subseteq V$ be measurable subsets. Suppose that K is symmetric and contains the zero vector in its interior. Let a, b be non-zero elements of $\mathbb{F}$. For any i.i.d. random variables X, Y taking values in V, we have*

$$\mathbb{P}(aX + bY \in F) \leq N(a, b, F, K) \cdot \mathbb{P}(X - Y \in K), \tag{2.72}$$

where the constant $N(a, b, F, K)$ is defined by

$$N(a, b, F, K) = \frac{1}{2}\left(N(a^{-1}F, K) + N(b^{-1}F, K)\right). \tag{2.73}$$

Proof. The proof is essentially the same as that of Theorem 2.3.1. Let $T = \{x_1, \cdots, x_m\}$ be a multi-set of V. For $s \in \mathbb{R}$, we define

$$T_+(F, s, a, b) = ms + |\{(i, j) : i \neq j, ax_i + bx_j \in F\}|. \tag{2.74}$$

By Proposition 2.3.1, we only need to prove the following combinatorial analogue

$$T_+(F, s, a, b) \leq N(a, b, F, K) \cdot T_-(K, 2s) \tag{2.75}$$

for $s \geq 2$, where $T_-(K, 2s)$ is defined the same as (2.53). We choose x^* in the same way as in Lemma 2.3.1 and set $T^* = T \setminus \{x^*\}$. The estimate of $T_-(K, 2s)$ in (2.55) still holds. Similar to (2.56), we have

$$T_+(F, s, a, b) \leq T_+^*(F, s, a, b) + s + |(-b^{-1}ax^* + b^{-1}F) \cap T| + |(-a^{-1}bx^* + a^{-1}F) \cap T|.$$

Applying the covering argument (2.57) to $-b^{-1}ax^* + b^{-1}F$, and using the definition of $N(b^{-1}F, K)$, we will have

$$|(-b^{-1}ax^* + b^{-1}F) \cap T| \leq (t + 1)N(b^{-1}F, K).$$

Similarly we can see that

$$|(-a^{-1}bx^* + a^{-1}F) \cap T| \leq (t + 1)N(a^{-1}F, K).$$

Thus we have

$$T_+(F, s, a, b) \leq T_+^*(F, s, a, b) + s + 2(t + 1)N(a, b, F, K). \tag{2.76}$$

Then the estimate (2.75) follows from (2.55) and (2.76). So we complete the proof. ■

The following result emerges as a special case of Theorem 2.3.5. It extends Theorem 2.1.1 and Theorem 2.1.2 in certain extent.

Corollary 2.3.1. *Let a, b, c, d be non-zero real numbers and $c, d > 0$. For any real-valued i.i.d. random variables X, Y, we have*

$$\mathbb{P}(|aX + bY| \leq c) \leq \frac{1}{2}\left(\left\lceil \frac{2c}{|a|d} \right\rceil + \left\lceil \frac{2c}{|b|d} \right\rceil\right) \mathbb{P}(|X - Y| \leq d). \tag{2.77}$$

Proof. We can take $F = [-c, c]$ and $K = [-d, d]$. Elementary geometric argument will yield

$$N(a^{-1}F, K) = \left\lceil \frac{2c}{|a|d} \right\rceil, \quad N(b^{-1}F, K) = \left\lceil \frac{2c}{|b|d} \right\rceil.$$

Then the result follows from Theorem 2.3.5. ■

Remark. If $a = b$, we can see that Theorem 2.3.5 and Corollary 2.3.1 match Theorem 2.3.1 and Theorem 2.1.2, respectively. However, for $a = -b$, Theorem 2.3.5 and Corollary 2.3.1 are weaker than Theorem 2.3.2 and Theorem 2.1.1, respectively.

Let $F \subseteq V$ be a measurable subset and let X be a random variable taking values in V. The generalized Lévy's concentration function of X is defined to be

$$Q(X, F) = \sup_{x \in V} \mathbb{P}(X \in x + F). \tag{2.78}$$

Then Theorem 2.3.5 can be used to bound the concentration function of $aX + bY$ by that of $X - Y$ for i.i.d. random variables X, Y.

Corollary 2.3.2. *Let $F, K \subseteq V$ be measurable subsets. Suppose that K is symmetric and contains the zero vector in its interior. Let a, b be non-zero elements of $\mathbb{F}$. For any i.i.d. random variables X, Y taking values in V, we have*

$$Q(aX + bY, F) \leq N(a, b, F, K) \cdot Q(X - Y, K). \tag{2.79}$$

Proof. By the definition of generalized Lévy's concentration function, for any $\epsilon > 0$, there exists $x \in V$ such that

$$\begin{aligned} Q(aX + bY, F) &< \mathbb{P}(aX + bY \in x + F) + \epsilon \\ &\leq N(a, b, F, K) \cdot \mathbb{P}(X - Y \in K) + \epsilon \\ &\leq N(a, b, F, K) \cdot Q(X - Y, K) + \epsilon. \end{aligned}$$

In the second inequality, we use Theorem 2.3.5. Since $\epsilon > 0$ is arbitrary, the statement follows by letting $\epsilon \to 0$. ■

The main study of concentration function is devoted to the sum of independent random variables in Banach spaces (mostly on Euclidean spaces) with $F = K$ taken to be normed balls, see [99, 37, 88, 137, 48, 49, 84, 70, 74, 123]. In the i.i.d. case, Theorem 2.3.1 can provide us a symmetrization technique to treat different sets and also general groups where no norm may exist. For a random variable X, we use $\widetilde{X}$ to denote the symmetrized random variable $X - Y$, where Y is an independent copy of X.

Corollary 2.3.3. *Let G be an abelian group and let $F, K \subseteq G$ be measurable subsets. Suppose that K is symmetric and contains the identity of G in its interior. For any i.i.d. random variables $X_1, \cdots, X_n$ taking values in G, we have*

$$Q(X_1 + \cdots + X_n, F) \leq N(F, K) \cdot Q(\widetilde{X}_1 + \cdots + \widetilde{X}_{\lfloor n/2 \rfloor}, K). \tag{2.80}$$

Proof. For independent random variables X, Y, (Y is not necessary an independent copy of X), it is not hard to see that

$$Q(X + Y, F) \leq Q(X, F).$$

Thus we have

$$\begin{aligned} Q(X_1 + \cdots + X_n, F) &\leq Q(X_1 + \cdots + X_{2\lfloor n/2 \rfloor}, F) \\ &\leq N(F, K) \cdot Q(\widetilde{X}_1 + \cdots + \widetilde{X}_{\lfloor n/2 \rfloor}, K). \end{aligned}$$

The second inequality follows from Theorem 2.3.1. ■

As another application of the combinatorial argument, we include the following result of Katona [76], which is related to Turán's theorem for triangle-free graph. We claim no contribution for the proof.

Theorem 2.3.6 (Katona [76]). *Let X, Y be i.i.d. random variables taking values in a Hilbert space V with the norm $\|\cdot\|$. Then we have*

$$(\mathbb{P}(\|X\| \geq 1))^2 \leq 2\mathbb{P}(\|X+Y\| \geq 1). \tag{2.81}$$

Proof. Let $F, K \subseteq V \times V$ be the subsets defined by

$$F = \{(x, y) : \|x\| \geq 1, \|y\| \geq 1\},$$

and

$$K = \{(x, y) : \|x + y\| \geq 1\}.$$

Given a multi-set $T = \{x_1, \cdots, x_m\}$ of V, we define

$$T_m(F) = |\{(i, j) : i \neq j, (x_i, x_j) \in F\}|,$$

and

$$T_m(K) = |\{(i, j) : i \neq j, (x_i, x_j) \in K\}|.$$

By Proposition 2.3.1, the theorem will hold if we can show that

$$T_m(F) \leq 2(m + T_m(K)). \tag{2.82}$$

Suppose that there are n elements of T with norms not less than 1. Then we have

$$T_m(F) = n^2 - n. \tag{2.83}$$

Let us consider a simple graph G on these n elements. (The notation G should not be confused with the notation used for group. For any two elements, we always think them as different vertices even if they have the same value). Two vertices x, y are adjacent if and only if $\|x + y\| \geq 1$. Then we have

$$T_m(K) \geq 2e(G), \tag{2.84}$$

where $e(G)$ is the number of edges of G. For any 3 vertices x, y, z, there exists at least a pair, say x, y, such that the angle between them is no more than $2\pi/3$, which implies

that $\|x+y\| \geq 1$. This fact implies that the complementary graph is triangle free. Using Turán's theorem, we have

$$e(G) \geq \binom{n}{2} - \frac{n^2}{4}. \tag{2.85}$$

Then the estimate (2.82) follows from (2.83), (2.84) and (2.85). ■

Remark. The above argument was used by Katona to prove the theorem for discrete random variables uniformly distributed on a finite subset of a Hilbert space. Then he made generalizations for arbitrary distributions based on extensions of discrete Turán-type theorems to the continuous setting. We can not see if this extension process is necessary, since the problem can be treated in a unified way according to Proposition 2.3.1.

Remark. In a series of papers [76, 77, 78, 79, 79, 80, 81, 82, 83], Katona studied the optimal estimate of $\mathbb{P}(\|X+Y\| \geq c)$ in terms of $\mathbb{P}(\|X\| \geq 1)$ for i.i.d. random variables X, Y taking values in a Hilbert space. The basic idea is to study this type of problems for uniformly distributed discrete random variables and make extensions to the continuous setting. In the discrete situation, it usually involves extremal combinatorial problems. Comprehensive results are given in the survey [83]. Similar results were independently obtained by Sidorenko [147], who also considered the estimate of $\mathbb{P}(\|aX+bY\| \geq c)$.

Remark. Proposition 2.3.1 provides a combinatorial argument for the comparison of two probabilities of the same magnitude, i.e. one can be bounded by the other one linearly. A general question is that do we have a similar approach when the probabilities are of different magnitudes. In another word, can we establish the combinatorial analogue for the following distribution-free inequality

$$\mathbb{P}((X_1, \cdots, X_k) \in F) \leq f(\mathbb{P}(X_1, \cdots, X_l) \in K),$$

where f is certain function, not necessary linear. The estimate in Theorem 2.3.6 is a particular example in this flavor.

2.4 Discussion of tightness

In this section, we study the near extremal distributions for the probabilistic estimates developed in previous sections. The discussion will mainly focus on Theorem 2.3.1 and Theorem 2.3.2 for random variables taking values in the Euclidean space $\mathbb{R}^d$. We will see their close connections with the sphere packing problem in geometry.

In general it is hard to compute the ratio of $\mathbb{P}(X \pm Y) \in F$ and $\mathbb{P}(X - Y) \in K$. If X, Y are assumed to be uniformly distributed on a finite set $T = \{x_1, x_2, \cdots, x_n\}$, then we have

$$\mathbb{P}(X \pm Y \in F) = \frac{1}{n} \sum_{i=1}^{n} |(\mp x_i + F) \cap T|,$$

and

$$\mathbb{P}(X - Y \in K) = \frac{1}{n} \sum_{i=1}^{n} |(x_i + K) \cap T|.$$

If the set T is K-separated, i.e. $x_i - x_j \notin K$ for $i \neq j$, we will have $|(x_i + K) \cap T| = 1$ for all $x_i \in T$. We can even make a further assumption that, except $o(n)$ of them, all the sets $\mp x_i + F$ contain the same number of elements of T. This is possible if T is selected to consist of certain lattice points. (So the random variables X, Y need to be in a topological vector space V). Under these assumptions, we have

$$\frac{\mathbb{P}(X \pm Y \in F)}{\mathbb{P}(X - Y \in K)} \to \max_{x \in T} |(\mp x + F) \cap T|, \quad \text{as } n \to \infty.$$

Then the estimate in Theorem 2.3.1 is tight if there exists a K-separated lattice $\mathcal{L}$ and a point $x \in V$ (not necessary a lattice point) such that $x + F$ contains $N(F, K)$ points of $\mathcal{L}$. We can take the support set T as the union of a subset of $\mathcal{L}$ and the reflection of this subset after certain shift. Similarly, Theorem 2.3.2 is tight if for every lattice point $x \in \mathcal{L}$ the set $x + (F \backslash K)$ contains $N(F \backslash K, K)$ points of $\mathcal{L}$. In this case, we only need to take T to be certain subset of the lattice $\mathcal{L}$. This idea can be used to produce near optimal examples for the estimates (1.2) and (1.1). For the estimate (1.2), we can take X to be uniformly distributed on $\{-(n-1)\delta - r, \cdots, -\delta - r, \delta, 2\delta, \cdots, n\delta\}$, where $r > 0$ and $\delta > a$. For any a, b, we can always choose appropriate parameters r, δ such that the ratio $\mathbb{P}(|X + Y| \leq b)/\mathbb{P}(|X - Y| \leq a)$ will approach $\lceil 2b/a \rceil$ as $n \to \infty$.

This example is essentially the same as the one given in [38]. To see the sharpness of (1.1), we can take X to be uniformly distributed on $\{\delta, 2\delta, \cdots, n\delta\}$ for certain $\delta > a$, which was given in [4].

In the Euclidean space $\mathbb{R}^d$, let us take F and K to be closed balls centered at the origin of radius r and 1, respectively. For simplicity, we use $N_+(r)$ and $N_-(r)$ to denote $N(F, K)$ and $N(F\backslash K, K)+1$, respectively. Then $N_+(r)$ represents the maximal number of points in a Euclidean ball of radius r with all mutual distances greater than 1. For $N_-(r)$ we put an extra restriction that one of these points should be at the center of the ball. These are the so-called sphere packing problems. The dual problem of $N_+(r)$ asks for the smallest radius of the ball to contain n points with mutual distances at least 1. We use $r_+(n)$ to denote this quantity. (Notice that they are not exactly dual to each other, since in the definition of $r_+(n)$ the mutual distances can be equal to 1). Similarly we can define $r_-(n)$ with the restriction that one of the points should be at the center of the ball.

For $d = 2$, instead of the radius function $r_+(n)$, Bateman and Erdős [10] studied the diameters of the extremal configurations of points. Using their results, we can get the corresponding radius function $r_+(n)$. Using the duality, we list the values of $N_+(r)$ for r in certain range.

$$N_+(r) = \begin{cases} 1, & \text{if } 0 < r \le 1/2 \\ 2, & \text{if } 1/2 < r \le \sqrt{3}/3 \\ 3, & \text{if } \sqrt{3}/3 < r \le \sqrt{2}/2 \\ 4, & \text{if } \sqrt{2}/2 < r \le \frac{1}{2}\csc(\pi/5) \\ 5, & \text{if } \frac{1}{2}\csc(\pi/5) < r \le 1 \\ 7, & \text{if } 1 < r \le 1+\epsilon, \text{ small } \epsilon > 0. \end{cases}$$

Since the extremal configurations given by Bateman and Erdős are lattice points, the listed values of $N_+(r)$ are tight for Theorem 2.3.1. It is not hard to see that $r_-(2) = \cdots = r_-(7) = 1$ with one point at the center of a unit circle and the rest points on this

circle. Bateman and Erdős also gave the values of $r_-(n)$ for $n = 8, 9, 10, 11$. Then we can get a list of values of $N_-(r)$.

$$N_-(r) = \begin{cases} 7, & \text{if } 1 < r \leq \frac{1}{2}\csc(\pi/7) \\ 8, & \text{if } \frac{1}{2}\csc(\pi/7) < r \leq \frac{1}{2}\csc(\pi/8) \\ 9, & \text{if } \frac{1}{2}\csc(\pi/8) < r \leq \frac{1}{2}\csc(\pi/9) \\ 10, & \text{if } \frac{1}{2}\csc(\pi/9) < r \leq \frac{1}{2}\csc(\pi/10). \end{cases}$$

which are tight for Theorem 2.3.2. For sphere packing problems, people are generally interested in the packing density. In $\mathbb{R}^2$, it is known that hexagonal lattice packing is optimal among all packings (not necessary lattice packings) with packing density $\sqrt{3}\pi/6 \approx 0.9069$. Then we have the following asymptotic behavior

$$N_+(r) = N_-(r) = (1 + o(1))\frac{2\sqrt{3}}{3}\pi r^2, \tag{2.86}$$

which is asymptotically tight for Theorem 2.3.1 and Theorem 2.3.2.

There is a long history on the sphere packing problem in three dimensional Euclidean space. Kepler conjectured that no arrangement of equally sized spheres can fill the space with a greater average density than that of the face-centered cubic and hexagonal close packing arrangements. The density of these arrangements is $\sqrt{2}\pi/6 \approx 0.7404$. It is proved by Gauss that Kepler's conjecture is true if the spheres have to be arranged in a regular lattice. The complete proof of Kepler's conjecture was given by Hales [65]. Thus we have the following asymptotic behavior

$$N_+(r) = N_-(r) = (1 + o(1))\frac{4\sqrt{2}}{3}\pi r^3, \tag{2.87}$$

which is asymptotically tight for Theorem 2.3.1 and Theorem 2.3.2.

In the very recent breakthrough work [162], Viazovskait proved that the E_8 lattice packing gives the optimal packing density in dimension 8, and the density is $\pi^4/384 \approx 0.025367$. Thus we have the following asymptotic behavior

$$N_+(r) = N_-(r) = (1 + o(1))\frac{2}{3}\pi^4 r^8. \tag{2.88}$$

Building on Viazovskait's work, it is shown in [30] that Leech lattice is the densest packing in 24 dimension, and the packing density is $\pi^{12}/12! \approx 0.00193$. Correspondingly we have the following asymptotic behavior

$$N_+(r) = N_-(r) = (1 + o(1))\frac{2^{24}}{12!}\pi^{12}r^{24}. \tag{2.89}$$

Another interesting problem related to our study is the kissing number problem. In three dimensions it asks how many billiard balls can be arranged so that they all just touch another billiard ball of the same size. This question was a subject of a famous discussion between Isaac Newton and David Gregory in 1694. Newton believed the answer was 12, while Gregory though that 13 might be possible. Generally we can define the d-dimensional kissing number τ_d as the maximal number of points on the unit sphere with Euclidean distances at least 1. For $1 < r < 1 + \epsilon_d$ with small $\epsilon_d > 0$, it is not hard to see the following relation

$$N_-(r) = \tau_d + 1. \tag{2.90}$$

The number $\tau_3 = 12$ was studied by various researchers in the nineteenth century. The best proof now available is due to Leech [97]. The answers $\tau_8 = 240$ and $\tau_{24} = 196,560$ are given by [130] and [98], respectively. It is somewhat surprising that they are technically easier to establish than τ_3. The correct answer $\tau_4 = 24$ was obtained much later by Musin [125]. For all these results, the extremal configurations follows from lattice packings. Using the relation (2.90), Theorem 2.3.2 can give explicit tight estimates for r slightly greater than 1 in corresponding dimensions. These are all the known values of the kissing number so far. In high dimensions, τ_d grows exponentially with unknown base. We refer to the monograph [31] for more discussions of sphere packing problems and their relations with number theory and coding theory.

2.5 Applications

In this section, we will apply the estimates developed in the previous section to study the comparison between $\mathbb{E}(\varphi(aX + bY))$ and $\mathbb{E}(\phi(X - Y))$ for certain functions φ and ϕ. In particular we will establish some moment inequalities.

