All You Need is Resistance: On the Equivalence of Effective Resistance and Certain Optimal Transport Problems on Graphs

Sawyer Robertson[*], Zhengchao Wan[†], and Alexander Cloninger[*†]

April 29, 2024

Abstract

The fields of effective resistance and optimal transport on graphs are filled with rich connections to combinatorics, geometry, machine learning, and beyond. In this article we put forth a bold claim: that the two fields should be understood as one and the same, up to a choice of p. We make this claim precise by introducing the parameterized family of p-Beckmann distances for probability measures on graphs and relate them sharply to certain Wasserstein distances. Then, we break open a suite of results including explicit connections to optimal stopping times and random walks on graphs, graph Sobolev spaces, and a Benamou-Brenier type formula for 2-Beckmann distance. We further explore empirical implications in the world of unsupervised learning for graph data and propose further study of the usage of these metrics where Wasserstein distance may produce computational bottlenecks.

Keywords: effective resistance, optimal transport, graph Laplacians, convex optimization, random walks
MSC: 65K10, 05C21, 90C25, 68R10, 05C50

Contents

1 Introduction and notation **2**
- 1.1 Introduction . . . 2
- 1.2 Outline of this paper . . . 3
- 1.3 Related work . . . 4
- 1.4 Mathematical background . . . 4

2 General properties of the Beckmann problem **8**
- 2.1 Duality theory for the p-Beckmann problem . . . 8
- 2.2 Comparing p-Beckmann and p-Wasserstein distances . . . 9
- 2.3 Worked examples: Paths and trees . . . 11

3 The 2-Beckmann problem: Three perspectives **13**
- 3.1 Effective resistance between measures . . . 13
- 3.2 Sobolev norms and a graph Benamou-Brenier formula . . . 18
- 3.3 A linearized optimal transportation distance . . . 20

4 Acknowledgements **24**

A Proofs **29**
- A.1 Proofs from section 2 . . . 30
- A.2 Proofs from section 3 . . . 32

[*]Department of Mathematics, University of California, San Diego
[†]Halıcıoğlu Data Science Institute, University of California, San Diego

1 Introduction and notation

1.1 Introduction

The theory of optimal transportation (OT), which traces its roots to the Monge formulation as early as 1781 [56] as well as the Kantorovich relaxation articulated in 1942 [41], is typically framed as an optimization problem for some convex or affine function which measures the cost of mass transportation over all feasible plans which provide recipes for transporting some mass distribution to another. OT has received significant attention due to its connections to geometry [85, 33], partial differential equations [7, 30, 29], applied mathematics [63, 71, 72, 32, 69], and many other fields. For discrete measures and domains in particular, OT and Wasserstein distance have been used and studied for a variety of purposes including document retrieval [44], statistics [81, 61, 11], image registration [38], distance approximation [75], political redistricting [55, 15, 1], graph neural networks [16, 17, 92], and graph Ollivier-Ricci curvature [83, 4, 70, 59].

When the underlying metric space of the transportation process is a graph, much attention has been focused on the 1-Wasserstein metric, which can be informally set up as a linear program of the following form:

$$\mathcal{W}_1(\alpha,\beta) = \inf\left\{\sum_{ij}\pi_{ij}d(i,j) : \pi\in\Pi(\alpha,\beta)\right\}, \tag{1}$$

where $\Pi(\alpha,\beta)$ consists of the couplings of two generic probability distributions α,β on the nodes of a graph and $d(i,j)$ is the shortest path metric on the vertex set V (see Definition 1.3 for a complete description). It so happens on graphs that this metric can also be written in a fashion often termed the Beckmann formulation:

$$\mathcal{W}_1(\alpha,\beta) = \inf\left\{\sum_{e}|J(e)|w_e : J : E\to\mathbb{R},\ BJ=\alpha-\beta\right\} \tag{2}$$

where B is the graph incidence matrix, J is any map on the edges of the graph and which is identified as a vector in $\mathbb{R}^{|E|}$, and w_e is the nonnegative weight for edge $e\in E$ (see Definition 1.4 for a complete description). This formulation is named after Martin Beckmann who studied dynamic formulations of OT beginning in the 1950s [5], and whose models for OT in the continuous setting Eq. (2) closely mirrors in a discrete sense. The variable J is sometimes called a feasbile flow between α and β (see Fig. 1), and as such Eq. (2) is also sometimes deemed a flow-based formulation for OT on graphs.

Seemingly separately, effective resistance (ER) on graphs has long been studied for its metric properties between nodes and its connections to many other problems in graph theory. Recall that for two nodes $i,j\in V$, and corresponding standard basis vectors $\delta_i,\delta_j\in\mathbb{R}^{|V|}$, the ER between them is given by $r_{ij} = (\delta_i-\delta_j)^T L^\dagger(\delta_i-\delta_j)$ where $L^\dagger$ is the Moore-Penrose psuedo inverse of the graph Laplacian. Applications and research on ER include foundational work by Spielman and Srivastava on spectral graph sparsifcation using ER [76], as well as spectral clustering models leveraging computational advantages of ER solvers [42]. ER has also been used for computing and studying statistics associated to random walks on graphs [25, 82, 50]. Interestingly, in the case of directed and connection graphs, the opposite is the case as random walk statistics have been used to obtain versions of ER [78, 19]. Moreover, there exist geometric applications involving ER as an ingredient for defining discrete Ricci curvature [22, 23]. For other references which study and utilize effective resistance in various contexts, see, e.g., [25, 40, 34, 91, 36, 12, 18, 10].

This paper is built on an observation which seems to have not yet appeared in the field: that effective resistance, when extended to probability measures on a graph G, can be realized as an OT distance obtained from a Beckmann-type problem with a 2-norm penalty. The resulting metric, while removed from a conventional coupling-based formulation of OT, is intricately connected to the theory of random walks on graphs, the graph Laplacian matrix, and graph Sobolev spaces.

Theorem 1.1 (Informal statement of Theorem 3.4)**.** *Let $G = (V, E, w)$ be a weighted connected graph with Laplacian matrix L and psuedoinverse $L^\dagger$. Denote the oriented vertex-edge incidence matrix of G by B, and let α, β be fixed probability measures on V regarded as vectors in $\mathbb{R}^{|V|}$. For $1 \leqslant p < \infty$, define $\mathcal{B}_p(\alpha, \beta)$ by the following convex optimization problem:*

$$\mathcal{B}_p(\alpha,\beta)^p = \inf\left\{\sum_{e\in E} |J(e)|^p w_e : J : E \to \mathbb{R},\ BJ = \alpha - \beta\right\}. \tag{3}$$

Then it holds:

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

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

The metric introduced in Theorem 1.1 for $p = 2$ is a (squared) norm distance, so its properties and structures are fundamentally different from the (nonlinear) Wasserstein metric in general. However, this metric seems to posses a rich array of properties and characterizations which demonstrate strong homophily with the classical results in OT. This metric also shows promise for use in kernel-based learning methods for data defined on graphs, and in terms of complexity, is significantly less expensive to compute at scale compared to the Wasserstein metric.

Our main contributions are summarized as follows.

1. We introduce the notion of p-Beckmann distance between probability measures, and derive various bounds between p-Beckmann distances and p-Wasserstein distances for measures on general graphs (e.g., Corollary 2.11).

2. We provide an explicit formula for $\mathcal{B}_p$ on trees (Proposition 2.15).

3. We obtain a generalized commute time formula for 2-Beckmann distance, or measure effective resistance, in terms of optimal access times of the simple random walk on a graph (Theorem 3.13).

4. We realize 2-Beckmann distance as a negative Sobolev-type semi-norm distance and use this property to obtain a Benamou-Brenier type formula for 2-Beckmann distance (Theorem 3.22).

5. We apply 2-Beckmann distance to an unsupervised kernel-based clustering method involving handwritten digit data, and compare the results with the 2-Wasserstein kernel; and thereby establish empirical evidence for the usage of a 2-Beckmann kernel on a drop-in basis in place of a 2-Wasserstein kernel for datasets defined on graphs (Fig. 6).

1.2 Outline of this paper

In Section 1.4, we provide relevant mathematical background and notation.

In Section 2, we delve into the general properties of the p-Beckmann problem, including duality theory in Section 2.1 and estimates between Beckmann and Wasserstein distances Section 2.2. In Section 2.3, we highlight some specialized results in the cases of paths and trees.

In Section 3 we focus exclusively in the properties and theory of the 2-Beckmann distance. In Section 3.1 we frame the metric as effective resistance between measures and relate it to optimal stopping times and access times between probability measures on G. In Section 3.2 we frame the metric as a negative Sobolev-type seminorm and obtain a graph Bemaou-Brenier-type formula. Finally, in Section 3.3 we apply the 2-Beckmann distance to an unsupervised learning problem and explore various empirical observations and implications.

Lastly, in Appendix A we provide proofs of the results mentioned in the main component of the paper.

1.3 Related work

In this section we wish to highlight some particularly relevant works on which this paper is based in parts, and in doing so, better situate the novelty of our contributions in a broader context.

The most relevant investigation of OT on graphs, including regularization and convex duality, appears in work by Essid and Solomon [28] and further studies in the field can be found in [75, 74, 62, 57, 67, 49, 65].

The introduction of p-Beckmann distances in Section 2 and its study as an effective resistance metric in Section 3.1 should be understood as extensions of the notion of p-resistances from work of Alamgir and Luxburg [2] to general probability measures (from the case of nodes, i.e., Dirac measures). Such generalizations of resistances and their implications in unsupervised models were also considered by Nguyen and Hamitsuka [60], and Saito and Herbster [68] but only between nodes. The notions of measure access times and optimal stopping times have been studied at length by Lovász, Winkler and Beveridge across several papers [50, 51, 52, 9, 8].

Separately, the notion of using negative Sobolev distance as a linearization of quadratic Wasserstein distance was studied by Greengard, Hoskins, Marshall, and Singer in [37] and is a homophilous framework appearing in the continuous setting. The proof of the Benamou-Brenier-type formula for 2-Beckmann distance is a direct discretization of the proof appearing in [35]. It is also notable that our model of 2-Beckmann distance as a negative Sobolev distance on graphs is parallel to the works by Le, T. Nguyen, Phung, V. A. Nguyen, and Fukumizo and appearing in [46, 47, 48]; but our definitions differ slightly from theirs. Linearized OT and its usage in the unsupervised setting is also explored by Moosmüller and AC [58], and results therein form a basis for our Theorem 3.25 on linear separation of measures in the 2-Beckmann metric space.

1.4 Mathematical background

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 cardinality $m \geqslant 0$, and $w = (w_{ij})_{i,j \in V}$ is a choice of real edge weights satisfying $w_{ij} \geqslant 0$, $w_{ij} = w_{ji}$, and $w_{ij} > 0$ if and only if $\{i, j\} \in E$. In order to ensure the feasibility of optimal transportation problems and simplify the exposition in places, we assume that G is finite, has no multiple edges or loops, and is connected.

For our purposes, 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 $0 \leqslant \ell \leqslant k-1$. If P is a path, $|P|$ is the (integral) length of the path, i.e., $|P| = k$, and $|P|_w$ is the weighted length of P, or the sum of the weights of edges it contains; i.e. $|P|_w = \sum_{\ell=0}^{k-1} w_{i_\ell i_{\ell+1}}$. For any two nodes $i, j \in V$, $\mathscr{P}(i,j)$ is the set of all paths which begin at i and end at j.

In this paper, we will consider several different metrics on V. A metric on V is a map $\kappa : V \times V \to \mathbb{R}$ such that

(i) (Nonnegativity) $\kappa(i,j) \geqslant 0$ for each $i, j \in V$,

(ii) (Definiteness) $\kappa(i,j) = 0$ if and only if $i = j$,

(iii) (Symmetry) $\kappa(i,j) = \kappa(j,i)$ for each $i, j \in V$,

(iv) (Triangle inequality) $\kappa(i,k) \leqslant \kappa(i,j) + \kappa(j,k)$ for each $i, j, k \in V$.

We will primarily use variants of the shortest path metric on V and the metrics induced by effective resistance.

Definition 1.2. Let $1 \leqslant p < \infty$. Define the **weighted p-shortest path metric** on V by the formula

$$d_p(i,j) = \inf \left\{ \left(\sum_{\ell=0}^{k-1} w_{i_\ell i_{\ell+1}}^p \right)^{1/p} : P = (i = i_0, i_1, \ldots, i_k = j) \in \mathscr{P}(i,j) \right\} \tag{4}$$

It is relatively straightforward to verify that d_p satisfies the properties necessary for a metric on V for each p, w.

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

$$A_{ij} = \begin{cases} w_{ij} & \text{if } i \sim j \\ 0 & \text{otherwise} \end{cases}. \tag{5}$$

where the notation $i \sim j$ means $\{i, j\} \in E$. 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 $D = \mathrm{diag}(d_1, \ldots, d_n) \in \mathbb{R}^{n\times n}$ and the diagonal edge weight matrix $W = \mathrm{diag}(w_{e_1}, \ldots, w_{e_m}) \in \mathbb{R}^{m\times m}$.

For technical purposes, we may ocassionally refer to the index-oriented and bi-oriented edge sets E' and E'', respectively, defined below:

$$E' = \{(i, j) : i, j \in V, i \sim j, i < j\} \tag{6}$$

$$E'' = \{(i, j), (j, i) : i, j \in V, i \sim j\}. \tag{7}$$

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:

$$B_{i,e_j} = \begin{cases} 1 & \text{if } e_j = (i, \cdot) \\ -1 & \text{if } e_j = (\cdot, i) \\ 0 & \text{otherwise} \end{cases}. \tag{8}$$

We also occasionally use the matrix $\widetilde{B} \in \mathbb{R}^{n\times 2m}$, with rows indexed by V and columns indexed by E'', defined by an identical formula as in Eq. (8). We define the Laplacian matrix L by the formula $L = D - A$, or $L = BWB^T$. Note that $\widetilde{B}\widetilde{W}\widetilde{B}^T = 2L$, where $\widetilde{W} \in \mathbb{R}^{2m\times 2m}$ is the diagonal matrix of edge weights of E''. We recall that L is symmetric, positive semi-definite, and has a set of nonnegative eigenvalues

$$0 = \lambda_0 \leqslant \lambda_1 \leqslant \ldots \leqslant \lambda_n. \tag{9}$$

We use $\Lambda = \mathrm{diag}(\lambda_0, \lambda_1, \ldots, \lambda_n) \in \mathbb{R}^{n\times n}$ for the diagonal matrix of eigenvalues, and $U = \begin{bmatrix} u_1 & u_2 & \cdots & u_n \end{bmatrix}$ for the respective orthonormal eigenvectors of L. The spectral decomposition of L is given by the equation $L = U\Lambda U^T$.

For any matrix $X \in \mathbb{R}^{\ell\times k}$, we will use the notation $X^\dagger \in \mathbb{R}^{k\times \ell}$ for its Moore-Penrose psuedoinverse. We recall that $X^\dagger$ is the unique matrix satisfying the following four properties [6]:

(i) $XX^\dagger X = X$

(ii) $X^\dagger X X^\dagger = X^\dagger$

(iii) $(XX^\dagger)^T = XX^\dagger$

(iv) $(X^\dagger X)^T = X^\dagger X$

where $(\cdot)^T$ indicates matrix transpose.

In the case of the Laplacian matrix L, we have

$$L^\dagger = U\Lambda^{-1}U^T$$

where, with an abuse of notation, we write

$$\Lambda^{-1} = \mathrm{diag}(\lambda_0 = 0, \lambda_1^{-1}, \ldots, \lambda_n^{-1}) \in \mathbb{R}^{n\times n}. \tag{10}$$

Similarly, we write $L^{-1/2} = U\Lambda^{-1/2}U^T$.

