$\mathcal{O}(\log N)$ Latent Dimension Suffices for Universal Approximation of Permutation-invariant Function

Min Zhou [1] **Enming Liang** [1] **Minghua Chen** [1 2]

Abstract

Learning permutation-invariant functions over sets of N elements, where the output is independent of the input ordering, is fundamental to many deep learning applications. While sum-decomposable architectures like DeepSets offer universal approximation for such functions, existing constructive bounds require a latent dimension of $\mathcal{O}(N)$, posing a significant scalability bottleneck. We break this barrier for *Wasserstein-stable* functions, i.e., those that are Lipschitz continuous with respect to the Wasserstein-1 metric on input distributions. We constructively prove that a latent dimension of $\mathcal{O}\left(C_D\varepsilon^{-(3D+4)}\log(N+\varepsilon^{-D})\right)$ suffices for uniform approximation as $\varepsilon \to 0$, where D is the element dimension and C_D is a constant depending only on D. We first discretize the input space into a finite net of measures with covering number polynomial in N. We then embed it via a multiscale Random Fourier Feature encoder that guarantees both Lipschitz stability and Hölder separation. Finally, we recover the target function via a McShane-extended Hölder decoder. This result advances the theoretical understanding of the expressivity and scalability of set-based neural architectures.

1. Introduction

Learning from unordered sets is a fundamental challenge in modern machine learning. Unlike images or text, which have a fixed grid or sequential structure (O'shea & Nash, 2015), many real-world data types are naturally represented as sets of N exchangeable elements. Prominent examples include point clouds in 3D vision (Qi et al., 2017; Guo et al., 2020) and scenario samples in data-driven optimization (Zhang et al., 2023; Zhou et al., 2025). A key requirement for neural networks in this domain is *permutation invariance*: the output must remain unchanged regardless of the order in which input elements are presented.

To achieve this symmetry, sum-decomposable architectures, most notably DeepSets (Zaheer et al., 2017), have become the standard approach. These models apply a shared transformation to each set element, sum the resulting latent representations, and process the aggregate to produce the final output. This design is both simple and theoretically principled: any continuous permutation-invariant function can be approximated arbitrarily well by such an architecture (Zaheer et al., 2017; Schmidt-Hieber, 2021; Wang et al., 2024; Tabaghi & Wang, 2024; Liu et al., 2025).

However, a critical gap persists between the existence of these approximations and their practical scalability. As summarized in Table 1, standard constructive proofs require the latent dimension—the width of the network before the summation—to scale at least linearly with the set size N (Zaheer et al., 2017; Wagstaff et al., 2019). This $\mathcal{O}(N)$ scaling poses a significant theoretical bottleneck. For large-scale applications such as high-resolution point cloud processing or large-batch stochastic optimization (Shapiro et al., 2021), requiring a latent width proportional to N becomes computationally prohibitive.

Yet in practice, latent dimensions much smaller than $\mathcal{O}(N)$ are routinely used without significant loss of expressivity (Qi et al., 2017; Zaheer et al., 2017). This gap suggests that *exact injectivity*, while necessary for distinguishing arbitrary multisets, is overly stringent for approximating well-behaved functions. For many tasks, the target function satisfies stability properties that make worst-case capacity unnecessary. This observation motivates a natural question: *for which function classes can we provably break the $\mathcal{O}(N)$ barrier?*

In this paper, we answer this question for Wasserstein-stable functions—a broad class of permutation-invariant functions that are Lipschitz continuous with respect to the Wasserstein-1 metric. We prove that a latent dimension logarithmic in N suffices for uniform approximation, advancing our understanding of the expressivity of set-based neural networks

[1]Department of Data Science, City University of Hong Kong [2]School of Data Science, The Chinese University of Hong Kong, Shenzhen. Correspondence to: Minghua Chen <minghua@cuhk.edu.cn>.

Proceedings of the 43^{rd} International Conference on Machine Learning, Seoul, South Korea. PMLR 306, 2026.

Table 1. Existing studies on latent embedding dimension for learning permutation-invariant functions.

Study	Input Dim.	Guarantees	Latent Dim. (m)	Sufficient Conditions
(Zaheer et al., 2017)	$D = 1$	Exact Rep.	$\mathcal{O}(N)$	–
(Wagstaff et al., 2019)	$D = 1$	Exact Rep.	$\mathcal{O}(N)$	–
(Qi et al., 2017)	$D \geq 1$	Uni. Approx.	$\mathcal{O}(\varepsilon^{-D})$	Hausdorff continuous
(Segol & Lipman, 2019)	$D \geq 1$	Uni. Approx.	$\mathcal{O}(N^D)$	–
(Wang et al., 2024)	$D \geq 1$	Exact Rep.	$\mathcal{O}(N^4 D^2)$	–
(Amir et al., 2023)	$D \geq 1$	Exact Rep.	$\mathcal{O}(ND)$	–
(Tabaghi & Wang, 2024)	$D \geq 1$	Exact Rep.	$\mathcal{O}(ND)$	Identifiable multisets
This Paper	$D \geq 1$	**Uni. Approx.**	$\mathcal{O}\Big(C_D \varepsilon^{-(3D+4)} \log(N + \varepsilon^{-D})\Big)$	**Wasserstein Stability**

