Diffusion Earth Mover's Distance and Distribution Embeddings

Alexander Tong* [1], Guillaume Huguet* [2,3], Amine Natik* [2,3], Kincaid Macdonald[4], Manik Kuchroo[5,1], Ronald Coifman[4], Guy Wolf† [2,3], and Smita Krishnaswamy† [5,1]

[1] Dept of Comp. Sci. Yale University; [2] Dept. of Math & Stat. Université de Montréal; [3] MILA Institute; [4] Dept. of Math. Yale University; [5] Dept of Genetics Yale University; *† Equal Contribution.

Summary

A challenge in modern machine learning is the analysis of collections of datasets. We tackle the problem of comparing many datasets sampled from an underlying manifold with the Earth Mover's Distance (EMD). We present Diffusion EMD which is topologically equivalent to EMD with a geodesic ground distance and has (nearest neighbor) complexity scaling linearly in both the number of points and the number of distributions.

Main Idea

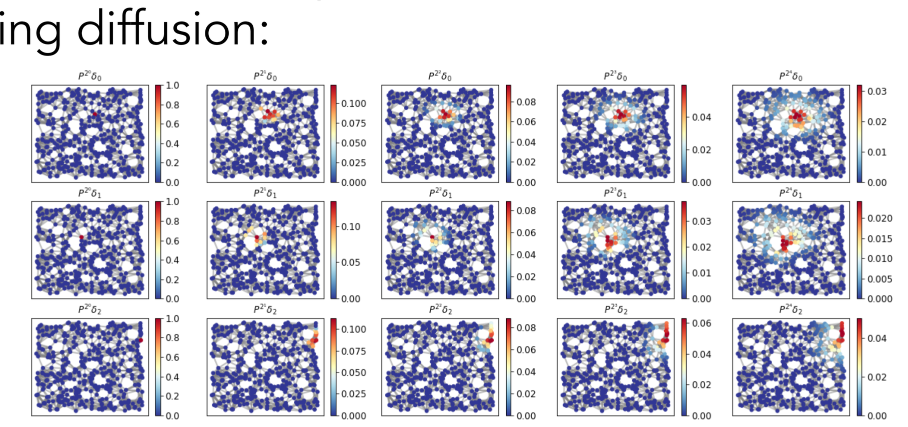
Embed distributions on a graph → vectors such that the L^1 norm is equivalent to the EMD between distributions.

Background & Theory

The Kantorovich Rubenstein Dual form of optimal transport has closed form in the wavelet domain leading to Wavelet EMD [1] which takes the differences of histograms in R^d and performs a weighted sum of wavelet coefficients.

$$\mathrm{WEMD}_\alpha(\mu,\nu) := \sum_j 2^{-j(\alpha+1/2)} \sum_k |\langle \mu-\nu, \psi_{j,k}\rangle|$$

We can define equivalent wavelets on the graph using diffusion:



We define Diffusion EMD based on these graph diffusion wavelets following theory in [2]:

$$\mathrm{DEMD}_{\alpha,K}(X_i,X_j) := \sum_{k=0}^{K} \|T_{\alpha,k}(X_i) - T_{\alpha,k}(X_j)\|_1; \quad 0<\alpha<1/2$$

$$T_{\alpha,k}(X_i) := \begin{cases} 2^{-(K-k-1)\alpha}(\mu_i^{(2^{k+1})} - \mu_i^{(2^k)}) & k<K \\ \mu_i^{(2^K)} & k=K \end{cases}$$

$$\mu_i^{(t)} := \frac{1}{n_i}\mathbf{P}^t \mathbf{1}_{X_i}$$

Further we show that as the number of points converge to continuous distributions on some manifold DEMD is topologically equivalent to EMD with geodesic ground distance:

$$\lim_{X_i\to\mu_i, X_j\to\mu_j} \mathrm{DEMD}_{\alpha,K}(X_i,X_j) \simeq \mathrm{EMD}(\mu_i,\mu_j) \text{ with a geodesic ground distance } d_{\mathcal{M}}^{2\alpha}$$

