Generalized similarity measures for text data.

Hubert Wagner (IST Austria)
Joint work with Herbert Edelsbrunner

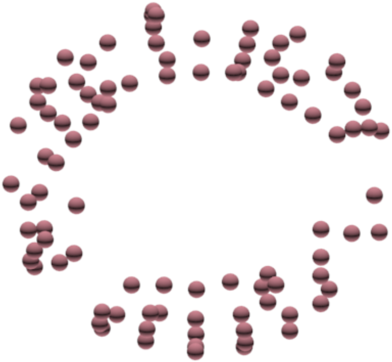
GETCO 2015, Aalborg

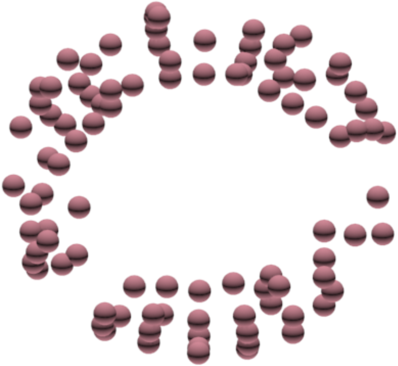
April 9, 2015

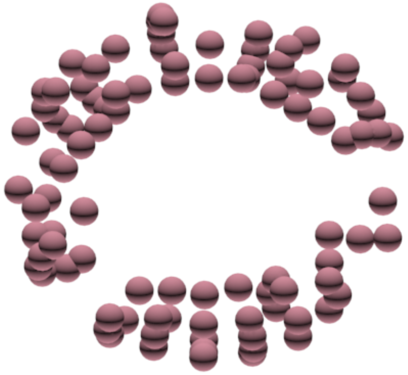
Plan

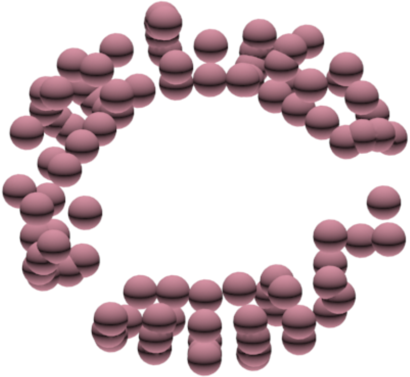
- Shape of data.
- Text as a point-cloud.
- Log-transform and similarity measure.
- Bregman divergence and topology.

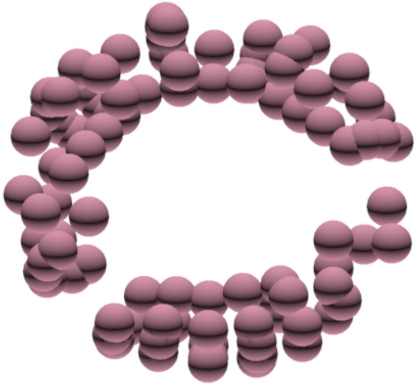
Shape of data.

