# Generalized Dimension Reduction Using Semi-Relaxed Gromov-Wasserstein Distance

**Ranthony A. Clark**[1], **Tom Needham**[2], **Thomas Weighill**[3]

[1] Department of Mathematics, Duke University, Durham NC
[2] Department of Mathematics, Florida State University, Tallahassee FL
[3] Department of Mathematics & Statistics, University of North Carolina at Greensboro, Greensboro NC
ranthony.clark@duke.edu, tneedham@fsu.edu, t_weighill@uncg.edu

## Abstract

Dimension reduction techniques typically seek an embedding of a high-dimensional point cloud into a low-dimensional Euclidean space which optimally preserves the geometry of the input data. Based on expert knowledge, one may instead wish to embed the data into some other manifold or metric space in order to better reflect the geometry or topology of the point cloud. We propose a general method for manifold-valued multidimensional scaling based on concepts from optimal transport. In particular, we establish theoretical connections between the recently introduced semi-relaxed Gromov-Wasserstein (srGW) framework and multidimensional scaling by solving the Monge problem in this setting. We also derive novel connections between srGW distance and Gromov-Hausdorff distance. We apply our computational framework to analyze ensembles of political redistricting plans for states with two Congressional districts, achieving an effective visualization of the ensemble as a distribution on a circle which can be used to characterize typical neutral plans, and to flag outliers.

**Code** — github.com/thomasweighill/srgw-embedding
**Extended version** — arxiv.org/abs/2405.15959

## 1 Introduction

Dimension reduction is a fundamental task in unsupervised learning and is frequently a first step in the data exploration pipeline. Typically, dimension reduction is framed as the process of determining an embedding of a finite metric space $(X, d)$ into a low-dimensional Euclidean space $\mathbb{R}^m$ which optimally preserves the geometry of the input data. As a concrete example, for $X = \{x_1, \dots, x_n\}$, the **metric multidimensional scaling (MDS) problem** seeks a point cloud $\mathrm{MDS}_m(X) \subset \mathbb{R}^m$ satisfying

$$\mathrm{MDS}_m(X) \in \operatorname*{argmin}_{y_1,\dots,y_n} \sum_{i,j=1}^{n} (d(x_i, x_j) - \|y_i - y_j\|)^2. \quad (1)$$

Based on prior knowledge, it may be natural to theorize that a dataset $X$ is noisily sampled from a distribution on a specific low-dimensional manifold in the ambient data space, and the practitioner may wish for this structure to be reflected in the dimension reduction process—indeed, this was the case for the geospatial data studied in Section 5, which catalyzed the ideas in this paper. A simple observation is that the MDS problem (1) still makes sense if $\mathbb{R}^m$ is replaced with some other low-dimensional Riemannian manifold, and the main goal of this paper is to develop a theoretical and computational framework for this manifold-valued variant of MDS. Another observation is that the objective of the MDS problem is very similar to that of the Gromov-Wasserstein (GW) distance from optimal transport theory (Mémoli 2011), and our main theoretical result makes this connection precise.

**Contributions.** The contributions of this paper are centered on theoretical results which tie together several ideas from classical and more recent literature. The impetus for the paper was the problem of visualizing complex redistricting datasets; our solution provides an extended computational example that illustrates how our framework can be used to gain important insights into non-Euclidean data. More specifically, our main contributions are:

- We extend the formulation of semi-relaxed Gromov-Wasserstein distance introduced in (Vincent-Cuaz et al. 2022) to a 1-parameter family of $L^p$-type distances defined for general metric measure spaces (including continuous spaces). We then show in Corollary 6 that the semi-relaxed GW problem generalizes the MDS problem (1) in several ways. This is based on Theorem 2, which shows that the semi-relaxed GW distance is realized by a Monge map in very general situations, thus adding to the growing recent literature on the existence of Monge maps in the GW framework (see Remark 4).
- We develop and unify the theory of generalized MDS and semi-relaxed GW distances by exhibiting connections to variants of the Gromov-Hausdorff (GH) distance that have appeared previously in the literature. Theorem 8 and Theorem 9 together show that a symmetrized version of the $p = \infty$ semi-relaxed GW distance is equal to the modified GH distance of (Mémoli 2012), which, in turn, appeared in classical work on generalized MDS (Bronstein, Bronstein, and Kimmel 2006).
- By drawing the connection between the MDS and srGW problems, we are able to design an efficient algorithm (SRGW+GD) for computing MDS embeddings into mani-

Copyright © 2025, Association for the Advancement of Artificial Intelligence (www.aaai.org). All rights reserved.

folds, consisting of an initial discretized optimal transport computation followed by a gradient descent stage. This allows general target spaces, and produces significantly better embeddings (quantitatively and qualitatively) than a naive gradient descent algorithm (Table 2 and Figure 1).

- Using experiments on the MNIST dataset, we show that SRGW+GD matches or outperforms SMACOF MDS for Euclidean target spaces. We demonstrate its effectiveness for embeddings into spheres on a dataset of rotated MNIST images, and a set of GPS coordinates of cities.
- In the final section, we apply SRGW+GD to ensembles of political districting plans for low-population states, achieving a natural and effective visualization of each ensemble as a distribution on a circle. These visualizations lead to pertinent insights into the distribution of possible districting plans for each state.

## 2 Semi-Relaxed Gromov-Wasserstein Distance and Multidimensional Scaling

**Gromov-Wasserstein Distances.** Let $(X, d_X, \mu_X)$ and $(Y, d_Y, \mu_Y)$ be **metric measure spaces (mm-spaces)**; that is, $(X, d_X)$ is a metric space, which we assume to be complete and separable, and $\mu_X$ is a Borel probability measure on $X$. We will sometimes abuse notation and simply write $X$ for $(X, d_X, \mu_X)$ when it is clear that there is an associated choice of metric and measure.

