Characterizing the Region of Entropic Vectors via Information Geometry

John MacLaren Walsh
Department of Electrical and Computer Engineering
Drexel University
Philadelphia, PA
jwalsh@ece.drexel.edu

Thanks to NSF CCF-1016588 & NSF CCF-1053702.
Outline

• Entropic Vectors Review
  – What Are They?
  – Why are They Important?
    * Unconstrained Importance in Network Coding Capacity Regions
    * Constrained Importance in Multiterminal Information Theory
  – What do we know about them? Open Problems/Issues

• Information Geometry “Review”
  – What is it?
  – Places it has been shown to be useful

• Relating These Two Disciplines
  – A information projection construction of the set of entropic vectors

• Conclusions
Outline

- **Entropic Vectors Review**
  - What Are They?
  - Why are They Important?
    - Unconstrained Importance in Network Coding Capacity Regions
    - Constrained Importance in Multiterminal Information Theory
  - What do we know about them? Open Problems/Issues

- **Information Geometry “Review”**
  - What is it?
  - Places it has been shown to be useful

- **Relating These Two Disciplines**
  - A information projection construction of the set of entropic vectors

- **Conclusions**
Entropic Vectors – What are they?

1. Let $X = (X_1, \ldots, X_N)$ be $N$ discrete random variables with finite support.
2. Let $h(X_A)$ be the entropy of the subset of rvs $X_A = (X_i, i \in A)$ for some non-empty subset $A \subseteq \{1, \ldots, N\} \equiv [N]$.
3. Let $h = (h(X_A), A \subseteq [N])$ be the vector of entropies of each non-empty subset $A \subseteq [N]$. Note $h$ has $2^N - 1$ entries.
   - Example: for $N = 3$, $h = (h_1, h_2, h_3, h_{12}, h_{13}, h_{23}, h_{123})$.
4. A vector $h \in \mathbb{R}^{2^N - 1}$ is called entropic if its elements are the entropies for some joint distribution $p_X$ on the $N$ rvs $X$.
5. The entropy vector region (EVR) $\bar{\Gamma}_N^*$ is the closure of the set of all entropic vectors. It’s a convex cone [1].
6. Normalize by the number of bits for the support $m$: $\tilde{h} = h / \log_2 m$, and define $\bar{\Omega}_N^*$ as the set of normalized entropy vectors [2, 3].
Entropic Vectors – Why are they Important?

- **Network Coding**: Capacity region of a multi-source network under network coding formed from a linear map of $\bar{\Gamma}_N^*$ intersected w/ a series of polyhedral constraints [4]. For every Non-Shannon face there is a network whose capacity region depends on this face [5, 6, 7].

- **Multiterminal Information Theory** More generally, if we allow extra constraints $\mathcal{C}$ on the random variables, then all multiterminal rate regions are expressible in terms of a linear map of $\bar{\Gamma}_N^*(\mathcal{C})$
Entropic Vectors – What do we know? – Outer Bounds

• Yeung & Zhang Non-Shannon [8, 9]
  – Showed that Shannon Outer bound

$$\Gamma_n := \left\{ h \mid h_A + h_B \geq h_{A \cap B} + h_{A \cup B} \quad A \subseteq B \implies h_A \leq h_B \right\}$$

(= matroid rank function cond. [1]) was not tight for $N \geq 4$ via new inequality

• Dougherty, Freiling, & Zeger [10, 11] & Others [12]
  – More Non-Shannon Information Inequalities
  – Construction of Codes via Representable Matroids

• Matùš [13, 14]
  – Showed that $\bar{\Gamma}_4^*$ is not polyhedral

• Technique for creating all of these Non-Shannons [15]:
  – Create one or more R.V.s in terms of the originals (d-copy over), and look at the implications of Shannon inequalities among this larger collection of variables on the subset of original variables.
Entropic Vectors – What do we know? – Inner Bounds

• Matroid Representation Based [16]:
  – Binary matroids: (convex hull of rank functions of) $\forall N$. Not tight $N \geq 4$.
  – Ternary Matroids: (convex hull of rank functions of) $\forall N$.
  – Regular Matroids: (both binary and ternary = rep. over any field)
    Algorithm: check all possible rank functions for spec. forbidden minors, then take convex hull of remaining

• convex hull of representable (over some field) matroids
  – explicitly known only for $N \leq 6$. (4=Ingleton [17, 18, 19], 5,6 new inequalities [20, 21, 22])
  – Not a fully tight inner bound for $N \geq 4$.

