Extremal Entropy: Information Geometry, Numerical Entropy Mapping, and Machine Learning Application of Associated Conditional Independences

A Thesis

Submitted to the Faculty

of

Drexel University

by

Yunshu Liu

in partial fulfillment of the

requirements for the degree

of

Doctor of Philosophy

April 2016

© Copyright 2016
Yunshu Liu. All Rights Reserved.

Acknowledgments

I would like to express the deepest appreciation to my advisor Dr. John MacLaren Walsh for his guidance, encouragement and enduring patience over the last few years. I would like to thank my committee members, Dr. Steven Weber, Dr. Naga Kandasamy, Dr. Hande Benson and Dr. Andrew Cohen for their advice and constant support. In addition, I want to thank my colleagues in the Adaptive Signal Processing and Information Theory Research Group (ASPITRG) for their support and all the fun we have had in the last few years. I also thank all the faculty, staffs and students at Drexel University who had helped me in my study, research and life. Last but not the least, I would like to thank my parents: my mother Han Zhao and my father Yongsheng Liu, for their everlasting love and support.

Table of Contents

LIST OF TABLES iv
LIST OF FIGURES v
ABSTRACT vii
1. Introduction 1
2. Bounding the Region of Entropic Vectors 4
   2.1 Outer Bounds of $\overline{\Gamma}^*_N$ 6
   2.2 Inner Bounds of $\overline{\Gamma}^*_N$ 8
   2.3 Semimatroids and Faces of Γ_N and Γ^*_N 10
   2.4 The gap between $\mathcal{I}_4$ and Γ_4 13
3. Understanding the structure of Entropic region by projection 16
   3.1 Derivation of a single nonlinear function 16
   3.2 Properties of $g^{up}_{\mathcal{A}}(\mathbf{h}_{\backslash\mathcal{A}})$ and $g^{low}_{\mathcal{A}}(\mathbf{h}_{\backslash\mathcal{A}})$ 20
4. Listing Canonical k-atom Supports and map the Entropy Region with them 26
   4.1 Non-isomorphic k-atom supports via Snakes and Ladders 29
   4.2 Maximal Ingleton Violation and the Four Atom Conjecture 34
   4.3 Optimizing Inner Bounds to Entropy from k-Atom Distributions 36
5. Perspectives on How to Use Information Geometry to Understand Entropy Geometry 44
   5.1 Review of Some Relevant Ideas from Information Geometry 45
   5.2 Information Geometric Structure of the Shannon Faces of the Region of Entropic Vectors 49
   5.3 Information Geometric Structure of Ingleton-Violating Entropic Vectors & their Distributions 62
6. Information Guided Dataset Partitioning for Supervised Learning 67
   6.1 Preliminaries, Related works and Problem setup 69
   6.2 Proposed Partition Criteria and Algorithms 78
   6.3 Experimental Results and Discussions 83
      6.3.1 Click-Through Rate(CTR) prediction 84
      6.3.2 Customer relationship prediction 87
      6.3.3 Discussion of Experimental Results 90
7. Conclusions 92
BIBLIOGRAPHY 94

List of Tables

4.1 # of non-isomorphic k-atom, N-variable supports. 33

4.2 # of non-isomorphic k-atom, 4-variable supports that can violate the Ingleton inequality. 34

4.3 The volume increase within pyramid G_4^{34} as more atoms are included 38

6.1 Frequent feature-value pairs of CTR data 85

6.2 Scores of different partitions for CTR data 86

6.3 Classification results of different partitions for CTR data 86

6.4 Frequent feature-value pairs of customer relationship data 88

6.5 Scores of different partitions for customer relationship data 89

6.6 Classification results of different partitions for customer relationship data . 90

List of Figures

3.1 The extreme rays of G_4^{34}. The top row is the ray f_{34}, and all of its coefficients except in the red column (corresponding to h_{123}) are the sum of the entries in the green rows. Hence $\pi_{\backslash h_{123}} G_4^{34}$ is entirely entropic. 18

3.2 The coefficients of the non-redundant inequalities in G_4^{34}. Note that in each column where $Ingleton_{34}$ has a non-zero coefficient, it is the only coefficient with its sign. 19

3.3 Entropic vector hyperplane with only h_1 and h_2 coordinate not fixed 25

4.1 Lattice of $\Pi(\mathbb{N}_1^4)$: the set of all set partitions for $k = 4$ 31

4.2 The surface is a projection of a 3D face of the inner bound to $\bar{\Gamma}_4^*$ created with the numerical process described in the text with 4-atom distributions. The blue circles are the from the entropic vectors from the extremal optimized four atom distributions, while the red Xs and the black squares are the additional extremal optimized k distributions for $k \in \{5, 6\}$, respectively. 40

4.3 Comparison of contour plots of inner bound created from $\leq k$ atom distributions for $k \in \{4, 5, 6\}$. For each contour value, the inner most line is $k = 4$ while the outermost line is $k = 6$. The numerical inner bounds generated from only four atom distributions are quite good in this space... 41

4.4 The surface is the bound created in [1] while the markers above it are those created from the generated non-isomorphic supports of size 4,5, and 6 in this section. The blue circles are from the entropic vectors from the extremal optimized four atom distributions, while the red Xs and the black squares are the additional extremal optimized k distributions for $k \in \{5, 6\}$, respectively. The low cardinality supports generate much of the surface [1], and even slightly improve it is some parts, despite being far lower cardinality and not even being optimized in this space..... 43

5.1 Pythagorean style relation. 49

5.2 Region of entropic vector Γ_2^*. 50

5.3 Submodularity of the entropy function is equivalent to the non-negativity of a divergence $D(\overrightarrow{\pi}_{\mathcal{E}^{\perp}_{\mathcal{A}\cup\mathcal{B}}}(p)||\ \overrightarrow{\pi}_{\mathcal{E}^{\leftrightarrow,\perp}_{\mathcal{A},\mathcal{B}}}(p))$ between two information projections, one projection $(\overrightarrow{\pi}_{\mathcal{E}^{\leftrightarrow,\perp}_{\mathcal{A},\mathcal{B}}}(p))$ is to a set that is a submanifold the other projection's $(\overrightarrow{\pi}_{\mathcal{E}^{\perp}_{\mathcal{A}\cup\mathcal{B}}}(p))$ set. A Pythagorean style relation shows that for such an arrangement $D(p||\ \overrightarrow{\pi}_{\mathcal{E}^{\leftrightarrow,\perp}_{\mathcal{A},\mathcal{B}}}(p)) = D(p||\ \overrightarrow{\pi}_{\mathcal{E}^{\perp}_{\mathcal{A}\cup\mathcal{B}}}(p)) + D(\overrightarrow{\pi}_{\mathcal{E}^{\perp}_{\mathcal{A}\cup\mathcal{B}}}(p)||\ \overrightarrow{\pi}_{\mathcal{E}^{\leftrightarrow,\perp}_{\mathcal{A},\mathcal{B}}}(p))$. 56

5.4 Pythagorean relationship on 3 discrete random variables 58

5.5 The mapping between some faces of Γ_4 and submanifold of four variable distribution in $\boldsymbol{\theta}$ coordinate 61

5.6 Manifold $\mathcal{D}_{4atom}$ in θ coordinate 64

5.7 G_4^{34}: one of the six gaps between $\mathcal{I}_4$ and Γ_4 65

5.8 Slice of $\mathcal{D}_{5atom}$ in θ coordinate 66

6.1 Example of a Bayesian Network 70

6.2 Example of a Markov Random Fields 71

6.3 Example of a sum-product network 73

6.4 Context-Specific Independence of a dataset in Sum-Product Network 79

6.5 Partition of a dataset viewed as Sum-Product Network 81

Abstract

Extremal Entropy: Information Geometry, Numerical Entropy Mapping, and Machine Learning Application of Associated Conditional Independences

Yunshu Liu

Advisor: John MacLaren Walsh, Ph.D.

Entropy and conditional mutual information are the key quantities information theory provides to measure uncertainty of and independence relations between random variables. While these measures are key to diverse areas such as physics, communication, signal processing, and machine learning, surprisingly there is still much about them that is yet unknown. This thesis explores some of this unknown territory, ranging from tackling fundamental questions involving the interdependence between entropies of different subsets of random variables via the characterization of the region of entropic vectors, to applied questions involving how conditional independences can be harnessed to improve the efficiency of supervised learning in discrete valued datasets.

The region of entropic vectors is a convex cone that has been shown to be at the core of many fundamental limits for problems in multiterminal data compression, network coding, and multimedia transmission. This cone has been shown to be non-polyhedral for four or more random variables, however its boundary remains unknown for four or more discrete random variables. We prove that only one form of nonlinear non-shannon inequality is necessary to fully characterize the region for four random variables. We identify this inequality in terms of a function that is the solution to an optimization problem. We also give some symmetry and convexity properties of this function which rely on the structure of the region of entropic vectors and Ingleton inequalities. Methods for specifying probability distributions that are in faces and on the boundary of the convex cone are derived, then utilized to map optimized

inner bounds to the unknown part of the entropy region. The first method utilizes tools and algorithms from abstract algebra to efficiently determine those supports for the joint probability mass functions for four or more random variables that can, for some appropriate set of non-zero probabilities, yield entropic vectors in the gap between the best known inner and outer bounds. These supports are utilized, together with numerical optimization over non-zero probabilities, to provide inner bounds to the unknown part of the entropy region. Next, information geometry is utilized to parameterize and study the structure of probability distributions on these supports yielding entropic vectors in the faces of entropy and in the unknown part of the entropy region.

In the final section of the thesis, we propose score functions based on entropy and conditional mutual information as components in partition strategies for supervised learning of datasets with discrete valued features. Partitioning the data enables a reduction in the complexity of training and testing on large datasets. We demonstrate that such partition strategies can also be efficient in the sense that when the training and testing datasets are split according to them, and the blocks in the partition are processed separately, the classification performance is comparable to, or better than, the performance when the data are not partitioned at all.

1. Introduction

Entropy and conditional mutual information are the key quantities information theory provides to measure uncertainty of and independence relations between random variables. While these measures are key to diverse areas such as physics, communication, signal processing, and machine learning, surprisingly there is still much about them that is yet unknown. This thesis explores some of this unknown territory, ranging from tackling fundamental questions involving the interdependence between entropies of different subsets of random variables via the characterization of the region of entropic vectors, to applied questions involving how conditional independences can be harnessed to improve the efficiency of supervised learning in datasets with discrete valued features.

The region of entropic vectors is a convex cone that is known to be at the core of many yet undetermined fundamental limits for problems in data compression, network coding, and multimedia transmission. This set has been shown to be non-polyhedral, but its boundaries remain unknown. Much of this dissertation develops methods of parameterizing probability distributions that yield entropic vectors in this unknown part of the entropy region.

In §2, after reviewing the definition of this set and its importance in applications, we discuss the best known outer and inner bounds for it, most of the latter of which are based on results for representable matroids and linear polymatroids. These inner bounds, which, as we review, are based on linear constructions, are unable to parameterize the boundary of the region of entropic vectors because non-linear dependence structures are necessary to parameterize the unknown part of it. We also review the concept of a semimatroid and its p-representability, and emphasize its relationship with studying the faces of the entropy region.

Next, in §3, we argue that the complete characterization of $\overline{\Gamma}_4^*$ can be seen as finding a single nonlinear inequality determined by a single nonlinear function. After defining the function as the solution to an optimization problem, we investigate some properties of it.

In §4, we provide a systematic and efficient method for searching for supports for joint probability mass functions which can, for appropriate choices of non-zero probabilities, yield entropic vectors in between the best known outer and inner bounds. Key to this study is determining equivalence classes of supports, within which each of the contained supports are guaranteed, when ranging over all non-zero probabilities, to give the same sets of entropy vectors. Only one representative support from each of these equivalence classes, the canonical representative, is necessary to be considered for the purposes of mapping entropies. These equivalence classes are formalized via the language of group actions, and an efficient algorithm for listing canonical supports is provided. With these canonical supports in hand, we set about determining which of the supports for four random variables yield entropic vectors in the gap between the best known inner and outer bounds on the entropy region. Via numerical optimization over the probabilities on these supports, tuned inner bounds to the unknown part of the entropy region are also provided.

Seeking to gain a better understanding of the structure of probability distributions that would yield extremal entropic vectors in this unknown part of the entropy region, we shift in §5 to studying the information geometric properties of probability distributions associated with entropic vectors that are extremal in the sense that they live in faces of the widely used outer bound. Furthermore, those distributions on the smallest support that can yield entropic vectors in the unknown part of the entropy region are shown to have a special information geometric structure, which we also show disappears for larger supports having this property.

Shifting to a more applied focus, inspired by the theory of conditional independences reviewed in the study of the entropy region, in the final section of the the thesis §6, we are proposing the use information measures like entropy and conditional mutual information in the design of score functions for partitioning discrete valued datasets. The aim of the partition strategy will not be for the purposes of clustering, rather, for supervised learning. Such a partition strategy will be deemed to be efficient if, when we split the original dataset based on the partition that gives optimal score, then training classifiers on the different parts separately, the classification accuracy on the test dataset either remains constant or improves relative to a scheme which does not partition the data. Such a scheme enables not only the parallelization of the training process, which is crucial for large datasets, but also adds diversity to training models which has the potential to improve the overall classification accuracy.

The thesis concludes in §7 with a number of interesting directions for further investigation.

2. Bounding the Region of Entropic Vectors

Consider N discrete random variables $\boldsymbol{X} = \boldsymbol{X}_{\mathcal{N}} = (X_1, \ldots, X_N)$, $\mathcal{N} = \{1, \ldots, N\}$ with joint probability mass function $p_{\boldsymbol{X}}(\boldsymbol{x})$. To every non-empty subset of these random variables $\boldsymbol{X}_{\mathcal{A}} := (X_n \mid n \in \mathcal{A})$, $\mathcal{A} \subseteq \mathcal{N}$, there is associated a Shannon entropy $H(\boldsymbol{X}_{\mathcal{A}})$ calculated from the marginal distribution $p_{\boldsymbol{X}_{\mathcal{A}}}(\boldsymbol{x}_{\mathcal{A}}) = \sum_{\boldsymbol{x}_{\mathcal{N}\setminus\mathcal{A}}} p_{\boldsymbol{X}_{\mathcal{N}}}(\boldsymbol{x})$ via

$$H(\boldsymbol{X}_{\mathcal{A}}) = \sum_{\boldsymbol{x}_{\mathcal{A}}} -p_{\boldsymbol{X}_{\mathcal{A}}}(\boldsymbol{x}_{\mathcal{A}}) \log_2 p_{\boldsymbol{X}_{\mathcal{A}}}(\boldsymbol{x}_{\mathcal{A}}) \tag{2.1}$$

One can stack these entropies of different non-empty subsets into a $2^N - 1$ dimensional vector $\boldsymbol{h} = (H(\boldsymbol{X}_{\mathcal{A}}) | \mathcal{A} \subseteq \mathcal{N})$, which can be clearly viewed as $\boldsymbol{h}(p_{\boldsymbol{X}})$, a function of the joint distribution $p_{\boldsymbol{X}}$. A vector $\boldsymbol{h}_? \in \mathbb{R}^{2^N-1}$ is said to be *entropic* if there is some joint distribution $p_{\boldsymbol{X}}$ such that $\boldsymbol{h}_? = \boldsymbol{h}(p_{\boldsymbol{X}})$. The region of entropic vectors is then the image of the set $\mathcal{D} = \{p_{\boldsymbol{X}} | p_{\boldsymbol{X}}(\boldsymbol{x}) \geq 0,\ \sum_{\boldsymbol{x}} p_{\boldsymbol{X}}(\boldsymbol{x}) = 1\}$ of valid joint probability mass functions under the function $\boldsymbol{h}(\cdot)$, and is denoted by

$$\Gamma_N^* = \boldsymbol{h}(\mathcal{D}) \subsetneq \mathbb{R}^{2^N-1} \tag{2.2}$$

It is known that the closure of this set $\overline{\Gamma}_N^*$ is a convex cone [2], but surprisingly little else is known about this cone for $N \geq 4$. Understanding the "shape" and boundaries of the set $\overline{\Gamma}_N^*$ is the subject of many sections in my thesis.

The fundamental importance of $\overline{\Gamma}_N^*$ lies in several contexts in signal processing, compression, network coding and information theory [2]. Firstly, the fundamental rate region limits of many data compression problems that otherwise remain unsolved can be directly expressed in terms of $\overline{\Gamma}_N^*$ and related entropy vector sets. In

particular, a class of problems studied by Zhang and Yeung under the name of distributed source coding for satellite communications [2, 3] (YZDSC) have rate regions directly expressible as projections of $\overline{\Gamma}_N^*$ after intersection with linear entropy equalities associated with certain Markov chain conditions. Secondly, the multiple multicast capacity region for a lossless network under network coding can also be directly expressed in terms of $\overline{\Gamma}_N^*$, as was proved in [4] and [2, 5]. The multi-source network coding (MSNC) problem generalizes the YZDSC problem by allowing for an arbitrary topology of intermediate encoding nodes between the source encoders and decoders. If one can determine the boundaries of $\overline{\Gamma}_N^*$ one can determine the capacity region of any network under network coding. Furthermore, Chan and Grant [6, 7, 8] proved that every part of the (unknown) boundary of $\overline{\Gamma}_N^*$ has a network whose capacity region depends on it. Hence, they have established that the problem of determining the boundaries of $\overline{\Gamma}_N^*$ is equivalent to the problem of calculating the capacity region of every network under network coding.

Even more generally, as *all* achievable rate regions in information theory are expressed in terms of information measures between random variables obeying certain distribution constraints, they can be expressed as linear projections of entropy vectors associated with these *constrained* random variables. In the case of network coding and the distributed source coding problems introduced in the previous two sections, these constraints are embodied solely as entropy constraints and conditional entropies being zero, and hence can handle, after a great deal of proof, the constraints solely in entropy space and work with the region of unconstrained entropic vectors. However, as one moves to more general multi-terminal information theory problems, particular distributions, either marginal or conditional, are typically specified, and in many cases there are also distortion/expectation of some function of a collection of the variables constraints. Such constraints which can only be expressed in probability distribu-

tion space make the problem even harder, and a modified region of entropic vectors $\overline{\Gamma}^*_N(\mathcal{C})$ will need to be employed incorporating distribution constraints $\mathcal{C}$. However, given the wide array of multiterminal information theory whose rate regions can be directly expressed in terms of the simpler unconstrained region $\overline{\Gamma}^*_N$, it makes sense to attempt to bound and characterize it first. Thus, in the next few sections we will attempt to review what is known about bounding $\overline{\Gamma}^*_N$, as sectionwell as introduce some new relevant ideas from abstract algebra, combinatorics, and information geometry to understand these bounds, and enabling the creation of new even better bounds.

2.1 Outer Bounds of $\overline{\Gamma}^*_N$

Viewed as a function $h_{\mathcal{A}} = H(\boldsymbol{X}_{\mathcal{A}})$ of the selected subset, with the convention that $h_{\emptyset} = 0$, entropy is *sub-modular* [2, 9], meaning that

$$h_{\mathcal{A}} + h_{\mathcal{B}} \geq h_{\mathcal{A}\cap\mathcal{B}} + h_{\mathcal{A}\cup\mathcal{B}} \quad \forall \mathcal{A}, \mathcal{B} \subseteq \mathcal{N}, \tag{2.3}$$

and is also *non-decreasing* and *non-negative*, meaning that

$$h_{\mathcal{K}} \geq h_{\mathcal{W}} \geq 0 \quad \forall\ \mathcal{W} \subseteq \mathcal{K} \subseteq \mathcal{N}. \tag{2.4}$$

Viewed as requirements for arbitrary set functions (not necessarily entropy) the inequalities (2.3) and (2.4) are known as the *polymatroidal axioms* [2, 9], and a function obeying them is called the *rank function* of a *polymatroid*. If a set function $\boldsymbol{f}$ that obeys the polymatroidal axioms (2.3) and (2.4) additionally obeys

$$f_{\mathcal{A}} \leq |\mathcal{A}|, \quad f_{\mathcal{A}} \in \mathbb{Z} \quad \forall \mathcal{A} \subseteq \mathcal{N} \tag{2.5}$$

then it is the rank function of a *matroid* on the ground set $\mathcal{N}$.

Since entropy must obey the polymatroidal axioms, the set of all rank functions of polymatroids forms an outer bound for the region of entropic vectors which is often denoted by

$$\Gamma_N = \left\{ \boldsymbol{h} \;\middle|\; \begin{array}{c} \boldsymbol{h} \in \mathbb{R}^{2^N-1} \\ h_{\mathcal{A}} + h_{\mathcal{B}} \geq h_{\mathcal{A}\cap\mathcal{B}} + h_{\mathcal{A}\cup\mathcal{B}} \;\forall \mathcal{A}, \mathcal{B} \subseteq \mathcal{N} \\ h_{\mathcal{K}} \geq h_{\mathcal{W}} \geq 0 \quad \forall\, \mathcal{W} \subseteq \mathcal{K} \subseteq \mathcal{N} \end{array} \right\} \tag{2.6}$$

In fact, any inequality in (2.6) can be expressed as a sum of the following two types of elemental inequalities[2]

$$\begin{aligned} & h_{\mathcal{N}} - h_{\mathcal{N}\backslash i} \geq 0, i \in \mathcal{N} \\ & h_{i\mathcal{K}} + h_{j\mathcal{K}} - h_{\mathcal{K}} - h_{ij\mathcal{K}} \geq 0, \text{for } i \neq j, \mathcal{K} \subset \mathcal{N}\backslash ij \end{aligned} \tag{2.7}$$

The inequalities (2.7) are the minimal, non-redudant, set of information inequalities for defining Γ_N. As we can see from the definition, Γ_N is a polyhedron, and this polyhedral set is often known as the Shannon outer bound for $\overline{\Gamma}_N^*$ [2, 9].

While in the low dimensional cases we have $\Gamma_2 = \Gamma_2^*$ and $\Gamma_3 = \overline{\Gamma}_3^*$, for $N \geq 4$, $\Gamma_N \neq \overline{\Gamma}_N^*$. Zhang and Yeung first showed this in [9] by proving a new inequality among 4 variables

$$2I(C;D) \leq I(A;B) + I(A;C,D) + 3I(C;D|A) + I(C;D|B) \tag{2.8}$$

which held for entropies and was not implied by the polymatroidal axioms, and which they dubbed a *non-Shannon type* inequality to distinguish it from inequalities implied by Γ_N. For roughly the next decade a few authors produced other new non-Shannon inequalities [10, 11]. In 2007, Matúš [12] showed that $\overline{\Gamma}_N^*$ is not a polyhedron for $N \geq 4$. The proof of this fact was carried out by constructing a sequence of non-Shannon inequalities, including

$$s[I(A;B|C)+I(A;B|D)+I(C;D)-I(A;B)] \tag{2.9}$$
$$+I(B;C|A)+\frac{s(s+1)}{2}[I(A;C|B)+I(A;B|C)]\geq 0$$

Notice (2.9) is the same as Zhang-Yeung inequality (2.8) when $s=1$. Additionally, the infinite sequence of inequalities was used with a curve constructed from a particular form of distributions to prove $\overline{\Gamma}_N^*$ is not a polyhedron. Despite this result, even $\overline{\Gamma}_4^*$ is still not fully understood. Since then, many authors has been investigating the properties of $\overline{\Gamma}_N^*$ with the hope of ultimately fully characterizing the region [1, 13, 14, 15, 16, 17].

2.2 Inner Bounds of $\overline{\Gamma}_N^*$

Shifting from outer bounds to bounding from the inside, the most common way to generate inner bounds for the region of entropic vectors is to consider special families of distributions for which the entropy function is known to have certain properties. [16, 18, 19, 20] focus on calculating inner bounds based on special properties of binary random variables. However, the most common way to generate inner bounds is based on inequalities for representable matroids [21], boolean polymatroids [22, 23] and subspace arrangements.

For the latter method, we first introduce some basics in linear polymatroids and the Ingeton inner bound. Fix a $N'\geq N$, and partition the set $\{1,\ldots,N'\}$ into N disjoint sets $\mathcal{Q}_1,\ldots,\mathcal{Q}_N$. Let $\boldsymbol{U}$ be a length m row vector whose elements are i.i.d. uniform over the finite field $GF(q)$, and let $\mathbf{T}$ be a particular $m\times N'$ deterministic

matrix with elements in $GF(q)$. Consider the N' dimensional vector

$$\boldsymbol{Y} = \boldsymbol{U}\mathbf{T}, \text{ and define } \boldsymbol{X}_i = \boldsymbol{Y}_{\mathcal{Q}_i}, \ i \in \{1, \ldots, N\}.$$

The subset entropies of the random variables $\{\boldsymbol{X}_i\}$ obey

$$H(\boldsymbol{X}_{\mathcal{A}}) = g(\mathcal{A}) \log_2(q) = \text{rank}\left([\mathbf{T}_{\mathcal{Q}_i} \,|i \in \mathcal{A}]\right) \log_2(q). \tag{2.10}$$

A set function $\boldsymbol{g}(\cdot)$ created in such a manner is called a linear polymatriod or a subspace rank function. It obeys the polymatroidal axioms, and is additionally proportional to an integer valued vector. If the $\mathcal{Q}_i$ are all singletons and $N' = N$, then this set function is proportional (via $\log q$) to the rank function of a representable matroid [24]. Alternatively, when the sets $\mathcal{Q}_i$ are not singletons and $N' > N$, such a construction is clearly related to a representable matroid on a larger ground set[24]. Indeed, the subspace rank function vector is merely formed by taking some of the elements from the $2^{N'} - 1$ representable matroid rank function vector associated with $\mathbf{T}$. That is, set function vectors created via (2.10) are ($\log_2 q$ scaled) projections of rank function vectors of representable matroids.

Set functions capable of being represented in this manner for some N', q and $\mathbf{T}$, are called subspace ranks in some contexts [25, 26, 27], while other papers effectively define a collection of vector random variables created in this manner a subspace arrangement [28].

Define $\mathcal{I}_N$ to be the conic hull of all subspace ranks for N subspaces. It is known that $\mathcal{I}_N$ is an inner bound for $\overline{\Gamma}^*_N$[25], which we name the subspace inner bound. So far $\mathcal{I}_N$ is only known for $N \leq 5$ [27, 28]. More specifically, $\mathcal{I}_2 = \overline{\Gamma}^*_2 = \Gamma_2$, $\mathcal{I}_3 = \overline{\Gamma}^*_3 = \Gamma_3$. As with most entropy vector sets, things start to get interesting at $N = 4$ variables (subspaces). For $N = 4$, $\mathcal{I}_4$ is given by the Shannon type inequalities

(i.e. the polymatroidal axioms) together with six additional inequalities known as *Ingleton*'s inequality [25, 26, 29] which states that for $N = 4$ random variables

$$Ingleton_{ij} \geq 0 \tag{2.11}$$

where

$$\begin{aligned} Ingleton_{ij} \quad &= I(X_k; X_l|X_i) + I(X_k; X_l|X_j) \\ &+ I(X_i; X_j) - I(X_k; X_l) \end{aligned}$$

Thus, $\mathcal{I}_4$ is usually called the Ingleton inner bound.

2.3 Semimatroids and Faces of Γ_N and Γ_N^*

In [26], the concept of *semimatroid* is introduced to help analyze the conditional independences among four random variables. In this section, we first review the definition of a *semimatroid*; then some results in [26] that are highly related to the structure of the region of entropic vectors on four variables will be presented; at last, we build the mapping between subset of extreme rays of Γ_4 and some particular *p-representable* semimatroids, which we will use information geometry to analyze in section §5.

Let $\mathcal{N} = \{1, 2, \cdots N\}$ and $\mathcal{S}$ be the family of all couples $(i, j|\mathcal{K})$, where $\mathcal{K} \subset \mathcal{N}$ and ij is the union of two singletons i and j in $\mathcal{N} \setminus \mathcal{K}$. If we include the cases when $i = j$, there are, for example, 18 such couples for three variables, and 56 such couples for $N = 4$. A relation $\mathcal{L} \subset \mathcal{S}$ is called *probabilistically representable* or *p-representable*

if there exists a system of N random variables $\mathbf{X} = \{X_i\}_{i\in\mathcal{N}}$ such that

$$\mathcal{L} = \{(i,j|\mathcal{K}) \in \mathcal{S}(N) | X_i \text{ is conditionally}$$
$$\text{independent of } X_j \text{ given } X_\mathcal{K} \text{ i.e. } I(X_i; X_j|X_\mathcal{K}) = 0 \ \}.$$

Definition 1: For $\boldsymbol{f} \in \Gamma_N$ we define $|[\boldsymbol{f}]|$ as

$$|[\boldsymbol{f}]| = \{(i,j|\mathcal{K}) \in \mathcal{S}(N) | f_{i\mathcal{K}} + f_{j\mathcal{K}} - f_{ij\mathcal{K}} - f_\mathcal{K} = 0\}. \tag{2.12}$$

A relation $\mathcal{L} \subset \mathcal{S}(N)$ is called a *semimatroid* if and only if $\mathcal{L} = |[\boldsymbol{f}]|$ for some $\boldsymbol{f} \in \Gamma_N$, the Shannon outer bound for N random variables.