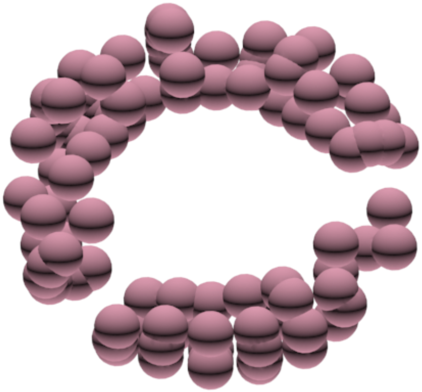


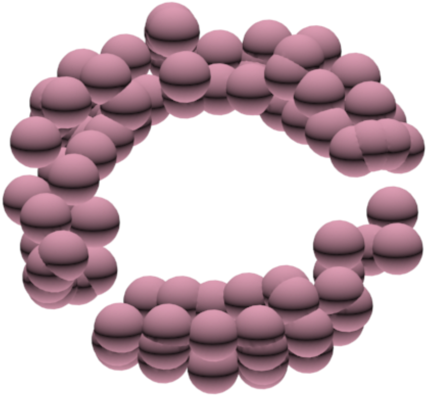


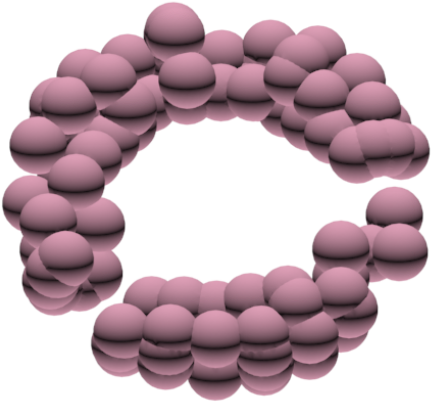


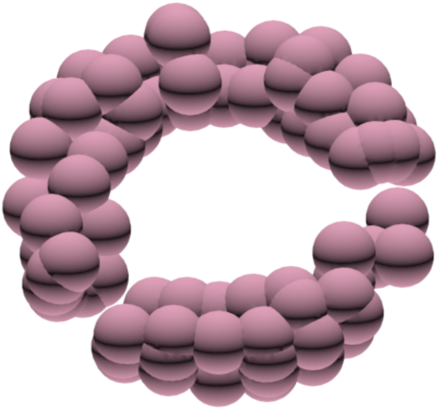


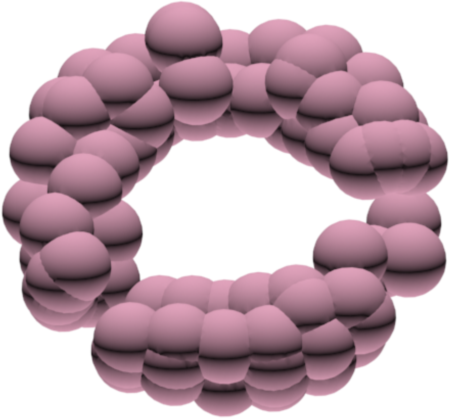


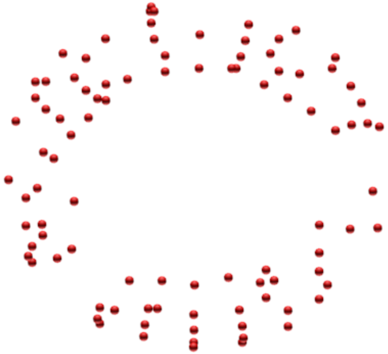


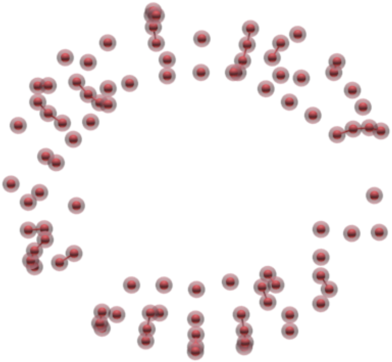


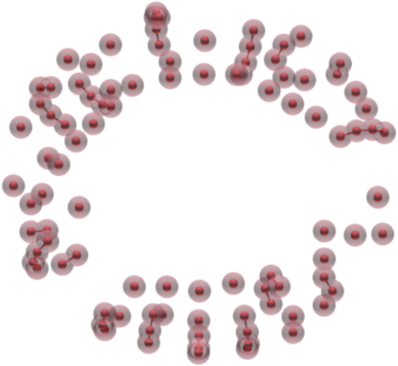


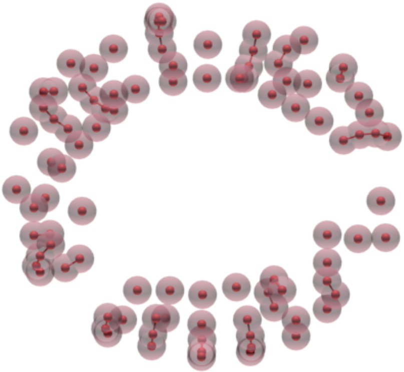


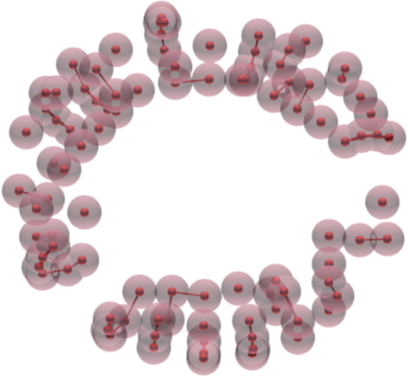