2.5.1 Hölder type inequalities

Let V be a vector space over the complex field $\mathbb{C}$. Let $\varphi = \|\cdot\|_1$ and $\phi = \|\cdot\|_2$ be two equivalent norms on V and let $I : (V, \|\cdot\|_1) \to (V, \|\cdot\|_2)$ be the identity operator. Its norm $\|I\|$ is defined in the usual way

$$\|I\| = \sup_{\|x\|_1=1} \|x\|_2. \tag{2.91}$$

Let $B_1(r)$, $B_2(r)$ be the closed balls centered at the origin of radius r under the gauges $\|\cdot\|_1$, $\|\cdot\|_2$, respectively. Then it is not hard to see the following geometric interpretation of $\|I\|$:

$$\|I\| = \inf\{r > 0 : B_1(1) \subseteq B_2(r)\}. \tag{2.92}$$

Theorem 2.5.1. *Let $a, b \in \mathbb{C}$ be non-zero complex numbers, and let $q \geq p$ be real numbers such that $pq > 0$. For any i.i.d. random variables X, Y taking values in V, we have*

$$(\mathbb{E}\|X - Y\|_2^p)^{1/p} \leq 2\|I\| \max\left\{|a|^{-1}, |b|^{-1}\right\} \cdot (\mathbb{E}\|aX + bY\|_1^q)^{1/q}. \tag{2.93}$$

Proof. We assume that the right hand side of (2.93) is finite. Otherwise the theorem yields a trivial result. By Hölder's inequality, we only need to prove the theorem for $q = p$. For $p > 0$ we have

$$\begin{aligned} \mathbb{E}\|aX + bY\|_1^p &= \int_0^\infty pt^{p-1}\mathbb{P}(\|aX + bY\|_1 > t)dt \\ &= \int_0^\infty pt^{p-1}\left(1 - \mathbb{P}(\|aX + bY\|_1 \leq t)\right)dt. \end{aligned}$$

By the geometric interpretation (2.92) of $\|I\|$ and Theorem 2.3.5, we have

$$\begin{aligned} \mathbb{P}(\|aX + bY\|_1 \leq t) &\leq \mathbb{P}(\|aX + bY\|_2 \leq t\|I\|) \\ &\leq \mathbb{P}(\|X - Y\|_2 \leq Ct), \end{aligned}$$

where the constant $C = 2\|I\| \max\{|a|^{-1}, |b|^{-1}\}$. Hence we have

$$\begin{aligned} \mathbb{E}\|aX + bY\|_1^p &\geq \int_0^\infty pt^{p-1}\left(1 - \mathbb{P}(\|X - Y\|_2 \leq Ct)\right)dt \\ &= \int_0^\infty pt^{p-1}\mathbb{P}(\|X - Y\|_2 > Ct)dt \\ &= C^{-p} \cdot \mathbb{E}\|X - Y\|_2^p, \end{aligned}$$

which is exactly the estimate (2.93) for $p = q$. For $p < 0$, we have

$$\begin{aligned}
\mathbb{E}\|aX + bY\|_1^p &= -p\int_0^\infty t^{-p-1}\mathbb{P}(\|aX + bY\|_1 \le t^{-1})dt \\
&\le -p\int_0^\infty t^{-p-1}\mathbb{P}(\|aX + bY\|_2 \le \|I\|t^{-1})dt \\
&\le -p\int_0^\infty t^{-p-1}\mathbb{P}(\|X - Y\|_2 \le Ct^{-1})dt \\
&= C^{-p}\left(-p\int_0^\infty t^{-p-1}\mathbb{P}(\|X - Y\|_2 \le t^{-1})dt\right) \\
&= C^{-p}\cdot\mathbb{E}\|X - Y\|_2^p
\end{aligned}$$

In the first and second inequalities, we use the geometric interpretation of $\|I\|$ and Theorem 2.3.5, respectively. We will get (2.93) by taking the $1/p$-th root of both sides. ■

The following result is an immediate consequence of the above theorem.

Corollary 2.5.1. *Let $(V, \|\cdot\|)$ be a normed vector space over the complex field $\mathbb{C}$. For any i.i.d. random variables X, Y taking values in V, we have*

$$\mathbb{E}\|X - Y\| \le 2\mathbb{E}\|X + Y\|, \tag{2.94}$$

and

$$\mathbb{E}\|X + Y\|^{-1} \le 2\mathbb{E}\|X - Y\|^{-1}. \tag{2.95}$$

For $p > 2$, Buja, et al [26] constructed an example such that

$$\mathbb{E}\|X - Y\|_p > \mathbb{E}\|X + Y\|_p, \tag{2.96}$$

where $\|\cdot\|_p$ is the l^p norm on $\mathbb{R}^d$. But the ratio is a little bit greater than 1. So we do not know whether the estimate (2.94) is tight. Let p, r be positive numbers such that $0 < \gamma \le p \le 2$, $1 \le p \le 2$. For $\mathbb{R}^d$-valued i.i.d. random variables X, Y, Buja, et al [26] also proved that

$$\mathbb{E}\|X - Y\|_p^\gamma \le \mathbb{E}\|X + Y\|_p^\gamma. \tag{2.97}$$

Another related result proved by Mattner [115] is that for $0 < \gamma \leq 2$ and any orthogonal map T on $\mathbb{R}^d$ we have

$$\mathbb{E}\|X - Y\|^\gamma \leq \mathbb{E}\|X - TY\|^\gamma, \tag{2.98}$$

where $\|\cdot\|$ is the Euclidean norm.

We call a function $\varphi : V \to \mathbb{R}$ unimodal if the super-level set $\{x \in V : \varphi(x) \geq t\}$ is convex for all $t \in \mathbb{R}$.

Theorem 2.5.2. *Let $a, b \in \mathbb{C}$ be non-zero complex numbers and let φ be a non-negative symmetric unimodal function on V. For any i.i.d. random variables X, Y taking values in V, we have*

$$\mathbb{E}\varphi(aX + bY) \leq \mathbb{E}\varphi\left(\frac{X - Y}{2\max\{|a|^{-1}, |b|^{-1}\}}\right). \tag{2.99}$$

Proof. Since φ is non-negative, we have

$$\begin{aligned}\mathbb{E}\varphi(aX + bY) &= \int_0^\infty \mathbb{P}\left(\varphi(aX + bY) > t\right) dt \\ &= \int_0^\infty \mathbb{P}(aX + bY \in \varphi^{-1}(t, \infty))dt\end{aligned}$$

where we use $\varphi^{-1}(t, \infty)$ to denote the set $\{x \in V : \varphi(x) > t\}$. Since φ is a symmetric unimodal function, we can see that $\varphi^{-1}(t, \infty)$ is a symmetric convex set. Using Theorem 2.3.5, we have

$$\mathbb{P}(aX + bY \in \varphi^{-1}(t, \infty)) \leq \mathbb{P}\left(X - Y \in 2\max\{|a|^{-1}, |b|^{-1}\}\varphi^{-1}(t, \infty)\right),$$

which implies

$$\begin{aligned}\mathbb{E}\varphi(aX + bY) &\leq \int_0^\infty \mathbb{P}\left(\frac{X - Y}{2\max\{|a|^{-1}, |b|^{-1}\}} \in \varphi^{-1}(t, \infty)\right) dt \\ &= \mathbb{E}\varphi\left(\frac{X - Y}{2\max\{|a|^{-1}, |b|^{-1}\}}\right),\end{aligned}$$

■

Remark. Take $\varphi(x) = \|x\|^p$ for $p < 0$. The above result will yield Theorem 2.5.1 when the two norms are the same. Let φ and ϕ be two non-negative symmetric unimodal functions. The comparison between $\mathbb{E}\varphi(aX + bY)$ and $\mathbb{E}\phi(X - Y)$ usually involves the comparison of the gauges $\|\cdot\|_{\varphi,t}$ and $\|\cdot\|_{\phi,t}$ induced by the symmetric convex sets $\varphi^{-1}(t, \infty)$ and $\phi^{-1}(t, \infty)$, respectively. That is related to the study of the entropy number $N(a, b, \varphi^{-1}(t, \infty), \psi^{-1}(t, \infty))$ defined in Theorem 2.3.5.

2.5.2 Reverse Hölder type inequalities

The reverse Hölder inequality asserts the equivalence of higher and lower moments of random variables. More precisely, there exists a constant $C(p, q)$ depending only on $q \geq p$ such that

$$(\mathbb{E}\|X\|^q)^{1/q} \leq C(p, q)(\mathbb{E}\|X\|^p)^{1/p} \tag{2.100}$$

holds for random variables X in certain normed measurable space. In general such an inequality does not hold. But it is well known that the reverse Hölder inequality holds for a large class of random variables, the so-called log-concave random variables. For example, Borell [21] showed the equivalence between the p-th and q-th moments of log-concave random variables for $q \geq p \geq 1$. It is demonstrated by Latała [93] that the constant $C(p, q)$ can be independent of p and the result also holds under semi-norm for $p \to 0$. Later Guédon [62] extended Latała's result to negative moments $p \in (-1, 0]$.

A finite Borel measure μ on $\mathbb{R}^n$ is called log-concave if we have

$$\mu(\lambda A + (1 - \lambda)B) \geq \mu(A)^\lambda \mu(B)^{1-\lambda} \tag{2.101}$$

for all $0 \leq \lambda \leq 1$ and all non-empty Borel sets $A, B \subseteq \mathbb{R}^n$. Here $\lambda A + (1 - \lambda)B$ stands for the Minkowski sum of λA and $(1 - \lambda)B$. A random variable is called log-concave if its distribution is log-concave. Log-concave distributions consist of a large class of distributions, such as Gaussian distribution, exponential distribution, and uniform distribution over any convex set. An important fact implied by Prékopa-Leindler inequality is that the sum and difference of independent log-concave random

variables are still log-concave. Therefore it is reasonable to expect reverse Hölder-type inequalities relating $aX + bY$ and $X - Y$ for i.i.d. log-concave random variables.

To prove such reverse Hölder-type inequalities, we need the following result of Guédon [62], which demonstrates the concentration phenomenon of log-concave probability measures.

Lemma 2.5.1 (Guédon [62]). *Let μ be a log-concave probability measure on $\mathbb{R}^d$, and let $K \subseteq \mathbb{R}^d$ be a symmetric convex body. For any $t \geq 1$, we have*

$$\mu((tK)^c) \leq (1 - \mu(K))^{\frac{t+1}{2}}. \tag{2.102}$$

For any $0 < t \leq 1$, we have

$$\mu(tK) \leq -2t \log(1 - \mu(K)). \tag{2.103}$$

Remark. Guédon's result (2.102) is a generalization of Borell's lemma, which says

$$\mu((tK)^c) \leq \mu(K) \left(\frac{1 - \mu(K)}{\mu(K)} \right)^{\frac{t+1}{2}}. \tag{2.104}$$

It is clear that Borell's lemma is non-trivial only when $\mu(K) > 1/2$.

Let $\|\cdot\|_1$ and $\|\cdot\|_2$ be two equivalent norms on $\mathbb{R}^n$.

Theorem 2.5.3. *Let $a, b \in \mathbb{R}$ be non-zero numbers and let $p, q \in \mathbb{R}$ such that $q \geq p > 0$ or $-1 < p \leq q < 0$. Then there is a constant $C(a, b, p, q)$ such that*

$$(\mathbb{E}\|X - Y\|_2^q)^{1/q} \leq C(a, b, p, q) \cdot (\mathbb{E}\|aX + bY\|_1^p)^{1/p} \tag{2.105}$$

holds for all i.i.d. log-concave random variables X, Y taking valued in $\mathbb{R}^n$.

Proof. If $q \geq p > 0$, we assume that $\mathbb{E}\|aX + bY\|_1^p = 1$. For $r_1 \geq 1$, Chebyshev's inequality implies

$$\mathbb{P}(\|aX + bY\|_1 \leq r_1) \geq 1 - r_1^{-p}.$$

By Theorem 2.3.5, for any $r_2 \geq 0$ we have

$$\mathbb{P}(\|X - Y\|_2 \leq r_2) \geq (1 - r_1^{-p}) N^{-1},$$

where

$$N = \frac{1}{2}\left(N(a^{-1}B_1(r_1), B_2(r_2)) + N(b^{-1}B_1(r_1), B_2(r_2))\right)$$

and we denote by $B_1(r)$, $B_2(r)$ the closed balls centered at the origin of radius r under the gauges $\|\cdot\|_1$ and $\|\cdot\|_2$, respectively. For any $s \geq 1$, Guédon's lemma (2.102) implies that

$$\mathbb{P}(\|X - Y\|_2 > r_2 s) \leq \Delta^{\frac{s+1}{2}}. \tag{2.106}$$

where $\Delta = 1 - (1 - r_1^{-p})N^{-1}$. Since $q > 0$, we have

$$\begin{aligned}\mathbb{E}\|X - Y\|_2^q &= \int_0^\infty q t^{q-1}\mathbb{P}(\|X - Y\|_2 > t)dt \\ &\leq r_2^q + \int_{r_2}^\infty q t^{q-1}\mathbb{P}\left(\|X - Y\|_2 > t\right)dt.\end{aligned}$$

Combine the estimate (2.106) with $s = tr_2^{-1}$, then elementary calculations will yield

$$\mathbb{E}\|X - Y\|_2^q \leq r_2^q\left(1 + \Gamma(q+1)\Delta^{1/2}\left(-\frac{\log\Delta}{2}\right)^{-q}\right).$$

That implies

$$(\mathbb{E}\|X - Y\|_2^q)^{1/q} \leq r_2\left(1 + \Gamma(q+1)\Delta^{1/2}\left(-\frac{\log\Delta}{2}\right)^{-q}\right)^{1/q}.$$

For $-1 < p \leq q < 0$, we assume $\mathbb{E}\|X - Y\|_2^q = 1$. For any $r_2 \in [0, 1]$, Chebyshev's inequality implies that

$$\mathbb{P}(\|X - Y\|_2 < r_2) \leq r_2^{-q}.$$

For $0 \leq s \leq 1$ and $r_1 \geq 0$, by Guédon's lemma (2.103) and Theorem 2.3.5, we have

$$\begin{aligned}\mathbb{P}(\|aX + bY\|_1 < sr_1) &\leq -2s\log\left(1 - \mathbb{P}(\|aX + bY\|_1 < r_1)\right) \\ &\leq -2s\log(1 - r_2^{-q}N),\end{aligned} \tag{2.107}$$

where N is the same as before. For $-1 < p < 0$, we have

$$\begin{aligned}\mathbb{E}\|aX + bY\|_1^p &= \int_0^\infty (-p)t^{-p-1}\mathbb{P}(\|aX + bY\|_1 < t^{-1})dt \\ &\leq r_1^p + \int_{r_1^{-1}}^\infty (-p)t^{-p-1}\mathbb{P}(\|aX + bY\|_1 < t^{-1})dt.\end{aligned}$$

Combine the estimate (2.107) with $s = (r_1 t)^{-1}$, then we have

$$\mathbb{E}\|aX + bY\|_1^p \leq r_1^p \left(1 + \frac{2p}{1+p} \log(1 - r_2^{-q} N)\right).$$

That implies

$$(\mathbb{E}\|aX + bY\|_1^p)^{1/p} \geq r_1 \left(1 + \frac{2p}{1+p} \log(1 - r_2^{-q} N)\right)^{1/p}.$$

■

Remark. There is no Hölder-type or reverse Hölder-type inequalities of the following form

$$(\mathbb{E}\|X + Y\|_1^q)^{1/q} \leq c \cdot (\mathbb{E}\|X - Y\|_2^p)^{1/p}.$$

To see this, we can take X, Y to be uniformly distributed on $[n, n+1]$.

2.5.3 Positive definite functions

In this section, we consider the estimate of $\mathbb{E}\varphi(aX + bY)$, where φ is a positive definite function. The study in the following is independent of the small ball inequalities developed in previous sections.

Let G be an abelian topological group. A Hermitian function $\varphi : G \to \mathbb{C}$ is called positive definite if, for any $x_1, \cdots, x_n \in G$ and $c_1, \cdots, c_n \in \mathbb{C}$, we have

$$\sum_{i,j=1}^{n} \varphi(x_i - x_j) c_i \overline{c_j} \geq 0. \tag{2.108}$$

Similarly the Hermitian function φ is called negative definite if the reversed inequality holds under the condition $\sum_{i=1}^n c_i = 0$. For example, for $0 < p \leq 2$, the function $e^{-\|x\|^p}$ is positive definite over the Euclidean space $\mathbb{R}^d$.

The famous Bochner's theorem asserts that a continuous positive definite function φ on a locally compact abelian group G can be uniquely represented as the Fourier transform of a positive finite Radon measure μ on the Pontryagin dual group G^*, i.e.

$$\varphi(x) = \int_{G^*} \xi(x) d\mu(\xi). \tag{2.109}$$

The counterpart of Bochner's theorem is the Lévy-Khinchin representation formula for a continuous negative definite function φ on $\mathbb{R}^d$, i.e.

$$\varphi(x) = c + i\langle y_0, x\rangle + q(x) + \int_{\mathbb{R}^d\setminus\{0\}} \left(1 - e^{-i\langle x,y\rangle} - \frac{i\langle x,y\rangle}{1+\|y\|^2}\right) d\mu(y) \tag{2.110}$$

where $c \in \mathbb{R}$, $y_0 \in \mathbb{R}^d$, $q(x)$ is some quadratic form on $\mathbb{R}^d$ and μ is a Lévy measure. The close relations between these two types of functions has been well studied. For example, a function φ is negative definite if and only if $e^{-t\varphi}$ is positive definite for all $t > 0$. This observation goes back to Schoenberg. They are also closely related to another important type of functions, the so-called completely monotone functions. We refer to [11, 12] for more details in this direction.

Theorem 2.5.4. *Let G be a locally compact abelian group and let $\varphi : G \to \mathbb{C}$ be a continuous positive definite function. For independent random variables X, Y taking values in G and $m, n \in \mathbb{Z}$, we have*

$$|\mathbb{E}\varphi(mX + nY)|^2 \leq \mathbb{E}\varphi(mX - mX')\mathbb{E}\varphi(nY - nY'), \tag{2.111}$$

where X', Y' are independent copies of X, Y, respectively.

Proof. Using Bochner's theorem (2.109), we have

$$|\mathbb{E}\varphi(mX + nY)| = \left|\mathbb{E}\int_{G^*} \xi(mX + nY)d\mu(\xi)\right| = \left|\int_{G^*} \mathbb{E}\xi(mX)\mathbb{E}\xi(nY)d\mu(\xi)\right|$$

The second equation follows from Fubini's theorem and the assumption that X, Y are

independent. By Cauchy-Schwartz inequality, we have

$$\begin{aligned}
&\leq \left(\int_{G^*} |\mathbb{E}\xi(mX)|^2 d\mu(\xi) \int_{G^*} |\mathbb{E}\xi(nY)|^2 d\mu(\xi)\right)^{1/2} \\
&= \left(\int_{G^*} \mathbb{E}\xi(mX)\overline{\mathbb{E}\xi(mX)} d\mu(\xi) \int_{G^*} \mathbb{E}\xi(nY)\overline{\mathbb{E}\xi(nY)} d\mu(\xi)\right)^{1/2} \\
&= \left(\int_{G^*} \mathbb{E}\xi(mX)\mathbb{E}\overline{\xi(mX)} d\mu(\xi) \int_{G^*} \mathbb{E}\xi(nY)\cdot\mathbb{E}\overline{\xi(nY)} d\mu(\xi)\right)^{1/2} \\
&= \left(\int_{G^*} \mathbb{E}\xi(mX)\mathbb{E}\xi(-mX) d\mu(\xi) \int_{G^*} \mathbb{E}\xi(nY)\mathbb{E}\xi(-nY) d\mu(\xi)\right)^{1/2} \\
&= \left(\int_{G^*} \mathbb{E}\xi(mX)\mathbb{E}\xi(-mX') d\mu(\xi) \int_{G^*} \mathbb{E}\xi(nY)\mathbb{E}\xi(-nY') d\mu(\xi)\right)^{1/2} \\
&= \left(\mathbb{E}\int_{G^*} \xi(mX-mX') d\mu(\xi)\mathbb{E}\int_{G^*} \xi(nY-nY') d\mu(\xi)\right)^{1/2} \\
&= \left(\mathbb{E}\varphi(mX-mX')\mathbb{E}\varphi(nY-nY')\right)^{1/2}.
\end{aligned}$$

We denote by $\overline{\mathbb{E}\xi(mX)}$ the conjugate of $\mathbb{E}\xi(mX)$, and the equation $\overline{\xi(mX)} = \xi(-mX)$ follows from the fact that $\xi \in G^*$. ■

Let V be a topological vector space over a field $\mathbb{F}$. Assume that V is locally compact. Then the following result is a consequence of the above theorem.

