FOURIER SLICED-WASSERSTEIN EMBEDDING FOR MULTISETS AND MEASURES

Tal Amir
Faculty of Mathematics
Technion–Israel Institute of Technology
Haifa, Israel
talamir@technion.ac.il

Nadav Dym
Faculty of Mathematics and Faculty of Computer Science
Technion–Israel Institute of Technology
Haifa, Israel
nadavdym@technion.ac.il

ABSTRACT

We present the *Fourier Sliced Wasserstein (FSW) embedding*—a novel method to embed multisets and measures over $\mathbb{R}^d$ into Euclidean space.

Our proposed embedding approximately preserves the sliced Wasserstein distance on distributions, thereby yielding geometrically meaningful representations that better capture the structure of the input. Moreover, it is injective on measures and *bi-Lipschitz* on multisets—a significant advantage over prevalent methods based on sum- or max-pooling, which are provably not bi-Lipschitz, and in many cases, not even injective. The required output dimension for these guarantees is near-optimal: roughly $2Nd$, where N is the maximal input multiset size.

Conversely, we prove that it is *impossible* to embed distributions over $\mathbb{R}^d$ into Euclidean space in a bi-Lipschitz manner. Thus, the metric properties of our embedding are, in a sense, the best possible.

Through numerical experiments, we demonstrate that our method yields superior multiset representations that improve performance in practical learning tasks. Specifically, we show that (a) a simple combination of the FSW embedding with an MLP achieves state-of-the-art performance in learning the (non-sliced) Wasserstein distance; and (b) replacing max-pooling with the FSW embedding makes PointNet significantly more robust to parameter reduction, with only minor performance degradation even after a 40-fold reduction.[1]

1 INTRODUCTION

Multisets are unordered collections of vectors that account for repetitions. They are the main mathematical tool for representing unordered data, with perhaps the most notable example being point clouds. As such, there is growing interest in developing architectures suited for learning tasks on multisets. To address this need, several permutation-invariant neural networks have been introduced, with applications for point-cloud classification (Qi et al., 2017a), chemical property prediction (Pozdnyakov & Ceriotti, 2023), and image deblurring (Aittala & Durand, 2018). Multiset aggregation functions also serve as key components in more complex architectures, such as Message Passing Neural Networks (MPNNs) for graphs (Gilmer et al., 2017), or setups with multiple permutation actions (Maron et al., 2020).

A central concept in the study of multiset functions, i.e. functions defined on multisets, is *injectivity*. Its importance is highlighted by the following observation: An architecture that cannot separate two distinct multisets $\boldsymbol{X} \neq \boldsymbol{X}'$, will not be able to approximate a target function f that distinguishes between these multisets, i.e. $f(\boldsymbol{X}) \neq f(\boldsymbol{X}')$. Conversely, a model that maps multisets injectively to vectors, composed with an MLP, can universally approximate *all* continuous multiset functions (Zaheer et al., 2017; Dym & Gortler, 2024). This observation has inspired works to study the injectivity of multiset architectures (Wagstaff et al., 2022; 2019; Tabaghi & Wang, 2024). Injectivity on multisets also plays a key role in the development of expressive MPNNs (Xu et al., 2018).

[1] Our code is available in https://github.com/tal-amir/FSW-embedding.

Common multiset architectures are typically based on simple building blocks of the form

$$E\left(\{x_1, \ldots, x_n\}\right) = \text{Pool}\left\{F\left(\boldsymbol{x}^{(1)}\right), \ldots, F\left(\boldsymbol{x}^{(n)}\right)\right\},$$

where F is usually an MLP, and Pool is a simple pooling operation such as maximum, sum, or mean. Xu et al. (2018) showed that multiset functions based on max- or mean-pooling are never injective, but injectivity can be achieved using sum-pooling, under the assumption that the vectors $\boldsymbol{x}^{(i)}$ come from a discrete domain, and an appropriate function F is used. Then it was shown by Zaheer et al. (2017); Maron et al. (2019) that injectivity over multisets with continuous elements can be achieved using sum pooling with a polynomial F. The more common case, in which F is a neural network, was studied in (Amir et al., 2023). There it was shown that injectivity on multisets and measures over $\mathbb{R}^d$ can be achieved using F that is a shallow MLP with random parameters and analytic, non-polynomial activations, such as Sigmoid or Softplus.

However, injectivity alone is not the strongest property one may desire for multiset functions. While an injective multiset embedding E can separate pairs of distinct multisets $\boldsymbol{X} \neq \boldsymbol{X}'$, this does not ensure the *quality* of separation. Ideally, if two multisets $\boldsymbol{X}, \boldsymbol{X}'$ are far apart in terms of some notion of distance, then one would expect $E(\boldsymbol{X}), E(\boldsymbol{X}') \in \mathbb{R}^m$ to also be far apart, and vice versa. The standard mathematical notion used to guarantee this behaviour is *bi-Lipschitzness*.

Definition. Let $E : \mathcal{D} \to \mathbb{R}^m$, where $\mathcal{D}$ is a collection of multisets or, more generally, distributions over $\mathbb{R}^d$. We say that E is *bi-Lipschitz* with respect to $\mathcal{W}_p$, if there exist constants $c, C > 0$ such that

$$c \cdot \mathcal{W}_p(\mu, \tilde{\mu}) \leq \|E(\mu) - E(\tilde{\mu})\| \leq C \cdot \mathcal{W}_p(\mu, \tilde{\mu}), \quad \forall \mu, \tilde{\mu} \in \mathcal{D}, \tag{1}$$

where $\mathcal{W}_p$ denotes the p-Wasserstein distance and $\|\cdot\|$ denotes the ℓ_2 norm.

The Wasserstein distance, defined in the next section, serves as a standard notion of distance for multisets and distributions. The ratio of the Lipschitz constants C/c in (1) represents a bound on the maximal distortion incurred by the map E, analogous to the condition number of a matrix.

Bi-Lipschitz embeddings enable the application of metric-based learning methods, such as nearest-neighbor search, data clustering, and multi-dimensional scaling, to non-Euclidean data types like multisets and measures. These methods can be more readily applied to the embedded Euclidean domain than to the original data domain, where metric computations are typically more expensive (Indyk & Thaper, 2003). The bi-Lipschitzness of the embedding often provides correctness guarantees for these applications, which depend on the Lipschitz constants c and C (Cahill et al., 2024).

Achieving bi-Lipschitzness is typically more challenging than injectivity and often requires different theoretical tools. Recently, in (Amir et al., 2023), we proved that methods based on average- or sum-pooling can never be bi-Lipschitz, while Xu et al. (2018) showed that methods based on max-pooling cannot even be injective. Thus, a bi-Lipschitz embedding would constitute a substantial improvement over the standard multiset embedding methods currently in use.

To date, two main approaches have been proposed for constructing bi-Lipschitz multiset embeddings: (1) *sort embedding* (Balan et al., 2022), which is based on sorting random projections of the multiset elements, and (2) *max filtering* (Cahill et al., 2022), which is based on computations of Wasserstein distances from 'template multisets', and is computationally expensive.

While sort-based methods have been used with some success (Zhang et al., 2019; 2018; Balan et al., 2022), it seems that their popularity in practical applications is still rather limited, despite their bi-Lipschitzness guarantees. Perhaps one of the main reasons for this is that these methods can only handle multisets of fixed size, and, to date, it remains unclear how to generalize them to multisets of varying size, let alone to distributions. This is a major limitation, since multisets of varying size arise naturally in numerous learning tasks, for example graph classification, where vertices may have neighbourhoods of different sizes. This problem is often circumvented via ad-hoc solutions such as padding (Zhang et al., 2018) or interpolation (Zhang et al., 2019), which do not preserve the original theoretical guarantees of the method. Moreover, even in the restricted setting of fixed-size multisets, the bi-Lipschitzness guarantees of these methods require prohibitively high embedding dimensions.

Our goal in this paper is to overcome these limitations by constructing a bi-Lipschitz embedding for the space of all multisets over $\mathbb{R}^d$ with at most N elements. We denote this space by $\mathcal{S}_{\leq N}\left(\mathbb{R}^d\right)$. Note that the assumption of bounded cardinality is necessary, as otherwise, even injectivity is impossible; see (Amir et al., 2023, Theorem C.3). We are also interested in the larger space of probability

distributions over $\mathbb{R}^d$ supported on at most N points, which we denote by $\mathcal{P}_{\leq N}(\mathbb{R}^d)$. This setting, in which the points may have non-uniform weights, can be particularly relevant for attention-based methods on sets (Lee et al., 2019), as well as graph architectures such as GCN (Kipf & Welling, 2016) or GAT (Veličković et al., 2018), which use non-uniform weights for vertex neighbourhoods. In summary, our main goal is:

Main goal For $\mathcal{D} = \mathcal{S}_{\leq N}(\mathbb{R}^d)$ and $\mathcal{D} = \mathcal{P}_{\leq N}(\mathbb{R}^d)$, construct an embedding $E : \mathcal{D} \to \mathbb{R}^m$ that is injective and, preferably, bi-Lipschitz.

Main results We propose an embedding method for multisets and distributions with finite support, which is a non-trivial generalization of the sort embedding. We observe that the Euclidean distance between the sort embeddings of two multisets can be interpreted as a finite Monte Carlo sampling of their *sliced Wasserstein* distance (Bonneel et al., 2015): in the special case where the input consists of multisets of fixed size, this sampling corresponds to the project-and-sort operations used in the sort embedding. Based on this interpretation, we extend beyond fixed-size multisets and propose an embedding method for both $\mathcal{S}_{\leq N}(\mathbb{R}^d)$ and $\mathcal{P}_{\leq N}(\mathbb{R}^d)$. Our method essentially operates as follows: (1) compute random one-dimensional projections (also called *slices*) of the input distribution; (2) compute the *quantile function* of each projected distribution; and (3) sample each quantile function at a random frequency in the Fourier domain. We name our method the *Fourier Sliced-Wasserstein (FSW) embedding* and denote it by E_m^{FSW}. The function $E_m^{\text{FSW}} : \mathcal{P}_{\leq N}(\mathbb{R}^d) \to \mathbb{R}^m$ is of the form

$$E_m^{\text{FSW}}(\mu) = E_m^{\text{FSW}}\left(\mu; \left(\boldsymbol{v}^{(k)}, \xi^{(k)}\right)_{k=1}^m\right).$$

It maps multisets and distributions into $\mathbb{R}^m$, and depends on parameters $\boldsymbol{v}^{(k)} \in \mathbb{R}^d$, $\xi^{(k)} \in \mathbb{R}$, $k \in [m]$, representing projection vectors and frequencies respectively. It has the following properties:

1. **Bi-Lipschitzness on multisets:** For $m \geq 2Nd + 1$, the map $E_m^{\text{FSW}} : \mathcal{S}_{\leq N}(\mathbb{R}^d) \to \mathbb{R}^m$ is bi-Lipschitz (and, in particular, injective) for almost any choice of the parameters $\left(\boldsymbol{v}^{(k)}, \xi^{(k)}\right)_{k=1}^m$ (Theorem 4.1 and Corollary 4.3).
2. **Injectivity on distributions and measures:** For $m \geq 2Nd + 2N - 1$, the map $E_m^{\text{FSW}} : \mathcal{P}_{\leq N}(\mathbb{R}^d) \to \mathbb{R}^m$ is injective (but not bi-Lipschitz) for almost any choice of parameters (Theorem 4.1). Moreover, by adding one more output coordinate, it can be extended to measures that may have total mass different from one, while preserving injectivity; see Appendix A.1. Furthermore, we prove that bi-Lipschitzness on $\mathcal{P}_{\leq N}(\mathbb{R}^d)$ is *impossible* for any Euclidean embedding (Theorem 4.4). This suggests that the metric properties of our embedding are, in a sense, the best possible.
3. **Sliced-Wasserstein approximation:** The expectation of $\frac{1}{m}\left\|E_m^{\text{FSW}}(\mu) - E_m^{\text{FSW}}(\tilde{\mu})\right\|^2$ over the parameters $\left(\boldsymbol{v}^{(k)}, \xi^{(k)}\right)_{k=1}^m$, drawn from our appropriately defined distribution, is exactly the squared sliced-Wasserstein distance between μ and $\tilde{\mu}$ (Corollary 3.3), with the standard error decreasing as $\mathcal{O}\left(\frac{1}{\sqrt{m}}\right)$.
4. **Piecewise smoothness:** The map E_m^{FSW} is continuous and piecewise smooth in both the input measure parameters $\left(\boldsymbol{x}^{(i)}, w_i\right)_{i=1}^N$ and the embedding parameters $\left(\boldsymbol{v}^{(k)}, \xi^{(k)}\right)_{k=1}^m$. Thus, it is amenable to gradient-based learning methods, and its parameters can be trained.
5. **Complexity:** The embedding $E_m^{\text{FSW}}(\mu)$ can be computed efficiently in $\mathcal{O}(mNd + mN \log N)$ time—the same complexity (up to the logarithmic term) as the most efficient methods used in practice (Theorem A.2, Appendix A.2).

In Properties 1 and 2 above, the required embedding dimension m is near-optimal, essentially up to a multiplicative factor of two.

Empirically, in Section 5, we demonstrate the practical promise of our method by evaluating it on the task of learning the (non-sliced) 1-Wasserstein distance function. We show that a simple composition of our embedding with an MLP outperforms the state-of-the-art while requiring shorter training times. Moreover, we show that replacing max-pooling in PointNet with our embedding significantly improves the architecture's robustness in the low-parameter regime, with only minor performance degradation even after a 40-fold parameter reduction.

2 PROBLEM SETTING

In this section, we describe the problem in detail and briefly review its theoretical background and existing approaches.

2.1 THEORETICAL BACKGROUND

We begin by defining the spaces of multisets and distributions of interest, along with the metrics we use on these spaces.

Multisets and distributions Denote by $\mathcal{P}_{\leq N}(\Omega)$ the collection of all probability distributions over a measurable domain $\Omega \subseteq \mathbb{R}^d$, supported on at most N points. Any distribution $\mu \in \mathcal{P}_{\leq N}(\Omega)$ can be parametrized by N points $\boldsymbol{x}^{(i)} \in \Omega$ and weights $w_i \geq 0$ such that

$$\mu = \sum_{i=1}^{N} w_i \delta_{\boldsymbol{x}^{(i)}}, \tag{2}$$

where $\boldsymbol{w} = (w_1, \ldots, w_N)$ is in the probability simplex $\Delta^N \subset \mathbb{R}^N$, and $\delta_{\boldsymbol{x}}$ is Dirac's delta function at $\boldsymbol{x}$. Distributions with less than N support points can be parameterized in this way by setting some of the weights w_i to zero and choosing the corresponding $\boldsymbol{x}^{(i)}$ arbitrarily.

Similarly, let $\mathcal{S}_{\leq N}(\Omega)$ be the collection of all multisets over $\Omega \subseteq \mathbb{R}^d$ with at most N elements. To enable computation of Wasserstein distances, which are defined on distributions, we modify the definition of $\mathcal{S}_{\leq N}(\Omega)$ by excluding the empty multiset and identifying each multiset $\boldsymbol{X} = \left\{\boldsymbol{x}^{(i)}\right\}_{i=1}^{n} \in \mathcal{S}_{\leq N}(\Omega)$ with the distribution $\mu[\boldsymbol{X}] \in \mathcal{P}_{\leq N}(\Omega)$ that assigns uniform weights $w_i = \frac{1}{n}$ to all $\boldsymbol{x}^{(i)}$, taking their multiplicities into account.[2] Throughout this work, our use of $\mathcal{S}_{\leq N}(\Omega)$ assumes this modification, and we regard $\mathcal{S}_{\leq N}(\Omega)$ as a subset of $\mathcal{P}_{\leq N}(\Omega)$.

Note that our embedding in its basic form, to be defined in the next section, will take input multisets from $\mathcal{S}_{\leq N}(\Omega)$. As such, it will not be defined on the empty multiset, and will not distinguish between multisets of different sizes if their element proportions are identical.[3] This can be easily remedied by augmenting the embedding with an additional output coordinate that encodes the multiset cardinality, or in the case of measures, the *total mass* $\sum_{i=1}^{N} w_i$; see Appendix A.1 for details.

Throughout this work, we focus on $\Omega = \mathbb{R}^d$ and only discuss finitely-supported multisets and distributions. Nonetheless, our embedding can accommodate general distributions over $\mathbb{R}^d$, while retaining its sliced-Wasserstein approximation property. Thus, in principle, our method can be applied to structures other than point clouds, for example polygonal meshes and volumetric data.

Wasserstein distance As a distance measure on $\mathcal{S}_{\leq N}\left(\mathbb{R}^d\right)$ and $\mathcal{P}_{\leq N}\left(\mathbb{R}^d\right)$, we use the aforementioned Wasserstein distance. Intuitively, this distance is the least amount of work required to 'transport' one distribution to another. For two distributions $\mu, \tilde{\mu} \in \mathcal{P}_{\leq N}\left(\mathbb{R}^d\right)$, parametrized by points $\boldsymbol{x}^{(i)}, \tilde{\boldsymbol{x}}^{(j)}$ and weights $w_i, \tilde{w}_j$ as in (2), the p-Wasserstein distance between μ and $\tilde{\mu}$ is defined by

$$\mathcal{W}_p(\mu, \tilde{\mu}) \coloneqq \left(\inf_{\pi \in \Pi(\mu, \tilde{\mu})} \sum_{i,j \in [N]} \pi_{ij} \left\| \boldsymbol{x}^{(i)} - \tilde{\boldsymbol{x}}^{(j)} \right\|^p \right)^{\frac{1}{p}} \qquad p \in [1, \infty),$$

$$\mathcal{W}_\infty(\mu, \tilde{\mu}) \coloneqq \inf_{\pi \in \Pi(\mu, \tilde{\mu})} \max \left\{ \left\| \boldsymbol{x}^{(i)} - \tilde{\boldsymbol{x}}^{(j)} \right\| \;\middle|\; i, j \in [N], \pi_{ij} > 0 \right\},$$

where $\|\cdot\|$ is the Euclidean norm, and $\Pi(\mu, \tilde{\mu})$ is the set of all *transport plans* from μ to $\tilde{\mu}$:

$$\Pi(\mu, \tilde{\mu}) \coloneqq \left\{ \pi \in \mathbb{R}^{N \times N} \;\middle|\; (\forall i, j \in [N])\, \pi_{ij} \geq 0 \wedge \sum_{j \in [N]} \pi_{ij} = w_i \wedge \sum_{i \in [N]} \pi_{ij} = \tilde{w}_j \right\}.$$

[2] For example, the multiset $\boldsymbol{X} = \{a, b, b\} \in \mathcal{S}_{\leq 3}(\mathbb{R})$ is identified with the distribution $\mu[\boldsymbol{X}] = \frac{1}{3}\delta_a + \frac{2}{3}\delta_b \in \mathcal{P}_{\leq 3}(\mathbb{R})$.

[3] e.g., $\boldsymbol{X} = \{a, b, b\}$ and $\boldsymbol{Y} = \{a, a, b, b, b, b\}$ are considered identical in $\mathcal{S}_{\leq 6}(\mathbb{R})$, since $\mu[\boldsymbol{X}] = \mu[\boldsymbol{Y}]$.

The numbers π_{ij} intuitively represent the amount of mass to be transported from point $\boldsymbol{x}^{(i)}$ to $\tilde{\boldsymbol{x}}^{(j)}$. Whenever p is omitted, we refer to $\mathcal{W}_p$ with $p = 2$. Similarly, $\|\cdot\|$ always denotes the ℓ_2 norm.

Computation of Wasserstein The Wasserstein distance can be computed in $\mathcal{O}\big(N^3 \log N\big)$ time via a linear program (Altschuler et al., 2017; Orlin, 1988). Alternatively, one may use the Sinkhorn approximation algorithm (Cuturi, 2013), which takes $\tilde{\mathcal{O}}\big(N^2\varepsilon^{-3}\big)$ time, ε being the error tolerance (Altschuler et al., 2017). This was improved to $\tilde{\mathcal{O}}\big(\min\big\{N^{2.25}\varepsilon^{-1}, N^2\varepsilon^{-2}\big\}\big)$ in (Dvurechensky et al., 2018). However, in the special case $d = 1$, it can be computed significantly faster.

Wasserstein in 1D In one dimension, the Wasserstein distance can be computed in only $\mathcal{O}(N \log N)$ time. If $\boldsymbol{x} = (x_1, \ldots, x_n)$, $\boldsymbol{y} = (y_1, \ldots, y_n)$ are two vectors in $\mathbb{R}^n$, then the distance between the two uniform distributions induced by the vector coordinates is given by

$$\mathcal{W}(\mu[\boldsymbol{x}], \mu[\boldsymbol{y}]) = \frac{1}{\sqrt{n}}\|\mathrm{sort}(\boldsymbol{x}) - \mathrm{sort}(\boldsymbol{y})\|, \tag{3}$$

with $\mathrm{sort} : \mathbb{R}^n \to \mathbb{R}^n$ being the function that returns the input coordinates sorted in increasing order.

For arbitrary distributions in $\mathcal{P}_{\leq N}(\mathbb{R})$, the Wasserstein distance can be computed similarly via the *quantile function*. For a distribution μ over $\mathbb{R}$, the quantile function $Q_\mu : [0, 1) \to \mathbb{R}$ is a continuous analog of the sort function, defined by

$$Q_\mu(t) \coloneqq \inf\{x \in \mathbb{R} \mid \mu((-\infty, x]) > t\}.$$

Figure 1 depicts the quantile functions for three distinct multisets.

The quantile function enables an explicit formula for the Wasserstein distance between two distributions over $\mathbb{R}$ (see e.g. Bayraktar & Guo (2021), Eq. 2.3 and the paragraph thereafter):

$$\mathcal{W}(\mu, \tilde{\mu}) = \sqrt{\int_0^1 \big(Q_\mu(t) - Q_{\tilde{\mu}}(t)\big)^2 dt}. \tag{4}$$

Note that when μ and $\tilde{\mu}$ are generated by multisets of the same cardinality (like the two multisets of cardinality 3 in Figure 1), the formulas (4) and (3) coincide.

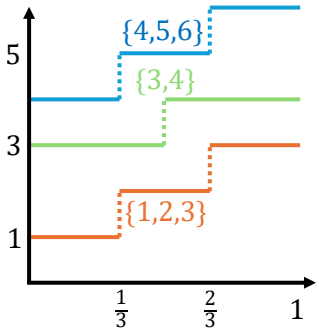

Figure 1: The quantile function of three different multisets

Sliced Wasserstein distance The *sliced Wasserstein distance*, proposed by Bonneel et al. (2015) as a surrogate for the Wasserstein distance, exploits the efficient computation of the latter for $d = 1$ to define a more computationally tractable distance for $d > 1$. It is defined as the average Wasserstein distance between all 1-dimensional projections (or *slices*) of the two input distributions. For a formal definition, we first define the projection of a distribution.

