FOURIER SLICED-WASSERSTEIN EMBEDDING FOR MULTISETS AND MEASURES

Anonymous authors
Paper under double-blind review

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 embedding 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 number of support points in the input.

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 achievable.

Through numerical experiments, we demonstrate that our method yields superior representations of input multisets and offers practical advantage for learning on multiset data. Specifically, we show that (a) the FSW embedding induces significantly lower distortion on the space of multisets, compared to the leading method for computing sliced-Wasserstein-preserving embeddings; and (b) a simple combination of the FSW embedding and an MLP achieves state-of-the-art performance in learning the (non-sliced) Wasserstein distance.

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., 2017), chemical property prediction (Pozdnyakov & Ceriotti, 2023), and image deblurring (Aittala & Durand, 2018). Multiset aggregation functions are also 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 that take multisets as input, is *injectivity*. The importance of injectivity is highlighted by the following observation: A multiset architecture that cannot separate two distinct multisets $\boldsymbol{X} \neq \boldsymbol{X}'$, will not be able to approximate a target function f that differentiates between these multisets, i.e. $f(\boldsymbol{X}) \neq f(\boldsymbol{X}')$. Conversely, a multiset 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 many works to study the injectivity properties of multiset models (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).

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

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

where F is usually an MLP, and Pool is a simple pooling operation such as maximum, mean, or sum. 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 discussed 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 and Softplus.

However, injectivity alone is not the strongest property one may desire for multiset functions. While an injective multiset embedding E can separate any pair 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 some collection of multisets, or more generally, distributions over $\mathbb{R}^d$. We say that E is *bi-Lipschitz* if there exist constants $0 < c \leq C < \infty$ 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, is used as a standard notion of distance between multisets and distributions. The ratio of Lipschitz constants C/c represents a bound on the maximal distortion incurred by the map E, akin to the condition number of a matrix.

Bi-Lipschitz embeddings can be used to apply metric-based learning methods, such as nearest-neighbor search, data clustering and multi-dimensional scaling, to the embedded Euclidean domain rather than the original domain of multisets and distributions, where metric calculations are more computationally demanding; see, for example, (Indyk & Thaper, 2003). The bi-Lipschitzness of the embedding provides correctness guarantees for this approach, which depend on the Lipschitz constants c, C; see (Cahill et al., 2024).

A guarantee of bi-Lipschitzness is typically more difficult to achieve than injectivity, and often requires a different set of theoretical tools. It was recently shown in (Amir et al., 2023) that multiset embeddings based on average- or sum-pooling can never be bi-Lipschitz, even if they are injective. Currently, there are two main approaches for constructing bi-Lipschitz embeddings for multisets: (1) the *max filtering* approach of Cahill et al. (2022) which is relatively computationally intensive as it requires multiple computations of Wasserstein distances from ‘template multisets’; and (2) the *sort embedding* approach of Balan et al. (2022), which is based on sorting random projections of the multiset elements.

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 is not clear how to generalize them to multisets of varying size, let alone 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 nonempty multisets over $\mathbb{R}^d$ with at most n elements. We denote this space by $\mathcal{S}_{\leq n}(\mathbb{R}^d)$. Note that the assumption of bounded cardinality is necessary, as otherwise, even injectivity is impossible; see, e.g. (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 $\mathcal{P}_{\leq n}\left(\mathbb{R}^d\right)$. 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 $\mathcal{D} = \mathcal{S}_{\leq n}\left(\mathbb{R}^d\right)$ or $\mathcal{P}_{\leq n}\left(\mathbb{R}^d\right)$, 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 finitely supported multisets and distributions, which is a non-trivial generalization of the sort embedding. We observe that the Euclidean distance between the sort embedding 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 $\mathcal{S}_{\leq n}\left(\mathbb{R}^d\right)$ and $\mathcal{P}_{\leq n}\left(\mathbb{R}^d\right)$. Our method essentially operates as follows: (1) compute random one-dimensional projections (also called *slices*) of the input distribution; (2) for each projected distribution, compute the *quantile function*; 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 $E_m^{\text{FSW}} : \mathcal{P}_{\leq n}\left(\mathbb{R}^d\right) \to \mathbb{R}^m$.

The function

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

maps multisets and distributions to $\mathbb{R}^m$, and depends on the parameters $\boldsymbol{v}^{(k)} \in \mathbb{R}^d$, $\xi^{(k)} \in \mathbb{R}$ for $k = 1, \ldots, m$, which represent projection vectors and frequencies respectively. It has the following properties:

1. **Bi-Lipschitzness on multisets:** For $m \geq 2nd+1$, the map $E_m^{\text{FSW}} : \mathcal{S}_{\leq n}\left(\mathbb{R}^d\right) \to \mathbb{R}^m$ is bi-Lipschitz (and hence also 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 measures:** For $m \geq 2nd + 2n - 1$, the map $E_m^{\text{FSW}} : \mathcal{P}_{\leq n}\left(\mathbb{R}^d\right) \to \mathbb{R}^m$ is injective (but is not bi-Lipschitz) for almost any choice of parameters (Theorem 4.1). Moreover, by adding one more output coordinate, it can be made injective on arbitrary measures supported on n points. We also prove that bi-Lipschitzness on $\mathcal{P}_{\leq n}\left(\mathbb{R}^d\right)$ is *impossible* for any Euclidean embedding (Theorem 4.4). Thus, the bi-Lipschitzness properties of E_m^{FSW} are in a sense the best possible.
3. **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 $\left(\boldsymbol{v}^{(k)}, \xi^{(k)}\right)_{k=1}^m$. Hence, it is amenable to gradient-based learning methods, and its parameters can be trained.
4. **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 $\mathcal{O}\left(\frac{1}{\sqrt{m}}\right)$.
5. **Complexity:** The embedding $E_m^{\text{FSW}}(\mu)$ can be computed efficiently in $\mathcal{O}(mnd + mn\log n)$ time, which matches the time complexity of the most efficient methods that are used in practice (up to the logarithmic factor).

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

Empirically, we show that our method embeds the space of input multisets with a significantly lower distortion than the leading method for computing sliced-Wasserstein-preserving embeddings. We also demonstrate the promise of our method for practical applications by evaluating it in the task of learning the (non-sliced) 1-Wasserstein distance function. We show that replacing the summation-based aggregation used in state-of-the-art methods with our FSW embedding leads to improved results with lower training times.

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 that we are interested in, and metrics over these spaces.

Multisets and distributions Following the notation of Amir et al. (2023), we use $\mathcal{P}_{\leq n}(\Omega)$ to denote the collection of all probability distributions over $\Omega \subseteq \mathbb{R}^d$ that are supported on at most n points. Any distribution $\mu \in \mathcal{P}_{\leq n}(\Omega)$ can be parametrized by points $\boldsymbol{x}^{(i)} \in \Omega$ and weights $w_i \geq 0$ such that $\sum_{i=1}^n w_i = 1$,

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

where $\delta_{\boldsymbol{x}}$ is Dirac's delta function at $\boldsymbol{x}$. Note that distributions supported on less than n points can be parameterized in this was by setting some of the weights w_i to zero and choosing the corresponding $\boldsymbol{x}^{(i)}$ arbitrarily. This parametrization is generally not unique.

Similarly, let $\mathcal{S}_{\leq n}(\Omega)$ be the collection of all nonempty multisets over $\Omega \subseteq \mathbb{R}^d$ with at most n points. We identify each multiset $\boldsymbol{X} = \left\{\boldsymbol{x}^{(i)}\right\}_{i\in[n]} \in \mathcal{S}_{\leq n}(\Omega)$ with the distribution $\mu[\boldsymbol{X}] \in \mathcal{P}_{\leq n}(\Omega)$ that assigns uniform weights $w_i = \frac{1}{n}$ to each $\boldsymbol{x}^{(i)}$, accounting for multiplicities.[1] With this identification, we can regard $\mathcal{S}_{\leq n}(\Omega)$ as a subset of $\mathcal{P}_{\leq n}(\Omega)$. Our embedding, at its basic form, described in the next section, considers $\mathcal{S}_{\leq n}(\Omega)$ with this identification and therefore does not distinguish between multisets of different cardinalities if their element proportions are identical.[2] This can be easily remedied by augmenting the embedding with an additional coordinate representing the multiset cardinality, or in the case of measures, the total mass $\sum_{i=1}^n w_i$; see discussion in Appendix A.1.

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 measure of distance on $\mathcal{S}_{\leq n}\left(\mathbb{R}^d\right)$ and $\mathcal{P}_{\leq n}\left(\mathbb{R}^d\right)$, we use the Wasserstein distance. Intuitively, the Wasserstein distance is the minimal amount of work required in order 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 $\boldsymbol{x}^{(i)}, \tilde{\boldsymbol{x}}^{(i)}$ and weights $w_i, \tilde{w}_i$ as in (2), the p-Wasserstein distance from μ to $\tilde{\mu}$ is defined by

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

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

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

Intuitively, π_{ij} denotes how much mass is to be transported from point $\boldsymbol{x}^{(i)}$ to point $\tilde{\boldsymbol{x}}^{(j)}$. For $p = \infty$, the Wasserstein distance is defined by

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

Whenever p is omitted, we refer to $\mathcal{W}_p$ with $p = 2$. Similarly, $\|\cdot\|$ always denotes the ℓ_2 norm.

[1] For example, if $\boldsymbol{X} = \{a, b, b\} \in \mathcal{S}_{\leq 3}(\mathbb{R})$, then $\mu[\boldsymbol{X}] = \frac{1}{3}\delta_a + \frac{2}{3}\delta_b$.

[2] 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 $\mu[\boldsymbol{X}] = \mu[\boldsymbol{Y}]$.

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

Wasserstein when $d = 1$ In the one-dimensional case, 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 sort : $\mathbb{R}^n \to \mathbb{R}^n$ being the function that returns the input coordinates sorted in increasing order.

