Neural Representations with Embedding Guarantees via Probabilistic Transformers

A. Kratsios, V. Debarnot, I. Dokmanić

What's up with: Machine Learning

1. Get Data,

What's up with: Machine Learning

1. Get Data,
2. Model Pattern from Data,

What's up with: Machine Learning

1. Get Data,
2. Model Pattern from Data,
3. Train Model,

What's up with: Machine Learning

1. Get Data,
2. Model Pattern from Data,
3. Train Model,
4. Get Predictions.

What's up with: Machine Learning

1. Get Data,
2. Model Pattern from Data,
3. Train Model,
4. Get Predictions.

Model Pattern **From Data**:

Data

What's up with: Machine Learning

1. Get Data,
2. Model Pattern from Data,
3. Train Model,
4. Get Predictions.

Model Pattern **From Data**:

$$\text{Data} \overset{\text{Rep. Learn.}}{\rightarrow} \text{Representation Space}$$

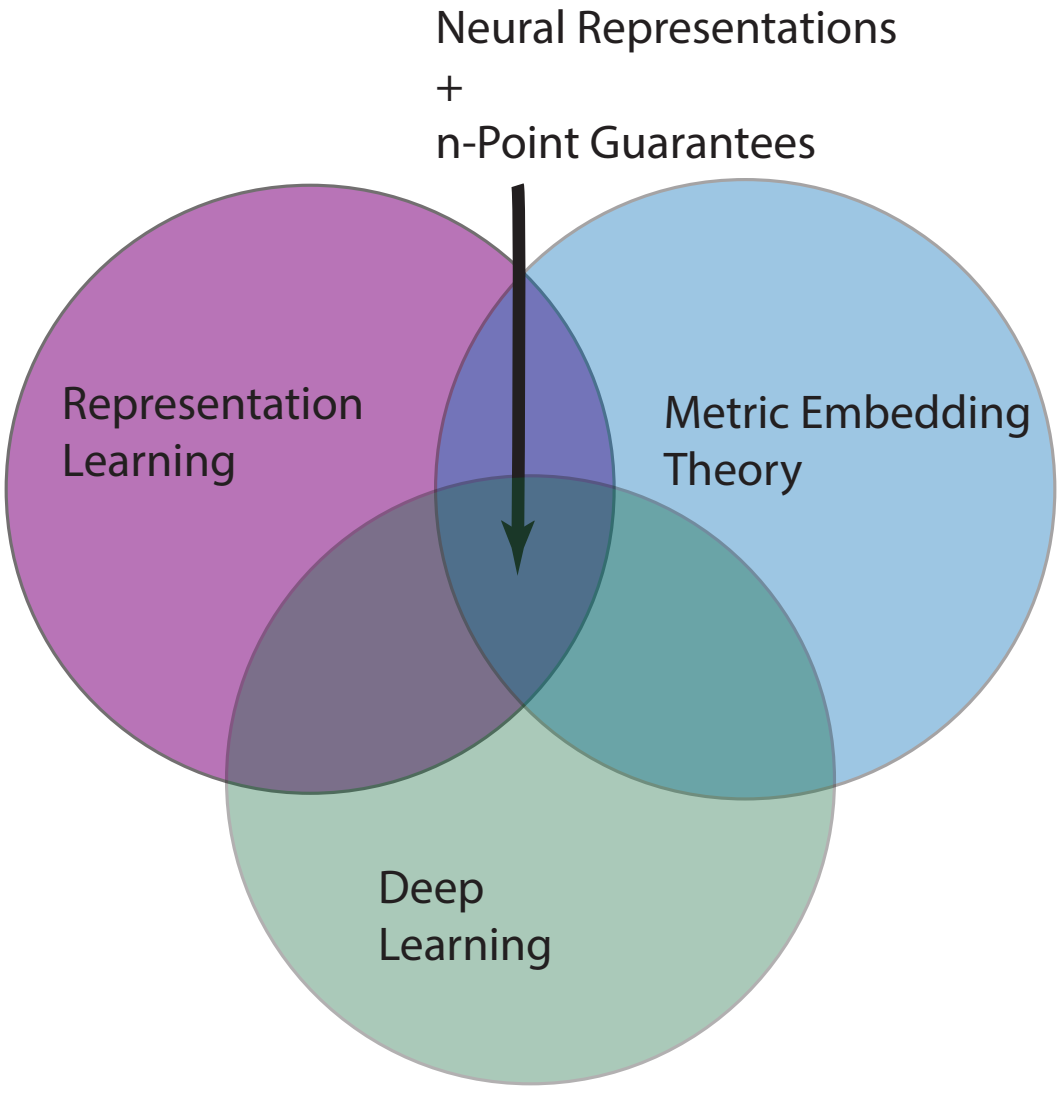
What's up with: Machine Learning

1. Get Data,
2. Model Pattern from Data,
3. Train Model,
4. Get Predictions.

Model Pattern **From Data**:

$$\text{Data} \overset{\text{Rep. Learn.}}{\rightarrow} \text{Representation Space} \overset{\text{Model}}{\rightarrow} \text{Prediction.}$$

Representation Learning with Guarantees



Representation Space = Important

$$\{x_i\}_{i=1}^n \subset \mathcal{X}$$

1. $\mathcal{X}$ "too big" for deep learning (i.e. no BAP; e.g. non sep.),

Representation Space = Important

$$\{x_i\}_{i=1}^n \subset \mathcal{X}$$

1. $\mathcal{X}$ "too big" for deep learning (i.e. no BAP; e.g. non sep.),
2. $\{x_i\}_{i=1}^n \subseteq \mathcal{X}' \subseteq \mathcal{X}$; $\mathcal{X}'$ may have small "effective dimension"
   - Unknown parametric family $\{\mathbb{P}_\theta\}_{\theta \in \mathbb{R}^d}$ in $\mathcal{P}(\mathbb{R})$,

Representation Space = Important

$$\{x_i\}_{i=1}^n \subset \mathcal{X}$$

1. $\mathcal{X}$ "too big" for deep learning (i.e. no BAP; e.g. non sep.),
2. $\{x_i\}_{i=1}^n \subseteq \mathcal{X}' \subseteq \mathcal{X}$; $\mathcal{X}'$ may have small "effective dimension"
   - Unknown parametric family $\{\mathbb{P}_\theta\}_{\theta \in \mathbb{R}^d}$ in $\mathcal{P}(\mathbb{R})$,
3. May not have any "good" deep learning model on $\mathcal{X}$
   - Good approximation properties,
   - Generalize well,
   - Can be easily trained.

Find:

1. Good representation space in which all finite data can be "well-embedded",

Find:

1. Good representation space in which all finite data can be "well-embedded",
2. Implement embeddings of large datasets with small neur. net models,

Find:

1. Good representation space in which all finite data can be "well-embedded",
2. Implement embeddings of large datasets with small neur. net models,
3. Embeddings have small "effective dimension" when data has "good geometric prior"

Data

Consider only a pair $(\mathcal{X}, d)$:

- Fix a *finite set* $\mathcal{X}_n := \{x_1, \ldots, x_n\}$ *of "important points"* in $\mathcal{X}$,

Data

Consider only a pair $(\mathcal{X}, d)$:

- Fix a *finite set* $\mathcal{X}_n := \{x_1, \ldots, x_n\}$ *of "important points"* in $\mathcal{X}$,
- A *metric* d quantifying dissimilarities between any pair of datums.