A **coupling** of $\mu_X$ and $\mu_Y$ is a Borel probability measure $\gamma$ on $X \times Y$ with marginals equal to $\mu_X$ and $\mu_Y$, respectively. Writing this symbolically, $(\pi_1)_{\#}\gamma = \mu_X$ and $(\pi_2)_{\#}\gamma = \mu_Y$, where $\pi_1 : X \times Y \to X$ and $\pi_2 : X \times Y \to Y$ are the coordinate projection maps and $(\pi_1)_{\#}\gamma$ denotes the pushforward measure. We denote the set of all measure couplings between $\mu_X$ and $\mu_Y$ as $\mathcal{C}(\mu_X, \mu_Y)$. For $p \in [1, \infty)$, the **Gromov-Wasserstein (GW) $p$-distance** (Memoli 2007; Mémoli 2011) is

$$d_{\mathrm{GW},p}(X, Y) = \inf_{\gamma \in \mathcal{C}(\mu_X, \mu_Y)} \mathrm{dis}_p(\gamma),$$

where the **$p$-distortion** of $\gamma$, $\mathrm{dis}_p(\gamma)$, is given by

$$\mathrm{dis}_p(\gamma) = \frac{1}{2}\|\Gamma_{X,Y}\|_{L^p(\gamma\otimes\gamma)} \tag{2}$$
$$= \frac{1}{2}\left(\iint |\Gamma_{X,Y}(x, y, x', y')|^p \, d\gamma(x, y)\, d\gamma(x', y')\right)^{1/p}$$

with $\Gamma_{X,Y}(x, y, x', y') = d_X(x, x') - d_Y(y, y')$.

This extends to the $p = \infty$ case, where the distortion is equal to

$$\mathrm{dis}_\infty(\gamma) = \frac{1}{2}\|\Gamma_{X,Y}\|_{L^\infty(\mathrm{supp}(\gamma)\times\mathrm{supp}(\gamma))}$$
$$= \frac{1}{2}\sup_{(x,y),(x',y')\in\mathrm{supp}(\gamma)} |\Gamma_{X,Y}(x, y, x', y')|,$$

with $\mathrm{supp}(\gamma)$ denoting the support of the measure $\gamma$.

The GW distance was introduced in (Memoli 2007), where it was shown that $d_{\mathrm{GW},p}$ defines a metric on the space of measure-preserving isometry classes of fully supported compact mm-spaces.

**Semi-Relaxed GW Distances.** GW distances have become a popular tool in the machine learning community, due to their ability to compare distinct data types (Peyré, Cuturi, and Solomon 2016; Chowdhury and Needham 2021; Xu, Luo, and Carin 2019). Many variants of GW distances have been introduced in recent years (Scetbon, Peyré, and Cuturi 2022; Chowdhury, Miller, and Needham 2021; Chowdhury et al. 2023; Redko et al. 2020; Vayer et al. 2020). Of particular interest here is the *semi-relaxed Gromov-Wasserstein (srGW) distance* of (Vincent-Cuaz et al. 2022). There, the $p = 2$ version was defined for finite spaces, and we give a natural generalization here.

**Definition 1** (Semi-Relaxed Gromov-Wasserstein Distance)**.** Let $X = (X, d_X, \mu_X)$ and $Y = (Y, d_Y, \mu_Y)$ be mm-spaces and let $p \in [1, \infty]$. A **semi-coupling** of $\mu_X$ and $Y$ is a Borel probability measure $\gamma$ on $X \times Y$ such that $(\pi_1)_{\#}\gamma = \mu_X$ (note that this definition doesn't depend on the measure on $Y$). The set of semi-couplings will be denoted $\mathcal{SC}(\mu_X, Y)$. The **semi-relaxed Gromov-Wasserstein (srGW) $p$-distance** is

$$d_{\mathrm{srGW},p}(X, Y) = \inf_{\gamma \in \mathcal{SC}(\mu_X, Y)} \mathrm{dis}_p(\gamma), \tag{3}$$

where the $p$-distortion is as defined in (2).

We call $d_{\mathrm{srGW},p}$ a "distance" in an informal sense. It is clearly asymmetric and, in fact, we will show in Section 3 that it satisfies the triangle inequality if and only if $p = \infty$. For now, we consider (3) as an interesting optimization problem. We show below that it is closely related to MDS (1).

**Monge Maps and Generalized MDS.** Let $(X, d_X, \mu_X)$ be a mm-space and $Y$ a Polish space. Given a measurable function $f : X \to Y$, we define the **semi-coupling induced by** $f$ to be the measure $\mu_f$ on $X \times Y$ given by $\mu_f := (\mathrm{id}_X \times f)_{\#}\mu_X$, where $\mathrm{id}_X \times f : X \to X \times Y$ is the function $x \mapsto (x, f(x))$. In the case that $X$ is finite, this measure is given explicitly by

$$\mu_f = \sum_{x \in X} \mu_X(x)\delta_{(x, f(x))}.$$

Here, and throughout the rest of the paper, we use $\delta_z$ to denote the Dirac mass at a point $z \in Z$.

We recall that a metric space is called **proper** if its closed and bounded sets are compact. An action of a group $G$ on a metric space $X$ is said to be **cocompact** if there exists a compact $K \subseteq X$ such that $X$ is covered by translates of $K$ under the $G$-action. The following is our main result.

**Theorem 2** (Existence of Monge Maps)**.** *Let $(X, d_X, \mu_X)$ be a metric measure space with $X$ finite and $\mu_X$ fully supported and let $(Y, d_Y)$ be a proper metric space with a cocompact action by isometries by some group $G$. Then for any $p \in [1, \infty]$, there exists a function $f : X \to Y$ such that*

$$d_{\mathrm{srGW},p}(X, Y) = \mathrm{dis}_p(\mu_f).$$

*Moreover, if $p < \infty$, then any semi-coupling with the same distortion as $\mu_f$ is induced by a function.*

The proof is included in the arXiv version. The main idea is that, from any initial coupling, one can construct a new one of the form $\mu_f$ with lower distortion via disintegration of the initial coupling.

**Remark 3.** *The theorem applies when $Y = \mathbb{R}^n$, endowed with Euclidean distance (a main motivating example), but also to a large class of spaces including compact metric spaces (e.g. closed Riemannian manifolds) or infinite binary trees. Regarding the source space $X$, our proof requires $X$ to be finite; removing this assumption, even in the Euclidean case, seems difficult – see e.g. (Murray and Pickarski 2024).*

**Remark 4** (The Monge Problem)**.** *We will refer to the map $f$ in Theorem 2 as a **Monge map**, in reference to the original formulation of optimal transport, due to Monge, where optimization was performed over measure-preserving maps, rather than over couplings (see (Villani 2021)). The **Monge problem** in OT theory is to determine conditions under which the Wasserstein distance is realized by a measure-preserving map; the problem is now well-understood, with several general results, e.g., (Brenier 1987) and (Villani et al. 2009, Theorem 10.41). The Monge problem in the GW setting is an active area of current research. The state-of-the-art results in this direction appear in (Dumont, Lacombe, and Vialard 2024), where the problem is solved for variants of the $p = 2$ GW distance between measures on Euclidean spaces with density. Other recent results for (variants of) the Monge problem for GW distance between more restrictive subclasses of mm-spaces appear in (Vayer 2020; Sturm 2023; Beinert, Heiss, and Steidl 2023; Salmona, Delon, and Desolneux 2022; Mémoli and Needham 2024).*

