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

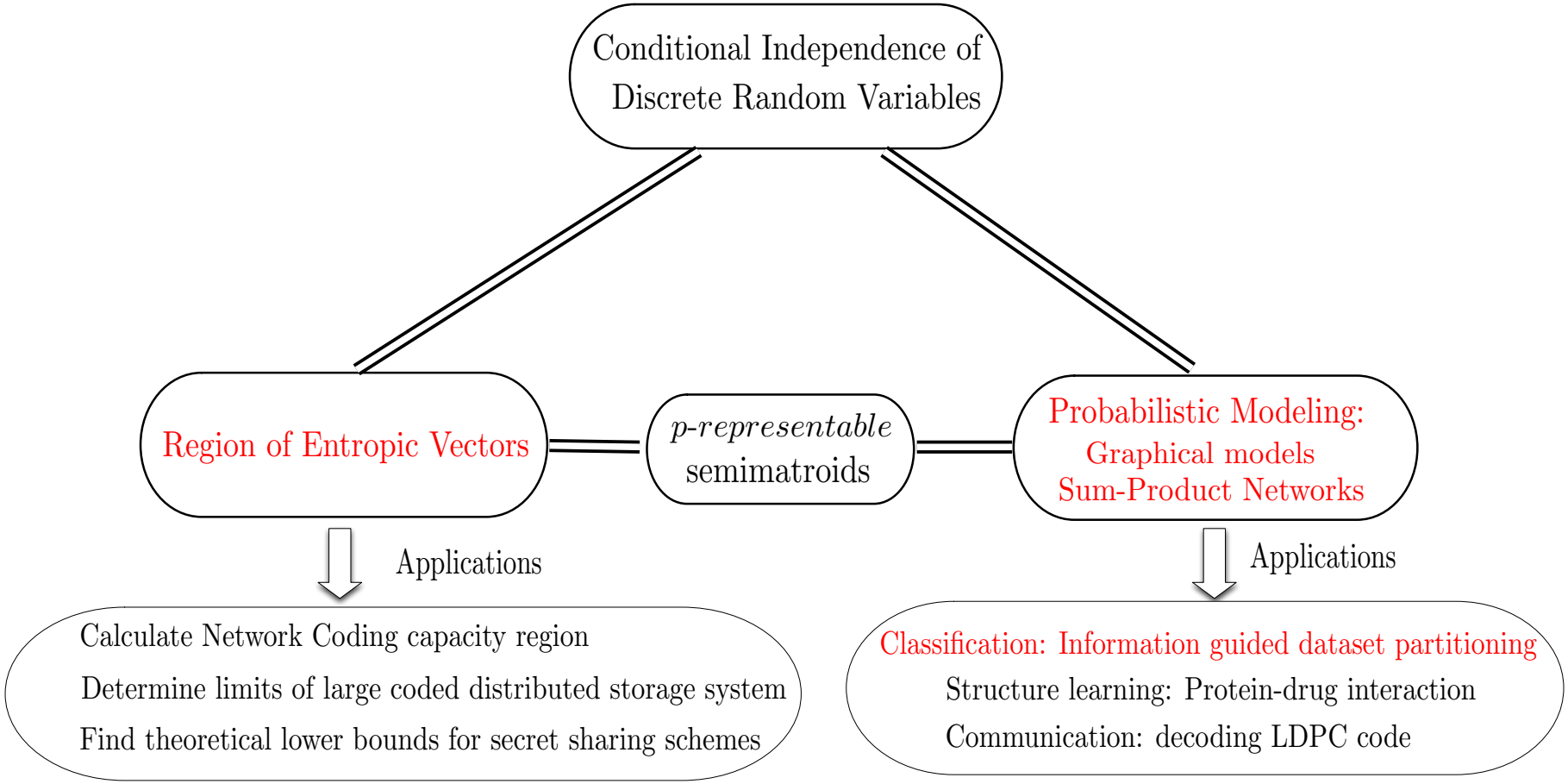
– Ph.D. Dissertation Defense

Yunshu Liu

Adaptive Signal Processing and Information Theory Research Group
ECE Department, Drexel University, Philadelphia, PA

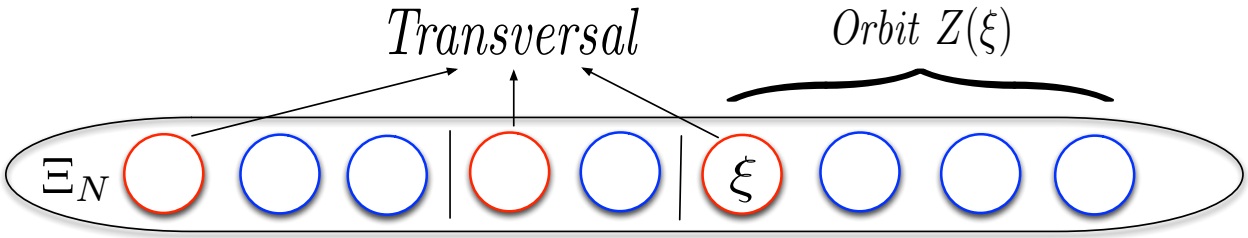
2016-04-06

Motivation

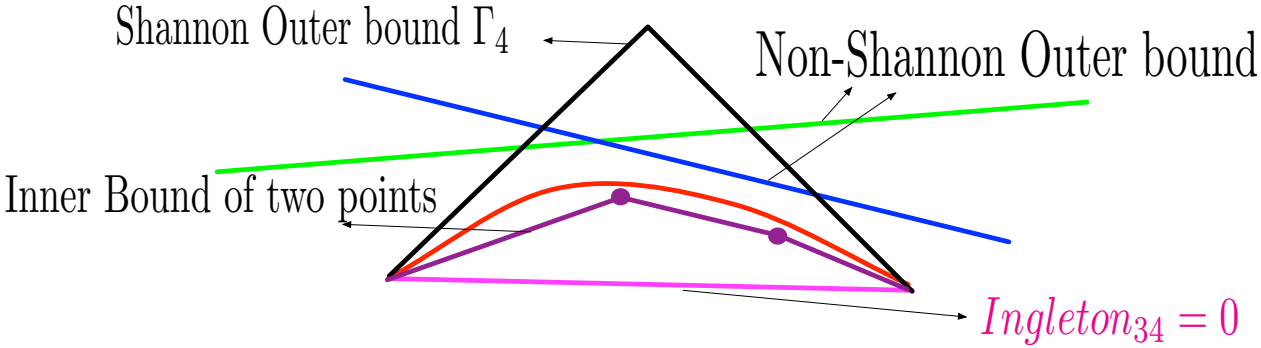

Conditional Independence of
Discrete Random Variables
Region of Entropic Vectors
p-representable
semimatroids
Probabilistic Modeling:
Graphical models
Sum-Product Networks
Applications
Applications
Calculate Network Coding capacity region
Determine limits of large coded distributed storage system
Find theoretical lower bounds for secret sharing schemes
Classification: Information guided dataset partitioning
Structure learning: Protein-drug interaction
Communication: decoding LDPC code

Main Contribution 1: Enumerating k-atom supports and map to entropic region

- Introducing the concept of non-isomorphic distribution support enumeration for entropy vectors

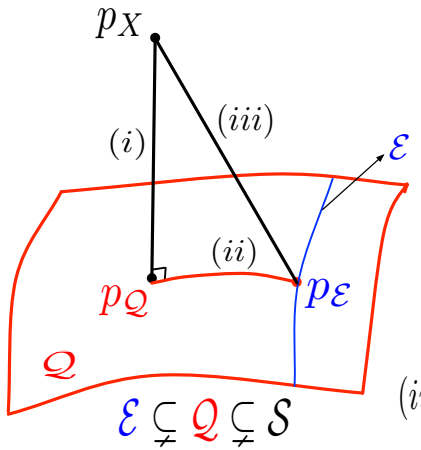


- Generating list of non-isomorphic k-atom, N-variable supports through the algorithm of snakes and ladders
- Optimizing entropy inner bounds from k-atom distribution



Main Contribution 2: Characterization of entropic region via Information Geometry

- Characterizing information inequalities with Information Geometry



Pythagorean relation:

$$D(p_X \| p_{\mathcal{E}}) = D(p_X \| p_{\mathcal{Q}}) + D(p_{\mathcal{Q}} \| p_{\mathcal{E}})$$

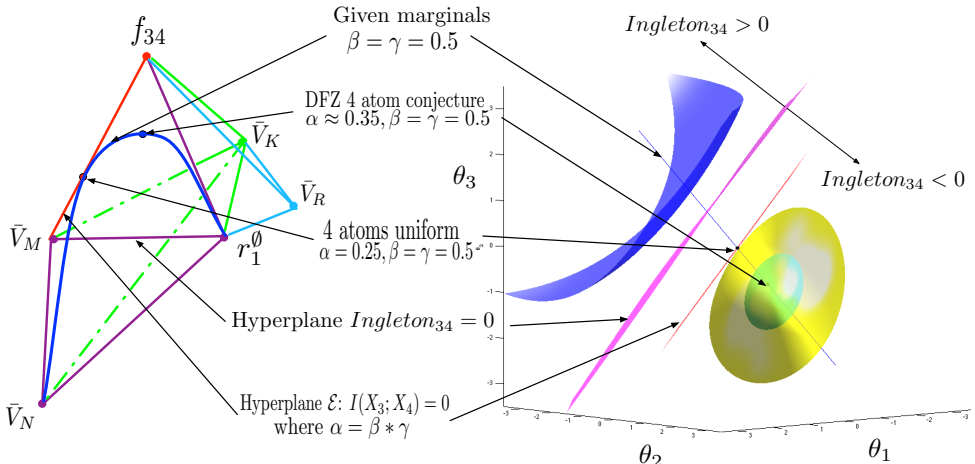
Markov chain in B:

$$\boldsymbol{X}_{\mathcal{A}\setminus\mathcal{B}} \leftrightarrow \boldsymbol{X}_{\mathcal{A}\cap\mathcal{B}} \leftrightarrow \boldsymbol{X}_{\mathcal{B}\setminus\mathcal{A}}$$

Entropy Submodularity:

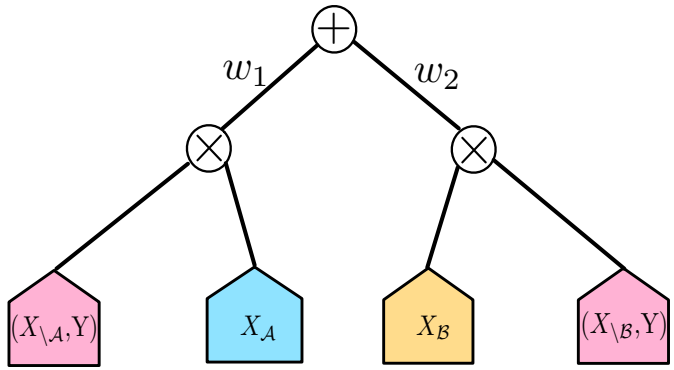
$$(ii)\ D(p_{\mathcal{Q}} \| p_{\mathcal{E}}) = h_{\mathcal{A}} + h_{\mathcal{B}} - h_{\mathcal{A}\cup\mathcal{B}} - h_{\mathcal{A}\cap\mathcal{B}} \geq 0$$

- Mapping k-atoms support distribution with Information Geometry

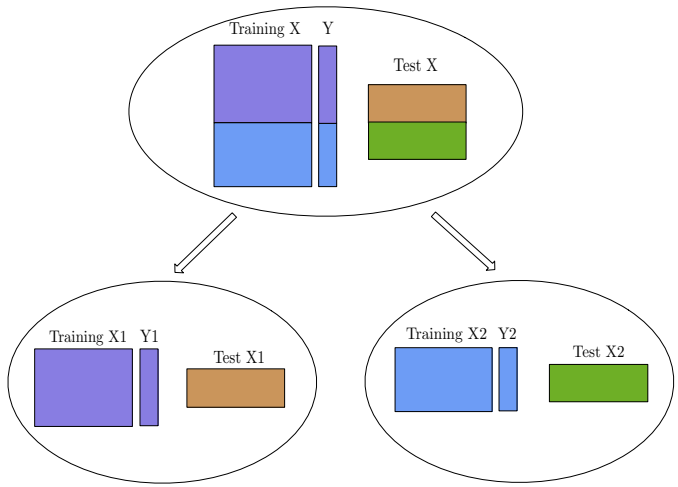


Main Contribution 3: Information guided dataset partitioning

- Exploit context-specific independence for supervised learning



- Design information theoretic score functions to partition dataset



Outline

- The Region of Entropic Vectors $\overline{\Gamma^*_N}$
- Enumerating k-atom supports and map to Entropic Region
  - Non-isomorphic k-atom supports
  - Maximal Ingleton Violation and the Four Atom Conjecture
  - Optimizing Entropy Inner Bounds from k-atom Distribution
- Characterization of Entropic Region via Information Geometry
  - Manifold of Probability distributions
  - e-autoparallel submanifold and m-autoparallel submanifold
  - Projections and Pythagorean Theorem
  - Information Geometry of k-atom support
- From Conditional Independence to Probabilistic Models
  - *p-representable* semimatroids
  - Probabilistic Models
  - Information guided dataset partitioning for supervised learning
  - Experiments on Information guided dataset partitioning

Outline

The Region of Entropic Vectors $\overline{\Gamma^*_N}$

Enumerating k-atom supports and map to Entropic Region

- Non-isomorphic k-atom supports
- Maximal Ingleton Violation and the Four Atom Conjecture
- Optimizing Entropy Inner Bounds from k-atom Distribution

Characterization of Entropic Region via Information Geometry

- Manifold of Probability distributions
- e-autoparallel submanifold and m-autoparallel submanifold
- Projections and Pythagorean Theorem
- Information Geometry of k-atom support

From Conditional Independence to Probabilistic Models

- *p-representable* semimatroids
- Probabilistic Models
- Information guided dataset partitioning for supervised learning
- Experiments on Information guided dataset partitioning

The Region of Entropic Vectors $\overline{\Gamma_N^*}$

Joint entropy viewed as a vector

For N discrete random variables, there is a $2^N - 1$ dimensional vector $\boldsymbol{h} = (h_{\mathcal{A}} | \mathcal{A} \subseteq \mathcal{N}) \in \mathbb{R}^{2^N-1}$ associated with it, we call it *entropic vectors*.

