Regularized 2-Wasserstein Barycenters on Discrete Metric Spaces

David Gentile, James M. Murphy
Department of Mathematics
Tufts University

***Abstract*—We apply entropically regularized Wasserstein geometry to the setting of a discrete metric space, where the cost functions are given by different choices of graph metric, which capture varying levels of connectivity information. We find that despite the degeneracy of the Monge-Kantorovich problem in the case of discrete metric spaces, regularized transport leads to geometric objects (e.g., geodesics and more generally barycenters) which convincingly resemble what one would expect from a Wasserstein-type geometry on the space of probability distributions defined for a network. We consider both synthesis (combining measures with respect to the regularized Wasserstein geometry) and analysis (decomposing signals with respect to fixed reference measures) on real geography data, demonstrating the utility of our approach. Our code is available on GitHub[1].**

I. Introduction

At this point in the development of data-driven scientific methods, the significance of network-structured data hardly bears remarking upon. Graphs arise naturally in models of social networks, discrete resource distribution, and segregation patterns [17], and there is growing interest in signal processing methods when the underlying space lacks a continuous structure [18]. In the past three decades, the topic of optimal transportation has become increasingly popular due (at least in part) to its myriad connections with a variety of mathematical fields (e.g., analysis, probability, PDE)[19, 11, 3], and its applications for problems in machine learning (ML) via, for example, Wasserstein GANs [4] or the barycentric coding model (BCM) [20]. While optimal transportation has proven to be a font of ideas giving rise to analytic techniques which have lead to developments in these fields and many others, applications of these ideas to the network setting have proven somewhat tricky, often relying either on restrictive assumptions about the nature of the networks in question [16], or requiring abstract machinery leading to cumbersome numerical methods [10] and hence with as-of-yet limited use.

A. Summary of Contributions

Here, we apply some work originally developed by [6] together with a family of graph metrics which in some sense encode connectivity to implement a version of the barycentric coding model for probability measures, a scheme for encoding probability measures as weighted geometric averages of a given family of "reference" measures in the 2-Wasserstein geometry, and we find that we are reliably able both to synthesize and analyze measures using these tools in a way which resembles the spirit of transport.

[1]https://github.com/dcgentile/fixed-support-barycenters

II. Geometry of Optimal Transport, Entropy Regularization, Degeneracy on a Discrete Metric Space

A. The Monge-Kantorovich Problem & The p-Wasserstein Distance

Letting (Ω, d) be some metric space, $c : \Omega \times \Omega \to \mathbb{R}^+$ be a cost function, and $\mu, \nu \in P(\Omega)$ be probability measures on Ω , the Kantorovich problem is to find a *coupling* γ of μ and ν (i.e., probability distributions on the product space $\Omega \times \Omega$ having marginals given by μ and ν, the set of which is denoted $\Pi(\mu, \nu)$) which solves

$$OT(\mu, \nu) := \inf \left\{ \int_{\Omega \times \Omega} c(x, y) \ \mathrm{d}\gamma : \gamma \in \Gamma(\mu, \nu) \right\}.$$

When $c = d^p$, for some $p \geq 1$ the p^{th} root of the value of the Kantorovich problem is known as the *p-Wasserstein* metric and is denoted by W_p. The 2-Wasserstein metric, in particular, induces a rich geometric structure in the space of probability measures, and of particular interest are geodesic curves and gradient flows in this geometry [12]. In particular, it offers a method of comparing probability distributions in a way that respects the underlying geometry of the base space. For example, if two probability distributions have disjoint support, the L^2 distance between them will be invariant, no matter where the (disjoint) supports lie; on the other hand, the Kullback-Liebler divergence will be infinite, and yield essentially no information at all. Because the 2-Wasserstein geometry preserves geometric information about the space on which the distributions are supported, it is a natural tool of interest for studying the dynamics of distributions on, for example, a fixed geographical region.