Results

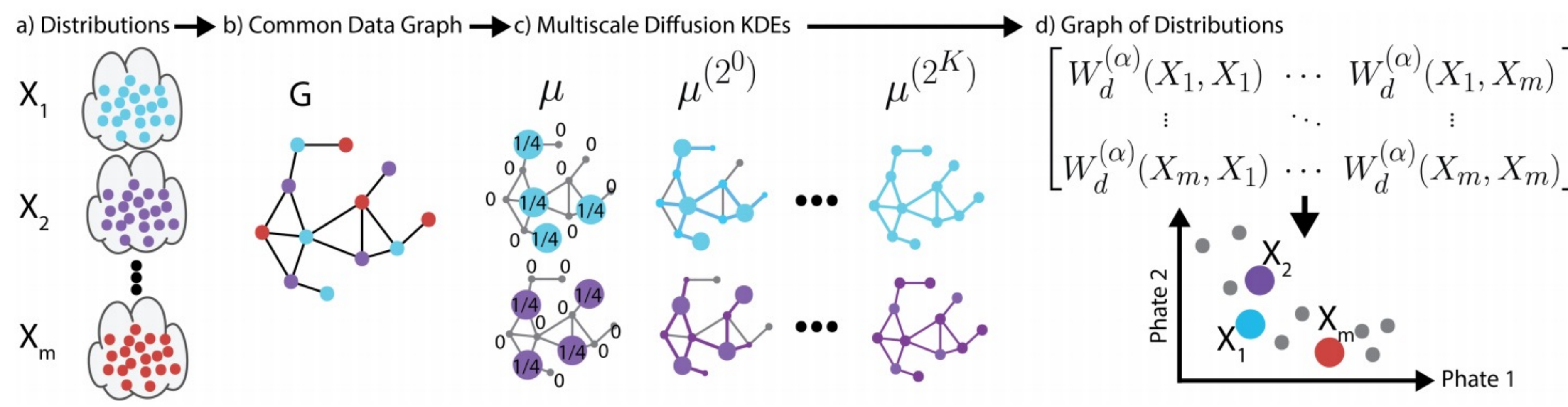


Fig 1. Diffusion EMD first embeds datasets into a common data graph, then takes multiscale diffusions for each dataset, these diffusions are then used to compute Diffusion EMD between the datasets through L_1 norm between vectors. This results in fast EMD-nearest-neighbor calculation between distributions.

Algorithm

Diffusion EMD can be calculated using Chebyshev approximation:

Algorithm 1 Chebyshev embedding

Input: $n \times n$ graph kernel $\boldsymbol{K}$, $n \times m$ distributions $\boldsymbol{\mu}$, maximum scale K, and snowflake constant α.
Output: $m \times (K+1)n$ distribution embeddings $\boldsymbol{b}$
$\boldsymbol{Q} \leftarrow Diag(\sum_i \boldsymbol{K}_{ij})$
$\boldsymbol{K}^{norm} \leftarrow \boldsymbol{Q}^{-1}\boldsymbol{K}\boldsymbol{Q}^{-1}$
$\boldsymbol{D} \leftarrow Diag(\sum_i \boldsymbol{K}^{norm}_{ij})$
$\boldsymbol{M} \leftarrow \boldsymbol{D}^{-1/2}\boldsymbol{K}^{norm}\boldsymbol{D}^{-1/2}$
$\boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{U}^T = \boldsymbol{M}$; $\boldsymbol{U}$ orthogonal, $\boldsymbol{\Sigma}$ Diagonal
$\boldsymbol{\mu}^{(2^0)} \leftarrow \boldsymbol{P}\boldsymbol{\mu} \leftarrow \boldsymbol{D}^{-1/2}\boldsymbol{M}\boldsymbol{D}^{1/2}\boldsymbol{\mu}$
for $k = 1$ **to** K **do**
  $\boldsymbol{\mu}^{(2^k)} \leftarrow \boldsymbol{P}^{2^k}\boldsymbol{\mu} \leftarrow \boldsymbol{D}^{-1/2}\boldsymbol{U}(\boldsymbol{\Sigma})^{2^k}\boldsymbol{U}^T\boldsymbol{D}^{1/2}\boldsymbol{\mu}$
  $\boldsymbol{b}_{k-1} \leftarrow 2^{(K-k-1)\alpha}(\boldsymbol{\mu}^{(2^k)} - \boldsymbol{\mu}^{(2^{k-1})})$
end for
$\boldsymbol{b}_K \leftarrow \boldsymbol{\mu}^{(2^K)}$
$\boldsymbol{b} \leftarrow [\boldsymbol{b}_0, \boldsymbol{b}_1, \ldots, \boldsymbol{b}_K]$