[1] **Abbreviations:** Dim. = Dimension; Rep. = Representation; Uni. Approx. = Universal Approximation.
[2] Previous works establish that for general or identifiable multisets, the latent dimension must scale at least linearly with set size N. An exception is PointNet (Qi et al., 2017), which relies on **Hausdorff continuity** to achieve an $\mathcal{O}(\varepsilon^{-D})$ latent dimension. However, Hausdorff continuity discards multiplicity information and is therefore unsuitable for mass-sensitive or statistical applications.
[3] We consider a function class satisfying **Wasserstein stability**, a distinct setting that respects the underlying probability measure. We prove that this stability condition suffices to guarantee universal approximation with latent width $\mathcal{O}(\varepsilon^{-(3D+4)} \log(N + \varepsilon^{-D}))$, achieving the approximation rate in Theorem 1. This bound is logarithmic in N when ε and D are fixed. Whether this dimension is necessary (i.e., whether a matching lower bound holds) remains an open question.
[4] Unlike the classical MLP universal approximation theorem (Hornik et al., 1989), which guarantees existence without specifying the required width, our universal approximation result is constructive: it provides an explicit latent dimension bound for Wasserstein-stable permutation-invariant functions.
[5] Our width bound also depends on the confidence level through the probabilistic construction, but this term is dominated by $\log N$ for fixed accuracy and confidence. See Section 4 for the complete bound with explicit constants.

and demonstrating that they are far more scalable than prior bounds suggest. Our analysis adopts a geometric perspective, treating input sets not as collections of vectors but as discrete probability measures in a Wasserstein space. By exploiting the geometric structure of this space, we make the following contributions:

▷ We derive a tight bound on the metric entropy (logarithm of the covering number) of the input space. By modeling input sets as discrete probability measures, we show that the metric entropy scales only logarithmically with the set size N. This reveals that the intrinsic "geometric volume" of the input space is far smaller than its combinatorial description suggests.

▷ Building on this insight, we provide a constructive approximation proof showing that a latent width logarithmic in N suffices for uniformly approximating Wasserstein-stable permutation-invariant functions. Our construction consists of two components: (i) a multiscale Random Fourier Feature mean encoder that is globally Lipschitz and separates points on the covering net of Wasserstein space with high probability; and (ii) a Hölder decoder constructed via the McShane extension theorem. The final uniform approximation error is controlled by combining the covering property with a triangle inequality argument.

Together, these results provide theoretical justification for the empirical success of compact set encoders and establish foundations for designing scalable network architectures.

2. Related Work

2.1. Permutation-invariant Architectures

Recently, the development of neural architectures for processing set-structured data has been an active research area (Dwivedi et al., 2021; Maron et al., 2018; Mialon et al., 2021). To effectively handle unordered sets, researchers have proposed diverse mechanisms that enforce permutation invariance while modeling complex element-wise dependencies. We categorize these mainstream approaches based on their aggregation strategies.

Sum-decomposition. Symmetric pooling operations are the standard mechanism for ensuring permutation invariance. Seminal frameworks such as DeepSets (Zaheer et al., 2017) and PointNet (Qi et al., 2017) leverage the Kolmogorov-Arnold representation theorem (Khesin & Tabachnikov, 2014) to establish that sum-decomposable architectures are universal approximators for continuous permutation-invariant functions. By processing elements independently before a global pooling step, the computational complexity of these methods scales linearly with set size ($\mathcal{O}(N)$ operations for fixed width), making them the predominant choice for large-scale applications.

Attention mechanisms. To capture complex dependencies between elements, attention-based models enable global information exchange (Vaswani et al., 2017; Liu et al., 2026; Li et al., 2025). Set Transformers adapt the standard Transformer architecture by removing positional encodings, al-

lowing every element to attend to every other and build context-aware representations before aggregation (Lee et al., 2019; Zhao et al., 2021). Slot Attention uses a competitive iterative process to group elements into distinct clusters (slots), effectively decomposing the input set into object-like components (Locatello et al., 2020). However, the pairwise interactions in these mechanisms typically incur quadratic complexity ($\mathcal{O}(N^2)$), limiting their scalability to large inputs.

Graph-based architectures. Explicit modeling of relational structure has been explored to capture dependencies beyond global statistics. Relation Networks (Hu et al., 2018; Sung et al., 2018; Hu et al., 2019) and Graph Neural Networks (GNNs) (Scarselli et al., 2008; Wu et al., 2020; Xu et al., 2018) treat the set as a fully connected graph, computing pairwise or neighborhood-based features to refine element representations. Janossy Pooling generalizes sum-decomposition by averaging over tractable subsets of permutations (Murphy et al., 2018). However, these structure-aware approaches often incur quadratic or combinatorial costs, making them less practical for large or high-dimensional sets.

Dimensionality bottleneck. Despite the empirical success of these architectures, a fundamental theoretical question remains: *what is the minimum latent dimension required for universal approximation?* Resolving this question is essential for establishing rigorous guarantees on representational capacity, guiding architecture design beyond heuristics, and enabling principled complexity analysis.

2.2. Theoretical Analysis of Latent Dimension

The minimum latent dimension required for universal approximation of permutation-invariant architectures has attracted significant recent attention (Maron et al., 2019a;b; Dym & Gortler, 2025; Amir et al., 2023; Zweig & Bruna, 2022). We summarize key theoretical results in Table 1.

Current analyses have established sufficiency conditions for both exact injectivity and universal approximation. For scalar inputs, (Wagstaff et al., 2019) proved that a latent width of $m \geq N$ is necessary to distinguish all multisets of size N, while (Zaheer et al., 2017) showed that $m = N + 1$ suffices for exact representation. In higher dimensions ($D > 1$), (Wang et al., 2024) demonstrated that DeepSets with a width polynomial in N and D can approximate arbitrary continuous permutation-invariant functions. (Tabaghi & Wang, 2024) extended this by proving that $m \approx 2DN$ suffices to distinguish identifiable multisets, providing a constructive bound for exact recovery. (Dym & Gortler, 2025) established that $m = 2ND + 1$ suffices to separate orbits, enabling exact representation of any continuous permutation-invariant function on a data manifold. Collectively, these results establish a *linear barrier*: to guarantee exact representation, the latent width must scale linearly with set size, imposing a prohibitive computational cost for large-scale applications.

This linear dependence stems from two strict algebraic requirements. First, existing works prioritize *exact injectivity*—the ability to uniquely distinguish every theoretically distinct multiset, regardless of geometric similarity. This reduces the problem to combinatorial separation, requiring the latent space to have sufficient degrees of freedom (typically matching N) to encode all element coordinates independently. Second, by targeting *arbitrary* continuous functions without stability constraints, these bounds must accommodate worst-case mappings that are highly sensitive to infinitesimal perturbations.