Data

Consider only a pair $(\mathcal{X}, d)$:

- Fix a *finite set* $\mathcal{X}_n := \{x_1, \ldots, x_n\}$ *of "important points"* in $\mathcal{X}$,
- A *metric* d quantifying dissimilarities between any pair of datums.

Representation

A map $\phi : (\mathcal{X}, d)$ into a "representation space" $\mathcal{R}$ with "distance function d_R" such that: $d(x_i, x_j) \approx d_R(\phi(x_i), \phi(x_j))$

Data

Consider only a pair $(\mathcal{X}, d)$:

- Fix a *finite set* $\mathcal{X}_n := \{x_1, \ldots, x_n\}$ *of "important points"* in $\mathcal{X}$,
- A *metric* d quantifying dissimilarities between any pair of datums.

Representation

A map $\phi : (\mathcal{X}, d)$ into a "representation space" $\mathcal{R}$ with "distance function d_R" such that: $d(x_i, x_j) \approx d_R(\phi(x_i), \phi(x_j))$

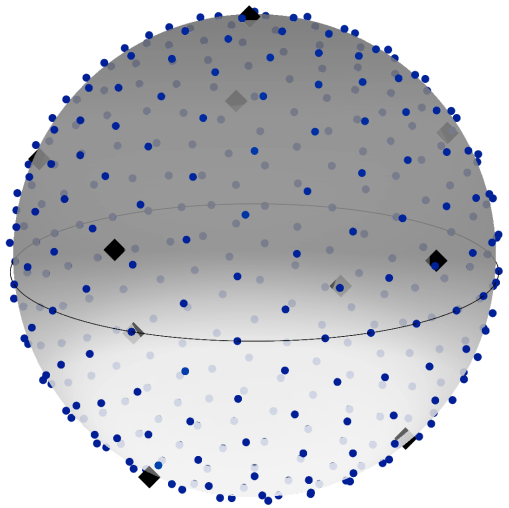
$$s\, d_{i,j}(x, \tilde{x}) \leq d_R(\phi(x_i), \phi(x_j)) \leq s D d(x_i, x_j).$$

Example: "Important Points" on Manifolds

Idea dating back to: [9] "isolated proofs" in [10].

Sample points $x_1, \ldots, x_N$ on the manifold $\mathcal{M}$ such that

$$\min_{i,j} d(x_i, x_j) \lesssim \text{Shortest non-contractible loop.}$$



What is *not* a good representation space?

Look for manifold with "good curvature" [14, 6, 8].

What is *not* a good representation space?

Look for manifold with "good curvature" [14, 6, 8].

Curvature is not enough

Let $\mathcal{R}$ be a complete connected Riemannian manifold. For every $n \in \mathbb{N}_+$, there exists an n-point metric space which cannot be bi-Lipschitz embedded into $\mathcal{R}$ with distortion less than $\mathcal{O}\left(\frac{\log(n)}{\log(\log(n))}\right)$

Because: Too small!

What is *not* a good representation space?

Look for manifold with "good curvature" [14, 6, 8].

Curvature is not enough

Let $\mathcal{R}$ be a complete connected Riemannian manifold. For every $n \in \mathbb{N}_+$, there exists an n-point metric space which cannot be bi-Lipschitz embedded into $\mathcal{R}$ with distortion less than $\mathcal{O}\left(\frac{\log(n)}{\log(\log(n))}\right)$

Because: Too small!

Infinite Dimensions are Not Enough: Bourgain [4] 1985

For every $n \in \mathbb{N}_+$ there exists an n-point metric space which cannot be bi-Lipschitz embedded in $L^2(\mathbb{R})$ with distortion less that $\mathcal{O}\left(\frac{\log(n)}{\log(\log(n))}\right)$.

Block: Too flat!

What *is* a good representation space?

What works?

1. [2] $\mathcal{W}_1(\mathbb{R}^d)$; $d \geq 3$ one can find bi-Hölder embeddings of any finite metric space whose distortion is arbitrarily small,
2. $\mathcal{W}_1(\mathbb{R})$ "exhibits positive curvature on all scales" [11, Section 4.1.1],
3. [11, Proposition 7.4] $\mathcal{W}_1(\mathbb{R})$ has infinite weak-rank (i.e. there are isometric embeddings of every finite-radius Euclidean ball into $\mathcal{W}_1(\mathbb{R})$).

What *is* a good representation space?

What works?

1. [2] $\mathcal{W}_1(\mathbb{R}^d)$; $d \geq 3$ one can find bi-Hölder embeddings of any finite metric space whose distortion is arbitrarily small,
2. $\mathcal{W}_1(\mathbb{R})$ "exhibits positive curvature on all scales" [11, Section 4.1.1],
3. [11, Proposition 7.4] $\mathcal{W}_1(\mathbb{R})$ has infinite weak-rank (i.e. there are isometric embeddings of every finite-radius Euclidean ball into $\mathcal{W}_1(\mathbb{R})$).

Where can we tweak [2]?

1. Work in 1 dimension, since $\mathcal{W}_2$ is hard to compute for $d > 1$,
2. Smaller space where we can exactly implement any probability measure therein

Proposal: Gaussian Mixtures!

1. $\mathcal{GM}_2(\mathbb{R})$: Univariate Gaussian mixtures: $\sum_{i=1}^{I} w_i \, N_1(\mu_i, \Sigma_i)$,
2. $\mathcal{WM}_2$: *"Gaussian Mixture" Optimal Transport Distance*

$$\mathcal{WM}_2(\mu,\nu) := \min_{\substack{\pi \in \mathcal{P}(\mathbb{R}^{d\times d}) \text{ Gauss. Mix.} \\ {}_{\mu,\nu} \text{ are marginals}}} \mathbb{E}_{(X_1,X_2)\sim\pi}\left[\|X_1 - X_2\|^2\right]^{1/2}$$

Proposal: Gaussian Mixtures!