**Example 5.** *Without the assumptions on $Y$ made in Theorem 2, there may be no Monge map from $X$ to $Y$. Indeed, let $X = \{0, 1\} \subseteq \mathbb{R}$ and let $Y = \{(0, 2n) \mid n \in \mathbb{N}\} \cup \{(1 + 2^{-n}, 2n) \mid n \in \mathbb{N}\}$. Since $X$ does not isometrically embed into $Y$, there is no zero distortion map from $X$ to $Y$. However, the maps $f_n : X \to Y$ given by $f(0) = (0, 2n),\ f(1) = (1 + 2^{-n}, 2n)$ have arbitrarily low distortion.*

We have the following corollary, showing that the srGW problem generalizes the MDS problem.

**Corollary 6.** *Let $(X, d)$ be a finite metric space with $X = \{x_1, \ldots, x_n\}$ and let $\mu$ be uniform measure on $X$. Let $f : X \to \mathbb{R}^m$ be a function which realizes $d_{\mathrm{srGW},2}(X, \mathbb{R}^m)$. Then the point cloud $f(x_1), \ldots, f(x_n) \in \mathbb{R}^m$ is a solution of* $\mathrm{MDS}_m(X)$.

**Related Work.** The semi-relaxed GW problem was first studied in (Vincent-Cuaz et al. 2022), where it was applied to graph machine learning problems such as dictionary learning and graph completion. The followup paper (Van Assel et al. 2024) considers theoretical aspects of srGW and, in particular, connections to dimension reduction. Their result (Van Assel et al. 2024, Theorem 3.2) is analogous to Corollary 6 and says that spectral methods of dimensional reduction can be realized as solutions to semi-relaxed GW problems. This covers, for example, classical multidimensional scaling (as opposed to metric MDS, studied here). These spectral methods are only applicable to Euclidean (or perhaps hyperbolic) embedding spaces, so that the target applications of (Van Assel et al. 2024) and this paper are fairly disjoint. Our theoretical results are related to and complementary with those of (Van Assel et al. 2024), but neither paper generalizes the other.

The problem of dimension reduction into a non-Euclidean space has been well-studied in the topological data analysis literature. These approaches are fundamentally different than the one used here; they rely on specific constructions of algebraic topology and are therefore only suited to embedding in specific classes of spaces such as circles (De Silva and Vejdemo-Johansson 2009; Paik and Park 2023), projective spaces (Perea 2018), or lens spaces (Polanco and Perea 2019). More importantly, the results of these algorithms are qualitatively different than ours, as they are based on persistent (co)homology rather than on geometry preservation. We compare our results to one of these methods, circular coordinates, in Section 4.

The Gromov-Hausdorff distance is another useful tool for shape comparison and analysis where shapes can be modeled as (compact) metric spaces, versus the Gromov-Wasserstein setting which considers the distributional properties of a sample via metric measure spaces. It is used in areas such as manifold learning, computer vision, computational geometry, and topological data analysis which emphasize the geometry and topological properties of data. Our work is closely related to the influential paper (Bronstein, Bronstein, and Kimmel 2006), which approaches the problem of partial surface matching through a certain "semi-relaxed" version of Gromov-Hausdorff distance. This connection inspired the work in the next section, which derives a precise relationship between the version of Gromov-Hausdorff distance considered in (Bronstein, Bronstein, and Kimmel 2006) and the srGW distance introduced of (Vincent-Cuaz et al. 2022).

## 3 Connections to Gromov-Hausdorff Distance

**Semi-Relaxed Gromov-Hausdorff Distance.** Let $X = (X, d_X)$ and $Y = (Y, d_Y)$ be metric spaces. Recall that the **Gromov-Hausdorff (GH) distance** between $X$ and $Y$ is defined by

$$d_{\mathrm{GH}}(X, Y) = \inf_{R \in \mathcal{R}(X,Y)} \mathrm{dis}(R), \tag{4}$$

where $\mathcal{R}(X, Y)$ is the set of **correspondences** between $X$ and $Y$—that is, the set of relations $R \subset X \times Y$ such that the coordinate projection maps take $R$ surjectively onto each component—and $\mathrm{dis}(R)$ is the **metric distortion** of the correspondence $R$, defined by

$$\mathrm{dis}(R) = \frac{1}{2} \sup_{(x,y),(x',y') \in R} |d_X(x, x') - d_Y(y, y')|. \tag{5}$$

Inspired by the semi-relaxed version of Gromov-Wasserstein distance considered above, we now define a semi-relaxed version of Gromov-Hausdorff distance.

**Definition 7** (Semi-Relaxed Gromov-Hausdorff Distance)**.** Let $X$ and $Y$ be metric spaces. A relation $R \subset X \times Y$ is called a **semi-correspondence** if the coordinate projection to $X$ takes $R$ surjectively onto $X$ (but we put no such condition on the projection map to $Y$). Let $\mathcal{SR}(X, Y)$ denote the set of semi-correspondences. The **semi-relaxed Gromov-Hausdorff (srGH) distance** is

$$d_{\mathrm{srGH}}(X, Y) = \inf_{R \in \mathcal{SR}(X,Y)} \mathrm{dis}(R),$$

with $\mathrm{dis}(R)$ defined as in (5). This is symmetrized as

$$\widehat{d}_{\mathrm{srGH}}(X,Y) = \max\{d_{\mathrm{srGH}}(X,Y), d_{\mathrm{srGH}}(Y,X)\}.$$