For a finite set S, we define $\ell_2(S)$ to be the linear space of functions $f : S \to \mathbb{R}$ with the standard Euclidean inner product $\langle f, g\rangle_{\ell_2(S)} = g^T f$. We identify $\ell_2(V)$ with $\mathbb{R}^n$ and $\ell_2(E')$ with $\mathbb{R}^m$, respectively.

For a vector $x \in \mathbb{R}^n$ and $1 \leqslant p < \infty$, we define the vector p-norm by

$$\|x\|_p = \left(\sum_{i=1}^{n} |x_i|^p\right)^{1/p}.$$

For any finite set S, in a similar manner as before, we can define $\ell_p(S)$ to be the linear linear space of functions $f : S \to \mathbb{R}$ with the norm $\|\cdot\|_p$. If in some cases we wish to refer simply to the linear space of functions on S, such as when we employ multiple norms at once, we use the notation $\ell(S)$ and specify the norm or inner product in context.

If $J \in \ell(E')$, then we can weight the norm $\|\cdot\|_p$ by the edge weights as follows:

$$\|J\|_{w,p} = \left(\sum_{e\in E'} |J(e)|^p w_e\right)^{1/p}, \tag{11}$$

and similarly for $\ell(E'')$. We will use the subscript of w to indicate the presence of edge weights.

As a matter of notation, functions in $\ell(V)$ will be denoted with lowercase Latin and Greek letters (in the case of measures), and functions in $\ell(E')$ and $\ell(E'')$ will be denoted with uppercase Latin letters.

Note that the matrices B, B^T, and L act on their respective function spaces by matrix multiplication. We summarize their pointwise formulas below.

$$(BJ)(i) = \sum_{e\in E':e=(i,\cdot)} J(e) - \sum_{e\in E':e=(\cdot,i)} J(e), \qquad i \in V, J \in \ell(E') \tag{12}$$

$$(B^T f)(e = (i,j)) = f(i) - f(j), \qquad e \in E', f \in \ell(V) \tag{13}$$

$$(Lf)(i) = \sum_{j\sim i}(f(i) - f(j))w_{ij}, \qquad i \in V, f \in \ell(V) \tag{14}$$

It is worth noting that in some contexts (e.g., [27, 26]), $B(\cdot)$ and $B^T(\cdot)$ are considered the graph equivalents of the divergence and gradient operators, respectively. We will also occasionally use the Dirichlet energy functional for $f \in \ell(V)$, defined by the quadratic form of the Laplacian matrix:

$$f^T L f = \sum_{(i,j)\in E'} w_{ij}(f(i) - f(j))^2 = \|B^T f\|_{w,2}^2 \tag{15}$$

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

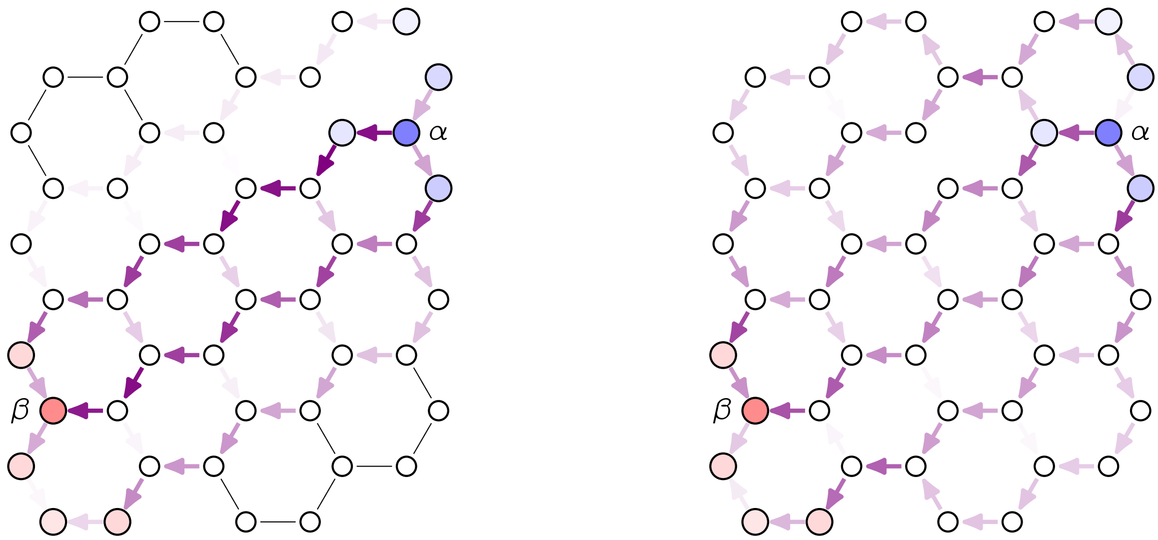
$$\mathcal{P}(V) := \left\{\alpha \in \ell(V) : \alpha \geqslant 0, \sum_{i\in V} \alpha(i) = 1\right\}. \tag{16}$$

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

The primary objects of study for this paper are two different optimal transportation distances between probability measures on V. The first, which is classical and extensively studied on graphs, is p-Wasserstein distance.

Definition 1.3. Let $\alpha, \beta \in \mathcal{P}(V)$, $1 \leqslant p < \infty$, and k any metric on V. Define the set of **transportation couplings** between α and β, denoted $\Pi(\alpha, \beta)$, by the following set

$$\Pi(\alpha,\beta) = \left\{\pi \in \mathbb{R}^{n\times n} : \pi \geqslant 0, \pi\mathbf{1} = \alpha, \mathbf{1}^T\pi = \beta^T\right\}, \tag{17}$$



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

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

Figure 1: A side-by-side comparison of optimal flows for the 1-Beckmann and 2-Beckmann problems on a 4×4 hexagonal lattice graph. The masses of α, β at each node are rendered proportionally to opacity; and similarly the optimal flow values at each edge are rendered proportionally to opacity. The arrows indicate orientation of the flow value; i.e., $\circ \to \circ'$ if the optimal flow J satisfies $J(\circ, \circ') > 0$ and $\circ \leftarrow \circ'$ if $J(\circ, \circ') < 0$.

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

$$\mathcal{W}_{k,p}(\alpha, \beta) = \inf \left\{ \left(\sum_{i,j \in V} \pi_{ij} k(i,j)^p \right)^{1/p} : \pi \in \Pi(\alpha, \beta) \right\}. \tag{18}$$

As a matter of convention, when $k = d_1$, that is, the weighted 1-shortest-path metric, we write $\mathcal{W}_{d_1,p} = \mathcal{W}_p$. The second optimal mass transportation distance is much less studied in its own right except in the cases of $p = 1$ and to a seemingly much lesser extent, $p = 2$.

Definition 1.4. Let $1 \leqslant p < \infty$ and $\alpha, \beta \in \mathcal{P}(V)$. Define the set of **feasible edge flows** between α and β, denoted $\mathcal{J}(\alpha, \beta)$, by the affine region

$$\mathcal{J}(\alpha, \beta) = \left\{ J \in \ell_2(E') : BJ = \alpha - \beta \right\}. \tag{19}$$

Then the p-**Beckmann distance** between α, β, denoted $\mathcal{B}_p(\alpha, \beta)$ is given by the following constrainted norm optimization problem:

$$\mathcal{B}_p(\alpha, \beta) = \inf \left\{ \|J\|_{w,p} : J \in \mathcal{J}(\alpha, \beta) \right\} \tag{20}$$

Note that $\mathcal{J}(\alpha, \beta) \neq \varnothing$ for any α, β since the column space of B is exactly the mean zero functions on V. The p-Wasserstein and p-Beckmann problems are essentially two different perspectives on optimal transportation for measures on graphs. The "Wasserstein philosophy" being that mass transportation is tracked between *all pairs of nodes*, and is accounted according to d_{ij}^p. We can also think of this philosophy as one of mass teleportation- the primal variable, or coupling, does not reveal *how* mass π_{ij} moves from i to j, only that it did so along a shortest path somewhere, or a combination of shortest paths. The "Beckmann philosophy" is the viewpoint that mass must move *along edges* and the transportation should be accounted in terms of edge flows, which are really just signed mass values on each oriented edge. In this formulation, the penalty lies on $|J(e)|^p$, so transportation flows can often be non-sparse, or occur along weighted combinations of shortest and non-shortest paths. Comparisons of optimal flows for values $p = 1, 2$ in an example setting are shown in Fig. 1.

Remark 1.5. A slightly different way to formulate Definition 1.4 would be through E'' with a nonnegativity constraint. That is, let

$$\mathcal{J}''(\alpha,\beta) = \left\{ J \in \ell_2(E'') : J \geqslant 0, \widetilde{B}J = \alpha - \beta, \right\}. \tag{21}$$

and then,

$$\mathcal{B}_p(\alpha,\beta) = \inf \left\{ \|J\|_{w,p} : J \in \mathcal{J}''(\alpha,\beta) \right\}. \tag{22}$$

It is not hard to show that these two formulations arrive at the same value. Definition 1.4 simply allows for signed flows; for example, a flow value of -1 along the oriented edge $(1, 2)$ represents one unit of mass transported from node 2 to node 1. We opt to drop the nonnegativity constraint for this paper and focus on the former formulation as it simplifies the calculations slightly, but we may have occassion to use this formulation instead.

In the Proposition below, we summarize two relevant properties of $\mathcal{W}_p$ and $\mathcal{B}_p$.

Proposition 1.6. *Let $\alpha, \beta \in \mathcal{P}(V)$.*

(i) The objective functions for both $\mathcal{W}_p(\alpha,\beta)$ and $\mathcal{B}_p(\alpha,\beta)$ are convex for all $1 \leqslant p < \infty$, and the infima are always achieved.

(ii) The objective for $\mathcal{B}_p(\alpha,\beta)$ is strictly convex for $p > 1$. The infimum is always achieved and is unique in the case $p > 1$.

(iii) $\mathcal{W}_p$ and $\mathcal{B}_p(\alpha,\beta)$ define metrics on $\mathcal{P}(V)$.

Proofs of items (i) and (ii) are straightforward since the objectives are defined either by affine functions or common norms whose convexity properties are well-understood. For (iii), a proof for $\mathcal{W}_p$ can be found in [62, Ch. 2], and the case of $\mathcal{B}_p$ is relatively straightforward.

2 General properties of the Beckmann problem

In this section, we look at a few interesting properties of the Beckman problem $\mathcal{B}_p(\alpha,\beta)$ for general $1 \leqslant p < \infty$. In Section 2.1 we derive the Lagrangian dual to the p-Beckmann problem, in Section 2.2 we obtain bounds that relate certain Beckmann and Wasserstein distances, and finally, in Section 2.3, we work through some example results on paths and trees to highlight examples of Wasserstein and Beckmann distances.

2.1 Duality theory for the p-Beckmann problem

In this subsection we study the convex duality of the p-Beckmann problem. In [2] Alamgir and Luxburg introduced formulations of p-resistances, that is, generalizations of effective resistance obtained by Beckmann-type min cost flow problems penalized by a p-norm. They studied the primal and dual forms of these problems, and thereby obtained so-called potential-based formulations of the resistance problems. As described in Section 1.3, the results presented in this section may be considered extensions of the results concerning p-resistances to the case of general probability measures; namely, we understand the p-resistance between generic nodes i, j to be p-Beckmann distance between Dirac measures at i, j. In Theorem 2.1 we obtain the dual formulation of the p-Beckmann optimization problem, and thus generalize the Kantorovich-Rubenstein duality on graphs to all $1 \leqslant p < \infty$ (see Remark 2.3).

Theorem 2.1. *Let $\alpha, \beta \in \mathcal{P}(V)$ and $1 < p < \infty$ be fixed. Let q be the conjugate of p so that $\frac{1}{p} + \frac{1}{q} = 1$. Then the dual to the p-Beckmann problem, given by*

$$\mathcal{B}_p(\alpha, \beta) = \inf \left\{ \|J\|_{w,p} : J \in \ell(E'), BJ = \alpha - \beta \right\}, \tag{23}$$

is given by the maximization problem

$$\mathcal{B}_p(\alpha, \beta) = \sup \left\{ \phi^T(\alpha - \beta) : \phi \in \ell(V), \left\| B^T \phi \right\|_{w^{1-q},q} \leqslant 1 \right\}, \tag{24}$$

and strong duality holds. In the special case of $p = 1$ and $q = \infty$, we have

$$\mathcal{B}_1(\alpha, \beta) = \inf \left\{ \|J\|_{w,1} : J \in \ell(E'), BJ = \alpha - \beta \right\} = \sup \left\{ \phi^T(\alpha - \beta) : \phi \in \ell(V), \left\| B^T \phi \right\|_{w^{-1},\infty} \leqslant 1 \right\} \tag{25}$$

We defer the proof to Appendix A.1.

Remark 2.2. If we specialize to the case where the edge weights are all unit value; i.e., G is unweighted, then the weights can be dropped from the preceding results. In particular, the dual norm to $\|\cdot\|_p$ is simply $\|\cdot\|_q$ and

$$\mathcal{B}_p(\alpha, \beta) = \inf \left\{ \|J\|_p : J \in \ell(E'), BJ = \alpha - \beta \right\} = \sup \left\{ \phi^T(\alpha - \beta) : \phi \in \ell(V), \left\| B^T \phi \right\|_q \leqslant 1 \right\} \tag{26}$$

Remark 2.3. The case $p = 1$ and $q = \infty$ is the conventional Kantorovich-Rubenstein duality on graphs [89], which can be more conventionally written as

$$\inf \left\{ \sum_{e \in E'} |J(e)| w_e : J \in \ell(E'), BJ = \alpha - \beta \right\} = \sup \left\{ \phi^T(\alpha - \beta) : \phi \in \ell(V), \|\phi\|_{\mathrm{Lip}} \leqslant 1 \right\} \tag{27}$$

where the Lipschitz seminorm $\|\phi\|_{\mathrm{Lip}} = \|B^T \phi\|_{w^{-1},\infty}$ on $\ell(V)$ is defined by

$$\|\phi\|_{\mathrm{Lip}} = \max \frac{|\phi(i) - \phi(j)|}{d(i,j)} = \max_{i \sim j} \frac{|\phi(i) - \phi(j)|}{w_{ij}}. \tag{28}$$

Note also that this duality occurs in a homophilous manner for Beckmann's minimal flow problem in the continuous setting [71, Sec 4.2], which is the namesake for the minimal flow formulations on graphs.

2.2 Comparing p-Beckmann and p-Wasserstein distances

In this section, we establish some bounds between $\mathcal{B}_p$ and $\mathcal{W}_p$ in certain situations. Section 2.2.1 introduces a technique for obtaining an edge flow from a given coupling with some control on the cost, and uses it to derive elementary bounds for general p; and in Section 2.2.2 we focus on the case of $\mathcal{W}_1$ and obtain sharp estimates related to $\mathcal{B}_2$, the latter of which is of particular interest in the ensuing sections.

2.2.1 From coupling to flow

An essential ingredient in comparing the optimal values of transportation metrics is having the ability to convert a coupling π to a feasible edge flow J, and vice-versa. In this subsection we will make use of the bi-oriented edge flow formulation of $\mathcal{B}_p$, which was discussed in Remark 1.5.