When considering arbitrary distributions in $\mathcal{P}_{\leq n}(\mathbb{R})$, the Wasserstein distance can be computed 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):

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

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

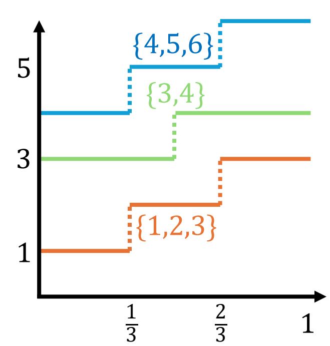

Figure 1: The quantile function of three different multisets

Sliced Wasserstein distance The *sliced Wasserstein distance*, proposed by Bonneel et al. (2015) as a surrogate for the Wasserstein distance, exploits the efficient calculation 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. To give a formal definition, we first define the projection of a distribution.

Definition. Let $\mu = \sum_{i=1}^n w_i \delta_{\boldsymbol{x}^{(i)}} \in \mathcal{P}_{\leq n}(\mathbb{R}^d)$. 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 $\mu, \tilde{\mu} \in \mathcal{P}_{\leq n}(\mathbb{R}^d)$ 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 $\mathcal{W}^2$ is the 2-Wasserstein distance squared, and the expectation $\mathbb{E}_{\boldsymbol{v}}[\,\cdot\,]$ is over the direction vector $\boldsymbol{v} \sim \mathrm{Uniform}(\mathbb{S}^{d-1})$, 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 $E : \mathcal{P}_{\leq n}(\mathbb{R}^d) \to \mathbb{R}^m$. In this subsection, we discuss existing embedding methods and some straightforward ideas to extend them. We then propose our method in the next section.

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

To extend this idea to multisets in $\mathcal{S}_{\leq n}(\mathbb{R})$ with possibly less than n elements, a naive approach would be to represent each multiset in $\mathcal{S}_{\leq n}(\mathbb{R})$ by a multiset of size N, with N being the *least* *common multiple* (LCM) of $\{1, 2, \ldots, n\}$. For example, for $n = 3$, 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 LCM$([n])$ 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 on a grid 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 probabilities w_i and sampling points t_k, 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 $\mathcal{P}_{\leq n}(\mathbb{R}^d)$ 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

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

This idea was discussed in (Balan et al., 2022; Zhang et al., 2019; Dym & Gortler, 2024; Balan & Tsoukanis, 2023b). It is rather straightforward to show that in expectation over the directions $\boldsymbol{v}_k$, this method gives a good approximation of the sliced Wasserstein distance. The relationship to the d-dimensional Wasserstein distance is a priori less clear. It was shown by Balan & Tsoukanis (2023a) that for m that is exponential in n, this mapping is injective and bi-Lipschitz for almost any choice of the directions $\boldsymbol{v}_1, \ldots, \boldsymbol{v}_m$. 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 the quantile function, to construct an embedding capable of handling arbitrary distributions in $\mathcal{P}_{\leq n}(\mathbb{R}^d)$ 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 preserves the sliced Wasserstein distance by embedding distributions into an infinite dimensional Hilbert space. In practice, a finite dimensional discretization is used, which does not maintain the injectivity guarantees. In contrast, our method is guaranteed injectivity with a finite and near-optimal embedding dimension of $\approx 2nd$.

Lastly, Haviv et al. (2024) recently proposed a neural architecture based on transformers that computes Euclidean embeddings for multisets and distributions. Their architecture, called the Wasserstein Wormhole, is trained to approximately preserve the Wasserstein distance. However, this method is not guaranteed to preserve the Wasserstein distance precisely. This limitation is particularly significant when generalizing to out-of-distribution samples.

In addition, there exist methods that compute sliced optimal-transport distances for *pairs* of input distributions (Deshpande et al., 2019; Kolouri et al., 2019; Nguyen et al., 2020). These methods have limited applicability to most learning tasks, which typically involve a single input distribution.

3 Proposed Method

Our method to embed a distribution μ essentially consists of computing random slices $\boldsymbol{v}^T\mu$ and, for each slice, taking one random sample of its quantile function $Q_{\boldsymbol{v}^T\mu}(t)$. Instead of sampling the function directly though, 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, $\tilde{\mu}$ will give us 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, the Fourier transform is smooth with respect to the frequencies, and thus so is our embedding. We shall 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* $E^{\text{FSW}}(\,\boldsymbol{\cdot}\,; \boldsymbol{v}, \xi) : \mathcal{P}_{\leq n}(\mathbb{R}^d) \to \mathbb{R}$ by

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

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

$$f_\xi(\xi) \coloneqq \begin{cases} \frac{1}{(1+\xi)^2} & \xi \geq 0 \\ 0 & \xi < 0. \end{cases}$$

We now show that this choice of E^{FSW} and $\mathcal{D}_\xi$ ensures that given two distributions $\mu, \tilde{\mu} \in \mathcal{P}_{\leq n}(\mathbb{R}^d)$, the expected distance between the samples equals the sliced Wasserstein distance between μ and $\tilde{\mu}$.

Theorem 3.2. [Proof in Appendix B.2] *Let $\mu, \tilde{\mu} \in \mathcal{P}_{\leq n}(\mathbb{R}^d)$, whose support points are all of ℓ_2-norm $\leq R$. Let $\boldsymbol{v} \sim$ Uniform$(\mathbb{S}^{d-1})$, $\xi \sim \mathcal{D}_\xi$.*

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

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

where $\mathbb{E}[\,\cdot\,]$ and $\text{STD}[\,\cdot\,]$ are the expectation and standard deviation. The result can be further stabilized by taking multiple samples. Building on this idea, we define the *Fourier Sliced Wasserstein (FSW) embedding* $E_m^{\text{FSW}} : \mathcal{P}_{\leq n}(\mathbb{R}^d) \to \mathbb{R}^m$, which aggregates multiple independent samples of the one-sample embedding:

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

where $\left(\boldsymbol{v}^{(k)},\xi^{(k)}\right)_{k=1}^m$ are drawn randomly i.i.d. from Uniform$(\mathbb{S}^{d-1}) \times \mathcal{D}_\xi$.

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

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

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

Note that the bounds in Corollary 3.3 are independent of both the number of points n and the dimension d. Thus, the estimation error is not affected by 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 maintaining a bounded standard estimation error, provided all distributions have supports contained 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 increases. Here we show that with a *finite* dimension, our embedding guarantees injectivity and bi-Lipschitzness, as outlined in the Main results paragraph of Section 1.

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

Theorem 4.1. [Proof on Page 21] *Let $E_m^{\text{FSW}} : \mathcal{P}_{\leq n}(\mathbb{R}^d) \to \mathbb{R}^m$ be as in (9), with $\left(\boldsymbol{v}^{(k)},\xi^{(k)}\right)_{k=1}^m$ sampled i.i.d. from* 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 essentially optimal up to a multiplicative factor of 2, for *any* continuous embedding, since any m smaller than nd precludes injectivity (Amir et al., 2023).

Proof idea. The proof relies on the *Finite Witness Theorem*—a result from the theory of σ-subanalytic functions presented in (Amir et al., 2023). The core idea is to use a dimension counting argument to show that for sufficiently large m, the set of embedding parameters $\left(\boldsymbol{v}^{(k)}, \xi^{(k)}\right)_{k=1}^{m}$ for which $E_m^{\text{FSW}}\left(\,\cdot\,; \left(\boldsymbol{v}^{(k)}, \xi^{(k)}\right)_{k=1}^{m}\right)$ does not uniquely determine all distributions in $\mathcal{S}_{\leq n}\left(\mathbb{R}^d\right)$ or $\mathcal{P}_{\leq n}\left(\mathbb{R}^d\right)$ is dimensionally deficient. $\square$

Next, we show that in the case of $\mathcal{S}_{\leq n}\left(\mathbb{R}^d\right)$, 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 refer to the distribution parametrized by 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 $\mathcal{D} = \mathcal{P}_{\leq n}\left(\mathbb{R}^d\right)$ or $\mathcal{D} = \mathcal{S}_{\leq n}\left(\mathbb{R}^d\right)$. 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 Page 28] *Let $E : \mathcal{P}_{\leq n}\left(\mathbb{R}^d\right) \to \mathbb{R}^m$ be injective and positively homogeneous. Let Δ^n be the probability simplex in $\mathbb{R}^n$. Suppose that the function $E(\boldsymbol{X}, \boldsymbol{w}) : \mathbb{R}^{d \times n} \times \Delta^n \to \mathbb{R}^m$ is piecewise linear in $\boldsymbol{X}$ for any fixed $\boldsymbol{w}$. Then for any fixed $\boldsymbol{w}, \tilde{\boldsymbol{w}} \in \Delta^n$, there exist constants $c, C > 0$ such that for all $\boldsymbol{X}, \tilde{\boldsymbol{X}} \in \mathbb{R}^{d \times n}$ and $p \in [1, \infty]$,*

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

Proof idea. For fixed $\boldsymbol{w}, \tilde{\boldsymbol{w}}$, both functions $\left\| E(\boldsymbol{X}, \boldsymbol{w}) - E\left(\tilde{\boldsymbol{X}}, \tilde{\boldsymbol{w}}\right) \right\|_1$ and $\mathcal{W}_1\left((\boldsymbol{X}, \boldsymbol{w}), \left(\tilde{\boldsymbol{X}}, \tilde{\boldsymbol{w}}\right)\right)$ are homogeneous and piecewise-linear with respect to $\left(\boldsymbol{X}, \tilde{\boldsymbol{X}}\right)$. The proof uses a topological argument to show that this property, combined with injectivity, implies bi-Lipschitzness. $\square$

The assumption that the weighs $\boldsymbol{w}, \tilde{\boldsymbol{w}}$ are fixed can be straightforwardly relaxed to allow for weights that come from a finite set. Based on this observation, the following corollary shows 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 $\mathcal{S}_{\leq n}\left(\mathbb{R}^d\right)$.*

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

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

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

The bi-Lipschitzness of the FSW embedding constitutes a significant advantage over prevalent methods for handling multisets. In contrast, methods based on sum- or average-pooling inevitably induce unbounded distortion on $\mathcal{S}_{\leq n}(\Omega)$, even when Ω is compact (Amir et al., 2023), and methods based on max-pooling are not even injective (Xu et al., 2018). In the next section we demonstrate how this theoretical advantage translates into practical improvements.

