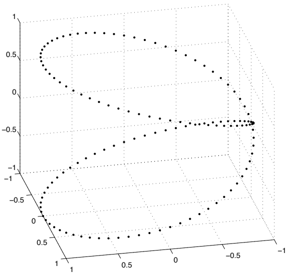
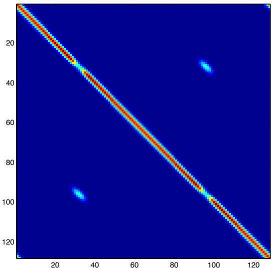
Dual Norms on Product Spaces

William Leeb
University of Minnesota, Twin Cities

Ronald Coifman
Yale University

- In many machine learning problems, we are given a family of kernels $a_t(x, y), t \geq 0$, measuring the similarity of two data points at scale t.





- On the right is an affinity matrix for the curve shown on the left, based on a local Gaussian kernel. (Red is high affinity, blue is low affinity.)

X will denote a measure space equipped with a family of integral operators $A_t, t \geq 0$, with kernels $a_t(x, y)$. The kernels are assumed to satisfy the following:

- (The semigroup property.) For all $t, s > 0$, $A_t A_s = A_{t+s}$. This property can be expressed in terms of the kernels $a_t(x, y)$ as

$$a_{t+s}(x, y) = \int_X a_t(x, w) a_s(w, y) dw.$$

- (The conservation property.) If $\mathbf{1}$ is the constant function 1 on X, then for all $t > 0$, $A_t \mathbf{1} = \mathbf{1}$. This property can be expressed in terms of the kernels $a_t(x, y)$ as

$$\int_X a_t(x, y) dy = 1.$$

- (The integrability property.) There is a constant $C > 0$ such that for all $t > 0$ and $x \in X$,

$$\int_X |a_t(x,y)|dy \leq C.$$

- (The regularity property.) There are constants $C > 0$ and $0 < \alpha < 1$ such that for every $1 \geq s \geq t > 0$ and every $x \in X$,

$$\int_X |a_t(x,y)| \cdot \|a_s(x,\cdot) - a_s(y,\cdot)\|_1 dy \leq C\left(\frac{t}{s}\right)^{\alpha}.$$

- The only strange-looking property is the regularity property

$$\int_X |a_t(x,y)| \cdot \|a_s(x,\cdot) - a_s(y,\cdot)\|_1 dy \le C\left(\frac{t}{s}\right)^\alpha.$$

- This holds in many cases of interest, including:
  - The heat kernel on a "nice" Riemannian manifold;
  - Radial semigroups $K_t(x-y)$ on $\mathbb{R}^n$ with Fourier transform

    $$\hat{K}_t(\xi) = e^{-t|\xi|^\theta}$$

    where $0 < \theta \le 2$ (this includes the Gaussian and Poisson kernels);
  - The heat kernel on fractals such as the Sierpinski Gasket;
  - And many more...

- In our setting, we are not given a distance $d(x, y)$ on X. However, we can use the affinity kernel $a_t(x, y)$ to *define* a distance between points x and y.

- A conceptually meaningful and robust distance is the *diffusion distance* introduced by Coifman and Lafon.

- For each time t, the diffusion distance is defined by

$$d_t(x, y) = \|a_t(x, \cdot) - a_t(y, \cdot)\|_2$$

  with respect to a suitably defined measure on X.

- We define a different metric, namely the weighted sum of L^1 diffusion distances over all scales from 0 to 1:

$$\rho_\alpha(x,y) = \int_0^1 t^{\alpha-1} \|a_t(x,\cdot) - a_t(y,\cdot)\|_1 dt$$

  where $0 < \alpha < 1$.

- This distance is equivalent to

$$d_\alpha(x,y) = \sum_{k=0}^{\infty} 2^{-k\alpha} \|a_{2^{-k}}(x,\cdot) - a_{2^{-k}}(y,\cdot)\|_1.$$

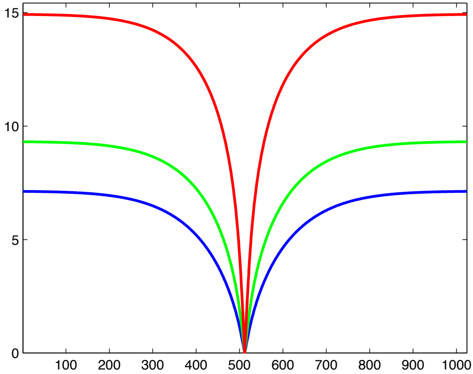
In many examples of interest, $d_\alpha(x,y) \sim \rho(x,y)^\delta$, where $0 < \delta < 1$ and $\rho(x,y)$ is the "natural" distance on X. For example:

- For radial semigroups $K_t(x-y)$ on $\mathbb{R}^n$ with Fourier transform $\hat{K}_t(\xi) = e^{-t|\xi|^\theta}$, $d_\alpha(x,y)$ is locally equivalent to $|x-y|^{\alpha\theta}$.

- If $a_t(x,y)$ is a product of such kernels with scaling parameter θ_i in the i^{th} variable, then $d_\alpha(x,y)$ is locally equivalent to the mixed-homogeneity distance $\sum_{i=1}^n |x_i - y_i|^{\alpha\theta_i}$.

- If $a_t(x,y)$ is the heat kernel on a "nice" Riemannian manifold $\mathcal{M}$, then $d_\alpha(x,y)$ is equivalent to $d_{geod}(x,y)^{2\alpha}$.

We plot the distances $d_\alpha(x, y)$ to a fixed point x on the real line using the Gaussian kernel. Red is $\alpha = .1$, green is $\alpha = .3$, and blue is $\alpha = .45$. The curve approaches the origin like $|y|^{2\alpha}$.



- We consider the space Λ_α of functions f that are Lipschitz with respect to the distance $d_\alpha(x,y)$; that is,

$$\|f\|_{\Lambda_\alpha} = \|f\|_\infty + \sup_{x \neq y} \frac{f(x) - f(y)}{d_\alpha(x,y)} < \infty.$$

- Since in most examples of interest $d_\alpha(x,y)$ is equivalent to $\rho(x,y)^\delta$ for some $0 < \delta < 1$, such functions are usually called *Hölder* functions; we call them *Hölder-Lipschitz* functions.

- The Hölder-Lipschitz space provides a convenient model for functions in non-parametric statistics and machine learning.

- Characterizing these spaces is useful for regression and signal denoising. For example, Donoho and Johnstone use the equivalence of the Hölder norm of f with

$$\sup_{j,k} 2^{k(\alpha+1/2)} |\langle f, \psi_{j,k} \rangle|$$

  for wavelet bases $\{\psi_{j,k}\}_{j,k}$ for optimal denoising.

- These and similar characterizations relate the variation of a function in space to its variation across scales.

- We show that the norm $\|f\|_{\Lambda_\alpha}$ is equivalent to the norms

$$\|f\|_\infty + \sup_{k\geq 0} 2^{k\alpha}\|\Delta_k f\|_\infty$$

and

$$\|f\|_\infty + \sup_{k\geq 0} 2^{k\alpha}\|\delta_k f\|_\infty$$

where

$$\Delta_k = A_{2^{-k}} - A_{2^{-(k+1)}}$$

and

$$\delta_k = I - A_{2^{-k}}.$$

- We also study the space Λ_α^* dual to Λ_α; this contains measures T such that

$$\|T\|_{\Lambda_\alpha^*} = \sup_{\|f\|_{\Lambda_\alpha} \le 1} \langle f, T \rangle < \infty.$$

- The norm $\|T\|_{\Lambda_\alpha^*}$ is equivalent to the norms

$$\|A_1^* T\|_1 + \sum_{k \ge 0} 2^{-k\alpha} \|\Delta_k^* T\|_1$$

and

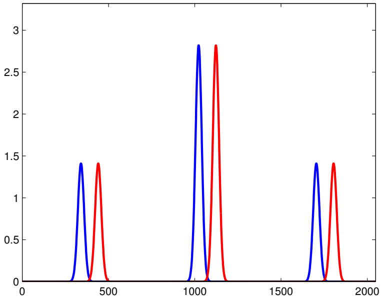
$$\|A_1^* T\|_1 + \sum_{k \ge 0} 2^{-k\alpha} \|D_k^* T\|_1$$

where

$$D_k = A_{2^{-k}} - A_1.$$

- The dual norm is related to the *Earth Mover's Distance (EMD)* between probability measures.

- Informally, the EMD between distributions p_1 and p_2 is the minimal cost of turning p_1 into p_2 by rearranging mass.