Corollary 2.5.2. *Let $\varphi : V \to \mathbb{C}$ be a continuous positive definite function. For independent random variables X, Y taking values in V and $a, b \in \mathbb{F}$, we have*

$$|\mathbb{E}\varphi(aX+bY)|^2 \leq \mathbb{E}\varphi(aX-aX')\mathbb{E}\varphi(bY-bY'), \tag{2.112}$$

where X', Y' are independent copies of X, Y, respectively.

Corollary 2.5.3. *Let G be a locally compact abelian group and let $\varphi : G \to \mathbb{C}$ be a continuous positive definite function. For i.i.d. random variables X, Y taking values in G, we have*

$$|\mathbb{E}\varphi(X+Y)| \leq \mathbb{E}\varphi(X-Y). \tag{2.113}$$

Remark. Let $\varphi : \mathbb{R}^d \to \mathbb{R}$ be a continuous negative definite function. For all i.i.d. random variables X, Y taking values in $\mathbb{R}^d$, Lifshits, et al [106] proved that

$$\mathbb{E}\varphi(X - Y) \leq \mathbb{E}\varphi(X + Y). \tag{2.114}$$

Their proof relies on the Lévy-Khinchin representation theorem for continuous negative definite functions. They also show that $\mathbb{E}\varphi(X+Y) - \mathbb{E}\varphi(X-Y)$ is the variance of an integrated centered Gaussian process.

Corollary 2.5.4. *Let G be a locally compact abelian group and let $\varphi : G \to \mathbb{C}$ be a continuous positive definite function. For any random variable X taking values in G, we have*

$$|\mathbb{E}\varphi(X)|^2 \leq \mathbb{E}\varphi(X - X'), \tag{2.115}$$

where X' is an independent copy of X.

Chapter 3

INFORMATION-THEORETIC INEQUALITIES

The second part devotes to the study of information theoretical inequalities analogous to sumset estimates in additive combinatorics. In particular, we consider the comparison between entropies of sum and difference of i.i.d. random variables. In Section 3.1, we show that entropies of sums (of i.i.d. random variables) are never greater than entropies of differences for random variables taking values in the cyclic group $\mathbb{Z}/3\mathbb{Z}$; however this fails for larger groups, and in particular we show that there always exist distributions on finite cyclic groups of order at least 21 such that $H(X+Y) > H(X-Y)$. In Section 3.2 and Section 3.3, we explore more quantitative questions– that is, we ask not only what the ordering of $H(X+Y)$ and $H(X-Y)$ may be, but how different these can be in either direction; the finding here is that on $\mathbb{Z}$, these can differ by any amount additively, but not too much multiplicatively. These results are closely related to the study of more-sum-than-difference (MSTD) sets in additive number theory. Finally we investigate polar codes for q-ary input channels using non-canonical kernels to construct the generator matrix, and present applications of our results to constructing polar codes with significantly improved error probability compared to the canonical construction. All the results in this part can be found in [3].

3.1 Basic examples

Let $p = (p_0, p_1, p_2)$ be a probability distribution on $\mathbb{Z}/3\mathbb{Z}$, and let $H(p)$ be its Shannon entropy. We denote by $\|p - U\|_2$ the Euclidean distance between p and the uniform distribution $U = (1/3, 1/3, 1/3)$. For any fixed $0 \leq t \leq \log 3$, the following lemma verifies the "triangular" shape of the entropy circle $H(p) = t$.

Lemma 3.1.1. *Let p be a probability distribution on the entropy circle $H(p) = t$ such that $p_0 \geq p_1 \geq p_2$. Then the distance $\|p - U\|_2$ is an increasing function of p_0.*

Proof. If $t = 0$, then p has to be the deterministic distribution $(1, 0, 0)$. In this case, we have $\|p - U\|_2 = \sqrt{2/3}$. If $t = \log 3$, we have $p = U$ and $\|p - U\|_2 = 0$. In the following, we may assume that $0 < t < \log 3$. The condition $p_0 + p_1 + p_2 = 1$ yields

$$1 + \frac{dp_1}{dp_0} + \frac{dp_2}{dp_0} = 0. \tag{3.1}$$

The entropy identity $H(p) = t$ implies

$$(\log p_0 + 1) + (\log p_1 + 1)\frac{dp_1}{dp_0} + (\log p_2 + 1)\frac{dp_2}{dp_0} = 0 \tag{3.2}$$

The above two identities give us that

$$\frac{dp_1}{dp_0} = \frac{\log p_0 - \log p_2}{\log p_2 - \log p_1} \tag{3.3}$$

and

$$\frac{dp_2}{dp_0} = \frac{\log p_0 - \log p_1}{\log p_1 - \log p_2}. \tag{3.4}$$

Using identities (3.1), (3.3) and (3.4), we have

$$\begin{aligned}
\frac{1}{2} \cdot \frac{d}{dp_0}\|p - U\|_2 &= \sum_{i=0}^{2} \left(p_i - \frac{1}{3}\right) \frac{dp_i}{dp_0} \\
&= p_0 + p_1 \frac{\log p_0 - \log p_2}{\log p_2 - \log p_1} + p_2 \frac{\log p_0 - \log p_1}{\log p_1 - \log p_2} \\
&= (p_0 - p_1)\frac{\log p_0 - \log p_2}{\log p_1 - \log p_2} - (p_0 - p_2)\frac{\log p_0 - \log p_1}{\log p_1 - \log p_2} \\
&= \frac{(p_0 - p_1)(p_0 - p_2)}{\log p_1 - \log p_2}\left(\frac{\log p_0 - \log p_2}{p_0 - p_2} - \frac{\log p_0 - \log p_1}{p_0 - p_1}\right) \\
&\geq 0
\end{aligned}$$

The last inequality follows from the assumption that $p_0 \geq p_1 \geq p_2$ and the concavity of the logarithmic function. ■

Now we can show that the entropy of the sum of two i.i.d. random variables taking values in $\mathbb{Z}/3\mathbb{Z}$ can never exceed the entropy of their difference. We use basic facts about the Fourier transform on finite groups, which can be found, e.g., in [151].

Theorem 3.1.1. *For any i.i.d. random variables X, Y taking values in $\mathbb{Z}/3\mathbb{Z}$, we have*

$$H(X+Y) \leq H(X-Y). \tag{3.5}$$

Proof. Let $p = (p_0, p_1, p_2)$ be the distribution of X. Since Y is an independent copy of X, we can see that $-Y$ has distribution $q = (p_0, p_2, p_1)$. Then the distributions of $X+Y$ and $X-Y$ can be written as $p \star p$ and $p \star q$, respectively, where '$\star$' is the convolution operation. Let $\widehat{p} = (\widehat{p}_0, \widehat{p}_1, \widehat{p}_2)$ be the Fourier transform of p with Fourier coefficients defined by

$$\widehat{p}_j = \sum_{k=0}^{2} p_k e^{-i2\pi jk/3}, \quad j = 0, 1, 2.$$

One basic property of the Fourier transform asserts that

$$\widehat{q}_j = \overline{\widehat{p}_j}, \tag{3.6}$$

where $\overline{\widehat{p}_j}$ is is the conjugate of $\widehat{p}_j$. We also have

$$(\widehat{p \star q})_j = \widehat{p}_j \cdot \widehat{q}_j, \tag{3.7}$$

which holds for general distributions q. The Parseval-Plancherel identity says

$$\|\widehat{p}\|_2^2 = 3\|p\|_2^2. \tag{3.8}$$

Using the identities (3.6), (3.7) and (3.8), we have

$$\|p \star p\|_2 = \|p \star q\|_2,$$

which implies

$$\|p \star p - U\|_2 = \|p \star q - U\|_2.$$

It is not hard to see that $X - Y$ is symmetric with $(p \star q)_0 \geq (p \star q)_1 = (p \star q)_2$. Using Lemma 3.1.1, we can see that the entropy circle passing through $p \star q$ lies inside the Euclidean circle centered at U with radius $\|p \star q - U\|_2$. Thus the distribution $p \star p$ is on an entropy circle with entropy not greater than $H(p \star q)$. Then we have the desired statement. ■

The property in Theorem 3.1.1 fails to hold for larger cyclic groups; we demonstrate this by discussing three specific examples of i.i.d. random variables X, Y such that the entropy of their sum is larger than the entropy of their difference.

1. For Conway's MSTD set $A = \{0, 2, 3, 4, 7, 11, 12, 14\}$, we have $|A + A| = 26$ and $|A - A| = 25$. Let X, Y be independent random variables uniformly distributed on A. Straightforward calculations show that
$$H(X + Y) - H(X - Y) = \frac{1}{64} \log \frac{282429536481}{215886856192} > 0.$$

2. The second example is based on the regular set $A = \{0, 1, 3, 4, 5, 6, 7, 10\}$ with $|A + A| = |A - A| = 19$. Let X, Y be independent random variables uniformly distributed on A. Then we have
$$H(X + Y) - H(X - Y) = \frac{1}{64} \log \frac{5^{10} \cdot 8^{10}}{3^6 \cdot 7^7} > 0.$$

3. The group $\mathbb{Z}/12\mathbb{Z}$ is the smallest cyclic group that contains a MSTD set. Let $A = \{0, 1, 2, 4, 5, 9\}$. It is easy to check that A is a MSTD set since $A + A = \mathbb{Z}_{12}$ and $A - A = (\mathbb{Z}/12\mathbb{Z})\backslash\{6\}$. We let X, Y be independent random variables uniformly distributed on A. Then we have
$$H(X + Y) - H(X - Y) = \frac{1}{36} \log \frac{3^{34}}{20^{10}} > 0.$$

Remark. Applying linear transformations, we can get infinitely many MSTD sets of $\mathbb{Z}$ from Conway's MSTD set. Correspondingly, one can get as many MSTD random variables as one please. The second example shows that one can always find MSTD random variables taking values in $\mathbb{Z}/n\mathbb{Z}$ for $n \geq 21$. These examples show that MSTD sets are useful in the construction of MSTD random variables. However we can indeed get MSTD random variables supported on non-MSTD sets as shown by the second example.

Remark. Hegarty [68] proved that there is no MSTD set in $\mathbb{Z}$ of size 7 and, up to linear transformations, Conway's set is the unique MSTD set in $\mathbb{Z}$ of size 8. We do not know the smallest support of MSTD random variables taking values in $\mathbb{Z}$.

3.2 Achievable differences

In the following, we briefly introduce the construction of Stein [152] of finite subsets $A_k \subset \mathbb{Z}$ such that the ratio $|A_k - A_k|/|A_k + A_k|$ can be arbitrarily large or small when k is large. Using this construction we will give an alternate proof of the result of Lapidoth and Pete [92], which asserts that $H(X-Y)$ can exceed $H(X+Y)$ by an arbitrarily large amount.

Let $A, B \subset \mathbb{Z}$ be two finite subsets. Suppose that the gap between any two consecutive elements of B is sufficiently large. For any $b \in B$, the set $b + A$ represents a relatively small fluctuation around b. Large gaps between elements of B will imply that $(b+A)\cap(b'+A) = \emptyset$ for distinct $b, b' \in B$. Then we will have $|A+B| = |A||B|$. For $m \in \mathbb{Z}$ large, this argument implies that $|A+m\cdot A| = |A|^2$, where $m\cdot A := \{ma : a \in A\}$. Therefore, the following equations hold simultaneously for sufficiently large $m_0 \in \mathbb{Z}$, which depends on $A, A-A$ and $A+A$,

$$|A + m_0 \cdot A| = |A|^2,$$

$$|(A + m_0 \cdot A) - (A + m_0 \cdot A)| = |(A - A) + m_0 \cdot (A - A)| = |A - A|^2,$$

and

$$|(A + m_0 \cdot A) + (A + m_0 \cdot A)| = |A + A|^2.$$

Repeating this argument, we can get a sequence of sets A_k, defined by

$$A_k = A_{k-1} + m_{k-1}A_{k-1}, \tag{3.9}$$

where $A_0 = A$, $m_{k-1} \in \mathbb{Z}$ sufficiently large, with the following properties

$$|A_k| = |A|^{2k}, \quad |A_k \pm A_k| = |A \pm A|^{2k}. \tag{3.10}$$

Now we are ready to reprove the result of Lapidoth and Pete [92].

Theorem 3.2.1. *[92] For any $M > 0$, there exists i.i.d. $\mathbb{Z}$-valued random variables X, Y with finite entropy such that*

$$H(X-Y) - H(X+Y) > M.$$

Proof. Recall the following basic property of Shannon entropy

$$0 \leq H(X) \leq \log |\text{range of } X|. \tag{3.11}$$

We let X_k, Y_k be independent random variables uniformly distributed on the set A_k obtained by the iteration equation (3.9). Using the right hand side of (3.11) and the properties given by (3.10), we have

$$H(X_k + Y_k) \leq \log |A_k + A_k| = 2k \log |A + A|. \tag{3.12}$$

Since X_k, Y_k are independent and uniform on A_k, for all $x \in A_k - A_k$, we have

$$\mathbb{P}(X_k - Y_k = x) \geq |A_k|^{-2}.$$

Notice the fact that $-t \log t$ is increasing over $(0, 1/e)$. When k is large enough, we have

$$\begin{aligned} H(X_k - Y_k) &\geq \frac{|A_k - A_k|}{|A_k|^2} \log |A_k|^2 \\ &= 4k \log |A| \left(\frac{|A - A|}{|A|^2} \right)^{2k}. \end{aligned} \tag{3.13}$$

For any $k \in \mathbb{Z}^+$, we can always find a set $A \subset \mathbb{Z}$ with k^2 elements such that the set $A - A$ achieves the possible maximal cardinality,

$$|A| = k^2, \ |A - A| = |A|^2 - |A| + 1. \tag{3.14}$$

Combining (3.12), (3.14) and the trivial bound

$$|A + A| \leq \frac{|A|(|A| + 1)}{2},$$

we have that for k large

$$\begin{aligned} H(X_k + Y_k) &\leq 2k \log \frac{|A|(|A| + 1)}{2} \\ &= 8k \log k - 2k \log 2 + 2k \log(1 + k^{-2}) \\ &= 8k \log k - 2k \log 2 + o(1). \end{aligned}$$

Combining (3.13) and (3.14), we have

$$\begin{aligned} H(X_k - Y_k) &\geq 8k \log k \left(1 - k^{-2} + k^{-4}\right)^{2k} \\ &= 8k \log k \exp(2k(-k^{-2} + O(k^{-4}))) \\ &= 8k \log k (1 - 2k^{-1} + O(k^{-2})) \\ &= 8k \log k - 16 \log k + o(1). \end{aligned}$$

Therefore we have

$$H(X_k - Y_k) - H(X_k + Y_k) = 2k \log 2 - 16 \log k + o(1).$$

Then the statement follows from that k can be arbitrarily large. ■

We observe that the following complementary result is also true.

Theorem 3.2.2. *For any $M > 0$, there exists i.i.d. $\mathbb{Z}$-valued random variables X, Y with finite entropy such that*

$$H(X + Y) - H(X - Y) > M.$$

Remark. The previous argument can not be used to prove this result. If we proceed the same argument, we will see that the lower bound of $H(X_k + Y_k)$ similar to (3.13) will be really bad. The reason is that

$$\left(\frac{|A + A|}{|A|^2}\right)^{2k} \to 0$$

exponentially fast. Both results can also be proved using a probabilistic construction of Ruzsa [141] on the existence of large additive sets A with $|A - A|$ very close to the maximal value $|A|^2$, but $|A + A| \leq n^{2-c}$ for some explicit absolute constant $c > 0$; and similarly with the roles of $A - A$ and $A + A$ reversed.

In fact, we have the following stronger result.

Theorem 3.2.3. *For any $M \in \mathbb{R}$, there exist i.i.d $\mathbb{Z}$-valued random variables X, Y with finite entropy such that*

$$H(X + Y) - H(X - Y) = M.$$

Proof. Let X be a random variable taking values in $\{0, 1, \cdots, n-1\} \subset \mathbb{Z}$. Then $H(X+Y) - H(X-Y)$ is a continuous function of n variables. We can assume that n is large enough if necessary. From the discussion in Section 3.1, we know that this function can take both positive and negative values. (For instance Theorem 3.1.1 implies that a binary distribution can give us negative difference, and the uniform distribution on Conway's MSTD set will yield positive difference). Since the function is continuous, the intermediate value theorem implies that its range must contain an open interval (a, b) with $a < 0 < b$. Let $X_1, \cdots X_k$ be k independent copies of X and we define $X' = (X_1, \cdots, X_k)$. Let Y' be an independent copy of X'. Then we have

$$H(X'+Y') - H(X'-Y') = k(H(X+Y) - H(X-Y)).$$

The range of $H(X'+Y')-H(X'-Y')$ will contain (ka, kb). The right hand side can take any real number since k can be arbitrarily large. The random variables X', Y' take finite values of $\mathbb{Z}^k$. Using the linear transformation $(x_1, \cdots, x_k) \to x_1 + dx_2 + \cdots + d^{k-1}x_k$, we can map X, Y to $\mathbb{Z}$-valued random variables. This map preserves entropy as d is large enough. So these $\mathbb{Z}$-valued random variables will have the desired property. ■

Recall that, for a continuous random variable X with the density function $f(x)$, the differential entropy $h(X)$ is defined by

$$h(X) = -\mathbb{E} \log f(X). \tag{3.15}$$

Theorem 3.2.4. *For any $M \in \mathbb{R}$, there exist i.i.d. real-valued random variables X, Y with finite differential entropy such that*

$$h(X+Y) - h(X-Y) = M. \tag{3.16}$$

Proof. From the above theorem we know that there exist $\mathbb{Z}$-valued random variables X', Y' with the desired property. Let U, V be independent random variables uniformly distributed on $(-1/4, 1/4)$, which are also independent of (X', Y'). Then we define $X = X' + U$ and $Y = Y' + V$. Elementary calculations will show that

$$h(X+Y) = H(X'+Y') + h(U+V),$$

and

$$h(X-Y) = H(X'-Y') + h(U-V).$$

Since U, V are symmetric, $U+V$ and $U-V$ have the same distribution. Therefore, we have

$$h(X+Y) - h(X-Y) = H(X'+Y') - H(X'-Y').$$

Then the theorem follows. ■

Remark. In the set cardinality setting, Nathanson [127] raised the question: what are the possible values of $|A+A|-|A-A|$ for finite subsets $A \subset \mathbb{Z}$? Martin and O'Bryant [114] proved that for any $k \in \mathbb{Z}$ there exists A such that $|A+A|-|A-A| = k$, that is independently obtained by Hegarty [68].

3.3 Entropy analogue of Freiman-Pigarev inequality

We proved that the entropies of the sum and difference of two i.i.d. random variables can differ by an arbitrarily large amount additively. However we will show that they do not differ too much multiplicatively.