Next, we explore whether it is possible to further improve by finding an embedding that is bi-Lipschitz on the entirety of $\mathcal{P}_{\leq n}\left(\mathbb{R}^d\right)$. For the broader class of distributions $\bigcup_{n \in \mathbb{N}} \mathcal{P}_{\leq n}\left(\mathbb{R}^d\right)$, 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 on Page 22] *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 $\mathcal{W}_p$.*

Proof idea. The proof is technically involved. The core idea is to create two distributions where an infinitesimally small mass is transported a small distance, such that their Wasserstein distance decreases linearly, whereas any embedding would produce quadratically-converging outputs. □

5 Numerical Experiments

In this section, we demonstrate how the theoretical advantages of our method translate into superior embeddings in practice and improved results in learning on multisets.

Empirical distortion evaluation This experiment evaluates the ability of our embedding to approximately preserve the sliced Wasserstein distance and compares it with PSWE, which is designed for the same purpose. In each trial, an instance of each embedding, $E : \mathcal{S}_{\leq n}(\mathbb{R}^d) \to \mathbb{R}^m$, was generated. A batch of 6000 point-clouds in $\mathbb{R}^d$ was either sourced from the `ModelNet-large` dataset or generated randomly, with n points uniformly distributed in the unit cube $[-1, 1]^d$. The sliced Wasserstein distance from each point cloud $\boldsymbol{X} \in \mathbb{R}^{d \times n}$ to the delta distribution at zero $\delta_{\mathbf{0}}$ is given by the explicit formula $\mathcal{SW}(\boldsymbol{X}, \delta_{\mathbf{0}}) = \frac{1}{\sqrt{d}} \mathcal{W}(\boldsymbol{X}, \delta_{\mathbf{0}}) = \frac{1}{\sqrt{nd}} \sqrt{\sum_{i=1}^{n} \left\| \boldsymbol{x}^{(i)} \right\|^2}$. The embedding $E(\boldsymbol{X})$ and the quantity $r(\boldsymbol{X}) = \frac{\|E(X) - \mathbf{0}\|}{\mathcal{SW}(\boldsymbol{X}, \delta_{\mathbf{0}})}$ were calculated, and the empirical distortion was taken as the ratio of the maximal to minimal $r(\boldsymbol{X})$ across the batch. As shown in Table 1, our embedding exhibits markedly lower distortion, with the improvement being particularly pronounced in real data.

Table 1: Empirical Distortion Evaluation

Dataset	d	n	Method	m = 10	50	200	1000
ModelNet	3	2047	PSWE	16.45	30.26	*OOM*	*OOM*
			FSW	2.47	1.46	1.2	1.08
Uniform	3	20	PSWE	10.03	4.5	3.55	*OOM*
			FSW	2.23	1.36	1.16	1.07
Uniform	10	20	PSWE	8.15	2.84	1.97	*OOM*
			FSW	2.29	1.41	1.18	1.08
Uniform	100	20	PSWE	7.88	2.74	1.62	*OOM*
			FSW	2.43	1.44	1.19	1.08
Uniform	1000	20	PSWE	7.97	2.7	1.66	*OOM*
			FSW	2.4	1.43	1.19	1.08

Empirical distortion with respect to the sliced Wasserstein distance, evaluated on real and synthetic data. In each trial, distortion was evaluated on 6000 point clouds. The numbers show the average over 200 independent trials. d, n: ambient dimension and number of points in each cloud; m: embedding dimension; OOM: Out of Memory—the method failed due to insufficient memory.

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, 2024; Kawano et al., 2020), which proposed architectures designed to approximate functions $F : \mathcal{S}_{\leq n}\left(\mathbb{R}^d\right) \times \mathcal{S}_{\leq n}\left(\mathbb{R}^d\right) \to \mathbb{R}$, 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 seems 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 $E_1 : \mathcal{P}_{\leq n}\left(\mathbb{R}^d\right) \to \mathbb{R}^{m_1}$ 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

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

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 that F is designed to approximate.

The experimental setting was replicated from (Chen & Wang, 2024), where the objective is to approximate the 1-Wasserstein distance $\mathcal{W}_1$. We used the following evaluation datasets, kindly provided to us by the authors: Three synthetic datasets `noisy-sphere-3`, `noisy-sphere-6` and `uniform`, consisting of random point clouds in $\mathbb{R}^3$, $\mathbb{R}^6$ and $\mathbb{R}^2$ respectively; two real datasets `ModelNet-small` and `ModelNet-large`, consisting of 3D point-clouds sampled from ModelNet40 objects (Wu et al., 2015); and the gene-expression dataset `RNAseq` (Yao et al., 2021), consisting of 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 Euclidean embeddings for input distributions, and $\mathcal{N}_{\text{ProductNet}}$, which further processes the two joined embeddings by an MLP (Chen & Wang, 2024); (b) a Siamese autoencoder called Wasserstein Point-Cloud Embedding network (WPCE) (Kawano et al., 2020); (c) the Sinkhorn algorithm (Cuturi, 2013), which computes an efficient approximation to $\mathcal{W}_p$ by adding an entropy regularization term. We also evaluated the PSWE embedding of Naderializadeh et al. (2021), by employing it in our architecture instead of E_1, E_2.

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.

6 Conclusion

In this paper, we introduced an embedding that offers strong bi-Lipschitzness and injectivity guarantees for multisets and measures respectively. Our experimental results indicate that our embedding produces representations that better preserve the original geometry of the data and can lead to improved performance in practical learning tasks.

In the future, we aim to explore the use of the FSW embedding as an aggregation function in graph neural networks, and to generalize the concepts described here to other 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, checkpoints, and actual numerical results presented in this paper, are available at the anonymous URL https://drive.filen.io/d/07bcfdb5-b7bf-41b0-96b3-265542caf1fa#3l6YAiOBzMA9lvwIpywqiCOZHsoIs57V. Reproduction instructions can be found in the file `readme.txt` at the root directory of the downloaded zip file. While 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.

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. volume 37, 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, 2024.

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.

Ishan Deshpande, Yuan-Ting Hu, Ruoyu Sun, Ayis Pyrros, Nasir Siddiqui, Sanmi Koyejo, Zhizhen Zhao, David Forsyth, and Alexander G Schwing. Max-sliced wasserstein distance and its use for gans. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 10648–10656, 2019.

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.

Soheil Kolouri, Kimia Nadjahi, Umut Simsekli, Roland Badeau, and Gustavo Rohde. Generalized sliced wasserstein distances. *Advances in neural information processing systems*, 32, 2019.

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.

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.

Khai Nguyen, Nhat Ho, Tung Pham, and Hung Bui. Distributional sliced-wasserstein and applications to generative modeling. *arXiv preprint arXiv:2002.07367*, 2020.

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 2017.

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.

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.

Denote by $\mathcal{M}_{\leq n}(\Omega)$ the collection of all measures μ over $\Omega \subseteq \mathbb{R}^d$ with at most n support points $\boldsymbol{x}^{(1)}, \ldots, \boldsymbol{x}^{(n)}$ and corresponding weights $w_1, \ldots, w_n \geq 0$,

$$\mu = \sum_{i=1}^{n} w_i \delta_{\boldsymbol{x}^{(i)}}. \tag{13}$$

The *total mass* of μ is the quantity $\mu(\Omega) = \sum_{i=1}^{n} w_i$.

A first step towards extending E^{FSW} to $\mathcal{M}_{\leq n}(\Omega)$ while maintaining injectivity is to simply add an extra coordinate in the output that represents the total mass of the input measure. Define $\tilde{E}_m^{\text{FSW}} : \mathcal{M}_{\leq n}(\Omega) \to \mathbb{R}^m$ for an input μ as in (13) by

$$\tilde{E}_m^{\text{FSW}}(\mu) \coloneqq \left(\mu(\Omega), E_{m-1}^{\text{FSW}}(\hat{\mu})\right), \tag{14}$$

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

$$\hat{\mu} = \sum_{i=1}^{n} \frac{w_i}{\sum_{j=1}^{n} w_j} \delta_{\boldsymbol{x}^{(i)}}. \tag{15}$$

It is easy to show that with the above definition, by Theorem 4.1, $\tilde{E}_m^{\text{FSW}}$ with output dimension $m \geq 2nd + 2n$ is injective on $\mathcal{M}_{\leq n}(\mathbb{R}^d)$ excluding the zero measure, for which (15) is not defined, and when restricted to nonempty multisets in $\mathcal{S}_{\leq n}(\mathbb{R}^d)$, it suffices to take $m \geq 2nd+2$ to ensure that $\tilde{E}_m^{\text{FSW}}$ differentiates between input multisets with different cardinalities but the same proportions of element multiplicities, as discussed in Page 4, Section 2.1.

One limitation of the definition in (14) is that $\tilde{E}_m^{\text{FSW}}$ is not well defined and has a pathological jump discontinuity at the input zero measure $\mu(\Omega) = 0$. This can be remedied by padding input measures whose total mass is below a chosen threshold with the complementary mass assigned to the zero vector. Namely, choose an arbitrary threshold $\rho > 0$ and adjust the definition in (14) to

$$\tilde{E}_m^{\text{FSW}}(\mu) \coloneqq \begin{cases} \left(\mu(\Omega), E_{m-1}^{\text{FSW}}(\hat{\mu})\right) & \mu(\Omega) \geq \rho, \\ \left(\mu(\Omega), E_{m-1}^{\text{FSW}}\left(\left(1 - \frac{\mu(\Omega)}{\rho}\right)\delta_{\boldsymbol{0}} + \sum_{i=1}^{n} \frac{w_i}{\rho} \delta_{\boldsymbol{x}^{(i)}}\right)\right) & \mu(\Omega) < \rho. \end{cases} \tag{16}$$

It is easy to show that with the definition (16), $\tilde{E}_m^{\text{FSW}}$ with the appropriate m as detailed above is well defined and injective on the whole of $\mathcal{M}_{\leq n}(\mathbb{R}^d)$.