- For example, the EMD between the blue and red distributions will be the size of the shift separating them.

- The *Kantorovich-Rubinstein Theorem* says that $\mathrm{EMD}(p_1, p_2)$ is equal to:

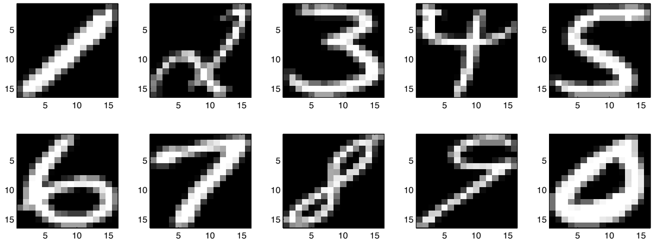
$$\sup_{f:|f(x)-f(y)|\leq d(x,y)} \left\{ \int f(x)(p_1(x) - p_2(x))dx \right\}$$

- This holds in great generality, and in particular for the metric/measure spaces we consider here.

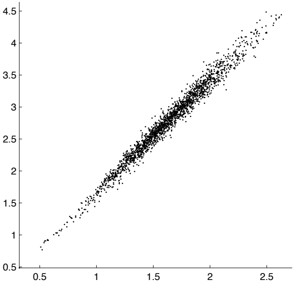
- The Kantorovich-Rubinstein Theorem says that the dual norm $\|p_1 - p_2\|_{\Lambda^*_\alpha}$ of $p_1 - p_2$ is equivalent to $\mathrm{EMD}(p_1, p_2)$.

- Our approximation to EMD(p_1, p_2) is a weighted ℓ_1 distance between the functions $A_{2^{-k}}(p_1)$ and $A_{2^{-k}}(p_2)$.
- We can exploit existing machinery for rapid application of $A_{2^{-k}}$ to get fast approximations to EMD – e.g. diffusion wavelets (Coifman and Maggioni), fast Gauss transform (Greengard and Strain), etc.
- Often, the entire computation of approximate EMD can be done with $\mathcal{O}(n \times \log^k n)$ operations.
- Furthermore, the heavy load is done for each distribution individually, yielding fast methods for computing all pairwise distances between $p_1, p_2, p_3, \ldots$.

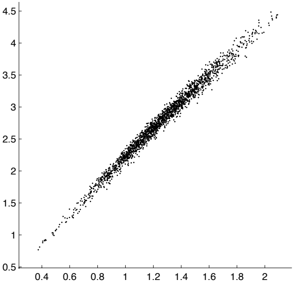
To illustrate the performance of these approximations, we took 2000 random pairs from the USPS dataset (16-by-16 pixel images of handwritten digits). We compared their true EMD to the Gaussian approximations, with $\alpha = .45$.



The left scatterplot is the Δ-approximation, the right is the D-approximation.

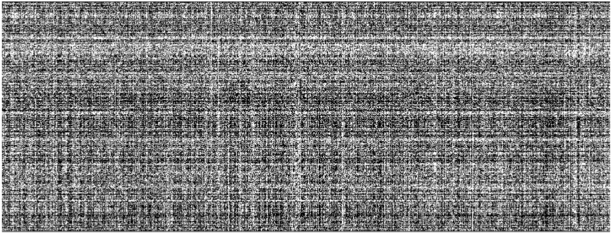


distortion$= 1.316$



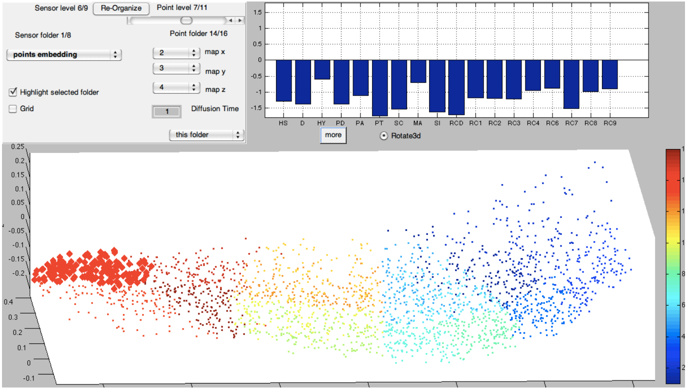
distortion$= 1.202$

- Comparing functions on datasets arises in matrix organization.
- Each row is a function over the columns, and each column is a function over the rows.



