Interpolating Optimal Transport Barycenters of Patient Manifolds

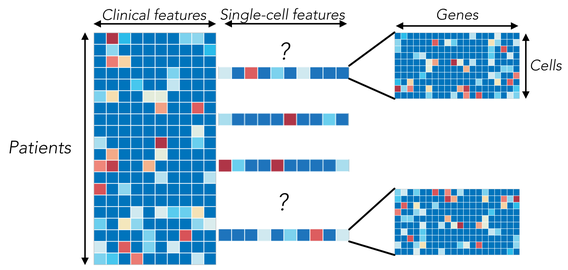
Alexander Tong[1] and Smita Krishnaswamy[1 2]

Yale University Departments of Computer Science[1] and Genetics[2]

Problem Statement

It is now common to get single-cell patient samples in large scale at many timepoints and across disease spectrum.

We tackle the problem of imputing single-cell samples across these states which can improve understanding of disease dynamics by modeling full dynamics and simplify interpretation by summarizing multiple samples.



Background

Wasserstein Barycenters [1] Generalizes averaging of points to averaging of distributions based on a ground distance between points.

Allows interpolation of a distribution from a weighted set of distributions.

$$\text{Weights: } \lambda = \{\lambda_1, \lambda_2, \ldots, \lambda_K\}$$

$$\text{Distance: } d(x,y) = \begin{cases} \|x-y\|_2 \\ \|x-y\|_1 \\ W_p^p(x,y) \end{cases}$$

$$W_p(x,y) = \left(\inf_{\pi \in \Pi(x,y)} \int_{\mathbb{R}^d} D(u,v)^p d\pi(u,v) \right)^{1/p}$$

$$\min_{x \in \mathcal{X}} \sum_{k=1}^{K} \lambda_k d(x, x_k)$$

Barycenters on geometric domains [2] Existing work generalized fast Sinkhorn approximation to the geometric using a geodesic distance on discrete domains.

Using repeated projection with the heat kernel barycenter calculation over a fixed domain is computationally efficient.

$$W_{2,H_t}^2(x,y) = \gamma \left[1 + \min_{\pi \in \Pi} \mathrm{KL}(\pi \mid H_t) \right]$$

Results

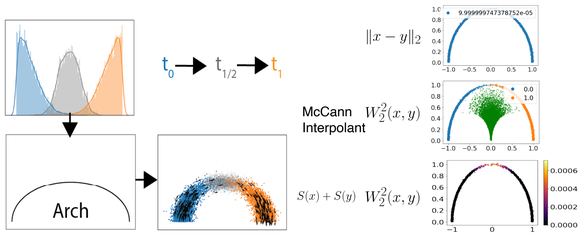


Fig. 1: The choice of support and distributional distance is important on curved manifolds. Here we perform optimal transport on a 1D manifold embedded in 2D

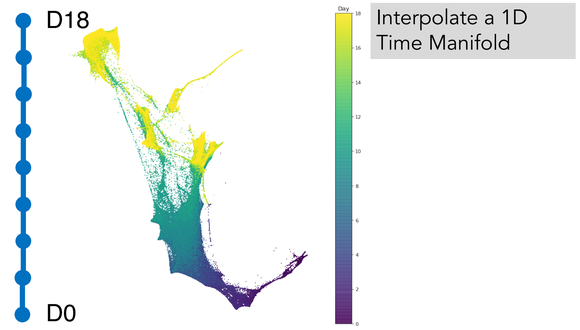


Fig. 2: Data from [3] 315,000 cells in 42 samples over 18 days reprogramming stem cells to diverse populations