Definition. Let $\mu = \sum_{i=1}^N w_i \delta_{\boldsymbol{x}^{(i)}} \in \mathcal{P}_{\leq N}\big(\mathbb{R}^d\big)$. The projection of μ in the direction $\boldsymbol{v} \in \mathbb{R}^d$, denoted by $\boldsymbol{v}^T\mu$, is the one-dimensional distribution in $\mathcal{P}_{\leq N}(\mathbb{R})$ defined by

$$\boldsymbol{v}^T\mu \coloneqq \sum_{i=1}^N w_i \delta_{\boldsymbol{v}^T\boldsymbol{x}^{(i)}}.$$

Using the above definition, the sliced Wasserstein distance between $\mu, \tilde{\mu} \in \mathcal{P}_{\leq N}\big(\mathbb{R}^d\big)$ is defined by

$$\mathcal{SW}(\mu, \tilde{\mu}) \coloneqq \sqrt{\mathbb{E}_{\boldsymbol{v}}[\mathcal{W}^2(\boldsymbol{v}^T\mu, \boldsymbol{v}^T\tilde{\mu})]}, \tag{5}$$

where $\mathcal{W}^2(\,\cdot\,,\,\cdot\,)$ is the squared 2-Wasserstein distance, and the expectation $\mathbb{E}_{\boldsymbol{v}}[\,\cdot\,]$ is over the direction vector $\boldsymbol{v} \sim \mathrm{Uniform}\big(\mathbb{S}^{d-1}\big)$, i.e. distributed uniformly over the unit sphere in $\mathbb{R}^d$.

2.2 Existing embedding methods

We now return to our main goal of constructing an embedding $E : \mathcal{P}_{\leq N}\big(\mathbb{R}^d\big) \to \mathbb{R}^m$. In this subsection, we discuss existing embedding methods and some straightforward ideas to extend them.

We first observe that, on the collection of multisets over $\mathbb{R}$ of fixed size n, it follows from (3) that the map $\{x_1, \ldots, x_n\} \mapsto \frac{1}{\sqrt{n}} \cdot \text{sort}(x_1, \ldots, x_n)$ is an isometry, i.e. (1) holds with $c = C = 1$.

To extend this idea to multisets in $\mathcal{S}_{\leq N}(\mathbb{R})$, a naive approach would be to represent each multiset in $\mathcal{S}_{\leq N}(\mathbb{R})$ by a multiset of size $\hat{N}$, with $\hat{N}$ being the *least common multiple* (LCM) of $\{1, 2, \ldots, N\}$. For example, for $N = 3$, $\text{LCM}(\{1, 2, 3\}) = 6$, and thus multisets in $\mathcal{S}_{\leq N}(\mathbb{R})$ of sizes 1 $\{a\}$, 2 $\{a, b\}$, and 3 $\{a, b, c\}$, would be represented, respectively, by $\{a, a, a, a, a, a\}$, $\{a, a, a, b, b, b\}$ and $\{a, a, b, b, c, c\}$. At this point, a sorting approach can be applied. However, as N increases, this method quickly becomes infeasible, both in terms of computation time as well as memory, since $\hat{N}$ grows exponentially in N. Moreover, this method cannot handle arbitrary distributions in $\mathcal{P}_{\leq N}(\mathbb{R})$, whose weights may be irrational.

One possible approach to embed general distributions $\mu \in \mathcal{P}_{\leq N}(\mathbb{R})$ is to sample $Q_\mu(t)$ at m points $t_1, \ldots, t_m \in [0, 1]$ equispaced or drawn uniformly at random. While this approach would indeed approximately preserve the Wasserstein distance, as follows from (4), it is easy to show that for any finite number of samples m, this embedding is not injective on $\mathcal{P}_{\leq N}(\mathbb{R})$. Moreover, it is discontinuous with respect to the sampling points t_k and input probabilities w_i, and thus not amenable to gradient-based learning methods. Our method, described in the next section, resolves these issues by sampling the quantile function in the frequency domain rather than in the t-domain.

When considering $\mathcal{P}_{\leq N}\left(\mathbb{R}^d\right)$ with $d > 1$, one natural idea is to take m one-dimensional projections of the input distribution, and then embed each of the projections using one of the methods described above for $\mathcal{P}_{\leq N}(\mathbb{R})$. In the case of multisets of fixed cardinality n, this corresponds to the mapping

$$\{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n\} \mapsto \frac{1}{\sqrt{n}} \cdot \text{rowsort}\Big(\left[\boldsymbol{v}_k^T \boldsymbol{x}_i\right]_{k \in [m], i \in [n]}\Big).$$

This idea was discussed in (Balan et al., 2022; Zhang et al., 2019; Dym & Gortler, 2024; Balan & Tsoukanis, 2023b). In expectation over the directions $\boldsymbol{v}_k$, this method indeed gives a good approximation of the sliced Wasserstein distance. The relationship to the d-dimensional Wasserstein distance is *a priori* less clear. Balan & Tsoukanis (2023a) showed that for m that is exponential in n, this mapping bi-Lipschitz for almost any choice of $\boldsymbol{v}_1, \ldots, \boldsymbol{v}_m$. Later, Dym & Gortler (2024) showed that $m = 2nd + 1$ is sufficient. In this paper, we combine this idea of using linear projections, with our idea of Fourier sampling of quantile functions, to construct an embedding defined on the whole of $\mathcal{P}_{\leq N}\left(\mathbb{R}^d\right)$ while maintaining theoretical guarantees and practical efficiency.

In a related line of work, Kolouri et al. (2015); Naderializadeh et al. (2021); Lu et al. (2024) developed a method that embeds continuous distributions into an infinite-dimensional Hilbert space isometrically with respect to the sliced Wasserstein distance. However, in practice, a finite-dimensional discretization is used, which breaks the method's injectivity, as we show empirically in Section 5. In contrast, our method is provably injective with a finite and near-optimal embedding dimension of $\approx 2Nd$.

Lastly, Haviv et al. (2024) recently proposed a transformer-based architecture called Wasserstein Wormhole, designed to compute Euclidean embeddings for multisets and distributions. This architecture can be trained to approximate the Wasserstein distance. However, the fact that it requires training is a key limitation, as it does not guarantee Wasserstein-preserving embeddings. This concern becomes particularly significant when generalizing to out-of-distribution samples.

3 Proposed Method

Our method for embedding a distribution μ essentially consists of computing random slices $\boldsymbol{v}^T\mu$ and, for each slice, taking a single random sample of its quantile function $Q_{\boldsymbol{v}^T\mu}(t)$. However, instead of sampling the function directly, we sample its *cosine transform*—a variant of the Fourier transform. Since the Fourier transform is a linear isometry, integrating the squared difference of these samples for two distributions μ, $\tilde{\mu}$ yields the squared sliced Wasserstein distance $\mathcal{SW}^2(\mu, \tilde{\mu})$, as we shall show next. We will also show that this sampling guarantees injectivity, unlike direct sampling of $Q_{\boldsymbol{v}^T\mu}(t)$. Lastly, since $Q_{\boldsymbol{v}^T\mu}(t)$ is a compactly supported step function, its Fourier transform is smooth with respect to the frequencies, and thus so is our embedding. We now discuss this in detail.

Definition 3.1. Given a *projection vector* $\boldsymbol{v} \in \mathbb{S}^{d-1}$ and a number $\xi \geq 0$ denoting a frequency, we define the *one-sample embedding* $E^{\text{FSW}}(\,\cdot\,; \boldsymbol{v}, \xi) : \mathcal{P}_{\leq N}(\mathbb{R}^d) \to \mathbb{R}$ by

$$E^{\text{FSW}}(\mu; \boldsymbol{v}, \xi) \coloneqq 2(1+\xi) \int_0^1 Q_{\boldsymbol{v}^T\mu}(t) \cos(2\pi\xi t) dt, \tag{6}$$

which is the *cosine transform* of $Q_{\boldsymbol{v}^T\mu}(t)$, sampled at frequency ξ and multiplied by $1+\xi$; see Appendix B.1 for further discussion. Details on the practical computation of E^{FSW} are in Appendix A.2.

Next, we define a probability distribution $\mathcal{D}_\xi$ for the frequency ξ, given by the PDF

$$f_\xi(\xi) \coloneqq (1+\xi)^{-2}, \qquad \xi \geq 0.$$

We now show that E^{FSW} preserves the sliced Wasserstein distance in expectation over $\boldsymbol{v}, \xi$.

Theorem 3.2. (Proof in Appendix B.2) *Let $\mu, \tilde{\mu} \in \mathcal{P}_{\leq N}(B_R)$, with $B_R \subset \mathbb{R}^d$ being the closed ℓ_2-norm ball of radius R, centered at zero. Let $\boldsymbol{v} \sim \text{Uniform}(\mathbb{S}^{d-1})$, $\xi \sim \mathcal{D}_\xi$. Then*

$$\mathbb{E}_{\boldsymbol{v},\xi}\left[\left|E^{\text{FSW}}(\mu; \boldsymbol{v}, \xi) - E^{\text{FSW}}(\tilde{\mu}; \boldsymbol{v}, \xi)\right|^2\right] = \mathcal{SW}^2(\mu, \tilde{\mu}), \tag{7}$$

$$\text{STD}_{\boldsymbol{v},\xi}\left[\left|E^{\text{FSW}}(\mu; \boldsymbol{v}, \xi) - E^{\text{FSW}}(\tilde{\mu}; \boldsymbol{v}, \xi)\right|^2\right] \leq 4\sqrt{10}R^2. \tag{8}$$

This result can be further stabilized by taking multiple samples. To this end, we define the *Fourier Sliced Wasserstein (FSW) embedding* $E_m^{\text{FSW}} : \mathcal{P}_{\leq N}(\mathbb{R}^d) \to \mathbb{R}^m$ by

$$E_m^{\text{FSW}}(\mu) \coloneqq \left(E^{\text{FSW}}\left(\mu; \boldsymbol{v}^{(1)}, \xi^{(1)}\right), \ldots, E^{\text{FSW}}\left(\mu; \boldsymbol{v}^{(m)}, \xi^{(m)}\right)\right), \tag{9}$$

with $\left(\boldsymbol{v}^{(k)}, \xi^{(k)}\right)_{k=1}^m$ drawn randomly i.i.d. from $\text{Uniform}(\mathbb{S}^{d-1}) \times \mathcal{D}_\xi$. Consequently, we have:

Corollary 3.3. *Under the assumptions of Theorem 3.2,*

$$\mathbb{E}_{\boldsymbol{v},\xi}\left[\tfrac{1}{m}\left\|E_m^{\text{FSW}}(\mu) - E_m^{\text{FSW}}(\tilde{\mu})\right\|^2\right] = \mathcal{SW}^2(\mu, \tilde{\mu}), \tag{10}$$

$$\text{STD}_{\boldsymbol{v},\xi}\left[\tfrac{1}{m}\left\|E_m^{\text{FSW}}(\mu) - E_m^{\text{FSW}}(\tilde{\mu})\right\|^2\right] \leq 4\sqrt{10}\frac{R^2}{\sqrt{m}}. \tag{11}$$

Notably, the error bound in (11) is independent of the number of points N and the dimension d. Thus, the estimation error does not suffer from the curse of dimensionality. By taking a sufficiently high embedding dimension, one can embed distributions of arbitrarily high dimension and with arbitrary (and possibly infinite) support cardinality while ensuring a bounded standard estimation error, as long as the distributions are supported within a fixed ball of radius R.

4 Theoretical results

In the previous section, we showed that our embedding approximately preserves the sliced Wasserstein distance in a probabilistic sense, with diminishing estimation error as the embedding dimension grows. Here we show that with a *finite* dimension, our embedding is guaranteed injectivity and bi-Lipschitzness, as outlined in the Main results part of Section 1.

First, we show that with a sufficiently high dimension m, our embedding is injective.

Theorem 4.1. (Proof in Appendix B.3) *Let $E_m^{\text{FSW}} : \mathcal{P}_{\leq N}(\mathbb{R}^d) \to \mathbb{R}^m$ be as in (9), with $\left(\boldsymbol{v}^{(k)}, \xi^{(k)}\right)_{k=1}^m$ drawn i.i.d. from $\text{Uniform}(\mathbb{S}^{d-1}) \times \mathcal{D}_\xi$. Then:*

1. *If $m \geq 2Nd + 1$, then with probability 1, E_m^{FSW} is injective on $\mathcal{S}_{\leq N}(\mathbb{R}^d)$.*
2. *If $m \geq 2Nd + 2N - 1$, then with probability 1, E_m^{FSW} is injective on $\mathcal{P}_{\leq N}(\mathbb{R}^d)$.*

These bounds are optimal essentially up to a multiplicative factor of 2, since an embedding dimension m smaller than Nd precludes injectivity for *any* continuous embedding (Amir et al., 2023).

Next, we show that in the case of $\mathcal{S}_{\leq N}(\mathbb{R}^d)$, the injectivity of E_m^{FSW} implies that it is in fact bi-Lipschitz. Our proof relies on the fact that E_m^{FSW} is piecewise linear and homogeneous in the input points, in a sense we shall now define. By a slight abuse of notation, we denote the distribution parametrized by the points $\boldsymbol{X} = \left(\boldsymbol{x}^{(1)}, \ldots, \boldsymbol{x}^{(N)}\right)$ and weights $\boldsymbol{w} = (w_1, \ldots, w_N)$ as $(\boldsymbol{X}, \boldsymbol{w})$.

Definition. Let $E : \mathcal{D} \to \mathbb{R}^m$ with $\mathcal{D} = \mathcal{P}_{\leq N}(\mathbb{R}^d)$ or $\mathcal{S}_{\leq N}(\mathbb{R}^d)$. We say that E is *positively homogeneous* if for any $\alpha \geq 0$ and any distribution $(\boldsymbol{X}, \boldsymbol{w}) \in \mathcal{D}$, $E(\alpha \boldsymbol{X}, \boldsymbol{w}) = \alpha E(\boldsymbol{X}, \boldsymbol{w})$.

The following theorem shows that any embedding that is injective, positively homogeneous and piecewise linear, is bi-Lipschitz when restricted to distributions with fixed weights.

Theorem 4.2. (Proof in Appendix B.3) *Let $E : \mathcal{P}_{\leq N}(\mathbb{R}^d) \to \mathbb{R}^m$ be injective and positively homogeneous. Let $\boldsymbol{w}, \tilde{\boldsymbol{w}} \in \Delta^N$ be fixed, and suppose that the functions $E(\boldsymbol{X}, \boldsymbol{w})$ and $E(\boldsymbol{X}, \tilde{\boldsymbol{w}})$ are piecewise linear in $\boldsymbol{X}$. Then there exist $c, C > 0$ such that for all $\boldsymbol{X}, \tilde{\boldsymbol{X}} \in \mathbb{R}^{d \times N}$ and $p \in [1, \infty]$,*

$$c \cdot \mathcal{W}_p\Big((\boldsymbol{X}, \boldsymbol{w}), \Big(\tilde{\boldsymbol{X}}, \tilde{\boldsymbol{w}}\Big)\Big) \leq \Big\| E(\boldsymbol{X}, \boldsymbol{w}) - E\Big(\tilde{\boldsymbol{X}}, \tilde{\boldsymbol{w}}\Big) \Big\| \leq C \cdot \mathcal{W}_p\Big((\boldsymbol{X}, \boldsymbol{w}), \Big(\tilde{\boldsymbol{X}}, \tilde{\boldsymbol{w}}\Big)\Big). \quad (12)$$

The assumption of fixed weights can be straightforwardly relaxed to allow for weights that come from a finite set. Using this observation, we now show that E_m^{FSW} is bi-Lipschitz on multisets.

Corollary 4.3. *Let E_m^{FSW} be as in (9) with $m \geq 2Nd + 1$. Then with probability 1, E_m^{FSW} is bi-Lipschitz on $\mathcal{S}_{\leq N}(\mathbb{R}^d)$.*

Proof. Any $\mu \in \mathcal{S}_{\leq N}(\mathbb{R}^d)$ can be represented by a parameter of the form $\big(\boldsymbol{X}, \boldsymbol{w}^{(n)}\big)$, where

$$\boldsymbol{w}^{(n)} = \Big(\overbrace{\tfrac{1}{n}, \ldots, \tfrac{1}{n}}^{n}, \overbrace{0, \ldots, 0}^{N-n}\Big), \qquad 1 \leq n \leq N. \quad (13)$$

For $n, r \in [N]$, let c_{nr}, C_{nr} be the Lipschitz constants c, C of (12) for E_m^{FSW} with the weight vectors $\boldsymbol{w} = \boldsymbol{w}^{(n)}$, $\tilde{\boldsymbol{w}} = \boldsymbol{w}^{(r)}$. Then it is easy to show that E_m^{FSW} is bi-Lipschitz on $\mathcal{S}_{\leq N}(\mathbb{R}^d)$ with the constants $c = \min_{n,r \in [N]} c_{nr} > 0$ and $C = \max_{n,r \in [N]} C_{nr} < \infty$. □

Next, we explore whether it is possible to further improve by finding an embedding that is bi-Lipschitz on the whole of $\mathcal{P}_{\leq N}(\mathbb{R}^d)$. For the broader class of distributions $\bigcup_{N \in \mathbb{N}} \mathcal{P}_{\leq N}(\mathbb{R}^d)$, Naor & Schechtman (2007) proved that no bi-Lipschitz embedding exists into the space $L^1([0, 1])$, and thus not into any finite-dimensional space. One may ask whether this can be remedied by restricting the distributions to a bounded number of support points that come from a fixed compact domain, namely, $\mathcal{P}_{\leq N}(\Omega)$ with a compact $\Omega \subset \mathbb{R}^d$. The following theorem shows that even this restricted class cannot be embedded in a bi-Lipschitz manner into a finite-dimensional Euclidean space.

Theorem 4.4. (Proof in Appendix B.3) *Let $E : \mathcal{P}_{\leq N}(\Omega) \to \mathbb{R}^m$, where $N \geq 2$ and $\Omega \subseteq \mathbb{R}^d$ has a nonempty interior. Then for all $p \in [1, \infty]$, E is not bi-Lipschitz on $\mathcal{P}_{\leq N}(\Omega)$ with respect to $\mathcal{W}_p$.*

5 Numerical experiments

In this section, we demonstrate how the theoretical strengths of our method translate into practical advantages. Specifically, we show that our method yields embeddings with superior distance preservation and enhances performance in practical learning tasks.

Comparison with PSWE This experiment compares our method with PSWE (Naderializadeh et al., 2021)—an embedding method developed with the same purpose of preserving the sliced Wasserstein distance. As mentioned in Section 2.2, PSWE embeds continuous distributions into an infinite-dimensional Hilbert space, and does so isometrically with respect to the sliced Wasserstein distance. Finite multisets and distributions are treated by this method as discretizations of continuous distributions, which causes different finite distributions to be identified with the same continuous distribution, consequently leading to a loss of injectivity.

To demonstrate this, we constructed a pair of multisets $\boldsymbol{X}_1, \boldsymbol{X}_2 \subset \mathbb{R}^3$, of sizes 5 and 200, respectively, that PSWE identifies as the same continuous distribution, and thus fails to distinguish; see Appendix C.1 for details. Table 1 reports the normalized Euclidean distances, $\frac{1}{\sqrt{m}} \|E(\boldsymbol{X}_1) - E(\boldsymbol{X}_2)\|$, where $E : \mathcal{S}_{\leq N}(\mathbb{R}^3) \to \mathbb{R}^m$ is either FSW or PSWE, for various embedding dimensions m. As shown in the table, PSWE fails to distinguish the two multisets regardless of m, whereas FSW succeeds even when $m = 1$. Moreover, as m increases, the distances produced by FSW converge to the sliced Wasserstein distance $\mathcal{SW}(\boldsymbol{X}_1, \boldsymbol{X}_2)$, as promised by Corollary 3.3.

Table 1: Embedding distance vs. embedding dimension

Method	Embedding dimension				
	1	10	100	1000	10000
FSW	0.125	0.0326	0.0639	0.0733	0.0737
PSWE	0.0	5.48e-8	1.47e-7	1.3e-7	1.44e-7

Embedding distances between a pair of multisets $\boldsymbol{X}_1, \boldsymbol{X}_2 \subset \mathbb{R}^3$ distinguished by FSW but not by PSWE. Here $|\boldsymbol{X}_1| = 5$, $|\boldsymbol{X}_2| = 200$, and $\mathcal{SW}(\boldsymbol{X}_1, \boldsymbol{X}_2) = 0.0734$.

Learning to approximate the Wasserstein distance One possible approach to overcome the high computation time of the Wasserstein distance for $d > 1$ is to try to estimate it using a neural network, trained on pairs of point-clouds for which the distance is known. This approach was used in previous works (Chen & Wang, 2023; Kawano et al., 2020), which proposed architectures designed to approximate functions $F : \mathcal{S}_{\leq N}(\mathbb{R}^d) \times \mathcal{S}_{\leq N}(\mathbb{R}^d) \to \mathbb{R}$, such as the Wasserstein distance function. These methods handle multisets using the traditional approach of sum- or average-pooling. Since our embedding is bi-Lipschitz with respect to the Wasserstein distance, it is *a priori* likely to be a more effective building block for architectures designed to learn it.

For this task, we used the following architecture: First, an FSW embedding $E_1 : \mathcal{P}_{\leq N}(\mathbb{R}^d) \to \mathbb{R}^{m_1}$ is applied to each of the two input distributions μ, $\tilde{\mu}$. Then, a second FSW embedding $E_2 : \mathcal{S}_{\leq 2}(\mathbb{R}^{m_1}) \to \mathbb{R}^{m_2}$ is applied to the multiset $\{E_1(\mu), E_1(\tilde{\mu})\}$. The output of E_2 is then fed to an MLP $\Phi : \mathbb{R}^{m_2} \to \mathbb{R}_+$; see Appendix C.2 for dimensions and technical details. Our full architecture is described by the formula

$$F(\mu, \tilde{\mu}) \coloneqq \Phi(E_2(\{E_1(\mu), E_1(\tilde{\mu})\})).$$

This formulation ensures that F is symmetric with respect to swapping μ and $\tilde{\mu}$. In addition, we used leaky-ReLU activations and no biases in Φ, which renders F scale-equivariant by design, i.e.

$$F\big((\alpha\boldsymbol{X}, \boldsymbol{w}), (\alpha\tilde{\boldsymbol{X}}, \tilde{\boldsymbol{w}})\big) = \alpha F\big((\boldsymbol{X}, \boldsymbol{w}), (\tilde{\boldsymbol{X}}, \tilde{\boldsymbol{w}})\big) \quad \forall \alpha > 0,$$

as is the Wasserstein distance function, which F is designed to approximate.

The experimental setting was replicated from (Chen & Wang, 2023). The objective is to approximate the 1-Wasserstein distance $\mathcal{W}_1$. The following evaluation datasets, kindly provided to us by the authors, were used: Three synthetic datasets `noisy-sphere-3`, `noisy-sphere-6` and `uniform`, with random point-clouds in $\mathbb{R}^3$, $\mathbb{R}^6$ and $\mathbb{R}^2$; two real datasets `ModelNet-small` and `ModelNet-large`, with 3D point-clouds sampled from ModelNet40 objects (Wu et al., 2015); and the gene-expression dataset `RNAseq` (Yao et al., 2021), with multisets in $\mathbb{R}^{2000}$.

We compared our architecture to the following methods: (a) $\mathcal{N}_{\text{SDeepSets}}$, a DeepSets-like architecture trained to compute $\mathcal{W}_1$-preserving embeddings, and $\mathcal{N}_{\text{ProductNet}}$, which further processes the sum of the two embeddings through an MLP (Chen & Wang, 2023); (b) a Siamese autoencoder called Wasserstein Point-Cloud Embedding network (WPCE) (Kawano et al., 2020); and (c) the Sinkhorn approximation algorithm (Cuturi, 2013). We also evaluated the PSWE embedding by employing it in our architecture instead of E_1.