We use $Semi(N)$ to denote the set of all semimatroids on N, and we say that semimatroid $\mathcal{L}$, $\mathcal{L} = |[\boldsymbol{f}]|$, *arises from* polymatroid vector $\boldsymbol{f}$. The *p-representable* semimatroids are just those semimatroids arising from an entropic vector $\boldsymbol{h}$. We use $P_{rep}(N)$ to denote the set of all *p-representable* relations on $\mathcal{N}$. For $N \leq 3$, since $\Gamma_N = \overline{\Gamma}_N^*$, we have $P_{rep}(N) = Semi(N)$. However, $P_{rep}(4) \subsetneq Semi(4)$, that is to say, there are semimatroids on four variables that are not *p-representable*. The main theorem of [30] lists all irreducible *p-representable* semimatroids over four variables. There are 120 such semimatroids of 16 types, and every *p-representable* semimatroid is at the intersection of some of these semimatroids. For $\mathcal{N} = \{1,2,3,4\}$, with $\mathcal{K} \subseteq \mathcal{N}$ and $0 \leq t \leq |\mathcal{N}\backslash\mathcal{K}|$, define $\boldsymbol{r}_t^\mathcal{K}$, $\boldsymbol{g}_i^{(2)}$ and $\boldsymbol{g}_i^{(3)}$ such that $\boldsymbol{r}_t^\mathcal{K}(\mathcal{W})$, $\boldsymbol{g}_i^{(2)}(\mathcal{W})$ and $\boldsymbol{g}_i^{(3)}(\mathcal{W})$ as follows:

$$\boldsymbol{r}_t^\mathcal{K}(\mathcal{W}) = \min\{t, |\mathcal{W}\backslash\mathcal{K}|\} \ with \ \mathcal{W} \subseteq \mathcal{N}$$

$$\boldsymbol{g}_i^{(2)}(\mathcal{W}) = \begin{cases} 2 & \text{if } \mathcal{W} = i \\ min\{2, |\mathcal{W}|\} & \text{if } \mathcal{W} \neq i \end{cases}$$

$$\boldsymbol{g}_i^{(3)}(\mathcal{W}) = \begin{cases} |\mathcal{W}| & \text{if } i \notin \mathcal{W} \\ min\{3, |\mathcal{W}|+1\} & \text{if } i \in \mathcal{W} \end{cases}$$

Now we present the main theorem of [30]

Theorem 1: (Matúš)[30] There are 120 irreducible *p-representable* semimatroids of sixteen types over four-element set N. Among which there are 36 ingleton semimatroids of 11 types: $|[0]|$, $|[\boldsymbol{r}_1^{N-i}]|$ for $i \in \mathcal{N}$, $|[\boldsymbol{r}_1^{ij}]|$ for $i, j \in \mathcal{N}$ distinct, $|[\boldsymbol{r}_1^{i}]|$ for $i \in \mathcal{N}$, $|[\boldsymbol{r}_1]|$, $|[\boldsymbol{r}_2^{i}]|$ for $i \in \mathcal{N}$, $|[\boldsymbol{r}_2^{i\|j}]|$ for $i, j \in \mathcal{N}$ distinct, $|[\boldsymbol{r}_2]|$, $|[\boldsymbol{r}_3]|$, $|[\boldsymbol{g}_i^{(2)}]|$ for $i \in \mathcal{N}$, $|[\boldsymbol{g}_i^{(3)}]|$ for $i \in \mathcal{N}$. There are also 84 non-Ingleton semimatroids of 5 types:

$$\begin{aligned}
\mathcal{L}_{ij}^{kl|\emptyset} &= \{(kl|i), (kl|j), (ij|\emptyset), (kl|ij)\} \\
&\quad \cup\{(k|ij), (l|ij), (i|jkl), (j|ikl), (k|ijl), (l|ijk)\} \\
\mathcal{L}_{ij}^{(ij|kl)} &= \{(ij|k), (ij|l), (kl|ij), (kl|i), (kl|j)\} \\
\mathcal{L}_{ij}^{(ik|jl)} &= \{(kl|ij), (ij|k), (ik|l), (kl|j), (l|ij), (l|ijk)\} \\
\mathcal{L}_{ij}^{ik|j} &= \{(ij|k), (ik|l), (kl|j), (i|jkl), (j|ikl)\} \\
&\quad \cup\{(k|ijl), (l|ijk)\} \\
\mathcal{L}_{ij}^{jl|\emptyset} &= \{(kl|i), (jl|k), (ij|\emptyset), (kl|ij)\} \\
&\quad \cup\{(k|ij), (l|ij), (i|jkl), (j|ikl), (k|ijl), (l|ijk)\}
\end{aligned}$$

The theorem not only solved *p-representability* of semimatroids over four variables, but also answered the question of which faces of Γ_4 have interior points that are entropic, which is stated in the following corollary:

Corollary 1: A relation $\mathcal{L}_0 \subseteq \mathcal{S}(N)$ is a *p-representable* semimatroid if and only if

$$\mathcal{F}_0 = \{\ \boldsymbol{h} \in \Gamma_N \mid h_{i\mathcal{K}} + h_{j\mathcal{K}} - h_{\mathcal{K}} - h_{ij\mathcal{K}} = 0,\ \forall (i, j|\mathcal{K}) \in \mathcal{L}\} \tag{2.13}$$

is a face of Γ_N such that there exists $\boldsymbol{h}_0$, a point in the relative interior of $\mathcal{F}_*$, satisfying $\boldsymbol{h}_0 \in \Gamma_N^*$.

Proof: $\Rightarrow$ Suppose relation $\mathcal{L}_0$ is a *p-representable* semimatroid. By definition (2.6) and (2.7),

$$\mathcal{F}_a = \{\ \boldsymbol{h} \in \Gamma_N \mid h_{i\mathcal{K}} + h_{j\mathcal{K}} - h_{\mathcal{K}} - h_{ij\mathcal{K}} = 0\} \tag{2.14}$$

defines a facet of Γ_N if $i \neq j$ or $i = j$ and $K = \mathcal{N} \setminus \{i, j\}$, and an intersection of such facets otherwise. By definition a semimatroid $\mathcal{L}_0$ must be a union of $(i, j|\mathcal{K})$ couples such that there is a polymatroid $\boldsymbol{f} \in \Gamma_N$ which obeys exclusively these independence relations. Thus $\mathcal{F}_0$ defined by any semimatroid must be face of Γ_N because it is a exhaustive list of facets whose intersection forms this face. Furthermore, since $\mathcal{L}_0$ is p-representibile, there exist some collection of random variables, generating an entropic vector in $\Gamma_n^* \cap \mathcal{F}_0$, that does not obey any additional conditional independence relations beyond $\mathcal{L}_0$. This vector is thus in the relative interior of $\mathcal{F}_0$ (for being on a relative boundary of this face would require living in additional facets of Γ_N and thus obeying additional conditional independence relations).
$\Leftarrow$ Now suppose for a given $\mathcal{F}_0$, which is a face of Γ_N, we have $\boldsymbol{h}_0$, a relative interior point of $\mathcal{F}_0$ such that $\boldsymbol{h}_0 \in \Gamma_N^*$. Then the relation corresponding to the information equalities satisfied by $\boldsymbol{h}_0$ must be *p-representable*. □

2.4 The gap between $\mathcal{I}_4$ and Γ_4

For the structure of the gap between $\mathcal{I}_4$ and Γ_4, we know Γ_4 is generated by 28 elemental Shannon type information inequalities[2]. As for $\mathcal{I}_4$, in addition to the the 28 Shannon type information inequalities, we also need six Ingleton's inequalities (2.11), thus $\mathcal{I}_4 \subsetneq \Gamma_4$. In [26] it is stated that Γ_4 is the disjoint union of $\mathcal{I}_4$ and six cones $\{\boldsymbol{h} \in \Gamma_4 | Ingleton_{ij} < 0\}$. The six cones $G_4^{ij} = \{\boldsymbol{h} \in \Gamma_4 | Ingleton_{ij} \leq 0\}$ are

symmetric due to the permutation of inequalities $Ingleton_{ij}$, so it is sufficient to study only one of the cones. Furthermore, [26] gave the extreme rays of G_4^{ij} in Lemma 1 by using $\boldsymbol{r}_t^{\mathcal{K}}$, $\boldsymbol{g}_i^{(2)}$, $\boldsymbol{g}_i^{(3)}$ and the following functions $\boldsymbol{f}_{ij}$:

For $\mathcal{N} = \{1, 2, 3, 4\}$, define $\boldsymbol{f}_{ij}(\mathcal{W})$ as follows:

$$\boldsymbol{f}_{ij}(\mathcal{W}) = \begin{cases} 3 & \text{if } \mathcal{W} \in \{ik, jk, il, jl, kl\} \\ min\{4, 2|\mathcal{W}|\} & \text{otherwise} \end{cases}$$

Lemma 1: (Matúš)[26] The cone $G_4^{ij} = \{\boldsymbol{h} \in \Gamma_4 | Ingleton_{ij} \leq 0, i, j \in \mathcal{N} \text{ distinct}\}$ is the convex hull of 15 extreme rays. They are generated by the 15 linearly independent functions $\boldsymbol{f}_{ij}$, $\boldsymbol{r}_1^{ijk}$, $\boldsymbol{r}_1^{ijl}$, $\boldsymbol{r}_1^{ikl}$, $\boldsymbol{r}_1^{jkl}$, $\boldsymbol{r}_1^{\emptyset}$, $\boldsymbol{r}_3^{\emptyset}$, $\boldsymbol{r}_1^{i}$, $\boldsymbol{r}_1^{j}$, $\boldsymbol{r}_1^{ik}$, $\boldsymbol{r}_1^{jk}$, $\boldsymbol{r}_1^{il}$, $\boldsymbol{r}_1^{jl}$, $\boldsymbol{r}_2^{k}$, $\boldsymbol{r}_2^{l}$, where $kl = \mathcal{N} \backslash ij$.

Note that among the 15 extreme rays of G_4^{ij}, 14 extreme rays $\boldsymbol{r}_1^{ijk}$, $\boldsymbol{r}_1^{ijl}$, $\boldsymbol{r}_1^{ikl}$, $\boldsymbol{r}_1^{jkl}$, $\boldsymbol{r}_1^{\emptyset}$, $\boldsymbol{r}_3^{\emptyset}$, $\boldsymbol{r}_1^{i}$, $\boldsymbol{r}_1^{j}$, $\boldsymbol{r}_1^{ik}$, $\boldsymbol{r}_1^{jk}$, $\boldsymbol{r}_1^{il}$, $\boldsymbol{r}_1^{jl}$, $\boldsymbol{r}_2^{k}$, $\boldsymbol{r}_2^{l}$ are also extreme rays of $\mathcal{I}_4$ and thus entropic, which leaves $\boldsymbol{f}_{ij}$ the only extreme ray in G_4^{ij} that is not entropic[26]. It is easily verified that $\overline{\Gamma}_4^*$ is known as long as we know the structure of six cones $\overline{\Gamma}_4^* \cap G_4^{ij}$. Due to symmetry, we only need to focus on one of the six cones $\overline{\Gamma}_4^* \cap G_4^{34}$, thus we define $P_4^{34} = \overline{\Gamma}_4^* \cap G_4^{34}$. We thus aims to study the properties of supports and probability distributions yielding entropic vectors in P_4^{34}, this gap between the best known inner and outer bounds for region of entropic vectors on four variables.

Next, let's examine the relationship between subset of the extreme rays of G_4^{34} and some particular *p-representable* semimatroids. There are 15 extreme rays in G_4^{34}: $\boldsymbol{r}_1^{13}$, $\boldsymbol{r}_1^{23}$, $\boldsymbol{r}_1^{123}$, $\boldsymbol{r}_1^{124}$, $\boldsymbol{r}_1^{134}$, $\boldsymbol{r}_1^{234}$, $\boldsymbol{r}_1^{\emptyset}$, $\boldsymbol{r}_1^{3}$, $\boldsymbol{r}_1^{4}$, $\boldsymbol{r}_1^{14}$, $\boldsymbol{r}_2^{1}$, $\boldsymbol{r}_1^{24}$, $\boldsymbol{r}_2^{2}$, $\boldsymbol{r}_3^{\emptyset}$, $\boldsymbol{f}_{34}$. We verified that none of the 56 $(i, j|\mathcal{K})$ couples is satisfied by all of the 15 extreme rays. If we remove $\boldsymbol{r}_1^{24}$, then $(1, 3|2)$ is the only relation satisfied by all the rest 14 extreme rays; if we remove both $\boldsymbol{r}_1^{24}$ and $\boldsymbol{r}_1^{14}$, then two relations $\{(1, 3|2)$ & $(2, 3|1)\}$ are satisfied by all

the rest 13 extreme rays; at last if we remove $\boldsymbol{r}_1^{24}$, $\boldsymbol{r}_1^{14}$ and $\boldsymbol{r}_1^{3}$, then the remaining 12 extreme rays all satisfy the set of three relations $\{(1,3|2)\ \&\ (2,3|1)\ \&\ (1,2|3)\}$. From Theorem 1, relations $\{(1,2|3)\}$, $\{(1,2|3)\ \&\ (2,3|1)\}$ and $\{(1,2|3)\ \&\ (1,3|2)\ \&\ (2,3|1)\}$ as semimatroids are all *p-representable*, we can say that the faces of Γ_4 generated by the corresponding subset of the extreme rays all have interior points that are *p-representable*. Furthermore, as we will see in 5.2, the probability distributions associated with these faces of Γ_4 are well characterized in Information Geometry.

3. Understanding the structure of Entropic region by projection

3.1 Derivation of a single nonlinear function

One way to propose the problem of characterizing the entropy region is by the following optimization problem

$$\gamma(\boldsymbol{a}) = \min_{h \in \Gamma_N^*} \sum_{\mathcal{A} \subseteq \mathcal{N}} a_{\mathcal{A}} h_{\mathcal{A}} \tag{3.1}$$

where $a_{\mathcal{A}} \in \mathcal{R}$ and $\boldsymbol{a} = [a_{\mathcal{A}} | \mathcal{A} \subseteq \mathcal{N}]$. The resulting system of inequalities $\{\boldsymbol{a}^T \mathbf{h} \geq \gamma(\boldsymbol{a}) | \quad \forall \boldsymbol{a} \in \mathbb{R}^{2^N - 1}\}$, has each inequality linear in $\mathbf{h}$, and the minimal, non-redundant, subset of these inequalities is uncountably infinite due to the non-polyhedral nature of $\bar{\Gamma}_N^*$. Hence, while solving the program in principle provides a characterization to the region of entropic vectors, the resulting characterization with uncountably infinite cardinality is likely to be very difficult to use.

By studying the conditions on the solution to 3.1, in [31], the authors defined the notion of a *quasi-uniform* distribution and made the following connection between Γ_n^* and Λ_n (the space of entropy vectors generated by quasi-uniform distributions).

Theorem 2: (Chan)[31] The closure of the cone of Λ_n is the closure of Γ_n^* : $\overline{con(\Lambda_n)} = \overline{\Gamma}_n^*$

From Theorem 2, we know finding all entropic vectors associated with quasi-uniform distribution are sufficient to characterize the entropy region, however, determining all quasi-uniform distributions is a hard combinatorial problem, while taking their conic hull and reaching a nonlinear inequality description of the resulting non-polyhedral set appears even harder, perhaps impossible. Thus new methods to simplify the optimization problem should be explored. Our main result in the next

theorem shows that in order to characterize $\overline{\Gamma}_4^*$, we can simplify the optimization problem (3.1) by utilizing extra structure of P_4^{34}.

Theorem 3 (Only one non-Shannon inequality is necessary)**:** To determine the structure of $\overline{\Gamma}_4^*$, it suffices to find a single nonlinear inequality. In particular, select any $h_{\mathcal{A}} \in Ingleton_{ij}$. The region P_4^{ij} is equivalently defined as:

$$P_4^{ij} = \left\{ \boldsymbol{h} \in \mathbb{R}^{15} \;\middle|\; \begin{array}{c} A\boldsymbol{h}_{\backslash\mathcal{A}} \leqslant \boldsymbol{b} \quad (= G_4^{ij} \text{ project out } h_{\mathcal{A}}) \\ h_{\mathcal{A}} \geqslant g_{\mathcal{A}}^{low}(\boldsymbol{h}_{\backslash\mathcal{A}}) \\ h_{\mathcal{A}} \leqslant g_{\mathcal{A}}^{up}(\boldsymbol{h}_{\backslash\mathcal{A}}) \end{array} \right\} \tag{3.2}$$

where $\boldsymbol{h}_{\backslash\mathcal{A}}$ is the 14 dimensional vector excluding $h_{\mathcal{A}}$,

$$g_{\mathcal{A}}^{low}(\boldsymbol{h}_{\backslash\mathcal{A}}) = \min_{[h_{\mathcal{A}} \;\; \boldsymbol{h}_{\backslash\mathcal{A}}^T]^T \in P_4^{ij}} h_{\mathcal{A}}, \tag{3.3}$$

$$g_{\mathcal{A}}^{up}(\boldsymbol{h}_{\backslash\mathcal{A}}) = \max_{[h_{\mathcal{A}} \;\; \boldsymbol{h}_{\backslash\mathcal{A}}^T]^T \in P_4^{ij}} h_{\mathcal{A}}. \tag{3.4}$$

Furthermore, if the coefficient of $h_{\mathcal{A}}$ in $Ingleton_{ij}$ is positive, $h_{\mathcal{A}} \leqslant g_{\mathcal{A}}^{up}(h_{\backslash\mathcal{A}})$ is the inequality $Ingleton_{ij} \leqslant 0$. Similarly, if the coefficient of $h_{\mathcal{A}}$ in $Ingleton_{ij}$ is negative, $h_{\mathcal{A}} \geqslant g_{\mathcal{A}}^{low}(h_{\backslash\mathcal{A}})$ is the inequality $Ingleton_{ij} \leqslant 0$.

Proof: We know G_4^{34} is a 15 dimensional polyhedral cone. Inside this cone, some of the points are entropic, some are not, that is to say, $P_4^{34} \subsetneq G_4^{34}$. From Lemma 1 we obtain the 15 extreme rays of G_4^{34}: f_{34}, r_1^{134}, r_1^{234}, r_1^{123}, r_1^{124}, $r_1^{\emptyset}$, $r_3^{\emptyset}$, r_1^3, r_1^4, r_1^{13}, r_1^{14}, r_1^{23}, r_1^{24}, r_2^1, r_2^2, where each of these extreme rays are 15 dimensional, corresponding to the 15 joint entropy $h_{\mathcal{A}}$ for $\mathcal{A} \subset \mathcal{N}$. The elements of these extreme rays are listed in Fig. 3.1. As shown in Fig. 3.1 with the green rows, if we project out h_{123} from these 15 extreme rays, the only ray which is not entropic, f_{34}, falls into the conic hull of the other 14 entropic extreme rays, that is to say, $\pi_{\backslash h_{123}} P_4^{34} = \pi_{\backslash h_{123}} G_4^{34}$. Furthermore,

h_1	h_2	h_{12}	h_3	h_{13}	h_{23}	h_{123}	h_4	h_{14}	h_{24}	h_{124}	h_{34}	h_{134}	h_{234}	h_{1234}
2	2	3	2	3	3	4	2	3	3	4	4	4	4	4
1	0	1	1	1	1	1	0	1	0	1	1	1	1	1
0	1	1	1	1	2	2	1	1	2	2	2	2	2	2
1	1	1	0	1	1	1	1	1	1	1	1	1	1	1
0	1	1	0	0	1	1	1	1	1	1	1	1	1	1
1	0	1	0	1	0	1	1	1	1	1	1	1	1	1
0	0	0	0	0	0	0	1	1	1	1	1	1	1	1
1	1	1	1	1	1	1	0	1	1	1	1	1	1	1
0	1	1	1	1	1	1	0	0	1	1	1	1	1	1
1	0	1	1	2	1	2	1	2	1	2	2	2	2	2
0	0	0	1	1	1	1	0	0	0	0	1	1	1	1
1	1	2	1	2	2	3	1	2	2	3	2	3	3	3
0	1	1	0	0	1	1	0	0	1	1	0	0	1	1
1	0	1	0	1	0	1	0	1	0	1	0	1	0	1
1	1	1	1	1	1	1	1	1	1	1	1	1	1	1

Figure 3.1: The extreme rays of G_4^{34}. The top row is the ray f_{34}, and all of its coefficients except in the red column (corresponding to h_{123}) are the sum of the entries in the green rows. Hence $\pi_{\backslash h_{123}} G_4^{34}$ is entirely entropic.

one can easily verify that the same statement holds if we drop any one of the 10 joint entropies $h_{\mathcal{A}} \in Ingleton_{34}$ by summing other extreme ray rows to get all but the dropped dimension. This then implies that for $h_{\mathcal{A}} \in Ingleton_{34}$ the projected polyhedron $\pi_{\backslash h_{\mathcal{A}}} G_4^{34}$ (from which the dimension $h_{\mathcal{A}}$ is dropped) is entirely entropic, and hence $\pi_{\backslash h_{\mathcal{A}}} P_4^{34} = \pi_{\backslash h_{\mathcal{A}}} G_4^{34}$. Hence, for some $h_{\mathcal{A}}$ with a non-zero coefficient in $Ingleton_{34}$, given any point $\boldsymbol{h}_{\backslash \mathcal{A}} \in \pi_{\backslash h_{\mathcal{A}}} G_4^{34} (= \pi_{\backslash h_{\mathcal{A}}} P_4^{34})$, the problem of determining whether or not $[h_{\mathcal{A}} \boldsymbol{h}_{\backslash \mathcal{A}}^T]^T$ is an entropic vector in P_4^{34} is equivalent to determining if $h_{\mathcal{A}}$ is compatible with the specified $\boldsymbol{h}_{\backslash \mathcal{A}}$, as P_4^{34} is convex. The set of such compatible $h_{\mathcal{A}}$s must be an interval $[g^{low}(\boldsymbol{h}_{\backslash \mathcal{A}}), g^{up}(\boldsymbol{h}_{\backslash \mathcal{A}})]$ with functions defined via (3.3) and (3.4). This concludes the proof of (3.2).

To see why one of the two inequalities in (3.4),(3.3) is just the Ingleton inequality $Ingleton_{34}$, observe that for the case of dropping out h_{123}, the only lower bound for h_{123} in G_4^{34} is given by $Ingleton_{34} \leqslant 0$ (all other inequalities have positive coefficients

h_1	h_2	h_{12}	h_3	h_{13}	h_{23}	h_{123}	h_4	h_{14}	h_{24}	h_{124}	h_{34}	h_{134}	h_{234}	h_{1234}
0	0	0	1	0	0	0	1	0	0	0	−1	0	0	0
−1	0	0	0	1	0	0	0	1	0	0	0	−1	0	0
−1	0	1	0	0	0	0	0	1	0	−1	0	0	0	0
−1	0	1	0	1	0	−1	0	0	0	0	0	0	0	0
0	−1	0	0	0	1	0	0	0	1	0	0	0	−1	0
0	−1	1	0	0	0	0	0	0	1	−1	0	0	0	0
0	−1	1	0	0	1	−1	0	0	0	0	0	0	0	0
0	0	0	−1	1	1	−1	0	0	0	0	0	0	0	0
0	0	0	0	0	0	−1	0	0	0	0	0	0	0	1
0	0	0	0	0	0	0	−1	1	1	−1	0	0	0	0
0	0	0	0	0	0	0	0	0	0	−1	0	0	0	1
0	0	0	0	0	0	0	0	0	0	0	−1	1	1	−1
0	0	0	0	0	0	0	0	0	0	0	0	−1	0	1
0	0	0	0	0	0	0	0	0	0	0	0	0	−1	1
1	1	−1	0	−1	−1	1	0	−1	−1	1	1	0	0	0

Figure 3.2: The coefficients of the non-redundant inequalities in G_4^{34}. Note that in each column where $Ingleton_{34}$ has a non-zero coefficient, it is the only coefficient with its sign.

for this variable in the non-redundant inequality description of G_4^{34} depicted in Fig. 3.2). Thus, if $\mathbf{h} \in P_4^{34}$, then $\mathbf{h} \in G_4^{34}$, and

$$h_{123} \geqslant g_{123}^{low}(\mathbf{h}_{\backslash 123}) \geqslant -h_1 - h_2 + h_{12} + h_{13} + h_{23} + h_{14} + h_{24} - h_{124} - h_{34}$$

Furthermore, $\{Ingleton_{34}{=}0 \cap G_4^{34}\} = \{Ingleton_{34}{=}0 \cap P_4^{34}\}$ since all $\{Ingleton_{34} = 0\}$ rays of the outer bound G_4^{34} are entropic, and there is only one ray with a non-zero $Ingleton_{34}$, so the extreme rays of $\{Ingleton_{34}{=}0 \cap G_4^{34}\}$ are all entropic. This means that for any $\mathbf{h}_{\backslash 123} \in \pi_{\backslash 123} G_4^{34}$, the minimum for h_{123} specified by $Ingleton_{34}$ is attainable, and hence $g_{123}^{low}(\mathbf{h}_{\backslash 123}) = -h_1 - h_2 + h_{12} + h_{13} + h_{23} + h_{14} + h_{24} - h_{124} - h_{34}$.

Thus, the problem of determining $\overline{\Gamma}_4^*$ is equivalent to determining a single non-linear function $g_{123}^{up}(\mathbf{h}_{\backslash 123})$. A parallel proof applied for other $h_{\mathcal{A}}$ with a non-zero coefficient in $Ingleton_{ij}$ yields the remaining conclusions. ■

From Theorem 3, we have ten nonlinear inequalities (depending on which $\mathcal{A}$ with $h_{\mathcal{A}}$ appearing in $Ingleton_{ij}$ is selected), any single one of which completely determines P_4^{ij}, and thus, with its six permutations, determine $\overline{\Gamma}_4^*$. This theorem largely simplifies the optimization problem of determining $\overline{\Gamma}_4^*$, in that we only need to work on maximizing or minimizing a single entropy $h_{\mathcal{A}}$ given any $\mathbf{h}_{\backslash\mathcal{A}}$ in the polyhedral cone $\pi_{\backslash h_{\mathcal{A}}} G_4^{ij}$, which is entirely entropic.

3.2 Properties of $g_{\mathcal{A}}^{up}(\mathbf{h}_{\backslash\mathcal{A}})$ and $g_{\mathcal{A}}^{low}(\mathbf{h}_{\backslash\mathcal{A}})$

Based on the analysis in the above section, once we know any one of the ten nonlinear functions, $g_1^{up}(\mathbf{h}_{\backslash 1})$, $g_2^{up}(\mathbf{h}_{\backslash 2})$, $g_{34}^{up}(\mathbf{h}_{\backslash 34})$, $g_{123}^{up}(\mathbf{h}_{\backslash 123})$, $g_{124}^{up}(\mathbf{h}_{\backslash 124})$, $g_{12}^{low}(\mathbf{h}_{\backslash 12})$, $g_{13}^{low}(\mathbf{h}_{\backslash 13})$, $g_{14}^{low}(\mathbf{h}_{\backslash 14})$, $g_{23}^{low}(\mathbf{h}_{\backslash 23})$, and $g_{24}^{low}(\mathbf{h}_{\backslash 24})$ we know P_4^{34} and hence $\overline{\Gamma}_4^*$.

In this section, we investigate the properties of these functions, including the properties of a single nonlinear function, as well as the relationship between different nonlinear functions. The first result is the convexity of $-g_{\mathcal{A}}^{up}(\mathbf{h}_{\backslash\mathcal{A}})$ and $g_{\mathcal{A}}^{low}(\mathbf{h}_{\backslash\mathcal{A}})$.

Lemma 2: The following functions corresponding to P_4^{34} are convex:

$$-g_1^{up}(\mathbf{h}_{\backslash 1}),\ -g_2^{up}(\mathbf{h}_{\backslash 2}),\ -g_{34}^{up}(\mathbf{h}_{\backslash 34}), -g_{123}^{up}(\mathbf{h}_{\backslash 123}),\ -g_{124}^{up}(\mathbf{h}_{\backslash 124})$$
$$g_{12}^{low}(\mathbf{h}_{\backslash 12}),\ g_{13}^{low}(\mathbf{h}_{\backslash 13}),\ g_{14}^{low}(\mathbf{h}_{\backslash 14}),\ g_{23}^{low}(\mathbf{h}_{\backslash 23}),\ g_{24}^{low}(\mathbf{h}_{\backslash 24})$$

Proof: Without loss of generality, we investigate the convexity of $g_1^{up}(\mathbf{h}_{\backslash 1})$. Let $\boldsymbol{h}^a = [h_1^a \quad \boldsymbol{h}_{\backslash 1}^a]^T$, $\boldsymbol{h}^b = [h_1^b \quad \boldsymbol{h}_{\backslash 1}^b]^T$ be any two entropic vectors in the pyramid P_4^{34}. Since $\overline{\Gamma}_4^*$ is a convex set, P_4^{34} is also convex. Thus for $\forall\ 0 \leqslant \lambda \leqslant 1$, we have $\lambda\boldsymbol{h}^a + (1-\lambda)\boldsymbol{h}^b \in P_4^{34}$. According to Theorem 3, we have

$$\lambda h_1^a + (1-\lambda)h_1^b \leqslant g_1^{up}(\lambda h_{\backslash 1}^a + (1-\lambda)h_{\backslash 1}^b) \tag{3.5}$$

Furthermore, for some $\boldsymbol{h}^a$ and $\boldsymbol{h}^b$ to make g_1^{up} tight, besides (3.5), the following two conditions also hold:

$$h_1^a \leqslant \boldsymbol{h}_1^a = g_1^{up}(\boldsymbol{h}_{\backslash 1}^a) \quad h_1^b \leqslant \boldsymbol{h}_1^b = g_1^{up}(\boldsymbol{h}_{\backslash 1}^b) \tag{3.6}$$

Combining (3.5) and (3.6), we get

$$\lambda h_1^a + (1-\lambda)h_1^b \leqslant \lambda \boldsymbol{h}_1^a + (1-\lambda)\boldsymbol{h}_1^b =$$
$$\lambda g_1^{up}(\boldsymbol{h}_{\backslash 1}^a) + (1-\lambda) g_1^{up}(\boldsymbol{h}_{\backslash 1}^b) \leqslant g_1^{up}(\lambda \boldsymbol{h}_{\backslash 1}^a + (1-\lambda)\boldsymbol{h}_{\backslash 1}^b)$$

Thus $g_1^{up}(h_{\backslash 1})$ is a concave function. Similarly we can prove the convexity of other functions listed above. ■

Next we would like to study the symmetry properties of $g_1^{up}(h_{\backslash 1})$. From the form of Ingleton inequality (2.11), we notice that for a given distribution, if we swap the position of X_i and X_j, the value calculated from $Ingleton_{ij}$ remain unchanged, same properties hold if we exchange X_k and X_l. However, for a given distribution which has its entropic vector $\boldsymbol{h}^a$ tight on g_1^{up} (thus $h_1^a = g_1^{up}(\boldsymbol{h}_{\backslash 1}^a)$), due to symmetry, the entropic vector $\boldsymbol{h}^b$ corresponding to the distribution swapping X_i and X_j(and/or swap X_k and X_l) will still be on the boundary and satisfy $h_1^b = g_1^{up}(\boldsymbol{h}_{\backslash 1}^b)$. Based on this fact, and that $g_1^{up}(\mathbf{h}_{\backslash 1})$ is a concave function, we get the following theorem.

Theorem 4: Suppose we have a distribution p_X with corresponding entropic vector $\boldsymbol{h}^a$ tight on g_1^{up}, and denote $\boldsymbol{h}^b$ the entropic vector from swapping X_3 and X_4 in $p_{\boldsymbol{X}}$, then $g_1^{up}(\boldsymbol{h}_{\backslash 1}^a) = g_1^{up}(\boldsymbol{h}_{\backslash 1}^b)$ and

$$\max_{\lambda \in [0,\ 1]} g_1^{up}(\lambda \boldsymbol{h}_{\backslash 1}^a + (1-\lambda)\boldsymbol{h}_{\backslash 1}^b) = g_1^{up}(\frac{1}{2}\boldsymbol{h}_{\backslash 1}^a + \frac{1}{2}\boldsymbol{h}_{\backslash 1}^b) \tag{3.7}$$

thus the maximum of g_1^{up} along $\lambda \boldsymbol{h}^a_{\backslash 1} + (1-\lambda)\boldsymbol{h}^b_{\backslash 1}$ must be obtained at entropic vectors satisfying $h_3 = h_4$, $h_{13} = h_{14}$, $h_{23} = h_{24}$ and $h_{123} = h_{124}$.