Definition 2.4. Let $i, j \in V$. Let $P \in \mathscr{P}(i,j)$ be a path between i, j. Write $P = (i = i_0, i_1, \ldots, i_k = j)$. Define

$$I_P = \sum_{\ell=0}^{k-1} \delta_{(i_\ell, i_{\ell+1})} \in \ell(E''). \tag{29}$$

Proposition 2.5. *Let $\alpha, \beta \in \mathcal{P}(V)$, $1 \leqslant p < \infty$, and suppose $\pi \in \Pi(\alpha, \beta)$. For each pair of nodes $i, j \in V$ with $i \neq j$, let $P_{ij} \in \mathscr{P}(i,j)$ be a fixed path between i, j. Define the edge flow $J^\pi = J^{\pi,(P_{ij})} \in \ell(E'')$ by the expression*

$$J^\pi = \sum_{i \neq j \in V} \pi_{ij} I_{P_{ij}} \tag{30}$$

Then J^π is a feasible flow for $\mathcal{B}_p(\alpha,\beta)$, i.e., $J^\pi \geqslant 0$ and $BJ^\pi = \alpha - \beta$, and the following estimate holds:

$$\|J^\pi\|_{w,p} \leqslant \sum_{i,j \in V} \pi_{ij} \|I_{P_{ij}}\|_{w,p}. \tag{31}$$

We include a proof in Appendix A.1. We can use Proposition 1.6 to derive at least two estimates which relate $\mathcal{B}_p$ to $\mathcal{W}_{k,p}$ for different choices of k. The first is somewhat brutal, but is informative for p small.

Theorem 2.6. *Let $\alpha, \beta \in \mathcal{P}(V)$, $1 \leqslant p < \infty$, and assume that $w_{ij} \geqslant 1$ for each $i, j \in V$ such that $i \sim j$. Then*

$$\mathcal{B}_p(\alpha, \beta) \leqslant \mathcal{W}_p(\alpha,\beta)^p \tag{32}$$

Proof. Let $\pi \in \Pi(\alpha,\beta)$ be an optimal coupling for $\mathcal{W}_p(\alpha,\beta)$. For each i, j, let $P_{ij} \in \mathscr{P}(i,j)$ be a choice of path which is minimal in the sense of d_1, and let $J^\pi = J^{\pi,(P_{ij})}$ be the feasible flow as in Proposition 2.5. Then

$$\mathcal{B}_p(\alpha,\beta) \leqslant \|J^\pi\|_{w,p} \leqslant \sum_{i,j\in V} \pi_{ij}\|I_{P_{ij}}\|_{w,p} \leqslant \sum_{i,j \in V} \pi_{ij} d_1(i,j)^p = \mathcal{W}_p(\alpha,\beta)^p \tag{33}$$

since $\|\cdot\|_{w,p} \leqslant \|\cdot\|_{w,1}^p$ provided $w_{ij} \geqslant 1$. $\square$

If we allow for some flexibility when picking the underlying metric k of $\mathcal{W}_{k,p}$, then we can provide a tighter estimate when $\mathcal{W}$ is large.

Theorem 2.7. *Let $\alpha, \beta \in \mathcal{P}(V)$, $1 < p < \infty$ and let q be the conjugate of p such that $\frac{1}{p} + \frac{1}{q} = 1$. Then*

$$\mathcal{B}_p(\alpha,\beta) \leqslant n^{2/q} \mathcal{W}_{d_p,p}(\alpha,\beta). \tag{34}$$

The proof of Theorem 2.7 follows mainly from Hölder's inequality and Proposition 2.5; we provide the proof in Appendix A.1. Note in particular that when $p = 2$, we have $q = 2$ and thus it holds $\mathcal{B}_2(\alpha,\beta) \leqslant n\mathcal{W}_{d_2,2}(\alpha,\beta)$.

2.2.2 Estimating from the $p = 1$ case

When $p = 1$ and the choice of metric is $k = d_1$ there is a direct overlap between $\mathcal{B}_1$ and $\mathcal{W}_1$ [62, 28].

Theorem 2.8. *Let $\alpha, \beta \in \mathcal{P}(V)$. Then*

$$\mathcal{B}_1(\alpha,\beta) = \mathcal{W}_1(\alpha,\beta).$$

For a proof of this result, we recommend the exposition in [62, Ch. 6]. Note that for $p \geqslant 1$ and $J \in \ell(E')$, we have that since $\|\cdot\|_p \leqslant \|\cdot\|_1$ in general,

$$\|J\|_{w,p} = \|W^{1/p}J\|_p \leqslant \|W^{1/p}J\|_1 \leqslant \|J\|_{w,1}\left(\max_{e\in E'} w_e^{1/p-1}\right).$$

Thus if $C_{w,p} = \max_{e \in E'} w_e^{1/p-1}$ we have $\|J\|_{w,p} \leqslant C_{w,p}\|J\|_{w,1}$. If J^* is an optimal flow for $\mathcal{B}_1(\alpha,\beta)$, we have

$$\mathcal{B}_p(\alpha,\beta) \leqslant \|J^*\|_{w,p} \leqslant C_{w,p}\|J^*\|_{w,1} = C_{w,p}\mathcal{B}_1(\alpha,\beta) \tag{35}$$

Therefore, we have the following two estimates for any choice of α, β and $p \geqslant 1$.

Corollary 2.9. *Let $\alpha, \beta \in \mathcal{P}(V)$ and $1 \leqslant p < \infty$. Then the following estimates hold:*

$$\mathcal{B}_p(\alpha,\beta) \leqslant C_{w,p}\mathcal{B}_1(\alpha,\beta) \tag{36}$$

$$\mathcal{B}_p(\alpha,\beta) \leqslant C_{w,p}\mathcal{W}_1(\alpha,\beta) \tag{37}$$

Moreover, via similar logic as before, since $\|\cdot\|_1 \leqslant m^{1-1/p}\|\cdot\|_p$ on $\ell(E')$, where we recall $m = |E|$, we have for $J \in \ell(E')$,

$$\|J\|_{w,1} = \|WJ\|_1 \leqslant m^{1-1/p}\|WJ\|_p \leqslant m^{1-1/p}\|W\mathbf{1}_{E'}\|_\infty^{\frac{p-1}{p}}\|J\|_{w,p} \tag{38}$$

Therefore if we set $C_{w,m,p} = m^{1-1/p}\|W\mathbf{1}_{E'}\|_\infty^{\frac{p-1}{p}}$, we have $\|J\|_{w,1} \leqslant C_{w,m,p}\|J\|_{w,p}$ and using a similar argument as before, the following corollary holds.

Corollary 2.10. *Let $\alpha, \beta \in \mathcal{P}(V)$ and $1 \leqslant p < \infty$. Then the following estimates hold:*

$$\mathcal{B}_1(\alpha,\beta) \leqslant C_{w,m,p}\mathcal{B}_p(\alpha,\beta) \tag{39}$$

$$\mathcal{W}_1(\alpha,\beta) \leqslant C_{w,m,p}\mathcal{B}_p(\alpha,\beta) \tag{40}$$

An application of this allows us to relate 2-Beckmann distance to $\mathcal{W}_1$, which we state as a corollary below.

Corollary 2.11. *Let $\alpha, \beta \in \mathcal{P}(V)$. Then*

$$\mathcal{B}_2(\alpha,\beta) \leqslant C_{w,2}\mathcal{W}_1(\alpha,\beta) \leqslant C_{w,2}C_{w,m,2}\mathcal{B}_2(\alpha,\beta). \tag{41}$$

If in particular the graph is unweighted, then we have that $C_{w,2} = 1$ and $C_{w,m,2} = m^{1/2}$, so that

$$\mathcal{B}_2(\alpha,\beta) \leqslant \mathcal{W}_1(\alpha,\beta) \leqslant m^{1/2}\mathcal{B}_2(\alpha,\beta). \tag{42}$$

Remark 2.12. Note that the bound in Corollary 2.11 is sharp without making additional assumptions on the structure of G or α, β. Suppose G is a path on n vertices with $m = n-1$, then the upper and lower bounds are achieved, respectively, when we have:

(i) $\alpha = \delta_1$, $\beta = \delta_n$, so that $\mathcal{B}_2(\delta_1, \delta_n) = \sqrt{n-1}$ and $\mathcal{W}_1(\delta_1,\delta_n) = n-1$ so $\mathcal{W}_1 = m^{1/2}\mathcal{B}_2$.

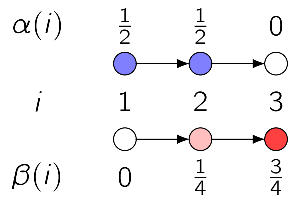
(ii) $\alpha = \delta_1$ and $\beta = \delta_2$, so that $\mathcal{B}_2 = \mathcal{W}_1 = 1$.

Proofs of these statements could be obtained directly or by way of Proposition 2.13 and Proposition 2.14, which we discuss in the next subsection. This leads to an as-yet open question: If one constrains the structure of G or α, β, can the bound in Corollary 2.11 be improved?

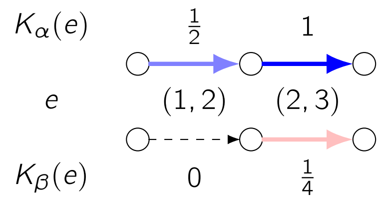
2.3 Worked examples: Paths and trees

In this subsection, we specialize to the case of the path graph P_n and derive explicit formulas for $\mathcal{W}_p$ and $\mathcal{B}_p$ so as to illustrate their connections to optimal transportation in the setting of the real line and to elucidate their differences when $p > 1$. We also take a look at $\mathcal{B}_p$ on tree graphs, and obtain a closed-form solution for all $1 \leqslant p < \infty$ which generalizes the known formula for 1-Wasserstein distance on trees. Proofs of Proposition 2.13, Proposition 2.14, and Proposition 2.15 are all included in Appendix A.1.

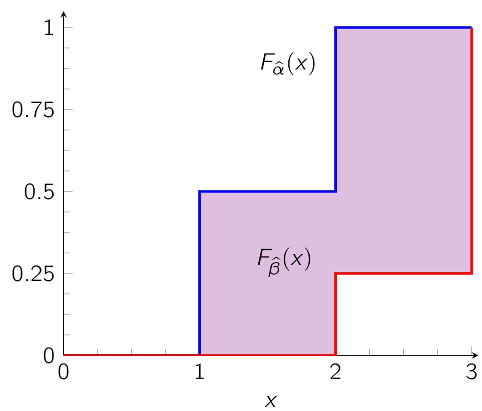
Let P_n be the unweighted path graph with n nodes ordered from 1 to n, and oriented edges E' of the form $(i, i+1)$ for $1 \leqslant i \leqslant n-1$. For $\alpha \in \mathcal{P}(V(P_n))$, let $\widehat{\alpha} = \sum_{i=1}^n \alpha(i)\delta_i$ be the empirical measure on $\mathbb{R}$ induced by α where, with a slight abuse of notation, δ_i is the Dirac measure on $\mathbb{R}$ at $x = i$. Let $F_{\widehat{\alpha}}$ be the cumulative distribution function of $\widehat{\alpha}$, and let $F_{\widehat{\alpha}}^{-1}$ be its psuedo-inverse or quantile function, i.e., $F_{\widehat{\alpha}}^{-1}(t) = \inf\{x : F_{\widehat{\alpha}}(x) \geqslant t\}$ for $t \in [0,1]$. The proposition below is an adaptation of the classical inverse cdf result for $\mathcal{W}_p$ on $\mathbb{R}$ to the case of path graphs; for more detail, see, e.g., [64, 71].



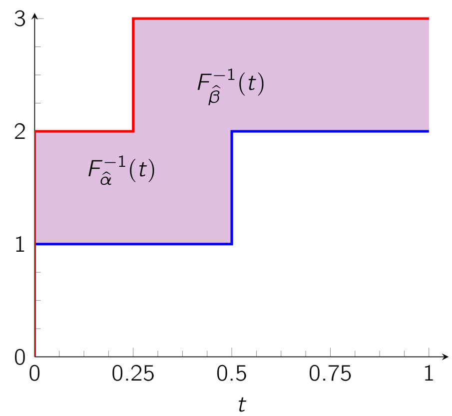
(a) Illustration of two copies of P_3 with arrows indicating orientation of the edges. Example probability measures α, β are shown with node color opacity proportional to mass.



(b) Illustration of two copies of P_3 with arrows indicating orientation of the edges. The edge cdfs K_α and K_β are shown with edge color opacity proportional to value.



(c) Plots of the cdfs $F_{\hat{\alpha}}, F_{\hat{\beta}}$ of the empirical measures $\hat{\alpha}, \hat{\beta}$ on $\mathbb{R}$. From Proposition 2.14 we have $\mathcal{B}_p(\alpha,\beta)^p = 0.5^p + 0.75^p$.



(d) Plots of the inverse cdfs $F_{\hat{\alpha}}^{-1}$ and $F_{\hat{\beta}}^{-1}$ of the empirical measure $\hat{\alpha}, \hat{\beta}$ on $\mathbb{R}$. From Proposition 2.13 we have $\mathcal{W}_p(\alpha,\beta)^p = 1^p(0.25) + 2^p(0.25) + 1^p(0.5)$.

Figure 2: **(a)-(b)** Illustrations of two measures α, β and their edge cdfs K_α, K_β for the fixed path graph P_3. **(c)-(d)** Plots of the empirical cdfs $F_{\hat{\alpha}}$ and $F_{\hat{\beta}}$, as well as their inverses. Note that the shaded regions are reflections of each other; and that for $p = 1$ their common area is $\mathcal{B}_1(\alpha,\beta) = \mathcal{W}_1(\alpha,\beta)$. This also demonstrates the divergence of the metrics for $p > 1$.

Proposition 2.13. *Let $\alpha, \beta \in \mathcal{P}(V(P_n))$ and $1 \leqslant p < \infty$. Then the p-Wasserstein distance $\mathcal{W}_p(\alpha,\beta)$ on P_n is given by*

$$\mathcal{W}_p(\alpha,\beta)^p = \int_0^1 \left| F_{\hat{\alpha}}^{-1}(t) - F_{\hat{\beta}}^{-1}(t) \right|^p \mathrm{d}t. \tag{43}$$

To characterize p-Beckmann distance on P_n, we introduce a bit more notation. For $\alpha \in \mathbb{P}(V)$ and an oriented edge $e = (i, i+1) \in E'$, let $K_\alpha \in \ell(E')$ be the "edge cdf," i.e.,

$$K_\alpha(i, i+1) = \sum_{j \leqslant i} \alpha(j). \tag{44}$$

Note that, naturally, $K_\alpha(i, i+1) = F_{\hat{\alpha}}(i)$.

Proposition 2.14. *Let $\alpha, \beta \in \mathcal{P}(V(P_n))$ and $1 \leqslant p < \infty$. Then the p-Beckmann distance $\mathcal{B}_p(\alpha,\beta)$ on P_n is given by*

$$\mathcal{B}_p(\alpha,\beta)^p = \|K_\alpha - K_\beta\|_p^p = \int_{\mathbb{R}} |F_{\hat{\alpha}}(t) - F_{\hat{\beta}}(t)|^p \mathrm{d}t. \tag{45}$$

Proposition 2.13 and Proposition 2.14 provide an alternate, albeit circuitous, proof of Theorem 2.8 on paths by applying a change of variables. This also demonstrates why we do not expect the two families of metrics to overlap for $p > 1$. A worked example of Proposition 2.13 and Proposition 2.14 in action is illustrated in Fig. 2.

It is also intriguing to note that much in the same way that $\mathcal{W}_1$, and hence $\mathcal{B}_1$, is determined explicitly on trees (a result which itself has been leveraged in other applications such as computer vision and natural language processing [79, 53, 90, 80, 31, 45]), $\mathcal{B}_p$ can be determined in an similar manner for general p on trees.