For example:

N = 2: $\boldsymbol{h} = \{h_1 \quad h_2 \quad h_{12}\}$

N = 3: $\boldsymbol{h} = \{h_1 \quad h_2 \quad h_{12} \quad h_3 \quad h_{13} \quad h_{23} \quad h_{123}\}$

The Region of Entropic vectors: all valid entropic vectors

$\Gamma_N^* = \{\boldsymbol{h} \in \mathbb{R}^{2^N-1} | \boldsymbol{h} \text{ is } \textit{entropic}\}$

Examples of two random variables Γ_2^*(Polyhedral cone):

$h_1 + h_2 \geqslant h_{12}$, $h_{12} \geqslant h_1$ and $h_{12} \geqslant h_2$

The Region of Entropic Vectors $\overline{\Gamma_N^*}$

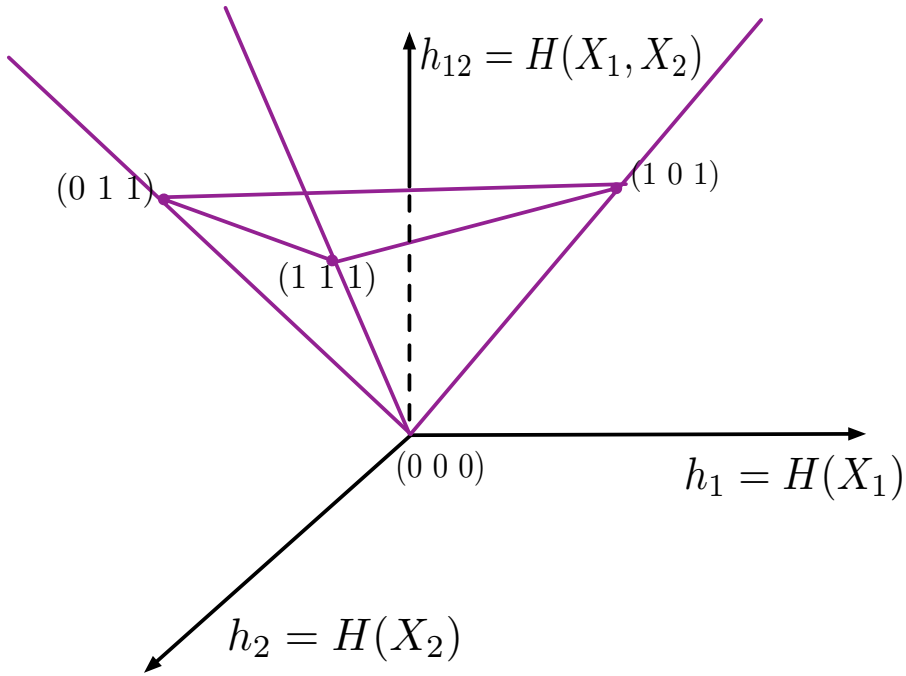
Shannon-Type Information inequality

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

Relationship between Γ_n^* and Γ_n

$\Gamma_2^* = \Gamma_2$ and $\overline{\Gamma_3^*} = \Gamma_3$:
All the Information inequalities on $N \leqslant 3$ variables are Shannon-Type.

For N = 2, Use $h_1 + h_2 \geqslant h_{12}$, $h_{12} \geqslant h_1$ and $h_{12} \geqslant h_2$, we get Γ_2^*

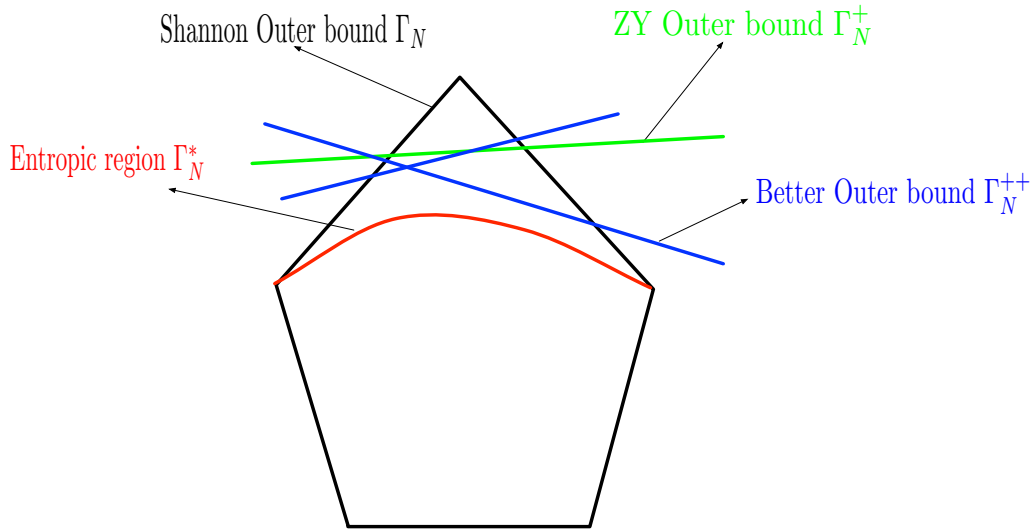


The Region of Entropic Vectors $\overline{\Gamma_N^*}$

- Γ_2^* and $\overline{\Gamma_3^*}$ are polyhedral cone and fully characterized.
- $\overline{\Gamma_4^*}$ is a non-polyhedral but convex cone, we use outer bound and inner bound to approximate it.

Outer bound for $\overline{\Gamma_N^*}$: non-Shannon type Inequality

For $N \geq 4$ we have $\Gamma_N \neq \overline{\Gamma_N^*}$, indicating non-Shannon type Information inequalities exist for $N \geqslant 4$



First Non-Shannon-Type Information inequality(Zhang-Yeung[1])

$$2I(X_3;X_4) \leq I(X_1;X_2)+I(X_1;X_3,X_4)+3I(X_3;X_4|X_1)+I(X_3;X_4|X_2)$$

[1] **Zhen Zhang and Raymond W. Yeung,**
On Characterization of Entropy Function via Information Inequalities,
IEEE Trans. on Information Theory July 1998.

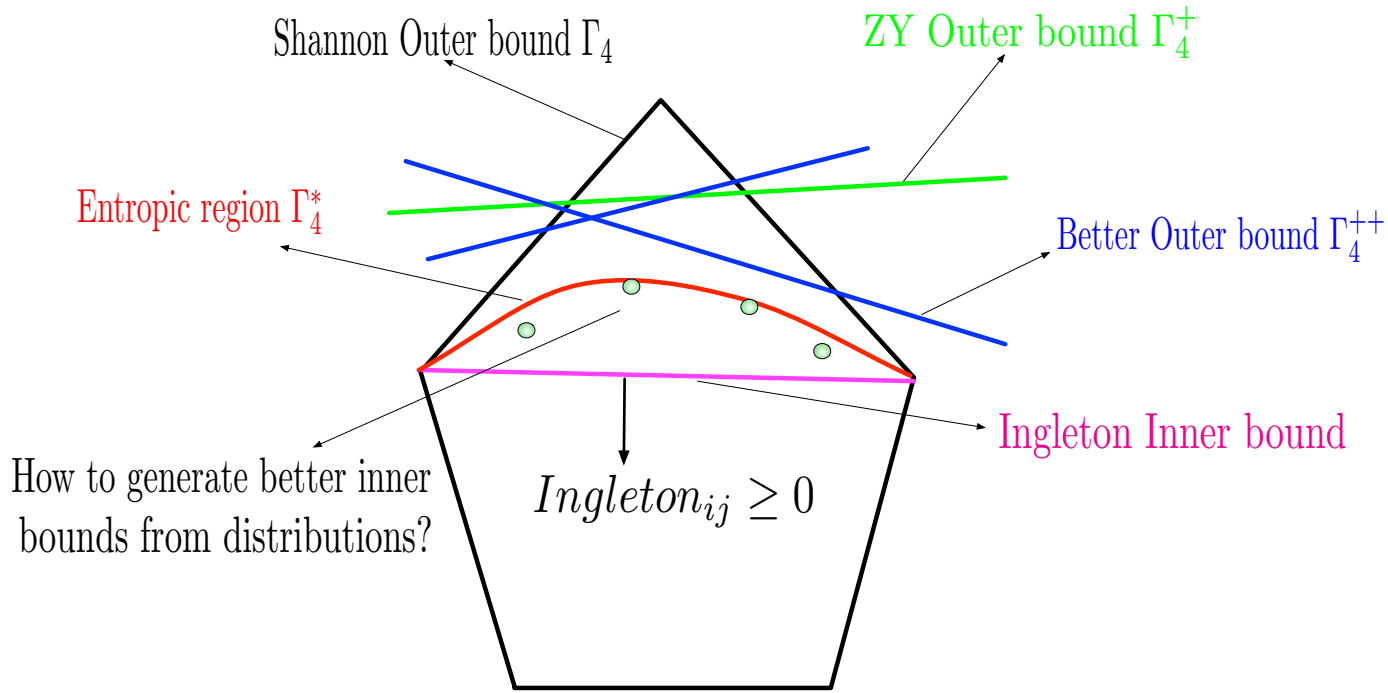
The Region of Entropic Vectors $\overline{\Gamma^*_N}$

Inner bound for $\overline{\Gamma^*_4}$ (N =4):

Ingleton Inequality: $Ingleton_{ij} \geq 0$

$$Ingleton_{ij} = I(X_k; X_l|X_i) + I(X_k; X_l|X_j) + I(X_i; X_j|\emptyset) - I(X_k; X_l|\emptyset)$$

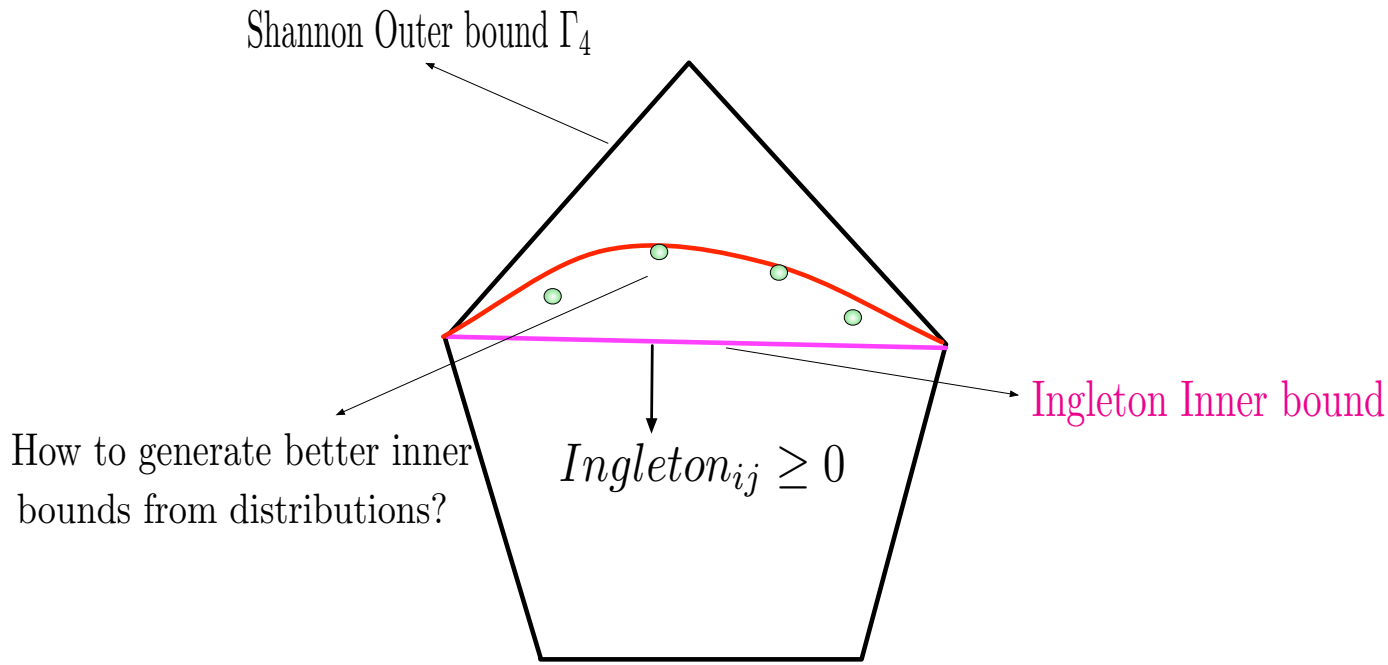
Ingleton Inequality hold for ranks(dimensions) of finite subsets of linear spaces, meaning linear codes give us $Ingleton_{ij} \geq 0$.



The Region of Entropic Vectors $\overline{\Gamma_N^*}$

Motivation of the main results in contribution 1 and contribution 2:

- What type of distributions generate entropy vectors in the gap? Distribution support Enumeration
- What are the properties of distributions on the boundary and in the gap? Information Geometric characterization



Outline

The Region of Entropic Vectors $\overline{\Gamma^*_N}$

Enumerating k-atom supports and map to Entropic Region

- Non-isomorphic k-atom supports
- Maximal Ingleton Violation and the Four Atom Conjecture
- Optimizing Entropy Inner Bounds from k-atom Distribution

Characterization of Entropic Region via Information Geometry

- Manifold of Probability distributions
- e-autoparallel submanifold and m-autoparallel submanifold
- Projections and Pythagorean Theorem
- Information Geometry of k-atom support

From Conditional Independence to Probabilistic Models

- *p-representable* semimatroids
- Probabilistic Models
- Information guided dataset partitioning for supervised learning
- Experiments on Information guided dataset partitioning

Enumerating k-atom supports and map to Entropic Region

References

- Yunshu Liu and John MacLaren Walsh, *Non-isomorphic Distribution Supports for Calculating Entropic Vectors*, in 53nd Annual Allerton Conference on Communication, Control, and Computing, Oct. 2015.
- Yunshu Liu and John MacLaren Walsh, *Mapping the Entropic Region with Support Enumeration & Information Geometry*, submitted to IEEE Transaction on Information Theory on Dec. 2015.

Outline

The Region of Entropic Vectors $\overline{\Gamma^*_N}$

Enumerating k-atom supports and map to Entropic Region

Non-isomorphic k-atom supports

Maximal Ingleton Violation and the Four Atom Conjecture

Optimizing Entropy Inner Bounds from k-atom Distribution

Characterization of Entropic Region via Information Geometry

Manifold of Probability distributions

e-autoparallel submanifold and m-autoparallel submanifold

Projections and Pythagorean Theorem

Information Geometry of k-atom support

From Conditional Independence to Probabilistic Models

p-representable semimatroids

Probabilistic Models

Information guided dataset partitioning for supervised learning

Experiments on Information guided dataset partitioning

Equivalence of k-atom Supports

An example of 3-variable 4-atom support

$$\begin{bmatrix} (0,0,0) - - > & \alpha \\ (0,2,2) - - > & \beta - \alpha \\ (1,0,1) - - > & \gamma - \alpha \\ (1,1,3) - - > & 1 + \alpha - \beta - \gamma \end{bmatrix} \tag{1}$$

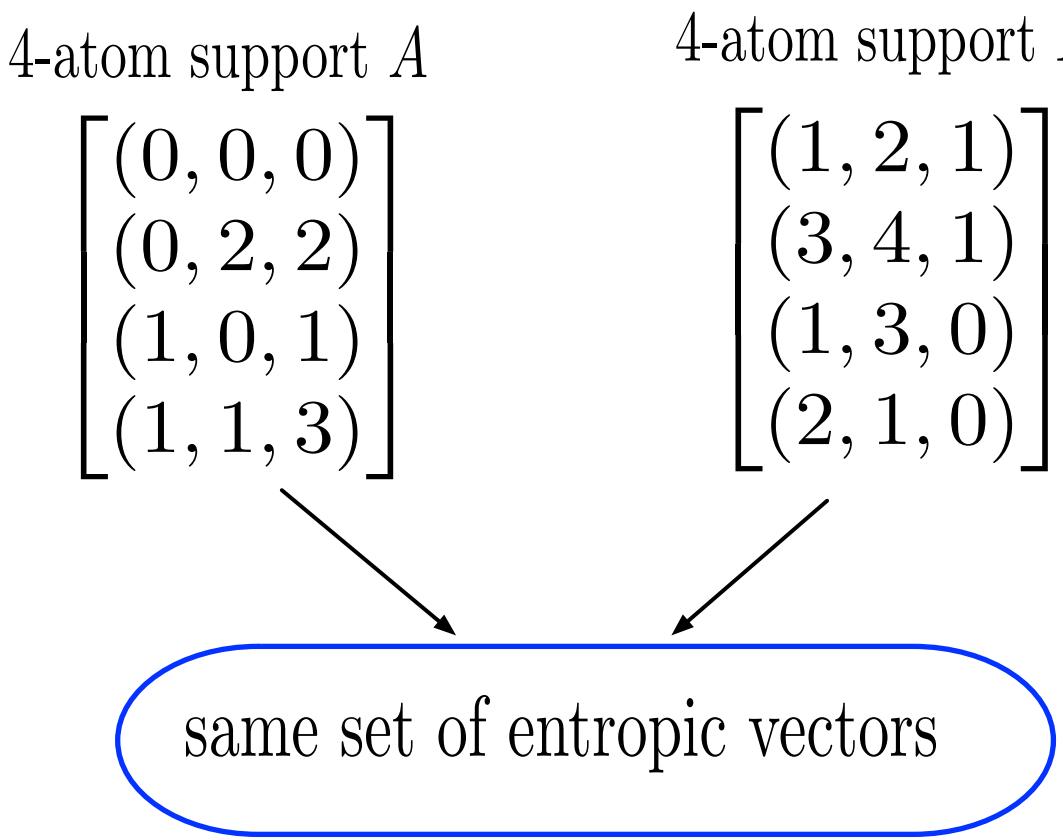
In distribution support (1), each row corresponding to one outcome/atom, each column represent a variable.