**Equivalence to Modified Gromov-Hausdorff Distance.** We will now show that the symmetrized srGH distance is a reformulation of a distance which has already appeared in the literature. Recall (see, e.g., (Burago, Burago, and Ivanov 2022)) that the GH distance can be expressed as

$$d_{\mathrm{GH}}(X,Y) = \inf_{f,g} \max\{\mathrm{dis}(f), \mathrm{dis}(g), \mathrm{codis}(f,g)\}, \quad (6)$$

where the infimum is over (not necessarily continuous) functions $f : X \to Y$ and $g : Y \to X$, the **function distortion** $\mathrm{dis}(f)$ is defined by

$$\mathrm{dis}(f) = \frac{1}{2} \sup_{x,x' \in X} |d_X(x,x') - d_Y(f(x), f(x'))|,$$

with $\mathrm{dis}(g)$ defined similarly, and where the **codistortion** $\mathrm{codis}(f,g)$ is defined by

$$\mathrm{codis}(f,g) = \frac{1}{2} \sup_{x \in X, y \in Y} |d_X(x, g(y)) - d_Y(y, f(x))|.$$

This formulation leads to a natural modification, where the maps $f$ and $g$ are decoupled by dropping the codistortion term in (6). The **modified Gromov-Hausdorff distance** is

$$\begin{aligned} d_{\mathrm{mGH}}(X,Y) &= \inf_{f,g} \max\{\mathrm{dis}(f), \mathrm{dis}(g)\} \\ &= \max\left\{ \inf_{f:X\to Y} \mathrm{dis}(f), \inf_{g:Y\to X} \mathrm{dis}(g) \right\}. \end{aligned}$$

This distance was introduced in (Mémoli 2012), where it was shown to be a metric on the space of isometry classes of compact metric spaces (Mémoli 2012, Theorem 4.1). Our next main result shows that it is the same as the symmetrized semi-relaxed GH distance.

**Theorem 8** (Equivalence of Gromov-Hausdorff Distances)**.** *The symmetrized semi-relaxed Gromov-Hausdorff distance $\widehat{d}_{\mathrm{srGH}}$ is equal to the modified Gromov-Hausdorff distance $d_{\mathrm{mGH}}$.*

The proof is given in the arXiv version. We also provide an additional characterization of semi-relaxed GH distance in terms of isometric embeddings in the arXiv version.

**Connection to Semi-Relaxed Gromov-Wasserstein Distance.** There is an apparent relationship between the $p = \infty$ version of (semi-relaxed) Gromov-Wasserstein distance and (semi-relaxed) Gromov-Hausdorff distance. Indeed, given mm-spaces $(X, d_X, \mu_X)$ and $(Y, d_Y, \mu_Y)$ with fully-supported measures, any coupling $\gamma \in \mathcal{C}(\mu_X, \mu_Y)$ induces a correspondence $\mathrm{supp}(\gamma)$, and it follows that

$$d_{\mathrm{GH}}(X,Y) \leq d_{\mathrm{GW},\infty}(X,Y), \quad (7)$$

where the quantity on the left is understood as the Gromov-Hausdorff distance between the underlying metric spaces. However, the inequality (7) is not an equality, in general—see (Mémoli 2011, Theorem 5.1). The following result shows that equality does hold in the semi-relaxed setting. We define the **symmetrized srGW distance** $\hat{d}_{\mathrm{srGW},p}$ by

$$\widehat{d}_{\mathrm{srGW},p}(X,Y) = \max\{d_{\mathrm{srGW},p}(X,Y), d_{\mathrm{srGW},p}(Y,X)\}.$$

**Theorem 9** (Equivalence of srGW and srGH)**.** *Let $X$ and $Y$ be mm-spaces such that $\mu_X$ has full support. Then*

$$\widehat{d}_{\mathrm{srGW},\infty}(X,Y) = \widehat{d}_{\mathrm{srGH}}(X,Y) = d_{\mathrm{mGH}}(X,Y).$$

*Moreover, $\widehat{d}_{\mathrm{srGW},p}$ defines a metric on the space of isometry classes of compact metric spaces which is topologically equivalent to Gromov-Hausdorff distance on any GH precompact family of compact metric spaces when $p = \infty$, but does not define a pseudometric for $p < \infty$.*

The theorem is proved in the arXiv version. It is based on the observation that any semi-correspondence can be approximated by a measurable semi-correspondence with an arbitrarily small change in distortion.

**Remark 10.** *The version of generalized MDS used in (Bronstein, Bronstein, and Kimmel 2006) that we mentioned in the related work section is the asymmetric modified GH problem, $\inf_{f:X\to Y} \mathrm{dis}(f)$ and its $\ell^p$ relaxation, which the authors of that paper apply to find embeddings of subsets of geodesic spaces to aid in partial surface matching. Our results Corollary 6 and Theorem 9 give a cohesive connection between various ideas: the embedding problem considered in (Bronstein, Bronstein, and Kimmel 2006) is an application of the modified GH distance of (Mémoli 2012), which is equal to the $p = \infty$ version of a semi-relaxed GW distance, whereas the $p = 2$ version of srGW first introduced in (Vincent-Cuaz et al. 2022) generalizes the standard MDS problem.*

# 4 Numerical Implementation and Experiments

**Implementation.** In this section we describe how to use the semi-relaxed Gromov Wasserstein distance to find an embedding of a finite metric space $(X, d)$ into a smooth Riemannian manifold $Y$. The algorithm is straightforward: in short, we solve a srGW problem to embed $X$ into a predefined finite subset of $Y$, then use this as initialization for gradient descent of the MDS functional (1), with $Y$ as the target metric space, which could be more general than $\mathbb{R}^m$. We now provide some details.

Given a finite metric space $(X, d)$, we construct an embedding $\hat{f} : X \to Y$ as follows. We first pick a discrete finite subset $S \subseteq Y$ to map into. In practice, we often use a grid in some coordinate system for $Y$, and perturb the points slightly to make it easier to solve the resulting optimization problem. We then solve the semi-relaxed Gromov-Wasserstein problem (3), $d_{\mathrm{srGW},2}(X,S)$, which, by Theorem 2, yields an optimal Monge map $f : X \to S$. The computation of $d_{\mathrm{srGW},2}(X,S)$ is approximated via the srGW implementation in the Python Optimal Transport package (Flamary et al. 2021); although this approximation is not guaranteed to yield a Monge map, we found that it does so in practice.

To compute the required embedding $\hat{f} : X \to Y$, we run a gradient descent, initialized with the embedding $f$ (sometimes with a small perturbation); as we will see below, this drastically improves the likelihood of finding a good local minimum by gradient descent. Let $X = \{x_1, \ldots, x_n\}$ and

initialize a set of $n$ points $(y_i)_{1\leq i\leq n}$ in $Y$ via $y_i = f(x_i)$. We then consider the distortion function $Y^n \to \mathbb{R}$ defined by

$$(y_1, \dots, y_n) \mapsto \frac{1}{2} \sum_{i,j=1}^{n} \left(d_X(x_i, x_j) - d_Y(y_i, y_j)\right)^2 \quad (8)$$

(cf. the MDS functional (1)) and use a gradient-based method on $Y$ to find a local minimum $(\hat{y}_1, \hat{y}_2, \dots, \hat{y}_n)$. Our embedding is then given by $\hat{f}(x_i) = \hat{y}_i$. We refer to this method as **SRGW+GD**. In our examples below, we use the Adam optimizer (Kingma and Ba 2014).