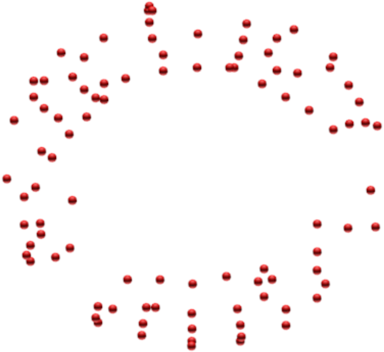
Main tools.

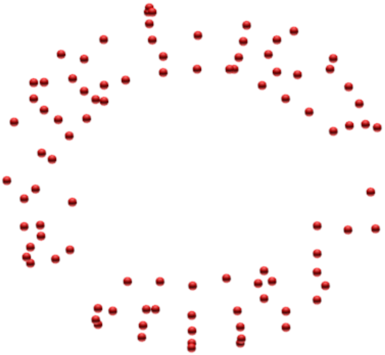
Rips and Cech simplicial complexes:

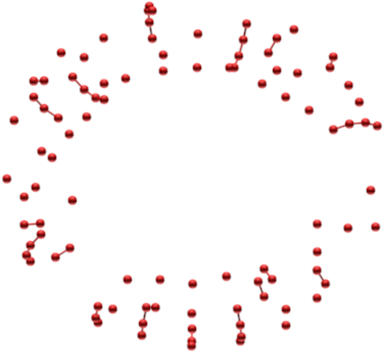
- Capture the shape of the union of balls.
- Combinatorial representation.

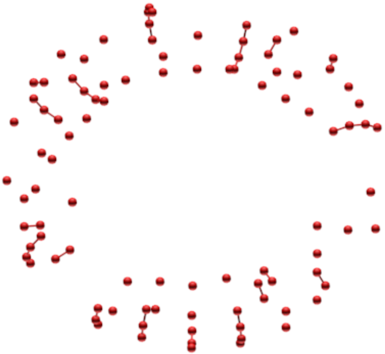
Persistence captures geometric-topological information of the data:

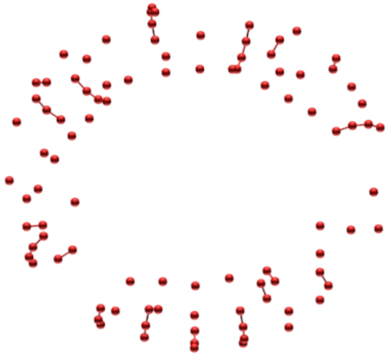
- Key property: stability!

