Information Inequalities: Subadditivity & Relationships with Additive Combinatorics

LAU, Chin Wa

A Thesis Submitted in Partial Fulfilment
of the Requirements for the Degree of
Doctor of Philosophy
in
Information Engineering

The Chinese University of Hong Kong
July 2025

Abstract of thesis entitled: *Information Inequalities: Subadditivity & Relationships with Additive Combinatorics*

Submitted by LAU, Chin Wa

for the degree of Doctor of Philosophy

at The Chinese University of Hong Kong in July 2025

Information inequalities play a fundamental role in establishing outer bounds of the achievable region in multi-user information theory. Beyond this specific application, these inequalities serve important roles in multidisciplinary research. This thesis focuses on utilizing information-theoretic tools to build connections and equivalence theorems across various fields including projection theory, functional analysis, and additive combinatorics. Additionally, we aim to utilize ideas from these fields to develop machinery and tools for proving new entropic inequalities.

We begin by studying a fundamental supermodularity inequality for mutual information, which is associated with the "compression" partial order from combinatorics. By constructing appropriate auxiliary random variables, we simplify proofs of existing results and generalize the supermodularity results to Fisher information, Kullback–Leibler divergence, and strong data processing inequality constant, among others.

By combining supermodularity with "rotation"-trick, we develop machinery for establishing the optimality of uniform distributions in optimizing entropic functionals over finite abelian groups. This approach leads to a discrete analogue of the entropy power inequality and has direct applications to the polynomial Freiman-Ruzsa conjecture.

We further extend our submodularity results to sumset inequalities through an equivalence theorem framework. By introducing copies of random variables with suitable Markovian structures, we establish new entropic inequalities via submodularity. We also propose an entropic formulation for magnification ratio in a bipartite graph, laying groundwork for deeper connections between sumset theory and information theory.

論文題為《資訊不等式：次可加性與加法組合學之關聯》之摘要

呈交者：劉展華

申請哲學博士學位

於香港中文大學，2025 年 5 月

資訊不等式在多用戶資訊理論中，對於建立可行區域的外界範圍起著基礎性作用。除了此特定應用外，這些不等式在跨學科研究中也擔當重要角色。本論文聚焦運用資訊理論工具，在投影理論、泛函分析與加法組合學等多個領域間建立聯繫與等價定理，並致力將這些領域的思想轉化為證明新型熵不等式的方法工具。

我們首先研究互資訊的基礎超模性不等式，其與組合學中「壓縮」偏序關係相關。通過構建適當輔助隨機變量，我們簡化了現有結果的證明，並將超模性結果推廣至費雪資訊量、Kullback–Leibler 散度及强數據處理不等式常量等領域。

結合超模性與「旋轉」技巧，我們建立了一套方法論用於證明有限阿貝爾群上均匀分布對熵泛函優化的最優性。此方法導出熵冪不等式的離散類比，並可直接應用於多項式形式的 Freiman–Ruzsa 猜想。

我們進一步通過等價定理框架將次模性結果延伸至和集不等式。藉由引入具有適當馬可夫結構的隨機變量副本，利用次模性建立新型熵不等式。同時提出二部圖擴展性的熵形式化表述，為和集理論與資訊理論的深度聯繫奠定基礎。

Acknowledgement

I am deeply grateful to my doctoral advisor, Chandra Nair, whose guidance has been foundational to my academic journey. Prior to encountering his mentorship during my undergraduate studies, I had intended to join the software industry immediately after graduation. It was his encouragement that inspired me to explore research, revealing a path I had not envisioned. He possesses a remarkable ability to recognize and cultivate students' potential, paired with humility, patience, and intellectual curiosity. He consistently welcomed discussions of my nascent ideas, no matter how abstract, and it was through this openness that connections between additive combinatorics and diverse areas of information theory came to light. Without his support, such interdisciplinary insights would have remained beyond reach.

My gratitude extends to all my graduate school mentors, whose teachings equipped me with the tools to thrive in research. Special thanks go to Amin Gohari, Farzan Farnia, Cheuk Ting Li, Pascal Vontobel, and Raymond Wai Ho Yeung for imparting their expertise in information theory, coding theory, learning theory, and automated theorem-proving frameworks. Their instruction solidified the technical foundation underlying my work.

I would like to express my sincere appreciation to my thesis committee members, Ioannis Kontoyiannis, Amin Gohari, Cheuk Ting Li, Chandra Nair, and Raymond Wai Ho Yeung, for their time and valuable feedback in refining this thesis. Their insights have enhanced both the rigor and accessibility of this work.

I am equally thankful for my colleagues, whose brilliant minds and friendship have enriched this journey. Collaborating with Yannan Wang, Wai Ho Ng, Jinpei Zhao, Yi Liu, Zijie Chen, Zhaobang Zhu, Qinghua Ding, Chaorui Yao, Daoyuan Chen, Yunrui Guan, Chun Hei Shiu, Chih Wei Ling, Yanxiao Liu, Yuwen Huang, and Junda Zhou has been both an honor and a source of motivation. They generously engaged with my ideas, shared their own innovations, and became cherished friends who offered wisdom beyond academia, ensuring I never faced challenges alone.

Above all, my heartfelt appreciation goes to my parents, Siu Lai Lau and Wai Keung Lau, for their unwavering support. Their steadfast belief in my aspirations gave me the courage to pursue a doctoral degree, and their unconditional love has been my constant anchor. It is to them that I dedicate my academic journey.

Lastly, I offer my most profound thanks to the Lord, whose boundless love and wisdom sustain me. Every insight gained and every obstacle overcome is a testament to His grace. To Him belongs all glory.

This thesis is dedicated to my mother and father.

Contents

Abstract

摘要 i

Acknowledgement ii

List of Publications viii

Notation ix

1 Introduction 1

1.1 Preliminaries of entropic algebra 2

1.2 Application for subadditive and submodular inequality 3

1.2.1 Shearer's Lemma and Han's inequality 4

1.2.2 Establishment of Gaussian optimality 5

1.2.3 Entropy power inequality 7

1.3 Connections to functional inequalities 9

1.3.1 Entropic proof of Hölder inequality 9

1.3.2 Forward Brascamp–Lieb inequality 12

1.4 Connections to additive combinatorics 14

1.4.1 Parallelism between sumset inequalities and entropic inequalities 15

1.4.2 Historical remark 17

1.4.3 Ruzsa equivalence theorem on sumset inequalities 18

1.5 Structure of the thesis 19

2 Supermodular Information Inequalities and their Applications 21

2.1 Preliminaries 22

2.2 Two families of perturbative auxiliaries 27

2.3 Some consequences of the supermodularity inequalities 31

2.3.1 Entropy power inequalities and Fisher information inequalities 31

2.3.2 Discrete convexity, strong data processing constant and maximal correlation 41

2.4 Connection with additive combinatorics 51

3 Rotation Trick in Discrete Spaces and Its Applications 54

3.1 Preliminaries 55

3.2 Discrete analogue of Unified BLI and EPI 56

3.2.1 Framework for establishing optimality of uniform distribution 59

3.3 Application in additive combinatorics 64

3.3.1 Entropic formulation of PFR conjecture 64

3.3.2 Elementary superadditive results for PFR functional . . . 66

3.3.3 A superadditivity proof for the optimality of uniform distribution in PFR functional 68

3.4 Discussion 74

4 Inequalities and Links to Additive Combinatorics 76

4.1 Generalized Ruzsa-type equivalence theorem 77

4.2 Application of generalized Ruzsa-type equivalence theorem 81

4.2.1 Fundamental maximal entropic coupling inequalities . . . 81

4.2.2 Sum-difference inequality 84

4.2.3 Connection to the covering lemma 94

4.3 Entropic formulation of magnification ratio 98

4.3.1 Overview of the proof framework 99

4.3.2 Properties of the optimal channel for the magnification ratio 101

4.3.3 Properties of the minimal input distribution for the magnification ratio . . . 103
4.4 Discussion . . . 110

A Preliminary inequalities for Section 3.3.3 112
A.1 Estimates related to PFR conjecture . . . 112

B Type counting argument 119
B.1 Discrete Sanov Theorem . . . 119
B.2 Maximal couplings . . . 128

C Proof of Theorem 4.2.15 132

Bibliography 136

List of Publications

The work presented in this thesis is based on the following papers.

1. K. Lau, C. Nair and D. Ng, "A mutual information inequality and some applications," *2022 IEEE International Symposium on Information Theory (ISIT)*, Espoo, Finland, 2022, pp. 951-956, doi: 10.1109/ISIT50566.2022.9834382.

2. C. W. Lau, C. Nair, and D. Ng, "A Mutual Information Inequality and Some Applications," *IEEE Transactions on Information Theory*, vol. 69, no. 10, pp. 6210—6220, 2023, doi: 10.1109/TIT.2023.3285928.

3. C. W. (Ken) Lau and C. Nair, "Information Inequalities via Ideas from Additive Combinatorics," in *2023 IEEE International Symposium on Information Theory (ISIT)*, 2023, pp. 2452—2457. doi: 10.1109/ISIT54713.2023.10206561.

4. C. W. Lau and C. Nair, "Information Inequalities via Ideas from Additive Combinatorics," in *IEEE Transactions on Information Theory*, doi: 10.1109/TIT.2025.3557796.

5. C. W. Lau and C. Nair, "An Entropic Inequality in Finite Abelian Groups Analogous to the Unified Brascamp-Lieb and Entropy Power Inequality," in *2024 IEEE International Symposium on Information Theory (ISIT)*, 2024, pp. 3588—3593.

6. C. W. Lau and C. Nair, "An entropic inequality in finite Abelian groups analogous to the unified Brascamp-Lieb and Entropy Power Inequality," to appear in *Andrew Barron's Festschrift Volume.*

Notation

This section provides a general guideline for the notation used throughout this thesis. Any deviations from these conventions are explicitly noted when they occur.

$\mathbb{Z}$	the set of integers
$\mathbb{Z}^d$	the d-dimensional integer lattice
$\mathbb{R}$	the set of real numbers
$\mathbb{R}^d$	d-dimensional Euclidean space
$\langle \cdot, \cdot \rangle$	the inner product in $\mathbb{R}^d$
$\lvert \cdot \rvert$	the Euclidean norm on $\mathbb{R}^d$
$q \ll p$	q is absolutely continuous with respect to p
$[a:b]$	the set of integers from a to b, inclusive
$\lvert S \rvert$	the cardinality of set S
$A \succeq 0$	A is a positive definite matrix
$A \succeq B$	$A - B \succeq 0$
$f, g, \ldots$	functions
$X, Y, \ldots$	random variables
$\mathcal{X}, \mathcal{Y}, \ldots$	the support of discrete random variables X, Y
$x, y, \ldots$	constants or values of random variables
X_S	the tuple of random variables $\{X_i\}_{i \in S}$
$\mathbb{P}(E)$	the probability of an event E
p_X	the probability mass function of discrete random variable X
$p_{X,Y}$	the joint probability mass function

	of discrete random variables X and Y
$p_{Y\|X}$	the conditional probability mass function of Y given X
$\Pi(X_1, \ldots, X_n)$	the collection of joint distributions $p_{X_1,\ldots,X_n}$
	that are consistent with the marginal distributions $p_{X_1}, \ldots, p_{X_n}$
W	transition channel between two random variables
μ_X	the probability density function of continuous random variable X
$X \sim p_X$	random variable X follows the probability mass function p_X
$\mathrm{supp}(p_X)$	the support of probability mass function p_X
$X \perp Y$	random variables X and Y are independent
$X \perp Y\|Z$	random variables X and Y are conditionally independent given Z
$X_1 \to \cdots \to X_n$	random variables $X_1, \ldots, X_n$ form a Markov chain
$\mathrm{E}[X]$	the expectation of random variable X
$\mathrm{E}[Y\|X]$	the conditional expectation of random variable Y given X
$H(X)$	the entropy of a discrete random variable X
	$H(X) := -\sum_x p_X(x) \log p_X(x)$
$H(Y\|X)$	the conditional entropy of a discrete random variable Y given X
	$H(Y\|X) := -\sum_{x,y} p_{X,Y}(x,y) \log p_{Y\|X}(y\|x)$
$h(X)$	the differential entropy of a continuous random variable X
	$h(X) := -\int_x \mu_X(x) \log \mu_X(x)$
$h(Y\|X)$	the differential entropy of a continuous random variable Y given X
	$h(Y\|X) := -\int_{x,y} \mu_{X,Y}(x,y) \log \mu_{Y\|X}(y\|x)$
$\rho_X(X)$	the score function of a continuous random variables X
	$\rho_X(X) := \nabla \log f_X$
$J(X)$	the Fisher information of a continuous random variable X
	$J(X) := \mathrm{E}[\\|\rho_X(X)\\|^2]$
$\frac{d\mu}{d\nu}$	the Radon–Nikodym derivative
	of distribution μ with respect to distribution ν
$D(\mu\\|\nu)$	the Kullback—Leibler divergence
	of distribution μ from distribution ν

	$D(\mu \Vert \nu) := \mathrm{E}_{\mu}\left[\log\left(\frac{d\mu}{d\nu}\right)\right]$
$I(X;Y)$	the mutual information between random variables X and Y
	$I(X;Y) := D(\mu_{X,Y} \Vert \mu_X \mu_Y)$
$I(X;Y \mid Z)$	the conditional mutual information
	between random variables X and Y given Z
	$I(X;Y \mid Z) := \mathrm{E}_{\mu_Z}\left[D(\mu_{X,Y \mid Z} \Vert \mu_{X \mid Z}\mu_{Y \mid Z})\right]$
$\mathbb{G}, \mathbb{H}, \ldots$	finite groups
$\mathbb{T}$	a finitely generated torsion-free Abelian group

Chapter 1

Introduction

Information inequalities play a fundamental role in establishing outer bounds of the achievable region in multi-user information theory. Beyond this specific application, these inequalities serve important roles in multidisciplinary research. This thesis focuses on utilizing information-theoretic tools to build connections and equivalence theorems across various fields including projection theory, functional analysis, and combinatorics. It aims to simplify and generalize existence results through a submodularity framework. Additionally, we seek to utilize ideas from these fields to develop machinery and tools for proving new entropic inequalities.

In this chapter, we provide necessary preliminaries on the fields that we intend to connect with information inequalities. We begin with Section 1.1, which covers the fundamental entropic inequalities that will appear frequently throughout this thesis.

By leveraging the basic information inequalities from entropic algebra, we can establish that entropy and differential entropy are both subadditive and submodular. In Section 1.2, we briefly explore several applications of submodular entropic inequalities, including Shearer's Lemma and the entropy power inequality. These results motivate our search for a more generalized family of supermodularity inequalities, and their relationships to different information quantities, such as Fisher information and KL divergence.

Through Legendre duality, we can establish equivalence between functional inequalities and entropic inequalities. In Section 1.3, we start with a simple entropic proof of Hölder's inequality. By combining the concepts of subadditivity and the "rotation" technique, we establish the Gaussian optimality of a family of differential entropic functionals for continuous random variables, which unifies the entropic formulation of the Brascamp–Lieb inequality and the entropy power inequality. In this thesis, we develop a discrete counterpart of this mechanism, which establishes the optimality of uniform distribution for a family of entropic functionals.

In Section 1.4, we introduce the concept of sumsets from additive combinatorics and the potential parallelism between sumset inequalities and entropic inequalities. We also discuss previous attempts to build connections between these two domains. Inspired by these attempts, we establish a generalized equivalence theorem between sumset inequalities and entropic inequalities, and provide standalone entropic proofs derived from combinatorial construction techniques.

In Section 1.5, we outline the overall structure of this thesis and summarize our contributions.

1.1 Preliminaries of entropic algebra

We now introduce fundamental inequalities in entropic algebra. For a comprehensive treatment and proofs of these properties, we refer the reader to Chapter 2 of Cover and Thomas [CT91].

- **Optimality of uniform distribution:** For a discrete random variable X with finite support $\mathcal{X}$, we have $H(X) \leq \log|\mathcal{X}|$, with equality if and only if X follows a uniform distribution on $\mathcal{X}$.

- **Non-negativity of entropy:** For any discrete random variable X, we have $H(X) \geq 0$.

- **Chain rule of entropy:** For a sequence of discrete random variables $X_1, \ldots, X_n$, we have $H(X_1, \ldots, X_n) = \sum_{i=1}^{n} H(X_i|X_1, \ldots, X_{i-1})$.

- **Conditional mutual information:** For discrete random variables X, Y, Z, we have $I(X;Y|Z) = H(X,Z)+H(Y,Z)-H(X,Y,Z)-H(Z)$. Furthermore, $I(X;Y|Z) \geq 0$.

- **Data processing inequality:** If $X \to Y \to Z$ forms a Markov chain, then $I(X;Y) \geq I(X;Z)$.

1.2 Application for subadditive and submodular inequality

In this section, we demonstrate applications of subadditivity and submodularity in various scenarios. In Section 1.2.1, we present direct consequences of the submodularity of entropy and its applications in projection theory and combinatorics. In Section 1.2.2, we show how a subadditivity approach establishes Gaussian optimality. By combining these results, we derive the entropy power inequality (EPI) using a novel proof technique in Section 2.3.1, following our introduction to different EPI formulations in Section 1.2.3.

First, we introduce the definitions of subadditivity and submodularity, then establish that both entropy and differential entropy satisfy these properties.

Definition 1.2.1 (Subadditivity and submodularity)**.** Let Ω be a finite set with power set $\mathcal{P}(\Omega)$. For a function $f : \mathcal{P}(\Omega) \to \mathbb{R}$, we say:

- f is subadditive if $f(A \cup B) \leq f(A) + f(B)$ for all subsets $A, B \subseteq \Omega$.

- f is submodular if $f(A \cup B) + f(A \cap B) \leq f(A) + f(B)$ for all subsets $A, B \subseteq \Omega$.

Proposition 1.2.2 (Subadditivity and submodularity of entropy and differential entropy)**.** *Let n be a positive integer and $\Omega = [1:n]$. For a tuple of discrete-*

valued random variables $(X_1, \ldots, X_n)$, define $f(A) = H(X_A) = H(\{X_i\}_{i \in A})$ where $A \subseteq \Omega$. Then f is both subadditive ($f(A \cup B) \leq f(A) + f(B)$) and submodular ($f(A \cup B) + f(A \cap B) \leq f(A) + f(B)$). Analogous results hold for $f(A) := h(X_A)$ (differential entropy) when $X_1, \ldots, X_n$ are continuous-valued random variables.

Proof. Both properties follow directly from information theory: subadditivity from the non-negativity of mutual information, and submodularity from the non-negativity of conditional mutual information. □

1.2.1 Shearer's Lemma and Han's inequality

A direct consequence of submodularity of entropy is Shearer's Lemma (prominent in theoretical computer science) or Han's inequality (well-established in information theory):

Lemma 1.2.3 ([CGFS86, Han78]). *Let $X_1, \ldots, X_n$ be random variables. For subsets $S_1, \ldots, S_m$ of $\Omega := [1 : n]$ where each element $i \in \Omega$ appears in at least r subsets, we have*

$$H(X_1, \ldots, X_n) \leq \frac{1}{r} \sum_{k=1}^{m} H(X_{S_k}).$$

Shearer's Lemma immediately yields the following equivalent projection inequality:

Theorem 1.2.4 ([BB12, Ruz09a]). *Let $K \subseteq \mathbb{R}^d$ be a finite set of points. Let $\Omega := [1 : d]$ and consider subsets $S_1, \ldots, S_m$ of Ω where each element $i \in \Omega$ appears in at least r subsets. For each subset S_j, define the projection K_{S_j} as:*

$$K_{S_j} := \{\pi_{S_j}(x) : x \in K\}$$

where $\pi_{S_j} : \mathbb{R}^d \to \mathbb{R}^{|S_j|}$ is the projection onto the coordinate subspace indexed by S_j. Specifically, for $x = (x_1, x_2, \ldots, x_d) \in K$, we have $\pi_{S_j}(x) = (x_i)_{i \in S_j}$.

Then the following inequality holds:

$$|K|^r \le \prod_{j=1}^{m} |K_{S_j}|.$$

Proof. For $X_1, \dots, X_n$ uniformly distributed over K, we have $H(X_1, \dots, X_n) = \log |K|$ and $H(X_{S_k}) \le \log |K_{S_k}|$. Applying Shearer's Lemma yields the inequality. The converse follows from information-theoretic typicality arguments detailed in Chapter 4. □

The submodularity of differential entropy extends to a continuous analog of Shearer's Lemma through the Loomis–Whitney inequality:

Theorem 1.2.5 (Loomis–Whitney inequality)**.** *Let $K \subseteq \mathbb{R}^d$ be a measurable set. For subsets $S_1, \dots, S_m$ of $\Omega := [1:d]$ where each element $i \in \Omega$ appears in at least r subsets, with projections $K_{S_1}, \dots, K_{S_m}$ onto their respective coordinate subspaces:*

$$(\mathrm{vol}(K))^r \le \prod_{k=1}^{m} (\mathrm{vol}(K_{S_k})),$$

where $\mathrm{vol}(K_{S_k})$ is the volume of the projection K_{S_k} in the respective space.

Shearer's Lemma has numerous applications in combinatorics; for further exploration, see [Fri04, Gal14, Rad03a]. Balister and Bollobás [BB12] unified Shearer's Lemma [CGFS86, Rad03b], Han's inequality [Han78], and the Madiman–Tetali inequality [MT10] through a partial ordering framework called "compression," which we examine in Section 2.1. Recent advances in submodularity-based information inequalities are documented in [IKBA22, Sas22, Tia11].

1.2.2 Establishment of Gaussian optimality

In [GN14], Geng and Nair employed a subadditivity framework with a "rotation trick" to establish Gaussian optimality of differential entropy functionals, subsequently determining the capacity region for Gaussian broadcast channels with common messages. Previous approaches typically utilized heat flow methods involving Fisher information inequalities, which required sophisticated analytical techniques.

We now demonstrate how subadditivity principles establish Gaussian optimality under covariance constraints:

Theorem 1.2.6. *Let X be a real-valued random variable satisfying the power constraint $\mathrm{E}[XX^T] \preceq K$, where K is a symmetric positive semidefinite matrix. The differential entropy $h(X)$ is maximized by a Gaussian distribution.*

Proof. Without going into the technical detail, we may assume the target optimization problem has a maximizer X^* and the corresponding maximum value of $V^* := \sup_{X:\mathrm{E}[XX^T]\preceq K} h(X)$.

Observed that $h(X)$ is subadditive, we can establish Gaussian optimality through a structured approach. First, we take X_1^* and X_2^* as independent and identical copies of the optimizer X^*.

Next, we apply a rotation to the vector (X_1^*, X_2^*), resulting in $\left(\frac{X_1^*+X_2^*}{\sqrt{2}}; \frac{X_1^*-X_2^*}{\sqrt{2}}\right)$. Our strategy is to show that both $X_1^* \perp X_2^*$ and $\frac{X_1^*+X_2^*}{\sqrt{2}} \perp \frac{X_1^*-X_2^*}{\sqrt{2}}$. If we can establish these independence relationships, the Darmois-Skitovich theorem (Theorem 3.1.1) implies that X_1^* and X_2^* must be Gaussians with identical covariance matrices.

This rotation preserves key properties: the differential entropy of $h(X_1^*, X_2^*)$ and the power constraints. Since X_1^* and X_2^* are independent and satisfy the power constraints, the rotated forms $\frac{X_1^*+X_2^*}{\sqrt{2}}$ and $\frac{X_1^*-X_2^*}{\sqrt{2}}$ will also satisfy the power constraints.

Therefore, our goal is to establish the independence relationship of the rotated form. In particular, we would like to show $I\left(\frac{X_1^*+X_2^*}{\sqrt{2}}; \frac{X_1^*-X_2^*}{\sqrt{2}}\right) = 0$.

$$2V^* = h(X_1^*) + h(X_2^*) \overset{(a)}{=} h(X_1^*, X_2^*) = h\left(\frac{X_1^*+X_2^*}{\sqrt{2}}, \frac{X_1^*-X_2^*}{\sqrt{2}}\right)$$

$$= h\left(\frac{X_1^* + X_2^*}{\sqrt{2}}\right) + h\left(\frac{X_1^* - X_2^*}{\sqrt{2}}\right) - I\left(\frac{X_1^* + X_2^*}{\sqrt{2}}; \frac{X_1^* - X_2^*}{\sqrt{2}}\right)$$

$$\overset{(b)}{\leq} 2V^* - I\left(\frac{X_1^* + X_2^*}{\sqrt{2}}; \frac{X_1^* - X_2^*}{\sqrt{2}}\right),$$

where (a) follows because X_1^* and X_2^* are independent, (b) follows because $\frac{X_1^*+X_2^*}{\sqrt{2}}$ and $\frac{X_1^*-X_2^*}{\sqrt{2}}$ satisfy the power constraint, so the value of their differential entropy must be less than or equal to the maximum value.

By using the non-negativity of the mutual information, this forces $I\left(\frac{X_1^*+X_2^*}{\sqrt{2}}; \frac{X_1^*-X_2^*}{\sqrt{2}}\right) = 0$, and we have established the Gaussian optimality. □

This framework extends generally to subadditive optimization functionals, establishing Gaussian optimality across various information-theoretic scenarios. For interested readers, please refer to [GN14, AJN22, LCCV18, SG22].

1.2.3 Entropy power inequality

In this subsection, we will introduce the celebrated *entropy power inequality* (EPI), a powerful tool that has found widespread applications in network information theory. It has been widely used to establish the capacity region in various multi-user information theory settings, such as broadcast channels with additive white Gaussian noise [Ber73], Gaussian wire-tap channels [LYCH78], and Gaussian MIMO broadcast channels with private messages [WSS06]. In Section 2.3.1, we will elaborate how to use a submodularity argument to obtain EPI and its generalization.

EPI was originally postulated by Shannon [Sha48] in the following formulation:

Theorem 1.2.7 (Entropy power inequality [Sha48, Sta59])**.** *Suppose X and Y are independent $\mathbb{R}^d$-valued random variables. The entropy power of X is defined as*

$$\mathcal{N}(X) = \frac{1}{2\pi e} \exp\left(\frac{2}{d} h(X)\right),$$

where $h(X)$ is the differential entropy of X.

Assume the differential entropy of X, Y, and $X+Y$ exists. Then the Entropy

Power Inequality states that

$$\mathcal{N}(X) + \mathcal{N}(Y) \le \mathcal{N}(X+Y),$$

where equality holds if and only if X and Y are Gaussians with proportional covariance matrices.

Stam [Sta59] showed that the EPI is a consequence of

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

An equivalent dimension-independent form of the Entropy Power Inequality was formulated by Lieb [Lie78].

Theorem 1.2.8. *Suppose X and Y are independent $\mathbb{R}^n$-valued random variables. For any $\lambda \in [0,1]$, we have*

$$\inf_{X,Y:X\perp Y} h(\sqrt{\lambda}X + \sqrt{1-\lambda}Y) - \lambda h(X) - (1-\lambda)h(Y) \ge 0,$$
$$\sup_{X,Y:X\perp Y} J(\sqrt{\lambda}X + \sqrt{1-\lambda}Y) - \lambda J(X) - (1-\lambda)J(Y) \le 0,$$

where the equality holds if and only if X and Y are Gaussians with identical covariance matrices.

In other words, the functional

$$f(\mu_X, \mu_Y) : h(\sqrt{\lambda}X + \sqrt{1-\lambda}Y) - \lambda h(X) - (1-\lambda)h(Y),$$

where $X \sim \mu_X$ and $Y \sim \mu_Y$ are independent random variables and are minimized by Gaussians with identical covariance matrices.

Several generalizations of the EPI have been proposed. Notably, Artstein, Ball, Barthe and Naor [ABBN04] showed that the quantity $h(\frac{X_1+\cdots+X_n}{n})$ is monotone in n when $X_1, \ldots, X_n$ are i.i.d. random variables. In [MB07], Madiman and Barron extended Stam's inequality for Fisher information [Sta59], which is applicable in giving a new proof of the Artstein–Ball–Barthe–Naor inequality. Later, Courtade proposed an elementary proof of monotonicity of entropy power and Fisher information. The details of these inequalities will be elaborated in Section 2.3.1, since a new proof will be proposed in the following subsections.

Several other proofs for the EPI were discovered by Guo–Shamai–Verdu [GSV06] (via MMSE), Rioul [Rio11], and Courtade [Cou16b]. For a comprehensive understanding of EPI, a survey of different versions of entropy power inequalities (forward and reverse) for Shannon entropy and Rényi entropy is presented in [MMX17].

1.3 Connections to functional inequalities

In this section, we would like to outline the connections between entropic inequalities and functional analysis. By using Legendre duality of entropy, we will be able to build a simple yet powerful connection between functional inequalities and entropic inequalities. We start with a simple example of the Hölder inequality, then we will extend the result to forward and backward Brascamp–Lieb inequalities.

1.3.1 Entropic proof of Hölder inequality

In this subsection, we present an elegant proof of the Hölder inequality using Legendre duality of entropy. This approach not only simplifies the traditional proof but also provides a framework that extends to the forward Brascamp–Lieb inequality and its entropic form (Theorem 1.3.4).

We begin with the Legendre duality of entropy:

Lemma 1.3.1 (Legendre duality of entropy (cf. [CCE09])). *For any function $f : \mathcal{X} \to \mathbb{R}$ where $\mathcal{X}$ is a finite set:*

$$\log\left(\sum_x \exp f(x)\right) = \sup_{p_X}\{\mathrm{E}[f(X)] + H(X)\}.$$

Proof. Fix the function f. Define a distribution q_X by

$$q_X(x) = \frac{\exp f(x)}{\sum_{x'} \exp f(x')}.$$

Since q_X has full support on $\mathcal{X}$, the Kullback–Leibler divergence $D(p_X \| q_X)$ is well-defined for any distribution p_X. Expanding the divergence:

$$\begin{aligned}
D(p_X \| q_X) &= -H_{p_X}(X) - \sum_x p_X(x) \log q_X(x) \\
&= -H_{p_X}(X) - \sum_x p_X(x) \left(f(x) - \log \left(\sum_{x'} \exp f(x') \right) \right) \\
&= -H_{p_X}(X) - \mathrm{E}_{p_X}[f(X)] + \log \left(\sum_x \exp f(x) \right).
\end{aligned}$$

By rearranging and using the non-negativity of KL divergence, we have

$$\log \left(\sum_x \exp f(x) \right) \geq \mathrm{E}_{p_X}[f(X)] + H_{p_X}(X).$$

Taking the supremum over all distributions p_X on the right-hand side preserves the inequality, and the equality is achieved when $p_X = q_X$, as this makes $D(p_X \| q_X) = 0$. Thus, the original equality holds. □

We now present an alternative characterization of entropy through Legendre duality that complements our previous result.

Lemma 1.3.2 (Alternate Legendre duality of entropy (cf. [CCE09]))**.** *Let p_X be a probability distribution of a random variable X with finite support $\mathcal{X}$. Then the entropy of X can be expressed as:*

$$H(X) = \inf_f \left\{ \log \left(\sum_x \exp f(x) \right) - \mathrm{E}[f(X)] \right\},$$

where the infimum is taken over all functions f defined on the support of $\mathcal{X}$.

Proof. The argument follows a similar structure to the proof of Lemma 1.3.1. For a fixed distribution p_X, we can establish

$$H_{p_X}(X) \leq \log\left(\sum_x \exp f(x)\right) - \mathrm{E}_{p_X}[f(X)].$$

When we take the infimum over all functions f on the right-hand side, the inequality is preserved. Equality is attained when $f(x) = \log p_X(x)$. Therefore, the original equality holds. □

The Legendre duality allows us to directly derive the Hölder inequality without resorting to Young's inequality:

Theorem 1.3.3 (Hölder inequality)**.** *For* $p, q \in (1, \infty)$ *with* $1/p + 1/q = 1$ *and functions* $f, g : [1:n] \to \mathbb{R}$*:*

$$\sum_{i=1}^n |f(i)g(i)| \leq \left(\sum_{i=1}^n |f(i)|^p\right)^{1/p} \left(\sum_{i=1}^n |g(i)|^q\right)^{1/q}.$$

Proof. Let $\hat{f}(i) = \log|f(i)|$ and $\hat{g}(i) = \log|g(i)|$. We need to show:

$$\log\left(\sum_{i=1}^n \exp(\hat{f}(i) + \hat{g}(i))\right) \leq \frac{1}{p}\log\left(\sum_{i=1}^n \exp(p\hat{f}(i))\right) + \frac{1}{q}\log\left(\sum_{i=1}^n \exp(q\hat{g}(i))\right).$$

Applying the Legendre duality, this inequality transforms into:

$$\begin{aligned}
&\sup_{r_X}\{\mathrm{E}[(\hat{f}+\hat{g})(X)] + H(X)\} \\
&\leq \frac{1}{p}\sup_{r_X}\{\mathrm{E}[p\hat{f}(X)] + H(X)\} + \frac{1}{q}\sup_{r_X}\{\mathrm{E}[q\hat{g}(X)] + H(X)\} \\
&= \sup_{r_X}\left\{\mathrm{E}[\hat{f}(X)] + \frac{1}{p}H(X)\right\} + \sup_{r_X}\left\{\mathrm{E}[\hat{g}(X)] + \frac{1}{q}H(X)\right\},
\end{aligned}$$

This inequality holds because $1/p + 1/q = 1$, completing an entropic proof of the Hölder inequality. □

1.3.2 Forward Brascamp–Lieb inequality

Brascamp–Lieb inequalities represent a powerful class of functional inequalities that unify and generalize many fundamental results in functional analysis. These include Hölder's inequality, the Loomis–Whitney inequality, the Prékopa–Leindler inequality, and sharp forms of Young's convolution inequalities [BCCT08]. Gardner's survey [Gar02] provides an excellent overview of the relationships between these inequalities.

Let us begin by presenting the functional form of the Brascamp–Lieb inequalities:

Theorem 1.3.4. *For $i \in [1:m]$, let E and E_i be Euclidean spaces, $A_i : E \to E_i$ be linear maps, c_i be positive real numbers, and f_i be non-negative integrable functions on E_i. Define the function $\mathcal{F}$ as*

$$\mathcal{F}(f_1, \dots, f_m) := \frac{\int_E \prod_{i=1}^m f_i^{c_i}(A_i x) dx}{\prod_{i=1}^m \left(\int_{E_i} f_i(x_i) dx_i \right)^{c_i}}.$$

The supremum of $\mathcal{F}$ over all non-negative and integrable functions f_i equals the supremum when restricted to centered Gaussian functions of the form $f_i(x_i) \propto \exp(-x_i^T B_i x_i)$, where each B_i is a positive semi-definite matrix.

The entropic formulation of this inequality, established by Carlen and Cordero–Erausquin [CCE09] using Legendre duality, provides an elegant information-theoretic perspective:

Theorem 1.3.5 (Theorem 2.1 of [CCE09])**.** *For $i \in [1:m]$, let E, E_i, A_i and c_i be as in Theorem 1.3.4. For a random variable X on E with well-defined differential entropy and finite second moment, define:*

$$f(X) := h(X) - \sum_{i=1}^m c_i h(A_i X).$$

The supremum of f over all qualifying random variables equals the supremum over Gaussian random variables.

This entropic formulation's proof relies on the superadditivity of Fisher information combined with heat-flow techniques, which is a sophisticated approach that highlights the deep connection between information theory and functional analysis.

A significant advancement came in [AJN22], where the authors unified the Brascamp–Lieb inequalities with the Entropy Power Inequality:

Definition 1.3.6 (BL datum)**.** For an integer $m > 0$, define an m-transformation as a triple $\mathbf{A} := (n, \{n_j\}_{j\in[1:m]}, \{A_j\}_{j\in[1:m]})$, where $n > 0$ is an integer, and for each $j \in [1 : m]$, $A_j : \mathbb{R}^n \to \mathbb{R}^{n_j}$ is a surjective linear transformation with $n_j \geq 0$.

An m-exponent is defined as an m-tuple $\mathbf{c} = \{c_j\}_{j\in[1:m]}$, where $c_j \geq 0$ for all $j \in [1 : m]$.

A Brascamp-Lieb datum (BL datum) is defined as a pair $(\mathbf{A}, \mathbf{c})$ where $\mathbf{A}$ is an m-transformation and $\mathbf{c}$ is an m-exponent.

Definition 1.3.7 (EPI datum)**.** For an integer $k > 0$, define a k-partition of n as $\mathbf{r} = \{r_i\}_{i\in[1:k]}$, where $r_i > 0$ are integers satisfying $\sum_{i\in[1:k]} r_i = n$.

A k-exponent is a tuple $\mathbf{d} = \{d_i\}_{i\in[1:k]}$ such that $d_i \geq 0$ for all $i \in [1 : k]$.

An Entropy Power Inequality datum (EPI datum) is a pair $(\mathbf{r}, \mathbf{d})$ where $\mathbf{r}$ is a k-partition and $\mathbf{d}$ is a k-exponent.

Definition 1.3.8 (BL-EPI datum)**.** For an integer $n > 0$, a BL-EPI datum is defined as $(\mathbf{A}, \mathbf{c}, \mathbf{r}, \mathbf{d})$ where $(\mathbf{A}, \mathbf{c})$ is a BL datum with some $m > 0$, and $(\mathbf{r}, \mathbf{d})$ is an EPI datum with some $k > 0$.

Definition 1.3.9 (Sets of Random Vectors)**.** Let $(\mathbf{A}, \mathbf{c}, \mathbf{r}, \mathbf{d})$ be a BL-EPI datum where $\mathbf{r}$ is a k-partition of n. Define $\mathcal{P}(\mathbf{r})$ to be the set of all $\mathbb{R}^n$-valued random vectors $X := (X_1, X_2, \ldots, X_k)$ such that:

1. For each $i \in [1 : k]$, the random vector X_i takes values in $\mathbb{R}^{r_i}$ and its density belongs to the convex set of probability densities $\{f : \int_{\mathbb{R}^{r_i}} f(x) \log(1 + f(x))\, dx < +\infty\}$;

2. The random vectors $X_1, X_2, \ldots, X_k$ are mutually independent;

3. $\mathbb{E}[X] = 0$ and $\mathbb{E}[\|X\|_2^2] < \infty$.

Furthermore, define $\mathcal{P}_g(\mathbf{r}) \subseteq \mathcal{P}(\mathbf{r})$ as the subset consisting of random vectors in $\mathcal{P}(\mathbf{r})$ where each component X_i follows a Gaussian distribution.

Theorem 1.3.10 (Unified EPI and BLI, [AJN22]). *Let $(\mathbf{A}, \mathbf{c}, \mathbf{r}, \mathbf{d})$ be a BL-EPI datum. Define:*

$$M_g := \sup_{Z \in \mathcal{P}_g(\mathbf{r})} \sum_{i=1}^{k} d_i h(Z_i) - \sum_{j=1}^{m} c_j h(A_j Z).$$

Then for any $X \in \mathcal{P}(\mathbf{r})$, the following inequality holds:

$$\sum_{i=1}^{k} d_i h(X_i) - \sum_{j=1}^{m} c_j h(A_j X) \leq M_g.$$

The authors of [AJN22] demonstrate that with appropriate $(\mathbf{A}, \mathbf{c}, \mathbf{r}, \mathbf{d})$ datum, Theorem 1.3.10 encompasses the standard EPI, BLI, and Zamir–Feder's EPI, establishing Gaussian optimality through an elegant rotation technique. This unification suggests rich possibilities for further exploration—particularly through different coupling structures between random variables, as the independent coupling assumption currently limits the scope of entropy inequalities and their connections to functional analysis.

1.4 Connections to additive combinatorics

In this section, we explore the rich interconnections between additive combinatorics and analogous entropic inequalities. Our discussion unfolds in three parts:

First, we present a simple yet powerful observation highlighting the natural parallelism between entropic inequalities and sumset inequalities. These parallels suggest fundamental mathematical structures common to both fields.

Second, we trace the historical development of the relationship between additive combinatorics and entropic inequalities. This overview provides essential context for understanding how these seemingly distinct mathematical areas have converged over time.

Third, we introduce the Ruzsa equivalence theorem that formally bridges a portion of sumset and entropic inequalities. This theorem reveals deeper insights into the intrinsic relationship between these mathematical frameworks, demonstrating how results in one domain can translate to meaningful discoveries in the other.

1.4.1 Parallelism between sumset inequalities and entropic inequalities

Several researchers have observed striking parallels between sumset inequalities in additive combinatorics and certain entropic inequalities. To explore these connections, we begin with the fundamental concept of sumsets:

Definition 1.4.1 (Sumset)**.** Let A and B be finite sets on a group $(\mathbb{G}, +)$. The sumset is defined as $A + B := \{a + b : a \in A, b \in B\}$. Similarly, we have $A - B := \{a - b : a \in A, b \in B\}$ and $k \cdot A := \{\sum_{i=1}^{k} a_i : a_i \in A\}$, where k is a positive integer.