Proposition 2.15. *Let $T = (V, E, w)$ be a weighted tree, $\alpha, \beta \in \mathcal{P}(V)$, and fix $1 \leqslant p < \infty$. For an oriented edge $e = (i, j) \in E'$, define a generalized version of Eq. (44) 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

$$\mathcal{B}_p(\alpha, \beta) = \|K_\alpha - K_\beta\|_{w,p}. \tag{46}$$

3 The 2-Beckmann problem: Three perspectives

This section is focused entirely on $\mathcal{B}_2$. The foundational premise of this section is that $\mathcal{B}_2$, while ostensibly a simple least squares optimization problem, actually posesses a rich supply of connections and contexts to other notions that already exist in graph theory. In Section 3.1 we view $\mathcal{B}_2$ as an effective resistance metric between measures, and connect it to the simple random walk on G and optimal stopping times for measures. In Section 3.2, we view $\mathcal{B}_2$ as a negative Sobolev norm for functions on graphs, and thereby obtain a Benamou-Brenier type formula for $\mathcal{B}_2$. Lastly, in Section 3.3, we view $\mathcal{B}_2$ as a linearized variant of $\mathcal{W}_2$, prove a convex separation result, and showcase its potential application in unsupervised learning problems for graph data.

3.1 Effective resistance between measures

In this subsection we consider $\mathcal{B}_2$ from the perspective of effective resistance. We begin with some supplemental background on effective resistance which will be needed, and then proceed with discussion of the main results.

3.1.1 Background on effective resistance

Definition 3.1. Let $i, j \in V$ be any two nodes. The **effective resistance** between i, j, denoted r_{ij}, is given by the formula

$$r_{ij} = (\delta_i - \delta_j)^T L^\dagger (\delta_i - \delta_j). \tag{47}$$

The effective resistance between the nodes of a graph has many different properties and representations, many of which intersect with the simple random walk on G. We give its definition and some of its properties below.

Definition 3.2. The **simple random walk** on G is the Markov chain $(X_t)_{t \geqslant 0}$ on the state space of nodes V with transition probability matrix $D^{-1}A$; that is,

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

Recall that X_t will admit a stationary distribution whenever G is connected and will be ergodic whenever G is also non-bipartite. The stationary distribution, which we denote ρ, is always proportional to the degree at each node. A useful tool is the vertex hitting time

$$T_j^s = \inf\{t \geqslant s : X_t = j\}, \quad s \geqslant 0, j \in V. \tag{48}$$

Also, let $\mathrm{vol}(G) = \sum_{i \in V} d_i$ denote the volume of the graph G.

Proposition 3.3. *Let $i, j \in V$ be any two nodes. Then the following observations hold*

(i) $(i,j) \mapsto r_{ij}$ *is a metric on* V,

(ii) $(i,j) \mapsto r_{ij}^{1/2}$ *is a metric on* V,

(iii) $r_{ij} = \inf\{f^T L f : Lf = \delta_i - \delta_j\}$,

(iv) $r_{ij} = \|L^{-1/2}(u_i - u_j)\|_2^2$, *where u_i is the i-th orthonormal eigenvector of the graph Laplacian L.*

(v) $r_{ij} = \frac{1}{\mathrm{vol}(G)}\left(\mathbb{E}[T_i^0 : X_0 = j] + \mathbb{E}[T_j^0 : X_0 = i]\right)$, *i.e., r_{ij} is proportional to the commute time between i, j.*

For proofs of these results, see, e.g., [40, 50, 25].

3.1.2 2-Beckmann as measure effective resistance

We begin with the observation that $\mathcal{B}_2(\alpha, \beta)^2$ can be realized as a type of effective resistance for measures. We make this precise in the theorem below.

Theorem 3.4. *Let $\alpha, \beta \in \mathcal{P}(V)$. Then*

$$\mathcal{B}_2(\alpha,\beta)^2 = (\alpha - \beta)^T L^\dagger (\alpha - \beta). \tag{49}$$

Proof. We can write

$$\begin{aligned} \mathcal{B}_2(\alpha,\beta)^2 &= \inf\{\|J\|_{w,2}^2 : BJ = \alpha - \beta\} && (50)\\ &= \inf\{\|W^{1/2}J\|_2^2 : BJ = \alpha - \beta\} && (51)\\ &= \inf\{\|S\|_2^2 : (BW^{-1/2})S = \alpha - \beta\} && (52)\end{aligned}$$

with $S = W^{1/2}J$. Recall that for matrices X, x it holds that $X^\dagger x = \mathrm{argmin}\{\|y\|_2 : Xy = x\}$ when x belongs to the column space of X [6], and thus we have

$$\begin{aligned} \mathcal{B}_2(\alpha,\beta)^2 &= \|(BW^{-1/2})^\dagger(\alpha - \beta)\|_2^2 && (53)\\ &= (\alpha-\beta)^T((BW^{-1/2})^\dagger)^T(BW^{-1/2})^\dagger(\alpha-\beta) && (54)\\ &= (\alpha-\beta)^T L^\dagger(\alpha - \beta) && (55)\end{aligned}$$

since $L = BWB^T$. ☐

One immediate consequence of Theorem 3.4 is a Brenier-type theorem for graphs, which we state as a corollary below. Based on the discussion following Eq. (12), in this theorem we opt to use the notation $B^T = \nabla$ to achieve homophily with the continuous setting.

Corollary 3.5. *Let $\alpha, \beta \in \mathcal{P}(V)$. The unique optimal edge flow J^* for $\mathcal{B}_2(\alpha, \beta)$ can be written $J^* = \nabla f$ where f is a mean-zero potential function of minimal ℓ_2 norm satisfying the Poisson-type equation $Lf = \alpha - \beta$. The optimal cost is $\|\nabla f\|_{w,2}$.*

Proof. It is enough to write $(\alpha-\beta)^T L^\dagger(\alpha-\beta) = (\alpha-\beta)^T L^\dagger L L^\dagger(\alpha-\beta)$ and $L = BWB^T$. The claim then follows from the proof of Theorem 3.4. □

Based on the Theorem 3.4, we introduce the following definition and notation of effective resistance for measures.

Definition 3.6. Let $\alpha, \beta \in \mathcal{P}(V)$. Then we define the **measure effective resistance** by

$$r_{\alpha\beta} = (\alpha-\beta)^T L^\dagger(\alpha-\beta). \tag{56}$$

In the subsequent sections, we will explore the properties of $r_{\alpha\beta}$ in its own right, with particular focus on its relationship to optimal stopping rules of the simple random walk.

3.1.3 Measure effective resistance and optimal stopping rules

We have already argued in Theorem 3.4 that measure effective resistance can be understood as a type of flow-based optimal transportation distance between the measures. The goal of this section is to essentially set that aside and think about $r_{\alpha\beta}$ in its own right. A natural approach along these lines is to take Proposition 3.3 as a starting point. Immediately, $r_{\alpha\beta}$ has several analagous properties, which we state below without proof.

Proposition 3.7. *Let $\alpha, \beta \in \mathcal{P}(V)$. Then the following statements hold:*

(i) $(\alpha, \beta) \mapsto r_{\alpha\beta}^{1/2}$ *is a metric on* $\mathcal{P}(V)$,

(ii) $r_{\alpha\beta} = \inf\left\{f^T L f : Lf = \alpha-\beta\right\}$,

(iii) $r_{\alpha\beta} = \|L^{-1/2}(\alpha-\beta)\|_2^2$.

To go deeper, we pose the following question along the lines of Proposition 3.3(v): To what extent can $r_{\alpha\beta}$ be understood as a generalized commute time between the measures α, β? To answer this question, we need to use results that concern stopping rules for the simple random walk.

Definition 3.8. 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 $(X_0, X_1, \ldots, X_\Gamma)$.

The mean length $\mathbb{E}[\Gamma]$ is the expected duration of the walk. If $\mathbb{E}[\Gamma] < \infty$ then the walk stops almost surely in finite time, so we define X_Γ to be the position of the random walk at the stopping time. Having defined stopping rules, we can define the generalized hitting time between α, β.

Definition 3.9. Let $\alpha, \beta \in \mathcal{P}(V)$. The **access time** $H(\alpha, \beta)$ is defined as

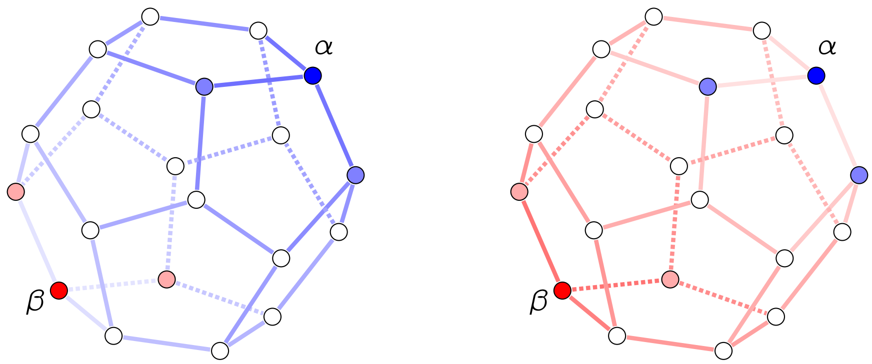
$$H(\alpha,\beta) = \inf\left\{\mathbb{E}[\Gamma | X_0 \sim \alpha] : \mathbb{E}[\Gamma] < \infty, \text{ and } X_\Gamma \sim \beta\right\}. \tag{57}$$

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**.

Remark 3.10. It is not hard to see that the set of feasible stopping rules in Definition 3.9 is nonempty. 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).



(a) Initial distribution α, stopping distribution β.

(b) Initial distribution β, stopping distribution α.

Figure 3: Two illustrations of 1000 simulated simple random walks on the dodecahedral graph with given initial distribution, illustrated with node opacity proportional to density; and naïve stopping rule according to the given stopping distribution, illustrated with opposite node color and opacity proportional to density. The edges are dashed only to indicate depth, and edge opacity is proportional to the total number of times the simulated random walks landed on each edge. The mean lengths of paths in (a) (resp. (b)) correspond to $H_n(\alpha, \beta)$ (resp. $H_n(\beta, \alpha)$).

Several examples of optimal stopping rules for general distributions are given in [52], and in particular, it is shown thereby that any such α, β admit an optimal stopping rule provided G is connected.

The access time $H(\alpha, \beta)$ has several notable properties, which we summarize in the Proposition below. The proofs of these results can be found in [52].

Proposition 3.11. *Let $\alpha, \beta \in \mathcal{P}(V)$. Then the following facts hold for $H(\alpha, \beta)$:*

(i) When β is concentrated at a node $j \in V$, it holds $H(\alpha, j) = \sum_{i \in V} \alpha_i H(i, j)$,

(ii) $H(\alpha, \beta) = \max_j \sum_{i \in V} (\alpha_i - \beta_i) H(i, j)$,

(iii) $H(\alpha, \beta)$ is convex in both of its arguments; namely if $\alpha, \alpha', \beta, \beta' \in \mathcal{P}(V)$ and $c \in [0, 1]$, we have

$$H(c\alpha + (1-c)\alpha', \beta) \leqslant cH(\alpha, \beta) + (1-c)H(\alpha', \beta) \tag{58}$$

$$H(\alpha, c\beta + (1-c)\beta') \leqslant cH(\alpha, \beta) + (1-c)H(\alpha, \beta') \tag{59}$$

Beveridge [8] also established a connection between the entries of the psuedoinverse of the graph Laplacian (treated as a discrete Green's function) and the access time between nodes and the stationary distribution of the random walk. We reproduce that formula here as a proposition, in a modified form to match our conventional choice of Laplacian matrix.

Theorem 3.12 (Access Time Formula for the Discrete Green's Function, Beveridge [8])**.** *Let ρ be the stationary distribution of the simple random walk on V. Then the i, j-th entry of $L^\dagger$ can be expressed as*

$$(L^\dagger)_{ij} = \frac{1}{\mathrm{vol}(G)} \left(H(\rho, j) - H(i, j) \right), \tag{60}$$

where $\mathrm{vol}(G) = \sum_{i \in V} d_i$.

For $\alpha \in \mathcal{P}(V)$, define $H_\alpha \in \ell(V)$ by $H_\alpha(i) = H(\alpha, i)$. We are ready to state our main result connecting $r_{\alpha\beta}$ to the access time.

Theorem 3.13 (Generalized Commute Time Formula)**.** *Let $\alpha, \beta \in \mathcal{P}(V)$. Then*

$$r_{\alpha\beta} = -\frac{1}{\mathrm{vol}(G)}(\alpha - \beta)^T (H_\alpha - H_\beta). \tag{61}$$

Or, in alternative expanded forms,

$$r_{\alpha\beta} = -\frac{1}{\mathrm{vol}(G)} \sum_{i \in V} (\alpha_i - \beta_i)(H(\alpha, i) - H(\beta, i)) \tag{62}$$

$$= -\frac{1}{\mathrm{vol}(G)} \sum_{i,k \in V} (\alpha_i - \beta_i)(\alpha_k - \beta_k) H(i, k). \tag{63}$$

We provide the proof in Appendix A.2. The reason we term Theorem 3.13 a generalized commute time formula is that in the case where α, β are Dirac measures, this recovers the result that $r_{ij} \propto (H(i,j) + H(j,i))$, since we have

$$r_i j = -\frac{1}{\mathrm{vol}(G)}(\delta_i - \delta_j)^T (H_{\delta_i} - H_{\delta_j}) \tag{64}$$

$$= -\frac{1}{\mathrm{vol}(G)}(H(i,i) - H(i,j) + H(j,j) - H(j,i) \tag{65}$$

$$= \frac{1}{\mathrm{vol}(G)}(H(i,j) + H(j,i)). \tag{66}$$

Theorem 3.13 shows, however, that in the more general case of measures $r_{\alpha\beta}$ is not exactly the same as a measure commute time $H(\alpha, \beta) + H(\beta, \alpha)$. Theorem 3.13 can be used to establish a relationship between the measure commute time and the measure resistance, as we show below.

Corollary 3.14 (Measure Commute Time Inequalities)**.** *Let $\alpha, \beta \in \mathcal{P}(V)$. Then $r_{\alpha\beta}$ satisfies the following two inequalities:*

$$r_{\alpha\beta} \leqslant \frac{2}{\mathrm{vol}(G)} \max\{H(\alpha, \beta), H(\beta, \alpha)\} \tag{67}$$

$$r_{\alpha\beta} \leqslant \frac{1}{\mathrm{vol}(G)}(H_n(\alpha, \beta) + H_n(\beta, \alpha)) \tag{68}$$

where $H_n(\alpha, \beta) = \mathbb{E}[\Gamma_n]$ (resp. $H_n(\beta, \alpha)$) is the expected duration of the naive stopping rule described in Remark 3.10 with initial distribution α (resp. β) and stopping node sampled from β (resp. α).

The proof of Corollary 3.14 is located in Appendix A.2. Note that the second half of the proof of Corollary 3.14 actually contains a related observation regarding $\mathcal{B}_2$ and $\mathcal{W}_{r,1}$, which we state as another corollary below, without proof.

Corollary 3.15. *Let $\alpha, \beta \in \mathcal{P}(V)$, and let r be the effective resistance metric on V. Then we have*

$$\mathcal{B}_2(\alpha, \beta) \leqslant \mathcal{W}_{r,1}(\alpha, \beta)^{1/2}.$$

A final observation that one can glean from the random walk and access time setting is an interpretation of the transportation potential $f = L^\dagger(\alpha - \beta)$ mentioned in Corollary 3.5.

Proposition 3.16. *Let $\alpha, \beta \in \mathcal{P}(V)$, and let Γ be any stopping rule such that $\mathbb{E}[\Gamma] < \infty$ and $X_\Gamma \sim \beta$. Let f_Γ be the degree-weighted exit frequencies*

$$f_\Gamma(i) = \mathbb{E}\left[\sum_{t=0}^{\Gamma - 1} \frac{1}{d_i} \delta_{\{X_t = i\}}(i) : X_0 \sim \alpha\right].$$

Then $Lf_\Gamma = \alpha - \beta$. In particular, the transportation potential $L^\dagger(\alpha - \beta)$ is given by the mean normalized function $f_\Gamma - \frac{1}{n} f_\Gamma^T \mathbf{1}_n$ for any suitable stopping rule Γ.

