The Geometry of Distributions II

Information Distances: Properties And Algorithms

Suresh Venkatasubramanian
University of Utah
Lecture Plan

- Information geometry
- Algorithms for information distances
- Spatially-aware information distances
Bregman divergence For convex $\phi : \mathbb{R}^d \rightarrow \mathbb{R}$

$$D_\phi(p, q) = \phi(p) - \phi(q) - \langle \nabla \phi(q), p - q \rangle$$

f-divergence For convex $f : \mathbb{R} \rightarrow \mathbb{R}, f(1) = 0$,

$$D_f(p, q) = \sum_i p_i f\left( \frac{q_i}{p_i} \right)$$

$\alpha$-divergence For $|\alpha| < 1$,

$$D_\alpha(p, q) = \frac{4}{1 - \alpha^2} \left[ 1 - \int p^{(1-\alpha)/2} q^{(1+\alpha)/2} \right]$$
Bregman divergence For convex $\phi : \mathbb{R}^d \to \mathbb{R}$

$$D_{\phi}(p, q) = \phi(p) - \phi(q) - \langle \nabla \phi(q), p - q \rangle$$

$\alpha$-divergence For $|\alpha| < 1$,

$$D_{\alpha}(p, q) = \frac{4}{1 - \alpha^2} [1 - \int p^{(1-\alpha)/2} q^{(1+\alpha)/2}]$$
Bregman Divergences
Properties: Asymmetry

\[ D_\phi(p, q) = \phi(p) - \phi(q) - \langle \nabla \phi(q), p - q \rangle \]

In general,

\[ D_\phi(p, q) \neq D_\phi(q, p) \]

However, let

\[ \phi^*(u) = \sup_p \langle u, p \rangle - \phi(p) \]

then

\[ D_\phi(p, q) = D_{\phi^*}(q^*, p^*) \]

where \( x^* = \nabla \phi(x) \)
Generalized Cosine Rule:

\[ D_\phi(p, q) + D_\phi(q, r) = D_\phi(p, r) + \langle \nabla \phi(r) - \nabla \phi(q), p - q \rangle \]

Compare with standard cosine rule:

\[ \frac{\|p - q\|^2}{2} + \frac{\|q - r\|^2}{2} = \frac{\|p - r\|^2}{2} + \langle r - q, p - q \rangle \]

- In general, \( D_\phi(p, q) \) behaves like square of a distance ("energy", not "distance")
- \( \sqrt{D_\phi(p, q)} \) is not a metric in general.
- \( D_\phi(p, q) \) is not a metric.
Bregman bisectors are halfplanes

\[ D_\phi(r, p) = D_\phi(r, q) \]
\[ \langle r, \nabla \phi(q) - \nabla \phi(p) \rangle = f(p, q) \]

Most combinatorial geometry carries over (Voronoi, Delaunay, arrangements, duality)[BNN10]
Euclidean Approximations

\[ D_\phi(p, q) = \phi(p) - \phi(q) - \langle \nabla \phi(q), p - q \rangle \]
Euclidean Approximations

\[ D_\phi(p, q) = \phi(p) - \phi(q) - \langle \nabla \phi(q), p - q \rangle \]

\[ D_\phi(p, q) = s^\top \nabla^2 \phi(r) s, s = p - q, r \in [p, q] \]
Euclidean Approximations

\[ D_\phi(p, q) = \phi(p) - \phi(q) - \langle \nabla \phi(q), p - q \rangle \]
\[ = s^\top \nabla^2 \phi(r)s, s = p - q, r \in [p, q] \]

As \( q \to p \), this becomes the Mahalanobis distance, for positive definite \( A \):

\[ \| p - q \|_A = \sqrt{(p - q)^\top A(p - q)} \]
\[
\lambda^*_{\text{max}} = \max_{x \in R} \lambda_{\text{max}}(\nabla^2 \phi(x)) \\
\lambda^*_{\text{min}} = \min_{x \in R} \lambda_{\text{min}}(\nabla^2 \phi(x)) \\
\mu = \frac{\lambda^*_{\text{max}}}{\lambda^*_{\text{min}}}
\]
A distance measure is $\mu$-similar if

$$d(a,b) + d(b,c) \geq \frac{1}{\mu} d(a,c)$$

**Lemma**

*For a given convex function $\phi$, there exists a bounded region $R(\mu)$ in which $D_\phi$ is $\mu$-similar.*

Proof follows by exploiting the connection to the Mahalanobis distance, and looking at the variation in eigenvalues of $\nabla^2 \phi$.