A compelling illustration of this parallelism appears in Ruzsa sum-difference inequality, which provides a powerful bound for sumset cardinality:

Theorem 1.4.2 ([Ruz96])**.** *Let A and B be finite subsets on an Abelian additive group $(\mathbb{G}, +)$. We have $|A||B||A + B| \leq |A - B|^3$.*

Tao later established an entropic counterpart that mirrors this inequality's structure:

Theorem 1.4.3 ([Tao10])**.** *Let $(\mathbb{G}, +)$ be an Abelian group, and let X and Y be independent random variables with finite support on $\mathbb{G}$. We have $H(X) + H(Y) + H(X + Y) \leq 3H(X - Y)$.*

These inequalities share clear structural similarities, prompting the question of whether deeper connections exist between them. Interestingly, despite their formal resemblance, no direct implication has been established.

The relationship becomes even more nuanced when considering submodular properties. Consider the following entropic inequality:

Theorem 1.4.4 ([Mad08]). *Let $(\mathbb{G},+)$ be an Abelian group, and let X, Y and Z be independent random variables with finite support on $\mathbb{G}$. We have $H(X) + H(X+Y+Z) \leq H(X+Y) + H(X+Z)$.*

Proof. Since X, Y, Z are independent random variables, we can apply the data-processing inequality to get $I(Z; X+Y+Z) \leq I(Z; X+Z)$, which is equivalent to

$$H(X) + H(X+Y+Z) \leq H(X+Y) + H(X+Z).$$

□

One might expect a direct sumset analogue of the form $|A||A+B+C| \leq |A+B||A+C|$ for sets A, B, C in $(\mathbb{G},+)$. However, counterexamples demonstrate this is not true. Instead, the corresponding sumset inequality takes a more conditional form:

Proposition 1.4.5 (Proposition 2.1 of [Pet12]). *Let A and B be finite sets in a group $(\mathbb{G},+)$. For any subset $S \subseteq A$ satisfying*

$$\frac{|S+B|}{|S|} \leq \frac{|T+B|}{|B|} \qquad \textit{for all } T \subseteq S,$$

and for all finite sets $C \subseteq \mathbb{G}$,

$$|C+S+B| \leq \frac{|C+S||S+B|}{|S|}.$$

This discrepancy reveals that submodular properties manifest differently in sumset and entropic contexts. The relationship between these domains is more subtle than simple translation of results from one to the other. These observations motivate a deeper investigation into the connections between sumset and entropic inequalities, particularly to identify which classes of inequalities in one domain have meaningful correspondences in the other.

1.4.2 Historical remark

Ruzsa provided a useful categorization of sumset inequalities in relation to entropic inequalities [Ruz09a], identifying three distinct scenarios:

a): There exists an equivalence form (see Theorem 1.4.7) and explicit implication between a combinatorial inequality and an associated entropic inequality.

b): A structural analog exists between combinatorial and entropic inequalities, but no direct equivalence is known. Sometimes, one-directional implication can be established.

c): There is a combinatorial/entropic inequality, but the correctness of the counterpart (analogous) inequality is unknown.

Most subsequent research has focused on the second scenario, which is developing analogous entropic inequalities without establishing formal equivalence. Tao's work [Tao10] made significant progress by establishing entropic analogs of the Plünnecke–Ruzsa–Frieman sumset and inverse sumset theory. In 2012, Madiman, Marcus, and Tetali [MMT12] developed both entropic analogs and equivalence theorems based on partition-determined functions of random variables.

Further expanding this connection, Kontoyiannis and Madiman explored the relationship between sumsets and differential entropies [KM14]. For readers interested in deeper explorations of these connections, we refer readers to the following works [Mad08], [LP08], and [MK10]. A comprehensive summary of the connections between combinatorial and entropic inequalities can also be found in [ED16].

Additionally, [TV] established a systematic analogous relationship between notation in sumset theory and information theory.

1.4.3 Ruzsa equivalence theorem on sumset inequalities

In [Ruz09a], Ruzsa made the first significant attempt to establish an equivalence relationship between sumset inequalities and entropic inequalities. Rather than focusing on conventional sumset notation, his equivalence theorem addresses inequalities involving G-restricted sumsets, defined as follows:

Definition 1.4.6. (G-restricted sumset [Ruz09a]) Suppose G is a subset of $A \times B$, where A, B are finite subsets of $(\mathbb{G}, +)$. We denote the G-restricted sumset and difference set of A and B as $A \overset{G}{+} B$ and $A \overset{G}{-} B$.

$$A \overset{G}{+} B = \{a + b : a \in A, b \in B, (a, b) \in G\},$$
$$A \overset{G}{-} B = \{a - b : a \in A, b \in B, (a, b) \in G\}.$$

Using a typicality-based argument, Ruzsa established the following equivalence theorem:

Theorem 1.4.7 (Ruzsa equivalence theorem, Equivalence Theorem 2 of [Ruz09a])**.** *Let $f, g_1, \ldots, g_k$ be linear functions in two variables with integer coefficients, and let $\alpha_1, \ldots, \alpha_k$ be positive real numbers. Let $(\mathbb{T}, +)$ be a finitely generated and torsion-free group. The following statements are equivalent:*

1. *For every finite $A \subseteq \mathbb{T} \times \mathbb{T}$ we have*

$$|f(A)| \leq \prod |g_i(A)|^{\alpha_i},$$

where $|f(A)|$ denotes the cardinality of the image $f(A)$.

2. *For every pair X, Y of (not necessarily independent) random variables with values in $(\mathbb{T}, +)$ such that the entropy of each $g_i(X, Y)$ is finite, the entropy*

of $f(X,Y)$ is also finite and it satisfies

$$H(f(X,Y)) \leq \sum \alpha_i H(g_i(X,Y)).$$

This equivalence theorem immediately yields several non-trivial entropic inequalities, such as the entropic formulation of Katz-Tao sumset inequalities (Theorem 4.2.14). However, the applicability of this theorem is limited because most sumset inequalities in additive combinatorics don't apply in the graph-restricted form and don't require the underlying group structure to be a finitely generated torsion-free group. Consequently, to further extend the equivalence relationships between entropic and sumset inequalities, new equivalence theorems are needed.

1.5 Structure of the thesis

In Chapter 2, we establish a family of supermodularity inequalities for mutual information involving auxiliary random variables and independent random variables. These inequalities arise naturally from "compression" operations. This can be viewed as a generalization of Shearer's Lemma discussed in Section 1.2.1. By introducing suitable auxiliary random variables and exploiting their structural properties, we extend these supermodularity results to various information measures, including Fisher information, Kullback–Leibler divergence, strong data processing inequality constants, and Hirschfeld–Gebelein–Rényi maximal correlation. A key contribution is our submodularity-based proof of the generalized Stam's inequality, which, through a convex duality framework, leads to the fractional Entropy Power Inequality, which is the most general version of the EPI known to date.

In Chapter 3, we develop a framework for establishing the optimality of uniform distributions in discrete entropic functional optimization problems. Using a discrete "rotation"-trick and superadditivity, we construct machinery directly inspired by the approach used to establish Gaussian optimality in differential entropy optimization problems through continuous "rotation"-trick and subadditivity, as described in Section 1.2.2. With this framework, we propose a discrete analogue to Theorem 1.3.10 applicable to any finite Abelian group, significantly generalizing the Entropy Power Inequality for discrete-valued random variables. Furthermore, we demonstrate how this technique proves the optimality of uniform distributions in an optimization problem related to the polynomial Freiman-Ruzsa conjecture, which is a longstanding open problem in additive combinatorics. This highlights the potential breadth of our approach.

In Chapter 4, we establish a generalized equivalence theorem connecting sumset inequalities and entropic inequalities that extends beyond the G-sumset inequality framework of the Ruzsa equivalence theorem (Theorem 1.4.7). We introduce a powerful information-theoretic result (Lemma 4.2.11), inspired by its combinatorial counterpart (Lemma 4.2.9), which enables proofs of several nontrivial entropic inequalities related to sumset theory. Additionally, we provide an entropic formulation of the magnification ratio, which is a central concept in Plünnecke-Ruzsa sumset theory, laying the groundwork for deeper connections between sumset theory and information theory.

Chapter 2

Supermodular Information Inequalities and their Applications

This chapter demonstrates how supermodularity inequalities enable both streamlined proofs of established results and novel generalizations in discrete convexity analysis, particularly concerning strong data processing constants, maximal correlation, and Kullback-Leibler divergence.

In Section 2.1, we systematically extend supermodularity principles from fundamental two-point inequalities to comprehensive informational inequality families. Our approach originates from a straightforward supermodular relationship involving auxiliary random variables and independent pairs, subsequently generalized through iterative applications of the compression framework introduced in [BB12].

Section 2.2 introduces two families of perturbative auxiliary variables crucial for estimating conditional expectations and KL divergence. These constructs significantly expand the operational scope of supermodular inequalities while enabling diverse corollary applications.

In Section 2.3.1, we combine these two ideas to present a novel proof of a generalized Stam's inequality with fractional partitions, originally established in [MG19] (Theorem 1). Unlike the two-variable case, the original proof linking this inequality to the fractional EPI superadditivity required a substantial technical effort. Through convex duality principles, we develop a streamlined argument that not only provides a new demonstration of this relationship but also reveals structural parallels with the simpler two-variable scenario.

Section 2.3.2 focuses on independent identically distributed systems to derive discrete convexity properties, generalizing key results about information-theoretic constants. Finally, Section 2.4 establishes foundational connections between submodular sumset inequalities and their entropic counterparts.

2.1 Preliminaries

To quantify the supermodular behavior of information measures, we introduce a partial ordering of fractional multisets through the concept of compression. These fractional multisets govern the coefficients in mutual information linear combinations, with the compression partial order inducing entropic inequalities that capture supermodularity properties.

Definition 2.1.1 (Fractional multiset)**.** Let n be a positive integer. An *n-fractional multiset* $\{\alpha_T\}_T$ is a finite sequence of non-negative real numbers α_T indexed by $T \subseteq [1:n]$.

Remark 2.1.2. The notion of n-fractional multisets is not new and has been used in [BB12] where the authors call n-fractional multisets to be "multisets of subsets of $[n]$". On the other hand, we view an n-fractional multiset as the finite sequence of its, potentially fractional, multiplicities.

Definition 2.1.3 (Elementary compression & Compression)**.** Let n be a positive integer and let $\{\alpha_T\}_T, \{\beta_T\}_T$ be two n-fractional multisets. We call $\{\beta_T\}_T$ an *elementary compression* of $\{\alpha_T\}_T$ if there exist $A, B \subseteq [1:n]$ with $A \not\subseteq B$ and

$B \not\subseteq A$, and $0 < \delta \leq \min\{\alpha_A, \alpha_B\}$ such that for all $T \subseteq [1:n]$ we have

$$\beta_T = \begin{cases} \alpha_T - \delta & \text{if } T = A \text{ or } T = B, \\ \alpha_T + \delta & \text{if } T = A \cup B \text{ or } T = A \cap B, \\ \alpha_T & \text{otherwise.} \end{cases}$$

The result of a finite sequence of elementary compressions of $\{\alpha_T\}_T$ is called a *compression* of $\{\alpha_T\}_T$.

Remark 2.1.4. As studied in [BB12], the relation "is a compression of" defines a partial order on the collection of n-fractional multisets. It is immediate that an n-fractional multiset $\{\beta_T\}_T$ is minimal under this partial order (i.e. cannot be further compressed) if and only if the set $\{T \subseteq [1:n] : \beta_T \neq 0\}$ is totally ordered under set inclusion.

The following lemma establishes a family of supermodularity inequalities for mutual information. The derivation originates from a fundamental two-point inequality corresponding to elementary compression, which we subsequently generalize into a family of inequalities governed by partial ordering under compression.

Lemma 2.1.5. *Let $X_1, \ldots, X_n$ be random variables that are mutually independent conditioned on a random variable $S_\emptyset$, and let U be any auxiliary random variable. Then the following hold:*

(i) $I(U; S_\emptyset, X_A) + I(U; S_\emptyset, X_B) \leq I(U; S_\emptyset, X_{A\cup B}) + I(U; S_\emptyset, X_{A\cap B})$ *for all* $A, B \subseteq [1:n]$.

(ii) $\sum_{T\subseteq[1:n]} \alpha_T I(U; S_\emptyset, X_T) \leq \sum_{T\subseteq[1:n]} \beta_T I(U; S_\emptyset, X_T)$, *for any n-fractional multisets $\{\alpha_T\}, \{\beta_T\}$ such that $\{\beta_T\}$ is a compression of $\{\alpha_T\}$.*

(iii) $\sum_{T\subseteq[1:n]} \beta_T I(U; S_\emptyset, X_T) \leq I(U; S_\emptyset, X_{[1:n]}) + (c-1) I(U; S_\emptyset)$, *where $\{\beta_T\}$ is an n-fractional multiset satisfying $\sum_{T\subseteq[1:n]:T\ni i} \beta_T \leq 1$ for all $i = 1, \ldots, n$, and $c := \sum_{T\subseteq[1:n]} \beta_T$.*

Proof. Suppose $A, B \subseteq [1:n]$. Then

$$\begin{aligned}
& I(U;S_\emptyset,X_B) - I(U;S_\emptyset,X_{A\cap B}) \\
&= I(U;X_{B\setminus A}|S_\emptyset,X_{A\cap B}) \\
&\le I(U,X_{A\setminus B};X_{B\setminus A}|S_\emptyset,X_{A\cap B}) \\
&\overset{(a)}{=} I(U,X_{A\setminus B};X_{B\setminus A}|S_\emptyset,X_{A\cap B}) \\
&\quad - I(X_{A\setminus B};X_{B\setminus A}|S_\emptyset,X_{A\cap B}) \\
&= I(U;X_{B\setminus A}|S_\emptyset,X_A) \\
&= I(U;S_\emptyset,X_{A\cup B}) - I(U;S_\emptyset,X_A).
\end{aligned}$$

where (a) holds by the mutual independence of the X_i's conditioned on $S_\emptyset$. Rearranging gives

$$I(U;S_\emptyset,X_A) + I(U;S_\emptyset,X_B) \le I(U;S_\emptyset,X_{A\cup B}) + I(U;S_\emptyset,X_{A\cap B}).$$

which is (i).

If $\{\beta_T\}$ is an elementary compression of $\{\alpha_T\}$, then the inequality in (ii) follows from (i) by canceling like terms on both sides. Since a compression is obtained as a sequence of elementary compressions, (ii) follows.

We will show (iii) by induction on n. Indeed the base case $n = 1$ is trivial. Note that (i) gives

$$I(U;S_\emptyset,X_{[1:n-1]}) + I(U;S_\emptyset,X_{T\cup\{n\}}) \le I(U;S_\emptyset,X_{[1:n]}) + I(U;S_\emptyset,X_T)$$

for all $T \subseteq [1:n-1]$. Suppose β_T ($T \subseteq [1:n]$) are non-negative real numbers satisfying $\sum_{T\subseteq[1:n]:T\ni i} \beta_T \le 1$ for all $i = 1,\ldots,n$. Then

$$\begin{aligned}
& \sum_{T\subseteq[1:n]} \beta_T I(U;S_\emptyset,X_T) \\
&= \sum_{T\subseteq[1:n-1]} \big(\beta_T I(U;S_\emptyset,X_T) + \beta_{T\cup\{n\}} I(U;S_\emptyset,X_{T\cup\{n\}})\big)
\end{aligned}$$

$$\begin{aligned}
&\leq \sum_{T\subseteq[1:n-1]} \Big(\beta_T I(U;S_\emptyset,X_T) + \beta_{T\cup\{n\}}(I(U;S_\emptyset,X_{[1:n]}) \\
&\quad - I(U;S_\emptyset,X_{[1:n-1]}) + I(U;S_\emptyset,X_T))\Big) \\
&\overset{(a)}{\leq} I(U;S_\emptyset,X_{[1:n]}) - I(U;S_\emptyset,X_{[1:n-1]}) + \sum_{T\subseteq[1:n-1]} (\beta_T + \beta_{T\cup\{n\}}) I(U;S_\emptyset,X_T) \\
&\overset{(b)}{\leq} I(U;S_\emptyset,X_{[1:n]}) - I(U;S_\emptyset,X_{[1:n-1]}) + I(U;S_\emptyset,X_{[1:n-1]}) + (c-1)I(U;S_\emptyset) \\
&= I(U;S_\emptyset,X_{[1:n]}) + (c-1)I(U;S_\emptyset).
\end{aligned}$$

where (a) holds since $\sum_{T\subseteq[1:n-1]}\beta_{T\cup\{n\}} \leq 1$, and (b) follows by applying the induction hypothesis to the non-negative real numbers $\big\{\beta_T + \beta_{T\cup\{n\}}\big\}_{T\subseteq[1:n-1]}$. □

We now introduce the notion of layered function family to extend the supermodularity results from random vectors to functions of random vectors.

Definition 2.1.6. Let X_i $(i = 1,\ldots,n)$ and S_T $(T \subseteq [1:n])$ be random variables. We call $\{S_T\}_T$ a *layered function family* on $X_1,\ldots,X_n$ if $S_\emptyset$ is independent of $X_{[1:n]}$, and for every non-empty $T \subseteq [1:n]$ and $i \in T$ there is a function $g_{T,i}$ such that $S_T = g_{T,i}(S_{T\setminus\{i\}}, X_i)$.

Remark 2.1.7. Clearly a trivial example of a layered function family is given by $S_T := (S_\emptyset, X_T)$. A canonical example of a layered function family is given by $S_T := S_\emptyset + \sum_{i\in T} f_i(X_i)$, where f_i's are functions taking values in some Abelian monoid (i.e. a set with a binary operation, which we denote by $+$, that is associative and commutative, and has an identity element). In particular,

(i) $S_T := S_\emptyset + \sum_{i\in T} X_i$, where $S_\emptyset, X_i \in \mathbb{R}^d$;

(ii) $S_T := \max(\{S_\emptyset\} \cup \{X_i\}_{i\in T})$, where $S_\emptyset, X_i \in \mathbb{R}$;

are examples of layered function families.

Remark 2.1.8. Layered function families play a similar role as that of *partition-determined functions* in [MMT12] and it may be possible that they are intrinsically trying to capture a similar behaviour and dependence structure. For our results, we prefer to stick with the definition of layered function families. Note that [MMT12] deals with dependent random variables while here our main focus is on mutually independent random variables.

The layered structure of function families guarantees consistent propagation of Markov structures across subset hierarchies.

Lemma 2.1.9. *Let $\{S_T\}_T$ be a layered function family on mutually independent random variables $X_1, \dots, X_n$. Suppose $U \to S_{[1:n]} \to (S_\emptyset, X_{[1:n]})$ forms a Markov chain. Then the following hold:*

(i) $U \to S_T \to (S_\emptyset, X_T)$ forms a Markov chain for all $T \subseteq [1:n]$.

(ii) $I(U; S_T) = I(U; S_\emptyset, X_T)$ for all $T \subseteq [1:n]$.

Proof. Suppose $T \subseteq [1:n]$. Consider

$$\begin{aligned}
0 &\overset{(a)}{=} I(U; S_\emptyset, X_{[1:n]} | S_{[1:n]}) \\
&= I(U; S_\emptyset, X_T, X_{[1:n]\setminus T} | S_{[1:n]}) \\
&\overset{(b)}{=} I(U; S_\emptyset, X_T, X_{[1:n]\setminus T}, S_T | S_{[1:n]}) \\
&\geq I(U; S_\emptyset, X_T | S_{[1:n]}, X_{[1:n]\setminus T}, S_T) \\
&\overset{(c)}{=} I(U; S_\emptyset, X_T | X_{[1:n]\setminus T}, S_T) \\
&\overset{(d)}{=} I(U; S_\emptyset, X_T | X_{[1:n]\setminus T}, S_T) + I(X_{[1:n]\setminus T}; S_\emptyset, X_T | S_T) \\
&= I(U, X_{[1:n]\setminus T}; S_\emptyset, X_T | S_T) \\
&\geq I(U; S_\emptyset, X_T | S_T) \\
&\geq 0.
\end{aligned}$$

where (a) holds since $U \to S_{[1:n]} \to (S_\emptyset, X_{[1:n]})$ forms a Markov chain, (b) holds since S_T is a function of $(S_\emptyset, X_T)$, (c) holds since $S_{[1:n]}$ is a function of $(S_T, X_{[1:n]\setminus T})$, and (d) holds since $X_{[1:n]\setminus T}$ and $(S_\emptyset, X_T, S_T)$ are independent. This shows (i). Furthermore,

$$I(U; S_T) \overset{(a)}{=} I(U; S_T, S_\emptyset, X_T) \overset{(b)}{=} I(U; S_\emptyset, X_T).$$

where (a) holds since $U \to S_T \to (S_\emptyset, X_T)$ forms a Markov chain, and (b) holds since S_T is a function of $(S_\emptyset, X_T)$. This shows (ii). □

We now state the main theorem in this chapter. The proof is an immediate application of Lemma 2.1.9 to Lemma 2.1.5.

Theorem 2.1.10. *Let $\{S_T\}_T$ be a layered function family on mutually independent random variables $X_1, \ldots, X_n$. Suppose $U \to S_{[1:n]} \to (S_\emptyset, X_{[1:n]})$ forms a Markov chain. Then the following hold:*

(i) $I(U; S_A) + I(U; S_B) \leq I(U; S_{A\cup B}) + I(U; S_{A\cap B})$ *for all* $A, B \subseteq [1:n]$.

(ii) $\sum_{T\subseteq[1:n]} \alpha_T I(U; S_T) \leq \sum_{T\subseteq[1:n]} \beta_T I(U; S_T)$, *for any n-fractional multisets $\{\alpha_T\}, \{\beta_T\}$ such that $\{\beta_T\}$ is a compression of $\{\alpha_T\}$.*

(iii) $\sum_{T\subseteq[1:n]} \beta_T I(U; S_T) \leq I(U; S_{[1:n]}) + (c-1)I(U; S_\emptyset)$, *where $\{\beta_T\}$ is an n-fractional multiset satisfying $\sum_{T\subseteq[1:n]:T\ni i} \beta_T \leq 1$ for all $i = 1, \ldots, n$, and $c := \sum_{T\subseteq[1:n]} \beta_T$.*

It turns out that the freedom in choosing the auxiliary random variable U plays a rather important role in the development of the inequalities.

2.2 Two families of perturbative auxiliaries

In this section, we introduce two auxiliary families that prove instrumental for deriving corollaries to Theorem 2.1.10. These families of auxiliary random variables interact with conditional expectations and KL divergence through perturbative methods, enabling extension of supermodularity properties to these quantities beyond mutual information.

Lemma 2.2.1. *Let $\{S_T\}_T$ be a layered function family on mutually independent random variables $X_1, \ldots, X_n$. Suppose f is an $\mathbb{R}^d$-valued bounded measurable function, defined on the set of values of $S_{[1:n]}$, such that $\mathrm{E}[f(S_{[1:n]})] = 0$. Then* *there exists a family of random variables $\{U^{(\epsilon)}\}_\epsilon$, indexed by small enough $\epsilon > 0$, such that $U^{(\epsilon)} \to S_{[1:n]} \to (S_\emptyset, X_{[1:n]})$ forms a Markov chain and*

$$I(U^{(\epsilon)}; S_T) = \frac{1}{2}\epsilon^2 \operatorname{E}[\| \operatorname{E}[f(S_{[1:n]})|S_T]\|^2] + O(\epsilon^3)$$

for all $T \subseteq [1:n]$.

Proof. Let $\tilde{p}(\cdot)$ be the probability mass function of the uniform distribution on the Boolean hypercube $\{\pm 1\}^d$. For small enough $\epsilon > 0$, define the random variable $U^{(\epsilon)}$ taking values in $\{\pm 1\}^d$, satisfying the Markov chain $U^{(\epsilon)} \to S_{[1:n]} \to (S_\emptyset, X_{[1:n]})$, according to

$$p_{U^{(\epsilon)}|S_{[1:n]}}(u|s) := \tilde{p}(u)(1 + \epsilon\langle f(s), u\rangle).$$

Note that $p_{U^{(\epsilon)}}(u) = \tilde{p}(u)$ (which follows from $\operatorname{E}[f(S_{[1:n]})] = 0$), $\operatorname{E}[U^{(\epsilon)}] = 0$, and $\operatorname{E}[U^{(\epsilon)}U^{(\epsilon)\intercal}] = I$. For any $T \subseteq [1:n]$, since $U^{(\epsilon)} \to S_{[1:n]} \to S_T$ forms a Markov chain,

$$p_{U^{(\epsilon)}|S_T}(u|S_T) = \operatorname{E}[p_{U^{(\epsilon)}|S_{[1:n]}}(u|S_{[1:n]})|S_T] = \tilde{p}(u)(1 + \epsilon\langle \operatorname{E}[f(S_{[1:n]})|S_T], u\rangle).$$

Then we compute:

$$\begin{aligned}
I(U^{(\epsilon)}; S_T) &= \operatorname{E}_{U^{(\epsilon)},S_T}\left[\log \frac{p(U^{(\epsilon)}|S_T)}{p(U^{(\epsilon)})}\right] \\
&= \operatorname{E}_{U^{(\epsilon)},S_T}\left[\log(1 + \epsilon\langle \operatorname{E}[f(S_{[1:n]})|S_T], U^{(\epsilon)}\rangle)\right] \\
&= \operatorname{E}_{S_T}\Big[\sum_u \tilde{p}(u)(1 + \epsilon\langle \operatorname{E}[f(S_{[1:n]})|S_T], u\rangle)\log(1 + \epsilon\langle \operatorname{E}[f(S_{[1:n]})|S_T], u\rangle)\Big] \\
&\overset{(a)}{=} \operatorname{E}_{S_T}\Big[\sum_u \tilde{p}(u)\big(\epsilon\langle \operatorname{E}[f(S_{[1:n]})|S_T], u\rangle + \frac{1}{2}\epsilon^2\langle \operatorname{E}[f(S_{[1:n]})|S_T], u\rangle^2 + O(\epsilon^3)\big)\Big].
\end{aligned}$$

The equality (a) is justified by Remark 2.2.2.

We now apply the linearity of expectation to obtain:

$$\begin{aligned} I(U^{(\epsilon)}; S_T) &= \frac{1}{2}\epsilon^2 \operatorname{tr}\left(\mathrm{E}[\mathrm{E}[f(S_{[1:n]})|S_T]\,\mathrm{E}[f(S_{[1:n]})|S_T]^\intercal] \sum_u \tilde{p}(u)uu^\intercal \right) + O(\epsilon^3) \\ &= \frac{1}{2}\epsilon^2\,\mathrm{E}[\|\,\mathrm{E}[f(S_{[1:n]})|S_T]\|^2] + O(\epsilon^3). \end{aligned}$$

This completes the proof. □

Remark 2.2.2. Note that one can show that

$$\left|(1+x)\log(1+x) - x - x^2/2\right| \le \frac{1}{3}|x|^3, \quad \text{for } x \in \left[-\frac{1}{2}, \frac{1}{2}\right].$$

Since f is a bounded, measurable function, $\langle f, u\rangle$ is also bounded for all unit vectors u, say by B. For any $0 < \epsilon < \frac{1}{2B}$, we have

$$\left|(1 + \epsilon\langle f, u\rangle)\log(1 + \epsilon\langle f, u\rangle) - \epsilon\langle f, u\rangle - \frac{1}{2}\epsilon^2\langle f, u\rangle^2\right| \le \frac{1}{3}(\epsilon B)^3.$$

Lemma 2.2.3. *Let $\{S_T\}_T$ be a layered function family on mutually independent random variables $X_1, \ldots, X_n$. Suppose $q(\cdot)$ is a distribution that is absolutely continuous and has a bounded Radon–Nikodym derivative with respect to the distribution of $S_{[1:n]}$. Then there exists a family of random variables $\{U^{(\epsilon)}\}_\epsilon$, indexed by small enough $\epsilon > 0$, such that $U^{(\epsilon)} \to S_{[1:n]} \to (S_\emptyset, X_{[1:n]})$ forms a Markov chain and*

$$I(U^{(\epsilon)}; S_T) = \epsilon D_{\mathrm{KL}}(p_{\tilde{S}_T} \| p_{S_T}) + O(\epsilon^2)$$

for all $T \subseteq [1:n]$, where the random variable $\tilde{S}_T$ is defined by

$$p_{\tilde{S}_T}(\tilde{s}) := \sum_s p_{S_T|S_{[1:n]}}(\tilde{s}|s)q(s).$$

Proof. Let $f(s) := q(s)/p_{S_{[1:n]}}(s)$ be the Radon–Nikodym derivative. For small enough $\epsilon > 0$, define the random variable $U^{(\epsilon)}$ taking values in $\{0, 1\}$, satisfying the Markov chain $U^{(\epsilon)} \to S_{[1:n]} \to (S_\emptyset, X_{[1:n]})$, according to

$$p_{U^{(\epsilon)}|S_{[1:n]}}(u|s) := \begin{cases} 1 - \epsilon f(s) & \text{if } u = 0, \\ \epsilon f(s) & \text{if } u = 1. \end{cases}$$

Note that $\mathrm{E}[f(S_{[1:n]})] = 1$ and

$$p_{U^{(\epsilon)}}(u) = \begin{cases} 1 - \epsilon & \text{if } u = 0, \\ \epsilon & \text{if } u = 1. \end{cases}$$

For any $T \subseteq [1:n]$, since $U^{(\epsilon)} \to S_{[1:n]} \to S_T$ forms a Markov chain,

$$p_{U^{(\epsilon)}|S_T}(u|S_T) = \mathrm{E}[p_{U^{(\epsilon)}|S_{[1:n]}}(u|S_{[1:n]})|S_T] = \begin{cases} 1 - \epsilon\,\mathrm{E}[f(S_{[1:n]})|S_T] & \text{if } u = 0, \\ \epsilon\,\mathrm{E}[f(S_{[1:n]})|S_T] & \text{if } u = 1. \end{cases}$$

Then we have

$$\begin{aligned} I(U^{(\epsilon)}; S_T) &= \mathrm{E}_{U^{(\epsilon)},S_T}\left[\log \frac{p(U^{(\epsilon)}|S_T)}{p(U^{(\epsilon)})}\right] \\ &= \mathrm{E}_{S_T}\Big[\epsilon\,\mathrm{E}[f(S_{[1:n]})|S_T] \log \mathrm{E}[f(S_{[1:n]})|S_T] \\ &\quad + (1 - \epsilon\,\mathrm{E}[f(S_{[1:n]})|S_T]) \log \frac{1 - \epsilon\,\mathrm{E}[f(S_{[1:n]})|S_T]}{1 - \epsilon}\Big] \\ &= \epsilon\,\mathrm{E}_{S_T}\left[\frac{p_{\tilde{S}_T}(S_T)}{p_{S_T}(S_T)} \log \frac{p_{\tilde{S}_T}(S_T)}{p_{S_T}(S_T)}\right] \\ &\quad + \mathrm{E}_{S_T}\left[(1 - \epsilon\,\mathrm{E}[f(S_{[1:n]})|S_T])(\epsilon(1 - \mathrm{E}[f(S_{[1:n]})|S_T]) + O(\epsilon^2))\right] \\ &= \epsilon D_{\mathrm{KL}}(p_{\tilde{S}_T} \| p_{S_T}) + O(\epsilon^2). \end{aligned}$$

Using an approach similar to that presented in Lemma 2.2.1, we can justify the $O(\epsilon^2)$ term. □

Remark 2.2.4. These two families of perturbative auxiliaries are not new here and have been used extensively in [AGKN13, AGKN14] and references therein.

2.3 Some consequences of the supermodularity inequalities

In this section we will outline some existing results, extensions of existing results, as well as the new ones that we obtain as consequences of Theorem 2.1.10.

2.3.1 Entropy power inequalities and Fisher information inequalities

Historical remark

Lieb's form of the EPI (Theorem 1.2.8) implies that (by taking $\lambda = 1/2$)

$$h\left(\frac{X+Y}{\sqrt{2}}\right) \geq \frac{1}{2}\left(h(X)+h(Y)\right).$$

Building on this result, Lieb [Lie78] conjectured that for any sequence $X_1, \dots, X_n$ of independent and identically distributed real-valued random variables, the entropy functional $h\left(\frac{X_1+\cdots+X_n}{\sqrt{n}}\right)$ exhibits monotonic non-decreasing behavior in n.

This conjecture was resolved by Artstein–Ball–Barthe–Naor [ABBN04] who showed the following inequality: If $a_1, \dots, a_{n+1} \geq 0$ satisfies $\sum_{i=1}^{n+1} a_i^2 = 1$ then

$$h\left(\sum_{i=1}^{n+1} a_i X_i\right) \geq \sum_{i=1}^{n+1} \frac{1-a_i^2}{n} h\left(\frac{1}{\sqrt{1-a_i^2}} \sum_{\substack{j=1 \\ j\neq i}}^{n+1} a_j X_j\right).$$

and in particular,

$$h\left(\frac{1}{\sqrt{n+1}} \sum_{i=1}^{n+1} X_i\right) \geq \frac{1}{n+1} \sum_{i=1}^{n+1} h\left(\frac{1}{\sqrt{n}} \sum_{\substack{j=1 \\ j\neq i}}^{n+1} X_j\right).$$

Their proof was simplified and extended in a series of works, e.g. Madiman–Barron [MB07] and Madiman–Ghassemi [MG19]. The best known version (see Theorem

1 in [MG19]) is the *fractional partition* form of the EPI:

$$\exp\left(\frac{2}{d}h\left(\sum_{i=1}^{n} X_i\right)\right) \geq \sum_{\substack{T\subseteq[1:n]\\ T\neq\emptyset}} \beta_T \exp\left(\frac{2}{d}h\left(\sum_{i\in T} X_i\right)\right).$$

for any mutually independent random variables $X_1 \ldots, X_n$ in $\mathbb{R}^d$ with differentiable densities, and *fractional partition* $\{\beta_T\}_T$, i.e. a finite collection indexed by $T \subseteq [1:n]$, $T \neq \emptyset$, of non-negative real numbers satisfying $\sum_{T\subseteq[1:n]:T\ni i} \beta_T = 1$ for every $i \in [1:n]$. This was derived as a consequence of the following Fisher information inequality, that we shall refer to as the *generalized Stam's inequality*:

$$\frac{1}{J(S_{[1:n]})} \geq \sum_{T\subseteq[1:n]} \beta_T \frac{1}{J(S_T)}.$$

where $S_T := \sum_{i\in T} X_i$.

Remark 2.3.1. Unlike the $n = 2$ setting, the implication that the generalized Stam's inequality implies the fractional partition form of the EPI did not have a straightforward proof. In this subsection, we use convex duality to show a straightforward proof of this implication.

Alternate proof of generalized Stam's inequality

In this subsection, we derive the generalized Stam's inequality involving Fisher information as an immediate consequence of our mutual information inequality. While a similar proof technique that we employ has been used by Courtade in [Cou16a] for the case of mutually independent and identically distributed random variables, as noted in [Joh20] (future work, item 4), the extension of the ideas to independent random variables is of independent interest.

Remark 2.3.2. To avoid technical issues, we will deal with random variables X with density function f_X that is smooth and rapidly decaying such that $|\log f_X|$ has at most polynomial growth at infinity.

Definition 2.3.3 (Score function)**.** Let X be a random variable in $\mathbb{R}^d$ with differentiable density f_X with respect to the Lebesgue measure. Assume that f_X is differentiable almost everywhere and that $f_X(x) > 0$ for all x in the support of X. The *score function* ρ_X of X is defined by

$$\rho_X := \frac{\nabla f_X}{f_X} = \nabla \log f_X.$$

for all x in the support of X where f_X is differentiable.

The *Fisher information* $J(X)$ of X is defined by

$$J(X) := \mathrm{E}[\|\rho_X(X)\|^2].$$

Remark 2.3.4. Let X, Z be independent random variables in $\mathbb{R}^d$ such that $Z \sim \mathcal{N}(0, I)$. We have the following basic properties of Fisher information:

(i) $J(aX) = a^{-2}J(X)$ for all $a > 0$.

(ii) If X has a finite second moment, then $\frac{1}{2}J(X + \sqrt{t}Z) = \frac{\partial}{\partial t}h(X + \sqrt{t}Z)$ for all $t \geq 0$.

(iii) If X has a (finite) covariance matrix then

$$h(X) = \frac{d}{2}\log 2\pi e - \frac{1}{2}\int_0^\infty \left(J(X + \sqrt{t}Z) - \frac{d}{1+t}\right) dt.$$

Property (ii) is also called de Bruijn's identity (e.g. [Sta59]). Property (iii) is a consequence of (ii) and is originally shown by Barron [Bar86] (cf. Lemma 3 of [MB07]).

Our proof employs the following theorem.

Theorem 2.3.5 (Stam [Sta59])**.** *Suppose $X_1, \ldots, X_n$ are mutually independent random variables in $\mathbb{R}^d$ with differentiable densities and their score functions are square-integrable, and write $S_k := X_1 + \cdots + X_k$. Then*

$$\rho_{S_n}(S_n) = \mathrm{E}[\rho_{S_k}(S_k)|S_n]$$

for all $k = 1, \dots, n$.

Consequently we have

$$\mathrm{E}[\| \mathrm{E}[\rho_{S_k}(S_k)|S_n]\|^2] = J(S_n).$$

We now use Cauchy–Schwarz inequality to obtain an upper bound on the squared norm of the reversed conditional expectation.

Lemma 2.3.6. *Let* $X_1, \dots, X_n$ *be mutually independent random variables in* $\mathbb{R}^d$ *with differentiable densities and their score functions are square-integrable. For* $k = 1, \dots, n$ *we write* $S_k := X_1 + \cdots + X_k$*. Then*

$$\mathrm{E}[\| \mathrm{E}[\rho_{S_n}(S_n)|S_k]\|^2] \geq \frac{J(S_n)^2}{J(S_k)}$$

for all $k = 1, \dots, n$.

Proof. Consider

$$\begin{aligned}
J(S_n) &= \mathrm{E}[\|\rho_{S_n}(S_n)\|^2] \\
&= \mathrm{E}[\langle \rho_{S_n}(S_n), \mathrm{E}[\rho_{S_k}(S_k)|S_n]\rangle] \\
&= \mathrm{E}[\mathrm{E}[\langle \rho_{S_n}(S_n), \rho_{S_k}(S_k)\rangle|S_n]] \\
&= \mathrm{E}[\langle \rho_{S_n}(S_n), \rho_{S_k}(S_k)\rangle] \\
&= \mathrm{E}[\mathrm{E}[\langle \rho_{S_n}(S_n), \rho_{S_k}(S_k)\rangle|S_k]] \\
&= \mathrm{E}[\langle \mathrm{E}[\rho_{S_n}(S_n)|S_k], \rho_{S_k}(S_k)\rangle] \\
&\overset{(a)}{\leq} \mathrm{E}[\| \mathrm{E}[\rho_{S_n}(S_n)|S_k]\|^2]^{1/2} \, \mathrm{E}[\|\rho_{S_k}(S_k)\|^2]^{1/2} \\
&= \mathrm{E}[\| \mathrm{E}[\rho_{S_n}(S_n)|S_k]\|^2]^{1/2} J(S_k)^{1/2}.
\end{aligned}$$

where (a) follows from the Cauchy-Schwarz inequality. This gives the result. □

Proposition 2.3.7 (Generalized Stam's inequality, Theorem 2 of [MB07])**.** *Let* $X_1, \dots, X_n$ *be mutually independent random variables in* $\mathbb{R}^d$ *with differentiable* *densities. Suppose β_T ($T \subseteq [1:n]$) are non-negative real numbers satisfying $\sum_{T\subseteq[1:n]:T\ni i} \beta_T \le 1$ for all $i = 1, \dots, n$. Then*

$$\frac{1}{J(S_{[1:n]})} \ge \sum_{T\subseteq[1:n]} \beta_T \frac{1}{J(S_T)}.$$

where $S_T := \sum_{i\in T} X_i$.

Proof. Without loss of generality we can assume $J(S_{[1:n]}) < +\infty$, since otherwise we also have $J(S_T) = +\infty$ for all $T \subseteq [1:n]$. Note that $S_\emptyset = 0$. Let us first assume that $\rho_{S_{[1:n]}}$ is bounded. An application of Lemma 2.2.1 (with $f = \rho_{S_{[1:n]}}$) gives the existence of a family of random variables $\{U^{(\epsilon)}\}_\epsilon$, indexed by small enough $\epsilon > 0$, such that $U^{(\epsilon)} \to S_{[1:n]} \to X_{[1:n]}$ forms a Markov chain and

$$I(U^{(\epsilon)}; S_T) = \frac{1}{2}\epsilon^2 \operatorname{E}[\| \operatorname{E}[\rho_{S_{[1:n]}}(S_{[1:n]})|S_T]\|^2] + O(\epsilon^3) \tag{2.1}$$

for all $T \subseteq [1:n]$. Then Theorem 2.1.10 (iii) implies

$$\sum_{T\subseteq[1:n]} \beta_T I(U^{(\epsilon)}; S_T) \le I(U^{(\epsilon)}; S_{[1:n]}). \tag{2.2}$$

Now consider

$$\begin{aligned} J(S_{[1:n]}) &= \operatorname{E}[\|\rho_{S_{[1:n]}}(S_{[1:n]})\|^2] \\ &\overset{(a)}{\ge} \sum_{T\subseteq[1:n]} \beta_T \operatorname{E}[\| \operatorname{E}[\rho_{S_{[1:n]}}(S_{[1:n]})|S_T]\|^2] \\ &\overset{(b)}{\ge} \sum_{T\subseteq[1:n]} \beta_T \frac{J(S_{[1:n]})^2}{J(S_T)}. \end{aligned}$$

where (a) is obtained by putting (2.1) into (2.2), dividing by $\frac{1}{2}\epsilon^2$ and then taking $\epsilon \to 0$, and (b) follows from Lemma 2.3.6. The result then follows from rearranging.