Assign variables from left to right as X_1, X_2 and X_3 such that $\boldsymbol{X} = (X_1, X_2, X_3)$. If we assume $|\mathcal{X}_1| = 2$, $|\mathcal{X}_2| = 3$ and $|\mathcal{X}_3| = 4$, then $\boldsymbol{X}$ has totally 24 outcomes, among these outcomes, only 4 of them have non-zero probability, we call it a 4-atom distribution support.

Equivalence of k-atom Supports

Equivalence of two k-atom supports

Two k-atom supports $\boldsymbol{S}, \boldsymbol{S}'$, $|\boldsymbol{S}| = |\boldsymbol{S}'| = k$, 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.



Equivalence of k-atom Supports

Outcome for each variables in a k-atom support is a *set partition*, entropy remain unchanged for set partitions under symmetric group $\mathbb{S}_k$

Consider the 4-atom support in (1)

$$\begin{bmatrix} (0,0,0) \\ (0,2,2) \\ (1,0,1) \\ (1,1,3) \end{bmatrix}$$

outcome of X_1 can be encoded as $\{\{1,2\},\{3,4\}\}$
outcome of X_2 can be encoded as $\{\{1,3\},\{2\},\{4\}\}$
outcome of X_3 can be encoded as $\{\{1\},\{2\},\{3\},\{4\}\}$

How to encode and find equivalent supports for multiple variables?

Orbit data structure

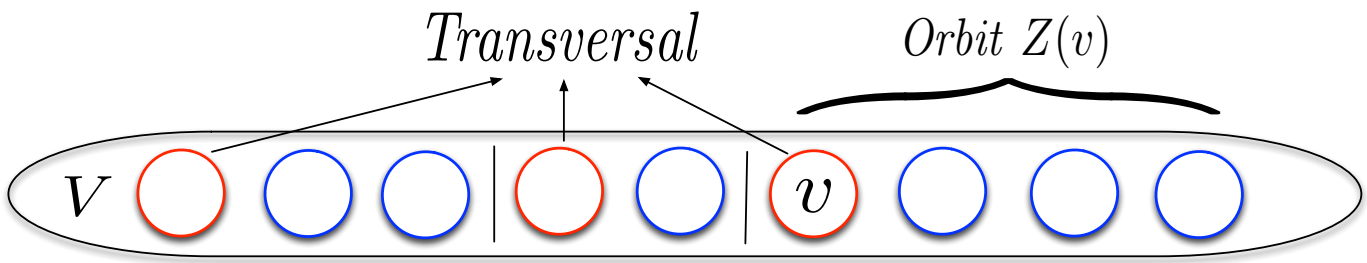
Orbit of elements under group action

Let a finite group Z acting on a finite set V, For $v \in V$, the *orbit* of v under Z is defined as

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

Transversal $\mathcal{T}$ of the Orbits under group Z

transversal is the collection of canonical representative of all the orbits, e.g. least under some ordering of V.



Non-isomorphic k-atom supports

set partitions

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 $\mathbb{N}_1^k$ 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 cardinality of $\Pi(\mathbb{N}_1^k)$ is commonly known as *Bell numbers*.

Examples of *set partitions*

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

$$\Pi(\mathbb{N}_1^4) = \{\{\{1,2,3,4\}\}, \{\{1,2,3\},\{4\}\}, \{\{1,2,4\},\{3\}\}, \{\{1,3,4\},\{2\}\}, \{\{2,3,4\},\{1\}\}, \{\{1,2\},\{3,4\}\}, \{\{1,3\},\{2,4\}\}, \{\{1,4\},\{2,3\}\}, \{\{1,2\},\{3\},\{4\}\}, \{\{1,3\},\{2\},\{4\}\}, \{\{1,4\},\{2\},\{3\}\}, \{\{2,3\},\{1\},\{4\}\}, \{\{2,4\},\{1\},\{3\}\}, \{\{3,4\},\{1\},\{2\}\}, \{\{1\},\{2\},\{3\},\{4\}\}\}.$$

Non-isomorphic k-atom supports

Use combinations of N set partitions to represent distribution supports for N variables

Let Ξ_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),

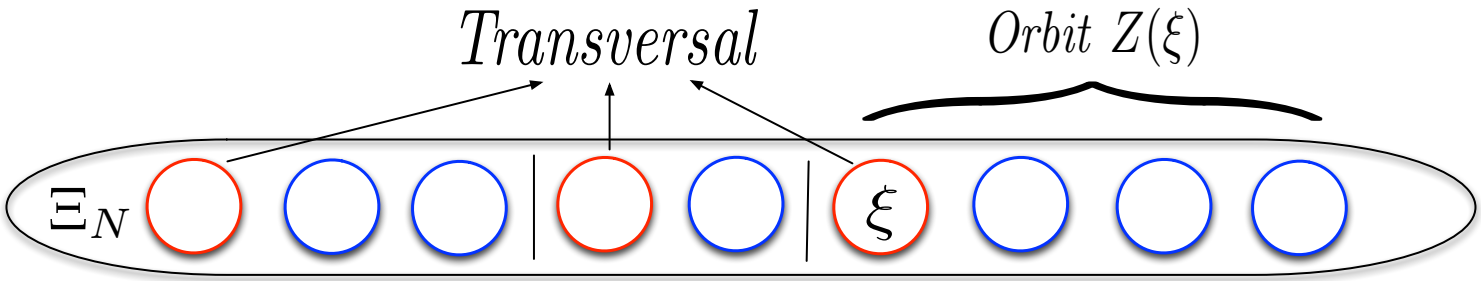
$$\Xi_N := \left\{ \xi \,\middle|\, \xi \subseteq \Pi(\mathbb{N}_1^k),\ |\xi| = N,\ \bigwedge_{\mathcal{B}\in\xi} \mathcal{B} = \bigcup_{i=1}^{N} \{\{i\}\} \right\}. \qquad (2)$$

where *meet* of two partitions $\mathcal{B}, \mathcal{B}'$ is defined as

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

Non-isomorphic k-atom supports

Ξ_N represent all the necessary distribution supports to calculate entropic vectors for N variables and k atoms, selecting one representative from each equivalence class will lead us to a list of all non-isomorphic k-atom, N-variable supports.



The problem of generating the list of all non-isomorphic k-atom, N-variable supports is equivalent to obtaining a transversal of the orbits of $\mathbb{S}_k$ acting on Ξ_N.

Non-isomorphic k-atom supports

Calculate transversal via the algorithm of Snakes and Ladders[2]

For given set $\mathbb{N}_1^k$, the algorithm of Snakes and Ladders first compute the transversal of Ξ_1, then it recursively increase the subsets size i, where the enumeration of Ξ_i is depending on the result of Ξ_{i-1}: $\Xi_1 \to \Xi_2 \to \cdots \to \Xi_{N-1} \to \Xi_N \cdots$

Number of non-isomorphic k-atom, N-variable supports

$N \backslash k$	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		

[2] **Anton Betten**, **Michael Braun**, **Harald Fripertinger etc.** , **Error-Correcting Linear Codes: Classification by Isometry and Applications**, **Springer 2006 pp. 709–710**.

Outline

The Region of Entropic Vectors $\overline{\Gamma^*_N}$

Enumerating k-atom supports and map to Entropic Region

Non-isomorphic k-atom supports

Maximal Ingleton Violation and the Four Atom Conjecture

Optimizing Entropy Inner Bounds from k-atom Distribution

Characterization of Entropic Region via Information Geometry

Manifold of Probability distributions

e-autoparallel submanifold and m-autoparallel submanifold

Projections and Pythagorean Theorem

Information Geometry of k-atom support

From Conditional Independence to Probabilistic Models

p-representable semimatroids

Probabilistic Models

Information guided dataset partitioning for supervised learning

Experiments on Information guided dataset partitioning

Maximal Ingleton Violation and the Four Atom Conjecture

Ingleton score[3]

Given a probability distribution, we define the Ingleton score to be

$$\text{Ingleton score} = \frac{Ingleton_{ij}}{H(X_i X_j X_k X_l)} \tag{3}$$

where

$$Ingleton_{ij} = I(X_k; X_l|X_i) + I(X_k; X_l|X_j) + I(X_i; X_j|\emptyset) - I(X_k; X_l|\emptyset)$$

Ingleton score determine how much a distribution violate one of the six Ingleton Inequality

[3] **Randall Dougherty**, **Chris Freiling**, **Kenneth Zeger**, **Non-Shannon Information Inequalities in Four Random Variables**, **arXiv:1104.3602v1**.

Maximal Ingleton Violation and the Four Atom Conjecture

Four Atom Conjecture: the lowest reported Ingleton score is approximately -0.0894, it is attained by (4)

$$\begin{bmatrix} (0,0,0,0) - - > 0.35 \\ (0,1,1,0) - - > 0.15 \\ (1,0,1,0) - - > 0.15 \\ (1,1,1,1) - - > 0.35 \end{bmatrix} \tag{4}$$

A even lower Ingleton score of -0.09243 was reported by F. Matúš and L. Csirmaz[4] through a transformation of entropic vectors, thus without a distribution associated with it.

[4] **F. Matus and L. Csirmaz**, **Entropy region and convolution**, **arXiv:1310.5957v1**.

Maximal Ingleton Violation and the Four Atom Conjecture

Number of non-isomorphic Ingleton violating supports for four variables

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

The only Ingleton violating 5-atom support that is not grown from 4-atom support (4) with a Ingleton score -0.02423

$$\begin{bmatrix} (0,0,0,0) \\ (0,0,1,1) \\ (0,1,1,0) \\ (1,0,1,0) \\ (1,1,1,0) \end{bmatrix} \tag{5}$$

Outline

The Region of Entropic Vectors $\overline{\Gamma^*_N}$

Enumerating k-atom supports and map to Entropic Region

- Non-isomorphic k-atom supports
- Maximal Ingleton Violation and the Four Atom Conjecture
- Optimizing Entropy Inner Bounds from k-atom Distribution

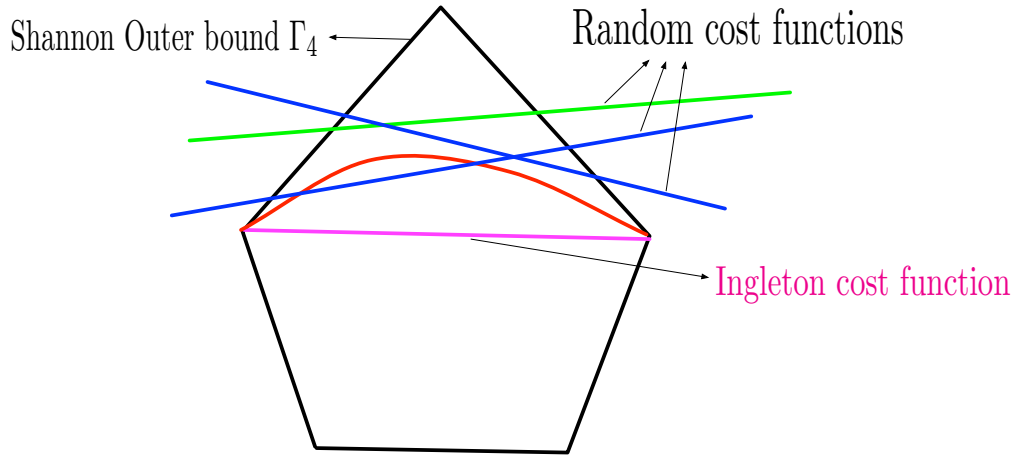
Characterization of Entropic Region via Information Geometry

- Manifold of Probability distributions
- e-autoparallel submanifold and m-autoparallel submanifold
- Projections and Pythagorean Theorem
- Information Geometry of k-atom support

From Conditional Independence to Probabilistic Models

- *p-representable* semimatroids
- Probabilistic Models
- Information guided dataset partitioning for supervised learning
- Experiments on Information guided dataset partitioning

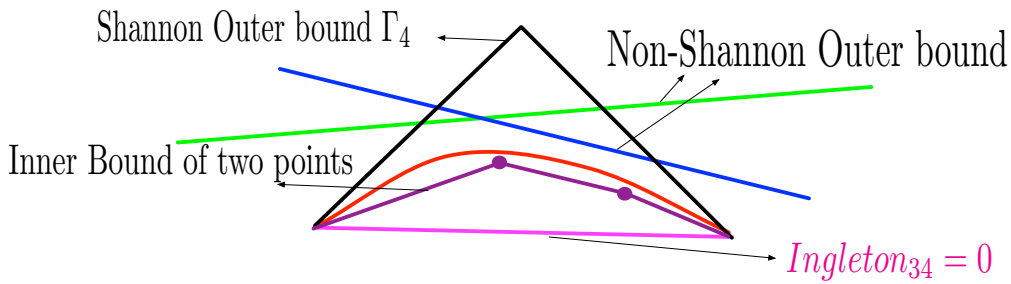
Optimizing Inner Bounds to Entropy from k-atom Distribution



k-atom inner bound generation for four variables

- Randomly generate enough cost functions from the gap between Shannon outer bound and Ingleton inner bound
- Given a k-atom support, find the distribution that optimizes each of the cost function, save it if it violates Ingleton
- Taking the convex hull of all the entropic vectors from the k-atom distributions that violate $Ingleton_{kl}$ to get the k-atom inner bound

Optimizing Inner Bounds to Entropy from k-atom Distribution

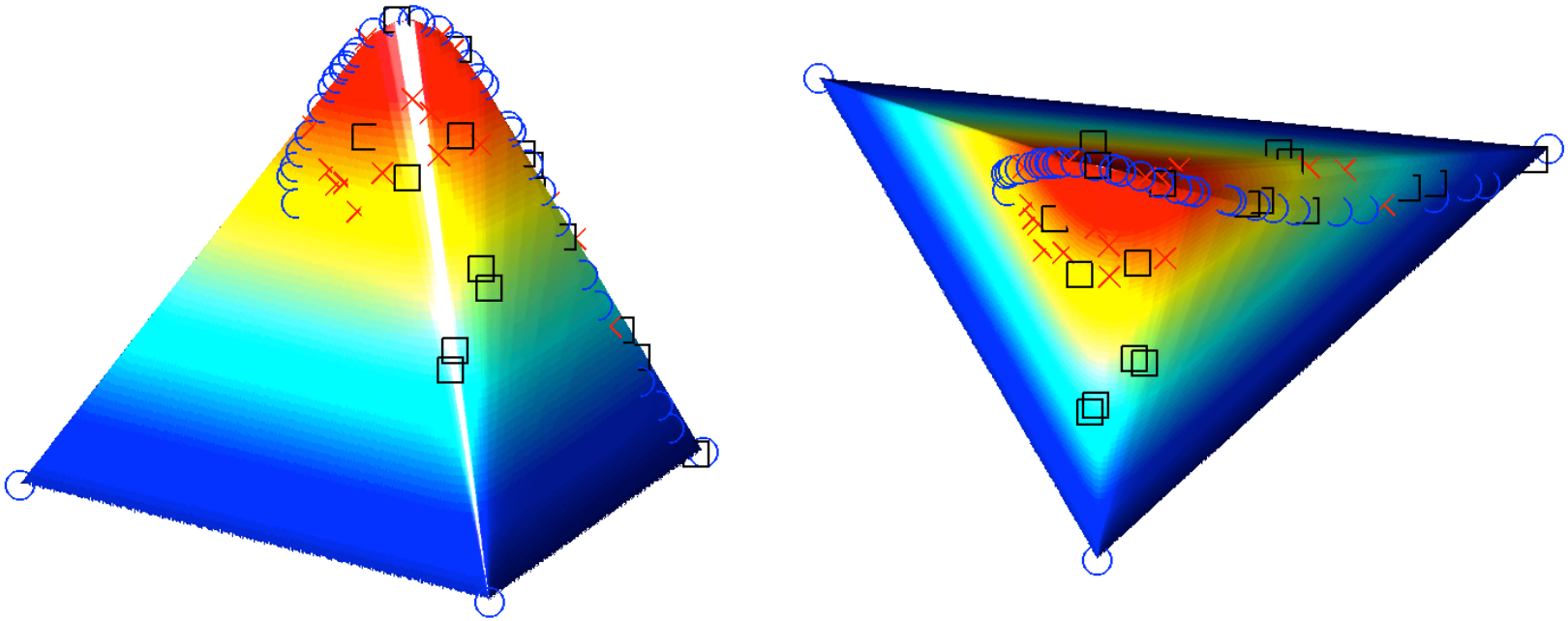


inner and outer bounds	percent of pyramid
Shannon	100
Outer bound from Dougherty etc.[5]	96.5
4,5,6 atoms inner bound	57.8
4,5 atoms inner bound	57.1
4 atoms inner bound	55.9
4 atom conjecture point only	43.5
3 atoms inner bound	0

[5] **Randall Dougherty**, **Chris Freiling**, **Kenneth Zeger**, **Non-Shannon Information Inequalities in Four Random Variables**, **arXiv:1104.3602v1**.

