Entropic Graphs

Alfred O. Hero
Dept. EECS, Dept Biomed. Eng., Dept. Statistics
University of Michigan - Ann Arbor
hero@eecs.umich.edu
http://www.eecs.umich.edu/~hero

Collaborators: H. Heemuchwala, J. Costa, B. Ma, O. Michel

• Motivating Applications
• Entropic Euclidean Graphs
• A variant: $K$-point entropic graphs
• Application to US image registration
Pattern Matching and Image Retrieval

QUERY

DATABASE
Adaptive radar sensor management

Figure 1: SAR clutter image, target on boundary at column 305.
Figure 2: A multidate image registration example


**Objective:** For given fitness criterion $Q$, find operator $T$ which minimizes/maximizes $Q$

**Our focus:** entropic fitness criterion $Q(f)$

$f$: feature density over $x \in \mathbb{R}^d$
Some Popular Entropic Q’s

1. Shannon Entropy of feature density $f$

$$Q(f) = H(f) = - \int f(x) \ln f(x) \, dx$$

2. KL Divergence between feature densities $f, g$

$$Q(f, g) = D(f||g) = \int f(x) \ln \left( \frac{f(x)}{g(x)} \right) \, dx$$

3. Jensen difference between feature densities $f, g$:

$$Q(f, g) = H(\varepsilon f + (1 - \varepsilon)g) - \varepsilon H(f) - (1 - \varepsilon)H(g)$$

4. Mutual information within joint feature density $f_{X,Y}$

$$Q(f_{X,Y}) = MI(X, Y) = \int \int f_{X,Y}(x,y) \ln \left( \frac{f_{X,Y}(x,y)}{f_X(x)f_Y(y)} \right) \, dx$$
**Issue**: How to estimate entropic $Q$ from measured data?

Some possibilities:

1. Assume parameteric models for $f, g, f_{X,Y}$
2. Quantize feature space and use histograms
3. Non-parameteric density estimation of $f, g, f_{X,Y}$

**Our Strategy**: construct “entropic graphs” on features
A Set of Feature Samples and a Euclidean Spanning Graph

128 random samples

MST
Minimal Euclidean Graphs: MST

Let $T_n = T(\mathcal{X}_n)$ denote the possible sets of edges in the class of acyclic graphs spanning $\mathcal{X}_n$ (spanning trees).

The Euclidean Power Weighted MST achieves

$$L_{\gamma}^{\text{MST}}(\mathcal{X}_n) = \min_{T_n} \sum_{e \in T_n} \|e\|^\gamma.$$
Minimal Euclidean graphs: $k$-NNG

Let $\mathcal{N}_{k,i}(X_n)$ denote the possible sets of $k$ edges connecting point $x_i$ to all other points in $X_n$.

The Euclidean Power Weighted $k$-NNG is

$$L_{\gamma}^{k-NNG}(X_n) = \sum_{i=1}^{n} \min_{\mathcal{N}_{k,i}(X_n)} \sum_{e \in \mathcal{N}_{k,i}(X_n)} |e|^\gamma$$
MST for Two Different Samples

Figure 3:
Large $n$ behavior of MST

Figure: MST and log MST weights as function of the number of samples.
Asymptotics: the BHH Theorem

Define the MST length functional

\[ L_\gamma(X_n) = \min_{T_n} \sum_{e \in T_n} \|e\|^{\gamma}. \]

**Theorem 1** [Beardwood, Halton&Hammersley:1959] Let \( X_n = \{X_1, \ldots, X_n\} \) be an i.i.d. realization from a Lebesgue density \( f \) with support \( S \subset [0, 1]^d. \)

\[
\lim_{n \to \infty} \frac{L_\gamma(X_n)}{n^{(d-\gamma)/d}} = \beta_{L_\gamma,d} \int_S f(x)^{(d-\gamma)/d} dx, \quad (a.s.)
\]

Or, letting \( \alpha = (d-\gamma)/d \)

\[
\frac{1}{1 - \alpha} \ln \frac{L_\gamma(X_n)}{n^\alpha} \to H_\alpha(f) + c \quad (a.s.)
\]
Rényi Entropy and Divergence