This proposition follows from applying the so-called conservation equation for exit frequencies, mentioned in, e.g., [52, 51]. Another way of putting it is that the unique optimal flow for $\mathcal{B}_2(\alpha, \beta)$ can be realized as the gradient of the degree-normalized exit frequencies of *any* finite-mean stopping rule which achieves a stopping distribution of β, having been intialized at α (since, moreover, as discussed in [51], any two degree-normalized exit frequencies for different stopping times differ by a constant, and thus have the same gradient).

3.2 Sobolev norms and a graph Benamou-Brenier formula

The relationship between Sobolev spaces and optimal transportation have been explored at length in the continuous setting [84, 63, 37, 77]. The goal of this subsection is to present the 2-Beckmann problem as an optimization problem on a negative Sobolev space for functions defined on graphs. In this section we follow the exposition laid out in [63] in the continuous setting by analogy to the discrete setting. As it turns out, there are at least two results that achieve strong homophily between the graph and continuous setting.

3.2.1 Background from the continuous setting

Recall that if $f : \mathbb{R}^n \to \mathbb{R}$ is a function with a square integrable derivative ∇f in the weak sense and μ is a Borel probability measure which is absolutely continuous with respect to the Lebesgue measure $\mathrm{d}x$; i.e. so that $\mathrm{d}\mu = g\mathrm{d}x$ for a density function g, we can define the Sobolev-type seminorm $\|\cdot\|^2_{\dot{H}^1(\mu)}$ by

$$\|f\|^2_{\dot{H}^1(\mu)} = \int_{\mathbb{R}^n} \|\nabla f\|_2^2 \mathrm{d}\mu. \tag{69}$$

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

$$\|h\mathrm{d}x\|_{\dot{H}^{-1}(\mu)} = \sup\left\{\int_{\mathbb{R}^n} fhd\mu : \|f\|_{\dot{H}^1(\mu)} \leqslant 1\right\}. \tag{70}$$

This setup leads to two results (among others) in the continuous setting, which are of interest in the graph setting as well.

Theorem 3.17 ($\dot{H}^{-1}$ - Linearization of $\mathcal{W}_2$ [63])**.** *If μ is a Borel probability measure on $\mathbb{R}^n$ and $\mathrm{d}\mu$ is an infintesimally small perturbation of μ, then*

$$\mathcal{W}_2(\mu, \mu + \mathrm{d}\mu) = \|\mathrm{d}\mu\|_{\dot{H}^{-1}(\mu)} + o(\mathrm{d}\mu).$$

The second result which interests us is the Sobolev form of the Benamou-Brenier formula, which we state below.

Theorem 3.18 (Benamou-Brenier formula, $\dot{H}^{-1}$ form [63])**.** *Let μ, ν be Borel probability measures on $\mathbb{R}^n$. Then it holds:*

$$\mathcal{W}_2(\mu, \nu) = \inf\left\{\int_0^1 \|\mathrm{d}\mu_t\|_{\dot{H}^{-1}(\mu_t)} : \mu_0 = \mu, \mu_1 = \nu\right\}. \tag{71}$$

3.2.2 Graph sobolev norms

Next we transition to defining the appropriate discrete analogues of $\|\cdot\|_{\dot{H}^1(\mu)}$ and $\|\cdot\|_{\dot{H}^{-1}(\mu)}$ on the graph G.

Definition 3.19. Let $f, g \in \ell(V)$. We define the **graph Sobolev seminorm** $\|\cdot\|_{\dot{H}^1(V)}$ by the equation

$$\|f\|_{\dot{H}^1(V)}^2 = \sum_{(i,j)\in E'} w_{ij} |\nabla f(i,j)|_2^2 = \sum_{(i,j)\in E'} w_{ij} |f(i) - f(j)|_2^2. \tag{72}$$

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

$$\|g\|_{\dot{H}^{-1}(V)}^2 = \sup\left\{ f^T g : \|f\|_{\dot{H}^1(V)} \leqslant 1 \right\}. \tag{73}$$

It is useful to note that 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). The first proposition in the section elucidates the relationship between graph Sobolev norms and the graph Laplacian matrix.

Proposition 3.20. *Let $f, g \in \ell(V)$. Then the following hold:*

(i) $\|f\|_{\dot{H}^1(V)}^2 = f^T L f$.

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

Proof. The first claim is straightforward. The second requires a closer look. We can work backwards and use the proof of Theorem 2.1 to obtain

$$(g^T L^\dagger g)^{1/2} = \|L^{-1/2} g\|_2 \tag{74}$$

$$= \inf\{\|f\|_2 : L^{1/2} f = g\} \tag{75}$$

$$= \sup\left\{ f^T g : \|L^{1/2} f\|_2 \leqslant 1 \right\}, \tag{76}$$

where $L^{\{-1/2,1/2\}} = U \Lambda^{\{-1/2,1/2\}} U^T$ as in Eq. (10). But $\|L^{1/2} f\|_2^2 = f^T L f = \|f\|_{\dot{H}^1(V)}^2$, so the claim follows. □

3.2.3 2-Beckmann as a negative Sobolev distance

Since our definitions have been set up and explained properly, we can derive the following result, which follows directly from Proposition 3.20 and Theorem 3.4. We consider this a graph analogue of Theorem 3.17 for 2-Beckmann distance.

Theorem 3.21. *Let $\alpha, \beta \in \mathcal{P}(V)$. Then*

$$\mathcal{B}_2(\alpha, \beta) = \|\alpha - \beta\|_{\dot{H}^{-1}(V)}.$$

Thus in particular if $\beta = \alpha + \mathrm{d}\alpha$ for a small perturbation $\mathrm{d}\alpha$, then $\mathcal{B}_2(\alpha, \alpha + \mathrm{d}\alpha) = \|\mathrm{d}\alpha\|_{\dot{H}^{-1}(V)}$.

Note that since the $\dot{H}^{-1}$ norm depends on the operator $L^{-1/2}$, the Beckmann distance $\mathcal{B}_2(\alpha, \alpha + \mathrm{d}\alpha)$ can be articulated in terms of the spectral coefficients of $\mathrm{d}\alpha$; to wit, if $\mathrm{d}\alpha = \sum_{\ell=2}^n c_\ell u_\ell$ is the spectral decomposition of a mean-zero perturbation $\mathrm{d}\alpha \in \ell(V)$ in terms of the eigenvectors u_ℓ of L, then

$$\mathcal{B}_2(\alpha, \alpha + \mathrm{d}\alpha)^2 = \|L^{-1/2} \mathrm{d}\alpha\|_2^2 = \sum_{\ell=2}^n \frac{c_\ell^2}{\lambda_\ell}. \tag{77}$$

Therefore in general, one has $\mathcal{B}_2(\alpha, \alpha + \mathrm{d}\alpha) \leqslant \lambda_2^{-1/2} \|\mathrm{d}\alpha\|_2$; but if the spectral properties of $\mathrm{d}\alpha$ are known, this estimate may be sharpened.

The next result which leverages the Sobolev norm perspective on graph optimal transport can be considered a graph analogue of the Benamou-Brenier formula. Let $\mu_t \in \ell(V)$ for each $t \in [0,1]$. 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 $\mathbb{R}^n$. Moreover, when μ_t admits a derivative, we write $\mathrm{d}\mu_t = \frac{\mathrm{d}}{\mathrm{d}s} \mu_s \big|_{s=t}$.

Theorem 3.22 (Graph Benamou-Brenier Formula)**.** *Let $\alpha, \beta \in \mathcal{P}(V)$. Then we have*

$$\mathcal{B}_2(\alpha, \beta)^2 = \inf \left\{ \int_0^1 \|\mathrm{d}\mu_t\|^2_{\dot{H}^{-1}(V)} \, \mathrm{d}t : \mu_t \in C^1([0,1]), \mu_0 = \alpha, \mu_1 = \beta \right\}. \tag{78}$$

See Appendix A.2 for a proof, which is based on the proof appearing in the lecture notes [35]. Note that the infimum is achieved by a linear line segment connecting α to β, which underlines the idea that $\mathcal{B}_2$ should be treated as a linearized variant of Wasserstein distance. We explore the linearization angle and its implications on graph learning tasks in the next section.

3.3 A linearized optimal transportation distance

A third and final perspective on 2-Beckmann distance is from the world of clustering and classification of graph data. We begin with a background discussion from the continuous setting, and then explore convex separation properties for graph data as well as an application involving a handwritten digit dataset.

3.3.1 Background from the continuous setting

A typical classification scenario usually consists of some data $\{x_i\} \subset \mathbb{R}^n$ which one wishes to separate into classes or clusters either without prior knowledge of the data labels (unsupervised) or with some information about the class labels of the dataset (semi-supervised or supervised). In many applications, e.g., [20, 93, 13], the data x_i can often occur not as vectors in $\mathbb{R}^n$ but as distributions μ_i on $\mathbb{R}^n$.

In such scenarios, Wasserstein distance, formulated in the Monge sense as

$$\mathcal{W}_2(\mu, \nu)^2 = \inf_{T : T_\sharp \mu = \nu} \int_{\mathbb{R}^n} \|T(x) - x\|_2^2 d\mu(x), \tag{79}$$

where $T_\sharp\mu$ is the push-forward of μ under the map T and the inf is over all such T exhibiting desired regularity, arises as a natural distance metric between probabilitiy measures on $\mathbb{R}^n$. However, this metric can be slow to compute at scale, and in the discrete setting can even suffer from being undefined.

One approach to resolving this issue involves the usage and study of Linearized Optimal Transport embeddings [58, 88, 54, 43, 21], which can be summarized as follows. Given data measures $\{\mu_i\}$, fix a reference measure μ, and define $T_i = \operatorname{argmin}_{T : T_\sharp \mu = \mu_i} \int_{\mathbb{R}^n} \|T(x) - x\|_2^2 d\mu(x)$. Then the measures are featurized in the form of the maps $T_i : \mathbb{R}^n \to \mathbb{R}^n$ and 2-Wasserstein distance between the data measures is then approximated by $\|T_i - T_j\|_{L^2}$.

Theoretically, the ability of Linearized Optimal Transport in approximating 2-Wasserstein distance can be captured by the following two results, the informal statements of which we restate from [58].

Theorem 3.23 (Informal Statement of Theorem 4.4 from [58])**.** *If $\mathscr{P} = \{\mu_i\}$ are ϵ-perturbations of shifts and scalings of μ, and $\mathscr{Q} = \{\nu_i\}$ are ϵ-perturbations of shifts and scalings of ν, and $\mathscr{P}$ and $\mathscr{Q}$ have small minimal distance depending on ϵ and satisfy a few technical assumptions, $\mathscr{P}$ and $\mathscr{Q}$ are linearly separable in LOT embedding space.*

Theorem 3.24 (Informal Statement of Theorem 4.1 from [58])**.** *If μ, ν are ϵ-perturbations of shifts and scalings of one another, then*

$$\mathcal{W}_2(\mu, \nu) \leqslant \mathcal{W}_2^{\mathrm{LOT}}(\mu, \nu) \leqslant \mathcal{W}_2(\mu, \nu) + C_\sigma \epsilon + C'_\sigma \epsilon^{1/2}. \tag{80}$$

In particular, when $\epsilon = 0$ and μ, ν are shifts and scalings of each other, LOT is an isometry.

It is also worth noting that in the situation where μ_i is some shift or scaling of μ, i.e., $\mu_i = T_\sharp\mu$ for an affine combination of scaling and translation T on $\mathbb{R}^n$, then the LOT embedding of μ_i will be the map T itself; i.e., T is optimal in the Wasserstein cost.

3.3.2 Linear separtation for 2-Beckmann distance

To set up the problem, suppose we have data sampled from C classes, and each class is associated to a canonical distribution $\alpha_i \in \mathcal{P}(V)$ so that our data may be represented in the form $\mathcal{D} = \{T\alpha_i\}_{i=1,\dots,C;T\in\mathcal{T}}$. That is, each data point is of the form $T\alpha_i$ where α_i is the underlying distribution of the class and the map $T : \mathcal{P}(V) \to \mathcal{P}(V)$ represents a perturbation of the class distribution among valid perturbations $\mathcal{T}$ (to be determined).

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. In our data model, we assume that each image may be expressed as a perturbation of a canonical distribution associated to the corresponding image class.

In the unsupervised setting, we simply have $\mathcal{D}$ and we wish to use some clustering technique, e.g., k-means or principal geodesic analysis, associated to a metric on $\mathcal{P}(V)$. In the supervised setting, we have data equipped with labels $\mathcal{D}' = \{(T\alpha_i, y_i)\}_{i=1,\dots,C;T\in\mathcal{T}}$ so that each label depends only on the class distribution and we seek to build some classifier, ideally linear, which facilitates predictive analysis.

We treat $\mathcal{B}_2$ as a linearization of $\mathcal{W}_2$ on graphs. In this setting, the featurization of each canonical class distribution is given by $\alpha_i \mapsto L^{-1/2}\alpha_i$ and similarly for the individual samples $T\alpha_i$.

For a set of measures $\mathcal{A} = \{\alpha_1, \dots, \alpha_k\} \subset \mathcal{P}(V)$, define the mutual support by

$$\mathscr{MS}(\mathcal{A}) = \bigcap_{i=1}^{k} \operatorname{supp}(\alpha_i) \subset V,$$

where $\operatorname{supp}(\alpha_i)$ is the subset of nodes in V on which $\alpha > 0$.

Theorem 3.25. *Let $\alpha_1 \neq \alpha_2 \in \mathcal{P}(V)$ be the canonical class distributions for a binary distributional dataset $\mathcal{D} = \{T\alpha_i\}_{i=1,2;T\in\mathcal{T}}$. Assume $\mathscr{MS}(\{\alpha_1,\alpha_2\}) \neq \varnothing$ and let $\mathcal{T}$ be the set of affine perturbation maps defined by*

$$T\alpha = \alpha + \mathrm{d}T$$

where $\mathrm{d}T \in \ell(V)$ is vector which satisfies the three conditions:

(i) $\mathrm{d}T$ *is mean zero, or* $\mathbf{1}_n^T \mathrm{d}T = 0$*;*

(ii) $\operatorname{supp}(\mathrm{d}T) \subset \mathscr{MS}(\{\alpha_1,\alpha_2\})$*; and*

(iii) $\|\mathrm{d}T\|_1 < \delta$*, where*

$$\delta = \min\left\{\frac{1}{3}\|\alpha_1 - \alpha_2\|_2, \min_{\substack{i=1,2\\ k\in\mathscr{MS}(\{\alpha_1,\alpha_2\})}} \alpha_i(k)\right\} > 0. \tag{81}$$

Then the sets $\mathcal{A}_1 = \{T\alpha_1\}_{T\in\mathcal{T}}$ and $\mathcal{A}_2 = \{T\alpha_2\}_{T\in\mathcal{T}}$, which belong to the metric space $(\mathcal{P}(V), \mathcal{B}_2)$, can be isometrically embedded into $\ell_2(V)$ under $L^{-1/2}$ so that their images are linearly separable as convex sets in $\mathbb{R}^n$.

See Appendix A.2 for the proof. Note that it is not hard to extend Theorem 3.23 to the case of $C > 2$ classes, by simply requiring that the perturbations be smaller than the mutual distances between the canonical class distributions, up to an appropriate constant, so that in the data model there are no overlapping perturbations. The main point of this result is that $L^{-1/2}$ *as an embedding of measures*, is well-behaved; in the sense that as long as the original collections of measures were separable, $L^{-1/2}$ will not imbue any pathology.