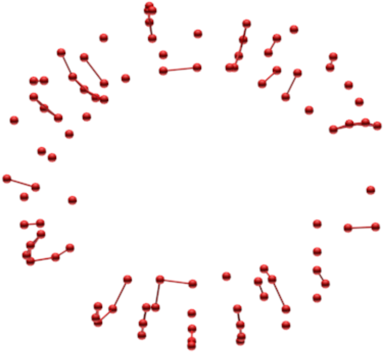


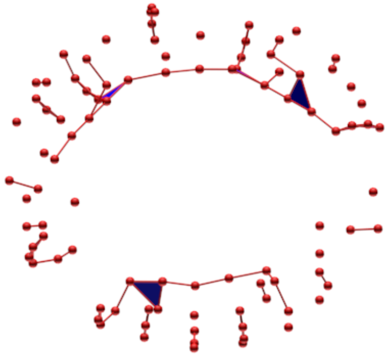


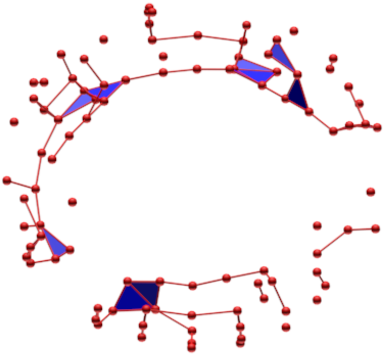


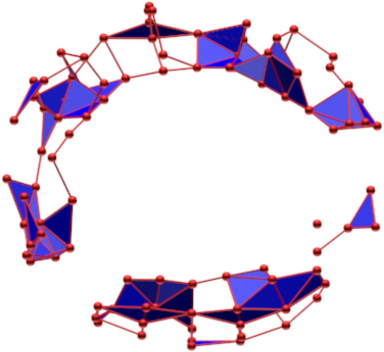


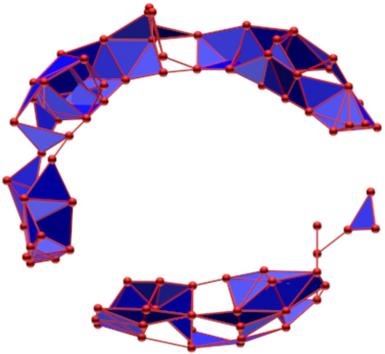


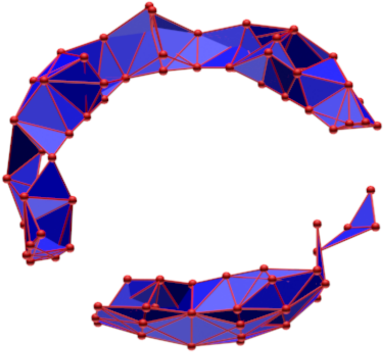


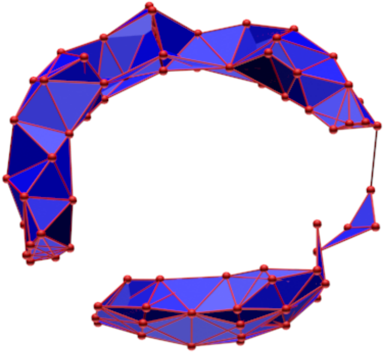


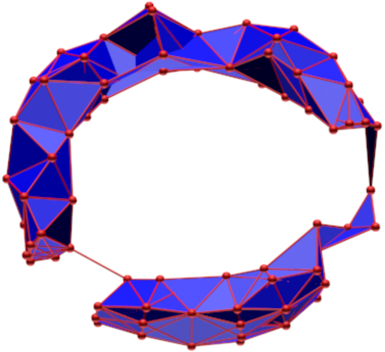


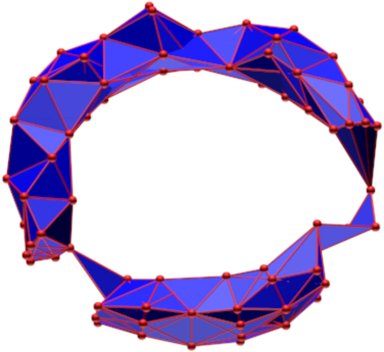


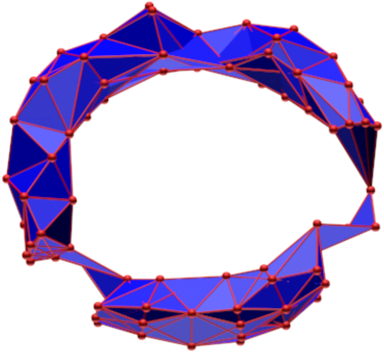


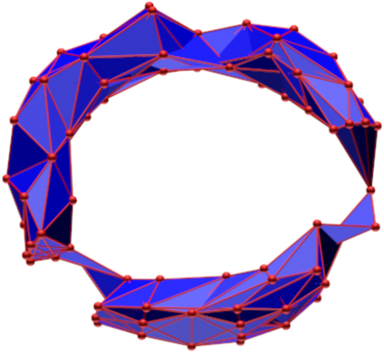


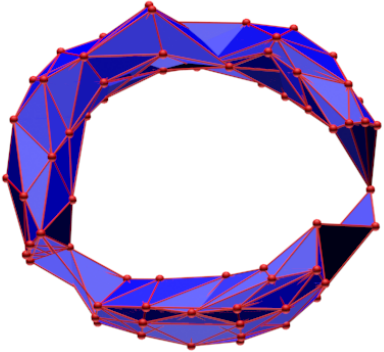


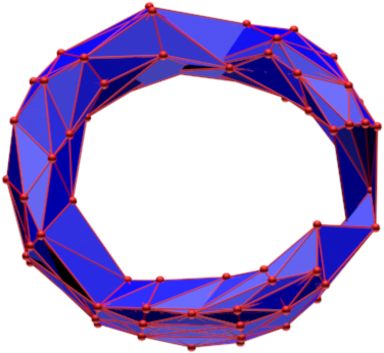




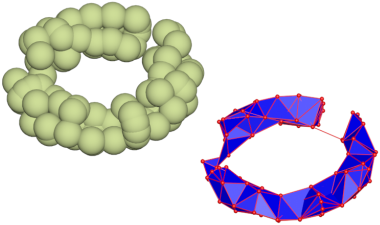








Interpretation of filtration values.



For a simplex $S = v_0, \ldots, v_k$, $f(S) = t$ means that at filtration threshold t, objects $v_0, \ldots, v_k$ are considered *close*.

Text as a point-cloud.