Further, the resulting embedding has mn(K+1) elements, which can be reduced particularly at higher scales using interpolative decomposition, with the idea that a few columns can be used to approximate the operator at high k.

When looking for k-nearest-neighbors in Wasserstein space, we can use locally sensitive hashing or other fast data structures that rely on underlying geometry.

Digit Distributions

	Accuracy	P@10	Spearman ρ	Time
Diff. EMD	**95.94**	**0.611**	**0.673**	34M
Cluster	91.91	0.393	0.484	30M
Quad	79.56	0.294	0.335	**16M**

Fig 2. In a similar amount of time, Diffusion EMD improves 1-nearest-neighbors classification on spherical MNIST digits over other multiscale EMD methods.

Distributions on a Swiss Roll

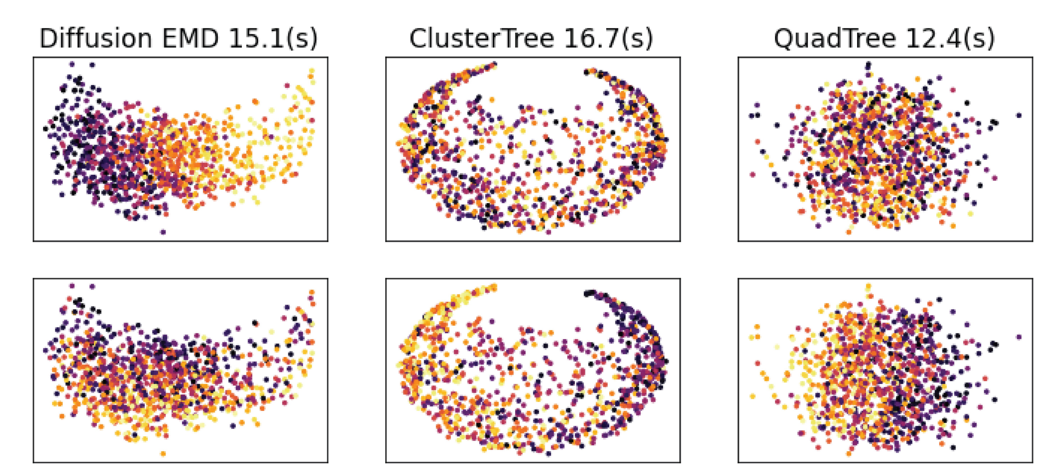


Fig 3. Embeddings of 1000 Gaussians over the swiss roll colored by x and t axes. Diffusion EMD preserves the swiss roll structure by diffusing along it.