1. $\mathcal{GM}_2(\mathbb{R})$: Univariate Gaussian mixtures: $\sum_{i=1}^{I} w_i \, N_1(\mu_i, \Sigma_i)$,
2. $\mathcal{WM}_2$: *"Gaussian Mixture" Optimal Transport Distance*

$$\mathcal{WM}_2(\mu,\nu) := \min_{\substack{\pi \in \mathcal{P}(\mathbb{R}^{d\times d}) \text{ Gauss. Mix.} \\ \mu,\nu \text{ are marginals}}} \mathbb{E}_{(X_1,X_2)\sim\pi}\left[\|X_1 - X_2\|^2\right]^{1/2}$$

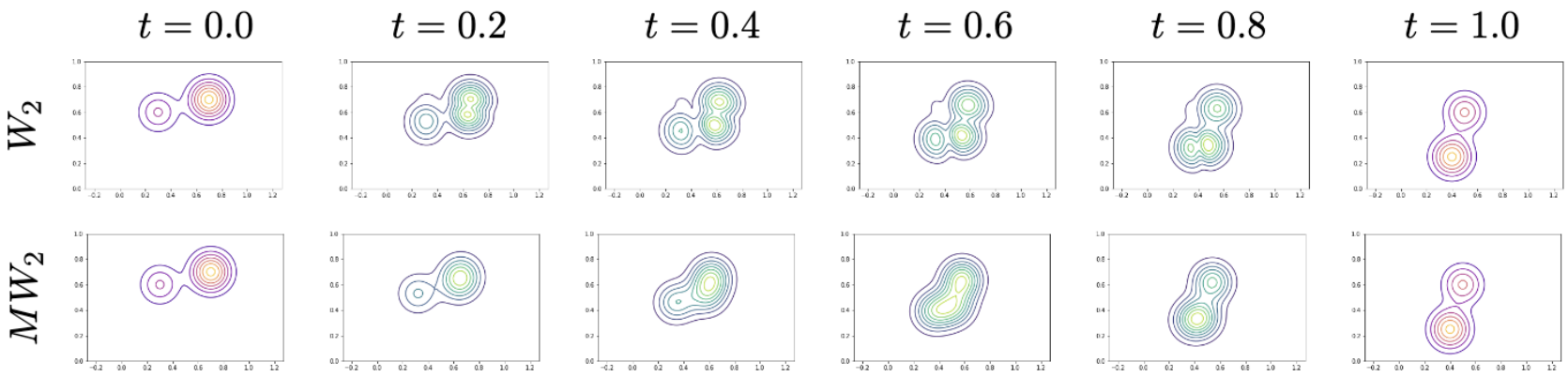
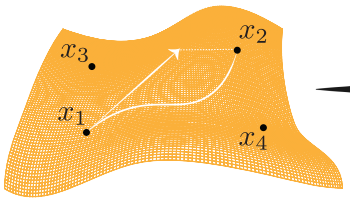


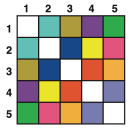
Figure: Geodesics in $\mathcal{W}_2(\mathbb{R})$ vs. in $\mathcal{GM}_2(\mathbb{R})$

Model: Probabilistic Graph Transformer

$\mathcal{X} = \{\text{the manifold}\}$

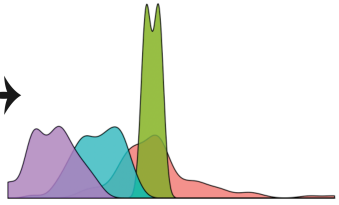


$\mathcal{X}_n = \{x_1, x_2, x_3, x_4\}$

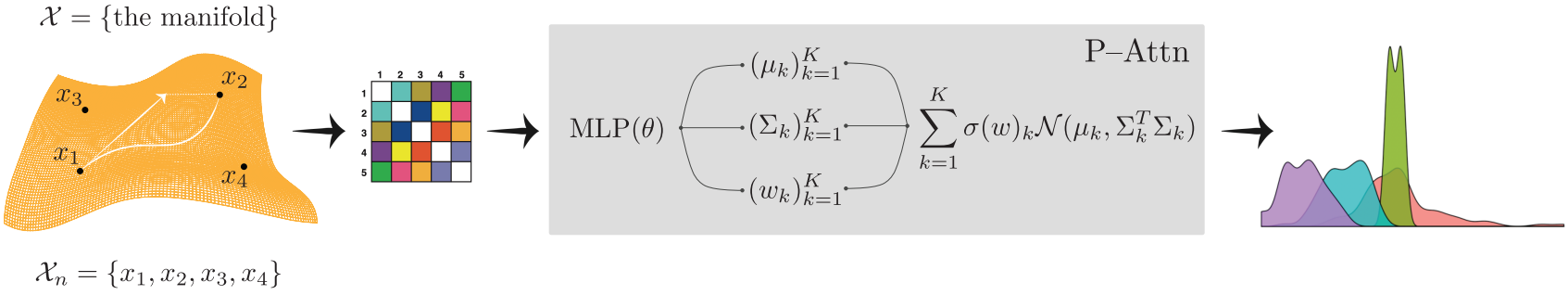


P–Attn

$$\text{MLP}(\theta) \left\{ \begin{array}{l} (\mu_k)_{k=1}^K \\ (\Sigma_k)_{k=1}^K \\ (w_k)_{k=1}^K \end{array} \right. \sum_{k=1}^{K} \sigma(w)_k \mathcal{N}(\mu_k, \Sigma_k^T \Sigma_k)$$



Model: Probabilistic Graph Transformer



1. (Unnormalized) Graph-Attention [16]

$$\overline{\text{G-Attn}} : \mathcal{X} \ni x \mapsto (d_n(x, x_l))_{l=1}^n \in \mathbb{R}^n,$$

2. Feedforward neural network $f : \mathbb{R}^n \to \mathbb{R}^{3K}$

$$u \mapsto W^J \circ \text{ReLU} \bullet \cdots \circ \text{ReLU} \bullet W^1(u) =: (\mu, \Sigma, w),$$

3. Probabilistic Attention:

$$\text{P-Attn}\left(w, (\mu_k, \Sigma_k, w_k)_{k=1}^K\right) := \sum_{k=1}^K \text{Softmax}_K(w)_k \, N(\mu_k, |\Sigma_k|) \in \mathcal{P}_2(\mathbb{R}).$$

What we know about transformers?

1. **Extremely High-Capacity:** Can approximate all "good" cond. dist. [12] + all "good" processes [1],

What we know about transformers?

1. **Extremely High-Capacity:** Can approximate all "good" cond. dist. [12] + all "good" processes [1],
2. **Super Flexible:** Can approximate any function *exactly encoding any constraints* [13]

What we know about transformers?

1. **Extremely High-Capacity:** Can approximate all "good" cond. dist. [12] + all "good" processes [1],
2. **Super Flexible:** Can approximate any function *exactly encoding any constraints* [13] (by leveraging randomness),

Main Result: PAC Embeddings

Given an n-point metric space $(\mathcal{X}_n, d_n)$ and the the uniform probability measure $\mathbb{P}$ on $\mathcal{X}_n$, every "distortion parameter" $D > 2$, there exists a "scale" $s > 0$ and a probabilistic transformer $\hat{T} : (\mathcal{X}_n, d_n) \hookrightarrow (\mathcal{MG}_2(\mathbb{R}), \mathcal{MW}_2)$ satisfying

$$\mathbb{P}\left(s d_n(x, \tilde{x}) \leq \mathcal{MW}_2(\hat{T}(x), \hat{T}(\tilde{x})) \leq s D^2 d_n(x, \tilde{x}) \right) = \frac{1}{n^{2-2\theta_D}},$$

Main Result: PAC Embeddings

Given an n-point metric space $(\mathcal{X}_n, d_n)$ and the the uniform probability measure $\mathbb{P}$ on $\mathcal{X}_n$, every "distortion parameter" $D > 2$, there exists a "scale" $s > 0$ and a probabilistic transformer $\hat{T} : (\mathcal{X}_n, d_n) \hookrightarrow (\mathcal{MG}_2(\mathbb{R}), \mathcal{MW}_2)$ satisfying

$$\mathbb{P}\left(s d_n(x, \tilde{x}) \leq \mathcal{MW}_2(\hat{T}(x), \hat{T}(\tilde{x})) \leq s D^2 d_n(x, \tilde{x}) \right) = \frac{1}{n^{2-2\theta_D}},$$

(i) **Width:** at-most $\mathcal{O}\left(\max\{ \frac{\theta_D \log_2(n)}{(D-2)^2}, n^{2\theta_D} \} \right)$,

(ii) **Depth:** $\approx \tilde{\mathcal{O}}\left(n^{\theta_D} \sqrt{\theta_D \, \log(n)} \left(1 + \frac{\log(2)}{\theta_D \, \log(n)} \right) \right)$,

(iii) **Effective Dimension:** $\mathcal{O}\left(\frac{\theta_D \log_2(n)}{(D-2)^2} \right)$.