A.2 PRACTICAL COMPUTATION

Here we present some formulas that facilitate the practical computation of E^{FSW}.

We start by developing some 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

$$\sigma(\boldsymbol{x}) = (\sigma_1(\boldsymbol{x}), \ldots, \sigma_n(\boldsymbol{x})) \in \mathrm{S}_n$$

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 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

$$Q_\mu(t) = x_{(k_{\min}(\sigma(\boldsymbol{x}), \boldsymbol{w}, t))}, \tag{17}$$

where $k_{\min}(\sigma, \boldsymbol{w}, t)$ is defined for $\sigma = (\sigma_1, \ldots, \sigma_n) \in \mathrm{S}_n$ by

$$k_{\min}(\sigma, \boldsymbol{w}, t) \coloneqq \min \{k \in [n] \mid w_{\sigma_1} + \cdots + w_{\sigma_k} > t\}. \tag{18}$$

It can be seen in (17) and (18) 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 $Q_\mu(1) = \operatorname{ess\,max}(\mu)$.

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 (17) and (18).

Using the identity (17), we can express $E(\mu; \boldsymbol{v}, \xi)$ as

$$\begin{aligned} E(\mu; \boldsymbol{v}, \xi) =& 2(1+\xi) \sum_{k=1}^n \int_{t=\sum_{i=1}^{k-1} w_{\sigma_i(\boldsymbol{v}^T\boldsymbol{X})}}^{\sum_{i=1}^{k} w_{\sigma_i(\boldsymbol{v}^T\boldsymbol{X})}} Q_{\boldsymbol{v}^T\mu}(t) \cos(2\pi\xi t) dt \\ =& 2(1+\xi) \sum_{k=1}^n \int_{t=\sum_{i=1}^{k-1} w_{\sigma_i(\boldsymbol{v}^T\boldsymbol{X})}}^{\sum_{i=1}^{k} w_{\sigma_i(\boldsymbol{v}^T\boldsymbol{X})}} \left(\boldsymbol{v}^T\boldsymbol{X}\right)_{(k)} \cos(2\pi\xi t) dt \\ =& 2\frac{1+\xi}{2\pi\xi} \sum_{k=1}^n \left(\boldsymbol{v}^T\boldsymbol{X}\right)_{(k)} \left[\sin(2\pi\xi t)\right]_{t=\sum_{i=1}^{k-1} w_{\sigma_i(\boldsymbol{v}^T\boldsymbol{X})}}^{\sum_{i=1}^{k} w_{\sigma_i(\boldsymbol{v}^T\boldsymbol{X})}}, \end{aligned} \tag{19}$$

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

$$E(\mu; \boldsymbol{v}, \xi) = 2\frac{1+\xi}{2\pi\xi} \sum_{k=1}^n \sin\left(2\pi\xi \sum_{i=1}^k w_{\sigma_i(\boldsymbol{v}^T\boldsymbol{X})}\right) \left[\left(\boldsymbol{v}^T\boldsymbol{X}\right)_{(k)} - \left(\boldsymbol{v}^T\boldsymbol{X}\right)_{(k+1)}\right], \tag{20}$$

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

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

B Proofs

B.1 The cosine transform

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

In the following discussion, L^p always 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

$$\hat{f}(\xi) \coloneqq 2 \int_0^\infty f(t) \cos(2\pi\xi t) dt \tag{21}$$

for $\xi \geq 0$.

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

$$\left\|\hat{f}\right\|_{L^\infty} \leq 2\|f\|_{L^1} \tag{22}$$

since

$$\left|\hat{f}(\xi)\right| \leq 2\int_0^\infty |f(t)|\cdot|\cos(2\pi\xi t)|dt \leq 2\int_0^\infty |f(t)|dt = 2\|f\|_{L^1}. \tag{23}$$

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*
$$\int_0^\infty (f(t))^2 dt = \int_0^\infty \left(\hat{f}(t)\right)^2 dt. \tag{24}$$

2. *Suppose that $f \in L^1 \cap L^\infty$, and that f is monotonous on an interval $(0,T)$ and vanishes almost everywhere outside of $(0,T)$. Then for any $\xi > 0$,*
$$\left|\hat{f}(\xi)\right| \leq \frac{3}{\pi\xi}\|f\|_{L^\infty}. \tag{25}$$

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

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

Then the Fourier transform of f_e is given by

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

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 to (21), 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

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

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 shall 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 $\tilde{f} : \mathbb{R} \to \mathbb{R}$

$$\tilde{f}(t) := \begin{cases} f(0^+) & t \le 0 \\ f(t) & t \in (0, T) \\ f(T^-) & t \ge T. \end{cases} \tag{26}$$

With this definition, $\tilde{f}$ coincides with f on I, is monotonous on $\mathbb{R}$, and it can be shown that

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

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

$$f_\varepsilon(t) \coloneqq \chi_I(t) \int_{\mathbb{R}} \tilde{f}(r)\phi_\varepsilon(t-r)dr = \chi_I(t) \int_{\mathbb{R}} \tilde{f}(t+r)\phi_\varepsilon(r)dr, \tag{27}$$

with χ_I denoting the characteristic function of I. From the rightmost part of (27), it is evident that the monotonicity of $\tilde{f}$ implies that f_ε is monotonous on I.

Also note that

$$\begin{aligned} |f_\varepsilon(t)| &\leq \chi_I(t) \int_{\mathbb{R}} \left|\tilde{f}(t+r)\right| \phi_\varepsilon(r)dr \\ &\leq \left\|\tilde{f}\right\|_{L^\infty} \int_{\mathbb{R}} \phi_\varepsilon(r)dr \\ &= \left\|\tilde{f}\right\|_{L^\infty} = \|f\|_{L^\infty}. \end{aligned} \tag{28}$$

Thus,

$$\|f_\varepsilon\|_{L^1} \leq T\|f\|_{L^\infty}, \qquad \|f_\varepsilon\|_{L^\infty} \leq \|f\|_{L^\infty}, \tag{29}$$

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

$$\left|\hat{f}_\varepsilon(\xi)\right| \leq \frac{3}{\pi\xi}\|f_\varepsilon\|_{L^\infty} \overset{\text{(a)}}{\leq} \frac{3}{\pi\xi}\|f\|_{L^\infty}, \tag{30}$$

with (a) due to (29). Thus,

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

In conclusion,

$$\frac{3}{\pi\xi}\|f\|_{L^\infty} \geq \text{(30)} \left|\hat{f}_\varepsilon(\xi)\right| \xrightarrow[\varepsilon\to 0]{} \left|\hat{f}(\xi)\right|$$

and therefore

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

$\square$

B.2 Probabilistic properties of $E(\mu; \boldsymbol{v}, \xi)$ and $\Delta(\mu, \nu; \boldsymbol{v}, \xi)$

In this proof, 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 'norm' for distributions in $\mathcal{P}_{\leq n}(\mathbb{R}^d)$ by

$$\|\mu\|_{\mathcal{W}_p} \coloneqq \mathcal{W}_p(\mu, 0), \quad p \in [1, \infty],$$

where 0 here denotes the distribution that assigns a mass of 1 to the point $0 \in \mathbb{R}^d$. Note that this is not a norm in the formal sense of the word, as $\mathcal{P}_{\leq n}(\mathbb{R}^d)$ is not a vector space.

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

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

$$\mathcal{SW}(\mu, \nu) \leq \mathcal{W}(\mu, \nu) \leq \|\mu\|_{\mathcal{W}_\infty} + \|\nu\|_{\mathcal{W}_\infty}. \tag{31}$$

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*

$$|E(\mu; \boldsymbol{v}, \xi)| \leq 3\|\mu\|_{\mathcal{W}_\infty} \quad \forall \xi \geq 0, \tag{32}$$

$$\mathbb{E}_\xi\left[\Delta^2(\mu, \nu; \boldsymbol{v}, \xi)\right] = \mathcal{W}^2\left(\boldsymbol{v}^T\mu, \boldsymbol{v}^T\nu\right), \tag{33}$$

$$\text{STD}_\xi\left[\Delta^2(\mu, \nu; \boldsymbol{v}, \xi)\right] \leq 3\left(\|\mu\|_{\mathcal{W}_\infty} + \|\nu\|_{\mathcal{W}_\infty}\right)\mathcal{W}\left(\boldsymbol{v}^T\mu, \boldsymbol{v}^T\nu\right). \tag{34}$$

Proof. By definition,

$$E(\mu; \boldsymbol{v}, \xi) = (1 + \xi)\hat{Q}_{\boldsymbol{v}^T\mu}(\xi). \tag{35}$$

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 (23),

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

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

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

and thus (32) holds. Note that since $E(\mu; \boldsymbol{v}, \xi)$ 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,

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

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, (33) holds.

