On Effective Resistance and Optimal Transportation on Graphs

Sawyer Robertson[1,2], Zhengchao Wan[3], Alexander Cloninger[2,3]

June 7, 2024

Networks Journal Club, UCLA

UC San Diego

[1] Presenting author
[2] Department of Mathematics, UCSD
[3] Halıcıoğlu Data Science Institute, UCSD

The Basic Ideas

- This talk will roughly follow our recent preprint[4]
- We primarily concern ourselves with two topics in graph theory: (1) optimal transportation metrics between probability measures on graphs, and (2) effective resistance between nodes on graphs
- We show that these notions can be "interpolated" in a natural way by a family of metrics between probability measures
- In the "$p = 2$" case of this family, there are several novel properties worth exploring
- And these lead to interesting implications for learning on graph data

[4] Sawyer Robertson, Zhengchao Wan, and Alexander Cloninger. "All You Need is Resistance: On the Equivalence of Effective Resistance and Certain Optimal Transport Problems on Graphs". In: *arXiv preprint arXiv:2404.15261* (2024).

Overview

Introduction
- Notation
- Optimal transportation on graphs
- Effective Resistance

Beckmann metrics

Commute Times

Graph Sobolev Spaces

Application: Digit Clustering

Introduction

Basic Notations

- Let $G = (V, E, w)$ be a graph, where $V = \{1, 2, \ldots, n\}$ is the set of vertices, $E \subset \binom{V}{2}$ is a set of undirected edges of size $m \geq 0$. E' are the oriented edges, i.e., $E' = \{(i,j) : i \sim j, i < j\}$.
- $w = (w_{ij})_{i,j \in V}$ is a choice of real edge weights satisfying $w_{ij} \geq 0$, $w_{ij} = w_{ji}$, and $w_{ij} > 0$ if and only if $\{i,j\} \in E$.
- We assume that G is finite, has no multiple edges or loops, and is connected.
- A path in G is an ordered sequence of nodes $P = (i_0, i_1, \ldots, i_k)$ such that $i_\ell \sim i_{\ell+1}$ for $0 \leq \ell \leq k-1$.
- For $i, j \in V$, $d(i,j)$ is the shortest-path distance between the nodes.

Useful matrices

- We define the Adjacency matrix $A \in \mathbb{R}^{n\times n}$ entrywise by

$$A_{ij} = \begin{cases} w_{ij} & \text{if } i \sim j \\ 0 & \text{otherwise} \end{cases}. \tag{1.1}$$

- For each $i \in V$ we define its degree $d_i = \sum_{j\sim i} w_{ij}$. We will also use the diagonal degree matrix $D = \operatorname{diag}(d_1, \dots, d_n) \in \mathbb{R}^{n\times n}$ and the diagonal edge weight matrix $W = \operatorname{diag}(w_{e_1}, \dots, w_{e_m}) \in \mathbb{R}^{m\times m}$.
- We define the incidence matrix $B \in \mathbb{R}^{n\times m}$, with rows indexed by V and columns indexed by E', by the formula:

$$B_{i,e_j} = \begin{cases} 1 & \text{if } e_j = (i, \cdot) \\ -1 & \text{if } e_j = (\cdot, i) \\ 0 & \text{otherwise} \end{cases}. \tag{1.2}$$

- We define the Laplacian matrix L by the formula $L = D - A$, or $L = BWB^T$.

Background: OT I

- Generally speaking, optimal transportation is a class of problems related to finding minimal-cost or minimal-energy methods for transporting mass distributed according to an initial probability measure α to a terminal measure β.[5][6]
- Many, many, many uses: Image processing[7], fluid mechanics[8], computer vision[9], ...

[5] Gabriel Peyré, Marco Cuturi, et al. "Computational optimal transport". In: *Center for Research in Economics and Statistics Working Papers* 1.2017-86 (2017).

[6] Filippo Santambrogio. "Optimal transport for applied mathematicians". In: *Birkäuser, NY* 55.58-63 (2015).

[7] Justin Solomon et al. "Earth mover's distances on discrete surfaces". In: *ACM Transactions on Graphics (ToG)* 33.4 (2014).

[8] Jean-David Benamou and Yann Brenier. "A computational fluid mechanics solution to the Monge-Kantorovich mass transfer problem". In: *Numerische Mathematik* 84.3 (2000), pp. 375–393.

[9] Caroline Moosmüller and Alexander Cloninger. "Linear optimal transport embedding: Provable wasserstein classification for certain rigid transformations and perturbations". In: *arXiv preprint arXiv:2008.09165* (2020).