Main Result: PAC Embeddings

Given an n-point metric space $(\mathcal{X}_n, d_n)$ and the the uniform probability measure $\mathbb{P}$ on $\mathcal{X}_n$, every "distortion parameter" $D > 2$, there exists a "scale" $s > 0$ and a probabilistic transformer $\hat{T} : (\mathcal{X}_n, d_n) \hookrightarrow (\mathcal{MG}_2(\mathbb{R}), \mathcal{MW}_2)$ satisfying

$$\mathbb{P}\left(s d_n(x, \tilde{x}) \leq \mathcal{MW}_2(\hat{T}(x), \hat{T}(\tilde{x})) \leq s D^2 d_n(x, \tilde{x}) \right) = \frac{1}{n^{2-2\theta_D}},$$

(i) **Width:** at-most $\mathcal{O}\Big(\max\{ \frac{\theta_D \log_2(n)}{(D-2)^2}, n^{2\theta_D} \} \Big)$,

(ii) **Depth:** $\approx \tilde{\mathcal{O}}\left(n^{\theta_D} \sqrt{\theta_D \, \log(n)} \left(1 + \frac{\log(2)}{\theta_D \, \log(n)} \right) \right)$,

(iii) **Effective Dimension:** $\mathcal{O}\big(\frac{\theta_D \log_2(n)}{(D-2)^2} \big)$.

where $\theta_D \approx \frac{D-2}{\log(1/(D-2)^2)}$ and $\text{aspect}(\mathcal{X}_n, d_n) := \frac{\max_{x, \tilde{x} \in \mathcal{X}_n} \, d_n(x, \tilde{x})}{\min_{x, \tilde{x} \in \mathcal{X}_n \, ; \, x \neq \tilde{x}} \, d_n(x, \tilde{x})}$.

Questions

Questions

1. Must the embeddings be probabilistic?

Questions

1. Must the embeddings be probabilistic?
2. Can we do better if our data has some latent geometry?

Questions

1. Must the embeddings be probabilistic?
2. Can we do better if our data has some latent geometry?

Answer to first Question: Deterministic Guarantees

No!

Answer to first Question: Deterministic Guarantees

No!

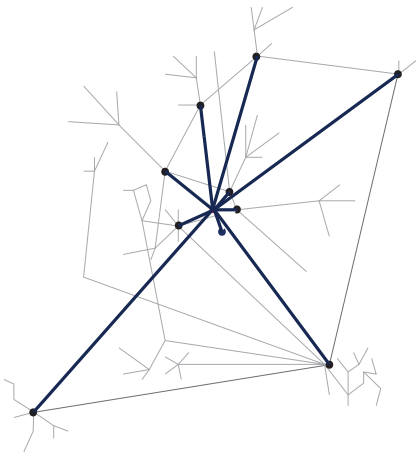


Figure: $\alpha = 1$, All priorities equal.

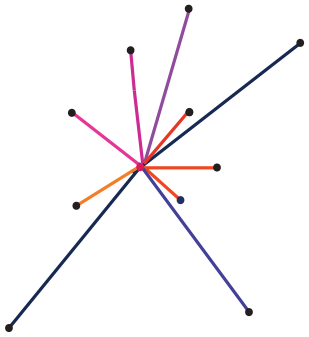


Figure: $\alpha \ll 1$, Priority to nearby data.

Figure: $d^{\alpha}(x_i, x_j) \approx \|\phi(x_i) - \phi(x_j)\|$

Deterministic Fractal Embeddings

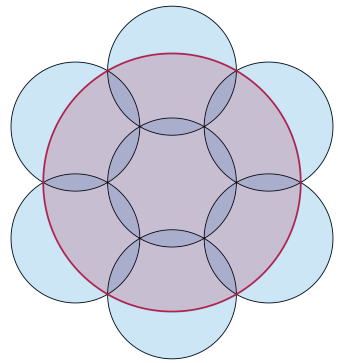
Given an n-point metric space $(\mathcal{X}_n, d_n)$. For every "geometric perturbation parameter" $1/2 < \alpha < 1$ there is a prob. transformer $\hat{T}$ satisfying

$$d_n^{\alpha}(x, \tilde{x}) \leq \mathcal{MW}_2(\hat{T}(x), \hat{T}(\tilde{x})) \left\lceil \left(\frac{12 \log(\mathrm{cap}(\mathcal{X}_n))}{(1-\alpha)} \right)^{1+\alpha} \right\rceil d_n^{\alpha}(x, \tilde{x}).$$

(i) **Width:** $\mathcal{O}\left(\max\{ \frac{\log(\mathrm{cap}(\mathcal{X}_n))}{\alpha}, n^2 \} \right)$,

(ii) **Depth:** $\tilde{\mathcal{O}} \left(n \left(1 + \frac{\log(2)\sqrt{n}}{\sqrt{\log(n)}} \right) \right)$,

(iii) **Effective Dimension:** $\lceil 12 C \alpha^{-1} (\log(\mathrm{cap}(\mathcal{X}_n))) \rceil$.



Application: 2-Hop Combinatorial Graphs

Let $G = (V, E)$ be a 2-Hop graph (e.g. complete bipartite graphs, cocktail graphs, friendship graphs, etc...) and let d_g be its combinatorial distance.
For every $\frac{1}{2} < \alpha < 1$, there is a $\hat{T}$ satisfying

$$d_n^\alpha(x, \tilde{x}) \leq \mathcal{MW}_2(\hat{T}(x), \hat{T}(\tilde{x})) \lesssim \left\lceil \left(\frac{12 \log(1+\rho(A_G)}{1-\alpha} \right)^{1+\alpha} \right\rceil d_n^\alpha(x, \tilde{x}).$$

Application: 2-Hop Combinatorial Graphs

Let $G = (V, E)$ be a 2-Hop graph (e.g. complete bipartite graphs, cocktail graphs, friendship graphs, etc...) and let d_g be its combinatorial distance.
For every $\frac{1}{2} < \alpha < 1$, there is a $\hat{T}$ satisfying

$$d_n^{\alpha}(x, \tilde{x}) \leq \mathcal{MW}_2(\hat{T}(x), \hat{T}(\tilde{x})) \lesssim \left\lceil \left(\frac{12 \log(1+\rho(A_G)}{1-\alpha} \right)^{1+\alpha} \right\rceil d_n^{\alpha}(x, \tilde{x}).$$