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

$$\begin{aligned}\mathrm{Var}_\xi\left[\Delta^2(\mu,\nu;\boldsymbol{v},\xi)\right] =& \mathbb{E}_\xi\left[\left(\Delta^2(\mu,\nu;\boldsymbol{v},\xi)\right)^2\right]-\left(\mathbb{E}_\xi\left[\Delta^2(\mu,\nu;\boldsymbol{v},\xi)\right]\right)^2\\ \overset{(33)}{=}& \mathbb{E}_\xi\left[\Delta^4(\mu,\nu;\boldsymbol{v},\xi)\right]-\left(\mathcal{W}^2\left(\boldsymbol{v}^T\mu,\boldsymbol{v}^T\nu\right)\right)^2\\ =& \mathbb{E}_\xi\left[\left(E(\mu;\boldsymbol{v},\xi)-E(\nu;\boldsymbol{v},\xi)\right)^2\cdot\Delta^2(\mu,\nu;\boldsymbol{v},\xi)\right]-\mathcal{W}^4\left(\boldsymbol{v}^T\mu,\boldsymbol{v}^T\nu\right)\\ \leq& \mathbb{E}_\xi\left[\left(|E(\mu;\boldsymbol{v},\xi)|+|E(\nu;\boldsymbol{v},\xi)|\right)^2\cdot\Delta^2(\mu,\nu;\boldsymbol{v},\xi)\right]-\mathcal{W}^4\left(\boldsymbol{v}^T\mu,\boldsymbol{v}^T\nu\right)\\ \overset{(32)}{\leq}& \mathbb{E}_\xi\left[\left(3\|\mu\|_{\mathcal{W}_\infty}+3\|\nu\|_{\mathcal{W}_\infty}\right)^2\cdot\Delta^2(\mu,\nu;\boldsymbol{v},\xi)\right]-\mathcal{W}^4\left(\boldsymbol{v}^T\mu,\boldsymbol{v}^T\nu\right)\\ =& 9\left(\|\mu\|_{\mathcal{W}_\infty}+\|\nu\|_{\mathcal{W}_\infty}\right)^2\cdot\mathbb{E}_\xi\left[\Delta^2(\mu,\nu;\boldsymbol{v},\xi)\right]-\mathcal{W}^4\left(\boldsymbol{v}^T\mu,\boldsymbol{v}^T\nu\right)\\ \overset{(33)}{=}& 9\left(\|\mu\|_{\mathcal{W}_\infty}+\|\nu\|_{\mathcal{W}_\infty}\right)^2\cdot\mathcal{W}^2\left(\boldsymbol{v}^T\mu,\boldsymbol{v}^T\nu\right)-\mathcal{W}^4\left(\boldsymbol{v}^T\mu,\boldsymbol{v}^T\nu\right)\\ \leq& 9\left(\|\mu\|_{\mathcal{W}_\infty}+\|\nu\|_{\mathcal{W}_\infty}\right)^2\cdot\mathcal{W}^2\left(\boldsymbol{v}^T\mu,\boldsymbol{v}^T\nu\right),\end{aligned}$$

and thus (34) holds.

This concludes the proof of Lemma B.4. □

Let us now prove Theorem 3.2.

Theorem 3.2. [Proof in Appendix B.2] *Let $\mu,\tilde{\mu}\in\mathcal{P}_{\leq n}\left(\mathbb{R}^d\right)$, whose support points are all of ℓ_2-norm $\leq R$. Let $\boldsymbol{v}\sim\mathrm{Uniform}\left(\mathbb{S}^{d-1}\right)$, $\xi\sim\mathcal{D}_\xi$.*

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

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

Proof. Eq. (7) holds since

$$\begin{aligned}\mathbb{E}_{\boldsymbol{v},\xi}\left[\Delta^2(\mu,\nu;\boldsymbol{v},\xi)\right] &= \mathbb{E}_{\boldsymbol{v}}\left[\mathbb{E}_{\xi|\boldsymbol{v}}\left[\Delta^2(\mu,\nu;\boldsymbol{v},\xi)\right]\right]\\ &\overset{(33)}{=} \mathbb{E}_{\boldsymbol{v}}\left[\mathcal{W}^2\left(\boldsymbol{v}^T\mu,\boldsymbol{v}^T\nu\right)\right]\\ &\overset{(5)}{=} \mathcal{SW}^2(\mu,\nu).\end{aligned}$$

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] \\
&= (33)\ \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] \\
&\le (34)\ \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] \\
&= (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] \\
&\le 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] \\
&\le (31)\ 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. [Proof on Page 21] *Let $E_m^{\text{FSW}} : \mathcal{P}_{\le n}(\mathbb{R}^d) \to \mathbb{R}^m$ be as in (9), with $\left(\boldsymbol{v}^{(k)},\xi^{(k)}\right)_{k=1}^m$ sampled i.i.d. from* Uniform$(\mathbb{S}^{d-1}) \times \mathcal{D}_\xi$*. Then:*

1. *If $m \ge 2nd+1$, then with probability 1, E_m^{FSW} is injective on $\mathcal{S}_{\le n}(\mathbb{R}^d)$.*
2. *If $m \ge 2nd+2n-1$, then with probability 1, E_m^{FSW} is injective on $\mathcal{P}_{\le n}(\mathbb{R}^d)$.*

Proof. This proof relies on the theory of σ-subanalytic functions, introduced in (Amir et al., 2023). The main result that we use from (Amir et al., 2023) is the *Finite Witness Theorem*, which is a tool to reduce an infinite set of equality constraints to a finite subset chosen randomly, while maintaining equivalence with probability 1. The Finite Witness Theorem is a useful tool to prove that certain functions are injective.

The theory defines a family of functions called σ*-subanalytic functions*. The full definition of this family is technically involved and requires heavy theoretical machinery, and thus we do not state here the full definition. However, we use the following properties of σ-subanalytic functions, proved in (Amir et al., 2023):

1. Piecewise-linear functions are σ-subanalytic.
2. Finite sums, products and compositions of σ-subanalytic functions are σ-subanalytic.

We first show that the function $E^{\text{FSW}}(\boldsymbol{X},\boldsymbol{p};\boldsymbol{v},\xi)$ is σ-subanalytic as a function of $(\boldsymbol{X},\boldsymbol{p},\boldsymbol{v},\xi)$. To see this, note that by (20), $E^{\text{FSW}}(\boldsymbol{X},\boldsymbol{p};\boldsymbol{v},\xi)$ is the sum over $k \in [n]$ of terms of the form

$$2\frac{1+\xi}{2\pi\xi}\sin\left(2\pi\xi\sum_{i=1}^{k} w_{\sigma_i(\boldsymbol{v}^T\boldsymbol{X})}\right)\left[\left(\boldsymbol{v}^T\boldsymbol{X}\right)_{(k)} - \left(\boldsymbol{v}^T\boldsymbol{X}\right)_{(k+1)}\right]. \tag{36}$$

Each term $\left[\left(\boldsymbol{v}^T\boldsymbol{X}\right)_{(k)} - \left(\boldsymbol{v}^T\boldsymbol{X}\right)_{(k+1)}\right]$ and $\sum_{i=1}^{k} w_{\sigma_i(\boldsymbol{v}^T\boldsymbol{X})}$ is piecewise linear in the product $\boldsymbol{v}^T\boldsymbol{X}$ and thus σ-subanalytic, as well as the product $2\pi\xi\sum_{i=1}^{k} w_{\sigma_i(\boldsymbol{v}^T\boldsymbol{X})}$, composition $\sin\left(2\pi\xi\sum_{i=1}^{k} w_{\sigma_i(\boldsymbol{v}^T\boldsymbol{X})}\right)$ and again product $2\frac{1+\xi}{2\pi\xi}\sin\left(2\pi\xi\sum_{i=1}^{k} w_{\sigma_i(\boldsymbol{v}^T\boldsymbol{X})}\right)$ and finally the product (36) and the finite sum of such.

We shall now show that $E^{\text{FSW}}(\boldsymbol{X},\boldsymbol{p};\boldsymbol{v},\xi)$ satisfies the dimension deficiency condition of the Finite Witness Theorem. Let $\mu,\tilde{\mu} \in \mathcal{P}_{\le n}(\mathbb{R}^d)$ be two fixed distributions. Let A be the set

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

and suppose that A is of full dimension. Then A contains a submanifold $B \times C$ of full dimension, where $B \subseteq \mathbb{S}^{d-1}$ and $C \subseteq (0, \infty)$. Thus, B and C are also of full dimension.

For any fixed $\boldsymbol{v} \in B$, the function $E^{\text{FSW}}(\mu; \boldsymbol{v}, \xi)$ is analytic on $(0, \infty)$ as a function of ξ, as can be seen in (36). Thus, the function

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

is also analytic on $(0, \infty)$. Since $f = 0$ on the set C of full dimension, $f = 0$ on all of $(0, \infty)$. By (33), this implies that

$$\mathcal{W}\big(\boldsymbol{v}^T \mu, \boldsymbol{v}^T \tilde{\mu}\big) = \sqrt{\mathbb{E}_\xi \Big[f(\xi)^2\Big]} = 0, \tag{37}$$

and thus $\boldsymbol{v}^T \mu = \boldsymbol{v}^T \tilde{\mu}$.

Since the above holds for all $\boldsymbol{v} \in B$, which is a set of full dimension, this implies that $\mu = \tilde{\mu}$. Hence, $E^{\text{FSW}}(\boldsymbol{X}, \boldsymbol{p}; \boldsymbol{v}, \xi)$ satisfies the dimension deficiency condition.

Lastly, note that $\dim\big(\mathcal{P}_{\leq n}(\mathbb{R}^d)\big) = nd + n - 1$ and $\dim\big(\mathcal{S}_{\leq n}(\mathbb{R}^d)\big) = nd$ and thus for $m \geq 2nd + 2n - 1$ and $m \geq 2nd + 1$ respectively, f qualifies for the Finite Witness Theorem on the domain $\mathcal{S}_{\leq n}(\mathbb{R}^d)$ and $\mathcal{P}_{\leq n}(\mathbb{R}^d)$ respectively. This finalizes our proof. $\square$

Theorem 4.4. [Proof on Page 22] *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.5. *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. $\mathcal{W}_p$. $\square$

Proof. Our proof of Theorem 4.4 consists of three steps. First, in Lemma B.6 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.7, 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

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

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

Lemma B.6. *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 $p \in [0, \infty]$, 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

$$0 < \theta_{t+1} \leq \tfrac{1}{2}\theta_t \leq 1 \quad \forall t \geq 1. \tag{38}$$

The set Ω contains a ball $B_r(0)$ by assumption. Choose $\boldsymbol{x} \neq 0$ in that ball. For $\theta \in [0, 1]$ we define

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

Note that for $1 \le p < \infty$

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

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

$$\frac{E(\mu(\theta_t)) - E(\delta_0)}{\mathcal{W}_p(\mu(\theta_t), \delta_0)} = \frac{1}{\|\boldsymbol{x}\|} \frac{E(\mu(\theta_t)) - E(\delta_0)}{\sqrt[p]{\theta_t}} = \frac{1}{\|\boldsymbol{x}\|} \frac{E(\mu(\theta_t))}{\sqrt[p]{\theta_t}}, \tag{39}$$

where for the last equality we used the homogeniety of E to show 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 above is uniformly bounded from above for all natural t. which implies that the exists a subsequence of θ_t for which this expression converges. Replacing θ_t with this subsequence, we note that this subsequence still satisfies (38), 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