Euclidean clustering algorithm $\Rightarrow$ $\mu$-similar Bregman clustering algorithm. [ABS10, AB10, ABS11, CM08, MR09]
Dimensionality Reduction
Dimensionality Reduction For Distributions [APV10, KPV10]

**Problem**

*Given n points on the simplex $\Delta_d$, find a mapping to $\Delta_k$, $k \ll d$ such that the "distance structure" of the points is preserved.*
The Hellinger Distance

\[ d_H^2(p, q) = \sum_i (\sqrt{p_i} - \sqrt{q_i})^2 \]

Hellinger distance:
- Natural geometric interpretation
- Local approximation of Fisher information, and closely related to Jensen-Shannon and Bhattacharya distance.
Hellinger Is Chordal Distance on Sphere

\[(p_1, p_2, \ldots, p_d) \rightarrow (\sqrt{p_1}, \sqrt{p_2}, \ldots, \sqrt{p_d})\]

Hellinger distance on simplex $\leftrightarrow$ Euclidean distance on sphere. Dimensionality reduction for simplex $\leftrightarrow$ Dimensionality reduction on “sphere”
The Plan
Let $f : \Delta_d \to \Delta_k$. Given set $X = \{p_1, \ldots, p_n\} \subset \Delta_d$, and $\epsilon > 0$

$$d_H(f(p_i), f(p_j)) \leq d_H(p_i, p_j) \leq (1 + \epsilon)d_H(f(p_i), f(p_j))$$

for all $i, j$.

- This measures “distortion”: the ratio of distances.
- Bounds are worst-case.
Preserving Distances Approximately

Problem

Given $n$ points on the simplex $\Delta_d$, find a mapping $f : \Delta_d \rightarrow \Delta_k$ such that

- $k$ is “small”
- All distances are preserved approximately to within $1 + \epsilon$. 
The Plan
The Plan
Theorem ([JL84])

Given $n$ points in $\mathbb{R}^d$ and $\epsilon > 0$, there exists a mapping $f : \mathbb{R}^d \to \mathbb{R}^k$, with $k = O(\log n / \epsilon^2)$ such that all distances are approximately preserved.
Theorem ([JL84])

Given $n$ points in $\mathbb{R}^d$ and $\epsilon > 0$, there exists a mapping $f : \mathbb{R}^d \rightarrow \mathbb{R}^k$, with $k = O(\log n / \epsilon^2)$ such that all distances are approximately preserved.

Theorem

Given $n$ points on $S^d$ and $\epsilon > 0$, there exists a mapping $f : S^d \rightarrow S^k$, with $k = O(\log n / \epsilon^2)$ such that all distances are approximately preserved.
Algorithm

1. Generate a random projection matrix $A \in M^{k \times d}$ and apply to $p$:

$$
\begin{pmatrix}
\hat{p}
\end{pmatrix} = 
\begin{pmatrix}
\cdots & A & \cdots
\end{pmatrix}
\begin{pmatrix}
p
\end{pmatrix}
$$

2. Normalize the resulting vector:

$$f(p) \triangleq \frac{\hat{p}}{||\hat{p}||}$$
Proof Idea I

- Projection to Euclidean space preserves distances approximately:
  \[ d(\hat{p}_i, \hat{p}_j) \approx d(p_i, p_j) \]

- Norms are preserved:
  \[ \|p\| = d(p, 0) \]