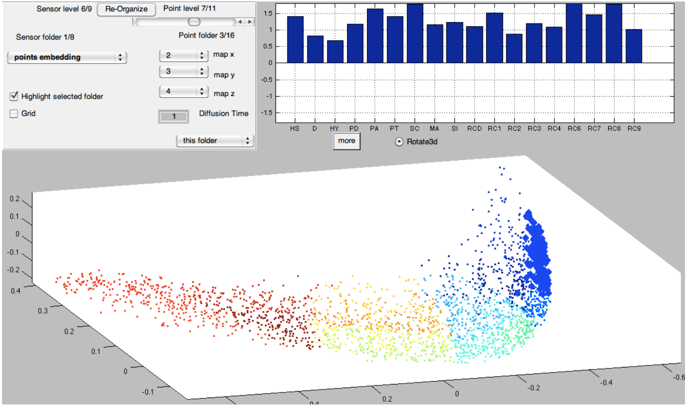
- This is the MMPI2 database of yes/no answers to a psychological questionnaire.

- We use EMD-based affinities to organize and embed the people from the MMPI2 database using diffusion maps:



- One on end of embedding are the clinically healthy people.

- On the other end are the clinically unhealthy people.



- Note that the embedding does not make use of the scores, but is only based on the questionnaire itself.

Our characterizations of the Hölder-Lipschitz space and its dual defined with respect to a single semigroup A_t on X can be extended to the setting of multiparameter semigroups.

- Here, there are two spaces X and Y, each with its own semigroup A_s and B_t, with kernels $a_s(x, x')$ and $b_t(y, y')$.

- The operators $A_sB_t, s \geq 0, t \geq 0$ are given by

$$A_sB_tf(x, y) = \int_Y \int_X a_s(x, x')b_t(y, y')f(x', y')dx'dy'$$

- We define metrics d_α on X and d_β on Y as in the one-parameter case, for $0 < \alpha < 1$ and $0 < \beta < 1$; specifically,

$$d_\alpha(x, x') = \sum_{k \geq 0} 2^{-k\alpha} \| a_{2^{-k}}(x, \cdot) - a_{2^{-k}}(x', \cdot) \|_1$$

and

$$d_\beta(y, y') = \sum_{l \geq 0} 2^{-l\beta} \| b_{2^{-l}}(y, \cdot) - b_{2^{-l}}(y', \cdot) \|_1$$

- The natural measure of a function's regularity in this context is its *mixed Hölder-Lipschitz norm*, the interesting term of which is

$$M(f) = \sup_{x \neq x', y \neq y'} \frac{f(x,y) - f(x,y') - f(x',y) + f(x',y')}{d_\alpha(x,x') d_\beta(y,y')}.$$

- We also require control on the size of the one-variable difference quotients

$$V_X(f) = \sup_{y, x \neq x'} \frac{B_1 f(x,y) - B_1 f(x',y)}{d_\alpha(x,x')}$$

and

$$V_Y(f) = \sup_{x, y \neq y'} \frac{A_1 f(x,y) - A_1 f(x,y')}{d_\beta(y,y')}$$

- We define the *mixed Hölder-Lipschitz space* $\Lambda_{\alpha,\beta}$ to be those functions f such that:

$$\|f\|_{\Lambda_{\alpha,\beta}} = M(f) + V_X(f) + V_Y(f) + \|f\|_\infty < \infty$$

- Mixed Lipschitz functions have better denoising and compressibility properties than ordinary Lipschitz functions, using sparse grids, tensor wavelet coefficients, etc.

- As for one-parameter semigroups, we have derived simple formulas equivalent to this norm and its dual.