In additive combinatorics, for a finite additive set A, the doubling constant $\sigma[A]$ is defined as

$$\sigma[A] = \frac{|A+A|}{|A|}. \tag{3.17}$$

Similarly the difference constant $\delta[A]$ is defined by

$$\delta[A] = \frac{|A-A|}{|A|}. \tag{3.18}$$

It was first observed by Ruzsa [140] that

$$\delta[A]^{1/2} \leq \sigma[A] \leq \delta[A]^3. \tag{3.19}$$

The upper bound can be improved down to $\delta[A]^2$ using Plünnecke inequalities. Thus a finite additive set has small doubling constant if and only if its difference constant is also small. In the entropy setting, we have

$$\frac{1}{2} \leq \frac{H(X+Y) - H(X)}{H(X-Y) - H(X)} \leq 2 \tag{3.20}$$

for i.i.d. random variables X, Y. The upper bound was proved by Madiman [107] and the lower bound was proved independently by Ruzsa [143] and Tao [158]. The inequalities also hold for differential entropy, see Madiman and Kontoyiannis [109]. In other words, after subtraction of $H(X)$, the entropies of the sum and the difference of two i.i.d. random variables are not too different. We observe that the entropy version (3.20) of the doubling-difference inequality implies the entropy analogue of the following result proved by Freiman and Pigarev [133]:

$$|A - A|^{3/4} \leq |A + A| \leq |A - A|^{4/3}. \tag{3.21}$$

Theorem 3.3.1. *Let X, Y be i.i.d. discrete random variables with finite entropy, then we have*

$$\frac{3}{4} < \frac{H(X+Y)}{H(X-Y)} < \frac{4}{3}. \tag{3.22}$$

Proof. The basic facts of Shannon entropy (1.5) and (1.6) imply that $H(X+Y) = 0$ if and only if $H(X-Y) = 0$. In this case, the above theorem is true if we define $0/0 = 1$. So we assume that neither $H(X+Y)$ nor $H(X-Y)$ is 0. For the upper bound, we have

$$\begin{aligned} \frac{H(X+Y)}{H(X-Y)} &= \frac{H(X+Y)}{H(X-Y) - H(X) + H(X)} \\ &\leq \frac{H(X+Y)}{(H(X+Y) - H(X))/2 + H(X)} \\ &= \frac{2H(X+Y)}{H(X+Y) + H(X)} \\ &< \frac{4}{3} \end{aligned}$$

The second step follows from the upper bound in (3.20) and the fact that Shannon entropy is non-negative. The last step uses (1.5) and the fact that, in the i.i.d. case, " $=$ " of the upper bound happens only when X takes on a single value, i.e. $H(X) = 0$. The lower bound can be proved in a similar way. ■

Remark. It is unknown if the inequality (3.20) is best possible. Suppose that, for some $\alpha \in (1,2)$, we have

$$\alpha^{-1} \leq \frac{H(X+Y)-H(X)}{H(X-Y)-H(X)} \leq \alpha.$$

Using the same argument, the above theorem can be improved to

$$\frac{\alpha+1}{2\alpha} < \frac{H(X+Y)}{H(X-Y)} < \frac{2\alpha}{\alpha+1}.$$

Remark. The above theorem does not hold for continuous random variables. For example, let X be an exponential random variable with parameter λ, and Y be an independent copy of X. Then $X+Y$ satisfies the Gamma distribution $\Gamma(2,\lambda^{-1})$ with the differential entropy

$$h(X+Y) = 1+\gamma-\log\lambda \approx 1.577-\log\lambda,$$

where γ is the Euler's constant. On the other hand, $X-Y$ has the Laplace distribution Laplace$(0,\lambda^{-1})$ with the differential entropy

$$h(X-Y) = 1+\log 2-\log\lambda \approx 1.693-\log\lambda.$$

We can see that

$$\lim_{\lambda\to(2e)^+} \frac{h(X+Y)}{h(X-Y)} = \infty,$$

and

$$\lim_{\lambda\to(2e)^-} \frac{h(X+Y)}{h(X-Y)} = -\infty.$$

3.4 Applications to polar codes

3.4.1 Introduction to polar codes

Polar codes, invented by Arıkan [6] in 2009, achieve the capacity of arbitrary binary-input symmetric discrete memoryless channels. Moreover, they have low encoding and decoding complexity and an explicit construction. Consequently they have attracted a great deal of attention in recent years. In order to discuss polar codes more precisely, we now recall some standard terminology from information and coding theory.

As a standard practice in information theory, we use U^k to denote $(U_1, \ldots, U_k)$, and $I(X;Y|Z)$ to denote the conditional mutual information between X and Y given Z, which is defined by

$$I(X;Y|Z) = H(X,Z) + H(Y,Z) - H(X,Y,Z) - H(Z).$$

It is well known, and also trivial to see, that the conditional entropy $H(X|Y)$, defined as the mean using the distribution of Y of $H(X|Y = y)$, satisfies the "chain rule" $H(Y) + H(X|Y) = H(X,Y)$, so that $I(X;Y|Z) = H(X|Z) - H(X|Y,Z)$. The mutual information between X and Y, namely $I(X;Y) = H(X) - H(X|Y)$, emerges in the case where there is no conditioning. In particular, $I(X;Y|Z) = 0$ if and only if X and Y are conditionally independent given Z. Furthermore, one also has the chain rule for mutual information, which states that $I(X;Y,Z) = I(X;Z) + I(X;Y|Z)$.

A major goal in coding theory is to obtain efficient codes that achieve the Shannon capacity on a discrete memoryless channel. A memoryless channel is defined first by a "one-shot" channel W, which is a stochastic kernel from an input alphabet $\mathcal{X}$ to an output alphabet $\mathcal{Y}$ (i.e., for each $x \in \mathcal{X}$, $W(\cdot|x)$ is a probability distribution on $\mathcal{Y}$), and the memoryless extension of W for length n vectors is defined by

$$W^{(n)}(y^n|x^n) = \prod_{i=1}^{n} W(y_i|x_i), \quad x^n \in \mathcal{X}^n, y^n \in \mathcal{Y}^n. \tag{3.23}$$

To simplify the notation, one often makes a slight abuse of notation, writing $W^{(n)}$ as W.

A linear code of block length n on an alphabet $\mathcal{X} = \mathbb{F}$ (which must be a field) is a subspace of $\mathbb{F}^n$. The vectors in the subspace are often called the codewords. A linear code is equivalently defined by a generator matrix, i.e., a matrix with entries in the field whose rows form a basis for the code. If the dimension of the code is k, and if G is a $k \times n$ generator matrix for the linear code, the codewords are given by the span of the rows of G, i.e., all multiplications uG where u is a $1 \times k$ vector over the field. We refer to [33, 138] for a more detailed introduction to information and coding theory.

In polar codes, the generator matrix of block length n is obtained by deleting[1] some rows of the matrix $G_n = \left[\begin{smallmatrix} 1 & 0 \\ 1 & 1 \end{smallmatrix}\right]^{\otimes \log_2 n}$. Which rows to delete depends on the channel and the targeted error probability (or rate). For a symmetric discrete memoryless channel W, the rows to be deleted are indexed by

$$\mathcal{B}_{\epsilon,n} := \{i \in [n] : I(U_i; Y^n U^{i-1}) \leq 1 - \epsilon\}, \tag{3.24}$$

where ϵ is a parameter governing the error probability, the vector U^n has i.i.d. components which are uniform on the input alphabet, $X^n = U^n G_n$, and Y^n is the output of n independent uses of W when X^n is the input.

To see the purpose of the transform G_n, consider the case $n = 2$ first. Applying G_2 to the vector (U_1, U_2) yields

$$X_1 = U_1 + U_2,$$
$$X_2 = U_2.$$

Transmitting X_1 and X_2 on two independent uses of a binary input channel W leads to two output variables Y_1 and Y_2; recall that this means that Y_1 (or Y_2) is a random variable whose distribution is given by $W(\cdot|x)$ where x is the realization of X_1 (or X_2). If we look at the mutual information between the vectors $X^2 = (X_1, X_2)$ and $Y^2 = (Y_1, Y_2)$, since the pair of components (X_1, Y_1) and (X_2, Y_2) are mutually independent, the chain rule yields

$$I(X^2; Y^2) = I(X_1; Y_1) + I(X_2; Y_2) = 2I(W), \tag{3.25}$$

where $I(W)$ is defined as the mutual information of the one-shot channel W with a uniformly distributed input. Further, since the transformation G_2 is one-to-one, and since the mutual information is clearly invariant under one-to-one transformations of

[1] If the channel is symmetric the generator matrix is indeed obtained by deleting rows, otherwise in addition to deleting rows one may also have to translate the codewords (affine code).

its arguments (the mutual information depends only on the joint distribution of its arguments), we have that

$$I(U^2;Y^2) = I(X^2;Y^2). \tag{3.26}$$

If we now apply the chain rule to the left hand side of previous equality, the dependencies in the components of U^2 obtained by mixing X^2 with G_2 lead this time to two different terms, namely,

$$I(U^2;Y^2) = I(U_1;Y^2) + I(U_2;Y^2,U_1). \tag{3.27}$$

Putting back (3.25), (3.26), and (3.27) together, we have that

$$I(W) = \frac{1}{2}\left(I(U_1;Y^2) + I(U_2;Y^2,U_1)\right). \tag{3.28}$$

Now, the above is interesting because the two terms in the right-hand side are precisely not equal. In fact, $I(U_2;Y^2,U_1)$ must be greater than its counter-part without the mixing of G_2, i.e., $I(U_2;Y^2,U_1) \geq I(X_2;Y_2) = I(W)$. To see this, note that

$$\begin{aligned} I(U_2;Y^2,U_1) &= H(U_2) - H(U_2|Y^2,U_1) \\ &\geq H(U_2) - H(U_2|Y^2) \\ &= H(X_2) - H(X_2|Y_2) \\ &= I(X_2;Y_2) \end{aligned}$$

where the inequality above uses the fact that conditioning can only reduce entropy, hence dropping the variable U_1 in $H(U_2|Y^2,U_1)$ can only increase the entropy. Further, one can check that besides for degenerated cases where W is deterministic or fully noisy (i.e., making input and output independent), $I(U_2;Y^2,U_1)$ is strictly larger than $I(X_2;Y_2)$. Thus, the two terms in the right-hand side of (3.28) are respectively lesser and greater that $I(W)$, but they average out to the original amount $I(W)$.

In summary, out of two independent copies of the channel W, the transform G_2 allows us to create two new synthetic channels

$$\begin{aligned} W^- &: U_1 \to Y_1, Y_2 \\ W^+ &: U_2 \to Y_1, Y_2, U_1 \end{aligned}$$

that have respectively a worse and better mutual information

$$I(W^-) \leq I(W) \leq I(W^+).$$

while overall preserving the total amount of mutual information

$$I(W) = \frac{1}{2}(I(W^+) + I(W^-)).$$

The key use of the above phenomena, is that if one wants to transmit only one bit (uniformly drawn), using W^+ rather than W leads to a lower error probability since the channel W^+ carries more information. One can then iterate this argument several times and hope obtaining a subset of channels of very high mutual information, on which bits can be reliably transmitted. After $\log_2 n$ iterations, one obtains the synthesized channels $U_i \mapsto (Y^n, U^{i-1})$. Thus, for a given number of information bits to be transmitted (i.e., for a given rate), one can select the channels with the largest mutual informations to minimize the error probability. As explained in the next section, the phenomenon of *polarization* happens in the sense that as n tends to infinity, the synthesized channels have mutual information tending to either 0 or 1 (besides for a vanishing fraction of exceptions). Hence, sending information bits through the high mutual information channels (equivalently, deleting rows of G_n corresponding to low mutual information channels) allows one to achieve communication rates as large as the mutual information of the original binary input channel. The construction extends to q-ary input alphabets when q is prime, using the same matrix $G_n = \left[\begin{smallmatrix} 1 & 0 \\ 1 & 1 \end{smallmatrix}\right]^{\otimes \log_2 n}$, while carrying the operations over $\mathbb{F}_q$.

It is tempting to investigate what happens if one keeps the Kronecker structure of the generator matrix but modifies the kernel $\left[\begin{smallmatrix} 1 & 0 \\ 1 & 1 \end{smallmatrix}\right]$. For binary input alphabets, there is no other interesting choice (up to equivalent permutations). In Mori and Tanaka [124], the error probability of non-binary polar codes constructed on the basis of Reed-Solomon matrices is calculated using numerical simulations on q-ary erasure channels. It is confirmed that 4-ary polar codes can have significantly better performance than binary polar codes. Our goal here is to investigate potential improvements at *finite block*

length using modified kernels over $\mathbb{F}_q$. We propose to pick kernels not by optimizing the polar code exponent as in [124] but by maximizing the polar martingale spread. This connects to the object of study in this paper, as explained next. The resulting improvements are illustrated with numerical simulations.

3.4.2 Polar martingale

In order to see that polarization happens, namely that

$$\frac{1}{n}|\{i \in [n] : I(U_i; Y^n, U^{i-1}) \in (\epsilon, 1-\epsilon)\}| \to 0, \tag{3.29}$$

it is helpful to rely on a random process having a uniform measure on the possible realizations of $I(U_i; Y^n U^{i-1})$. Then, counting the number of such mutual informations in $(\epsilon, 1-\epsilon)$ can be obtained by evaluating the probability that the process lies in this interval. The process is defined by taking $\{B_n\}_{n\geq 1}$ to be i.i.d. random variables uniform on $\{-,+\}$ and the binary (or q-ary with q prime) random input channels $\{W_n,\ n \geq 0\}$ are defined by

$$\begin{aligned} W_0 &:= W, \\ W_n &:= W_{n-1}^{B_n}, \quad \forall n \geq 1. \end{aligned} \tag{3.30}$$

Then the polarization result can be expressed as

$$\mathbb{P}\{I(W_n) \in (\epsilon, 1-\epsilon)\} \to 0. \tag{3.31}$$

The process $I(W_n)$ is particularly handy as it is a bounded martingale with respect to the filtration B_n. This is a consequence of the balance equation derived in (3.28). Therefore, $I(W_n)$ converges almost surely, which means that almost surely, for any $\epsilon > 0$ and n large enough, $|I(W_{n+1}) - I(W_n)| = I(W_n^+) - I(W_n) < \epsilon$. Since for q-ary input channels (q prime), the only channels for which $I(W^+) - I(W)$ is arbitrarily small is when $I(W)$ is arbitrarily close to 0 or 1, the conclusion of polarization follows. The key point is that the martingale $I(W_n)$ is a random walk in $[0,1]$ and it is 'unstable at any point $I(W) \in (0,1)$ as it must move at least $I(W^+) - I(W) > 0$ in this range. The plot of $I(W^+) - I(W) > 0$ for different values of $I(W)$ is provided in Figure 3.1.

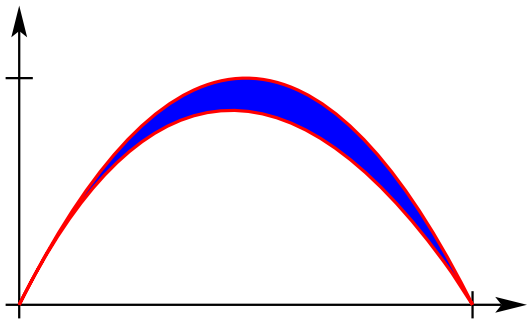

Figure 3.1: Plot of $I(W)$ (horizontal axis) vs. $I(W^+) - I(W)$ for all possible binary input channels (the tick on the horizontal axis is at 1 and the tick on vertical axis is at 1/4).

Thus, the larger the spread $I(W^+) - I(W)$, the more unstable the martingale is at non-extremal points and the faster it should converge to the extremes (i.e., polarized channels). To see why this is connected to the object of study of this paper, we need one more aspect about polar codes.

When considering channels that are 'additive noise', polarization can be understood in terms of the noise process rather than the actual channels W_n. Consider for example the the binary symmetric channel. When transmitting a codeword c^n on this channel, the output is $Y^n = c^n + Z^n$, where Z^n has i.i.d. Bernoulli components. The polar transform can then be carried over the noise Z^n. Since

$$I(U_i; Y^n U^{i-1}) = 1 - H((G_n Z^n)_i | (G_n Z^n)^{i-1}), \tag{3.32}$$

the mutual information of the polarized channels is directly obtained from the conditional entropies of the polarized noise vector $G_n Z^n$. The counter-part of this polarization phenomenon is called source polarization [7]. It is extended in [2] to multiple correlated source. For $n = 2$, the spread of the two conditional entropies is exactly given by $H(Z + Z') - H(Z)$, where Z, Z' are i.i.d. under the noise distribution. In Arıkan and Telatar [8], the rate of convergence of the polar martingale is studied as a function of the block length. Our goal here is to investigate the performance at

finite block length, motivated by maximizing the spread at block length $n = 1$. When considering non-binary polar codes, that spread is governed by the entropy of a linear combination of i.i.d. variables. Preliminary results on this approach were presented in [1], while the error exponent and scaling law of polar codes have been studied in particular in [67] and references therein.

3.4.3 Kernels with maximal spread

Being interested in the performance of polar codes at finite block length, we start with the optimization of the kernel matrix over $\mathbb{F}_q$ of block length $n = 2$. Namely, we investigate the following optimization problem:

$$K^*(W) = \arg \max_{K \in M_2(\mathbb{F}_q)} I(W^+(W, K)), \tag{3.33}$$

where $W^+(W, K)$ is the channel $u_2 \mapsto Y_1 Y_2 u_1$, and (Y_1, Y_2) are the output of two independent uses of W when $(x_1, x_2) = (u_1, u_2)K$ are the inputs. We call K^* the 2-optimal kernel for W.

A general kernel is a 2×2 invertible matrix over $\mathbb{F}_q$. Let $K = \left[\begin{smallmatrix} a & b \\ c & d \end{smallmatrix}\right]$ be such a matrix and let (U_1, U_2) be i.i.d. under μ over $\mathbb{F}_q$ and $(X_1, X_2) = (U_1, U_2)K$. Since K is invertible, we have

$$2H(\mu) = H(U_1, U_2) = H(X_1, X_2) = H(X_1) + H(X_2|X_1) \tag{3.34}$$

and

$$H(X_1) - H(\mu) = H(\mu) - H(X_2|X_1) \tag{3.35}$$

which is the entropy spread gained by using the transformation K. To maximize the spread, one may maximize $H(X_1) = H(aU_1 + cU_2)$ over the choice of a and c, or simply $H(U_1 + cU_2)$ over the choice of c. Hence, the maximization problem depends only on the variable c, (a can be set to 1, and b, d only need to ensure that K is invertible), which leads to a kernel of the form $K = \left[\begin{smallmatrix} 1 & 0 \\ c & 1 \end{smallmatrix}\right]$. Note that to maximize the spread, one may alternatively minimize $H(X_2|X_1) = H(U_2|U_1 + cU_2)$.

We consider in particular channels which are 'additive noise', in which case one can equivalently study the 'source' version of this problem as follows:

$$\lambda^*(\mu) = \arg\max_{\lambda \in \mathbb{F}_q} H(U_1 + \lambda U_2), \tag{3.36}$$

where U_1, U_2 are i.i.d. under μ. As discussed above, this is related with the previous problem by choosing

$$K^*(W) = \begin{bmatrix} 1 & 0 \\ \lambda^*(\mu) & 1 \end{bmatrix},$$

where μ is the distribution of the noise of the channel W.

Corollary 3.4.1. *For a probability distribution μ over $\mathbb{F}_3$,*

$$\lambda^*(\mu) = 2$$

if $\mu(1) \neq \mu(2)$, and $\lambda^*(\mu) = \{1, 2\}$ if $\mu(1) = \mu(2)$.