In this work, we show that the linear barrier can be circumvented for Wasserstein-stable functions by shifting from combinatorial injectivity to metric entropy, proving that a latent dimension logarithmic in N suffices. This does not contradict existing lower bounds, which apply to exact representation rather than approximation. We provide further discussion of the measure-theoretic universality literature and its relationship to our work in Appendix A.

3. Problem Setup and Theoretical Background

3.1. Problem Setup: Invariance and Architectures

Let $\mathcal{K} \subset \mathbb{R}^D$ be a compact domain, e.g., $[0,1]^D$, for analysis without loss of generality. We consider the problem of learning a function f defined on a *multiset* $\mathbf{x} = \{x_1, \ldots, x_N\}$ with $x_i \in \mathcal{K}$. A multiset is an unordered collection in which repeated elements are allowed, and their multiplicities matter. Without loss of generality, we assume f is scalar-valued[1]. Since the input is a multiset, the target function depends only on the composition of the set, not on element ordering. To formalize this symmetry, let $\mathfrak{S}_N$ denote the symmetric group of permutations on $\{1, \ldots, N\}$. We require f to be *permutation-invariant*:

$$f(x_1, \ldots, x_N) = f(x_{\pi(1)}, \ldots, x_{\pi(N)}), \quad \forall \pi \in \mathfrak{S}_N. \quad (1)$$

To learn such functions, we focus on sum-decomposable model structure, known as DeepSets (Zaheer et al., 2017). This architecture ensures invariance by processing elements independently and aggregating via a commutative operation. We adopt mean-pooling to align with the normalized nature of probability measures. Specifically, a sum-decomposable function takes the form:

$$h(\mathbf{x}) = \rho\left(\frac{1}{N}\sum_{i=1}^{N}\phi(x_i)\right), \quad (2)$$

[1]The analysis extends to vector-valued mappings by applying the derived bounds component-wise to each output dimension.

where $\phi : \mathcal{K} \to \mathbb{R}^m$ is the *encoder* that maps elements to a latent feature space, and $\rho : \mathbb{R}^m \to \mathbb{R}$ is the *decoder* that maps the aggregated representation to the output.

The latent dimension m is the critical bottleneck for model complexity. Existing theory establishes that *exact representation* of arbitrary multisets requires $m = \mathcal{O}(N)$ (Wagstaff et al., 2019; Tabaghi & Wang, 2024). This linear scaling renders standard DeepSets theoretically impractical for applications where N is large (e.g., $N > 10^5$). Yet in practice, latent dimensions much smaller than $\mathcal{O}(N)$ are routinely used without significant loss of expressivity (Qi et al., 2017; Zaheer et al., 2017). This gap suggests that *exact injectivity*, while necessary for distinguishing arbitrary multisets, is overly pessimistic for approximating well-behaved functions. In the following sections, we prove that a latent dimension logarithmic in N suffices for universal approximation of Wasserstein-stable functions.

3.2. Measure-Theoretic Representation

Existing lower bounds on latent dimension require exact separation of all distinct multisets (Zaheer et al., 2017; Wang et al., 2024). However, this strict injectivity is unnecessary for Wasserstein-stable functions, where discriminating between infinitesimally close inputs incurs prohibitive capacity costs without improving approximation quality. We instead adopt a measure-theoretic representation that identifies sets with empirical measures, equipping the domain with the geometric structure induced by Wasserstein distances. This perspective enables us to exploit the intrinsic compactness of probability space and derive sharper approximation bounds.

We represent any input set $\mathbf{x}$ as an empirical measure, as illustrated in Figure 1:

$$\mu_{\mathbf{x}} = \frac{1}{N} \sum_{i=1}^{N} \delta_{x_i}, \tag{3}$$

where δ_{x_i} denotes the Dirac mass at x_i.[2] To ensure this identification is rigorous, we invoke the following result, which guarantees that the measure representation preserves all information in the permutation-invariant structure.

Proposition 3.1 (Fundamental Identification)**.** *A function $f : \mathcal{K}^N \to \mathbb{R}$ is permutation-invariant if and only if there exists a unique function $F : \mathcal{P}_N(\mathcal{K}) \to \mathbb{R}$ such that $f(\mathbf{x}) = F(\mu_{\mathbf{x}})$, where $\mathcal{P}_N(\mathcal{K})$ denotes the set of empirical measures of the form $\frac{1}{N}\sum_{i=1}^{N} \delta_{x_i}$, i.e., measures supported on at most N points with masses in multiples of $1/N$.*

Building on this identification (proven in Appendix B), we now formalize the notion of stability. In physical systems and stochastic optimization, small perturbations in the input should yield small changes in the output. We capture this requirement through Lipschitz continuity with respect to the Wasserstein-1 metric.

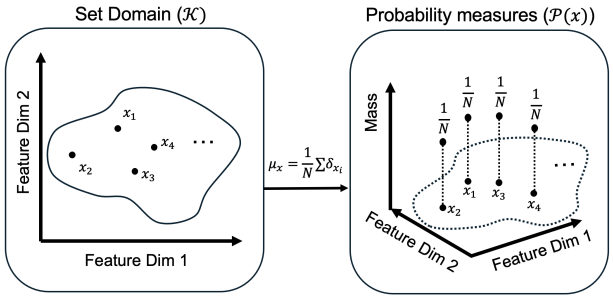

Figure 1. **Measure-Theoretic Representation.** (Left) The input set $\mathbf{x} = \{x_1, \ldots, x_N\}$ resides in a domain with standard vector space structure. (Right) Mapping it to the empirical measure $\mu_{\mathbf{x}} = \frac{1}{N}\sum_{i=1}^{N} \delta_{x_i}$ embeds the data into probability measure space. This transformation enables the use of Wasserstein geometry to quantify stability, bridging discrete inputs and continuous analysis.