(i) **Width:** $\mathcal{O}\big(\max\{ \frac{\log(1+\rho(A_G))}{\alpha}, n(n-1) \} \big)$,

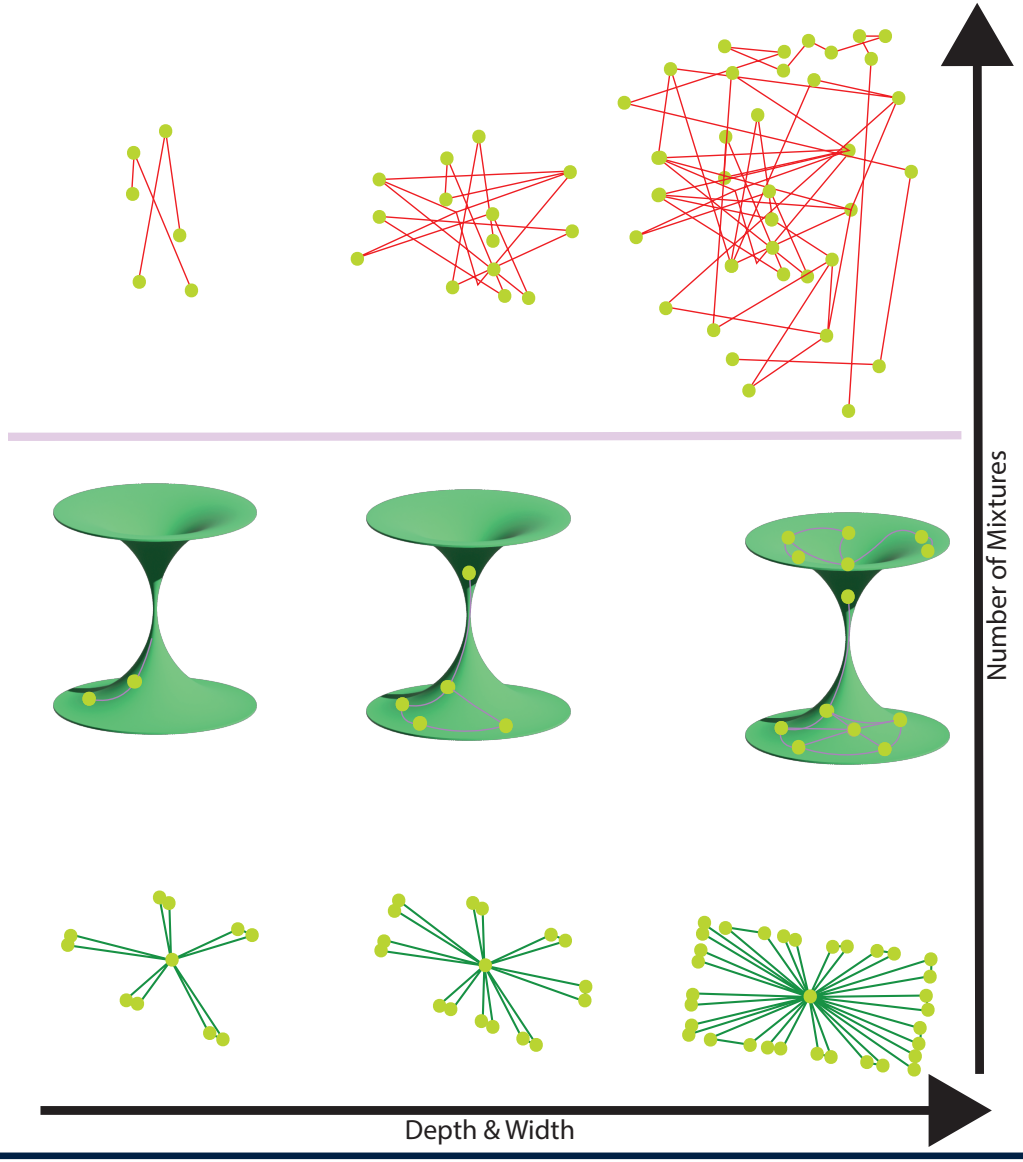
(ii) **Depth:** $\mathrm{Depth}(\hat{T})$ is $\tilde{\mathcal{O}} \left(n \left(1 + \frac{\log(2)\sqrt{n}}{\sqrt{\log(n)}} \right) \right)$,

(iii) **Effective Dimension:** $\lceil 12 C \alpha^{-1} \left(\log(1 + \rho(A_G)) \right) \rceil$.

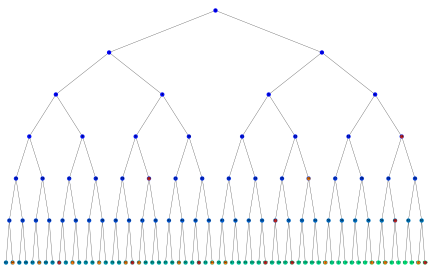
Questions

1. Must the embeddings be probabilistic?
2. Can we do better (bi-Lipschitz) if our data has some latent geometry?

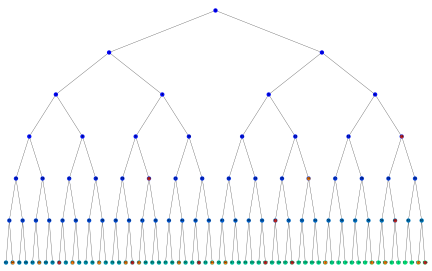
Question 2: Embedding Landscape

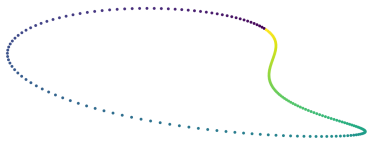


What's a geometric Prior?

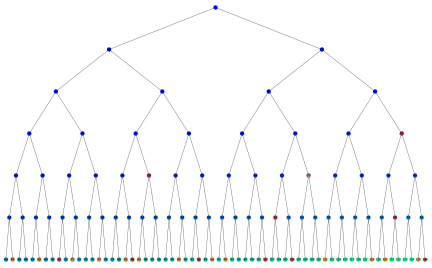


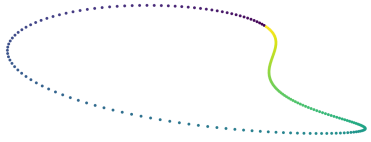
What's a geometric Prior?

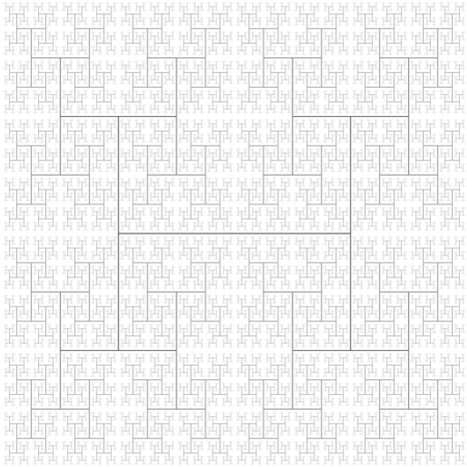




What's a geometric Prior?