It is an unfortunate fact, however, that the Kantorovich problem leads to a degenerate geometry when the underlying metric space is discrete. The geometry so induced is one in which there are no non-trivial geodesic curves; in fact, one can show that for a simple

graph on 2 vertices connected by a single edge, the p-Wasserstein geometry, as characterized by couplings of probability distributions, is isometric to the metric space $([0,1], |x-y|^{1/p})$, from which it follows that every geodesic curve is p-Hölder continuous, with $p \geq 1$ and hence constant (outside the special case where $p = 1$) — in other words, for such a geometry, a curve is a geodesic in $P(\Omega)$ if and only if its image is a singleton. While this problem can be tamed by approaching from a frame of reference beginning with the celebrated Benamou-Brenier formula, and suitable choices and assumptions will lead to the analagous results one might hope to see [13], the computational tools required for working in this regime are totally distinct from the popular methods we will describe below, and are much harder to implement[10].

B. Entropic Regularization of the Monge-Kantorovich Problem

The Kantorovich problem is a linear program, and when the distributions have finite support, standard convex optimization techniques may be employed to identify optimal couplings. Although polynomial in time, algorithms for solving such linear programs have poor scaling, roughly cubic in the support of the measures; however, a convex regularization of the problem, obtained by simultaneously computing the Kullback-Liebler divergence of a given coupling and the product coupling, converts this linear program into a strongly convex one. Consequently, a host of new tools become available from the realm of (strongly) convex analysis, and this results in a dramatic speed-up, shaving off an entire order of magnitude in compute time [9]. If we fix $\varepsilon > 0$ and suppose that $\mu = \sum_{i=1}^{N} \alpha_i \delta_{x_i}, \nu = \sum_{j=1}^{M} \beta_j \delta_{y_j}$ and take C to be a matrix whose entries are given by $C_{ij} = d^2(x_i, y_j)$, one writes for the entropy-regularized transport problem,

$$OT_\varepsilon(\mu, \nu) := \inf \{\langle \gamma, C \rangle + \varepsilon H(\gamma)\}$$

where $\langle \cdot, \cdot \rangle$ is the Frobenius inner product on matrices and $H(\gamma) := -\sum_{i,j} \gamma_{ij}(\log \gamma_{ij} - 1)$. In the limit of $\varepsilon \to 0^+$, we recover the optimal transport cost associated with the measures μ and ν [5]. In figure 1, we illustrate the effect of entropic regularization on the optimal coupling. Without regularization, the plan is simply a permutation matrix, assigning one point in the first distribution to another in the second; with regularization, diffusion is introduced from the entropic term.

C. Wasserstein Barycenters of Measures

Now, given p reference measures $\{\mu_i\}_{i=1}^p \subset P(\Omega)$ and a point in the p-simplex $\vec{\lambda} \in \Delta^p$, a *barycenter* in the 2-Wasserstein metric is a minimizer of the variance functional

$$J_{\vec{\lambda}}(\nu) = \sum_{i=1}^{p} \frac{\lambda_i}{2} W_2^2(\nu, \mu_i).$$

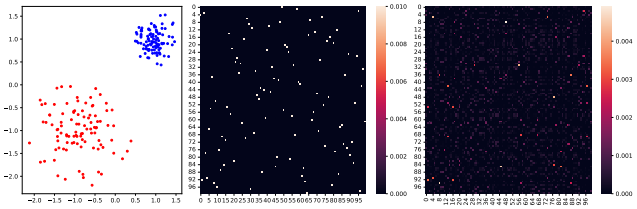

Fig. 1. *Left:* in red, 100 samples drawn from a random variable $X \sim \mathcal{N}(\bar{\mu}_X, \Sigma_X)$, with $\bar{\mu}_X = [-1, -1], \Sigma_X = 0.25 * I$, and in blue 100 samples drawn from a random variable $Y \sim \mathcal{N}(\bar{\mu}_Y, \Sigma_Y)$, with $\bar{\mu}_Y = [1, 1], \Sigma_Y = 0.05 * I$. *Middle*: optimal transport plan for transport red samples to blue samples. *Right*: entropic optimal transport plan taking red samples to blue samples, with regularizing parameter $\varepsilon = 0.05$.