Proof: First we need to point out the symmetry between $\boldsymbol{h}^a$ and $\boldsymbol{h}^b$ caused by the exchange of X_3 and X_4. For

$$\boldsymbol{h}^a = [h^a_1\ h^a_2\ h^a_{12}\ h^a_3\ h^a_{13}\ h^a_{23}\ h^a_{123} h^a_4 h^a_{14}\ h^a_{24}\ h^a_{124}\ h^a_{34}\ h^a_{134}\ h^a_{234}\ h^a_{1234}] \tag{3.8}$$

it can be easily verified that

$$\boldsymbol{h}^b = [h^a_1\ h^a_2\ h^a_{12}\ h^a_4\ h^a_{14}\ h^a_{24}\ h^a_{124} h^a_3 h^a_{13}\ h^a_{23}\ h^a_{123}\ h^a_{34}\ h^a_{134}\ h^a_{234}\ h^a_{1234}] \tag{3.9}$$

Since both $\boldsymbol{h}^a$ and $\boldsymbol{h}^b$ are tight on g_1^{up},

$$h^a_1 = g_1^{up}(\boldsymbol{h}^a_{\backslash 1}) \quad h^b_1 = g_1^{up}(\boldsymbol{h}^b_{\backslash 1})$$

Thus $\boldsymbol{h}^a_1 = \boldsymbol{h}^b_1$ implies $g_1^{up}(\boldsymbol{h}^a_{\backslash 1}) = g_1^{up}(\boldsymbol{h}^b_{\backslash 1})$, which also guarantee $g_1^{up}(\lambda \boldsymbol{h}^a_{\backslash 1} + (1-\lambda)\boldsymbol{h}^b_{\backslash 1}) = g_1^{up}((1-\lambda)\boldsymbol{h}^a_{\backslash 1} + \lambda \boldsymbol{h}^b_{\backslash 1})$.

Now we proof (3.7) by contradiction, suppose $\exists \lambda' \in [0,\ 1], \lambda' \neq \frac{1}{2}$ such that

$$g_1^{up}(\lambda' \boldsymbol{h}^a_{\backslash 1} + (1-\lambda')\boldsymbol{h}^b_{\backslash 1}) > g_1^{up}(\frac{1}{2}\boldsymbol{h}^a_{\backslash 1} + \frac{1}{2}\boldsymbol{h}^b_{\backslash 1}) \tag{3.10}$$

Since $g_1^{up}(h_{\backslash 1})$ is a concave function,

$$\begin{aligned}
&g_1^{up}(\lambda' \boldsymbol{h}^a_{\backslash 1} + (1-\lambda')\boldsymbol{h}^b_{\backslash 1}) \\
&= g_1^{up}((1-\lambda')\boldsymbol{h}^a_{\backslash 1} + \lambda' \boldsymbol{h}^b_{\backslash 1}) \\
&= \frac{1}{2} g_1^{up}(\lambda' \boldsymbol{h}^a_{\backslash 1} + (1-\lambda')\boldsymbol{h}^b_{\backslash 1}) + \frac{1}{2} g_1^{up}((1-\lambda')\boldsymbol{h}^a_{\backslash 1} + \lambda' \boldsymbol{h}^b_{\backslash 1}) \\
&\leqslant g_1^{up}(\frac{1}{2}[\lambda' \boldsymbol{h}^a_{\backslash 1} + (1-\lambda')\boldsymbol{h}^b_{\backslash 1}] + \frac{1}{2}[(1-\lambda')\boldsymbol{h}^a_{\backslash 1} + \lambda' \boldsymbol{h}^b_{\backslash 1}]) \\
&= g_1^{up}(\frac{1}{2}\boldsymbol{h}^a_{\backslash 1} + \frac{1}{2}\boldsymbol{h}^b_{\backslash 1})
\end{aligned}$$

which contradicts the assumption, and proves (3.7). Because of the symmetry between $\boldsymbol{h}^a$ in (3.8) and $\boldsymbol{h}^b$ in (3.9), entropic vector $\frac{1}{2}\boldsymbol{h}^a_{\backslash 1} + \frac{1}{2}\boldsymbol{h}^b_{\backslash 1}$ will have the properties that $h_3 = h_4$, $h_{13} = h_{14}$, $h_{23} = h_{24}$ and $h_{123} = h_{124}$. ■

Next we are going to investigate the relationship between g_1^{up} and g_2^{up} by swapping X_1 and X_2 OR swapping both X_1, X_2 and X_3, X_4. For a distribution p_X with corresponding entropic vector $\boldsymbol{h}^a$ tight on g_1^{up}, we denote $\boldsymbol{h}^c$ the entropic vector from swapping X_1 and X_2 in p_X, $\boldsymbol{h}^d$ be entropic vector from swapping both X_1, X_2 and X_3, X_4. For

$$\boldsymbol{h}^a = [h^a_1\ h^a_2\ h^a_{12}\ h^a_3\ h^a_{13}\ h^a_{23}\ h^a_{123} h^a_4 h^a_{14}\ h^a_{24}\ h^a_{124}\ h^a_{34}\ h^a_{134}\ h^a_{234}\ h^a_{1234}]$$

it can be easily verified that

$$\boldsymbol{h}^c = [h^a_2\ h^a_1\ h^a_{12}\ h^a_3\ h^a_{23}\ h^a_{13}\ h^a_{123} h^a_4 h^a_{24}\ h^a_{14}\ h^a_{124}\ h^a_{34}\ h^a_{234}\ h^a_{134}\ h^a_{1234}] \tag{3.11}$$

$$\boldsymbol{h}^d = [h^a_2\ h^a_1\ h^a_{12}\ h^a_4\ h^a_{24}\ h^a_{14}\ h^a_{124} h^a_3 h^a_{23}\ h^a_{13}\ h^a_{123}\ h^a_{34}\ h^a_{234}\ h^a_{134}\ h^a_{1234}] \tag{3.12}$$

Thus from $h_1^a = h_2^c = h_2^d$ we get

$$g_1^{up}(\boldsymbol{h}_{\setminus 1}^a) = g_2^{up}(\boldsymbol{h}_{\setminus 2}^c) = g_2^{up}(\boldsymbol{h}_{\setminus 2}^d) \tag{3.13}$$

which leads to the following theorem:

Theorem 5: Suppose we have a distribution p_X with corresponding entropic vector $\boldsymbol{h}^a$ tight on g_1^{up}, we denote by $\boldsymbol{h}^c$ the entropic vector from swapping X_1 and X_2 in p_X, and $\boldsymbol{h}^d$ the entropic vector from permuting both X_1, X_2 and X_3,X_4. Then

$$g_1^{up}(\boldsymbol{h}_{\setminus 1}^a) = g_2^{up}(\boldsymbol{h}_{\setminus 2}^c) = g_2^{up}(\boldsymbol{h}_{\setminus 2}^d) \tag{3.14}$$

Furthermore, if the entropic vector $\boldsymbol{h}^e$ associated with some distribution p_X satisfies $h_{13} = h_{23}$, $h_{14} = h_{24}$ and $h_{134} = h_{234}$, then $g_1^{up}(\boldsymbol{h}_{\setminus 1}^e) = g_2^{up}(\boldsymbol{h}_{\setminus 1}^e)$; if the entropic vector h^{fE} associated with some distribution p_X satisfies $h_3 = h_4$, $h_{13} = h_{24}$, $h_{14} = h_{23}$, $h_{123} = h_{124}$ and $h_{134} = h_{234}$, then $g_1^{up}(h_{\setminus 1}^{fE}) = g_2^{up}(h_{\setminus 1}^{fE})$.

Example 1: In order to explain Theorem 5, we consider the example such that we fix the last 13 dimension of entropic vector to $V = [3\ 2\ 3\ 3\ 4\ 2\ 3\ 3\ 4\ 4\ 4\ 4\ 4]$ and only consider the first two dimensions h_1 and h_2, which is shown in Figure 3.3. Since Γ_4^* is a 15 dimensional convex cone, if we fixed 13 dimensional to V, only h_1 and h_2 should be considered, thus we can easily plot the constrained region for visualization.

In Figure 3.3, f is the one of the 6 bad extreme rays(extreme rays of Γ_4 that are not entropic). The rectangle formed by connecting (0,0), (2,0), (0,2) and f is the mapping of Shannon outer bound Γ_4 onto this plane. The green line connecting a and e is the projection of $Ingleton_{34}$ onto the plane. Notice we also plot inequality (2.8) and (2.9) for some values of s in the figure for the comparison between Ingleton inner bound, Shannon outer bound and non-Shannon outer bound. The red dot

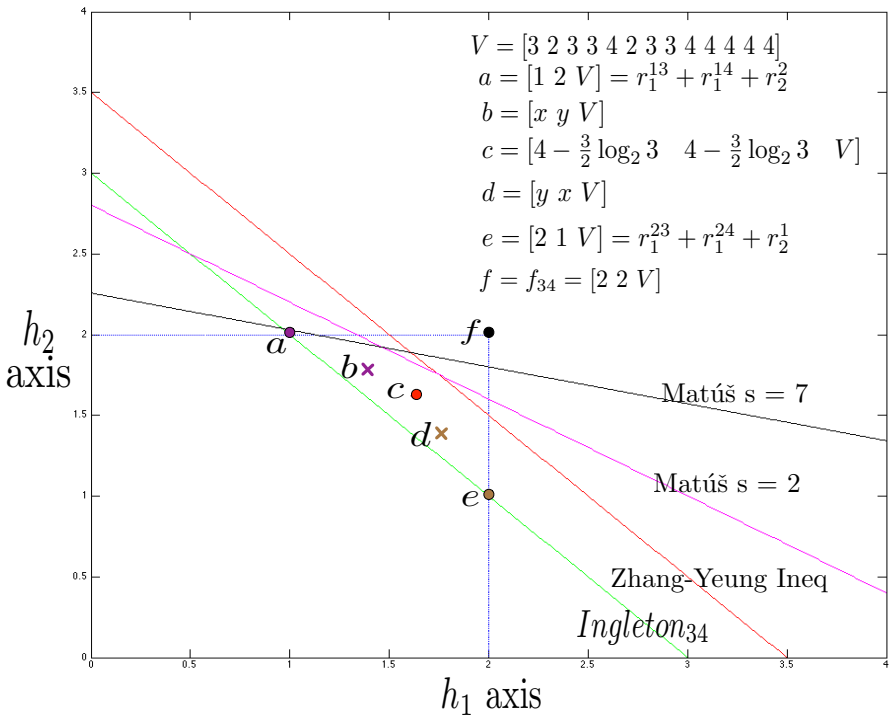

Figure 3.3: Entropic vector hyperplane with only h_1 and h_2 coordinate not fixed

point c is the entropic vector of the binary distribution with only four outcomes: (0000)(0110)(1010)(1111), each of the outcomes occur with probability $\frac{1}{4}$, and following from the convention of [32], we call it the 4 atom uniform point.

Since we already know $a = [1\ 2\ V]$ and $e = [2\ 1\ V]$ must lie on the boundary of P_4^{34}, thus $g_1^{up}([2\ V]) = g_2^{up}([2\ V])$ and $g_1^{up}([1\ V]) = g_2^{up}([1\ V])$. More generally, for any entropic vector $b = [x\ y\ V]$ on the boundary, we have $g_1^{up}([x\ V]) = g_2^{up}([x\ V])$ and $g_1^{up}([y\ V]) = g_2^{up}([y\ V])$. Thus we can say that when we constrain the last 13 dimension of entropic vector to $V = [3\ 2\ 3\ 3\ 4\ 2\ 3\ 3\ 4\ 4\ 4\ 4\ 4]$, the two function g_1^{up} and g_2^{up} always give us the same value, that is to say they are identical when fixed in this hyperplane.

4. Listing Canonical k-atom Supports and map the Entropy Region with them

A first key question when studying the part of the entropy region associated with the gap between its best known inner and outer bounds described in the prior chapter is which supports for joint probability mass functions for the random variables can yield entropic vectors in this region. In fact, results from Chan[31] have shown that, with infinite computational power, to determine the whole entropy region, it would suffice to consider the conic hull of entropic vectors associated with only the *quasi-uniform* probability distributions, which are completely specified via their support. This question is further motivated by the observation that previous campaigns that have attempted to numerically map unknown parts of this region have empirically observed that the joint probability mass functions associated with the extremal entropic vectors produced, while not quasi-uniform, do have many of their probabilities zero [1, 32].

To begin our study in this arena, we must formally introduce the concept of a k-atom support and define the equivalence of two k-atom supports. Consider the probability distributions for a random vector $\boldsymbol{X} = (X_1, \ldots, X_N)$ taking values on the Cartesian product $\boldsymbol{\mathcal{X}}^{\times} = \mathcal{X}_1 \times \mathcal{X}_2 \times \cdots \times \mathcal{X}_N$, where $\mathcal{X}_n$ is a finite set with values $i \in \{1, \ldots, |\mathcal{X}_n|\}$. To a particular probability mass function $p_{\boldsymbol{X}}$ we can associate a length $\prod_{n=1}^{N} |\mathcal{X}_n| - 1$ vector by listing the probabilities of all but one of the outcomes in $\boldsymbol{\mathcal{X}}^{\times}$ into a vector

$$\boldsymbol{\eta} = \left[p_{\mathbf{X}}(i_1, \ldots, i_N) \;\middle|\; \begin{array}{c} i_k \in \{1, 2, \ldots, |\mathcal{X}_k|\}, \\ k \in \{1, \ldots, N\}, \\ \sum_{k=1}^{N} i_k \neq N. \end{array} \right]. \tag{4.1}$$

Notice we are only listing outcomes such that $\sum_k i_k \neq N$, $p_{\mathbf{X}}(\mathbf{1}) = p_{\mathbf{X}}(i_1 = 1, \ldots, i_N = 1)$ will be the only outcome that is left out. $\boldsymbol{\eta}$ in (4.1) can be determined uniquely from the probability mass function $p_{\boldsymbol{X}}$, and owing to the fact that the probability mass function must sum to one, the omitted probability $p_{\mathbf{X}}(\mathbf{1})$ can be calculated, and hence the probability mass function can be determined from $\boldsymbol{\eta}$.

Specifying the joint probability distribution via the vector $\boldsymbol{\eta}$ enables all outcomes to have nonzero probabilities, however, entropic vectors are often extremized by selecting some of the elements of $\boldsymbol{\mathcal{X}}^{\times}$ to have zero probability. For this reason, rather than specifying some probabilities in their cartesian product to be zero, it is of interest to instead specify a support $\boldsymbol{\mathcal{X}} \subset \boldsymbol{\mathcal{X}}^{\times}$, no longer a cartesian product, on which the probabilities will be non-zero. Equivalently, if we take $|\boldsymbol{\mathcal{X}}| = k$, we are considering only those probability spaces $(\Omega, \mathcal{F}, \mathbb{P})$ with $|\Omega| = k$ to define the random variables $\boldsymbol{X} : \Omega \to \boldsymbol{\mathcal{X}}^{\times}$ on. A probability support $\boldsymbol{\mathcal{X}}$ satisfying $|\boldsymbol{\mathcal{X}}| = k$ is called a k-atom support, and a joint distribution created this way will be called a k-atom distribution.

Two k-atom supports $\boldsymbol{\mathcal{X}}, \boldsymbol{\mathcal{X}}'$, $|\boldsymbol{\mathcal{X}}| = |\boldsymbol{\mathcal{X}}'| = k$, will be said to be *equivalent*, for the purposes of tracing out the entropy region, if they yield the same set of entropic vectors, up to a permutation of the random variables. In other words, $\boldsymbol{\mathcal{X}}$ and $\boldsymbol{\mathcal{X}}'$ are equivalent, if, for every probability mass function $p_{\mathbf{X}} : \boldsymbol{\mathcal{X}} \to [0, 1]$, there is another probability mass function $p_{\mathbf{X}'} : \boldsymbol{\mathcal{X}}' \to [0, 1]$ and a bijection $\pi : \mathcal{N} \to \mathcal{N}$ such that

$$h_{\mathcal{A}}(p_{\mathbf{X}}) = h_{\pi(\mathcal{A})}(p_{\mathbf{X}'}) \quad \forall \mathcal{A} \subseteq \mathcal{N}.$$

Take $N = 4$ and $|\boldsymbol{\mathcal{X}}| = 1$ as a trivial example, since we only have one outcome/atom, it will have the probability of 1. In this way, different 1-atom supports like $[(0, 0, 0, 0)]$, $[(0, 0, 1, 2)]$, $[(0, 1, 2, 3)]$ and $[(2, 5, 7, 9)]$ are equivalent because they all map to the same 15 dimensional entropic vector with all zero elements.

The goal of §4.1 is to formalize this notion of equivalent supports with the use of

tools from abstract algebra, then describe some methods for enumerating and listing one representative from each equivalence class of supports. With the list of canonical k supports obtained through the method described in the §4.1 in hand, the matter turns to how to exploit them to better numerically map out the unknown parts of the entropy region. We study this problem from two angles, the first, in §4.2, aims to solely focus on optimizing the Ingleton score, while the second, in §4.3 describes a process for obtaining numerically optimized inner bounds to the entropy region.

In this regard, the following definitions from the theory of group actions will be helpful.

Definition 2: Let Z be a finite group acting on a finite set $\mathcal{V}$, a *group action* is a mapping

$$Z \times \mathcal{V} \to \mathcal{V} : (z, v) \mapsto zv$$

such that if e is the identity in Z, $ev = v \ \forall v \in \mathcal{V}$ and for any $z_1, z_2 \in Z$, $z_2 z_1 v = (z_2 z_1)v$ for all $v \in V$. For $v \in \mathcal{V}$, the *orbit* of v under Z is defined as

$$Z(v) = \{zv \mid z \in Z\}$$

the *stabilizer* subgroup of v in $\mathcal{V}$ is defined as

$$Z_v = \{z \in Z \mid zv = v\}$$

Suppose there is some ordering of $\mathcal{V}$, and let v be the element of $Z(v)$ that is least under this ordering, i.e. the *canonical representative* of the orbit $Z(v)$. For another $v' \in Z(v)$, an element $z \in Z$ is called a *transporter element* for v' if $zv' = v$.

Definition 3: (orbit data structure [33]) Let Z be a group which acts on the finite set

$\mathcal{V}$. The triple

$$orbit(Z, \mathcal{V}) = (\mathcal{T}, \sigma, \varphi)$$

is the *orbit data structure* for Z acting on $\mathcal{V}$ provided that

$$1.\mathcal{T} \text{ is a transversal of the } Z\text{-orbits on } \mathcal{V}$$

$$2.\sigma : \mathcal{V} \to L(Z) : v \mapsto Z_v$$

$$3.\varphi : \mathcal{V} \to Z : v \mapsto z \text{ with } zv \in \mathcal{T}$$

Here, $L(Z)$ denotes the lattice of subgroups of Z, we call σ the *stabilizer map* and φ the *transporter map*.

In next section, we will show that listing non-isomorphic distribution supports is equivalent to calculating the orbit data structure associated with the symmetric group acting a particular finite set.

4.1 Non-isomorphic k-atom supports via Snakes and Ladders

The key to list non-isomorphic distribution supports is to realize that a random variable on a probability space with $|\Omega| = k$ can be viewed as a *set partition* [34] of $\mathbb{N}_1^k = \{1, \ldots, k\}$ for the purpose of calculating entropy. A *set partition* of $\mathbb{N}_1^k$ is a set $\mathcal{B} = \{B_1, \ldots, B_t\}$ consisting of t subsets $B_1, \ldots, B_t$ of $\mathbb{N}_1^k$, called the *blocks* of the partition, that are pairwise disjoint $B_i \cap B_j = \emptyset, \forall i \neq j$, and whose union is $\mathbb{N}_1^k$, so that $\mathbb{N}_1^k = \bigcup_{i=1}^{t} B_i$. Let $\Pi(\mathbb{N}_1^k)$ denote the set of all set partitions of $\mathbb{N}_1^k$. The cardinalities of $\Pi(\mathbb{N}_1^k)$ for different k are commonly known as *Bell numbers*. For instance, there are 5 different set partitions for $k = 3$, that is $|\Pi(\mathbb{N}_1^3)| = 5$ and

$$\Pi(\mathbb{N}_1^3) = \{\{\{1,2,3\}\}, \{\{1,2\},\{3\}\}, \{\{1,3\},\{2\}\},$$
$$\{\{2,3\},\{1\}\}, \{\{1\},\{2\},\{3\}\}\},$$

while for $k = 4$, $|\Pi(\mathbb{N}_1^4)| = 15$ and $\Pi(\mathbb{N}_1^4)$ is the set

$$\{\{\{1,2,3,4\}\},\{\{1,2,3\},\{4\}\},\{\{1,2,4\},\{3\}\},$$
$$\{\{1,3,4\},\{2\}\},\{\{2,3,4\},\{1\}\},\{\{1,2\},\{3,4\}\},$$
$$\{\{1,3\},\{2,4\}\},\{\{1,4\},\{2,3\}\},\{\{1,2\},\{3\},\{4\}\},$$
$$\{\{1,3\},\{2\},\{4\}\},\{\{1,4\},\{2\},\{3\}\},\{\{2,3\},\{1\},\{4\}\},$$
$$\{\{2,4\},\{1\},\{3\}\},\{\{3,4\},\{1\},\{2\}\},\{\{1\},\{2\},\{3\},\{4\}\}\}.$$

A set partition $\mathcal{B} \in \Pi(\mathbb{N}_1^k)$ is said to *refine* a set partition $\mathcal{B}' \in \Pi(\mathbb{N}_1^k)$ if all of the blocks in $\mathcal{B}'$ can be written as the union of some blocks in $\mathcal{B}$. The *meet* of two partitions $\mathcal{B}, \mathcal{B}' \in \Pi(\mathbb{N}_1^k)$, denoted by $\mathcal{B} \wedge \mathcal{B}'$ is the partition of $\mathbb{N}_1^k$ formed by all of the non-empty intersections of a block from $\mathcal{B}$ and a block from $\mathcal{B}'$:

$$\mathcal{B} \wedge \mathcal{B}' = \left\{ B_i \cap B'_j \,\middle|\, B_i \in \mathcal{B}, B'_j \in \mathcal{B}', B_i \cap B'_j \neq \emptyset \right\}$$

Refinement and meet set up a partial order on $\Pi(\mathbb{N}_1^k)$ which enable it to be identified as a *lattice*, the lattice of $\Pi(\mathbb{N}_1^4)$ is shown in Figure 4.1

Let $\mathbf{\Xi}_N$ be the collection of all sets of N set partitions of $\mathbb{N}_1^k$ whose meet is the finest partition (the set of singletons),

$$\mathbf{\Xi}_N := \left\{ \xi \,\middle|\, \xi \subseteq \Pi(\mathbb{N}_1^k),\ |\xi| = N,\ \bigwedge_{\mathcal{B} \in \xi} \mathcal{B} = \bigcup_{i=1}^{N} \{\{i\}\} \right\}. \tag{4.2}$$

The symmetric group $\mathbb{S}_k$ induces a natural group action on a set partition $\mathcal{B} \in \Pi(\mathbb{N}_1^k)$, $\mathcal{B} = \{B_1, \ldots, B_t\}$: representing an element $\pi \in \mathbb{S}_k$ as a permutation $\pi : \mathbb{N}_1^k \to \mathbb{N}_1^k$, we have

$$\pi(\mathcal{B}) := \{\pi(B_1), \ldots, \pi(B_t)\}. \tag{4.3}$$

This action on set partitions induces, again in a natural manner, a group action of

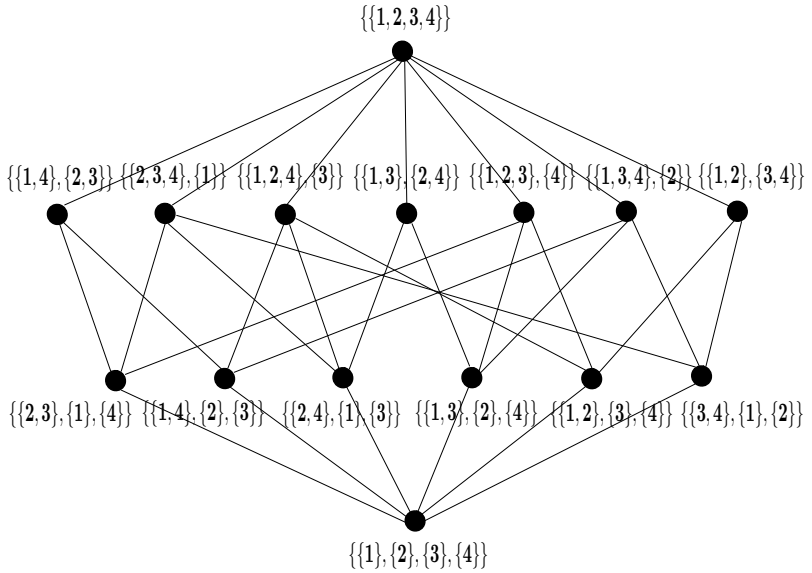

Figure 4.1: Lattice of $\Pi(\mathbb{N}_1^4)$: the set of all set partitions for $k = 4$

$\mathbb{S}_k$ on the set $\mathbf{\Xi}_N$ of subsets of N partitions from $\Pi(\mathbb{N}_1^k)$ whose meet is the singletons: $\pi \in \mathbb{S}_k$ acts on a set of partitions $\xi \in \mathbf{\Xi}_N$, $\xi = \{\mathcal{B}_1, \ldots, \mathcal{B}_N\}$ via

$$\pi(\xi) := \{\pi(\mathcal{B}_1), \ldots, \pi(\mathcal{B}_N)\}. \tag{4.4}$$

The group action (4.3) on set partition and group action (4.4) on sets of set partitions enable us to enumerate the non-isomorphic k-atom supports by calculating the orbit data structure of symmetry group acting on a well defined set, the result is summarized in Theorem 6.

Theorem 6: The problem of generating the list of all non-isomorphic k-atom, N-variable supports, that is, selecting one representative from each equivalence class of isomorphic supports, is equivalent to obtaining a transversal of the orbits $\mathbf{\Xi}_N//\mathbb{S}_k$ of $\mathbb{S}_k$ acting on $\mathbf{\Xi}_N$, the set of all subsets of N set partitions of the set $\mathbb{N}_1^k$ whose meets are the set of singletons $\{\{1\}, \{2\}, \ldots, \{N\}\}$.

Proof: A random variable introduces a partition on the sample space based on its inverse image. The joint distributions of several random variables is created from the meet of these partitions. The joint entropy is insensitive to the labeling of elements in the sample space, as well as the labeling of the outcomes of the random variable, hence it is only the partitions and their meets that matter in determining the joint entropies. Since an appropriate notion of isomorphism between supports also recognizes that it does not matter which random variable is labeled as the first random variable and so on, and there is no need to duplicate random variables when enumerating supports, rather than a N tuple of partitions, the support of a collection of random variables is best thought of, then, as a set of such set-partitions. The requirement that the meet is the singletons follows from the fact that if it is not, there is a k' atom distribution with $k' < k$ whose atoms are the sets in the meet partition, which gives equivalent entropies, and hence such a situation is better labelled as a k'-atom distribution. ■

Theorem 6 sets up the theoretical framework for obtaining the list of non-isomorphic k-atom, N-variable supports for the purpose of calculating entropic vectors: one must calculate the orbits data structure of the symmetric group $\mathbb{S}_k$ acting on $\mathbf{\Xi}_N$. One way to carry out this computation is to directly calculate the orbit data structure on $\mathbf{\Xi}_N$ using the default subroutine in GAP. However, this approach quickly becomes intractable when k and N are larger than four, as both CPU time and memory usage go beyond the resonable capacity of a single computer. Alternatively, one can use a recursive breadth-first search style algorithm named Leiterspiel or "Snakes and Ladders"[33, 35] to efficiently calculate the orbit data structure.

Suppose we have a group Z acting on a set $\mathcal{V}$, the algorithm Snakes and Ladders, see e.g. [33] pp. 709–710, is an algorithm which enables one to compute orbit data

$N\backslash k$	3	4	5	6	7
2	2	8	18	48	112
3	2	31	256	2437	25148
4	1	75	2665	105726	5107735
5	0	132	22422	3903832	
6	0	187	161118		

Table 4.1: # of non-isomorphic k-atom, N-variable supports.

structure of group Z on the set $\mathcal{P}_i(\mathcal{V})$ of all subsets of the set $\mathcal{V}$ of cardinality i. For a given set $\mathcal{V}$, the algorithm first computes the orbit data structure on the set of subsets of size $i = 1$, then it recursively increase the subsets size i, where the computation to determine the orbit data structure for subsets of size i is created from manipulations with the orbit data structure on subsets of size $i - 1$.

To apply this problem to the non-isomorphic support enumeration problem, one selects the set $\mathcal{V}$ to be the set of all set partitions of the set $\mathbb{N}_1^k$, ordered lexicographically, and the group Z to be the symmetric group $\mathbb{S}_k$. Once the step in Snakes and Ladders associated with subsets (of set partitions) of size N is reached, each element of the transversal is checked to determine if the meet of its partitions is the set of singletons, and the resulting canonical sets of N set partitions yield the non-isomorphic supports.

The snakes and ladders algorithm was applied to enumerate non-isomorphic supports in this manner, and the resulting numbers of non-isomorphic supports obtained are displayed in Table 4.1. As shown, for $N = 4$ variables, only one support is needed for calculating entropic vectors of 3-atom distribution, however, there are 75 non-isomorphic supports for $k = 4$, and the number of non-isomorphic supports grow rapidly in the number of atoms k.

In the next section, we will utilize these non-isomorphic k-atom supports together with numerical optimization to obtain inner bounds for entropy.

number of atoms k	3	4	5	6	7
all supports	1	75	2665	105726	5107735
Ingleton violating	0	1	29	1255	60996

Table 4.2: # of non-isomorphic k-atom, 4-variable supports that can violate the Ingleton inequality.

4.2 Maximal Ingleton Violation and the Four Atom Conjecture

Given that the unknown part of $\bar{\Gamma}_4^*$ is associated with violating the Ingleton inequality, substantial research effort has been exerted towards determining distributions on $N = 4$ random variables that violate the Ingleton inequality (2.11). Dougherty, Freiling, and Zeger[32] defined a normalized function called the *Ingleton score* to measure the degree of Ingleton violation for 4 random variables, and they also make the *Four-Atom Conjecture* which states that the Ingleton score of 4 random variables can not be lower than -0.08937. After the Four-Atom conjecture was proposed, Ingleton violation was studied extensively with finite groups[36, 37, 38], then in [1], the conjecture was refuted by transforming a distribution obtaining Ingleton score of -0.078277 through a operation which preserves the property of almost entropic to a vector with Ingleton score -0.09243. In this section, we study the number of k atom supports that can, for some probability distribution, violate Ingleton, as well as the Ingleton scores they can attain.