Optimizing Inner Bounds to Entropy from k-atom Distribution

Three dimensional projection of 4-atom inner bound(3-D projection introduced by F. Matúš and L. Csirmaz[3])



Outline

The Region of Entropic Vectors $\overline{\Gamma^*_N}$

Enumerating k-atom supports and map to Entropic Region

- Non-isomorphic k-atom supports
- Maximal Ingleton Violation and the Four Atom Conjecture
- Optimizing Entropy Inner Bounds from k-atom Distribution

Characterization of Entropic Region via Information Geometry

- Manifold of Probability distributions
- e-autoparallel submanifold and m-autoparallel submanifold
- Projections and Pythagorean Theorem
- Information Geometry of k-atom support

From Conditional Independence to Probabilistic Models

- *p-representable* semimatroids
- Probabilistic Models
- Information guided dataset partitioning for supervised learning
- Experiments on Information guided dataset partitioning

Characterization of Entropic Region via Information Geometry

References

- Yunshu Liu, John MacLaren Walsh, *Bounding the entropic region via information geometry*, in IEEE Information Theory Workshop, Sep. 2013.
- Yunshu Liu and John MacLaren Walsh, *Mapping the Entropic Region with Support Enumeration & Information Geometry*, submitted to IEEE Transaction on Information Theory on Dec. 2015.

Outline

The Region of Entropic Vectors $\overline{\Gamma^*_N}$

Enumerating k-atom supports and map to Entropic Region

- Non-isomorphic k-atom supports
- Maximal Ingleton Violation and the Four Atom Conjecture
- Optimizing Entropy Inner Bounds from k-atom Distribution

Characterization of Entropic Region via Information Geometry

- Manifold of Probability distributions
- e-autoparallel submanifold and m-autoparallel submanifold
- Projections and Pythagorean Theorem
- Information Geometry of k-atom support

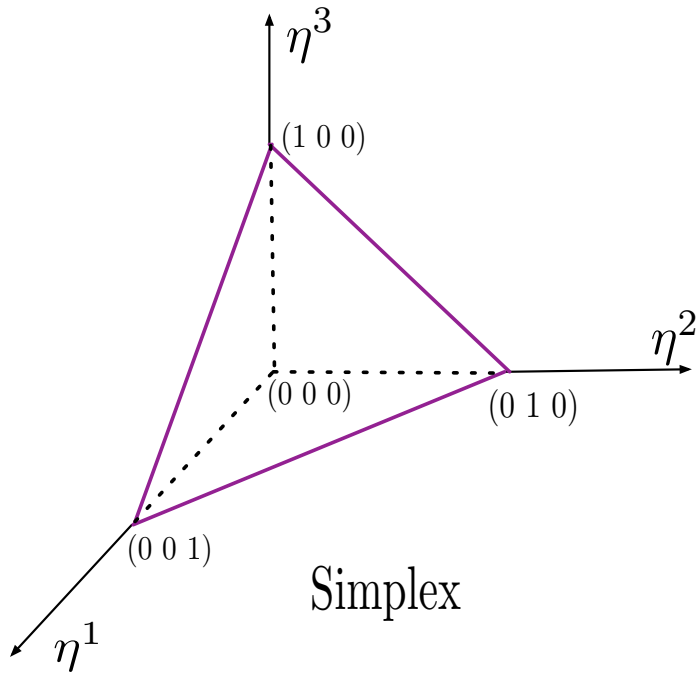
From Conditional Independence to Probabilistic Models

- *p-representable* semimatroids
- Probabilistic Models
- Information guided dataset partitioning for supervised learning
- Experiments on Information guided dataset partitioning

Manifold of Probability distributions

Example of two binary discrete random variables

Let $\boldsymbol{X} = \{X_1, X_2\}$, $X_i = 1$ *or* -1,
consider the family of all the
probability distributions $\mathcal{S} =$
$\{p(\boldsymbol{X}) \mid p(\boldsymbol{X}) > 0, \sum_i p(\boldsymbol{X} = C_i) = 1\}$
Outcomes:
$C_0 = \{-1, -1\}, C_1 = \{-1, 1\}$
$C_2 = \{1, -1\}, \ C_3 = \{1, 1\}$
Parameterization:
$p(\boldsymbol{X} = C_i) = \eta^i$ *for* $i = 1, 2, 3$
$p(\boldsymbol{X} = C_0) = 1 - \sum_{j=1}^{3} \eta^j$
$\boldsymbol{\eta} = (\eta^1 \ \eta^2 \ \eta^3)$ is called the
m-coordinate of $\mathcal{S}$.



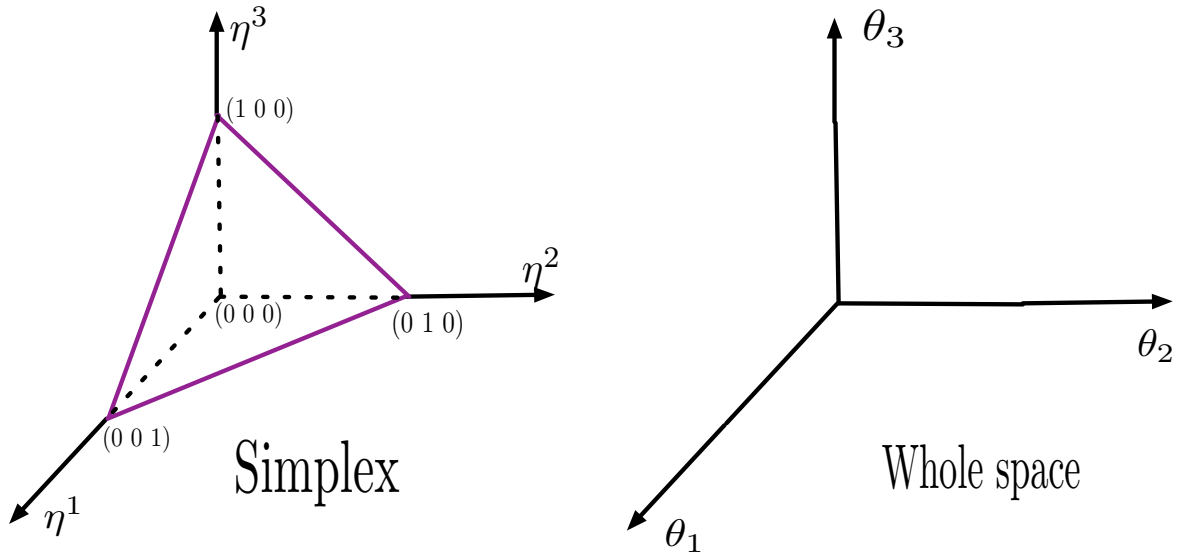
Manifold of Probability distributions

Example of two binary discrete random variables

A new coordinate $\{\theta_i\}$, which is defined as

$$\theta_i = ln \frac{\eta^i}{1 - \sum_{j=1}^{3} \eta^j} \quad for\ i = 1, 2, 3. \tag{6}$$

$\boldsymbol{\theta} = (\theta_1\ \theta_2\ \theta_3)$ is called the e-coordinate of $\mathcal{S}$.



Outline

The Region of Entropic Vectors $\overline{\Gamma^*_N}$

Enumerating k-atom supports and map to Entropic Region

- Non-isomorphic k-atom supports
- Maximal Ingleton Violation and the Four Atom Conjecture
- Optimizing Entropy Inner Bounds from k-atom Distribution

Characterization of Entropic Region via Information Geometry

- Manifold of Probability distributions
- e-autoparallel submanifold and m-autoparallel submanifold
- Projections and Pythagorean Theorem
- Information Geometry of k-atom support

From Conditional Independence to Probabilistic Models

- *p-representable* semimatroids
- Probabilistic Models
- Information guided dataset partitioning for supervised learning
- Experiments on Information guided dataset partitioning

e-autoparallel submanifold and m-autoparallel submanifold

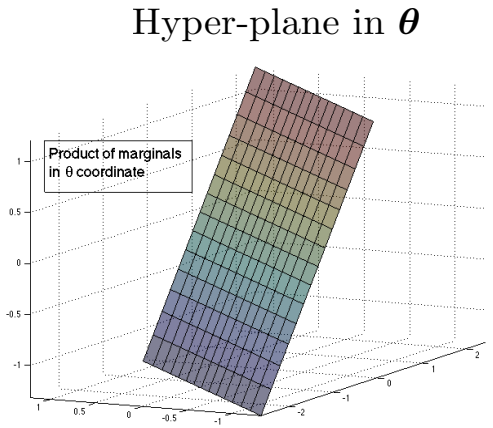
e-autoparallel submanifold

A subfamily $\mathcal{E}$ is said to be e-autoparallel submanifold of $\mathcal{S}$ if for some coordinate $\boldsymbol{\lambda}$ there exist a matrix A and a vector b such that for the e-coordinate $\boldsymbol{\theta}$ of $\mathcal{S}$

$$\boldsymbol{\theta}(p) = \mathbf{A}\boldsymbol{\lambda}(p) + \mathbf{b} \quad (\forall p \in \mathcal{E}) \tag{7}$$

If $\boldsymbol{\lambda}$ is one-dimensional, it is called a *e-geodesic*.

An e-autoparallel submanifold of $\mathcal{S}$ is a hyperplane in e-coordinate;
A e-geodesic of $\mathcal{S}$ is a straight line in e-coordinate.



e-autoparallel submanifold and m-autoparallel submanifold

m-autoparallel submanifold

A subfamily $\mathcal{M}$ is said to be m-autoparallel submanifold of $\mathcal{S}$ if for some coordinate γ there exist a matrix A and a vector b such that for the m-coordinate η of $\mathcal{S}$

$$\eta(p) = \mathbf{A}\gamma(p) + \mathbf{b} \quad (\forall p \in \mathcal{M}) \tag{8}$$

If γ is one-dimensional, it is called a *m-geodesic*.

e-autoparallel submanifold and m-autoparallel submanifold

Two independent bits

The set of all product distributions, which is defined as

$$\mathcal{E} = \{p(\boldsymbol{X}) \mid p(\boldsymbol{X}) = p_{X_1}(x_1)p_{X_2}(x_2)\} \tag{9}$$

$p(X_1 = 1) = p_1$, $p(X_2 = 1) = p_2$

Parameterization

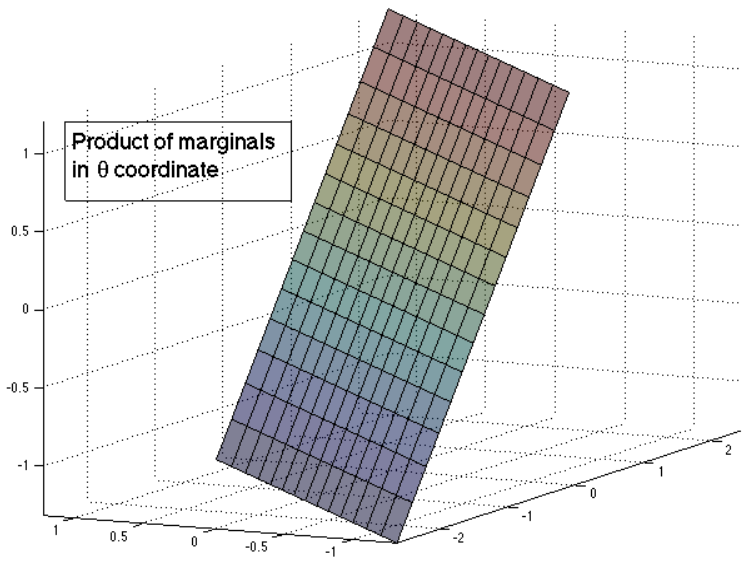
Define $\lambda_i = \frac{1}{2}\, ln\, \frac{p_i}{1-p_i}$ for i = 1, 2

Then $\mathcal{E}$ is an e-autoparallel submanifold(plane in θ coordinate) of S:

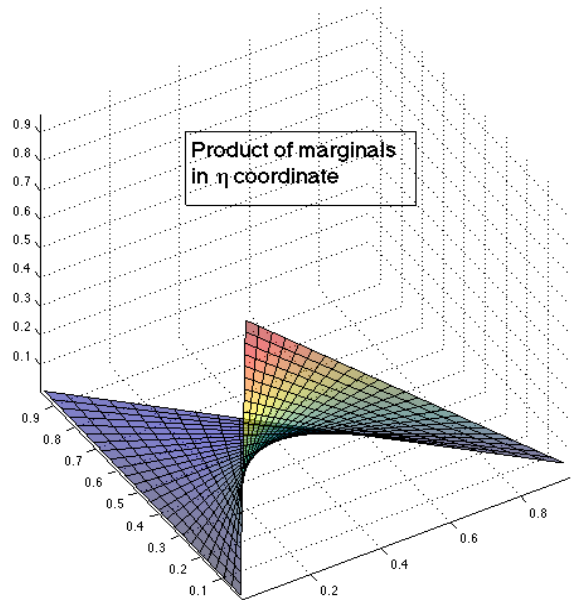
$$\boldsymbol{\theta}(p) = \begin{pmatrix} \theta_1 \\ \theta_2 \\ \theta_3 \end{pmatrix} = \begin{pmatrix} 2 & 2 \\ 2 & 0 \\ 0 & 2 \end{pmatrix} \bullet \begin{pmatrix} \lambda_1 \\ \lambda_2 \end{pmatrix} = \begin{pmatrix} 2 & 2 \\ 2 & 0 \\ 0 & 2 \end{pmatrix} \bullet \boldsymbol{\lambda}(p)$$

e-autoparallel submanifold and m-autoparallel submanifold

Two independent bits: e-autoparallel submanifold but not m-autoparallel submanifold