Table 2: 1-Wasserstein approximation: Relative error

Dataset	d	set size	Ours	PSWE	$\mathcal{N}_{\text{ProductNet}}$	WPCE	$\mathcal{N}_{\text{SDeepSets}}$	Sinkhorn
noisy-sphere-3	3	100–299	**1.4 %**	2.2 %	4.6 %	34.1 %	36.2 %	18.7 %
noisy-sphere-6	6	100–299	**1.3 %**	1.4 %	1.5 %	26.9 %	29.1 %	13.7 %
uniform	2	256	2.4 %	**2.1 %**	9.7 %	12.0 %	12.3 %	7.3 %
ModelNet-small	3	20–199	**2.9 %**	5.7 %	8.4 %	7.7 %	10.5 %	10.1 %
ModelNet-large	3	2047	2.6 %	**2.4 %**	14.0 %	15.9 %	16.6 %	14.8 %
RNAseq	2000	20–199	**1.1 %**	1.2 %	1.2 %	47.7 %	48.2 %	4.0 %

Mean relative error in approximating the 1-Wasserstein distance between point sets.

As seen in Table 2, our architecture achieves the best accuracy on most evaluation datasets. Training times are in Table 3. Further details on this experiment appear in Appendix C.2.

Table 3: 1-Wasserstein approximation: Training time

Dataset	Ours	PSWE	$\mathcal{N}_{\text{ProductNet}}$	WPCE	$\mathcal{N}_{\text{SDeepSets}}$
noisy-sphere-3	2.2 min	33 min	6 min	1 h 46 min	9 min
noisy-sphere-6	4 min	1 h	12 min	4 h 6 min	1 h 38 min
uniform	3 min	51 min	7 min	3 h 36 min	1 h 27 min
ModelNet-small	3 min	48 min	7 min	1 h 23 min	12 min
ModelNet-large	14.2 min	1 h 19 min	8 min	3 h 5 min	40 min
RNAseq	4 min	50 min	15 min	14 h 26 min	3 h 1 min

Training times for the different architectures.

Robustness to parameter reduction Next, we evaluate our embedding on the `ModelNet40` object classification benchmark (Wu et al., 2015). This dataset consists of 3D point clouds representing objects coming from 40 different classes.

To demonstrate the improved representational power of our method, we used a simple architecture of the form $F(\mu) = \Phi(E(\mu))$, where $E : \mathcal{S}_{\leq N}(\mathbb{R}^d) \to \mathbb{R}^m$ is the FSW embedding, $\Phi : \mathbb{R}^m \to \mathbb{R}^C$ is an MLP, and $C = 40$ is the number of classes. We compared this architecture with PointNet (Qi et al., 2017a), which applies an MLP elementwise, followed by max pooling and another MLP. We evaluated different sizes of both architectures to assess their robustness to parameter reduction.

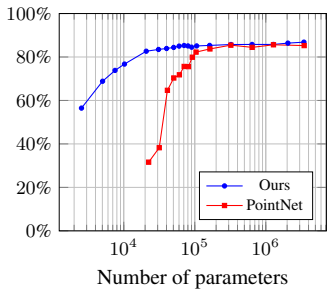

Figure 2: ModelNet40 classification accuracy

With 300 k parameters or more, the two methods perform comparably: our model achieves 85.7 %–86.8 % accuracy and PointNet[4] achieves 84.4 %–85.6 %. However, our model is considerably more robust when reducing the number of parameters, as shown in Figure 2. For example, with 30 k parameters, our model achieves 83.47 % accuracy, whereas PointNet achieves 38.25 %.

We note that simple methods such as PointNet and ours do not achieve the best results on ModelNet40. More complex methods achieve superior results by applying PointNet-based embeddings to local neighborhoods of each point. For example, PointNet++ (Qi et al., 2017b) achieved 90.7 % accuracy, and (Mohammadi et al., 2021) achieved an impressive 95.4 %. Based on our preliminary results, we believe that combining our embedding with such methods, by employing it in their multiset aggregation steps, may enhance robustness to parameter reduction. This property can be crucial in practical applications.

6 Conclusion

In this paper, we introduced the first injective embedding for multisets and measures that is provably bi-Lipschitz on multisets. Our embedding is computationally efficient, and both theoretical and empirical results suggest that it better preserves the original geometry of the data, consequently leading to improved performance in practical learning tasks.

Our embedding has already been employed in subsequent work (Sverdlov et al., 2024) to develop the first graph neural network with bi-Lipschitz Weisfeiler–Leman (WL) separation power.

Future research directions include further exploring the FSW embedding as a building block for additional learning tasks over non-Euclidean data, and extending the concepts introduced here to broader notions of distance, such as partial and unbalanced optimal transport.

Reproducibility Statement All experiments in this paper are fully reproducible. The code for training and evaluation, along with the datasets and actual numerical results presented in this paper, are available at the URL `https://github.com/tal-amir/FSW-embedding`. While

[4]These results were obtained using a standard PyTorch implementation of PointNet (Xia, 2019). The original result of PointNet's TensorFlow implementation is 89.2 % accuracy.

we did not use fixed random seeds for our experiments, the results are consistent across multiple runs. For further technical details regarding the experimental setup and parameters, please refer to Appendix C.

Acknowledgements TA and ND are partially funded by ISF grant 272/23. We thank Samantha Chen (Chen & Wang, 2023) for her help in reproducing her experiments.

REFERENCES

Miika Aittala and Fredo Durand. Burst image deblurring using permutation invariant convolutional neural networks. In *The European Conference on Computer Vision (ECCV)*, September 2018.

Jason Altschuler, Jonathan Niles-Weed, and Philippe Rigollet. Near-linear time approximation algorithms for optimal transport via sinkhorn iteration. *Advances in neural information processing systems*, 30, 2017.

Tal Amir, Steven Gortler, Ilai Avni, Ravina Ravina, and Nadav Dym. Neural injective functions for multisets, measures and graphs via a finite witness theorem. *Advances in Neural Information Processing Systems*, 36:42516–42551, 2023.

Radu Balan and Efstratios Tsoukanis. G-invariant representations using coorbits: Bi-lipschitz properties, 2023a.

Radu Balan and Efstratos Tsoukanis. Relationships between the phase retrieval problem and permutation invariant embeddings. In *2023 International Conference on Sampling Theory and Applications (SampTA)*, pp. 1–6. IEEE, 2023b.

Radu Balan, Naveed Haghani, and Maneesh Singh. Permutation invariant representations with applications to graph deep learning. *arXiv preprint arXiv:2203.07546*, 2022.

Erhan Bayraktar and Gaoyue Guo. Strong equivalence between metrics of wasserstein type, 2021.

Mary L Boas. *Mathematical methods in the physical sciences*. John Wiley & Sons, 2006.

Nicolas Bonneel, Julien Rabin, Gabriel Peyré, and Hanspeter Pfister. Sliced and radon wasserstein barycenters of measures. *Journal of Mathematical Imaging and Vision*, 51:22–45, 2015.

Jameson Cahill, Joseph W Iverson, Dustin G Mixon, and Daniel Packer. Group-invariant max filtering. *arXiv preprint arXiv:2205.14039*, 2022.

Jameson Cahill, Joseph W. Iverson, and Dustin G. Mixon. Towards a bilipschitz invariant theory, 2024.

Samantha Chen and Yusu Wang. Neural approximation of wasserstein distance via a universal architecture for symmetric and factorwise group invariant functions. *Advances in Neural Information Processing Systems*, 36:9506–9517, 2023.

Marco Cuturi. Sinkhorn distances: Lightspeed computation of optimal transport. In C.J. Burges, L. Bottou, M. Welling, Z. Ghahramani, and K.Q. Weinberger (eds.), *Advances in Neural Information Processing Systems*, volume 26. Curran Associates, Inc., 2013. URL https://proceedings.neurips.cc/paper_files/paper/2013/file/af21d0c97db2e27e13572cbf59eb343d-Paper.pdf.

George Cybenko. Approximation by superpositions of a sigmoidal function. *Mathematics of control, signals and systems*, 2(4):303–314, 1989.

Pavel Dvurechensky, Alexander Gasnikov, and Alexey Kroshnin. Computational optimal transport: Complexity by accelerated gradient descent is better than by sinkhorn's algorithm. In *International conference on machine learning*, pp. 1367–1376. PMLR, 2018.

Nadav Dym and Steven J. Gortler. Low-dimensional invariant embeddings for universal geometric learning. 2024. doi: 10.1007/s10208-024-09641-2. Publisher Copyright: © The Author(s) 2024.

Rémi Flamary, Nicolas Courty, Alexandre Gramfort, Mokhtar Z. Alaya, Aurélie Boisbunon, Stanislas Chambon, Laetitia Chapel, Adrien Corenflos, Kilian Fatras, Nemo Fournier, Léo Gautheron, Nathalie T.H. Gayraud, Hicham Janati, Alain Rakotomamonjy, Ievgen Redko, Antoine Rolet, Antony Schutz, Vivien Seguy, Danica J. Sutherland, Romain Tavenard, Alexander Tong, and Titouan Vayer. Pot: Python optimal transport. *Journal of Machine Learning Research*, 22(78): 1–8, 2021. URL http://jmlr.org/papers/v22/20-451.html.

Justin Gilmer, Samuel S. Schoenholz, Patrick F. Riley, Oriol Vinyals, and George E. Dahl. Neural message passing for quantum chemistry. In Doina Precup and Yee Whye Teh (eds.), *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pp. 1263–1272. PMLR, 06–11 Aug 2017. URL https://proceedings.mlr.press/v70/gilmer17a.html.

Branko Grünbaum. *Convex Polytopes*, volume 221. Springer Science & Business Media, 2003.

Doron Haviv, Russell Zhang Kunes, Thomas Dougherty, Cassandra Burdziak, Tal Nawy, Anna Gilbert, and Dana Pe'Er. Wasserstein wormhole: Scalable optimal transport distance with transformers. *ArXiv*, 2024.

Piotr Indyk and Nitin Thaper. Fast image retrieval via embeddings. *ICCV '03: Proceedings of the 3rd International Workshop on Statistical and Computational Theories of Vision*, 2003.

Frank Jones. *Lebesgue integration on Euclidean space*. Jones & Bartlett Learning, 2001.

Keisuke Kawano, Satoshi Koide, and Takuro Kutsuna. Learning wasserstein isometric embedding for point clouds. In *2020 International Conference on 3D Vision (3DV)*, pp. 473–482. IEEE, 2020.

Thomas N Kipf and Max Welling. Semi-supervised classification with graph convolutional networks. In *International Conference on Learning Representations*, 2016.

Soheil Kolouri, Se Rim Park, and Gustavo K Rohde. The radon cumulative distribution transform and its application to image classification. *IEEE transactions on image processing*, 25(2):920–934, 2015.

Juho Lee, Yoonho Lee, Jungtaek Kim, Adam Kosiorek, Seungjin Choi, and Yee Whye Teh. Set transformer: A framework for attention-based permutation-invariant neural networks. In *International conference on machine learning*, pp. 3744–3753. PMLR, 2019.

Yuzhe Lu, Xinran Liu, Andrea Soltoggio, and Soheil Kolouri. Slosh: Set locality sensitive hashing via sliced-wasserstein embeddings. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pp. 2566–2576, 2024.

Haggai Maron, Heli Ben-Hamu, Hadar Serviansky, and Yaron Lipman. Provably powerful graph networks. *Advances in neural information processing systems*, 32, 2019.

Haggai Maron, Or Litany, Gal Chechik, and Ethan Fetaya. On learning sets of symmetric elements. In Hal Daumé III and Aarti Singh (eds.), *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pp. 6734–6744. PMLR, 13–18 Jul 2020. URL https://proceedings.mlr.press/v119/maron20a.html.

Seyed Saber Mohammadi, Yiming Wang, and Alessio Del Bue. Pointview-gcn: 3d shape classification with multi-view point clouds. In *2021 IEEE International Conference on Image Processing (ICIP)*, pp. 3103–3107. IEEE, 2021.

Navid Naderializadeh, Joseph F Comer, Reed Andrews, Heiko Hoffmann, and Soheil Kolouri. Pooling by sliced-wasserstein embedding. *Advances in Neural Information Processing Systems*, 34: 3389–3400, 2021.

Assaf Naor and Gideon Schechtman. Planar earthmover is not in l_1. *SIAM Journal on Computing*, 37(3):804–826, 2007.

James Orlin. A faster strongly polynomial minimum cost flow algorithm. In *Proceedings of the Twentieth annual ACM symposium on Theory of Computing*, pp. 377–387, 1988.

Sergey Pozdnyakov and Michele Ceriotti. Smooth, exact rotational symmetrization for deep learning on point clouds. In A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine (eds.), *Advances in Neural Information Processing Systems*, volume 36, pp. 79469–79501. Curran Associates, Inc., 2023. URL https://proceedings.neurips.cc/paper_files/paper/2023/file/fb4a7e3522363907b26a86cc5be627ac-Paper-Conference.pdf.

Charles R. Qi, Hao Su, Kaichun Mo, and Leonidas J. Guibas. Pointnet: Deep learning on point sets for 3d classification and segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, July 2017a.

Charles Ruizhongtai Qi, Li Yi, Hao Su, and Leonidas J Guibas. Pointnet++: Deep hierarchical feature learning on point sets in a metric space. *Advances in neural information processing systems*, 30, 2017b.

Yonatan Sverdlov, Yair Davidson, Nadav Dym, and Tal Amir. FSW-GNN: A bi-Lipschitz WL-equivalent graph neural network, 2024. URL https://arxiv.org/abs/2410.09118.

Puoya Tabaghi and Yusu Wang. Universal representation of permutation-invariant functions on vectors and tensors. In Claire Vernade and Daniel Hsu (eds.), *Proceedings of The 35th International Conference on Algorithmic Learning Theory*, volume 237 of *Proceedings of Machine Learning Research*, pp. 1134–1187. PMLR, 25–28 Feb 2024. URL https://proceedings.mlr.press/v237/tabaghi24a.html.

Petar Veličković, Guillem Cucurull, Arantxa Casanova, Adriana Romero, Pietro Liò, and Yoshua Bengio. Graph attention networks. In *International Conference on Learning Representations*, 2018.

Edward Wagstaff, Fabian Fuchs, Martin Engelcke, Ingmar Posner, and Michael A Osborne. On the limitations of representing functions on sets. In *International Conference on Machine Learning*, pp. 6487–6494. PMLR, 2019.

Edward Wagstaff, Fabian B Fuchs, Martin Engelcke, Michael A Osborne, and Ingmar Posner. Universal approximation of functions on sets. *Journal of Machine Learning Research*, 23(151):1–56, 2022.

Larry Wasserman. *All of statistics: a concise course in statistical inference*, volume 26. Springer, 2004.

Zhirong Wu, Shuran Song, Aditya Khosla, Fisher Yu, Linguang Zhang, Xiaoou Tang, and Jianxiong Xiao. 3d shapenets: A deep representation for volumetric shapes. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 1912–1920, 2015.

Fei Xia. Pointnet.pytorch. https://github.com/fxia22/pointnet.pytorch, 2019.

Keyulu Xu, Weihua Hu, Jure Leskovec, and Stefanie Jegelka. How powerful are graph neural networks? In *International Conference on Learning Representations*, 2018.

Zizhen Yao, Cindy TJ Van Velthoven, Thuc Nghi Nguyen, Jeff Goldy, Adriana E Sedeno-Cortes, Fahimeh Baftizadeh, Darren Bertagnolli, Tamara Casper, Megan Chiang, Kirsten Crichton, et al. A taxonomy of transcriptomic cell types across the isocortex and hippocampal formation. *Cell*, 184(12):3222–3241, 2021.

Manzil Zaheer, Satwik Kottur, Siamak Ravanbhakhsh, Barnabás Póczos, Ruslan Salakhutdinov, and Alexander J Smola. Deep sets. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, NIPS'17, pp. 3394–3404, Red Hook, NY, USA, 2017. Curran Associates Inc. ISBN 9781510860964.

Muhan Zhang, Zhicheng Cui, Marion Neumann, and Yixin Chen. An end-to-end deep learning architecture for graph classification. In Sheila A. McIlraith and Kilian Q. Weinberger (eds.), *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence, (AAAI-18), the 30th innovative Applications of Artificial Intelligence (IAAI-18), and the 8th AAAI Symposium on Educational Advances in Artificial Intelligence (EAAI-18), New Orleans, Louisiana, USA, February 2-7, 2018*, pp. 4438–4445. AAAI Press, 2018. doi: 10.1609/AAAI.V32I1.11782. URL https://doi.org/10.1609/aaai.v32i1.11782.

Yan Zhang, Jonathon Hare, and Adam Prügel-Bennett. FSPool: Learning set representations with featurewise sort pooling. 2019. URL https://arxiv.org/abs/1906.02795.

A Further Details on the FSW Embedding

A.1 Extension to measures with arbitrary total mass

We now discuss how to extend the definition of the FSW embedding to input measures that are not necessarily probability measures, that is, measures whose weights do not sum to 1. This extension is crucial for distinguishing multisets of different sizes when their element proportions are identical (as in the example in footnote 3, Page 4), as well as for handling the empty multiset as input.

The latter case is particularly important when using the embedding as a message aggregation function in graph neural networks, where isolated vertices aggregate messages over the empty multiset. Common graph architectures that require weight normalization typically address this issue by adding self-loops to the graph, thereby ensuring that no vertex has an empty neighborhood. The solution presented here does not require modifications to the graph structure.

Denote by $\mathcal{M}_{\leq N}(\Omega)$ the collection of all measures with at most N support points over a measurable domain $\Omega \subseteq \mathbb{R}^d$. Any measure $\mu \in \mathcal{M}_{\leq N}(\Omega)$ can be represented by points $\boldsymbol{x}^{(1)}, \ldots, \boldsymbol{x}^{(N)} \in \Omega$ and corresponding weights $w_1, \ldots, w_N \geq 0$ as a weighted sum of Dirac's delta functions

$$\mu = \sum_{i=1}^{N} w_i \delta_{\boldsymbol{x}^{(i)}}. \tag{14}$$

Definition. Let μ be as in (14). The *total mass* of μ is defined as

$$\mu(\Omega) = \sum_{i=1}^{N} w_i.$$

A first step towards extending E^{FSW} from $\mathcal{P}_{\leq N}(\Omega)$ to $\mathcal{M}_{\leq N}(\Omega)$ while maintaining injectivity is to simply add an extra output coordinate that encodes the total mass of the input measure. Define $\hat{E}^{\text{FSW}}_m : \mathcal{M}_{\leq N}(\Omega) \setminus \mathbf{0} \to \mathbb{R}^m$ for an input μ as in (14) by

$$\hat{E}^{\text{FSW}}_m(\mu) \coloneqq \left[\mu(\Omega), E^{\text{FSW}}_{m-1}(\hat{\mu})\right], \tag{15}$$

where $\hat{\mu}$ is the measure μ normalized to have a total mass of 1:

$$\hat{\mu} \coloneqq \frac{\mu}{\mu(\Omega)} = \sum_{i=1}^{N} \left(\frac{w_i}{\sum_{j=1}^{N} w_j} \right) \delta_{\boldsymbol{x}^{(i)}}. \tag{16}$$

It is easy to verify that with the above definition, by Theorem 4.1, if $m \geq 2Nd + 2N$ then $\hat{E}^{\text{FSW}}_m$ is injective[5] on $\mathcal{M}_{\leq N}\left(\mathbb{R}^d\right) \setminus \mathbf{0}$, i.e., the collection of all measures in $\mathcal{M}_{\leq N}\left(\mathbb{R}^d\right)$ excluding the zero measure $\mu = \mathbf{0}$, for which $\hat{\mu}$ is not defined. Moreover, it further follows from the theorem that $m \geq 2Nd + 2$ suffices to ensure that $\hat{E}^{\text{FSW}}_m$ distinguishes between pairs of input multisets of size $\leq N$ with different cardinalities, even if their element proportions are identical.

One limitation of the definition in (15) is that $\hat{E}^{\text{FSW}}_m$ is not defined at the zero measure $\mu = 0$ and has a jump discontinuity there. This issue can be remedied by padding input measures whose total mass is below a chosen threshold, assigning the complementary mass to the zero vector in $\mathbb{R}^d$.

Thus, we choose a threshold $\rho > 0$ and define the *regularized measure* μ_ρ for $\mu \in \mathcal{M}_{\leq N}\left(\mathbb{R}^d\right)$ by

$$\mu_\rho \coloneqq \begin{cases} \frac{\mu}{\mu(\Omega)} & \mu(\Omega) \geq \rho, \\ \left(1 - \frac{\mu(\Omega)}{\rho}\right)\delta_{\mathbf{0}} + \frac{\mu}{\rho} & \mu(\Omega) < \rho, \end{cases} \tag{17}$$

where $\delta_{\mathbf{0}}$ is Dirac's delta function at $0 \in \mathbb{R}^d$. We then redefine $\hat{E}^{\text{FSW}}_m$ to use μ_ρ:

$$\hat{E}^{\text{FSW}}_m(\mu) \coloneqq \left[\mu(\Omega),\ E^{\text{FSW}}_{m-1}(\mu_\rho)\right]. \tag{18}$$

[5] Strictly speaking, it is injective for *almost any choice* of the embedding parameters $\left(\boldsymbol{v}^{(k)}, \xi^{(k)}\right)_{k=1}^{m-1}$, as in the statement of Theorem 4.1. We ignore this here for brevity.

With the definition in (18), $\hat{E}_m^{\text{FSW}}$ is well-defined on the whole of $\mathcal{M}_{\leq N}(\mathbb{R}^d)$. Moreover, it is continuous and has finite directional derivatives everywhere, including the zero measure. As discussed above, if $m \geq 2Nd + 2N$, then $\hat{E}_m^{\text{FSW}}$ is injective, and for injectivity on multisets, $m \geq 2Nd + 2$ suffices.

Note that ρ does not have to be small. In fact, very small values of ρ lead to large gradients of $\hat{E}_m^{\text{FSW}}$ near the zero measure. In our experiments in (Sverdlov et al., 2024), we found that values in the range $0.1 \leq \rho \leq 1$ work well in practice.

Maintaining homogeneity In some settings it is desired to have an embedding that is positively homogeneous with respect to the input points[6] while still encoding the input multiset size. This is particularly useful for architectures designed to approximate multiset functions that are known to be homogeneous, while maintaining their ability to distinguish between input multisets of different sizes.

Note that this necessitates a slight compromise of the injectivity, as any such embedding inevitably maps all measures supported on the zero vector $0 \in \mathbb{R}^d$ to the zero vector $0 \in \mathbb{R}^m$, regardless of their total mass.

To achieve this homogeneity, a straightforward solution is to use a slight modification of (18) and multiply the first coordinate $\mu(\Omega)$ by the norm of the internal embedding $E_{m-1}^{\text{FSW}}(\mu_\rho)$:

$$\hat{E}_m^{\text{FSW}}(\mu) \coloneqq \left[\left\| E_{m-1}^{\text{FSW}}(\mu_\rho) \right\| \cdot \mu(\Omega),\ E_{m-1}^{\text{FSW}}(\mu_\rho) \right]. \tag{19}$$

With the formulation (19), the positive homogeneity of E_{m-1}^{FSW} implies that $\hat{E}_m^{\text{FSW}}$ is also positively homogeneous. Moreover, it can be verified that $\hat{E}_m^{\text{FSW}}$ with this formulation still distinguishes between any pair of distinct input measures, except for the case that both measures are supported at $0 \in \mathbb{R}^d$.

A.2 Practical computation

Here we present formulas that facilitate the practical computation of E^{FSW}, and then discuss its computational complexity.

We start by introducing notation that shall be used to express quantile functions of distributions in $\mathcal{P}_{\leq N}(\mathbb{R})$.