- Rényi Entropy of order $\alpha$ [Rényi:61,70]

$$H_\alpha(f) = \frac{1}{1-\alpha} \ln \int_S f^\alpha(x) dx$$

- Rényi $\alpha$-divergence of fractional order $\alpha \in [0, 1]$

$$D_\alpha(f_1 \parallel f_0) = \frac{1}{\alpha-1} \ln \int_S f_0 \left( \frac{f_1}{f_0} \right)^\alpha dx$$

$$= \frac{1}{\alpha-1} \ln \int_S f_1^\alpha f_0^{1-\alpha} dx$$

- $\alpha$-Divergence vs. Kullback-Liebler divergence

$$\lim_{\alpha \to 1} D_\alpha(f_1 \parallel f_0) = \int f_1 \ln \frac{f_1}{f_0} dx.$$
Clustering via K-MST

Assume $f$ is a mixture density of the form

$$f = (1 - \epsilon)f_1 + \epsilon f_o,$$

where

- $f_o$ is a known (uniform) outlier density
- $f_1$ is an unknown target density
- $\epsilon \in [0, 1]$ is unknown mixture parameter
$K$-point Minimal Spanning Tree ($K$-MST)

Figure 4: Clustering an annulus density from uniform noise via $k$-MST.
Figure 5: Left: k-MST curve for 2D annulus density with addition of uniform “outliers” has a knee in the vicinity of $n - k = 35$. 
Greedy partitioning approximation to K-MST

Figure 6: A smallest subset $B_k^m$ is the union of the two cross hatched cells shown for the case of $m = 5$ and $k = 17$.
Extended BHH Theorem for Greedy K-MST

Fix $\rho \in [0, 1]$. If $k/n \to \rho$ then the length of the greedy partitioning K-MST satisfies [Hero&Michel:IT99]

$$L_\gamma(X_{n,k}^*)/(|\rho n|)^\alpha \to \beta_{L_\gamma,d} \min_{A: \int_A f \geq \rho} \int_S f^\alpha(x|x \in A) \, dx \quad (a.s.)$$

or, alternatively, with

$$H_\alpha(f|x \in A) = \frac{1}{1 - \alpha} \ln \int_S f^\alpha(x|x \in A) \, dx$$

$$\frac{1}{1 - \alpha} \ln L_\gamma(X_{n,k}^*)/(|\rho n|)^\alpha \to \beta_{L_\gamma,d} \min_{A: \int_A f \geq \rho} H_\alpha(f|x \in A) \quad (a.s.)$$
Figure 7: Waterpouring contraction of minimum entropy density.
**k-MST Influence Function**

Figure 8: MST and k-MST influence curves for Gaussian density on the plane.
Extension of BHH to Divergence Estimation?

Question: How to go from

\[ \frac{1}{1 - \alpha} \ln \int f^\alpha(x) \, dx \to \frac{1}{\alpha - 1} \ln \int f^\alpha(x) g^{1-\alpha}(x) \, dx \]

- \( g(x) \): a reference density on \( \mathbb{R}^d \)
- Assume \( f \ll g \), i.e. for all \( x \) such that \( g(x) = 0 \) we have \( f(x) = 0 \).
- Make measure transformation \( M(x) \) such that \( dx \to g(x) \, dx \) on \([0, 1]^d\).

Then for \( \mathcal{Y}_n = M(\mathcal{X}_n) \)

\[ L_\gamma(\mathcal{Y}_n)/n^\alpha \to \beta_{L_\gamma,d} \int \left( \frac{f(x)}{g(x)} \right)^\alpha g(x) \, dx, \quad (a.s.) \]
Figure 9: Top Left: i.i.d. sample from triangular distribution, Top Right: exact transformation, Bottom: after application of exact and empirical transformations.
Clustering Example

\( \mathcal{X}_n \) is a sample from the mixture

\[
f(x) = (1 - \varepsilon)g(x) + \varepsilon h(x)
\]

\( h(x) \) is uniform density on \([0, 1]^2\)

\( g(x) \) is triangular density on \([0, 1]^2\)