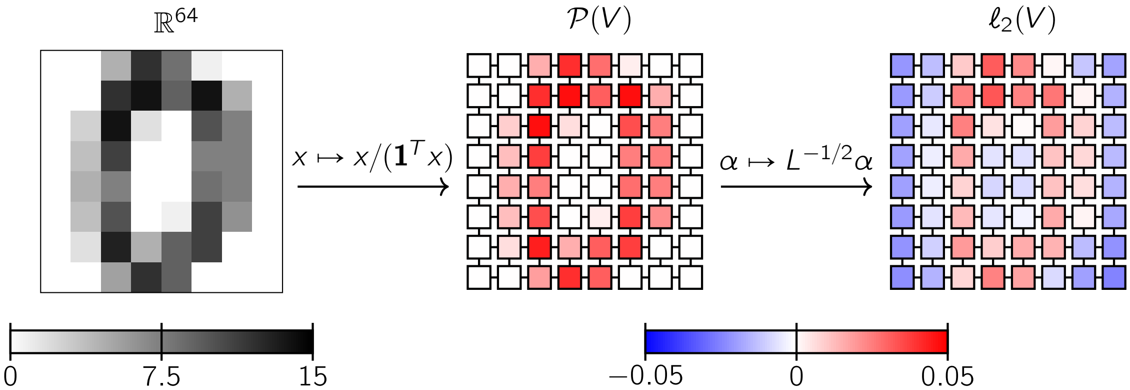


Figure 4: An illustration of the preprocessing pipeline for the digits data, with an example from the class of handwritten zeros. The first step is a mass normalization to convert the pixel values into a fixed-sum distribution viewed on the nodes V of the 8×8 lattice graph. The second step is an embedding $\alpha \mapsto L^{-1/2}\alpha$, 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.

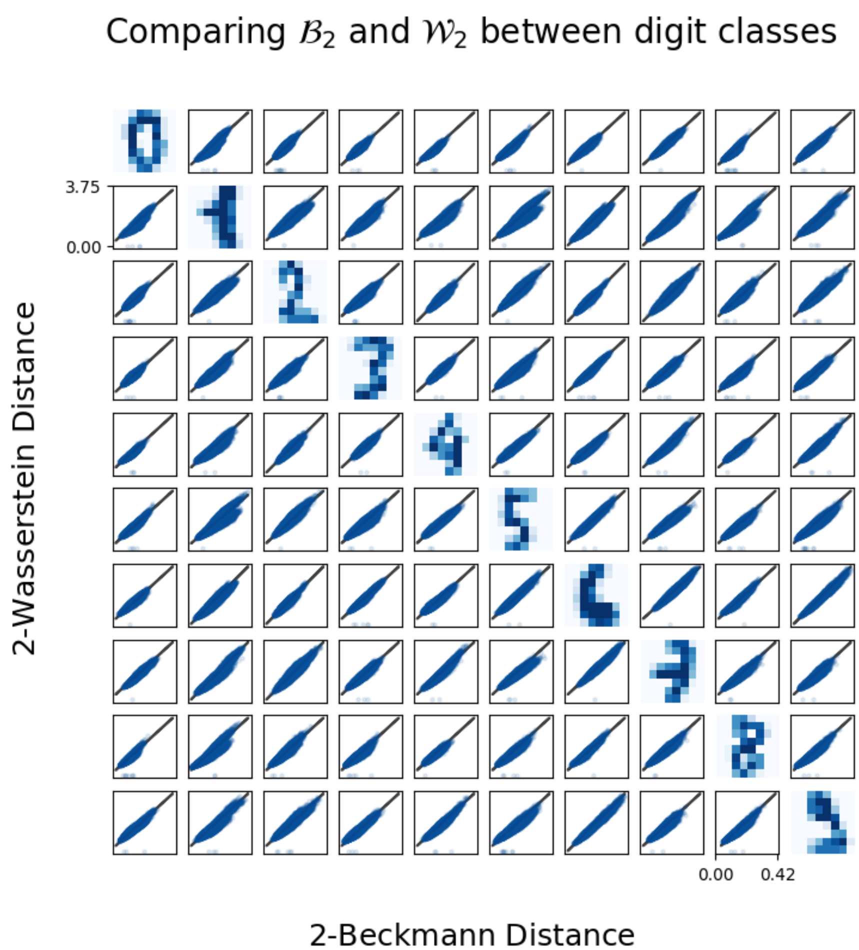
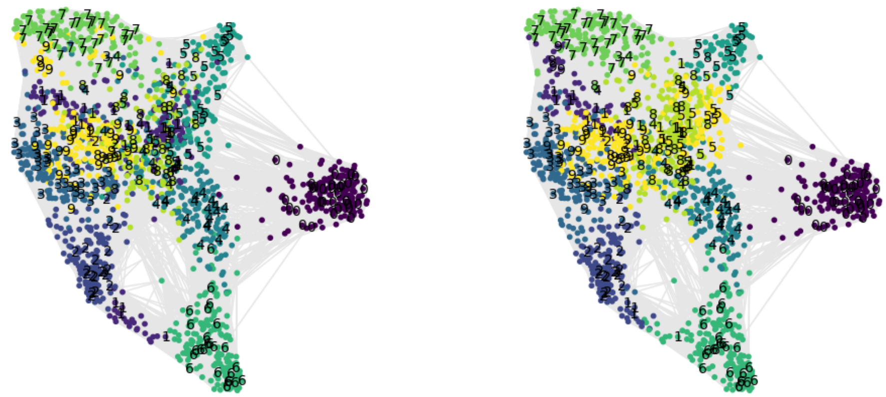
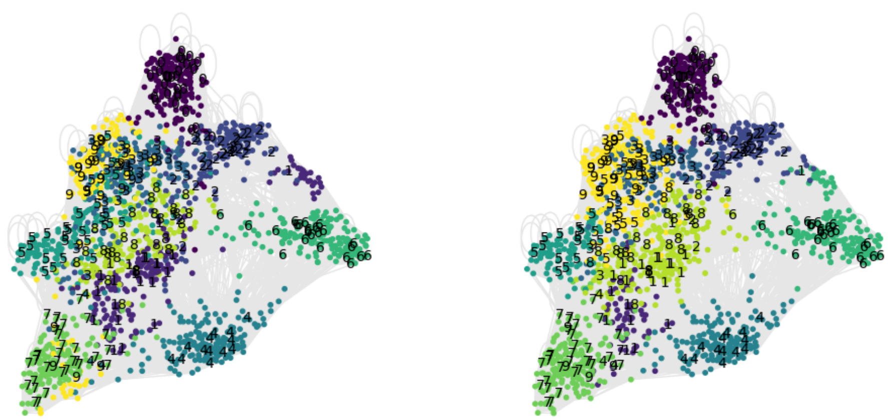


Figure 5: Using the digits dataset, and for each pair of digit classes, we computed the pairwise 2-Beckmann and 2-Wasserstein distances for each pair of samples originating from the respective digit classes (with around 30,000 pairs of distances per pair of digit classes). Within each tile of the grid, 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 $\mathcal{W}_2 \approx 8.446\mathcal{B}_2$.



Nodes colored by ground truth clusters | Nodes colored by spectral clustering

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



Nodes colored by ground truth clusters | Nodes colored by spectral clustering

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

	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

(c) Performance metrics

Figure 6: **(a)-(b)** Using the digits dataset, we demonstrate the results of an unsupervised clustering algorithm with different choices of similarity kernel. Each node corresponds to an image of a digit, which we featurize as a distribution on the 8×8 square lattice graph. We built a $k = 42$ nearest neighbor graph on the nodes (shown), and then apply spectral clustering [87] 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). **(c)** 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) [39]; mutual information (MI) and adjusted mutual information (AMI) [86]; and homogeneity (Hom) and completeness (Com) [66]. 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.

3.3.3 Experimental results

To further explore the implications of our results on learning with graph data, we conducted two experiments which illuminate some of the potential applications and underlying properties of 2-Beckmann distance between measures. Both examples make use of the "Pen-Based Recognition of Handwritten Digits" dataset [3], accessed and processed via the Sklearn Python package [14]. The dataset contains 1797 grayscale handwritten digits of resolution 8×8 and pixel values ranging from $0 - 15$. The reason for choosing this dataset is that the 2-Wasserstein kernel is computationally prohibitive at scale without approximation, so rather than regularize the Wasserstein metric, we opt for a lighter-weight image dataset. Specifically, if $\mathcal{W}_2$ is obtained via, e.g., a Hungarian algorithm or a linear program, the time complexity to obtain a kernel matrix for N samples, each defined on a graph of n nodes, is roughly $O(n^3N^2)$. A Beckmann kernel matrix, on the other hand, runs $O(n^3 + (n^2 + n)N^2)$ since only a single SVD calculation of $L^\dagger$ is required at the beginning, and then subsequent entries of the kernel matrix are obtained through matrix-vector multiplication and pairwise comparison. The Wasserstein metric $\mathcal{W}_2(\alpha, \beta)$ was directly evaluated for each pair of examples α, β by using the CVXPY convex optimization Python package [24] with initialization determined by the trivial coupling $\beta\alpha^T$. Fig. 4 illustrates the pre-processing pipeline for an example image of a handwritten zero.

The first experiment, described in detail in Fig. 5, is a comparative study of $\mathcal{B}_2$ and $\mathcal{W}_2$ in which we show that the two metrics are well-correlated for the dataset in question. The second experiment, described in Fig. 6, is a comparative study of $\mathcal{B}_2$ and $\mathcal{W}_2$ within the context of unsupervised learning on the digit dataset. In this setup, we show that the two kernels perform almost identically on the dataset with respect to several metrics (with a slight edge toward $\mathcal{B}_2$), and thus there is some empirical evidence for drop-in usage of $\mathcal{B}_2$ where $\mathcal{W}_2$, or perhaps $\mathcal{W}_p$ more generally, may present a computational bottleneck.

4 Acknowledgements

The authors would like to acknowledge Yusu Wang for her inspirational discussions early on in this project, Stefan Steinerberger for his help placing the preliminary results within the context of the broader OT space, Zachary Lubberts for his helpful discussions on applications of graph OT to discrete geometry, and Caroline Moosmüller for her helpful feedback on linearized Wasserstein distance and Sobolev transport in the continuous setting. SR wishes to acknowledge Evangelos Nikitopoulos, as well as 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. ZW was supported by NSF CCF 2217058.

References

[1] Tara Abrishami, Nestor Guillen, Parker Rule, Zachary Schutzman, Justin Solomon, Thomas Weighill, and Si Wu. Geometry of graph partitions via optimal transport. *SIAM Journal on Scientific Computing*, 42(5), 2020.

[2] Morteza Alamgir and Ulrike Luxburg. Phase transition in the family of p-resistances. *Advances in neural information processing systems*, 24, 2011.

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

[4] Frank Bauer, Jürgen Jost, and Shiping Liu. Ollivier-ricci curvature and the spectrum of the normalized graph laplace operator. *arXiv preprint arXiv:1105.3803*, 2011.

[5] Martin Beckmann. A continuous model of transportation. *Econometrica: Journal of the Econometric Society*, 1952.

[6] Adi Ben-Israel and A Charnes. Contributions to the theory of generalized inverses. *Journal of the Society for Industrial and Applied Mathematics*, 11(3), 1963.

[7] Jean-David Benamou and Yann Brenier. A computational fluid mechanics solution to the monge-kantorovich mass transfer problem. *Numerische Mathematik*, 84(3):375–393, 2000.

[8] Andrew Beveridge. A hitting time formula for the discrete green's function. *Combinatorics, Probability and Computing*, 25(3), 2016.

[9] Andrew Beveridge and László Lovász. Exit frequency matrices for finite markov chains. *Combinatorics, Probability and Computing*, 19(4), 2010.

[10] Robi Bhattacharjee, Alexander Cloninger, Yoav Freund, and Andreas Oslandsbotn. Effective resistance in metric spaces. *arXiv preprint arXiv:2306.15649*, 2023.

[11] Jérémie Bigot. Statistical data analysis in the wasserstein space. *ESAIM: Proceedings and Surveys*, 68, 2020.

[12] Mitchell Black, Zhengchao Wan, Amir Nayyeri, and Yusu Wang. Understanding oversquashing in gnns through the lens of effective resistance. In *International Conference on Machine Learning*, pages 2528–2547. PMLR, 2023.

[13] Robert V Bruggner, Bernd Bodenmiller, David L Dill, Robert J Tibshirani, and Garry P Nolan. Automated identification of stratifying signatures in cellular subpopulations. *Proceedings of the National Academy of Sciences*, 111(26), 2014.

[14] Lars Buitinck, Gilles Louppe, Mathieu Blondel, Fabian Pedregosa, Andreas Mueller, Olivier Grisel, Vlad Niculae, Peter Prettenhofer, Alexandre Gramfort, Jaques Grobler, Robert Layton, Jake VanderPlas, Arnaud Joly, Brian Holt, and Gaël Varoquaux. API design for machine learning software: experiences from the scikit-learn project. In *ECML PKDD Workshop: Languages for Data Mining and Machine Learning*, 2013.

[15] Sarah Cannon, Moon Duchin, Dana Randall, and Parker Rule. Spanning tree methods for sampling graph partitions. *arXiv preprint arXiv:2210.01401*, 2022.

[16] Benson Chen, Gary Bécigneul, Octavian-Eugen Ganea, Regina Barzilay, and Tommi Jaakkola. Optimal transport graph neural networks. *arXiv preprint arXiv:2006.04804*, 2020.

[17] Samantha Chen, Sunhyuk Lim, Facundo Mémoli, Zhengchao Wan, and Yusu Wang. Weisfeiler-lehman meets gromov-wasserstein. In *International Conference on Machine Learning*, pages 3371–3416. PMLR, 2022.

[18] Fan Chung, Wenbo Zhao, and Mark Kempton. Ranking and sparsifying a connection graph. *Internet Mathematics*, 10(1-2), 2014.

[19] Alex Cloninger, Gal Mishne, Andreas Oslandsbotn, Sawyer Jack Robertson, Zhengchao Wan, and Yusu Wang. Random walks, conductance, and resistance for the connection graph laplacian. *arXiv preprint arXiv:2308.09690*, 2023.

[20] Alexander Cloninger, Brita Roy, Carley Riley, and Harlan M Krumholz. People mover's distance: Class level geometry using fast pairwise data adaptive transportation costs. *Applied and Computational Harmonic Analysis*, 47(1), 2019.

[21] Alexander Cloninger, Keaton Hamm, Varun Khurana, and Caroline Moosmüller. Linearized wasserstein dimensionality reduction with approximation guarantees. *arXiv preprint arXiv:2302.07373*, 2023.

[22] Karel Devriendt and Renaud Lambiotte. Discrete curvature on graphs from the effective resistance. *Journal of Physics: Complexity*, 3(2), 2022.

[23] Karel Devriendt, Andrea Ottolini, and Stefan Steinerberger. Graph curvature via resistance distance. *Discrete Applied Mathematics*, 348, 2024.

[24] Steven Diamond and Stephen Boyd. CVXPY: A Python-embedded modeling language for convex optimization. *Journal of Machine Learning Research*, 2016.

[25] Peter G Doyle and J Laurie Snell. *Random walks and electric networks*, volume 22. American Mathematical Soc., 1984.

[26] Abderrahim Elmoataz, Olivier Lezoray, and Sébastien Bougleux. Nonlocal discrete regularization on weighted graphs: a framework for image and manifold processing. *IEEE transactions on Image Processing*, 17(7), 2008.

[27] Abderrahim Elmoataz, Matthieu Toutain, and Daniel Tenbrinck. On the p-laplacian and infinity-laplacian on graphs with applications in image and data processing. *SIAM Journal on Imaging Sciences*, 8(4), 2015.