In practice, the desired embedding space $Y$ may only be known up to certain hyperparamters, such as scale. Our main example below will be when $Y$ is a circle of unknown radius. If $Y$ depends on a scale factor (or multiple scale factors), we add the scale factor as an additional variable in (8).

**MNIST.** We first benchmark SRGW+GD against other dimension reduction methods in the case when the target space is Euclidean. We use the MNIST dataset consisting of 70,000 $28 \times 28$ grayscale images of handwritten digits, which we view as vectors in $\mathbb{R}^{784}$, separated into ten smaller datasets (MNIST0 to MNIST9), one for each digit. We embed each dataset into $\mathbb{R}^2$ using PCA, SMACOF MDS, and SRGW+GD. Table 1 reports the distortion ($\mathrm{dis}_2$) values for each embedding. We see that SRGW+DG and SMACOF achieve similar distortion in all cases, with SRGW+GD performing slightly better. In terms of compute time, we find that SRGW+GD is between 3 and 14 times faster then SMACOF MDS (using scikit-learn).

| | PCA | SMACOF MDS | SRGW+GD |
|---|---|---|---|
| MNIST0 | 2.580 | 1.556 | 1.554 |
| MNIST1 | 1.290 | 0.794 | 0.777 |
| MNIST2 | 3.070 | 1.761 | 1.751 |
| MNIST3 | 2.792 | 1.594 | 1.586 |
| MNIST4 | 2.668 | 1.523 | 1.518 |
| MNIST5 | 2.702 | 1.602 | 1.594 |
| MNIST6 | 2.549 | 1.495 | 1.477 |
| MNIST7 | 2.346 | 1.350 | 1.342 |
| MNIST8 | 2.901 | 1.646 | 1.623 |
| MNIST9 | 2.421 | 1.405 | 1.388 |

Table 1: Distortion for embeddings of MNIST data into $\mathbb{R}^2$.

**Rotated MNIST.** In order to demonstrate the performance of SRGW+GD when the target space is non-Euclidean, we artifically introduce non-linear structure into the MNIST datasets. We apply a random rotation between $0^\circ$ and $360^\circ$ to each image in each MNIST dataset (filling gaps with black pixels) to create new datasets R-MNIST0 to R-MNIST9. We embed these datasets into $\mathbb{R}^2$ using t-SNE, PCA, SMACOF MDS and SRGW+GD. We also embed each dataset into a circle of unknown radius using SRGW+GD and the following comparison methods:

- **GD** denotes minimizing (8) with a random initalization and the Adam optimizer. We use 10 random initializations and report the min and max distortion over all trials.
- **CC** is the circular coordinate method introduced in (De Silva and Vejdemo-Johansson 2009), which constructs a map from a finite metric space $X$ to $S^1$ using persistent cohomology. We used the density-robust version introduced in (Paik and Park 2023). Since the method only produces circular coordinates and not a radius, we estimate the radius as $\max_{x,y\in X} d(x, y)/\pi$.

Distortion values are contained in Table 2.[1] For embeddings into $\mathbb{R}^2$, we again find that SRGW+GD and SMACOF MDS have similar distortion values, all significantly lower than PCA. SRGW+GD achieves higher distortion on $S^1$ than on $\mathbb{R}^2$ (since the target space has one less dimension and thus captures less variation), but still achieves a lower distortion on $S^1$ than PCA achieves on $\mathbb{R}^2$. SRGW+GD achieves a significantly lower distortion on $S^1$ than CC does, and lower distortion than GD in all cases except for some trials on R-MNIST1.

We can also gain some insight into how SRGW+GD differs qualitatively from other methods. In Figure 1, we show the image of the embedding for the dataset R-MNIST9, colored by the true angle of rotation. We also plot the true angle against the inferred angular coordinate (for $\mathbb{R}^2$ this is taken to be the angle of the point from the $x$-axis). We note that CC maps true rotation angles to angular coordinates in a roughly injective way, thereby capturing the rotation process accurately. t-SNE maps the rotation angle roughly injectively onto a thickened curve in the plane, but inferring an angular coordinate with our naive method does not recover the rotation angle correctly. PCA, MDS and both versions of SRGW+GD map true rotation angles to angular coordinates via a roughly degree two map. This better captures the global geometry, since an image of a 9 is often closer to its $180^\circ$ rotation than its $90^\circ$ rotation. Finally, we see that GD recovers none of the rotation structure in the dataset.

In general, it can be hard to infer an angular coordinate from a planar embedding (see t-SNE in Figure 1). Even our naive method above requires finding an appropriate center for the data, which might not be the mean if the data is distributed very unevenly (e.g. in the redistricting application below). The advantage of choosing a circle as the target space is that it produces a well-defined angular coordinate. Our experiments demonstrate that regardless of which target space is preferred, SRGW+GD effectively preserves global geometry. They also demonstrate the necessity of srGW embeddings as an initialization point for gradient descent.

**Cities.** To demonstrate a non-Euclidean embedding where approximate isometric embedding is possible, we use a list of the 20 largest cities[2], with the geodesic distance on the Earth between every pair of cities as ground truth (this distance does not assume the Earth is a perfect sphere, and instead uses the WGS-84 ellipsoid). Using SRGW+GD, we embed this dataset into a sphere of radius 6371 (the average radius of the Earth in kilometers). Figure 2 shows the embedding. SRGW+GD achieves an embedding that is almost

[1] While t-SNE does not aim to reduce distortion and thus is not a fair comparison, it is a helpful contrast for the qualitative behavior of SRGW+GD; we include distortion values for completeness