Assumption 1 (Wasserstein Stability)**.** The target function $F : \mathcal{P}_N(\mathcal{K}) \to \mathbb{R}$ is Lipschitz continuous with respect to the Wasserstein-1 metric: there exists a constant $L_F > 0$ such that for all $\mu, \nu \in \mathcal{P}_N(\mathcal{K})$,

$$|F(\mu) - F(\nu)| \leq L_F \cdot W_1(\mu, \nu).$$

The Wasserstein-1 distance quantifies the minimum cost of transporting mass from one distribution to another.[3] Formally, for $\mu, \nu \in \mathcal{P}_N(\mathcal{K})$, it is defined via the Kantorovich-Rubinstein duality (Villani et al., 2009):

$$W_1(\mu, \nu) = \sup_{s \in \mathrm{Lip}_1(\mathcal{K})} \left| \int_{\mathcal{K}} s(x)\, d\mu(x) - \int_{\mathcal{K}} s(x)\, d\nu(x) \right|, \tag{4}$$

where the supremum is over all 1-Lipschitz functions on $\mathcal{K}$, and $\int s(x)\, d\mu(x)$ denotes the integral of s with respect to the measure μ.

To demonstrate that Wasserstein stability is not merely a mathematical convenience but a common feature of real-world problems, we examine two applications: data-driven stochastic optimization and robust point cloud processing.

Case 1: Data-Driven Stochastic Optimization (DDSO). DDSO problems arise in power system operations (Guo et al., 2018; Zhou et al., 2025) and portfolio management (Boyd et al., 2017), where the goal is to map a distribution of uncertainty to an optimal decision.

[2] While we use uniform weights $1/N$ to represent standard multisets, this formulation extends naturally to weighted measures for applications involving element-wise importance or confidence scores.

[3] For fixed-size sets, the Wasserstein distance is mathematically equivalent to a symmetrized, normalized Euclidean distance (Levin et al., 2026). We adopt the Wasserstein formulation because it explicitly captures the permutation-invariant structure and aligns naturally with our measure-theoretic proof techniques.

Proposition 3.2 (Stability of Stochastic Optimization)**.** *Consider a stochastic optimization problem where the goal is to find a decision $y \in \mathcal{Y}$ that minimizes the expected loss under an uncertain distribution μ:*

$$y^*(\mu) := \arg\min_{y \in \mathcal{Y}} \mathbb{E}_{\xi \sim \mu}[\ell(y, \xi)], \tag{5}$$

where $\mathcal{Y} \subset \mathbb{R}^d$ is closed and convex, and ξ is the uncertain parameter drawn from μ. Suppose $\ell(y, \xi)$ is α-strongly convex in y (ensuring a unique optimal solution), differentiable with integrable gradient $\nabla_y \ell$, and has L_g-Lipschitz continuous gradients in ξ uniformly over $y \in \mathcal{Y}$. Then the solution mapping $y^* : \mu \mapsto y^*(\mu)$ is Lipschitz continuous in the Wasserstein-1 metric:

$$\|y^*(\mu) - y^*(\nu)\| \leq \frac{L_g}{\alpha} W_1(\mu, \nu). \tag{6}$$

Proposition 3.2 (proved in Appendix C) confirms that Wasserstein stability is a naturally arising condition. By establishing that the input-to-solution mapping of stochastic optimization satisfies this property, we demonstrate that our stability assumption is consistent with well-defined data-driven tasks, ensuring that our logarithmic width bound applies to a broad class of practical learning problems.

Case 2: Robust Point Cloud Processing. In 3D computer vision, point clouds are acquired from LiDAR or depth sensors subject to additive Gaussian noise, a common perturbation in safety-critical perception systems such as autonomous driving (Qi et al., 2017).

Proposition 3.3 (Wasserstein Stability under Gaussian Perturbation)**.** *Let P be a probability distribution on $\mathbb{R}^d$ with finite second moment, and let $\zeta \sim \mathcal{N}(0, \sigma^2 I_d)$ be independent Gaussian noise. Define the perturbed distribution $Q = P * \mathcal{N}(0, \sigma^2 I_d)$ as the law of $Y = X + \zeta$ where $X \sim P$. Then for any L_F-Lipschitz function F with respect to W_1:*

$$|F(P) - F(Q)| \leq L_F \cdot W_1(P, Q) \leq L_F \cdot \sigma\sqrt{d}. \tag{7}$$

Proposition 3.3 (proved in Appendix D) shows that for any Wasserstein-stable model, the output perturbation is proportional to the noise level σ and grows at most as $\sqrt{d}$ with the ambient dimension. This provides a rigorous robustness guarantee for safety-critical applications and directly motivates Assumption 1.

In summary, these two cases illustrate that Wasserstein stability is not an artificial constraint but a naturally arising property in diverse application domains, from decision-making under uncertainty to robust perception systems. Crucially, this stability assumption also enables a fundamental shift in our theoretical analysis: from exact separation to geometric approximation. Prior work (Zaheer et al., 2017; Wagstaff et al., 2019) enforces exact injectivity, treating even infinitesimally close multisets as distinct entities that must be separated. This rigid requirement necessitates a latent dimension proportional to the set size, i.e., $\mathcal{O}(N)$, to encode every element independently. In contrast, the stability condition relaxes this to an ε-covering problem: since small Wasserstein perturbations yield small output changes, we effectively compress the input space, replacing the linear cost of exact injectivity with the logarithmic growth of metric entropy.

Remark. PointNet (Qi et al., 2017) adopts Hausdorff continuity as a sufficient condition for universal approximation, achieving a latent dimension of $\mathcal{O}(\varepsilon^{-D})$ independent of N. However, Hausdorff continuity ignores element multiplicity, making it blind to the distribution of mass. For example, the multisets $X_1 = \{1, 0\}$ and $X_2 = \{1, 1, 0\}$ have Hausdorff distance zero since they share the same unique elements, yet the set mean—a distribution-sensitive invariant—yields different values ($0.5 \neq 0.67$). In contrast, the Wasserstein metric treats inputs as empirical measures, explicitly capturing both geometry and mass. Wasserstein-stable functions, therefore, represent a distinct and practically vital class for tasks such as DDSO and point cloud processing.