- Normalizing vectors introduces error:
Proof Idea II

- Close-by points are not distorted too much.
- Projection also preserves norms (vectors aren’t too long)
- “Simulate” addition of extra points to ensure that long distances are preserved.
The Plan
The Plan
(p_1, p_2, \ldots, p_d) \rightarrow (\sqrt{p_1}, \sqrt{p_2}, \ldots, \sqrt{p_d})

Mapping from simplex to sphere only invertible for points in positive orthant!
Need projection that

- Preserves all distances approximately
- Takes points on positive orthant to points on (lower-dimensional) positive orthant.
A Restricted Result

Fix center $c = (1/d, 1/d, \ldots, 1/d)^\top$.

$$P = \{ x \in \mathbb{S}_d^+ \mid \langle x, c \rangle \geq \sqrt{\frac{d-1}{d}} \}$$

$$R = \{ r \in \mathbb{S}^d \mid \langle r, c \rangle = \frac{1}{\sqrt{d}} \}$$
A Restricted Result

Fix center $c = (1/d, 1/d, \ldots, 1/d)^\top$.

$$P = \{ x \in S^+_d \mid \langle x, c \rangle \geq \sqrt{\frac{d-1}{d}} \}$$

$$\mathcal{R} = \{ r \in S^d \mid \langle r, c \rangle = \frac{1}{\sqrt{d}} \}$$
If all points lie in $P$, and projection vectors are chosen randomly from $\mathcal{R}$, then all projected points lie in positive orthant and distances are preserved.
Proof Idea

Lemma

If $r \in \mathbb{R}$ and $u$ is any unit vector, then $u \cdot r$ is "subgaussian": if $T \sim \mathcal{N}(0, 1/d)$, then

$$E[(u \cdot r)^{2m}] \leq E[T^{2m}]$$

Coupled with a result due to Achlioptas[Ach03]:

Lemma

If $r$ is as above, then picking $r_1, \ldots, r_k$ randomly and projecting onto these preserves distances approximately.
Challenges
Bregman Near Neighbors[Cay08, ZOPT09]

Need ball packing bounds as well as relaxed triangle inequality for approximation.
Bregman Near Neighbors [Cay08, ZOPT09]

Need ball packing bounds as well as relaxed triangle inequality for approximation.
Bregman Near Neighbors [Cay08, ZOPT09]

Need ball packing bounds as well as relaxed triangle inequality for approximation.
Need ball packing bounds as well as relaxed triangle inequality for approximation.
Jensen-Shannon Dimensionality Reduction

\[ d_{JS}(p, q) = (1/2) \sum_i p_i \log \frac{2p_i}{p_i + q_i} + q_i \log \frac{2q_i}{p_i + q_i} \]

- This is known to be the square of a metric (by reproducing kernel arguments) [BF].
- There exists positive definite kernel \( K \) such that

\[ d_{JS}(p, q) = K(p, p) + K(q, q) - 2K(p, q) \]

- Can we do dimensionality reduction? (see next lecture)
Spatial Sensitivity

Stay Tuned!
Spatial Sensitivity

Stay Tuned!
Marcel R. Ackermann and Johannes Blömer. 
Bregman clustering for separable instances. 
In Haim Kaplan, editor, SWAT, volume 6139 of Lecture 

Marcel R. Ackermann, Johannes Blömer, and Christian 
Sohler. 
Clustering for metric and nonmetric distance measures. 

Marcel R. Ackermann, Johannes Blömer, and Christoph 
Scholz. 
Hardness and non-approximability of bregman clustering 
problems. 
Electronic Colloquium on Computational Complexity (ECCC), 
18:15, 2011.


Rasmus J. Kyng, Jeff M. Phillips, and Suresh Venkatasubramanian. 
Johnson-Lindenstrauss dimensionality reduction on the simplex. 
In 20th Fall Workshop on Computational Geometry, October 2010.

Bodo Manthey and Heiko Röglin. 
Worst-case and smoothed analysis of -means clustering with bregman divergences. 
Zhenjie Zhang, Beng Chin Ooi, Srinivasan Parthasarathy, and Anthony K. H. Tung.
Similarity search on bregman divergence: Towards non-metric indexing.