OT on Graphs

We define the probability measure simplex $\mathcal{P}(V)$ by the set

$$\mathcal{P}(V) := \left\{ \alpha \in \ell(V) : \alpha \geq 0, \sum_{i \in V} \alpha(i) = 1 \right\}. \tag{1.3}$$

For $i \in V$, δ_i is the Dirac or unit measure at node i, identified with the i-th standard basis vector in $\mathbb{R}^n$.

Definition

Let $\alpha, \beta \in \mathcal{P}(V)$, $1 \leq p < \infty$. Define the set of **couplings** between α and β, denoted $\Pi(\alpha, \beta)$, by the following set

$$\Pi(\alpha, \beta) = \left\{ \pi \in \mathbb{R}^{n \times n} : \pi \geq 0, \pi \mathbf{1} = \alpha, \mathbf{1}^T \pi = \beta^T \right\}, \tag{1.4}$$

where $\mathbf{1} \in \mathbb{R}^n$ is the vector containing all ones. We define the p-**Wasserstein distance** between two probability measures, denoted $\mathcal{W}_p(\alpha, \beta)$ by the following optimization problem:

$$\mathcal{W}_p(\alpha, \beta) = \inf \left\{ \left(\sum_{i,j \in V} \pi_{ij} d(i,j)^p \right)^{1/p} : \pi \in \Pi(\alpha, \beta) \right\}. \tag{1.5}$$

OT on Graphs II

- $\mathcal{W}_p$ is a metric on the probability simplex for all $1 \leq p < \infty$
- For this talk we are primarily interested in the case $p = 1, 2$.
- On graphs specifically, $\mathcal{W}_1$ has been used for a variety of things; including graph Ricci curvature[10] and clustering models that use it[11], graph-based approximations to $\mathcal{W}_1$ on other spaces[12], image processing[13], ...
- Something fun happens on graphs for the $\mathcal{W}_1$ problem...

[10] Frank Bauer, Jürgen Jost, and Shiping Liu. "Ollivier-Ricci curvature and the spectrum of the normalized graph Laplace operator". In: *arXiv preprint arXiv:1105.3803* (2011).

[11] Yu Tian, Zachary Lubberts, and Melanie Weber. "Curvature-based clustering on graphs". In: *arXiv preprint arXiv:2307.10155* (2023).

[12] Tam Le et al. "Tree-sliced variants of Wasserstein distances". In: *Advances in neural information processing systems* 32 (2019).

[13] Ernest K Ryu et al. "Vector and matrix optimal mass transport: theory, algorithm, and applications". In: *SIAM Journal on Scientific Computing* 40.5 (2018).

Min Cost Flow Approach

$$\mathcal{W}_1(\alpha, \beta) = \inf \left\{ \sum_e |J(e)| w_e : J : E' \to \mathbb{R},\ BJ = \alpha - \beta \right\} \tag{1.6}$$

- It turns out that $\mathcal{W}_1$ can be expressed as a min cost flow problem, which is a classical computer science problem related to linear programming, network simplex algorithms, ...
- This formulation is sometimes called the "Beckmann problem" on graphs, owing to flow-based formulations of Martin Beckmann[14].
- Gives us access to new approaches in primal-dual methods, regularization, and beyond

[14] Martin Beckmann. "A continuous model of transportation". In: *Econometrica: Journal of the Econometric Society* (1952).

Min Cost Flow Approach

$$\mathcal{W}_1(\alpha, \beta) = \inf \left\{ \sum_e |J(e)| w_e : J : E' \to \mathbb{R},\ BJ = \alpha - \beta \right\} \tag{1.6}$$

- It turns out that $\mathcal{W}_1$ can be expressed as a min cost flow problem, which is a classical computer science problem related to linear programming, network simplex algorithms, ...
- This formulation is sometimes called the “Beckmann problem" on graphs, owing to flow-based formulations of Martin Beckmann[14].
- Gives us access to new approaches in primal-dual methods, regularization, and beyond
- “Why don’t you just square the penalty?" – Too many people

[14] Beckmann, “A continuous model of transportation”.

Effective Resistance

Definition

Let $i, j \in V$ be any two nodes. The **effective resistance** between i, j, denoted r_{ij}, is given by the formula

$$r_{ij} = (\delta_i - \delta_j)^T L^\dagger (\delta_i - \delta_j), \tag{1.7}$$

where $L^\dagger$ is the Moore-Penrose psuedoinverse of L.