If $\rho_{S_{[1:n]}}$ is not bounded, then we define $f_B := \min\left\{1, \frac{B}{\|\rho_{S_{[1:n]}}\|}\right\} \rho_{S_{[1:n]}}$ and the proof proceeds as before with $\rho_{S_{[1:n]}}$ replaced by $\hat{f}_B := f_B - \operatorname{E}[f_B(S_{[1:n]})]$ until inequality (a). Now, via the dominated convergence theorem, we let $B \to +\infty$ to recover the form as above with the score functions. □

From generalized Stam's inequality to fractional entropy power inequality

In this section, we provide a new argument based on convex duality that shows that the fractional super-addivity of the EPI follows from the generalized Stam's inequality. The first two lemmas that we present below are well-known (see [MG19] and the references therein) and are the "Lieb-type-equivalent" forms of the fractional EPI and the generalized Stam's inequality. We present a proof of these here for completeness. Lemma 2.3.10 is the crucial observation that leads to the new argument. This lemma is used to show that by restricting our attention to optimal fractional partitions, we can essentially extend the proof for $n = 2$ to larger values of n.

Lemma 2.3.8. *Let $X_1, \ldots, X_n$ be mutually independent random variables in $\mathbb{R}^d$. Let $S_T := \sum_{i \in T} X_i$. Suppose β_T ($T \subseteq [1:n]$, $T \neq \emptyset$) are non-negative real numbers satisfying $\sum_{T \subseteq [1:n]: T \ni i} \beta_T \leq 1$ for all $i \in [1:n]$. Suppose that the differential entropy of S_T is well-defined for all non-empty subsets $T \subseteq [1:n]$. Then the following are equivalent.*

(i) It holds that

$$\exp\left(\frac{2}{d} h(S_{[1:n]})\right) \geq \sum_{\substack{T \subseteq [1:n] \\ T \neq \emptyset}} \beta_T \exp\left(\frac{2}{d} h(S_T)\right).$$

(ii) For all non-negative real numbers w_T ($T \subseteq [1:n]$, $T \neq \emptyset$) with $\sum_{\substack{T \subseteq [1:n] \\ T \neq \emptyset}} w_T = 1$, it holds that

$$h(S_{[1:n]}) \geq \sum_{\substack{T \subseteq [1:n] \\ T \neq \emptyset}} w_T h\left(\sqrt{\frac{\beta_T}{w_T}} S_T\right).$$

Proof. We first show (i) implies (ii). Indeed,

$$\begin{aligned}
\sum_{T\subseteq[1:n] T\neq\emptyset} w_T h\left(\sqrt{\frac{\beta_T}{w_T}}S_T\right) &\overset{(a)}{\leq} \frac{d}{2}\log\left(\sum_{\substack{T\subseteq[1:n]\\ T\neq\emptyset}} w_T \exp\left(\frac{2}{d}h\left(\sqrt{\frac{\beta_T}{w_T}}S_T\right)\right)\right)\\
&= \frac{d}{2}\log\left(\sum_{\substack{T\subseteq[1:n]\\ T\neq\emptyset}} \beta_T \exp\left(\frac{2}{d}h(S_T)\right)\right)\\
&\overset{(b)}{\leq} h(S_{[1:n]}).
\end{aligned}$$

where (a) follows from concavity of $\log(\cdot)$ and (b) follows from (i).

Now we show (ii) implies (i). Set $w_T := \beta_T e^{\frac{2}{d}h(S_T)}\left(\sum_{\substack{\tilde{T}\subseteq[1:n]\\ \tilde{T}\neq\emptyset}} \beta_{\tilde{T}} \exp\left(\frac{2}{d}h(S_{\tilde{T}})\right)\right)^{-1}$. Note that

$$h\left(\sqrt{\frac{\beta_T}{w_T}}S_T\right) = \frac{d}{2}\log\frac{\beta_T\exp\left(\frac{2}{d}h(S_T)\right)}{w_T} = \frac{d}{2}\log\left(\sum_{\substack{\tilde{T}\subseteq[1:n]\\ \tilde{T}\neq\emptyset}} \beta_{\tilde{T}} \exp\left(\frac{2}{d}h(S_{\tilde{T}})\right)\right)$$

is independent of the choice of T, and hence (i) follows immediately from (ii). □

Lemma 2.3.9. *Let $X_1, \dots, X_n$ be mutually independent random variables in $\mathbb{R}^d$. Let $S_T := \sum_{i\in T} X_i$. Suppose β_T ($T \subseteq [1:n]$, $T\neq\emptyset$) are non-negative real numbers satisfying $\sum_{T\subseteq[1:n]:T\ni i}\beta_T \leq 1$ for all $i\in[1:n]$. Suppose that the Fisher information of S_T is well-defined for all non-empty subsets $T\subseteq[1:n]$. Then the following are equivalent.*

(i) It holds that

$$\frac{1}{J(S_{[1:n]})} \geq \sum_{\substack{T\subseteq[1:n]\\ T\neq\emptyset}} \beta_T \frac{1}{J(S_T)}.$$

(ii) For all non-negative real numbers w_T ($T\subseteq[1:n]$, $T\neq\emptyset$) with $\sum_{\substack{T\subseteq[1:n]\\ T\neq\emptyset}} w_T =$

1, *it holds that*

$$J(S_{[1:n]}) \leq \sum_{\substack{T \subseteq [1:n] \\ T \neq \emptyset}} w_T J\left(\sqrt{\frac{\beta_T}{w_T}} S_T\right).$$

Proof. We first show (i) implies (ii). Indeed,

$$\begin{aligned}
\sum_{\substack{T \subseteq [1:n] \\ T \neq \emptyset}} w_T J\left(\sqrt{\frac{\beta_T}{w_T}} S_T\right) &\overset{(a)}{\geq} \left(\sum_{\substack{T \subseteq [1:n] \\ T \neq \emptyset}} w_T \frac{1}{J\left(\sqrt{\frac{\beta_T}{w_T}} S_T\right)}\right)^{-1} \\
&= \left(\sum_{\substack{T \subseteq [1:n] \\ T \neq \emptyset}} \beta_T \frac{1}{J(S_T)}\right)^{-1} \\
&\overset{(b)}{\geq} J(S_{[1:n]}).
\end{aligned}$$

where (a) follows from convexity of $(\cdot)^{-1}$ and (b) follows from (i).

Now we show (ii) implies (i). Set $w_T := \beta_T \frac{1}{J(S_T)} \left(\sum_{\substack{\tilde{T} \subseteq [1:n] \\ \tilde{T} \neq \emptyset}} \beta_{\tilde{T}} \frac{1}{J(S_{\tilde{T}})}\right)^{-1}$. Note that

$$J\left(\sqrt{\frac{\beta_T}{w_T}} S_T\right) = \frac{w_T}{\beta_T} J(S_T) = \left(\sum_{\substack{\tilde{T} \subseteq [1:n] \\ \tilde{T} \neq \emptyset}} \beta_{\tilde{T}} \frac{1}{J(S_{\tilde{T}})}\right)^{-1}$$

is independent of the choice of T, and hence (i) follows immediately from (ii). □

We now present a simple but powerful observation that allows us to simplify the proof that the generalized Stam's inequality implies the fractional superadditivity of EPI.

Lemma 2.3.10. *Let w_T ($T \subseteq [1:n]$, $T \neq \emptyset$) be non-negative real numbers. Then the maximization*

$$\max_{\substack{\beta_T \geq 0 \\ \sum_{T \ni i} \beta_T \leq 1}} \sum_{\substack{T \subseteq [1:n] \\ T \neq \emptyset}} w_T \log \beta_T$$

is attained at $\beta_T = \frac{w_T}{\sum_{i\in T}\lambda_i}$*, for some* $\lambda_i > 0$ *(*$i \in [1:n]$*), with* $\sum_{T\subseteq[1:n]:T\ni i}\beta_T = 1$ *for all* $i \in [1:n]$*.*

Proof. Consider

$$\begin{aligned}
&\max_{\substack{\beta_T\geq 0\\ \sum_{T\ni i}\beta_T\leq 1}} \sum_{\substack{T\subseteq[1:n]\\ T\neq\emptyset}} w_T \log \beta_T \\
&\overset{(a)}{=} \min_{\lambda_i\geq 0}\max_{\beta_T\geq 0}\left(\sum_{\substack{T\subseteq[1:n]\\ T\neq\emptyset}} w_T\log\beta_T + \sum_{i=1}^n \lambda_i\left(1-\sum_{T\ni i}\beta_T\right)\right)\\
&= \min_{\lambda_i\geq 0}\left(\sum_{i=1}^n\lambda_i + \max_{\beta_T\geq 0}\sum_{\substack{T\subseteq[1:n]\\ T\neq\emptyset}}\left(w_T\log\beta_T - \beta_T\sum_{i\in T}\lambda_i\right)\right)\\
&\overset{(b)}{=} \min_{\lambda_i\geq 0}\left(\sum_{i=1}^n\lambda_i + \sum_{\substack{T\subseteq[1:n]\\ T\neq\emptyset}}\left(w_T\log\frac{w_T}{\sum_{i\in T}\lambda_i} - w_T\right)\right).
\end{aligned}$$

where (a) holds by strong duality since Slater's condition (see Theorem 3.2.8 in [BL05] for instance) is satisfied for the maximization on the left hand side, and (b) holds since the maximum is attained at $\beta_T = \frac{w_T}{\sum_{i\in T}\lambda_i}$. The minimization on the last line is a convex problem and is attained at some λ_i^*'s satisfying the first-order condition $\sum_{T\ni i}\frac{w_T}{\sum_{j\in T}\lambda_j^*} = 1$ $(i\in[1:n])$. Let $\beta_T^* := \frac{w_T}{\sum_{i\in T}\lambda_i^*}$. Then

$$\begin{aligned}
&\max_{\substack{\beta_T\geq 0\\ \sum_{T\ni i}\beta_T\leq 1}} \sum_{\substack{T\subseteq[1:n]\\ T\neq\emptyset}} w_T \log \beta_T \\
&\leq \sum_{i=1}^n\lambda_i^* + \sum_{\substack{T\subseteq[1:n]\\ T\neq\emptyset}}\left(w_T\log\beta_T^* - \beta_T^*\sum_{i\in T}\lambda_i^*\right)\\
&= \sum_{\substack{T\subseteq[1:n]\\ T\neq\emptyset}} w_T\log\beta_T^* + \sum_{i=1}^n\lambda_i^* - \sum_{i=1}^n\left(\lambda_i^*\sum_{T\ni i}\beta_T^*\right)\\
&= \sum_{\substack{T\subseteq[1:n]\\ T\neq\emptyset}} w_T\log\beta_T^*.
\end{aligned}$$

hence the maximization on the left hand side of the first line is attained at $\beta_T =$

β_T^*. □

The following lemma shows that the dual variables λ_i in the proof of Lemma 2.3.10 represent the variances of the Gaussians while extending the proof from $n = 2$ to larger n using an approach of calculus of variations.

Lemma 2.3.11. *Let $X_1, \ldots, X_n$ be mutually independent random variables in $\mathbb{R}^d$. Let $S_T := \sum_{i \in T} X_i$. Let w_T ($T \subseteq [1:n]$, $T \neq \emptyset$) be non-negative real numbers satisfying $\sum_{\substack{T \subseteq [1:n] \\ T \neq \emptyset}} w_T = 1$. Let β_T ($T \subseteq [1:n]$, $T \neq \emptyset$) be non-negative real numbers satisfying $\sum_{T \subseteq [1:n]: T \ni i} \beta_T \leq 1$ for all $i \in [1:n]$. Suppose that the Fisher information of S_T is well-defined for all non-empty subsets $T \subseteq [1:n]$. Then (i) implies (ii).*

(i) For all $X_1, \ldots, X_n$, $\{w_T\}$ and $\{\beta_T\}$ it holds that

$$J(S_{[1:n]}) \leq \sum_{\substack{T \subseteq [1:n] \\ T \neq \emptyset}} w_T J\left(\sqrt{\frac{\beta_T}{w_T}} S_T\right).$$

(ii) For all $X_1, \ldots, X_n$, $\{w_T\}$ and $\{\beta_T\}$ it holds that

$$h(S_{[1:n]}) \geq \sum_{\substack{T \subseteq [1:n] \\ T \neq \emptyset}} w_T h\left(\sqrt{\frac{\beta_T}{w_T}} S_T\right).$$

Proof. It suffices to show that (ii) holds for the β_T's that maximize the right-hand side. In view of Lemma 2.3.10 we can write $\beta_T = \frac{w_T}{\sum_{i \in T} \lambda_i}$ for some $\lambda_i > 0$ ($i \in [1:n]$) such that $\sum_{T \subseteq [1:n]: T \ni i} \beta_T = 1$ is satisfied for all $i \in [1:n]$. Consequently, we have

$$\sum_{i=1}^{n} \lambda_i = \sum_{i=1}^{n} \left(\lambda_i \sum_{T \ni i} \beta_T\right) = \sum_{\substack{T \subseteq [1:n] \\ T \neq \emptyset}} \left(\beta_T \sum_{i \in T} \lambda_i\right) = \sum_{\substack{T \subseteq [1:n] \\ T \neq \emptyset}} w_T = 1.$$

Now for $t \in [0,1]$ define

$$f(t) := h\left(\sqrt{1-t}S_{[1:n]} + \sqrt{t}Z\right) - \sum_{\substack{T\subseteq[1:n]\\ T\neq\emptyset}} w_T h\left(\sqrt{\frac{\beta_T}{w_T}}\sqrt{1-t}S_T + \sqrt{t}Z\right).$$

where $Z \sim \mathcal{N}(0,1)$. Note that $f(1) = 0$ and hence it suffices to show $f'(t) \le 0$ for all $0 \le t \le 1$. Indeed

$$\begin{aligned}
f'(t) &= \frac{1}{2}\frac{1}{1-t}\Bigg(J\left(\sqrt{1-t}S_{[1:n]} + \sqrt{t}Z\right) \\
&\quad - \sum_{\substack{T\subseteq[1:n]\\ T\neq\emptyset}} w_T J\left(\sqrt{\frac{\beta_T}{w_T}}\sqrt{1-t}S_T + \sqrt{t}Z\right)\Bigg) \\
&= \frac{1}{2}\frac{1}{1-t}\Bigg(J\left(\sqrt{1-t}S_{[1:n]} + \sqrt{\sum_{i=1}^n \lambda_i}\sqrt{t}Z\right) \\
&\quad - \sum_{\substack{T\subseteq[1:n]\\ T\neq\emptyset}} w_T J\left(\sqrt{\frac{\beta_T}{w_T}}\sqrt{1-t}S_T + \sqrt{\frac{\beta_T}{w_T}\sum_{i\in T}\lambda_i}\sqrt{t}Z\right)\Bigg) \\
&= \frac{1}{2}\frac{1}{1-t}\left(J\left(\sum_{i=1}^n X_{i,t}\right) - \sum_{\substack{T\subseteq[1:n]\\ T\neq\emptyset}} w_T J\left(\sqrt{\frac{\beta_T}{w_T}}\sum_{i\in T} X_{i,t}\right)\right) \\
&\overset{(a)}{\le} 0.
\end{aligned}$$

where we have set $X_{i,t} := \sqrt{1-t}X_i + \sqrt{\lambda_i t}Z_i$, where $Z_i \sim \mathcal{N}(0,1)$, and (a) follows from (i). □

2.3.2 Discrete convexity, strong data processing constant and maximal correlation

In this subsection, we establish some discrete convexity results and consequently some results about strong data processing constants and maximal correlations of joint distributions, generalizing results in [KN15] and [DKS01].

The following is a subclass of layered function families that we will also be considering in this section.

Definition 2.3.12 (Symmetric layered function family)**.** Let $\{S_T\}_T$ be a layered function family on mutually independent and identically distributed random variables $X_1, \ldots, X_n$. We call the layered function family $\{S_T\}_T$ *symmetric* if for all permutations π of $[1 : n]$ the distributions of $(S_{[1:n]}, S_\emptyset, X_1, \ldots, X_n)$ and $(S_{[1:n]}, S_\emptyset, X_{\pi(1)}, \ldots, X_{\pi(n)})$ are the same.

Remark 2.3.13. If $X_1, \ldots, X_n$ are mutually independent and identically distributed random variables, Remark 2.1.7 (i) and (ii) are examples of symmetric layered function families.

Lemma 2.3.14 (Discrete convexity)**.** *Suppose φ_k $(k = 0, 1, \ldots, n)$ are real numbers satisfying*

$$\varphi_{k-1} + \varphi_{k+1} \geq 2\varphi_k \tag{2.3}$$

for all $k = 1, \ldots, n-1$. Then

$$\varphi_k \leq \frac{n-k}{n-l}\varphi_l + \frac{k-l}{n-l}\varphi_n$$

for all $l = 0, 1, \ldots, n-1$, and k satisfying $l \leq k \leq n$.

Proof. Note that $k = n$ and $l = k$ are immediate, so we assume $l < k < n$. Observe that $\varphi_k - \varphi_{k-1}$ is nondecreasing in k. Then

$$\begin{aligned}
\varphi_n - \varphi_k &= (\varphi_n - \varphi_{n-1}) + (\varphi_{n-1} - \varphi_{n-2}) + \cdots + (\varphi_{k+1} - \varphi_k) \\
&\geq (n-k)(\varphi_{k+1} - \varphi_k) \\
&\geq (n-k)(\varphi_k - \varphi_{k-1}) \\
&\geq \frac{n-k}{k-l}((\varphi_k - \varphi_{k-1}) + (\varphi_{k-1} - \varphi_{k-2}) + \cdots + (\varphi_{l+1} - \varphi_l)) \\
&= \frac{n-k}{k-l}(\varphi_k - \varphi_l).
\end{aligned}$$

The result follows by rearranging. □

By leveraging the permutation invariance of the given joint distribution, we demonstrate that the quantity $I(U; S_T)$ both depends exclusively on the cardinality of T and satisfies discrete convexity with respect to $|T|$.

Proposition 2.3.15. *Let $\{S_T\}_T$ be a symmetric layered function family on mutually independent and identically distributed random variables $X_1, \ldots, X_n$. Suppose U is a random variable such that $U \to S_{[1:n]} \to (S_\emptyset, X_{[1:n]})$ forms a Markov chain. Then $I(U; S_T)$ is a function of $|T|$, and we have*

$$I(U; S_T) + I(U; S_{T \cup \{i,j\}}) \geq I(U; S_{T \cup \{i\}}) + I(U; S_{T \cup \{j\}})$$

for all $T \subseteq [1:n]$ and distinct elements i, j in $[1:n] \setminus T$. Furthermore,

$$I(U; S_T) \leq \frac{n - |T|}{n} I(U; S_\emptyset) + \frac{|T|}{n} I(U; S_{[1:n]})$$

for all $T \subseteq [1:n]$.

Proof. We first show that $I(U; S_T)$ is a function of $|T|$. It suffices to establish $I(U; S_T) = I(U; S_{[1:|T|]})$ for all $T \subseteq [1:n]$. Take a permutation π of $[1:n]$, that is increasing on $[1:|T|]$, such that $T = \{\pi(i)\}_{i=1,\ldots,|T|}$. From the definition of symmetric layered function family and the Markov chain $U \to S_{[1:n]} \to (S_\emptyset, X_1, \ldots, X_n)$, we have that the distributions of $(U, S_\emptyset, X_1, \ldots, X_n)$ and $\left(U, S_\emptyset, X_{\pi(1)}, \ldots, X_{\pi(n)}\right)$ are the same. In particular, the distributions of $\left(U, S_\emptyset, X_{[1:|T|]}\right)$ and $(U, S_\emptyset, X_T)$ are the same. Hence Lemma 2.1.9 (ii) gives

$$I(U; S_T) = I(U; S_\emptyset, X_T) = I(U; S_\emptyset, X_{[1:|T|]}) = I(U; S_{[1:|T|]}).$$

Now we show that $\varphi_k := I(U; S_T)$, where T is any subset of $[1:n]$ of cardinality k, satisfies (2.3). For any $k = 1, \ldots, n-1$, take any $T \subseteq [1:n]$ with $|T| = k-1$ and distinct elements i, j in $[1:n] \setminus T$, and we have

$$\begin{aligned}\varphi_{k-1} + \varphi_{k+1} &= I(U; S_T) + I(U; S_{T \cup \{i,j\}}) \\ &\overset{(a)}{\geq} I(U; S_{T \cup \{i\}}) + I(U; S_{T \cup \{j\}})\end{aligned}$$

$$= 2\varphi_k.$$

where (a) follows from (i) of Theorem 2.1.10. Hence (2.3) is satisfied. Then an application of Lemma 2.3.14 (with $l = 0$) yields

$$\varphi_k \leq \frac{n-k}{n}\varphi_0 + \frac{k}{n}\varphi_n.$$

or equivalently,

$$I(U; S_T) \leq \frac{n-|T|}{n} I(U; S_\emptyset) + \frac{|T|}{n} I(U; S_{[1:n]})$$

for all $T \subseteq [1:n]$. □

Using the discrete convexity result above, we can establish the following estimates for conditional expectations and divergences.

Corollary 2.3.16. *Let $\{S_T\}_T$ be a symmetric layered function family on mutually independent and identically distributed random variables $X_1, \ldots, X_n$. Then the following hold:*

(i) Suppose f is an $\mathbb{R}^d$-valued bounded measurable function, defined on the set of values of $S_{[1:n]}$, such that $\mathrm{E}[f(S_{[1:n]})] = 0$. Then

$$\mathrm{E}[\| \mathrm{E}[f(S_{[1:n]})|S_T]\|^2] \leq \frac{n-|T|}{n} \mathrm{E}[\| \mathrm{E}[f(S_{[1:n]})|S_\emptyset]\|^2] + \frac{|T|}{n} \mathrm{E}[\|f(S_{[1:n]})\|^2]$$

for all $T \subseteq [1:n]$.

(ii) Suppose $q(\cdot)$ is a distribution absolutely continuous and with bounded Radon–Nikodym derivative with respect to the distribution of $S_{[1:n]}$. For $T \subseteq [1:n]$ let the random variable $\tilde{S}_T$ be defined by

$$p_{\tilde{S}_T}(\tilde{s}) := \sum_s p_{S_T|S_{[1:n]}}(\tilde{s}|s) q(s).$$

Then

$$\begin{aligned} &D_{\mathrm{KL}}(p_{\tilde{S}_T} \| p_{S_T}) + D_{\mathrm{KL}}(p_{\tilde{S}_{T\cup\{i,j\}}} \| p_{S_{T\cup\{i,j\}}}) \\ &\quad \geq D_{\mathrm{KL}}(p_{\tilde{S}_{T\cup\{i\}}} \| p_{S_{T\cup\{i\}}}) + D_{\mathrm{KL}}(p_{\tilde{S}_{T\cup\{j\}}} \| p_{S_{T\cup\{j\}}}) \end{aligned}$$

for all $T \subseteq [1:n]$ *and distinct elements* i, j *in* $[1:n] \setminus T$*. Furthermore,*

$$D_{\mathrm{KL}}(p_{\tilde{S}_T} \| p_{S_T}) \leq \frac{n-|T|}{n} D_{\mathrm{KL}}(p_{\tilde{S}_\emptyset} \| p_{S_\emptyset}) + \frac{|T|}{n} D_{\mathrm{KL}}(p_{\tilde{S}_{[1:n]}} \| p_{S_{[1:n]}})$$

for all $T \subseteq [1:n]$*.*

Proof. (i) and (ii) are direct applications of Lemma 2.2.1 and 2.2.3, respectively, to Proposition 2.3.15. □

A weakened version of the symmetric layer function family is the cyclically symmetric layer function family, which only requires the joint distribution to remain invariant under cyclic shifts. This yields a weaker yet meaningful discrete convexity result in Proposition 2.3.19.

Definition 2.3.17 (Cyclically symmetric layer function family)**.** Let S be a function on mutually independent and identically distributed random variables $X_1, \ldots, X_n$. We call S *cyclically symmetric* if for all cyclic shifts π of $[1:n]$ the distributions of $(S, X_1, \ldots, X_n)$ and $(S, X_{\pi(1)}, \ldots, X_{\pi(n)})$ are the same.

Remark 2.3.18. The function $S := \sum_{i=1}^n X_i X_{i+1}$ (with $X_{n+1} := X_1$), where X_i's are mutually independent and identically distributed random variables in $\mathbb{R}$, is an example of cyclically symmetric function.

Proposition 2.3.19. *Let S be a cyclically symmetric function on mutually independent and identically distributed random variables $X_1, \ldots, X_n$. Suppose U is a random variable such that $U \to S \to X_{[1:n]}$ forms a Markov chain. Then for all $k = 1, \ldots, n-1$ we have*

$$I(U; X_{[1:k-1]}) + I(U; X_{[1:k+1]}) \geq 2I(U; X_{[1:k]}).$$

Furthermore,

$$I(U; X_{[1:k]}) \leq \frac{k}{n} I(U; S)$$

for all $k = 0, 1, \ldots, n$.

Proof. Since $U \to S \to X_{[1:n]}$ forms a Markov chain and S is a function of $X_{[1:n]}$, we have $I(U; S) = I(U; X_{[1:n]})$. Further from the cyclic symmetry of S and the Markov chain $U \to S \to X_{[1:n]}$, we have that the distributions of $(U, S, X_1, X_2, \ldots, X_n)$ and $(U, S, X_n, X_1, \ldots, X_{n-1})$ are the same. Consequently, for all $k = 0, \ldots, n-1$ we have $I(U; X_{[1:k+1]}) = I(U; X_{[1:k] \cup \{n\}})$. Hence for $k = 1, \ldots, n-1$,

$$\begin{aligned}
&I(U; X_{[1:k+1]}) - I(U; X_{[1:k]}) \\
&= I(U; X_{[1:k] \cup \{n\}}) - I(U; X_{[1:k]}) \\
&= I(U; X_n | X_{[1:k]}) \\
&\overset{(a)}{=} I(U; X_n | X_{[1:k]}) + I(X_k; X_n | X_{[1:k-1]}) \\
&= I(U, X_k; X_n | X_{[1:k-1]}) \\
&\geq I(U; X_n | X_{[1:k-1]}) \\
&= I(U; X_{[1:k-1] \cup \{n\}}) - I(U; X_{[1:k-1]}) \\
&= I(U; X_{[1:k]}) - I(U; X_{[1:k-1]}).
\end{aligned}$$

where (a) holds since X_k is independent of $X_{[1:k-1] \cup \{n\}}$. Now $\varphi_k := I(U; X_{[1:k]})$ satisfies (2.3) and hence by Lemma 2.3.14 (with $l = 0$) we have

$$I(U; X_{[1:k]}) \leq \frac{k}{n} I(U; X_{[1:n]}) = \frac{k}{n} I(U; S)$$

as required. □

Strong data processing constant

We establish a clean upper bound for the strong data processing constant through the construction of specific joint distributions over general symmetric layer function families, utilizing discrete convexity results.

Definition 2.3.20. The *strong data processing constant* $s_*(X;Y)$ of two random variables X, Y is defined by

$$s_*(X;Y) := \sup_{\substack{p(u|x) \\ I(U;X)\neq 0}} \frac{I(U;Y)}{I(U;X)}.$$

Corollary 2.3.21. *Let $\{S_T\}_T$ be a symmetric layered function family on mutually independent and identically distributed random variables $X_1, \ldots, X_n$. Then*

$$s_*(S_{[1:n]}; S_T) \leq \frac{n-|T|}{n} s_*(S_{[1:n]}; S_\emptyset) + \frac{|T|}{n}$$

for all $T \subseteq [1:n]$.

Proof. Fix any U satisfying the Markov chain $U \to S_{[1:n]} \to S_T$. Define a random variable $\tilde{U}$, satisfying the Markov chain $\tilde{U} \to S_{[1:n]} \to (S_\emptyset, X_{[1:n]})$, according to

$$p_{\tilde{U}|S_{[1:n]}}(u|s) := p_{U|S_{[1:n]}}(u|s).$$

Indeed $\tilde{U}$ also satisfies the Markov chain $\tilde{U} \to S_{[1:n]} \to S_T$ since S_T is a function of $(S_\emptyset, X_{[1:n]})$. Hence the distributions of $(U, S_{[1:n]}, S_T)$ and $(\tilde{U}, S_{[1:n]}, S_T)$ are the same. Therefore,

$$\begin{aligned}
\frac{I(U;S_T)}{I(U;S_{[1:n]})} &= \frac{I(\tilde{U};S_T)}{I(\tilde{U};S_{[1:n]})} \\
&\overset{(a)}{\leq} \frac{n-|T|}{n} \frac{I(\tilde{U};S_\emptyset)}{I(\tilde{U};S_{[1:n]})} + \frac{|T|}{n} \\
&\leq \frac{n-|T|}{n} s_*(S_{[1:n]}; S_\emptyset) + \frac{|T|}{n}.
\end{aligned}$$

where (a) is an application of Proposition 2.3.15. □

Remark 2.3.22. Observe that this result generalizes the one in [KN15] from sums of mutually independent and identically distributed random variables to the more general symmetric layered function families. The proof technique used here is clearly motivated by the arguments in [KN15].

Corollary 2.3.23. *Let S be a cyclically symmetric function on mutually independent and identically distributed random variables $X_1, \dots, X_n$. Then $s_*(S; X_{[1:k]}) \leq \frac{k}{n}$ for all $k = 1, \dots, n$.*

Proof. This is immediate from Proposition 2.3.19. □

Maximal correlation

The Hirschfeld–Gebelein–Rényi maximal correlation quantifies dependence between two random variables within general probability spaces. First introduced by Hirschfeld [Hir35] and Gebelein [Geb41], this measure was later studied by Rényi [Rén59]. By employing an auxiliary random variable that encodes conditional expectation structures, we derive analogous upper bounds for the strong data processing inequality through discrete convexity methods.

Definition 2.3.24. The *Hirschfeld–Gebelein–Rényi maximal correlation* $\rho_m(X;Y)$ of two random variables X, Y is defined by

$$\rho_m(X;Y) := \sup_{\substack{f,g \text{ real-valued measurable} \\ \mathrm{E}[f(X)]=\mathrm{E}[g(Y)]=0 \\ \mathrm{E}[f(X)^2]=\mathrm{E}[g(X)^2]=1}} \mathrm{E}[f(X)g(Y)].$$

An alternative expression for the quantity is formulated by Rényi [Rén59] as follows.

Proposition 2.3.25 (Rényi [Rén59]). *Let X, Y be random variables. Then*

$$\rho_m(X;Y) = \sup_{\substack{f \textit{ real-valued measurable} \\ \mathrm{E}[f(X)]=0 \\ \mathrm{E}[f(X)^2]=1}} \mathrm{E}[\mathrm{E}[f(X)|Y]^2]^{1/2}.$$

Corollary 2.3.26. *Let $\{S_T\}_T$ be a symmetric layered function family on mutually independent and identically distributed random variables $X_1, \dots, X_n$. Then*

$$\rho_m(S_{[1:n]}; S_T)^2 \leq \frac{n - |T|}{n} \rho_m(S_{[1:n]}; S_\emptyset)^2 + \frac{|T|}{n}$$

for all $T \subseteq [1:n]$.

Proof. By Corollary 2.3.16 (i), for any bounded real-valued measurable function f such that $\mathrm{E}[f(S_{[1:n]})] = 0$ and $\mathrm{E}[f(S_{[1:n]})^2] = 1$ we have

$$\begin{aligned} &\mathrm{E}[\mathrm{E}[f(S_{[1:n]})|S_T]^2] \\ \leq\ &\frac{n-|T|}{n}\, \mathrm{E}[\mathrm{E}[f(S_{[1:n]})|S_\emptyset]^2] + \frac{|T|}{n}\, \mathrm{E}[f(S_{[1:n]})^2] \\ \leq\ &\frac{n-|T|}{n} \rho_m(S_{[1:n]}; S_\emptyset)^2 + \frac{|T|}{n}. \end{aligned}$$

Taking supremum over f yields the result. □

KL divergence inequality

A direct consequence of Corollary 2.3.16(ii) establishes convexity properties for KL divergence. Through selection of $X_1, \dots, X_n$ as following a fixed Poisson distribution, we demonstrate novel convexity characteristics in the KL divergence between binomial and Poisson distributions by constructing an associated symmetric layer function family over $X_1, \dots, X_n$.

Our findings have a similar favour with Yu's conjecture (Conjecture 1 of [Yu09]), which posits complete monotonicity for $N \mapsto D_{\mathrm{KL}}\left(\mathrm{Binomial}\left(N, \frac{\lambda}{N}\right) \middle\| \mathrm{Poisson}\left(\lambda\right)\right)$. Notably, even proving basic convexity for this function remains an open problem.

The following lemma is well-known and we present a proof here for completeness.

Lemma 2.3.27. *Suppose $X_1 \sim \mathrm{Poisson}(\lambda_1)$ and $X_2 \sim \mathrm{Poisson}(\lambda_2)$ are indepen-*

dent and $Y \sim \mathrm{Binomial}(N, \mu)$*. Then the random variable* $\tilde{Y}$ *defined by*

$$p_{\tilde{Y}}(\tilde{y}) := \sum_{y} p_{X_1|X_1+X_2}(\tilde{y}|y) p_Y(y)$$

satisfies $\tilde{Y} \sim \mathrm{Binomial}\left(N, \frac{\lambda_1}{\lambda_1+\lambda_2}\mu\right)$*.*

Proof. We first compute

$$p_{X_1|X_1+X_2}(\tilde{y}|y) = \frac{p_{X_1}(\tilde{y})p_{X_2}(y-\tilde{y})}{p_{X_1+X_2}(y)} = \binom{y}{\tilde{y}} \frac{\lambda_1^{\tilde{y}}\lambda_2^{y-\tilde{y}}}{(\lambda_1+\lambda_2)^y}.$$

Then

$$\begin{aligned}
p_{\tilde{Y}}(\tilde{y}) &= \sum_{y} p_{X_1|X_1+X_2}(\tilde{y}|y) p_Y(y) \\
&= \sum_{y=\tilde{y}}^{N} \binom{y}{\tilde{y}} \frac{\lambda_1^{\tilde{y}}\lambda_2^{y-\tilde{y}}}{(\lambda_1+\lambda_2)^y} \binom{N}{y} \mu^y (1-\mu)^{N-y} \\
&= \binom{N}{\tilde{y}} \left(\frac{\lambda_1}{\lambda_1+\lambda_2}\mu\right)^{\tilde{y}} \sum_{y=\tilde{y}}^{N} \binom{N-\tilde{y}}{y-\tilde{y}} \left(\frac{\lambda_2}{\lambda_1+\lambda_2}\mu\right)^{y-\tilde{y}} (1-\mu)^{N-y} \\
&= \binom{N}{\tilde{y}} \left(\frac{\lambda_1}{\lambda_1+\lambda_2}\mu\right)^{\tilde{y}} \left(1-\mu+\frac{\lambda_2}{\lambda_1+\lambda_2}\mu\right)^{N-\tilde{y}} \\
&= \binom{N}{\tilde{y}} \left(\frac{\lambda_1}{\lambda_1+\lambda_2}\mu\right)^{\tilde{y}} \left(1-\frac{\lambda_1}{\lambda_1+\lambda_2}\mu\right)^{N-\tilde{y}}
\end{aligned}$$

as required. □

Corollary 2.3.28. *Let* $N \geq 0$*,* $\tilde{\lambda}, \lambda \geq 0$ *and* $0 \leq \mu \leq 1$*. For* $k = 0, 1, \dots, n$ *let*

$$\varphi_k := \quad D_{\mathrm{KL}}\left(\mathrm{Binomial}\left(N, \frac{\tilde{\lambda}+\lambda k}{\tilde{\lambda}+\lambda n}\mu\right) \middle\| \mathrm{Poisson}\left(\tilde{\lambda}+\lambda k\right)\right).$$

Then

$$\varphi_{k-1} + \varphi_{k+1} \geq 2\varphi_k$$

for all $k = 1, \ldots, n-1$, *and*

$$\varphi_k \leq \frac{n-k}{n}\varphi_0 + \frac{k}{n}\varphi_n$$

for all $k = 0, 1, \ldots, n$.

Proof. Let $S_\emptyset \sim \text{Poisson}(\tilde{\lambda})$ and $X_1, \ldots, X_n \sim \text{Poisson}(\lambda)$ be mutually independent random variables. Let $S_T := S_\emptyset + \sum_{i \in T} X_i$ for non-empty $T \subseteq [1:n]$. Note that $\{S_T\}_T$ forms a symmetric layered function family on $X_1, \ldots, X_n$. Also note that $S_T \sim \text{Poisson}(\tilde{\lambda} + \lambda|T|)$ and $S_{[1:n]} - S_T \sim \text{Poisson}(\lambda(n - |T|))$ are independent. Let $\tilde{S}_T$ be defined as in Corollary 2.3.16 (ii) (with $q(\cdot) \sim \text{Binomial}(N, \mu)$). Applying Lemma 2.3.27, we have $\tilde{S}_T \sim \text{Binomial}\left(N, \frac{\tilde{\lambda}+\lambda|T|}{\tilde{\lambda}+\lambda n}\mu\right)$. The result then follows from Corollary 2.3.16 (ii). □

Corollary 2.3.29. *For all* $N \geq 0$ *and* $\lambda \geq 0$, *the function*

$$t \mapsto D_{\text{KL}}\left(\text{Binomial}\left(N, t\right) \| \, \text{Poisson}\left(\lambda t\right)\right)$$

is convex on $[0, 1]$.

Proof. This is immediate from Corollary 2.3.28 (with $\tilde{\lambda} = 0$ and $\mu = 1$) and continuity. □

2.4 Connection with additive combinatorics

One potential application of our main result lies in revealing connections between sumset inequalities in additive combinatorics and entropic inequalities in information theory. To contextualize this relationship, we first recall a fundamental submodularity (or submultiplicativity) property observed in Abelian semigroups:

Theorem 2.4.1 (Theorem 1.2 of [GMR10]). *Let* $A_1, \ldots, A_n$ *be finite, non-empty* *sets in an arbitrary Abelian semigroup. Define $S_T = \sum_{i \in T} A_i$. Then*

$$|S_{[1:n]}|^{n-1} \leq \prod_{i=1}^{n} |S_{[1:n] \setminus \{i\}}|.$$

This combinatorial result finds an information-theoretic counterpart through our framework. The following entropic analogue emerges naturally as a consequence of Theorem 2.1.10:

Corollary 2.4.2. *Let $X_1, \ldots, X_n$ be mutually independent random variables taking values in an arbitrary Abelian semigroup, with $S_T = \sum_{i \in T} X_i$. Then*

$$(n-1)H(S_{[1:n]}) \leq \sum_{i=1}^{n} H(S_{[1:n] \setminus \{i\}}).$$

While no direct implication exists between Theorem 2.4.1 and Corollary 2.4.2, their structural similarity in establishing submodularity highlights a profound parallelism between combinatorial and entropic inequalities. We will further explore this relationship by presenting an entropic equivalent of Theorem 2.4.1 in subsequent chapters.

The pursuit of generalized submodularity properties extends beyond Abelian structures. Ruzsa conjectured the following non-Abelian generalization:

Conjecture 2.4.3 (Conjecture 3.13 of [MMT12])**.** *For finite, non-empty sets $A_1, \ldots, A_n$ in an arbitrary group, we conjecture:*

$$\prod_{i=1}^{n} \max_{a_i \in A_i} |A_1 \circ \cdots \circ A_{i-1} \circ a_i \circ A_{i+1} \circ \cdots \circ A_n| \geq |A_1 \circ \cdots \circ A_n|^{n-1}. \tag{2.4}$$

where $\circ$ denotes the group operation, and $A \circ B := \{a \circ b : a \in A, b \in B\}$ for any subsets A, B of the group.

This conjecture is known to hold for Abelian groups (Theorem 9.3, Chapter 1 of [Ruz09a]). For non-Abelian groups, it has been verified for $n \leq 3$ (Corollary 3.12 of [MMT12]), while general cases remain open (Problem 9.4, Chapter 1 of [Ruz09a]).