- This norm is equivalent to the following two other norms:

$$\|f\|_\infty + \sup_{k\geq 0} 2^{k\alpha}\|\Delta_{A,k}f\|_\infty + \sup_{l\geq 0} 2^{l\beta}\|\Delta_{B,l}f\|_\infty \\ + \sup_{k,l\geq 0} 2^{k\alpha+l\beta}\|\Delta_{A,k}\Delta_{B,l}f\|_\infty$$

and

$$\|f\|_\infty + \sup_{k\geq 0} 2^{k\alpha}\|\delta_{A,k}f\|_\infty + \sup_{l\geq 0} 2^{l\beta}\|\delta_{B,l}f\|_\infty \\ + \sup_{k,l\geq 0} 2^{k\alpha+l\beta}\|\delta_{A,k}\delta_{B,l}f\|_\infty$$

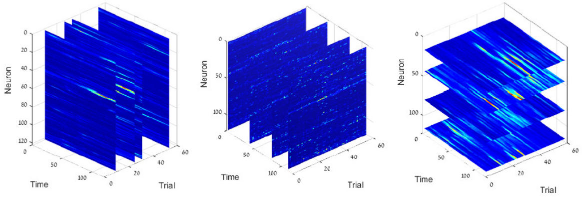
- The norm of a measure T in the dual space is equivalent to the norms

$$\|A_1^* T\|_1 + \sum_{k\geq 0} 2^{-k\alpha} \|\Delta_{A,k}^* T\|_1 + \sum_{l\geq 0} 2^{-l\beta} \|\Delta_{B,l}^* T\|_1 + \sum_{k,l\geq 0} 2^{-(k\alpha+l\beta)} \|\Delta_{A,k}^* \Delta_{B,l}^* T\|_1$$

and

$$\|A_1^* T\|_1 + \sum_{k\geq 0} 2^{-k\alpha} \|D_{A,k}^* T\|_1 + \sum_{l\geq 0} 2^{-l\beta} \|D_{B,l}^* T\|_1 + \sum_{k,l\geq 0} 2^{-(k\alpha+l\beta)} \|D_{A,k}^* D_{B,l}^* T\|_1$$

- In recent work, Mishne et al (2016) use equivalent dual metrics for organizing three-dimensional databases.



- They have a three-dimensional array $\mathbf{X}[r, t, j]$, where r is a neuron, t a short time scale, and j the experiment number. The 3-D structure is organized by comparing 2-D slices using the dual norm.

Calderón, A.P. (1964)
Intermediate spaces and interpolation: the complex method
Studia Mathematica 24(2), 113-190.

Coifman, R.R., and Lafon, S. (2006)
Diffusion maps
Applied and Computational Harmonic Analysis 21(1), 5-30.

Coifman, R.R., and Maggioni, M. (2006)
Diffusion wavelets
Applied and Computational Harmonic Analysis 21(1), 53-94.

Donoho, D. L., & Johnstone, I. M. (1996).
Neo-classical minimax problems, thresholding and adaptive function estimation
Bernoulli, 39-62.

Greengard, L., and Strain, J. (1991)
The fast Gauss transform
SIAM Journal on Scientific and Statistical Computing 12(1), 79-94.

Kuroiwa, S., Tsuge, S., Kita, M., and Ren, F. (2007)
Speaker Identification Method Using Earth Mover's Distance for CCC Speaker Recognition Evaluation 2006
Computational Linguistics and Chinese Language Processing 12(3), 239-254.

Meyer, Y. (1992)
Wavelets and Operators
Vol. 37. Cambridge: Cambridge University Press.

Mishne, G., Talmon, R., Meir, R., Schiller, J., Lavzin, M., Dubin, U., Coifman, R. R. (2016)
Hierarchical Coupled-Geometry Analysis for Neuronal Structure and Activity Pattern Discovery
IEEE Journal of Selected Topics in Signal Processing 10(7), 1238-1253.

Pele, O., and Werman, M. (2009)
Fast and robust earth mover's distances
Computer vision, 2009 IEEE 12th international conference on 460-467.

Rubner, Y., Tomasi, C., and Guibas, L.J. (2000)
The Earth Mover's Distance as a Metric for Image Retrieval
International Journal of Computer Vision 40(2), 99-121.

Sandler, R. and Lindenbaum, M. (2009)
Non-negative Matrix Factorization with Earth Mover's Distance metric
2009 IEEE Conference on Computer Vision and Pattern Recognition 1873-1880.

Shirdhonkar, S., and Jacobs, D.W. (2008)
Approximate earth mover's distance in linear time
Computer Vision and Pattern Recognition. CVPR 2008. IEEE Conference on 1-8.

Typke, R., Wiering, F., and Veltkamp, R.C. (2007)
Transportation distances and human perception of melodic similarity
Musicae Scientae Discussion Forum 4A, 153-181.

Villani, C. (2003)
Topics in Optimal Transportation
No. 58. American Mathematical Soc.

Thank you