$\mathcal{E}$ in $\boldsymbol{\theta}$ coordinate



$\mathcal{E}$ in $\boldsymbol{\eta}$ coordinate

Outline

The Region of Entropic Vectors $\overline{\Gamma^*_N}$

Enumerating k-atom supports and map to Entropic Region

- Non-isomorphic k-atom supports
- Maximal Ingleton Violation and the Four Atom Conjecture
- Optimizing Entropy Inner Bounds from k-atom Distribution

Characterization of Entropic Region via Information Geometry

- Manifold of Probability distributions
- e-autoparallel submanifold and m-autoparallel submanifold
- Projections and Pythagorean Theorem
- Information Geometry of k-atom support

From Conditional Independence to Probabilistic Models

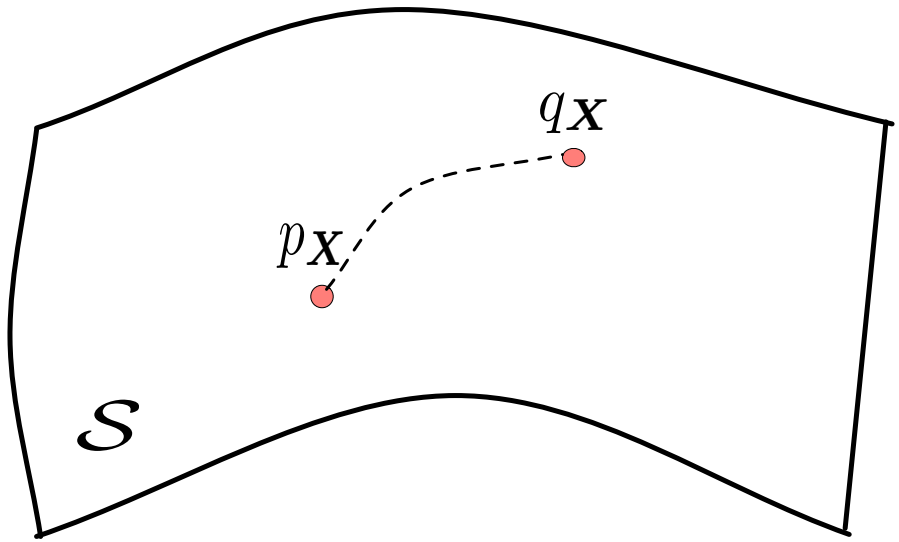
- *p-representable* semimatroids
- Probabilistic Models
- Information guided dataset partitioning for supervised learning
- Experiments on Information guided dataset partitioning

Projections and Pythagorean Theorem

KL-divergence

On the manifold of probability mass functions for random variables taking values in the set $\mathcal{X}$, we can also define the Kullback Leibler divergence, or relative entropy, measured in bits, according to

$$D(p_{\boldsymbol{X}}||q_{\boldsymbol{X}}) = \sum_{\boldsymbol{x}\in\mathcal{X}} p_{\boldsymbol{X}}(\boldsymbol{x}) \log\left(\frac{p_{\boldsymbol{X}}(\boldsymbol{x})}{q_{\boldsymbol{X}}(\boldsymbol{x})}\right) \tag{10}$$



Projections and Pythagorean Theorem

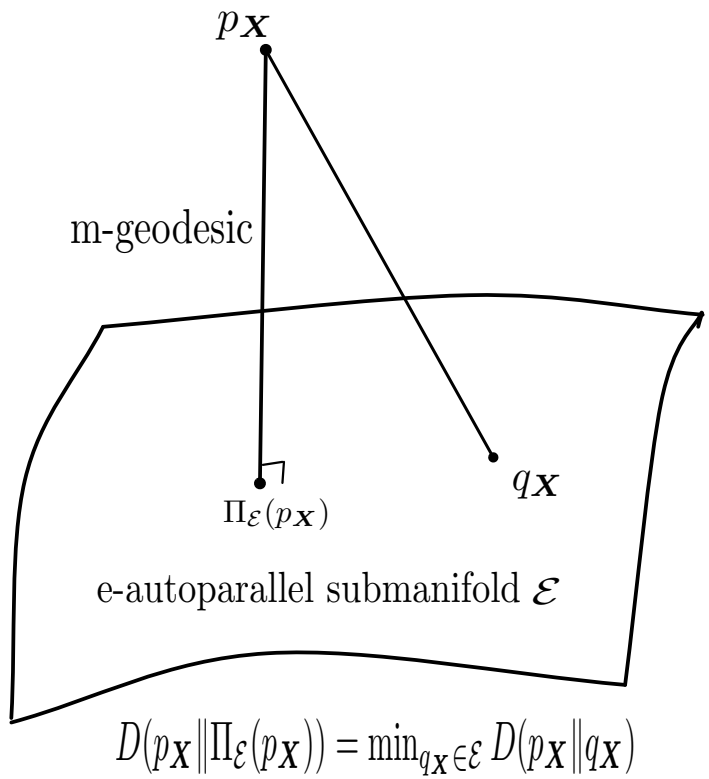
Information Projection

Let p be a point in $\mathcal{S}$ and let $\mathcal{E}$ be a e-autoparallel submanifold of $\mathcal{S}$. We find projection point $\Pi_{\mathcal{E}}(p_{\boldsymbol{X}})$ in $\mathcal{E}$ minimizes the KL-divergence

$$D(p_{\boldsymbol{X}} \| \Pi_{\mathcal{E}}(p_{\boldsymbol{X}})) = \min_{q_{\boldsymbol{X}} \in \mathcal{E}} D(p_{\boldsymbol{X}} \| q_{\boldsymbol{X}}) \tag{11}$$

The m-geodesic connecting $p_{\boldsymbol{X}}$ and $\Pi_{\mathcal{E}}(p_{\boldsymbol{X}})$ is orthogonal to $\mathcal{E}$ at $\Pi_{\mathcal{E}}(p_{\boldsymbol{X}})$[5].

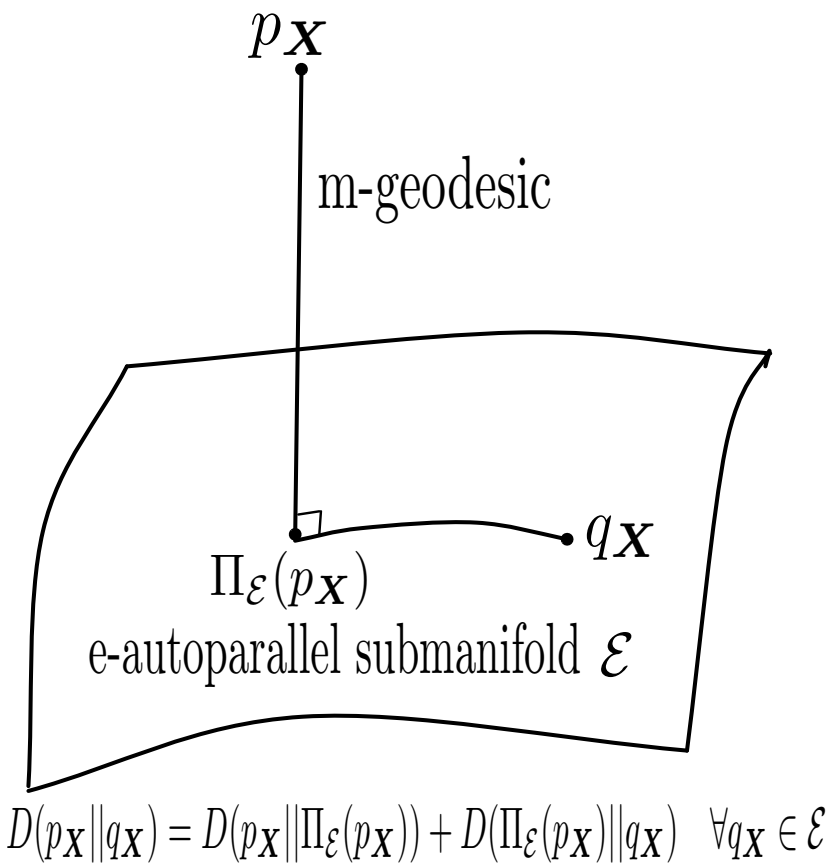
- In geometry, orthogonal usually means right angle
- In information geometry, orthogonal usually means certain values are uncorrelated



Projections and Pythagorean Theorem

For $\forall q_{\boldsymbol{X}} \in \mathcal{E}$, we have the following *Pythagorean relation*[6].

$$D(p_{\boldsymbol{X}}||q_{\boldsymbol{X}}) = D(p_{\boldsymbol{X}}||\Pi_{\mathcal{E}}(p_{\boldsymbol{X}})) + D(\Pi_{\mathcal{E}}(p_{\boldsymbol{X}})||q_{\boldsymbol{X}}) \quad \forall q_{\boldsymbol{X}} \in \mathcal{E} \quad (12)$$



[6] **Shun-ichi Amari and Hiroshi Nagaoka, Methods of Information Geometry, Oxford University Press, 2000.**

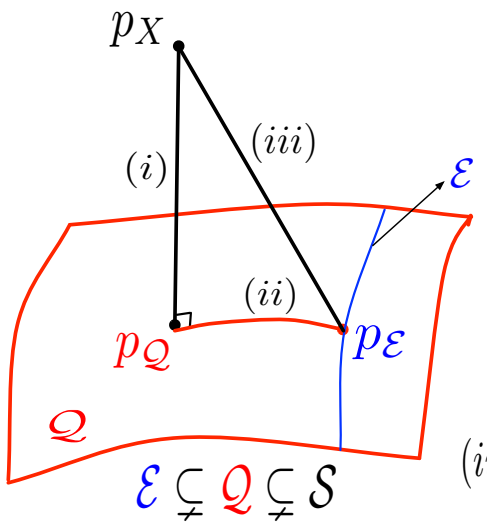
Projections and Pythagorean Theorem

From Pythagorean relation to Information Inequalities

For $\mathcal{S} = \{p(\boldsymbol{X}) \mid p(\boldsymbol{X}) = q_X(x_1, \ldots, x_N)\}$, consider submanifolds

- $\mathcal{Q} = \left\{ p_{\boldsymbol{X}}(\cdot) \,\middle|\, p_{\boldsymbol{X}} = q_{\boldsymbol{X}_{\mathcal{A}\cup\mathcal{B}}} \cdot q_{\boldsymbol{X}_{(\mathcal{A}\cup\mathcal{B})^c}} \quad \forall \boldsymbol{x} \right\}$
- $\mathcal{E} = \left\{ p_{\boldsymbol{X}}(\cdot) \,\middle|\, p_{\boldsymbol{X}} = q_{\boldsymbol{X}_{\mathcal{A}\setminus\mathcal{B}}|\boldsymbol{X}_{\mathcal{A}\cap\mathcal{B}}} \cdot q_{\boldsymbol{X}_{\mathcal{B}\setminus\mathcal{A}}} \cdot q_{\boldsymbol{X}_{(\mathcal{A}\cup\mathcal{B})^c}} \quad \forall \boldsymbol{x} \right\}$

$\mathcal{Q}$ and $\mathcal{E}$ are e-autoparallel submanifold of $\mathcal{S}$ and $\mathcal{E} \subsetneq \mathcal{Q} \subsetneq \mathcal{S}$.
$p_{\mathcal{Q}}$ and $p_{\mathcal{E}}$: Projection of $p_{\boldsymbol{X}}$ onto $\mathcal{Q}$ and $\mathcal{E}$ respectively



Pythagorean relation:

$$D(p_X \| p_{\mathcal{E}}) = D(p_X \| p_{\mathcal{Q}}) + D(p_{\mathcal{Q}} \| p_{\mathcal{E}})$$

Markov chain in B:

$$\boldsymbol{X}_{\mathcal{A}\setminus\mathcal{B}} \leftrightarrow \boldsymbol{X}_{\mathcal{A}\cap\mathcal{B}} \leftrightarrow \boldsymbol{X}_{\mathcal{B}\setminus\mathcal{A}}$$

Entropy Submodularity:

$$(ii)\ D(p_{\mathcal{Q}} \| p_{\mathcal{E}}) = h_{\mathcal{A}} + h_{\mathcal{B}} - h_{\mathcal{A}\cup\mathcal{B}} - h_{\mathcal{A}\cap\mathcal{B}} \geq 0$$

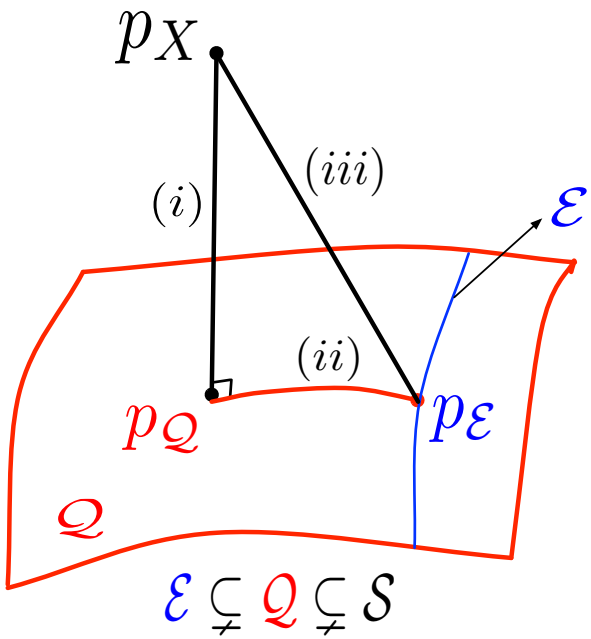
Projections and Pythagorean Theorem

From Pythagorean to Information Inequalities: example

For $S = \{p(\boldsymbol{X}) \mid p(\boldsymbol{X}) = q_X(x_1, x_2, x_3)\}$, consider submanifolds

- $\mathcal{Q} = \{p(\boldsymbol{X}) \mid p(\boldsymbol{X}) = q_{X_1X_3}(x_1, x_3)q_{X_2}(x_2)\}$
- $\mathcal{E} = \{p(\boldsymbol{X}) \mid p(\boldsymbol{X}) = q_{X_1}(x_1)q_{X_2}(x_2)q_{X_3}(x_3)\}$

Projection of $p_{\boldsymbol{X}}$ onto $\mathcal{Q}$: $p_{\mathcal{Q}} = p_{X_1X_3}(x_1, x_3)p_{X_2}(x_2)$
Projection of $p_{\boldsymbol{X}}$ onto $\mathcal{E}$: $p_{\mathcal{E}} = p_{X_1}(x_1)p_{X_2}(x_2)p_{X_3}(x_3)$



Pythagorean relation

$$D(p_X\|p_{\mathcal{E}}) = D(p_X\|p_{\mathcal{Q}}) + D(p_{\mathcal{Q}}\|p_{\mathcal{E}})$$

$$(iii) = (i) + (ii)$$

(i) $D(p_X\|p_{\mathcal{Q}}) = h_{13} + h_2 - h_{123} \geq 0$

(ii) $D(p_{\mathcal{Q}}\|p_{\mathcal{E}}) = h_1 + h_3 - h_{13} \geq 0$

(iii) $D(p_X\|p_{\mathcal{E}}) = h_1 + h_2 + h_3 - h_{123} \geq 0$

Outline

The Region of Entropic Vectors $\overline{\Gamma^*_N}$

Enumerating k-atom supports and map to Entropic Region

- Non-isomorphic k-atom supports
- Maximal Ingleton Violation and the Four Atom Conjecture
- Optimizing Entropy Inner Bounds from k-atom Distribution