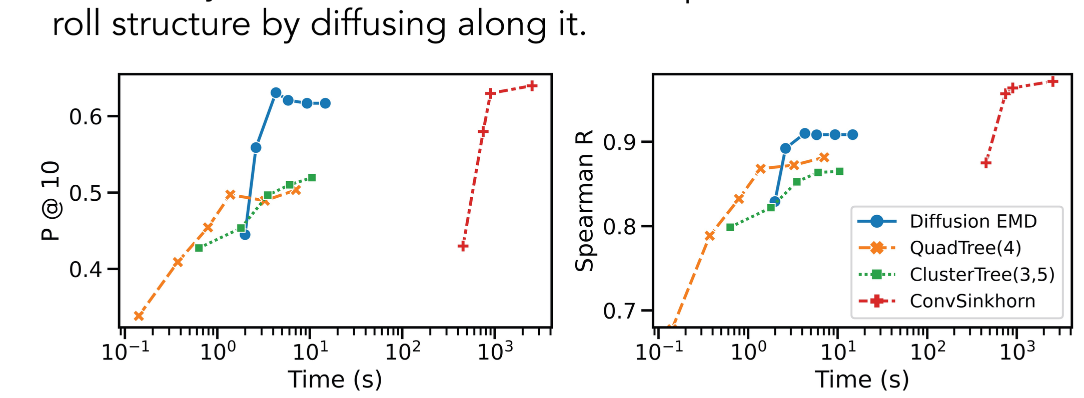


Fig 4. On the Swiss Roll, Diffusion EMD is more accurate than multiscale methods for the same time budget, and faster than convolutional Sinkhorn for the same performance.

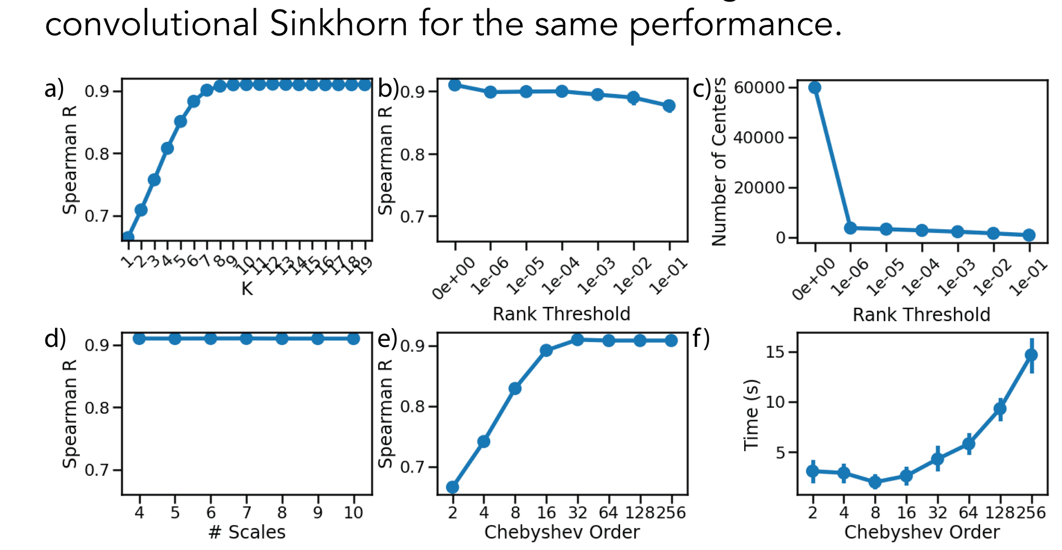


Fig 5. Ablating parameters of Diffusion EMD we find robustness to the largest scale (a), and that the size of the embedding can be reduced significantly with interpolative decomposition.

Distributions on a Line graph

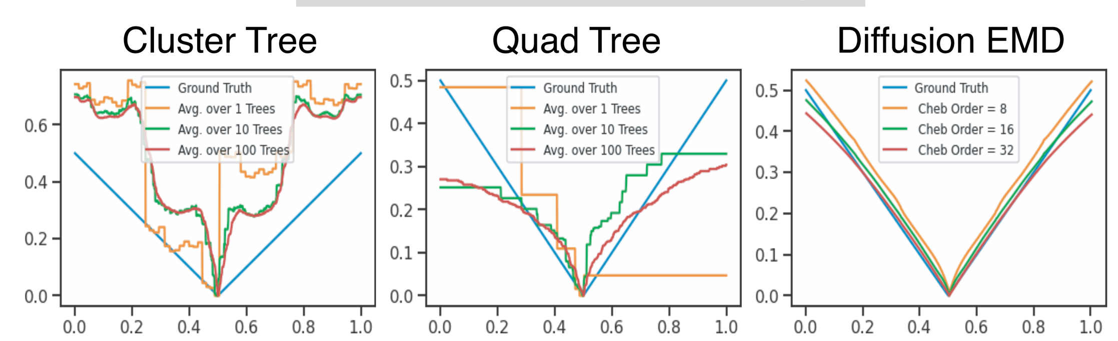


Fig 6. Other linear time multi-scale approximations using trees instead of wavelets produce non-smooth distance functions and require averaging over many trees for a reasonable distance even on a line graph.