[28] Montacer Essid and Justin Solomon. Quadratically regularized optimal transport on graphs. *SIAM Journal on Scientific Computing*, 40(4), 2018.

[29] Lawrence C Evans. Partial differential equations and monge-kantorovich mass transfer. *Current developments in mathematics*, 1997(1), 1997.

[30] Lawrence C Evans and Wilfrid Gangbo. *Differential equations methods for the Monge-Kantorovich mass transfer problem*. American Mathematical Soc., 1999.

[31] Zhongxi Fang, Jianming Huang, Xun Su, and Hiroyuki Kasai. Wasserstein graph distance based on l1–approximated tree edit distance between weisfeiler–lehman subtrees. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, 2023.

[32] Alfred Galichon. *Optimal transport methods in economics*. Princeton University Press, 2018.

[33] Wilfrid Gangbo and Robert J McCann. The geometry of optimal transportation. *Acta Mathematica*, 177, 1996.

[34] Jun Ge and Fengming Dong. Spanning trees in complete bipartite graphs and resistance distance in nearly complete bipartite graphs. *Discrete Applied Mathematics*, 283, 2020.

[35] Augusto Gerolin. Benamou-brenier's approach for ott. In *Optimal Transport and Applications*. Summer School, Lake Arrowhead, 2013.

[36] Severino V Gervacio. Resistance distance in complete n-partite graphs. *Discrete Applied Mathematics*, 203, 2016.

[37] Philip Greengard, Jeremy G Hoskins, Nicholas F Marshall, and Amit Singer. On a linearization of quadratic wasserstein distance. *arXiv preprint arXiv:2201.13386*, 2022.

[38] Steven Haker, Lei Zhu, Allen Tannenbaum, and Sigurd Angenent. Optimal mass transport for registration and warping. *International Journal of computer vision*, 60, 2004.

[39] Lawrence Hubert and Phipps Arabie. Comparing partitions. *Journal of classification*, 2, 1985.

[40] Palle Jorgensen and Erin PJ Pearse. *Operator theory and analysis of infinite networks*, volume 7. World Scientific, 2023.

[41] Leonid V Kantorovich. On the translocation of masses. In *Dokl. Akad. Nauk. USSR (NS)*, volume 37, 1942.

[42] Nguyen Lu Dang Khoa and Sanjay Chawla. Large scale spectral clustering using resistance distance and spielman-teng solvers. In *Discovery Science: 15th International Conference, DS 2012, Lyon, France, October 29-31, 2012. Proceedings 15*. Springer, 2012.

[43] Varun Khurana, Harish Kannan, Alexander Cloninger, and Caroline Moosmüller. Supervised learning of sheared distributions using linearized optimal transport. *Sampling Theory, Signal Processing, and Data Analysis*, 21(1), 2023.

[44] Matt Kusner, Yu Sun, Nicholas Kolkin, and Kilian Weinberger. From word embeddings to document distances. In *International conference on machine learning*. PMLR, 2015.

[45] Tam Le, Makoto Yamada, Kenji Fukumizu, and Marco Cuturi. Tree-sliced variants of wasserstein distances. *Advances in neural information processing systems*, 32, 2019.

[46] Tam Le, Truyen Nguyen, Dinh Phung, and Viet Anh Nguyen. Sobolev transport: A scalable metric for probability measures with graph metrics. In *International Conference on Artificial Intelligence and Statistics*, pages 9844–9868. PMLR, 2022.

[47] Tam Le, Truyen Nguyen, and Kenji Fukumizu. Scalable unbalanced sobolev transport for measures on a graph. In *International Conference on Artificial Intelligence and Statistics*. PMLR, 2023.

[48] Tam Le, Truyen Nguyen, and Kenji Fukumizu. Generalized sobolev transport for probability measures on a graph. *arXiv preprint arXiv:2402.04516*, 2024.

[49] Haibin Ling and Kazunori Okada. An efficient earth mover's distance algorithm for robust histogram comparison. *IEEE transactions on pattern analysis and machine intelligence*, 29(5), 2007.

[50] László Lovász. Random walks on graphs. *Combinatorics, Paul erdos is eighty*, 2(1-46), 1993.

[51] 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.

[52] László Lovász and Peter Winkler. Mixing of random walks and other diffusions on a graph. *London Mathematical Society Lecture Note Series*, 1995.

[53] Maxime Mathey-Prevot and Alain Valette. Wasserstein distance and metric trees. *L'Enseignement Mathématique*, 69(3), 2023.

[54] Quentin Mérigot, Alex Delalande, and Frederic Chazal. Quantitative stability of optimal transport maps and linearization of the 2-wasserstein space. In *International Conference on Artificial Intelligence and Statistics*. PMLR, 2020.

[55] Stacy Miller. The problem of redistricting: the use of centroidal voronoi diagrams to build unbiased congressional districts. *Senior project, Whitman College*, 2007.

[56] Gaspard Monge. Mémoire sur la théorie des déblais et des remblais. *Mem. Math. Phys. Acad. Royale Sci.*, 1781.

[57] LUIGI Montrucchio and Giovanni Pistone. Kantorovich distance on a weighted graph. *arXiv preprint arXiv:1905.07547*, 1420, 2019.

[58] Caroline Moosmüller and Alexander Cloninger. Linear optimal transport embedding: Provable wasserstein classification for certain rigid transformations and perturbations. *arXiv preprint arXiv:2008.09165*, 2020.

[59] Florentin Münch and Radosław K Wojciechowski. Ollivier ricci curvature for general graph laplacians: heat equation, laplacian comparison, non-explosion and diameter bounds. *Advances in Mathematics*, 356, 2019.

[60] Canh Hao Nguyen and Hiroshi Mamitsuka. New resistance distances with global information on large graphs. In *Artificial intelligence and statistics*. PMLR, 2016.

[61] Victor M Panaretos and Yoav Zemel. Statistical aspects of wasserstein distances. *Annual review of statistics and its application*, 6, 2019.

[62] Gabriel Peyré, Marco Cuturi, et al. Computational optimal transport. *Center for Research in Economics and Statistics Working Papers*, 1(2017-86), 2017.

[63] Rémi Peyre. Comparison between w2 distance and h- 1 norm, and localization of wasserstein distance. *ESAIM: Control, Optimisation and Calculus of Variations*, 24(4), 2018.

[64] Aaditya Ramdas, Nicolás García Trillos, and Marco Cuturi. On wasserstein two-sample testing and related families of nonparametric tests. *Entropy*, 19(2), 2017.

[65] Sawyer Robertson, Dhruv Kohli, Gal Mishne, and Alexander Cloninger. On a generalization of wasserstein distance and the beckmann problem to connection graphs. *arXiv preprint arXiv:2312.10295*, 2023.

[66] Andrew Rosenberg and Julia Hirschberg. V-measure: A conditional entropy-based external cluster evaluation measure. In *Proceedings of the 2007 joint conference on empirical methods in natural language processing and computational natural language learning (EMNLP-CoNLL)*, 2007.

[67] Ernest K Ryu, Yongxin Chen, Wuchen Li, and Stanley Osher. Vector and matrix optimal mass transport: theory, algorithm, and applications. *SIAM Journal on Scientific Computing*, 40(5), 2018.

[68] Shota Saito and Mark Herbster. Multi-class graph clustering via approximated effective p-resistance. In *International Conference on Machine Learning*. PMLR, 2023.

[69] Tim Salimans, Han Zhang, Alec Radford, and Dimitris Metaxas. Improving gans using optimal transport. *arXiv preprint arXiv:1803.05573*, 2018.

[70] Areejit Samal, RP Sreejith, Jiao Gu, Shiping Liu, Emil Saucan, and Jürgen Jost. Comparative analysis of two discretizations of ricci curvature for complex networks. *Scientific reports*, 8(1), 2018.

[71] Filippo Santambrogio. Optimal transport for applied mathematicians. *Birkäuser, NY*, 55(58-63), 2015.

[72] Geoffrey Schiebinger, Jian Shu, Marcin Tabaka, Brian Cleary, Vidya Subramanian, Aryeh Solomon, Joshua Gould, Siyan Liu, Stacie Lin, Peter Berube, et al. Optimal-transport analysis of single-cell gene expression identifies developmental trajectories in reprogramming. *Cell*, 176(4), 2019.

[73] Maurice Sion. On general minimax theorems. *Pacific J. Math.*, 8(4), 1958.

[74] Justin Solomon. Optimal transport on discrete domains. *AMS Short Course on Discrete Differential Geometry*, 2018.

[75] Justin Solomon, Raif Rustamov, Leonidas Guibas, and Adrian Butscher. Earth mover's distances on discrete surfaces. *ACM Transactions on Graphics (ToG)*, 33(4), 2014.

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

[77] Stefan Steinerberger. A wasserstein inequality and minimal green energy on compact manifolds. *Journal of Functional Analysis*, 281(5), 2021.

[78] Tomohiro Sugiyama and Kazuhiro Sato. Kron reduction and effective resistance of directed graphs. *SIAM Journal on Matrix Analysis and Applications*, 44(1), 2023.

[79] Yuki Takezawa, Ryoma Sato, and Makoto Yamada. Supervised tree-wasserstein distance. In *International Conference on Machine Learning*. PMLR, 2021.

[80] Yuki Takezawa, Ryoma Sato, Zornitsa Kozareva, Sujith Ravi, and Makoto Yamada. Fixed support tree-sliced wasserstein barycenter. *IEICE Technical Report; IEICE Tech. Rep.*, 122(325), 2022.

[81] MINH TANG, AVANTI ATHREYA, DANIEL L SUSSMAN, VINCE LYZINSKI, and CAREY E PRIEBE. A nonparametric two-sample hypothesis testing problem for random graphs. *Bernoulli*, 2017.

[82] Prasad Tetali. Random walks and the effective resistance of networks. *Journal of Theoretical Probability*, 4, 1991.

[83] Yu Tian, Zachary Lubberts, and Melanie Weber. Curvature-based clustering on graphs. *arXiv preprint arXiv:2307.10155*, 2023.

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

[85] Cédric Villani et al. *Optimal transport: old and new*, volume 338. Springer, 2009.

[86] NX Vinh, J Epps, and J Bailey. Information theoretic measures for clusterings comparison: Variants. *Properties, Normalization and Correction for Chance*, 18, 2009.

[87] Ulrike Von Luxburg. A tutorial on spectral clustering. *Statistics and computing*, 17, 2007.

[88] Wei Wang, Dejan Slepčev, Saurav Basu, John A Ozolek, and Gustavo K Rohde. A linear optimal transportation framework for quantifying and visualizing variations in sets of images. *International journal of computer vision*, 101, 2013.

[89] Nik Weaver. *Lipschitz algebras*. World Scientific, 2018.

[90] Makoto Yamada, Yuki Takezawa, Ryoma Sato, Han Bao, Zornitsa Kozareva, and Sujith Ravi. Approximating 1-wasserstein distance with trees. *Transactions on Machine Learning Research*, 2022.

[91] Luzh Ye and Weigen Yan. Resistance between two vertices of almost complete bipartite graphs. *Discrete Applied Mathematics*, 257, 2019.

[92] Ze Ye, Tengfei Ma, Chien-Chun Ni, Kin Sum Liu, Jie Gao, and Chao Chen. Ricci-{gnn}: Defending against structural attacks through a geometric approach, 2021. URL https://openreview.net/forum?id=_qoQkWNEhS.

[93] Yin Zhang, Rong Jin, and Zhi-Hua Zhou. Understanding bag-of-words model: a statistical framework. *International journal of machine learning and cybernetics*, 1, 2010.

A Proofs

This appendix contains deferred proofs from the remainder of the paper.

A.1 Proofs from section 2

Proof of Theorem 2.1. We can derive the Lagrangian dual in a fairly straightforward manner. We rewrite the constraint $BJ = \alpha - \beta$ using the expresion

$$\sup_{\phi \in \ell(V)} \phi^T(\alpha - \beta) - \phi^T(BJ) = \begin{cases} 0 & \text{if } BJ = \alpha - \beta \\ +\infty & \text{otherwise.} \end{cases} \tag{82}$$

Thus the Beckmann problem may be rewritten as the following

$$\inf_{BJ = \alpha - \beta} \|J\|_{w,p} = \inf_J \left\{ \|J\|_{w,p} + \sup_\phi \phi^T(\alpha - \beta) - \phi^T(BJ) \right\} \tag{83}$$

$$= \sup_\phi \left\{ \phi^T(\alpha - \beta) + \inf_J \|J\|_{w,p} - \phi^T(BJ) \right\}, \tag{84}$$

where we exchange the sup and inf using, e.g., Sion's minimax theorem [73] or Slater's condition. Now we have that, using the Legendre transform of $\|\cdot\|_p$,

$$\inf_J \|J(e)\|_{w,p} - \phi^T(BJ) = \inf_J \|J\|_{w,p} - J^T(B^T\phi) \tag{85}$$

$$= \begin{cases} 0 & \text{if } \|B^T\phi\|_{w^{1-q},q} \leqslant 1 \\ -\infty & \text{otherwise} \end{cases} \tag{86}$$

since the dual norm to $\|\cdot\|_{w,p}$ is $\|\cdot\|_{w^{1-q},q}$. Therefore, the Lagrangian dual becomes

$$\sup \left\{ \phi^T(\alpha - \beta) : \phi \in \ell(V), \left\|B^T\phi\right\|_{w^{1-q},q} \leqslant 1 \right\} \tag{87}$$

The special case of $p = 1$ and $q = \infty$ follows from the same proof, with the only change being that the dual norm to $\|\cdot\|_{w,1}$ is $\|\cdot\|_{w^{-1},\infty}$, which can be shown directly or through a limiting argument. $\square$

Proof of Proposition 2.5. Let $i, j \in V$ be fixed distinct vertices. We begin by noting simply that by inspection,

$$\widetilde{B}(\pi_{ij} I_{P_{ij}}) = \pi_{ij}(\delta_i - \delta_j)$$

and thus that

$$\widetilde{B}\left(\sum_{i,j \in V} \pi_{ij} I_{P_{ij}}\right) = \sum_{i,j \in V} \pi_{ij}(\delta_i - \delta_j) \tag{88}$$

$$= \sum_{i \in V} \delta_i (\pi \mathbf{1})_i - \sum_{j \in V} \delta_j (\mathbf{1}^T \pi)_j \tag{89}$$

$$= \alpha - \beta. \tag{90}$$

Since $J^\pi \geqslant 0$ by construction, J^π is a feasible edge flow for $\mathcal{B}_p(\alpha, \beta)$. The estimate on $\|J^\pi\|_{w,p}$ follows immediately from the triangle inequality. $\square$

Proof of Theorem 2.7. Suppose $\pi \in \Pi(\alpha, \beta)$ is an optimal coupling for $\mathcal{W}_{d_p,p}(\alpha, \beta)$. For each i, j, let $P_{ij} \in \mathscr{P}(i, j)$ be a choice of path which is minimal in the sense of d_p, and let $J^\pi = J^{\pi,(P_{ij})}$ be the feasible flow as in Proposition 2.5. Then by Hölder's inequality,

$$
\begin{aligned}
\mathcal{B}_p(\alpha,\beta) &\leqslant \|J^\pi\|_{w,p} && (91)\\
&\leqslant \sum_{i,j\in V} \pi_{ij}\|I_{P_{ij}}\|_{w,p} && (92)\\
&\leqslant \sum_{i,j\in V} \pi_{ij} d_p(i,j) && (93)\\
&\leqslant \|\mathbf{1}_{V\times V}\|_q \left(\sum_{i,j\in V} \pi_{ij}^p d_p(i,j)^p\right)^{1/p} && (94)\\
&\leqslant \|\mathbf{1}_{V\times V}\|_q \left(\sum_{i,j\in V} \pi_{ij} d_p(i,j)^p\right)^{1/p} = n^{2/q}\mathcal{W}_{d_p,p}(\alpha,\beta) && (95)
\end{aligned}
$$

since $\pi_{ij} \leqslant 1$ and $p > 1$. □

Proof of Proposition 2.13. The proof of this result is straightforward and comes in two parts. First, we observe that $\mathcal{W}_p(\alpha,\beta) = \mathcal{W}_p(\widehat{\alpha},\widehat{\beta})$ where

$$
\mathcal{W}_p(\widehat{\alpha},\widehat{\beta})^p = \inf\left\{\int_{\mathbb{R}\times\mathbb{R}} |x-y|^p \mathrm{d}\pi(x,y) : \pi \in \Pi(\widehat{\alpha},\widehat{\beta})\right\}. \tag{96}
$$

In this case, $\Pi(\widehat{\alpha},\widehat{\beta})$ is the set of all Borel probability measures on $\mathbb{R}\times\mathbb{R}$ with respective marginals α,β. Equality here holds because any such $\pi \in \Pi(\widehat{\alpha},\widehat{\beta})$, since it has discrete marginals, must itself be discrete and supported on the Cartesian product $\{1,\dots,n\}\times\{1,\dots,n\}$; i.e., it can be written as a matrix. Thus, there is a one-to-one correspondence between the feasible couplings on $V\times V$ and $\mathbb{R}\times\mathbb{R}$, and the cost of any such pair of corresponding couplings is identical since shortest path distance on $V(P_n)$ is $|i-j|$.