- The name originates from its usage in electrical network models, random walks, and the like
- $r_{ij} = \|L^{-1/2}(\delta_i - \delta_j)\|_2^2$ where $L^{-1/2}$ is (abusively) defined as the square root of $L^\dagger$, can also be written spectrally in a nice way
- r_{ij} is a metric on the nodes- not obvious, and very useful
- Amuse-bouche: graph sparsification methods[15], GNNs[16], graph Ricci curvature[17], ...

[15] Daniel A Spielman and Nikhil Srivastava. "Graph sparsification by effective resistances". In: *Proceedings of the fortieth annual ACM symposium on Theory of computing*. 2008.

[16] Mitchell Black et al. "Understanding oversquashing in gnns through the lens of effective resistance". In: *International Conference on Machine Learning*. PMLR. 2023, pp. 2528–2547.

[17] Karel Devriendt and Renaud Lambiotte. "Discrete curvature on graphs from the effective resistance". In: *Journal of Physics: Complexity* 3.2 (2022).

ER and Random Walks

Definition

The **simple random walk** on G is the Markov chain $(X_t)_{t\geq 0}$ on the state space of nodes V with transition probability matrix $D^{-1}A$; that is,

$$\mathbb{P}[X_{t+1} = j | X_t = i] = \begin{cases} \frac{w_{ij}}{d_i} & \text{if } i \sim j \\ 0 & \text{otherwise} \end{cases}.$$

- For $i \in V$, $T_i = \inf\{t \geq 0 : X_t = i\}$ is the hitting time for node i.
- For $i, j \in V$, the commute time is defined by:

$$C(i,j) = \mathbb{E}\left[T_i : X_0 = j\right] + \mathbb{E}\left[T_j : X_0 = i\right].$$

- It holds: $r_{ij} = \frac{1}{\text{vol}(G)} C(i,j)$.

Beckmann metrics

p-Beckmann distance

- The p-Beckmann distance has a simple motivation: what if instead of studying an ℓ_1 penalty in the min cost flow problem between $\alpha, \beta \in \mathcal{P}(V)$, we study an ℓ_p penalty?
- We lose touch with the coupling-based optimal transportation formulation
- And obtain a family of interesting optimal transport metrics

p-Beckmann distance

- The p-Beckmann distance has a simple motivation: what if instead of studying an ℓ_1 penalty in the min cost flow problem between $\alpha, \beta \in \mathcal{P}(V)$, we study an ℓ_p penalty?
- We lose touch with the coupling-based optimal transportation formulation
- And obtain a family of interesting optimal transport metrics

Definition

Let $1 \leq p < \infty$ and $\alpha, \beta \in \mathcal{P}(V)$. Then the **$p$-Beckmann distance** between α, β, denoted $\mathcal{B}_p(\alpha, \beta)$ is given by the following constrained norm optimization problem:

$$\mathcal{B}_p(\alpha, \beta) = \inf \left\{ \left(\sum_{e \in E} |J(e)|^p w_e \right)^{1/p} : J : E' \to \mathbb{R}, \quad BJ = \alpha - \beta \right\} \tag{2.1}$$

Comparing $p = 1$ and $p = 2$

Theorem

Let $\alpha, \beta \in \mathcal{P}(V)$. *It holds:*

1. *When* $p = 1$, $\mathcal{B}_1(\alpha, \beta) = \mathcal{W}_1(\alpha, \beta)$.
2. *When* $p = 2$, $\mathcal{B}_2(\alpha, \beta)^2 = (\alpha - \beta)^T L^\dagger (\alpha - \beta)$.

[18] Peyré, Cuturi, et al., "Computational optimal transport".

Comparing $p = 1$ and $p = 2$

Theorem

Let $\alpha, \beta \in \mathcal{P}(V)$. It holds:

1. *When $p = 1$, $\mathcal{B}_1(\alpha, \beta) = \mathcal{W}_1(\alpha, \beta)$.*
2. *When $p = 2$, $\mathcal{B}_2(\alpha, \beta)^2 = (\alpha - \beta)^T L^\dagger (\alpha - \beta)$.*

- The proof of (1) is well-known and can be found in[18], and (2) is in our preprint; it's short- apply a change of variables using the formula $L = BWB^T$, and then the result follows from properties of $L^\dagger$.
- In some sense, $\mathcal{B}_p$ is an interpolation between $\mathcal{W}_1$ and **effective resistance between probability measures**
- Which also raises the question: What is ER between probability measures? Are there interesting theoretical properties there?

[18] Peyré, Cuturi, et al., "Computational optimal transport".