Characterization of Entropic Region via Information Geometry

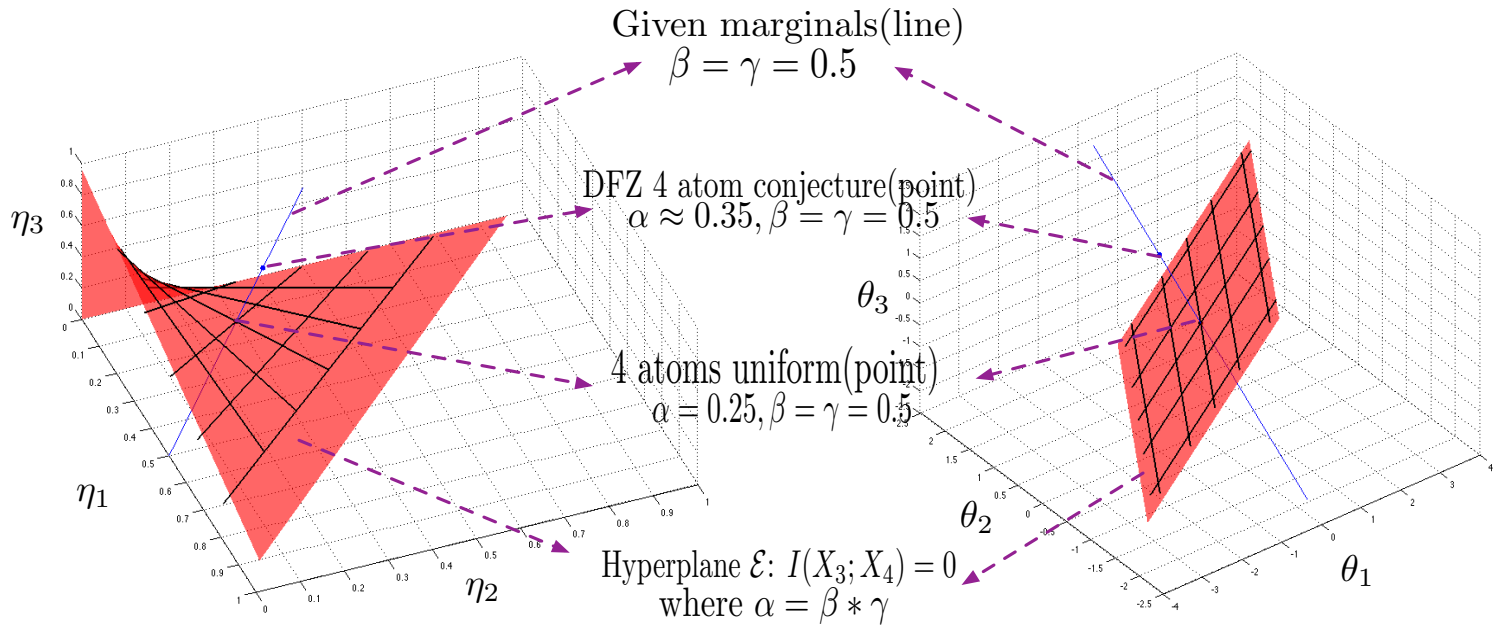
- Manifold of Probability distributions
- e-autoparallel submanifold and m-autoparallel submanifold
- Projections and Pythagorean Theorem
- Information Geometry of k-atom support

From Conditional Independence to Probabilistic Models

- *p-representable* semimatroids
- Probabilistic Models
- Information guided dataset partitioning for supervised learning
- Experiments on Information guided dataset partitioning

Characterization of 4-atoms support for Four variables

Recall 4-atom support in (4) [(0000)(0110)(1010)(1111)]
Left: 4-atom support in m-coordinate
Right: 4-atom support in e-coordinate



The whole 3D space

$p(0000) = \alpha$
$p(0110) = \beta - \alpha$
$p(1010) = \gamma - \alpha$
$p(1111) = 1 + \alpha - \gamma - \beta$

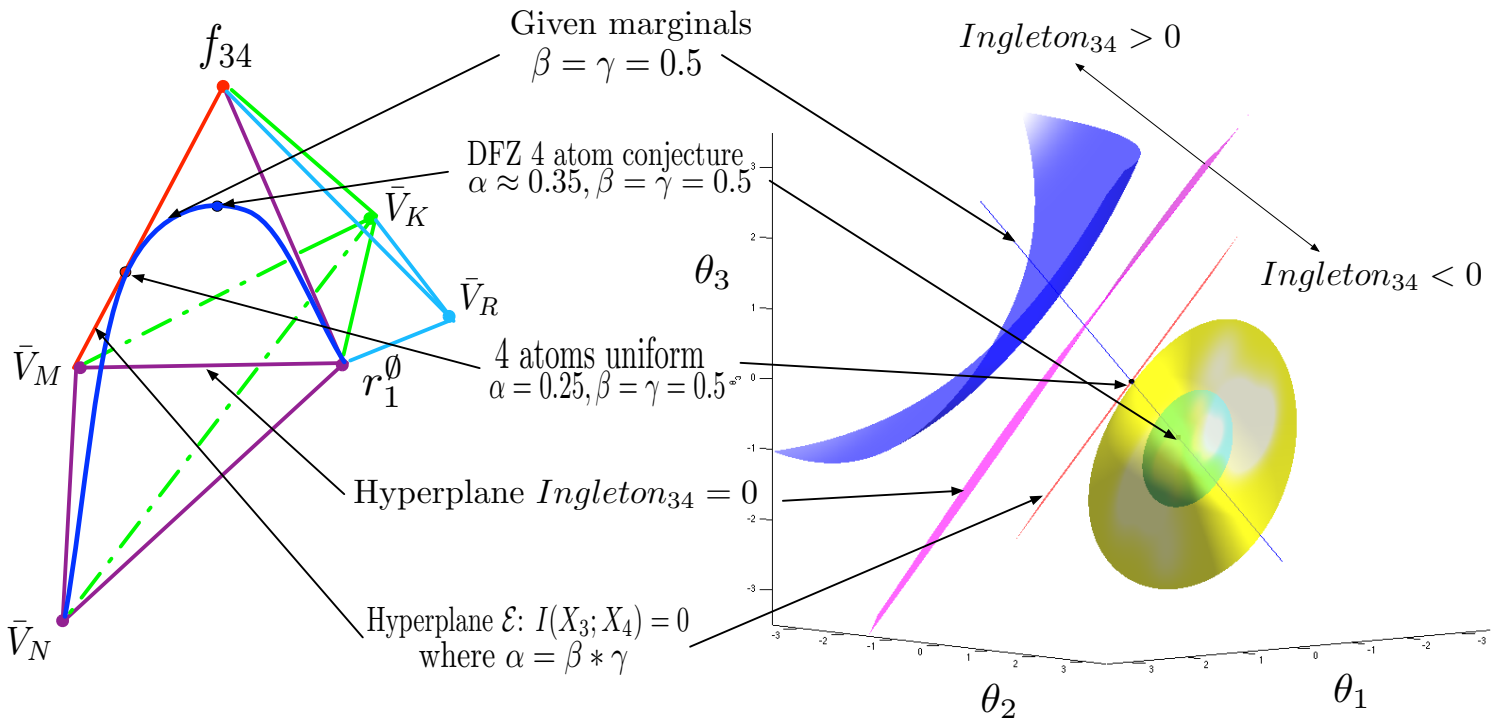
Marginal distribution of X_3 and X_4

$p(X_3 = 0) = \gamma$

$p(X_4 = 0) = \beta$

Characterization of 4-atoms support for Four variables

Ingleton$_{34} = 0$ is *e*-autoparallel in θ coordinate for the given 4-atom support

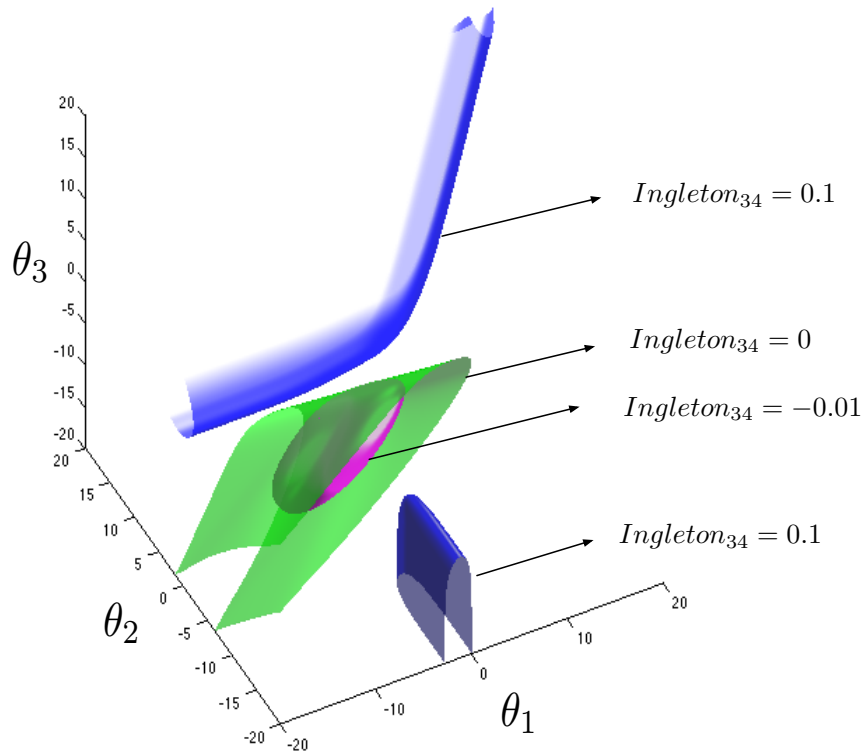


Characterization of 5-atoms support for Four variables

Recall 5-atom support in Equation (5)

[(0,0,0,0)(0,0,1,1)(0,1,1,0)(1,0,1,0)(1,1,1,0)]

Fixing one coordinate θ_0 in e-coordinate to arbitrary value, to plot the resulting 3 dimensional manifold:



The result shows the *e*-autoparallel property of $Ingleton_{34} = 0$ is unique to the given 4-atom support

Outline

The Region of Entropic Vectors $\overline{\Gamma^*_N}$

Enumerating k-atom supports and map to Entropic Region

- Non-isomorphic k-atom supports
- Maximal Ingleton Violation and the Four Atom Conjecture
- Optimizing Entropy Inner Bounds from k-atom Distribution

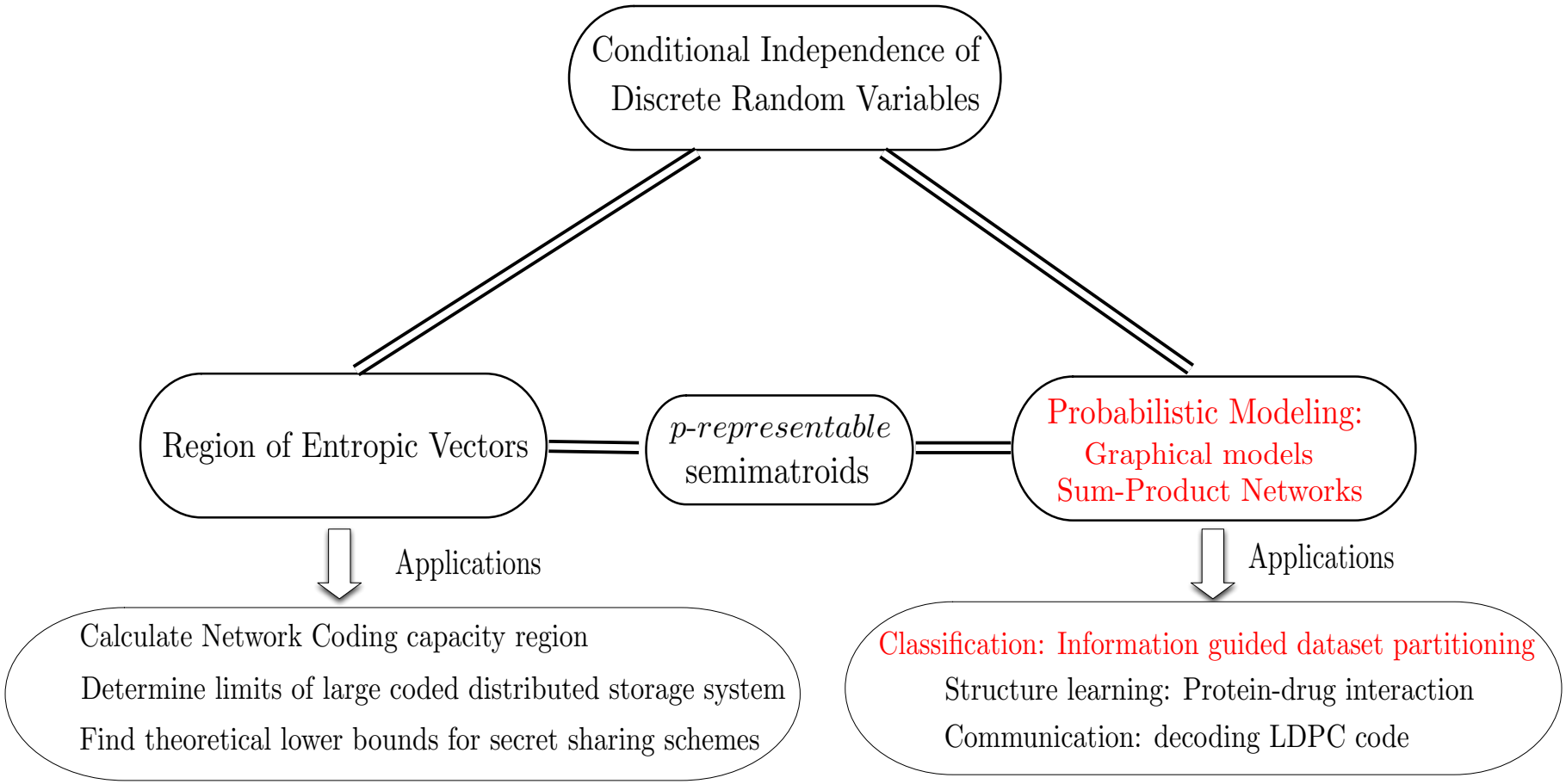
Characterization of Entropic Region via Information Geometry

- Manifold of Probability distributions
- e-autoparallel submanifold and m-autoparallel submanifold
- Projections and Pythagorean Theorem
- Information Geometry of k-atom support

From Conditional Independence to Probabilistic Models

- *p-representable* semimatroids
- Probabilistic Models
- Information guided dataset partitioning for supervised learning
- Experiments on Information guided dataset partitioning

Motivation


Conditional Independence of
Discrete Random Variables
Region of Entropic Vectors
p-representable
semimatroids
Probabilistic Modeling:
Graphical models
Sum-Product Networks
Applications
Applications
Calculate Network Coding capacity region
Determine limits of large coded distributed storage system
Find theoretical lower bounds for secret sharing schemes
Classification: Information guided dataset partitioning
Structure learning: Protein-drug interaction
Communication: decoding LDPC code

From Conditional Independence to Probabilistic Models

References

- Yunshu Liu and John MacLaren Walsh, Information Guided Dataset Partitioning for Supervised Learning, Conference on Uncertainty in Artificial Intelligence, submitted on March 01, 2016.