Figure 3.2 illustrates the improvements of the error probability of a polar code using the kernel $\left[\begin{smallmatrix} 1 & 0 \\ 2 & 1 \end{smallmatrix}\right]$ instead of $\left[\begin{smallmatrix} 1 & 0 \\ 1 & 1 \end{smallmatrix}\right]$ for a block length $n = 1024$ when the channel is an additive noise channel over $\mathbb{F}_3$ with noise distribution $\{0.7, 0.3, 0\}$.

When μ is over $\mathbb{F}_q$ with $q \geq 5$, $\lambda^*(\mu)$ varies with μ. For example, one can check numerically that for the distribution $\{0.8, 0.1, 0.1, 0, 0\}$ we have $\lambda^* = 4$, whereas for the distribution $\{0.7, 0.2, 0.1, 0, 0\}$ we have $\lambda^* = \{2, 3\}$. Thus finding a solution to the problem of determining $\lambda^*(\mu)$ for general probability distributions μ on $\mathbb{F}_q$ seems not so easy. Nonetheless, for a certain class of probability distributions μ, we can identify $\lambda^*(\mu)$ explicitly using the following observation.

Proposition 3.4.1. *Let μ be a probability distribution over $\mathbb{F}_q$ with support S_μ. If there exists $\gamma \in \mathbb{F}_q$ such that*

$$|S_\mu + \gamma S_\mu| = |S_\mu|^2 \tag{3.37}$$

then

$$H(U_2 | U_1 + \gamma U_2) = 0 \tag{3.38}$$

where U_1, U_2 are i.i.d. under μ.

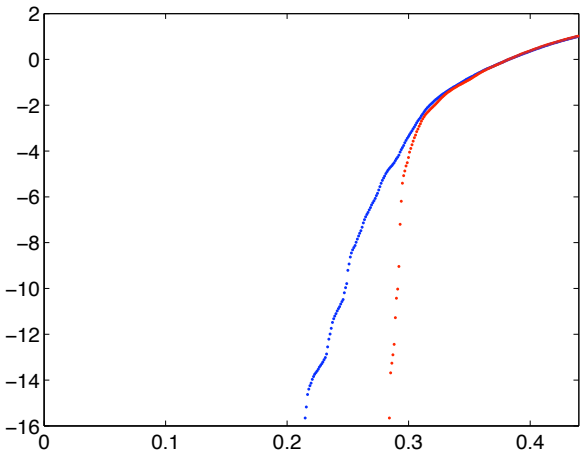

Figure 3.2: Block error probability (in $\log_{10}$ scale) of a polar code using the 2-optimal kernel (red curve – lower curve) vs. original kernel (blue curve) for a block length of $n = 1024$ and an additive noise channel over $\mathbb{F}_3$ with noise distribution $\{0.7, 0.3, 0\}$.

Proof. The condition $|S_\mu + \gamma S_\mu| = |S_\mu|^2$ ensures that knowing $u_1 + \gamma u_2$ with $u_1, u_2 \in S_\mu$ allows to exactly recover both u_1 and u_2. ■

Remark. The condition on the support could be simplified but as such it makes the conclusion of Proposition 3.4.1 immediate. Also note that γ such that $H(U_2|U_1+\gamma U_2) = 0$ is clearly optimal to maximize the spread, i.e., it maximizes $H(U_1 + \gamma U_2)$.

Let us consider some examples of distributions satisfying (3.37):

1. Let μ over $\mathbb{F}_5$ be such that $S_\mu = \{0, 1\}$. Picking $\gamma = 2$, one obtains $2S_\mu = \{0, 2\}$ and $S_\mu + 2S_\mu = \{0, 1, 2, 3\}$, and (3.37) is verified. In this case, using $\gamma = 1$ can only provide a strictly smaller spread since it will not set $H(U_2|U_1 + \gamma U_2) = 0$. It is hence better to use the 2-optimal kernel $\left[\begin{smallmatrix} 1 & 0 \\ 2 & 1 \end{smallmatrix}\right]$ rather than the original kernel $\left[\begin{smallmatrix} 1 & 0 \\ 1 & 1 \end{smallmatrix}\right]$. As illustrated in Figure 3.3, this leads to significant improvements in the error probability at finite block length. Also note that a channel with noise μ satisfying (3.37) has positive zero-error capacity, which is captured by the 2-optimal kernel as shown with the rapid drop of the error probability (it is 0 at low enough rates since half of the synthesized channels have noise entropy exactly zero). If μ is close to a distribution satisfying (3.37), the error probability can also be significantly improved with respect to the original kernel $\left[\begin{smallmatrix} 1 & 0 \\ 1 & 1 \end{smallmatrix}\right]$.

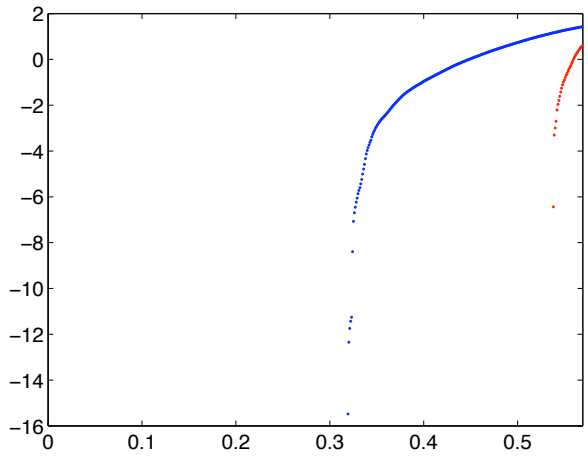

Figure 3.3: Block error probability (in $\log_{10}$ scale) of a polar code using the 2-optimal kernel (red curve – lower curve) vs. original kernel (blue curve) for a block length of $n = 1024$ and an additive noise channel over $\mathbb{F}_5$ with noise distribution $\{0.5, 0.5, 0, 0, 0\}$. This channel takes any symbol of $\mathbb{F}_5$ to itself with probability $1/2$ and shifts any symbol circularly with probability $1/2$.

2. Over $\mathbb{F}_{11}$, let μ be such that $S_\mu = \{0, 1, 2\}$. Picking $\gamma = 2$, one obtains $2S_\mu = \{0, 2, 4\}$ and (3.37) does not hold. However, picking $\gamma = 3$ leads to $3S_\mu = \{0, 3, 6\}$ and (3.37) holds. Therefore, the choice of γ varies with respect to q.

3. Over general $\mathbb{F}_q$, let $k = \lfloor \sqrt{q-1} \rfloor$. If $S_\mu = \{0, 1, \ldots, k-1\}$, we can see that $\gamma = k$ will satisfy (3.37).

In conclusion, we have shown that over $\mathbb{F}_q$, the martingale spread can be significantly enlarged by using 2-optimal kernels rather than the original kernel $\left[\begin{smallmatrix} 1 & 0 \\ 1 & 1 \end{smallmatrix}\right]$. Moreover, we have observed that this can lead to significant improvements on the error probability of polar codes, even at low block length ($n = 1024$). For additive noise channels, while the improvement is significant when the noise distribution is concentrated on "small" support, the improvement may not be as significant for distributions that are more more spread out.

Chapter 4

CONCENTRATION OF INFORMATION CONTENT

This chapter devotes to the study of concentration properties of information content. Sharp exponential deviation estimates for the information content as well as a sharp bound on the varentropy are obtained for convex probability measures on Euclidean spaces. These provide, in a sense, a nonasymptotic equipartition property for convex measures even in the absence of stationarity-type assumptions. In Section 4.1, we show that the exponential deviation of a functional follows from the log-concavity of normalized moments of that functional. The log-concavity of the normalized moments of s-concave functions are studied in Section 4.2. Optimal concentration and sharp variance bound of the information content of κ-concave random variables are obtained in Section 4.3 by combining the results of the preceding two sections. All the results can be found in [52], which will form one part of a larger paper [13].

4.1 General principle for exponential deviation

Let X be a random variable taking values in $\mathbb{R}^n$. Suppose that it has density f with respect to the Lebesgue measure on $\mathbb{R}^n$. Let $\varphi : \mathbb{R}^n \to \mathbb{R}$ be a real-valued function. One natural way to show the exponential deviation of $\varphi(X)$ from its mean is to prove the finiteness of the moment generating function $\mathbb{E}e^{\alpha\varphi(X)}$ for certain α. The logarithmic moment generating function $L(\alpha)$ is defined by

$$L(\alpha) = \log \mathbb{E}e^{\alpha\varphi(X)}. \tag{4.1}$$

The following observation is a well known fact about exponential families in statistics.

Lemma 4.1.1. *Let $a, b > 0$ be certain positive real numbers. Suppose that $L(\alpha) < \infty$ for $\alpha \in (-a, b)$. Then we have*

$$L'(\alpha) = \mathbb{E}\varphi(X_\alpha), \quad L''(\alpha) = Var(\varphi(X_\alpha)), \tag{4.2}$$

where X_α is a random variable with density

$$f_\alpha(x) = \frac{e^{\alpha\varphi(x)} f(x)}{\int_{\mathbb{R}^n} e^{\alpha\varphi(x)} f(x) dx}. \tag{4.3}$$

For $\alpha = 0$, one particularly has $L'(0) = \mathbb{E}\varphi(X)$ and $L''(0) = Var(\varphi(X))$.

Proof. The assumption $L(\alpha) < \infty$ for $\alpha \in (-a, b)$ guarantees that $L(\alpha)$ is infinitely differentiable with respect to $\alpha \in (-a, b)$ and that we can freely change the order of differentiation and expectation. Then we have

$$L'(\alpha) = \frac{\int_{\mathbb{R}^n} e^{\alpha\varphi(x)} f(x)\varphi(x) dx}{\int_{\mathbb{R}^n} e^{\alpha\varphi(x)} f(x) dx} = \mathbb{E}\varphi(X_\alpha).$$

Differentiate $L'(\alpha)$ one more time. We have

$$L''(\alpha) = \frac{\int_{\mathbb{R}^n} e^{\alpha\varphi(x)} f(x)\varphi^2(x) dx}{\int_{\mathbb{R}^n} e^{\alpha\varphi(x)} f(x) dx} - \left(\frac{\int_{\mathbb{R}^n} e^{\alpha\varphi(x)} f(x)\varphi(x) dx}{\int_{\mathbb{R}^n} e^{\alpha\varphi(x)} f(x) dx} \right)^2 = \mathrm{Var}(\varphi(X_\alpha)).$$

■

A function $f : \mathbb{R}^n \to \mathbb{R}_+$ is called *log-concave* if we have

$$f((1-\lambda)x + \lambda y) \geq f(x)^{1-\lambda} f(y)^\lambda \tag{4.4}$$

for all $x, y \in \mathbb{R}^n$ and all $\lambda \in [0, 1]$. The following lemma tells us that the upper bound of $\mathbb{E}e^{\alpha\varphi(X)}$ emerges as a consequence of the log-concavity of $L(\alpha)$ after certain normalization.

Lemma 4.1.2. *Let $c(\alpha)$ be a smooth function such that $e^{-c(\alpha)}\mathbb{E}e^{\alpha\varphi(X)}$ is log-concave for $-a < \alpha < b$. Then we have*

$$\mathbb{E}e^{\alpha(\varphi(X) - \mathbb{E}\varphi(X))} \leq e^{\psi_c(\alpha)}, \tag{4.5}$$

where $\psi_c(\alpha) = c(\alpha) - c(0) - c'(0)\alpha$.

Proof. Since $e^{-c(s)}\mathbb{E}e^{s\varphi(X)}$ is log-concave, we have $L''(s) \leq c''(s)$. For any $0 < t < \alpha < b$, integrating the inequality over $(0, t)$ we have

$$L'(t) - L'(0) \leq c'(t) - c'(0).$$

Integrating both sides over $(0, \alpha)$, we have

$$L(\alpha) - L(0) - L'(0)\alpha \leq c(\alpha) - c(0) - c'(0)\alpha. \tag{4.6}$$

Similarly we can show that the estimate also holds for $-a < \alpha < 0$. Notice that $L(0) = 0$ and $L'(0) = \mathbb{E}\varphi(X)$. Then the lemma follows from exponentiating both sides of (4.6). ■

Remark. From Lemma 4.1.1 and Lemma 4.1.2, we can see that the study of upper bound of $\mathrm{Var}(\varphi(X_\alpha))$ is equivalent to that of the normalizing function for $\mathbb{E}e^{\alpha\varphi(X)}$ to be log-concave. We can get one from the other by differentiating or integrating twice. That is why variance bounds can imply exponential deviation inequalities when moment generating functions exist.

Let $f : \mathbb{R} \to \mathbb{R} \cup \{\infty\}$ be a real-valued function. For $x \in \mathbb{R}$, its *Fenchel-Legendre transform* $f^*(x)$ is defined as

$$f^*(x) = \sup_y (xy - f(y)). \tag{4.7}$$

Let $\psi_{c,+}(\alpha)$ and $\psi_{c,-}(\alpha)$ be the restrictions of $\psi_c(\alpha)$ on the positive and negative half axis, respectively.

Corollary 4.1.1. *Under the assumptions and notations of Lemma 4.1.2, for any $t > 0$ we have*

$$\mathbb{P}(\varphi(X) - \mathbb{E}\varphi(X) > t) \leq e^{-\psi^*_{c,+}(t)}, \tag{4.8}$$

and

$$\mathbb{P}(\varphi(X) - \mathbb{E}\varphi(X) < -t) \leq e^{-\psi^*_{c,-}(-t)}, \tag{4.9}$$

where $\psi^*_{c,+}$, $\psi^*_{c,-}$ are Fenchel-Legendre transforms of $\psi_{c,+}$, $\psi_{c,-}$, respectively.

Proof. The proof follows from the so-called Cramér-Chernoff method: using Markov inequality in conjunction with optimization of the resulting bound. For the upper tail, we have for $0 < \alpha < b$ and $t > 0$ that

$$\begin{aligned}\mathbb{P}(\varphi(X) - \mathbb{E}(\varphi(X)) > t) &= \mathbb{P}(e^{\alpha(\varphi(X)-\mathbb{E}\varphi(X))} > e^{\alpha t}) \\ &\leq e^{-\alpha t} \cdot \mathbb{E}e^{\alpha(\varphi(X)-\mathbb{E}\varphi(X))} \\ &\leq e^{-(\alpha t - \psi_{c,+}(\alpha))}.\end{aligned}$$

We use Lemma 4.1.2 in the second inequality. Then the upper tail estimate follows by taking the infimum of the right hand side over $0 < \alpha < b$. The lower tail estimate follows from the same argument for $-a < \alpha < 0$. ■

4.2 Log-concavity of Moments of s-concave functions

In this section, we study the log-concavity of the (normalized) moments of s-concave functions, which, in conjugation with the results from the previous section, will enable us to obtain optimal concentration of the information content for convex measures.

Definition 4.2.1. For $s \in \mathbb{R}$, a function $f : \mathbb{R}^n \to \mathbb{R}_+$ is called s-concave if we have

$$f((1-\lambda)x + \lambda y) \geq ((1-\lambda)f(x)^s + \lambda f(y)^s)^{1/s} \tag{4.10}$$

for all x, y such that $f(x)f(y) > 0$ and for all $\lambda \in [0, 1]$.

For $s = 0$, the right hand side is defined by continuity, which corresponds to log-concave functions defined in (4.4). For $s > 0$, the previous definition is equivalent to that f^s is concave on its support; while for $s < 0$, it is equivalent to that f^s is convex on its support.

Recall that for $x > 0$, the Gamma function $\Gamma(x)$ is defined by

$$\Gamma(x) = \int_0^\infty t^{x-1}e^{-t}dt.$$

For $x, y > 0$, the Beta function $B(x, y)$ is defined by

$$B(x, y) = \int_0^1 t^{x-1}(1-t)^{y-1}dt.$$

The following result is proved by Borell [20] for $s > 0$, except that the function φ is assumed to be decreasing. It was then noticed by some people and available for example in Guédon, Nayar and Tkocz [63] that the result remains true without any monotonicity hypothesis. For $s < 0$, it is proved by Fradelizi, Guédon and Pajor [51], and the case $s = 0$ follows by taking the limits (or reproducing the mechanics of the proof).

Proposition 4.2.1. *Let $s \in \mathbb{R}$ and let $\varphi : [0, +\infty) \to [0, +\infty)$ be an s-concave integrable function.*
1) If $s > 0$, then $p \mapsto B(p, s^{-1} + 1)^{-1} \int_0^{+\infty} t^{p-1}\varphi(t)dt$ is log-concave for $p > 0$.
2) If $s = 0$, then $p \mapsto \Gamma(p)^{-1} \int_0^{+\infty} t^{p-1}\varphi(t)dt$ is log-concave for $p > 0$.
3) If $s < 0$, then $p \mapsto B(p, -s^{-1} - p)^{-1} \int_0^{+\infty} t^{p-1}\varphi(t)dt$ is log-concave for $0 < p < -1/s$.

Let us define the function $\varphi_s(t) = (1 - st)_+^{1/s} \mathbf{1}_{\mathbb{R}_+}$ for $s \neq 0$, and $\varphi_0(t) = e^{-t}\mathbf{1}_{\mathbb{R}_+}$. Then the preceding proposition may be expressed in the following way: if $\varphi : [0, +\infty) \to [0, +\infty)$ is s-concave, then the function

$$p \mapsto \frac{\int_0^{+\infty} t^{p-1}\varphi(t)dt}{\int_0^{+\infty} t^{p-1}\varphi_s(t)dt}$$

is log-concave for p such that $1/p > \max(0, -s)$. Using the preceding proposition, we can prove the following theorem which unifies and partially extends previous results of Borell [20], Bobkov and Madiman [15], and Fradelizi, Madiman and Wang [53]. A weaker log-concavity statement was also obtained by Nguyen[129].

Theorem 4.2.1. *Let $s \in \mathbb{R}$ and let $f : \mathbb{R}^n \to \mathbb{R}_+$ be an integrable s-concave function. Then the function*

$$p \mapsto (p + s) \cdots (p + ns) \int_{\mathbb{R}^n} f(x)^p dx \tag{4.11}$$

is log-concave for $p > \max(0, -ns)$.