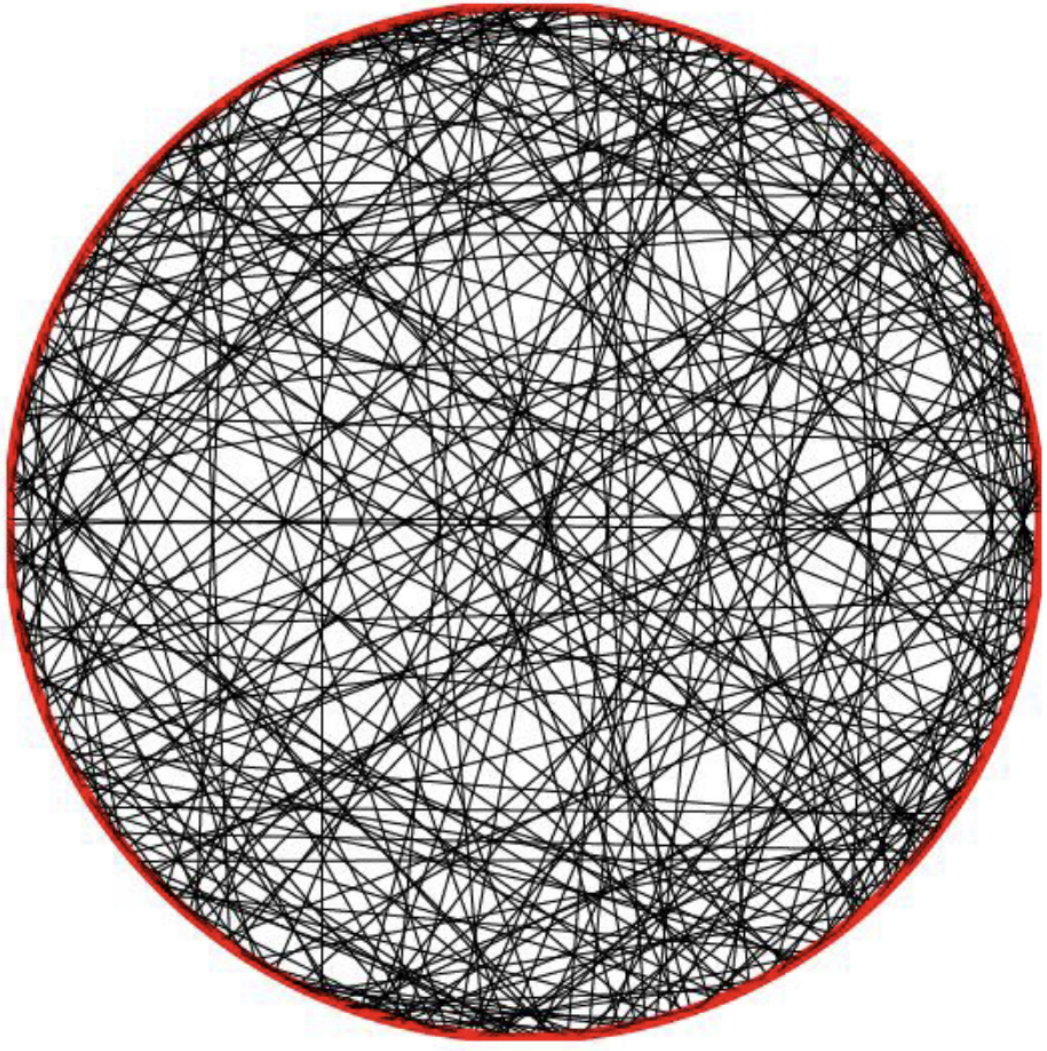






What's not a geometric prior...

"The Leftovers"



If G is an n-node combinatorial tree.

If G is an n-node combinatorial tree. For every $M \geq 2$ there is a $\hat{T}$ with *effective dimension* M satisfying

$$d_n(x, \tilde{x}) \leq \mathcal{MW}_2(\hat{T}(x), \hat{T}(\tilde{x})) \lesssim n^{1/(M-1)} d_n(x, \tilde{x}).$$

If (M, g) is a d-dimensional compact connected Riemannian manifold

If (M, g) is a d-dimensional compact connected Riemannian manifold there is a $\hat{T}$ of *effective dimension* $2d$ satisfying

$$d_n(x, \tilde{x}) \leq \mathcal{MW}_2(\hat{T}(x), \hat{T}(\tilde{x})) \lesssim n^{1/(M-1)} d_n(x, \tilde{x}).$$

Benchmark:

1. [15] all "good" trees embed "well" into the hyperbolic plane,

Benchmark:

1. [15] all "good" trees embed "well" into the hyperbolic plane,
2. [7, 5] neural nets. into hyperbolic space do well for implementing trees,

Benchmark:

1. [15] all "good" trees embed "well" into the hyperbolic plane,
2. [7, 5] neural nets. into hyperbolic space do well for implementing trees,
3. [3] Represent hyperbolic plane as Gaussian mixtures with "Information Geometry",

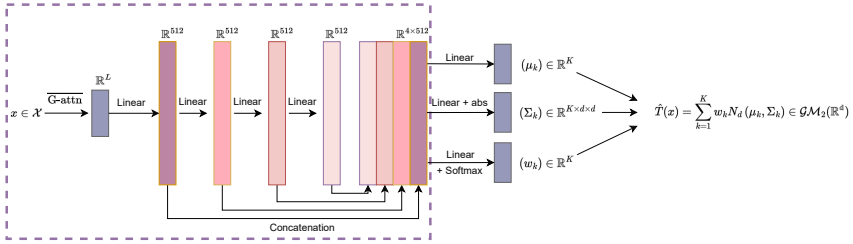
Benchmark:

1. [15] all "good" trees embed "well" into the hyperbolic plane,
2. [7, 5] neural nets. into hyperbolic space do well for implementing trees,
3. [3] Represent hyperbolic plane as Gaussian mixtures with "Information Geometry",
4. [1, Example 11] Transformer universal model into hyperbolic space

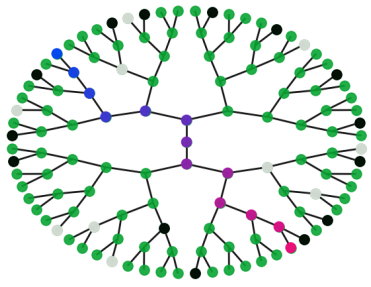
Does it work in practice?