Our framework yields new insights into this longstanding problem. By applying 2.1.10 (iii) with $U := X_1 \circ \cdots \circ X_n$ and $S_T := X_T$, we obtain a non-Abelian entropic analogue:

Corollary 2.4.4. *Let $X_1, \ldots, X_n$ be mutually independent random variables with finite support in an arbitrary group. Then*

$$\sum_{i=1}^{n} H(X_1 \circ \cdots \circ X_n | X_i) \geq (n-1) H(X_1 \circ \cdots \circ X_n).$$

This result suggests potential strategies for the proof of Conjecture 2.4.3. The established entropic formulation not only parallels the combinatorial conjecture but also opens avenues for cross-disciplinary proof techniques. Our subsequent work will formalize this connection through an entropic equivalent of Conjecture 2.4.3, potentially enabling new approaches to this fundamental problem in additive combinatorics.

Chapter 3

Rotation Trick in Discrete Spaces and Its Applications

In this chapter, we develop a framework to establish uniform distribution optimality for entropic optimization problems through superadditivity principles and rotational tricks. This approach provides alternative insights into proving the polynomial Freiman–Ruzsa (PFR) functional conjecture.

In Section 3.1, we first review various discrete analogues of EPI that are applicable to different specific families of random variables. We then introduce the Darmois–Skitovich theorem, which plays a central role in proving Gaussian optimality in the continuous EPI. Following this, we present the discrete counterpart of the Darmois–Skitovich theorem, which serves as the motivation for our results.

In Section 3.2, we construct a discrete entropic analogue of Theorem 1.3.10. Our methodology adapts continuous-case proofs with substantial modifications: first identifying superadditive functionals for Theorem 3.2.1's optimization target, then establishing independence relations through perturbed variational problems. The discrete rotation technique emerges via Lemma 3.2.3, which is a variant of Feldman-type theorem, ultimately yielding uniform distribution optimality proofs.

In Section 3.3, we will apply this discrete rotation trick framework to show the optimality of uniform distribution for an entropic functionals, which has been show as equivalent to the PFR conjecture for characteristics 2. This shows the potential of this superadditivity framework in additive combinatorics.

3.1 Preliminaries

Since entropic power inequality plays an important role in network information theory (Section 2.3.1), there are various versions to formulate the discrete analogue of EPI. Shamai and Wyner, [SW90], established a discrete analog of EPI for the binary random variables. Harremoës and Vignaet, [HV03], discovered a discrete analog of EPI for a particular family of binomial random variables. Sharma, Das, and Muthukrishnan, [SDM11] based on the work of [HV03], establish another version of the discrete EPI. On the other hand, there have been generalizations of Mrs. Gerber's Lemma (Wyner and Ziv [WZ73]); for example, Jog and Anantharam have shown a generalization of Mrs. Gerber's Lemma for random variables on the Abelian group with order 2^n [JA14]. These formulations leverage the underlying structure of the specific families of random variables considered.

In our attempt to find a discrete analogue of EPI for general Abelian groups, one approach is to identify the corresponding Lieb's formulation for discrete random variables. This motivates us to investigate the inequality that unifies EPI and BLI (Theorem 1.3.10). The key step in the original argument is establishing Gaussian optimality by utilizing the Darmois-Skitovich theorem (Section 1.2.2).

Theorem 3.1.1 (Darmois-Skitovich theorem [Dar53, Ski53])**.** *Let $X_1, \ldots, X_n$ be independent random variables. Let $\alpha_1, \ldots, \alpha_n, \beta_1, \ldots, \beta_n$ be non-zero constants for each coordinate. If the linear statistics $L_1 = \sum_{i=1}^n \alpha_i X_i$ and $L_2 = \sum_{i=1}^n \beta_i X_i$ are independent, then all random variables $X_1, \ldots, X_n$ are Gaussians.*

A finite Abelian group analog of this was discovered by Feldman [Fel99].

Theorem 3.1.2 (Feldman [Fel99])**.** *Let $\mathbb{G}$ be a finite Abelian group, and X_1, X_2 be independent random variables with values in $\mathbb{G}$. Let $\alpha_1, \alpha_2, \beta_1, \beta_2$ be automorphisms of the group $\mathbb{G}$. Then if the linear statistics $L_1 = \alpha_1(X_1) + \alpha_2(X_2)$ and $L_2 =$*

$\beta_1(X_1) + \beta_2(X_2)$ are independent, then X_1 and X_2 are shifts of a Haar distribution of some subgroup $\mathbb{H}$ of $\mathbb{G}$, or equivalently, X_1 and X_2 are uniform distributions on a coset of some subgroup $\mathbb{H}$ of the group $\mathbb{G}$.

Remark 3.1.3. The uniform distribution on a coset of some subgroup $\mathbb{H}$ of a finite Abelian group $\mathbb{G}$ has very similar properties to that of Gaussians in the respective from the above theorem. By shifting the mean, we see that Haar distributions (uniform distributions) on subgroups play an analogous role to Gaussian distributions.

Therefore, it is natural to guess that Gaussians can be replaced by uniform distributions on a coset (corresponding to a shift in the mean) of some subgroup when working for the discrete analogue of the unified EPI and LBI under finite Abelian groups. However, while this intuition is correct, we show a way to overcome some technical hassles (different from the continuous case) in our proof. Furthermore, just like the rotation trick in the continuous case, we believe this argument can find several other applications to establish the optimality of uniform distributions.

3.2 Discrete analogue of Unified BLI and EPI

The main result of this section (Theorem 3.2.1) is a discrete analog (in finite Abelian groups) of Theorem 1.3.10. Further, we demonstrate that the proof technique in [AJN22] can be essentially mimicked (modulo some differences in the technical arguments) in this setting.

Theorem 3.2.1. *Let $X_1, \ldots, X_n$ be independent random variables taking values in some subgroup $\mathbb{H}_1, \ldots, \mathbb{H}_n$ of a finite Abelian group $\mathbb{G}$. Let $a_1, \ldots, a_n$, and $b_1, \ldots, b_\ell$ be positive constants, and $c_{i,j}^{(1)}, \ldots, c_{i,j}^{(m_j)}$ be integers. Then, the following optimization problem*

$$\max_{\prod_{i=1}^n p_{X_i}} \sum_{i=1}^n a_i H(X_i) - \sum_{j=1}^{\ell} b_j H\left(\sum_{i=1}^n c_{i,j}^{(1)} X_i, \ldots, \sum_{i=1}^n c_{i,j}^{(m_j)} X_i\right),$$

has an optimizer $(X_1^*, \dots, X_n^*)$ of the form, each X_i^* has an uniform distribution on a coset of a subgroup $\mathbb{K}_i \subseteq \mathbb{H}_i$.

Remark 3.2.2. The following points are worth noting:

1. One can relax the assumption on the sign of a_i. Note that, if any $a_k \leq 0$, it is immediate that an optimal choice is to set the corresponding X_k to be a constant random variable. To see this one observes that

$$H\left(\sum_{i=1}^{n} c_{i,j}^{(1)} X_i, \dots, \sum_{i=1}^{n} c_{i,j}^{(m_j)} X_i,\right) \geq H\left(\sum_{i=1}^{n} c_{i,j}^{(1)} X_i, \dots, \sum_{i=1}^{n} c_{i,j}^{(m_j)} X_i | X_k\right)$$

.

2. Unlike the continuous case, where Lieb's formulation of EPI was known, the extremality of the uniform distribution of a coset of some subgroup for $a_1 H(X_1) + a_2 H(X_2) - H(X_1 + X_2)$ was not known. There have been conjectures (and some results), [JA14], of a similar flavor.

3. The optimization problem is the Lagrangian dual of the following:

$$f(x, y) = \min_{H(X)=x, H(Y)=y} H(X + Y)$$

In [JA14], $f(x, y)$ is shown to be convex in x for fixed y, and convex in y for fixed x when the underlying group has order 2^n.

4. Since the underlying group is an Abelian group, we define the random variable kX as $\Pr(kX = y) = \sum_{x:kx=y} \Pr(X = x)$ for all $y \in \mathbb{G}$, where $kx = \underbrace{x + \cdots + x}_{k \text{ times}}$ when k is positive, $kx = 0$ when $k = 0$, and $kx = -|k|x$ if k is negative.

We establish the following lemma before providing proof of Theorem 3.2.1. This is the analogous result of the Darmois-Skitovich theorem we need in our proof.

Lemma 3.2.3. *Let X_A and X_B be two independent random variables taking values in some finite Abelian group $\mathbb{H}$. Let S denote the support of the probability distribution of X_B. Let $\mathbb{D}$ denote the subgroup generated by the pairwise differences of the elements of* $\mathrm{supp}(X_B)$.

For $X_A + X_B$ to be independent of X_B, it is necessary and sufficient that $\mathrm{P}(X_A = h_1) = \mathrm{P}(X_A = h_2)$ *whenever h_1, h_2 belong to the same coset of $\mathbb{D}$ (in other words, p_{X_A} is uniformly distributed conditioned on it taking values in a given coset of $\mathbb{D}$). Consequently* $|\mathrm{supp}(X_A)| = k|\mathbb{D}| \geq k|\mathrm{supp}(X_B)|$ *for some $k \in \mathbb{N}$ satisfying $1 \leq k \leq \frac{|\mathbb{H}|}{|\mathbb{D}|}$, and $k = 1$ only if X_A is uniformly distributed on a coset of $\mathbb{D}$.*

Proof. First, assume that X_A is uniform on the cosets of $\mathbb{D}$. Let T be a set of coset representatives, i.e., a transversal of the collection of cosets of $\mathbb{D}$. Therefore, any element $h \in H$ can be uniquely represented as $h = t + d$, for some $t \in T$ and $d \in \mathbb{D}$. If X_A is uniform on the cosets of $\mathbb{D}$, then $\mathrm{P}(X_A = h) = \mathrm{P}(X_A = t + d) = \frac{1}{|\mathbb{D}|}\mathrm{P}(T = t)$ for some arbitrary distribution on the transversal. If X_A and X_B are independent, note that $\mathrm{P}(X_A + X_B = h + b, X_B = b) = \mathrm{P}(X_A = h)\mathrm{P}(X_B = b) = \frac{1}{|\mathbb{D}|}\mathrm{P}(T = t)\mathrm{P}(X_B = b)$.

On the other hand $\mathrm{P}(X_A + X_B = h + b) = \sum_{\hat{b} \in S} \mathrm{P}(X_A = h + b - \hat{b})\mathrm{P}(X_B = \hat{b})$. Since $b - \hat{b} \in D$, $h + b - \hat{b}$ belongs to the same coset as h. Therefore, for all $\hat{b}$, we have $\mathrm{P}(X_A = h + b - \hat{b}) = \frac{1}{|\mathbb{D}|}\mathrm{P}(T = t)$. Consequently, $\mathrm{P}(X_A + X_B = h + b) = \frac{1}{|\mathbb{D}|}\mathrm{P}(T = t) \sum_{\hat{b} \in S} \mathrm{P}(X_B = \hat{b}) = \frac{1}{|\mathbb{D}|}\mathrm{P}(T = t)$. Therefore $\mathrm{P}(X_A + X_B = h + b, X_B = b) = \mathrm{P}(X_A = h)\mathrm{P}(X_B = b) = \frac{1}{|\mathbb{D}|}\mathrm{P}(T = t)\mathrm{P}(X_B = b) = \mathrm{P}(X_A + X_B = h + b)\mathrm{P}(X_B = b)$. This implies that $X_A + X_B$ is also independent of X_B.

Conversely, let us assume that X_A and X_B are independent, and additionally, $X_A + X_B$ is also independent of X_B. Therefore $\mathrm{P}(X_A + X_B = h + b)\mathrm{P}(X_B = b) = \mathrm{P}(X_A + X_B = h + b, X_B = b) = \mathrm{P}(X_A = h)\mathrm{P}(X_B = b)$. This implies that for all $b \in S$, we have $\mathrm{P}(X_A = h) = \mathrm{P}(X_A + X_B = h + b) = \sum_{\hat{b} \in S} \mathrm{P}(X_A = h + b - \hat{b})\mathrm{P}(X_B = \hat{b})$. Rewriting h as $h - b$, we see that $\mathrm{P}(X_A = h - b) = \sum_{\hat{b} \in S} \mathrm{P}(X_A = h - \hat{b})\mathrm{P}(X_B = \hat{b})$. Since the right-hand-side does not depend on b, we obtain that $\mathrm{P}(X_A = h - b_1) = \mathrm{P}(X_A = h - b_2)$, for all $b_1, b_2 \in S$ and $h \in \mathbb{H}$. Replacing $h - b_1$ by h, we note that $\mathrm{P}(X_A = h) = \mathrm{P}(X_A = h + b_1 - b_2)$. Since the pairwise differences $b_i - b_j$ generate $\mathbb{D}$, and from above p_{X_A} is invariant under a shift by a pairwise difference, it follows that p_{X_A} is invariant under a shift by an element in $\mathbb{D}$. In other words, X_A is uniform on the cosets of $\mathbb{D}$.

Finally note that $|\mathrm{supp}(X_A)| = |\mathrm{supp}(T)||\mathbb{D}|$, and $|\mathrm{supp}(X_A)| = |\mathbb{D}|$ only if T is a constant random variable, implying that X_A is uniform on a coset of $\mathbb{D}$. We also have that $|\mathbb{D}| \geq |\mathrm{supp}(X_B)|$, since $b \mapsto b - b_0$ is an injection from $\mathrm{supp}(X_B)$ to $\mathbb{D}$, where b_0 is an arbitrary fixed element from $\mathrm{supp}(X_B)$. □

Remark 3.2.4. The proof is similar to that in [Tao10, Section 5]. In [Tao10], X_A and X_B are assumed to be identically distributed.

3.2.1 Framework for establishing optimality of uniform distribution

Identifying a superadditive functional

The first step in proving the optimality of the uniform distribution of a coset of some subgroup is to identify a superadditive functional. To this end, given an n-tuple of distributions $(p_{X_1}, \ldots, p_{X_n})$, such that X_i has support on $\mathbb{H}_i$, let us define:

$$F(X_1, \ldots, X_n) := \sup_{\substack{p_{U|X_1,\ldots,X_n}: \\ p_{X_1,\ldots,X_n|U} = \prod_{i=1}^n p_{X_i|U}}} \sum_{i=1}^n a_i H(X_i|U) - \sum_{j=1}^{\ell} b_j H\left(\sum_{i=1}^n c_{i,j}^{(1)} X_i, \ldots, \sum_{i=1}^n c_{i,j}^{(m_j)} X_i | U\right).$$

Observe that the maximum value of $F(X_1, \ldots, X_n)$ is the same as the value of the optimization problem in Theorem 3.2.1, as the average is always less than the maximum (the other direction is immediate by taking $X_1, \ldots, X_n$ to be mutually independent and U to be a constant).

Remark 3.2.5. This is essentially the same function as the one employed in [AJN22].

Now consider an n-tuple of distributions $(p_{X_1,\hat{X}_1},\dots,p_{X_n,\hat{X}_n})$, such that $(X_i,\hat{X}_i)$ has support on $\mathbb{H}_i \times \hat{\mathbb{H}}_i$, let us define (ignoring the abuse of notation):

$$
\begin{aligned}
F((X_1,\hat{X}_1),\dots,(X_n,\hat{X}_n)) := \sup_{\substack{p_{U|(X_1,\hat{X}_1),\dots,(X_n,\hat{X}_n)}: \\ p_{(X_1,\hat{X}_1),\dots,(X_n,\hat{X}_n)|U} = \prod_{i=1}^n p_{(X_i,\hat{X}_i)|U}}} & \sum_{i=1}^n a_i H(X_i,\hat{X}_i|U) \\
- \sum_{j=1}^{\ell} b_j H\left(\sum_{i=1}^n c_{i,j}^{(1)} X_i,\dots,\sum_{i=1}^n c_{i,j}^{(m_j)} X_i, \sum_{i=1}^n c_{i,j}^{(1)} \hat{X}_i,\dots,\sum_{i=1}^n c_{i,j}^{(m_j)} \hat{X}_i|U\right)&.
\end{aligned}
$$

Observe that

$$
\begin{aligned}
&\sum_{i=1}^n a_i H(X_i,\hat{X}_i|U) - \sum_{j=1}^{\ell} b_j H\left(\sum_{i=1}^n c_{i,j}^{(1)} X_i,\dots,\sum_{i=1}^n c_{i,j}^{(m_j)} X_i, \sum_{i=1}^n c_{i,j}^{(1)} \hat{X}_i,\dots,\sum_{i=1}^n c_{i,j}^{(m_j)} \hat{X}_i|U\right) \\
&= \sum_{i=1}^n a_i H(X_i|U) - \sum_{j=1}^{\ell} b_j H\left(\sum_{i=1}^n c_{i,j}^{(1)} X_i,\dots,\sum_{i=1}^n c_{i,j}^{(m_j)} X_i|U\right) \\
&\quad + \sum_{i=1}^n a_i H(\hat{X}_i|U,X_i) - \sum_{j=1}^{\ell} b_j H\left(\sum_{i=1}^n c_{i,j}^{(1)} \hat{X}_i,\dots,\sum_{i=1}^n c_{i,j}^{(m_j)} \hat{X}_i|U, \sum_{i=1}^n c_{i,j}^{(1)} X_i,\dots,\sum_{i=1}^n c_{i,j}^{(m_j)} X_i\right) \\
&\overset{(a)}{=} \sum_{i=1}^n a_i H(X_i|U) - \sum_{j=1}^{\ell} b_j H\left(\sum_{i=1}^n c_{i,j}^{(1)} X_i,\dots,\sum_{i=1}^n c_{i,j}^{(m_j)} X_i|U\right) \\
&\quad + \sum_{i=1}^n a_i H(\hat{X}_i|U,\mathbf{X}) - \sum_{j=1}^{\ell} b_j H\left(\sum_{i=1}^n c_{i,j}^{(1)} \hat{X}_i,\dots,\sum_{i=1}^n c_{i,j}^{(m_j)} \hat{X}_i|U, \sum_{i=1}^n c_{i,j}^{(1)} X_i,\dots,\sum_{i=1}^n c_{i,j}^{(m_j)} X_i\right) \\
&\overset{(b)}{\leq} \sum_{i=1}^n a_i H(X_i|U) - \sum_{j=1}^{\ell} b_j H\left(\sum_{i=1}^n c_{i,j}^{(1)} X_i,\dots,\sum_{i=1}^n c_{i,j}^{(m_j)} X_i|U\right) \\
&\quad + \sum_{i=1}^n a_i H(\hat{X}_i|U,\mathbf{X}) - \sum_{j=1}^{\ell} b_j H\left(\sum_{i=1}^n c_{i,j}^{(1)} \hat{X}_i,\dots,\sum_{i=1}^n c_{i,j}^{(m_j)} \hat{X}_i|U,\mathbf{X}\right) \\
&\overset{(c)}{\leq} F(X_1,\dots,X_n) + F(\hat{X}_1,\dots,\hat{X}_n).
\end{aligned}
$$

In the above $\mathbf{X} = (X_1,\dots,X_n)$. Equality (a) follows, as conditioned on U, $\{(X_i,\hat{X}_i)\}$ are mutually independent and equality (b) follows from data-processing inequality as $(U,\sum_{i=1}^n c_{i,j}^{(1)} X_i,\dots,\sum_{i=1}^n c_{i,j}^{(m_j)} X_i) \to (U,\mathbf{X}) \to (U,\sum_{i=1}^n c_{i,j}^{(1)} \hat{X}_i,\dots,$ $\sum_{i=1}^n c_{i,j}^{(m_j)} \hat{X}_i)$ is Markov. Finally inequality (c) follows since conditioned on U, the random variables $\{X_i\}$ are mutually independent, and conditioned on $(U,\mathbf{X})$, the random variables $\{\hat{X}_i\}$ are mutually independent.

Remark 3.2.6. The next step in the proof (in the continuous case) is to argue that rotated versions of two independent copies of the maximizers are independent. In the continuous case, this involves showing the existence of the maximizers and then (sometimes) considering a perturbed function to deduce the independence of the rotated versions. In the finite alphabet case, the existence of the maximizers is immediate but one still needs to consider a perturbed function to deduce the independence.

Establish discrete analogues for "rotation" trick

In the next part of the proof, we will argue that certain linear forms of the maximizer are independent. To this end, consider the two maximization problems listed below:

$$\max_{\prod_{i=1}^n p_{X_i}} \sum_{i=1}^n a_i H(X_i) - \sum_{j=1}^{\ell} b_j H\left(\sum_{i=1}^n c_{i,j}^{(1)} X_i, \ldots, \sum_{i=1}^n c_{i,j}^{(m_j)} X_i\right),$$
$$\max_{\prod_{i=1}^n p_{\hat{X}_i}} \sum_{i=1}^n a_i H(\hat{X}_i) - \sum_{j=1}^{\ell} b_j H\left(\sum_{i=1}^n c_{i,j}^{(1)} \hat{X}_i, \ldots, \sum_{i=1}^n c_{i,j}^{(m_j)} \hat{X}_i\right) - \sum_{i=1}^n \epsilon H(\hat{X}_i).$$

In the above two problems, the random variables X_i and $\hat{X}_i$ are assumed to take values in the subgroup $\mathbb{H}_i$. Let $(X_1^*, \ldots, X_n^*)$ and $(\hat{X}_{1,\epsilon}^*, \ldots, \hat{X}_{n,\epsilon}^*)$ be maximizers of the two optimization problems respectively and V, V_ϵ be the maximum value attained by the two optimization problems. Further, let us assume that among all possible maximizers of the first problem, $(X_1^*, \ldots, X_n^*)$ minimizes the function $\prod_{i=1}^n (1 + |\mathrm{supp}(X_i)|)$.

It is immediate that $V_\epsilon \to V$ and $\epsilon \to 0$ (as the difference between the objective functions at any point is bounded by $\epsilon \left(\sum_{i=1}^n \log |\mathbb{H}_i|\right)$. Furthermore, by the compactness of the probability simplex and continuity of the function, we know that there is a sequence of maximizers $(\hat{X}_{1,\epsilon_m}^*, \ldots, \hat{X}_{n,\epsilon_m}^*)$ that converge to a maximizer of the first optimization problem.

Finally, we define

$$F_\epsilon(X_1, \ldots, X_n) := \sup_{\substack{p_{U|X_1,\ldots,X_n}: \\ p_{X_1,\ldots,X_n|U} = \prod_{i=1}^n p_{X_i|U}}} \sum_{i=1}^n a_i H(X_i|U)$$

$$- \sum_{j=1}^{\ell} b_j H \left(\sum_{i=1}^{n} c_{i,j}^{(1)} X_i, \ldots, \sum_{i=1}^{n} c_{i,j}^{(m_j)} X_i | U \right) - \sum_{i=1}^{n} \epsilon H(X_i|U).$$

We have $F_\epsilon(X_1, \ldots, X_n) \leq V_\epsilon$.

Observe that by taking independent copies of the maximizers $(X_1^*, \ldots, X_n^*)$ and $(\hat{X}_{1,\epsilon}^*, \ldots, \hat{X}_{n,\epsilon}^*)$, we obtain

$$\begin{aligned}
&V + V_\epsilon \\
&= \sum_{i=1}^{n} a_i H(X_i^*) - \sum_{j=1}^{\ell} b_j H \left(\sum_{i=1}^{n} c_{i,j}^{(1)} X_i^*, \ldots, \sum_{i=1}^{n} c_{i,j}^{(m_j)} X_i^* \right) \\
&\quad + \sum_{i=1}^{n} a_i H(\hat{X}_{i,\epsilon}^*) - \sum_{j=1}^{\ell} b_j H \left(\sum_{i=1}^{n} c_{i,j}^{(1)} \hat{X}_{i,\epsilon}^*, \ldots, \sum_{i=1}^{n} c_{i,j}^{(m_j)} \hat{X}_{i,\epsilon}^* \right) - \sum_{i=1}^{n} \epsilon H(\hat{X}_{i,\epsilon}^*) \\
&\overset{(a)}{=} \sum_{i=1}^{n} a_i H(X_i^*, \hat{X}_{i,\epsilon}^*) - \sum_{i=1}^{n} \epsilon H(\hat{X}_{i,\epsilon}^*) \\
&\quad - \sum_{j=1}^{\ell} b_j H \left(\sum_{i=1}^{n} c_{i,j}^{(1)} X_i^*, \ldots, \sum_{i=1}^{n} c_{i,j}^{(m_j)} X_i^*, \sum_{i=1}^{n} c_{i,j}^{(1)} \hat{X}_{i,\epsilon}^*, \ldots, \sum_{i=1}^{n} c_{i,j}^{(m_j)} \hat{X}_{i,\epsilon}^* \right) \\
&\overset{(b)}{=} \sum_{i=1}^{n} a_i H(X_i^* + \hat{X}_{i,\epsilon}^*, \hat{X}_{i,\epsilon}^*) - \sum_{i=1}^{n} \epsilon H(\hat{X}_{i,\epsilon}^*) \\
&\quad - \sum_{j=1}^{\ell} b_j H \left(\sum_{i=1}^{n} c_{i,j}^{(1)} (X_i^* + \hat{X}_{i,\epsilon}^*), \ldots, \sum_{i=1}^{n} c_{i,j}^{(m_j)} (X_i^* + \hat{X}_{i,\epsilon}^*), \sum_{i=1}^{n} c_{i,j}^{(1)} \hat{X}_{i,\epsilon}^*, \ldots, \sum_{i=1}^{n} c_{i,j}^{(m_j)} \hat{X}_{i,\epsilon}^* \right) \\
&= \sum_{i=1}^{n} a_i H(X_i^* + \hat{X}_{i,\epsilon}^*) - \sum_{j=1}^{\ell} b_j H \left(\sum_{i=1}^{n} c_{i,j}^{(1)} (X_i^* + \hat{X}_{i,\epsilon}^*), \ldots, \sum_{i=1}^{n} c_{i,j}^{(m_j)} (X_i^* + \hat{X}_{i,\epsilon}^*) \right) \\
&\quad + \sum_{i=1}^{n} a_i H(\hat{X}_{i,\epsilon}^* | X_i^* + \hat{X}_{i,\epsilon}^*) - \sum_{i=1}^{n} \epsilon H(\hat{X}_{i,\epsilon}^* | X_i^* + \hat{X}_{i,\epsilon}^*) - \sum_{i=1}^{n} \epsilon I(\hat{X}_{i,\epsilon}^*; X_i^* + \hat{X}_{i,\epsilon}^*) \\
&\quad - \sum_{j=1}^{\ell} b_j H \left(\sum_{i=1}^{n} c_{i,j}^{(1)} \hat{X}_{i,\epsilon}^*, \ldots, \sum_{i=1}^{n} c_{i,j}^{(m_j)} \hat{X}_{i,\epsilon}^* | \sum_{i=1}^{n} c_{i,j}^{(1)} (X_i^* + \hat{X}_{i,\epsilon}^*), \ldots, \sum_{i=1}^{n} c_{i,j}^{(m_j)} (X_i^* + \hat{X}_{i,\epsilon}^*) \right) \\
&\overset{(c)}{\leq} \sum_{i=1}^{n} a_i H(X_i^* + \hat{X}_{i,\epsilon}^*) - \sum_{j=1}^{\ell} b_j H \left(\sum_{i=1}^{n} c_{i,j}^{(1)} (X_i^* + \hat{X}_{i,\epsilon}^*), \ldots, \sum_{i=1}^{n} c_{i,j}^{(m_j)} (X_i^* + \hat{X}_{i,\epsilon}^*) \right) \\
&\quad + \sum_{i=1}^{n} a_i H(\hat{X}_{i,\epsilon}^* | \mathbf{X}^* + \hat{\mathbf{X}}_\epsilon^*) - \sum_{j=1}^{\ell} b_j H \left(\sum_{i=1}^{n} c_{i,j}^{(1)} \hat{X}_{i,\epsilon}^*, \ldots, \sum_{i=1}^{n} c_{i,j}^{(m_j)} \hat{X}_{i,\epsilon}^* | \mathbf{X}^* + \hat{\mathbf{X}}_\epsilon^* \right) \\
&\quad - \sum_{i=1}^{n} \epsilon H(\hat{X}_{i,\epsilon}^* | \mathbf{X}^* + \hat{\mathbf{X}}_\epsilon^*) - \sum_{i=1}^{n} \epsilon I(\hat{X}_{i,\epsilon}^*; X_i^* + \hat{X}_{i,\epsilon}^*) \\
&\overset{(d)}{\leq} F(X_1^* + \hat{X}_{1,\epsilon}^*, \ldots, X_n^* + \hat{X}_{n,\epsilon}^*) + F_\epsilon(\hat{X}_{1,\epsilon}^*, \ldots, \hat{X}_{n,\epsilon}^*) - \sum_{i=1}^{n} \epsilon I(\hat{X}_{i,\epsilon}^*; X_i^* + \hat{X}_{i,\epsilon}^*) \\
&\overset{(e)}{\leq} V + V_\epsilon - \sum_{i=1}^{n} \epsilon I(\hat{X}_{i,\epsilon}^*; X_i^* + \hat{X}_{i,\epsilon}^*).
\end{aligned}$$

Here $\mathbf{X}^* + \hat{\mathbf{X}}^*_\epsilon$ stands for the vector $(X_1^* + \hat{X}^*_{1,\epsilon}, \ldots, X_n^* + \hat{X}^*_{n,\epsilon})$. In the above, equality (a) follows from the independence of $\mathbf{X}^*$ and $\hat{\mathbf{X}}^*_\epsilon$ and equality (b) follows from $H(X_1, X_2) = H(X_1 + X_2, X_2)$. Equality (c) follows from data-processing and the independence of the components of $(\mathbf{X}^* + \hat{\mathbf{X}}^*_\epsilon)$, and (d) follows from the definition of F and F_ϵ as elaborated next. Note that $(X_1^* + \hat{X}^*_{1,\epsilon}, \ldots, X_n^* + \hat{X}^*_{n,\epsilon})$ satisfies the support constraints and is a valid input for the function F (with U taken to be a constant). Now take $U = \mathbf{X}^* + \hat{\mathbf{X}}^*_\epsilon$ and use independence of the components of $(\mathbf{X}^* + \hat{\mathbf{X}}^*_\epsilon)$ to justify that this choice is a valid extension$p_{U|\hat{\mathbf{X}}^*_\epsilon}$ in the definition of F_ϵ. Finally, we note that the maximum of F and F_ϵ are V and V_ϵ to justify the inequality (e).

For $\epsilon > 0$, note that the above manipulations imply that $I(\hat{X}^*_{i,\epsilon}; X_i^* + \hat{X}^*_{i,\epsilon}) = 0$ using the non-negativity of mutual information, or in other words, that $X_i^* + \hat{X}^*_{i,\epsilon}$ is independent of $\hat{X}^*_{i,\epsilon}$. Since X_i^* was independent of $\hat{X}^*_{i,\epsilon}$ by construction, note that we can apply Lemma 3.2.3 to deduce that the distribution of X_i^* is uniform on the cosets of $\mathbb{D}_{i,\epsilon}$. Here $\mathbb{D}_{i,\epsilon}$ is the subgroup of $\mathbb{H}_i$ generated by the pairwise differences of the support of $\hat{X}^*_{i,\epsilon}$. Further $|\text{supp}(X_i^*)| = k_{i,\epsilon}|\mathbb{D}_{i,\epsilon}|$ for some $k_{i,\epsilon} \in \mathbb{N}$ satisfying $1 \leq k_{i,\epsilon} \leq \frac{|\mathbb{H}_i|}{|\mathbb{D}_{i,\epsilon}|}$.

As argued earlier, we have a sequence of optimizers $\hat{\mathbf{X}}^*_{\epsilon_m}$ such that as $\epsilon_m \downarrow 0$ and $\hat{\mathbf{X}}^*_{\epsilon_m}$ converges to a maximizer, say $\tilde{\mathbf{X}}^*$, of the problem with $\epsilon = 0$. Now, we have for any $\epsilon > 0$,

$$\begin{aligned}\prod_{i=1}^n (1 + k_{i,\epsilon}|\mathbb{D}_{i,\epsilon}|) &= \prod_{i=1}^n (1 + |\text{supp}(X_i^*)|) \leq \prod_{i=1}^n (1 + |\text{supp}(\tilde{X}_i^*)|) \\ &= \lim_{m\to\infty} \prod_{i=1}^n (1 + |\text{supp}(\hat{X}^*_{i,\epsilon_m})|) \leq \lim_{m\to\infty} \prod_{i=1}^n (1 + |\mathbb{D}_{i,\epsilon_m}|).\end{aligned}$$

The second assertion holds because we assumed that $\mathbf{X}^*$ minimizes $\prod_{i=1}^n (1 + |\text{supp}(X_i)|)$ among all the maximizers of the optimization problem. This forces, for each $1 \leq i \leq n$, the sequence $k_{i,\epsilon_m} \to 1$ as $m \to \infty$. Therefore for some large, enough m, have $k_{i,\epsilon_m} = 1$ for all i, where $1 \leq i \leq m$. Therefore, again invoking Lemma 3.2.3, we see that X_i^* is uniformly distributed on some coset of a subgroup

$\mathbb{D}_{i,\epsilon_m} \subseteq \mathbb{H}_i$. This completes the proof of Theorem 3.2.1.

Remark 3.2.7. Note that the above argument also establishes some properties of the maximizers of the optimization problem. Suppose $\mathbf{X}_a^*$ is another maximizer such that $\prod_{i=1}^n (1 + |\mathrm{supp}(X_{a,i}^*)|) > \prod_{i=1}^n (1 + |\mathrm{supp}(X_i^*)|)$. Then, the above argument implies that one cannot have a sequence of maximizers of the perturbed problem that converges to $\mathbf{X}_a^*$.

3.3 Application in additive combinatorics

Recent work by Gowers, Green, Manners, and Tao [GGMT23] established uniform distribution optimality for discrete information functionals - equivalent to resolving the Polynomial Freiman–Ruzsa (PFR) conjecture in characteristic 2 groups. Our analysis progresses through three key stages:

In Section 3.3.1, we introduce the entropic functional $\tau(X, Y; X^0, Y^0)$, demonstrating that we aim to prove uniform distributions (X^*, Y^*) minimize τ when fixing reference distributions (X^0, Y^0). Section 3.3.2 develops preliminary superadditivity properties for τ, though these prove insufficient for full optimality characterization.

The conclusive Section 3.3.3 presents an advanced superadditivity argument establishing uniform distribution optimality for PFR functionals, albeit with relaxed constant constraints.

3.3.1 Entropic formulation of PFR conjecture

Definition 3.3.1 (Independent entropic Ruzsa distance, [GGMT23])**.** Suppose X, Y are $\mathbb{G}$-valued random variables. The independent entropic Ruzsa distance between X and Y is defined as

$$d(X, Y) = H(X' + Y') - \frac{1}{2}H(X) - \frac{1}{2}H(Y),$$

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

Remark 3.3.2. This is sometimes defined as $H(X'-Y') - \frac{1}{2}H(X) - \frac{1}{2}H(Y)$. For groups with characteristic 2, these two definitions are equivalent. There is also another "entropic Ruzsa distance" (defined in [KLN23]) where

$$d_{\text{coupling}}(X,Y) = \max_{\Pi(p_x,p_y)} H(X'-Y') - \frac{1}{2}H(X) - \frac{1}{2}H(Y),$$

where $\Pi(p_x,p_y)$ denotes the set of couplings with fixed marginals. Note that none of the above definitions is a distance. When $p_x = p_y$, it does not hold that $d(X,Y) = 0$.

Lemma 3.3.3. *The independent entropic Ruzsa distance satisfies the triangle inequality, i.e.* $d(X,Z) \leq d(X,Y) + d(Y,Z)$.

Proof. Let (X,Y,Z) be independent. What we need to show is equivalent to

$$H(X+Z) + H(Y) \leq H(X+Y) + H(Y+Z).$$

This can be rewritten as

$$I(X;X+Y+Z) \leq I(X;X+Y) + I(Y;X+Y+Z).$$

By data-processing inequality, as $X \to X+Y \to X+Y+Z$ is Markov, $I(X;X+Y+Z) \leq I(X;X+Y)$ and the lemma follows. □

Definition 3.3.4 (Conditionally-independent entropic Ruzsa distance)**.** Suppose X, Y are $\mathbb{G}$-valued random variables. The conditionally-independent entropic Ruzsa distance between X and Y is defined as

$$d(X,Y|U) = H(X'+Y'|U) - \frac{1}{2}H(X|U) - \frac{1}{2}H(Y|U)$$

where $(U,X') \sim (U,X)$, $(U,Y') \sim (U,Y)$, and $X' \to U \to Y'$ is Markov.

Definition 3.3.5 (Polynomial Freiman–Ruzsa functional)**.** [GGMT23, Equation 2.1] For any random variables X^0, Y^0 with support contained inside $\mathbb{G}$, a finite

Abelian group with characteristic 2, define the functional

$$\begin{aligned}\tau(X,Y;X^0,Y^0) := &\left(H(X+Y)-\frac{1}{2}H(X)-\frac{1}{2}H(Y)\right)\\ &+\eta\left(H(X+X^0)-\frac{1}{2}H(X)-\frac{1}{2}H(X^0)\right)\\ &+\eta\left(H(Y+Y^0)-\frac{1}{2}H(Y)-\frac{1}{2}H(Y^0)\right),\end{aligned}$$

where X, Y, X^0, Y^0 are mutually independent. Here X, Y also take values in $\mathbb{G}$.

It was shown in [GGMT23, Proposition 2.1] that all minimizers of $\tau(X,Y)$ must be uniform distributions on a coset of a subgroup for all X^0, Y^0 with support in $\mathbb{G}$, when $\eta \leq \frac{1}{9}$.

3.3.2 Elementary superadditive results for PFR functional

A natural question to ask is whether there is a related superadditive function and whether one can use the machinery developed in the first part of the chapter to deduce the optimality of the uniform distribution. The answer to the former part is yes, while the latter part seems to be not as straightforward.

Let us consider a slight modification of the above functional.

Definition 3.3.6 (Conditional PFR functional)**.** Let X^0 and Y^0 be fixed $\mathbb{G}$-valued random variables. Suppose U, X, Y are $\mathbb{G}$-valued random variables. We require the triple $(U,X,Y), X^0, Y^0$ are independent. We define the conditional PFR functional as below

$$\begin{aligned}&\tau(X,Y;X^0,Y^0|U)\\ &:= d(X,Y|U)+\eta d(X,X^0|U)+\eta d(Y,Y^0|U)\\ &= H(X'+Y'|U)-\frac{1+\eta}{2}H(X|U)-\frac{1+\eta}{2}H(Y|U)\\ &\qquad +\eta H(X+X^0|U)+\eta H(Y+Y^0|U)-\frac{\eta}{2}H(X^0)-\frac{\eta}{2}H(Y^0)\end{aligned}$$

where $(U,X')\sim(U,X)$, $(U,Y')\sim(U,Y)$, and $X'\to U\to Y'$ is Markov.

Define the two-letter form

$$\begin{aligned}
&T((X_a, X_b), (Y_a, Y_b); (X_a^0, Y_a^0), (X_b^0, Y_b^0)) \\
&:= \min_{\substack{p_{U|X_a,X_b,Y_a,Y_b}: \\ p_{X_a,X_b,Y_a,Y_b|U} = p_{X_a,X_b|U} p_{Y_a,Y_b|U}}} H(X_a + Y_a, X_b + Y_b|U) - \frac{1}{2}H(X_a, X_b|U) - \frac{1}{2}H(Y_a, Y_b|U) \\
&\quad + \eta \left(H(X_a + X_a^0, X_b + X_b^0|U) - \frac{1}{2}H(X_a, X_b|U) - \frac{1}{2}H(X_a^0, X_b^0|U) \right) \\
&\quad + \eta \left(H(Y_a + Y_a^0, Y_b + Y_b^0|U) - \frac{1}{2}H(Y_a, Y_b|U) - \frac{1}{2}H(Y_a^0, Y_b^0|U) \right),
\end{aligned}$$

where the tuple $(U, (X_a, X_b), (Y_a, Y_b)), X_a^0, X_b^0, Y_a^0$, and Y_b^0 are mutually independent.

Lemma 3.3.7. *For any $\eta \geq 0$, following superadditivity inequality holds:*

$$T((X_a, X_b), (Y_a, Y_b); (X_a^0, Y_a^0), (X_b^0, Y_b^0)) \geq T(X_a, Y_a; X_a^0, Y_a^0) + T(X_b, Y_b; X_b^0, Y_b^0)$$

Proof. Observe that the following holds:

$$\begin{aligned}
&H(X_a + Y_a, X_b + Y_b|U) - \frac{1}{2}H(X_a, X_b|U) - \frac{1}{2}H(Y_a, Y_b|U) \\
&= H(X_a + Y_a|U) - \frac{1}{2}H(X_a|U) - \frac{1}{2}H(Y_a|U) + H(X_b + Y_b|U, X_a - Y_a) \\
&\qquad - \frac{1}{2}H(X_b|U, X_a) - \frac{1}{2}H(Y_b|U, Y_a) \\
&\overset{(a)}{=} H(X_a + Y_a|U) - \frac{1}{2}H(X_a|U) - \frac{1}{2}H(Y_a|U) + H(X_b + Y_b|U, X_a + Y_a) \\
&\qquad - \frac{1}{2}H(X_b|U, X_a, Y_a, X_a^0, Y_a^0) - \frac{1}{2}H(Y_b|U, X_a, Y_a, X_a^0, Y_a^0) \\
&\geq H(X_a + Y_a|U) - \frac{1}{2}H(X_a|U) - \frac{1}{2}H(Y_a|U) + H(X_b + Y_b|U, X_a, Y_a, X_a^0, Y_a^0) \\
&\qquad - \frac{1}{2}H(X_b|U, X_a, Y_a, X_a^0, Y_a^0) - \frac{1}{2}H(Y_b|U, X_a, Y_a, X_a^0, Y_a^0).
\end{aligned}$$

Here (a) follows from the independence and the Markov structure of the random variables.