For a particular k atom support for N variables, the Ingleton score can be numerically optimized via fine grid search and numerical gradient optimization. Doing so for each four variable support with 7 or fewer atoms yielded the results in Table 4.2, which shows that only a small fraction of the canonical supports can violate Ingleton.

Among all the 75 non-isomorphic 4-atom distribution supports, only one can be

assigned a series of probabilities to violate Ingleton, that is the 4-atom support (4.5),

$$\begin{bmatrix} (0,0,0,0) \\ (0,1,1,0) \\ (1,0,1,0) \\ (1,1,1,1) \end{bmatrix}, \tag{4.5}$$

which is the support achieving Ingleton score $-.08937$ associated with the Four atom conjecture.

Each row in (4.5) is a vector in $\boldsymbol{\mathcal{X}}$ and corresponds to one outcome/atom for the collection of random variables, so the number of columns in (4.5) is the same as the number of random variables, in this case 4. Among the 29 5-atom supports that can violate Ingleton, 28 of them obtain the same minimal Ingleton score of -0.08937, with one atom's probability shrinking to zero. These 28 thus all shrink to the 4-atom support (4.5), achieving the same Ingleton score. The remaining one support,

$$\begin{bmatrix} (0,0,0,0) \\ (0,0,1,1) \\ (0,1,1,0) \\ (1,0,1,0) \\ (1,1,1,0) \end{bmatrix} \tag{4.6}$$

only achieves a minimal Ingleton score of -0.02423. For the 1255 6-atom supports, 58 of them get a minimal Ingleton score strictly less than -0.08937, while the remainder of the supports yield minimal scores of -0.08937. Experiments with the 7-atom supports have shown that, in keeping with the findings of the four atom conjecture and the attempts to refute it, no k-atom support with $k \leq 7$ is capable of directly beating the four atom distributions score. These exhaustive results substantiate findings

from other researchers that suggest that if it is indeed possible to give a probability distribution which directly (without any almost entropic preserving transformation on the entropic vectors as utilized in [1]) violates the four atom conjecture, at least a large support will be required.

4.3 Optimizing Inner Bounds to Entropy from k-Atom Distributions

For the purpose of generating better inner bounds for the region of entropic vectors, minimizing only the Ingleton score is far from enough, since it is only optimizing the distribution to a cost function of certain hyperplane defined by the Ingleton inequality. Bearing this in mind, one can define cost functions different from the Ingleton score, but still yielding optimized points that are in the unknown part of the entropy region associated with violating Ingleton. We will first describe a simple procedure to randomly generate such cost functions, and then their numerical optimization over each of the Ingleton violating 4, 5, and 6 atom supports. The resulting entropic vectors are then collected to generate inner bounds to $\bar{\Gamma}_4^*$ based on distributions with 4, 5, and 6 atom supports.

Lemma 1 defined the 15 extreme rays of the pyramid G_4^{ij}, and, without loss of generality, it suffices to consider G_4^{34}. Among these 15 rays, the 14 extreme rays that lie on the hyperplane of $Ingleton_{34} = 0$ are $\boldsymbol{r}_1^{134}$, $\boldsymbol{r}_1^{234}$, $\boldsymbol{r}_1^{123}$, $\boldsymbol{r}_1^{124}$, $\boldsymbol{r}_1^{\emptyset}$, $\boldsymbol{r}_3^{\emptyset}$, $\boldsymbol{r}_1^{3}$, $\boldsymbol{r}_1^{4}$, $\boldsymbol{r}_1^{13}$, $\boldsymbol{r}_1^{14}$, $\boldsymbol{r}_1^{23}$, $\boldsymbol{r}_1^{24}$, $\boldsymbol{r}_2^{1}$, $\boldsymbol{r}_2^{2}$, while the only extreme ray in G_4^{34} that is not entropic is $\boldsymbol{f}_{34}$. For generating cost functions, among several options, we found the following one gives us the best inner bound. First a random vector $\Lambda = \{\lambda_1, \lambda_2, \cdots \lambda_{14}\}$ of 14 dimension is generated, where λ_i takes value from 0 to a large positive value. Then for each of the 14 extreme rays on the hyperplane of $Ingleton_{34} = 0$, one new ray is generated

through the following equation

$$\boldsymbol{r}_i^{new} = \frac{\boldsymbol{r}_i^{base} + \lambda_i \boldsymbol{f}_{34}}{1 + \lambda_i} \tag{4.7}$$

where $\boldsymbol{r}_i^{base} \in \{\boldsymbol{r}_1^{134}, \boldsymbol{r}_1^{234}, \boldsymbol{r}_1^{123}, \boldsymbol{r}_1^{124}, \boldsymbol{r}_1^{\emptyset}, \boldsymbol{r}_3^{\emptyset}, \boldsymbol{r}_1^{3}, \boldsymbol{r}_1^{4}, \boldsymbol{r}_1^{13}, \boldsymbol{r}_1^{14}, \boldsymbol{r}_1^{23}, \boldsymbol{r}_1^{24}, \boldsymbol{r}_2^{1}, \boldsymbol{r}_2^{2}\}$. After obtaining the 14 new rays $\mathbf{r}^{\mathbf{new}} = \{\boldsymbol{r}_1^{new}, \cdots, \boldsymbol{r}_{14}^{new}\}$, the hyperplane defined by these new rays, which in turn defines a new cost function, can be easily calculated. Notice if we let $\lambda_i = 0$ for $i = 1, 2, \cdots 14$, we will get the hyperplane of $Ingleton_{34} = 0$.

Computer experiments were run to generate more than 1000 cost functions in this manner by random selection of the λ_is. For each of these cost functions, numerical optimization of the distribution for each Ingleton violating k-atom support was performed, and a k-atom inner bound was generated by taking the convex hull of the entropic vectors corresponding to the optimized distribution.

The progress of characterizing Γ_4^* while performing these experiments was in part estimated by the volume of the inner bound as compared to the total volume of the pyramid G_4^{34}, as summarized in Table 4.3. For the purpose of comparison, we also list there the calculated volume ratio of the best outer bound in [32]. Note the volume of the k-atom inner bound obtained through the process describe above is only a estimated value and a lower bound to the true volume fraction, because only a finite number of cost functions and a finite number of entropic vectors were generated through the random cost function generation process. In principle one can generate as many entropic vectors as one wants through this process by growing the number of random cost functions selected, however calculating volume for many extreme points in high dimension can become computationally intractable. A key observation from the process is that while growing the support helps, from a volume standpoint the improvement after four atoms is somewhat small.

However, volume is just one metric for an inner bound, which can also be hard

inner and outer bounds	percent of pyramid
Shannon	100
Outer bound from [27]	96.5
4,5,6 atoms inner bound	57.8
4,5 atoms inner bound	57.1
4 atoms inner bound	55.9
4 atom conjecture point only	43.5
3 atoms inner bound	0

Table 4.3: The volume increase within pyramid G_4^{34} as more atoms are included

to visualize in high dimensions. For this reason, we would like to visualize the k-atom inner bound in lower dimension. In this regard, a certain 3 dimensional subset of it selected in [1] will be utilized. In order to perform the transformation, each numerically obtained 15 dimensional vector $\boldsymbol{h} \in G_4^{34}$ is first transformed into its *tight* component by subtracting its *modular* component which is defined by

$$\boldsymbol{h}^m(\mathcal{W}) = \sum_{i \in I} [\boldsymbol{h}(N) - \boldsymbol{h}(N \backslash i)] \quad \mathcal{W} \subseteq \mathcal{N}$$

Next $\boldsymbol{h}^{ti}$ was pushed onto the hyperplane such that $I(X_3, X_4) = 0$ and $I(X_1; X_2 | X_3, X_4) = 0$ through the linear mapping

$$\begin{aligned} \boldsymbol{h}_{AB} \quad &= A_{34} B_{34,1} \boldsymbol{h}^{ti} \\ &= \boldsymbol{h}^{ti} + (h_3^{ti} + h_4^{ti} - h_{34}^{ti})(\boldsymbol{r}_1^3 - \boldsymbol{r}_1^{\emptyset}) \\ &+ (h_{123}^{ti} + h_{124}^{ti} - h_{34}^{ti} - h_{1234}^{ti})(\boldsymbol{r}_2^1 - \boldsymbol{r}_3^{\emptyset}) \end{aligned}$$

After that, another linear mapping C_{34} is used to further reduce the dimension of G_4^{34} to 4.

$$\boldsymbol{h}_C = C_{34}\boldsymbol{h}_{AB} = -Ingleton_{34}(\boldsymbol{h}^{ti})\boldsymbol{f}_{34}$$

$$+(h_3^{ti} + h_4^{ti} - h_{34}^{ti})\boldsymbol{r}_1^{\emptyset} + (h_{123}^{ti} + h_{124}^{ti} - h_{34}^{ti} - h_{1234}^{ti})\boldsymbol{r}_3^{\emptyset}$$

$$+\tfrac{1}{2}(h_{13}^{ti} + h_{23}^{ti} - h_3^{ti} - h_{123}^{ti} + h_{14}^{ti} + h_{24}^{ti} - h_4^{ti} - h_{124}^{ti})(\boldsymbol{r}_1^3 + \boldsymbol{r}_1^4)$$

$$+\tfrac{1}{2}(h_{13}^{ti} + h_{14}^{ti} - h_1^{ti} - h_{134}^{ti} + h_{23}^{ti} + h_{24}^{ti} - h_2^{ti} - h_{234}^{ti})(\boldsymbol{r}_2^1 + \boldsymbol{r}_2^2)$$

$$+\tfrac{1}{4}(h_{12}^{ti} + h_{14}^{ti} - h_1^{ti} - h_{124}^{ti} + h_{12}^{ti} + h_{13}^{ti} - h_1^{ti} - h_{123}^{ti} +$$

$$h_{12}^{ti} + h_{24}^{ti} - h_2^{ti} - h_{124}^{ti} + h_{12}^{ti} + h_{23}^{ti} - h_2^{ti} - h_{123}^{ti})(\boldsymbol{r}_1^{13} +$$

$$\boldsymbol{r}_1^{14} + \boldsymbol{r}_1^{23} + \boldsymbol{r}_1^{24})$$

If we further normalize the last dimension to equal to one, by dividing through by it, the resulting polytope associated with the convex hull in the remaining three dimensions is three dimensional. In order to coordinatize it, define $\boldsymbol{\alpha} = \frac{1}{4}\boldsymbol{f}_{34}$, $\boldsymbol{\beta} = \frac{1}{2}(\boldsymbol{r}_1^3 + \boldsymbol{r}_1^4)$, $\boldsymbol{\gamma} = \frac{1}{4}(\boldsymbol{r}_2^1 + \boldsymbol{r}_2^2)$ and $\boldsymbol{\delta} = \frac{1}{4}(\boldsymbol{r}_1^{13} + \boldsymbol{r}_1^{14} + \boldsymbol{r}_1^{23} + \boldsymbol{r}_1^{24})$, for any given transformed $\boldsymbol{g}$, we can write

$$\boldsymbol{g} = \bar{\alpha}_h\boldsymbol{\alpha} + \bar{\beta}_h\boldsymbol{\beta} + \bar{\gamma}_h\boldsymbol{\gamma} + \bar{\delta}_h\boldsymbol{\delta}$$

where $\bar{\alpha}_h + \bar{\beta}_h + \bar{\gamma}_h + \bar{\delta}_h = 1$. So in three dimensional space, we consider $\boldsymbol{\alpha}$ to be $(0, 0, 0)$, $\boldsymbol{\beta}$ to be $(\frac{1}{2}, \frac{\sqrt{3}}{2}, 0)$, $\boldsymbol{\gamma}$ to be $(1, 0, 0)$, $\boldsymbol{\delta}$ to be $(\frac{1}{2}, \frac{\sqrt{3}}{6}, \frac{\sqrt{6}}{3})$, so we can make the plot using $\bar{\beta}_h + \bar{\delta}_h$, $\bar{\gamma}_h + \bar{\delta}_h$ and $\bar{\alpha}_h$ as the three coordinate.

See Figure 4.2 for a surface plot of the inner bound generated by 4-atom supports, where we also plot the extremal points of 5-atom(red X) and 6-atom(black squares) inner bounds for comparison. Since we are transforming entropic vector from 15 dimension to 3 dimension, lots of extreme points of our 15 dimensional inner bound actually become redundant in this three dimensional space, so the number of points we can plot is significantly less than the number of extreme points we get from numerical

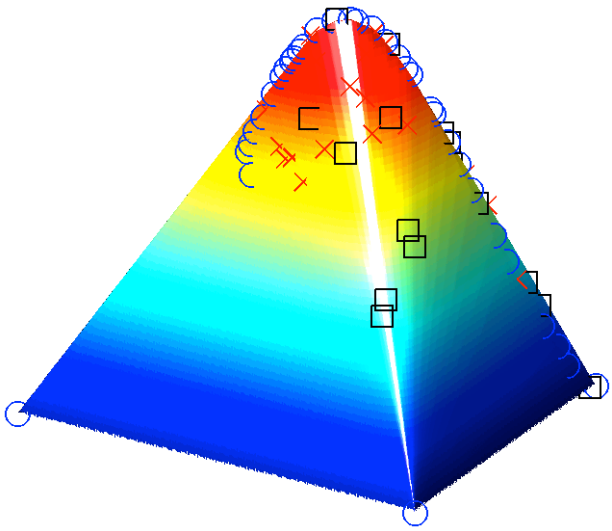

Figure 4.2: The surface is a projection of a 3D face of the inner bound to $\bar{\Gamma}_4^*$ created with the numerical process described in the text with 4-atom distributions. The blue circles are the from the entropic vectors from the extremal optimized four atom distributions, while the red Xs and the black squares are the additional extremal optimized k distributions for $k \in \{5, 6\}$, respectively.

optimization. As can be seen in Figure 4.2, the extreme points of 4-atom inner bound mostly lies in a curve, and there are some 5-atom extreme points away from the convex hull generated with this curve, and some of the 6-atom extreme points can get even further away.

In order to better visualize the difference between the various inner bounds, we also compared the contour of inner bound generated by $\leq k$ atom supports for $k \in \{4, 5, 6\}$, see Figure 4.3 for this comparison plot where *blue* line is $k = 4$, *red* line is

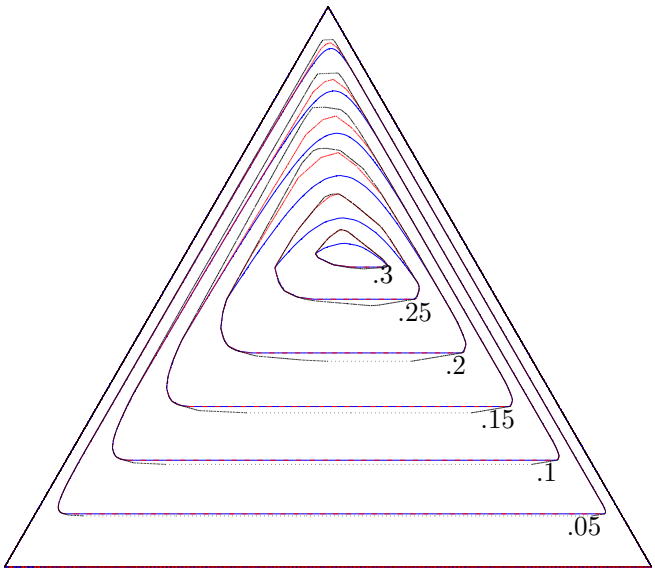

Figure 4.3: Comparison of contour plots of inner bound created from $\leq k$ atom distributions for $k \in \{4, 5, 6\}$. For each contour value, the inner most line is $k = 4$ while the outermost line is $k = 6$. The numerical inner bounds generated from only four atom distributions are quite good in this space.

$k = 5$ and *black* line is $k = 6$. As you can see, the contour is larger as more atoms are involved, meaning we can constantly get better inner bounds by increasing the number of atoms. This is because as the number of atoms are increased, a greater and greater variety of distribution supports are found that can violate the Ingleton inequality. The increase of the inner bound from five to six atom is smaller than the increase from four to five atoms, which is consistent with the full dimensional volume calculation from Table 4.3.

Finally, a comparison of the three dimensional bound in Fig. 4.2 with the three dimensional inner bound created in [1] is in order. In this respect, it is important to note that [1] numerically optimized within this lower dimensional space directly, and when performing the search, used distributions on very large supports and did not utilize a systematic search over probabilities to set to zero. This is to be contrasted

with our bounds which utilize supports of cardinality ≤ 6, and also are optimized with random cost functions in the higher 15-dimesional space then projected, rather than being directly optimized in the 3-dimensional constrained space being plotted. Via personal communication with the authors of [1], we obtained the points in the three dimensional space used to generate their version of Fig. 4.2 with these larger supports. Fig.4.4, which plots the same surface for different angles, superposes the extremal points obtained by our 4, 5, and 6 supports, which were optimized in 15-dimensions and then mapped into the 3-dimensional space via the process described above, over top of the surface [1] obtained with far larger supports and direct numerical optimization in these 3-dimensions. It shows that this systematic search of supports introduced in this section is able to find, and in some instances slightly improve upon, many of the extremal parts in this 3-dimensional region of [1], while other parts of this region are evidently unreachable with these smaller supports, or were not found our process of randomly sampling 15-dimensional cost functions. This argues strongly for pushing the systematic support enumeration process followed by numerical optimization employed in this section to higher support sizes and higher numbers of cost functions, rather than attempting to find effectively low cardinality supports through numerical optimization over higher supports.

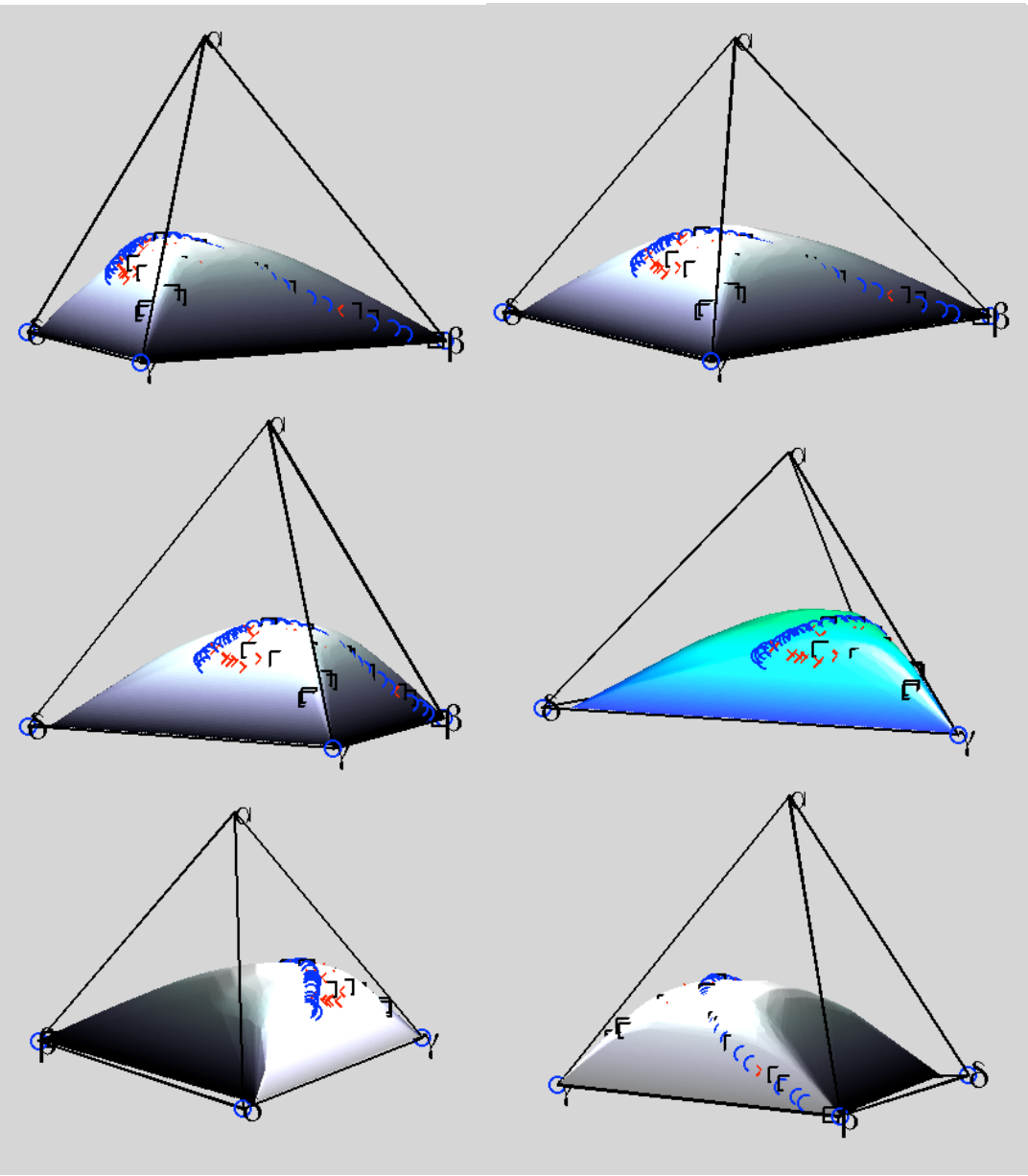

Figure 4.4: The surface is the bound created in [1] while the markers above it are those created from the generated non-isomorphic supports of size 4,5, and 6 in this section. The blue circles are from the entropic vectors from the extremal optimized four atom distributions, while the red Xs and the black squares are the additional extremal optimized k distributions for $k \in \{5, 6\}$, respectively. The low cardinality supports generate much of the surface [1], and even slightly improve it is some parts, despite being far lower cardinality and not even being optimized in this space.

5. Perspectives on How to Use Information Geometry to Understand Entropy Geometry

The previous chapter utilized support enumeration followed by numerical optimization to map the unknown parts of the entropy. While this did provide some numerical insight into which supports can have distributions that violate Ingleton and map to the unknown part of the entropy region, it did not give any analytical insight into how to parametrize probabilities so that the associated entropic vector would be in this unknown part, or be extremal. Bearing this in mind, in this chapter we aim to study properties of probability distributions associated with entropic vectors that are extremal, in the sense that they give entropic vectors lying in faces of the Shannon outer bound, and also that violate Ingleton. To do so, we will make use of information geometry, a discipline that in part studies ways to coordinatize probability distributions in a manner that relates properties such as independence to affine subspaces of the parameter space.

Information geometry is a discipline in statistics which endows the manifold of probability distribution with a special "dually flat" differential geometric structure created by selecting a Riemannian metric based on the Fisher information and a family of affine connections called the α-connections.

The natural divergence functions arising in information geometry, which include the Kullback Leibler divergence, can easily be related to the Shannon entropy by measuring the divergence with the uniform distribution. Additionally, one could think of entropic vectors as characterizing simultaneous properties (entropies) of marginal distributions on subsets of variables, and this simultaneous marginal distribution relationship has a well studied structure in information geometry.

Building on these observations, here we wish to sketch out a couple of ideas in

an attempt to show that characterizing the boundary of $\overline{\Gamma}_N^*$ could be thought of as a problem in information geometry, and that information geometry may in fact be a convenient framework to utilize to calculate these boundaries. The perspective we will provide is by no means unique, however, the aim here is to provide a preliminary link between these disciplines and enough evidence about their relationship in the hope that we or other researchers in these areas may one day solve the apparently very difficult, yet fundamental and very important, problem of explicitly determining $\overline{\Gamma}_N^*$ by exploiting information geometric tools.

Along these lines, we will first review in a somewhat rough manner in §5.1 some concepts from information geometry related to entropic vectors. Next, we will provide in §5.2 an information geometric characterization of those probability distributions associated with Shannon facets and some Shannon faces of the region of entropic vectors. We then present an information geometric characterization of submodularity as a corollary. Finally, we will close this chapter in §5.3 with characterization of the 4 atom distributions (4.5) which violate Ingleton via Information Geometry, and show a natural information geometric parameterization for them.

5.1 Review of Some Relevant Ideas from Information Geometry

Information geometry endows a manifold of probability distributions $p(x;\boldsymbol{\xi})$, parameterized by a vector of real numbers $\boldsymbol{\xi} = [\xi_i]$, with a Riemannian metric, or inner product between tangent vectors, given by the Fisher information:

$$g_{i,j}(\xi) \triangleq \mathbb{E}_\xi[\frac{\partial \log p(x;\boldsymbol{\xi})}{\partial \xi_i} \frac{\partial \log p(x;\boldsymbol{\xi})}{\partial \xi_j}] \tag{5.1}$$

This allows us to calculate an inner product between two tangent vectors $c = \sum_i c_i \partial_{\xi_i}$ and $d = \sum_i d_i \partial_{\xi_i}$ at a particular point ξ in the manifold of probability distributions,

as

$$< c, d >_{\xi} = \sum_{i,j} c_i d_j g_{i,j}(\xi) \tag{5.2}$$

Selecting this Riemannian metric (indeed, we could have selected others), and some additional structure, allows the differential geometric structure to be related to familiar properties of exponential families and mixture families of probability distributions. The additional structure, called an affine connection, is given to the manifold to allow us establish a correspondence, called parallel translation, between tangent vectors living the tangent space at two different points along a curve in the manifold by solving a system of ordinary differential equations involving given functions on the manifold called connection coefficients. Just as with a Riemannian metric, we have an infinite number possible of choices for the affine connection (embodied by selecting connection coefficients with respect to a particular parametrization), but if we choose particular ones, called the α-connections, certain differential geometric notions like flatness and autoparallel submanifolds can be related to familiar notions of probability distribution families/subfamilies like exponential families and mixture families. While the resulting theory is very elegant, it is also somewhat complex, and hence we must omit the general details, referring the interested reader to [39], and only introduce a small subset of the concepts that can be utilized via a series of examples.

In particular, let's focus our attention on the manifold $\mathcal{O}(\boldsymbol{\mathcal{X}})$ of probability distributions for a random vector $\boldsymbol{X} = (X_1, \ldots, X_N)$ taking values on the Cartesian product $\boldsymbol{\mathcal{X}}^{\times} = \mathcal{X}_1 \times \mathcal{X}_2 \times \cdots \times \mathcal{X}_N$. We already defined a $\prod_{n=1}^{N} |\mathcal{X}_n| - 1$ dimensional vector in (4.1) as $\boldsymbol{\eta}$-coordinates, which we also call m-coordinate in information geometry. Alternatively we could parameterize the probability mass function for such

a joint distribution with a vector $\boldsymbol{\theta}$, whose $\prod_{n=1}^{N} |\mathcal{X}_n| - 1$ elements take the form

$$\boldsymbol{\theta} = \left[\log\left(\frac{p_{\mathbf{X}}(i_1, \ldots, i_N)}{p_{\mathbf{X}}(1, \ldots, 1)} \right) \middle| \begin{array}{c} i_k \in \{1, 2, \ldots, |\mathcal{X}_k|\}, \\ k \in \{1, \ldots, N\}, \\ \sum_{k=1}^{N} i_k \neq N. \end{array} \right] \tag{5.3}$$

where $\mathcal{X}_n$ is a finite set with values denoted by i_k. These coordinates provide an alternate unique way of specifying the joint probability mass function $p_{\mathbf{X}}$, called the e-coordinates or $\boldsymbol{\theta}$ coordinates.

A subfamily of these probability mass functions associated with those $\boldsymbol{\eta}$ coordinates that take the form

$$\boldsymbol{\eta} = \mathbf{A}\boldsymbol{p} + \mathbf{b} \tag{5.4}$$

for some $\boldsymbol{p}$ for any particular fixed $\mathbf{A}$ and $\mathbf{b}$, that is, that lie in an affine submanifold of the $\boldsymbol{\eta}$ coordinates, are said to form a *m-autoparallel* submanifold of probability mass functions. This is a not a definition, but rather a consequence of a theorem involving a great deal of additional structure which must omit here [39].

Similarly, a subfamily of these probability mass functions associated with those $\boldsymbol{\theta}$ coordinates that take the form

$$\boldsymbol{\theta} = \mathbf{A}\boldsymbol{\lambda} + \mathbf{b} \tag{5.5}$$

for some $\boldsymbol{\lambda}$ for any particular fixed $\mathbf{A}$ and $\mathbf{b}$, that is, that lie in an affine submanifold of the $\boldsymbol{\theta}$ coordinates, are said to form a *e-autoparallel* submanifold of probability mass functions.

An e-autoparallel submanifold (resp. m-autoparallel submanifold) that is one dimensional, in that its $\boldsymbol{\lambda}$ (resp. $\boldsymbol{p}$) parameter vector is in fact a scalar, is called a *e-geodesic* (resp. *m-geodesic*).

On this manifold of probability mass functions for random variables taking values

in the set $\boldsymbol{\mathcal{X}}$, we can also define the Kullback Leibler divergence, or relative entropy, measured in bits, according to

$$D(p_{\boldsymbol{X}}||q_{\boldsymbol{X}}) = \sum_{\boldsymbol{x}\in\boldsymbol{\mathcal{X}}} p_{\boldsymbol{X}}(\boldsymbol{x}) \log_2 \left(\frac{p_{\boldsymbol{X}}(\boldsymbol{x})}{q_{\boldsymbol{X}}(\boldsymbol{x})} \right) \tag{5.6}$$

Note that in this context $D(p||q) \geq 0$ with equality if and only if $p = q$, and hence this function is a bit like a distance, however it does *not* in general satisfy symmetry or triangle inequality.