4. Main Results

In this section, we derive a theoretical upper bound on the latent dimension required for DeepSets to approximate Wasserstein-stable functions. We provide a constructive proof demonstrating that, under the assumption of Wasserstein stability, the required latent dimension scales only logarithmically with N to achieve universal approximation.

4.1. Logarithmic Width Sufficiency

Theorem 1 (Logarithmic Width Sufficiency)**.** *Let $\mathcal{K} = [0, 1]^D$ and let $F : \mathcal{P}_N(\mathcal{K}) \to \mathbb{R}$ be L_F-Lipschitz with respect to W_1. Then with high probability, there exist a pointwise encoder $\psi : \mathcal{K} \to \mathbb{R}^m$ and a continuous decoder $\rho : \mathbb{R}^m \to \mathbb{R}$ defining a DeepSets architecture*

$$h(\mathbf{x}) = \rho\left(\frac{1}{N}\sum_{i=1}^{N} \psi(x_i)\right) \tag{8}$$

such that

$$\sup_{\mathbf{x} \in \mathcal{K}^N} |f(\mathbf{x}) - h(\mathbf{x})| \leq \widehat{C}_D \, L_F \, \varepsilon^{\beta}, \tag{9}$$

with $\beta = 2/(D + 6)$, provided the latent width satisfies

$$m = \mathcal{O}\Big(C_D \varepsilon^{-(3D+4)} \log(N + \varepsilon^{-D})\Big). \tag{10}$$

Here C_D and $\widehat{C}_D$ are constants depending only on the dimension D. For fixed ε and D, this width is logarithmic in N.

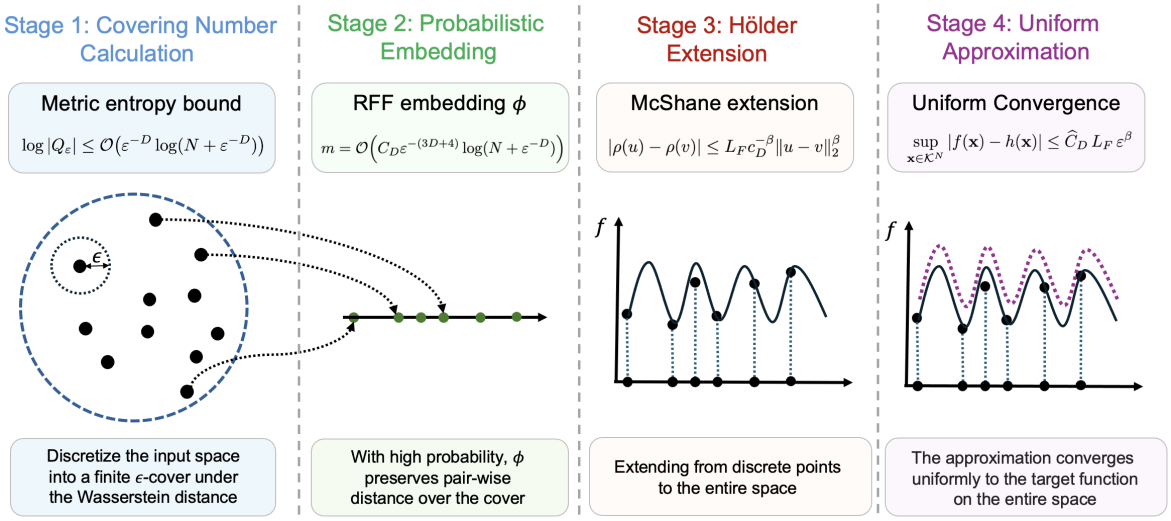

Figure 2. **Proof sketch of Theorem 1.** The constructive proof proceeds in four stages. (**Stage 1**) The space of empirical measures $\mathcal{P}_N(\mathcal{K})$ is discretized into a finite ε-cover via a maximal ε-separated net with log-cardinality $\mathcal{O}(\varepsilon^{-D}\log(N+\varepsilon^{-D}))$. (**Stage 2**) A multiscale Random Fourier Feature encoder embeds this net into $\mathbb{R}^m$, providing Lipschitz stability and Hölder separation guarantees. (**Stage 3**) A Hölder-continuous decoder constructed via the McShane extension theorem interpolates the target values on the net and extends them to the entire latent space. (**Stage 4**) The total error is decomposed via the triangle inequality into a discretization error and a latent stability error, yielding the uniform bound in (9).

The bound in Theorem 1 has three notable implications.

First, the logarithmic dependence on N. The width $m = \mathcal{O}(\varepsilon^{-(3D+4)}\log(N+\varepsilon^{-D}))$ grows only logarithmically in the set size N for fixed accuracy ε and input dimension D, effectively decoupling model capacity from set cardinality. This provides a rigorous theoretical justification for the empirical success of DeepSets in applications such as point cloud processing (Qi et al., 2017) and particle flow networks in high-energy physics (Komiske et al., 2019), which employ fixed latent dimensions (e.g., $m = 256$) regardless of set size. While algebraic bounds indicate that linear scaling is required to strictly distinguish all possible multisets, our geometric framework explains why fixed-width architectures remain effective in practice: for Wasserstein-stable functions, information content saturates according to the metric entropy of the input space, rendering linear width scaling unnecessary even for massive sets (e.g., $N \sim 10^5$).