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

Since $\frac{\theta_t}{\theta_{t-1}} \le \frac{1}{2}$, and x is contained in a ball in Ω. the measure $\tilde{\mu}_t$ is indeed in $\mathcal{P}_{\le n}(\Omega)$. We wish to lower-bound the p-Wasserstein distance from $\mu(\theta_t)$ to $\tilde{\mu}_t$ for $t \ge 2$. 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 non-zero 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. Therefore a transporting $\tilde{\mu}_t$ to $\mu(\theta_t)$ requires transporting at least $\theta_{t-1} - \theta_t$ mass from $\sqrt[p]{\frac{\theta_t}{\theta_{t-1}}}\boldsymbol{x}$ to 0, so that for all $1 \le p < \infty$

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

We obtained that

$$\mathcal{W}_p(\mu(\theta_t), \tilde{\mu}_t) \ge \sqrt[p]{\theta_t/2} \|\boldsymbol{x}\| \tag{40}$$

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

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

where (a) is by (40) 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.6. $\square$

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

Lemma B.7. *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 $0 < c \leq C < \infty$,

$$c \cdot \mathcal{W}_p(\mu, \tilde{\mu}) \leq \|E(\mu) - E(\tilde{\mu})\| \leq C \cdot \mathcal{W}_p(\mu, \tilde{\mu}), \qquad \forall \mu, \tilde{\mu} \in \mathcal{P}_{\leq n}(\Omega). \tag{41}$$

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

$$\tilde{E}(\mu) \coloneqq E(\mu) - E(0),$$

then E satisfies (41) if and only if $\tilde{E}$ satisfies (41).

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 $0 < \|\tilde{\mu}\|_{\mathcal{W}_p} \leq 1$,*

$$\left\| E\left(\frac{\tilde{\mu}}{\|\tilde{\mu}\|_{\mathcal{W}_p}}\right) - E(\tilde{\mu}) \right\| \leq C \cdot \left(1 - \|\tilde{\mu}\|_{\mathcal{W}_p}\right) \leq C \cdot \mathcal{W}_p(\mu, \tilde{\mu}). \tag{42}$$

Proof. By (41),

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

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 $\tilde{\mu} = \sum_{i=1}^n p_i \delta_{\tilde{\boldsymbol{x}}_i}$ 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

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

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 $p < \infty$,

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

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, (42) holds. $\square$

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

$$\begin{cases} \hat{E}(\mu) := \left[\|\mu\|_{\mathcal{W}_p}, \|\mu\|_{\mathcal{W}_p} E\left(\frac{\mu}{\|\mu\|_{\mathcal{W}_p}}\right) \right], & \mu \neq 0 \\ 0 & \mu = 0. \end{cases} \tag{43}$$

Clearly $\hat{E}$ is positively homogeneous. By Lemma B.6, $\hat{E}$ 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

$$\frac{\left\| \hat{E}(\mu_t) - \hat{E}(\tilde{\mu}_t) \right\|}{\mathcal{W}_p(\mu_t, \tilde{\mu}_t)} \xrightarrow[t\to\infty]{} L, \tag{44}$$

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 $\tilde{\mu}_t = 0$ 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 (44) reads as

$$\frac{\left\| \hat{E}(\mu_t) - \hat{E}(0) \right\|}{\mathcal{W}_p(\mu_t, 0)} = \frac{\|E(\mu_t) - E(0)\|}{\mathcal{W}_p(\mu_t, 0)} \xrightarrow[t\to\infty]{} L.$$

This contradicts the bi-Lipschitzness of E. Therefore, $\tilde{\mu}_t = 0$ 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

$$1 = \|\mu_t\|_{\mathcal{W}_p} \geq \|\tilde{\mu}_t\|_{\mathcal{W}_p} > 0 \quad \text{for all } t \geq 1. \tag{45}$$

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 (42). Therefore, by (44) combined with the fact that $\tilde{\mu}_t > 0 \; \forall t$, we must have that

$$\frac{\left\| \|\mu_t\|_{\mathcal{W}_p} E\left(\frac{\mu_t}{\|\mu_t\|_{\mathcal{W}_p}}\right) - \|\tilde{\mu}_t\|_{\mathcal{W}_p} E\left(\frac{\tilde{\mu}_t}{\|\tilde{\mu}_t\|_{\mathcal{W}_p}}\right) \right\|}{\mathcal{W}_p(\mu_t, \tilde{\mu}_t)} \xrightarrow[t\to\infty]{} \infty. \tag{46}$$

On the other hand,

$$\begin{aligned} &\left\| \|\mu_t\|_{\mathcal{W}_p} E\left(\frac{\mu_t}{\|\mu_t\|_{\mathcal{W}_p}}\right) - \|\tilde{\mu}_t\|_{\mathcal{W}_p} E\left(\frac{\tilde{\mu}_t}{\|\tilde{\mu}_t\|_{\mathcal{W}_p}}\right) \right\| \overset{\text{(a)}}{=} \left\| E(\mu_t) - \|\tilde{\mu}_t\|_{\mathcal{W}_p} E\left(\frac{\tilde{\mu}_t}{\|\tilde{\mu}_t\|_{\mathcal{W}_p}}\right) \right\| \\ &\overset{\text{(b)}}{\leq} \|E(\mu_t) - E(\tilde{\mu}_t)\| + \left\| E(\tilde{\mu}_t) - E\left(\frac{\tilde{\mu}_t}{\|\tilde{\mu}_t\|_{\mathcal{W}_p}}\right) \right\| + \left\| E\left(\frac{\tilde{\mu}_t}{\|\tilde{\mu}_t\|_{\mathcal{W}_p}}\right) - \|\tilde{\mu}_t\|_{\mathcal{W}_p} E\left(\frac{\tilde{\mu}_t}{\|\tilde{\mu}_t\|_{\mathcal{W}_p}}\right) \right\|, \end{aligned} \tag{47}$$

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,

$$\|E(\mu_t) - E(\tilde{\mu}_t)\| \leq C \cdot \mathcal{W}_p(\mu_t, \tilde{\mu}_t) \tag{48}$$

by (41). Second,

$$\left\| E(\tilde{\mu}_t) - E\left(\frac{\tilde{\mu}_t}{\|\tilde{\mu}_t\|_{\mathcal{W}_p}}\right) \right\| \leq C \cdot \mathcal{W}_p(\mu_t, \tilde{\mu}_t) \tag{49}$$

by (42). Lastly,

$$\left\| E\left(\frac{\tilde{\mu}_t}{\|\tilde{\mu}_t\|_{\mathcal{W}_p}}\right) - \|\tilde{\mu}_t\|_{\mathcal{W}_p} E\left(\frac{\tilde{\mu}_t}{\|\tilde{\mu}_t\|_{\mathcal{W}_p}}\right) \right\| = \left(1 - \|\tilde{\mu}_t\|_{\mathcal{W}_p}\right) \cdot \left\| E\left(\frac{\tilde{\mu}_t}{\|\tilde{\mu}_t\|_{\mathcal{W}_p}}\right) - 0 \right\|$$
$$= \left(1 - \|\tilde{\mu}_t\|_{\mathcal{W}_p}\right) \cdot \left\| E\left(\frac{\tilde{\mu}_t}{\|\tilde{\mu}_t\|_{\mathcal{W}_p}}\right) - E(0) \right\|$$
$$\overset{\text{(a)}}{\leq} \left(1 - \|\tilde{\mu}_t\|_{\mathcal{W}_p}\right) \cdot C \cdot \mathcal{W}_p\left(\frac{\tilde{\mu}_t}{\|\tilde{\mu}_t\|_{\mathcal{W}_p}}, 0\right) \tag{50}$$
$$= \left(1 - \|\tilde{\mu}_t\|_{\mathcal{W}_p}\right) \cdot C \cdot \left\| \frac{\tilde{\mu}_t}{\|\tilde{\mu}_t\|_{\mathcal{W}_p}} \right\|_{\mathcal{W}_p}$$
$$= C \cdot \left(1 - \|\tilde{\mu}_t\|_{\mathcal{W}_p}\right) \overset{\text{(b)}}{\leq} C \cdot \mathcal{W}_p(\mu_t, \tilde{\mu}_t),$$

where (a) is by (41) and (b) is by (42). Inserting (48)-(50) into (47) 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 (46).

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 (44) 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

$$1 - \|\tilde{\mu}_t\|_{\mathcal{W}_p} = \left| \|\mu_t\|_{\mathcal{W}_p} - \|\tilde{\mu}_t\|_{\mathcal{W}_p} \right| = o(d_t). \tag{51}$$

By the triangle inequality,

$$\|E(\mu_t) - E(\tilde{\mu}_t)\| \leq \tag{52}$$
$$\left\| E(\mu_t) - \|\tilde{\mu}_t\|_{\mathcal{W}_p} E\left(\frac{\tilde{\mu}_t}{\|\tilde{\mu}_t\|_{\mathcal{W}_p}}\right) \right\| + \left\| \|\tilde{\mu}_t\|_{\mathcal{W}_p} E\left(\frac{\tilde{\mu}_t}{\|\tilde{\mu}_t\|_{\mathcal{W}_p}}\right) - E\left(\frac{\tilde{\mu}_t}{\|\tilde{\mu}_t\|_{\mathcal{W}_p}}\right) \right\| + \left\| E\left(\frac{\tilde{\mu}_t}{\|\tilde{\mu}_t\|_{\mathcal{W}_p}}\right) - E(\tilde{\mu}_t) \right\|. \tag{53}$$