Some Precedent

- Alamgir and von Luxburg proved a “Dirac" version of this result; but were only focused on nodes as opposed to measures[19].

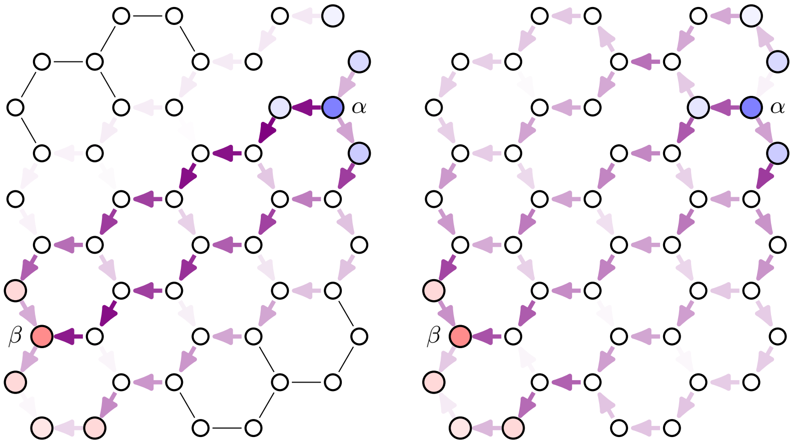
Theorem (Alamgir, von Luxburg, 2011)

Let $i, j \in V$. For brevity put $r_{ij}^{(p)} = \mathcal{B}_p(\delta_i, \delta_j)$. Then:

1. $(p = 1)$, $r_{ij}^{(1)}$ *is the (weighted) shortest path distance between* i, j;
2. $(p = 2)$, $r_{ij}^{(2)}$ *is the effective resistance between* i, j;
3. $(p = \infty)$, *As* $p \to \infty$, $r_{ij}^{(p)} \to 1/mincut(i, j)$.

[19] Morteza Alamgir and Ulrike Luxburg. “Phase transition in the family of p-resistances”. In: *Advances in neural information processing systems* 24 (2011).

Example



(a) $p = 1$,
$\mathcal{B}_1(\alpha, \beta) = \mathcal{W}_1(\alpha, \beta) \approx 9.3$.

(b) $p = 2$, $\mathcal{W}_2(\alpha, \beta) \approx 1.225$,
$\mathcal{B}_2(\alpha, \beta) \approx 1.499$

An informative estimate

Proposition

Let $\alpha, \beta \in \mathcal{P}(V)$. Then

$$\mathcal{B}_2(\alpha,\beta) \le C_1 \mathcal{W}_1(\alpha,\beta) \le C_2 \mathcal{B}_2(\alpha,\beta) \tag{2.2}$$

for some constants C_1, C_2 that do not depend on α, β. If in particular the graph is unweighted, then we have that $C_1 = 1$ and $C_2 = m^{1/2}$, so that

$$\mathcal{B}_2(\alpha,\beta) \le \mathcal{W}_1(\alpha,\beta) \le m^{1/2} \mathcal{B}_2(\alpha,\beta). \tag{2.3}$$

- This estimate shows that $\mathcal{W}_1$ and $\mathcal{B}_2$ are equivalent as metrics.
- Although this bound is a bit brutal, it is sharp. Suppose G is a path on n vertices with $m = n-1$, then the upper and lower bounds are achieved, respectively, when we have:
  1. $\alpha = \delta_1$, $\beta = \delta_n$, so that $\mathcal{B}_2(\delta_1, \delta_n) = \sqrt{n-1}$ and $\mathcal{W}_1(\delta_1,\delta_n) = n-1$ so $\mathcal{W}_1 = m^{1/2}\mathcal{B}_2$.
  2. $\alpha = \delta_1$ and $\beta = \delta_2$, so that $\mathcal{B}_2 = \mathcal{W}_1 = 1$.

Example: Trees

Proposition

Let $T = (V, E, w)$ be a weighted tree, $\alpha, \beta \in \mathcal{P}(V)$, and fix $1 \leq p < \infty$. For an edge $e = (i, j) \in E'$, define K_α by

$$K_\alpha(e = (i,j)) = \sum_{k \in V^*(i;e)} \alpha(k),$$

where $V^*(i; e) \subset V$ is the set of nodes belonging to the subtree with root i obtained from T by removing the edge e (and similarly for K_β). Then it holds

$$\mathcal{B}_p(\alpha, \beta) = \|K_\alpha - K_\beta\|_{w,p}. \tag{2.4}$$

Commute Times

Motivation

- As mentioned earlier, an intresting question comes up: if we “forget" the background of transportation distances, are there things we can say about effective resistance between measures, as opposed to nodes?
- Namely, define $r_{\alpha\beta} = (\alpha - \beta)^T L^\dagger (\alpha - \beta)$
- A useful tool are stopping rules and access times for measures. These are studied in detail by Lovász and Winkler[20], and later, Beveridge[21] across a series of papers from the mid-1990s through the 2010s, most recently.