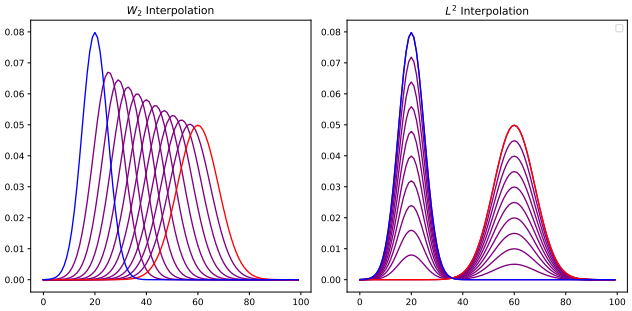

Fig. 2. *Left:* McCann Interpolation of Gaussian measures. *Right:* L^2 Interpolation of Gaussian measures

The theory of 2-Wasserstein barycenters was first explicated in [1], wherein the authors demonstrate existence and uniqueness for the case of distributions are supported on $\Omega = \mathbb{R}^d$, and it supplies us with a notion of geometric mean with respect to the 2-Wasserstein geometry. In that paper, the authors also show that Wasserstein barycenters coincide exactly with the McCann interpolation of measures [14], i.e., for $\lambda = (t, 1-t)$, we recover the unit speed geodesic between the two reference measures in the 2-Wasserstein geometry. Thus Wasserstein barycenters give us a notion of measure interpolation which respects the underlying geometry of the domain, as illustrated in figure 2 — with the 2-Wasserstein geometry, there is a "continuous" movement of mass from one position to the other, where as in the L^2 geometry the mass is "teleported" between its initial and terminal configurations. The computation of 2-Wasserstein barycenters is in general an NP-hard problem [2], but once again entropic regularization has made the computation of regularized barycenters feasible, particularly the realization that they may be computed via Bregman projections [5, 6].

D. The Barycentric Coding Model in Wasserstein Space

Finally, the *the barycentric coding model* for measures in the Wasserstein space presents the problem of identifying a suitable "barycentric basis set" of measures such

that the set

$$Bary(\{\mu_i\}_{i=1}^p) = \left\{\nu : \nu \text{ solves } \inf\left\{J_{\vec{\lambda}}\right\} \text{ for some } \vec{\lambda} \in \Delta^p\right\}$$

adequately represents a dataset comprised of measures. For the BCM, there are two essential operations which are necessary: the so-called "synthesis" and "analysis" problems. For the former, we mean the problem of computing the barycenter ν^* of a given family of reference measures $\{\mu_i\}_{i=1}^p$ for a given weight vector $\vec{\lambda} \in \Delta^p$, and for the latter we mean finding, for a given family of reference measures and a target measures ν, a weight vector $\vec{\lambda}^*$ such that ν is the barycenter of the μ_i with weights $\vec{\lambda}^*$. One can think of this as a starting point for a kind of principal component analysis in the Wasserstein space — although there is no nice linear structure on the space to exploit, Wasserstein barycenters give a method for encoding a family of measures as the "barycentric combinations" of a finite family of measures. The problem of identifying suitable basis measures is also of interest, but not the primary concern of this work.

III. Shortest Path and Diffusion Distances

Since the essence of Wasserstein geometry follows from the metric structure of the underlying space, we should take care to specify what the precise metrics of interest are on our networks. For the remainder, let $\mathcal{X} = (V, E)$ be a graph with vertex set V and edge set $E \subset V \times V$, and let $\omega : E \to \mathbb{R}^+$ be a weighting of the edges. The *shortest path metric* on $\mathcal{X}$ will be defined to be

$$d(u, v) := \inf\left\{\sum_k \omega_{i_k, i_{k+1}}\right\}$$

where the infimum is taken over all uv walks. If $|V| < +\infty$, then the shortest path between any two nodes can be computed in $\mathcal{O}(|V|^2)$ via Dijkstra's algorithm [15]. An alternative metric on the nodes is given by the so-called *diffusion distances*, a family of metrics parameterized by $t > 0$, and which intuitively measures the likelihood that a random walker on the network traverses from u to v in time t [8, 7]. Fix $t > 0$, let Q be the natural Markov transition matrix induced by the structure of the graph, that is,

$$Q_{ij} = \begin{cases} \deg(i)^{-1} & \text{if } (i, j) \in E \\ 0 & \text{otherwise} \end{cases} \tag{1}$$

and let π be the steady state distribution, i.e., the eigenvector of Q with unit eigenvalue, and let $p(y, t|x)$ be the probability of a random walker landing on node y t steps after landing on node x. Then the (squared) diffusion distance between nodes $u, v \in V$ at time t is given by

$$d_t^2(u, v) = \sum_w \frac{(p(w, t|u) - p(w, t|v))^2}{\pi(w)}.$$

As we see in figure 3, as $t \to \infty$, the diffusion distance between a given node and its neighbors decreases in a way reminiscient of heat dispersion on the graph.