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

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

$$\left\| E(\mu_t) - \|\tilde{\mu}_t\|_{\mathcal{W}_p} E\left(\frac{\tilde{\mu}_t}{\|\tilde{\mu}_t\|_{\mathcal{W}_p}}\right) \right\| = \left\| \|\mu_t\|_{\mathcal{W}_p} E\left(\frac{\mu_t}{\|\mu_t\|_{\mathcal{W}_p}}\right) - \|\tilde{\mu}_t\|_{\mathcal{W}_p} E\left(\frac{\tilde{\mu}_t}{\|\tilde{\mu}_t\|_{\mathcal{W}_p}}\right) \right\| \tag{54}$$
$$= \left\| \hat{E}(\mu_t) - \hat{E}(\tilde{\mu}_t) \right\|,$$

which is $o(d_t)$ by (44). For the second term,

$$\begin{aligned}
&\left\| \|\tilde{\mu}_t\|_{\mathcal{W}_p} E\left(\frac{\tilde{\mu}_t}{\|\tilde{\mu}_t\|_{\mathcal{W}_p}}\right) - E\left(\frac{\tilde{\mu}_t}{\|\tilde{\mu}_t\|_{\mathcal{W}_p}}\right) \right\| \\
&= \left(1 - \|\tilde{\mu}_t\|_{\mathcal{W}_p}\right) \cdot \left\| E\left(\frac{\tilde{\mu}_t}{\|\tilde{\mu}_t\|_{\mathcal{W}_p}}\right) \right\| \\
&= \left(1 - \|\tilde{\mu}_t\|_{\mathcal{W}_p}\right) \cdot \left\| E\left(\frac{\tilde{\mu}_t}{\|\tilde{\mu}_t\|_{\mathcal{W}_p}}\right) - 0 \right\| \\
&\overset{(a)}{\leq} \left(1 - \|\tilde{\mu}_t\|_{\mathcal{W}_p}\right) \cdot C \cdot \mathcal{W}\left(\frac{\tilde{\mu}_t}{\|\tilde{\mu}_t\|_{\mathcal{W}_p}}, 0\right) \\
&= \left(1 - \|\tilde{\mu}_t\|_{\mathcal{W}_p}\right) \cdot C \cdot \left\| \frac{\tilde{\mu}_t}{\|\tilde{\mu}_t\|_{\mathcal{W}_p}} \right\|_{\mathcal{W}_p} \\
&= \left(1 - \|\tilde{\mu}_t\|_{\mathcal{W}_p}\right) C \overset{(b)}{=} o(d_t),
\end{aligned} \tag{55}$$

where (a) is by (41) and (b) is by (51).

Finally, by (42),

$$\left\| E\left(\frac{\tilde{\mu}_t}{\|\tilde{\mu}_t\|_{\mathcal{W}_p}}\right) - E(\tilde{\mu}_t) \right\| \leq C \cdot \left(1 - \|\tilde{\mu}_t\|_{\mathcal{W}_p}\right) = o(d_t). \tag{56}$$

Therefore, by (54)-(56) and (52), 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.7. ☐

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 $E : \mathcal{P}_{\leq n}(\Omega) \to \mathbb{R}^m$ with $n \geq 2$, 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.7, and thus there exist two sequences of distributions $\mu_t, \tilde{\mu}_t \in \mathcal{P}_{\leq n}(\Omega_0 - x_0)$, $t \geq 1$ 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 $\{\mu_t + \boldsymbol{x}_0\}_{t \geq 1}$ 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. [Proof in Page 28] *Let $E : \mathcal{P}_{\leq n}(\mathbb{R}^d) \to \mathbb{R}^m$ be injective and positively homogeneous. Let Δ^n be the probability simplex in $\mathbb{R}^n$. Suppose that the function $E(\boldsymbol{X}, \boldsymbol{w}) : \mathbb{R}^{d\times n} \times \Delta^n \to \mathbb{R}^m$ is piecewise linear in $\boldsymbol{X}$ for any fixed $\boldsymbol{w}$. Then for any fixed $\boldsymbol{w}, \tilde{\boldsymbol{w}} \in \Delta^n$, there exist constants $c, C > 0$ 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 $\boldsymbol{p}, \boldsymbol{q} \in \Delta^n$ there exists a constant $\beta > 0$ such that for all $\boldsymbol{X}, \boldsymbol{Y} \in \mathbb{R}^{d\times n}$,

$$\mathcal{W}_1((\boldsymbol{X}, \boldsymbol{p}), (\boldsymbol{Y}, \boldsymbol{q})) \geq \beta \cdot \mathcal{W}_\infty((\boldsymbol{X}, \boldsymbol{p}), (\boldsymbol{Y}, \boldsymbol{q})). \tag{57}$$

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

$$\mathcal{W}_1((\boldsymbol{X}, \boldsymbol{p}), (\boldsymbol{Y}, \boldsymbol{q})) \leq \mathcal{W}_p((\boldsymbol{X}, \boldsymbol{p}), (\boldsymbol{Y}, \boldsymbol{q})) \leq \mathcal{W}_\infty((\boldsymbol{X}, \boldsymbol{p}), (\boldsymbol{Y}, \boldsymbol{q})).$$

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

$$\mathcal{W}_1((\boldsymbol{X}, \boldsymbol{p}), (\boldsymbol{Y}, \boldsymbol{q})) = \min_{\pi \in \Pi(\boldsymbol{p}, \boldsymbol{q})} \sum_{i,j\in[n]} \pi_{ij} \Big\| \boldsymbol{x}^{(i)} - \boldsymbol{y}^{(j)} \Big\|, \tag{58}$$

where the set $\Pi(\boldsymbol{p}, \boldsymbol{q})$ of admissible transport plans from $(\boldsymbol{X}, \boldsymbol{p})$ to $(\boldsymbol{Y}, \boldsymbol{q})$ is given by

$$\Pi(\boldsymbol{p}, \boldsymbol{q}) = \left\{ \pi \in [0,1]^{n\times n} \;\middle|\; \forall i \in [n] \sum_{j=1}^{n} \pi_{ij} = p_i \bigwedge \forall j \in [n] \sum_{i=1}^{n} \pi_{ij} = q_j \right\}.$$

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

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:

$$\widetilde{\mathcal{W}}_1((\boldsymbol{X}, \boldsymbol{p}), (\boldsymbol{Y}, \boldsymbol{q})) \coloneqq \min_{\pi \in \Pi(\boldsymbol{p}, \boldsymbol{q})} \sum_{i,j\in[n]} \pi_{ij} \Big\| \boldsymbol{x}^{(i)} - \boldsymbol{y}^{(j)} \Big\|_1. \tag{59}$$

Note that since

$$\|\boldsymbol{x}\|_2 \leq \|\boldsymbol{x}\|_1 \leq \sqrt{d}\|\boldsymbol{x}\|_2 \quad \forall \boldsymbol{x} \in \mathbb{R}^d, \tag{60}$$

we have

$$\mathcal{W}_1((\boldsymbol{X}, \boldsymbol{p}), (\boldsymbol{Y}, \boldsymbol{q})) \leq \widetilde{\mathcal{W}}_1((\boldsymbol{X}, \boldsymbol{p}), (\boldsymbol{Y}, \boldsymbol{q})) \leq \sqrt{d} \cdot \mathcal{W}_1((\boldsymbol{X}, \boldsymbol{p}), (\boldsymbol{Y}, \boldsymbol{q})). \tag{61}$$

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

$$f(\boldsymbol{X}, \boldsymbol{Y}) \coloneqq \begin{bmatrix} \|E(\boldsymbol{X}, \boldsymbol{p}) - E(\boldsymbol{Y}, \boldsymbol{q})\|_1 \\ \widetilde{\mathcal{W}}_1((\boldsymbol{X}, \boldsymbol{p}), (\boldsymbol{Y}, \boldsymbol{q})) \end{bmatrix}.$$

To achieve the desired result, we first show that f is piecewise linear in $(\boldsymbol{X}, \boldsymbol{Y})$. The first component of f, $\|E(\boldsymbol{X}, \boldsymbol{p}) - E(\boldsymbol{Y}, \boldsymbol{q})\|_1$, 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 $\widetilde{\mathcal{W}}_1((\boldsymbol{X}, \boldsymbol{p}), (\boldsymbol{Y}, \boldsymbol{q}))$ is also piecewise linear. For any fixed $\boldsymbol{X}$ and $\boldsymbol{Y}$, the optimization problem in (59) is a linear program in π, with the set of feasible solutions being the compact polytope $\Pi(\boldsymbol{p}, \boldsymbol{q})$[3]. Thus, the optimal solution must be attained at one of the vertices of $\Pi(\boldsymbol{p}, \boldsymbol{q})$. As any polytope has a finite number of vertices[4], let $\pi^{(1)}, \ldots, \pi^{(K)}$ be the vertices of $\Pi(\boldsymbol{p}, \boldsymbol{q})$, and recall that these vertices do not depend on $(\boldsymbol{X}, \boldsymbol{Y})$. Therefore, (59) can be reformulated as

$$\widetilde{\mathcal{W}}_1((\boldsymbol{X}, \boldsymbol{p}), (\boldsymbol{Y}, \boldsymbol{q})) = \min_{k\in[K]} \sum_{i,j\in[n]} \pi_{ij}^{(k)} \Big\| \boldsymbol{x}^{(i)} - \boldsymbol{y}^{(j)} \Big\|_1. \tag{62}$$

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

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

From (62) it can be seen that $\widetilde{\mathcal{W}}_1((\boldsymbol{X},\boldsymbol{p}),(\boldsymbol{Y},\boldsymbol{q}))$ is piecewise linear in $(\boldsymbol{X},\boldsymbol{Y})$, 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 $f(\boldsymbol{X},\boldsymbol{Y})$ is piecewise linear.

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

$$A \coloneqq \left\{ f(\boldsymbol{X},\boldsymbol{Y}) \;\middle|\; \boldsymbol{X},\boldsymbol{Y} \in \mathbb{R}^{d\times n} \right\}.$$

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 $\boldsymbol{X},\boldsymbol{Y}$ such that $E(\boldsymbol{X},\boldsymbol{p}) = E(\boldsymbol{Y},\boldsymbol{q})$ and $\widetilde{\mathcal{W}}_1((\boldsymbol{X},\boldsymbol{p}),(\boldsymbol{Y},\boldsymbol{q})) = 1$, which contradicts the injectivity of E. Similarly, if $(1,0) \in A$, then there exist $\boldsymbol{X},\boldsymbol{Y}$ such that on one hand $\widetilde{\mathcal{W}}_1((\boldsymbol{X},\boldsymbol{p}),(\boldsymbol{Y},\boldsymbol{q})) = 0$, which implies that $(\boldsymbol{X},\boldsymbol{p})$ and $(\boldsymbol{Y},\boldsymbol{q})$ represent the same distribution, but on the other hand $E(\boldsymbol{X},\boldsymbol{p}) \neq E(\boldsymbol{Y},\boldsymbol{q})$. 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.