Outline

The Region of Entropic Vectors $\overline{\Gamma^*_N}$

Enumerating k-atom supports and map to Entropic Region

- Non-isomorphic k-atom supports
- Maximal Ingleton Violation and the Four Atom Conjecture
- Optimizing Entropy Inner Bounds from k-atom Distribution

Characterization of Entropic Region via Information Geometry

- Manifold of Probability distributions
- e-autoparallel submanifold and m-autoparallel submanifold
- Projections and Pythagorean Theorem
- Information Geometry of k-atom support

From Conditional Independence to Probabilistic Models

- *p-representable* semimatroids
- Probabilistic Models
- Information guided dataset partitioning for supervised learning
- Experiments on Information guided dataset partitioning

p-representable semimatroids

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 i, j are two singletons 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 subset $\mathcal{L} \subset \mathcal{S}$ is called *p-representable semimatroids* (probabilistically representable)if there exists a system of N random variables $\mathbf{X} = \{X_i\}_{i \in \mathcal{N}}$ such that

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

p-representable semimatroids

- František Matúš and Milan Studený find the minimal set of configurations of conditional independence that can occur within four discrete random variables.
- Milan Studený studied the minimal set of conditional independent structures that can be described by undirected graphs, directed acyclic graphs and chain graphs for four variables.
- František Matúš and Milan Studený point out there are huge difference between the above two results, meaning lots of conditional independence structures can not be represented by traditional graph-based models.

Outline

The Region of Entropic Vectors $\overline{\Gamma^*_N}$

Enumerating k-atom supports and map to Entropic Region

- Non-isomorphic k-atom supports
- Maximal Ingleton Violation and the Four Atom Conjecture
- Optimizing Entropy Inner Bounds from k-atom Distribution

Characterization of Entropic Region via Information Geometry

- Manifold of Probability distributions
- e-autoparallel submanifold and m-autoparallel submanifold
- Projections and Pythagorean Theorem
- Information Geometry of k-atom support

From Conditional Independence to Probabilistic Models

- *p-representable* semimatroids
- Probabilistic Models
- Information guided dataset partitioning for supervised learning
- Experiments on Information guided dataset partitioning

Probabilistic Modeling: Bayesian networks

Bayesian networks: directed graphical model

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:

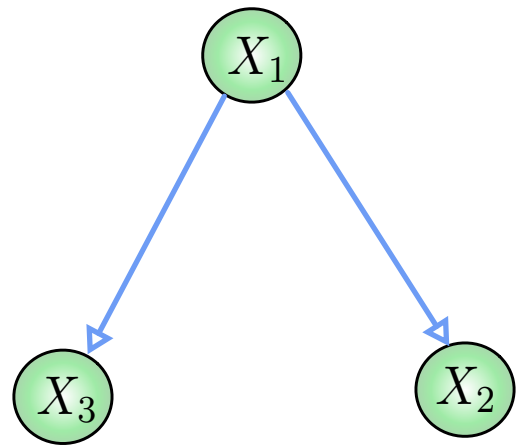
$$p(\boldsymbol{X}) = p(X_1, \cdots, X_K) = \prod_{k \in K} p(X_k | \mathbf{pa}_k)$$

where $\mathbf{pa}_k$ is the direct parents nodes of X_k.

Examples of Bayesian networks

Consider joint distribution $p(\boldsymbol{X}) = p(X_1, X_2, X_3)$ over three variables, we can write:

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



Probabilistic Modeling: limitations of Graphical Models

Problems with Graphical Models

- Only able to represent limit number of conditional independent relations
- Inference is usually intractable beyond tree structured graphs
- Not able to represent conditional independent relations exhibit exclusively in certain contexts.

Probabilistic Modeling: Sum-product Network

Sum-product Network: Definition

A sum-product network(SPN) is defined as follows.

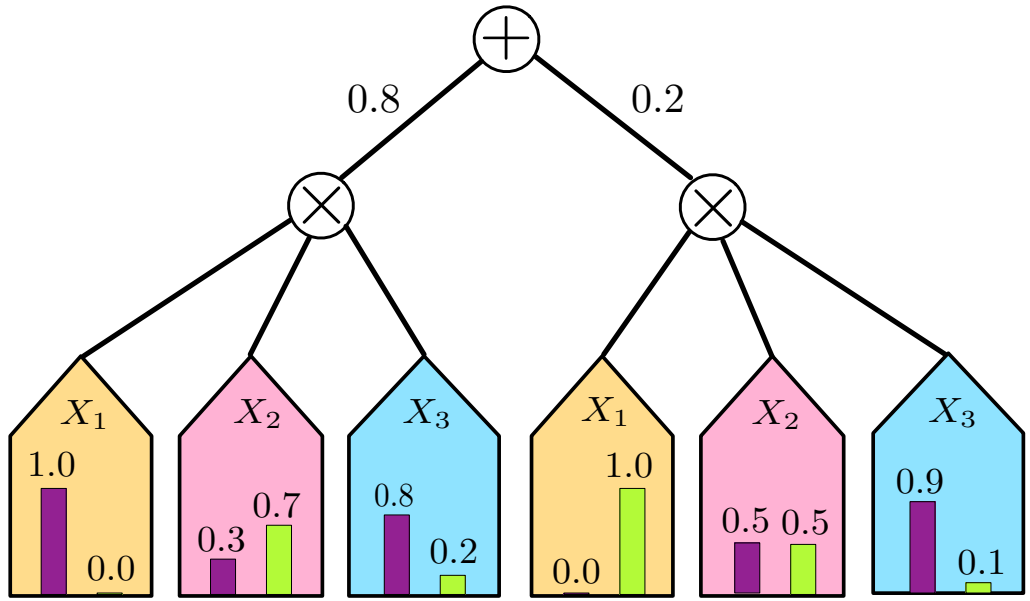
- (1) A *tractable univariate distribution* is SPN.
- (2) A *product* of SPNs with disjoint scopes is SPN.
- (3) A *weighted sum* of SPNs with the same scope is SPN, provided all weights are positive.
- (4) Nothing else is SPN.

Note 1: A univariate distribution is tractable iff its partition function and its mode can be computed in $\mathcal{O}(1)$ time.
Note 2: The *scope* of SPN is the set of variables that appear in it.

Probabilistic Modeling: Sum-product Network

Sum-product Network: Representation

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.

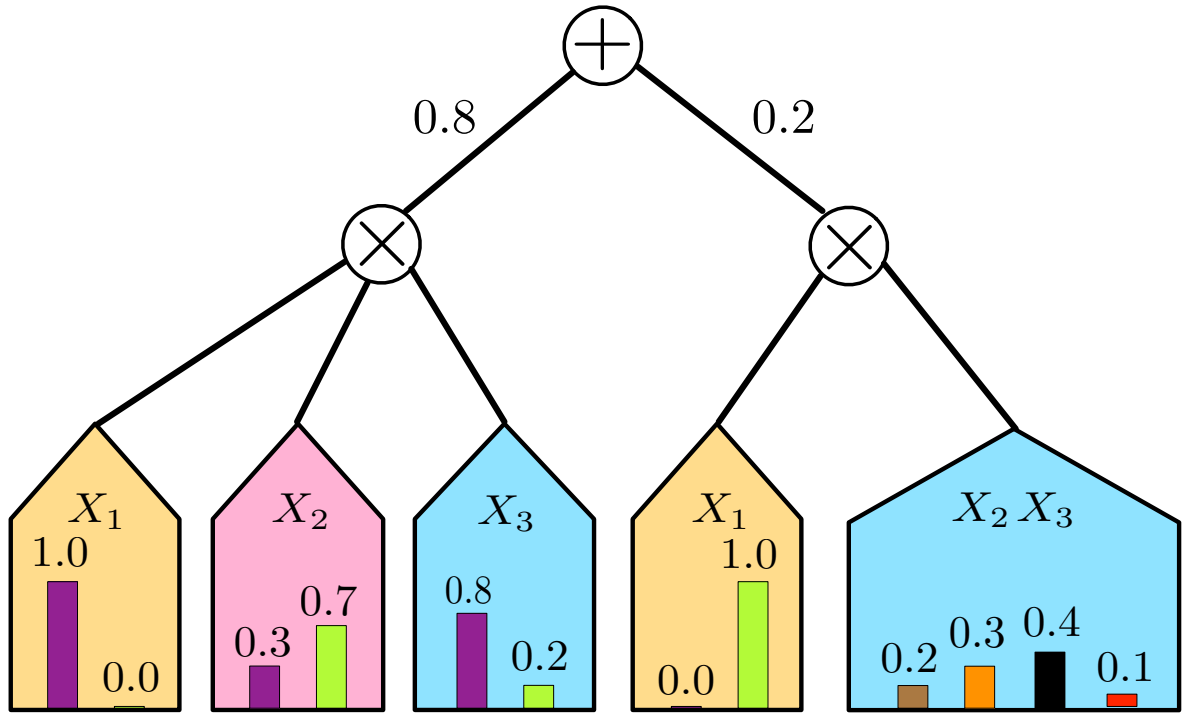


$$P(\mathbf{x}) = 0.8(1.0x_1 + 0.0\bar{x}_1)(0.3x_2 + 0.7\bar{x}_2)(0.8x_3 + 0.2\bar{x}_3) + 0.2(0.0x_1 + 1.0\bar{x}_1)(0.5x_2 + 0.5\bar{x}_2)(0.9x_3 + 0.1\bar{x}_3).$$

Probabilistic Modeling: Sum-product Network

Sum-product Network: Representation

SPN can represent context-specific independence



Outline

The Region of Entropic Vectors $\overline{\Gamma^*_N}$

Enumerating k-atom supports and map to Entropic Region

Non-isomorphic k-atom supports

Maximal Ingleton Violation and the Four Atom Conjecture

Optimizing Entropy Inner Bounds from k-atom Distribution

Characterization of Entropic Region via Information Geometry

Manifold of Probability distributions

e-autoparallel submanifold and m-autoparallel submanifold

Projections and Pythagorean Theorem

Information Geometry of k-atom support

From Conditional Independence to Probabilistic Models

p-representable semimatroids

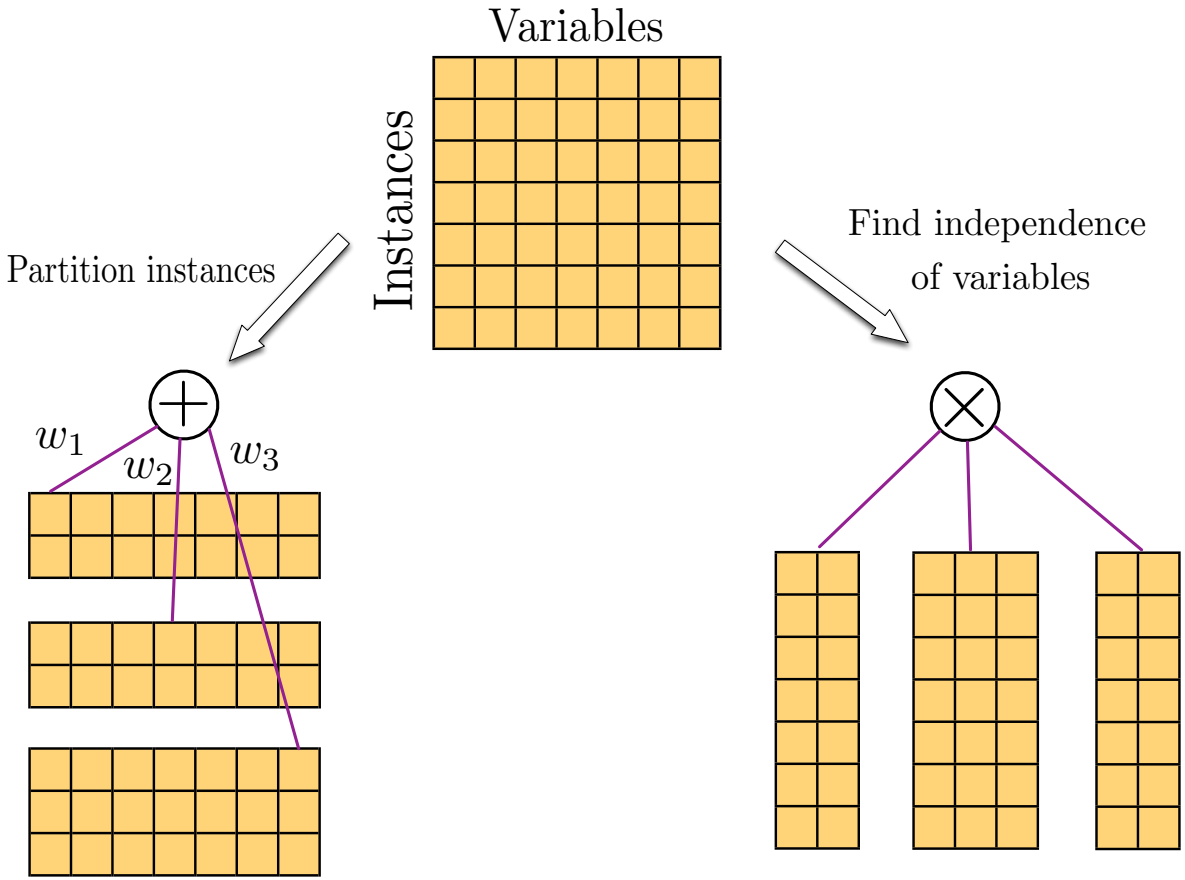
Probabilistic Models

Information guided dataset partitioning for supervised learning

Experiments on Information guided dataset partitioning

Information guided dataset partitioning for supervised learning

What we learned from Structure learning of Sum-product Network



Information guided dataset partitioning

Supervised learning algorithm : generalize from the training data to unseen situations in a ”reasonable” way

Partition data for supervised learning

For supervised learning of discrete dataset, we have N feature variables X_i for $i \in \mathcal{N}$ and label variable Y. Denote $X_{\mathcal{A}}$,$X_{\mathcal{B}}$ as subsets of all variables for $\mathcal{A} \subset \mathcal{N}$ and $\mathcal{B} \subset \mathcal{N}$.

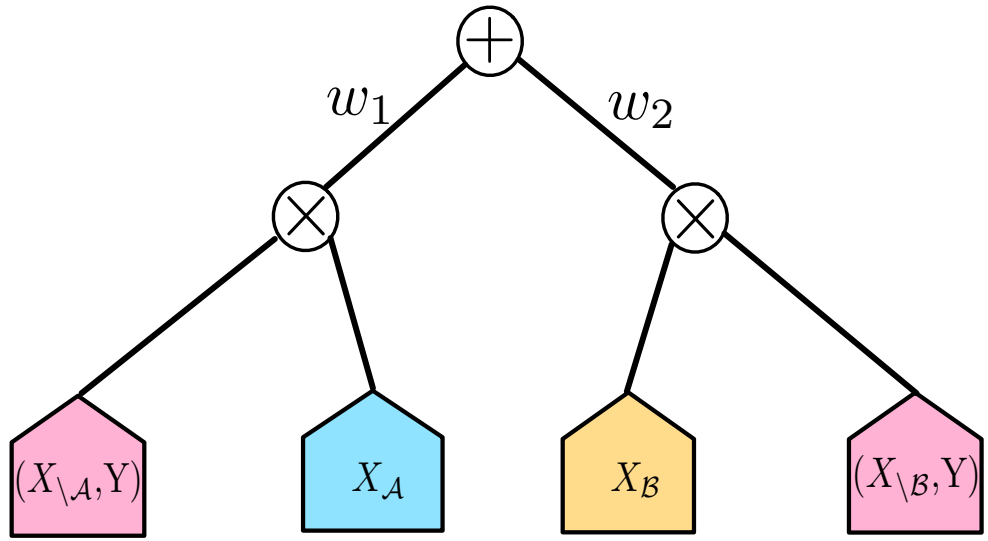


Figure : Context-Specific Independence of a dataset in Sum-Product Network

Information guided dataset partitioning

Literatures on divide and conquer machine learning algorithms: designed to only work on particular machine learning algorithms e.g. Support Vector Machines[7], Ridge Regression[8]...