Let $\mathcal{E}$ be a particular e-autoparallel submanifold, and consider a probability distribution $p_{\boldsymbol{X}}$ not necessarily in this submanifold. The problem of finding the point $\overrightarrow{\Pi}_{\mathcal{E}}\ (p_{\boldsymbol{X}})$ in $\mathcal{E}$ closest in Kullback Leibler divergence to $p_{\boldsymbol{X}}$ defined by

$$\overrightarrow{\Pi}_{\mathcal{E}}\ (p_{\boldsymbol{X}}) \triangleq \arg\min_{q_{\boldsymbol{X}}\in\mathcal{E}} D(p_{\boldsymbol{X}}||q_{\boldsymbol{X}}) \tag{5.7}$$

is well posed, and is characterized in the following two ways (here $\overrightarrow{\Pi}_{\mathcal{E}}$ with a right arrow means we are minimizing over the second argument $q_{\boldsymbol{X}}$). The tangent vector of the m-geodesic connecting $p_{\boldsymbol{X}}$ to $\overrightarrow{\Pi}_{\mathcal{E}}\ (p_{\boldsymbol{X}})$ is orthogonal, in the sense of achieving Riemannian metric value 0, at $\overrightarrow{\Pi}_{\mathcal{E}}\ (p_{\boldsymbol{X}})$ to the tangent vector of the e-geodesic connecting $\overrightarrow{\Pi}_{\mathcal{E}}\ (p_{\boldsymbol{X}})$ and any other point in $\mathcal{E}$. Additionally, for any other point $q \in \mathcal{E}$, we have the Pythagorean like relation

$$D(p_{\boldsymbol{X}}||q_{\boldsymbol{X}}) = D(p_{\boldsymbol{X}}||\ \overrightarrow{\Pi}_{\mathcal{E}}\ (p_{\boldsymbol{X}})) + D(\overrightarrow{\Pi}_{\mathcal{E}}\ (p_{\boldsymbol{X}})||q_{\boldsymbol{X}}). \tag{5.8}$$

This relationship, which is an important one in information geometry [39], is depicted in Fig. 5.1.

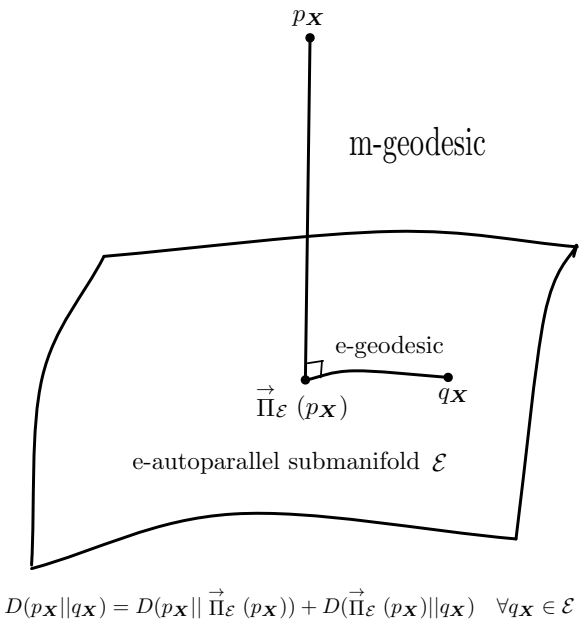

Figure 5.1: Pythagorean style relation.

5.2 Information Geometric Structure of the Shannon Faces of the Region of Entropic Vectors

As identified in §2.1, $\Gamma_2 = \Gamma_2^*$ and $\Gamma_3 = \overline{\Gamma}_3^*$, implying that Γ_2^* and $\overline{\Gamma}_3^*$ are fully characterized by Shannon type information inequalities. For example, when N = 2, Γ_2^* is a 3-dimensional polyhedral cone characterized by $H(X_1, X_2) - H(X_1) \geqslant 0$, $H(X_1, X_2) - H(X_2) \geqslant 0$ and $I(X_1; X_2) = H(X_1) + H(X_2) - H(X_1, X_2) \geqslant 0$ as depicted in Fig. 5.2.

For $N = 4$, even though $\Gamma_4 \neq \overline{\Gamma}_4^*$ and the region is not a polyhedral cone, there are still many exposed faces of $\overline{\Gamma}_N^*$ defined by attaining equality in a particular Shannon type information inequality of the form (2.3) or (2.4). Such exposed faces of $\overline{\Gamma}_N^*$ could be referred to as the "Shannon facets" of entropy, and in this section we will first aim to characterize the distributions associated with these Shannon facets via information geometry.

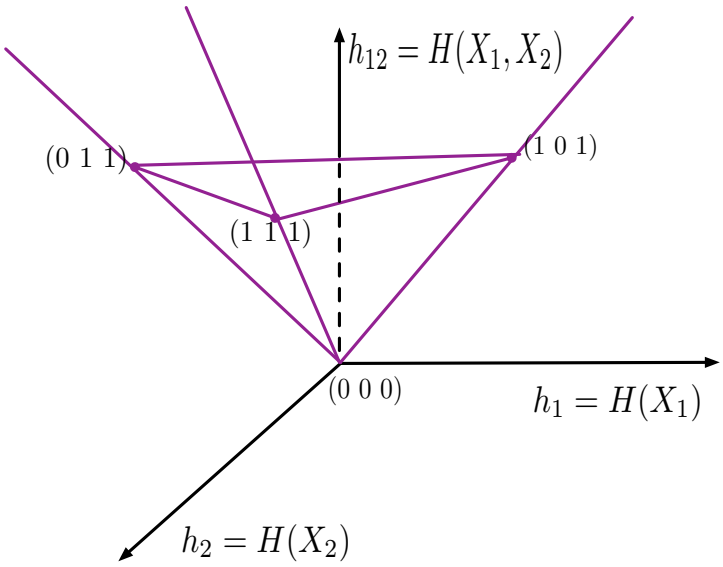

Figure 5.2: Region of entropic vector Γ_2^*.

Let $\mathcal{E}_{\mathcal{A}}^{\perp}$ be the submanifold of probability distributions for which $\boldsymbol{X}_{\mathcal{A}}$ and $\boldsymbol{X}_{\mathcal{A}^c} = \boldsymbol{X}_{\mathcal{N}\setminus\mathcal{A}}$ are independent

$$\mathcal{E}_{\mathcal{A}}^{\perp} = \{p_{\boldsymbol{X}}(\cdot) \mid p_{\boldsymbol{X}}(\boldsymbol{x}) = p_{\boldsymbol{X}_{\mathcal{A}}}(\boldsymbol{x}_{\mathcal{A}}) p_{\boldsymbol{X}_{\mathcal{A}^c}}(\boldsymbol{x}_{\mathcal{A}^c}) \quad \forall \boldsymbol{x}\} \tag{5.9}$$

then we define $\mathcal{E}_{\mathcal{A}\cup\mathcal{B}}^{\perp}$, $\mathcal{E}_{\mathcal{A},\mathcal{B}}^{\leftrightarrow,\perp}$ and $\mathcal{M}_{\mathcal{A}}$ as follows

$$\begin{aligned}
\mathcal{E}_{\mathcal{A}\cup\mathcal{B}}^{\perp} &= \left\{ p_{\boldsymbol{X}}(\cdot) \,\middle|\, p_{\boldsymbol{X}} = p_{\boldsymbol{X}_{(\mathcal{A}\cup\mathcal{B})}} p_{\boldsymbol{X}_{(\mathcal{A}\cup\mathcal{B})^c}} \right\} \\
\mathcal{E}_{\mathcal{A},\mathcal{B}}^{\leftrightarrow,\perp} &= \left\{ p_{\boldsymbol{X}}(\cdot) \,\middle|\, p_{\boldsymbol{X}} = p_{\boldsymbol{X}_{\mathcal{A}}} p_{\boldsymbol{X}_{\mathcal{B}\setminus\mathcal{A}}|\boldsymbol{X}_{\mathcal{A}\cap\mathcal{B}}} p_{\boldsymbol{X}_{(\mathcal{A}\cup\mathcal{B})^c}} \right\} \\
\mathcal{M}_{\mathcal{A}}^{w} &= \{p_{\boldsymbol{X}}(\cdot) \mid p_{\boldsymbol{X}} = p_{\boldsymbol{X}_{\mathcal{A}}} \cdot \delta_{\mathcal{A}^c}^{w}\}
\end{aligned}$$

where $\delta_{\mathcal{A}^c}^{w} = \begin{cases} 1 & \text{if } \boldsymbol{X}_{\mathcal{A}^c} = w(\boldsymbol{X}_{\mathcal{A}}) \\ 0 & \text{otherwise} \end{cases}$ for some fixed function $w : \boldsymbol{\mathcal{X}}_{\mathcal{A}} \to \boldsymbol{\mathcal{X}}_{\mathcal{A}^c}$.

Note $\mathcal{E}^{\leftrightarrow,\perp}_{\mathcal{A},\mathcal{B}}$ is a submanifold of $\mathcal{E}^{\perp}_{\mathcal{A}\cup\mathcal{B}}$ such that the random variables $\boldsymbol{X}_{\mathcal{A}\cup\mathcal{B}}$, in addition to being independent from $\boldsymbol{X}_{(\mathcal{A}\cup\mathcal{B})^c}$ form the Markov chain $\boldsymbol{X}_{\mathcal{A}\setminus\mathcal{B}} \leftrightarrow \boldsymbol{X}_{\mathcal{A}\cap\mathcal{B}} \leftrightarrow \boldsymbol{X}_{\mathcal{B}\setminus\mathcal{A}}$. These sets of distributions will be useful because $I(\boldsymbol{X}_{\mathcal{A}\cup\mathcal{B}}; \boldsymbol{X}_{(\mathcal{A}\cup\mathcal{B})^c}) = h_{\mathcal{A}\cup\mathcal{B}} + h_{(\mathcal{A}\cup\mathcal{B})^c} - h_{\mathcal{N}} = 0$ for every distribution in $\mathcal{E}^{\perp}_{\mathcal{A}\cup\mathcal{B}}$ and $\mathcal{E}^{\leftrightarrow,\perp}_{\mathcal{A},\mathcal{B}}$, $I(\boldsymbol{X}_{\mathcal{A}\setminus\mathcal{B}}; \boldsymbol{X}_{\mathcal{B}\setminus\mathcal{A}}|\boldsymbol{X}_{\mathcal{A}\cap\mathcal{B}}) = h_{\mathcal{A}} + h_{\mathcal{B}} - h_{(\mathcal{A}\cap\mathcal{B})} - h_{(\mathcal{A}\cup\mathcal{B})} = 0$ for every distribution in $\mathcal{E}^{\leftrightarrow,\perp}_{\mathcal{A},\mathcal{B}}$, and $h_{\mathcal{N}} - h_{\mathcal{A}} = 0$ for every distribution in $\mathcal{M}^{w}_{\mathcal{A}}$ for any w.

We will show that the sets of distributions $\mathcal{E}^{\perp}_{\mathcal{A}\cup\mathcal{B}}$, $\mathcal{E}^{\leftrightarrow,\perp}_{\mathcal{A},\mathcal{B}}$, $\mathcal{M}^{w}_{\mathcal{A}}$ are of interest because they correspond to every Shannon facet. This is because every facet of Γ_N, which must correspond to equality in an inequality of the form (2.3) or (2.4), can be regarded as some conditional entropy, mutual information, or conditional mutual information being identically zero. For example, as depicted in Fig. 5.2, Γ^*_2 has three facets corresponding to set of distributions $\mathcal{E}^{\perp}_{1\cup 2}$, $\{\mathcal{M}^{w}_{1}\}$ and $\{\mathcal{M}^{w}_{2}\}$ respectively. Surprisingly, while the entropy vector is a nonlinear function of the joint distribution, and these families of distributions correspond to the intersection of affine sets with the region of entropic vectors, they are themselves, when regarded with the correct information geometric parameterization, *associated with affine sets of distributions*.

In order to prove the information geometric properties of Shannon facets, we first give the equivalent condition of $I(\boldsymbol{X}_{\mathcal{A}\setminus\mathcal{B}}; \boldsymbol{X}_{\mathcal{B}\setminus\mathcal{A}}|\boldsymbol{X}_{\mathcal{A}\cap\mathcal{B}}) = 0$ in distribution space, especially in $\boldsymbol{\theta}$ coordinate as mentioned in (5.5).

Lemma 3: Let $\boldsymbol{X} = \{X_1, X_2, \cdots, X_N\}$, $\mathcal{A}, \mathcal{B} \subset \mathcal{N} = \{1, 2, \cdots N\}$ where $|\mathcal{X}_{\mathcal{A}\setminus\mathcal{B}}| = m$, $|\mathcal{X}_{\mathcal{B}\setminus\mathcal{A}}| = n$ and $|\mathcal{X}_{\mathcal{A}\cap\mathcal{B}}| = q$. For each assignment C_0 of $\boldsymbol{X}_{\mathcal{A}\cap\mathcal{B}}$, define the following equations for $i = 1, 2, \cdots, m-1$ and $j = 1, 2, \cdots, n-1$ for the probabilities $p_{\mathbf{X}_{\mathcal{A}\setminus\mathcal{B}}, \mathbf{X}_{\mathcal{B}\setminus\mathcal{A}}, \mathcal{X}_{\mathcal{A}\cap\mathcal{B}}}$:

$$\frac{p_{0jC_0} p_{i0C_0}}{p_{00C_0} p_{ijC_0}} = 1 \tag{5.10}$$

which is equivalent with hyperplane (5.11) in $\boldsymbol{\theta}$ coordinates for $p_{\mathbf{X}_{\mathcal{A}\cup\mathcal{B}}}$,

$$\theta_{0jC_0} + \theta_{i0C_0} - \mathbb{1}[C_0 \neq \mathbf{0}] \cdot \theta_{00C_0} - \theta_{ijC_0} = 0 \tag{5.11}$$

then $I(\boldsymbol{X}_{\mathcal{A}\setminus\mathcal{B}}; \boldsymbol{X}_{\mathcal{B}\setminus\mathcal{A}}|\boldsymbol{X}_{\mathcal{A}\cap\mathcal{B}}) = 0$ and the set of equations (5.10), (5.11) are equivalent.

Proof: From Equation (5.10), it can be verified that

$$\frac{p_{wjC_0}p_{ieC_0}}{p_{weC_0}p_{ijC_0}} = 1 \tag{5.12}$$

for any distinct i, j, w and e such that $w, i \in \{0, 1, \cdots, m-1\}$ and $j, e \in \{0, 1, \cdots, n-1\}$. More specifically, the equations in (5.12) but not in (5.10) all can be derived by combining some of the $(m-1)(n-1)$ equations in (5.10).

$\Rightarrow$ Since $I(\boldsymbol{X}_{\mathcal{A}\setminus\mathcal{B}}; \boldsymbol{X}_{\mathcal{B}\setminus\mathcal{A}}|\boldsymbol{X}_{\mathcal{A}\cap\mathcal{B}}) = 0$, for each assignment C_0 of $\boldsymbol{X}_{\mathcal{A}\cap\mathcal{B}}$, it can be verified [2] that

$$\begin{aligned} &p_{\boldsymbol{X}_{\mathcal{A}\setminus\mathcal{B}}\boldsymbol{X}_{\mathcal{B}\setminus\mathcal{A}}|\boldsymbol{X}_{\mathcal{A}\cap\mathcal{B}}}(x_{\mathcal{A}\setminus\mathcal{B}} = i, x_{\mathcal{B}\setminus\mathcal{A}} = j, x_{\mathcal{A}\cap\mathcal{B}} = C_0) \\ &= p_{\boldsymbol{X}_{\mathcal{A}\setminus\mathcal{B}}|\boldsymbol{X}_{\mathcal{A}\cap\mathcal{B}}}(x_{\mathcal{A}\setminus\mathcal{B}} = i|x_{\mathcal{A}\cap\mathcal{B}} = C_0) \\ &\quad p_{\boldsymbol{X}_{\mathcal{B}\setminus\mathcal{A}}|\boldsymbol{X}_{\mathcal{A}\cap\mathcal{B}}}(x_{\mathcal{B}\setminus\mathcal{A}} = j|x_{\mathcal{A}\cap\mathcal{B}} = C_0) \end{aligned}$$

which make sure the nominator and denominator cancelled with each other in (5.10) and (5.12).

$\Leftarrow$ Now suppose (5.10) , (5.11) and (5.12) hold for $w, i \in \{0, 1, \cdots, m-1\}$ and $j, e \in \{0, 1, \cdots, n-1\}$, it is suffices to show for each assignment C_0 of $\boldsymbol{X}_{\mathcal{A}\cap\mathcal{B}}$,

$$H(\boldsymbol{X}_{\mathcal{A}\setminus\mathcal{B}}|\boldsymbol{X}_{\mathcal{A}\cap\mathcal{B}}) + H(\boldsymbol{X}_{\mathcal{B}\setminus\mathcal{A}}|\boldsymbol{X}_{\mathcal{A}\cap\mathcal{B}}) = H(\boldsymbol{X}_{\mathcal{A}\setminus\mathcal{B}}, \boldsymbol{X}_{\mathcal{B}\setminus\mathcal{A}}|\boldsymbol{X}_{\mathcal{A}\cap\mathcal{B}}) \tag{5.13}$$

We can calculate:

$$H(\boldsymbol{X}_{\boldsymbol{\mathcal{A}}\setminus\boldsymbol{\mathcal{B}}}|\boldsymbol{X}_{\boldsymbol{\mathcal{A}}\cap\boldsymbol{\mathcal{B}}}=C_0)=\sum_{i=0}^{m-1}[(\sum_{j=0}^{n-1}p_{ijC_0})\log(\sum_{j=0}^{n-1}p_{ijC_0})] \tag{5.14}$$

$$H(\boldsymbol{X}_{\boldsymbol{\mathcal{B}}\setminus\boldsymbol{\mathcal{A}}}|\boldsymbol{X}_{\boldsymbol{\mathcal{A}}\cap\boldsymbol{\mathcal{B}}}=C_0)=\sum_{j=0}^{n-1}[(\sum_{i=0}^{m-1}p_{ijC_0})\log(\sum_{i=0}^{m-1}p_{ijC_0})] \tag{5.15}$$

$$H(\boldsymbol{X}_{\boldsymbol{\mathcal{A}}\setminus\boldsymbol{\mathcal{B}}},\boldsymbol{X}_{\boldsymbol{\mathcal{B}}\setminus\boldsymbol{\mathcal{A}}}|\boldsymbol{X}_{\boldsymbol{\mathcal{A}}\cap\boldsymbol{\mathcal{B}}}=C_0)=\sum_{i=0}^{m-1}\sum_{j=0}^{n-1}p_{ijC_0}\log p_{ijC_0} \tag{5.16}$$

From (5.14) and (5.15) we have

$$\begin{aligned}H(\boldsymbol{X}_{\boldsymbol{\mathcal{A}}\setminus\boldsymbol{\mathcal{B}}}|\boldsymbol{X}_{\boldsymbol{\mathcal{A}}\cap\boldsymbol{\mathcal{B}}}=C_0)+H(\boldsymbol{X}_{\boldsymbol{\mathcal{B}}\setminus\boldsymbol{\mathcal{A}}}|\boldsymbol{X}_{\boldsymbol{\mathcal{A}}\cap\boldsymbol{\mathcal{B}}}=C_0)=\\ \sum_{i=0}^{m-1}\sum_{j=0}^{n-1}p_{ijC_0}\log[(\sum_{e=0}^{n-1}p_{ieC_0})(\sum_{w=0}^{m-1}p_{wjC_0})]\end{aligned} \tag{5.17}$$

where $(\sum_{e=0}^{n-1}p_{ieC_0})(\sum_{w=0}^{m-1}p_{wjC_0})=p_{ijC_0}$ when (5.12) are satisfied, in which case (5.13) hold. □

Now let's take $I(X_1;X_2|X_3)=0$ as a example:

Example 1: Let $\boldsymbol{X}=\{X_1,X_2,X_3\}$, where $|\mathcal{X}_1|=m$, $|\mathcal{X}_2|=n$ and $|\mathcal{X}_3|=q$. For mnq different values of $\boldsymbol{X}$, if we want to parameterize this manifold $\mathcal{O}_3$ in $\boldsymbol{\theta}$ coordinate, we will need $(mnq-1)$ different parameters: θ_1, θ_2, $\cdots$, θ_{mnq-1}, which can be defined as in (5.3). We define submanifold $\mathcal{E}_3$ to satisfy the constraint $I(X_1;X_2|X_3)=0$. It is easy to verify that the dimension of $\mathcal{O}_3$ is $(mnq-1)$, the dimension of $\mathcal{E}_3$ is $((m+n-1)q-1)$. We know any hyperplane $\sum_i^{mnq-1}\lambda_i\theta_i=0$ is a $(mnq-2)$ dimensional e-autoparallel submanifold in $\mathcal{O}_3$. Then subtracting $((m+n-1)q-1)+1$ from $mnq-2$, we know $\mathcal{E}_3$ is the intersection of $(m-1)(n-1)q$ different hyperplanes, which are $\sum_{i=1}^{mnq-1}\lambda_i^k\theta_i=0$

for $k = 1, 2, \cdots, (m-1)(n-1)q$. For $i = 1, 2, \cdots, m-1$, $j = 1, 2, \cdots, n-1$ and $r = 1, 2, \cdots, q$, each hyperplane $\sum_{i=1}^{mnq-1} \lambda_i^k \theta_i = 0$ corresponding to a constraint on the probability:

$$\frac{p_{0jr} p_{i0r}}{p_{00r} p_{ijr}} = 1 \tag{5.18}$$

which can be written as

$$\theta_{0jr} + \theta_{i0r} - \mathbb{1}[r \neq 0] \cdot \theta_{00r} - \theta_{ijr} = 0 \tag{5.19}$$

where $\mathbb{1}[r \neq 0] = 0$ when $r = 0$.

Now we can use Lemma 3 to prove Theorem 7, then use the relationship between m-autoparallel submanifold and affine subspace to prove Theorem 8:

Theorem 7: $\mathcal{E}_{\mathcal{A},\mathcal{B}}^{\leftrightarrow,\perp} \subseteq \mathcal{E}_{\mathcal{A}\cup\mathcal{B}}^{\perp} \subseteq \mathcal{O}(\boldsymbol{\mathcal{X}})$, $\mathcal{E}_{\mathcal{A},\mathcal{B}}^{\leftrightarrow,\perp}$ is an e-autoparallel submanifold of $\mathcal{E}_{\mathcal{A}\cup\mathcal{B}}^{\perp}$ and $\mathcal{E}_{\mathcal{A}\cup\mathcal{B}}^{\perp}$ is an e-autoparallel submanifold of $\mathcal{O}(\boldsymbol{\mathcal{X}})$.

Proof: Follows directly from Bayes' rule, we can rewrite $\mathcal{E}_{\mathcal{A}\cup\mathcal{B}}^{\perp}$ and $\mathcal{O}(\boldsymbol{\mathcal{X}})$ as

$$\mathcal{E}_{\mathcal{A}\cup\mathcal{B}}^{\perp} = \left\{ p_{\boldsymbol{X}}(\cdot) \;\middle|\; p_{\boldsymbol{X}} = p_{\boldsymbol{X}_{\mathcal{A}}} p_{\boldsymbol{X}_{\mathcal{B}\setminus\mathcal{A}}|\boldsymbol{X}_{\mathcal{A}}} p_{\boldsymbol{X}_{(\mathcal{A}\cup\mathcal{B})^c}} \right\}$$
$$\mathcal{O}(\boldsymbol{\mathcal{X}}) = \left\{ p_{\boldsymbol{X}}(\cdot) \;\middle|\; p_{\boldsymbol{X}} = p_{\boldsymbol{X}_{(\mathcal{A}\cup\mathcal{B})}} p_{\boldsymbol{X}_{(\mathcal{A}\cup\mathcal{B})^c|(\mathcal{A}\cup\mathcal{B})}} \right\}$$

then we can easily verify $\mathcal{E}_{\mathcal{A},\mathcal{B}}^{\leftrightarrow,\perp} \subseteq \mathcal{E}_{\mathcal{A}\cup\mathcal{B}}^{\perp} \subseteq \mathcal{O}(\boldsymbol{\mathcal{X}})$. Since from Lemma 3, the equivalent condition of $I(\boldsymbol{X}_{\mathcal{A}\setminus\mathcal{B}}; \boldsymbol{X}_{\mathcal{B}\setminus\mathcal{A}}|\boldsymbol{X}_{\mathcal{A}\cap\mathcal{B}}) = 0$ in e-coordinate for $p_{\mathbf{X}_{\mathcal{A}\cup\mathcal{B}}}$ is the intersection of a sequences of hyperplanes (5.11) and the case of $\mathcal{A} \cap \mathcal{B} = \emptyset$ can be considered as a special case of Lemma 3. Then from the definition of autoparallel submanifold, $\mathcal{E}_{\mathcal{A},\mathcal{B}}^{\leftrightarrow,\perp}$ is e-autoparallel in $\mathcal{E}_{\mathcal{A}\cup\mathcal{B}}^{\perp}$, and $\mathcal{E}_{\mathcal{A}\cup\mathcal{B}}^{\perp}$ is e-autoparallel in $\mathcal{O}(\boldsymbol{\mathcal{X}})$. □

Theorem 8: Let $\mathcal{A} \subseteq \mathcal{B} \subseteq \mathcal{N} = \{1, 2, \cdots N\}$, then $\mathcal{M}_{\mathcal{A}}^{w} \subseteq \mathcal{M}_{\mathcal{B}}^{w'} \subseteq \mathcal{O}(\boldsymbol{\mathcal{X}})$, $\mathcal{M}_{\mathcal{A}}^{w}$ is a

m-autoparallel submanifold of $\mathcal{M}_{\mathcal{B}}^{w'}$ and $\mathcal{M}_{\mathcal{B}}^{w'}$ is a m-autoparallel submanifold of $\mathcal{O}(\boldsymbol{\mathcal{X}})$.

Proof: It follows directly from the definition of m-autoparallel submanifold since for the m-affine coordinate of $\mathcal{O}(\boldsymbol{\mathcal{X}})$ and $\mathcal{M}_{\mathcal{B}}^{w'}$, we can easily find the corresponding matrix A and vector B such that $\mathcal{M}_{\mathcal{B}}^{w'}$ is affine subspace of $\mathcal{O}(\boldsymbol{\mathcal{X}})$; similarly for the m-affine coordinate of $\mathcal{M}_{\mathcal{A}}^{w}$ and $\mathcal{M}_{\mathcal{B}}^{w'}$, we can easily find the corresponding matrix $\mathbf{A}$ and vector $\mathbf{b}$ such that $\mathcal{M}_{\mathcal{A}}^{w}$ is affine subspace of $\mathcal{M}_{\mathcal{B}}^{w}$. □

Theorem 7 and Theorem 8, have shown that Shannon facets are associated with affine subsets of the family of probability distribution, when it is regarded in an appropriate parameterization. In fact, as we shall presently show, all Shannon Type information inequalities correspond to the positivity of divergences of projections to these submanifolds. Note that $\mathcal{E}_{\mathcal{A},\mathcal{B}}^{\leftrightarrow,\perp}$ is set up so that the difference between the right and left sides in the entropy submodularity inequality (2.3) is zero for every probability distribution in $\mathcal{E}_{\mathcal{A},\mathcal{B}}^{\leftrightarrow,\perp}$. The nested nature of these e-autoparallel submanifolds, and the associated Pythagorean relation, is one way to view the submodularity of entropy, as we shall now explain with Figure 5.3 and Corollary 2.

Corollary 2: The submodularity (2.3) of the entropy function can be viewed as a consequence of Pythagorean style information projection relationships depicted in Figure 5.3. In particular, submodularity is equivalent to the inequality

$$D\left(\overrightarrow{\Pi}_{\mathcal{E}_{\mathcal{A}\cup\mathcal{B}}^{\perp}}(p_{\boldsymbol{X}})||\,\overrightarrow{\Pi}_{\mathcal{E}_{\mathcal{A},\mathcal{B}}^{\leftrightarrow,\perp}}(p_{\boldsymbol{X}})\right)\geq 0 \tag{5.20}$$

since

$$\begin{aligned} D\left(\overrightarrow{\Pi}_{\mathcal{E}_{\mathcal{A}\cup\mathcal{B}}^{\perp}}(p_{\boldsymbol{X}})||\,\overrightarrow{\Pi}_{\mathcal{E}_{\mathcal{A},\mathcal{B}}^{\leftrightarrow,\perp}}(p_{\boldsymbol{X}})\right) &= H(\boldsymbol{X}_{\mathcal{A}})+H(\boldsymbol{X}_{\mathcal{B}}) \\ &\quad -H(\boldsymbol{X}_{\mathcal{A}\cap\mathcal{B}})-H(\boldsymbol{X}_{\mathcal{A}\cup\mathcal{B}}) \end{aligned}$$

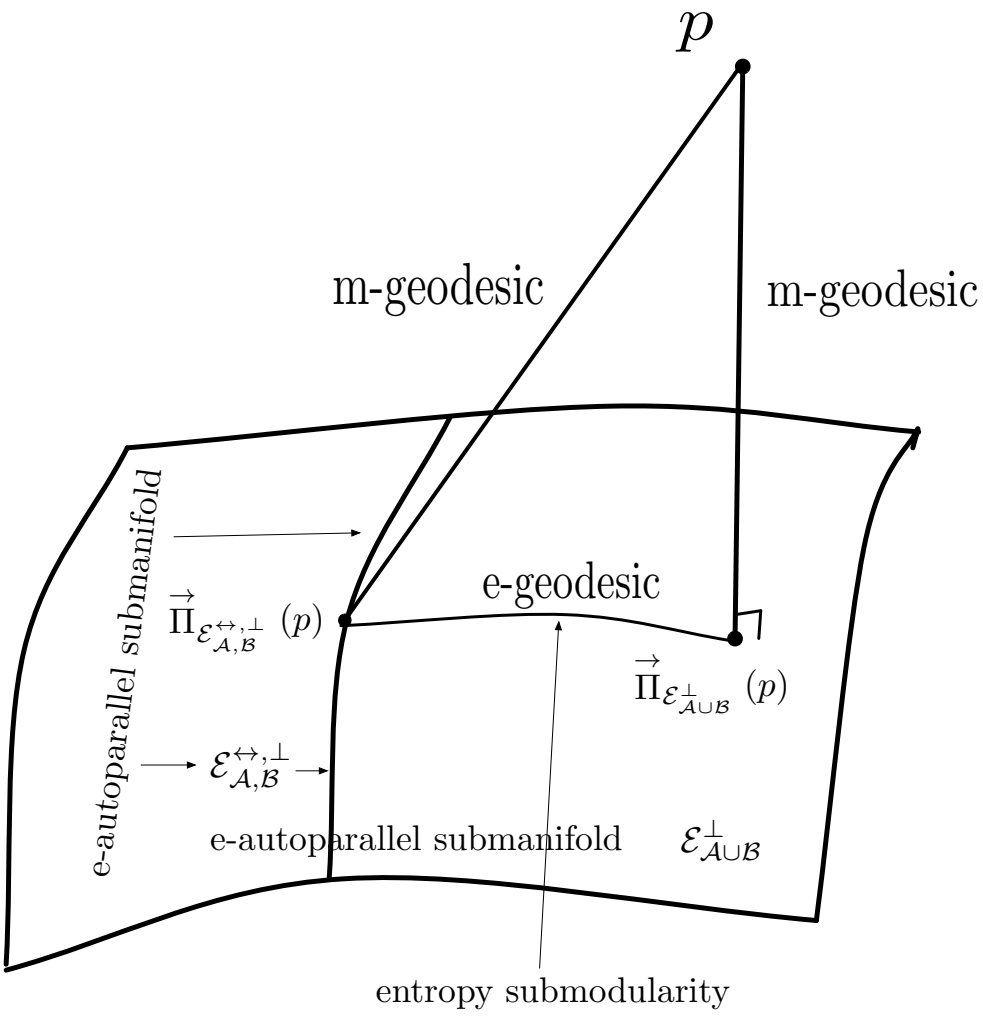

Figure 5.3: Submodularity of the entropy function is equivalent to the non-negativity of a divergence $D(\overrightarrow{\pi}_{\mathcal{E}^{\perp}_{\mathcal{A}\cup\mathcal{B}}}(p)||\ \overrightarrow{\pi}_{\mathcal{E}^{\leftrightarrow,\perp}_{\mathcal{A},\mathcal{B}}}(p))$ between two information projections, one projection $(\overrightarrow{\pi}_{\mathcal{E}^{\leftrightarrow,\perp}_{\mathcal{A},\mathcal{B}}}(p))$ is to a set that is a submanifold the other projection's $(\overrightarrow{\pi}_{\mathcal{E}^{\perp}_{\mathcal{A}\cup\mathcal{B}}}(p))$ set. A Pythagorean style relation shows that for such an arrangement $D(p||\ \overrightarrow{\pi}_{\mathcal{E}^{\leftrightarrow,\perp}_{\mathcal{A},\mathcal{B}}}(p)) = D(p||\ \overrightarrow{\pi}_{\mathcal{E}^{\perp}_{\mathcal{A}\cup\mathcal{B}}}(p)) + D(\overrightarrow{\pi}_{\mathcal{E}^{\perp}_{\mathcal{A}\cup\mathcal{B}}}(p)||\ \overrightarrow{\pi}_{\mathcal{E}^{\leftrightarrow,\perp}_{\mathcal{A},\mathcal{B}}}(p))$.