Proof. The case $s = 1$ is due to Borell [20] and the case $s > 0$ deduces directly by applying Borell's result to f^s. The case $s = 0$ was proved by Fradelizi, Madiman and

Wang [53]. The case $s = -1$ is due to Bobkov and Madiman [15][1], except that the range was $p > n+1$. In the same way, the case $s < 0$ deduces from the case $s = -1$ by applying it to $f^{|s|}$. So we only need to prove the extension of the range for $s = -1$. Let us assume that $s = -1$. Thus f is -1-concave, which means that $g = f^{-1}$ is convex on its support. As done by Bobkov and Madiman [15], we write

$$\int_{\mathbb{R}^n} f(x)^p dx = \int_{\mathbb{R}^n} g(x)^{-p} dx = \int_0^{+\infty} p t^{p-1} \psi(1/t) dt,$$

where $\psi(t) = |\{x \in \mathbb{R}^n : g(x) \leq t\}|_n$ is the Lebesgue measure of the sub-level set $\{x \in \mathbb{R}^n : g(x) \leq t\}$. Using Brunn-Minkowski theorem, we can see that ψ is a $1/n$-concave function. Using the properties of the perspective function, we can deduce that the function $\varphi(t) = t^n \psi(1/t)$ is also a $1/n$-concave function. Thus it follows that

$$\int_{\mathbb{R}^n} f(x)^p dx = p \int_0^{+\infty} t^{p-n-1} \varphi(t) dt.$$

Applying Proposition 4.2.1 to $s = 1/n$ and p replaced by $p - n$ we get that

$$B(p-n, n+1)^{-1} \int_0^{+\infty} t^{p-1-n} \varphi(t) dt$$

is log-concave on $(n, +\infty)$. Then we can conclude the proof using the following identity

$$B(p-n, n+1)^{-1} = \frac{p(p-1)\cdots(p-n)}{\Gamma(n+1)}.$$

■

The fact that Theorem 4.2.1 is optimal can be seen from the following example. Let $U : \mathbb{R}^n \to [0, \infty]$ be a positively homogeneous convex function of degree 1, i.e. that $U(tx) = tU(x)$ for all $x \in \mathbb{R}^n$ and all $t > 0$. We define $f_{s,U} = (1 - sU)_+^{1/s}$ for $s \neq 0$ and $f_{0,U} = e^{-U}$ for $s = 0$. Then we have

$$\int_{\mathbb{R}^n} f_{s,U}(x)^p dx = \frac{C_U n!}{(p+s)\cdots(p+ns)},$$

[1] The details of this proof were omitted from [15] because of space considerations, and are being presented here. A complete presentation will appear in [13].

where C_U is the Lebesgue measure of the sub-level set $\{x \in \mathbb{R}^n : U(x) \leq 1\}$. We only check the identity for $s > 0$, and the other two cases can be proved similarly.

$$\begin{aligned}
\int_{\mathbb{R}^n} f_{s,U}(x)^p dx &= p \int_0^1 t^{p-1} |\{x \in \mathbb{R}^n : (1 - sU(x))_+^{1/s} > t\}| dt \\
&= p \int_0^1 t^{p-1} |\{x \in \mathbb{R}^n : U(x) < (1 - t^s)/s\}| dt \\
&= C_U p \int_0^1 t^{p-1} ((1 - t^s)/s)^n dt \\
&= C_U s^{-n-1} p B(p/s, n + 1)
\end{aligned}$$

In the third equation, we use the homogeneity of U and the property of Lebesgue measure. Then we can prove the identity using the following fact

$$B(p/s, n + 1) = \frac{n!}{(p/s + n) \cdots p/s}.$$

Thus the preceding theorem can be written in the following way: if $f : \mathbb{R}^n \to \mathbb{R}_+$ is an integrable s-concave function, then

$$p \mapsto \frac{\int_{\mathbb{R}^n} f(x)^p dx}{\int_{\mathbb{R}^n} f_{s,U}(x)^p dx}$$

is log-concave for $p > \max(0, -ns)$.

4.3 Concentration of information content

Now we are ready to study the concentration property of information content for convex measures introduced and studied by Borell [21, 22].

Definition 4.3.1. Let $-\infty \leq \kappa \leq \infty$. A finite Borel measure μ on $\mathbb{R}^n$ is called κ-concave if we have

$$\mu((1 - \lambda)A + \lambda B) \geq ((1 - \lambda)\mu(A)^\kappa + \lambda \mu(B)^\kappa)^{1/\kappa} \tag{4.12}$$

for all $\lambda \in [0, 1]$ and all Borel sets $A, B \subseteq \mathbb{R}^n$ such that $\mu(A)\mu(B) > 0$.

Here $(1 - \lambda)A + \lambda B = \{(1 - \lambda)x + \lambda y : x \in A, y \in B\}$ stands for the Minkowski sum of two sets. The limit cases are interpreted by continuity. Thus the right hand

side of (4.12) is equal to $\min(\mu(A), \mu(B))$ for $\kappa = -\infty$; $\mu(A)^{1-\lambda}\mu(B)^{\lambda}$ for $\kappa = 0$; and $\max(\mu(A), \mu(B))$ for $\kappa = \infty$. Note that the inequality (4.12) becomes stronger as κ increases. For $\kappa = -\infty$, we obtain the largest class, whose members are called *convex or hyperbolic* measures. The case $\kappa = 0$ describes log-concave measures. If μ is a convex measure on $\mathbb{R}^n$ then it is absolutely continuous with respect to the Lebesgue measure on the subspace generated by its support and its density has a concavity property. For example, if μ is κ-concave and has a density f on $\mathbb{R}^n$ then $\kappa \leq 1/n$ and f is $-1/\beta$-concave with $\beta = n - 1/\kappa$.

We say that a $\mathbb{R}^n$-valued random variable X is κ-concave if the probability measure induced by X is κ-concave. In this section, we let X be a κ-concave random variable with density f and $\kappa < 0$. Then Borell's characterization implies that there is a convex function V such that $f = V^{-\beta}$. In the following, we will study the deviation of $\widetilde{h}(X)$ from its mean $h(X)$, that is corresponding to taking $\varphi = -\log f$ in Section 4.1. Then the moment generating function is

$$\mathbb{E}f^{-\alpha}(X) = \int_{\mathbb{R}^n} f(x)^{1-\alpha} dx.$$

The integral is finite as long as $(1-\alpha)\beta > n$, i.e. that $\alpha < 1 - n/\beta$.

Proposition 4.3.1. *Let $\beta > n$ and let X be a random variable in $\mathbb{R}^n$ with density f being $-1/\beta$-concave. Then the function*

$$\alpha \mapsto \prod_{i=1}^{n}((1-\alpha)\beta - i)\mathbb{E}f^{-\alpha}(X) \tag{4.13}$$

is log-concave for $\alpha < 1 - n/\beta$.

Proof. It easily follows from Theorem 4.2.1 with p replaced by $1-\alpha$ and s replaced by $-1/\beta$. ■

Following Lemma 4.1.2, we can set

$$c(\alpha) = -\sum_{i=1}^{n} \log((1-\alpha)\beta - i). \tag{4.14}$$

Corollary 4.3.1. *Under the conditions and notations of Proposition 4.3.1, we have*

$$Var(\widetilde{h}(X)) \leq \beta^2 \sum_{i=1}^{n} (\beta - i)^{-2}. \tag{4.15}$$

Proof. By Lemma 4.1.1, we know that $\mathrm{Var}(\widetilde{h}(X_\alpha)) = L''(\alpha)$, where X_α is a random variable with density proportional to $f^{1-\alpha}$ and $L(\alpha) = \log \mathbb{E} f^{-\alpha}(X)$ is the logarithmic moment generating function. By Proposition 4.3.1, we know that $L''(\alpha) \leq c''(\alpha)$, where $c(\alpha)$ is defined in (4.14). Then the variance bound (4.15) follows by differentiating $c(\alpha)$ twice and setting $\alpha = 0$. ∎

Remark. The variance bound is sharp. Suppose X has density $f = (1 + U/\beta)_+^{-\beta}$ with U being a positively homogeneous convex function of degree 1. In this case, the function in Proposition 4.3.1 is log-affine, i.e. $L''(\alpha) = c''(\alpha)$. Then we have equality in the above variance bound. In particular, it includes the Pareto distribution with density

$$f(x) = \frac{1}{Z_n(a, \beta)} (a + x_1 + \cdots + x_n)^{-\beta}, \ x_i > 0, \tag{4.16}$$

where $a > 0$ and $Z_n(a, \beta)$ is a normalizing constant.

Let $\beta > n + 2$ and let X be a random variable in $\mathbb{R}^n$ with density f being $-1/\beta$-concave. In this case, we have $\mathbb{E}|X|^2 < \infty$ and the covariance matrix Σ is defined by

$$\Sigma = \mathbb{E}(X - \mathbb{E}X) \otimes (X - \mathbb{E}X). \tag{4.17}$$

Then we have

$$\begin{aligned} n &= \int_{\mathbb{R}^n} \langle x - \mathbb{E}X, -\nabla \log f(x) \rangle f(x) dx \\ &\leq \left(\int_{\mathbb{R}^n} |x - \mathbb{E}X|^2 f(x) dx \cdot \int_{\mathbb{R}^n} |\nabla \log f(x)|^2 f(x) dx \right)^{1/2} \\ &= \sqrt{\mathrm{tr}(\Sigma) J(X)}, \end{aligned}$$

where $\mathrm{tr}(\Sigma)$ is the trace of Σ and $J(X)$ is the Fisher information defined by

$$J(X) = \int_{\mathbb{R}^n} \frac{|\nabla f|^2}{f} dx. \tag{4.18}$$

Combining with Corollary 4.3.1 we have the following result.

Corollary 4.3.2. *Let $\beta > n+2$ and let X be a random variable in $\mathbb{R}^n$ with density f being $-1/\beta$-concave. Then we have*

$$Var(\widetilde{h}(X)) \leq \frac{tr(\Sigma)\beta^2}{n^2} \sum_{i=1}^{n} (\beta - i)^{-2} J(X). \tag{4.19}$$

In particular, if X is isotropic, i.e. that $\mathbb{E}X = 0$ and Σ is the identity matrix, we have

$$\mathrm{Var}(\widetilde{h}(X)) \leq \frac{\beta^2}{n} \sum_{i=1}^{n} (\beta - i)^{-2} J(X). \tag{4.20}$$

Taking $\beta \to \infty$ yields the analogue for log-concave random variables, namely

$$\mathrm{Var}(\widetilde{h}(X)) \leq J(X), \tag{4.21}$$

which was observed by Nguyen [129].

Theorem 4.3.1. *Let $\beta > n$ and let X be a random variable in $\mathbb{R}^n$ with density f being $-1/\beta$-concave. Then we have*

$$\mathbb{E}e^{\alpha(\widetilde{h}(X) - h(X))} \leq e^{\psi_c(\alpha)} \tag{4.22}$$

for $\alpha < 1 - n/\beta$, where

$$\psi_c(\alpha) = -\alpha\beta \sum_{i=1}^{n} (\beta - i)^{-1} - \sum_{i=1}^{n} \log \frac{(1-\alpha)\beta - i}{\beta - i}. \tag{4.23}$$

Particularly, one has equality for Pareto distributions.

Proof. The moment generating function bound (4.22) easily follows from Lemma 4.1.2 and Proposition 4.3.1. Some easy calculations will show the equality case for Pareto distributions. Essentially that is due to the identity $L''(\alpha) = c''(\alpha)$, where $c(\alpha)$ is defined in (4.14).

■

Corollary 4.3.3. *Under the conditions and notations of Theorem 4.3.1, for any $t > 0$ we have*

$$\mathbb{P}(\widetilde{h}(X) - h(X) > t) \leq e^{-\psi^*_{c,+}(t)}, \tag{4.24}$$

and

$$\mathbb{P}(\widetilde{h}(X) - h(X) < -t) \leq e^{-\psi_{c,-}^*(-t)}, \tag{4.25}$$

where $\psi_{c,+}^*$ and $\psi_{c,-}^*$ are Fenchel-Legendre transforms of $\psi_{c,+}$ and $\psi_{c,-}$, respectively.

In general we do not have explicit expressions for $\psi_{c,+}^*$ or $\psi_{c,-}^*$. The following result was obtained by Bobkov and Madiman [15] with the assumption $\beta \geq n+1$, which can be relaxed to $\beta > n$. It basically says that the entropy of an κ-concave distribution can not exceed that of the Pareto distribution with the same maximal density value.

Corollary 4.3.4. *Under the conditions and notations of Theorem 4.3.1, we have*

$$h(X) \leq -\log \|f\|_\infty + \beta \sum_{i=1}^{n} (\beta - i)^{-1}, \tag{4.26}$$

where we denote by $\|f\|_\infty$ the essential supremum. We have equality for Pareto distributions.

Proof. As a function of α, we have

$$(-\alpha t - \psi_c(\alpha))' = -t + \beta \sum_{i=1}^{n} (\beta - i)^{-1} - \beta \sum_{i=1}^{n} ((1-\alpha)\beta - i)^{-1}.$$

For any $t > \beta \sum_{i=1}^{n} (\beta - i)^{-1}$, we can see that $-\alpha t - \psi_c(\alpha)$ is a decreasing function of $\alpha < 1 - n/\beta$. It is clear that $\lim_{\alpha \to -\infty} (-\alpha t - \psi_c(\alpha)) = \infty$. Therefore we have $\psi_{c,-}^*(-t) = \infty$ for $t > \beta \sum_{i=1}^{n} (\beta - i)^{-1}$. Using the lower tail estimate in Corollary 4.3.3, almost surely we have

$$\widetilde{h}(X) - h(X) \geq -\beta \sum_{i=1}^{n} (\beta - i)^{-1}.$$

Taking the supremum over all realizable values of X yields

$$-\log \|f\|_\infty - h(X) \geq -\beta \sum_{i=1}^{n} (\beta - i)^{-1}.$$

That is equivalent to the desired statement. ∎

Remark. We can get corresponding estimates for log-concave random variables (see [15, 53]) by taking the limit $\beta \to \infty$.

The following result is an improvement of Proposition 5.1 of Bobkov and Madiman [18]. Its analogue for log-concave probability measures was first observed by Klartag and Milman [87], with refinement made by [53, Corollary 4.7].

Corollary 4.3.5. *Under the conditions and notations of Theorem 4.3.1, we have*

$$\mathbb{P}(f(X) \geq c_0^n \|f\|_\infty) \geq 1 - c_1^n \tag{4.27}$$

for $0 < c_0 < 1$ *such that* $n \log c_0 < -\beta \sum_{i=1}^n (\beta - i)^{-1}$ *and some* $0 < c_1 < 1$ *depending on* c_0 *and* β.

Proof. Note that

$$\begin{aligned}
\mathbb{P}(f(X) \leq c_0^n \|f\|_\infty) &= \mathbb{P}(\log f(X) \leq \log \|f\|_\infty + n \log c_0) \\
&= \mathbb{P}(\widetilde{h}(X) \geq -\log \|f\|_\infty - n \log c_0) \\
&\leq \mathbb{P}\left(\widetilde{h}(X) \geq h(X) - \beta \sum_{i=1}^n (\beta - i)^{-1} - n \log c_0\right)
\end{aligned}$$

We use Corollary 4.3.4 in the above inequality. Applying the upper tail estimate of Corollary 4.3.3 with

$$t = -n \log c_0 - \beta \sum_{i=1}^n (\beta - i)^{-1} \tag{4.28}$$

yields

$$\mathbb{P}(f(X) \leq c_0^n \|f\|_\infty) \leq e^{-\psi_{c,+}^*(t)}. \tag{4.29}$$

As a function of α, we have

$$(\alpha t - \psi_c(\alpha))' = -n \log c_0 - \sum_{i=1}^n \frac{\beta}{(1-\alpha)\beta - i},$$

from which we can see $(\alpha t - \psi_c(\alpha))'(0) = t > 0$ and $(\alpha t - \psi_c(\alpha))'(1 - n/\beta) = -\infty$. In addition we can see the concavity of $\alpha t - \psi_c(\alpha)$ from

$$(\alpha t - \psi_c(\alpha))'' = -\sum_{i=1}^{n} \frac{\beta^2}{((1-\alpha)\beta - i)^2} < 0.$$

Therefore we have

$$\psi^*_{c,+}(t) = \alpha^* t - \psi_c(\alpha^*), \tag{4.30}$$

where α^* is a positive number such that $(\alpha t - \psi_c(\alpha))'(\alpha^*) = 0$, i.e. that

$$\sum_{i=1}^{n} \frac{\beta}{(1-\alpha^*)\beta - i} = -n \log c_0. \tag{4.31}$$

Using the definitions of $\psi_c(\alpha)$ and t in (4.23) and (4.28), respectively, we have

$$\psi^*_{c,+}(t) = -n\alpha^* \log c_0 + \sum_{i=1}^{n} \log \frac{(1-\alpha^*)\beta - i}{\beta - i}. \tag{4.32}$$

Combining with (4.29), we have

$$\mathbb{P}(f(X) \le c_0^n \|f\|_\infty) \le c_1^n, \tag{4.33}$$

where

$$c_1 = c_0^{\alpha^*} \left(\prod_{i=1}^{n} \frac{\beta - i}{(1-\alpha^*)\beta - i} \right)^{1/n}. \tag{4.34}$$

That is equivalent to the desired statement. To see that $c_1 < 1$, we take the logarithm of c_1,

$$\begin{aligned}
\log c_1 &= \alpha^* \log c_0 + \frac{1}{n} \sum_{i=1}^{n} \log \frac{\beta - i}{(1-\alpha^*)\beta - i} \\
&= -\frac{1}{n} \sum_{i=1}^{n} \frac{\alpha^* \beta}{(1-\alpha^*)\beta - i} + \frac{1}{n} \sum_{i=1}^{n} \log \frac{\beta - i}{(1-\alpha^*)\beta - i} \\
&= -\frac{1}{n} \sum_{i=1}^{n} \left(\frac{\alpha^* \beta}{(1-\alpha^*)\beta - i} - \log \left(1 + \frac{\alpha^* \beta}{(1-\alpha^*)\beta - i} \right) \right) \\
&< 0.
\end{aligned}$$

We use the equation (4.31) in the second identity. The last inequality follows from the fact that $\log(1+x) < x$ for $x > 0$. ■

Chapter 5

FUTURE WORK

In the last part, we discuss several specific projects or directions that I would like to explore in the future.

Discrete EPI. The role of the entropy power inequality (EPI), which states that

$$e^{2h(X+Y)/n} \geq e^{2h(X)/n} + e^{2h(Y)/n}$$

for independent $\mathbb{R}^n$-valued random variables X and Y, in information and communication theory, as well as in physics and in probability theory, is now well known. The problem of searching for discrete analogues of the EPI has been long studied, with various interesting partial results [145, 64, 71, 164, 166]. When we talk about discrete analogues of the EPI, what we really mean is looking for lower bounds on the entropy of a sum of independent random variables that take values in a finite or countable group G. We define the minimal entropy function

$$f_G(s,t) = \inf H(X+Y),$$

where the infimum is taken over all independent G-valued random variables X, Y such that $H(X) = s$ and $H(Y) = t$. For simplicity we use $f_n(s,t)$ instead for the cyclic group $G = \mathbb{Z}_n$. For the special case of $\mathbb{Z}_2$, Wyner and Ziv [168] made the interesting observation that they termed "Mrs. Gerber's lemma": f_2 is convex in each of its arguments if the other is held fixed. Shamai and Wyner [145] used this to obtain a binary analogue of the EPI. Recently, Jog and Anantharam [71] made the following remarkable observation: If G is a group of order 2^n for some natural number n, then $f_G(s,t)$ is convex in each argument. More importantly, they found that $f_G(s,t)$ has

certain interesting structure. Mrs. Gerber's Lemma generally fails, even for $\mathbb{Z}_3$. But we believe that $f_G(s,t)$ has certain structure property for cyclic groups with orders of prime powers.

Conjecture 5.0.1. For any prime number p, we have

$$f_{p^n}(s,t) = \begin{cases} k\log p + f_p(s-k\log p, t-k\log p), \\ \qquad \text{if } k\log p \le s,t < (k+1)\log p, \\ \max\{s,t\}, \qquad \text{otherwise.} \end{cases}$$

Numerical experiments show that the formula appears to hold for $p = 3, 5$. There exist random variables X, Y such that $H(X+Y)$ can reach the lower bound. To see this, let G' be any subgroup of $\mathbb{Z}_{p^n}$ of order p^k. For the first case, we can let X, Y be random variables such that both are uniform when conditioned on each coset of G'. For the second case, we always have $f_{p^n}(s,t) \ge \max\{s,t\}$. Suppose that $s < k\log p \le t < (k+1)\log p$ for some $1 \le k \le n-1$. To see that the equality can happen, we can let X be a random variable supported on G' and Y be a random variable such that it is uniform when conditioned on each coset of G'. For such X, Y, it is not hard to check that $H(X+Y) = H(Y)$, which implies the second case.