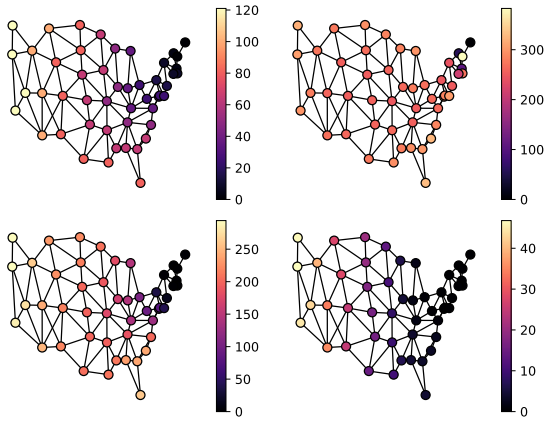

Fig. 3. Color intensity represents the cost of moving from to one fixed node (Maine). Upper left: squared shortest path metric; Upper right: squared diffusion metric at $t = 8$; Lower left: squared diffusion metric at $t = 32$; Lower right: squared diffusion metric at $t = 1024$

IV. Methodology & Numerical Results

Following [6], we use gradient descent to compute barycenters of measures defined on a fixed-support, with the cost matrix realized by 1) the shortest path metric induced by the connectivity of the graph and 2) the diffusion distances between the nodes for various values of t. In other words, fixing a graph, a family of refence measures $\{\mu_i\}_{i=1}^p$, and a prescribed set of barycentric coordinates $\vec{\lambda} = (\lambda_1, \ldots, \lambda_p)$, we compute the measure ν^* which minimizes $J_{\vec{\lambda}}$. With ν^* in hand, we then attempt to recover the prescribed coordinates $\vec{\lambda}$. The network is defined by the 48 contiguous states of the U.S.[2], and demonstrates irregular connectivity (i.e., the degree of the vertices is not constant). Correctness is assured in our computations because the algorithm laid out in [6] simultaneously computes an approximation of the regularized barycenter and the optimal barycentric coordinates for barycentric representation over the dictionary set, and we can check directly that the recovered coordinates of the output barycenter match the input coordinates.

Now, let $\hat{\nu}$ be the barycenter with coordinates $\hat{\lambda}$ be a step along our sequence of iterates in the gradient descent scehem. In minimizing the "barycentric loss" of the coordinates $\hat{\lambda}$ during the process of coordinate recovery (that is, the difference between $\hat{\nu}$ and the target barycenter ν^*, a choice of loss function must be made. Suitable choices are the L^1 and L^2 norms, the Kullback-Liebler divergence, and the total variation between the measures — here we choose the L^2 norm for the loss.

In figures 4, 5, and 6, we solve both the analysis and synthesis problems for some chosen measures on the network. In figures 4 and 5, we have two measures which are concentrated in disparate geographical regions, and

[2]https://hub.arcgis.com/datasets/usdot::states/

using this Wasserstein-like interpolation, transport mass from one locale to the other. Interpreting this geodesic as having barycentric coordinates $(t, 1-t)$, we check our ability to recover the coordinates for $t = 1/3, 2/3$ and find that we do so with modest error. In figure 6 we append an additional reference measure, concentrated away from either of the other two, and synthesize a barycenter as an uneven Wasserstein-like interpolation of the three. As with the geodesic, the algorithm accurately recovers the input coordinates.

We observe natural-looking barycenters arising, particularly for the case $p = 2$, corresponding to geodesic curves in the space of probability measures, and that the barycentric coordinates of these synthesized measures are accurately recovered, with degeneracy occurring in the limit as $\varepsilon \to 0^+$. In figures 4 and 5 we see that