Proof: The projections are

$$\overrightarrow{\Pi}_{\mathcal{E}^{\perp}_{\mathcal{A}\cup\mathcal{B}}}(p_{\boldsymbol{X}}) = p_{\boldsymbol{X}_{\mathcal{A}\cup\mathcal{B}}} p_{\boldsymbol{X}_{(\mathcal{A}\cup\mathcal{B})^c}} \tag{5.21}$$

and

$$\overrightarrow{\Pi}_{\mathcal{E}^{\leftrightarrow,\perp}_{\mathcal{A},\mathcal{B}}}(p_{\boldsymbol{X}}) = p_{\boldsymbol{X}_{\mathcal{A}\setminus\mathcal{B}}|\boldsymbol{X}_{\mathcal{A}\cap\mathcal{B}}} p_{\boldsymbol{X}_{\mathcal{B}}} p_{\boldsymbol{X}_{(\mathcal{A}\cup\mathcal{B})^c}} \tag{5.22}$$

since for $\overrightarrow{\Pi}_{\mathcal{E}^{\perp}_{\mathcal{A}\cup\mathcal{B}}}(p_{\boldsymbol{X}})$ given by (5.21) for every $q_{\boldsymbol{X}} \in \mathcal{E}^{\perp}_{\mathcal{A}\cup\mathcal{B}}$ we can reorganize the divergence as

$$\begin{aligned} D(p_{\boldsymbol{X}}||q_{\boldsymbol{X}}) &= D(p_{\boldsymbol{X}}||\overrightarrow{\Pi}_{\mathcal{E}^{\perp}_{\mathcal{A}\cup\mathcal{B}}}(p_{\boldsymbol{X}})) \\ &+ D(\overrightarrow{\Pi}_{\mathcal{E}^{\perp}_{\mathcal{A}\cup\mathcal{B}}}(p_{\boldsymbol{X}})||q_{\boldsymbol{X}_{(\mathcal{A}\cup\mathcal{B})}} q_{\boldsymbol{X}_{(\mathcal{A}\cup\mathcal{B})^c}}) \end{aligned}$$

and for $\overrightarrow{\Pi}_{\mathcal{E}^{\leftrightarrow,\perp}_{\mathcal{A},\mathcal{B}}}(p_{\boldsymbol{X}})$ given by (5.22) for every $q_{\boldsymbol{X}} \in \mathcal{E}^{\leftrightarrow,\perp}_{\mathcal{A},\mathcal{B}}$ we can reorganize the divergence as

$$\begin{aligned} D(p_{\boldsymbol{X}}||q_{\boldsymbol{X}}) &= D(p_{\boldsymbol{X}}||\overrightarrow{\Pi}_{\mathcal{E}^{\leftrightarrow,\perp}_{\mathcal{A},\mathcal{B}}}(p_{\boldsymbol{X}})) \\ &+ D(\overrightarrow{\Pi}_{\mathcal{E}^{\leftrightarrow,\perp}_{\mathcal{A},\mathcal{B}}}(p_{\boldsymbol{X}})||q_{\boldsymbol{X}_{\mathcal{A}\setminus\mathcal{B}}|\boldsymbol{X}_{\mathcal{A}\cap\mathcal{B}}} q_{\boldsymbol{X}_{\mathcal{B}}} q_{\boldsymbol{X}_{(\mathcal{A}\cup\mathcal{B})^c}}) \end{aligned}$$

The remainder of the corollary is proved by substituting (5.21) and (5.22) in the equation for the divergence. □

In order to further illustrate these ideas, we will demonstrate Theorem 7, Theorem 8 and Corollary 2 in the context of $\overline{\Gamma}^*_3$.

Example 2: Setting $N = 3$ we have $\boldsymbol{\mathcal{X}} = \mathcal{X}_1 \times \mathcal{X}_2 \times \mathcal{X}_3$. Let $\mathcal{N} = \{1, 2, 3\}$, and denote $\mathcal{A}$ and $\mathcal{B}$ as some subset of $\mathcal{N}$ that is not equal to $\emptyset$ and $\mathcal{N}$, then we consider the

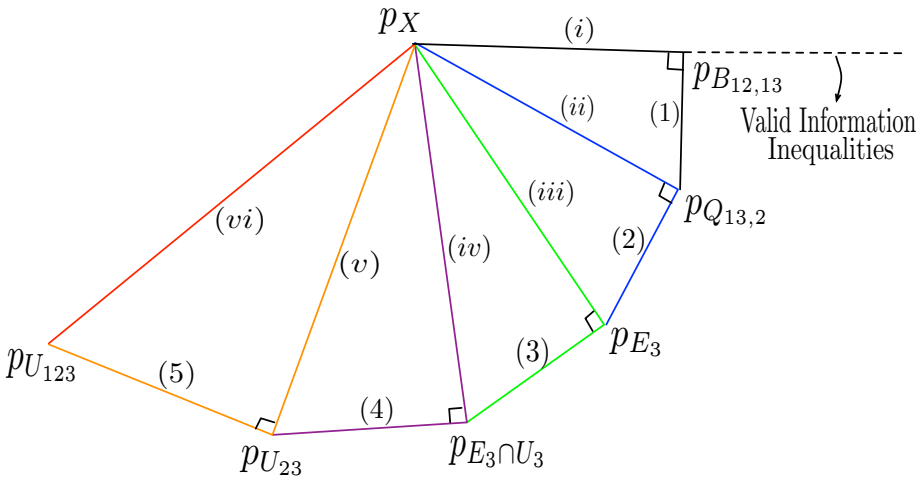

Figure 5.4: Pythagorean relationship on 3 discrete random variables

following submanifolds of $\mathcal{O}(\boldsymbol{\mathcal{X}})$:

$$B_{\mathcal{A},\mathcal{B}} = \left\{ p_{\boldsymbol{X}}(\cdot) \,\middle|\, p_{\boldsymbol{X}} = p_{\boldsymbol{X}_{\mathcal{A}}} \cdot p_{\boldsymbol{X}_{\mathcal{B}\setminus\mathcal{A}|(\mathcal{A}\cap\mathcal{B})}} \right\}$$
$$Q_{\mathcal{A},\mathcal{B}} = \left\{ p_{\boldsymbol{X}}(\cdot) \,|\, p_{\boldsymbol{X}} = p_{\boldsymbol{X}_{\mathcal{A}}} \cdot p_{\boldsymbol{X}_{\mathcal{B}}} \right\}$$
$$E_3 = \left\{ p_{\boldsymbol{X}}(\cdot) \,\middle|\, p_{\boldsymbol{X}} = \prod_{i=1}^{3} p_{X_i} \right\}$$
$$U_{\mathcal{A}} = \left\{ p_{\boldsymbol{X}}(\cdot) \,\middle|\, p_{\boldsymbol{X}} = \frac{1}{\prod_{i\in\mathcal{A}} |\chi_i|} p_{\boldsymbol{X}_{\mathcal{A}^c}} \right\}$$

Theorem 7 and Theorem 8 show that $B_{\mathcal{A},\mathcal{B}}$, $Q_{\mathcal{A},\mathcal{B}}$, E_3 and $U_{\mathcal{A}}$ are all nested e-autoparallel submanifolds of $\mathcal{O}(\boldsymbol{\mathcal{X}})$. For example, $U_{123} \subset U_{23} \subset E_3 \cap U_3 \subset E_3 \subset Q_{13,2} \subset B_{12,13} \subset S_3$, and in this chain, the previous one is the e-autoparallel submanifold of the later one. If we denote $p_{B_{12,13}}$, $p_{Q_{13,2}}$, p_{E_3}, $p_{E_3\cap U_3}$, $p_{U_{23}}$ and $p_{U_{123}}$ the m-projection of a point $p \in S_3 \setminus B_{12,13}$ onto $B_{12,13}$, $Q_{13,2}$, E_3, $E_3 \cap U_3$, U_{23} and U_{123} respectively, we will have the Pythagorean relation as shown in Figure 5.4. From these pythagorean relations, we will get a series of

Information inequalities

$$
\begin{aligned}
&(i) \quad h_{12} + h_{13} \geqslant h_1 + h_{123} \\
&(ii) \quad h_{13} + h_2 \geqslant h_{123} \\
&(iii) \quad h_1 + h_2 + h_3 \geqslant h_{123} \\
&(iv) \quad h_1 + h_2 + \log|\chi_3| \geqslant h_{123} \\
&(v) \quad h_1 + \log|\chi_2| + \log|\chi_3| \geqslant h_{123} \\
&(vi) \quad \log|\chi| \geqslant h_{123}
\end{aligned}
$$

as well as

$$
\begin{aligned}
&(1) \quad h_1 + h_2 \geqslant h_{12} \\
&(2) \quad h_1 + h_3 \geqslant h_{13} \\
&(3) \quad \log|\chi_3| \geqslant h_3 \\
&(4) \quad \log|\chi_2| \geqslant h_2 \\
&(5) \quad \log|\chi_1| \geqslant h_1
\end{aligned}
$$

where each inequality corresponds to the positivity of the divergence with the associated label in Figure 5.4. As these are the basic Shannon type inequalities for $N = 3$, this diagram yields an information geometric interpretation for the region of entropic vectors $\overline{\Gamma}_3^*$.

Lemma 3 can also be used to derive information geometric results for certain Shannon faces.

Theorem 9: Let relation $\mathcal{L}' \subseteq \mathcal{S}(N)$ be the containment minimal *p-representable* semimatroid containing a set $\mathcal{L} \subseteq \mathcal{S}(N)$ such that $\{i \cup j \cup \mathcal{K}\} = \mathcal{N}' \subseteq \mathcal{N}$ for $\forall (i, j|\mathcal{K}) \in$

$\mathcal{L}$. Denote $\mathcal{O}(\boldsymbol{\mathcal{X}}')$ the manifold of probability distributions for random vector $\boldsymbol{X}_{\mathcal{N}'}$. Let $N' = |\mathcal{N}'|$ and define $\mathcal{F}_{\mathcal{N}'}$ in (5.23),

$$\mathcal{F}_{\mathcal{N}'} = \{\ \boldsymbol{h} \in \Gamma_{N'} \mid h_{i\mathcal{K}} + h_{j\mathcal{K}} - h_{\mathcal{K}} - h_{ij\mathcal{K}} = 0,\ \forall (i,j|\mathcal{K}) \in \mathcal{L}'\} \tag{5.23}$$

then the set of distributions corresponding to $\mathcal{F}_{\mathcal{N}'}$ form a e-autoparallel submanifold in $\mathcal{O}(\boldsymbol{\mathcal{X}}')$.

Proof: From Corollary 1 we know $\mathcal{F}_{\mathcal{N}'}$ is a face of $\Gamma_{N'}$. As a result of Lemma 3, each $(i,j|\mathcal{K}) \in \mathcal{L}$ correspond to a series of hyperplanes (5.11) in $\boldsymbol{\theta}$ coordinate, we use $E_{eq}^{\mathcal{F}_0}$ to denote the intersection of all the hyperplanes representing any $(i,j|\mathcal{K})$ pairs in $\mathcal{L}$. Since for the submanifold

$$\mathcal{E}^{\perp} = \left\{ p_{\boldsymbol{X}}(\cdot) \,\middle|\, p_{\boldsymbol{X}}(\boldsymbol{x}) = \prod_{i \in \mathcal{N}'} p_{X_i}(x_i) \right\}$$

we have $\boldsymbol{h}(\mathcal{E}^{\perp}) \subseteq \mathcal{F}_{\mathcal{N}'}$, which means the intersection $E_{eq}^{\mathcal{F}_0}$ will always be non-empty, and hence there will always exist a p-representable semimatroid containing $\mathcal{L}$. The containment minimal p-representable matroid will contain any other conditional independence relations implied by those in $\mathcal{L}$, and hence any point in $E_{eq}^{\mathcal{F}_0}$ will obey these too. Furthermore, $E_{eq}^{\mathcal{F}_0}$ is e-autoparallel in $\mathcal{O}(\boldsymbol{\mathcal{X}}')$ since it is the non-empty interactions of hyperplanes in $\boldsymbol{\theta}$ coordinate according to Lemma 3. □

Example 3: As a example, recall in the last paragraph of section §2.4 we study the relationship between subset of the extreme rays of G_4^{34} and $p\text{-}representable$ semimatroids $\{(1,3|2)\}$, $\{(1,3|2)\ \&\ (2,3|1)\}$ and $\{(1,2|3)\ \&\ (1,3|2)\ \&\ (2,3|1)\}$. Since all these three semimatroids are $p\text{-}representable$, there are distributions satisfy the conditional independent relations corresponding to these $(i,j|\mathcal{K})$ couples. For example, the distributions corresponding to semimatroid $\{(1,2|3)\ \&\ (1,3|2)\ \&\ (2,3|1)\}$ satisfied the constraints

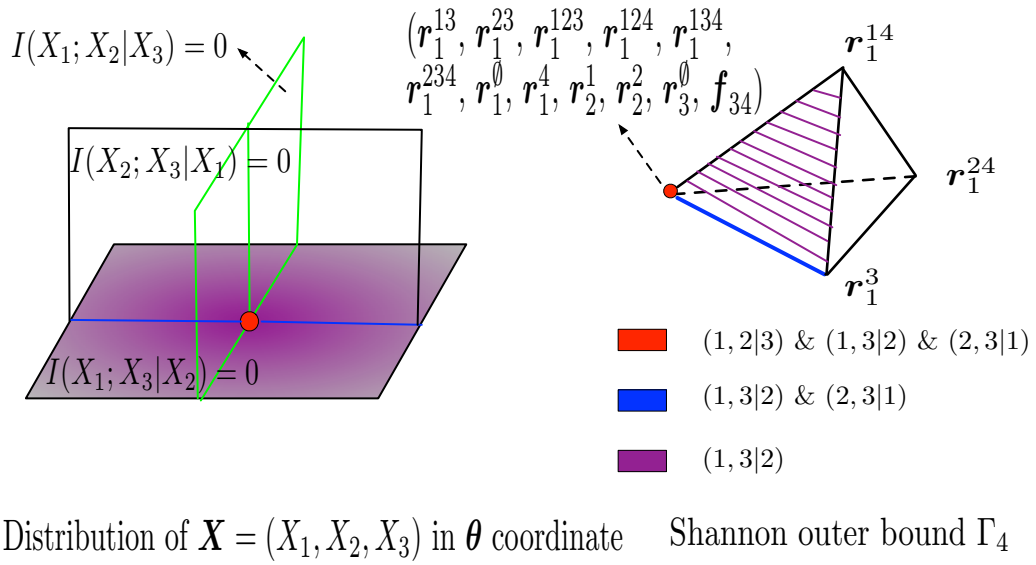

Figure 5.5: The mapping between some faces of Γ_4 and submanifold of four variable distribution in $\boldsymbol{\theta}$ coordinate

$I(X_1;X_2|X_3)=0$, $I(X_1;X_3|X_2)=0$ and $I(X_2;X_3|X_1)=0$. Then according to Lemma 3, the distributions which corresponding to semimatroid $\{(1,2|3)\ \&\ (1,3|2)\ \&\ (2,3|1)\}$ must obey (5.11) defined for each of the three $(i,j|\mathcal{K})$ couples. In $\boldsymbol{\theta}$ coordinate, the mapping can be illustrated by the plot on the left of Fig. 5.5, in which we only consider three random variables X_1, X_2 and X_3. Since each of three set of distributions is e-autoparallel, any their intersections are also e-autoparallel. On the right of Fig. 5.5 we plot the relationship among the Shannon faces corresponding to the three *p-representable* semimatroids $\{(1,3|2)\}$, $\{(1,3|2)\ \&\ (2,3|1)\}$ and $\{(1,2|3)\ \&\ (1,3|2)\ \&\ (2,3|1)\}$ in the gap G_4^{34} of Γ_4. As shown in Fig. 5.5, the polyhedral cone constructed from $\boldsymbol{r}_1^3$, $\boldsymbol{r}_1^{14}$, $\boldsymbol{r}_1^{13}$, $\boldsymbol{r}_1^{23}$, $\boldsymbol{r}_1^{123}$, $\boldsymbol{r}_1^{124}$, $\boldsymbol{r}_1^{134}$, $\boldsymbol{r}_1^{234}$, $\boldsymbol{r}_1^{\emptyset}$, $\boldsymbol{r}_1^4$, $\boldsymbol{r}_2^1$, $\boldsymbol{r}_2^2$, $\boldsymbol{r}_3^{\emptyset}$ and $\boldsymbol{f}_{34}$ corresponding to semimatroids $\{(1,3|2)\}$; removing $\boldsymbol{r}_1^{14}$ from the list, we get semimatroids $\{(1,3|2)\ \&\ (2,3|1)\}$; finally, the intersection of three faces corresponding to the polyhedral cone constructed from $\boldsymbol{r}_1^{13}$, $\boldsymbol{r}_1^{23}$, $\boldsymbol{r}_1^{123}$, $\boldsymbol{r}_1^{124}$, $\boldsymbol{r}_1^{134}$, $\boldsymbol{r}_1^{234}$, $\boldsymbol{r}_1^{\emptyset}$, $\boldsymbol{r}_1^4$, $\boldsymbol{r}_2^1$, $\boldsymbol{r}_2^2$, $\boldsymbol{r}_3^{\emptyset}$ and $\boldsymbol{f}_{34}$.

As for the relations between *p-representable* semimatroids and the conditions of $\mathcal{L}$

specified in Theorem 9, by examine the 120 irreducible *p-representable* semimatroids listed in Theorem 1, we get the following Corollary:

Corollary 3: Let relation $\mathcal{L} \subseteq \mathcal{S}(4)$ be a *p-representable* semimatroid such that $\mathcal{L}$ contains at least two $(i,j|\mathcal{K})$ pairs. If $\{i \cup j \cup \mathcal{K}\} = \mathcal{N}' \subseteq \{1,2,3,4\}$ for $\forall (i,j|\mathcal{K}) \in \mathcal{L}$, then $\mathcal{L}$ is an Ingleton semimatroid.

In this sense, the information geometric parametrization presented thus far can only reach those faces of $\bar{\Gamma}_4^*$ shared with the Ingleton inner bound. In part for this reason, in the next section we will study the information geometric structure of certain Ingleton violating distributions.

5.3 Information Geometric Structure of Ingleton-Violating Entropic Vectors & their Distributions

As discussed in §4.2, every k-atom distribution for $k < 4$ is incapable of violating Ingleton, and every k-atom distribution for $k \in \{5,6,7\}$ are incapable of exceeding the Ingleton score achieved by the 4-atom distribution on support $\boldsymbol{\mathcal{X}}_4$ given in (4.5). Here we investigate study the information geometric structure of those distributions violating Ingleton on this special 4-atom support, and show that they admit a nice information geometric. We will also show that larger k-atom supports that are capable of violating Ingleton, but do not yield better Ingleton scores, do not share this special information geometric structure.

Denote by $\mathcal{O}(\boldsymbol{\mathcal{X}}_4)$ the manifold of all probability mass functions for four binary random variables with the support (4.5). Parameterize these distributions with the parameters $\eta_1 = \mathbb{P}(\boldsymbol{X} = 0000) = \alpha$, $\eta_2 = \mathbb{P}(\boldsymbol{X} = 0110) = \beta - \alpha$, $\eta_3 = \mathbb{P}(\boldsymbol{X} = 1010) = \gamma - \alpha$, $\eta_4 = \mathbb{P}(\boldsymbol{X} = 1111) = 1 - \sum_{i=1}^{3} \eta_i = 1 + \alpha - \beta - \gamma$, yielding the m-coordinates of $\mathcal{O}(\boldsymbol{\mathcal{X}}_4)$. Let the associated e-coordinates can be calculated as $\theta_i = \log_2 \frac{\boldsymbol{\eta}_i}{\boldsymbol{\eta}_4}$ for $i = 1,2,3$. Next, we consider the submanifold $\mathcal{D}_{u1} = \{p_x \in \mathcal{O}(\boldsymbol{\mathcal{X}}_4) | I(X_3; X_4) = 0\}$,

by Theorem 7, $\mathcal{O}(\boldsymbol{\mathcal{X}}_4)$ is a e-autoparallel submanifold of $\mathcal{D}_{4atom}$. In fact, an equivalent definition is $\mathcal{D}_{u1} = \{p_x \in \mathcal{O}(\boldsymbol{\mathcal{X}}_4) | -\theta_1 + \theta_2 + \theta_3 = 0\}$. Numerical calculation illustrated in the Figure 5.6 lead to the following proposition, which we have verified numerically.

Proposition 1: Let $\mathcal{D}_{m1} = \{p_x \in \mathcal{O}(\boldsymbol{\mathcal{X}}_4) | Ingleton_{34} = 0\}$, then $\mathcal{D}_{m1}$ is a e-autoparallel submanifold of $\mathcal{O}(\boldsymbol{\mathcal{X}}_4)$ and is parallel with $\mathcal{D}_{u1}$ in e-coordinate. In fact, an equivalent definition of $\mathcal{D}_{m1}$ is $\mathcal{D}_{m1} = \{p_x \in \mathcal{O}(\boldsymbol{\mathcal{X}}_4) | -\theta_1 + \theta_2 + \theta_3 = \log_2(\frac{0.5-\alpha_0}{\alpha_0})^2\}$, where α_0 is the solution of $-\alpha \log_2 \alpha - (1-\alpha)\log_2(1-\alpha) = \frac{1+2\alpha}{2}$ in $0 < \alpha < \frac{1}{2}$.

In fact, using this equivalent definition, we can also determine all the distributions in $\mathcal{O}(\boldsymbol{\mathcal{X}}_4)$ that violate $Ingleton_{34} \geqslant 0$ as the submanifold $\mathcal{D}_{Vio} = \{p_x \in \mathcal{O}(\boldsymbol{\mathcal{X}}_4) | -\theta_1 + \theta_2 + \theta_3 < \log_2(\frac{0.5-\alpha_0}{\alpha_0})^2\}$. Because we are dealing with $\mathcal{O}(\boldsymbol{\mathcal{X}}_4)$, a 3 dimensional manifold, we can use a plot to visualize our results in Fig. 5.6. In Fig. 5.6, besides $\mathcal{D}_{u1}$ and $\mathcal{D}_{m1}$, we also plot the following submanifolds and points:

$$
\begin{aligned}
\mathcal{D}_{q1} &= \{p_x \in \mathcal{O}(\boldsymbol{\mathcal{X}}_4) \mid Ingleton_{34} = 0.1\} \\
\mathcal{D}_{s1} &= \{p_x \in \mathcal{O}(\boldsymbol{\mathcal{X}}_4) \mid Ingleton_{34} = -0.126\} \\
\mathcal{D}_{w1} &= \{p_x \in \mathcal{O}(\boldsymbol{\mathcal{X}}_4) \mid Ingleton_{34} = -0.16\} \\
\mathcal{D}_{n1} &= \{p_x \in \mathcal{O}(\boldsymbol{\mathcal{X}}_4) \mid \beta = \gamma = 0.5\} \\
p_u &= \{p_x \in \mathcal{O}(\boldsymbol{\mathcal{X}}_4) \mid \alpha = 0.25, \beta = \gamma = 0.5\} \\
p_m &= \{p_x \in \mathcal{O}(\boldsymbol{\mathcal{X}}_4) \mid \alpha \approx 0.33, \beta = \gamma = 0.5\}
\end{aligned}
$$

As we can see from Fig. 5.6, $\mathcal{D}_{m1}$ and $\mathcal{D}_{u1}$ are e-autoparallel and parallel to each other. As $Ingleton_{34}$ goes from 0 to negative values, the hyperplane becomes ellipsoid-like, and as $Ingleton_{34}$ becomes smaller and smaller, the ellipsoid-like surface shrinks, finally shrinking to a single point p_m at $Ingleton_{34} \approx -0.1699$, the point associated with the four atom conjecture in [32]. Also for each e-autoparallel submanifold $\mathcal{D}_{\forall e} \subset$

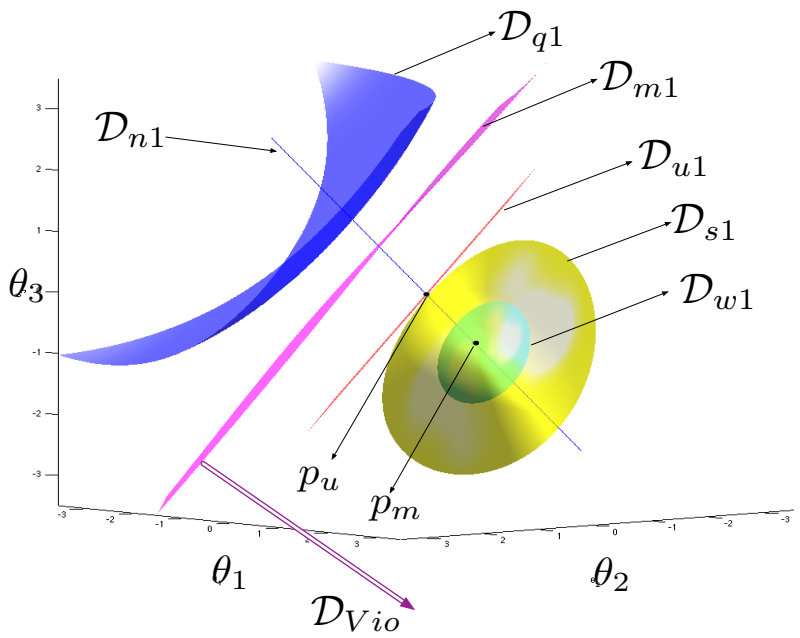

Figure 5.6: Manifold $\mathcal{D}_{4atom}$ in θ coordinate

$\mathcal{D}_{Vio}$ that is parallel to $\mathcal{D}_{m1}$ in e-coordinate, the minimum Ingleton of $\mathcal{D}_{\forall e}$ is achieved at point $\mathcal{D}_{\forall e} \cap \mathcal{D}_{n1}$, where $\mathcal{D}_{n1}$ is the *e-geodesic* in which marginal distribution of X_3 and X_4 are uniform, i.e. $\beta = \gamma = 0.5$.

Next, we map some submanifolds of $\mathcal{O}(\boldsymbol{\mathcal{X}}_4)$ to the entropic region. From Lemma 1, G_4^{34}, one of the six gaps between $\mathcal{I}_4$ and Γ_4 is characterized by extreme rays $\overline{\boldsymbol{V}}_P = (\overline{\boldsymbol{V}}_M, \overline{\boldsymbol{V}}_R, \boldsymbol{r}_1^{\emptyset}, \boldsymbol{f}_{34})$, where we group several extreme rays in to sets: $\overline{\boldsymbol{V}}_M = (\boldsymbol{r}_1^{13}, \boldsymbol{r}_1^{14}, \boldsymbol{r}_1^{23}, \boldsymbol{r}_1^{24}, \boldsymbol{r}_2^{1}, \boldsymbol{r}_2^{2})$ and $\overline{\boldsymbol{V}}_R = (\boldsymbol{r}_1^{123}, \boldsymbol{r}_1^{124}, \boldsymbol{r}_1^{134}, \boldsymbol{r}_1^{234}, \boldsymbol{r}_1^{3}, \boldsymbol{r}_1^{4}, \boldsymbol{r}_3^{\emptyset})$. In addition, we define $\overline{\boldsymbol{V}}_N = (\boldsymbol{r}_1^{1}, \boldsymbol{r}_1^{2}, \boldsymbol{r}_1^{12})$, and use Fig. 5.7 to help visualize G_4^{34}.

In Fig. 5.7, $\boldsymbol{f}_{34}$ is one of the six Ingleton-violating extreme ray of Γ_4, $\overline{\boldsymbol{V}}_M$,$\overline{\boldsymbol{V}}_R$ and $\boldsymbol{r}_1^{\emptyset}$ are all extreme rays of $\mathcal{I}_4$ that make $Ingleton_{34} = 0$. Based on the information geometric characterization, the mapping from $\mathcal{O}(\boldsymbol{\mathcal{X}}_4)$ to Γ_4^* is straight forward: the curve $\tilde{\boldsymbol{h}}_{n1} = \boldsymbol{h}(\mathcal{D}_{n1})$, the straight line $\tilde{\boldsymbol{h}}_{m1} = \boldsymbol{h}(\mathcal{D}_{m1})$, the point $\boldsymbol{h}_u = \boldsymbol{h}(p_u)$ and the point $\boldsymbol{h}_m = \boldsymbol{h}(p_m)$.