Definition A.1. For a vector $\boldsymbol{x} = (x_1, \ldots, x_N) \in \mathbb{R}^N$, the *order statistics* $x_{(1)}, \ldots, x_{(N)}$ are the coordinates of $\boldsymbol{x}$ sorted in increasing order: $x_{(1)} \leq \ldots \leq x_{(N)}$. We define the sorting permutation

$$\sigma(\boldsymbol{x}) = (\sigma_1(\boldsymbol{x}), \ldots, \sigma_N(\boldsymbol{x})) \in S_N$$

to be a permutation that satisfies $x_{\sigma_i(\boldsymbol{x})} = x_{(i)}$ for all $i \in [N]$, with ties broken arbitrarily.

We now show how $Q_\mu(t)$ can be expressed explicitly in terms of the order statistics of the support of μ. Let

$$\mu = \sum_{i=1}^{N} w_i x_i \in \mathcal{P}_{\leq N}(\mathbb{R})$$

and denote $\boldsymbol{x} = (x_1, \ldots, x_N)$, $\boldsymbol{w} = (w_1, \ldots, w_N)$. Then for all $t \in [0, 1)$, it can be shown that

$$Q_\mu(t) = x_{(k_{\min}(\sigma(\boldsymbol{x}), \boldsymbol{w}, t))}, \tag{20}$$

where $k_{\min}(\sigma, \boldsymbol{w}, t)$ is defined for $\sigma = (\sigma_1, \ldots, \sigma_N) \in S_N$ by

$$k_{\min}(\sigma, \boldsymbol{w}, t) \coloneqq \min \{k \in [N] \mid w_{\sigma_1} + \cdots + w_{\sigma_k} > t\}. \tag{21}$$

It can be seen in (20) and (21) that $Q_\mu(t)$ is monotone increasing with respect to t. Moreover,

$$Q_\mu(0) = \operatorname{ess\,min}(\mu) \quad \text{and} \quad \lim_{t \nearrow 1} Q_\mu(t) = \operatorname{ess\,max}(\mu),$$

with $\operatorname{ess\,min}(\mu)$ and $\operatorname{ess\,max}(\mu)$ denoting the essential minimum and maximum of the distribution μ. We thus augment the definition of Q_μ to $[0, 1]$ by setting $Q_\mu(1) = \operatorname{ess\,max}(\mu)$.

[6] See definition in Section 4

Note. In the following discussion we treat quantile functions only in terms of their integrals, and thus we only need their values at *almost* every $t \in [0,1]$. Still it's worth noting that under the above definition, $Q_\mu(t)$ is right-continuous on $[0,1]$, is continuous at both end points, and since it is monotone increasing, it only has jump discontinuities. Lastly, we note that $Q_\mu(t)$ indeed depends only on the distribution μ and not on its particular representation $\sum_{i=1}^N p_i x_i$, which can be verified from (20) and (21).

Using the identity (20), we can express $E(\mu; \boldsymbol{v}, \xi)$ as

$$\begin{aligned} E(\mu; \boldsymbol{v}, \xi) =& 2(1+\xi) \sum_{k=1}^N \int_{t=\sum_{i=1}^{k-1} w_{\sigma_i(\boldsymbol{v}^T \boldsymbol{X})}}^{\sum_{i=1}^{k} w_{\sigma_i(\boldsymbol{v}^T \boldsymbol{X})}} Q_{\boldsymbol{v}^T \mu}(t) \cos(2\pi\xi t) dt \\ =& 2(1+\xi) \sum_{k=1}^N \int_{t=\sum_{i=1}^{k-1} w_{\sigma_i(\boldsymbol{v}^T \boldsymbol{X})}}^{\sum_{i=1}^{k} w_{\sigma_i(\boldsymbol{v}^T \boldsymbol{X})}} \left(\boldsymbol{v}^T \boldsymbol{X}\right)_{(k)} \cos(2\pi\xi t) dt \\ =& 2\frac{1+\xi}{2\pi\xi} \sum_{k=1}^N \left(\boldsymbol{v}^T \boldsymbol{X}\right)_{(k)} \left[\sin(2\pi\xi t)\right]_{t=\sum_{i=1}^{k-1} w_{\sigma_i(\boldsymbol{v}^T \boldsymbol{X})}}^{\sum_{i=1}^{k} w_{\sigma_i(\boldsymbol{v}^T \boldsymbol{X})}}, \end{aligned} \tag{22}$$

under the notion $\sum_{i=1}^0 w_{\sigma_i(\boldsymbol{v}^T \boldsymbol{X})} = 0$. Rearranging terms gives us the alternative formula

$$E(\mu; \boldsymbol{v}, \xi) = 2\frac{1+\xi}{2\pi\xi} \sum_{k=1}^N \sin\left(2\pi\xi \sum_{i=1}^k w_{\sigma_i(\boldsymbol{v}^T \boldsymbol{X})}\right) \left[\left(\boldsymbol{v}^T \boldsymbol{X}\right)_{(k)} - \left(\boldsymbol{v}^T \boldsymbol{X}\right)_{(k+1)}\right], \tag{23}$$

with the definition of $\left(\boldsymbol{v}^T \boldsymbol{X}\right)_{(k)}$ extended to $k = N+1$ by

$$\left(\boldsymbol{v}^T \boldsymbol{X}\right)_{(N+1)} \coloneqq 0.$$

Using (23), the following theorem specifies the complexity of computing $E_m^{\text{FSW}}\left(\mu; \left(\boldsymbol{v}^{(k)}, \xi^{(k)}\right)_{k=1}^m\right)$.

Theorem A.2. *The function* $E_m^{\text{FSW}}\left(\mu; \left(\boldsymbol{v}^{(k)}, \xi^{(k)}\right)_{k=1}^m\right) : \mathcal{P}_{\leq N}\left(\mathbb{R}^d\right) \to \mathbb{R}^m$ *can be computed in* $\mathcal{O}(mNd + mN\log N)$ *time.*

Proof. Computing $E_m^{\text{FSW}}\left(\mu; \left(\boldsymbol{v}^{(k)}, \xi^{(k)}\right)_{k=1}^m\right)$ consists of the computation of $E^{\text{FSW}}\left(\mu; \boldsymbol{v}^{(k)}, \xi^{(k)}\right)$ independently for $k = 1, \ldots, m$. In (23) it can be seen that the most expensive operations in each such computation are: (1) the computation of N inner products in $\mathbb{R}^d$, and (2) finding the sorting permutation of these inner products. Therefore, each such computation takes $\mathcal{O}(Nd + N\log N)$ time. □

B PROOFS

B.1 THE COSINE TRANSFORM

The cosine transform takes a major role in our proofs. We now define it and present some of its properties. The results in this subsection appear in standard textbooks such as (Jones, 2001; Boas, 2006). We include them here for completeness.

In the following discussion, L^p denotes the space $L^p(\mathbb{R})$, defined by

$$L^p(\mathbb{R}) \coloneqq \{f : \mathbb{R} \to \mathbb{R} \mid f \text{ is Lebesgue measurable and } \|f\|_{L^p} < \infty\},$$

with

$$\|f\|_{L^p} \coloneqq \begin{cases} \left[\int_{\mathbb{R}} |f(t)|^p dt\right]^{1/p} & p \in [1, \infty) \\ \operatorname{ess\,sup}_{t\in\mathbb{R}} |f(t)| & p = \infty. \end{cases}$$

Definition B.1. Let $f \in L^1$ such that $f(t) = 0$ for all $t < 0$. The cosine transform of f is

$$\hat{f}(\xi) \coloneqq 2\int_0^\infty f(t)\cos(2\pi\xi t)dt \tag{24}$$

for $\xi \geq 0$.

Note that if $f \in L^1$, then

$$\left\|\hat{f}\right\|_{L^\infty} \leq 2\|f\|_{L^1} \tag{25}$$

since

$$\left|\hat{f}(\xi)\right| \leq 2\int_0^\infty |f(t)| \cdot |\cos(2\pi\xi t)|dt \leq 2\int_0^\infty |f(t)|dt = 2\|f\|_{L^1}. \tag{26}$$

Thus, $\hat{f} \in L^\infty$. The following lemma proves a better bound as $\xi \to \infty$ if f is monotonous, and shows that the cosine transform preserves the L^2-norm.

Lemma B.2 (Properties of the cosine transform)**.** *Let $f \in L^1$ such that $f(t) = 0$ for all $t < 0$. Then:*

1. *If $f \in L^1 \cap L^2$ then*
$$\int_0^\infty (f(t))^2 dt = \int_0^\infty \left(\hat{f}(t)\right)^2 dt. \tag{27}$$
2. *Suppose that $f \in L^1 \cap L^\infty$, and that f is monotonous on an interval $I = (0, T)$ and vanishes almost everywhere outside of I. Then for any $\xi > 0$,*
$$\left|\hat{f}(\xi)\right| \leq \frac{3}{\pi\xi}\|f\|_{L^\infty}. \tag{28}$$

Proof. We start from part 1. Let $f_e(t)$ be the even part of f,

$$f_e(t) \coloneqq \tfrac{1}{2}(f(t) + f(-t)) = \tfrac{1}{2}f(|t|).$$

Then the Fourier transform of f_e is given by

$$\begin{aligned}
\widehat{f_e}(\xi) &\coloneqq \int_{-\infty}^\infty f_e(t)e^{-2\pi i\xi t}dt \overset{\text{(a)}}{=} \int_{-\infty}^\infty f_e(t)\cos(-2\pi\xi t)dt \\
&= \int_{-\infty}^\infty \tfrac{1}{2}(f(t) + f(-t))\cos(-2\pi\xi t)dt \\
&= \tfrac{1}{2}\int_{-\infty}^0 (f(t) + f(-t))\cos(-2\pi\xi t)dt + \tfrac{1}{2}\int_0^\infty (f(t) + f(-t))\cos(-2\pi\xi t)dt \\
&= \tfrac{1}{2}\int_{-\infty}^0 f(-t)\cos(-2\pi\xi t)dt + \tfrac{1}{2}\int_0^\infty f(t)\cos(-2\pi\xi t)dt \\
&\underset{r=-t}{=\!=\!=} \tfrac{1}{2}\int_\infty^0 f(r)\cos(2\pi\xi r)(-dr) + \tfrac{1}{2}\int_0^\infty f(t)\cos(2\pi\xi t)dt \\
&= \int_0^\infty f(t)\cos(2\pi\xi t)dt = \tfrac{1}{2}\hat{f}(\xi),
\end{aligned}$$

with (a) holding since the Fourier transform of a real even function is real. Thus,

$$\hat{f}(\xi) = 2\widehat{f_e}(\xi).$$

Now extend the definition of $\hat{f}(\xi)$ to negative values of ξ, according (24), namely $\hat{f}(\xi) = \hat{f}(-\xi)$. Then

$$\begin{aligned}\int_0^\infty \left(\hat{f}(\xi)\right)^2 d\xi &= \tfrac{1}{2}\left\|\hat{f}\right\|_{L^2}^2 \\ &= 2\left\|\widehat{f_e}\right\|_{L^2}^2 \overset{\text{(a)}}{=} 2\|f_e\|_{L^2}^2 \overset{\text{(b)}}{=} \|f\|_{L^2}^2 \\ &= \int_{-\infty}^\infty (f(t))^2 dt = \int_0^\infty (f(t))^2 dt,\end{aligned}$$

with (a) holding by the Plancherel theorem, and (b) holding since

$$\begin{aligned}\|f_e\|_{L^2}^2 &= \int_{-\infty}^\infty (f_e(t))^2 dt = \int_{-\infty}^\infty \left(\tfrac{1}{2}(f(t)+f(-t))\right)^2 dt \\ &= \int_{-\infty}^\infty \left[\tfrac{1}{4}(f(t))^2 + \tfrac{1}{2}f(t)f(-t) + \tfrac{1}{4}(f(-t))^2\right] dt \\ &= \tfrac{1}{4}\int_{-\infty}^\infty \left[(f(t))^2 + (f(-t))^2\right] dt \\ &= \tfrac{1}{4}\int_0^\infty (f(t))^2 dt + \tfrac{1}{4}\int_{-\infty}^0 (f(-t))^2 dt \\ &= \tfrac{1}{2}\int_0^\infty (f(t))^2 dt = \tfrac{1}{2}\int_{-\infty}^\infty (f(t))^2 dt = \tfrac{1}{2}\|f\|_{L^2}^2.\end{aligned}$$

We now prove part 2. Suppose first that f is differentiable on I. Using integration by parts, we have

$$\begin{aligned}\hat{f}(\xi) &= 2\int_0^T f(t)\cos(2\pi\xi t)dt \\ &= \frac{1}{\pi\xi}\overbrace{\Big[f(t)\sin(2\pi\xi t)\Big]_{t=0}^T}^{A_1} - \frac{1}{\pi\xi}\overbrace{\int_0^T f'(t)\sin(2\pi\xi t)dt}^{A_2}.\end{aligned}$$

Let us now bound A_1 and A_2.

$$|A_1| = |f(T)\sin(2\pi\xi T)| \le |f(T)| \le \|f\|_{L^\infty},$$

and

$$\begin{aligned}|A_2| &= \left|\int_0^T f'(t)\sin(2\pi\xi t)dt\right| \\ &\le \int_0^T |f'(t)|\cdot|\sin(2\pi\xi t)|dt \\ &\le \int_0^T |f'(t)|dt \overset{\text{(a)}}{=} \left|\int_0^T f'(t)dt\right| \\ &= |f(T)-f(0)| \le 2\|f\|_{L^\infty},\end{aligned}$$

with (a) holding since f' does not change sign on $(0,T)$ due to the monotonicity of f.

In conclusion, we have

$$\left|\hat{f}(\xi)\right| \le \frac{1}{\pi\xi}(|A_1|+|A_2|) \le \frac{3}{\pi\xi}\|f\|_{L^\infty}.$$

To remove the differentiability assumption on f, we use the technique of *mollifying*; namely, replace f by a sequence of smooth functions that converges to it in L^1; see Chapter 7, Section C.3 of (Jones, 2001).

For the smooth functions to be monotonous, we first define a modified function $\tilde{f} : \mathbb{R} \to \mathbb{R}$

$$\tilde{f}(t) := \begin{cases} f(0^+) & t \leq 0 \\ f(t) & t \in (0, T) \\ f(T^-) & t \geq T, \end{cases} \tag{29}$$

where $f(0^+)$ and $f(T^-)$ are the one-sided limits of $f(t)$ as t approaches $0, T$ respectively from within the interval $(0, T)$.

With this definition, $\tilde{f}$ coincides with f on I, is monotonous on $\mathbb{R}$, and by construction,

$$\left\|\tilde{f}\right\|_{L^\infty} = \|f\|_{L^\infty}.$$

Let $\phi_\varepsilon : \mathbb{R} \to \mathbb{R}$ for $\varepsilon > 0$ be the mollifying function defined in (Jones, 2001), page 176. We now list a few properties of ϕ_ε.

1. ϕ_ε is infinitely differentiable and compactly supported.
2. ϕ_ε is radial, i.e. $\phi_\varepsilon(t) = \phi_\varepsilon(-t)$.
3. $\phi_\varepsilon(t) \geq 0$ for all t, and $\phi_\varepsilon(t) > 0$ iff $|t| < \varepsilon$.
4. $\int_{\mathbb{R}} \phi_\varepsilon(t)dt = 1$.

Let $f_\varepsilon : \mathbb{R} \to \mathbb{R}$ for $\varepsilon > 0$ be defined by

$$f_\varepsilon(t) := \chi_I(t) \int_{\mathbb{R}} \tilde{f}(r)\phi_\varepsilon(t-r)dr = \chi_I(t) \int_{\mathbb{R}} \tilde{f}(t+r)\phi_\varepsilon(r)dr, \tag{30}$$

with χ_I denoting the characteristic function of I. From the rightmost part of (30), it is evident that the monotonicity of $\tilde{f}$ implies that f_ε is monotonous on I.

Also note that

$$\begin{aligned} |f_\varepsilon(t)| &\leq \chi_I(t) \int_{\mathbb{R}} \left|\tilde{f}(t+r)\right| \phi_\varepsilon(r)dr \\ &\leq \left\|\tilde{f}\right\|_{L^\infty} \int_{\mathbb{R}} \phi_\varepsilon(r)dr \\ &= \left\|\tilde{f}\right\|_{L^\infty} = \|f\|_{L^\infty}. \end{aligned} \tag{31}$$

Thus,

$$\|f_\varepsilon\|_{L^1} \leq T\|f\|_{L^\infty}, \qquad \|f_\varepsilon\|_{L^\infty} \leq \|f\|_{L^\infty}, \tag{32}$$

and hence $f_\varepsilon \in L^1 \cap L^\infty$.

From the discussion in (Jones, 2001), f_ε satisfies:

1. $f_\varepsilon \in C^\infty(I)$
2. $\lim_{\varepsilon \to 0} \|f_\varepsilon - f\|_{L^1} = 0$

So far we have shown that for any $\varepsilon > 0$, f_ε is in $L^1 \cap L^\infty$, is monotonous and smooth on I, and vanishes outside of I. Therefore its cosine transform satisfies

$$\left|\hat{f}_\varepsilon(\xi)\right| \leq \frac{3}{\pi\xi}\|f_\varepsilon\|_{L^\infty} \overset{\text{(a)}}{\leq} \frac{3}{\pi\xi}\|f\|_{L^\infty}, \tag{33}$$

with (a) due to (32). Thus,

$$\begin{aligned}\tfrac{1}{2}\left|\hat{f}_\varepsilon(\xi)-\hat{f}(\xi)\right| =& \left|\int_0^T (f_\varepsilon(t)-f(t))\cos(2\pi\xi t)dt\right| \\ \leq& \|f_\varepsilon - f\|_{L^1}\|\cos(2\pi\xi t)\|_{L^\infty} \\ \leq& \|f_\varepsilon - f\|_{L^1} \underset{\varepsilon\to 0}{\longrightarrow} 0.\end{aligned}$$

In conclusion,

$$\frac{3}{\pi\xi}\|f\|_{L^\infty} \overset{(33)}{\geq} \left|\hat{f}_\varepsilon(\xi)\right| \underset{\varepsilon\to 0}{\longrightarrow} \left|\hat{f}(\xi)\right|$$

and therefore

$$\left|\hat{f}(\xi)\right| \leq \frac{3}{\pi\xi}\|f\|_{L^\infty}.$$

□

B.2 Probabilistic properties of $E^{\text{FSW}}(\mu;\boldsymbol{v},\xi)$

Here we use the notation

$$\Delta(\mu,\tilde{\mu};\boldsymbol{v},\xi) \coloneqq \left|E^{\text{FSW}}(\mu;\boldsymbol{v},\xi) - E^{\text{FSW}}(\tilde{\mu};\boldsymbol{v},\xi)\right|.$$

We define a *pseudonorm* for distributions in $\mathcal{P}_{\leq N}(\mathbb{R}^d)$ by

$$\|\mu\|_{\mathcal{W}_p} \coloneqq \mathcal{W}_p(\mu,\delta_0), \quad p\in[1,\infty],$$

where δ_0 denotes the distribution that assigns a mass of 1 to the point $0\in\mathbb{R}^d$. Note that $\|\cdot\|_{\mathcal{W}_p}$ is not truly a norm, since $\mathcal{P}_{\leq N}(\mathbb{R}^d)$ is not a vector space.

The following claim provides a useful bound on the Wasserstein and sliced Wasserstein distances.

Claim B.3. *For any $\mu,\nu\in\mathcal{P}_{\leq N}(\mathbb{R}^d)$,*

$$\mathcal{SW}(\mu,\nu) \leq \mathcal{W}(\mu,\nu) \leq \|\mu\|_{\mathcal{W}_\infty} + \|\nu\|_{\mathcal{W}_\infty}. \tag{34}$$

Proof. The left inequality is a well-known property of the sliced Wasserstein distance; see e.g. Eq. (3.2) of (Bayraktar & Guo, 2021). The right inequality is easy to see by considering the transport plans that transport each of the distributions to δ_0, and applying the triangle inequality. □

To prove Theorem 3.2, we first prove the following lemma.

Lemma B.4. *Let $\mu,\nu\in\mathcal{P}_{\leq N}(\mathbb{R}^d)$ and $\boldsymbol{v}\in\mathbb{S}^{d-1}$. Let $\xi\sim\mathcal{D}_\xi$. Then*

$$\left|E^{\text{FSW}}(\mu;\boldsymbol{v},\xi)\right| \leq 3\|\mu\|_{\mathcal{W}_\infty} \quad \forall\xi\geq 0, \tag{35}$$

$$\mathbb{E}_\xi\left[\Delta^2(\mu,\nu;\boldsymbol{v},\xi)\right] = \mathcal{W}^2\left(\boldsymbol{v}^T\mu,\boldsymbol{v}^T\nu\right), \tag{36}$$

$$\text{STD}_\xi\left[\Delta^2(\mu,\nu;\boldsymbol{v},\xi)\right] \leq 3\left(\|\mu\|_{\mathcal{W}_\infty} + \|\nu\|_{\mathcal{W}_\infty}\right)\mathcal{W}\left(\boldsymbol{v}^T\mu,\boldsymbol{v}^T\nu\right). \tag{37}$$

Proof. By definition,

$$E^{\text{FSW}}(\mu;\boldsymbol{v},\xi) = (1+\xi)\hat{Q}_{\boldsymbol{v}^T\mu}(\xi). \tag{38}$$

From part 2 of Lemma B.2,

$$\left|\hat{Q}_{\boldsymbol{v}^T\mu}(\xi)\right| \leq \frac{3}{\pi\xi}\left\|Q_{\boldsymbol{v}^T\mu}\right\|_{L^\infty} \leq \frac{3}{\pi\xi}\|\mu\|_{\mathcal{W}_\infty}$$

and from (26),

$$\left|\hat{Q}_{\boldsymbol{v}^T\mu}(\xi)\right| \leq 2\left\|Q_{\boldsymbol{v}^T\mu}\right\|_{L^1} \overset{(a)}{\leq} 2\left\|Q_{\boldsymbol{v}^T\mu}\right\|_{L^\infty} = 2\|\mu\|_{\mathcal{W}_\infty},$$

with (a) holding since $Q_{\boldsymbol{v}^T\mu}$ is supported on $[0,1]$. Thus,

$$\left|\hat{Q}_{\boldsymbol{v}^T\mu}(\xi)\right| \leq \min\left\{2, \frac{3}{\pi\xi}\right\}\|\mu\|_{\mathcal{W}_\infty},$$

which implies

$$\left|E^{\text{FSW}}(\mu;\boldsymbol{v},\xi)\right| \leq (1+\xi)\min\left\{2, \frac{3}{\pi\xi}\right\}\|\mu\|_{\mathcal{W}_\infty} \leq \left(2+\frac{3}{\pi}\right)\|\mu\|_{\mathcal{W}_\infty} \leq 3\|\mu\|_{\mathcal{W}_\infty},$$

and thus (35) holds. Note that since $E^{\text{FSW}}(\mu;\boldsymbol{v},\xi)$ is bounded as a function of ξ, so is $\Delta^2(\mu,\nu;\boldsymbol{v},\xi)$, and therefore both have finite moments of all orders with respect to ξ.