Basic concepts

Corpus:

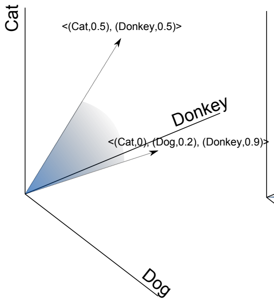
- (Large) collection of text documents.

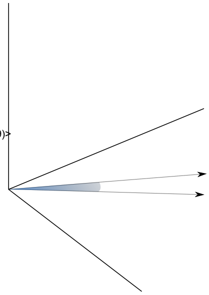
Term-vector:

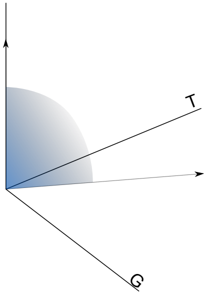
- Weighted vector of key-words or *terms*.
- Summarizes the topic of a single document.
- Higher weight means higher *importance*.

Concept: Vector Space Model

- *Vector Space Model* maps a corpus K to $\mathbb{R}^d$.
- Each distinct *term* in K becomes a direction, so d can be high (10s thousands).
- Each document is represented by its *term-vector*.

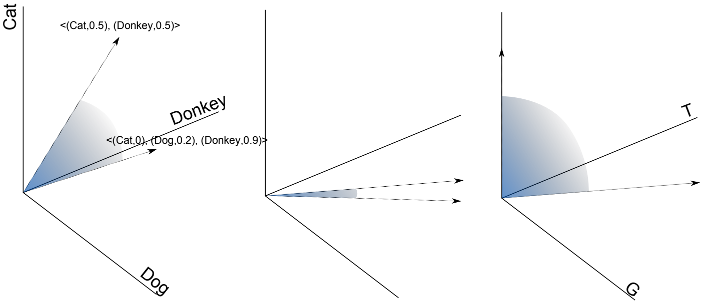






Concept: Similarity measures

- *Cosine similarity* compares two documents.
- Distance (dissimilarity) $d(a, b) := 1 - sim(a, b)$.
- This d is *not a metric*.



Geometry-topological tools.