Benchmark:

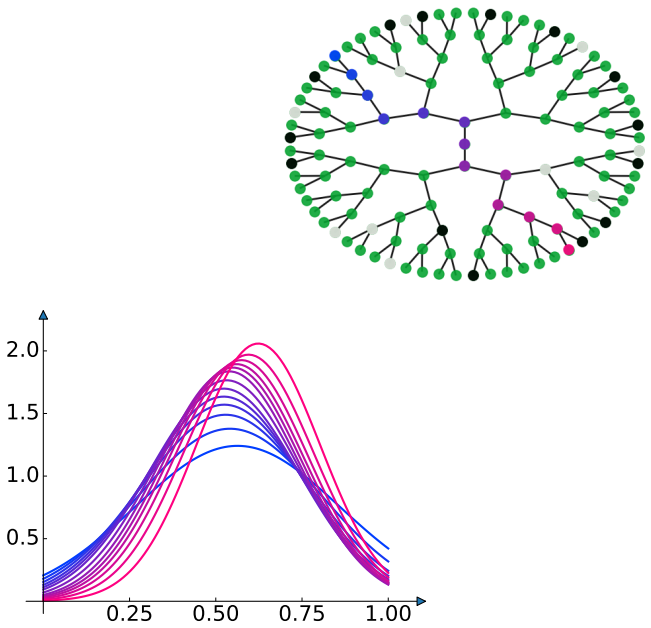
1. [15] all "good" trees embed "well" into the hyperbolic plane,
2. [7, 5] neural nets. into hyperbolic space do well for implementing trees,
3. [3] Represent hyperbolic plane as Gaussian mixtures with "Information Geometry",
4. [1, Example 11] Transformer universal model into hyperbolic space



Comparing with Hyperbolic Embeddings

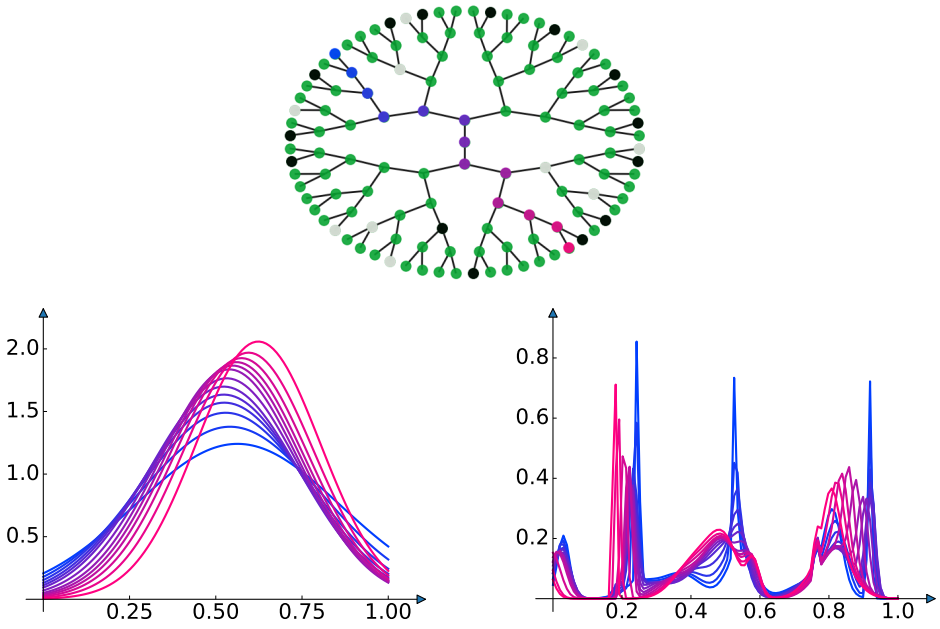


Comparing with Hyperbolic Embeddings



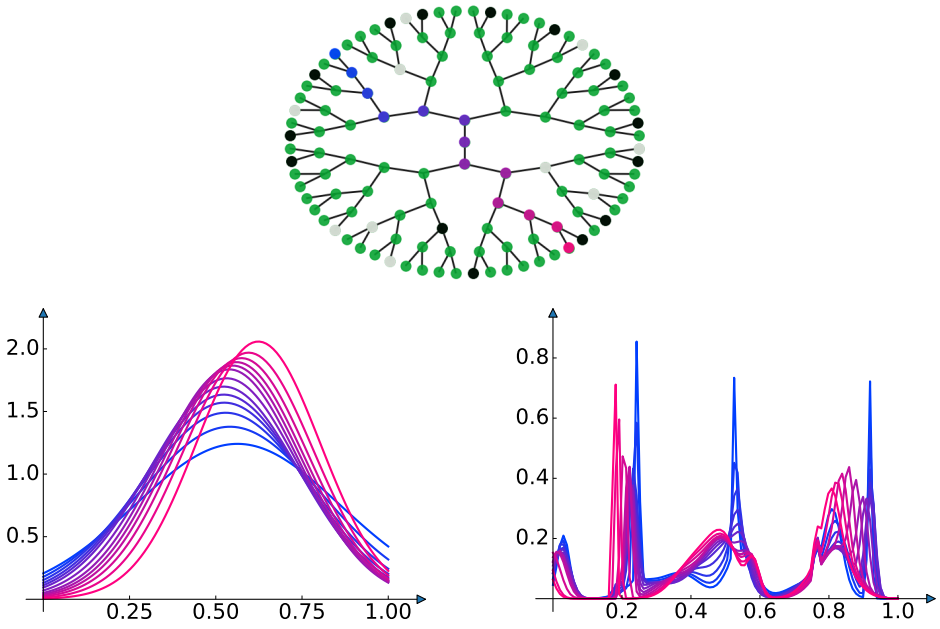
(a) Hyperbolic

Comparing with Hyperbolic Embeddings



(a) Hyperbolic (b) Prob. Trans

Comparing with Hyperbolic Embeddings



(a) Hyperbolic (b) Prob. Trans

Comparting with Hyperbolic Embeddings

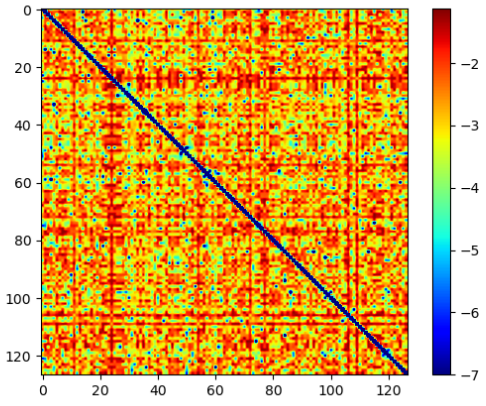


Figure: Hyperbolic Embedding

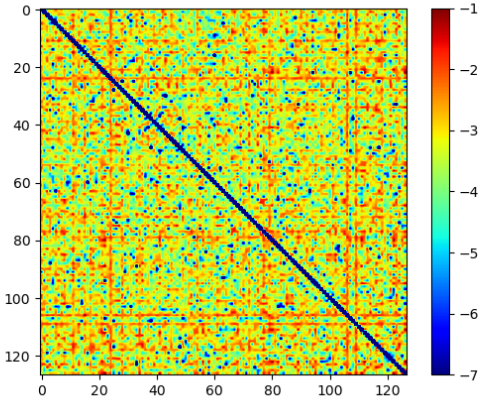


Figure: Probabilistic Transformer

How Stable are the Embeddings?