Now,

$$\begin{aligned}
\mathbb{E}_\xi\left[\Delta^2(\mu,\nu;\boldsymbol{v},\xi)\right] &= \mathbb{E}_\xi\left[\left(E^{\text{FSW}}(\mu;\boldsymbol{v},\xi) - E^{\text{FSW}}(\nu;\boldsymbol{v},\xi)\right)^2\right] \\
&= \int_0^\infty \frac{1}{(1+\xi)^2}\left((1+\xi)^2\left(\hat{Q}_{\boldsymbol{v}^T\mu}(\xi) - \hat{Q}_{\boldsymbol{v}^T\nu}(\xi)\right)^2\right)d\xi \\
&= \int_0^\infty \left(\hat{Q}_{\boldsymbol{v}^T\mu}(\xi) - \hat{Q}_{\boldsymbol{v}^T\nu}(\xi)\right)^2 d\xi \\
&\overset{\text{(a)}}{=} \int_0^\infty \left(Q_{\boldsymbol{v}^T\mu}(t) - Q_{\boldsymbol{v}^T\nu}(t)\right)^2 dt \\
&= \int_0^1 \left(Q_{\boldsymbol{v}^T\mu}(t) - Q_{\boldsymbol{v}^T\nu}(t)\right)^2 dt \\
&\overset{\text{(b)}}{=} \mathcal{W}^2\left(\boldsymbol{v}^T\mu, \boldsymbol{v}^T\nu\right),
\end{aligned}$$

with (a) following from part 1 of Lemma B.2 and the linearity of the cosine transform, and (b) holding by the identity (4). Thus, (36) holds.

To bound the variance of $\Delta^2(\mu,\nu;\boldsymbol{v},\xi)$, note that

$$\begin{aligned}
\text{Var}_\xi\left[\Delta^2(\mu,\nu;\boldsymbol{v},\xi)\right] &= \mathbb{E}_\xi\left[\left(\Delta^2(\mu,\nu;\boldsymbol{v},\xi)\right)^2\right] - \left(\mathbb{E}_\xi\left[\Delta^2(\mu,\nu;\boldsymbol{v},\xi)\right]\right)^2 \\
&\overset{(36)}{=} \mathbb{E}_\xi\left[\Delta^4(\mu,\nu;\boldsymbol{v},\xi)\right] - \left(\mathcal{W}^2\left(\boldsymbol{v}^T\mu, \boldsymbol{v}^T\nu\right)\right)^2 \\
&= \mathbb{E}_\xi\left[\left(E^{\text{FSW}}(\mu;\boldsymbol{v},\xi) - E^{\text{FSW}}(\nu;\boldsymbol{v},\xi)\right)^2 \cdot \Delta^2(\mu,\nu;\boldsymbol{v},\xi)\right] - \mathcal{W}^4\left(\boldsymbol{v}^T\mu, \boldsymbol{v}^T\nu\right) \\
&\leq \mathbb{E}_\xi\left[\left(\left|E^{\text{FSW}}(\mu;\boldsymbol{v},\xi)\right| + \left|E^{\text{FSW}}(\nu;\boldsymbol{v},\xi)\right|\right)^2 \cdot \Delta^2(\mu,\nu;\boldsymbol{v},\xi)\right] - \mathcal{W}^4\left(\boldsymbol{v}^T\mu, \boldsymbol{v}^T\nu\right) \\
&\overset{(35)}{\leq} \mathbb{E}_\xi\left[\left(3\|\mu\|_{\mathcal{W}_\infty} + 3\|\nu\|_{\mathcal{W}_\infty}\right)^2 \cdot \Delta^2(\mu,\nu;\boldsymbol{v},\xi)\right] - \mathcal{W}^4\left(\boldsymbol{v}^T\mu, \boldsymbol{v}^T\nu\right) \\
&= 9\left(\|\mu\|_{\mathcal{W}_\infty} + \|\nu\|_{\mathcal{W}_\infty}\right)^2 \cdot \mathbb{E}_\xi\left[\Delta^2(\mu,\nu;\boldsymbol{v},\xi)\right] - \mathcal{W}^4\left(\boldsymbol{v}^T\mu, \boldsymbol{v}^T\nu\right) \\
&\overset{(36)}{=} 9\left(\|\mu\|_{\mathcal{W}_\infty} + \|\nu\|_{\mathcal{W}_\infty}\right)^2 \cdot \mathcal{W}^2\left(\boldsymbol{v}^T\mu, \boldsymbol{v}^T\nu\right) - \mathcal{W}^4\left(\boldsymbol{v}^T\mu, \boldsymbol{v}^T\nu\right) \\
&\leq 9\left(\|\mu\|_{\mathcal{W}_\infty} + \|\nu\|_{\mathcal{W}_\infty}\right)^2 \cdot \mathcal{W}^2\left(\boldsymbol{v}^T\mu, \boldsymbol{v}^T\nu\right),
\end{aligned}$$

and thus (37) holds.

This concludes the proof of Lemma B.4. □

Let us now prove Theorem 3.2.

Theorem 3.2. (Statement in Section 3) *Let $\mu, \tilde{\mu} \in \mathcal{P}_{\leq N}(B_R)$, with $B_R \subset \mathbb{R}^d$ being the closed ℓ_2-norm ball of radius R, centered at zero. Let $\boldsymbol{v} \sim \text{Uniform}\left(\mathbb{S}^{d-1}\right)$, $\xi \sim \mathcal{D}_\xi$. Then*

$$\mathbb{E}_{\boldsymbol{v},\xi}\left[\left|E^{\text{FSW}}(\mu;\boldsymbol{v},\xi) - E^{\text{FSW}}(\tilde{\mu};\boldsymbol{v},\xi)\right|^2\right] = \mathcal{SW}^2(\mu,\tilde{\mu}), \tag{7}$$

$$\text{STD}_{\boldsymbol{v},\xi}\left[\left|E^{\text{FSW}}(\mu;\boldsymbol{v},\xi) - E^{\text{FSW}}(\tilde{\mu};\boldsymbol{v},\xi)\right|^2\right] \leq 4\sqrt{10}R^2. \tag{8}$$

Proof. Eq. (7) holds since

$$\begin{aligned}\mathbb{E}_{\boldsymbol{v},\xi}\left[\Delta^2(\mu,\nu;\boldsymbol{v},\xi)\right] &= \mathbb{E}_{\boldsymbol{v}}\left[\mathbb{E}_{\xi|\boldsymbol{v}}\left[\Delta^2(\mu,\nu;\boldsymbol{v},\xi)\right]\right] \\ &\overset{(36)}{=} \mathbb{E}_{\boldsymbol{v}}\left[\mathcal{W}^2\left(\boldsymbol{v}^T\mu,\boldsymbol{v}^T\nu\right)\right] \\ &\overset{(5)}{=} \mathcal{SW}^2(\mu,\nu).\end{aligned}$$

We now prove (8).

$$\begin{aligned}\mathrm{Var}_{\boldsymbol{v},\xi}\left[\Delta^2(\mu,\nu;\boldsymbol{v},\xi)\right] &\overset{(a)}{=} \mathbb{E}_{\boldsymbol{v}}\left[\mathrm{Var}_{\xi|\boldsymbol{v}}\left[\Delta^2(\mu,\nu;\boldsymbol{v},\xi)\right]\right] + \mathrm{Var}_{\boldsymbol{v}}\left[\mathbb{E}_{\xi|\boldsymbol{v}}\left[\Delta^2(\mu,\nu;\boldsymbol{v},\xi)\right]\right] \\ &\overset{(36)}{=} \mathbb{E}_{\boldsymbol{v}}\left[\mathrm{Var}_{\xi|\boldsymbol{v}}\left[\Delta^2(\mu,\nu;\boldsymbol{v},\xi)\right]\right] + \mathrm{Var}_{\boldsymbol{v}}\left[\mathcal{W}^2\left(\boldsymbol{v}^T\mu,\boldsymbol{v}^T\nu\right)\right] \\ &\overset{(37)}{\leq} \mathbb{E}_{\boldsymbol{v}}\left[9\left(\|\mu\|_{\mathcal{W}_\infty}+\|\nu\|_{\mathcal{W}_\infty}\right)^2\mathcal{W}^2\left(\boldsymbol{v}^T\mu,\boldsymbol{v}^T\nu\right)\right] + \mathrm{Var}_{\boldsymbol{v}}\left[\mathcal{W}^2\left(\boldsymbol{v}^T\mu,\boldsymbol{v}^T\nu\right)\right] \\ &= 9\left(\|\mu\|_{\mathcal{W}_\infty}+\|\nu\|_{\mathcal{W}_\infty}\right)^2\mathbb{E}_{\boldsymbol{v}}\left[\mathcal{W}^2\left(\boldsymbol{v}^T\mu,\boldsymbol{v}^T\nu\right)\right] + \mathrm{Var}_{\boldsymbol{v}}\left[\mathcal{W}^2\left(\boldsymbol{v}^T\mu,\boldsymbol{v}^T\nu\right)\right] \\ &\overset{(5)}{=} 9\left(\|\mu\|_{\mathcal{W}_\infty}+\|\nu\|_{\mathcal{W}_\infty}\right)^2\mathcal{SW}^2(\mu,\nu) + \mathrm{Var}_{\boldsymbol{v}}\left[\mathcal{W}^2\left(\boldsymbol{v}^T\mu,\boldsymbol{v}^T\nu\right)\right] \\ &\leq 9\left(\|\mu\|_{\mathcal{W}_\infty}+\|\nu\|_{\mathcal{W}_\infty}\right)^2\mathcal{SW}^2(\mu,\nu) + \mathbb{E}_{\boldsymbol{v}}\left[\mathcal{W}^4\left(\boldsymbol{v}^T\mu,\boldsymbol{v}^T\nu\right)\right] \\ &\overset{(34)}{\leq} 9\left(\|\mu\|_{\mathcal{W}_\infty}+\|\nu\|_{\mathcal{W}_\infty}\right)^2\left(\|\mu\|_{\mathcal{W}_\infty}+\|\nu\|_{\mathcal{W}_\infty}\right)^2 + \mathbb{E}_{\boldsymbol{v}}\left[\left(\|\mu\|_{\mathcal{W}_\infty}+\|\nu\|_{\mathcal{W}_\infty}\right)^4\right] \\ &= 10\left(\|\mu\|_{\mathcal{W}_\infty}+\|\nu\|_{\mathcal{W}_\infty}\right)^4,\end{aligned}$$

where (a) is by (Wasserman, 2004, Theorem 3.27, pg. 55). Thus, (8) holds. □

B.3 Injectivity and bi-Lipschitzness

Theorem 4.1. (Statement in Section 4) *Let* $E_m^{\mathrm{FSW}} : \mathcal{P}_{\leq N}\left(\mathbb{R}^d\right) \to \mathbb{R}^m$ *be as in* (9)*, with* $\left(\boldsymbol{v}^{(k)},\xi^{(k)}\right)_{k=1}^m$ *drawn i.i.d. from* $\mathrm{Uniform}\left(\mathbb{S}^{d-1}\right)\times\mathcal{D}_\xi$*. Then:*

1. *If* $m \geq 2Nd+1$*, then with probability 1,* E_m^{FSW} *is injective on* $\mathcal{S}_{\leq N}\left(\mathbb{R}^d\right)$*.*
2. *If* $m \geq 2Nd+2N-1$*, then with probability 1,* E_m^{FSW} *is injective on* $\mathcal{P}_{\leq N}\left(\mathbb{R}^d\right)$*.*

Proof. Our proof relies on the theory of σ-subanalytic functions, introduced in (Amir et al., 2023), and its main result—the *Finite Witness Theorem*—which enables reducing an infinite family of equality constraints to a finite sub-family chosen randomly, while maintaining equivalence with probability 1. The Finite Witness Theorem is a powerful tool for proving the injectivity of certain functions.

The theory defines a class of sets called *σ-subanalytic sets* and a class of functions called *σ-subanalytic functions*. Since these definitions are technically involved and require elaborate construction, we do not include them here. However, our proof is self-contained, and we state here the relevant properties of these constructs. For further details, we refer the interested reader to (Amir et al., 2023, Appendix A).

We begin by stating the full version of the Finite Witness Theorem, which will be our main tool for proving Theorem 4.1.

Theorem B.5 (Finite Witness Theorem)**.** *Let* $\mathbb{M}\subseteq\mathbb{R}^p$ *and* $\mathbb{W}\subseteq\mathbb{R}^q$ *be σ-subanalytic sets of dimensions D and $D_{\boldsymbol{\theta}}$ respectively. Let* $F:\mathbb{M}\times\mathbb{W}\to\mathbb{R}$ *be a σ-subanalytic function. Define the set*

$$\mathcal{N} := \{\boldsymbol{z}\in\mathbb{M} \mid F(\boldsymbol{z};\boldsymbol{\theta})=0,\ \forall\boldsymbol{\theta}\in\mathbb{W}\}.$$

Suppose that for all $\boldsymbol{z}\in\mathbb{M}\setminus\mathcal{N}$*,*

$$\dim\{\boldsymbol{\theta}\in\mathbb{W} \mid F(\boldsymbol{z};\boldsymbol{\theta})=0\} \leq D_{\boldsymbol{\theta}}-1. \tag{39}$$

Then for generic $\left(\boldsymbol{\theta}^{(1)},\ldots,\boldsymbol{\theta}^{(D+1)}\right)\in\mathbb{W}^{D+1}$*,*

$$\mathcal{N} = \left\{\boldsymbol{z}\in\mathbb{M} \;\middle|\; F\left(\boldsymbol{z};\boldsymbol{\theta}^{(i)}\right)=0,\ \forall i=1,\ldots,D+1\right\}.$$

Moreover, if $\mathbb{W}$ is an open and connected subset of $\mathbb{R}^q$, and $F(\boldsymbol{z};\boldsymbol{\theta})$ is analytic as a function of $\boldsymbol{\theta}$ for all fixed $\boldsymbol{z} \in \mathbb{M}$, then condition (39) is not required, as it is automatically satisfied.

The function $F(\boldsymbol{z};\boldsymbol{\theta})$ in the theorem can be regarded as representing a parametric family of constraints on $\boldsymbol{z}$, parametrized by $\boldsymbol{\theta}$: a point $\boldsymbol{z} \in \mathbb{M}$ belongs to the feasible set $\mathcal{N}$ if it satisfies the constraints $F(\boldsymbol{z};\boldsymbol{\theta}) = 0$ for all parameter vectors $\boldsymbol{\theta} \in \mathbb{W}$. The vectors $\boldsymbol{\theta}$ are called *witness vectors*.

The theorem essentially states that, if one draws a sufficiently large number of witness vectors from $\mathbb{W}$ at random, then with probability 1, that set of vectors is sufficient to fully determine $\mathcal{N}$.

The number of required vectors depends on the dimension of $\mathbb{M}$. The notion of dimension used in the theorem is the *Hausdorff dimension*, which coincides with the standard notion of dimension for vector spaces and smooth manifolds. The condition (39) is called the *dimension deficiency condition*.

Our next step is to show *how* Theorem 4.1 follows from Theorem B.5, assuming the latter's assumptions are satisfied. We then establish that these assumptions indeed hold. First, we prove part 2 of Theorem 4.1. To this end, we must choose the appropriate $\mathbb{M}, \mathbb{W}$ and F. First, set

$$\mathbb{M} = \mathbb{R}^{d\times N} \times \mathbb{R}^{d\times N} \times \Delta_N \times \Delta_N,$$

where Δ_N is the probability simplex in $\mathbb{R}^N$. Namely, $\mathbb{M}$ is the set of parameters required to describe *pairs* of distributions $\mu, \tilde{\mu} \in \mathcal{P}_{\leq N}\big(\mathbb{R}^d\big)$:

$$\mathbb{M} = \Big\{ \Big(\boldsymbol{X}, \tilde{\boldsymbol{X}}, \boldsymbol{w}, \tilde{\boldsymbol{w}} \Big) \;\Big|\; \boldsymbol{X}, \tilde{\boldsymbol{X}} \in \mathbb{R}^{d\times N},\ \boldsymbol{w}, \tilde{\boldsymbol{w}} \in \Delta_N \Big\}.$$

Throughout our proof, we use the convention

$$\mu = \sum_{i=1}^{N} w_i \delta_{\boldsymbol{x}^{(i)}}, \quad \tilde{\mu} = \sum_{i=1}^{N} \tilde{w}_i \delta_{\tilde{\boldsymbol{x}}^{(i)}}.$$

Since $\dim \Delta_N = N-1$, the dimension of $\mathbb{M}$ is

$$\dim \mathbb{M} = 2Nd + 2N - 2.$$

Second, set $\mathbb{W}$ to be the space of embedding parameters

$$\mathbb{W} = \big\{ (\boldsymbol{v}, \xi) \;\big|\; \boldsymbol{v} \in \mathbb{S}^{d-1}, \xi \in (0,\infty) \big\}.$$

Now, define the function $F : \mathbb{M} \times \mathbb{W} \to \mathbb{R}$,

$$F\Big(\boldsymbol{X}, \tilde{\boldsymbol{X}}, \boldsymbol{w}, \tilde{\boldsymbol{w}}; \boldsymbol{v}, \xi \Big) \coloneqq E^{\text{FSW}}(\boldsymbol{X}, \boldsymbol{w}; \boldsymbol{v}, \xi) - E^{\text{FSW}}\Big(\tilde{\boldsymbol{X}}, \tilde{\boldsymbol{w}}; \boldsymbol{v}, \xi \Big), \tag{40}$$

where, by abuse of notation, $E^{\text{FSW}}(\boldsymbol{X}, \boldsymbol{w}; \boldsymbol{v}, \xi)$ denotes $E^{\text{FSW}}(\mu; \boldsymbol{v}, \xi)$ and $E^{\text{FSW}}\Big(\tilde{\boldsymbol{X}}, \tilde{\boldsymbol{w}}; \boldsymbol{v}, \xi \Big)$ denotes $E^{\text{FSW}}(\tilde{\mu}; \boldsymbol{v}, \xi)$.

Essentially, $F\Big(\boldsymbol{X}, \tilde{\boldsymbol{X}}, \boldsymbol{w}, \tilde{\boldsymbol{w}}; \boldsymbol{v}, \xi \Big)$ compares the two input distributions $\mu, \tilde{\mu}$ by considering their one-sample embeddings $E^{\text{FSW}}(\,\cdot\,; \boldsymbol{v}, \xi)$ with the parameters $\boldsymbol{v}, \xi$; specifically, $F\Big(\boldsymbol{X}, \tilde{\boldsymbol{X}}, \boldsymbol{w}, \tilde{\boldsymbol{w}}; \boldsymbol{v}, \xi \Big)$ equals zero if and only if both distributions μ and $\tilde{\mu}$ "appear identical" to $E^{\text{FSW}}(\,\cdot\,; \boldsymbol{v}, \xi)$ with the embedding parameters $\boldsymbol{v}, \xi$.

The full embedding $E_m^{\text{FSW}}\Big(\cdot\,; \big(\boldsymbol{v}^{(k)}, \xi^{(k)}\big)_{k=1}^m \Big)$ uses m pairs of parameters $\big(\boldsymbol{v}^{(k)}, \xi^{(k)}\big)_{k=1}^m$ drawn independently at random. Recall that part 2 of the theorem assumes that

$$m \geq 2Nd + 2N - 1 = \dim \mathbb{M} + 1.$$

Thus, provided that the rest of the assumptions of Theorem B.5 are satisfied, the theorem states that, with probability 1 on the embedding parameters $\big(\boldsymbol{v}^{(k)}, \xi^{(k)}\big)_{k=1}^m$ the following holds: for any input pair of distributions $\mu, \tilde{\mu} \in \mathcal{P}_{\leq N}\big(\mathbb{R}^d\big)$, if $E_m^{\text{FSW}}\Big(\cdot\,; \big(\boldsymbol{v}^{(k)}, \xi^{(k)}\big)_{k=1}^m \Big)$ does not distinguish between μ and $\tilde{\mu}$, then

$$E^{\text{FSW}}(\mu; \boldsymbol{v}, \xi) = E^{\text{FSW}}(\tilde{\mu}; \boldsymbol{v}, \xi)$$

for *all* parameters $(\boldsymbol{v}, \xi) \in \mathbb{W}$. This, combined with (7), implies that $\mathcal{SW}(\mu, \tilde{\mu}) = 0$, and therefore $\mu = \tilde{\mu}$. Thus, with probability 1, E_m^{FSW} is injective on $\mathcal{P}_{\leq N}(\mathbb{R}^d)$.

To prove part 1 of the theorem, namely, injectivity on $\mathcal{S}_{\leq N}(\mathbb{R}^d)$, we restrict $\mathbb{M}$ to parameters that describe pairs of multisets in $\mathcal{S}_{\leq N}(\mathbb{R}^d)$. That is,

$$\mathbb{M} = \left\{ \left(\boldsymbol{X}, \tilde{\boldsymbol{X}}, \boldsymbol{w}, \tilde{\boldsymbol{w}} \right) \;\middle|\; \boldsymbol{X}, \tilde{\boldsymbol{X}} \in \mathbb{R}^{d \times N},\ \boldsymbol{w}, \tilde{\boldsymbol{w}} \in \left\{ \boldsymbol{w}^{(1)}, \ldots, \boldsymbol{w}^{(N)} \right\} \right\},$$

where each $\boldsymbol{w}^{(n)}$ for $n \in [N]$ is a uniform weight vector corresponding to a multiset of size n, as defined in (13). Note that with this definition, $\mathbb{M}$ is a finite union of affine spaces of dimension $2Nd$, and thus

$$\dim \mathbb{M} = 2Nd.$$

Recall that part 1 of the theorem assumes that

$$m \geq 2Nd + 1 = \dim \mathbb{M} + 1,$$

and the rest of the proof follows as for part 2.

It is now left for us to show that F indeed satisfies the assumptions of Theorem B.5.

First, we argue that $\mathbb{M}$ and $\mathbb{W}$ are σ-subanalytic sets. The reason is that both sets are semialgebraic, which is easy to verify, and any semialgebraic set is σ-subanalytic; see Figure 3 in (Amir et al., 2023, Appendix A).

Next, we want to show that F is a σ-subanalytic function. To this end, we state a few properties of σ-subanalytic functions.

Proposition B.6. *Properties of σ-subanalytic functions:*

1. *Analytic functions defined on open domains are σ-subanalytic.*
2. *Any semialgebraic function, and in particular piecewise-linear functions, are σ-subanalytic.*
3. *Finite sums, products and compositions of σ-subanalytic functions are σ-subanalytic.*

Property 1 follows from (Amir et al., 2023, Lemma A.14). For Property 2, see Figure 3 in (Amir et al., 2023, Appendix A). Property 3 is established in (Amir et al., 2023, Proposition A.15).

To show that F is σ-subanalytic, we first show that $E^{\text{FSW}}(\boldsymbol{X}, \boldsymbol{w}; \boldsymbol{v}, \xi)$ is σ-subanalytic as a function of $(\boldsymbol{X}, \boldsymbol{w}, \boldsymbol{v}, \xi)$. To see this, note that by (23), $E^{\text{FSW}}(\boldsymbol{X}, \boldsymbol{w}; \boldsymbol{v}, \xi)$ is the sum of terms of the form

$$2\frac{1+\xi}{2\pi\xi} \sin\left(2\pi\xi \sum_{i=1}^{k} w_{\sigma_i(\boldsymbol{v}^T \boldsymbol{X})} \right) \left[\left(\boldsymbol{v}^T \boldsymbol{X}\right)_{(k)} - \left(\boldsymbol{v}^T \boldsymbol{X}\right)_{(k+1)} \right], \tag{41}$$

with the sum taken over $k \in [N]$. The product $\boldsymbol{v}^T \boldsymbol{X}$ is semialgebraic, and thus is σ-subanalytic. Each term $\left[\left(\boldsymbol{v}^T \boldsymbol{X}\right)_{(k)} - \left(\boldsymbol{v}^T \boldsymbol{X}\right)_{(k+1)} \right]$ is piecewise linear in $\boldsymbol{v}^T \boldsymbol{X}$ and thus σ-subanalytic. Each $\sum_{i=1}^{k} w_{\sigma_i(\boldsymbol{v}^T \boldsymbol{X})}$ is semialgebraic and thus is also σ-analytic. The product $2\pi\xi \sum_{i=1}^{k} w_{\sigma_i(\boldsymbol{v}^T \boldsymbol{X})}$, composition with the analytic $\sin$ function $\sin\left(2\pi\xi \sum_{i=1}^{k} w_{\sigma_i(\boldsymbol{v}^T \boldsymbol{X})} \right)$, and again product $2\frac{1+\xi}{2\pi\xi} \sin\left(2\pi\xi \sum_{i=1}^{k} w_{\sigma_i(\boldsymbol{v}^T \boldsymbol{X})} \right)$ is also σ-subanalytic. Finally, the product (41) and the finite sum of such over $k \in [N]$ is σ-subanalytic.