Interpreting Rips

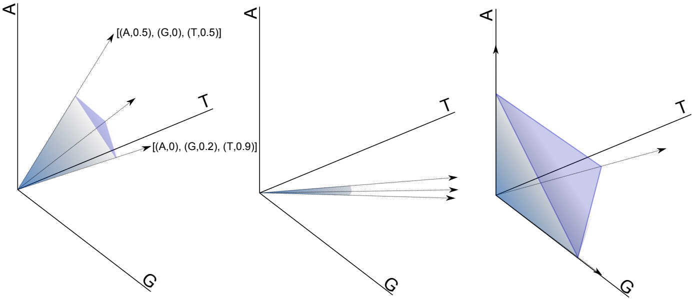
A simplex is added immediately after its boundary:

- $d(a, b)$ – the dissimilarity.
- For triangle $d(a, b, c) = max(d(a, b), d(a, c), d(b, c))$.
- Is this the *filtering function* we want?

Generalized similarity

Goal:

- Extend similarity from pairs to *larger subsets of documents*.
- Its persistence should be stable.
- As a bonus, the resulting complex will be smaller.



Simple example.

For simplicity, let us work with binary term-vectors (or sets of terms).

- $sim_J(X_1, dots, X_d) = \frac{\mathrm{card} \cap_i X_i}{\mathrm{card} \cup_i X_i}$.
- Generalizes the *Jaccard index*.

cat	dog	donkey
1	1	0
0	1	1
1	0	1

New direction.

Flawed generalized cosine measure:

$$R_{\cos}(p^0, p^1, \ldots, p^k) = \sum_{j=1}^{n} \prod_{i=0}^{k} p_j^i. \qquad (1)$$

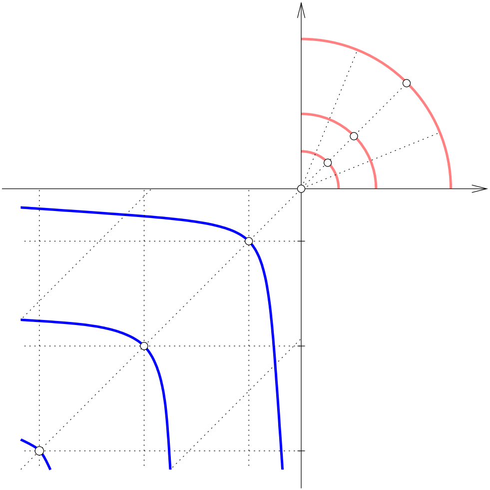
Another option: the length of the geometric mean:

$$R_{\mathrm{gm}}(p^0, p^1, \ldots, p^k) = \left(\sum_{j=1}^{n} \left(\prod_{i=0}^{k} p_j^i \right)^{\frac{2}{k+1}} \right)^{\frac{1}{2}}. \qquad (2)$$

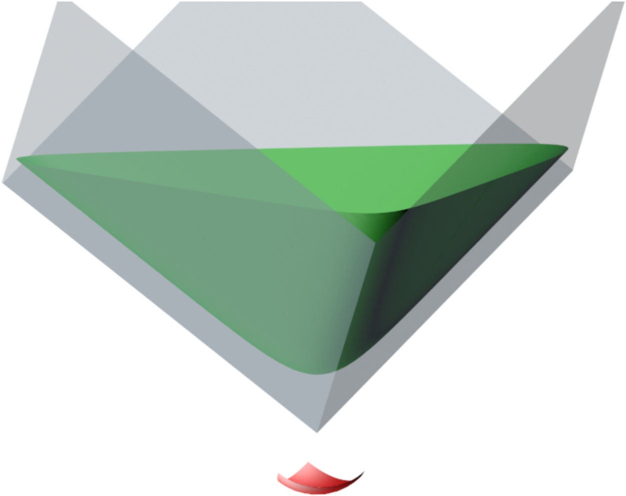
Log-transform

We study the N-dimensional log-transform and related distances.