• Binary entropic vectors
  – Membership test via a finite terminating numerical algorithm for any $N$ [23, 24, 25, 26].
  – Contains points outside Ingleton (Representable matroids) at $N = 4$.
  – together w/ vertex enumeration can list extreme points of any outer bound which are extreme points of convex hull of binary entropic vectors.
Outline

• Entropic Vectors Review
  – What Are They?
  – Why are They Important?
    * Unconstrained Importance in Network Coding Capacity Regions
    * Constrained Importance in Multiterminal Information Theory
  – What do we know about them? Open Problems/Issues

• Information Geometry “Review”
  – What is it?
  – Places it has been shown to be useful

• Relating These Two Disciplines
  – A information projection construction of the set of entropic vectors

• Conclusions
Information Geometry [27] – What is it? - Notation

• Overall idea: treat family of probability distributions as a differentiable manifold: \( p(x; \xi) \) is parameterized by \( \xi \)

• Endow w/ Riemannian metric (inner product between Tangent vectors) given by Fisher Information Matrix \( g_{i,j}(\xi) = \mathbb{E}_\xi[\partial_i \ell_\xi \partial_j \ell_\xi] \) w/ \( \ell_\xi = \log p(x; \xi) \), \( \partial_i = \frac{\partial}{\partial \xi_i} \).

• Select \( \alpha \)-affine connections \( \nabla^{(\alpha)} \) such that \( \left\langle \nabla^{(\alpha)} \partial_i, \partial_k \right\rangle = \Gamma^{(\alpha)}_{ij,k} \)

\[
\Gamma^{(\alpha)}_{ij,k} = \mathbb{E} \left[ \left( \partial_i \partial_j \ell_\xi + \frac{1 - \alpha}{2} \partial_i \ell_\xi \partial_j \ell_\xi \right) (\partial_k \ell_\xi) \right] \tag{2}
\]

• purpose of affine connection: define parallel translation \( \Pi_{p,p'} : T_p \to T_{p'} \) to correspond tangent vectors along curves \( \gamma : [a,b] \to \mathcal{P} \)

\[
\Pi_{\gamma(t),\gamma(t+dt)}(X(t)) = \sum_{ijk} \left\{ X^k(t) - dt \gamma^i(t) X^j(t) (\Gamma_{ij,k})_{\gamma(t)} \right\} (\partial_k)_{\gamma(t+dt)} \tag{3}
\]

• Curve w/ tangent vector transported by parallel transl. w/ \( \nabla^{(\alpha)} \) is \( \nabla^{(\alpha)} \) geodesic

• If there is a coordinate system in which every parallel translation under \( \nabla^{(\alpha)} \) leaves coefficients in Tangent vector unchanged, the manifold is said to be \( \alpha \)-flat, and associated coordinate system is an affine coordinate system.

• \( \nabla^{(\alpha)} \) has property \( \langle X, Y \rangle_p = \langle \Pi^{(\alpha)}_{p,p'}(X), \Pi^{(-\alpha)}_{p,p'}(Y) \rangle_{p'} \)
Information Geometry [27] – What is it? - Picture - Parallel Translation

\[ \nabla \partial_i \partial_j = \sum_k \Gamma_{i,j,k} \partial_k \quad \Gamma_{i,j,k} = 0 \text{ if “flat”} \]
Information Geometry [27] – What is it? - Picture - Information Projection

\[ D^{(\alpha)}(p_X \mid \mid q) = D^{(\alpha)}(p_X \mid \mid q^*) + D^{(\alpha)}(q^* \mid \mid q) \]
Information Geometry [27] – What is it? - Examples

• 2 flat coordinate systems (associated with $\alpha = -1, 1$) for finite discrete $X \in \mathcal{X} = \{x_0, x_1, x_2, \ldots, x_N\}$
  
  - e-flat: exponential family: $q_1(x), q_2(x) \in \mathcal{E} \implies c(\lambda)q_1^\lambda(x)q_2^{1-\lambda}(x) \in \mathcal{E}$
    
    \[ p_X(x) = \exp \left( \theta^T t(x) - \psi(\theta) \right) \]  

    with $\theta_i = \log \frac{p_X(x_i)}{p_X(x_0)}$, $i \in \{1, \ldots, N\}$, $\psi(\theta) = \log (1 + \| \exp(\theta) \|_1)$
  
  - m-flat: mixture family: $q_1(x), q_2(x) \in \mathcal{M} \implies \lambda q_1(x) + (1 - \lambda)q_2(x) \in \mathcal{M}$
    
    \[ p_X(x) = \eta \cdot t(x) + (1 - 1^T \eta)1_{x=x_0} \quad \eta_i = p_X(x_i) \]  

• Legendre Transform Relationship

• KL Divergence (Relative Entropy)
Information Geometry [27] – Examples, Cont’d

• **e-flat submanifold:** set of all product distributions

\[ \mathcal{E}_0 = \left\{ p_X \mid p_X(x_1, \ldots, x_N) = \prod_{i=1}^{N} p_X(x_i) \right\} \tag{7} \]

• **m-flat submanifold:** set of joint distributions with given marginals

\[ \mathcal{M}_0 = \left\{ p_X \mid \sum_{x \setminus i} p_X(x) = q_i(x_i) \quad \forall i \in \{1, \ldots, N\} \right\} \tag{8} \]

• **Information Projections & Pythagorean Relation:**

\[ q^* = \arg \min_{q \in \mathcal{E}_0} D(p_X \| q), \quad D(p_X \| q) = D(p_X \| q^*) + D(q^* \| q) \quad \forall q \in \mathcal{E}_0 \tag{9} \]

\[ q^* = \arg \min_{q \in \mathcal{M}_0} D(q \| p_X), \quad D(q \| p_X) = D(q^* \| p_X) + D(q \| q^*) \quad \forall q \in \mathcal{M}_0 \tag{10} \]
Information Geometry [27] – What has it been used for?