Now, let $\boldsymbol{X},\boldsymbol{Y} \in \mathbb{R}^{d\times n}$ such that $\mathcal{W}_1((\boldsymbol{X},\boldsymbol{p}),(\boldsymbol{Y},\boldsymbol{q})) > 0$. Then by (61), $\widetilde{\mathcal{W}}_1((\boldsymbol{X},\boldsymbol{p}),(\boldsymbol{Y},\boldsymbol{q})) > 0$. Denote

$$\nu \coloneqq \left[\widetilde{\mathcal{W}}_1((\boldsymbol{X},\boldsymbol{p}),(\boldsymbol{Y},\boldsymbol{q}))\right]^{-1}.$$

Then

$$\widetilde{\mathcal{W}}_1((\nu\boldsymbol{X},\boldsymbol{p}),(\nu\boldsymbol{Y},\boldsymbol{q})) = 1,$$

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

$$\begin{aligned}
&\frac{\|E(\boldsymbol{X},\boldsymbol{p}) - E(\boldsymbol{Y},\boldsymbol{q})\|_1}{\widetilde{\mathcal{W}}_1((\boldsymbol{X},\boldsymbol{p}),(\boldsymbol{Y},\boldsymbol{q}))} = \frac{\|E(\nu\boldsymbol{X},\boldsymbol{p}) - E(\nu\boldsymbol{Y},\boldsymbol{q})\|_1}{\widetilde{\mathcal{W}}_1((\nu\boldsymbol{X},\boldsymbol{p}),(\nu\boldsymbol{Y},\boldsymbol{q}))} = \|E(\nu\boldsymbol{X},\boldsymbol{p}) - E(\nu\boldsymbol{Y},\boldsymbol{q})\|_1 \\
&= \left\| \begin{bmatrix} \|E(\nu\boldsymbol{X},\boldsymbol{p}) - E(\nu\boldsymbol{Y},\boldsymbol{q})\|_1 \\ 1 \end{bmatrix} - \begin{bmatrix} 0 \\ 1 \end{bmatrix} \right\|_2 \\
&= \left\| \begin{bmatrix} \|E(\nu\boldsymbol{X},\boldsymbol{p}) - E(\nu\boldsymbol{Y},\boldsymbol{q})\|_1 \\ \widetilde{\mathcal{W}}_1((\nu\boldsymbol{X},\boldsymbol{p}),(\nu\boldsymbol{Y},\boldsymbol{q})) \end{bmatrix} - \begin{bmatrix} 0 \\ 1 \end{bmatrix} \right\|_2 \\
&= \left\| f(\nu\boldsymbol{X},\nu\boldsymbol{Y}) - \begin{bmatrix} 0 \\ 1 \end{bmatrix} \right\|_2 \geq \operatorname{dist}(A,\{(0,1),(1,0)\}) = \alpha.
\end{aligned} \tag{63}$$

Therefore,

$$\begin{aligned}
\frac{\|E(\boldsymbol{X},\boldsymbol{p}) - E(\boldsymbol{Y},\boldsymbol{q})\|_2}{\mathcal{W}_1((\boldsymbol{X},\boldsymbol{p}),(\boldsymbol{Y},\boldsymbol{q}))} &\overset{\text{(a)}}{\geq} \tfrac{1}{\sqrt{m}} \frac{\|E(\boldsymbol{X},\boldsymbol{p}) - E(\boldsymbol{Y},\boldsymbol{q})\|_1}{\mathcal{W}_1((\boldsymbol{X},\boldsymbol{p}),(\boldsymbol{Y},\boldsymbol{q}))} \\
&\overset{\text{(b)}}{\geq} \tfrac{1}{\sqrt{m}} \frac{\|E(\boldsymbol{X},\boldsymbol{p}) - E(\boldsymbol{Y},\boldsymbol{q})\|_1}{\widetilde{\mathcal{W}}_1((\boldsymbol{X},\boldsymbol{p}),(\boldsymbol{Y},\boldsymbol{q}))} \geq \tfrac{\alpha}{\sqrt{m}},
\end{aligned} \tag{64}$$

where (a) is by the $\ell_1 - \ell_2$ norm inequality over $\mathbb{R}^m$, (b) is by (61), and (c) is by (63).

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

$$\nu \coloneqq \|E(\boldsymbol{X},\boldsymbol{p}) - E(\boldsymbol{Y},\boldsymbol{q})\|_1^{-1}.$$

Since E is homogeneous,

$$\|E(\nu\boldsymbol{X},\boldsymbol{p}) - E(\nu\boldsymbol{Y},\boldsymbol{q})\|_1 = 1$$

and thus

$$
\begin{aligned}
&\frac{\widetilde{\mathcal{W}}_1((\boldsymbol{X},\boldsymbol{p}),(\boldsymbol{Y},\boldsymbol{q}))}{\|E(\boldsymbol{X},\boldsymbol{p})-E(\boldsymbol{Y},\boldsymbol{q})\|_1}=\frac{\widetilde{\mathcal{W}}_1((\nu\boldsymbol{X},\boldsymbol{p}),(\nu\boldsymbol{Y},\boldsymbol{q}))}{\|E(\nu\boldsymbol{X},\boldsymbol{p})-E(\nu\boldsymbol{Y},\boldsymbol{q})\|_1}=\widetilde{\mathcal{W}}_1((\nu\boldsymbol{X},\boldsymbol{p}),(\nu\boldsymbol{Y},\boldsymbol{q}))\\
&=\left\|\begin{bmatrix}1\\ \widetilde{\mathcal{W}}_1((\nu\boldsymbol{X},\boldsymbol{p}),(\nu\boldsymbol{Y},\boldsymbol{q}))\end{bmatrix}-\begin{bmatrix}1\\0\end{bmatrix}\right\|_2\\
&=\left\|\begin{bmatrix}\|E(\nu\boldsymbol{X},\boldsymbol{p})-E(\nu\boldsymbol{Y},\boldsymbol{q})\|_1\\ \widetilde{\mathcal{W}}_1((\nu\boldsymbol{X},\boldsymbol{p}),(\nu\boldsymbol{Y},\boldsymbol{q}))\end{bmatrix}-\begin{bmatrix}1\\0\end{bmatrix}\right\|_2\\
&=\left\|f(\nu\boldsymbol{X},\nu\boldsymbol{Y})-\begin{bmatrix}1\\0\end{bmatrix}\right\|_2\geq\operatorname{dist}(A,\{(0,1),(1,0)\})=\alpha.
\end{aligned}
\tag{65}
$$

Therefore,

$$
\begin{aligned}
\frac{\|E(\boldsymbol{X},\boldsymbol{p})-E(\boldsymbol{Y},\boldsymbol{q})\|_2}{\mathcal{W}_1((\boldsymbol{X},\boldsymbol{p}),(\boldsymbol{Y},\boldsymbol{q}))}&\overset{\text{(a)}}{\leq}\frac{\|E(\boldsymbol{X},\boldsymbol{p})-E(\boldsymbol{Y},\boldsymbol{q})\|_1}{\mathcal{W}_1((\boldsymbol{X},\boldsymbol{p}),(\boldsymbol{Y},\boldsymbol{q}))}\\
&\overset{\text{(b)}}{\leq}\sqrt{d}\frac{\|E(\boldsymbol{X},\boldsymbol{p})-E(\boldsymbol{Y},\boldsymbol{q})\|_1}{\widetilde{\mathcal{W}}_1((\boldsymbol{X},\boldsymbol{p}),(\boldsymbol{Y},\boldsymbol{q}))}\overset{\text{(c)}}{\leq}\frac{\sqrt{d}}{\alpha},
\end{aligned}
\tag{66}
$$

where (a) is since $\|\cdot\|_2\leq\|\cdot\|_1$, (b) is by (61), and (c) is by (65). Hence, from (64) and (66), we have

$$
\frac{\alpha}{\sqrt{m}}\leq\frac{\|E(\boldsymbol{X},\boldsymbol{p})-E(\boldsymbol{Y},\boldsymbol{q})\|_2}{\mathcal{W}_1((\boldsymbol{X},\boldsymbol{p}),(\boldsymbol{Y},\boldsymbol{q}))}\leq\frac{\sqrt{d}}{\alpha}.
\tag{67}
$$

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 (57) holds with some constant $\beta>0$ assuming that $\boldsymbol{p}$ and $\boldsymbol{q}$ 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,

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

where (a) is by (61); (b) is by (62); (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 $\Pi(\boldsymbol{p},\boldsymbol{q}) \supseteq \left\{\pi^{(k)}\right\}_{k=1}^{[K]}$; and (j) is by the definition of $\mathcal{W}_\infty$. Hence, (57) 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 Empirical distortion evaluation

In some instances during this experiment, particularly with high embedding dimensions m or a large number of points n, the PSWE method failed due to insufficient memory. To mitigate this issue, we computed the PSWE embeddings for each input multiset sequentially rather than processing them in batches. This approach resolved most cases, although memory limitations persisted in the instances marked as *OOM* in Table 1.

In the particularly challenging case of $n = 2047$ with $m = 200$ or 1000 (right-hand half of the top row in Table 1), applying our method to entire batches also resulted in insufficient memory. We resolved this by using our implementation's support for processing slices sequentially instead of in parallel, thereby parallelizing over the embedding dimension m rather than the batch size of 6000. This adjustment allowed us to complete all test cases without the need for sequential batch processing.

Due to the different parallelization strategies, a fair comparison of computation times between the two methods in this experiment is not possible.

C.2 Learning to approximate the 1-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`).