Patient Distributions

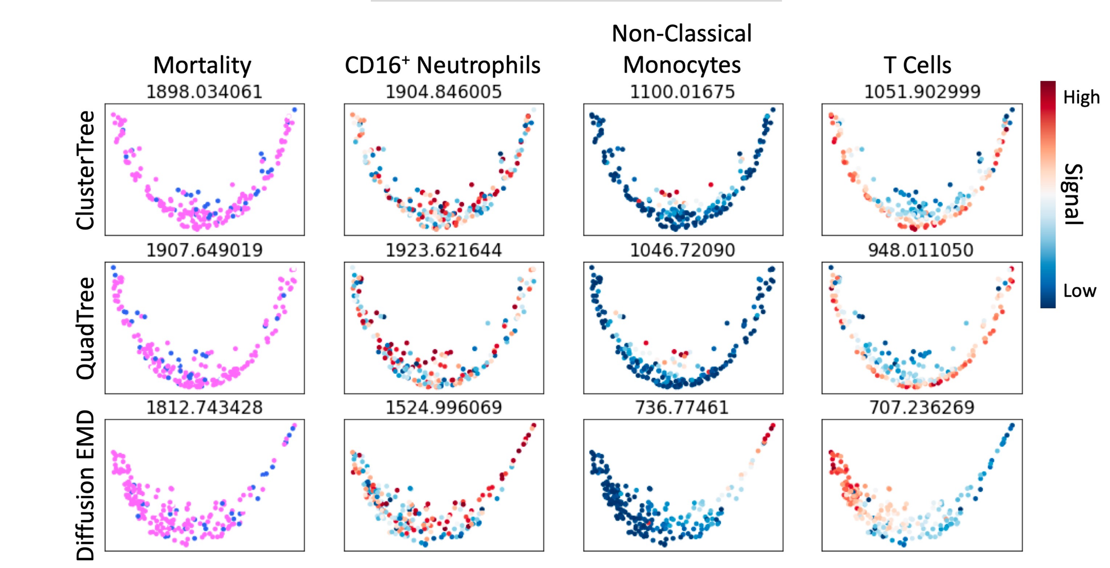


Fig 7. On a dataset of 216 patient samples (distributions) with 23m total cells, Diffusion EMD recapitulates known structure between patients better than other multiscale methods, by taking advantage of the graph structure between cells.

Conclusions

Diffusion EMD approximates EMD with a manifold ground distance efficiently using an embedding into L_1.

Nearest neighbor calculation scales linearly in points and dimensions, and several tricks such as Chebyshev approximation, interpolative decomposition, and data structures reduce computation even further.

References

[1] Shirdhonkar, S. & Jacobs, D. W. Approximate earth mover's distance in linear time. in *CVPR* (IEEE, 2008)
[2] Leeb, W. & Coifman, R. Hölder–Lipschitz Norms and Their Duals on Spaces with Semigroups, with Applications to Earth Mover's Distance. *J Fourier Anal Appl,* (2016).
[3] Tong, A., Huguet, G., Natik, A., MacDonald, K., Kuchroo, M., Coifman, R., Wolf, G. & Krishnaswamy, S. Diffusion Earth Mover's Distance and Distribution Embeddings. in *ICML* (2021).

Further information

Email: alexander.tong@yale.edu

Code: github.com/KrishnaswamyLab/DiffusionEMD

Supported by the IVADO, CIFAR, CZI and NIH.