\( \varepsilon \) is unknown

Objective: Detect deviation of \( f \) from triangular and cluster the uniform variates in the sample
Figure 10: Left: A sample from triangle-uniform mixture density with $\varepsilon = 0.9$ in the transformed domain $\gamma_n$. Right: ROC curves of thresholded $\alpha$-divergence test for deviation from $g$. Curves are decreasing in $\varepsilon$ over the range $\varepsilon \in \{0.1, 0.3, 0.5, 0.7, 0.9\}$.
Figure 11: Left: the K-MST implemented on the transformed scatterplot $\mathcal{Y}_n$ with $k = 230$. Right: same K-MST displayed in the original data domain.
Bounds on Minimax Convergence Rate

**Theorem 2 (Hero, Costa & Ma 2001)** Let $d \geq 2$ and $1 \leq \gamma \leq d - 1$. Assume $X_1, \ldots, X_n$ are i.i.d. random vectors over $[0, 1]^d$ with density $f \in \Sigma_d(\beta, l)$, $\beta, l > 0$, having support $S \subset [0, 1]^d$. Assume also that $f^{\frac{1}{2} - \frac{\gamma}{d}}$ is integrable. Then,

$$O \left(n^{-r_1(d, \beta)}\right) \leq \sup_{f \in \Sigma_d(\beta, l)} E \left[ \left\| L_{\gamma}(X_1, \ldots, X_n)/n^{(d-\gamma)/d} - \beta L_{\gamma,d} \int_S f^{(d-\gamma)/d}(x) dx \right\|^p \right]^{1/p} \leq O \left(n^{-r_2(d, \gamma)}\right),$$

where

$$r_1(d, \beta) = \min\{\frac{4\beta}{4\beta + d}, 1/2\} \quad r_2(d, \gamma) = \frac{\alpha}{\alpha + 1} \frac{1}{d}$$

and $\alpha = \frac{d-\gamma}{d}$. 

27
Extension to Partition Approximations

\[ L^m_\gamma (X_n) = \sum_{i=1}^{m^d} L_\gamma (X_n \cap Q_i) + b(m), \]

Figure 12: Partition approximation.
Theorem 3 (Hero, Costa & Ma 2001) Let $L^m_\gamma(X_n)$ be a partition approximation to $L_\gamma(X_n)$. Under the same hypotheses as in the previous proposition, if $b(m) = O(m^{d-\gamma})$

$$O\left(n^{-r_1(d,\beta)}\right) \leq \sup_{f \in \Sigma_d(\beta,l)} E \left[\left| L^m_\gamma(X_1, \ldots, X_n) / n^{(d-\gamma)/d} - \beta L_\gamma \int_S f^{(d-\gamma)/d}(x) dx \right|^p \right]^{1/p} \leq O\left(n^{-r_3(d,\gamma)}\right),$$

where

$$r_3(d, \gamma) = \frac{\alpha}{d-1} \frac{1}{\frac{\gamma}{\alpha} + 1} \frac{1}{d}.$$
Figure 13: Three ultrasound breast scans. From top to bottom are: case 151, case 142 and case 162.
MI Registration of Gray Levels (Viola&Wells:ICCV95)

- $X$: a $N \times N$ image (lexicographically ordered)
- $X(k)$: image gray level at pixel location $k$
- $X_0$ and $X_1$: primary and secondary images to be registered

**Hypothesis:** $\{(X_0(k),X_i(k))\}_{k=1}^{N^2}$ are i.i.d. r.v.’s with j.p.d.f

$$f_{0,i}(x_0,x_1), \quad x_0, x_1 \in \{0, 1, \ldots, 255\}$$

**Mutual Information (MI) criterion:** $T = \arg\max_T \hat{M}I$

where $\hat{M}I$ is an estimate of

$$M{I}(f_{0,i}) = \int \int f_{0,i}(x_0,x_1)\ln f_{0,i}(x_0,x_1)/(f_0(x_0)f_i(x_1))dx_1dx_0. \quad (1)$$
Figure 14: Single Pixel Coincidences (Left and right: reference image $I^R$ at $0^o$ and rotated image $I^T$ at $8^o$)
Figure 15: Grey level scatterplots. 1st Col: target=reference slice. 2nd Col: target = reference+1 slice.
Higher Level Features

Disadvantages of single-pixel features:

- Only depends on histogram of single pixel pairs
- Insensitive to spatial reording of pixels in each image
- Difficult to select out grey level anomalies (shadows, speckle)
- Spatial discriminants fall outside of single pixel domain