In an identical fashion, we can also show that

$$H(X_a + X_a^0, X_b + X_b^0|U) - \frac{1}{2}H(X_a, X_b|U) - \frac{1}{2}H(X_a^0, X_b^0|U)$$

$$
\begin{aligned}
&\geq H(X_a + X_a^0|U) - \frac{1}{2}H(X_a|U) - H(X_a^0|U) + H(X_b + X_b^0|U, X_a, Y_a, X_a^0, Y_a^0)\\
&\quad - \frac{1}{2}H(X_b|U, X_a, Y_a, X_a^0, Y_a^0) - \frac{1}{2}H(X_b^0|U, X_a, Y_a, X_a^0, Y_a^0),
\end{aligned}
$$

and

$$
\begin{aligned}
&H(Y_a + Y_a^0, Y_b - Y_b^0|U) - \frac{1}{2}H(Y_a, Y_b|U) - \frac{1}{2}H(Y_a^0, Y_b^0|U)\\
&\geq H(Y_a + Y_a^0|U) - \frac{1}{2}H(Y_a|U) - H(Y_a^0|U) + H(Y_b + Y_b^0|U, X_a, Y_a, X_a^0, Y_a^0)\\
&\quad - \frac{1}{2}H(Y_b|U, X_a, Y_a, X_a^0, Y_a^0) - \frac{1}{2}H(Y_b^0|U, X_a, Y_a, X_a^0, Y_a^0).
\end{aligned}
$$

Denote $U_a = U$, and observe that $p_{X_aY_a|U_a} = p_{X_a|U_a}p_{Y_a|U_a}$ and $(U_a, X_a, Y_a), X_a^0$, and Y_a^0 are mutually independent. Denote $U_b = (U, X_a, Y_a, X_a^0, Y_a^0)$, and observe that $p_{X_bY_b|U_b} = p_{X_b|U_b}p_{Y_b|U_b}$ and $(U_b, X_b, Y_b), X_a^0$, and Y_a^0 are mutually independent. Putting the above inequalities together, the requisite superadditivity follows. □

However, we cannot do the transformation $(X_a + X_a^0, X_b + X_b^0) \mapsto (X_a + X_a^0 + X_b + X_b^0, X_b + X_b^0)$ as this would replace X_a^0 by $X_a^0 + X_b^0$. This is not permitted as X_a^0 is a fixed distribution. Instead, one can place X_a, X_b, Y_a, Y_b at the minimizer by alternate linear forms and use the minimality to force an independence of some linear forms.

3.3.3 A superadditivity proof for the optimality of uniform distribution in PFR functional

We believe that it will be illustrative to revisit the arguments in [GGMT23] in light of superadditivity. For the purpose of illustration of the ideas, we will try to keep our estimates rather elementary (the ideas are still borrowed, in many cases verbatim, from [GGMT23]). We will establish the following (weaker) result.

Theorem 3.3.8. *Let X^0, Y^0 be any pair of independent random variables with* *support contained inside $\mathbb{G}$, a finite Abelian group with characteristic 2. Let*

$$\begin{aligned}\tau(X,Y) := &\left(H(X+Y) - \frac{1}{2}H(X) - \frac{1}{2}H(Y)\right)\\ &+\eta\left(H(X+X^0) - \frac{1}{2}H(X) - \frac{1}{2}H(X^0)\right)\\ &+\eta\left(H(Y+Y^0) - \frac{1}{2}H(Y) - \frac{1}{2}H(Y^0)\right),\end{aligned}$$

where X, Y, X^0, Y^0 are mutually independent. Here X, Y also take values in $\mathbb{G}$. Then, all minimizers of $\tau(X,Y)$ must be uniform distributions on a coset of a subgroup of $\mathbb{G}$, when $\eta \leq \eta_0$, where $\eta_0 = \frac{\sqrt{1452}-36}{26}$.

We will divide the proof of Theorem 3.3.8 into some components. Some of the required inequalities will be established in the Appendix.

Superadditivity and rotation in PFR functional

Suppose (X^*, Y^*) is a minimizer of $\tau(X,Y;X^0,Y^0)$. Without loss of generality, we may assume X^* and Y^* are independent. Let (X_A, Y_A) and (X_B, Y_B) are independent copies of (X^*, Y^*). The minimality of (X^*, Y^*) implies that

$$\begin{aligned}&\tau(X_A+U_A, Y_A+V_A; X_A^0, Y_A^0|W_A) + \tau(X_B+U_B, Y_B+V_B; X_B^0, Y_B^0|W_B)\\ &\quad\geq \tau(X_A, Y_A; X_A^0, Y_A^0) + \tau(X_B, Y_B; X_B^0, Y_B^0) \qquad (3.1)\end{aligned}$$

for any valid choice that $X_A+U_A \to W_A \to Y_A+V_A$ and $X_B+U_B \to W_B \to Y_B+V_B$. Here, $(U_A, V_A, W_A, U_B, V_B, W_B, X_A, Y_A, X_B, Y_B)$ is assumed to be independent of $(X_A^0, Y_A^0, X_B^0, Y_B^0)$.

Set $U_A = X_B, V_A = Y_B, W_B = (X_A+X_B, Y_A+Y_B), W_A = U_B = V_B = \emptyset$. Then (3.1) reduces to

$$\begin{aligned}&I(X_A+X_B; X_B+Y_B|X_A+Y_A+X_B+Y_B)\\ &\leq \eta I(X_B; X_A+X_B+X_A^0) + \eta I(Y_B; Y_A+Y_B+Y_A^0)\\ &\qquad - \eta I(X_A+X_B; X_B+X_B^0) - \eta I(Y_A+Y_B; Y_B+Y_B^0)\end{aligned}$$

$$\begin{aligned} &\leq \eta I(X_B; X_A + X_B + X_A^0) + \eta I(Y_B; Y_A + Y_B + Y_A^0) \\ &\leq \eta I(X_B; X_A + X_B) + \eta I(Y_B; Y_A + Y_B). \end{aligned} \tag{3.2}$$

The last inequality is due to $(X_A^0, Y_A^0) \perp (X_A, X_B, Y_A, Y_B)$.

Similarly, by setting $U_A = Y_B, V_A = X_B, W_B = (X_A + Y_B, Y_A + X_B), W_A = U_B = V_B = \emptyset$, (3.1) yields

$$I(X_A + Y_B; X_B + Y_B | X_A + Y_A + X_B + Y_B) \leq \eta I(X_B; Y_A + X_B) + \eta I(Y_B; X_A + Y_B). \tag{3.3}$$

Finally, by setting, $U_A = X_A, V_A = Y_B, W_B = (X_A + X_B, Y_A + Y_B), W_A = U_B = V_B = \emptyset$, (3.1) yields

$$I(X_A + X_B; X_A + Y_B | X_A + Y_A + X_B + Y_B) \leq \eta I(X_A; X_A + X_B) + \eta I(Y_B; Y_A + Y_B). \tag{3.4}$$

Remark 3.3.9. We have employed three different linear transformations on the superadditive function and obtained three constraints (equations (3.2),(3.3),(3.4)) that has to be satisfied by the minimizer. Following the approach in the earlier sections, we need to use these inequalities to deduce some independence of linear forms, which would imply that minimizers need to be uniform. The choice of the identifications (three inequalities) above is directly motivated from [[GGMT23], Equations 3.1—3.4]. It may be possible that one could use other linear transformations and obtain additional constraints that implies the independence for a lower η but this is left for future work.

Lemma 3.3.10 ([GGMT23], Equation 5.9)**.** *Let $S = (X_A + X_B) + (Y_A + Y_B)$.*

$$H(S) - \frac{1}{2}H(X) - \frac{1}{2}H(Y) \leq (2 + \eta)d(X, Y).$$

Proof. By optimality of (X, Y), we have $\tau(X_A, Y_A; X^0, Y^0 | X_A + Y_B, Y_A + X_B) \geq$

$\tau(X, Y; X^0, Y^0)$, which is equivalent to

$$\begin{aligned} d(X_A, Y_A|X_A + Y_B, Y_A + X_B) \geq d(X, Y) &- \eta(d(X^0, X_A|X_A + Y_B) - d(X^0, X_A)) \\ &- \eta(d(Y^0, Y_A|Y_A + X_B) - d(Y^0, Y_A)). \end{aligned}$$

This implies

$$\begin{aligned} &d(X_A + Y_B, Y_A + X_B) \\ &\overset{(A.4)}{=} 2d(X, Y) - d(X_A, Y_A|X_A + Y_B, Y_A + X_B) \\ &\quad - I(X_A + Y_A; Y_A + X_B|X_A + X_B + Y_A + Y_B) \\ &\leq d(X, Y) + \eta(d(X^0, X_A|X_A + Y_B) - d(X^0, X_A)) \\ &\quad + \eta(d(Y^0, Y_A|Y_A + X_B) - d(Y^0, Y_A)) \\ &\quad - I(X_A + Y_A; Y_A + X_B|X_A + X_B + Y_A + Y_B) \\ &\overset{(a)}{\leq} d(X, Y) + \eta\left(\frac{1}{2}H(X_A + Y_B) - \frac{1}{2}H(Y_B) + \frac{1}{2}H(Y_A + X_B) - \frac{1}{2}H(X_B)\right) \\ &\quad - I(X_A + Y_A; Y_A + X_B|X_A + X_B + Y_A + Y_B) \\ &= (1 + \eta)d(X, Y) - I(X_A + Y_A; Y_A + X_B|X_A + X_B + Y_A + Y_B), \end{aligned}$$

where (a) follows from Family 3 of Lemma A.1.4. Therefore,

$$d(X_A + Y_B; Y_A + X_B) \leq (1 + \eta)d(X, Y).$$

This implies that

$$\begin{aligned} &H(S) - \frac{1}{2}H(X_A) - \frac{1}{2}H(Y_A) \\ &= d(X_A + Y_B; Y_A + X_B) + \frac{1}{2}I(Y_B; X_A + Y_B) + \frac{1}{2}I(X_B; Y_A + X_B) \\ &= d(X_A + Y_B; Y_A + X_B) + d(X, Y) \\ &\leq (2 + \eta)d(X, Y). \end{aligned}$$

□

Lemma 3.3.11 (Inspired by [GGMT23], Equation 7.2)**.** *Let* $S = (X_A + X_B) + (Y_A + Y_B)$*. Let* $T_1 = X_A + X_B$*,* $T_2 = X_B + Y_B$*,* $T_3 = X_A + Y_B$*.*

$$\begin{aligned} &I(T_1; T_2|S) + I(T_2; T_3|S) + I(T_1; T_3|S) \\ &\qquad \le 2\eta\Big(d(X^*, X^*) + d(Y^*, Y^*) + d(X^*, Y^*)\Big) \le 10\eta d(X^*, Y^*). \end{aligned}$$

Proof. From the superadditivity estimation, i.e. (3.2),(3.3),(3.4), we have

$$\begin{aligned} &I(T_1; T_2|S) + I(T_2; T_3|S) + I(T_1; T_3|S) \\ &\le \eta I(X_B; X_A + X_B) + \eta I(Y_B; Y_A + Y_B) + \eta I(X_B; Y_A + X_B) \\ &\quad + \eta I(Y_B; X_A + Y_B) + \eta I(X_A; X_A + X_B) + \eta I(Y_B; Y_A + Y_B) \\ &= \eta(2H(X_A + X_B) + 2H(Y_A + Y_B) + H(X_A + X_B) + H(X_B + Y_A) - 3H(X^*) - 3H(Y^*)) \\ &= 2\eta(d(X^*, X^*) + d(Y^*, Y^*) + d(X^*, Y^*)) \\ &\le 10\eta d(X^*, Y^*) \end{aligned}$$

The last part of the inequality follows by the triangle inequality, Lemma 3.3.3, of the independent entropic Ruzsa distance, i.e. $d(X, X) \le d(X, Y) + d(Y, X) = 2d(X, Y)$. □

Inducing independent relationships for minimizers

We have, if (X, Y) is the minimizer for $\tau(X, Y; X_0, Y_0)$, then

$$\begin{aligned} &\tau(X, Y; X^0, Y^0) \\ &\le \frac{1}{6}\Big(\tau(T_1, T_2; X^0, Y^0|T_3, S) + \tau(T_2, T_3; X^0, Y^0|T_1, S) + \tau(T_3, T_1; X^0, Y^0|T_2, S) \\ &\qquad + \tau(T_2, T_1; X^0, Y^0|T_3, S) + \tau(T_3, T_2; X^0, Y^0|T_1, S) + \tau(T_1, T_3; X^0, Y^0|T_2, S)\Big) \\ &= \frac{1}{3}\Big(d(T_1, T_2|T_3, S) + d(T_2, T_3|T_1, S) + d(T_3, T_1|T_2, S)\Big) \\ &\qquad + \frac{\eta}{6}\sum_{i=1}^{3}\sum_{j \ne i}\Big(d(X^0, T_i|T_j, S) + d(Y^0, T_i|T_j, S)\Big) \end{aligned}$$

$$\begin{aligned}
&\overset{(a)}{\leq} I(T_1;T_2|S) + I(T_2;T_3|S) + I(T_1;T_3|S) + \frac{\eta}{3}\sum_{i=1}^{3}\Big(d(X^0,T_i|S) + d(Y^0,T_i|S)\Big)\\
&\qquad + \frac{\eta}{3}\Big(I(T_1;T_2|S) + I(T_2;T_3|S) + I(T_1;T_3|S)\Big)\\
&\overset{(b)}{\leq} \left(1+\frac{\eta}{3}\right)\Big(I(T_1;T_2|S) + I(T_2;T_3|S) + I(T_1;T_3|S)\Big)\\
&\qquad + \eta\Big(d(X,X^0) + d(Y,Y^0) + H(S) - \frac{1}{2}H(X) - \frac{1}{2}H(Y)\Big),
\end{aligned}$$

where (a) follows by Corollary A.1.3 and Lemma A.1.1, (b) follows by Families 1 and 2 of Lemma A.1.4.

By the definition of the PFR functional, we have

$$\begin{aligned}
d(X^*,Y^*) &\leq \left(1+\frac{\eta}{3}\right)\Big(I(T_1;T_2|S) + I(T_2;T_3|S) + I(T_1;T_3|S)\Big)\\
&\qquad + \eta\Big(H(S) - \frac{1}{2}H(X) - \frac{1}{2}H(Y)\Big)\\
&\leq \left(1+\frac{\eta}{3}\right)10\eta d(X^*,Y^*) + \eta(2+\eta)d(X^*,Y^*),
\end{aligned}$$

where the last inequality is a consequence of Lemma 3.3.10 and Lemma 3.3.11. Therefore, if $1 > \left(1+\frac{\eta}{3}\right)10\eta + \eta(2+\eta)$, for some $\eta > 0$, then $d(X^*,Y^*) = 0$. Note that $\eta_0 = \frac{\sqrt{1452}-36}{26}$, is the positive root of $1 = \left(1+\frac{\eta}{3}\right)10\eta + \eta(2+\eta)$. Therefore, for $\eta < \eta_0$, $d(X^*,Y^*) = 0$, or in other words, $0 = H(X^*+Y^*) - \frac{1}{2}H(X^*) - \frac{1}{2}H(Y^*) = \frac{1}{2}I(X^*;X^*+Y^*) + \frac{1}{2}I(Y^*;X^*+Y^*)$. This implies that X^* is independent of X^*+Y^* and Y^* is independent of X^*+Y^*.

Establishing optimality of uniform distributions for PFR functional

From Lemma 3.2.3 and that X^* is independent of X^*+Y^*, we have $|\mathrm{supp}(X^*)| \geq k|\mathrm{supp}(Y^*)|$ for some $k \in \mathbb{N}$, and from Y^* is independent of X^*+Y^*, we have $|\mathrm{supp}(Y^*)| \geq k|\mathrm{supp}(X^*)|$, $k \in \mathbb{N}$. This implies that $|\mathrm{supp}(X^*)| = |\mathrm{supp}(Y^*)|$. Further, from Lemma 3.2.3, we can also conclude that $|\mathbb{D}| = |\mathrm{supp}(Y^*)|$, where $\mathbb{D}$ denote the subgroup generated by pairwise differences of the elements of $\mathrm{supp}(Y^*)$. This implies that Y^* is supported on a coset of $\mathbb{D}$. Further, from Lemma 3.2.3, and

$|\mathrm{supp}(X^*)| = |\mathbb{D}|$, we can also infer that X^* is uniform on a coset of $\mathbb{D}$. Reversing the roles of X^* and Y^*, we can also infer the Y^* is uniform over its support. Therefore, X^* and Y^* are uniformly distributed on cosets of the same subgroup.

3.4 Discussion

As referenced in Section 2.3.1, numerous extensions of the Entropy Power Inequality (EPI) exist in contemporary literature. One notable extension concerns monotonicity properties of $h\left(\frac{X_1+\cdots+X_n}{\sqrt{n}}\right)$, raising natural questions about discrete analogs. Following this direction, Tao's conjecture [Tao10] proposes a discrete EPI analogue for torsion-free groups:

Conjecture 3.4.1. *Suppose $X_1, \ldots, X_{n+1}$ are identically distributed and independent random variables on some torsion-free group $\mathbb{T}$. Then, for any $\epsilon > 0$, as long as $H(X)$ is sufficiently large (depending on n, ϵ), we have*

$$H(X_1 + \cdots + X_{n+1}) \geq H(X_1 + \cdots + X_n) + \frac{1}{2} \log \frac{n+1}{n} - \epsilon.$$

While validated for $n = 1$, the general case remains open. Gavalakis' recent work [Gav23] proves this conjecture under log-concave distribution assumptions. Our superadditivity framework proves inapplicable here due to Lemma 3.2.3's limitations in infinite Abelian groups, exacerbated by torsion-free groups containing only trivial finite subgroups. This challenges us to develop Lemma 3.2.3 adaptations for torsion-free settings.

On the other hand, the discrete rotation technique suggests promising applications in characterizing capacity regions for network information theory problems. Consider the Z-interference Gaussian channel capacity problem, featuring an information-theoretic functional analogous to our discrete entropy analysis. Current conjectures propose Gaussian distribution optimality under specific parameter constraints - mirroring uniform distribution optimality in discrete settings with small η thresholds.

Notably, the methodological framework developed for the PFR conjecture proof (Section 3.3.3) might extend to this context, potentially resolving long-standing multiuser information theory challenges. This cross-domain adaptation could bridge discrete and continuous entropy optimization paradigms.

Chapter 4

Inequalities and Links to Additive Combinatorics

In this chapter, we aim to establish formal equivalence relationships between entropic inequalities in information theory and sumset inequalities in additive combinatorics. Unlike previous chapters, which focused on building analogies and parallelism between two communities, this chapter will reveal deeper insights into why such parallelism holds and provide a systematic perspective on these connections.

We begin by establishing a formal equivalence theorem (Theorem 4.1.1) between combinatorial and entropic inequalities in Section 4.1. This equivalence theorem relies heavily on an entropic quantity—the maximal entropic coupling—which is central to building equivalence relationships with sumset theory. The entropic inequalities involving maximal entropic coupling differ slightly in form from analogous entropic inequalities studied by earlier researchers. In some cases, the analogous entropic inequalities are stronger (Remark 4.2.6); in others, even analogous ones fail to imply their equivalent counterparts (Remark 4.2.29), and vice versa.

In Section 4.2, we use Theorem 4.1.1 to establish various inequalities involving maximal entropic coupling, demonstrating its similarity to independent coupling between random variables. We also provide purely information-theoretic arguments to derive properties of maximal entropic couplings. An entropic equality (Lemma 4.2.11), motivated by an analogous combinatorial lemma, has proven repeatedly useful in our arguments. This lemma shares similarities with the copy lemma and aids in establishing non-trivial relationships between maximal entropic couplings.

Finally, in Section 4.3, we prove an information-theoretic characterization of the magnification ratio (Theorem 4.3.3). This result, which serves as a foundational primitive for broader families of sumset inequalities (as evidenced in Ruzsa' s lecture notes [Ruz09b]), lays the groundwork for future efforts to establish comprehensive equivalence theorems between deeper results in sumset theory and information theory.

4.1 Generalized Ruzsa-type equivalence theorem

We state a simple fact below. There exists a trivial equivalence between cardinality inequalities and entropy inequalities through the observation that $\log |A+B| = \max_{p_{XY}} H(X+Y)$, where X takes values in A and Y takes values in B. This equality is achieved by taking a uniform distribution over the support of $A+B$. However, our focus lies on non-trivial versions of equivalence theorems.

In this section, we state the main theorem, which establishes an equivalence between families of entropic inequalities and sumset inequalities using the notion of maximal entropic coupling. This framework yields a broad family of new entropic inequalities, as detailed in the following subsections.

Theorem 4.1.1. *(Generalized Ruzsa-type equivalence theorem)*

Let $(\mathbb{T}, +)$ be a finitely generated torsion-free Abelian group. Let $f_1, \ldots, f_k$ and $g_1, \ldots, g_\ell$ be linear functions on $\mathbb{T}^n$ with integer coefficients, and let $\alpha_1, \ldots, \alpha_k$, $\beta_1, \ldots, \beta_\ell$ be positive real numbers. For the linear function f_i, let $S_i \subseteq [1:n]$ denote the index set of non-zero coefficients. Similarly, for g_i let $T_i \subseteq [1:n]$ denote the *corresponding index set of non-zero coefficients. For any subset $S \subseteq [1:n]$, define $\mathbb{T}_S$ as the projection of $\mathbb{T}^n$ onto the coordinates indexed by S. (So, effectively, f_i and g_i are linear functions on $\mathbb{T}_{S_i}$ and $\mathbb{T}_{T_i}$ respectively). Further, let us assume that $\{S_i\}$ is a pairwise disjoint collection of sets. The following statements are equivalent:*

a) *For any $A_1, A_2, \ldots, A_n$ that are finite subsets of $\mathbb{T}$, we have*

$$\prod_{i=1}^{k} |f_i(A_{S_i})|^{\alpha_i} \leq \prod_{i=1}^{\ell} |g_i(A_{T_i})|^{\beta_i},$$

where $A_S = \otimes_{i \in S} A_i$.

b) *For any $m \in \mathbb{N}$, and for any $\hat{A}_1, \hat{A}_2, \ldots, \hat{A}_n$ that are finite subsets of $\mathbb{T}^m$, we have*

$$\prod_{i=1}^{k} |\hat{f}_i(\hat{A}_{S_i})|^{\alpha_i} \leq \prod_{i=1}^{\ell} |\hat{g}_i(\hat{A}_{T_i})|^{\beta_i},$$

where $\hat{A}_S = \otimes_{i \in S} \hat{A}_i$, and $\hat{f}_i$ (and $\hat{g}_i$) are the natural coordinate-wise extensions of f_i (and g_i) respectively, mapping points in

$$\underbrace{\mathbb{T}^m \times \mathbb{T}^m \times \cdots \times \mathbb{T}^m}_{n \text{ times}} \mapsto \mathbb{T}^m.$$

c) *For every sequence of random variables $(X_1, \ldots, X_n)$, with fixed marginals p_{X_i} and having finite support in $\mathbb{T}$, we have*

$$\sum_{i=1}^{k} \alpha_i \max_{\Pi(X_{S_i})} H(f_i(X_{S_i})) \leq \sum_{i=1}^{\ell} \beta_i \max_{\Pi(X_{T_i})} H(g_i(X_{T_i})),$$

where $\Pi(X_S)$ is collection of joint distributions p_{X_S} that are consistent with the marginals $p_{X_i}, i \in S$.

Remark 4.1.2. It may be worthwhile mentioning a key difference between Theorem 4.1.1 and Theorem 1.4.7. The equivalence in Theorem 1.4.7 follows when the sumset inequalities hold for every G-restricted sumset. On the other hand, most of the inequalities in literature are established for the Minkowski sum of sets, and Theorem 4.1.1 holds under such a situation.

Proof. We will show that $a) \implies b)$, $b) \implies c)$, and $c) \implies a)$. We make a brief remark on the three implications. That $a) \implies b)$ Ruzsa has essentially established in [Ruz09a], and this is where the requirements that the functions be linear and that the ambient group is finitely generated and torsion-free play a crucial role. Now $b) \implies c)$ is a rather standard argument in the information theory literature using the method of types (see Chapter 2 of [CK11]) and Sanov's theorem (we provide an outline in the Appendix for completeness). Finally, $c) \implies a)$ is immediate by taking specific marginal distributions that induce uniform distributions on the support of $f_i(X_{S_i})$ and is where the requirement that S_i be pairwise disjoint plays a role.

$a) \implies b)$: We outline the method used by Ruzsa in [Ruz09a]. By the classification theorem of finitely generated Abelian groups, we know that a torsion-free finitely generated Abelian group is isomorphic to $\mathbb{Z}^d$, for a finite d. We denote t to be a generic element in $\mathbb{T}$, (or equivalently $\mathbb{Z}^d$). Let a linear function with integer coefficients $f : \mathbb{T}^n \mapsto \mathbb{T}$, be defined by $f(t_1, \dots, t_n) = \sum_{i=1}^n a_i t_i$. (In the context of our discussion, the locations of the non-zero values of a_i determine the support of f). Similarly we denote $\mathbf{t} = (t_1, \dots, t_m)$ to be a generic element in $\mathbb{T}^m$. Therefore, we have $\hat{f}(\mathbf{t}_1, \dots, \mathbf{t}_n) = \sum_{i=1}^n a_i \mathbf{t}_i$. Let ψ_q be a linear mapping from $\mathbb{T}^m$ to $\mathbb{T}$ defined as

$$\psi_q(\mathbf{t}) := t_1 + t_2 q + \cdots + t_m q^{m-1}.$$

Observe that, by linearity,

$$\psi_q(\hat{f}(\mathbf{t}_1, .., \mathbf{t}_n)) = \psi_q\left(\sum_{i=1}^n a_i \mathbf{t}_i\right) = f(\psi_q(\mathbf{t}_1), ..., \psi_q(\mathbf{t}_n)). \tag{4.1}$$

Given the finite subsets $\hat{A}_1, \ldots, \hat{A}_n$ of $\mathbb{T}^m$, and the linear functions $f_1, \ldots, f_k$ and $g_1, \ldots, g_\ell$, we can choose a q large enough that $\psi_q(\hat{f}_i(\hat{A}_{S_i}))$ and $\psi_q(\hat{g}_i(\hat{A}_{T_i}))$ are injections. Now set $A_i = \psi_q(\hat{A}_i)$. Therefore we have

$$|\hat{f}_i(\hat{A}_{S_i})| = |\psi_q(\hat{f}_i(\hat{A}_{S_i}))| \overset{(a)}{=} |f_i(\{\psi_q(\hat{A}_k)\}_{k \in S_i})| = |f_i(A_{S_i})|,$$

where (a) follows from (4.1). A similar equality holds for g's as well. With these equalities, we have that $a) \implies b)$.

$b) \implies c)$: We are given a set of marginal distributions $p_{X_1}, \ldots, p_{X_n}$ whose supports are finite subsets of $\mathbb{T}$, say $\mathcal{X}_1, \ldots, \mathcal{X}_n$. Consider a non-negative sequence $\{\delta_m\}$, where $\delta_m \to 0$ and $\sqrt{m} \cdot \delta_m \to \infty$ as $m \to \infty$. For every m, we construct the strongly typical sets $\mathsf{T}_{(m, p_{X_i}, \delta_m)}$, for $1 \leq i \leq n$, where

$$\mathsf{T}_{(m, p_{X_i}, \delta_m)} := \left\{ \mathbf{x} \in \mathcal{X}_i^m : \left| \frac{1}{m} N(a|\mathbf{x}) - p_{X_i}(a) \right| \leq \delta_m \cdot p_{X_i}(a) \text{ for any } a \in \mathcal{X}_i \right\}.$$

Here $N(a|\mathbf{x}) = \sum_{i=1}^m 1_{\{\mathbf{x}_i = a\}}$, the number of occurrences of the symbol a in $\mathbf{x}$. Suppressing dependence on other variables, let $\hat{A}_i = \mathsf{T}_{(m, p_{X_i}, \delta_m)}$ for $1 \leq i \leq n$. Now consider a linear function $f : \mathbb{T}_S \to \mathbb{T}$ and let $\hat{f}$ be the coordinate-wise extension of it to $(\mathbb{T}^m)_S$. Define $Y = f(X_S)$, $S \subseteq [1:n]$, and let $\mathcal{M}_Y$ denote the set of probability distributions of Y induced by all couplings $\Pi(X_S)$ that are consistent with the marginals p_{X_i} for $i \in S$. Let q_Y be the uniform distribution on $\mathcal{Y}$, and by a routine application[1] of Sanov's theorem we obtain that

$$\lim_{m \to \infty} \frac{1}{m} \log \frac{|\hat{f}(\hat{A}_S)|}{|\mathcal{Y}|^m} = \max_{P_Y \in \mathcal{M}_Y} H(Y) - \log |\mathcal{Y}| = \max_{\Pi(X_S)} H(f(X_S)) - \log |\mathcal{Y}|.$$

Therefore, we have

$$\lim_{m \to \infty} \frac{1}{m} \log |\hat{f}(\hat{A}_S)| = \max_{\Pi(X_S)} H(f(X_S)).$$

[1] This is standard in certain information theory circles. For completeness, we outline a proof of the maximum coupling by the discrete Sanov theorem in Appendix B.1 and Appendix B.2.

Thus, the implication $b) \implies c)$ is established.

$c) \implies a)$: This is rather immediate. Since S_i's are pairwise disjoint, let $p_{X_{S_i}}$ induce a uniform distribution on $f(A_{S_i})$ and let p_{X_i} be the induced marginals. Then it is clear that $\max_{\Pi(X_{S_i})} H(f_i(X_{S_i})) = \log |f(A_{S_i})|$ and $\max_{\Pi(X_{T_i})} H(g_i(X_{T_i})) \leq \log |g(A_{T_i})|$ and this completes the proof. □

4.2 Application of generalized Ruzsa-type equivalence theorem

The following corollaries to Theorem 4.1.1 lead to some entropic inequalities. Some of the sumset inequalities in literature are stated using Ruzsa-distance, and the equivalent entropic inequalities can be stated using a similar distance between distributions.

4.2.1 Fundamental maximal entropic coupling inequalities

In this subsection, we introduce the entropic Ruzsa distance, an analogue of the Ruzsa distance between finite sets, which is a fundamental quantity in additive combinatorics. We then establish several inequalities related to maximal entropic coupling via the application of Theorem 4.1.1 to sumset inequalities. Of particular note is Corollary 4.2.7, a novel fundamental entropic inequality for which no stand-alone information-theoretic proof is currently known; this remains a significant open problem.

Definition 4.2.1 (Ruzsa distance between finite sets, [Ruz96])**.** The Ruzsa distance between two finite subsets A, B on an Abelian group $(\mathbb{G}, +)$ is defined as

$$d_R(A, B) := \log \frac{|A - B|}{|A|^{1/2}|B|^{1/2}}.$$

Remark 4.2.2. It is clear that $d_R(A, B) = d_R(B, A)$ and that $d_R(A, A) \geq 0$.

Definition 4.2.3 (Entropic Ruzsa distance)**.** The entropic Ruzsa "distance" between two distributions p_X, p_Y taking values in $(\mathbb{G}, +)$ is defined as

$$d_{HR}(X,Y) := \max_{p_{XY} \in \Pi(p_X,p_Y)} H(X-Y) - \frac{1}{2}H(X) - \frac{1}{2}H(Y),$$

where $\Pi(p_X, p_Y)$ is the set of all couplings with the given marginals.

Remark 4.2.4. The following remarks are worth noting with regard to the entropic Ruzsa distance:

1. As with the abuse of notation in information theory $d_{HR}(X,Y)$ is a function of p_X, p_Y and not of X and Y.

2. Just like the original Ruzsa distance between two sets, we have $d_{HR}(X,Y) \geq 0$ (this follows by observing that when $p_{XY} = p_X p_Y$, we have $H(X-Y) \geq \max\{H(X), H(Y)\}$ as $0 \leq I(X; X-Y) = H(X-Y) - H(Y)$). Further it is immediate that $d_{HR}(X,Y) = d_{HR}(Y,X)$.

3. There is no ordering between $d_{HR}(X,Y)$ and $d_R(A,B)$ where A is the support of p_X and B is the support of p_Y.

   - Consider P_X and P_Y such that they are uniform on sets A and B respectively. Thus for any $P_{XY} \in \Pi(P_X, P_Y)$ we have $H(X-Y) \leq \log|A-B|$ and consequently $d_{HR}(X,Y) \leq d_R(A,B)$ (and the inequality can be strict).

   - Consider a joint distribution p_{XY} that is uniform on $A - B$ and let p_X and p_Y be its induced marginal distributions on sets A and B respectively. Then as $H(X) \leq \log|A|$ and $H(Y) \leq \log|B|$, we have $d_{HR}(X,Y) \geq d_R(A,B)$ (and the inequality can be strict).

4. This definition is different from that of Tao [Tao10], where he defines the similar quantity using independent coupling of p_X and p_Y. An advantage of our definition is that we have a formal equivalence between the two inequalities (one in sumset and one in entropy). Independent of this work, in

[GMT23, Equation 1.4] the authors also defined the same notion of distance and called it the *maximal entropic Ruzsa distance.*

Theorem 4.1.1 immediately implies the following entropic inequalities from the corresponding sumset inequalities.

Corollary 4.2.5. *For any distributions p_X, p_Y, p_Z with finite support on a finitely generated torsion-free Abelian group $(\mathbb{T}, +)$, we have*

$$\begin{aligned} &d_{HR}(X,Z) \le d_{HR}(X,Y) + d_{HR}(Y,Z), \\ \textit{or equivalently}: \ &H(Y) + \max_{\Pi(X,Z)} H(X-Z) \le \max_{\Pi(X,Y)} H(X-Y) + \max_{\Pi(Y,Z)} H(Y-Z). \end{aligned} \tag{4.2}$$

Proof. In [Ruz96], Ruzsa showed that for any finite A, B, C on a finitely generated torsion-free Abelian group $(\mathbb{T}, +)$, we have $d_R(A,C) \le d_R(A,B) + d_R(B,C)$, or equivalently $|B||A-C| \le |A-B||B-C|$. We will obtain the desired inequality by applying Theorem 4.1.1.

Remark 4.2.6. The entropic inequality in (4.2) can also be obtained as a direct consequence of a stronger entropic inequality that was established in [MMT12]. There, it was established that if Y and (X,Z) are independent and taking values in an ambient Abelian group $(\mathbb{G}, +)$, then one has $H(Y) + H(X-Z) \le H(X-Y) + H(Y-Z)$. To see this, observe that $H(Y, X-Z) = H(X-Y, Y-Z) - I(X; Y-Z|X-Z)$, and the requisite inequality is immediate.

□

The following corollary presents a novel entropic inequality derived from a direct application of the Plünnecke—Ruzsa inequality.

Corollary 4.2.7. *For distributions p_X, p_Y, p_Z with finite support on a finitely generated torsion-free Abelian group $(\mathbb{T}, +)$, we have*

$$H(X) + \max_{\Pi(Y,Z)} H(Y+Z) \le \max_{\Pi(X,Y)} H(X+Y) + \max_{\Pi(X,Z)} H(X+Z). \tag{4.3}$$

Proof. In [Ruz96], Ruzsa showed that for any finite A, B, C on a finitely generated torsion-free Abelian group $(\mathbb{T}, +)$, we have

$$|A||B+C| \leq |A+B||A+C|. \tag{4.4}$$

We obtain the desired entropic inequality by applying Theorem 4.1.1. □

Remark 4.2.8. The authors are unaware of a stand-alone information-theoretic proof of the above inequality. Our results in Section 4.3 are a step toward building an information-theoretic counterpart to the sumset arguments used to establish this. When X, Y, and Z are mutually independent, an entropic analog has been established in [Mad08, MMT12]. Note that in this case, by the data-processing inequality, we have $I(Z; X+Y+Z) \leq I(Z; X+Z)$ implying

$$H(X) + H(Y+Z) \leq H(X) + H(X+Y+Z) \leq H(X+Y) + H(X+Z).$$

A relaxation of this proof to the case, when X is independent of (Y, Z), would have yielded (4.3); however, this relaxation does not seem immediate.

4.2.2 Sum-difference inequality

In the following subsection, we will demonstrate the entropic proofs of the Katz-Tao sum-difference inequality and the Ruzsa sum-difference inequality. The key is to construct various copies of random variables and establish a desirable joint distribution with Markovian structures, as formalized in Lemma 4.2.11. This reveals the potential to discover new entropic inequalities by constructing desirable algebraic relationships through suitable copies of random variables.

Katz-Tao sum-difference inequality

The proof of the Katz-Tao sum-difference inequality begins with the combinatorial lemma below, which provides an estimate for the number of tuples satisfying specific algebraic constraints.

Lemma 4.2.9 (Lemma 2.1 of [KT99])**.** *Let A and $B_1, \ldots, B_{n-1}$ be finite sets for some positive n. Let $f_i : A \to B_i$ be a function for all $i \in [1 : n-1]$. Then*

$$|\{(a_1, \ldots, a_n) \in A^n : f_i(a_i) = f_i(a_{i+1}) \quad \text{for all } i \in [1 : n-1]\}| \geq \frac{|A|^n}{\prod_{i=1}^{n-1} |B_i|}.$$

Motivated by this lemma, we will prove an information-theoretic version (which would imply the combinatorial version) and will turn out to be useful in several of our arguments. We will first present a lemma in a more general form.

Lemma 4.2.10. *Suppose the following Markov chain holds:*

$$X_1 \to U_1 \to X_2 \to U_2 \to \cdots \to X_{n-1} \to U_{n-1} \to X_n.$$

Then,

$$\begin{aligned} H(X_1, \ldots, X_n, U_1, \ldots, U_{n-1}) + \sum_{i=1}^{n-1} I(X_i; U_i) \\ + \sum_{i=1}^{n-1} I(U_i; X_{i+1}) = \sum_{i=1}^{n} H(X_i) + \sum_{i=1}^{n-1} H(U_i). \end{aligned}$$

Proof. This lemma is an immediate consequence of the Chain Rule for entropy as follows. Note that the chain rule and the Markov Chain assumption yield

$$\begin{aligned} &H(X_1, \ldots, X_n, U_1, \ldots, U_{n-1}) \\ &= H(X_1) + \sum_{i=1}^{n-1} H(U_i|X_i) + \sum_{i=1}^{n-1} H(X_{i+1}|U_i) \\ &= H(X_1) + \sum_{i=1}^{n-1} \big(H(U_i) - I(U_i; X_i)\big) + \sum_{i=1}^{n-1} \big(H(X_{i+1}) - I(U_i; X_{i+1})\big). \end{aligned}$$

Now, rearranging yields the desired equality. □

As a special case of Lemma 4.2.10 we obtain the following version that is useful in this subsection.

Lemma 4.2.11. *Let $(X_i)_{i=1}^n$ be a sequence of finite-valued random variables (defined on some common probability space) and $(f_i, g_i)_{i=1}^{n-1}$ be a sequence of functions that take a finite set of values in some space $\mathcal{S}$ such that: $f_i(X_i) = g_i(X_{i+1}) (=: U_i)$ and the following Markov chain holds,*

$$X_1 \to U_1 \to X_2 \to U_2 \to \cdots \to X_{n-1} \to U_{n-1} \to X_n.$$

Then,

$$H(X_1, \ldots, X_n) + \sum_{i=1}^{n-1} H(U_i) = \sum_{i=1}^{n} H(X_i).$$

Proof. Note that $H(X_1, \ldots, X_n) = H(X_1, \ldots, X_n, U_1, \ldots, U_{n-1})$ since U_i is determined by X_i (and also by X_{i+1}). Further we also have $I(U_i; X_i) = I(U_i; X_{i+1}) = H(U_i)$ for $1 \leq i \leq n-1$. Hence, the desired consequence follows from Lemma 4.2.10. □

Remark 4.2.12. The following remarks are worth noting:

- Lemma 4.2.11 seems to play a similar role as the copy lemma [ZY98] used in deriving several non-Shannon type inequalities.
- Note that Lemma 4.2.11 will imply Lemma 4.2.9 directly. It suffices to construct random variables $X_1, \ldots, X_n$ with each X_i uniform on A, satisfying $f_i(X_i) = f_i(X_{i+1})$ for all $i \in [1 : n-1]$, such that the joint distribution of $(X_1, \ldots, X_n)$ is supported on the set

  $$C = \{(a_1, \ldots, a_n) \in A^n : f_i(a_i) = f_i(a_{i+1}) \quad \text{for all } i \in [1 : n-1]\}.$$

  Define X_1 to be uniformly distributed on A. Proceeding inductively, for each $k \in [1 : n-1]$, assume X_k is defined and uniform on A. Set $U_k := f_k(X_k)$

and define the conditional distribution of X_{k+1} given the history via

$$\begin{aligned}&\Pr(X_{k+1} = x_{k+1} | X_1 = x_1, \ldots, X_k = x_k, U_1 = u_1, \ldots, U_k = u_k)\\ &\quad := \Pr(X_k = x_{k+1} | U_k = u_k).\end{aligned}$$

This construction ensures three properties:

- The joint distribution of (X_{k+1}, U_k) matches that of (X_k, U_k), preserving uniformity so X_{k+1} is uniform on A;
- The sequence forms a Markov chain $X_1 \to U_1 \to X_2 \to \cdots \to U_{n-1} \to X_n$;
- The equality $f_k(X_k) = U_k = f_k(X_{k+1})$ holds almost surely for each k, confirming $(X_1, \ldots, X_n) \in C$ with probability 1.