Thus, since F is the difference of two σ-subanalytic functions, it is also σ-subanalytic.

Since F is not analytic as a function of $(\boldsymbol{v}, \xi)$ for all fixed $\left(\boldsymbol{X}, \tilde{\boldsymbol{X}}, \boldsymbol{w}, \tilde{\boldsymbol{w}} \right)$, we need to explicitly show that F satisfies the dimension deficiency condition of Theorem B.5. Let $\mu, \tilde{\mu} \in \mathcal{P}_{\leq N}(\mathbb{R}^d)$ be two fixed distributions and suppose that $\mu \neq \tilde{\mu}$. Let $A \subseteq \mathbb{W}$ be the set

$$A \coloneqq \left\{ (\boldsymbol{v}, \xi) \in \mathbb{S}^{d-1} \times (0, \infty) \;\middle|\; E^{\text{FSW}}(\mu; \boldsymbol{v}, \xi) = E^{\text{FSW}}(\tilde{\mu}; \boldsymbol{v}, \xi) \right\}.$$

We need to show that A is dimension deficient; that is, $\dim A < \dim \mathbb{W}$. Suppose, by contradiction, that A is of full dimension. Since E^{FSW} is continuous, A is closed. Since A is closed and of full-dimension, it has a nonempty interior in the product topology of $\mathbb{W} = \mathbb{S}^{d-1} \times (0, \infty)$. By basic topology, this implies that A contains a subset of the form $B \times C$, where $B \subseteq \mathbb{S}^{d-1}$ and $C \subseteq (0, \infty)$ are open and nonempty.

For any fixed $\boldsymbol{v} \in B$, the function $E^{\text{FSW}}(\mu; \boldsymbol{v}, \xi)$ is analytic as a function of ξ on $(0, \infty)$, as can be seen in (41). Thus, the function $f_{\boldsymbol{v}} : (0, \infty) \to \mathbb{R}$ given by

$$f_{\boldsymbol{v}}(\xi) = E^{\text{FSW}}(\mu; \boldsymbol{v}, \xi) - E^{\text{FSW}}(\tilde{\mu}; \boldsymbol{v}, \xi)$$

is also analytic. Since $f_{\boldsymbol{v}} = 0$ on the nonempty open set C, $f_{\boldsymbol{v}} = 0$ on all of $(0, \infty)$. By (36), this implies that

$$\mathcal{W}\big(\boldsymbol{v}^T \mu, \boldsymbol{v}^T \tilde{\mu}\big) = \sqrt{\mathbb{E}_{\xi}\Big[f_{\boldsymbol{v}}(\xi)^2\Big]} = 0, \tag{42}$$

and thus $\boldsymbol{v}^T \mu = \boldsymbol{v}^T \tilde{\mu}$; that is, the two projected distributions $\boldsymbol{v}^T \mu$ and $\boldsymbol{v}^T \tilde{\mu}$ are identical.

Note that the above holds for all $\boldsymbol{v} \in B$, which is an open subset of the unit sphere $\mathbb{S}^{d-1}$. We shall now show that this implies that $\mu = \tilde{\mu}$, which is a contradiction.

To facilitate analysis, we replace the set B, which is confined to the unit sphere, by its *back-projection* to $\mathbb{R}^d$,

$$\tilde{B} \coloneqq \left\{ \boldsymbol{v} \in \mathbb{R}^d \setminus 0 \;\middle|\; \frac{\boldsymbol{v}}{\|\boldsymbol{v}\|} \in B \right\}.$$

Since B is open in $\mathbb{S}^{d-1}$ and the projection operator $\boldsymbol{v} \to \frac{\boldsymbol{v}}{\|\boldsymbol{v}\|}$ is continuous, we have that $\tilde{B}$ is open in $\mathbb{R}^d$. Since $\tilde{B} \supset B$, and B is nonempty, then so is $\tilde{B}$. Moreover, for any $\boldsymbol{v} \in \tilde{B}$, $\boldsymbol{v}^T \mu = \boldsymbol{v}^T \tilde{\mu}$.

Now, let $S : \mathbb{R} \to \mathbb{R}$ be the sigmoid function

$$S(t) \coloneqq \frac{1}{1 + \exp(-t)}.$$

By Lemma 1 of (Cybenko, 1989), the function S is *discriminatory*; that is, for any two distinct distributions μ, $\tilde{\mu}$ over $\mathbb{R}^d$, there exist $\boldsymbol{v} \in \mathbb{R}^d$ and $b \in \mathbb{R}$ such that

$$\int_{\mathbb{R}^d} S(\langle \boldsymbol{v}, \boldsymbol{x} \rangle + b) d\mu(\boldsymbol{x}) \neq \int_{\mathbb{R}^d} S(\langle \boldsymbol{v}, \boldsymbol{x} \rangle + b) d\tilde{\mu}(\boldsymbol{x}). \tag{43}$$

Since by our assumption $\mu \neq \tilde{\mu}$, there exist $\boldsymbol{v} \in \mathbb{R}^d$, $b \in \mathbb{R}$ that satisfy (43).

Define the function $f(\boldsymbol{v}, b) : \mathbb{R}^d \times \mathbb{R} \to \mathbb{R}$ by

$$\begin{aligned} f(\boldsymbol{v}, b) &= \int_{\mathbb{R}^d} S(\langle \boldsymbol{v}, \boldsymbol{x} \rangle + b) d\mu(\boldsymbol{x}) - \int_{\mathbb{R}^d} S(\langle \boldsymbol{v}, \boldsymbol{x} \rangle + b) d\tilde{\mu}(\boldsymbol{x}) \\ &= \sum_{i=1}^{N} w_i S\Big(\Big\langle \boldsymbol{v}, \boldsymbol{x}^{(i)} \Big\rangle + b\Big) - \sum_{i=1}^{N} \tilde{w}_i S\Big(\Big\langle \boldsymbol{v}, \tilde{\boldsymbol{x}}^{(i)} \Big\rangle + b\Big). \end{aligned}$$

It is easy to see that f is analytic. The two integral terms in $f(\boldsymbol{v}, b)$ depend only on the projected distributions $\boldsymbol{v}^T \mu$ and $\boldsymbol{v}^T \tilde{\mu}$ respectively. Since these two distributions are identical for any $\boldsymbol{v} \in \tilde{B}$, f is annihilated on the set $\tilde{B} \times \mathbb{R}$, which is open in $\mathbb{R}^d \times \mathbb{R}$ and nonempty. Therefore, f is identically zero, thereby contradicting (43). Thus, we have shown that F satisfies the dimension deficiency condition. This concludes our proof. □

Theorem 4.4. (Statement in Section 4) *Let $E : \mathcal{P}_{\leq N}(\Omega) \to \mathbb{R}^m$, where $N \geq 2$ and $\Omega \subseteq \mathbb{R}^d$ has a nonempty interior. Then for all $p \in [1, \infty]$, E is not bi-Lipschitz on $\mathcal{P}_{\leq N}(\Omega)$ with respect to $\mathcal{W}_p$.*

Before proving the theorem, we note that it implies that most practical embeddings of $\mathcal{P}_{\leq N}(\Omega)$ are likely to fail in lower-Lipschitzness, since it is reasonable to expect most such embeddings to be upper Lipschitz. This is formulated in the following corollary.

Corollary B.7. *Under the above assumptions, if $E : \mathcal{P}_{\leq N}(\Omega) \to \mathbb{R}^m$ is upper-Lipschitz with respect to $\mathcal{W}_1$, then it is not lower-Lipschitz with respect to any $\mathcal{W}_p$ with $p \in [1, \infty]$.*

Proof. If E is upper-Lipschitz w.r.t. $\mathcal{W}_1$, then by Theorem 4.4 it is not lower-Lipschitz w.r.t. $\mathcal{W}_1$. Since $\mathcal{W}_p(\mu, \tilde{\mu}) \geq \mathcal{W}_1(\mu, \tilde{\mu})$ for any $p \geq 1$, E is thus not lower-Lipschitz w.r.t. $\mathcal{W}_p$. □

Proof. Our proof of Theorem 4.4 consists of three steps. First, in Lemma B.8 below, we prove the theorem for the special case that E is positively homogeneous and Ω is an open ball centered at zero. Then, in Lemma B.9, we release the homogeneity assumption by considering a homogenized version of E. Finally, we generalize to arbitrary Ω with a nonempty interior in a straightforward manner.

Before we state and prove our results, we define the operation of scalar multiplication of distributions in $\mathcal{P}_{\leq N}(\Omega)$.

Definition. For $\mu = \sum_{i=1}^{N} w_i \delta_{\boldsymbol{x}^{(i)}} \in \mathcal{P}_{\leq N}(\mathbb{R}^d)$ and a scalar $\alpha \in \mathbb{R}$, we define the distribution $\alpha\mu \in \mathcal{P}_{\leq N}(\mathbb{R}^d)$ by

$$\alpha\mu \coloneqq \sum_{i=1}^{N} w_i \delta_{\alpha \boldsymbol{x}^{(i)}}.$$

Note that this definition is *not* the same as the scalar multiplication operation on measures used in Appendix A.1, e.g. in equation (16), which multiplies the measure weights rather than the support points.

Let us begin with the special case of a positively homogeneous E.

Lemma B.8. *Let $E : \mathcal{P}_{\leq N}(\Omega) \to \mathbb{R}^m$, with $\Omega \subseteq \mathbb{R}^d$ being an open ball centered at zero, $N \geq 2$ and $m \geq 1$. Suppose that E is* positively homogeneous*, i.e. $E(\alpha\mu) = \alpha E(\mu)$ for any $\mu \in \mathcal{P}_{\leq N}(\Omega)$, $\alpha \geq 0$. Then for all $p \in [1, \infty]$, E is not bi-Lipschitz with respect to $\mathcal{W}_p$.*

Proof. Let $\{\theta_t\}_{t=1}^{\infty}$ be a sequence of real numbers such that

$$0 < \theta_{t+1} \leq \tfrac{1}{2}\theta_t \leq 1 \quad \forall t \geq 1. \tag{44}$$

By assumption, the set Ω contains a ball $B_r(0)$ of radius $r > 0$ centered at zero. Choose $\boldsymbol{x} \neq 0$ in that ball. For $\theta \in [0, 1]$, we define the distribution

$$\mu(\theta) = (1 - \theta)\delta_0 + \theta\delta_{\boldsymbol{x}},$$

where δ_0 is Dirac's delta function at $0 \in \mathbb{R}^d$. That is, $\mu(\theta)$ assigns weights of $1 - \theta$ and θ to the points 0 and $\boldsymbol{x}$ in $\mathbb{R}^d$ respectively. Note that for $1 \leq p < \infty$

$$\mathcal{W}_p(\mu(\theta_t), \delta_0) = [\theta_t \|\boldsymbol{x}\|^p]^{1/p} = \sqrt[p]{\theta_t}\|\boldsymbol{x}\|$$

This holds for $p = \infty$ too, if we denote $\sqrt[p]{\theta_t} = 1$ in this case. Therefore, for all $t \in \mathbb{N}$,

$$\frac{E(\mu(\theta_t)) - E(\delta_0)}{\mathcal{W}_p(\mu(\theta_t), \delta_0)} = \frac{1}{\|\boldsymbol{x}\|} \frac{E(\mu(\theta_t)) - E(\delta_0)}{\sqrt[p]{\theta_t}} = \frac{1}{\|\boldsymbol{x}\|} \frac{E(\mu(\theta_t))}{\sqrt[p]{\theta_t}}, \tag{45}$$

where for the last equality we used the homogeniety of E, which implies that $E(\delta_0) = 0$.

We can assume that E in upper-Lipschitz, since otherwise there is nothing to prove. Under this assumption, the norm of the expression in (45) is uniformly upper-bounded for all $t \in \mathbb{N}$, which implies that the exists a subsequence of θ_t for which this expression converges. Replacing θ_t with this subsequence, we note that the new subsequence still satisfies (44), and that for an appropriate vector L,

$$\lim_{t \to \infty} \frac{E(\mu(\theta_t))}{\sqrt[p]{\theta_t}} = L$$

Now, consider the sequence of distributions

$$\tilde{\mu}_t \coloneqq \sqrt[p]{\frac{\theta_t}{\theta_{t-1}}}\mu(\theta_{t-1}), \quad t \geq 2.$$

Since $\frac{\theta_t}{\theta_{t-1}} \leq \frac{1}{2}$, and $\boldsymbol{x} \in B_r(0) \subseteq \Omega$, the measure $\tilde{\mu}_t$ indeed belongs to $\mathcal{P}_{\leq N}(\Omega)$. We wish to lower-bound the p-Wasserstein distance from $\mu(\theta_t)$ to $\tilde{\mu}_t$ for $t \geq 2$. Note that both measures split their mass between zero and an additional vector. The measure $\tilde{\mu}_t$ assigns a mass of θ_{t-1} to the nonzero point $\sqrt[p]{\frac{\theta_t}{\theta_{t-1}}}\boldsymbol{x}$, whereas the other measure $\mu(\theta_t)$ assigns a smaller mass of θ_t to a non-zero point $\boldsymbol{x}$. Therefore, transporting $\tilde{\mu}_t$ to $\mu(\theta_t)$ requires transporting at least a mass of $\theta_{t-1} - \theta_t$ from $\sqrt[p]{\frac{\theta_t}{\theta_{t-1}}}\boldsymbol{x}$ to 0, so that for all $1 \leq p < \infty$

$$\begin{aligned}
\mathcal{W}_p^p(\mu(\theta_t), \tilde{\mu}_t) &\geq (\theta_{t-1} - \theta_t)\| \sqrt[p]{\frac{\theta_t}{\theta_{t-1}}}\boldsymbol{x} - 0\|^p \\
&= \theta_t(1 - \frac{\theta_t}{\theta_{t-1}})\|\boldsymbol{x}\|^p \\
&\geq \frac{1}{2}\theta_t\|\boldsymbol{x}\|^p.
\end{aligned}$$

We obtained that

$$\mathcal{W}_p(\mu(\theta_t), \tilde{\mu}_t) \geq \sqrt[p]{\theta_t/2}\|\boldsymbol{x}\| \tag{46}$$

for $p < \infty$, and the same argument can be used to verify that this is the case for $p = \infty$ as well. We conclude that

$$\begin{aligned}
\frac{\|E(\mu(\theta_t)) - E(\tilde{\mu}_t)\|}{\mathcal{W}_p(\mu(\theta_t), \tilde{\mu}_t)} &\overset{\text{(a)}}{\leq} \frac{\sqrt[p]{\frac{1}{\theta_t}}\left\|E(\mu(\theta_t)) - \sqrt[p]{\frac{\theta_t}{\theta_{t-1}}}E(\mu(\theta_{t-1}))\right\|}{\sqrt[p]{1/2}\|\boldsymbol{x}\|} \\
&= \frac{\left\|\sqrt[p]{\frac{1}{\theta_t}}E(\mu(\theta_t)) - \sqrt[p]{\frac{1}{\theta_{t-1}}}E(\mu(\theta_{t-1}))\right\|}{\sqrt[p]{1/2}\|\boldsymbol{x}\|} \to 0
\end{aligned}$$

where (a) is by (46) and the homogeniety of E, and the convergence to zero is because both expressions in the numerator converge to the same limit L. This shows that E is not lower-Lipschitz, which concludes the proof of Lemma B.8. □

The following lemma shows that the homogeneity assumption on E can be released.

Lemma B.9. *Let $E : \mathcal{P}_{\leq N}(\Omega) \to \mathbb{R}^m$, with $\Omega \subseteq \mathbb{R}^d$ being an open ball centered at zero, $N \geq 2$ and $m \geq 1$. Then for all $p \in [1, \infty]$, E is not bi-Lipschitz with respect to $\mathcal{W}_p$.*

Proof. Let $p \in [1, \infty]$ and suppose by contradiction that E is bi-Lipschitz with constants $0 < c \leq C < \infty$,

$$c \cdot \mathcal{W}_p(\mu, \tilde{\mu}) \leq \|E(\mu) - E(\tilde{\mu})\| \leq C \cdot \mathcal{W}_p(\mu, \tilde{\mu}), \qquad \forall \mu, \tilde{\mu} \in \mathcal{P}_{\leq N}(\Omega). \tag{47}$$

We can assume without loss of generality that $E(\delta_0) = 0$, since otherwise, let

$$\tilde{E}(\mu) \coloneqq E(\mu) - E(\delta_0),$$

then E satisfies (47) if and only if $\tilde{E}$ satisfies (47).

We first prove an auxiliary claim.

Claim. *For any $\mu, \tilde{\mu} \in \mathcal{P}_{\leq N}(\Omega)$ with $\|\mu\|_{\mathcal{W}_p} = 1$ and $0 < \|\tilde{\mu}\|_{\mathcal{W}_p} \leq 1$,*

$$\left\|E\left(\frac{\tilde{\mu}}{\|\tilde{\mu}\|_{\mathcal{W}_p}}\right) - E(\tilde{\mu})\right\| \leq C \cdot \left(1 - \|\tilde{\mu}\|_{\mathcal{W}_p}\right) \leq C \cdot \mathcal{W}_p(\mu, \tilde{\mu}). \tag{48}$$

Proof. By (47),

$$\left\|E\left(\frac{\tilde{\mu}}{\|\tilde{\mu}\|_{\mathcal{W}_p}}\right) - E(\tilde{\mu})\right\| \leq C \cdot \mathcal{W}_p\left(\frac{\tilde{\mu}}{\|\tilde{\mu}\|_{\mathcal{W}_p}}, \tilde{\mu}\right).$$

We shall now show that

$$\mathcal{W}_p\left(\frac{\tilde{\mu}}{\|\tilde{\mu}\|_{\mathcal{W}_p}}, \tilde{\mu}\right) \leq 1 - \|\tilde{\mu}\|_{\mathcal{W}_p}.$$

Let $\tilde{\mu} = \sum_{i=1}^N w_i \delta_{\tilde{\boldsymbol{x}}_i}$ be a parametrization of $\tilde{\mu}$. Consider the transport plan $\pi = (\pi_{ij})_{i,j\in[N]}$ from $\tilde{\mu}$ to $\frac{\tilde{\mu}}{\|\tilde{\mu}\|_{\mathcal{W}_p}}$ given by

$$\pi_{ij} = \begin{cases} w_i & i = j \\ 0 & i \neq j. \end{cases}$$

By definition, $\mathcal{W}_p\left(\frac{\tilde{\mu}}{\|\tilde{\mu}\|_{\mathcal{W}_p}}, \tilde{\mu}\right)$ is smaller or equal to the cost of transporting $\tilde{\mu}$ to $\frac{\tilde{\mu}}{\|\tilde{\mu}\|_{\mathcal{W}_p}}$ according to π. Thus, for $p < \infty$,

$$\begin{aligned}
\mathcal{W}_p^p\left(\frac{\tilde{\mu}}{\|\tilde{\mu}\|_{\mathcal{W}_p}}, \tilde{\mu}\right) &\leq \sum_{i=1}^N w_i \left\| \tfrac{1}{\|\tilde{\mu}\|_{\mathcal{W}_p}} \tilde{\boldsymbol{x}}_i - \tilde{\boldsymbol{x}}_i \right\|^p = \sum_{i=1}^N w_i \left\| \left(\tfrac{1}{\|\tilde{\mu}\|_{\mathcal{W}_p}} - 1 \right) \tilde{\boldsymbol{x}}_i \right\|^p \\
&= \left(\tfrac{1}{\|\tilde{\mu}\|_{\mathcal{W}_p}} - 1 \right)^p \sum_{i=1}^N w_i \|\tilde{\boldsymbol{x}}_i\|^p = \left(\tfrac{1}{\|\tilde{\mu}\|_{\mathcal{W}_p}} - 1 \right)^p \|\tilde{\mu}\|_{\mathcal{W}_p}^p \\
&= \left(1 - \|\tilde{\mu}\|_{\mathcal{W}_p}\right)^p,
\end{aligned}$$

and thus

$$\mathcal{W}_p\left(\frac{\tilde{\mu}}{\|\tilde{\mu}\|_{\mathcal{W}_p}}, \tilde{\mu}\right) \leq \left(1 - \|\tilde{\mu}\|_{\mathcal{W}_p}\right).$$

Both sides of the above inequality are continuous in p, including at the limit $p \to \infty$. Thus, the above inequality also holds for $p = \infty$. Now, to show that

$$1 - \|\tilde{\mu}\|_{\mathcal{W}_p} \leq \mathcal{W}_p(\mu, \tilde{\mu}),$$

note that

$$1 - \|\tilde{\mu}\|_{\mathcal{W}_p} = \|\mu\|_{\mathcal{W}_p} - \|\tilde{\mu}\|_{\mathcal{W}_p} = \mathcal{W}_p(\mu, 0) - \mathcal{W}_p(\tilde{\mu}, 0) \leq \mathcal{W}_p(\mu, \tilde{\mu}),$$

where the last inequality is the reverse triangle inequality, since $\mathcal{W}_p(\,\cdot\,,\,\cdot\,)$ is a metric. Thus, (48) holds. □

Now we define the *homogenized* function $\hat{E} : \mathcal{P}_{\leq N}(\Omega) \to \mathbb{R}^{m+1}$ by

$$\begin{cases} \hat{E}(\mu) := \left[\|\mu\|_{\mathcal{W}_p}, \|\mu\|_{\mathcal{W}_p} E\left(\frac{\mu}{\|\mu\|_{\mathcal{W}_p}} \right) \right], & \mu \neq 0 \\ 0 & \mu = 0. \end{cases} \tag{49}$$

Clearly $\hat{E}$ is positively homogeneous. By Lemma B.8, $\hat{E}$ it is not bi-Lipschitz with respect to $\mathcal{W}_p$, and thus there exist two sequences of distributions $\mu_t, \tilde{\mu}_t \in \mathcal{P}_{\leq N}(\Omega)$, $t \geq 1$, such that

$$\frac{\left\| \hat{E}(\mu_t) - \hat{E}(\tilde{\mu}_t) \right\|}{\mathcal{W}_p(\mu_t, \tilde{\mu}_t)} \xrightarrow[t\to\infty]{} L, \tag{50}$$

with $L = 0$ or $L = \infty$. Since $\hat{E}$ is positively homogeneous, we can assume without loss of generality that

$$1 = \|\mu_t\|_{\mathcal{W}_p} \geq \|\tilde{\mu}_t\|_{\mathcal{W}_p} \quad \text{for all } t \geq 1.$$

This can be seen by dividing each μ_t and $\tilde{\mu}_t$ by $\max\left\{ \|\mu_t\|_{\mathcal{W}_p}, \|\tilde{\mu}_t\|_{\mathcal{W}_p} \right\}$ and swapping μ_t and $\tilde{\mu}_t$ for all t for which $\|\mu_t\|_{\mathcal{W}_p} < \|\tilde{\mu}_t\|_{\mathcal{W}_p}$.