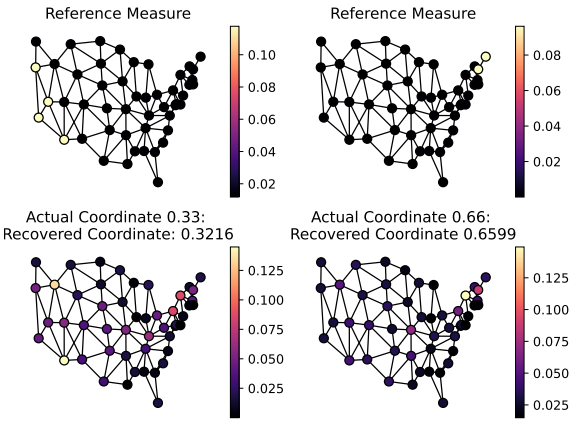

Fig. 4. Wasserstein-like interpolation of measures on graph, with cost given by squared shortest-path distance, $\varepsilon = 0.125$; Upper left, right: reference measures, lower right: computed geodesic with weights $(1/3, 2/3)$; lower right: computed geodesic with weights $(2/3, 1/3)$

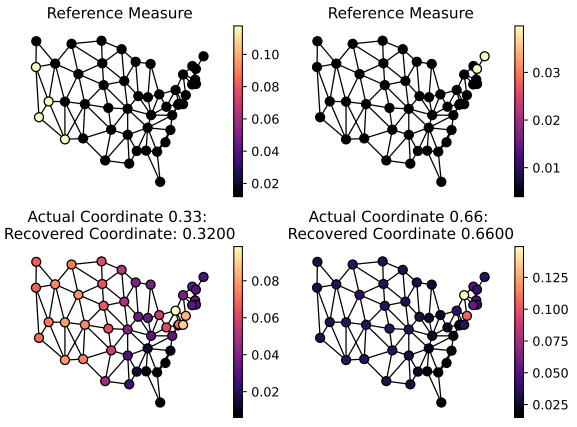

Fig. 5. Wasserstein-like interpolation of measures on graph, with cost given by squared diffusion distance, $t = 16, \varepsilon = 0.125$; Upper left, right: reference measures, lower right: computed geodesic with weights $(1/3, 2/3)$; lower right: computed geodesic with weights $(2/3, 1/3)$

the geodesics induced by this Wasserstein-like geometry inherit the diffusivity of the ground metric. It is interesting that strictly speaking there is no direct access to edge/connectivity information in these metrics, yet we still obtain a motion that has a "continuous" appearance.

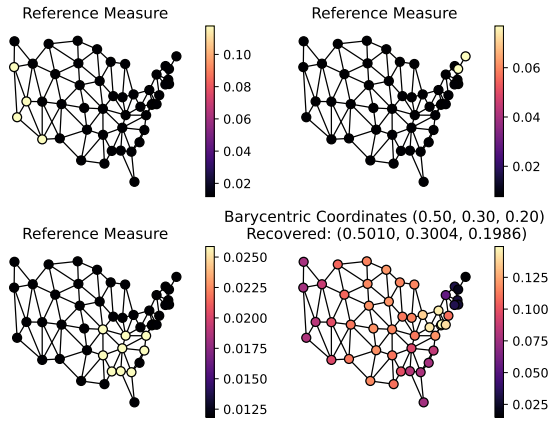

Fig. 6. Wasserstein-like barycenter of measures on graph, with cost given by squared diffusion distance, $t = 16, \varepsilon = 0.125$; Upper left, upper right, lower left: reference measures, lower right: computed barycenter with weights $0.5, 0.3, 0.2)$, and recovered coordinates (0.5010, 0.3004, 0.1986)

V. Discussion

Our experiments demonstrate that despite the degeneracy of the Wasserstein geometry in the graph setting, a measure interpolation that respects the geometry and connectivity of a graph can still be achieved via a combination of entropy-regularized optimal transportation and choice of suitable ground cost, and indeed that the BCM may be deployed both for measure synthesis and decomposition in this setting. This is a promising baseline for the development of transport-based signal processing tools on networks. As a matter of future work, it is of interest to develop an algorithm that can learn an optimal set of dictionary atoms for measure representation in the barycentric coding model. It would also be of interest to investigate any possible relationship between barycenters obtained via regularization of Kantorovich problem and those that respect the formal Riemannian structure on the probability simplex induced by the Benamou-Brenier approach to network transportation, where the Wasserstein-like geometry so developed depends directly on the graph structure rather than on metric information encoding the connectivity.

VI. Acknowledgements

The authors acknowledge generous support from NSF-DMS 2318894.