[2] https://simplemaps.com/data/world-cities, CC-BY 4.0 license

| | $\mathbb{R}^2$ embeddings | | | | $S^1$ embeddings | | |
|---|---|---|---|---|---|---|---|
| | t-SNE | PCA | SMACOF MDS | SRGW+GD | CC | GD (min,max) | SRGW+GD |
| R-MNIST0 | 28.196 | 3.508 | 1.989 | 1.987 | 5.467 | (2.810, 2.812) | 2.534 |
| R-MNIST1 | 33.631 | 2.032 | 1.200 | 1.202 | 1.818 | (1.709, 2.247) | 1.702 |
| R-MNIST2 | 30.199 | 3.615 | 2.020 | 2.019 | 5.586 | (2.854, 2.856) | 2.591 |
| R-MNIST3 | 30.410 | 3.261 | 1.873 | 1.873 | 5.459 | (2.787, 2.790) | 2.439 |
| R-MNIST4 | 29.236 | 3.593 | 1.880 | 1.828 | 4.846 | (2.488, 2.490) | 2.329 |
| R-MNIST5 | 28.459 | 3.159 | 1.812 | 1.812 | 5.325 | (2.734, 2.736) | 2.401 |
| R-MNIST6 | 32.086 | 3.598 | 1.942 | 1.900 | 2.758 | (2.621, 2.623) | 2.437 |
| R-MNIST7 | 32.798 | 3.347 | 1.828 | 1.796 | 2.594 | (2.537, 2.539) | 2.324 |
| R-MNIST8 | 28.310 | 3.296 | 1.870 | 1.869 | 5.368 | (2.735, 2.737) | 2.418 |
| R-MNIST9 | 32.468 | 3.307 | 1.783 | 1.775 | 2.638 | (2.464, 2.466) | 2.280 |

Table 2: Distortion ($\mathrm{dis}_2$) for embeddings of randomly rotated MNIST data using various methods.

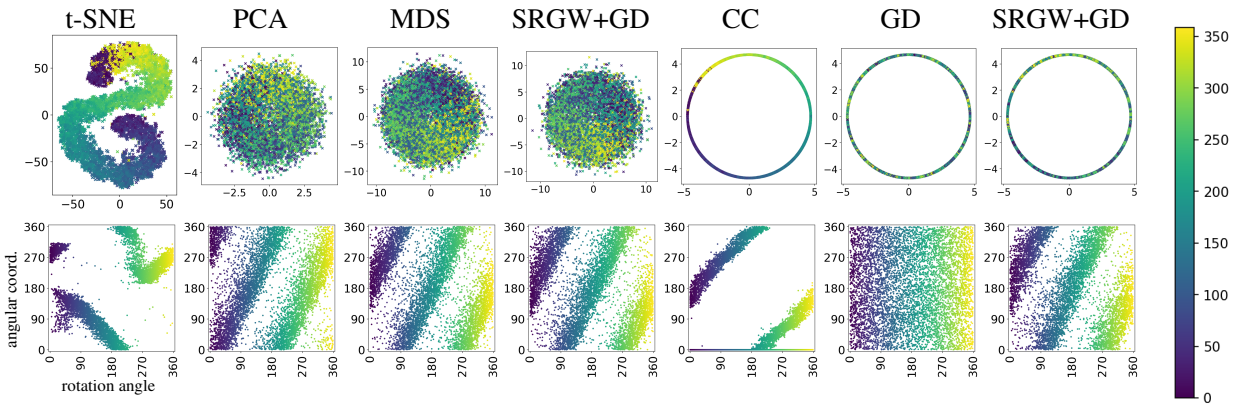



Figure 1: Circle and planar embeddings for the R-MNIST9 dataset (above), and plots comparing the true angle of rotation vs the inferred angular coordinate from the embedding (below). Color indicates the true angle of rotation.

isometric; the pairwise distances between the embedded points never differ by more than 14 kilometers from the true distance. By contrast, an MDS embedding into $\mathbb{R}^3$ achieves a distortion of 264.724, indicating the benefit of choosing a target space with the appropriate metric (a geodesic sphere).

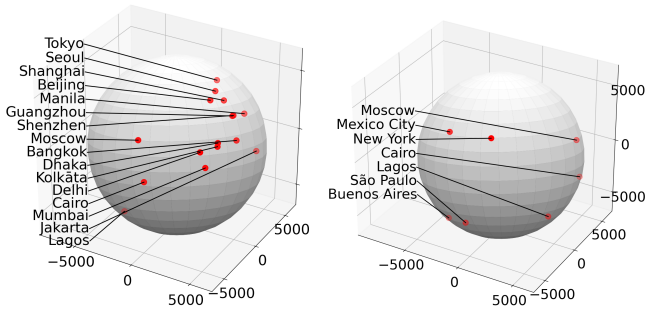



Figure 2: Embedding onto a geodesic sphere of 20 world cities.

## 5 Application to Redistricting

We now demonstrate how an embedding into a natural non-Euclidean target can enable visualization of a complex data set, resulting in important insights, using computational redistricting as our area of application.

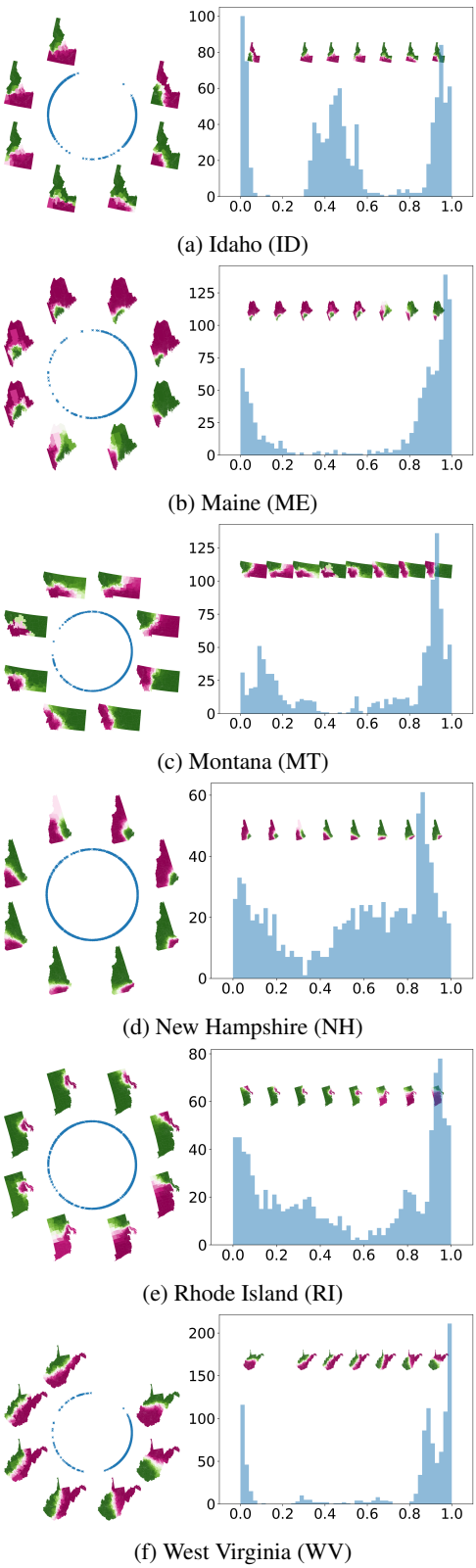
**Background and Data.** Redistricting is the process of dividing a region into contiguous, equal population districts for the purposes of electing representatives. There has been a lot of recent attention on generating redistricting **ensembles** – large samples from the space of valid redistricting plans for a given U.S. state (Chen and Rodden 2015; Chikina, Frieze, and Pegden 2017; Herschlag et al. 2020; DeFord, Duchin, and Solomon 2021; Duchin, Needham, and Weighill 2022). When analyzed, ensembles can uncover baseline expectations for a typical plan, or be used to flag outliers (some of which might be so-called gerrymanders, i.e. unfair maps). We will use our SRGW+GD method to visualize ensembles of two-district plans in order to achieve both these goals, similar to the approach in (Abrishami et al. 2020).