Definition

A **stopping rule** is a map Γ that associates to each finite path $\omega = (X_0, X_1, \ldots, X_k)$ on G a number $\Gamma(\omega)$ in $[0, 1]$. We can think of $\Gamma(\omega)$ as the probability that we continue a random walk given that ω is the walk so far observed. Alternatively, Γ can be considered a random variable taking values in $\{0, 1, 2, \ldots\}$ whose distribution depends only on the steps $(X_0, X_1, \ldots, X_\Gamma)$.

---

[20] László Lovász and Peter Winkler. “Efficient stopping rules for Markov chains”. In: *Proceedings of the twenty-seventh annual ACM symposium on Theory of computing*. 1995.

[21] Andrew Beveridge. “A hitting time formula for the discrete Green’s function”. In: *Combinatorics, Probability and Computing* 25.3 (2016).

Access Times

Definition

Let $\alpha, \beta \in \mathcal{P}(V)$. The **access time** $H(\alpha, \beta)$ is defined as

$$H(\alpha, \beta) = \inf \{\mathbb{E}[\Gamma | X_0 \sim \alpha] : \Gamma \text{ is a stopping rule and } X_\Gamma \sim \beta\} . \tag{3.1}$$

where for any random variable Y on V, we say $Y \sim \alpha$ if $\mathbb{P}[Y = i] = \alpha(i)$ for $i \in V$. In other words, $H(\alpha, \beta)$ is the minimum mean length of walks that originate with distribution α and terminate according to a stopping rule that achieves distribution β at stopping time. If Γ achieves the inf in $H(\alpha, \beta)$, then Γ is said to be an **optimal stopping rule**.

- The so-called "naïve" stopping rule Γ_n can be obtained from the following construction: at the beginning of the random walk, sample $j \sim \beta$, and stop the walk when $X_{\Gamma_n} = j$. It is readily verified that $X_{\Gamma_n} \sim \beta$, and that

$$\mathbb{E}[\Gamma_n] = \sum_{i,j \in V} \alpha_i \beta_j H(i,j)$$

where for $i, j \in V$, the hitting time $H(i,j)$ is defined by $H(i,j) = H(\delta_i, \delta_j)$ (or, the mean number of steps to reach j from i).

Properties of Access Times

- If we set $H(i,j) = H(\delta_i, \delta_j)$, recall that $r_{ij} = \frac{1}{\mathrm{vol}(G)}(H(i,j) + H(j,i))$.
- This construction leads to a natural conjecture given some known properties of (node) effective resistance...

Conjecture

Let $\alpha, \beta \in \mathcal{P}(V)$. Then does it hold that

$$r_{\alpha\beta} = \frac{1}{\mathrm{vol}(G)}\left(H(\alpha,\beta) + H(\beta,\alpha)\right) \quad ?$$

Properties of Access Times

- If we set $H(i,j) = H(\delta_i, \delta_j)$, recall that $r_{ij} = \frac{1}{\mathrm{vol}(G)}(H(i,j) + H(j,i))$.
- This construction leads to a natural conjecture given some known properties of (node) effective resistance...

Conjecture

Let $\alpha, \beta \in \mathcal{P}(V)$. Then does it hold that

$$r_{\alpha\beta} = \frac{1}{\mathrm{vol}(G)} \left(H(\alpha,\beta) + H(\beta,\alpha)\right) \quad ?$$

- It turns out this is **false**. But we were able to obtain a formula for $r_{\alpha\beta}$ in terms of access times.

Theorem (Generalized Commute Time Formula)

Let $\alpha, \beta \in \mathcal{P}(V)$. Then it holds that:

$$r_{\alpha\beta} = -\frac{1}{\mathrm{vol}(G)} \sum_{i \in V} (\alpha_i - \beta_i)(H(\alpha, i) - H(\beta, i)) \tag{3.2}$$

Access Time Inequalities

- We term the preceding result a "generalized commute time formula," since when α, β are concentrated at nodes $i, j \in V$, it reduces to the commute time representation of r_{ij}.
- Although the conjecture is not true, is it close? Sort of...