From Lemma 4.2.11, that

$$n \log |A| = \sum_{i=1}^{n} H(X_i) = H(X_1, \ldots, X_n) + \sum_{i=1}^{n-1} H(U_i) \leq \log |C| + \sum_{i=1}^{n-1} \log |B_i|.$$

The main intent of the remainder of the section is to demonstrate the role of Lemma 4.2.11 to establish various entropic sum-difference inequalities.

Theorem 4.2.13. *(Katz-Tao sum-difference inequality [KT99])*

For any finite subsets $A, B \subseteq \mathbb{T}$ and $G \subseteq A \times B$,

$$|A \overset{G}{-} B| \leq |A|^{2/3} |B|^{2/3} |A \overset{G}{+} B|^{1/2}.$$

Ruzsa obtained the following entropy version of Katz-Tao sum-difference inequality by applying Theorem 1.4.7 to Theorem 4.2.13 [Ruz09a].

Theorem 4.2.14. *[Ruz09a] Suppose X and Y are random variables with finite support on $(\mathbb{T}, +)$, we have*

$$H(X - Y) \leq \frac{2}{3} H(X) + \frac{2}{3} H(Y) + \frac{1}{2} H(X + Y). \tag{4.5}$$

The theorem derived from Theorem 1.4.7 imposes the requirement that the underlying group be a finitely generated torsion-free Abelian group. This restriction, however, is unnecessary. In the following theorem, we demonstrate through a purely entropic argument that the condition can be relaxed to apply to any Abelian group.

Theorem 4.2.15 (Entropic Katz-Tao inequality, Proposition 3.6 of [TV])**.** *Suppose X and Y are random variables with finite support on an ambient Abelian group $\mathbb{G}$, we have*

$$\frac{1}{2}I(X;X-Y)+\frac{1}{2}I(Y;X-Y)\leq\frac{3}{2}I(X;X+Y)+\frac{3}{2}I(Y;X+Y)+3I(X;Y).$$

The proof of this theorem will be presented in the Appendix for the sake of completeness. The only minor difference between the arguments is using Lemma 4.2.10 instead of the submodularity argument used in [TV].

Remark 4.2.16. The following remarks and acknowledgments may be of interest to the careful reader.

- Initially, the authors were unaware of an entropic argument by Tao and Vu (see [TV]) for the result in Theorem 4.2.15. This connection was brought to our attention by Prof. Ben Green shortly after we uploaded a preliminary version of this work to arXiv.

- The formulation in Theorem 4.2.15 employs mutual information rather than entropies. Consequently, the inequality extends immediately to continuous random variables or those with non-finite support, requiring no additional adjustments.

- An earlier version of this result, framed as in Corollary 4.2.17, was presented at ISIT in July 2023. Following discussions with Lampros Gavalakis and Ioannis Kontoyannis regarding potential continuous-variable generalizations, we adapted our original proof to establish Theorem 4.2.15. It was only afterward that we recognized the overlap with Tao and Vu's unpublished result in [TV].

Corollary 4.2.17. *Suppose X and Y are random variables with finite support on an ambient Abelian group $\mathbb{G}$. We have*

$$H(X-Y) \leq \frac{2}{3}H(X) + \frac{2}{3}H(Y) + \frac{1}{2}H(X+Y). \tag{4.6}$$

Proof. Following the equivalent form of the result in (C.4), we have

$$\begin{aligned} 0 &\geq 5H(X,Y) - 4H(X) - 4H(Y) - 3H(X+Y) + H(X-Y) \\ &\geq 6H(X-Y) - 4H(X) - 4H(Y) - 3H(X+Y). \end{aligned}$$

The second inequality holds if and only if (X, Y) is a function of $X - Y$. □

Ruzsa sum-difference inequality

We can regard the Ruzsa sum-difference inequality as correlated with the Katz-Tao sum-difference inequality. We first recall the Ruzsa sum-difference inequality as follows:

Theorem 4.2.18 (Ruzsa sum-difference inequality, Theorem 5.3 of [Ruz96])**.** *The Ruzsa distance between two finite subsets A, B on an Abelian group $(\mathbb{G}, +)$ satisfies*

$$\begin{aligned} &d_R(A, -B) \leq 3d_R(A, B), \\ &\textit{or equivalently} \;\; |A+B||A||B| \leq |A-B|^3. \end{aligned} \tag{4.7}$$

An entropic analogue of the Ruzsa sum-difference inequality, which requires X and Y to be independent, can be immediately derived as a corollary of Theorem 4.2.15.

Corollary 4.2.19. *[Tao10, Theorem 1.10] If X and Y are independent discrete-valued random variables*

$$H(X-Y) \leq 3H(X+Y) - H(X) - H(Y).$$

[KM14, Theorem 3.7] If X and Y are independent continuous-valued random variables with well-defined differential entropies

$$h(X-Y) \leq 3h(X+Y) - h(X) - h(Y).$$

Proof. The independence between X and Y reduces the inequality established in Theorem 4.2.15 to

$$\frac{1}{2}I(X;X-Y) + \frac{1}{2}I(Y;X-Y) \leq \frac{3}{2}I(X;X+Y) + \frac{3}{2}I(Y;X+Y),$$

which is equivalent to

$$\begin{aligned} H(X-Y) &\leq 3H(X+Y) - \frac{3}{2}H(X+Y|X) - \frac{3}{2}H(X+Y|Y) \\ &\quad + \frac{1}{2}H(X-Y|X) + \frac{1}{2}H(X-Y|Y) \\ &= 3H(X+Y) - H(X) - H(Y). \end{aligned}$$

The proof for the continuous case is identical. □

In the following, we first present the proof of the entropic version of the Ruzsa sum-difference inequality. We then establish a generalized Ruzsa sum-difference inequality in sumset theory, along with its corresponding entropic formulation.

Proposition 4.2.20 (Entropic Ruzsa sum-difference inequality)**.** *Let $X_1, Y_1, X_2, Y_2, X_3, Y_3$ be random variables (on a common probability space) with finite support on an Abelian group $(\mathbb{G}, +)$ such that $X_1 - Y_1 = X_2 - Y_2$ $(=: U)$ and also satisfies that $(X_1, Y_1) \to U \to (X_2, Y_2)$ forms a Markov chain. Further, suppose (X_1, Y_1, X_2, Y_2)*

and (X_3, Y_3) *are independent. Then the following inequality holds:*

$$\begin{aligned} &H(X_1, Y_1) + H(X_2, Y_2) + H(X_3 + Y_3) \\ &\quad \leq H(X_1 - Y_1) + H(X_1, Y_2, X_2 - Y_3, X_3 - Y_1). \end{aligned} \tag{4.8}$$

Remark 4.2.21. This proposition and the proof below are essentially identical to that of Proposition 2.4 in [TV]. The only (minor) difference is that we do not assume that X and Y are independent.

Proof. Since $U = X_1 - Y_1 = X_2 - Y_2$ and $(X_1, Y_1) \to U \to (X_2, Y_2)$ forms a Markov chain, from Lemma 4.2.11 we have

$$H(X_1, Y_1, X_2, Y_2) + H(U) = H(X_1, Y_1) + H(X_2, Y_2) \tag{4.9}$$

We now decompose $H(X_1, Y_1, X_2, Y_2, X_3, Y_3 | X_3 + Y_3)$ in two ways. Firstly, since (X_1, Y_1, X_2, Y_2) and (X_3, Y_3) are independent, we have

$$\begin{aligned} &H(X_1, Y_1, X_2, Y_2, X_3, Y_3 | X_3 + Y_3) \\ &= H(X_1, Y_1, X_2, Y_2) + H(X_3, Y_3 | X_3 + Y_3) \\ &\overset{(a)}{=} H(X_1, Y_1) + H(X_2, Y_2) - H(U) + H(X_3, Y_3 | X_3 + Y_3), \end{aligned}$$

where (a) follows by (4.9).

On the other hand, we have

$$\begin{aligned} &H(X_1, Y_1, X_2, Y_2, X_3, Y_3 | X_3 + Y_3) \\ &= H(X_1, Y_2, X_2 - Y_3, X_3 - Y_1, X_3, Y_3 | X_3 + Y_3) \\ &\leq H(X_1, Y_2, X_2 - Y_3, X_3 - Y_1 | X_3 + Y_3) + H(X_3, Y_3 | X_3 + Y_3) \\ &= H(X_1, Y_2, X_2 - Y_3, X_3 - Y_1, X_3 + Y_3) - H(X_3 + Y_3) + H(X_3, Y_3 | X_3 + Y_3) \\ &= H(X_1, Y_2, X_2 - Y_3, X_3 - Y_1) - H(X_3 + Y_3) + H(X_3, Y_3 | X_3 + Y_3). \end{aligned}$$

The last equality is a consequence of the observation that $(X_1, Y_2, X_2 - Y_3, X_3 - Y_1)$

implies $(X_1, Y_2, X_2+Y_1-(X_3+Y_3))$. However as $X_1+Y_2 = X_2+Y_1$ by assumption, we observe that $H(X_3+Y_3|X_1, Y_2, X_2-Y_3, X_3-Y_1) = 0$ and thus justifying the equality.

By combining these two decompositions, we obtain

$$H(X_1, Y_1) + H(X_2, Y_2) + H(X_3+Y_3) \leq H(U) + H(X_1, Y_2, X_2-Y_3, X_3-Y_1).$$

□

Remark 4.2.22. The arguments here are also motivated by similar arguments in the sumset literature [Gre09] and in Tao's work on a similar inequality in [Tao10].

Corollary 4.2.23. *In addition to the assumptions on $X_1, Y_1, X_2, Y_2, X_3, Y_3$ imposed in Proposition 4.2.20, let us assume that X_1 is independent of Y_1 and X_2 independent of Y_2. Then we have*

$$H(X_2) + H(Y_1) + H(X_3+Y_3) \leq H(X_1-Y_1) + H(X_3-Y_1) + H(X_2-Y_3).$$

Proof. The proof is immediate from Proposition 4.2.20 along with the observation that the assumptions imply $H(X_1, Y_1) = H(X_1) + H(Y_1)$, $H(X_2, Y_2) = H(X_2) + H(Y_2)$, and using the subadditivity of entropy applied to $H(X_1, Y_2, X_2-Y_3, X_3-Y_1)$. □

Remark 4.2.24. Suppose X and Y are independent random variables having finite support on $\mathbb{G}$, and random variables X_3, Y_3 also have finite support on $\mathbb{G}$, then observe that we can always construct a coupling $(X_1, Y_1, X_2, Y_2, X_3, Y_3)$ satisfying the assumptions of Corollary 4.2.23, so that (X_1, Y_1) and (X_2, Y_2) are distributed as (X, Y).

Corollary 4.2.25 (Generalized Ruzsa sum-difference inequality)**.** *Let A, B, C, D be finite subsets of an Abelian group $(\mathbb{G}, +)$. Then the following sumset inequality holds:*

$$|A||B||C+D| \leq |A-B||C-B||A-D|,$$

or equivalently

$$d_R(C,-D) \leq d_R(C,B) + d_R(B,A) + d_R(A,D).$$

Proof. Suppose X is a uniform distribution on A and Y is a uniform distribution on B. Further let X_3, Y_3 be taking values on C, D (respectively) such that $X_3 + Y_3$ is uniform on $C + D$. Let $(X_1, Y_2, X_2, Y_2, X_3, Y_3)$ be the coupling according to Remark 4.2.24 and observe that Corollary 4.2.23 implies that

$$\begin{aligned}
&\log|A| + \log|B| + \log|C+D| \\
&\quad \leq H(U) + H(X_3 - Y_1) + H(X_2 - Y_3) \\
&\quad \leq \log|A-B| + \log|C-B| + \log|A-D|.
\end{aligned}$$

Here, the second inequality used the fact that the entropy of a finite valued random variable is upper bounded by the logarithm of its support size. □

Remark 4.2.26. Setting $C = A$ and $D = B$, we can see that the above is a generalization of Theorem 4.2.18.

Corollary 4.2.27. *For any distributions p_U, p_V, p_X, p_Y with finite support on a finitely generated torsion-free group* $(\mathbb{T}, +)$, *we have*

$$\begin{aligned}
&H(X) + H(Y) + \max_{\Pi(U,V)} H(U+V) \\
&\quad \leq \max_{\Pi(X,Y)} H(X-Y) + \max_{\Pi(X,U)} H(X-U) + \max_{\Pi(V,Y)} H(V-Y).
\end{aligned}$$

Proof. From Corollary 4.2.23, for any finite A, B, C, D on a finitely generated torsion-free Abelian group $(\mathbb{T}, +)$, we have

$$|A||B||C+D| \leq |A-B||A-D||C-B|.$$

We will obtain the desired inequality by applying Theorem 4.1.1. □

Remark 4.2.28. Setting $U = Y$ and $V = X$ from the above result. We will obtain an entropic analog of sum-difference inequality

$$d_{HR}(X, -Y) \leq 3d_{HR}(X, Y),$$

or equivalently

$$H(X) + H(Y) + \max_{\Pi(X,Y)} H(X + Y) \leq 3 \max_{\Pi(X,Y)} H(X - Y).$$

Remark 4.2.29. There seems to be no direct implication between these two statements:

- Suppose X and Y are independent, we have $H(X) + H(Y) + H(X + Y) \leq 3H(X - Y)$. This was the previously considered analogous form of the sum-difference inequality (4.7), established in [Tao10].
- For any p_X, p_Y, we have

  $$H(X) + H(Y) + \max_{\Pi(X,Y)} H(X + Y) \leq 3 \max_{\Pi(X,Y)} H(X - Y).$$

  This entropic inequality can be derived from the combinatorial inequality (4.7).

4.2.3 Connection to the covering lemma

In this subsection, we establish a non-trivial entropic inequality rooted in the conceptual framework of covering lemmas from additive combinatorics. We present an information-theoretic proof inspired by the combinatorial construction via the Green—Ruzsa covering lemma.

First, we recall the Green—Ruzsa covering lemma and a non-trivial sumset inequality derived from it.

Lemma 4.2.30 (Green—Ruzsa covering lemma, [GR06])**.** *Let A and B be additive sets with common ambient group. Then there exists an additive set $X \subseteq B$ with $|X| \leq 2\frac{|A+B|}{|A|} - 1$ such that for every $y \in B$ there are at least $|A|/2$ triplets*

$(x, a, a) \in X \times A \times A$ *with* $x + a—a = y$. *More informally,* $A—A + X$ *covers* B *with multiplicity at least* $|A|/2$. *Furthermore, we have*

$$B - B \subseteq A - A + X - X.$$

Similar claims hold if $\frac{|A+B|}{|A|}$ *is replaced by* $\frac{|A-B|}{|A|}$.

Theorem 4.2.31 ([Ruz96])**.** *Let* A, B *be additive sets in an ambient group. Then*

$$|(B + B) - (B + B)| \leq \frac{|A + B|^4 |A - A|}{|A|^4}.$$

In [TV], an entropic analogue has been presented using a submodularity argument; however, this formulation has no direct relationship with the corresponding sumset theorem.

Theorem 4.2.32 ([TV])**.** *Let* X, Y *be independent discrete random variables taking values in an additive groups* $(\mathbb{G}, +)$*, let* Y_1, Y_2, Y_3, Y_4 *be independent trials of* Y*, and let* X_5, X_6 *be independent trials of* X*. We have*

$$H(Y_1 - Y_2 - Y_3 + Y_4) \leq 4H(X + Y) + H(X_5 - X_6) - 4H(X).$$

In the following section, we establish an entropic formulation of the sumset inequality that directly follows from its combinatorial counterpart. Furthermore, we develop an entropic proof requiring sophisticated joint distribution constructions.

Theorem 4.2.33. *Let* X, Y *be discrete random variables taking values in an additive group* $(\mathbb{G}, +)$*. We have*

$$\max_{\Pi(Y,Y,Y,Y)} H(Y_1 - Y_2 - Y_3 + Y_4) \leq 4 \max_{\Pi(X,Y)} H(X + Y) + \max_{\Pi(X,X)} H(X_5 - X_6) - 4H(X).$$

Remark 4.2.34. By applying Theorem 4.1.1, we immediately establish this inequality for any torsion-free finitely generated Abelian group. However, the subsequent analysis presents a purely information-theoretic proof that offers deeper insight into translating combinatorial constructions into joint distributional frameworks for entropic arguments. Notably, this approach simultaneously extends the inequality's validity to arbitrary Abelian groups.

Entropic proof. We establish a strengthened version of the above statement. For any joint distribution $p(y_1, y_2, y_3, y_4, x_5, x_6)$ satisfying:

- $p_{Y_1} = p_{Y_2} = p_{Y_3} = p_{Y_4} = p_Y$
- $p_{X_5} = p_{X_6} = p_X$

there exists an extended joint distribution $q(y_1, \ldots, y_6, x_3, \ldots, x_6)$ with:

- $q_{Y_i} = p_Y$ for $i = 1, \ldots, 6$
- $q_{X_j} = p_X$ for $j = 3, \ldots, 6$
- $q_{Y_1 - Y_2 - Y_3 + Y_4} = p_{Y_1 - Y_2 - Y_3 + Y_4}$

and the following inequality holds:

$$H(Y_1 - Y_2 - Y_3 + Y_4) \leq 4 \max_{\Pi(X,Y)} H(X+Y) + \max_{\Pi(X,X)} H(X_5 - X_6) - 4H(X).$$

We develop the joint distribution q through sequential construction. First, define the marginal distribution $q(y_1, y_2, y_3, y_4)$ satisfying:

$$\begin{aligned} q_{Y_1 - Y_2 - Y_3 + Y_4} &= p_{Y_1 - Y_2 - Y_3 + Y_4}, \\ q(y_1, y_2, y_3, y_4) &= q(y_1 - y_2, y_3 - y_4) \cdot q(y_1, y_2 | y_1 - y_2) \cdot q(y_3, y_4 | y_3 - y_4), \end{aligned}$$

inducing the Markov chain structure $(Y_1, Y_2) \to Y_1 - Y_2 \to Y_3 - Y_4 \to (Y_3, Y_4)$.

We then extend this to the complete joint distribution through the factorization:

$$\begin{aligned} q(y_1, \ldots, y_6, x_3, \ldots, x_6) =& p(x_5, y_5 | x_3 + y_2) p(x_3) p(y_1, y_2, y_3, y_4) \\ & \cdot p(x_4) p(x_6, y_6 | x_4 + y_4), \end{aligned}$$

with conditional distributions defined by:

$$p(x_5, y_5|x_3 + y_2) = p(x_3, y_2|x_3 + y_2),$$
$$p(x_6, y_6|x_4 + y_4) = p(x_4, y_4|x_4 + y_4).$$

From the algebraic relationships $X_3 + Y_2 = X_5 + Y_5$ and $X_4 + Y_4 = X_6 + Y_6$, we derive the entropy upper bound through the following chain of inequalities:

$$\begin{aligned}
&H(X_3 + Y_1, X_5, Y_5, X_4 + Y_3, X_6, Y_6|Y_1 - Y_2, Y_3 - Y_4)\\
&\overset{(a)}{=} H(X_3 + Y_1, X_5 - X_6, Y_5, X_4 + Y_3, Y_6|Y_1 - Y_2, Y_3 - Y_4)\\
&\leq H(X_3 + Y_1, X_5 - X_6, Y_5, X_4 + Y_3, Y_6|Y_1 - Y_2 - Y_3 + Y_4)\\
&\overset{(b)}{=} H(X_3 + Y_1, X_5 - X_6, Y_5, X_4 + Y_3, Y_6) - H(Y_1 - Y_2 - Y_3 + Y_4)\\
&\leq H(X_3 + Y_1) + H(X_5 - X_6) + H(Y_5) + H(X_4 + Y_3) + H(Y_6)\\
&\quad - H(Y_1 - Y_2 - Y_3 + Y_4).
\end{aligned}$$

The equality (a) follows from the substitutions $X_5 = (X_3+Y_1)-(Y_1-Y_2)-Y_5$ and $X_6 = (X_4+Y_3)-(Y_3-Y_4)-Y_6$, while (b) emerges from the identity $Y_1-Y_2-Y_3+Y_4 = (X_3+Y_1)-Y_5-(X_5-X_6)-(X_4+Y_3)+Y_6$ combined with the chain rule of entropy.

Utilizing the Markovian structure $(X_3 + Y_1, X_5, Y_5) \to Y_1 - Y_2 \to Y_3 - Y_4 \to (X_4 + Y_3, X_6, Y_6)$, we decompose the conditional entropy as:

$$\begin{aligned}
&H(X_3 + Y_1, X_5, Y_5, X_4 + Y_3, X_6, Y_6|Y_1 - Y_2, Y_3 - Y_4)\\
&\quad = H(X_3 + Y_1, X_5, Y_5|Y_1 - Y_2) + H(X_4 + Y_3, X_6, Y_6|Y_3 - Y_4).
\end{aligned}$$

To establish a lower bound, we analyze the first component:

$$\begin{aligned}
&H(X_3 + Y_1, X_5, Y_5|Y_1 - Y_2)\\
&\geq H(X_3 + Y_1, X_5, Y_5|Y_1, Y_2) = H(X_3, X_5, Y_5|Y_1, Y_2)\\
&\overset{(a)}{=} H(X_3, X_5, Y_5|Y_2) = H(X_3, Y_2, X_5, Y_5) - H(Y_2)
\end{aligned}$$

$$
\begin{aligned}
&\overset{(b)}{=} H(X_3, Y_2) + H(X_5, Y_5) - H(X_3 + Y_2) - H(Y_2)\\
&\overset{(c)}{=} H(X_3) + H(Y_2) + H(X_5) + H(Y_5) - H(X_3 + Y_2) - H(Y_2)\\
&= 2H(X) + H(Y) - H(X_3 + Y_2),
\end{aligned}
$$

where (a) follows from the Markov relation $Y_1 \to Y_2 \to (X_3, X_5, Y_5)$, (b) applies Lemma 4.2.11 to the coupling $(X_3, Y_2) \to X_3 + Y_2 \to (X_5, Y_5)$, (c) uses independence $X_3 \perp Y_2$ and $X_5 \perp Y_5$.

A parallel argument for the second component yields $H(X_4 + Y_3, X_6, Y_6|Y_3 - Y_4) \geq 2H(X) + H(Y) - H(X_4 + Y_4)$.

Combining all estimates, we derive the target inequality:

$$
\begin{aligned}
&H(Y_1 - Y_2 - Y_3 + Y_4)\\
&\leq H(X_3 + Y_1) + H(X_5 - X_6) + H(X_4 + Y_3)\\
&\quad + H(X_3 + Y_4) + H(X_4 + Y_4) - 4H(X)\\
&\leq 4 \max_{\Pi(X,Y)} H(X + Y) + \max_{\Pi(X,X)} H(X_5 - X_6) - 4H(X).
\end{aligned}
$$

□

4.3 Entropic formulation of magnification ratio

Even though several equivalence theorems have been established between entropic inequalities and sumset inequalities (e.g., Theorem 1.4.7 and Theorem 4.1.1), there are still a large number of sumset inequalities that do not yet have entropic equivalents, such as the Plünnecke—Ruzsa inequality (though some entropic analogs have been established in [Tao10, KM14]). A combinatorial primitive that frequently occurs in combinatorial proofs is the notion of a magnification ratio (see the lecture notes: [Ruz09b]).

In this section, we introduce an entropic characterization of the magnification ratio through a min-max optimization framework. Subsequent subsections first outline the proof strategy, then analyze properties of the optimal channel in the inner maximization problem, and finally examine the optimal input distribution in the outer minimization problem by leveraging channel properties.

Beyond its potential for deriving new entropic equivalences, this framework may hold independent significance to the combinatorics community.

In the following section, we let $G \subseteq A \times B$ be a finite bipartite graph with no isolated vertices in A or B. For every $S \subseteq A$, let $\mathcal{N}(S) \subseteq B$ denote the set of neighbors of S.

4.3.1 Overview of the proof framework

Definition 4.3.1 (Magnification ratio)**.** The magnification ratio of G from A to B is defined as

$$\mu_{A\to B}(G) = \min_{S\subseteq A, S\neq\emptyset} \frac{|\mathcal{N}(S)|}{|S|}.$$

Definition 4.3.2. (Channel consistent with a bipartite graph) Let $\mathcal{W}$ be the set of all possible channels (or probability transition matrices) from A to B. Given a bipartite graph $G \subseteq A \times B$, we define

$$\mathcal{W}(G) := \{W \in \mathcal{W} : W(Y = b|X = a) = 0 \text{ if } (a,b) \notin G\},$$

to be the set of all channels consistent with the bipartite graph G. Note that $\mathcal{W}(G)$ is a closed and compact set.

In the above, we think of X (taking values in A) as the input and Y (taking values in B) as the output of a channel $W_{Y|X}$. Given an input distribution p_X, we define

$$\lambda_{A\to B}(G; p_X) := \max_{W\in\mathcal{W}(G)} (H(Y) - H(X)).$$

Given a fixed p_X, it is rather immediate that $H(Y)$ is concave in $W_{Y|X}$. Let

$W^*(G;p_X) \in \mathcal{W}(G)$ denote a corresponding optimizer, i.e.

$$W^*(G;p_X) := \operatorname*{arg\,max}_{W \in \mathcal{W}(G)} (H(Y) - H(X)).$$

If the optimizer is a convex set, we define it to be an arbitrary element of this set.

Finally, we define the quantity

$$\lambda_{A\to B}(G) := \min_{p_X} \lambda_{A\to B}(G;p_X) = \min_{p_X} \max_{W\in\mathcal{W}(G)} (H(Y) - H(X)). \tag{4.10}$$

The main result of this section is the following result.

Theorem 4.3.3 (Entropic characterization of the magnification ratio)**.**

$$\log \mu_{A\to B}(G) = \lambda_{A\to B}(G), \text{ *or equivalently,*}$$
$$\log \mu_{A\to B}(G) = \min_{p_X} \max_{W\in\mathcal{W}(G)} (H(Y) - H(X)).$$

Proof. We first establish that $\lambda_{A\to B}(G) \leq \log \mu_{A\to B}(G)$. This direction is rather immediate. Let

$$A^* := \operatorname*{arg\,min}_{S\subseteq A, S\neq\emptyset} \frac{|\mathcal{N}(S)|}{|S|}.$$

So we have $\mu_{A\to B}(G) = \frac{|\mathcal{N}(A^*)|}{|A^*|}$. Let p_X be the uniform distribution on A^*. Then note that

$$\begin{aligned}
\lambda_{A\to B}(G) &\leq \lambda_{A\to B}(G;p_X)\\
&= \max_{W\in\mathcal{W}(G)} (H(Y) - H(X))\\
&= \max_{W\in\mathcal{W}(G)} (H(Y) - \log|A^*|)\\
&\leq \log|\mathcal{N}(A^*)| - \log|A^*| = \log \mu_{A\to B}(G).
\end{aligned}$$

This completes this direction.

We next establish that $\mu_{A\to B}(G) \leq \log \lambda_{A\to B}(G)$. This direction is comparatively rather involved whose main ingredient is the following lemma:

Lemma 4.3.4. *There exists a p_X^*, an optimizer of the outer minimization problem in*

$$\min_{p_X} \max_{W \in \mathcal{W}(G)} (H(Y) - H(X)),$$

such that the inner optimizer $W^*(G; p_X^*)$ induces a uniform output distribution on $\mathcal{N}(S^*)$.

Now, let S^* be the support of p_X^*. If so, one would have

$$\begin{aligned}\lambda_{A\to B}(G) = H(Y) - H(X) = \log|\mathcal{N}(S^*)| - H(X)\\ \geq \log\frac{|\mathcal{N}(S^*)|}{|S^*|} \geq \min_{S\subseteq A, S\neq\emptyset} \frac{|\mathcal{N}(S)|}{|S|} = \mu_{A\to B}(G),\end{aligned}$$

and the proof is complete. □

4.3.2 Properties of the optimal channel for the magnification ratio

In this subsection, we use an optimization framework to characterize key properties of the optimal channel in the inner maximization of Equation (4.10). These results are essential for proving Lemma 4.3.4.

Definition 4.3.5. Given an input distribution p_X and a bipartite graph G, we define an edge $(a, b) \in G$ to be *active* under $W^*(G; p_X)$ if $W^*(b|a) > 0$. Otherwise, it is said to be *inactive*.

Lemma 4.3.6. *Let S be the support of p_X.*

1. *Any maximizer $W^*(G; p_X)$ induces an output distribution, p_Y, such that the support of p_Y is $\mathcal{N}(S)$.*

2. *Let $a_1 \in S$ and $(a_1, b_1), (a_1, b_2)$ be edges in G.*

    (a) *If the edges (a_1, b_1) and (a_1, b_2) are active under $W^*(G; p_X)$, then $p_Y(b_1) = p_Y(b_2)$.*

(b) If (a_1, b_1) is active and (a_1, b_2) is inactive under $W^*(G; p_X)$, then $p_Y(b_1) \leq p_Y(b_2)$.

Proof. The proof of part 1) proceeds by contradiction. Assume that there exists $b_1 \in \mathcal{N}(S)$ such that $p_Y(b_1) = 0$. This implies that there exists $a_1 \in S$, such that $(a_1, b_1) \in G$ and $W^*_{Y|X}(b_1|a_1) = 0$ as $p_Y(b_1) = 0$. Further since $p_X(a_1) > 0$, there exists $b_2 \in \mathcal{N}(S)$ with $(a_1, b_2) \in G$ and $W^*_{Y|X}(b_2|a_1) > 0$. For $\alpha \geq 0$ and sufficiently small, define W_α as follows:

$$W_{Y|X,\alpha}(b|a) = \begin{cases} W^*_{Y|X}(b|a) + \alpha = \alpha, & (a,b) = (a_1, b_1) \\ W^*_{Y|X}(b|a) - \alpha, & (a,b) = (a_1, b_2) \\ W^*_{Y|X}(b|a), & \text{otherwise} \end{cases} \cdot$$

Define $f(\alpha) := H(Y_\alpha) - H(X)$, where p_{Y_α} is the output distribution of p_X under W_α. Note that

$$f'(\alpha) = p_X(a_1) \log\left(\frac{p_Y(b_2) - \alpha p_X(a_1)}{\alpha p_X(a_1)}\right).$$

By assumption, $W_0 = W^*$ is a maximizer of $f(\alpha)$. However, $f'(\alpha) \to +\infty$ as $\alpha \to 0^+$, yielding the requisite contradiction.

We now establish part 2). Note that $H(Y)$ is concave in $\mathcal{W}(G)$ and all constraints in $\mathcal{W}(G)$ are linear under $\mathcal{W}$. Therefore, Karush—Kuhn—Tucker(KKT) conditions are the necessary and sufficient conditions for optimality for $W_{Y|X}$. We rewrite the optimization problem as follows,

$$\begin{aligned} \max_{W \in \mathcal{W}(G)} \quad & H(Y) \\ \text{subject to} \quad & W(b|a) \geq 0, a \in S, (a,b) \in G \\ & \textstyle\sum_b W(b|a) = 1, a \in S \end{aligned} \cdot$$

Define the Lagrangian as follows,

$$\mathcal{L}(W) := H(Y) + \mu_{a,b} W(b|a) + \sum_{a} \lambda_a \left(\sum_{b} W(b|a) - 1 \right).$$

The KKT conditions for optimality implies that for $W \in \mathcal{W}$, $a \in S$, and $(a, b) \in G$, we have

$$\begin{aligned} \frac{\partial \mathcal{L}}{\partial W(b|a)} = -p_X(a)(\log p_Y(b) + 1) + \mu_{a,b} + \lambda_a = 0, \\ \mu_{a,b} W(b|a) = 0, \\ \mu_{a,b} \geq 0. \end{aligned}$$

By solving the above conditions, we have

$$p_Y(b) = \exp\left(-\frac{(\tilde{\lambda}_a - \mu_{a,b})}{p_X(a)} \right),$$

where $\tilde{\lambda}_a = p_X(a) - \lambda_a$.

a) Suppose (a_1, b_1) and (a_1, b_2) are active. This implies that $\mu_{a_1,b_1} = \mu_{a_1,b_2} = 0$, and forces $p_Y(b_1) = p_Y(b_2)$.

b) Suppose (a_1, b_1) is active and (a_1, b_2) is inactive. We have $\mu_{a_1,b_1} = 0$ and $\mu_{a_1,b_2} \geq 0$, this implies $p_Y(b_1) \leq p_Y(b_2)$.

This establishes part 2) of the lemma. □

4.3.3 Properties of the minimal input distribution for the magnification ratio

In this subsection, we first introduce an equivalence relation on the support of the output distribution p_Y induced by the optimal channel in Equation (4.10), where elements within the same equivalence class are assigned identical probabilities. We then demonstrate via an inductive argument that the minimal input distribution p_X in the outer minimization problem of Equation (4.10) possesses only one equivalence class. Consequently, the output distribution of p_X must be uniform, which completes the proof of Lemma 4.3.4.

Equivalence relationship among output distribution elements

Based on p_X (with support S) and the properties of the maximizer $W^*(G; p_X)$, we induce equivalence relationships between the elements in $\mathcal{N}(S)$, and another one between the elements in S. Let p_Y be the distribution on $\mathcal{N}(S)$ induced by P_X and $W^*(G; p_X)$. For $b_1, b_2 \in \mathcal{N}(S)$, we say that $b_1 \sim b_2$ if $p_Y(b_1) = p_Y(b_2)$. We use the above to induce an equivalence relationship on S as follows: For $a_1, a_2 \in S$, we say that $a_1 \sim a_2$ if there exists $b_1, b_2 \in \mathcal{N}(S)$ such that the edges (a_1, b_1) and (a_2, b_2) are active (see Definition 4.3.5) and $b_1 \sim b_2$.

Remark 4.3.7. The main observation is that the active edges in $W^*(G; p_X)$ partition the graph into disconnected components and further there is a one-to-one correspondence between the equivalence classes in $\mathcal{N}(S)$ and the equivalence classes in S. To see this: consider an equivalence class $T \subset \mathcal{N}(S)$ and let $\hat{S} = \{a \in S : (a, b) \text{ is active for some } b \in T\}$. From Lemma 4.3.6, we see that all elements in $\hat{S}$ are equivalent to each other and there is no active edge (a, b) where $a \in \hat{S}$ and $b \notin T$. Further if $a_1 \in S \setminus \hat{S}$, then observe that a_1 is not equivalent to any element in $\hat{S}$.

Total order on equivalence classes

Let $T_1, \ldots, T_k$ be the partition of $\mathcal{N}(S)$ into equivalence classes and let $S_1, \ldots, S_k$ be the corresponding partition of S into equivalence classes. We can define a total order on the equivalence classes of $\mathcal{N}(S)$ as follows: we say $T_{i_1} \geq T_{i_2}$ if $p_Y(b_{i_1}) \geq p_Y(b_{i_2})$. This also induces a total order on the equivalence classes on S. Further, without loss of generality, let us assume that $T_1, \ldots, T_k$ (and correspondingly $S_1, \ldots, S_k$) be monotonically decreasing according to the order defined above.

Proof of Lemma 4.3.4. Let p_X^* be a minimizer of the outer minimization problem in (4.10) and let S^* be its support. Further, let $S_1, \dots, S_k$ be the equivalence classes (that form a partition of S) induced by $W^*(G; p_X^*)$. If $k = 1$, i.e. there is only one equivalence class, then Lemma 4.3.6 implies that p_X^* and $W^*(G; p_X^*)$ induces a uniform output distribution on $\mathcal{N}(S^*)$. Therefore, our goal is to show the existence of an optimizer p_X^* that induces exactly one equivalence class.

Let S_1 and S_2 be the largest and second largest elements under the total ordering mentioned previously. Let $m_\ell = |S_\ell|$, $n_\ell = |T_\ell|$, and for $1 \leq i \leq k$, let $s_{i,j}, 1 \leq j \leq m_i$ be an enumeration of the elements of S_i and $t_{i,j}, 1 \leq j \leq n_i$ be an enumeration of the elements of T_i. Further let $p_{i,j} = p_X^*(s_{i,j})$ and $p_i = \sum_{j=1}^{m_i} p_{i,j}$. Since the induced output probabilities on the elements of T_i are uniform (by the definition of equivalence class), observe that $q_{i,j} := p_Y^*(t_{i,j}) = \frac{p_i}{n_i}$ for all $1 \leq j \leq n_i$. By construction of the equivalence class, $\frac{p_i}{n_i}$, is strictly decreasing in i, $i \in [1:k]$.

By the grouping property of entropy, we have

$$\begin{aligned}
H(X) &= H(p_{1,1}, \dots, p_{1,m_1}, p_{2,1}, \dots, p_{2,m_2}, p_{3,1}, \dots, p_{k,m_k}) \\
&= p_1 H\left(\frac{p_{1,1}}{p_1}, \dots, \frac{p_{1,m_1}}{p_1}\right) + p_2 H\left(\frac{p_{2,1}}{p_2}, \dots, \frac{p_{2,m_2}}{p_2}\right) \\
&\quad + (p_1 + p_2) H\left(\frac{p_1}{p_1 + p_2}, \frac{p_2}{p_1 + p_2}\right) \\
&\quad + H(p_1 + p_2, p_{3,1}, \dots, p_{k,m_k}).
\end{aligned}$$

Similarly,

$$\begin{aligned}
H(Y) &= p_1 H\left(\frac{1}{n_1}, \dots, \frac{1}{n_1}\right) + p_2 H\left(\frac{1}{n_2}, \dots, \frac{1}{n_2}\right) \\
&\quad + (p_1 + p_2) H\left(\frac{p_1}{p_1 + p_2}, \frac{p_2}{p_1 + p_2}\right) \\
&\quad + H(p_1 + p_2, q_{3,1}, \dots, q_{k,n_k}).
\end{aligned}$$

Define a parameterized family of input distributions $\tilde{p}_{X(\alpha)}$ as follows:

$$\tilde{p}_{X(\alpha)}(s_{i,j}) = \begin{cases} \left(1 - \frac{\alpha}{p_1}\right) p_{i,j}, & i = 1 \\ \left(1 + \frac{\alpha}{p_2}\right) p_{i,j}, & i = 2 \\ p_{i,j}, & \text{otherwise.} \end{cases}$$

By Lemma 4.3.9 (presented after this proof) we know that for $\alpha \in [\alpha_{\min}, \alpha_{\max}]$, where

$$\alpha_{\max} := \frac{p_1 n_2 - p_2 n_1}{n_1 + n_2} > 0 > n_2 \left(\frac{p_3}{n_3} - \frac{p_2}{n_2}\right) =: \alpha_{\min},$$

$W^*(G; p_X^*)$ remain the optimal channel. Observe that the induced output distribution is

$$\tilde{p}_{Y(\alpha)}(t_{i,j}) = \begin{cases} \left(1 - \frac{\alpha}{p_1}\right) q_{i,j} = \frac{p_i}{n_i} - \frac{\alpha}{n_i}, & i = 1 \\ \left(1 + \frac{\alpha}{p_2}\right) q_{i,j} = \frac{p_i}{n_i} + \frac{\alpha}{n_i}, & i = 2 \\ q_{i,j}, & \text{otherwise.} \end{cases}$$

This implies $\lambda_{A\to B}(G; \tilde{P}_{X(\alpha)}) = H(\tilde{Y}(\alpha)) - H(\tilde{X}(\alpha))$. Note that

$$\begin{aligned} &\lambda_{A\to B}(G; \tilde{P}_{X(\alpha)}) := H(\tilde{Y}(\alpha)) - H(\tilde{X}(\alpha)) \\ &= (p_1 - \alpha)\left(H\left(\frac{1}{n_1}, \ldots, \frac{1}{n_1}\right) - H\left(\frac{p_{1,1}}{p_1}, \ldots, \frac{p_{1,m_1}}{p_1}\right)\right) \\ &\quad + (p_2 + \alpha)\left(H\left(\frac{1}{n_2}, \ldots, \frac{1}{n_2}\right) - H\left(\frac{p_{2,1}}{p_2}, \ldots, \frac{p_{2,m_2}}{p_2}\right)\right) \\ &\quad + H(p_1 + p_2, q_{3,1}, \ldots, q_{k,n_k}) \\ &\quad - H(p_1 + p_2, p_{3,1}, \ldots, p_{k,m_k}) \\ &= (p_1 - \alpha) f_1 + (p_2 + \alpha) f_2 + H(p_1 + p_2, q_{3,1}, \ldots, q_{k,n_k}) \\ &\quad - H(p_1 + p_2, p_{3,1}, \ldots, p_{k,m_k}), \end{aligned}$$

where

$$\begin{aligned} f_1 &= H\left(\frac{1}{n_1},\ldots,\frac{1}{n_1}\right) - H\left(\frac{p_{1,1}}{p_1},\ldots,\frac{p_{1,m_1}}{p_1}\right), \\ f_2 &= H\left(\frac{1}{n_2},\ldots,\frac{1}{n_2}\right) - H\left(\frac{p_{2,1}}{p_2},\ldots,\frac{p_{2,m_2}}{p_2}\right). \end{aligned}$$

Thus, $\lambda_{A\to B}(G;\tilde{p}_{X(\alpha)})$ is linear in α. At $\alpha=0$, note that $\tilde{p}_{X(\alpha)} = P^*_X$, and hence is a minimizer of $\lambda_{A\to B}(G;\tilde{p}_{X(\alpha)})$. Therefore, this necessitates that $f_1 = f_2$, and for $\alpha \in [\alpha_{\min}, \alpha_{max}]$ we have that $\lambda_{A\to B}(G;\tilde{p}_{X(\alpha)})$ is a constant. Consequently, both $\tilde{p}_{X(\alpha_{\min})}$ and $\tilde{p}_{X(\alpha_{\max})}$ are also minimizers of the outer minimization problem.