Second, the covering complexity on $1/\varepsilon$ and D. The factor $\varepsilon^{-(3D+4)}$ reflects the geometric covering complexity of the feature space and constitutes the dominant cost of stable approximation. While this suggests an exponential worst-case dependence on the ambient dimension D, the scaling can be significantly reduced when input data lies on a low-dimensional manifold with intrinsic dimension $d \ll D$—a common scenario in computer vision (Pope et al., 2021; Chen et al., 2024). In such cases, the covering number satisfies $M \le C\varepsilon^{-d}$ for a geometric constant $C > 0$ (Baraniuk & Wakin, 2009), and substituting this into our derivation reduces the effective width to $\mathcal{O}(\varepsilon^{-(3d+4)}\log(N+\varepsilon^{-d}))$, governed strictly by the intrinsic complexity of the data geometry. This provides a rigorous justification for the efficacy of DeepSets in high-dimensional regimes, guaranteeing that learning remains tractable when the data is on a low-dimensional manifold.

Third, the dependence on the Lipschitz constant L_F. The required width is also governed by the regularity of the target function through L_F. To maintain a fixed approximation accuracy τ, the error bound $\widehat{C}_D L_F \varepsilon^\beta \le \tau$ requires the covering resolution to scale as $\varepsilon \propto L_F^{-1/\beta}$, where $\beta = 2/(D+6)$. Substituting into (10) yields $m \propto L_F^{(3D+4)/\beta} = L_F^{(3D+4)(D+6)/2}$ up to logarithmic factors. This reveals a fundamental trade-off between function regularity and model size: in the easy regime where L_F is small (e.g., global mean), compact embeddings suffice; in the hard regime where the target is sensitive to distributional shifts (e.g., high-resolution kernel density estimation), a larger L_F necessitates finer resolution and a wider latent space to capture high-frequency geometric detail.

Proof Sketch: As illustrated in Figure 2, our proof strategy proceeds in four stages.

1. *Metric Entropy (Sec. 4.2):* We discretize the infinite space of empirical measures into a finite net Q_ε by constructing a maximal ε-separated subset of $\mathcal{P}_N(\mathcal{K})$. Maximality guarantees that Q_ε is simultaneously an ε-cover, while separation ensures every distinct pair is at least ε apart in W_1. The log-cardinality of this net scales as $\mathcal{O}(\varepsilon^{-D}\log(N+\varepsilon^{-D}))$, growing only

logarithmically in N.

2. *Multiscale Random Features (Sec. 4.3):* We construct a mean-pooling encoder Ψ from cosine features sampled at different frequencies with decaying weights. The decaying weights make Ψ globally Lipschitz with respect to W_1, while the multiscale embedding guarantees that for any pair of measures in Q_ε, the matched frequency scale provides a Hölder separation bound of strength W_1^α with $\alpha = (D+6)/2$. The separation guarantee is enforced simultaneously over all pairs in Q_ε on the high-probability event stated in Lemma 4.3. Together, these give a latent dimension of $\mathcal{O}(\varepsilon^{-(3D+4)} \log(N + \varepsilon^{-D}))$ sufficient to embed the geometry of the net with high probability.

3. *Hölder Extension (Sec. 4.4):* The separation bound converts the W_1-Lipschitz regularity of F into Hölder regularity of the induced map on the embedded net. The same lower separation event also makes Ψ injective on Q_ε, so the induced map $g(\Psi(\nu)) = F(\nu)$ is well-defined. We then apply the McShane extension theorem to construct a globally Hölder-continuous decoder $\rho : \mathbb{R}^m \to \mathbb{R}$ that exactly interpolates the target values on the net and extends them continuously to the entire latent space.

4. *Uniform Approximation (Sec. 4.5):* We extend the guarantee from the finite net to the entire domain. For any input measure, we bound the total error via the triangle inequality by comparing to its nearest net point. The Lipschitz continuity of F controls the discretization error by at most $L_F\varepsilon$, and the Hölder continuity of ρ together with the Lipschitz stability of Ψ controls the latent stability error by at most a constant multiple of $L_F\varepsilon^\beta$, yielding the uniform bound in (9).

4.2. Metric Entropy

Standard algebraic approaches to permutation-invariant function approximation prioritize exact injectivity and require latent dimension scaling linearly with set size N. We circumvent this bottleneck by adopting a metric entropy framework that exploits Wasserstein stability, relaxing exact separation to approximate covering, thereby reducing complexity from the raw set cardinality to the covering number of the Wasserstein space.

The following lemma shows that metric entropy grows only logarithmically in N for fixed resolution. We construct Q_ε as a *maximal* ε-separated subset, which simultaneously serves as an ε-cover and guarantees that every distinct pair is at least ε apart in W_1—a separation property essential for the matched-scale argument in Section 4.3.

Lemma 4.1 (Metric Entropy of Empirical Measures)**.** *Let $\mathcal{K} = [0,1]^D$. For any $\varepsilon \in (0,1)$, let $Q_\varepsilon \subset \mathcal{P}_N(\mathcal{K})$ be a maximal ε-separated subset with respect to W_1. Then Q_ε is an ε-cover of $\mathcal{P}_N(\mathcal{K})$, and its metric entropy (log-cardinality) satisfies*

$$\log |Q_\varepsilon| \le \mathcal{O}\big(\varepsilon^{-D} \log(N + \varepsilon^{-D})\big)\,. \tag{11}$$

We establish this bound via a constructive quantization argument. Partitioning $\mathcal{K}$ into an ε-net of cardinality [4] $M \asymp \varepsilon^{-D}$, we then project empirical measures onto the net's centroids. Since Q_ε is ε-separated, its cardinality is bounded by the covering number at resolution $\varepsilon/2$, which reduces to counting distinct allocations of N indistinguishable masses across M bins. By the stars-and-bars formula [5], this equals $\binom{N+M-1}{M-1}$, whose logarithm scales as $\mathcal{O}(M \log(N+M)) = \mathcal{O}(\varepsilon^{-D} \log(N+\varepsilon^{-D}))$, confirming logarithmic growth in N. See Appendix E for the complete derivation.