References

[1] Martial Agueh and Guillaume Carlier. "Barycenters in the Wasserstein Space". In: *SIAM Journal on Mathematical Analysis* 43.2 (2010), pp. 904–924.

[2] Jason M. Altschuler and Enric Boix-Adserà. "Wasserstein Barycenters Are NP-Hard to Compute". In: *SIAM Journal on Mathematics of Data Science* 4.1 (2022), pp. 179–203.

[3] Luigi Ambrosio, Elia Brué, and Daniele Semola. *Lectures on Optimal Transport*. Vol. 130. UNITEXT. Cham: Springer International Publishing, 2021. ISBN: 978-3-030-72161-9 978-3-030-72162-6.

[4] Martin Arjovsky, Soumith Chintala, and Léon Bottou. "Wasserstein generative adversarial networks". In: *International conference on machine learning*. PMLR. 2017, pp. 214–223.

[5] Jean-David Benamou et al. "Iterative Bregman projections for regularized transportation problems". In: *SIAM Journal on Scientific Computing* 37.2 (2015), A1111–A1138.

[6] Nicolas Bonneel, Gabriel Peyré, and Marco Cuturi. "Wasserstein Barycentric Coordinates: Histogram Regression Using Optimal Transport". In: *ACM Transactions on Graphics* 35.4 (2016), pp. 1–10. ISSN: 0730-0301, 1557-7368.

[7] R. R. Coifman et al. "Geometric Diffusions as a Tool for Harmonic Analysis and Structure Definition of Data: Diffusion Maps". In: *Proceedings of the National Academy of Sciences* 102.21 (2005), pp. 7426–7431.

[8] Ronald R. Coifman and Stéphane Lafon. "Diffusion Maps". In: *Applied and Computational Harmonic Analysis*. Special Issue: Diffusion Maps and Wavelets 21.1 (2006), pp. 5–30. ISSN: 1063-5203.

[9] Marco Cuturi. "Sinkhorn Distances: Lightspeed Computation of Optimal Transport". In: *Advances in Neural Information Processing Systems*. Vol. 26. Curran Associates, Inc., 2013.

[10] Matthias Erbar et al. "Computation of Optimal Transport on Discrete Metric Measure Spaces". In: *Numerische Mathematik* 144.1 (2020), pp. 157–200. ISSN: 0945-3245.

[11] Alessio Figalli and Federico Glaudo. *An Invitation to Optimal Transport, Wasserstein Distances, and Gradient Flows*. 2021. ISBN: 9783985470105 9783985475100.

[12] Richard Jordan, David Kinderlehrer, and Felix Otto. "The Variational Formulation of the Fokker–Planck Equation". In: *SIAM Journal on Mathematical Analysis* 29.1 (1998), pp. 1–17. ISSN: 0036-1410.

[13] Jan Maas. "Gradient Flows of the Entropy for Finite Markov Chains". In: *Journal of Functional Analysis* 261.8 (2011), pp. 2250–2292. ISSN: 0022-1236.

[14] Robert J. McCann. "A Convexity Principle for Interacting Gases". In: *Advances in Mathematics* 128.1 (1997), pp. 153–179. ISSN: 00018708.

[15] Alexander Schrijver. "A Course in Combinatorial Optimization". In: *CWI, Kruislaan* 413 (2003), p. 1098.

[16] Justin Solomon. "Optimal Transport on Discrete Domains". In: *AMS Short Course on Discrete Differential Geometry* 3 (2018).

[17] Ljubisa Stankovic et al. *Graph Signal Processing – Part I: Graphs, Graph Spectra, and Spectral Clustering*. 2019. arXiv: 1907.03467.

[18] Ljubisa Stankovic et al. *Graph Signal Processing – Part II: Processing and Analyzing Signals on Graphs*. 2019. arXiv: 1909.10325.

[19] Cédric Villani. *Optimal Transport*. Ed. by M. Berger et al. Vol. 338. Grundlehren Der Mathematischen Wissenschaften. Berlin, Heidelberg: Springer, 2009. ISBN: 978-3-540-71049-3 978-3-540-71050-9.

[20] Matthew Werenski et al. "Measure Estimation in the Barycentric Coding Model". In: *Proceedings of the 39th International Conference on Machine Learning*. PMLR, 2022, pp. 23781–23803.