Corollary (Measure Commute Time Inequalities)

Let $\alpha, \beta \in \mathcal{P}(V)$. Then $r_{\alpha\beta}$ satisfies the following two inequalities:

$$r_{\alpha\beta} \leq \frac{2}{\text{vol}(G)} \max\{H(\alpha, \beta), H(\beta, \alpha)\} \tag{3.3}$$

$$r_{\alpha\beta} \leq \frac{1}{\text{vol}(G)} (H_n(\alpha, \beta) + H_n(\beta, \alpha)) \tag{3.4}$$

where $H_n(\alpha, \beta) = \mathbb{E}[\Gamma_n]$ (resp. $H_n(\beta, \alpha)$) is the expected duration of the naïve stopping rule with initial distribution α (resp. β) and stopping node sampled from β (resp. α).

Graph Sobolev Spaces

Some Background

- Another perspective on the ℓ_2 problem is through graph Sobolev-type spaces.
- For a bit of background from the continuous setting, we follow Villani[22]
- Recall that if $f : \mathbb{R}^n \to \mathbb{R}$ is a function with a square integrable weak derivative ∇f and $\mathrm{d}\mu = g\mathrm{d}x$ is a Borel probability measure which is absolutely continuous with respect to the Lebesgue measure we can define the Sobolev-type seminorm $\|\cdot\|^2_{\dot{H}^1(\mu)}$ by

$$\|f\|^2_{\dot{H}^1(\mathrm{d}\mu)} = \int_{\mathbb{R}^n} \|\nabla f\|_2^2 \mathrm{d}\mu. \tag{4.1}$$

- The dot $\dot{H}^1(\mathrm{d}\mu)$ serves to distinguish $\|\cdot\|^2_{\dot{H}^1(\mathrm{d}\mu)}$ from a true Sobolev norm, which include a contribution from $\|\cdot\|_{L^2}$.
- We can then define the possibly infinite dual norm to $\|f\|^2_{\dot{H}^1(\mathrm{d}\mu)}$, denoted $\|\cdot\|_{\dot{H}^{-1}(\mathrm{d}\mu)}$ by the following, for any $\mathrm{d}x$-absolutely continuous signed measure $\mathrm{d}\nu = h\mathrm{d}x$:

$$\|\mathrm{d}\nu\|_{\dot{H}^{-1}(\mu)} = \sup\left\{ \int_{\mathbb{R}^n} fh\mathrm{d}\mu : \|f\|_{\dot{H}^1(\mathrm{d}\mu)} \leq 1 \right\}. \tag{4.2}$$

[22] Cédric Villani. *Topics in optimal transportation*. Vol. 58. American Mathematical Soc., 2021.

A Benamou-Brenier-type formula

- Benamou and Brenier[23] introduced an approach to the 2-Wasserstein problem in the continuous setting and obtain a formulation in terms of minimal energy time-dependent density and velocity fields which satisfy certain transport equations.
- Their formula can also be written in an $\dot{H}^{-1}$ form, i.e., as a minimum of Sobolev norms over arcs of measures which satisfify μ, ν initial and terminal conditions. For more, see[24].

Theorem (Benamou-Brenier formula, $\dot{H}^{-1}$ form)

Let μ, ν be Borel probability measures on $\mathbb{R}^n$. Then it holds:

$$\mathcal{W}_2(\mu, \nu) = \inf \left\{ \int_0^1 \|\mathrm{d}\mu_t\|_{\dot{H}^{-1}(\mu_t)} : \mu_0 = \mu, \mu_1 = \nu \right\}. \tag{4.3}$$

[23] Benamou and Brenier, "A computational fluid mechanics solution to the Monge-Kantorovich mass transfer problem".

[24] Rémi Peyre. "Comparison between W2 distance and H- 1 norm, and localization of wasserstein distance". In: *ESAIM: Control, Optimisation and Calculus of Variations* 24.4 (2018).

The Graph setup

- Recall that the map $f \mapsto B^T f : \ell(V) \to \ell(E')$, defined locally by

$$(B^T f)(e = (i,j)) = f(i) - f(j)$$

is often considered the graph gradient operator $B^T = \nabla$ (and similarly, $B = \text{div}$); namely, since $BWB^T = \text{div}\, W\nabla = L$.