If we consider $\tilde{p}_{X(\alpha_{\max})}$ observe that we have $\tilde{p}_{Y(\alpha_{\max})}(t_{1,j}) = \tilde{p}_{Y(\alpha_{\max})}(t_{2,j})$. Therefore $t_{1,j}\sim t_{2,j}$ and this causes T_1 and T_2 to merge into a new equivalence class. Therefore, we have a minimizer of the outer minimization problem with $k-1$ equivalence classes. We can proceed by induction till we get a single equivalence class. Note that the output elements in an equivalence class have the same probability, and the support of the induced output distribution is the neighborhood of the support of p^*_X (see Lemma 4.3.6). Therefore, establishing that p^*_X induces a single equivalence class establishes Lemma 4.3.4.

Alternately, if we consider $\tilde{p}_{X(\alpha_{\min})}$ observe that we have $\tilde{p}_{Y(\alpha_{\max})}(t_{2,j}) = \tilde{p}_{Y(\alpha_{\max})}(t_{3,j})$. Therefore $t_{2,j}\sim t_{3,j}$ and this causes T_2 and T_3 to merge into a new equivalence class. Therefore, again, we have a minimizer of the outer minimization problem with $k-1$ equivalence classes. Proceeding as above, we can reduce to a single equivalence class and establish Lemma 4.3.4. □

Remark 4.3.8. The argument above can be used to infer (with minimal modifications) that any minimizer p^*_X of the outer minimization problem must have $f_i = f_j$, where

$$\begin{aligned} f_i &= H\left(\frac{1}{n_i},\ldots,\frac{1}{n_i}\right) - H\left(\frac{p_{i,1}}{p_i},\ldots,\frac{p_{i,m_i}}{p_i}\right), \\ f_j &= H\left(\frac{1}{n_j},\ldots,\frac{1}{n_j}\right) - H\left(\frac{p_{j,1}}{p_j},\ldots,\frac{p_{j,m_j}}{p_j}\right). \end{aligned}$$

Further $\lambda_{A\to B}(G;\tilde{p}_{X^*}) = \sum_{i=1}^{k} p_i f_i$. Since all f_i's are identical, we have $\lambda_{A\to B}(G) = f_1$. Therefore, the restriction of $\tilde{p}_{X^*}$ to the first equivalence class is also a minimizer of the outer minimization problem, and observe that the induced output is uniform in T_1.

Lemma 4.3.9 (Reweighting input equivalence class probabilities preserves the optimality of the channel)**.** *Let the partition $S_1 \geq S_2 \geq \cdots \geq S_k$ (of S, the support of p_X) be the monotonically decreasing order of equivalence classes induced by $W^*(G;p_X)$. Define a parameterized family of input distributions $\tilde{p}_{X(\alpha)}$ as follows*

$$\tilde{p}_{X(\alpha)}(s_{i,j}) = \begin{cases} \left(1 - \frac{\alpha}{p_1}\right) p_{i,j}, & i = 1 \\ \left(1 + \frac{\alpha}{p_2}\right) p_{i,j}, & i = 2 \\ p_{i,j}, & \textit{otherwise.} \end{cases}$$

Then $W^*(G;p_X)$ continues to be an optimal channel under $\tilde{p}_{X(\alpha)}$ for $\alpha \in [\alpha_{\min}, \alpha_{\max}]$, where

$$\alpha_{\max} := \frac{p_1 n_2 - p_2 n_1}{n_1 + n_2} \geq 0 \geq n_2 \left(\frac{p_3}{n_3} - \frac{p_2}{n_2}\right) =: \alpha_{\min}.$$

Proof. We recall the KKT conditions (from the proof of Lemma 4.3.6), which are necessary and sufficient for the inner optimization problem to verify the optimality of $W^*(G;p_X)$. The KKT conditions for optimality states that for $a \in S$ and $(a,b) \in G$, the optimizers must satisfy

$$\begin{aligned} -p_X(a)(\log p_Y(b) + 1) + \mu_{a,b} + \lambda_a &= 0, \\ \mu_{a,b} W(b|a) &= 0, \\ \mu_{a,b} &\geq 0, \end{aligned}$$

for some dual parameters $\{\lambda_a\}$ and $\{\mu_{a,b}\}$.

For $a \in S$ and $(a,b) \in G$, let $\lambda_a, \mu_{a,b}$ denote the dual parameters that certify the optimality of $W^*(G; p_X)$ for p_X^*. Now define

$$\lambda_a(\alpha) = \begin{cases} \left(1 - \frac{\alpha}{p_1}\right)\left(\lambda_a + p_X^*(a)\log\left(1 - \frac{\alpha}{p_1}\right)\right) & a \in S_1 \\ \left(1 + \frac{\alpha}{p_2}\right)\left(\lambda_a + p_X^*(a)\log\left(1 + \frac{\alpha}{p_2}\right)\right) & a \in S_2 \\ \lambda_a, & \text{otherwise.} \end{cases}$$

Using the channel $W^*(G; p_X)$, the induced output distribution of $\tilde{p}_{X(\alpha)}$, is given by

$$\tilde{p}_{Y(\alpha)}(t_{i,j}) = \begin{cases} \left(1 - \frac{\alpha}{p_1}\right) q_{i,j} = \frac{p_i}{n_i} - \frac{\alpha}{n_i}, & i = 1 \\ \left(1 + \frac{\alpha}{p_2}\right) q_{i,j} = \frac{p_i}{n_i} + \frac{\alpha}{n_i}, & i = 2 \\ q_{i,j} = \frac{p_i}{n_i}, & \text{otherwise.} \end{cases}$$

Observe that if (a, b_a) is an active edge under $W^*(G; p_X)$, then note that $p_{Y(\alpha)}(b_a)$ only depends on a, or rather only on the equivalence class that a (or equivalently b_a) belongs to. Define

$$\mu_{a,b}(\alpha) = p_{X(\alpha)}(a)(\log p_{Y(\alpha)}(b) - \log p_{Y(\alpha)}(b_a)).$$

Note that $\mu_{a,b}(\alpha) \geq 0$ as long as

$$1 \geq \frac{p_1}{n_1} - \frac{\alpha}{n_1} \geq \frac{p_2}{n_2} + \frac{\alpha}{n_2} \geq \frac{p_3}{n_3},$$

or the ordering of equivalence classes remains unchanged. (Note that if $k = 2$, i.e. there are only two partitions, then we set $p_3 = 0$.) This is equivalent to $\alpha \geq \max\{n_2\left(\frac{p_3}{n_3} - \frac{p_2}{n_2}\right), p_1 - n_1\}$ and $\alpha \leq \frac{p_1 n_2 - p_2 n_1}{n_1 + n_2}$. Since $n_1 \geq 1$, and by our ordering of equivalence classes, we have $\frac{p_1}{n_1} \geq \frac{p_2}{n_2} \geq \frac{p_3}{n_3}$; a moments reflection implies the following:

$$\frac{p_1 n_2 - p_2 n_1}{n_1 + n_2} \geq 0 \geq n_2\left(\frac{p_3}{n_3} - \frac{p_2}{n_2}\right) \geq p_1 - n_1.$$

Therefore $\alpha \in [\alpha_{\min}, \alpha_{\max}]$ preserves the ordering of equivalence classes. A simple substitution shows that the dual variables $\lambda_a(\alpha)$ and $\mu_{a,b}(\alpha)$ defined above serve as witnesses for the optimality of $W^*(G; p_X)$ for $p_{X(\alpha)}$. This completes the proof of the lemma. □

Remark 4.3.10. The idea of the above proof is the following. The reweighting of the input classes preserves the uniformity of the output probabilities within each equivalence class and the ordering between the output probabilities between equivalence classes. This happens to be the KKT conditions for the maximality of the channel. The limits are achieved with the output probability in an equivalence class equals the value in its adjacent class. At this point, there are potentially multiple optimizers for the inner problem, and the active and inactive edges could be rearranged as you change α further.

4.4 Discussion

In this chapter, we introduce a framework for bridging concepts in information theory and additive combinatorics through formal equivalence relationships (Theorem 4.1.1). Key contributions include new information-theoretic tools (Lemma 4.2.11), motivated by a combinatorial lemma (Lemma 4.2.9), and an entropic characterization of the magnification ratio. These advances aim to foster interdisciplinary dialogue between the two fields.

Despite progress, numerous combinatorial results still lack entropic counterparts. A notable example arises in [Pet12], where the following elegant lemma necessitates an entropic equivalence in Proposition 1.4.5. By choosing S to attain the magnification ratio in the bipartite graph between A and $A + B$, the Plünnecke—Ruzsa inequality emerges naturally through induction:

Theorem 4.4.1. *Let A and B be finite sets in an Abelian group $(\mathbb{G}, +)$. If*

$|A+B| \le \alpha|A|$, *there exists* $S \subseteq A$ *such that*

$$|S+kB| \le \alpha^k |S| \quad \textit{for all positive } k,$$

and consequently,

$$|kB-\ell B| \le \alpha^{k+\ell}|A| \quad \textit{for all positive } k+\ell > 1.$$

The primary challenge lies in extending Sanov’s theorem to address subset existence problems. While partitioning strong typical sets into conditional typical sets offers a potential pathway, substantial technical hurdles must be overcome to fully connect combinatorial insights with information-theoretic methods.

Appendix A

Preliminary inequalities for Section 3.3.3

A.1 Estimates related to PFR conjecture

In this appendix, we present some preliminary inequalities that were (essentially) established in [GGMT23].

Lemma A.1.1 (Adapted from [GGMT23], Lemma 5.2)**.** *Suppose X is independent of Y, Z, we have*

$$d(X, Y|Z) \leq d(X, Y) + \frac{1}{2} I(Y; Z).$$

Proof. We have

$$\begin{aligned} d(X, Y|Z) &= H(X + Y|Z) - \frac{1}{2} H(X) - \frac{1}{2} H(Y|Z) \\ &= d(X, Y) + \frac{1}{2} I(Y; Z) - I(X + Y; Z). \end{aligned}$$

□

Lemma A.1.2 ([GGMT23], Lemma A.2)**.** *Let A, B, S be jointly distributed on $\mathbb{G}$.*

$$d(A, B|Z, S) \leq 3I(A; B|S) + 2H(A - B|S) - H(A|S) - H(B|S).$$

Proof. Let $Z = A - B$. Given S, construct two "copies" of (A, B), labeled as (A_1, B_1) and (A_2, B_2) such that $A_1 - B_1 = Z = A_2 - B_2$ and their joint law satisfies $p_S p_{Z|S} p_{A_1,B_1|Z,S} p_{A_2,B_2|Z,S}$.

We have (from sub-modularity)

$$\begin{aligned}
&H(A_1 + B_2, A_1, B_1|S) + H(A_1 + B_2|S) \\
&\leq H(A_1 + B_2, A_1|S) + H(A_1 + B_2, B_1|S) \\
&= H(A_1, B_2|S) + H(A_2 + B_1, B_1|S) \\
&= H(A_1, B_2|S) + H(A_2, B_1|S) \qquad \text{(A.1)}
\end{aligned}$$

Copy lemma, Lemma 4.2.11, yields

$$H(A_1, A_2, B_1, B_2|S) + H(A_1 - B_1|S) = H(A_1, B_1|S) + H(A_2, B_2|S)$$

However we also have that $(A_1 + B_2, A_1, B_1)$ determines and is determined by A_1, B_1, A_2, B_2. Therefore

$$\begin{aligned}
&H(A_1, B_1|S) + H(A_2, B_2|S) \\
&= H(A_1, B_1, A_2, B_2|S) + H(A_1 - B_1|S) \\
&= H(A_1 + B_2, A_1, B_1|S) + H(A_1 - B_1|S) \\
&\overset{\text{(A.1)}}{\leq} H(A_1, B_2|S) + H(A_2, B_1|S) - H(A_1 + B_2|S) + H(A_1 - B_1|S)
\end{aligned}$$

Rearranging yields,

$$H(A_1+B_2|S) \leq H(A_1, B_2|S)+H(A_2, B_1|S)-H(A_1, B_1|S)-H(A_2, B_2|S)+H(A_1-B_1|S). \qquad \text{(A.2)}$$

We also have $H(A|A - B, S) = H(B|A - B, S) = H(A, B|S) - H(A - B|S)$. Therefore,

$$H(A_1 + B_2|S) - \frac{1}{2}H(A_1|A_1 - B_1, S) - \frac{1}{2}H(B_2|A_2 - B_2, S)$$

$$= H(A_1 + B_2|S) - \frac{1}{2}H(A_1, B_1|S) - \frac{1}{2}H(A_2, B_2|S) + H(A - B|S)$$
$$\overset{(A.2)}{\leq} H(A_1, B_2|S) + H(A_2, B_1|S) - \frac{3}{2}H(A_1, B_1|S) - \frac{3}{2}H(A_2, B_2|S) + 2H(A - B|S)$$
$$\leq 3I(A; B|S) + 2H(A - B|S) - H(A|S) - H(B|S). \tag{A.3}$$

From definition

$$\begin{aligned} d(A, B|Z, S) &= H(A_1 + B_2|Z, S) - \frac{1}{2}H(A_1|Z, S) - \frac{1}{2}H(B_2|Z, S) \\ &\leq H(A_1 + B_2|S) - \frac{1}{2}H(A_1|Z, S) - \frac{1}{2}H(B_2|Z, S). \end{aligned}$$

Equation (A.3) completes the proof. □

Corollary A.1.3 (From the arguments in [GGMT23], Lemma 7.2)**.** *Let (S, T_1, T_2, T_3) be jointly distributed on a group of characteristic two such that $T_1 + T_2 + T_3 = 0$. Then, we have*

$$\begin{aligned} &d(T_2, T_3|T_1, S) + d(T_3, T_1|T_2, S) + d(T_1, T_2|T_3, S) \\ &\qquad \leq 3I(T_1; T_2|S) + 3I(T_2; T_3|S) + 3I(T_1; T_3|S) \end{aligned}$$

Proof. Under the characteristic two assumption, Lemma A.1.2 yields

$$\begin{aligned} d(T_2, T_3|T_1, S) &\leq 3I(T_1; T_2|S) + 2H(T_2 - T_3|S) - H(T_2|S) - H(T_3|S) \\ &= 3I(T_1; T_2|S) + 2H(T_2 + T_3|S) - H(T_2|S) - H(T_3|S) \\ &= 3I(T_1; T_2|S) + 2H(T_1|S) - H(T_2|S) - H(T_3|S). \end{aligned}$$

The corollary follows by adding the cyclic shifts of this inequality. □

Lemma A.1.4 (Adapted from [GGMT23], Lemma 7.1 and Section 7)**.** *Let X and Y be two independent random variables defined on a field of characteristic two. Let (X_A, Y_A) and (X_B, Y_B) be independent copies of (X, Y). Let $S = (X_A + X_B) + (Y_A + Y_B)$. Let $T_1 = (X_A + X_B)$, $T_2 = (X_B + Y_B)$, $T_3 = X_A + Y_B$. Let (X^0, Y^0) be independent of the other random variables.*

1. *Family 1: The following inequalities hold*

$$\begin{aligned} d(X^0, T_2|S) - d(X^0, X) &\leq \frac{1}{2}H(S) - \frac{1}{2}H(X), \\ d(X^0, T_3|S) - d(X^0, X) &\leq \frac{1}{2}H(S) - \frac{1}{2}H(X), \\ d(Y^0, T_2|S) - d(Y^0, Y) &\leq \frac{1}{2}H(S) - \frac{1}{2}H(Y), \\ d(Y^0, T_3|S) - d(Y^0, Y) &\leq \frac{1}{2}H(S) - \frac{1}{2}H(Y), \end{aligned}$$

2. *Family 2: The following inequalities hold*

$$\begin{aligned} d(X^0, T_1|S) - d(X^0, X) &\leq \frac{1}{2}H(S) + \frac{1}{2}H(X_A + X_B) - \frac{1}{2}H(Y_A + Y_B) - \frac{1}{2}H(X) \\ d(Y^0, T_1|S) - d(Y^0, Y) &\leq \frac{1}{2}H(S) + \frac{1}{2}H(Y_A + Y_B) - \frac{1}{2}H(X_A + X_B) - \frac{1}{2}H(Y). \end{aligned}$$

3. *Family 3: The following inequalities hold*

$$\begin{aligned} d(X^0, X_A|X_A + Y_B) - d(X^0, X_A) &\leq \frac{1}{2}H(X_A + Y_B) - \frac{1}{2}H(Y_B), \\ d(Y^0, Y_A|Y_A + X_B) - d(Y^0, Y_A) &\leq \frac{1}{2}H(Y_A + X_B) - \frac{1}{2}H(X_B). \end{aligned}$$

4. *The following equality holds:*

$$\begin{aligned} &d(X_A + Y_B, Y_A + X_B) + d(X_A, Y_A|X_A + Y_B, Y_A + X_B) \\ &\quad + I(X_A + Y_A; Y_A + X_B|X_A + X_B + Y_A + Y_B) \\ &\quad = 2d(X, Y). \end{aligned} \tag{A.4}$$

Proof. Family 1: The proofs of the inequalities in the first family are similar. We provide the details of the first. It suffices to show that

$$H(X_B + Y_B + X^0|S) - \frac{1}{2}H(X_B + Y_B|S) - H(X + X^0) + \frac{1}{2}H(X) \leq \frac{1}{2}H(S) - \frac{1}{2}H(X).$$

This is equivalent to showing that

$$H(X_B + Y_B + X^0|S) - \frac{1}{2}H(X_B + Y_B, S) \le I(X^0; X + X^0).$$

Observe that

$$H(X_B + Y_B, S) = H(X_A + Y_A) + H(X_B + Y_B) = 2H(X_B + Y_B).$$

Therefore, it suffices to show that

$$H(X_B + Y_B + X^0|S) - H(X_B + Y_B) \le I(X^0; X + X^0)$$

Note that

$$\begin{aligned} H(X_B + Y_B + X^0|S) - H(X_B + Y_B) &\le H(X_B + Y_B + X_0) - H(X_B + Y_B) \\ &= I(X_0; X_B + Y_B + X_0) \le I(X_0; X + X_0). \end{aligned}$$

The last inequality is a consequence of the data-processing inequality as Y_B is independent of (X_0, X_B). This establishes the first inequality.

Family 2: The proofs of the inequalities in the second family are similar. We only prove the first one. We wish to show that

$$\begin{aligned} H(T_1 + X^0|S) - \frac{1}{2}H(T_1|S) - H(X + X^0) + \frac{1}{2}H(X) \\ \le \frac{1}{2}H(S) + \frac{1}{2}H(X_A + X_B) - \frac{1}{2}H(Y_A + Y_B) - \frac{1}{2}H(X), \end{aligned}$$

or equivalently

$$H(T_1 + X^0|S) \le \frac{1}{2}H(T_1, S) + \frac{1}{2}H(X_A + X_B) - \frac{1}{2}H(Y_A + Y_B) + H(X + X^0) - H(X),$$

Since $H(T_1, S) + H(S) = H(X_A + X_B, Y_A + Y_B) = H(X_A + X_B) + H(Y_A + Y_B)$

(the random variables are independent),we wish to show that

$$H(T_1 + X^0|S) \le H(X_A + X_B) + H(X + X^0) - H(X).$$

Therefore, it suffices to prove the stronger inequality that

$$H(T_1 + X^0) \le H(X_A + X_B) + H(X + X^0) - H(X),$$

or equivalently

$$I(X^0; X_A + X_B + X^0) \le I(X^0; X_A + X^0) = I(X^0; X + X^0).$$

This inequality is a consequence of the data-processing inequality as X_B is independent of (X_0, X_A). This establishes the desired inequality.

Family 3: The proofs of the inequalities in the third family are similar. We only prove the first one. We wish to show that

$$\begin{aligned}&H(X^0 + X_A|X_A + Y_B) - \frac{1}{2}H(X_A|X_A + Y_B) - H(X^0 + X_A) + \frac{1}{2}H(X_A)\\ &\quad\le \frac{1}{2}H(X_A + Y_B) - \frac{1}{2}H(Y_B).\end{aligned}$$

This is equivalent to showing that

$$\begin{aligned}&H(X^0 + X_A|X_A + Y_B) - H(X^0 + X_A)\\ &\quad\le \frac{1}{2}H(X_A|X_A + Y_B) + \frac{1}{2}H(X_A + Y_B) - \frac{1}{2}H(X_A) - \frac{1}{2}H(Y_B).\end{aligned}$$

Note that $H(X_A|X_A + Y_B) + H(X_A + Y_B) = H(X_A, Y_B) = H(X_A) + H(Y_B)$. Therefore, the right-hand-side of the desired inequality is zero. On the other hand, $H(X^0 + X_A|X_A + Y_B) \le H(X^0 + X_A)$ is immediate, establishing the desired inequality.

Final Equality: Observe that

$$\begin{aligned}
&d(X_A+Y_B, Y_A+X_B) + d(X_A, Y_A|X_A+Y_B, Y_A+X_B)\\
&\qquad + I(X_A+Y_A; Y_A+X_B|X_A+X_B+Y_A+Y_B)\\
&= H(X_A+Y_B+Y_A+X_B) - \frac{1}{2}H(X_A+Y_B) - \frac{1}{2}H(Y_A+X_B)\\
&\qquad + H(X_A+X_B|X_A+Y_B, Y_A+X_B) - \frac{1}{2}H(X_A|X_A+Y_B) - \frac{1}{2}H(Y_A|Y_A+X_B)\\
&\qquad + I(X_A+Y_A; Y_A+X_B|X_A+X_B+Y_A+Y_B)\\
&= H(X_A+Y_B+Y_A+X_B) + H(X_A+X_B|X_A+Y_B, Y_A+X_B)\\
&\qquad - \frac{1}{2}H(X_A, X_A+Y_B) - \frac{1}{2}H(Y_A, Y_A+X_B)\\
&\qquad + H(X_A+Y_A|X_A+X_B+Y_A+Y_B) - H(X_A+Y_A|X_A+Y_B, Y_A+X_B)\\
&= H(X_A+Y_A, X_B+Y_B) - \frac{1}{2}H(X_A, Y_B) - \frac{1}{2}H(Y_A, X_B) = 2d(X,Y).
\end{aligned}$$

□

Appendix B

Type counting argument

B.1 Discrete Sanov Theorem

For completeness, we will prove the finite alphabet version of Sanov's theorem with most of the arguments borrowed from Chapter 2 of [DZ98].

Let Σ be a finite space. For the purposes of this section, let $\Sigma = \{1, 2, \ldots, M\}$, or equivalently $\Sigma = [1 : M]$. Denote $\mathcal{M}(\Sigma)$ as the set of probability mass functions on Σ. For a given probability mass function μ, let us denote $\Sigma_\mu = \{i : \mu(i) > 0\}$ to be the support of μ. Thus $\Sigma_\mu \subseteq \Sigma$.

Given a sequence $y^n \in \Sigma^n$, we define the **type** of y^n, $\mathsf{T}_{y^n} \in \mathcal{M}(\Sigma)$, as the probability mass function given by

$$\mathsf{T}_{y^n}(i) := \frac{1}{n}\sum_{k=1}^{n} 1_{\{y_k = i\}}, \quad 1 \le i \le M.$$

It is, equivalently, the empirical measure induced by the sequence y^n. Let $\mathsf{T}_n \subset \mathcal{M}(\Sigma)$ denote the collection of all types, i.e.

$$\mathsf{T}_n := \{\mu : \mu = \mathsf{T}_{y^n} \text{ for some } y^n \in \Sigma^n\}.$$

Lemma B.1.1. *The following statements hold:*

(i) $|\mathsf{T}_n| = \binom{n+M-1}{M-1} \le (n+1)^{M-1}$.

(*ii*) *For any* $\mu \in \mathcal{M}(\Sigma)$, *there exists* $\nu \in \mathsf{T}_n$ *such that* $|\mu(i) - \nu(i)| \leq \frac{1}{n}$. *Consequently,* $d_{TV}(\mu, \mathsf{T}_n) \leq \frac{M}{2n}$, *where* $d_{TV}(\mu, \mathsf{T}_n) := \min_{\nu \in \mathsf{T}_n} d_{TV}(\mu, \nu)$ *and* $d_{TV}(\mu, \nu) = \frac{1}{2}\sum_{i=1}^{M} |\mu(i) - \nu(i)|$. *Further* $\Sigma_\nu \subseteq \Sigma_\mu$.

Proof. Note that every $\mu \in \mathsf{T}_n$ is in one-to-one correspondence with non-negative integer sequences $\{a_1, \ldots, a_M\}$ such that $\sum_{i=1}^{M} a_i = n$. The count of the latter is a problem in elementary combinatorics, and the count is essentially a bijection to choosing the identities of $M-1$ dividers from $n+M-1$ locations. Note that $\binom{n+M-1}{M-1} \leq (n+1)^{M-1}$ is an equality for $M = 1$ and for $M > 1$, we have

$$\binom{n+M-1}{M-1} = \prod_{k=1}^{M-1} \frac{n+k}{k} \leq \prod_{k=1}^{M-1} (n+1) = (n+1)^{M-1},$$

as $\frac{n+k}{k} \leq n+1, \forall k \geq 1$.

Given a $\mu \in \mathcal{M}(\Sigma)$, let us define two non-negative integer sequences according to

$$k_l(i) = \lfloor n\mu(i) \rfloor, \quad k_u(i) = \lceil n\mu(i) \rceil, \quad 1 \leq i \leq M.$$

The following estimates are clear:

$$n\mu(i) - 1 \leq k_l(i) \leq n\mu(i) \leq k_u(i) \leq n\mu(i) + 1.$$

Summing up over i, we obtain

$$n - M \leq \sum_{i=1}^{M} k_l(i) \leq n \leq \sum_{i=1}^{M} k_u(i) \leq n + M.$$

Therefore, we can find a sequence of non-negative integers, $k_{int}(i)$ such that $k_l(i) \leq k_{int}(i) \leq k_u(i)$ such that $\sum_{i=1}^{M} k_{int}(i) = n$. This is essentially like a discrete intermediate-value-theorem, which a greedy algorithm (starting from k_l and increasing value at each coordinate by one while obeying the bounds) can easily establish. Now define $\nu(i) = \frac{k_{int}(i)}{n}$. Note that $\nu \in \mathsf{T}_n$. We know that

$$n\mu(i) - 1 \leq k_l(i) \leq k_{int}(i) \leq k_u(i) \leq n\mu(i) + 1.$$

Therefore $|\nu(i) - \mu(i)| \leq \frac{1}{n}$ and $d_{TV}(\mu, \nu) \leq \frac{M}{2n}$. If $\mu(i) = 0$, then observe that $k_u(i) = 0$ implying $k_{int}(i) = 0$ and hence $\nu(i) = 0$. This establishes the relationship between the supports. □

For $\nu \in \mathsf{T}_n$, we define the **type-class** by $\mathcal{Y}_n(\nu) = \{y^n \in \Sigma^n : \mathsf{T}_{y^n} = \nu\}$. Note that $\mathcal{Y}_n(\nu)$ is the collection of permutations of a generic string y^n whose empirical measure is ν, and the cardinality of $\mathcal{Y}_n(\nu)$ is the multinomial co-efficient associated with the empirical counts, i.e. $|\mathcal{Y}_n(\nu)| = \binom{n}{n\nu(1), n\nu(2), \cdots, n\nu(M)}$.

Let $\mathbb{P}_\mu$ be the probability law associated with an infinite sequence of i.i.d. random variables $Y_1, Y_2, \ldots$, distributed according to $\mu \in \mathcal{M}(\Sigma)$.

In the following:

$$H(\nu) = \sum_{i=1}^{N} -\nu(i) \log_2 \nu(i),$$
$$D(\nu\|\mu) = \sum_{i=1}^{N} \nu(i) \log_2 \frac{\nu(i)}{\mu(i)},$$

with the convention: $0 \log_2 0 = 0$, and if $\nu \not\ll \mu$, then $D(\nu\|\mu) = \infty$.

Lemma B.1.2.

$$\mathbb{P}_\mu[(Y_1, Y_2, \ldots, Y_n) = y^n] = 2^{-n(H(\mathsf{T}_{y^n}) + D(\mathsf{T}_{y^n}\|\mu))}.$$

Proof. Note that

$$\begin{aligned}\mathbb{P}_\mu[(Y_1, Y_2, \ldots, Y_n) = y^n] &= \prod_{i=1}^{M} (\mu(i))^{n\mathsf{T}_{y^n}(i)} \\ &= 2^{-n\left(-\sum_{i=1}^{M} \mathsf{T}_{y^n}(i) \log_2 \mathsf{T}_{y^n}(i) + \sum_{i=1}^{M} \mathsf{T}_{y^n}(i) \log_2 \frac{\mathsf{T}_{y^n}(i)}{\mu(i)}\right)} \\ &= 2^{-n(H(\mathsf{T}_{y^n}) + D(\mathsf{T}_{y^n}\|\mu))}.\end{aligned}$$

□

Lemma B.1.3. *Let $m, l \in \mathbb{N}$. Then $\frac{m!}{l!} \leq l^{m-l}$.*

Proof. If $m > l$, then $\frac{m!}{l!} = \prod_{k=l+1}^{m} k \geq l^{m-l}$. If $m < l$, then $\frac{m!}{l!} = \prod_{k=m+1}^{l} \frac{1}{k} \geq \frac{1}{l}^{l-m} = l^{m-l}$. Finally, equality holds for $m = l$. □

Corollary B.1.4. *For $\gamma, \nu \in \mathsf{T}_n$,*

$$\frac{|\mathcal{Y}_n(\nu)|}{|\mathcal{Y}_n(\gamma)|} \geq 2^{n(H(\nu)-D(\gamma\|\nu)-H(\gamma))}.$$

Proof. Note that

$$\frac{|\mathcal{Y}_n(\nu)|}{|\mathcal{Y}_n(\gamma)|} = \frac{\binom{n}{n\nu(1),n\nu(2),\cdots,n\nu(M)}}{\binom{n}{n\gamma(1),n\gamma(2),\cdots,n\gamma(M)}} = \prod_{i=1}^{M} \frac{(n\gamma(i))!}{(n\nu(i))!}$$
$$\geq \prod_{i=1}^{M} (n\nu(i))^{n(\gamma(i)-\nu(i))} = \prod_{i=1}^{M} (\nu(i))^{n(\gamma(i)-\nu(i))}.$$

It is immediate that,

$$\prod_{i=1}^{M} (\nu(i))^{n(\gamma(i)-\nu(i))} = 2^{n(H(\nu)-D(\gamma\|\nu)-H(\gamma))}.$$

□

Lemma B.1.5. *For every $\nu \in \mathsf{T}_n$,*

$$\frac{1}{|\mathsf{T}_n|} 2^{nH(\nu)} \leq |\mathcal{Y}_n(\nu)| \leq 2^{nH(\nu)}$$

Proof. $\mathbb{P}_\nu$ be the probability law associated with an infinite sequence of i.i.d. random variables $Y_1, Y_2, \ldots,$ distributed according to ν.

$$\sum_{y^n \in \mathcal{Y}_n(\nu)} \mathbb{P}_\nu[(Y_1, Y_2, \ldots, Y_n) = y^n] \leq 1,$$

implying (from Lemma B.1.2) that

$$|\mathcal{Y}_n(\nu)| 2^{-nH(\nu)} \leq 1.$$

Now, we also have

$$\begin{aligned}
1 &= \sum_{\gamma \in \mathsf{T}_n} \sum_{y^n \in \mathcal{Y}_n(\gamma)} \mathbb{P}_\nu[(Y_1, Y_2, \dots, Y_n) = y^n] \\
&= \sum_{\gamma \in \mathsf{T}_n} |\mathcal{Y}_n(\gamma)| 2^{-n(H(\gamma)+D(\gamma\|\nu))} \\
&\overset{(a)}{\leq} \sum_{\gamma \in \mathsf{T}_n} |\mathcal{Y}_n(\nu)| 2^{-n(H(\nu)-D(\gamma\|\nu)-H(\gamma))} 2^{-n(H(\gamma)+D(\gamma\|\nu))} \\
&= \sum_{\gamma \in \mathsf{T}_n} |\mathcal{Y}_n(\nu)| 2^{-nH(\nu)} \\
&= |\mathsf{T}_n| |\mathcal{Y}_n(\nu)| 2^{-nH(\nu)}.
\end{aligned}$$

Here, (a) follows from Corollary B.1.4. □

Lemma B.1.6. *For any $\nu, \mu \in \mathsf{T}_n$,*

$$\frac{1}{|\mathsf{T}_n|} 2^{-nD(\nu\|\mu)} \leq \mathbb{P}_\mu(\mathsf{T}_{Y^n} = \nu) \leq 2^{-nD(\nu\|\mu)}.$$

Proof. From Lemma B.1.2, we see that

$$\mathbb{P}_\mu(\mathsf{T}_{y^n} = \nu) = |\mathcal{Y}_n(\nu)| 2^{-n(H(\nu)+D(\nu\|\mu))}.$$

The proof is completed by applying Lemma B.1.5. □

Lemma B.1.7. *Let $\Sigma_{\nu_n}, \Sigma_\nu \subseteq \Sigma_\mu$ and $\nu_n \to \nu$. Then $D(\nu_n\|\mu) \to D(\nu\|\mu)$.*

Proof. This follows as, for $i = 1, \dots, M$, $\nu_n(i) \to \nu(i)$ and hence $\nu_n(i) \log \frac{\nu_n(i)}{\mu(i)} \to \nu(i) \log \frac{\nu(i)}{\mu(i)}$. □

Theorem B.1.8 (Sanov)**.** *For every $\Gamma \subseteq \mathcal{M}(\Sigma_\mu)$,*

$$\begin{aligned}
-\inf_{\nu \in \Gamma^r} D(\nu\|\mu) &\leq \liminf_n \frac{1}{n} \log \mathbb{P}_\mu(\mathsf{T}_{Y^n} \in \Gamma) \\
&\leq \limsup_n \frac{1}{n} \log \mathbb{P}_\mu(\mathsf{T}_{Y^n} \in \Gamma) \leq -\inf_{\nu \in \Gamma} D(\nu\|\mu).
\end{aligned}$$

Here, $\Gamma^r = \{\nu \in \Gamma : \exists \nu_n \in \mathsf{T}_n \cap \Gamma, \nu_n \to \nu\}$.

Proof. From Lemma B.1.6, we have

$$\begin{aligned}\mathbb{P}_\mu(\mathsf{T}_{Y^n} \in \Gamma) &= \sum_{\gamma \in \mathsf{T}_n \cap \Gamma} \mathbb{P}_\mu(\mathsf{T}_{Y^n} = \gamma) \\ &\leq \sum_{\gamma \in \mathsf{T}_n \cap \Gamma} 2^{-nD(\gamma\|\mu)} \\ &\leq \sum_{\gamma \in \mathsf{T}_n \cap \Gamma} 2^{-n \inf_{\nu\in\Gamma} D(\nu\|\mu)} \\ &= |\Gamma \cap \mathsf{T}_n| 2^{-n \inf_{\nu\in\Gamma} D(\nu\|\mu)} \\ &\leq (n+1)^{(M-1)} 2^{-n \inf_{\nu\in\Gamma} D(\nu\|\mu)}.\end{aligned}$$

Therefore,

$$\frac{1}{n} \log \mathbb{P}_\mu(\mathsf{T}_{Y^n} \in \Gamma) \leq \frac{M-1}{n} \log(n+1) - \inf_{\nu\in\Gamma} D(\nu\|\mu).$$

Taking $\limsup_n$ on both sides yields the upper bound.

Given $\nu \in \Gamma^r$. Let $\hat{\nu}_n \in \mathsf{T}_n \cap \Gamma$, such that $\hat{\nu}_n \to \nu$. Then

$$\begin{aligned}\mathbb{P}_\mu(\mathsf{T}_{Y^n} \in \Gamma) &= \sum_{\gamma \in \mathsf{T}_n \cap \Gamma} \mathbb{P}_\mu(\mathsf{T}_{Y^n} = \gamma) \\ &\geq \mathbb{P}_\mu(\mathsf{T}_{Y^n} = \hat{\nu}_n) \\ &\geq \frac{1}{|\mathsf{T}_n|} 2^{-nD(\hat{\nu}_n\|\mu)},\end{aligned}$$

where the last inequality follows from Lemma B.1.6.

Therefore

$$\begin{aligned}\frac{1}{n} \log \mathbb{P}_\mu(\mathsf{T}_{Y^n} \in \Gamma) &\geq -\frac{1}{n} \log |\mathsf{T}_n| - D(\hat{\nu}_n\|\mu) \\ &\geq -\frac{M-1}{n} \log(n+1) - D(\hat{\nu}_n\|\mu).\end{aligned}$$

Taking lim inf and using Lemma B.1.7, we obtain

$$\liminf_n \frac{1}{n} \log \mathbb{P}_\mu(\mathsf{T}_{Y^n} \in \Gamma) \geq -D(\nu\|\mu).$$

Since, this holds for all $\nu \in \Gamma^r$, we get

$$\liminf_n \frac{1}{n} \log \mathbb{P}_\mu(\mathsf{T}_{Y^n} \in \Gamma) \geq -\inf_{\nu \in \Gamma^r} D(\nu \| \mu).$$

□

Given two sets A and B, the Hausdorff distance is defined as

$$d_H(A, B) = \max\{\sup_{x \in A} \inf_{y \in B} d(x, y), \sup_{y \in B} \inf_{x \in A} d(x, y)\}.$$

In the space of probability distributions, let us consider the underlying metric to be the total variation distance.

Lemma B.1.9. *Let* $\{\Gamma_n\}_{n \geq 1}, \Gamma \subseteq \mathcal{M}(\Sigma_\mu)$*, and* $d_H(\Gamma_n, \Gamma) \to 0$*. Then*

$$\lim_n \inf_{\nu \in \Gamma_n} D(\nu \| \mu) = \inf_{\nu \in \Gamma} D(\nu \| \mu).$$

Proof. Let $\nu^* \in \Gamma$ be such that $D(\nu^* || \mu) \leq \inf_{\nu \in \Gamma} D(\nu || \mu) + \epsilon$. Note that $d_H(\Gamma_n, \Gamma) \geq \inf_{\hat{\nu} \in \Gamma_n} d_{TV}(\hat{\nu}, \nu^*)$. Since $d_H(\Gamma_n, \Gamma) \to 0$, there exists a sequence $\hat{\nu}_n \in \Gamma_n$ such that $\hat{\nu}_n \to \nu^*$. Hence from Lemma B.1.7, $\lim_n D(\hat{\nu}_n || \mu) \to D(\nu^* || \mu)$. Now

$$\begin{aligned} \limsup_n \inf_{\nu \in \Gamma_n} D(\nu || \mu) &\leq \limsup_n D(\hat{\nu}_n || \mu) \\ &= D(\nu^* || \mu) \leq \inf_{\nu \in \Gamma} D(\nu || \mu) + \epsilon, \end{aligned}$$

implying $\limsup_n \inf_{\nu \in \Gamma_n} D(\nu || \mu) \leq \inf_{\nu \in \Gamma} D(\nu || \mu)$.