A more general question is to study the property of $f_G(s,t)$ for any abelian group G. That might be doable by the decomposition of every finitely generated abelian group into the direct sum of primary cyclic groups and infinite cyclic groups. It is actually analogous to the Cauchy-Davenport problem in additive combinatorics: looking for the lower bound of $|A+B|$ over all subsets $A, B \subset G$ with fixed cardinalities. Eliahou, Kervaire and Plagne [44] proved that the lower bound only depends on the order of G rather than the group structure. It is reasonable to expect similar properties in the entropy setting.

Sum-product phenomena. Let $A \subset \mathbb{Z}$ be a finite subset. Recall the definition of sumset $A+A = \{a+b : a,b \in A\}$. Similarly the product set $A \cdot A$ is defined to be $A \cdot A = \{a \cdot b : a,b \in A\}$. If A is an arithmetic progression, it is not hard to see that

$|A+A| \approx |A|$ and $|A \cdot A| \approx |A|^2$ up to some constant factors. If A is a geometric progression, we have $|A \cdot A| \approx |A|$ and $|A+A| \approx |A|^2$ up to some constant factors. The sum-product phenomenon roughly says that a finite subset of $\mathbb{Z}$ can not behave as an arithmetic progression and a geometric progression simultaneously. The most general setting we can talk about the sum-product problem is for subsets of a ring. In this general setting, the sum-product phenomenon says that if a finite set A is not close to a subring, then either the sumset $A+A$ or the product set $A \cdot A$ must be considerably larger than A. In another word, it is difficult to make A closed under addition and multiplication simultaneously unless that A is close to a subring.

This problem was initiated by Erdős and Szemerédi [47] for integers. They proved that $\max\{|A+A|, |A \cdot A|\} \geq c|A|^{1+\delta}$ for a small but positive number δ. It is conjectured that δ can be arbitrarily close to 1 as long as $|A|$ is large enough. Elekes [43] proved that $\delta \geq 1/4$ by using the *Szemerédi-Trotter* theorem in an ingenious way. More importantly, it opens the gate of using tools from incidence geometry to study the sum-product problem. For real numbers, the state of the art is due to Solymosi [149]: one can take δ arbitrarily close to $1/3$. Motivated by finite field *Kakeya conjecture*, Wolff [165] formulated the finite field version of sum-product problem, and the breakthrough work is done by Bourgain, Katz and Tao [24]. Improvements are made in [66, 59, 163]. Similar problems, such as difference-product, sum-ratio and difference-ratio, as well as other generalizations are also studied for rational functions and elliptic curves. The sum-product phenomenon has deep connections with and applications to many other areas, such as incidence geometry, number theory, combinatorics, spectral graph theory, complexity theory, pseudo randomness, probabilistic checkable proofs, and cryptography.

The sum-product problem is one of my favorite problems. In the current project [102], we are trying to use the entropy method to tackle this problem. Its entropy analog asserts that for i.i.d real-valued random variables X, Y, we have

$$\max\{H(X+Y), H(X \cdot Y)\} \geq (1+\delta)H(X),$$

and the constant $\delta > 0$ can be arbitrarily close to 1, as long as $H(X)$ is large enough. Its entropy version is actually stronger than the original conjecture, which can be seen by taking X, Y to be uniform on a finite subset A. The key idea of [149] is to upper bound the multiplicative energy by the sumset. But the bound is not tight for geometric progressions. Our study involves the estimate of a quantity, which is a kind of "mixed" energy. Partial results are obtained and some equivalent formulations are also considered.

Entropy inverse sumset theory. Inverse sumset estimate is another fundamental part of additive combinatorics. It seeks to conclude the structural statement about additive sets provided that the sumsets are small or large. The famous Freiman-Green-Ruzsa theorem says that an additive set A with small doubling constant $\sigma[A]$ (i.e. $\sigma[A] \ll \log |A|$) is contained in a generalized arithmetic progression. The entropy analog developed by Tao [158] asserts that a discrete random variable X with small doubling constant $\sigma[X]$ (i.e. $\sigma[X] \leq K$ for some constant K) is roughly uniform on a generalized arithmetic progression. The continuous extension is made by Kontoyannis and Madiman [89], i.e. a continuous random variable with small doubling constant is close to Gaussian.

Our understanding of additive sets with large doubling constant (i.e. $|A|^\epsilon \ll \sigma[A] \leq |A|$) is quite poor. An additive set with distinct pairwise sums is called a *Sidon set*. The structure of Sidon sets is unclear so far. In the entropy setting, we are interested in the classification of random variables with large doubling constants. Suppose that X, Y are i.i.d. random variables such that $H(X+Y) \geq 2H(X) - K$ for some constant K. Our goal is to conclude the structure of the distribution of X and Y. If X is supported on a Sidon set, the ambiguity in the pair (X, Y) given the sum $X+Y$ is at most 1 bit. In another word, we have $H(X+Y) \geq 2H(X)-1$ bit. However it can not guarantee that X has large doubling constant even if we know the range(X) has large doubling constant. It is possible that the subset of range(X) which plays the key role in sumset estimate may contribute little in the entropy sense. We can indeed

construct such random variables. But it is reasonable to believe that range(X) must contain a subset with large doubling constant and X restricted on this subset makes the main contribution to the entropy of X. Then the classification of random variables with large doubling constants should be closely related to the study of additive sets with large doubling constants.

Entropy method in combinatorics. Entropy-based argument has been proved very useful in many combinatorial enumeration problems. Erdős and Rényi [46] gave the first combinatorial application of entropy to deriving a lower bound of the size of the smallest distinguishing family of a set. The so-called *Shearer's lemma* was introduced by Chung, Frankl, Graham and Shearer [28] to bound the size of intersecting families. Another application of entropy method is Radhakrishnan's entropy proof of Brégman's theorem [135]. Various other applications of entropy method can be found in the study of the number of embeddings of one graph in another Friedgut and Kahn [55], the number of independent sets in a regular bipartite graph Kahn [72], the number of graph homomorphisms Galvin and Tetali [58], the number of Hamilton cycles in a tournament Friedgut and Kahn [56], the number of matchings and independent sets of fixed size Carroll, Galvin and Tetali [27], counting graph homomorphisms and zero-error codes Madiman and Tetali [111]. For general background, we refer to the nice survey by Radhakrishnan [136]. In addition, entropy sumset inequalities of non-independent pair of random variables have great utility in many classical problems in combinatorics, such as the *Kakeya problem* and *Erdős distance problem*.

Slicing problem in asymptotic convex geometry. One of the central questions in convex geometry is called *Hyperplane Conjecture* or *Slicing Problem*. It asserts that for every convex body $K \subset \mathbb{R}^n$ of volume 1, there exists a $(n-1)$-dimensional hyperplane H such that $\mathrm{Vol}_{n-1}(K \cap H) \geq c$ for some dimension-free constant $c > 0$. This question was raised by Bourgain [23] and the best known lower bound is $cn^{-1/4}$ by Klartag [86]. It is implied by the *Thin Shell Conjecture* [5, 17], which again trivially implied by the *Kannan-Lovsz-Simonovits Conjecture* [73]. There are many equivalent formulations of

this problem [121, 85, 15, 54]. The information theoretical formulation in [15] asserts that the slicing problem is equivalent to the estimate of how (dimension-free) closeness of log-concave measure to a Gaussian measure. It is also equivalent to finding a lower bound of entropies of log-concave random variables with fixed covariance matrix. Even for real-valued random variables with fixed variance, it is still unknown which log-concave density has the minimum entropy. To figure out this extreaml density among one-dimensional log-concave densities is a topic we are studying in the current project [103].

BIBLIOGRAPHY

[1] E. Abbe. Polar martingale of maximal spread. *International Zurich Seminar*, 2012.

[2] E. Abbe. Randomness and dependencies extraction via polarization. In *Proc. Information Theory and Applications Workshop (ITA)*, pages 1–7, Feb. 2011.

[3] E. Abbe, J. Li, and M. Madiman. Entropies of weighted sums in cyclic groups and applications to polar codes. *submitted*.

[4] N. Alon and R. Yuster. The 123 theorem and its extensions. *J. Combin. Theory Ser. A*, 72(2):322–331, 1995.

[5] M. Anttila, K. Ball, and I. Perissinaki. The central limit problem for convex bodies. *Trans. Amer. Math. Soc.*, 355(12):4723–4735 (electronic), 2003.

[6] E. Arıkan. Channel polarization: a method for constructing capacity-achieving codes for symmetric binary-input memoryless channels. *IEEE Trans. Inform. Theory*, 55(7):3051–3073, 2009.

[7] E. Arıkan. Source polarization. In *Proc. IEEE Intl. Symp. Inform. Theory*, pages 899–903, Austin, June 2010.

[8] E. Arıkan and E. Telatar. On the rate of channel polarization. In *Proc. IEEE Intl. Symp. Inform. Theory*, pages 1493–1495, Seoul, Korea, 2009.

[9] A. R. Barron. The strong ergodic theorem for densities: generalized Shannon-McMillan-Breiman theorem. *Ann. Probab.*, 13(4):1292–1303, 1985.

[10] P. Bateman and P. Erdös. Geometrical extrema suggested by a lemma of Besicovitch. *Amer. Math. Monthly*, 58:306–314, 1951.

[11] C. Berg. Stieltjes-pick-bernstein-schoenberg and their connection to complete monotonicity. *Positive Definite Functions: From Schoenberg to Space-Time Challenges*, pages 15–45, 2008.

[12] C. Berg, J. P. R. Christensen, and P. Ressel. *Harmonic analysis on semigroups*, volume 100 of *Graduate Texts in Mathematics*. Springer-Verlag, New York, 1984.

[13] S. Bobkov, M. Fradelizi, J. Li, and M. Madiman. When can one invert hölder's inequality? (and why one may want to). *Preprint*, 2016.

[14] S. Bobkov and M. Madiman. Concentration of the information in data with log-concave distributions. *Ann. Probab.*, 39(4):1528–1543, 2011.

[15] S. Bobkov and M. Madiman. The entropy per coordinate of a random vector is highly constrained under convexity conditions. *IEEE Trans. Inform. Theory*, 57(8):4940–4954, 2011.

[16] S Bobkov and M. Madiman. An equipartition property for high-dimensional log-concave distributions. In *Proc. 50th Annual Allerton Conf. on Conmmunication, Control, and Computing*, pages 482–488, Monticello, Illinois: IEEE, October 2012.

[17] S. G. Bobkov and A. Koldobsky. On the central limit property of convex bodies. In *Geometric aspects of functional analysis*, volume 1807 of *Lecture Notes in Math.*, pages 44–52. Springer, Berlin, 2003.

[18] Sergey Bobkov and Mokshay Madiman. Reverse Brunn-Minkowski and reverse entropy power inequalities for convex measures. *J. Funct. Anal.*, 262(7):3309–3339, 2012.

[19] V. I. Bogachev. *Gaussian measures*, volume 62 of *Mathematical Surveys and Monographs*. American Mathematical Society, Providence, RI, 1998.

[20] C. Borell. Complements of Lyapunov's inequality. *Math. Ann.*, 205:323–331, 1973.

[21] C. Borell. Convex measures on locally convex spaces. *Ark. Mat.*, 12:239–252, 1974.

[22] C. Borell. Convex set functions in d-space. *Period. Math. Hungar.*, 6(2):111–136, 1975.

[23] J. Bourgain. On high-dimensional maximal functions associated to convex bodies. *Amer. J. Math.*, 108(6):1467–1476, 1986.

[24] J. Bourgain, N. Katz, and T. Tao. A sum-product estimate in finite fields, and applications. *Geom. Funct. Anal.*, 14(1):27–57, 2004.

[25] L. Breiman. The individual ergodic theorem for information theory. *Ann. Math. Sat.*, 28:809–810, 1960.

[26] A. Buja, B. F. Logan, J. A. Reeds, and L. A. Shepp. Inequalities and positive-definite functions arising from a problem in multidimensional scaling. *Ann. Statist.*, 22(1):406–438, 1994.

[27] T. Carroll, D. Galvin, and P. Tetali. Matchings and independent sets of a fixed size in regular graphs. *J. Combin. Theory Ser. A*, 116(7):1219–1227, 2009.

[28] F. R. K. Chung, R. L. Graham, P. Frankl, and J. B. Shearer. Some intersection theorems for ordered sets and graphs. *J. Combin. Theory Ser. A*, 43(1):23–37, 1986.

[29] A. S. Cohen and R. Zamir. Entropy amplification property and the loss for writing on dirty paper. *IEEE Trans. Inform. Theory*, 54(4):1477–1487, 2008.

[30] H. Cohn, A. Kumar, S. Miller, D. Radchenko, and M. S. Viazovska. The sphere packing problem in dimension 24.

[31] J. H. Conway and N. J. A. Sloane. *Sphere packings, lattices and groups*, volume 290. Springer-Verlag, New York, third edition, 1999.

[32] T. M. Cover and S. Pombra. Gaussian feedback capacity. *IEEE Trans. Inform. Theory*, 35(1):37–43, 1989.

[33] Thomas M. Cover and Joy A. Thomas. *Elements of information theory*. Wiley-Interscience [John Wiley & Sons], Hoboken, NJ, second edition, 2006.

[34] A. Dembo and O. Zeitouni. *Large deviations techniques and applications*, volume 38 of *Stochastic Modelling and Applied Probability*. Springer-Verlag, Berlin, 2010.

[35] J.-M. Deshouillers, G. A. Freiman, and A. A. Yudin. On bounds for the concentration function. II. *J. Theoret. Probab.*, 14(3):813–820, 2001.

[36] J-M. Deshouillers and Sutanto. On the rate of decay of the concentration function of the sum of independent random variables. *Ramanujan J.*, 9(1-2):241–250, 2005.

[37] W. Doeblin. Sur les sommes d'un grand nombre des variables alétoires indépendantes. *Bull. Sci. Math.*, 63:23–32, 35–64, 1939.

[38] Z. Dong, J. Li, and W. V. Li. A note on distribution-free symmetrization inequalities. *J. Theoret. Probab.*, 28(3):958–967, 2015.

[39] M. D. Donsker and S. R. S. Varadhan. Asymptotic evaluation of certain Markov process expectations for large time. I. II. *Comm. Pure Appl. Math.*, 28:1–47; ibid. 28 (1975), 279–301, 1975.

[40] M. D. Donsker and S. R. S. Varadhan. Asymptotic evaluation of certain Markov process expectations for large time. III. *Comm. Pure Appl. Math.*, 29(4):389–461, 1976.

[41] M. D. Donsker and S. R. S. Varadhan. Asymptotic evaluation of certain Markov process expectations for large time. IV. *Comm. Pure Appl. Math.*, 36(2):183–212, 1983.

[42] A. Dvoretzky. Some results on convex bodies and Banach spaces. In *Proc. Internat. Sympos. Linear Spaces (Jerusalem, 1960)*, pages 123–160. Jerusalem Academic Press, Jerusalem; Pergamon, Oxford, 1961.

[43] G. Elekes. On the number of sums and products. *Acta Arith.*, 81(4):365–367, 1997.

[44] S. Eliahou, M. Kervaire, and A. Plagne. Optimally small sumsets in finite abelian groups. *J. Number Theory*, 101(2):338–348, 2003.

[45] Yu. S. Eliseeva and A. Yu. Zaitsev. Estimates of the concentration functions of weighted sums of independent random variables. *Theory Probab. Appl.*, 57(4):670–678, 2013.

[46] P. Erdős and A. Rényi. On two problems of information theory. *Magyar Tud. Akad. Mat. Kutató Int. Közl.*, 8:229–243, 1963.

[47] P. Erdős and E. Szemerédi. On sums and products of integers. In *Studies in pure mathematics*, pages 213–218. Birkhäuser, Basel, 1983.

[48] C. G. Esseen. On the Kolmogorov-Rogozin inequality for the concentration function. *Z. Wahrscheinlichkeitstheorie und Verw. Gebiete*, 5:210–216, 1966.

[49] C. G. Esseen. On the concentration function of a sum of independent random variables. *Z. Wahrscheinlichkeitstheorie und Verw. Gebiete*, 9:290–308, 1968.

[50] R. H. Etkin and E. Ordentlich. The degrees-of-freedom of the K-user Gaussian interference channel is discontinuous at rational channel coefficients. *IEEE Trans. Inform. Theory*, 55(11):4932–4946, 2009.

[51] M. Fradelizi, O. Guédon, and A. Pajor. Thin-shell concentration for convex measures. *Studia Math.*, 223(2):123–148, 2014.

[52] M. Fradelizi, J. Li, and M. Madiman. Concentration of information content and other functionals under convex measures. *Preprint*.

[53] M. Fradelizi, M. Madiman, and L. Wang. Optimal concentration of information content for log-concave densities. *High Dimensional Probability*, VII, 2015.

[54] D. Fresen. The floating body and the hyperplane conjecture. *Arch. Math. (Basel)*, 98(4):389–397, 2012.

[55] E. Friedgut and J. Kahn. On the number of copies of one hypergraph in another. *Israel J. Math.*, 105:251–256, 1998.

[56] E. Friedgut and J. Kahn. On the number of Hamiltonian cycles in a tournament. *Combin. Probab. Comput.*, 14(5-6):769–781, 2005.

[57] O. Friedland and S. Sodin. Bounds on the concentration function in terms of the Diophantine approximation. *C. R. Math. Acad. Sci. Paris*, 345(9):513–518, 2007.

[58] D. Galvin and P. Tetali. On weighted graph homomorphisms. In *Graphs, morphisms and statistical physics*, volume 63 of *DIMACS Ser. Discrete Math. Theoret. Comput. Sci.*, pages 97–104. Amer. Math. Soc., Providence, RI, 2004.

[59] M. Z. Garaev. An explicit sum-product estimate in $\mathbb{F}_p$. *Int. Math. Res. Not. IMRN*, (11):Art. ID rnm035, 11, 2007.

[60] M. Gromov and V. D. Milman. A topological application of the isoperimetric inequality. *Amer. J. Math.*, 105(4):843–854, 1983.

[61] M. Gromov and V. D. Milman. Generalization of the spherical isoperimetric inequality to uniformly convex Banach spaces. *Compositio Math.*, 62(3):263–282, 1987.

[62] O. Guédon. Kahane-Khinchine type inequalities for negative exponent. *Mathematika*, 46(1):165–173, 1999.

[63] O. Guédon, P. Nayar, and T. Tkocz. Concentration inequalities and geometry of convex bodies. In *Analytical and probabilistic methods in the geometry of convex bodies*, volume 2 of *IMPAN Lect. Notes*, pages 9–86. Polish Acad. Sci. Inst. Math., Warsaw, 2014.

[64] S. Haghighatshoar, E. Abbe, and I. Emre Telatar. A new entropy power inequality for integer-valued random variables. *IEEE Trans. Inform. Theory*, 60(7):3787–3796, 2014.

[65] T. C. Hales. A proof of the Kepler conjecture. *Ann. of Math. (2)*, 162(3):1065–1185, 2005.

[66] D. Hart, A. Iosevich, and J. Solymosi. Sum-product estimates in finite fields via Kloosterman sums. *Int. Math. Res. Not. IMRN*, (5):Art. ID rnm007, 14, 2007.

[67] H. Hassani. *Polarization and spatial coupling: two techniques to boost performance*. PhD thesis, Informatique et Communications, EPFL, n. 5706, 2013.

[68] P. Hegarty. Some explicit constructions of sets with more sums than differences. *Acta Arith.*, 130(1):61–77, 2007.

[69] P. Hegarty and S. J. Miller. When almost all sets are difference dominated. *Random Structures Algorithms*, 35(1):118–136, 2009.