If $\tilde{\mu}_t = \delta_0$ for an infinite subset of indices t, then redefine μ_t and $\tilde{\mu}_t$ to be the corresponding subsequences with those indices, and now (50) reads as

$$\frac{\left\| \hat{E}(\mu_t) - \hat{E}(\delta_0) \right\|}{\mathcal{W}_p(\mu_t, \delta_0)} = \frac{\|E(\mu_t) - E(\delta_0)\|}{\mathcal{W}_p(\mu_t, \delta_0)} \xrightarrow[t\to\infty]{} L.$$

This contradicts the bi-Lipschitzness of E. Therefore, $\tilde{\mu}_t = \delta_0$ at most at a finite subset of indices t. By skipping those indices in μ_t and $\tilde{\mu}_t$, we can assume without loss of generality that

$$1 = \|\mu_t\|_{\mathcal{W}_p} \geq \|\tilde{\mu}_t\|_{\mathcal{W}_p} > 0 \quad \text{for all } t \geq 1. \tag{51}$$

Let us first handle the case $L = \infty$. The first component of $\hat{E}(\mu_t) - \hat{E}(\tilde{\mu}_t)$ is bounded by

$$\left| \|\mu_t\|_{\mathcal{W}_p} - \|\tilde{\mu}_t\|_{\mathcal{W}_p} \right| = 1 - \|\tilde{\mu}_t\|_{\mathcal{W}_p} \leq \mathcal{W}_p(\mu_t, \tilde{\mu}_t)$$

according to (48). Therefore, by (50) combined with the fact that $\|\tilde{\mu}_t\|_{\mathcal{W}_p} > 0 \ \forall t$, we must have that

$$\frac{\left\| \|\mu_t\|_{\mathcal{W}_p} E\left(\frac{\mu_t}{\|\mu_t\|_{\mathcal{W}_p}}\right) - \|\tilde{\mu}_t\|_{\mathcal{W}_p} E\left(\frac{\tilde{\mu}_t}{\|\tilde{\mu}_t\|_{\mathcal{W}_p}}\right) \right\|}{\mathcal{W}_p(\mu_t, \tilde{\mu}_t)} \xrightarrow[t \to \infty]{} \infty. \tag{52}$$

On the other hand,

$$\begin{aligned} &\left\| \|\mu_t\|_{\mathcal{W}_p} E\left(\frac{\mu_t}{\|\mu_t\|_{\mathcal{W}_p}}\right) - \|\tilde{\mu}_t\|_{\mathcal{W}_p} E\left(\frac{\tilde{\mu}_t}{\|\tilde{\mu}_t\|_{\mathcal{W}_p}}\right) \right\| \overset{\text{(a)}}{=} \left\| E(\mu_t) - \|\tilde{\mu}_t\|_{\mathcal{W}_p} E\left(\frac{\tilde{\mu}_t}{\|\tilde{\mu}_t\|_{\mathcal{W}_p}}\right) \right\| \\ &\overset{\text{(b)}}{\leq} \|E(\mu_t) - E(\tilde{\mu}_t)\| + \left\| E(\tilde{\mu}_t) - E\left(\frac{\tilde{\mu}_t}{\|\tilde{\mu}_t\|_{\mathcal{W}_p}}\right) \right\| + \left\| E\left(\frac{\tilde{\mu}_t}{\|\tilde{\mu}_t\|_{\mathcal{W}_p}}\right) - \|\tilde{\mu}_t\|_{\mathcal{W}_p} E\left(\frac{\tilde{\mu}_t}{\|\tilde{\mu}_t\|_{\mathcal{W}_p}}\right) \right\|, \end{aligned} \tag{53}$$

where (a) holds since $\|\mu_t\|_{\mathcal{W}_p} = 1$ and (b) is by the triangle inequality. We shall now bound the three above terms.

First,

$$\|E(\mu_t) - E(\tilde{\mu}_t)\| \leq C \cdot \mathcal{W}_p(\mu_t, \tilde{\mu}_t) \tag{54}$$

by (47). Second,

$$\left\| E(\tilde{\mu}_t) - E\left(\frac{\tilde{\mu}_t}{\|\tilde{\mu}_t\|_{\mathcal{W}_p}}\right) \right\| \leq C \cdot \mathcal{W}_p(\mu_t, \tilde{\mu}_t) \tag{55}$$

by (48). Lastly,

$$\begin{aligned} \left\| E\left(\frac{\tilde{\mu}_t}{\|\tilde{\mu}_t\|_{\mathcal{W}_p}}\right) - \|\tilde{\mu}_t\|_{\mathcal{W}_p} E\left(\frac{\tilde{\mu}_t}{\|\tilde{\mu}_t\|_{\mathcal{W}_p}}\right) \right\| &= \left(1 - \|\tilde{\mu}_t\|_{\mathcal{W}_p}\right) \cdot \left\| E\left(\frac{\tilde{\mu}_t}{\|\tilde{\mu}_t\|_{\mathcal{W}_p}}\right) - 0 \right\| \\ &= \left(1 - \|\tilde{\mu}_t\|_{\mathcal{W}_p}\right) \cdot \left\| E\left(\frac{\tilde{\mu}_t}{\|\tilde{\mu}_t\|_{\mathcal{W}_p}}\right) - E(\delta_0) \right\| \\ &\overset{\text{(a)}}{\leq} \left(1 - \|\tilde{\mu}_t\|_{\mathcal{W}_p}\right) \cdot C \cdot \mathcal{W}_p\left(\frac{\tilde{\mu}_t}{\|\tilde{\mu}_t\|_{\mathcal{W}_p}}, \delta_0\right) \\ &= \left(1 - \|\tilde{\mu}_t\|_{\mathcal{W}_p}\right) \cdot C \cdot \left\| \frac{\tilde{\mu}_t}{\|\tilde{\mu}_t\|_{\mathcal{W}_p}} \right\|_{\mathcal{W}_p} \\ &= C \cdot \left(1 - \|\tilde{\mu}_t\|_{\mathcal{W}_p}\right) \overset{\text{(b)}}{\leq} C \cdot \mathcal{W}_p(\mu_t, \tilde{\mu}_t), \end{aligned} \tag{56}$$

where (a) is by (47) and (b) is by (48). Inserting (54)-(56) into (53) yields

$$\left\| \|\mu_t\|_{\mathcal{W}_p} E\left(\frac{\mu_t}{\|\mu_t\|_{\mathcal{W}_p}}\right) - \|\tilde{\mu}_t\|_{\mathcal{W}_p} E\left(\frac{\tilde{\mu}_t}{\|\tilde{\mu}_t\|_{\mathcal{W}_p}}\right) \right\| \leq 3C \cdot \mathcal{W}_p(\mu_t, \tilde{\mu}_t),$$

which contradicts (52).

Let us now handle the case $L = 0$. For two sequences of numbers $a_t, b_t \in \mathbb{R}$, $t \geq 1$, we say that

$$a_t = o(b_t)$$

if

$$\lim_{t\to\infty} \frac{a_t}{b_t} = 0.$$

Denote

$$d_t \coloneqq \mathcal{W}_p(\mu_t, \tilde{\mu}_t).$$

According to (50) with $L = 0$, the first component of $\hat{E}(\mu_t) - \hat{E}(\tilde{\mu}_t)$, which equals $\|\mu_t\|_{\mathcal{W}_p} - \|\tilde{\mu}_t\|_{\mathcal{W}_p}$, satisfies

$$\frac{\left|\|\mu_t\|_{\mathcal{W}_p} - \|\tilde{\mu}_t\|_{\mathcal{W}_p}\right|}{\mathcal{W}_p(\mu_t, \tilde{\mu}_t)} \xrightarrow[t\to\infty]{} 0,$$

and thus

$$1 - \|\tilde{\mu}_t\|_{\mathcal{W}_p} = \left|\|\mu_t\|_{\mathcal{W}_p} - \|\tilde{\mu}_t\|_{\mathcal{W}_p}\right| = o(d_t). \tag{57}$$

By the triangle inequality,

$$\|E(\mu_t) - E(\tilde{\mu}_t)\| \leq \tag{58}$$

$$\left\|E(\mu_t) - \|\tilde{\mu}_t\|_{\mathcal{W}_p} E\left(\frac{\tilde{\mu}_t}{\|\tilde{\mu}_t\|_{\mathcal{W}_p}}\right)\right\| + \left\|\|\tilde{\mu}_t\|_{\mathcal{W}_p} E\left(\frac{\tilde{\mu}_t}{\|\tilde{\mu}_t\|_{\mathcal{W}_p}}\right) - E\left(\frac{\tilde{\mu}_t}{\|\tilde{\mu}_t\|_{\mathcal{W}_p}}\right)\right\| + \left\|E\left(\frac{\tilde{\mu}_t}{\|\tilde{\mu}_t\|_{\mathcal{W}_p}}\right) - E(\tilde{\mu}_t)\right\|. \tag{59}$$

We shall show that each of the three above terms is $o(d_t)$.

First, since $\|\mu_t\|_{\mathcal{W}_p} = 1$,

$$\begin{aligned}\left\|E(\mu_t) - \|\tilde{\mu}_t\|_{\mathcal{W}_p} E\left(\frac{\tilde{\mu}_t}{\|\tilde{\mu}_t\|_{\mathcal{W}_p}}\right)\right\| &= \left\|\|\mu_t\|_{\mathcal{W}_p} E\left(\frac{\mu_t}{\|\mu_t\|_{\mathcal{W}_p}}\right) - \|\tilde{\mu}_t\|_{\mathcal{W}_p} E\left(\frac{\tilde{\mu}_t}{\|\tilde{\mu}_t\|_{\mathcal{W}_p}}\right)\right\| \\ &= \left\|\hat{E}(\mu_t) - \hat{E}(\tilde{\mu}_t)\right\|,\end{aligned} \tag{60}$$

which is $o(d_t)$ by (50). For the second term,

$$\begin{aligned}&\left\|\|\tilde{\mu}_t\|_{\mathcal{W}_p} E\left(\frac{\tilde{\mu}_t}{\|\tilde{\mu}_t\|_{\mathcal{W}_p}}\right) - E\left(\frac{\tilde{\mu}_t}{\|\tilde{\mu}_t\|_{\mathcal{W}_p}}\right)\right\| \\ &= \left(1 - \|\tilde{\mu}_t\|_{\mathcal{W}_p}\right) \cdot \left\|E\left(\frac{\tilde{\mu}_t}{\|\tilde{\mu}_t\|_{\mathcal{W}_p}}\right)\right\| \\ &= \left(1 - \|\tilde{\mu}_t\|_{\mathcal{W}_p}\right) \cdot \left\|E\left(\frac{\tilde{\mu}_t}{\|\tilde{\mu}_t\|_{\mathcal{W}_p}}\right) - E(\delta_0)\right\| \\ &\overset{\text{(a)}}{\leq} \left(1 - \|\tilde{\mu}_t\|_{\mathcal{W}_p}\right) \cdot C \cdot \mathcal{W}\left(\frac{\tilde{\mu}_t}{\|\tilde{\mu}_t\|_{\mathcal{W}_p}}, \delta_0\right) \\ &= \left(1 - \|\tilde{\mu}_t\|_{\mathcal{W}_p}\right) \cdot C \cdot \left\|\frac{\tilde{\mu}_t}{\|\tilde{\mu}_t\|_{\mathcal{W}_p}}\right\|_{\mathcal{W}_p} \\ &= \left(1 - \|\tilde{\mu}_t\|_{\mathcal{W}_p}\right) C \overset{\text{(b)}}{=} o(d_t),\end{aligned} \tag{61}$$

where (a) is by (47) and (b) is by (57).

Finally, by (48),

$$\left\|E\left(\frac{\tilde{\mu}_t}{\|\tilde{\mu}_t\|_{\mathcal{W}_p}}\right) - E(\tilde{\mu}_t)\right\| \leq C \cdot \left(1 - \|\tilde{\mu}_t\|_{\mathcal{W}_p}\right) = o(d_t). \tag{62}$$

Therefore, by (60)-(62) and (58), we have that

$$\|E(\mu_t) - E(\tilde{\mu}_t)\| = o(d_t),$$

and thus E is not lower-Lipschitz. This concludes the proof of Lemma B.9. □

To finish the proof of Theorem 4.4, suppose that $\Omega \subseteq \mathbb{R}^d$ is an arbitrary set with a nonempty interior. Let $\Omega_0 \subseteq \Omega$ be an open ball contained in Ω, and let $\boldsymbol{x}_0$ be the center of Ω_0. Then $\Omega_0 - \boldsymbol{x}_0$ is an open ball centered at zero.

Given $E : \mathcal{P}_{\leq N}(\Omega) \to \mathbb{R}^m$, define $\tilde{E} : \mathcal{P}_{\leq N}(\Omega_0 - \boldsymbol{x}_0) \to \mathbb{R}^m$ by

$$\tilde{E}(\mu) \coloneqq E(\mu + \boldsymbol{x}_0).$$

Then $\tilde{E}$ satisfies the assumptions of Lemma B.9, and thus there exist two sequences of distributions $\mu_t, \tilde{\mu}_t \in \mathcal{P}_{\leq N}(\Omega_0 - \boldsymbol{x}_0)$, $t \geq 1$ such that

$$\frac{\left\|\tilde{E}(\mu_t) - \tilde{E}(\tilde{\mu}_t)\right\|}{\mathcal{W}_p(\mu_t, \tilde{\mu}_1)} \xrightarrow[t\to\infty]{} L,$$

with $L = 0$ or $L = \infty$. Note that the sequences of shifted distributions $\{\mu_t + \boldsymbol{x}_0\}_{t\geq 1}$ and $\{\tilde{\mu}_t + \boldsymbol{x}_0\}_{t\geq 1}$ are in $\mathcal{P}_{\leq N}(\Omega_0)$ and thus in $\mathcal{P}_{\leq N}(\Omega)$. Since

$$\mathcal{W}_p(\mu_t + \boldsymbol{x}_0, \tilde{\mu}_1 + \boldsymbol{x}_0) = \mathcal{W}_p(\mu_t, \tilde{\mu}_1),$$

we have that

$$\begin{aligned}\frac{\|E(\mu_t + \boldsymbol{x}_0) - E(\tilde{\mu}_t + \boldsymbol{x}_0)\|}{\mathcal{W}_p(\mu_t + \boldsymbol{x}_0, \tilde{\mu}_1 + \boldsymbol{x}_0)} &= \frac{\|E(\mu_t + \boldsymbol{x}_0) - E(\tilde{\mu}_t + \boldsymbol{x}_0)\|}{\mathcal{W}_p(\mu_t, \tilde{\mu}_1)} \\ &= \frac{\left\|\tilde{E}(\mu_t) - \tilde{E}(\tilde{\mu}_t)\right\|}{\mathcal{W}_p(\mu_t, \tilde{\mu}_1)} \xrightarrow[t\to\infty]{} L,\end{aligned}$$

which implies that E is not bi-Lipschitz on $\mathcal{P}_{\leq N}(\Omega_0)$, and thus not on $\mathcal{P}_{\leq N}(\Omega)$. □

Theorem 4.2. (Statement in Section 4) *Let $E : \mathcal{P}_{\leq N}(\mathbb{R}^d) \to \mathbb{R}^m$ be injective and positively homogeneous. Let $\boldsymbol{w}, \tilde{\boldsymbol{w}} \in \Delta^N$ be fixed, and suppose that the functions $E(\boldsymbol{X}, \boldsymbol{w})$ and $E(\boldsymbol{X}, \tilde{\boldsymbol{w}})$ are piecewise linear in $\boldsymbol{X}$. Then there exist $c, C > 0$ such that for all $\boldsymbol{X}, \tilde{\boldsymbol{X}} \in \mathbb{R}^{d\times N}$ and $p \in [1, \infty]$,*

$$c \cdot \mathcal{W}_p\Big((\boldsymbol{X}, \boldsymbol{w}), \Big(\tilde{\boldsymbol{X}}, \tilde{\boldsymbol{w}}\Big)\Big) \leq \Big\|E(\boldsymbol{X}, \boldsymbol{w}) - E\Big(\tilde{\boldsymbol{X}}, \tilde{\boldsymbol{w}}\Big)\Big\| \leq C \cdot \mathcal{W}_p\Big((\boldsymbol{X}, \boldsymbol{w}), \Big(\tilde{\boldsymbol{X}}, \tilde{\boldsymbol{w}}\Big)\Big). \tag{12}$$

Proof. The proof is outlined as follows: First we show that there exist constants $\tilde{c}, \tilde{C} > 0$ for which (12) holds in the special case $p = 1$. Then we show that for any fixed $\boldsymbol{w}, \tilde{\boldsymbol{w}} \in \Delta^N$ there exists a constant $\beta > 0$ such that for all $\boldsymbol{X}, \tilde{\boldsymbol{X}} \in \mathbb{R}^{d\times N}$,

$$\mathcal{W}_1\Big((\boldsymbol{X}, \boldsymbol{w}), \Big(\tilde{\boldsymbol{X}}, \tilde{\boldsymbol{w}}\Big)\Big) \geq \beta \cdot \mathcal{W}_\infty\Big((\boldsymbol{X}, \boldsymbol{w}), \Big(\tilde{\boldsymbol{X}}, \tilde{\boldsymbol{w}}\Big)\Big). \tag{63}$$

This will imply that for the given pair $\boldsymbol{w}, \tilde{\boldsymbol{w}}$, (12) holds with the constants $c = \beta\tilde{c}$ and $C = \tilde{C}$ for all $p \in [1, \infty]$, since

$$\mathcal{W}_1\Big((\boldsymbol{X}, \boldsymbol{w}), \Big(\tilde{\boldsymbol{X}}, \tilde{\boldsymbol{w}}\Big)\Big) \leq \mathcal{W}_p\Big((\boldsymbol{X}, \boldsymbol{w}), \Big(\tilde{\boldsymbol{X}}, \tilde{\boldsymbol{w}}\Big)\Big) \leq \mathcal{W}_\infty\Big((\boldsymbol{X}, \boldsymbol{w}), \Big(\tilde{\boldsymbol{X}}, \tilde{\boldsymbol{w}}\Big)\Big).$$

Let us begin by proving that (12) holds for $p = 1$. The 1-Wasserstein distance between two distributions parametrized by $(\boldsymbol{X}, \boldsymbol{w})$ and $\Big(\tilde{\boldsymbol{X}}, \tilde{\boldsymbol{w}}\Big)$ can be expressed by

$$\mathcal{W}_1\Big((\boldsymbol{X}, \boldsymbol{w}), \Big(\tilde{\boldsymbol{X}}, \tilde{\boldsymbol{w}}\Big)\Big) = \min_{\pi\in\Pi(\boldsymbol{w},\tilde{\boldsymbol{w}})} \sum_{i,j\in[N]} \pi_{ij} \Big\|\boldsymbol{x}^{(i)} - \tilde{\boldsymbol{x}}^{(j)}\Big\|, \tag{64}$$

where the set $\Pi(\boldsymbol{w}, \tilde{\boldsymbol{w}})$ of admissible transport plans from $(\boldsymbol{X}, \boldsymbol{w})$ to $\Big(\tilde{\boldsymbol{X}}, \tilde{\boldsymbol{w}}\Big)$ is given by

$$\Pi(\boldsymbol{w}, \tilde{\boldsymbol{w}}) = \left\{\pi \in [0,1]^{N\times N} \;\middle|\; \forall i \in [N] \sum_{j=1}^N \pi_{ij} = w_i \bigwedge \forall j \in [N] \sum_{i=1}^N \pi_{ij} = \tilde{w}_j\right\}.$$

In particular, $\Pi(\boldsymbol{w}, \tilde{\boldsymbol{w}})$ depends only on $\boldsymbol{w}$ and $\tilde{\boldsymbol{w}}$ and not on the points $\boldsymbol{X}, \tilde{\boldsymbol{X}}$.

Let $\widetilde{\mathcal{W}}_1$ be a modified 1-Wasserstein distance that uses the ℓ_1-norm rather than ℓ_2 as its basic cost function:

$$\widetilde{\mathcal{W}}_1\Big((\boldsymbol{X},\boldsymbol{w}),\Big(\tilde{\boldsymbol{X}},\tilde{\boldsymbol{w}}\Big)\Big) \coloneqq \min_{\pi\in\Pi(\boldsymbol{w},\tilde{\boldsymbol{w}})} \sum_{i,j\in[N]} \pi_{ij}\Big\|\boldsymbol{x}^{(i)}-\tilde{\boldsymbol{x}}^{(j)}\Big\|_1. \tag{65}$$

Note that since

$$\|\boldsymbol{x}\|_2 \le \|\boldsymbol{x}\|_1 \le \sqrt{d}\|\boldsymbol{x}\|_2 \quad \forall \boldsymbol{x}\in\mathbb{R}^d, \tag{66}$$

we have

$$\mathcal{W}_1\Big((\boldsymbol{X},\boldsymbol{w}),\Big(\tilde{\boldsymbol{X}},\tilde{\boldsymbol{w}}\Big)\Big) \le \widetilde{\mathcal{W}}_1\Big((\boldsymbol{X},\boldsymbol{w}),\Big(\tilde{\boldsymbol{X}},\tilde{\boldsymbol{w}}\Big)\Big) \le \sqrt{d}\cdot\mathcal{W}_1\Big((\boldsymbol{X},\boldsymbol{w}),\Big(\tilde{\boldsymbol{X}},\tilde{\boldsymbol{w}}\Big)\Big). \tag{67}$$

Let $f:\mathbb{R}^{d\times N}\times\mathbb{R}^{d\times N}\to\mathbb{R}^2$ be the function given by

$$f\Big(\boldsymbol{X},\tilde{\boldsymbol{X}}\Big) \coloneqq \begin{bmatrix} \Big\|E(\boldsymbol{X},\boldsymbol{w})-E\Big(\tilde{\boldsymbol{X}},\tilde{\boldsymbol{w}}\Big)\Big\|_1 \\ \widetilde{\mathcal{W}}_1\Big((\boldsymbol{X},\boldsymbol{w}),\Big(\tilde{\boldsymbol{X}},\tilde{\boldsymbol{w}}\Big)\Big) \end{bmatrix}.$$

To achieve the desired result, we first show that f is piecewise linear in $\Big(\boldsymbol{X},\tilde{\boldsymbol{X}}\Big)$. The first component of f, $\Big\|E(\boldsymbol{X},\boldsymbol{w})-E\Big(\tilde{\boldsymbol{X}},\tilde{\boldsymbol{w}}\Big)\Big\|_1$, is clearly piecewise linear, as it is the composition of the ℓ_1-norm with a piecewise-linear function. We shall now show that the second component $\widetilde{\mathcal{W}}_1\Big((\boldsymbol{X},\boldsymbol{w}),\Big(\tilde{\boldsymbol{X}},\tilde{\boldsymbol{w}}\Big)\Big)$ is also piecewise linear. For any fixed $\boldsymbol{X}$ and $\tilde{\boldsymbol{X}}$, the optimization problem in (65) is a linear program in π, with the set of feasible solutions being the compact polytope $\Pi(\boldsymbol{w},\tilde{\boldsymbol{w}})$[7]. Thus, the optimal solution must be attained at one of the vertices of $\Pi(\boldsymbol{w},\tilde{\boldsymbol{w}})$. As any polytope has a finite number of vertices[8], let $\pi^{(1)},\dots,\pi^{(K)}$ be the vertices of $\Pi(\boldsymbol{w},\tilde{\boldsymbol{w}})$, and recall that these vertices do not depend on $\Big(\boldsymbol{X},\tilde{\boldsymbol{X}}\Big)$. Therefore, (65) can be reformulated as

$$\widetilde{\mathcal{W}}_1\Big((\boldsymbol{X},\boldsymbol{w}),\Big(\tilde{\boldsymbol{X}},\tilde{\boldsymbol{w}}\Big)\Big) = \min_{k\in[K]} \sum_{i,j\in[N]} \pi^{(k)}_{ij}\Big\|\boldsymbol{x}^{(i)}-\tilde{\boldsymbol{x}}^{(j)}\Big\|_1. \tag{68}$$

From (68) it can be seen that $\widetilde{\mathcal{W}}_1\Big((\boldsymbol{X},\boldsymbol{w}),\Big(\tilde{\boldsymbol{X}},\tilde{\boldsymbol{w}}\Big)\Big)$ is piecewise linear in $\Big(\boldsymbol{X},\tilde{\boldsymbol{X}}\Big)$, as it is the minimum of a finite number of piecewise-linear functions. Since the concatenation of piecewise-linear functions is also piecewise linear, we have that $f\Big(\boldsymbol{X},\tilde{\boldsymbol{X}}\Big)$ is piecewise linear.