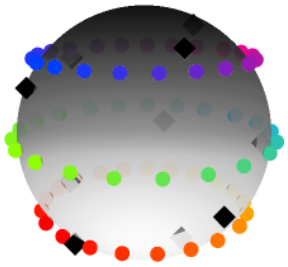


Figure: Point on Sphere

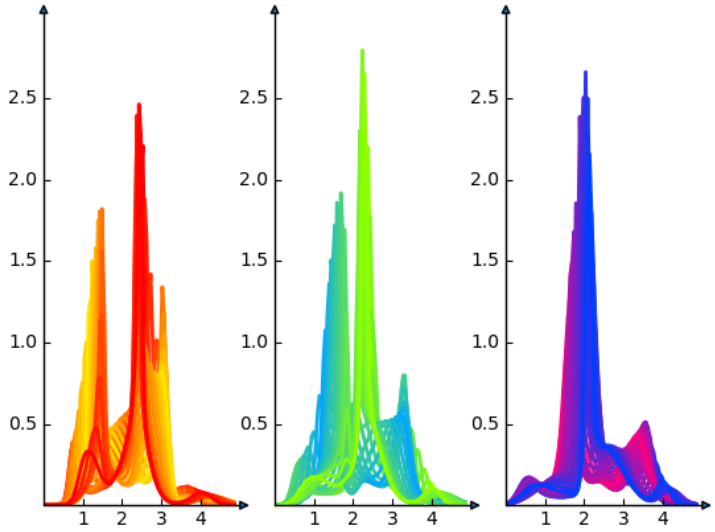
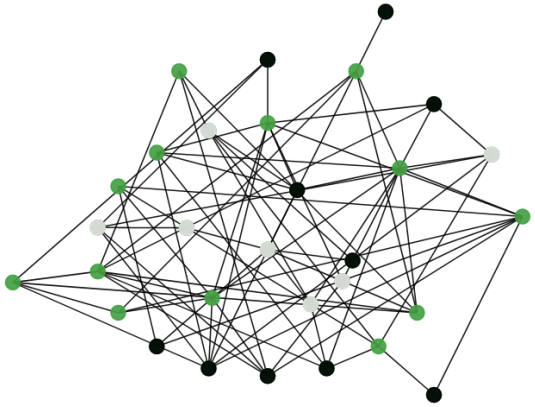


Figure: Feature Space Representation

What Parameters Matter?



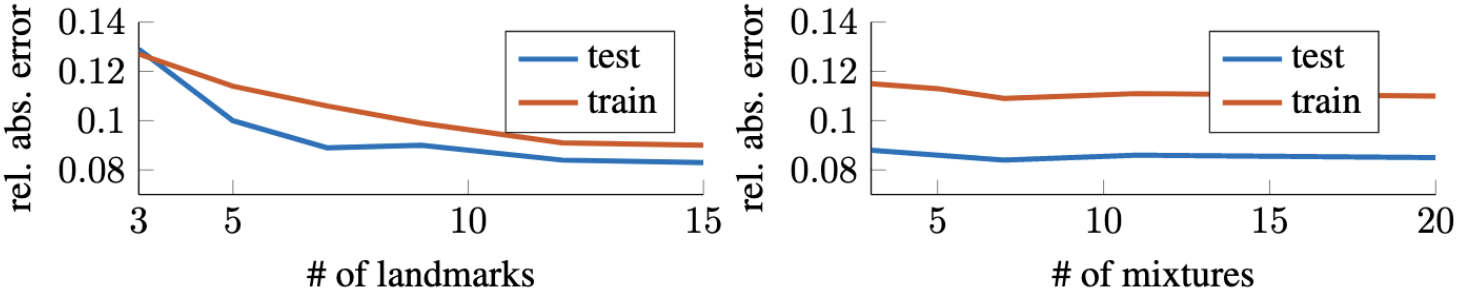


Figure 10: Ablation of parameters on embedding quality.

Summary

1. Represent entire metric space in “good common space”,
2. Implement embeddings with deep neural model,
3. n-point embedding guarantees for “important points”.
4. New proof techniques applicable to other embedding spaces ;)

[1] B. Acciaio, A. Kratsios, and G. Pammer. Metric hypertransformers are universal adapted maps, 2022.

[2] A. Andoni, A. Naor, and O. Neiman. Snowflake universality of Wasserstein spaces. *Ann. Sci. Éc. Norm. Supér. (4)*, 51(3):657–700, 2018. ISSN 0012-9593. doi: 10.24033/asens.2363. URL https://doi.org/10.24033/asens.2363.

[3] N. Ay, J. Jost, H. V. Lê, and L. Schwachhöfer. *Information geometry*, volume 64 of *Ergebnisse der Mathematik und ihrer Grenzgebiete. 3. Folge. A Series of Modern Surveys in Mathematics [Results in Mathematics and Related Areas. 3rd Series. A Series of Modern Surveys in Mathematics]*. Springer, Cham, 2017.

[4] J. Bourgain. On Lipschitz embedding of finite metric spaces in Hilbert space. *Israel J. Math.*, 52(1-2):46–52, 1985. ISSN 0021-2172. doi: 10.1007/BF02776078. URL https://doi.org/10.1007/BF02776078.

[5] C. Cruceru, G. Becigneul, and O.-E. Ganea. Computationally tractable riemannian manifolds for graph embeddings. *Proceedings of the AAAI Conference on Artificial Intelligence*, 35(8):7133–7141, May 2021. URL https://ojs.aaai.org/index.php/AAAI/article/view/16877.

[6] M. Eidi and J. Jost. Ollivier ricci curvature of directed hypergraphs. *Scientific Reports*, 10(1):12466, 2020. doi: 10.1038/s41598-020-68619-6. URL https://doi.org/10.1038/s41598-020-68619-6.

[7] O. Ganea, G. Becigneul, and T. Hofmann. Hyperbolic neural networks. In S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 31, pages 5345–5355. Curran Associates, Inc., 2018.

[8] F. D. Giovanni, G. Luise, and M. Bronstein. Heterogeneous manifolds for curvature-aware graph embedding, 2022.

[9] M. Gromov. Filling Riemannian manifolds. *J. Differential Geom.*, 18(1):1–147, 1983. ISSN 0022-040X. URL http://projecteuclid.org/euclid.jdg/1214509283.

[10] K. U. Katz and M. G. Katz. Bi-Lipschitz approximation by finite-dimensional imbeddings. *Geom. Dedicata*, 150: 131–136, 2011. ISSN 0046-5755. doi: 10.1007/s10711-010-9497-4. URL https://doi.org/10.1007/s10711-010-9497-4.

[11] B. R. Kloeckner et al. A geometric study of wasserstein spaces: isometric rigidity in negative curvature. *International Mathematics Research Notices*, 2016(5): 1368–1386, 2016.

[12] A. Kratsios. Universal regular conditional distributions, 2021.

[13] A. Kratsios, B. Zamanlooy, T. Liu, and I. Dokmanić. Universal approximation under constraints is possible with transformers. In *International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=JGO8CvG5S9.

[14] A. Samal, R. P. Sreejith, J. Gu, S. Liu, E. Saucan, and J. Jost. Comparative analysis of two discretizations of ricci curvature for complex networks. *Scientific Reports*, 8(1): 8650, 2018. doi: 10.1038/s41598-018-27001-3. URL https://doi.org/10.1038/s41598-018-27001-3.

[15] R. Sarkar. Low distortion delaunay embedding of trees in hyperbolic plane. In *International Symposium on Graph Drawing*, pages 355–366. Springer, 2011.

[16] P. Veličković, G. Cucurull, A. Casanova, A. Romero, P. Liò, and Y. Bengio. Graph attention networks. In *International Conference on Learning Representations*, 2018. URL https://openreview.net/forum?id=rJXMpikCZ.