4.3. Probabilistic Embeddings

Having reduced the effective input domain to a finite net Q_ε (Lemma 4.1), we now construct a mapping from probability measures into a finite-dimensional vector space that preserves their geometric relationships.

Random Fourier Features (RFF) (Rahimi & Recht, 2007) embed probability measures into $\mathbb{R}^m$ by averaging randomly sampled cosine features over the measure. A feature

$$\phi(x) = \sqrt{2}\cos(a^\top x + b),$$

with phase $b \sim \mathrm{Unif}[0, 2\pi]$, probes oscillations at frequency $\|a\|_2$, and therefore at spatial scale roughly $1/\|a\|_2$. In our construction, a frequency block with bandwidth A samples $a \sim \mathrm{Unif}(B(0,A))$. Small bandwidth produces smoother, more stable features, but lacks the frequency content needed to reliably resolve pairs whose Wasserstein separation is much below $1/A$. Large bandwidth can detect finer-scale differences, but it also increases the Lipschitz sensitivity of the encoder. Since Q_ε contains pairs across many W_1 scales, no single bandwidth simultaneously gives both global stability and lower separation for all pairs.

We resolve this by constructing a *multiscale* encoder $\Psi : \mathcal{P}_N(\mathcal{K}) \to \mathbb{R}^m$ that samples cosine features at dyadic frequency scales $A_s = 2^s$ for $s = 0, 1, \ldots, S$ with $2^S \asymp \varepsilon^{-1}$, weighted by geometrically decaying scale weights $w_s = 2^{-2s}$. The multiscale encoder with RFF is defined as

$$\Psi(\mu) = \bigoplus_{s=0}^{S} \frac{w_s}{\sqrt{m_s}} \Phi_s(\mu), \qquad \Phi_s(\mu) = \frac{1}{N}\sum_{i=1}^{N} \phi_s(x_i),$$

[4] Here $a \asymp b$ means that a and b are comparable up to positive constant factors independent of the main scaling parameters.

[5] The stars-and-bars formula counts the number of ways to distribute N indistinguishable objects among M distinguishable bins. Equivalently, the number of nonnegative integer solutions to $k_1 + \cdots + k_M = N,\ k_j \in \mathbb{Z}_{\ge 0}$ is $\binom{N+M-1}{M-1}$.

where the ℓ-th component of ϕ_s is $\phi_{s,\ell}(x) = \sqrt{2}\cos(a_{s,\ell}^\top x + b_{s,\ell})$ with $a_{s,\ell} \sim \mathrm{Unif}(B(0, 2^s))$ and $b_{s,\ell} \sim \mathrm{Unif}[0, 2\pi]$, and $\ell \in [1, \ldots, m_s]$. The symbol $\oplus$ denotes concatenation of the feature blocks over all scales, so that the total latent dimension is $m = \sum_{s=0}^{S} m_s$.

The decaying weights resolve the stability-separation challenge in two complementary ways. For *stability* (Lemma 4.2), they suppress high-frequency contributions, making the total Lipschitz constant a convergent geometric series independent of S. For *separation* (Lemma 4.3), each pair (ν_j, ν_k) with W_1 distance r is detected at a matched scale $s^* \asymp \log_2(r^{-1})$, where cosine features have just the right bandwidth; the weight $w_{s^*} \asymp r^2$ contributes the correct power of r to produce the Hölder exponent $\alpha = (D+6)/2$.

We now formalize these two properties. The first lemma establishes global Lipschitz continuity of the encoder.

Lemma 4.2 (Upper Stability)**.** *The multiscale encoder Ψ is Lipschitz with respect to W_1: for any $\mu, \nu \in \mathcal{P}_N(\mathcal{K})$,*

$$\|\Psi(\mu) - \Psi(\nu)\|_2 \le \sqrt{8/3} \cdot W_1(\mu, \nu), \tag{12}$$

Proof Sketch. The proof applies Kantorovich–Rubinstein duality to bound each cosine feature's contribution, then sums over scales. A feature at scale 2^s has Lipschitz constant of order 2^s, while its block weight is $w_s = 2^{-2s}$, so the squared contribution at scale s is of order 2^{-2s}. The per-scale contributions thus form the convergent series $\sum_s 2^{-2s} = 4/3$, yielding a global Lipschitz constant independent of ε. See Appendix F for details.

The second lemma establishes that well-separated measures in Q_ε remain distinguishable in the latent space.

Lemma 4.3 (Lower Separation)**.** *Let Q_ε be the maximal ε-separated net from Lemma 4.1. With probability at least $1 - \delta$, the multiscale encoder Ψ satisfies the following Hölder separation bound simultaneously for all distinct pairs $\nu_j, \nu_k \in Q_\varepsilon$:*

$$\|\Psi(\nu_j) - \Psi(\nu_k)\|_2 \ge c_D \cdot W_1(\nu_j, \nu_k)^\alpha, \tag{13}$$

Here $\alpha = (D+6)/2$ and the constant $c_D > 0$ depends only on the domain and ambient dimension. This guarantee holds provided

$$m \ge C_D \left(\varepsilon^{-(3D+4)} \log(N + \varepsilon^{-D}) + \varepsilon^{-(2D+4)} \log \frac{1}{\delta} \right),$$

which simplifies to $m = \mathcal{O}(C_D \varepsilon^{-(3D+4)} \log(N + \varepsilon^{-D}))$ for fixed confidence level δ.

Proof Sketch. The argument proceeds in four steps. First, for each pair ν_j, ν_k with $r = W_1(\nu_j, \nu_k) \ge \varepsilon$, we select a matched frequency scale s^* such that $2^{s^*} \asymp r^{-1}$. Second, Kantorovich–Rubinstein duality combined with Plancherel's theorem shows that the signed measure $\nu_j - \nu_k$ has nontrivial Fourier energy at frequencies below $O(r^{-1})$, yielding a deterministic lower bound of order r^{D+2}. Third, since the random cosine features are independent and bounded, Hoeffding's inequality ensures the empirical norm concentrates around its expectation. Fourth, combining the block weight $w_{s^*} \asymp r^2$ with the concentration bound yields the Hölder separation, and a union bound over $|Q_\varepsilon|^2$ pairs gives the stated dimension requirement. See Appendix G for the complete derivation.