Definition

Let $f, g \in \ell(V)$. We define the **graph Sobolev seminorm** $\|\cdot\|_{\dot{H}^1(V)}$ by the equation

$$\|f\|^2_{\dot{H}^1(V)} = \sum_{(i,j)\in E'} w_{ij} |\nabla f(i,j)|^2_2 = \sum_{(i,j)\in E'} w_{ij} |f(i) - f(j)|^2_2. \tag{4.4}$$

We define the (possibly infinite) **dual graph Sobolev norm** $\|\cdot\|_{\dot{H}^{-1}(V)}$ by the supremum

$$\|g\|^2_{\dot{H}^{-1}(V)} = \sup \left\{ f^T g : \|f\|_{\dot{H}^1(V)} \leq 1 \right\}. \tag{4.5}$$

Graph Benamou-Brenier Formula

- For mean zero functions, $\|\cdot\|_{\dot{H}^1(V)}$ and $\|\cdot\|_{\dot{H}^{-1}(V)}$ will be true norms (in particular, the former will be definite and the latter will be finite).

Proposition

Let $f, g \in \ell(V)$. Then the following hold:

1. $\|f\|^2_{\dot{H}^1(V)} = f^T L f$.
2. If $\mathbf{1}^T g = 0$, then $\|g\|^2_{\dot{H}^{-1}(V)} = g^T L^\dagger g$.

- We also observe here that $\mathcal{B}_2(\alpha, \beta) = \|\alpha - \beta\|_{\dot{H}^{-1}(V)}$.
- We say $\mu_t \in C^1([0,1])$ if the map $t \mapsto \mu_t : [0,1] \to \ell(V)$ is continuously differentiable as a map from $[0,1]$ to $\mathbb{R}^n$. We write $\mathrm{d}\mu_t = \frac{\mathrm{d}}{\mathrm{d}s}\mu_s\Big|_{s=t}$.

Theorem (Graph Benamou-Brenier Formula)

Let $\alpha, \beta \in \mathcal{P}(V)$. Then we have

$$\mathcal{B}_2(\alpha, \beta)^2 = \inf\left\{\int_0^1 \|\mathrm{d}\mu_t\|^2_{\dot{H}^{-1}(V)}\,\mathrm{d}t : \mu_t \in C^1([0,1]), \mu_0 = \alpha, \mu_1 = \beta\right\}. \tag{4.6}$$

Application: Digit Clustering

Measures as Data

- A typical classification scenario usually consists of some data $\{x_i\} \subset \mathbb{R}^n$ which one wishes to separate into classes or clusters.
- In many applications[25][26][27], the data x_i can often occur not as vectors in $\mathbb{R}^n$ but as distributions μ_i on $\mathbb{R}^n$.
- For example, consider a hypothetical dataset of images with resolution $k \times \ell$. In this setting, G is the $k \times \ell$ lattice graph, and after normalization each image can be understood as a distribution on G.
- Transportation metrics can be used as the basis for kernel functions or other unsupervised or semi-supervised techniques for differentiating the images.

[25] Alexander Cloninger et al. "People mover's distance: Class level geometry using fast pairwise data adaptive transportation costs". In: *Applied and Computational Harmonic Analysis* 47.1 (2019).
[26] Yin Zhang, Rong Jin, and Zhi-Hua Zhou. "Understanding bag-of-words model: a statistical framework". In: *International journal of machine learning and cybernetics* 1 (2010).
[27] Robert V Bruggner et al. "Automated identification of stratifying signatures in cellular subpopulations". In: *Proceedings of the National Academy of Sciences* 111.26 (2014).