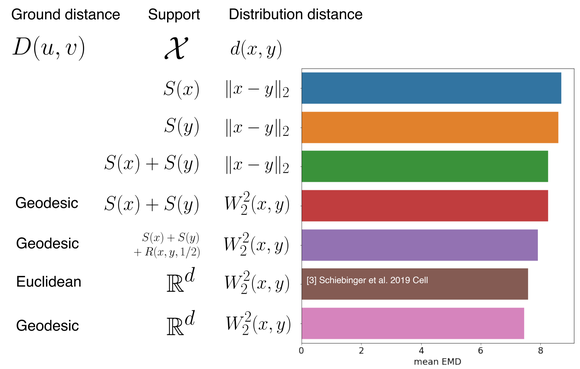


Fig. 3: Shows the mean EMD over all interpolated timepoints with different choices of ground distance, support of interpolated distribution, and the distributional distance

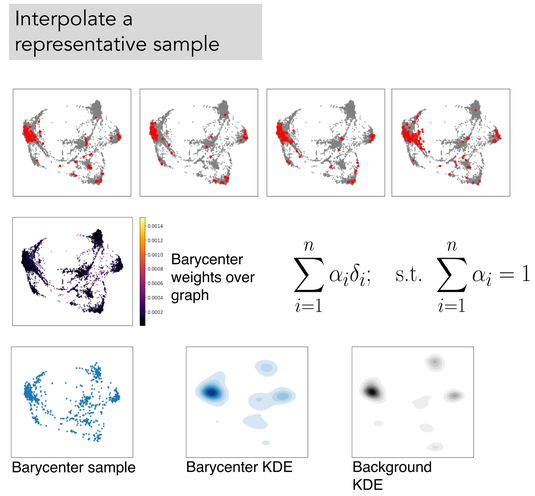


Fig. 4: FACS dataset of Monocytes from 209 samples of healthy and COVID-19 positive samples. Here we compute the barycenter using geodesic ground distance on the support of observed cells of 53 healthy samples to develop a baseline sample.

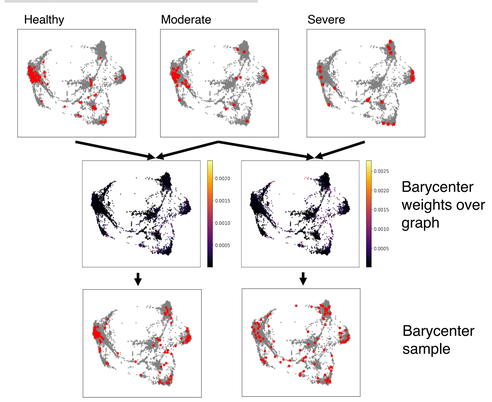


Fig. 5: Interpolating samples along disease state between samples from three patients exhibiting, no symptoms, moderate symptoms and severe symptoms.

Method

We interpolate between multiple distributions using Wasserstein barycenters. There are three major adaptations to the single cell domain:

1. Choice of ground distance

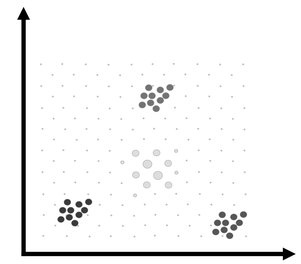
   We use Euclidean distance following [3] and diffusion distance along a constructed graph

2. Choice of distribution distance

   We use the 2-Wasserstein distance for computation and attractive properties on low-dimensional curved manifolds.

3. Support of imputed distribution

   We use the support of input distributions, random interpolations between distributions, and a graph of existing samples.



Conclusions

Barycenters are weighted averages of distributions computed on some support

General barycenter calculation is computationally challenging

For single-cell analysis, support can follow manifold structure improving interpolation and simplifying computation

References

[1] W. S. Chen *et al.*, "Uncovering axes of variation among single-cell cancer specimens," *Nat Methods*, 2020

[2] J. Solomon, *et al.*, "Convolutional Wasserstein Distances: Efficient Optimal Transportation on Geometric Domains," *SIGGRAPH*, 2015.

[3] G. Schiebinger *et al.*, "Optimal-Transport Analysis of Single-Cell Gene Expression Identifies Developmental Trajectories in Reprogramming," *Cell*, 2019.

Further information

Email: alexander.tong@yale.edu

Website: https://www.krishnaswamylab.org

Supported by the Chan-Zuckerberg Initiative