Now, let $A\subseteq\mathbb{R}^2$ be the image of f:

$$A \coloneqq \Big\{ f\Big(\boldsymbol{X},\tilde{\boldsymbol{X}}\Big) \;\Big|\; \boldsymbol{X},\tilde{\boldsymbol{X}}\in\mathbb{R}^{d\times N}\Big\}.$$

Since f is piecewise linear, it maps the space $\mathbb{R}^{d\times N}\times\mathbb{R}^{d\times N}$ to a finite union of closed polytopes (some of which may be unbounded). Hence, A is a finite union of closed sets, and thus is closed.

Now we show that the points $(0,1)$ and $(1,0)$ do not belong to A. If $(0,1)\in A$, then there exist $\boldsymbol{X},\tilde{\boldsymbol{X}}$ such that $E(\boldsymbol{X},\boldsymbol{w}) = E\Big(\tilde{\boldsymbol{X}},\tilde{\boldsymbol{w}}\Big)$ and $\widetilde{\mathcal{W}}_1\Big((\boldsymbol{X},\boldsymbol{w}),\Big(\tilde{\boldsymbol{X}},\tilde{\boldsymbol{w}}\Big)\Big) = 1$, which contradicts the injectivity of E. Similarly, if $(1,0)\in A$, then there exist $\boldsymbol{X},\tilde{\boldsymbol{X}}$ such that on one hand $\widetilde{\mathcal{W}}_1\Big((\boldsymbol{X},\boldsymbol{w}),\Big(\tilde{\boldsymbol{X}},\tilde{\boldsymbol{w}}\Big)\Big) = 0$, which implies that $(\boldsymbol{X},\boldsymbol{w})$ and $\Big(\tilde{\boldsymbol{X}},\tilde{\boldsymbol{w}}\Big)$ represent the same distribution, but on the other hand $E(\boldsymbol{X},\boldsymbol{w}) \ne E\Big(\tilde{\boldsymbol{X}},\tilde{\boldsymbol{w}}\Big)$. This contradicts the assumption that E depends only on the input distribution and not on its particular representation.

Let α be the ℓ_2-distance between the compact set $\{(0,1),(1,0)\}$ and the closed set A. As the distance between a compact and a closed set is always attained, we have that $\alpha>0$, otherwise, $\{(0,1),(1,0)\}$ and A would intersect.

[7] Here we denote by *polytope* any finite intersection of closed half-spaces.

[8] See (Grünbaum, 2003), Theorem 3, page 32, and the definition of polyhedral sets on page 26 therein.

Now, let $\boldsymbol{X}, \tilde{\boldsymbol{X}} \in \mathbb{R}^{d\times N}$ such that $\mathcal{W}_1\Big((\boldsymbol{X},\boldsymbol{w}),\Big(\tilde{\boldsymbol{X}},\tilde{\boldsymbol{w}}\Big)\Big) > 0$. Then by (67), $\widetilde{\mathcal{W}}_1\Big((\boldsymbol{X},\boldsymbol{w}),\Big(\tilde{\boldsymbol{X}},\tilde{\boldsymbol{w}}\Big)\Big) > 0$. Denote

$$\nu := \Big[\widetilde{\mathcal{W}}_1\Big((\boldsymbol{X},\boldsymbol{w}),\Big(\tilde{\boldsymbol{X}},\tilde{\boldsymbol{w}}\Big)\Big)\Big]^{-1}.$$

Then

$$\widetilde{\mathcal{W}}_1\Big((\nu\boldsymbol{X},\boldsymbol{w}),\Big(\nu\tilde{\boldsymbol{X}},\tilde{\boldsymbol{w}}\Big)\Big) = 1,$$

and since E and $\widetilde{\mathcal{W}}_1$ are homogeneous, we have

$$\begin{aligned}
&\frac{\Big\|E(\boldsymbol{X},\boldsymbol{w}) - E\Big(\tilde{\boldsymbol{X}},\tilde{\boldsymbol{w}}\Big)\Big\|_1}{\widetilde{\mathcal{W}}_1\Big((\boldsymbol{X},\boldsymbol{w}),\Big(\tilde{\boldsymbol{X}},\tilde{\boldsymbol{w}}\Big)\Big)} = \frac{\Big\|E(\nu\boldsymbol{X},\boldsymbol{w}) - E\Big(\nu\tilde{\boldsymbol{X}},\tilde{\boldsymbol{w}}\Big)\Big\|_1}{\widetilde{\mathcal{W}}_1\Big((\nu\boldsymbol{X},\boldsymbol{w}),\Big(\nu\tilde{\boldsymbol{X}},\tilde{\boldsymbol{w}}\Big)\Big)} = \Big\|E(\nu\boldsymbol{X},\boldsymbol{w}) - E\Big(\nu\tilde{\boldsymbol{X}},\tilde{\boldsymbol{w}}\Big)\Big\|_1 \\
&= \left\| \begin{bmatrix} \Big\|E(\nu\boldsymbol{X},\boldsymbol{w}) - E\Big(\nu\tilde{\boldsymbol{X}},\tilde{\boldsymbol{w}}\Big)\Big\|_1 \\ 1 \end{bmatrix} - \begin{bmatrix} 0 \\ 1 \end{bmatrix} \right\|_2 \\
&= \left\| \begin{bmatrix} \Big\|E(\nu\boldsymbol{X},\boldsymbol{w}) - E\Big(\nu\tilde{\boldsymbol{X}},\tilde{\boldsymbol{w}}\Big)\Big\|_1 \\ \widetilde{\mathcal{W}}_1\Big((\nu\boldsymbol{X},\boldsymbol{w}),\Big(\nu\tilde{\boldsymbol{X}},\tilde{\boldsymbol{w}}\Big)\Big) \end{bmatrix} - \begin{bmatrix} 0 \\ 1 \end{bmatrix} \right\|_2 \\
&= \left\| f\Big(\nu\boldsymbol{X},\nu\tilde{\boldsymbol{X}}\Big) - \begin{bmatrix} 0 \\ 1 \end{bmatrix} \right\|_2 \geq \mathrm{dist}(A, \{(0,1),(1,0)\}) = \alpha.
\end{aligned} \tag{69}$$

Therefore,

$$\begin{aligned}
\frac{\Big\|E(\boldsymbol{X},\boldsymbol{w}) - E\Big(\tilde{\boldsymbol{X}},\tilde{\boldsymbol{w}}\Big)\Big\|_2}{\mathcal{W}_1\Big((\boldsymbol{X},\boldsymbol{w}),\Big(\tilde{\boldsymbol{X}},\tilde{\boldsymbol{w}}\Big)\Big)} &\overset{\text{(a)}}{\geq} \tfrac{1}{\sqrt{m}} \frac{\Big\|E(\boldsymbol{X},\boldsymbol{w}) - E\Big(\tilde{\boldsymbol{X}},\tilde{\boldsymbol{w}}\Big)\Big\|_1}{\mathcal{W}_1\Big((\boldsymbol{X},\boldsymbol{w}),\Big(\tilde{\boldsymbol{X}},\tilde{\boldsymbol{w}}\Big)\Big)} \\
&\overset{\text{(b)}}{\geq} \tfrac{1}{\sqrt{m}} \frac{\Big\|E(\boldsymbol{X},\boldsymbol{w}) - E\Big(\tilde{\boldsymbol{X}},\tilde{\boldsymbol{w}}\Big)\Big\|_1}{\widetilde{\mathcal{W}}_1\Big((\boldsymbol{X},\boldsymbol{w}),\Big(\tilde{\boldsymbol{X}},\tilde{\boldsymbol{w}}\Big)\Big)} \geq \tfrac{\alpha}{\sqrt{m}},
\end{aligned} \tag{70}$$

where (a) is by the $\ell_1 - \ell_2$ norm inequality over $\mathbb{R}^m$, (b) is by (67), and (c) is by (69).

We now prove a converse bound using a similar argument. Since $\mathcal{W}_1\Big((\boldsymbol{X},\boldsymbol{w}),\Big(\tilde{\boldsymbol{X}},\tilde{\boldsymbol{w}}\Big)\Big) > 0$ and E is injective, $E(\boldsymbol{X},\boldsymbol{w}) \neq E\Big(\tilde{\boldsymbol{X}},\tilde{\boldsymbol{w}}\Big)$. Redefine ν to be

$$\nu := \Big\|E(\boldsymbol{X},\boldsymbol{w}) - E\Big(\tilde{\boldsymbol{X}},\tilde{\boldsymbol{w}}\Big)\Big\|_1^{-1}.$$

Since E is homogeneous,

$$\Big\|E(\nu\boldsymbol{X},\boldsymbol{w}) - E\Big(\nu\tilde{\boldsymbol{X}},\tilde{\boldsymbol{w}}\Big)\Big\|_1 = 1$$

and thus

$$\begin{aligned}
&\frac{\widetilde{\mathcal{W}}_1\Big((\boldsymbol{X},\boldsymbol{w}),\Big(\tilde{\boldsymbol{X}},\tilde{\boldsymbol{w}}\Big)\Big)}{\Big\|E(\boldsymbol{X},\boldsymbol{w}) - E\Big(\tilde{\boldsymbol{X}},\tilde{\boldsymbol{w}}\Big)\Big\|_1} = \frac{\widetilde{\mathcal{W}}_1\Big((\nu\boldsymbol{X},\boldsymbol{w}),\Big(\nu\tilde{\boldsymbol{X}},\tilde{\boldsymbol{w}}\Big)\Big)}{\Big\|E(\nu\boldsymbol{X},\boldsymbol{w}) - E\Big(\nu\tilde{\boldsymbol{X}},\tilde{\boldsymbol{w}}\Big)\Big\|_1} = \widetilde{\mathcal{W}}_1\Big((\nu\boldsymbol{X},\boldsymbol{w}),\Big(\nu\tilde{\boldsymbol{X}},\tilde{\boldsymbol{w}}\Big)\Big) \\
&= \left\| \begin{bmatrix} 1 \\ \widetilde{\mathcal{W}}_1\Big((\nu\boldsymbol{X},\boldsymbol{w}),\Big(\nu\tilde{\boldsymbol{X}},\tilde{\boldsymbol{w}}\Big)\Big) \end{bmatrix} - \begin{bmatrix} 1 \\ 0 \end{bmatrix} \right\|_2 \\
&= \left\| \begin{bmatrix} \Big\|E(\nu\boldsymbol{X},\boldsymbol{w}) - E\Big(\nu\tilde{\boldsymbol{X}},\tilde{\boldsymbol{w}}\Big)\Big\|_1 \\ \widetilde{\mathcal{W}}_1\Big((\nu\boldsymbol{X},\boldsymbol{w}),\Big(\nu\tilde{\boldsymbol{X}},\tilde{\boldsymbol{w}}\Big)\Big) \end{bmatrix} - \begin{bmatrix} 1 \\ 0 \end{bmatrix} \right\|_2 \\
&= \left\| f\Big(\nu\boldsymbol{X},\nu\tilde{\boldsymbol{X}}\Big) - \begin{bmatrix} 1 \\ 0 \end{bmatrix} \right\|_2 \geq \mathrm{dist}(A, \{(0,1),(1,0)\}) = \alpha.
\end{aligned} \tag{71}$$

Therefore,

$$\frac{\left\|E(\boldsymbol{X},\boldsymbol{w})-E\left(\tilde{\boldsymbol{X}},\tilde{\boldsymbol{w}}\right)\right\|_2}{\mathcal{W}_1\left((\boldsymbol{X},\boldsymbol{w}),\left(\tilde{\boldsymbol{X}},\tilde{\boldsymbol{w}}\right)\right)} \overset{\text{(a)}}{\leq} \frac{\left\|E(\boldsymbol{X},\boldsymbol{w})-E\left(\tilde{\boldsymbol{X}},\tilde{\boldsymbol{w}}\right)\right\|_1}{\mathcal{W}_1\left((\boldsymbol{X},\boldsymbol{w}),\left(\tilde{\boldsymbol{X}},\tilde{\boldsymbol{w}}\right)\right)} \overset{\text{(b)}}{\leq} \sqrt{d}\frac{\left\|E(\boldsymbol{X},\boldsymbol{w})-E\left(\tilde{\boldsymbol{X}},\tilde{\boldsymbol{w}}\right)\right\|_1}{\widetilde{\mathcal{W}}_1\left((\boldsymbol{X},\boldsymbol{w}),\left(\tilde{\boldsymbol{X}},\tilde{\boldsymbol{w}}\right)\right)} \overset{\text{(c)}}{\leq} \tfrac{\sqrt{d}}{\alpha}, \tag{72}$$

where (a) is since $\|\cdot\|_2 \leq \|\cdot\|_1$, (b) is by (67), and (c) is by (71). Hence, from (70) and (72), we have

$$\tfrac{\alpha}{\sqrt{m}} \leq \frac{\left\|E(\boldsymbol{X},\boldsymbol{w})-E\left(\tilde{\boldsymbol{X}},\tilde{\boldsymbol{w}}\right)\right\|_2}{\mathcal{W}_1\left((\boldsymbol{X},\boldsymbol{w}),\left(\tilde{\boldsymbol{X}},\tilde{\boldsymbol{w}}\right)\right)} \leq \tfrac{\sqrt{d}}{\alpha}. \tag{73}$$

Thus, (12) holds for the case $p=1$ with the constants $c=\frac{\alpha}{\sqrt{m}}$, $C=\frac{\sqrt{d}}{\alpha}$.

To finish the proof, it is left to show that (63) holds with some constant $\beta>0$ assuming that $\boldsymbol{w}$ and $\tilde{\boldsymbol{w}}$ are constant. To this end, define the sets $I_k \subseteq [N]^2$ for $k\in[K]$,

$$I_k \coloneqq \left\{(i,j)\in[N]^2 \;\middle|\; \pi_{ij}^{(k)}>0\right\},$$

and let

$$\delta_k \coloneqq \min_{(i,j)\in I_k} \pi_{ij}^{(k)}, \quad k\in[K].$$

By definition, $\delta_k>0$ for all $k\in[K]$. Let

$$\delta_{\min} \coloneqq \min_{k\in[K]} \delta_k > 0.$$

Therefore,

$$\begin{aligned}
&\sqrt{d}\cdot\mathcal{W}_1\left((\boldsymbol{X},\boldsymbol{w}),\left(\tilde{\boldsymbol{X}},\tilde{\boldsymbol{w}}\right)\right) \overset{\text{(a)}}{\geq} \widetilde{\mathcal{W}}_1\left((\boldsymbol{X},\boldsymbol{w}),\left(\tilde{\boldsymbol{X}},\tilde{\boldsymbol{w}}\right)\right)\\
&\overset{\text{(b)}}{=} \min_{k\in[K]} \sum_{i,j\in[N]} \pi_{ij}^{(k)}\left\|\boldsymbol{x}^{(i)}-\tilde{\boldsymbol{x}}^{(j)}\right\|_1 \overset{\text{(c)}}{\geq} \min_{k\in[K]} \sum_{i,j\in[N]} \pi_{ij}^{(k)}\left\|\boldsymbol{x}^{(i)}-\tilde{\boldsymbol{x}}^{(j)}\right\|_2\\
&\overset{\text{(d)}}{=} \min_{k\in[K]} \sum_{(i,j)\in I_k} \pi_{ij}^{(k)}\left\|\boldsymbol{x}^{(i)}-\tilde{\boldsymbol{x}}^{(j)}\right\|_2 \overset{\text{(e)}}{\geq} \min_{k\in[K]} \sum_{(i,j)\in I_k} \delta_k\left\|\boldsymbol{x}^{(i)}-\tilde{\boldsymbol{x}}^{(j)}\right\|_2\\
&\overset{\text{(f)}}{\geq} \min_{k\in[K]} \sum_{(i,j)\in I_k} \delta\left\|\boldsymbol{x}^{(i)}-\tilde{\boldsymbol{x}}^{(j)}\right\|_2 \geq \min_{k\in[K]} \max_{(i,j)\in I_k} \delta\left\|\boldsymbol{x}^{(i)}-\tilde{\boldsymbol{x}}^{(j)}\right\|_2\\
&\overset{\text{(g)}}{=} \delta\cdot \min_{k\in[K]} \max\left\{\left\|\boldsymbol{x}^{(i)}-\tilde{\boldsymbol{x}}^{(j)}\right\|_2 \;\middle|\; ij\in[N], \pi_{ij}^{(k)}>0\right\}\\
&\overset{\text{(h)}}{=} \delta\cdot \min_{\pi\in\left\{\pi^{(k)}\right\}_{k=1}^{[K]}} \max\left\{\left\|\boldsymbol{x}^{(i)}-\tilde{\boldsymbol{x}}^{(j)}\right\|_2 \;\middle|\; ij\in[N], \pi_{ij}>0\right\}\\
&\overset{\text{(i)}}{\geq} \delta\cdot \min_{\pi\in\Pi(\boldsymbol{w},\tilde{\boldsymbol{w}})} \max\left\{\left\|\boldsymbol{x}^{(i)}-\tilde{\boldsymbol{x}}^{(j)}\right\|_2 \;\middle|\; ij\in[N], \pi_{ij}>0\right\}\\
&\overset{\text{(j)}}{=} \delta\cdot \mathcal{W}_\infty\left((\boldsymbol{X},\boldsymbol{w}),\left(\tilde{\boldsymbol{X}},\tilde{\boldsymbol{w}}\right)\right).
\end{aligned}$$

where (a) is by (67); (b) is by (68); (c) is since $\|\cdot\|_1 \geq \|\cdot\|_2$; (d) is since $\pi_{ij}^{(k)}=0$ whenever $(i,j)\notin I_k$; (e) and (f) are by the definition of δ_k and δ respectively; (g) is by the definition of I_k; (h) is a simple reformulation; (i) is since the minimum is taken over a larger set $\Pi(\boldsymbol{w},\tilde{\boldsymbol{w}}) \supseteq \left\{\pi^{(k)}\right\}_{k=1}^{[K]}$; and (j) is by the definition of $\mathcal{W}_\infty$. Hence, (63) holds with $\beta=\frac{\delta}{\sqrt{d}}$ and the theorem is proven. □

C Experiment Details

Hardware All experiments were conducted on a single Nvidia A40 GPU.

C.1 Comparison with PSWE

In our comparison with PSWE, shown in Table 1, we used pairs of multisets of the form $\boldsymbol{X} = \left\{\frac{i}{n+1}\mathbf{1}\right\}_{i=1}^{n}$ for two different values of n, n_1, n_2, where $\mathbf{1} \in \mathbb{R}^d$ is the vector whose entries all equal 1. For example, one could choose

$$\boldsymbol{X}_1 = \left\{\tfrac{1}{3}, \tfrac{2}{3}\right\}, \quad \boldsymbol{X}_2 = \left\{\tfrac{1}{4}, \tfrac{2}{4}, \tfrac{3}{4}\right\}.$$

Despite the fact that these multisets are distinct, the interpolation used by PSWE (see e.g., Naderializadeh et al. (2021)) identifies these two multisets with the same continuous distribution, and hence assigns them the same embedding. In Table 1 we used $d = 3$ and $n_1 = 5$ and $n_2 = 200$, and found, as expected, that PSWE assigns the same value to both multisets, up to numerical error, regardless of the embedding dimension m. In contrast, our method successfully separates $\boldsymbol{X}_1$ and $\boldsymbol{X}_2$ even with $m = 1$, and gives a good approximation of the Sliced Wasserstein distance as m increases. Similar results to those shown in the table were obtained for other choices of d, n_1, n_2.

Note that computing the sliced Wasserstein distance between such pairs is straightforward, since, for any two multisets $\boldsymbol{X}_1, \boldsymbol{X}_2$ in $\mathbb{R}^d$ with all points belonging to the same line, it can be shown that

$$\mathcal{SW}(\boldsymbol{X}_1, \boldsymbol{X}_2) = \frac{1}{\sqrt{d}}\mathcal{W}(\boldsymbol{X}_1, \boldsymbol{X}_2),$$

and $\mathcal{W}(\boldsymbol{X}_1, \boldsymbol{X}_2)$ can be computed using a linear programming solver.

The PSWE method takes two input parameters—the number of slices L and embedding dimension N, which, unlike our method, need not be identical. We set both parameters to the embedding dimension m.

C.2 Learning to approximate the Wasserstein distance

In this experiment we used embedding dimensions $m_1 = m_2 = 1000$. The MLP consisted of three layers with a hidden dimension of 1000. With this choice of hyperparameters, our model has roughly 3 million learnable parameters and 5 million parameters in total. These hyperparameters were picked manually. The performance of our architecture did not exhibit high sensitivity to the choice of hyperparameters: on most datasets, similar results were obtained with MLPs consisting of 2 to 8 layers, and with hidden dimensions of 500, 1000, 2000 and 4000.

We used fixed parameters for the first embedding E_1 and learnable parameters for the second embedding E_2. This choice was made since E_1 is, in most cases, supposed to handle arbitrary input point clouds, whereas the input to E_2 is more specific, in that it is always a set of two vectors that are outputs of E_1. Thus, in principle the architecture may benefit from tuning E_2 to its particular input structure. In practice, using fixed parameters in both embeddings did not significantly impair performance.

Remarkably, applying an MLP to the input points prior to embedding them via E_1 (i.e. adding a feature transform), as well as applying an MLP to the two outputs of E_1 prior to embedding them via E_2, *impaired* rather than improved the performance. This indicates that our embedding is expressive enough to encode all the required information from the input multisets in a way that facilitates processing by the MLP Φ, thus making additional processing at intermediate steps unnecessary.

Inference times for one pair of multisets were less than half a second for the `ModelNet-large` dataset, and less than 0.2 seconds for the rest of the datasets. The training times of the competing models appear in Table 3.

Training was performed on an Nvidia A40 GPU, whereas the rest of the methods were trained over an Nvidia RTX A6000 GPU, both of which have similar performance on 32-bit floating point (37.4 and 38.7 TFLOPS).

Exact computation of the 1-Wasserstein distance using the `ot.emd2()` function of the Python Optimal Transport package (Flamary et al., 2021) was up to 2.5 times slower than our method (2 to

5 ms vs 1.9 ms) on small multisets (less than 300 elements) and 150 times slower (640 ms vs 4.2 ms) on large multisets (`ModelNet-large`).

C.3 Robustness to parameter reduction

In both architectures, the input and output dimension of each MLP, as well as that of the FSW embedding, were multiplied by a scaling factor to obtain the desired number of parameters. Training our model in all problem instances took between 60 to 65 minutes. Training PointNet took between 4:43 hours to 5 hours, and was done using the original code of (Xia, 2019).