- **Alternative**: Aggregate spatial features
(a) Image $I^R$  \hspace{1cm} (b) Image $I^T$

Figure 16: Local Tag Coincidences
Generalization: \(\alpha\)-MI Registration of Coincident Features

- \(X\): a \(N \times N\) US image (lexicographically ordered)
- \(Z = Z(X)\): a general image feature vector in a \(d\)-dimensional feature space

Let \(\{Z_0(k)\}_{k=1}^K\) and \(\{Z_i(k)\}_{k=1}^K\) be features extracted from \(X_0\) and \(X_i\) at \(K\) pairs of identical spatial locations

\(\alpha\)-MI coincident-feature criterion

\[
T = \text{argmax}_T \hat{\text{MI}}_\alpha
\]

where \(\hat{\text{MI}}_\alpha\) is an estimate of

\[
\text{MI}_\alpha(f_{0,i}) = \frac{1}{\alpha - 1} \log \int \int f_{0,i}^\alpha(z_0, z_1) f_{0}^{1-\alpha}(z_0) f_{i}^{1-\alpha}(z_1) dz_1 dz_0. \tag{2}
\]
**α-MI and Decision Theoretic Error Exponents**

\[ H_0 : \ Z_0(k), Z_i(k) \ \text{independent} \]

\[ H_1 : \ Z_0(k), Z_i(k) \ \text{o.w.} \]

Bayes probability of error

\[ P_e(n) = \beta(n)P(H_1) + \alpha(n)P(H_0) \]

Chernoff bound

\[ \liminf_{n \to \infty} \frac{1}{n} \log P_e(n) = - \sup_{\alpha \in [0,1]} \{(1 - \alpha)\text{MI}_\alpha(f_{0,i})\} \]
ICA Features

Decomposition of $M \times M$ tag images $Y(k)$ acquired at $k = 1, \ldots, K$ spatial locations

$$Y(k) = \sum_{p=1}^{P} a_{kp}S_p$$

- $\{S_k\}_{k=1}^{P}$: statistically independent components
- $a_{kp}$: projection coefficients of tag $Y(k)$ onto component $S_p$
- $\{S_k\}_{k=1}^{P}$ and $P$: selected via FastICA
- Feature vector for coincidence processing:

$$Z(k) = [a_{k1}, \ldots, a_{kP}]^T$$
ICA feature basis for US breast images

Figure 17: Estimated ICA basis set using FastICA
**Simpler Objective Function: $\alpha$-Jensen Difference**

1. Extract features from reference and transformed target images:

   $$X_m = \{X_i\}_{i=1}^m \quad \text{and} \quad \mathcal{Y}_n = \{Y_i\}_{i=1}^n$$

2. Construct following MST function on $X_m$ and $\mathcal{Y}_n$

   $$\Delta L = \ln L_\gamma(X_m \cup \mathcal{Y}_n) / (n+m)^\alpha - \frac{m}{n+m} \ln L_\gamma(X_m) / m^\alpha - \frac{n}{n+m} \ln L_\gamma(\mathcal{Y}_n) / n^\alpha$$

3. Minimize $\Delta L_\gamma$ over transformations producing $\mathcal{Y}_n$.

   $$(1 - \alpha)^{-1} \Delta L \rightarrow H_\alpha(\varepsilon f_x + (1 - \varepsilon) f_y) - \varepsilon H_\alpha(f_x) - (1 - \varepsilon) H_\alpha(f_y)$$

   where $\varepsilon = \frac{m}{m+n}$
Figure 18: MST demonstration for misaligned images
Figure 19: MST for aligned images. “x” denotes reference while “o” denotes a candidate image in the DEM database.
Quantitative Performance Comparisons for US Registration

Effect of Additive Noise on peak of objective function

- alphaJensen Diff MST on 8D–ICA
- alphaMI on single pixels w/ Hitograms
- alphaJensen on 8D–ICA using Histograms
- alphaJensen on single pixels w/ MST

Figure 20: US registration MSE comparisons.
Conclusions

1. Entropic graphs can be used to estimate $\alpha$-entropy and $\alpha$-divergence

2. MST and k-NN applied to high dimensional feature-based image registration

3. Clustering using entropic $K$-point graphs

4. Extensions to larger class of continuous quasi-additive graphs (Yukich)