- re-interpretation of EM algorithm [27]
- acceleration of Blahut Arimoto algorithm [28]
- learning algorithms in Neural Networks [29]
- analysis of Belief propagation & Turbo Decoding [30, 31, 32, 33]
Outline

• Entropic Vectors Review
  – What Are They?
  – Why are They Important?
    * Unconstrained Importance in Network Coding Capacity Regions
    * Constrained Importance in Multiterminal Information Theory
  – What do we know about them? Open Problems/Issues

• Information Geometry “Review”
  – What is it?
  – Places it has been shown to be useful

• Relating These Two Disciplines
  – A information projection construction of the set of entropic vectors

• Conclusions
Relating These – Constructing Entropic Vectors via Information Geometry

Easy to relate Shannon entropy to rel. entropy/ KL Divergence:

\[
D(p_X|\mathcal{U}|\mathcal{X}|) = \sum_{x \in \mathcal{X}} p_X(x) \log_2 \left( \frac{p_X(x)}{1/|\mathcal{X}|} \right) \tag{11}
\]

\[
= \log_2 (|\mathcal{X}|) - H(p_X) = H(\mathcal{U}|\mathcal{X}) - H(p_X) \tag{12}
\]

...perhaps put solution to minimization here as well ..
Next consider the family of distributions

$$H_i := \left\{ p_X \left| p(X) = \frac{1}{|\mathcal{X}_i|} q(X_{\backslash i}), \text{some } q(X_{\backslash i}) \right. \right\}$$

Observe:

- $U_X \in H_i$
- $H_i$ is both an e-flat and m-flat submanifold.
- Defining $q^*_{H_i}(p_X) = \arg \min_{q \in H_i} D(p_X \| q)$, have Pythagorean relation:

$$D(p_X \| U_X) = D(p_X \| q^*_{H_i}(p_X)) + \underbrace{D(q^*_{H_i}(p_X) \| U_X)}_{\log_2 |\mathcal{X}_i| - H(X_i | X_{\backslash i})} \underbrace{\log_2 |\mathcal{X}| - \log_2 |\mathcal{X}_i| - H(X_{\backslash i})}_{(erm... \ H(X) = H(X_i) + H(X_{\backslash i} | X_i) \text{ tyco})}$$

Moving this around, we have

$$H(X_{\backslash i}) = D(p_X \| q^*_{H_i}(p_X)) - D(p_X \| U_X) + \log_2 |\mathcal{X}| - \log_2 |\mathcal{X}_i|$$
Relating These – Constructing Entropic Vectors via Information Geometry

Generalizing this idea, consider the family of distributions

$$\bigcap_{i \in A^c} \mathcal{H}_i = \left\{ p_X = \frac{q(X_A)}{\prod_{i \in A^c} |X_i|} \right\}$$  \hspace{1cm} (16)

Observe:

- $\mathcal{U}X \in \cap_{i \in A^c} \mathcal{H}_i$
- $\cap_{i \in A^c} \mathcal{H}_i$ is both an e-flat and m-flat submanifold
- Defining $q^*_A(p_X) = \arg\min_{q \in \cap_{i \in A^c} \mathcal{H}_i} D(p_X || q)$, have Pythagorean relation:

$$D(p_X || \mathcal{U}X) = D(p_X || q^*_A(p_X)) + D(q^*_A(p_X) || \mathcal{U}X)$$  \hspace{1cm} (17)

(erm... $H(X) = H(X_A) + H(X_{A^c} | X_A)$ tyco)

From which we observe that

$$H(X_A) = D(p_X || q^*_A(p_X)) - D(p_X || \mathcal{U}X) - \sum_{i \in A^c} \log_2 |X_i| + \log_2 |X|$$  \hspace{1cm} (18)
Relating These – Constructing Entropic Vectors via Information Geometry

Defining the set function (then stack into a vector $d$)

$$d_A := \min_{q \in \bigcap_{i \in A^c} \mathcal{H}_i} D(p_X \| q) = D(p_X \| q_A(p_X))$$  \hspace{1cm} (19)

It is evident from the relation we derived

$$H(X_A) = D(p_X \| q^*_A(p_X)) - D(p_X \| U_X) - \sum_{i \in A^c} \log_2 |X_i| + \log_2 |X|$$  \hspace{1cm} (20)

that

$$h_A = d_A - d_{[N]} - \sum_{i \in A^c} \log_2 |X_i| + \log_2 |X|$$  \hspace{1cm} (21)

thus we can express entropic vector in terms of $d$ via

$$h(d) = Ad + b$$  \hspace{1cm} (22)

Region of entropic vectors is affine transformation of region of simultaneous divergences between submanifolds $\mathcal{H}_i$ and their intersections!
References