Image Measures on a Lattice

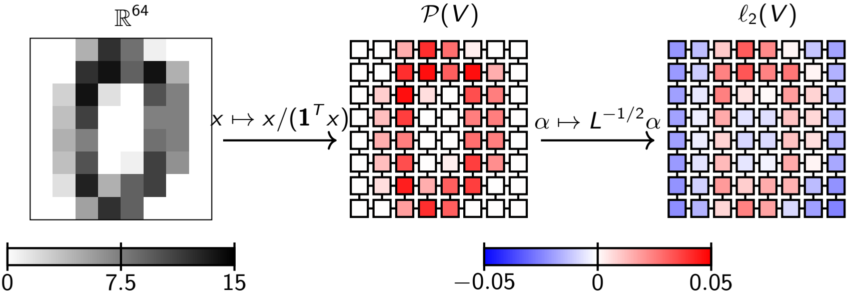


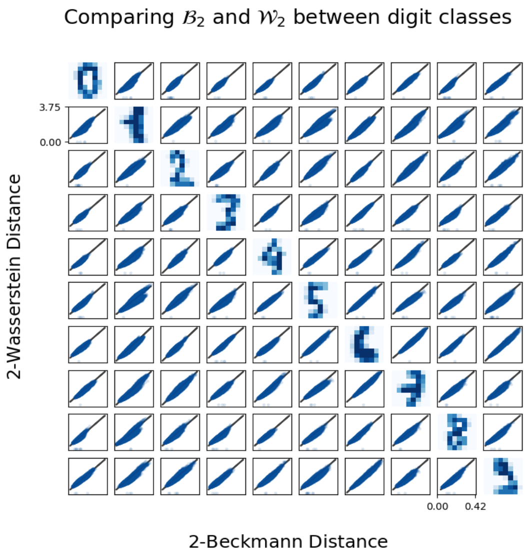
Figure: An illustration of the preprocessing pipeline for the digits data[28], with an example from the class of handwritten zeros. The first step is a mass normalization to convert the pixel values into a fixed-sum distribution viewed on the nodes V of the 8×8 lattice graph. The second step is an embedding $\alpha \mapsto L^{-1/2}\alpha$, such that ℓ_2 distance in the target corresponds to 2-Beckmann distance in $\mathcal{P}(V)$. When computing $\mathcal{W}_2$, we omit the final step.

[28] E. Alpaydin and Fevzi. Alimoglu. *Pen-Based Recognition of Handwritten Digits*. UCI Machine Learning Repository. 1998.

$\mathcal{B}_2$ vs. $\mathcal{W}_2$

- Using the digits dataset, and for each pair of digit classes, we computed the pairwise 2-Beckmann and 2-Wasserstein distances for each pair of samples originating from the respective digit classes (with around 30,000 pairs of distances per pair of digit classes). Within each tile of the grid, we render a scatterplot of the distances over the *overall* linear regression between $\mathcal{B}_2$ and $\mathcal{W}_2$ for the experiment given by $\mathcal{W}_2 \approx 8.446\mathcal{B}_2$.

$\mathcal{B}_2$ vs. $\mathcal{W}_2$



Clustering the Digits

- Using the digits dataset, we demonstrate the results of an unsupervised clustering algorithm with different choices of similarity kernel.
- We built a $k = 42$ nearest neighbor graph on the nodes, and then apply spectral clustering to create predicted classes.
- The text labels of the nodes correspond to the ground truth classes, i.e., digit values. The colors of the nodes on the left (resp. right) are given by the ground truth classes (resp. predicted classes).

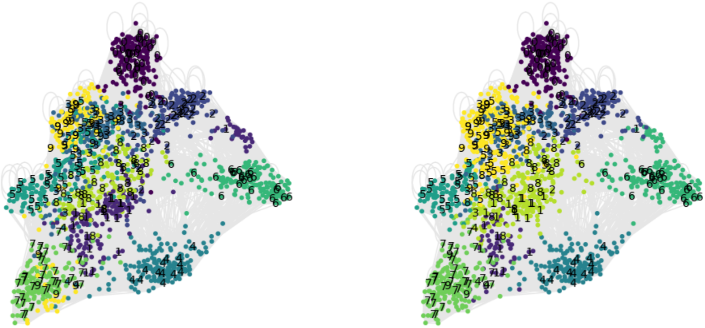
Clustering Performance

Nodes colored by ground truth clusters

Nodes colored by spectral clustering

Figure: Similarity kernel between each image is given by $\exp\{-\mathcal{B}_2(\cdot,\cdot)^2\}$

Clustering Performance II



Nodes colored by ground truth clusters

Nodes colored by spectral clustering

Figure: Similarity kernel between each image is given by $\exp\{-\mathcal{W}_2(\cdot,\cdot)^2\}$

Clustering Performance III

- We evaluate the performance of the unsupervised clustering alogrithm for each kernel. We compare across several metrics, including Rand index (RI) and adjusted Rand index (ARI) ; mutual information (MI) and adjusted mutual information (AMI); and homogeneity (Hom) and completeness (Com).
- In all such cases other than MI, a value of 1.0 corresponds to perfect clustering as compared to the ground truth. Since the predictions depend on a random initialization in the k-means step, we simulated 100 runs of the algorithm and reported the best result for each kernel across the six metrics.

	RI	ARI	MI	AMI	Hom	Com
$\mathcal{B}_2$	0.940	0.685	1.782	0.783	0.774	0.797
$\mathcal{W}_2$	0.935	0.656	1.719	0.755	0.747	0.775

Acknowledgements

- SR wishes to acknowledge financial support from the Halıcıoğlu Data Science Institute through their Graduate Prize Fellowship
- AC was funded by NSF DMS 2012266 and a gift from Intel. GM was supported by NSF CCF-2217058.

- Thanks to Mason for arranging my visit!