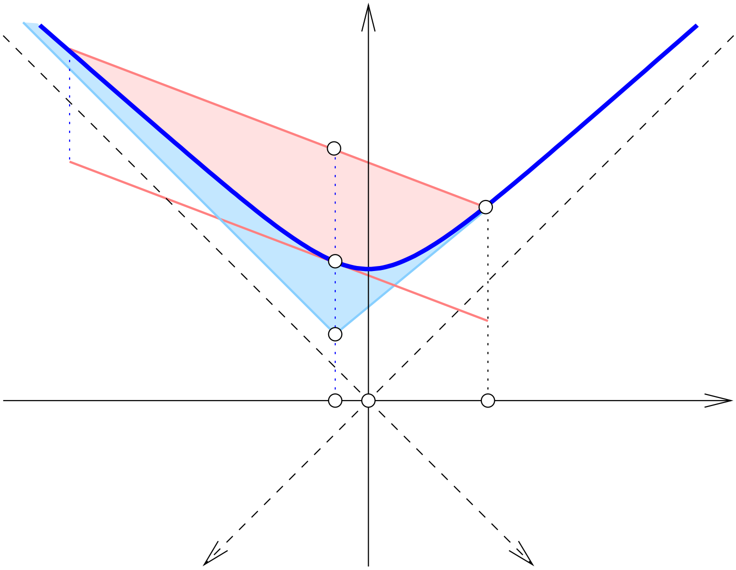
Log-transform



Log-transform in 3D



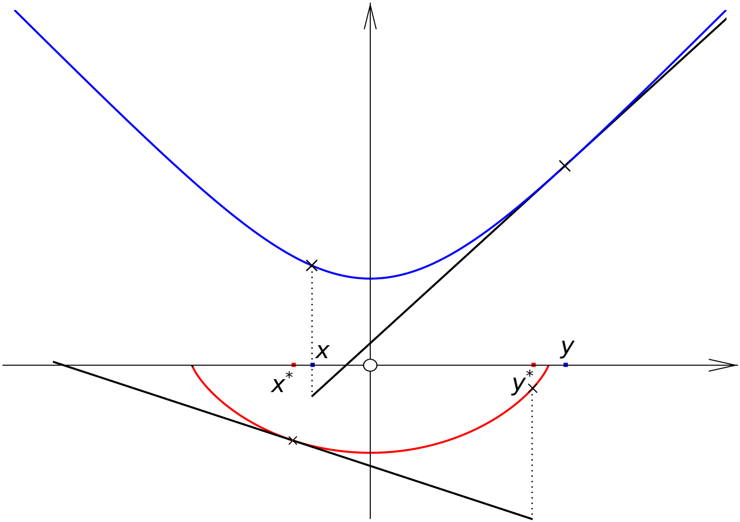
Log-distance



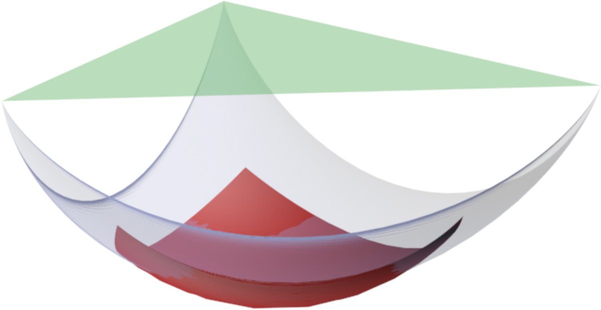
Log-distance: formula

Let $x, y \in \mathbb{R}^{n-1}$, $s = (x, F_1(x))$ and $t = (y, F_1(y))$.
Then the log-distance from x to y is
$D(x, y) = \sum_{j=1}^{n} (t_j - s_j) e^{2t_j}$.