- 1 : Randomly divide samples evenly and uniformly at random, or divide samples through clustering.
- 2 : Compute local cost function, build model on local partitioned data.
- 3 : Directly merge local solutions together or combined to yield a better initialization for the global problem.

---

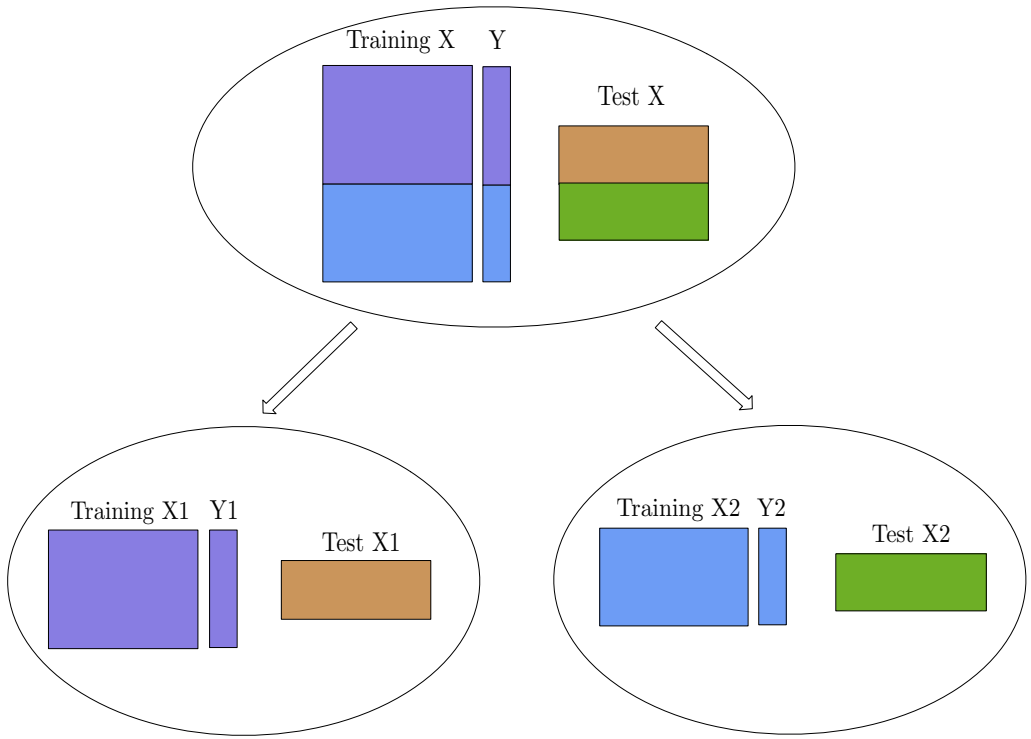
[7] **Cho-Jui Hsieh**, **Si Si**, **Inderjit S. Dhillon**, **A Divid-and-Conquer Solver for Kernel Support Vector Machines**, **ICML 2014**.

[8] **Yuchen Zhang**, **John C. Duchi**, **Martin J. Wainwright**, **Divid and Conquer Kernel Ridge Regression**, **JMLR 2015**.

Information guided dataset partitioning

Partition data for supervised learning

- Task: divide both training and test data into blocks of instances and conquer them completely separately.
- Goal: generate comparable classification performance to the performance of the data what is not partitioned.
- Challenge: huge number of possible partitions



Information guided dataset partitioning

Our motivations to partition data for supervised learning

- Able to run machine learning algorithms on very large dataset in parallel without adding communication overheads.
- Finding multiple partition strategies giving comparable scores, ensembling the different resulting classifiers across different partitioning strategies together to get further performance improvements.

Information guided dataset partitioning for supervised learning

Partition by frequent values of feature variables

1. Frequent values of feature variables as candidate partition criteria
2. Design score functions which calculating entropy and conditional mutual information among X_i for $i \in \mathcal{N}$ and Y
3. Divide dataset based on partitions giving lower scores.

Score function

Consider certain K-partition of data as a hidden variable C taking values of $c_1, c_2, \cdots c_K$.

$$\min_C S(C) \tag{13}$$

Information guided dataset partitioning for supervised learning

Designing score functions with Information measures

minimize $I(X_i; Y|C)$

minimize $I(X_i; Y|C = c_j)$

minimize $I(X_i; Y|C = c_j) - I(X_i; Y)$

larger $H(C)$ to encourage more even partition

larger $H(Y|C)$ to make sure partition not based on value of Y

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

Information guided dataset partitioning for supervised learning

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

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

Further more, we can calculate the sum of k smallest of $\mathcal{N}$ context-specific conditional mutual information

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

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

Information guided dataset partitioning for supervised learning

If only 2-parition is allowed, $C = \{c_1, c_2\}$, following score function can be used where we calculate the root-mean-square error (RMSE) to measure the difference of context-specific conditional mutual information on all feature variables.

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

Outline

The Region of Entropic Vectors $\overline{\Gamma^*_N}$

Enumerating k-atom supports and map to Entropic Region

- Non-isomorphic k-atom supports
- Maximal Ingleton Violation and the Four Atom Conjecture
- Optimizing Entropy Inner Bounds from k-atom Distribution

Characterization of Entropic Region via Information Geometry

- Manifold of Probability distributions
- e-autoparallel submanifold and m-autoparallel submanifold
- Projections and Pythagorean Theorem
- Information Geometry of k-atom support

From Conditional Independence to Probabilistic Models

- *p-representable* semimatroids
- Probabilistic Models
- Information guided dataset partitioning for supervised learning
- Experiments on Information guided dataset partitioning

Experiment 1: Click-Through Rate(CTR) prediction

- Results are tested on one of the online advertising click-through rate(CTR) datasets from Kaggle
- The task is to predict the probability of a user to click an advertisement based on historical training data.
- The training data have 40 million instances, test data have 4 million instances

Partition ID	feature-value	ratio in datasets
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

Table : Feature-value pairs of click-through rate data

Experiment 1: Click-Through Rate(CTR) prediction

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 : Scores of different partitions

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

Table : Classification results of different partitions(in log loss where $logloss(\boldsymbol{y}, \boldsymbol{p}) = -\frac{1}{N}\sum_{i=1}^{N}(y_i \log(p_i) + (1-y_i)\log(1-p_i)))$

Experiment 1: Click-Through Rate(CTR) prediction

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 : Scores of different partitions

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

Table : Classification results of different partitions(in log loss where $logloss(\boldsymbol{y}, \boldsymbol{p}) = -\frac{1}{N}\sum_{i=1}^{N}(y_i \log(p_i) + (1-y_i)\log(1-p_i))$)

Experiment 2: Customer relationship prediction

- Results are tested on KDD Cup 2009 customer relationship prediction dataset
- The task is to predict the probability of telecom company users to buy upgrades or add-ons
- The results can be used to analyze customers' buying preferences, helping the company offer better customer service

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

Table : Frequent feature-value pairs of customer relationship data

Experiment 2: Customer relationship prediction

Partition ID	$S_1(C)$	$S_2(C)$	$S_3(C, 10)$	$S_4(C, 10)$	$S_5(C)$	AUC of GBDT
1	-0.9998	-1.5853	0.8998	-11.1167	-1.2258	0.87083
2	-0.9998	-1.1277	1.4646	-7.2676	-1.0927	0.87056
3	-0.9785	-1.5325	1.3605	-6.4532	-1.0490	0.86630
4	-0.7098	-0.9634	1.6077	-6.6235	-1.0256	0.86911
5	-0.9433	-1.3324	1.7303	-3.8762	-0.9828	0.86821
6	-0.9303	-1.1234	1.0210	-4.7927	-0.9913	0.86748
7	-0.8564	-1.2235	1.2457	-5.5225	-0.9052	0.86759
8	-0.8774	-1.1143	1.5317	-4.7792	-0.9329	0.86257

Table : Scores of different partitions for customer relationship data

- AUC of GBDT: Area Under the ROC Curve(AUC) as evaluation metric, Gradient Boosting Decision Tree as evaluation algorithm
- AUC of original dataset without partitioning: 0.87014
- Ensemble the prediction results from top-4 partition strategies give us AUC of 0.8721

Experiment 2: Customer relationship prediction

Partition ID	$S_1(C)$	$S_2(C)$	$S_3(C, 10)$	$S_4(C, 10)$	$S_5(C)$	AUC of GBDT
1	-0.9998	-1.5853	0.8998	-11.1167	-1.2258	0.87083
2	-0.9998	-1.1277	1.4646	-7.2676	-1.0927	0.87056
3	-0.9785	-1.5325	1.3605	-6.4532	-1.0490	0.86630
4	-0.7098	-0.9634	1.6077	-6.6235	-1.0256	0.86911
5	-0.9433	-1.3324	1.7303	-3.8762	-0.9828	0.86821
6	-0.9303	-1.1234	1.0210	-4.7927	-0.9913	0.86748
7	-0.8564	-1.2235	1.2457	-5.5225	-0.9052	0.86759
8	-0.8774	-1.1143	1.5317	-4.7792	-0.9329	0.86257

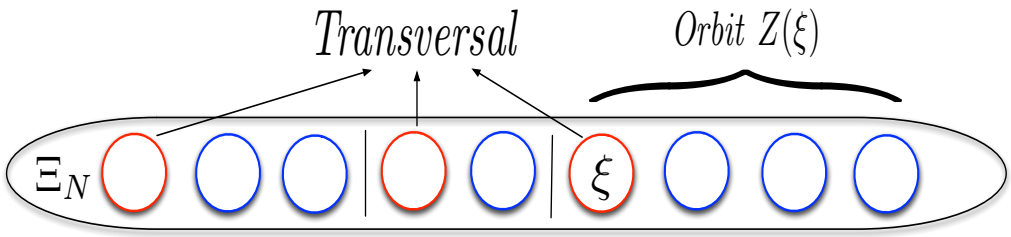
Table : Scores of different partitions for customer relationship data

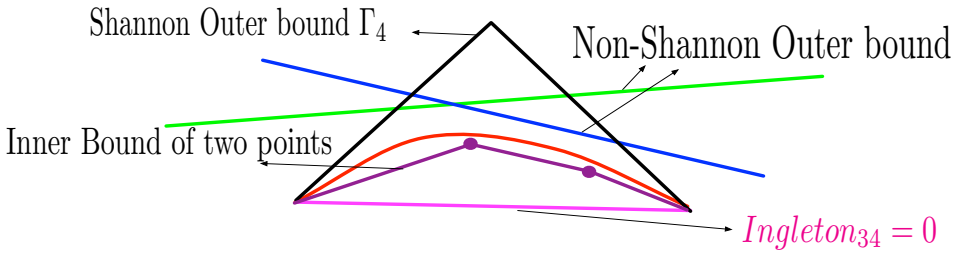
- AUC of GBDT: Area Under the ROC Curve(AUC) as evaluation metric, Gradient Boosting Decision Tree as evaluation algorithm
- AUC of original dataset without partitioning: 0.87014
- Ensemble the prediction results from top-4 partition strategies give us AUC of 0.8721

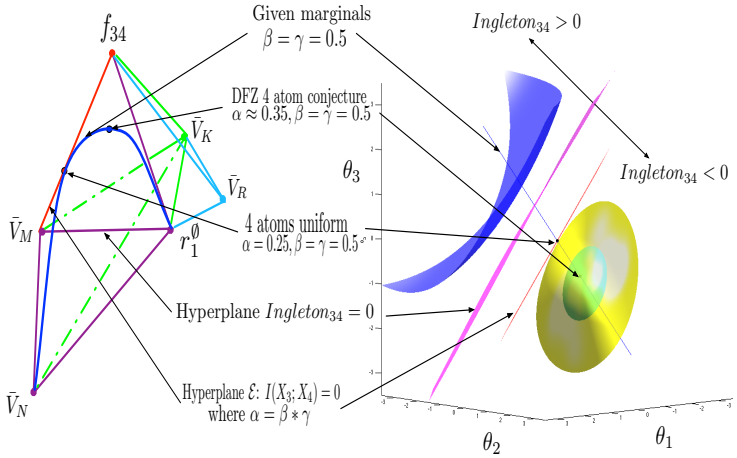
Publications

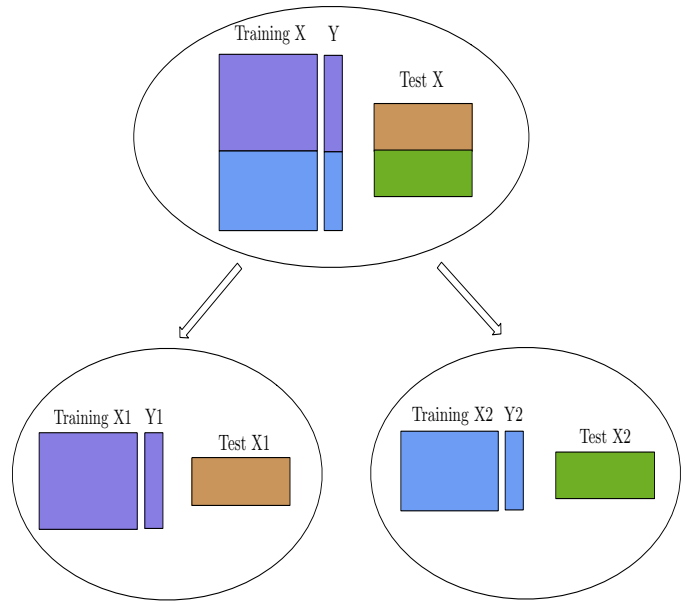
- Yunshu Liu and John MacLaren Walsh, Mapping the Region of Entropic Vectors with Support Enumeration & Information Geometry, IEEE Trans. Inform. Theory, submitted on December 08, 2015.
- Yunshu Liu and John MacLaren Walsh, Information Guided Dataset Partitioning for Supervised Learning, Conference on Uncertainty in Artificial Intelligence, submitted on March 01, 2016.
- Yunshu Liu, John MacLaren Walsh, Only One Nonlinear Non-Shannon Inequality is Necessary for Four Variables, in Proceedings of the 2nd Int. Electronic Conference on Entropy and Its Applications, Nov. 2015
- Yunshu Liu and John MacLaren Walsh, Non-isomorphic Distribution Supports for Calculating Entropic Vectors, in 53nd Annual Allerton Conference on Communication, Control, and Computing, Oct. 2015.
- Yunshu Liu, John MacLaren Walsh, Bounding the entropic region via information geometry, in IEEE Information Theory Workshop, Sep. 2013.

Summary


Transversal
Orbit $Z(\xi)$
Ξ_N
ξ


Shannon Outer bound Γ_4
Non-Shannon Outer bound
Inner Bound of two points
$Ingleton_{34} = 0$


f_{34}
Given marginals
$\beta = \gamma = 0.5$
$Ingleton_{34} > 0$
DFZ 4 atom conjecture
$\alpha \approx 0.35, \beta = \gamma = 0.5$
$\bar{V}_K$
$\bar{V}_R$
θ_3
$Ingleton_{34} < 0$
$\bar{V}_M$
$r_1^{\emptyset}$
4 atoms uniform
$\alpha = 0.25, \beta = \gamma = 0.5$
Hyperplane $Ingleton_{34} = 0$
$\bar{V}_N$
Hyperplane $\mathcal{E}$: $I(X_3; X_4) = 0$
where $\alpha = \beta * \gamma$
θ_2
θ_1


Training X
Y
Test X
Training X1
Y1
Test X1
Training X2
Y2
Test X2

Thanks!

Questions?