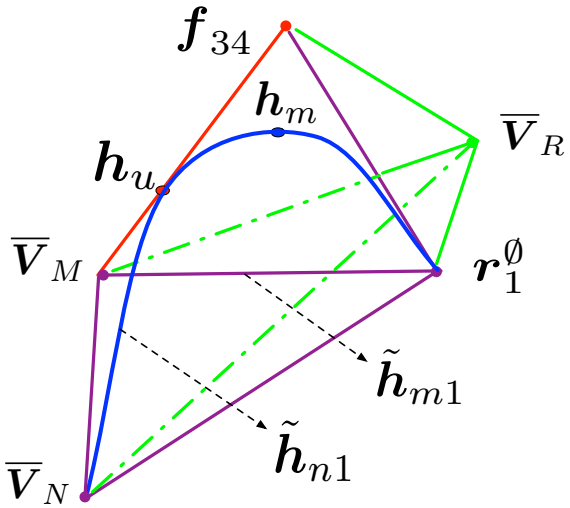

Figure 5.7: G_4^{34}: one of the six gaps between $\mathcal{I}_4$ and Γ_4

Given this nice information geometric property of 4-atom support that Ingleton violation corresponds to a half-space, a natural question to ask is if the e-autoparallel property of $\mathcal{D}_{m1}$ can be extended to five and more atoms. Since we already obtain the list of 29 nonisomorphic 5-atom distribution supports, we know among the 29 supports, only one of them as defined in (4.6) is not a direct extension of 4-atom distributions. For this 5-atom support, if we fix one coordinate θ_0 in e-coordinate, the resulting manifold will be three dimension, thus we can plot the hyperplane of $Ingleton_{34} = 0$ to check if it is e-autoparallel. The result is shown in Figure 5.8. As we can see, the curvature of $Ingleton_{34} = 0$ is non-zero, thus it can not be e-autoparallel. However, the property that distributions of $Ingleton_{34} > 0$ on one side of $Ingleton_{34} = 0$, and distributions of $Ingleton_{34} < 0$ on the other side of $Ingleton_{34} = 0$ still hold.

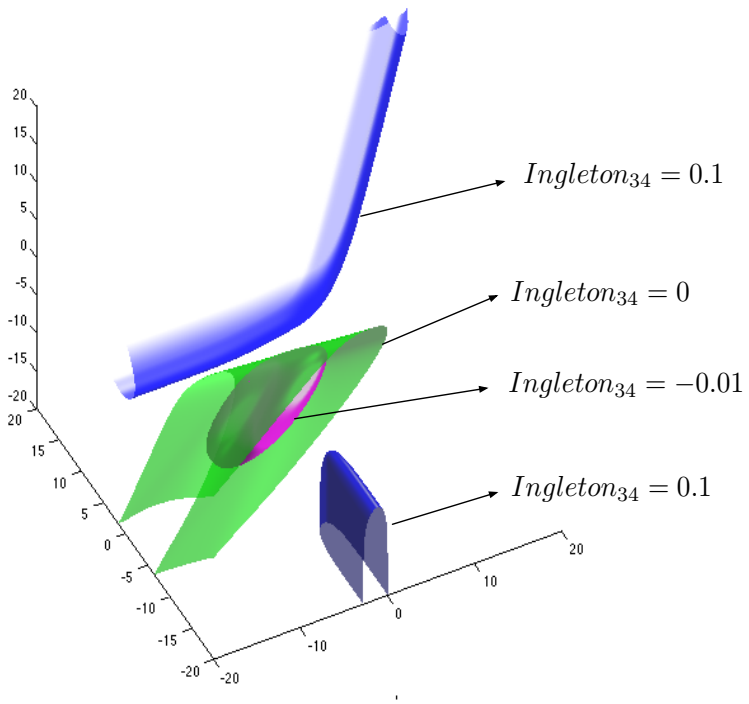

Figure 5.8: Slice of $\mathcal{D}_{5atom}$ in θ coordinate

6. Information Guided Dataset Partitioning for Supervised Learning

In the previous chapters, we have seen how entropy, joint entropy, and conditional independence relations are involved in the characterization of the region of entropic vectors. More specifically, §2.3 reviewed an idealogical framework developed by Matus and Studeny, which studies the consistency among conditional independence relations of discrete variables in terms of *p-representable* semimatroids, then showed how these considerations are equivalent to understanding the relative interior of certain faces of the region of entropic vectors. In the meantime, as we have seen from the information geometric characterization of conditional independence relations and Kullback-Leibler divergence in §5, these information measures are crucial to the understanding of the relations among random variables and measuring the difference between general distributions and those that obey a collection of conditional independence relations.

A major theme in statistical machine learning is the exploitation of these conditional independence structures to reduce the complexity of calculations for inference [40]. Indeed, graphical models [41, 42], Structural Imsets [43], and other probabilistic models all seek to address the problem of simultaneous occurrence of conditional independence among random variables, which we call the problem of *probabilistic representation*. A series of papers by František Matúš and Milan Studený[26, 30, 44] solved the problem of probabilistic representation of four discrete random variables by providing list of irreducible *p-representable* semimatroids(see Theorem 1), which means all configurations of conditional independence that can occur within four discrete random variables are completely characterized. The genesis of the notion of *p-representable* semimatroids was to understand the limitations of graphical models[43, 45, 46, 47] in terms of encoding the conditional independence relations a collection of discrete random variables may have. It was shown in Section 3.6 of [43] that the conditional

independence which can be represented by graphical methods are rather limited when comparing to all possible conditional independence configurations for random variable. In order to address this problem of graphical models, *Structural imsets* [43] was developed and used to learn the structure of graphical models [43, 48]. However, a key issue with graphical models and structural imsets is that some datasets and real probability distributions may not have many of the types of conditional independences that graphical models encode well, but rather may exhibit independence exclusively in certain contexts, for instance, for certain values of some random variables [49]. As we will review later in this chapter, one of the key marginal benefits of the more recent machine learning framework of sum-product networks over traditional graphical models is their ability to capture these context specific independence properties [50].

Bearing this in mind, in this chapter, we are aiming to use information measures like entropy and mutual information to discover context specific independence relationships among discrete random variables useful for supervised learning in certain datasets via a process of *information guided dataset partitioning*. By partition we mean the partition of data sample/instances in the dataset but not the partition of features. The partition selected reflects a series of different contexts in which to train separate classifiers, each of which can exploit independence encountered exclusively within their own context. Such context specific patterns, which will be reflected in the design of our partition score functions, will be used to decide whether or not we should train a generalized supervised learning model on the entire dataset or we could divide the data into parts and treat them as separated supervised learning problem. Partitioning instances in the datasets also enables us to divide the task into individual smaller task and deploy parallel learning algorithms more efficiently. The remainder of this chapter is organized as follows. In §6.1 we will first review some concepts and related works and introduce our work. Then, information-theoretic partition criteria

will be introduced under the framework of sum-product networks in §6.2. Next, some experimental results in §6.3 will demonstrate the utility of the information guided partitioning technique.

6.1 Preliminaries, Related works and Problem setup

Along our line of works in the region of entropic vectors, we have provided the information geometric properties of probability distributions associated with conditional independence relations, where in Lemma 3 $I(\boldsymbol{X}_{\mathcal{A}\setminus\mathcal{B}}; \boldsymbol{X}_{\mathcal{B}\setminus\mathcal{A}}|\boldsymbol{X}_{\mathcal{A}\cap\mathcal{B}}) = 0$ was shown to be equivalent with intersections of a set of hyperplanes (5.11) in $\boldsymbol{\theta}$ coordinates. We have also studied in section 2.3 and section 5.2 the minimal *p-representable* semimatroids containing sets of conditional independence relations that can be simultaneously satisfied by a set of discrete random variables. We learned that *p-representable* semimatroids can be viewed as a concise way to represent relationships among random variables. This lead us to the research of probabilistic models where the goal is to compactly encode joint distributions in a way that allows the model to be constructed and utilized effectively. Among such probabilistic models, graphical models [42] are used to find a graph-based representation of the joint distributions of a set of random variables. Such a representation utilizes conditional independence relations among random variables to help build a graph which can be exploited to achieve faster inference and learning. The graph is called a Bayesian Network (BN) when it is directed acyclic graph (DAG) and called Markov Random Fields (MRF) if it is undirected. A *Bayesian network* consists of a collection of probability distributions P over $\boldsymbol{X} = \{X_1, \cdots, X_K\}$ that *factorize* over a directed acyclic graph (DAG) in the following way:

$$p(\boldsymbol{X}) = p(X_1, \cdots, X_K) = \prod_{k \in K} p(X_k|\mathbf{pa}_k)$$

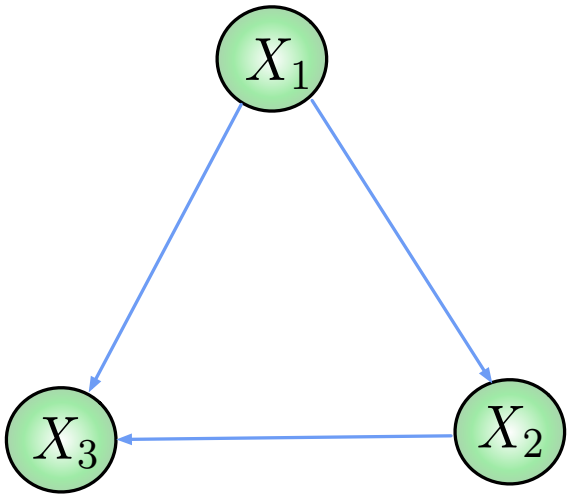

Figure 6.1: Example of a Bayesian Network

where $\mathbf{pa}_k$ is the direct parents nodes of X_k. Consider an arbitrary joint distribution $p(\boldsymbol{X}) = p(X_1, X_2, X_3)$ over three variables, we can write:

$$\begin{aligned} p(X_1, X_2, X_3) &= p(X_3|X_1, X_2)p(X_1, X_2) \\ &= p(X_3|X_1, X_2)p(X_2|X_1)p(X_1) \end{aligned}$$

which can be expressed in the directed graph in Figure 6.1

For the definition of *Markov Random Fields*, denote $\mathcal{C}$ as a clique, $\boldsymbol{X}_\mathcal{C}$ the set of variables in clique $\mathcal{C}$ and $\psi_\mathcal{C}(\boldsymbol{X}_\mathcal{C})$ a nonnegative potential function associated with clique $\mathcal{C}$. Then a Markov random field is a collection of distributions that *factorize* as a product of potential functions $\psi_\mathcal{C}(\boldsymbol{X}_\mathcal{C})$ over the *maximal cliques* of the graph:

$$p(\boldsymbol{X}) = \frac{1}{Z} \prod_\mathcal{C} \psi_\mathcal{C}(\boldsymbol{X}_\mathcal{C})$$

where normalization constant $Z = \sum_{\boldsymbol{X}} \prod_\mathcal{C} \psi_\mathcal{C}(\boldsymbol{X}_\mathcal{C})$ sometimes called the partition function.

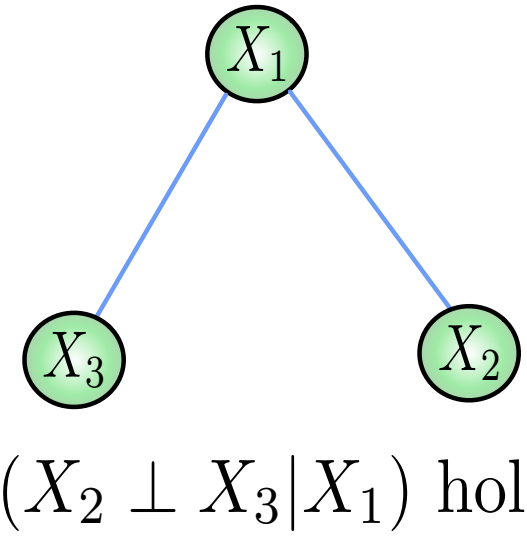

Figure 6.2: Example of a Markov Random Fields

For the undirected graph in Figure 6.2, we can write the joint distribution corresponding to it as

$$p(\boldsymbol{X}) = \frac{1}{Z}\psi_{12}(X_1, X_2)\psi_{13}(X_1, X_3)$$

where $\psi_{12}(X_1, X_2)$ and $\psi_{13}(X_1, X_3)$ are the potential functions and Z is the partition function that make sure $p(\boldsymbol{X})$ satisfy the conditions to be a probability distribution.

Graphical models provide a commonly utilized method to exploit conditional independence relations to reduce the complexity of inference in probabilistic models. However, in lots of real application cases, graphical models prove to be an insufficient language to represent the complex conditional independence relations exhibited by the data. First of all, the set of conditional independence structures that can be described by undirected graphs, directed acyclic graphs, is already a very small subset of the list of all possible conditional independence structures real probability distributions can exhibit (i.e. p-representable semimatroids), even four only 4 random variables [26, 30, 44]. Structural imsets[43] enable extra conditional independences to be encoded beyond traditional graphical models, however, the extra conditional

independences this model encodes cannot be utilized to make inference faster. While structural imsets can be utilized to learn the structure of Bayesian Networks [48], the framework does not provide a method of building *simple* Bayesian Networks, which is key to tractable inference.

A deeper major issue with these existing models, both classical graphical models and structural imsets, is that they can not encode context-specific independence[49]. A context-specific independence is a type of independence that only holds in certain *contexts*, for instance for certain values of a subset of the random variables. To overcome this limitation of graphical models, [49] introduced tree-structured Conditional Probability Tables (CPTs) to capture context-specific independence. Alternatively, [51, 52] used Bayesian multinets, in which we have a network such that the existence of edges depends on values of certain nodes in the graph. A simple example of a Bayesian multinet is a class-conditional naive Bayes classifier. Finally, sum-product networks[50, 53] are a form of recently proposed probabilistic model that can also be viewed as a way to exploit context-specific independence. As we shall see, the construction of a sum-product network from data can be done by partitioning instances and variables, which make them ideal as the framework for the criteria and algorithms we will be proposing.

A sum-product network(SPN) is defined as follows [54]:

- (1) A *tractable univariate distribution* is a SPN.
- (2) A *product* of SPNs with disjoint scopes is a SPN.
- (3) A *weighted sum* of SPNs with the same scope is SPN, provided all weights are positive.
- (4) Nothing else is SPN.

In this definition, a univariate distribution is tractable if and only if its partition function and its mode can be computed in $\mathcal{O}(1)$ time, and the *scope* of a SPN is the set of variables that appear in it.

A SPN can be represented as a tree with univariate distributions as leaves, sums and products as internal nodes, and the edges from a sum node to its children labeled with the corresponding weights. See Figure 6.3 for a sum-product network of three binary variables for which the joint probability distribution can be written as $p(\boldsymbol{X}) = 0.8(1.0x_1 + 0.0\bar{x}_1)(0.3x_2 + 0.7\bar{x}_2)(0.8x_3 + 0.2\bar{x}_3) + 0.2(0.0x_1 + 1.0\bar{x}_1)(0.5x_2 + 0.5\bar{x}_2)(0.9x_3 + 0.1\bar{x}_3)$, where x_i and $\bar{x}_i$ are indicator functions such that $x_i = 1$ and $\bar{x}_i = 0$ when the value of X_i is 1; $x_i = 0$ and $\bar{x}_i = 1$ when the value of X_i is 0.

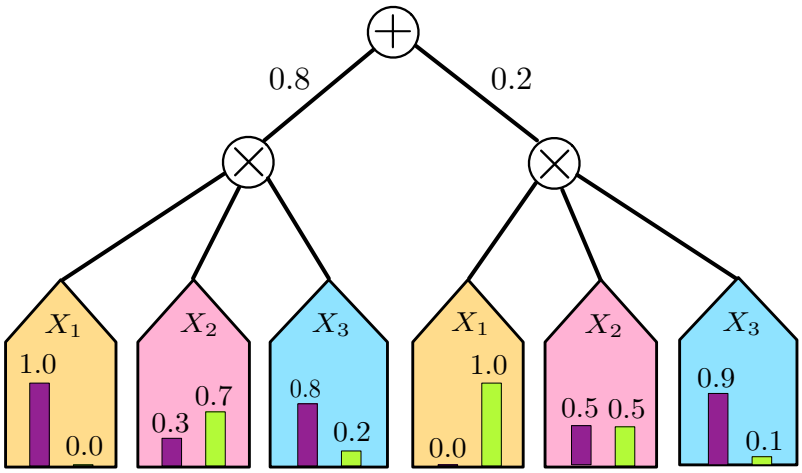

Figure 6.3: Example of a sum-product network

By exploiting context-specific independence, a sum-product network can be quite expressive, while additionally having the merit that inference is always tractable due to the construction of the tree-like structure. However, as mentioned before, in our line of work, we are not using sum-product network for inference or learning, but rather

using its unique structure as the framework for our algorithms which are classifier and model agnostic. As pointed out in [54], the structure of a sum-product network can be learned by recursively partitioning datasets into different mixtures and dividing variables into roughly independent sets. When the goal is to do structure learning for datasets without labels, such partitioning can be done via clustering algorithms, while the variable separation can be done via independence tests. However, in the setting of supervised learning, when the goal is to do supervised learning for each partition of instances, existing algorithms are no longer suitable. This motivates us to propose new partition criteria based on information theoretic measures which can be used in supervised learning. Before moving on to introduce our main results, we will first review some applications of information theoretic measures in machine learning, particularly in partitioning of datasets.

Information theoretic measures [2, 55] have been extensively used in machine learning for the purpose of feature selection, clustering, and classification. In feature selection[56], mutual information and conditional mutual information provide natural tools for measuring the relevance between features and classification labels. Such feature selection algorithms can be quite useful when the number of features are much more than the number of instances in the datasets. We refer the readers who are interested in this field to [57] for a unified framework of information theoretic feature selection algorithms.

In clustering, however, the datasets of interest usually do not have labels, our task is to automatically grouping instances into clusters such that instances in the same cluster are more similar(under some metric) to each others than to instances from other clusters. Information theoretic clustering algorithms[58, 59, 60, 61] provides criteria that are different from the most common geometric framework, which give them a different yet useful perspective.

As an important branch of classification and regression algorithms, decision tree learning algorithms [62, 63, 64, 65, 66] use entropy based criteria as instance selection measures to recursively split datasets. Among all the decision tree learning algorithms, ID3 [62] and C4.5 [62] was created by Ross Quinlan to generate a decision tree which can be used for supervised learning. Information gain was first used in ID3 to select instances for splitting data, then a modified version called Information Gain Ratio was introduced in C4.5 to overcome a drawback of Information gain that prefer instances taking on a large number of distinct values. Today, entropy based instance selection measures are used in the most popular decision tree algorithms like Random Forest [64] and Gradient Boosting Machines [65].

Information theoretic measures are used in clustering and decision tree learning both as criteria for partitioning datasets, more specifically, partitioning instances rather than variables. In clustering, the partition is done in a unsupervised manner, where as in decision tree learning, classification labels of training data are used in the calculation of mutual information such that after the recursive split, most of the instance that falls into a given partition will have the same classification label (and hence belongs to the same class). Here in our work, we also propose to use information theoretic measures to partition datasets, for the purpose that will be discussed shortly after. Our partition scheme will be using score functions to find good splitting variables, where the score functions are based on Information theoretical measures between values of feature variables and classification labels. After finding a good feature-value based partition strategy via calculating score functions on the training data, both the training and test datasets will be split into parts through this partition strategy, and treated separately for classification. Such a scheme can also be done recursively, although many iterations of partitions are likely to damage the classification performance, especially when the dataset is relatively small.

In a decision tree learning algorithm, where a greedy top-down induction is applied, the first few splits of the decision tree are crucial to the performance of the algorithm, usually the best score will be obtained when after splitting data into groups, most groups will only have instances from the same class. Our partition strategy differs from that of decision tree in that the score functions are designed to add an entropy penalty terms on the label variable conditioned on the group so that after the partition, each group will still have enough instances from different classes. Additionally, the metrics we select for the partitions aim to maximize context specific conditional independences, effectively reducing the complexity of the classification task on each of the subproblems associated with each block in the partition. In this way, we can train different parts separately, using the tuned parameters that works best for each individual parts. Such a partition strategy is also strictly differently from clustering, since mutual information between feature variables and label variables will play an important role in the score functions, resulting in a partition strategy very different from clustering. Since our algorithm generates a partition of the data instances, when we are building models with decision tree algorithms, it is similar to adding a hidden variable to the data and using this hidden variable to make the first split of the decision tree. However, such a hidden variable usually can not be directly learned by the various splitting metrics used by decision tree algorithms, as we are selecting it on the basis of maximizing context specific conditional independences, thereby reducing the complexity of each of the classification tasks on each block of the partition. Indeed, the key contribution of this work is that we provide information theoretic measure based score functions which recommend good partition strategies, in the sense of detecting and exploiting context specific conditional independences. If such partition strategies revealing context specific conditional independences can be found, then in essence the complexity of the sub-classification problems on each

block involve fewer relevant variables are simpler. We demonstrate with real datasets that the classification performance can increase even when classifiers are trained independently for different partition blocks.

An additional motivation of this work was that a key issue in machine learning in recent years has been the design of learning architectures which admit scaling to massive datasets, wherein terabytes or even petabytes of data need to be processed. In this setting, a divide and conquer algorithm associated with a dataset partition can be quite useful in that each part can be processed in parallel more efficiently without damaging the performance of classification. Recent works [67] used such divide and conquer methods for kernel support vector machines (SVM), wherein unsupervised kernel k-means was used to partition the data, after which, each partition was conquered separately, and the resulting solutions to these subproblems are used to initialize the global SVM solver so that it converges faster.

Finding a good partition criterion is important and non-trivial in any divide and conquer based method of classification. The remainder of the manuscript will explain how information theoretic measures like entropy and conditional mutual information on the training data can guide the selection of a dataset partitioning that maximizes context specific conditional independences. After partitioning, classifiers can be trained and tested on each partition block independently. Experiments with datasets from online machine learning competitions then show that these information guided partitioning strategies can simultaneously enable both performance improvements and parallel training, presumably because the context specific conditional independences being captured by the partition reduces the complexities of the sub-classification problems.

6.2 Proposed Partition Criteria and Algorithms

Let's first define some notations for this chapter. Again let $\mathcal{N} = \{1, 2, \cdots N\}$, and let D_N^M be a discrete valued training dataset consisting of N variables $X_1, X_2, \cdots X_N$ and M instances $\boldsymbol{d}_1, \boldsymbol{d}_2, \cdots \boldsymbol{d}_M$ where each $\boldsymbol{d}_i$ for $1 \leq i \leq M$ is a vector of N discrete valued (categorical) instances. Since the dataset is used for supervised learning, a separate label vector $\boldsymbol{y}_M$ of length M is also provided for the dataset. The features $X_1, X_2, \cdots X_N$ can be treated as discrete random variables and the label as a random variable Y. For classification purpose, a test dataset $D_N^{M'}$ consisting of the same N variables and M' instances will be provided together with the label vector $\boldsymbol{y}_{M'}$, which can be used exclusively for evaluating classification results. Now we can build a machine learning algorithms from D_N^M and $\boldsymbol{y}_M$, then use the resulting model to make predictions of variable Y on test dataset $D_N^{M'}$. Suppose the predicted probability or labels are $\hat{\boldsymbol{y}}_{M'}$, the performance of the supervised classification algorithm is measured through certain score function/cost function: $Function(\boldsymbol{y}_{M'}, \hat{\boldsymbol{y}}_{M'})$.

In a divide and conquer setting, we want to partition datasets into K groups $\boldsymbol{C} = \{C_1, C_2, \cdots C_K\}$. After partitioning, the original training dataset D_N^M are divided into K parts $D_N^M(C_1), D_N^M(C_2), \cdots D_N^M(C_K)$, original test dataset $D_N^{M'}$ are also divided into K parts $D_N^{M'}(C_1), D_N^{M'}(C_2), \cdots D_N^{M'}(C_K)$. We can view $\boldsymbol{C}$ as a K-partition of the indices $1, 2, \cdots, m$ for D_N^m and $1, 2, \cdots, m'$ for $D_N^{m'}$, and the partition itself can be viewed as a hidden variable C taking values from $\{c_1, c_2, \cdots c_K\}$. In the rest of this chapter, we will use variable C taking value of c_i to denote partition $C_i \in \boldsymbol{C}$. Here for simplicity, suppose we are making hard decisions when partitioning datasets, which means instances are assigned solely to a single group. Then we can do classification on K training-test dataset pairs $(D_N^M(c_1), D_N^{M'}(c_1))$, $(D_N^M(c_2), D_N^{M'}(c_2)), \cdots$ $(D_N^M(c_K), D_N^{M'}(c_K))$ separately. The prediction results on test data $D_N^{M'}(c_i)$ for $i \in \{1, 2, \cdots K\}$ of different parts can be simply concatenated to-

gether or as features stacked into the original datasets.

One of the concerns with such a scheme is whether or not the classification performance on the divided datasets is comparable to the case when we do not partition the data. Another question is that, since different partitions are likely to give different classification performances, if there are so many ways to do the split, it is unrealistic to test performance of every splitting strategy in large-scaled problems, how do we know which variables to use to do the partition? In an attempt to answer the first question, we analyze the problem in the framework of sum-product networks. The second question will be answered by proposing several good score functions to recommend variables for data partitioning.

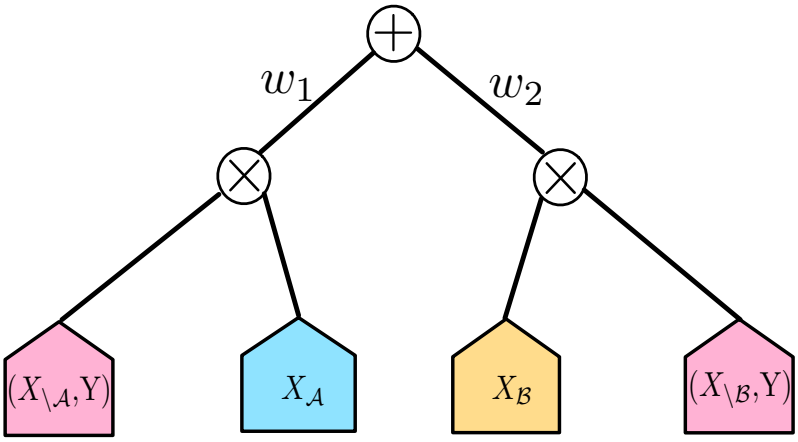

Figure 6.4: Context-Specific Independence of a dataset in Sum-Product Network

Sum-Product networks are a probabilistic model that is mainly used for reducing complexity in inference, however, the structure of a Sum-Product networks infers that mixtures in different branches of a sum node represent different context, and

variables in different branches of a product node are independent. The ability to explicitly encode context-specific independence is one of the major advantages of sum-product networks. Here in our partition scheme, we are only using the structure of sum-product networks to illustrate that different context specific independences can be exploited in different part of data, but not using it to do inference or learning. Since the partitions of dataset D_N^M or $D_N^{M'}$ can be treated as branches of a sum node in a sum-product networks, for each branch, we can exploit independence that are unique to that partition. For example, suppose we have a categorical dataset $(D_N^M, \boldsymbol{y}_M)$ with binary labels, that is to say random variable Y associated with label vector $\boldsymbol{y}_M$ only takes values of 0 or 1. After implementing a 2-partition(K = 2) on the datasets, suppose there exist $\mathcal{A} \subset \mathcal{N}$ and $\mathcal{B} \subset \mathcal{N}$ such that the set of feature variables $X_\mathcal{A}$ are no longer useful for the classification of Y for group $C = c_1$, meanwhile, the set of feature variables $X_\mathcal{B}$ are no longer useful for the classification of Y for group $C = c_2$. That is to say we can ignore $X_\mathcal{A}$ or $X_\mathcal{B}$ under the context of $C = c_1$ and $C = c_2$ separately. However, since $X_\mathcal{B}$ can still be useful for group c_1 and $X_\mathcal{A}$ for group c_2, it might not be a good idea to completely discard $X_{\mathcal{A}\cup\mathcal{B}}$. The example is illustrated in Figure 6.4, where w_i and w_2 denote the ratio of number of instances in each group which means $p(C = c_1) = w_1$, $p(C = c_2) = w_2$ and $w_1 + w_2 = 1$.

After such a partition, each groups can have different distribution for Y, corresponding to w_{11}, w_{12}, w_{21}, w_{22} in Figure 6.5, and have different relations between feature variables X and label variable Y, thus training them separately will be a good alternative.

For our work, we are especially interested in the Context-Specific conditional mutual information $I(X, Y|C = c_i)$ for group c_i. Our goal is to find a partition strategy over the space of all possible partitions, which is a very large set. However, since we are dealing with discrete valued features in the datasets, we can work with

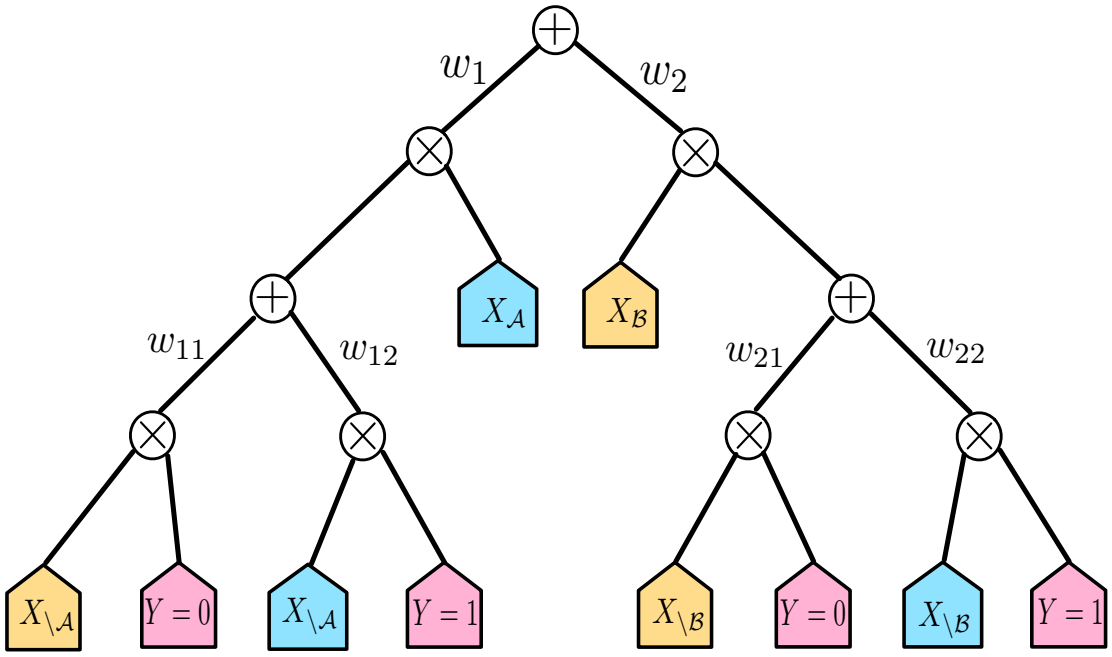

Figure 6.5: Partition of a dataset viewed as Sum-Product Network

value-based partitions that depend on certain values of feature variables. It is also a natural idea to use the most frequent values of feature variables or use several values of a feature variable that have ratios larger than a predefined threshold. In this way, we can avoid lots of extreme cases where the calculation of information measures become less informative. The problem is then formulated as a discrete optimization problem where we minimize a score function $S(C)$ over candidate partition strategies C:

$$\min_{C} S(C) \tag{6.1}$$

The first score function we are proposing is shown in (6.2)

$$S_0(C) = \min_{i \in \mathcal{N}} \frac{I(X_i; Y|C)}{H(Y|C) + 0.1} - H(C) \tag{6.2}$$

where we minimize conditional mutual information $I(X_i; Y|C)$, at the same time maximizing $H(Y|C)$ so that we will not be in a situation where most of the partitions only have instances from a single class. A constant 0.1 was in the denominator to make sure we are not dividing zero. A regularization term $H(C)$ was added so the score function will prefer simple and more evenly partitions.