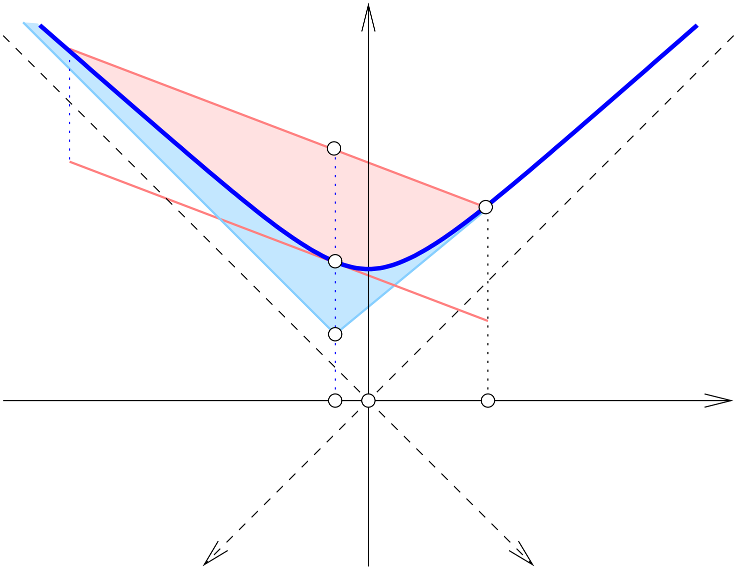
Log-distance: conjugate



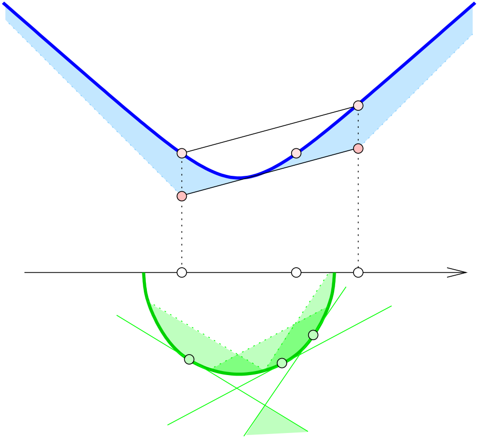
Log-distance: conjugate in 3D



Log Ball



Log Cech complex



$$\mathrm{Cech}_r(X) = \{\xi \subseteq X \mid \bigcap_{x\in\xi} \mathbb{B}_r(x) \neq \emptyset\}. \qquad (3)$$

Generalized measure.

For each simplex $\xi \in \Delta(X)$, there is a smallest radius for which ξ belongs to the Čech complex:

$$r_{\mathrm{C}}(\xi) = \min\{r \mid \xi \in \mathrm{Cech}_r(X)\}. \tag{4}$$

We call $r_{\mathrm{C}} \colon \Delta(X) \to \mathbb{R}$ the *Čech radius function* of X.

In the original coordinate space, we get the desired similarity measure:

$$R_{\mathrm{C}}(\xi) = e^{-r_{\mathrm{C}}(\xi)/\sqrt{n}} \tag{5}$$

Bregman divergences

Bregman divergences

Bregman distance from x to y:

$$D_F(x, y) = F(x) - [F(y) + \langle \nabla F(y), x - y \rangle] \,; \quad (6)$$

Bregman divergences

F can be *any* strictly convex function!

- It covers the Sq. Eucl. distance, squared Mahalanobis distance, Kullback-Leibler divergence, Itakura-Saito distance.
- Extensive use in machine learning.
- Links to statistics via [regular] exponential family (of distributions).

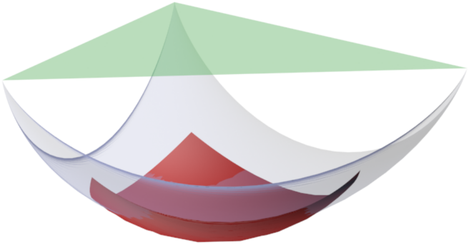
Further connections

- Bregman-based Voronoi [Nielsen at el].
- Information Geometry.
- Collapsibility Cech$\to$Delunay [Bauer, Edelsbrunner].
- Persistence stability for geometric complexes [Chazal, de Silva, Oudot]

Summary

- New, *stable* and relevant distance (dissimilarity measure) for texts.
- It serves as an interpretation of text data.
- Link between TDA and Bregman divergences.

Thank you!



Research partially supported by the TOPOSYS project