There are currently six states in the contiguous United States with two Congressional districts: Idaho (ID), Maine (ME), Montana (MT), New Hampshire (NH), Rhode Island (RI) and West Virginia (WV). For each of these states we

obtained Census blockgroups from (Manson et al. 2023) and generated an ensemble of 1,000 redistricting plans using the ReCom algorithm (DeFord, Duchin, and Solomon 2021). As our distance between plans we chose a **Hamming distance** where the distance between two redistricting plans is defined as the minimum number of Census blockgroups that must be reassigned to change the first plan into the second. Treating the ensemble as a 1,000-point metric space with this distance, we then embed the ensemble into a circle with SRGW+GD. For each embedded ensemble, we plot the image of the embedding as a set of points on the circle, as shown in Figure 3. We also display the average division for each part of the circle (see the arXiv version for details), and histograms showing the distribution of circular coordinates in each ensemble. In the arXiv version, we try other nonlinear planar embeddings and find that none of them reveal the circle structure within the data across all states.

**Results.** In general, we see that circular coordinates roughly parameterize the angle of the boundary: from a north-south division, round to an east-west division and then back to a north-south division. This is most easily visible for West Virginia and Montana. This is strong evidence that the circle is a good choice of target space for embedding these ensembles. The boundary does not always look linear for states with a very uneven population distribution such as Maine (where most of the population is in the south of the state), though we still see a smooth rotation.

A general trend we can observe is a preference for divisions of the state with short (internal) boundary length. Boundary length is one possible measure of "compactness", a redistricting criterion often written into legislation around redistricting. The ReCom algorithm is known to favor low boundary lengths, where the boundary is measured by the number of Census blockgroups (or other geographic units) on the boundary of the districts (DeFord, Duchin, and Solomon 2021; Procaccia and Tucker-Foltz 2022). Preference for short boundaries can be observed in West Virginia, where the northwest-southeast division requires a long boundary and is thus not likely to be drawn by the ReCom algorithm. In Idaho, we see a distribution with at least two modes: a northwest-southeast division and a northeast-southwest division. We should note that in Idaho, a straight, vertical boundary is so unlikely that it doesn't even show up on the heat maps; this is likely because most of the large boundary length this would require. In Maine, most plans are concentrated around northeast-southwest split. In New Hampshire most plans are concentrated around a roughly north-south split, though there is more variance than for Maine. These results for Maine and New Hampshire align with the analysis in (Asgari et al. 2020) of redistricting in these states. In particular, the authors of (Asgari et al. 2020) conclude that the enacted west-east redistricting plan in New Hampshire was an outlier compared to their ensemble, and propose that incumbent protection may have played a role in the drawing of that map. Our analysis also suggests that a east-west division of the state would be an outlier when compared to the ensemble.



(a) Idaho (ID)

(b) Maine (ME)

(c) Montana (MT)

(d) New Hampshire (NH)

(e) Rhode Island (RI)

(f) West Virginia (WV)

Figure 3: Circle embeddings of 1,000-plan ensembles using SRGW+GD. Heat maps indicate the average district location for plans in each part of the circle.

## Acknowledgments

This material is based upon work supported by the National Science Foundation under the following grants. T. Needham was supported by NSF grants DMS–2107808 and DMS–2324962. R. A. Clark was supported by the NSF MPS Ascending Postdoc Award 2138110. We would like to thank the anonymous reviewers for their helpful suggestions during the review process.

## References

Abrishami, T.; Guillen, N.; Rule, P.; Schutzman, Z.; Solomon, J.; Weighill, T.; and Wu, S. 2020. Geometry of graph partitions via optimal transport. *SIAM Journal on Scientific Computing*, 42(5): A3340–A3366.

Asgari, S.; Basewitz, Q.; Bergmann, E.; Brogsol, J.; Cox, N.; Davis, D.; Kampel, M.; Keating, B.; Knox, K.; Lam, A.; et al. 2020. Assessing congressional districting in Maine and New Hampshire. *arXiv preprint arXiv:2011.06555*.

Beinert, R.; Heiss, C.; and Steidl, G. 2023. On assignment problems related to Gromov–Wasserstein distances on the real line. *SIAM Journal on Imaging Sciences*, 16(2): 1028–1032.

Brenier, Y. 1987. Décomposition polaire et réarrangement monotone des champs de vecteurs. *CR Acad. Sci. Paris Sér. I Math.*, 305: 805–808.

Bronstein, A. M.; Bronstein, M. M.; and Kimmel, R. 2006. Generalized multidimensional scaling: a framework for isometry-invariant partial surface matching. *Proceedings of the National Academy of Sciences*, 103(5): 1168–1172.

Burago, D.; Burago, Y.; and Ivanov, S. 2022. *A course in metric geometry*, volume 33. American Mathematical Society.

Chen, J.; and Rodden, J. 2015. Cutting through the thicket: Redistricting simulations and the detection of partisan gerrymanders. *Election Law Journal*, 14(4): 331–345.

Chikina, M.; Frieze, A.; and Pegden, W. 2017. Assessing significance in a Markov chain without mixing. *Proceedings of the National Academy of Sciences*, 114(11): 2860–2864.