In order to fully exploit context-specific independence, we can build a cost function where we decompose conditional mutual information into context-specific conditional mutual information, which result in the score function in (6.3)

$$S_1(C) = \sum_{\forall c_j} \min_{i \in \mathcal{N}} \frac{I(X_i; Y|C = c_j)}{H(Y|C = c_j) + 0.1} - H(C) \tag{6.3}$$

In both $S_0(C)$ and $S_1(C)$, we are trying to find variables that are irrelevant with the label Y under the partition C. However, some feature variables can have little or no use for classifying Y even in the entire original datasets, $S_2(C)$ in (6.4) takes this into account so we only measuring the *explaining away residual(EAR)* [[52]]:

$$S_2(C) = \sum_{\forall c_j} \min_{i \in \mathcal{N}} \frac{I(X_i; Y|C = c_j) - I(X_i; Y)}{H(Y|C = c_j) + 0.1} - H(C) \tag{6.4}$$

Furthermore, we can calculate the sum of the k smallest context-specific conditional mutual informations, where it is natural to select k based on the total number of features in the datasets. Such score functions will not decide the best partitions based on a single feature but on a group of features that give small context-specific conditional mutual information. The minimum k version of $S_1(C)$ and $S_2(C)$ are given in $S_3(C, k)$ and $S_4(C, k)$ respectively.

$$S_3(C, k) = \sum_{\forall c_j} \Big(\sum_{smallest\ k} \frac{I(X_i; Y|C = c_j)}{H(Y|C = c_j) + 0.1} \Big) - H(C) \tag{6.5}$$

$$S_4(C,k) = \sum_{\forall c_j} \Big(\sum_{smallest\ k} \frac{I(X_i;Y|C=c_j) - I(X_i;Y)}{H(Y|C=c_j)+0.1} \Big) - H(C) \tag{6.6}$$

Since our goal is to measure the difference among different partitions, if only 2-partitions is allowed, we can use root-mean-square error (RMSE) to measure the difference of context-specific conditional mutual information on all feature variables. Such a score function is shown in (6.7)

$$S_5(C) = -\sqrt{\frac{\sum_{i\in\mathcal{N}}[I(X_i;Y|C=c_1) - I(X_i;Y|C=c_2)]^2}{N}} - H(C) \tag{6.7}$$

Since $K = 2$, partition variable C can only take value from $\{c_1, c_2\}$.

Many of the optimization based partition algorithms apply greedy sequential algorithms to sequentially reassign data instances into different partitions in order to minimize the score function. For a value-based partition, when there are lots of frequent values for a certain feature variable X_i, we can apply the same sequential algorithm to move around small groups of instances having same value $X_i = x_j$. Such an iterative algorithm will run until it converges to a local optimum.

In the experiments we will show in the next section, however, we find good partitions (in terms of classification performance) were usually found through frequent feature-value pairs. So we simply tried all candidate partitions generated from high frequency feature-value pairs and compared the classification performance on them.

6.3 Experimental Results and Discussions

In this section we describe experiments employing our divide and conquer algorithms on an online advertising click-through rate prediction dataset and a telecom company customer behavior prediction dataset.

6.3.1 Click-Through Rate(CTR) prediction

We conducted our first evaluation of information guided dataset partitioning on the Avazu click-through rate competition from Kaggle.[1] Click-through rate is the ratio of users who click an advertisement (usually a link to an advertisement) among all the users who viewed the same page. It is an important measure in the multi-billion dollar online advertising industry. The idea is that if we can more accurately predict the click-through rate, then we can do better at recommending advertisements to users, which will yield better click-through rate and make more profit. Multiple algorithms have been designed specifically for predicting click-through rate, which work better than traditional of-the-shell machine learning algorithms. One of these algorithms, which we are using in the experiments, is an online learning algorithm called Follow-the-Regularized-Leader-Proximal(FTRL-Proximal) algorithm [68, 69]. This algorithm belongs to the class of online gradient descent, and thus has fast running time, which is crucial when the dataset is huge. Another algorithm we are using is Field-aware Factorization Machines [70, 71]. Factorization machines can be seen as a generalization of a linear model where factorized parametrization are used and is able to do reliable parameter estimation when the dataset is very sparse. Since online advertising data usually has lots of missing values, factorization machine tend to perform very well on it.

An import aspect to measure the estimation accuracy of CTR is the selected evaluation metric. In most cases, the performance of CTR are estimated by Logarithmic loss (logloss), for which a smaller value is generally better.

The dataset on which we are conducting the experiments consists of 22 feature variables. The training data has 40 million instances while the test data has 4 million instances. We determined every feature-value pair that has a counting ratio of

[1]We would like to gratefully acknowledge Avazu Inc. for permitting us to use the dataset and Kaggle Inc. for hosting the competition.

more than 20 % in all training and test instances and tested the divide and conquer classification scheme on them.

Table 6.1: Frequent feature-value pairs of CTR data

Partition ID	feature-value	ratio
1	`app_id-ecad2386`	0.6389
2	`device_id-a99f214a`	0.8251
3	`site_category-50e219e0`	0.4090
4	`C15-320`	0.9327
5	`C18-0`	0.4189
6	`C19-35`	0.3010

These feature-value pairs are shown in Table 6.1. In the table we give each splitting feature-value pair a Partition ID which we will be using to represent that partition strategy in later experiments. In the table, all six partition strategies are 2-partitions, so `app_id-ecad2386` means we are splitting training and test dataset based on whether or not the value of feature `app_id` is `ecad2386`.

We calculated score functions $S_1(C)$, $S_2(C)$, $S_3(C, 3)$, $S_4(C, 3)$ and $S_5(C)$ for all six partition strategies, resulting in the metrics that are shown in Table 6.2

Next, we use FTRL-Proximal and Factorization Machines to test the performance of the partitions, here we simply concatenate the result of each individual partitions to test the logloss score on the entire test datasets. The results are summarized in Table 6.3.

In the last row of the table, we also provide the logloss score on the original datasets for comparison. In order to give a fair comparison, all the experiments on FTRL-Proximal and Factorization Machines use the same set of parameters, which

Table 6.2: Scores of different partitions for CTR data

Partition ID	$S_1(C)$	$S_2(C)$	$S_3(C,3)$	$S_4(C,3)$	$S_5(C)$
1	-0.9433	-2.6590	-0.9433	-6.0905	-1.2754
2	-0.1422	-1.7900	2.2659	-2.26773	-0.8324
3	-0.2060	-1.7175	1.1888	-3.3455	-1.1560
4	0.3215	-0.9478	2.8434	-0.9644	-0.6312
5	-0.2389	-1.9567	2.4944	-2.6588	-1.0977
6	-0.0886	-1.5523	2.5760	-1.8150	-0.9588

Table 6.3: Classification results of different partitions for CTR data

Partition ID	FTRL-Proximal	Factorization Machines
1	0.3906360	0.3876771
2	0.3943988	0.3884056
3	0.3931437	0.3887977
4	0.3945613	0.3895983
5	0.3951105	0.3896538
6	0.3943222	0.3892699
original	0.3925731	0.3887780

was tuned for the original datasets by cross-validation. Since the metric is logloss, smaller value means better model. As we can see, Partition 1 gives better results in both FTRL-Proximal algorithm and Factorization Machines when comparing to other partition strategy and compare to not splitting data, while Partition 1 also produce better scores than other partition strategies in the tested score functions in Table 6.2. Splitting datasets based on the partition strategy recommended by our score function, we can not only save running time by parallel training process for each partition, but also get good classification results. We can also see the advantage of Factorization Machines over FTRL-Proximal algorithm, which due to the fact that

Factorization Machines is good at capture complicated interactions among feature variables. We will need to manually create lots of useful feature interactions or stack in results from other classifiers to get comparable results for FTRL-Proximal.

6.3.2 Customer relationship prediction

Our partition strategy was also tested on the KDD Cup 2009 customer relationship prediction dataset.[2] The task is to estimate the probability of users to buy new products/services or buy upgrades or add-ons that were proposed to them. The prediction result can be used to analyze customers' buying preferences, helping the associated company offer better customer service and drive sales growth. In most application cases, user data are flowing into the database very quickly, hence customer relationship prediction models should be updated very frequently, and thus time efficiency is a crucial point. Our divide and conquer scheme has the advantage of saving both training and prediction time when we parallel the model building process of each partition. So if the performance of the partition model is comparable to the original model, it should be still preferred to split the data.

This dataset contains 50000 examples, one large dataset with 14740 features and one small dataset with 230 features are provided. However, since some of the links to the chunks of large dataset were broken, we will use the small dataset as test case. Since there are some important features only presented in the large dataset, the classification results on the small dataset will be slightly worse. There are 190 numerical variables and 40 categorical variables in the small dataset, we choose to predict the upselling probability of users, which means to predict whether or not a user will purchase more expensive items or upgrade their service to a better version. Performance of the model are evaluated by Area Under the ROC Curve(AUC) where only the

[2]We would like to gratefully acknowledge the organizers of KDD Cup 2009 as well as Orange Inc. for making the datasets available.

ranking of the data instance in the test data matters. For this particular problem, Gradient Boosting Machines [65] shown supreme performance, thereby we will use gradient boosting decision tree(GBDT) algorithm from the open-source project xgboost [72] as the testing machine learning algorithm. Since we are comparing the performance of different split of the data with performance of the original data, we decide not to do any feature engineering on the data, with the exception that we removed feature variables Var209, Var214 and Var230 (both Var209 and Var230 contains only NaN in all data instances, and Var214 is co-linear with Var200). The high frequency feature-value pairs used for partitioning the dataset are shown in Table 6.4.

Table 6.4: Frequent feature-value pairs of customer relationship data

Partition ID	feature-value	ratio
1	`Var200-NaN`	0.5081
2	`Var218-cJvF`	0.5063
3	`Var212-NhsEn4L`	0.5860
4	`Var211-L84s`	0.8059
5	`Var205-09_Q`	0.2314
6	`Var228-F2FyR07IdsN7I`	0.6540
7	`Var193-RO12`	0.7192
8	`Var227-RAYp`	0.7031

Score functions $S_1(C)$, $S_2(C)$, $S_3(C, 10)$, $S_4(C, 10)$ and $S_5(C)$ were calculated for each of the eight partition strategies, Table 6.5 gives you the summary of the score function results. Unlike the scores of previous click-through rate data, scores of different partition ID in the customer relationship data are very close for most partitions, thus we will not likely to see a very strong connection between the score function results and classification results.

Table 6.5: Scores of different partitions for customer relationship data

Partition ID	$S_1(C)$	$S_2(C)$	$S_3(C,10)$	$S_4(C,10)$	$S_5(C)$
1	-0.9998	-1.5853	0.8998	-11.1167	-1.2258
2	-0.9998	-1.1277	1.4646	-7.2676	-1.0927
3	-0.9785	-1.5325	1.3605	-6.4532	-1.0490
4	-0.7098	-0.9634	1.6077	-6.6235	-1.0256
5	-0.9433	-1.3324	1.7303	-3.8762	-0.9828
6	-0.9303	-1.1234	1.0210	-4.7927	-0.9913
7	-0.8564	-1.2235	1.2457	-5.5225	-0.9052
8	-0.8774	-1.1143	1.5317	-4.7792	-0.9329

In the classification step, we are using xgboost with logistic loss function and using AUC as the criterion for early stopping the training process. The same random seed is used for all experiments, key parameters like *max_depth*, *subsample*, *column_samplebytree* and *min_child_weight* of the algorithm were tuned to get the best AUC score for the original dataset, so it is possible to obtain better performance for different partitions if we fine tune parameters for them. Classification results are summarized in Table 6.6, wherein the AUC score of original dataset is listed in the last row. As we can see from Table 6.6, partition strategies with high scores of $S_1(C)$, $S_2(C)$, $S_3(C,10)$, $S_4(C,10)$ and $S_5(C)$ give us comparable classification AUC with the original dataset. However, the correlation between score function results in Table 6.5 and classification results in Table 6.6 are not very strong, partially due to the fact that the scores for most partitions are very close.

Since at the end of the experiment, we get prediction results from the original dataset and all different partition strategies, we ensemble the results from top-4 partitions and get a AUC of 0.8721. This result demonstrates that for small sized dataset, our partition strategy can not only be used as a distributed scheme for training machine learning models, but also be used to generate a set of partitions to get better

classification performance when ensemble the prediction results. This can be especially useful when a dataset has a huge number of frequent feature-value pairs and one does not have enough time or machine power to train models on all the different partitions.

Table 6.6: Classification results of different partitions for customer relationship data

Partition ID	AUC of GBDT
1	0.87083
2	0.87056
3	0.86630
4	0.86911
5	0.86821
6	0.86748
7	0.86759
8	0.86257
original	0.87014

6.3.3 Discussion of Experimental Results

Comparable or even better classification results were shown in the above two experiments when we partition the dataset according to some feature-value pairs that give better scores in the score functions. While not all datasets necessarily exhibit context specific conditional independences between class variables and features based on partitions with such high frequency feature-value pairs, our score functions were designed to detect and exhibit these patterns in the distributions of the dataset when they are present. As the experiments demonstrate, when they are present, the associated partitioning strategy can lead to comparable or even better classification performance.

The click-through rate dataset we used in the experiments has over 40 millions instances, which was down sampled from a even larger dataset. For such large problems, most nonlinear algorithms suffer from disk space and memory constraints. Compare to traditional distributed machine learning algorithms, the proposed partitioning strategy can overcome problems like communication overheads among cluster nodes in a distributed system, make it possible to run machine learning algorithms more efficiently. Furthermore, as shown in the second experiment, for small sized problems, the information guided partitioning strategy can be utilized to find multiple partition strategies giving comparable scores, subsequently ensembling the different resulting classifiers across different partitioning strategies together can result in further performance improvements.

We wish to close this section by noting that our method has also been tested with satisfactory results on private datasets for insurance quote prediction and a store sales forecasting problem. We were unable to get permission to present results with these datasets, however some characteristics of our partitioning design for them merit general discussion. The store sales forecast problem involved a regression problem with continuous labels. To adapt the information guided partitioning to such a situation with continuous labels, our approach was to calculate a n-bit quantization of the continuous label and use the discrete entropy to approximate the differential entropy. The resulting partitioning scheme was shown to yield both complexity reduction and performance improvement. We expect a similar method estimating conditional mutual information with a continuous class label through quantization to be applicable in other regression oriented datasets.

7. Conclusions

In this thesis, we first reviewed the importance of the region of entropic vectors in applications ranging from multiterminal data compression to network coding to multimedia transmission. The best known bounds for this set, which primarily are based on constructions involving representable matroids and hence linear constructions, were reviewed.

We proved that the problem of characterizing the region of entropic vectors was equivalent to finding a single non-linear inequality solving one of ten interchangeable optimization problems. Additionally, we investigated some symmetry and convexity based properties of the functions that are the solutions to these optimizations problem.

In order to provide an exhaustive search of more complicated constructions, we proposed and solved the problem of listing non-isomorphic distribution supports for the purpose of calculating entropic vectors. This is carried out by fixing k, the number of atoms and N, the number of random variables, so we can grow in k or N to see the progress we make towards the characterization of entropy region. Along the way, a recursive algorithm, Snakes and Ladders, was used to efficiently enumerate the unique supports. The concept of inner bounds based on k-atom distributions was introduced to aid understanding the structure of the entropic vector region. We experimentally generated k-atom inner bounds for $k = 4$, 5, and 6, calculated the volume of these inner bounds, and visualized them via a certain three dimensional projection. A future research direction in this line of work is to study the conditional independence relations specified by the k-atom support that violate Ingleton, and to explore the algebraic structure of k-atom supports.

The next part of the thesis shifted away from numerical distribution searches toward analytical characterization of properties of distributions which enabled them to

be extremal in the sense achieving entropic vectors of living in certain faces of the Shannon outer bound, as well as for violating Ingleton. These analytical characterizations made use of Information geometric parameterizations. It was shown that the set of distributions on the support associated with the yet-best Ingleton score achievable with an (not-almost) entropic vector which violate Ingleton correspond to a half-space in an appropriate coordinate system. This property was shown not be be shared by a larger 5-atom support achieving a strictly poorer Ingleton score.

In the last chapter, we showed how we can use entropy and conditional independence measures to help find good partitions to divide datasets with discrete valued features for the purpose of supervised learning. The selected information theoretic measures were designed to select partitions to maximize context specific conditional independence between a feature and the class label conditioned on the partition block. Several splitting criteria based on this intuition have been proposed and tested on click through rate and customer relationship datasets. Important advantages of partitioning the datasets include distributing the machine learning training tasks to different machines without adding communication overhead, as well as obtaining improved classification results by ensembling together predictions from different partition strategies. The results in the chapter inspire several interesting avenues for further research. A particularly compelling direction is to determine how to enable optimization over partitions built from criteria other than frequent feature-value pairs while still yielding a tractable subsequent optimization problem. Another possible future research direction is to allow soft decisions when assigning instances to partition blocks, so that classification results will not only depend on one partition but is the probabilistic combinations of results from several partitions.

Bibliography

[1] F. Matúš and L. Csirmaz, "Entropy region and convolution," Oct. 2013, arXiv:1310.5957v1.

[2] Raymond W. Yeung, *Information Theory and Network Coding*. Springer, 2008.

[3] R. W. Yeung and Zhen Zhang, "Distributed source coding for satellite communications," *IEEE Trans. on Information Theory*, vol. 45, no. 4, pp. 1111–1120, 1999.

[4] Xijin Yan, Raymond W. Yeung, and Zhen Zhang, "The Capacity Region for Multi-source Multi-sink Network Coding," in *IEEE International Symposium on Information Theory (ISIT)*, Jun. 2007, pp. 116 – 120.

[5] ——, "An Implicit Characterization of the Achievable Rate Region for Acyclic Multisource Multisink Network Coding," *IEEE Trans. on Information Theory*, vol. 58, no. 9, pp. 5625–5639, Sep. 2012.

[6] T. Chan and A. Grant, "Entropy Vectors and Network Codes," in *IEEE International Symposium on Information Theory*, Jun. 2007.

[7] T. Chan and A. Grant, "Dualities between entropy functions and network codes," in *Fourth Workshop on Network Coding, Theory and Applications (NetCod)*, January 2008.

[8] ——, "Mission impossible: computing the network coding capacity region," in *IEEE International Symposium on Information Theory (ISIT)*, July 2008, pp. 320–324.

[9] Zhen Zhang and Raymond W. Yeung, "On Characterization of Entropy Function via Information Inequalities," *IEEE Trans. on Information Theory*, vol. 44, no. 4, Jul. 1998.

[10] K. Makarychev, Y. Makarychev, A. Romashchenko, and N. Vereshchagin, "A new class of non-Shannon-type inequalities for entropies," *Communication in Information and Systems*, vol. 2, no. 2, pp. 147–166, December 2002.

[11] R. Dougherty, C. Freiling, and K. Zeger, "Six new non-Shannon information inequalities," in *IEEE International Symposium on Information Theory (ISIT)*, July 2006, pp. 233–236.

[12] František Matúš, "Infinitely Many Information Inequalities," in *IEEE Int. Symp. Information Theory (ISIT)*, Jun. 2007, pp. 41–44.

[13] T. Kaced, "Equivalence of two proof techniques for non-shannon-type inequalities," in *IEEE International Symposium on Information Theory (ISIT)*, July 2013.

[14] B. Hassibi and S. Shadbakht, "On a Construction of Entropic Vectors Using Lattice-Generated Distributions," in *IEEE International Symposium on Information Theory (ISIT)*, Jun. 2007, pp. 501 – 505.

[15] Weidong Xu, Jia Wang, Jun Sun, "A projection method for derivation of non-Shannon-type information inequalities," in *IEEE International Symposium on Information Theory (ISIT)*, 2008, pp. 2116 – 2120.

[16] J. M. Walsh and S. Weber, "A Recursive Construction of the Set of Binary Entropy Vectors and Related Inner Bounds for the Entropy Region," *IEEE Trans. Inf. Theory*, vol. 57, no. 10, Oct. 2011.

[17] Yunshu Liu and J. M. Walsh, "Bounding the entropic region via information geometry," in *IEEE Information Theory Workshop*, Seville, Spain, Sep. 2013, pp. 577–581.

[18] J. M. Walsh and S. Weber, "A Recursive Construction of the Set of Binary Entropy Vectors," in *Forty-Seventh Annual Allerton Conference on Communication, Control, and Computing*, Sep. 2009, pp. 545–552.

[19] ——, "Relationships Among Bounds for the Region of Entropic Vectors in Four Variables," in *2010 Allerton Conference on Communication, Control, and Computing*, Sep. 2010. [Online]. Available: http://www.ece.drexel.edu/walsh/allerton10.pdf

[20] Congduan Li, John MacLaren Walsh, Steven Weber, "A computational approach for determining rate regions and codes using entropic vector bounds," in *50th Annual Allerton Conference on Communication, Control and Computing*, Oct. 2012. [Online]. Available: http://www.ece.drexel.edu/walsh/Allerton2012LJW.pdf

[21] Babak Hassibi, Sormeh Shadbakht, Matthew Thill, "On Optimal Design of Network Codes," in *Information Theory and Applications*, UCSD, Feb. 2010, presentation.

[22] Martí-Farré and J. Padró, "On Secret Sharing Schemes, Matroids and Polymatroids," in *4th Theory of Cryptography Conference(TCC)*, Feb. 2007, pp. 273–290.

[23] L. Csirmaz, "The size of a share must be large," *Journal of Cryptology*, vol. 10, pp. 223–231, Sep. 1997.

[24] James Oxley, *Matroid Theory, 2nd. Ed.* Oxford University Press, 2011.

[25] D. Hammer, A. Romashschenko, A. Shen, N. Vereshchagin, "Inequalities for Shannon Entropy and Kolmogorov Complexity," *Journal of Computer and System Sciences*, vol. 60, pp. 442–464, 2000.

[26] F. Matúš and M. Studený, "Conditional Independences among Four Random Variables I," *Combinatorics, Probability and Computing*, vol. 4, no. 3, pp. 269–278, Sep. 1995.

[27] Randall Dougherty, Chris Freiling, Kenneth Zeger, "Linear rank inequalities on five or more variables," submitted to SIAM J. Discrete Math. arXiv:0910.0284.

[28] Ryan Kinser, "New Inequalities for Subspace Arrangements," *J. of Comb. Theory Ser. A*, vol. 188, no. 1, pp. 152–161, Jan. 2011.

[29] A. W. Ingleton, "Representation of Matroids," in *Combinatorial Mathematics and its Applications*, D. J. A. Welsh, Ed. San Diego: Academic Press, 1971, pp. 149–167.

[30] F. Matúš, "Conditional Independences among Four Random Variables III: Final Conclusion," *Combinatorics, Probability and Computing*, vol. 8, no. 3, pp. 269–276, May 1999.

[31] T. Chan and R. Yeung, "On a relation between information inequalities and group theory," *IEEE Trans. on Information Theory*, vol. 48, no. 7, pp. 1992 – 1995, Jul. 2002.

[32] Randall Dougherty, Chris Freiling, Kenneth Zeger, "Non-Shannon Information Inequalities in Four Random Variables," Apr. 2011, arXiv:1104.3602v1.

[33] A. Betten, M. Braun, H. Fripertinger, A. Kerber, A. Kohnert, and A. Wassermann, *Error-Correcting Linear Codes: Classification by Isometry and Applications*. Springer, 2006.

[34] D. E. Knuth, *The Art of Computer Programming, Volume 4A*. Addison-Wesley Professional, 2011.

[35] Bernd Schmalz, "t-Designs zu vorgegebener Automorphismengruppe," *Bayreuther Mathematische Schriften*, no. 41, pp. 1–164, 1992, Dissertation, Universität Bayreuth, Bayreuth.

[36] Wei Mao, Matthew Thill, and Babak Hassibi, "On group network codes: Ingleton-bound violations and independent sources," in *IEEE International Symposium on Information Theory (ISIT)*, Jun. 2010.

[37] Nigel Boston and Ting-Ting Nan, "Large Violations of the Ingleton Inequality," in *50th Annual Allerton Conference on Communication, Control, and Computing*, Sep. 2012.

[38] Pirita Paajanen, "Finite p-Groups, Entropy Vectors, and the Ingleton Inequality for Nilpotent Groups," *IEEE Trans. on Information Theory*, vol. 60, no. 7, pp. 3821–3824, Jul. 2014.

[39] S. Amari and H. Nagaoka, *Methods of Information Geometry*. American Mathematical Society Translations of Mathematical Monographs, 2004, vol. 191.

[40] Judea Pearl, *Probabilistic Reasoning in Intelligent Systems: networks of plausible inference*. Morgan Kaufmann Publishers, 1988.

[41] Steffen Lauritzen, *Graphical Models*. Oxford University Press New York, 1996.

[42] Daphne Koller and Nir Friedman, *Probabilistic Graphical Models: principles and techniques*. The MIT Press, 2009.

[43] Milan Studený, *On Probabilistic Conditional Independence Structures*. Springer-Verlag London, 2005.

[44] F. Matúš, "Conditional Independences among Four Random Variables II," *Combinatorics, Probability and Computing*, vol. 4, no. 4, pp. 407–417, Dec. 1995.

[45] Petr Šimeček, "A Short Note on Discrete Representability of Independence Models," *Proccedings of the European Workshop on Probabilistic Graphical Models*, 2006.

[46] Radim Lněnička and František Matúš, "On Gaussian conditional independence structures," *Kybernetika*, vol. 43, 2007.

[47] Remco Bouckaert, Raymond Hemmecke, Silvia Lindner, Milan Studený, "Efficient Algorithms for Conditional Independence Inference," *Journal of Machine Learning Research (JMLR)*, vol. 11, 2010.

[48] Raymond Hemmecke, Silvia Lindner and Milan Studený, "Characteristic imsets for learning Bayesian network structure," *International Journal of Approximate Reasoning*, vol. 53, 2012.

[49] Craig Boutilier, Nir Friedman, Moises Goldszmidt and Daphne Koller, "Context-specific independence in Bayesian networks," *12th Conference on Uncertainty in Artificial Intelligence (UAI)*, 1996.

[50] Robert Gens and Pedro Domingos, "Discriminative Learning of Sum-Product Networks," *25th Advances in Neural Information Processing Systems (NIPS)*, 2012.

[51] Dan Geiger and David Heckerman, "Knowledge representation and inference in similarity networks and Bayesian multinets," *Artificial Intelligence*, vol. 82, 1996.

[52] Jeff A. Blimes, "Dynamic Bayesian Multinets," *16th Conference on Uncertainty in Artificial Intelligence (UAI)*, 2000.

[53] Hoifung Poon and Pedro Domingos, "Sum-Product Networks: A New Deep Architecture," *27th Conference on Uncertainty in Artificial Intelligence (UAI)*, 2011.

[54] Robert Gens and Pedro Domingos, "Learning the Structure of Sum-Product Networks," *30th Internaltional Conference on Machine Learning (ICML)*, 2013.

[55] Thomas M. Cover, Joy A. Thomas, *Elements of Information Theory*. Wiley-Interscience, 2006.

[56] Isabelle Guyon and André Elisseeff, "An Introduction to Variable and Feature Selection," *Journal of Machine Learning Research (JMLR)*, vol. 3, 2003.

[57] Gavin Brown, Adam Pocock, Ming-Jie Zhao and Mikel Luján, "Conditional Likelihood Maximisation: A Unifying Framework for Information Theoretic Feature Selection," *Journal of Machine Learning Research (JMLR)*, vol. 13, 2012.

[58] Hans H. Bock, *Probabilistic Aspects in Cluster Analysis*. Berlin:Springer-verlag, 1989.

[59] Gilles Celeux, Gérard Govaert, "Clustering criteria for discrete data and latent class models," *Journal of Classification*, vol. 8, 1991.

[60] Tao Li, Sheng Ma and Mitsunori Ogihara, "Entropy-Based Criterion in Categorical Clustering," *21th Internaltional Conference on Machine Learning (ICML)*, 2004.

[61] Lev Faivishevsky and Jacob Goldberger, "A Nonparametric Information Theoretic Clustering Algorithm," *27th Internaltional Conference on Machine Learning (ICML)*, 2010.

[62] John R. Quinlan, "Induction of decision trees," *Machine Learning*, vol. 1, no. 1, 1986.

[63] Leo Breiman, Jerome H. Friedman, Charles J. Stone and R.A. Olshen, *Classification and Regression Trees*. New York:Chapman and Hall, 1984.

[64] Leo Breiman, "Random Forests," *Machine Learning*, vol. 45, no. 1, 2001.

[65] Jerome H. Friedman, "Greedy function approximation: A gradient boosting machine," *The Annals of Statistics*, vol. 29, no. 5, 2001.

[66] Carolin Strobl, James Malley, Gerhard Tutz, "An Introduction to Recursive Partitioning: Rationale, Application and Characteristics of Classification and Regression Trees, Bagging and Random Forests," *Psychol Methods*, vol. 14, no. 5, 2009.

[67] Cho-Jui Hsieh, Si Si and Inderjit S. Dhillon, "A Divid-and-Conquer Solver for Kernel Support Vector Machines," *31th Internaltional Conference on Machine Learning (ICML)*, 2014.

[68] Brendan H. McMahan, "Follow-the-Regularized-Leader and Mirror Descent: Equivalence Theorems and L1 Regularization," *14th International Conference on Artificial Intelligence and Statistics (AISTATS)*, 2011.

[69] Brendan H. McMahan etc., "Ad Click Prediction: a View from the Trenches," *19th ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD)*, 2013.

[70] Steffen Rendle, "Factorization Machines," *IEEE 10th International Conference on Data Mining (ICDM)*, 2010.

[71] YuChin Juan, Wei-Sheng Chin and Yong Zhuang, "LIBFFM: A Library for Field-aware Factorization Machines." [Online]. Available: https://www.csie.ntu.edu.tw/~cjlin/libffm

[72] Tianqi Chen etc., "XGBoost eXtreme Gradient Boosting." [Online]. Available: https://github.com/dmlc/xgboost