Let n_k be a subsequence such that $\inf_{\nu \in \Gamma_{n_k}} D(\nu \| \mu) \overset{k \to \infty}{\to} \liminf_n \inf_{\nu \in \Gamma_n} D(\nu \| \mu)$. Consider $\nu_k \in \Gamma_{n_k}$ such that $D(\nu_k \| \mu) \leq \inf_{\nu \in \Gamma_{n_k}} D(\nu \| \mu) + \frac{\epsilon}{k}$. Since $\mathcal{M}(\Sigma_\mu)$ is compact, there exists a convergent subsequence $\{k_l\}$, i.e. $\nu_{k_l} \to \nu^*$, for some $\nu^* \in \mathcal{M}(\Sigma_\mu)$. By construction,

$$\limsup_l D(\nu_{k_l} \| \mu) = \limsup_l D(\nu_{k_l} \| \mu) - \frac{\epsilon}{k_l}$$

$$\begin{aligned}&\leq \limsup_k D(\nu_k\|\mu) - \frac{\epsilon}{k} \leq \limsup_k \inf_{\nu\in\Gamma_{n_k}} D(\nu\|\mu)\\&= \liminf_n \inf_{\nu\in\Gamma_n} D(\nu\|\mu).\end{aligned}$$

Since $\nu_{k_l} \to \nu^*$, Lemma B.1.7 yields $D(\nu_{k_l}\|\mu) \to D(\nu^*\|\mu)$. Therefore

$$D(\nu^*\|\mu) \leq \liminf_n \inf_{\nu\in\Gamma_n} D(\nu\|\mu).$$

As $d_H(\Gamma_{n_k}, \Gamma) \geq \inf_{\hat{\nu}\in\Gamma} d_{TV}(\hat{\nu}, \nu_k)$, let $\nu_k^\dagger \in \Gamma$, satisfy $d_{TV}(\nu_k^\dagger, \nu_k) \leq d_H(\Gamma_{n_k}, \Gamma) + \frac{\epsilon}{k}$. Note that

$$\begin{aligned}d_{TV}(\nu_{k_l}^\dagger, \nu^*) &\leq d_{TV}(\nu_{k_l}^\dagger, \nu_{k_l}) + d_{TV}(\nu_{k_l}, \nu^*)\\&\leq d_H(\Gamma_{n_{k_l}}, \Gamma) + \frac{\epsilon}{k_l} + d_{TV}(\nu_{k_l}, \nu^*).\end{aligned}$$

Therefore, taking $l \to \infty$, we see that $\nu_{k_l}^\dagger \to \nu^*$. Finally as $\nu_{k_l}^\dagger \in \Gamma$,

$$\inf_{\nu\in\Gamma} D(\nu\|\mu) \leq \liminf_l D(\nu_{k_l}^\dagger\|\mu) = D(\nu^*\|\mu).$$

Putting all this together, we obtain

$$\begin{aligned}\inf_{\nu\in\Gamma} D(\nu\|\mu) \leq D(\nu^*\|\mu) &\leq \liminf_n \inf_{\nu\in\Gamma_n} D(\nu\|\mu)\\&\leq \limsup_n \inf_{\nu\in\Gamma_n} D(\nu\|\mu) \leq \inf_{\nu\in\Gamma} D(\nu\|\mu),\end{aligned}$$

establishing the lemma. □

The following lemma can be considered as a limiting version of the discrete Sanov's theorem.

Theorem B.1.10 (Limiting Sanov). *Let $\{\Gamma_n\}_{n\geq 1}, \Gamma \subseteq \mathcal{M}(\Sigma_\mu)$, and $d_H(\Gamma_n, \Gamma) \to 0$. Then,*

$$-\inf_{\nu\in\Gamma^r} D(\nu\|\mu) \leq \liminf_n \frac{1}{n} \log \mathbb{P}_\mu(\mathsf{T}_{Y^n} \in \Gamma_n)$$

$$\leq \limsup_n \frac{1}{n} \log \mathbb{P}_\mu(\mathsf{T}_{Y^n} \in \Gamma_n) \leq - \inf_{\nu \in \Gamma} D(\nu \| \mu). \tag{B.1}$$

Here, $\Gamma^r = \{\nu \in \Gamma : \exists \nu_n \in \mathsf{T}_n \cap \Gamma_n, \nu_n \to \nu\}$*. Clearly* $\Gamma^r \supseteq \Gamma^o$*, where* Γ^o *is the interior of* Γ *considered as a subset of* $\mathcal{M}(\Sigma_\mu)$*. In particular, if* $\Gamma_n \subseteq \mathsf{T}_n$*, then* $\Gamma^r = \Gamma$*, and*

$$\lim_n \frac{1}{n} \log \mathbb{P}_\mu(\mathsf{T}_{Y^n} \in \Gamma_n) = - \inf_{\nu \in \Gamma} D(\nu \| \mu).$$

Proof. Note that

$$\begin{aligned}
\mathbb{P}_\mu(\mathsf{T}_{Y^n} \in \Gamma_n) &= \sum_{\gamma \in \mathsf{T}_n \cap \Gamma_n} \mathbb{P}_\mu(\mathsf{T}_{Y^n} = \gamma) \\
&\overset{(a)}{\leq} \sum_{\gamma \in \mathsf{T}_n \cap \Gamma_n} 2^{-nD(\gamma \| \mu)} \\
&\leq \sum_{\gamma \in \mathsf{T}_n \cap \Gamma_n} 2^{-n \inf_{\nu \in \Gamma_n} D(\nu \| \mu)} \\
&= |\Gamma_n \cap \mathsf{T}_n| 2^{-n \inf_{\nu \in \Gamma_n} D(\nu_n \| \mu)} \\
&\leq (n+1)^{(M-1)} 2^{-n \inf_{\nu \in \Gamma_n} D(\nu \| \mu)}.
\end{aligned}$$

Here (a) follows from Lemma B.1.6. Therefore,

$$\frac{1}{n} \log \mathbb{P}_\mu(\mathsf{T}_{Y^n} \in \Gamma_n) \leq \frac{M-1}{n} \log(n+1) - \inf_{\nu \in \Gamma_n} D(\nu \| \mu).$$

Taking $\limsup_n$ on both sides and using Lemma B.1.9 yields the upper bound.

Given $\nu \in \Gamma^r$, let $\hat{\nu}_n \in \mathsf{T}_n \cap \Gamma, \hat{\nu}_n \to \nu$. Then

$$\begin{aligned}
\mathbb{P}_\mu(\mathsf{T}_{Y^n} \in \Gamma_n) &= \sum_{\gamma \in \mathsf{T}_n \cap \Gamma_n} \mathbb{P}_\mu(\mathsf{T}_{Y^n} = \gamma) \\
&\geq \mathbb{P}_\mu(\mathsf{T}_{Y^n} = \hat{\nu}_n) \\
&\overset{(a)}{\geq} \frac{1}{|\mathsf{T}_n|} 2^{-nD(\hat{\nu}_n \| \mu)}.
\end{aligned}$$

Here, again, (a) follows from Lemma B.1.6. Therefore

$$\frac{1}{n}\log\mathbb{P}_{\mu}(\mathsf{T}_{Y^n}\in\Gamma_n)\geq-\frac{1}{n}\log|\mathsf{T}_n|-D(\hat{\nu}_n\|\mu).$$

Taking $\liminf$ and using Lemma B.1.7, we obtain

$$\liminf_n\frac{1}{n}\log\mathbb{P}_{\mu}(\mathsf{T}_{Y^n}\in\Gamma_n)\geq-D(\nu\|\mu).$$

Since, this holds for all $\nu\in\Gamma^r$, we get

$$\liminf_n\frac{1}{n}\log\mathbb{P}_{\mu}(\mathsf{T}_{Y^n}\in\Gamma_n)\geq-\inf_{\nu\in\Gamma^r}D(\nu\|\mu).$$

This proves (B.1).

If $\Gamma_n\subseteq\mathsf{T}_n$, then as $d_H(\Gamma_n,\Gamma)\to0$ it is immediate that for every $\nu\in\Gamma,\exists\nu_n\in\Gamma_n=\Gamma_n\cap\mathsf{T}_n$ such that $d_{TV}(\nu_n,\nu)\to0$. This implies that $\Gamma^r=\Gamma$. □

B.2 Maximal couplings

In this section, we will show the derivation of maximal coupling from the discrete Sanov theorem. Let p_X,p_Y be two distributions supported on a finite alphabet Σ. Let $\{\omega_n\}$ be a non-negative sequence such that $\omega_n\to0$ as $n\to\infty$ and $\omega_n\sqrt{n}\to\infty$ as $n\to\infty$. Define $A_n\subseteq\Sigma^n$ as

$$A_n=\{g^n\in\Sigma^n:|\mathsf{T}_{g^n}(a)-p_X(a)|\leq\omega_np_X(a),\forall a\in\Sigma\}.$$

Remark B.2.1. The set A_n is usually called the typical sequences (or strongly-typical sequences) corresponding to distribution p_X. These sets play an important role in network information theory, particularly in the proofs of the channel coding theorems.

Similarly, let

$$B_n = \{g^n \in \Sigma^n : |\mathsf{T}_{g^n}(a) - p_Y(a)| \leq \omega_n p_Y(a), \forall a \in \Sigma\}.$$

Define $\hat{\Gamma}_n = \{\nu \in \mathcal{M}(\Sigma \times \Sigma) : \nu = \mathsf{T}_{(g_1^n, g_2^n)}$ for some $(g_1^n, g_2^n) \in A_n \times B_n\}$ and $\hat{\Gamma} = \Pi(p_X, p_Y)$, the set of all couplings with the given marginals.

Lemma B.2.2. *Let $\hat{\Gamma}_n$ and $\hat{\Gamma}$ be as described above. Then*

$$d_H(\hat{\Gamma}_n, \hat{\Gamma}) \to 0.$$

Proof. Let $\nu \in \hat{\Gamma}$. Then we know, from Lemma B.1.1 that there exists $\hat{\nu}_n \in \mathsf{T}_n$ such that $|\hat{\nu}_n(a,b) - \nu(a,b)| \leq \frac{1}{n}$ for all $(a,b) \in \Sigma_A \times \Sigma_B$, and if $\nu(a,b) = 0$, then $\hat{\nu}_n(a,b) = 0$. Now

$$\left| \sum_{b \in \Sigma_B} (\hat{\nu}_n(a,b) - \nu(a,b)) \right| \leq \sum_{b \in \Sigma_B} |\hat{\nu}_n(a,b) - \nu(a,b)| \leq \frac{|\Sigma_B|}{n}.$$

Note that $\sqrt{n}\omega_n \nu(a,b) \to \infty$, for all $(a,b) : \nu(a,b) > 0$. Therefore, for large n, $\frac{\Sigma_B}{n} \leq \omega_n p_X(a) = \omega_n \sum_b \nu(a,b)$. Similarly, for large n,

$$\begin{aligned} \left| \sum_{a \in \Sigma_A} (\hat{\nu}_n(a,b) - \nu(a,b)) \right| &\leq \sum_{a \in \Sigma_A} |\hat{\nu}_n(a,b) - \nu(a,b)| \leq \frac{|\Sigma_A|}{n} \\ &\leq \omega_n p_Y(b) = \omega_n, \sum_a \nu(a,b). \end{aligned}$$

Therefore, for large n, any (g_1^n, g_2^n), such that $\mathsf{T}_{g_1^n, g_2^n} = \hat{\nu}_n(a,b)$, is an element of $A_n \times B_n$, or that $\hat{\nu}_n \in \hat{\Gamma}_n$. Further, note that, $d_{TV}(\hat{\nu}_n, \nu) \leq \frac{|\Sigma_A||\Sigma_B|}{2n}$. Since this holds for any $\nu \in \hat{\Gamma}$, we obtain that $\sup_{\nu \in \hat{\Gamma}} \inf_{\nu_n \in \hat{\Gamma}_n} d_{TV}(\nu, \nu_n) \to 0$ as $n \to \infty$.

Now suppose $\sup_{\nu_n \in \hat{\Gamma}_n} \inf_{\nu \in \hat{\Gamma}} d_{TV}(\nu, \nu_n) \not\to 0$. There, there is a subsequence n_k and $\epsilon > 0$ such that

$$\sup_{\nu_k \in \hat{\Gamma}_{n_k}} \inf_{\nu \in \hat{\Gamma}} d_{TV}(\nu, \nu_k) > \epsilon.$$

Therefore, there is a sequence $\nu_k \in \hat{\Gamma}_{n_k}$ such that $\inf_{\nu \in \hat{\Gamma}} d_{TV}(\nu_k, \nu) > \frac{\epsilon}{2}$. As

$\mathcal{M}(\Sigma_A \times \Sigma_B)$ is a compact set, and $\hat{\Gamma}_{n_k} \subset \mathcal{M}(\Sigma_A \times \Sigma_B)$, we have a convergent subsequence $\nu_{k_l} \to \nu^\dagger$. By definition of $\hat{\Gamma}_{n_{k_l}}$ we have

$$\left| \left(\sum_b \nu_{k_l}(a,b) \right) - p_X(a) \right| \leq \omega_{n_{k_l}} p_X(a).$$

As $\omega_{n_{k_l}} \to 0$, $\sum_b \nu^\dagger(a,b) = p_X(a)$. Similarly, $\sum_a \nu^\dagger(a,b) = p_Y(b)$. Therefore $\nu^\dagger \in \hat{\Gamma}$. Therefore $d_{TV}(\nu_k, \nu^\dagger) \to 0$, contradicting $\inf_{\nu \in \hat{\Gamma}} d_{TV}(\nu_k, \nu) > \frac{\epsilon}{2}$. This shows that $\sup_{\nu_n \in \hat{\Gamma}_n} \inf_{\nu \in \hat{\Gamma}} d_{TV}(\nu, \nu_n) \to 0$ as desired. □

Lemma B.2.3 (Data processing)**.** *Let $W_{Y|X}$ be a stochastic mapping (channel). Let p_X, q_X be two distributions on $\mathcal{X}$ and $p_Y = \sum_x W_{Y|X} p_X$ and $q_Y = \sum_X W_{Y|X} q_X$ be the two induced distributions on $\mathcal{Y}$. Then $d_{TV}(p_X, q_X) \geq d_{TV}(p_Y, q_Y)$.*

Proof. Observe the following:

$$\begin{aligned}
\sum_y |p_Y(y) - q_Y(y)| &= \sum_y \left| \sum_x W_{Y|X}(y|x)(p_X(x) - q_X(x)) \right| \\
&\leq \sum_{x,y} W_{Y|X}(y|x) |p_X(x) - q_X(x)| \\
&= \sum_x |p_X(x) - q_X(x)|
\end{aligned}$$

□

Theorem B.2.4. *Let p_X and p_Y be two distributions having finite support on an Abelian group $\mathbb{G}$. Let $\{\omega_n\}$ be a non-negative sequence such that $\omega_n \to 0$ as $n \to \infty$ and $\omega_n \sqrt{n} \to \infty$ as $n \to \infty$. Define $A_n \subseteq \mathbb{G}^n$, as*

$$A_n = \{g^n \in \mathbb{G}^n : |k_{g^n}(a) - np_X(a)| \leq np_X(a)\omega_n, \forall a \in \mathbb{G}\}.$$

Here $k_{g^n}(a) := |\{i : g_i = a, 1 \leq i \leq n\}|$. Similarly, let

$$B_n = \{g^n \in \mathbb{G}^n : |k_{g^n}(a) - np_Y(a)| \leq np_Y(a)\omega_n, \forall a \in \mathbb{G}\}.$$

$$\lim_n \frac{1}{n} \log |A_n + B_n| = \max_{q \in \Pi(p_X, p_Y)} H_q(X+Y).$$

Here $\Pi(p_X, p_Y)$ *is the set of joint distributions (couplings) such that the marginals are* p_X *and* p_Y*, respectively.*

Proof. Let S be a finite subset of $\mathbb{G}$ such that S contains support of p_X, p_Y and p_{X+Y}. Let μ be the uniform distribution on S. Let $C_n = A_n + B_n$, and $\Gamma_n = \{\nu \in \mathcal{M}(S) : \nu = \mathsf{T}_{g^n}$ for some $g^n \in C_n\}$. Let $\Gamma = \{\nu \in \mathcal{M}(\mathbb{G}) : \nu = q_{X+Y}, q_{X,Y} \in \Pi(p_X, p_Y)\}$. Define $\hat{\Gamma}_n = \{\nu \in \mathcal{M}(S \times S) : \nu = \mathsf{T}_{(g_1^n, g_2^n)}$ for some $(g_1^n, g_2^n) \in A_n \times B_n\}$ and $\hat{\Gamma} = \Pi(p_X, p_Y)$, the set of all couplings with the given marginals.

For $\nu \in \mathsf{T}_n$, we had defined the type-class by $\mathcal{Y}_n(\nu) = \{y^n \in S^n : \mathsf{T}_{y^n} = \nu\}$. Hence, $\nu \in \Gamma_n$ if and only if $\mathcal{Y}_n(\nu) \subseteq C_n$. Note that, by definition, the sets A_n, B_n, and C_n are permutation invariant. Therefore

$$\mathbb{P}_\mu(\mathsf{T}_{Y^n} \in \Gamma_n) = \mathbb{P}_\mu(Y^n \in C_n) = \frac{|C_n|}{|S|^n}.$$

From Lemma B.2.2, $d_H(\hat{\Gamma}_n, \hat{\Gamma}) \to 0$, and since $\hat{\Gamma}_n \subseteq \mathsf{T}_n$, $\hat{\Gamma} = \hat{\Gamma}^r$. Considering $(X, Y) \mapsto X + Y$, by Lemma B.2.3, we obtain $d_H(\Gamma_n, \Gamma) \to 0$ and similarly as $\Gamma_n \subseteq \mathsf{T}_n$, $\Gamma = \Gamma^r$. Therefore, we apply Theorem B.1.10 to obtain

$$\lim_n \frac{1}{n} \log \mathbb{P}_\mu(\mathsf{T}_{Y^n} \in \Gamma_n) = -\inf_{\nu \in \Gamma} D(\nu \| \mu)$$

or equivalently

$$\begin{aligned}\lim_n \frac{1}{n} \log \frac{|C_n|}{|S|^n} &= -\inf_{\nu \in \Gamma} (\log |S| - H_\nu(X+Y)) \\ &= \sup_{\nu \in \Gamma} H_\nu(X+Y) - \log |S|.\end{aligned}$$

Since Γ and $\hat{\Gamma}$ are compact, and $\sup_{\nu \in \Gamma} H_\nu(X+Y) = \max_{q \in \Pi(p_X, p_Y)} H_q(X+Y)$, we are done. □

Appendix C

Proof of Theorem 4.2.15

Proof. The arguments here are directly motivated by those for establishing the sumset inequality in [KT99] and are essentially identical to the one employed in [TV]. We still present it here to highlight the role played by Lemma 4.2.11. Consider a joint distribution $(X, Y, Y^\dagger)$ such that $Y \to X \to Y^\dagger$ forms a Markov chain and (X, Y) shares the same marginal as $(X, Y^\dagger)$. From Lemma 4.2.11 (considering $(X, Y) - X - (X, Y^\dagger)$) we have

$$\begin{aligned} H(X, Y, Y^\dagger) &= H(X, Y) + H(X, Y^\dagger) - H(X) \\ &= 2H(X, Y) - H(X). \end{aligned} \tag{C.1}$$

Here, the last equality comes from the assumption that $(X, Y) \overset{(d)}{=} (X, Y^\dagger)$.

Define three functions: $f_1(x, y, y^\dagger) = (x+y, x+y^\dagger)$, $f_2(x, y, y^\dagger) = (y, y^\dagger)$, $f_3(x, y, y^\dagger) = (x+y, y^\dagger)$. Consider a joint distribution of $(X_1, Y_1, Y_1^\dagger, X_2, Y_2, Y_2^\dagger, X_3, Y_3, Y_3^\dagger, X_4, Y_4, Y_4^\dagger)$ such that the following three conditions are satisfied:

1. $(X_i, Y_i, Y_i^\dagger)$ shares the same marginal as $(X, Y, Y^\dagger)$ for $1 \le i \le 4$.

2. $f_i(X_i, Y_i, Y_i^\dagger) = f_i(X_{i+1}, Y_{i+1}, Y_{i+1}^\dagger)$ for $1 \le i \le 3$.

3. $(X_1, Y_1, Y_1^\dagger) \to f_1(X_1, Y_1, Y_1^\dagger) \to (X_2, Y_2, Y_2^\dagger) \to f_2(X_2, Y_2, Y_2^\dagger) \to (X_3, Y_3, Y_3^\dagger) \to f_3(X_3, Y_3, Y_3^\dagger) \to (X_4, Y_4, Y_4^\dagger)$ forms a Markov chain.

Now by Lemma 4.2.11, we have

$$\begin{aligned} &H(X_1, Y_1, Y_1^\dagger, X_2, Y_2, Y_2^\dagger, X_3, Y_3, Y_3^\dagger, X_4, Y_4, Y_4^\dagger) \\ &\quad = 4H(X, Y, Y^\dagger) - H(X+Y, X+Y^\dagger) - H(Y, Y^\dagger) - H(X+Y, Y^\dagger). \qquad \text{(C.2)} \end{aligned}$$

From condition 2) above and the definition of f_1, f_2, f_3, we have the following equalities:

$$\begin{aligned} X_1 + Y_1 = X_2 + Y_2, &\quad X_1 + Y_1^\dagger = X_2 + Y_2^\dagger, \\ Y_2 = Y_3, &\quad Y_2^\dagger = Y_3^\dagger, \\ X_3 + Y_3 = X_4 + Y_4, &\quad Y_3^\dagger = Y_4^\dagger. \end{aligned}$$

From this, we obtain the following:

$$Y_1 - Y_1^\dagger = Y_2 - Y_2^\dagger = Y_3 - Y_3^\dagger.$$

Consequently, we have

$$\begin{aligned} X_4 - Y_4^\dagger &= (X_4 + Y_4) - Y_4 - Y_4^\dagger = (X_3 + Y_3) - Y_4^\dagger - Y_4 \\ &= X_3 + (Y_3 - Y_3^\dagger) - Y_4 = X_3 + Y_1 - Y_1^\dagger - Y_4. \end{aligned}$$

Therefore $X_4 - Y_4^\dagger$ is a function of $(X_1, Y_1, Y_1^\dagger, X_3, Y_4)$. Therefore,

$$\begin{aligned} &H(X_1, Y_1, Y_1^\dagger, X_2, Y_2, Y_2^\dagger, X_3, Y_3, Y_3^\dagger, X_4, Y_4, Y_4^\dagger \\ &\qquad |X_1, Y_1, Y_1^\dagger, X_3, Y_4) \\ &= H(X_1, Y_1, Y_1^\dagger, X_2, Y_2, Y_2^\dagger, X_3, Y_3, Y_3^\dagger, X_4, Y_4, Y_4^\dagger \\ &\qquad |X_1, Y_1, Y_1^\dagger, X_3, Y_4, X_4 - Y_4^\dagger) \\ &= H(X_4 | X_1, Y_1, Y_1^\dagger, X_3, Y_4, X_4 - Y_4^\dagger) \\ &\quad + H(X_1, Y_1, Y_1^\dagger, X_2, Y_2, Y_2^\dagger, X_3, Y_3, Y_3^\dagger, X_4, Y_4, Y_4^\dagger \end{aligned}$$

$$
\begin{aligned}
&\qquad |X_1, Y_1, Y_1^\dagger, X_3, X_4, Y_4, Y_4^\dagger) \\
&\le H(X_1, Y_1, Y_1^\dagger, X_2, Y_2, Y_2^\dagger, X_3, Y_3, Y_3^\dagger, X_4, Y_4, Y_4^\dagger \\
&\qquad |X_1, Y_1, Y_1^\dagger, X_3, X_4, Y_4, Y_4^\dagger) + H(X_4|X_4 - Y_4^\dagger) \\
&= H(X_2, Y_2, Y_2^\dagger, Y_3, Y_3^\dagger \\
&\qquad |X_1, Y_1, Y_1^\dagger, X_3, X_4, Y_4, Y_4^\dagger) + H(X_4|X_4 - Y_4^\dagger).
\end{aligned}
$$

To complete the argument, observe that $Y_2 = Y_3 = X_4 + Y_4 - X_3$, $Y_2^\dagger = Y_3^\dagger = Y_4^\dagger$, and $X_2 = X_1 + Y_1 - Y_2 = X_1 + Y_1 + X_3 - X_4 - Y_4$. This implies that $(X_2, Y_2, Y_2^\dagger, Y_3, Y_3^\dagger)$ is a function of $(X_1, Y_1, Y_1^\dagger, X_3, X_4, Y_4, Y_4^\dagger)$. Therefore

$$H(X_2, Y_2, Y_2^\dagger, Y_3, Y_3^\dagger | X_1, Y_1, Y_1^\dagger, X_3, X_4, Y_4, Y_4^\dagger) = 0,$$

implying that

$$H(X_1, Y_1, Y_1^\dagger, X_2, Y_2, Y_2^\dagger, X_3, Y_3, Y_3^\dagger, X_4, Y_4, Y_4^\dagger | X_1, Y_1, Y_1^\dagger, X_3, Y_4) \le H(X_4|X_4 - Y_4^\dagger).$$

Thus, we have

$$
\begin{aligned}
&H(X_1, Y_1, Y_1^\dagger, X_2, Y_2, Y_2^\dagger, X_3, Y_3, Y_3^\dagger, X_4, Y_4, Y_4^\dagger) \\
&\qquad \le H(X_1, Y_1, Y_1^\dagger, X_3, Y_4) + H(X_4|X_4 - Y_4^\dagger). \qquad \text{(C.3)}
\end{aligned}
$$

By using (C.2) and (C.3), we have

$$
\begin{aligned}
0 &\ge 4H(X, Y, Y^\dagger) - H(X + Y, X + Y^\dagger) - H(Y, Y^\dagger) \\
&\quad - H(X + Y, Y^\dagger) - H(X_1, Y_1, Y_1^\dagger, X_3, Y_4) - H(X_4|X_4 - Y_4^\dagger) \\
&= 3H(X, Y, Y^\dagger) - H(X + Y, X + Y^\dagger) - H(Y, Y^\dagger) \\
&\quad - H(X + Y, Y^\dagger) - H(X_3, Y_4|X_1, Y_1, Y_1^\dagger) - H(X_4|X_4 - Y_4^\dagger).
\end{aligned}
$$

Now using (C.1) to replace $H(X, Y, Y^\dagger)$ we have

$$\begin{aligned}
0 &\geq 6H(X,Y) - 3H(X) - H(X+Y, X+Y^\dagger) - H(Y, Y^\dagger) \\
&\quad - H(X+Y, Y^\dagger) - H(X_3, Y_4 | X_1, Y_1, Y_1^\dagger) - H(X_4 | X_4 - Y_4^\dagger) \\
&\geq 6H(X,Y) - 3H(X) - 3H(Y) - 3H(X+Y) - H(X_3) \\
&\quad - H(Y_4) - H(X_4, Y_4^\dagger) + H(X_4 - Y_4^\dagger) \\
&= 5H(X,Y) - 4H(X) - 4H(Y) - 3H(X+Y) + H(X-Y) \qquad \text{(C.4)} \\
&= \frac{1}{2}I(X; X-Y) + \frac{1}{2}I(Y; X-Y) - \frac{3}{2}I(X; X+Y) - \frac{3}{2}I(Y; X+Y) - 3I(X;Y).
\end{aligned}$$

This completes the proof of the theorem. □

Bibliography

[ABBN04] Shiri Artstein, Keith Ball, Franck Barthe, and Assaf Naor, *Solution of Shannon's problem on the monotonicity of entropy*, Journal of the American Mathematical Society **17** (2004), no. 4, 975–982.

[AGKN13] Venkat Anantharam, Amin Aminzadeh Gohari, Sudeep Kamath, and Chandra Nair, *On maximal correlation, hypercontractivity, and the data processing inequality studied by Erkip and Cover*, CoRR **abs/1304.6133** (2013).

[AGKN14] Venkat Anantharam, Amin Gohari, Sudeep Kamath, and Chandra Nair, *On hypercontractivity and a data processing inequality*, 2014 IEEE International Symposium on Information Theory (ISIT'2014) (Honolulu, USA), June 2014, pp. 3022–3026.

[AJN22] Venkat Anantharam, Varun Jog, and Chandra Nair, *Unifying the Brascamp-Lieb inequality and the entropy power inequality*, IEEE Transactions on Information Theory **68** (2022), no. 12, 7665–7684.

[Bar86] Andrew R. Barron, *Entropy and the central limit theorem*, The Annals of Probability **14** (1986), no. 1, 336–342.

[BB12] Paul Balister and Béla Bollobás, *Projections, entropy and sumsets*, Combinatorica **32** (2012), no. 2, 125–141.

[BCCT08] J. Bennett, A. Carbery, M. Christ, and T. Tao, *The Brascamp-Lieb* *inequalities: Finiteness, structure and extremals*, Geometric and Functional Analysis **17** (2008), no. 5, 1343–1415.

[Ber73] P F Bergmans, *Coding theorem for broadcast channels with degraded components*, IEEE Trans. Info. Theory **IT-15** (March, 1973), 197–207.

[BL05] J.M. Borwein and A.S. Lewis, *Convex analysis and nonlinear optimization: Theory and examples*, CMS Books in Mathematics, Springer New York, 2005.

[CCE09] Eric A. Carlen and Dario Cordero-Erausquin, *Subadditivity of the entropy and its relation to Brascamp-Lieb type inequalities*, Geometric and Functional Analysis **19** (2009), no. 2, 373–405 (English).

[CGFS86] F.R.K Chung, R.L Graham, P Frankl, and J.B Shearer, *Some intersection theorems for ordered sets and graphs*, Journal of Combinatorial Theory, Series A **43** (1986), no. 1, 23–37.

[CK11] Imre Csiszár and János Körner, *Information theory: Coding theorems for discrete memoryless systems*, Cambridge University Press, 1 2011.

[Cou16a] Thomas A. Courtade, *Monotonicity of entropy and Fisher information: a quick proof via maximal correlation*, Commun. Inf. Syst. **16** (2016), no. 2, 111–115.

[Cou16b] Thomas A. Courtade, *Strengthening the entropy power inequality*, 2016 IEEE International Symposium on Information Theory (ISIT), 2016, pp. 2294–2298.

[CT91] T Cover and J Thomas, *Elements of information theory*, Wiley Interscience, 1991.

[Dar53] G. Darmois, *Analyse générale des liaisons stochastiques: etude particulière de l'analyse factorielle linéaire*, Revue de l'Institut International de Statistique / Review of the International Statistical Institute **21** (1953), no. 1/2, pp. 2–8 (English).

[DKS01] Amir Dembo, Abram Kagan, and Lawrence A. Shepp, *Remarks on the maximum correlation coefficient*, Bernoulli **7** (2001), no. 2, 343–350.

[DZ98] A. Dembo and O. Zeitouni, *Large deviations techniques and applications*, Applications of mathematics, Springer, 1998.

[ED16] Alberto Espuny Díaz, *Entropy methods for sumset inequalities*, Master's thesis, Universitat Politècnica de Catalunya, 2016.

[Fel99] Gennadiy Feldman, *More on the Skitovich-Darmois theorem for finite abelian groups*, Theory of Probability and Its Applications **45** (1999).

[Fri04] Ehud Friedgut, *Hypergraphs, entropy, and inequalities*, The American Mathematical Monthly **111** (2004), no. 9, 749–760.

[Gal14] D. Galvin, *Three tutorial lectures on entropy and counting*, arXiv preprint arXiv:1406.7872 (2014).

[Gar02] R. Gardner, *The Brunn-Minkowski inequality*, Bulletin of the American Mathematical Society **39** (2002), no. 3, 355–405.

[Gav23] Lampros Gavalakis, *Discrete generalised entropy power inequalities for log-concave random variables*, 2023 IEEE International Symposium on Information Theory (ISIT), 2023, pp. 42–47.

[Geb41] H. Gebelein, *Das statistische problem der korrelation als variations- und eigenwert-problem und sein zusammenhang mit der ausgleichungsrechnung*, Zeitschrift für angew. Math. und Mech. **21** (1941), 364–379.

[GGMT23] WT Gowers, Ben Green, Freddie Manners, and Terence Tao, *On a conjecture of Marton*, arXiv preprint arXiv:2311.05762 (2023).

[GMR10] Katalin Gyarmati, Máté Matolcsi, and Imre Z Ruzsa, *A superadditivity and submultiplicativity property for cardinalities of sumsets*, Combinatorica **30** (2010), no. 2, 163–174.

[GMT23] Ben Green, Freddie Manners, and Terence Tao, *Sumsets and entropy revisited*, arXiv:2306.13403 (2023).

[GN14] Yanlin Geng and Chandra Nair, *The capacity region of the two-receiver Gaussian vector broadcast channel with private and common messages*, IEEE Transactions on Information Theory **60** (2014), no. 4, 2087–2104.

[GR06] Ben Green and Imre Z. Ruzsa, *Sets with small sumset and rectification*, Bulletin of the London Mathematical Society **38** (2006), no. 1, 43—52.

[Gre09] Ben Green, *Additive combinatorics - chapter 2*, 2009.

[GSV06] Dongning Guo, Shlomo Shamai, and Sergio Verdu, *Proof of entropy power inequalities via MMSE*, 2006 IEEE International Symposium on Information Theory, 2006, pp. 1011–1015.

[Han78] Te Sun Han, *Nonnegative entropy measures of multivariate symmetric correlations*, Information and Control **36** (1978), no. 2, 133–156.

[Hir35] O. Hirschfeld, *A connection between correlation and contingency*, Mathematical Proceedings of the Cambridge Philosophical Society **31** (1935), 520–524.

[HV03] Peter Harremoës and Christophe Vignat, *An entropy power inequality for the binomial family*, JIPAM. Journal of Inequalities in Pure & Applied Mathematics [electronic only] **4** (2003).

[IKBA22] Rishabh Iyer, Ninad Khargonkar, Jeff Bilmes, and Himanshu Asnani, *Generalized submodular information measures: Theoretical properties,* *examples, optimization algorithms, and applications*, IEEE Transactions on Information Theory **68** (2022), no. 2, 752–781.

[JA14] Varun Jog and Venkat Anantharam, *The entropy power inequality and Mrs. Gerber's lemma for groups of order* 2^n, IEEE Transactions on Information Theory **60** (2014), no. 7, 3773–3786.

[Joh20] Oliver Johnson, *Maximal correlation and the rate of Fisher information convergence in the central limit theorem*, IEEE Transactions on Information Theory **66** (2020), no. 8, 4992–5002.

[KLN23] Chin Wa Ken Lau and Chandra Nair, *Information inequalities via ideas from additive combinatorics*, 2023 IEEE International Symposium on Information Theory (ISIT), 2023, pp. 2452–2457.

[KM14] Ioannis Kontoyiannis and Mokshay Madiman, *Sumset and inverse sumset inequalities for differential entropy and mutual information*, IEEE transactions on information theory **60** (2014), no. 8, 4503–4514.

[KN15] Sudeep Kamath and Chandra Nair, *The strong data processing constant for sums of i.i.d. random variables*, Information Theory (ISIT), 2015 IEEE International Symposium on, June 2015, pp. 2550–2552.

[KT99] Nets Hawk Katz and Terence Tao, *Bounds on arithmetic projections, and applications to the Kakeya conjecture*, Mathematical Research Letters **6** (1999), no. 6, 625–630.

[LCCV18] Jingbo Liu, Thomas A Courtade, Paul W Cuff, and Sergio Verdú, *A forward-reverse Brascamp-Lieb inequality: Entropic duality and Gaussian optimality*, Entropy **20** (2018), no. 6, 418.

[Lie78] Elliott H. Lieb, *Proof of an entropy conjecture of Wehrl*, Comm. Math. Phys. **62** (1978), no. 1, 35–41.

[LP08] Amos Lapidoth and Gábor Pete, *On the entropy of the sum and of the difference of independent random variables*, 2008 IEEE 25th Convention of Electrical and Electronics Engineers in Israel, IEEE, 2008, pp. 623–625.

[LYCH78] S. Leung-Yan-Cheong and M. Hellman, *The Gaussian wire-tap channel*, Information Theory, IEEE Transactions on **24** (1978), no. 4, 451–456.

[Mad08] Mokshay Madiman, *On the entropy of sums*, 2008 IEEE Information Theory Workshop, IEEE, 2008, pp. 303–307.

[MB07] Mokshay M. Madiman and Andrew R. Barron, *Generalized entropy power inequalities and monotonicity properties of information*, IEEE Trans. Inf. Theory **53** (2007), no. 7, 2317–2329.

[MG19] Mokshay Madiman and Farhad Ghassemi, *Combinatorial entropy power inequalities: A preliminary study of the Stam region*, IEEE Transactions on Information Theory **65** (2019), no. 3, 1375–1386.

[MK10] Mokshay Madiman and Ioannis Kontoyiannis, *The entropies of the sum and the difference of two iid random variables are not too different*, 2010 IEEE International Symposium on Information Theory, IEEE, 2010, pp. 1369–1372.

[MMT12] Mokshay Madiman, Adam W Marcus, and Prasad Tetali, *Entropy and set cardinality inequalities for partition-determined functions*, Random Structures & Algorithms **40** (2012), no. 4, 399–424.

[MMX17] Mokshay Madiman, James Melbourne, and Peng Xu, *Forward and reverse entropy power inequalities in convex geometry*, Convexity and Concentration (New York, NY) (Eric Carlen, Mokshay Madiman, and Elisabeth M. Werner, eds.), Springer New York, 2017, pp. 427–485.

[MT10] Mokshay Madiman and Prasad Tetali, *Information inequalities for joint distributions, with interpretations and applications*, IEEE Transactions on Information Theory **56** (2010), no. 6, 2699–2713.

[Pet12] Giorgis Petridis, *New proofs of Plünnecke-type estimates for product sets in groups*, Combinatorica **32** (2012), no. 6, 721–733.

[Rad03a] J. Radhakrishnan, *Entropy and counting*, Computational Mathematics, Modelling and Algorithms **146** (2003).

[Rad03b] Jaikumar Radhakrishnan, *Entropy and counting*, Computational mathematics, modelling and algorithms **146** (2003).

[Rén59] A. Rényi, *On measures of dependence*, Acta. Math. Acad. Sci. Hung. **10** (1959), 441–451.

[Rio11] O. Rioul, *Information theoretic proofs of entropy power inequalities*, Information Theory, IEEE Transactions on **57** (2011), no. 1, 33–55.

[Ruz96] Imre Z Ruzsa, *Sums of finite sets, number theory (new york, 1991–1995), 281–293*, 1996.

[Ruz09a] ______, *Sumsets and entropy*, Random Structures & Algorithms **34** (2009), no. 1, 1–10.

[Ruz09b] ______, *Sumsets and structure*, Combinatorial number theory and additive group theory (2009), 87–210.

[Sas22] Igal Sason, *Information inequalities via submodularity and a problem in extremal graph theory*, Entropy **24** (2022), no. 5.

[SDM11] Naresh Sharma, Smarajit Das, and Siddharth Muthukrishnan, *Entropy power inequality for a family of discrete random variables*, 2011 IEEE International Symposium on Information Theory Proceedings, 2011, pp. 1945–1949.

[SG22] Erixhen Sula and Michael Gastpar, *The Gray-Wyner network and Wyner's common information for Gaussian sources*, IEEE Transactions on Information Theory **68** (2022), no. 2, 1369–1384.

[Sha48] C E Shannon, *A mathermatical theory of communication*, Bell System Technical Journal **27** (July and October, 1948), 379–423 and 623–656.

[Ski53] Viktor P Skitovitch, *On a property of the normal distribution*, DAN SSSR **89** (1953), 217–219.

[Sta59] A.J. Stam, *Some inequalities satisfied by the quantities of information of Fisher and Shannon*, Information and Control **2** (1959), no. 2, 101–112.

[SW90] S. Shamai and A.D. Wyner, *A binary analog to the entropy-power inequality*, IEEE Transactions on Information Theory **36** (1990), no. 6, 1428–1430.

[Tao10] Terence Tao, *Sumset and inverse sumset theory for Shannon entropy*, Combinatorics, Probability and Computing **19** (2010), no. 4, 603–639.

[Tia11] Chao Tian, *Inequalities for entropies of sets of subsets of random variables*, 2011 IEEE International Symposium on Information Theory Proceedings, 2011, pp. 1950–1954.

[TV] Terence Tao and Van Vu, *Deleted scenes: entropy sumset estimates*, available at: https://terrytao.wordpress.com/books/additive-combinatorics/.

[WSS06] H. Weingarten, Y. Steinberg, and S. S. Shamai, *The capacity region of the Gaussian multiple-input multiple-output broadcast channel*, IEEE Transactions on Information Theory **52** (2006), no. 9, 3936–3964.

[WZ73] A. Wyner and J. Ziv, *A theorem on the entropy of certain binary* *sequences and applications: Part I*, IEEE Trans. Inform. Theory **IT-19** (1973), no. 6, 769–772.

[Yu09] Yaming Yu, *Monotonic convergence in an information-theoretic law of small numbers*, IEEE Transactions on Information Theory **55** (2009), no. 12, 5412–5422.

[ZY98] Z. Zhang and R.W. Yeung, *On characterization of entropy function via information inequalities*, IEEE Transactions on Information Theory **44** (1998), no. 4, 1440–1452.