[70] W. Hengartner and R. Theodorescu. *Concentration functions*. Number 20 in Probability and Mathematical Statistics Monograph. Academic Press Inc, New York–London, 1974.

[71] V. Jog and V. Anantharam. The entropy power inequality and Mrs. Gerber's lemma for groups of order 2^n. *Information Theory, 2013 IEEE International Symposium on*, pages 594–598, 2013.

[72] J. Kahn. An entropy approach to the hard-core model on bipartite graphs. *Combin. Probab. Comput.*, 10(3):219–237, 2001.

[73] R. Kannan, L. Lovász, and M. Simonovits. Isoperimetric problems for convex bodies and a localization lemma. *Discrete Comput. Geom.*, 13(3-4):541–559, 1995.

[74] M. Kanter. Probability inequalities for convex sets and multidimensional concentration functions. *J. Multivariate Anal.*, 6(2):222–236, 1976.

[75] D. Katona and B. S. Stečkin. Combinatorial numbers, geometric constants and probabilistic inequalities. *Dokl. Akad. Nauk SSSR*, 251(6):1293–1296, 1980.

[76] G. O. H. Katona. Graphs, vectors and probabilistic inequalities. *Math. Lapok*, 20:123–127, 1969.

[77] G. O. H. Katona. Inequalities for the distribution of the length of a sum of random vectors. *Teor. Verojatnost. i Primenen.*, 22(3):466–481, 1977.

[78] G. O. H. Katona. Inequalities for the distribution of the length of random vector sums. *Theory of Probability & Its Applications*, 22(3):450–464, 1978.

[79] G. O. H. Katona. Continuous versions of some extremal hypergraph problems. II. *Acta Math. Acad. Sci. Hungar.*, 35(1-2):67–77, 1980.

[80] G. O. H. Katona. Sums of vectors and Turán's problem for 3-graphs. *European J. Combin.*, 2(2):145–154, 1981.

[81] G. O. H. Katona. "Best" estimations on the distribution of the length of sums of two random vectors. *Z. Wahrsch. Verw. Gebiete*, 60(3):411–423, 1982.

[82] G. O. H. Katona. Sums of vectors and turán's graph problem. *Ann. Discrete Math.*, 17:377–382, 1983.

[83] G. O. H. Katona. Probabilistic inequalities from extremal graph results (a survey). volume 118, pages 159–170. North-Holland, Amsterdam, 1985.

[84] H. Kesten. A sharper form of the Doeblin-Lévy-Kolmogorov-Rogozin inequality for concentration functions. *Math. Scand.*, 25:133–144, 1969.

[85] B. Klartag. An isomorphic version of the slicing problem. *J. Funct. Anal.*, 218(2):372–394, 2005.

[86] B. Klartag. On convex perturbations with a bounded isotropic constant. *Geom. Funct. Anal.*, 16(6):1274–1290, 2006.

[87] B. Klartag and V. D. Milman. Geometry of log-concave functions and measures. *Geom. Dedicata*, 112:169–182, 2005.

[88] A. Kolmogorov. Sur les propriétés des fonctions de concentrations de M. P. Lévy. *Ann. Inst. H. Poincaré*, 16:27–34, 1958.

[89] I. Kontoyiannis and M. Madiman. Sumset and inverse sumset inequalities for differential entropy and mutual information. *IEEE Trans. Inform. Theory*, 60(8):4503–4514, 2014.

[90] J. Kuelbs and W. V. Li. Metric entropy and the small ball problem for Gaussian measures. *J. Funct. Anal.*, 116(1):133–157, 1993.

[91] J. Kuelbs, W. V. Li, and M. Talagrand. Lim inf results for gaussian samples and chung's functional lil. *Ann. Probab.*, 22:1879–1903, 1994.

[92] A. Lapidoth and G. Pete. On the entropy of the sum and of the difference of two independent random variables. *Proc. IEEEI 2008, Eilat, Israel*, 2008.

[93] R Latała. On the equivalence between geometric and arithmetic means for log-concave measures. In *Convex geometric analysis (Berkeley, CA, 1996)*, volume 34 of *Math. Sci. Res. Inst. Publ.*, pages 123–127. Cambridge Univ. Press, Cambridge, 1999.

[94] M. Ledoux. Isoperimetry and Gaussian analysis. In *Lectures on probability theory and statistics (Saint-Flour, 1994)*, volume 1648 of *Lecture Notes in Math.*, pages 165–294. Springer, Berlin, 1996.

[95] M. Ledoux. *The concentration of measure phenomenon*, volume 89 of *Mathematical Surveys and Monographs*. American Mathematical Society, Providence, RI, 2001.

[96] M. Ledoux and M. Talagrand. *Probability on Banach spaces*. Springer, Berlin, 1991.

[97] J. Leech. The problem of the thirteen spheres. *Math. Gaz.*, 40:22–23, 1956.

[98] V. I. Levenshten. On bounds for packing in n-dimensional euclidean space. *Dokl. Akad. Nauk SSSR*, 245:1299–1303, 1979.

[99] P. Lévy. *Théorie de l'addition des variables aléatoires*. Paris, 1937.

[100] P. Lévy. *Problèmes Concrets d'Analyse Fonctionelle*. Gauthier-Villars, Paris, 1951.

[101] J. Li and M. Madiman. A combinatorial approach to small ball inequalities for sums and differences. *submitted*.

[102] J. Li and M. Madiman. Sum-product type estimates for shannon entropy. *In preparation*.

[103] J. Li, M. Madiman, L. Wang, and J. O. Woo. The 1-d slicing problem. *In preparation*.

[104] W. V. Li and W. Linde. Approximation, metric entropy and small ball estimates for Gaussian measures. *Ann. Probab.*, 27(3):1556–1578, 1999.

[105] W. V. Li and Q. Shao. Gaussian processes: inequalities, small ball probabilities and applications. In *Stochastic processes: theory and methods*, volume 19 of *Handbook of Statist.*, pages 533–597. North-Holland, Amsterdam, 2001.

[106] M. Lifshits, R. L. Schilling, and I. Tyurin. A probabilistic inequality related to negative definite functions. *Progress in Probability*, 66(73–80), 2013.

[107] M. Madiman. On the entropy of sums. *Proc. IEEE Infor. Theory Workshop*, pages 303–307, 2008.

[108] M. Madiman and I. Kontoyiannis. Entropy bounds on abelian groups and the ruzsa divergence. *http://arxiv.org/abs/1508.04089*.

[109] M. Madiman and I. Kontoyiannis. The entropies of the sum and the difference of two iid random variables are not too different. *Proc. IEEE Intl. Symp. Inform. Theory*, 2010.

[110] M. Madiman, A. Marcus, and P. Tetali. Entropy and set cardinality inequalities for partition-determined functions. *Random Structures Algorithms*, 40(4):399–424, 2012.

[111] M. Madiman and P. Tetali. Information inequalities for joint distributions, with interpretations and applications. *IEEE Trans. Inform. Theory*, 56(6):2699–2713, 2010.

[112] M. Madiman, L. Wang, and S Bobkov. Some applications of the nonasymptotic equipartitin property of log-concave distributions. *Preprint*, 2016.

[113] J. Marica. On a conjecture of Conway. *Canad. Math. Bull.*, 12:233–234, 1969.

[114] G. Martin and K. O'Bryant. Many sets have more sums than differences. In *Additive combinatorics*, volume 43 of *CRM Proc. Lecture Notes*, pages 287–305. Amer. Math. Soc., Providence, RI, 2007.

[115] L. Mattner. Strict definiteness of integrals via complete monotonicity of derivatives. *Trans. Amer. Math. Soc.*, 349(8):3321–3342, 1997.

[116] B. Maurey. Some deviation inequalities. *Geom. Funct. Anal.*, 1(2):188–197, 1991.

[117] B. McMillan. The basic theorems of information theory. *Ann. Math. Statistics*, 24:196–219, 1953.

[118] S. J. Miller, B. Orosz, and D. Scheinerman. Explicit constructions of infinite families of MSTD sets. *J. Number Theory*, 130(5):1221–1233, 2010.

[119] V. D. Milman. A new proof of A. Dvoretzky's theorem on cross-sections of convex bodies. *Funkcional. Anal. i Prilozhen*, 5(4):28–37, 1971.

[120] V. D. Milman. The heritage of P. Lévy in geometrical functional analysis. *Astérisque*, (157-158):273–301, 1988. Colloque Paul Lévy sur les Processus Stochastiques (Palaiseau, 1987).

[121] V. D. Milman and A. Pajor. Isotropic position and inertia ellipsoids and zonoids of the unit ball of a normed n-dimensional space. In *Geometric aspects of functional analysis (1987–88)*, volume 1376 of *Lecture Notes in Math.*, pages 64–104. Springer, Berlin, 1989.

[122] V. D. Milman and G. Schechtman. *Asymptotic theory of finite-dimensional normed spaces*, volume 1200 of *Lecture Notes in Mathematics*. Springer-Verlag, Berlin, 1986. With an appendix by M. Gromov.

[123] A. L. Mirošnikov and B. A. Rogozin. Inequalities for concentration functions. *Teor. Veroyatnost. i Primenen.*, 25(1):178–183, 1980.

[124] R. Mori and T. Tanaka. Non-binary polar codes using reed-solomon codes and algebraic geometry codes. In *Proc. IEEE Inform. Theory Workshop*, Dublin, 2010.

[125] O. R. Musin. The problem of the twenty-five spheres. *Uspekhi Mat. Nauk*, 58(4(352)):153–154, 2003.

[126] M. B. Nathanson. Problems in additive number theory. I. In *Additive combinatorics*, volume 43 of *CRM Proc. Lecture Notes*, pages 263–270. Amer. Math. Soc., Providence, RI, 2007.

[127] M. B. Nathanson. Sets with more sums than differences. *Integers*, 7:A5, 24, 2007.

[128] H. H. Nguyen and V. Vu. Small ball probability, inverse theorems, and applications. In *Erdös centennial*, volume 25 of *Bolyai Soc. Math. Stud.*, pages 409–463. János Bolyai Math. Soc., Budapest, 2013.

[129] V. H. Nguyen. Dimensional variance inequalities of Brascamp-Lieb type and a local approach to dimensional Prékopa's theorem. *J. Funct. Anal.*, 266(2):931–955, 2014.

[130] A. M. Odlyzko and N. J. A. Sloane. New bounds on the number of unit spheres that can touch a unit sphere in n dimensions. *J. Combin. Theory Ser. A*, 26(2):210–214, 1979.

[131] S. Orey. On the shannon-perez-moy theorem. In *Particle systems, random media and large deviations (Brunswick, Maine, 1984)*, pages 319–327. Providence, R.I.: Amer. Math. Soc.,, 1985.

[132] Sophie Piccard. *Sur les ensembles de distances des ensembles de points d'un espace Euclidien*. Mém. Univ. Neuchâtel, vol. 13. Secrétariat de l'Université, Neuchâtel, 1939.

[133] V. P. Pigarev and G. A. Freĭman. The relation between the invariants R and T. In *Number-theoretic studies in the Markov spectrum and in the structural theory of set addition (Russian)*, pages 172–174. Kalinin. Gos. Univ., Moscow, 1973.

[134] G. Pisier. *The volume of convex bodies and Banach space geometry*, volume 94 of *Cambridge Tracts in Mathematics*. Cambridge University Press, Cambridge, 1989.

[135] J. Radhakrishnan. An entropy proof of Bregman's theorem. *J. Combin. Theory Ser. A*, 77:161–164, 1997.

[136] J. Radhakrishnan. Entropy and counting. In *Computational Mathematics, Modeling and Algorithms*. Narosa, 2003.

[137] B. A. Rogozin. On the increase of dispersion of sums of independent random variables. *Teor. Verojatnost. i Primenen*, 6:106–108, 1961.

[138] R. M. Roth. *Introduction to coding theory*. Cambridge Univ. Press, Cambridge, 2006.

[139] M. Rudelson and R. Vershynin. The Littlewood-Offord problem and invertibility of random matrices. *Adv. Math.*, 218(2):600–633, 2008.

[140] I. Z. Ruzsa. On the cardinality of $A + A$ and $A - A$. In *Combinatorics (Proc. Fifth Hungarian Colloq., Keszthely, 1976), Vol. II*, volume 18 of *Colloq. Math. Soc. János Bolyai*, pages 933–938. North-Holland, Amsterdam-New York, 1978.

[141] I. Z. Ruzsa. On the number of sums and differences. *Acta Math. Hungar.*, 59(3-4):439–447, 1992.

[142] I. Z. Ruzsa. Sums of finite sets. In *Number theory (New York, 1991–1995)*, pages 281–293. Springer, New York, 1996.

[143] I. Z. Ruzsa. Sumsets and entropy. *Random Structures Algorithms*, 34(1):1–10, 2009.

[144] G. Schechtman. Concentration results and applications. In *Handbook of the geometry of Banach spaces, Vol. 2*, pages 1603–1634. North-Holland, Amsterdam, 2003.

[145] S. Shamai and A. Wyner. A binary analog to the entropy-power inequality. *IEEE Trans. Inform. Theory*, 36(6):1428–1430, 1990.

[146] C. E. Shannon. A mathematical theory of communication. *Bell System Tech. J.*, 27:379–423, 623–656, 1948.

[147] A. F. Sidorenko. Extremal estimates of probability measures and their combinatorial nature. *Izv. Akad. Nauk SSSR Ser. Mat.*, 46(3):535–568, 671, 1982.

[148] R. Siegmund-Schultze and H. von Weizsäcker. Level crossing probabilities. I. One-dimensional random walks and symmetrization. *Adv. Math.*, 208(2):672–679, 2007.

[149] J. Solymosi. Bounding multiplicative energy by the sumset. *Adv. Math.*, 222(2):402–408, 2009.

[150] A. J. Stam. Some inequalities satisfied by the quantities of information of Fisher and Shannon. *Information and Control*, 2:101–112, 1959.

[151] E. M. Stein and R. Shakarchi. *Fourier analysis: an introduction*. Princeton Univ. Press, Princeton, NJ, 2003.

[152] S. K. Stein. The cardinalities of $A+A$ and $A-A$. *Canad. Math. Bull.*, 16:343–345, 1973.

[153] S. J. Szarek and D. Voiculescu. Volumes of restricted Minkowski sums and the free analogue of the entropy power inequality. *Comm. Math. Phys.*, 178(3):563–570, 1996.

[154] M. Talagrand. An isoperimetric theorem on the cube and the Kintchine-Kahane inequalities. *Proc. Amer. Math. Soc.*, 104(3):905–909, 1988.

[155] M. Talagrand. A new isoperimetric inequality and the concentration of measure phenomenon. In *Geometric aspects of functional analysis (1989–90)*, volume 1469 of *Lecture Notes in Math.*, pages 94–124. Springer, Berlin, 1991.

[156] M. Talagrand. New Gaussian estimates for enlarged balls. *Geom. Funct. Anal.*, 3(5):502–526, 1993.

[157] M. Talagrand. Concentration of measure and isoperimetric inequalities in product spaces. *Inst. Hautes Études Sci. Publ. Math.*, (81):73–205, 1995.

[158] T. Tao. Sumset and inverse sumset theory for Shannon entropy. *Combin. Probab. Comput.*, 19(4):603–639, 2010.

[159] M. Tribus and E. C. Mclrvine. Energy and information. *Scientific American*, 224, 1971.

[160] S. R. S. Varadhan. Asymptotic probabilities and differential equations. *Comm. Pure Appl. Math.*, 19:261–286, 1966.

[161] S. R. S. Varadhan. *Large deviations and applications*, volume 46 of *CBMS-NSF Regional Conference Series in Applied Mathematics*. Society for Industrial and Applied Mathematics, 1984.

[162] M. S. Viazovska. The sphere packing problem in dimension 8.

[163] L. A. Vinh. The szemerédi-trotter type theorem and the sum-product estimate in finite fields. *European J. Combin.*, 32:1177–1181, 2011.

[164] L. Wang, J. O. Woo, and M. Madiman. A lower bound on the Rényi entropy of convolutions in the integers. In *Proc. IEEE Intl. Symp. Inform. Theory*, Honolulu, Hawaii, July 2014.

[165] T. Wolff. Recent work connected with the Kakeya problem. In *Prospects in mathematics (Princeton, NJ, 1996)*, pages 129–162. Amer. Math. Soc., Providence, RI, 1999.

[166] J. O. Woo and M. Madiman. A discrete entropy power inequality for uniform distributions. In *Proc. IEEE Intl. Symp. Inform. Theory*, Hong Kong, China, June 2015.

[167] Y. Wu, S. Shamai, and S. Verdú. Degrees of freedom of the interference channel: a general formula. *Proc. IEEE Intl. Symp. Inform. Theory*, pages 1344–1348, 2011.

[168] A. D. D. Wyner and J. Ziv. A theorem on the entropy of certain binary sequences and applications. I. *IEEE Trans. Information Theory*, IT-19:769–772, 1973.

[169] Y. Xiao. Fractal measures of the sets assiciated to gaussian random fields. In *Trends in Probability and Related Analysis: Proceedings of the Symposium on Analysis and Probability*, pages 311–324. World Scientific, 1997.

[170] Y. Zhao. Constructing MSTD sets using bidirectional ballot sequences. *J. Number Theory*, 130(5):1212–1220, 2010.

[171] Y. Zhao. Counting MSTD sets in finite abelian groups. *J. Number Theory*, 130(10):2308–2322, 2010.

Appendix A

PERMISSION LETTER

Copyright Clearance Center

Note: Copyright.com supplies permissions but not the copyrighted content itself.

1 PAYMENT 2 REVIEW **3 CONFIRMATION**

Step 3: Order Confirmation

Thank you for your order! A confirmation for your order will be sent to your account email address. If you have questions about your order, you can call us at +1.855.239.3415 Toll Free, M-F between 3:00 AM and 6:00 PM (Eastern), or write to us at info@copyright.com. This is not an invoice.

Confirmation Number: 11559373
Order Date: 04/28/2016

If you paid by credit card, your order will be finalized and your card will be charged within 24 hours. If you choose to be invoiced, you can change or cancel your order until the invoice is generated.

Payment Information

Jiange Li
lijiange@udel.edu
+1 (302)3679989
Payment Method: n/a

Order Details

Journal of theoretical probability

Order detail ID:	69704437
Order License Id:	3857751310107
ISSN:	0894-9840
Publication Type:	Journal
Volume:	
Issue:	
Start page:	
Publisher:	SPRINGER NEW YORK LLC

Permission Status: Granted
Permission type: Republish or display content
Type of use: Thesis/Dissertation

Requestor type	Author of requested content
Format	Electronic
Portion	chapter/article
Number of pages in chapter/article	13
Title or numeric reference of the portion(s)	Chapter 2, sections 2.1 and 2.2
Title of the article or chapter the portion is from	A note on distribution-free symmetrization inequalities
Editor of portion(s)	N/A
Author of portion(s)	Jiange Li

Volume of serial or monograph	N/A
Page range of portion	2, 15-26
Publication date of portion	September, 2016
Rights for	Main product
Duration of use	Current edition and up to 5 years
Creation of copies for the disabled	no
With minor editing privileges	no
For distribution to	United States
In the following language(s)	Original language of publication
With incidental promotional use	no
Lifetime unit quantity of new product	Up to 499
Made available in the following markets	education
The requesting person/organization	Jiange Li
Order reference number	
Author/Editor	Jiange Li
The standard identifier	NA
Title	Some topics in probability theory, combinatorics and information theory
Publisher	ProQuest
Expected publication date	Sep 2016
Estimated size (pages)	112