Chowdhury, S.; Miller, D.; and Needham, T. 2021. Quantized Gromov-Wasserstein. In *Machine Learning and Knowledge Discovery in Databases. Research Track: European Conference, ECML PKDD 2021, Bilbao, Spain, September 13–17, 2021, Proceedings, Part III 21*, 811–827. Springer.

Chowdhury, S.; and Needham, T. 2021. Generalized spectral clustering via Gromov-Wasserstein learning. In *International Conference on Artificial Intelligence and Statistics*, 712–720. PMLR.

Chowdhury, S.; Needham, T.; Semrad, E.; Wang, B.; and Zhou, Y. 2023. Hypergraph co-optimal transport: Metric and categorical properties. *Journal of Applied and Computational Topology*, 1–60.

De Silva, V.; and Vejdemo-Johansson, M. 2009. Persistent cohomology and circular coordinates. In *Proceedings of the twenty-fifth annual symposium on Computational geometry*, 227–236.

DeFord, D.; Duchin, M.; and Solomon, J. 2021. Recombination: A family of Markov chains for redistricting. *Harvard Data Science Review*, 3(1): 3.

Duchin, M.; Needham, T.; and Weighill, T. 2022. The (homological) persistence of gerrymandering. *Foundations of Data Science*, 4(4): 581–622.

Dumont, T.; Lacombe, T.; and Vialard, F.-X. 2024. On the existence of Monge maps for the Gromov–Wasserstein problem. *Foundations of Computational Mathematics*, 1–48.

Flamary, R.; Courty, N.; Gramfort, A.; Alaya, M. Z.; Boisbunon, A.; Chambon, S.; Chapel, L.; Corenflos, A.; Fatras, K.; Fournier, N.; et al. 2021. Pot: Python optimal transport. *Journal of Machine Learning Research*, 22(78): 1–8.

Herschlag, G.; Kang, H. S.; Luo, J.; Graves, C. V.; Bangia, S.; Ravier, R.; and Mattingly, J. C. 2020. Quantifying gerrymandering in north carolina. *Statistics and Public Policy*, 7(1): 30–38.

Kingma, D. P.; and Ba, J. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

Manson, S.; Schroeder, J.; Riper, D. V.; Knowles, K.; Kugler, T.; Roberts, F.; and Ruggles, S. 2023. IPUMS National Historical Geographic Information System: Version 18.0 [dataset]. http://doi.org/10.18128/D050.V18.0.

Memoli, F. 2007. On the use of Gromov-Hausdorff Distances for Shape Comparison. In Botsch, M.; Pajarola, R.; Chen, B.; and Zwicker, M., eds., *Eurographics Symposium on Point-Based Graphics*. The Eurographics Association. ISBN 978-3-905673-51-7.

Mémoli, F. 2011. Gromov–Wasserstein distances and the metric approach to object matching. *Foundations of computational mathematics*, 11: 417–487.

Mémoli, F. 2012. Some properties of Gromov–Hausdorff distances. *Discrete & Computational Geometry*, 48: 416–440.

Mémoli, F.; and Needham, T. 2024. Comparison results for Gromov–Wasserstein and Gromov–Monge distances. *ESAIM: Control, Optimisation and Calculus of Variations*, 30: 78.

Murray, R.; and Pickarski, A. 2024. On Probabilistic Embeddings in Optimal Dimension Reduction. *arXiv preprint arXiv:2408.02433*.

Paik, T.; and Park, J. 2023. Circular Coordinates for Density-Robust Analysis. *arXiv preprint arXiv:2301.12742*.

Perea, J. A. 2018. Multiscale projective coordinates via persistent cohomology of sparse filtrations. *Discrete & Computational Geometry*, 59: 175–225.

Peyré, G.; Cuturi, M.; and Solomon, J. 2016. Gromov-Wasserstein averaging of kernel and distance matrices. In *International conference on machine learning*, 2664–2672. PMLR.

Polanco, L.; and Perea, J. A. 2019. Coordinatizing Data With Lens Spaces and Persistent Cohomology. In *Proceedings of the 31 st Canadian Conference on Computational Geometry (CCCG)*, 49–57.

Procaccia, A. D.; and Tucker-Foltz, J. 2022. Compact redistricting plans have many spanning trees. In *Proceedings of the 2022 Annual ACM-SIAM Symposium on Discrete Algorithms (SODA)*, 3754–3771. SIAM.

Redko, I.; Vayer, T.; Flamary, R.; and Courty, N. 2020. Co-optimal transport. *Advances in neural information processing systems*, 33: 17559–17570.

Salmona, A.; Delon, J.; and Desolneux, A. 2022. Gromov-Wasserstein Distances between Gaussian Distributions. *Journal of Applied Probability*, 59(4).

Scetbon, M.; Peyré, G.; and Cuturi, M. 2022. Linear-time Gromov Wasserstein distances using low rank couplings and costs. In *International Conference on Machine Learning*, 19347–19365. PMLR.

Sturm, K.-T. 2023. *The Space of Spaces: Curvature Bounds and Gradient Flows on the Space of Metric Measure Spaces*, volume 290. American Mathematical Society.

Van Assel, H.; Vincent-Cuaz, C.; Courty, N.; Flamary, R.; Frossard, P.; and Vayer, T. 2024. Distributional Reduction: Unifying Dimensionality Reduction and Clustering with Gromov-Wasserstein Projection. *arXiv preprint arXiv:2402.02239*.

Vayer, T. 2020. A contribution to optimal transport on incomparable spaces. *arXiv preprint arXiv:2011.04447*.

Vayer, T.; Chapel, L.; Flamary, R.; Tavenard, R.; and Courty, N. 2020. Fused Gromov-Wasserstein distance for structured objects. *Algorithms*, 13(9): 212.

Villani, C. 2021. *Topics in optimal transportation*, volume 58. American Mathematical Soc.

Villani, C.; et al. 2009. *Optimal transport: old and new*, volume 338. Springer.

Vincent-Cuaz, C.; Flamary, R.; Corneli, M.; Vayer, T.; and Courty, N. 2022. Semi-relaxed Gromov Wasserstein divergence with applications on graphs. In *International Conference on Learning Representations*.

Xu, H.; Luo, D.; and Carin, L. 2019. Scalable Gromov-Wasserstein learning for graph partitioning and matching. *Advances in neural information processing systems*, 32.