The second part of the proof is classical, and it suffices to observe that the p-Wasserstein distance between probability measures on $\mathbb{R}$ is simply the p-norm distance between their inverse cdfs. For a proof we recommend, e.g., [64]. □

Proof of Proposition 2.14. We note that the signed incidence matrix B, which is of shape $n\times n-1$, must have no kernel- for the rank of B agrees with the rank of $BB^T = L$, which is $n-1$ for any connected graph on n nodes. Therefore, for *any* $1\leqslant p<\infty$, $\mathcal{B}_p(\alpha,\beta)$ is the p-norm of the unique flow J such that $BJ = \alpha-\beta$ and which does not depend on p. We therefore need only show that $B(K_\alpha - K_\beta) = \alpha-\beta$. To wit, if $i=1$, note that

$$
BK_\alpha(1) = K_\alpha(1,2) = \alpha(1),
$$

and if $1<i<n$,

$$
BK_\alpha(i) = K_\alpha(i,i+1) - K_\alpha(i-1,i) = \alpha(i),
$$

and lastly if $i=n$,

$$
BK_\alpha(n) = -K_\alpha(n-1,n) = \alpha(n)-1.
$$

Therefore, with a simple cancellation, it holds $B(K_\alpha - K_\beta) = \alpha-\beta$ for all $i\in V$. The first equality in the proposition therefore follows, and the second follows immediately upon inspection. □

Proof of Proposition 2.15. Once again the proof of this result is a matter of rigid feasibility: there can only really be one feasible flow $J\in\ell(E')$ satisfying $BJ=\alpha-\beta$ owing to, e.g., rank considerations for Laplacians on trees. Thus it is enough only to establish that $B(K_\alpha-K_\beta)=\alpha-\beta$. Let $i\in V$ be fixed, and suppose i has N incoming oriented edges, and M outgoing oriented edges. At the root of each incoming oriented edge is a subtree I_ℓ, and at the head of each outgoing oriented edge is a subtree O_ℓ. This setup is illustrated in

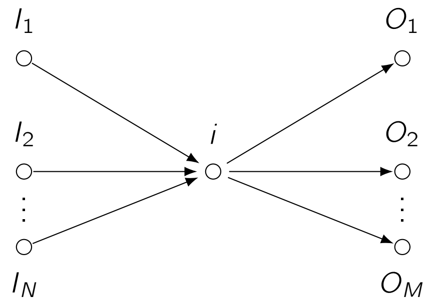


Figure 7: A sketch of a vertex neighborhood in an oriented tree.

Fig. 7. For each $1 \leqslant \ell \leqslant N$, let $\alpha(I_\ell)$ be the total mass of α on that component, and similarly for $\alpha(O_\ell)$. We split up the argument into a few cases depending on M, N. Assume for a moment that $M \geqslant 2$ and $N \geqslant 1$; then by inspection we have

$$\begin{aligned}
(BK_\alpha)(i) &= -(\alpha(I_1) + \alpha(I_2) + \ldots + \alpha(I_N)) && (97)\\
&\quad + \left(\alpha(i) + \sum_{\ell=1}^{N} \alpha(I_\ell) + \sum_{\ell \neq 1} \alpha(O_\ell)\right) + \cdots + \left(\alpha(i) + \sum_{\ell=1}^{N} \alpha(I_\ell) + \sum_{\ell \neq M} \alpha(O_\ell)\right) && (98)\\
&= M\alpha(i) + (M-1)\sum_{\ell=1}^{N} \alpha(I_\ell) + (M-1)\sum_{\ell=1}^{M} \alpha(O_\ell) && (99)\\
&= \alpha(i) + M - 1. && (100)
\end{aligned}$$

If $M = 1$ and $N \geqslant 1$, then $(BK_\alpha)(i) = \alpha(i)$. If $M = 0$ and $N \geqslant 1$, then $(BK_\alpha)(i) = \alpha(i) - 1$. If $N = 0$, then $(BK_\alpha)(i) = \alpha(i) + M - 1$, as before, since the only difference is the lack of contribution of the I_ℓ's. Therefore, $BK_\alpha = \alpha + c$ where $c \in \mathbb{R}^n$ is a vector which does not depend on α, and thus it holds that $B(K_\alpha - K_\beta) = \alpha - \beta$. The claim follows. □

A.2 Proofs from section 3

Proof of Theorem 3.13. We start by considering the pointwise value of the vector $L^\dagger(\alpha - \beta)$. To this end, a useful observation is that for any coupling $\pi \in \Pi(\alpha, \beta)$, we have

$$\alpha - \beta = \sum_{i,j \in V} \pi_{ij}(\delta_i - \delta_j).$$

With this in hand alongside Theorem 3.12, we can calculate

$$L^\dagger(\alpha-\beta)_i = \sum_{j\in V}\left((L^\dagger)_{ij}\sum_{k,\ell\in V}\pi_{k\ell}(\delta_k(j)-\delta_\ell(j))\right) \tag{101}$$

$$= \frac{1}{\mathrm{vol}(G)}\sum_{j,k,\ell\in V}(H(\rho,j)-H(i,j))\,\pi_{k\ell}(\delta_k(j)-\delta_\ell(j)) \tag{102}$$

$$= \frac{1}{\mathrm{vol}(G)}\sum_{k,\ell\in V}\pi_{k\ell}\,(H(\rho,k)-H(i,k)-H(\rho,\ell)+H(i,\ell)) \tag{103}$$

$$= \frac{1}{\mathrm{vol}(G)}\sum_k \alpha_k(H(\rho,k)-H(i,k)) - \sum_\ell \beta_\ell(H(\rho,\ell)-H(i,\ell)) \tag{104}$$

$$= \frac{1}{\mathrm{vol}(G)}\sum_k(\alpha_k-\beta_k)(H(\rho,k)-H(i,k)) \tag{105}$$

$$= c - \frac{1}{\mathrm{vol}(G)}\sum_k(\alpha_k-\beta_k)H(i,k) \tag{106}$$

where by introducing the placeholder c for some $c\in\mathbb{R}$, we are separating the first piece of the summand which does not depend on i and which will be eliminated in the subsequent calculation. Then, multiplying by $(\alpha-\beta)^T$, and using the fact that $(\alpha-\beta)^T\mathbf{1}_n = 0$, we get

$$(\alpha-\beta)^T L^\dagger(\alpha-\beta) = -\frac{1}{\mathrm{vol}(G)}\sum_{i,k\in V}(\alpha_i-\beta_i)(\alpha_k-\beta_k)H(i,k) \tag{107}$$

$$= -\frac{1}{\mathrm{vol}(G)}\sum_{k\in V}(\alpha_k-\beta_k)\sum_{i\in V}(\alpha_i H(i,k)-\beta_i H(i,k)) \tag{108}$$

$$= -\frac{1}{\mathrm{vol}(G)}\sum_{k\in V}(\alpha_k-\beta_k)(H(\alpha,k)-H(\beta,k)) \tag{109}$$

using Proposition 3.11(i) in the final line. The claim follows. □

Proof of Corollary 3.14. From the proof of Theorem 3.13 and Hölder's inequality, we have

$$r_{\alpha\beta} = -\frac{1}{\mathrm{vol}(G)}\sum_{i,k\in V}(\alpha_i-\beta_i)(\alpha_k-\beta_k)H(i,k) \tag{110}$$

$$= \frac{1}{\mathrm{vol}(G)}\sum_k(\beta_k-\alpha_k)\sum_i(\alpha_i-\beta_i)H(i,k) \tag{111}$$

$$\leqslant \frac{1}{\mathrm{vol}(G)}\|\alpha-\beta\|_1\left\|\sum_i(\alpha_i-\beta_i)H(i,k)\right\|_\infty \tag{112}$$

$$\leqslant \frac{2}{\mathrm{vol}(G)}\max\{H(\alpha,\beta),H(\beta,\alpha)\}. \tag{113}$$

where in the final line we make use of Proposition 3.11(ii). The second inequality in the Corollary follows from the observation in the proof of Theorem 3.13 that if $\pi\in\Pi(\alpha,\beta)$ is any coupling, $\alpha-\beta=\sum_{i,j\in V}\pi_{ij}(\delta_i-\delta_j)$. Then since $x\mapsto\|L^{-1/2}x\|_2^2$ is convex in x,

$$(\alpha-\beta)^T L^\dagger(\alpha-\beta) = \left\|L^{-1/2}(\alpha-\beta)\right\|_2^2 \tag{114}$$

$$= \left\|\sum_{i,j}\pi_{ij}L^{-1/2}(\delta_i-\delta_j)\right\|_2^2 \tag{115}$$

$$\leqslant \sum_{i,j}\pi_{ij}r_{ij}. \tag{116}$$

Thus in the special case where $\pi = \alpha\beta^T$,

$$r_{\alpha\beta} \leqslant \frac{1}{\mathrm{vol}(G)} \sum_{i,j \in V} \alpha_i \beta_j (H(i,j) + H(j,i)) \tag{117}$$

$$= \frac{1}{\mathrm{vol}(G)} (H_n(\alpha,\beta) + H_n(\beta,\alpha)). \tag{118}$$

□

Proof of Theorem 3.22. Let $\mu_t \in C^1([0,1])$ such that $\mu_0 = \alpha$ and $\mu_1 = \beta$. Then we estimate, using Proposition 3.20 and the fact that $L^\dagger = L^\dagger L L^\dagger$,

$$\int_0^1 \|\mathrm{d}\mu_t\|^2_{\dot{H}^{-1}(V)}\, \mathrm{d}t = \int_0^1 (\mathrm{d}\mu_t)^T L^\dagger \mathrm{d}\mu_t \mathrm{d}t \tag{119}$$

$$= \int_0^1 (L^\dagger \mathrm{d}\mu_t)^T L (L^\dagger \mathrm{d}\mu_t) \mathrm{d}t \tag{120}$$

$$= \int_0^1 \sum_{(i,j) \in E'} w_{ij} \left| L^\dagger \mathrm{d}\mu_t(i) - L^\dagger \mathrm{d}\mu_t(j) \right|^2 \mathrm{d}t \tag{121}$$

$$= \sum_{(i,j) \in E'} w_{ij} \int_0^1 \left| L^\dagger \mathrm{d}\mu_t(i) - L^\dagger \mathrm{d}\mu_t(j) \right|^2 \mathrm{d}t \tag{122}$$

$$\geqslant \sum_{(i,j) \in E'} w_{ij} \left| \int_0^1 (L^\dagger \mathrm{d}\mu_t(i) - L^\dagger \mathrm{d}\mu_t(j)) \mathrm{d}t \right|^2 \tag{123}$$

where the final inequality follows from Jensen's inequality. Then, focusing on the inner term and letting $e = (i,j) \in E'$ we evaluate

$$\int_0^1 (L^\dagger \mathrm{d}\mu_t(i) - L^\dagger \mathrm{d}\mu_t(j)) \mathrm{d}t = B^T \left(\int_0^1 L^\dagger \mathrm{d}\mu_t \mathrm{d}t \right)(e) \tag{124}$$

$$= B^T L^\dagger \left(\int_0^1 \mathrm{d}\mu_t \mathrm{d}t \right)(e) \tag{125}$$

$$= B^T L^\dagger (\beta - \alpha)(e). \tag{126}$$

(127)

Hence,

$$\sum_{(i,j) \in E'} w_{ij} \left| \int_0^1 (L^\dagger \mathrm{d}\mu_t(i) - L^\dagger \mathrm{d}\mu_t(j)) \mathrm{d}t \right|^2 = \sum_{(i,j) \in E'} w_{ij} \left| B^T L^\dagger (\beta - \alpha)(i,j) \right|^2 \tag{128}$$

$$= (L^\dagger(\beta - \alpha))^T L (L^\dagger (\beta - \alpha)) \tag{129}$$

$$= (\alpha - \beta)^T L^\dagger (\alpha - \beta), \tag{130}$$

and therefore,

$$\inf \left\{ \int_0^1 \|\mathrm{d}\mu_t\|^2_{\dot{H}^{-1}(V)}\, \mathrm{d}t : \mu_t \in C^1([0,1]), \mu_0 = \alpha, \mu_1 = \beta \right\} \geqslant (\alpha - \beta)^T L^\dagger (\alpha - \beta) = \mathcal{B}_2(\alpha,\beta)^2. \tag{131}$$

For the reverse inequality, it is sufficient to consider the special curve $\mu_t = (1-t)\alpha - t\beta$, which satisfies $\int_0^1 \|\mathrm{d}\mu_t\|^2_{\dot{H}^{-1}(V)}\, \mathrm{d}t = (\alpha - \beta)^T L^\dagger (\alpha - \beta) = \mathcal{B}_2(\alpha,\beta)^2$. □

Proof of Theorem 3.25. We observe first that the collection of vectors $\{dT\}_{T\in\mathcal{T}} \subset \ell(V)$ is convex; for the first two conditions (i) and (ii) are equivalent to the intersection of a finite number of linear constraints; and (iii) is stable under convex combinations. Since $L^{-1/2}$ is a linear map, the image $L^{-1/2}\mathcal{A}_i$ will be convex for $i = 1, 2$. Separately, we have for $S, T \in \mathcal{T}$:

$$\|T\alpha_1 - S\alpha_2\|_2 = \|(\alpha_1 - \alpha_2) + (dT - dS)\|_2 \tag{132}$$

$$\geqslant |\|\alpha_1 - \alpha_2\|_2 - \|dT - dS\|_2| \tag{133}$$

$$\geqslant \|\alpha_1 - \alpha_2\|_2 - 2\delta > 0. \tag{134}$$

Thus,

$$\min_{S,T\in\mathcal{T}} \|T\alpha_1 - S\alpha_2\|_2 > 0 \tag{135}$$

and hence $\mathcal{A}_1 \cap \mathcal{A}_2 = \varnothing$. Note moreover that $L^{-1/2}$ (and more generally, any power of L) will act nonsingularly on mean zero vectors since its kernel is exactly the constant functions. Therefore, since $T\alpha_1 - S\alpha_2$ is mean zero for any T, S, and since $\mathcal{A}_1$ and $\mathcal{A}_2$ are disjoint, the sets $L^{-1/2}\mathcal{A}_1$ and $L^{-1/2}\mathcal{A}_2$ are disjoint as well; and thus they are linearly separable. $\square$