4.4. Hölder Extension

Let $\mathcal{Z}_\varepsilon = \{z = \Psi(\nu) : \nu \in Q_\varepsilon\} \subset \mathbb{R}^m$ denote the embedded net. We then construct a decoder $\rho : \mathbb{R}^m \to \mathbb{R}$ that maps latent embeddings z to target values. Throughout this construction we condition on the event from Lemma 4.3, which holds with probability at least $1 - \delta$. On this event, the lower separation bound implies that Ψ is *injective* on Q_ε, i.e., the inverse embedding Ψ^{-1} exists on $\mathcal{Z}_\varepsilon$ such that $\nu = \Psi^{-1}(z)$.

Therefore, we can define a decoder $g : \mathcal{Z}_\varepsilon \to \mathbb{R}$ by $g(\Psi(\nu)) = F(\nu)$ for each $\nu \in Q_\varepsilon$. The lower separation bound (Lemma 4.3) implies $W_1(\nu_j, \nu_k) \le c_D^{-1/\alpha} \|\Psi(\nu_j) - \Psi(\nu_k)\|_2^{1/\alpha}$. Since F is L_F-Lipschitz in W_1, for $z_j, z_k \in \mathcal{Z}_\varepsilon$, the induced decoder g satisfies

$$\begin{aligned} |g(z_j) - g(z_k)| = |F(\nu_j) - F(\nu_k)| &\le L_F W_1(\nu_j, \nu_k) \\ &\le L_F c_D^{-\beta} \|z_j - z_k\|_2^\beta, \end{aligned}$$

where $\beta = \frac{1}{\alpha} = \frac{2}{D+6}$, and g is β-Hölder continuous on $\mathcal{Z}_\varepsilon$. Since $\beta \in (0, 1)$, the function $d_\beta(u, v) = \|u - v\|_2^\beta$ defines a valid metric on $\mathbb{R}^m$, and g is Lipschitz with respect to d_β. Applying the McShane–Whitney extension theorem (McShane, 1934) in this metric yields a global decoder $\rho : \mathbb{R}^m \to \mathbb{R}$ satisfying $\rho(\Psi(\nu)) = F(\nu)$ for all $\nu \in Q_\varepsilon$, with

$$|\rho(u) - \rho(v)| \le L_F c_D^{-\beta} \|u - v\|_2^\beta \quad \forall\, u, v \in \mathbb{R}^m. \tag{14}$$

The decoder admits the explicit construction $\rho(u) = \inf_{z \in \mathcal{Z}_\varepsilon} \{g(z) + L_F c_D^{-\beta} \|u - z\|_2^\beta\}$.

4.5. Uniform Approximation Analysis

The final stage combines the metric entropy bound (Sec. 4.2), the encoder stability and separation guarantees (Sec. 4.3), and the Hölder decoder (Sec. 4.4) to establish a uniform approximation bound over $\mathcal{P}_N(\mathcal{K})$.

For any $\mu \in \mathcal{P}_N(\mathcal{K})$, let $\nu \in Q_\varepsilon$ satisfy $W_1(\mu, \nu) \le \varepsilon$. Using the interpolation property $\rho(\Psi(\nu)) = F(\nu)$ and the

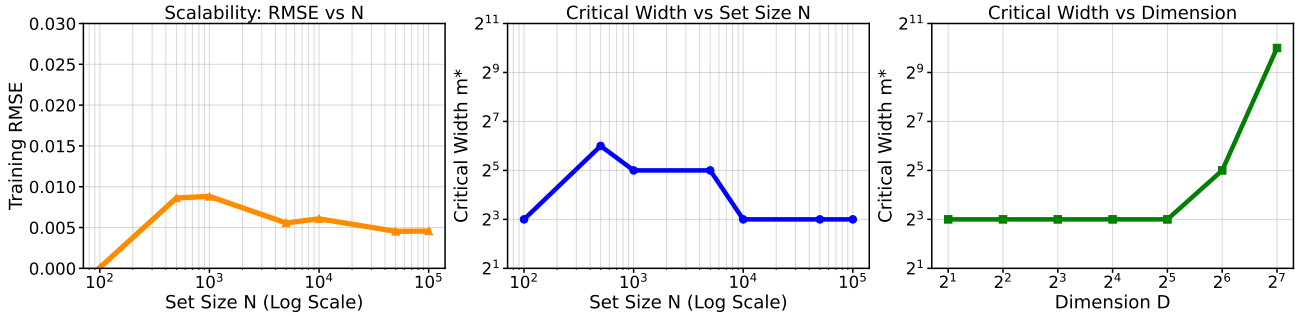

Figure 3. **Empirical scaling of representational capacity for geometric median estimation.** (**Left**) Training RMSE versus set size N at fixed $m = 256$ and $D = 16$. Error remains below 0.01 across $N \in [10^2, 10^5]$, confirming that approximation quality is stable as N increases by three orders of magnitude, consistent with our theory. (**Middle**) Critical latent dimension m^* (RMSE $< 10^{-2}$) versus N at fixed $D = 16$. After initial fluctuation, m^* stabilizes at 8 for $N \geq 10^4$, consistent with the predicted logarithmic dependence on N. (**Right**) Critical width m^* versus ambient dimension D at fixed $N = 10^4$. The required width remains small ($m^* \leq 16$) for $D \leq 32$ but grows sharply beyond $D = 64$, reflecting the exponential dependence on D from metric entropy.

triangle inequality:

$$|F(\mu) - \rho(\Psi(\mu))| \leq \underbrace{|F(\mu) - F(\nu)|}_{\text{(I) Discretization}} + \underbrace{|\rho(\Psi(\nu)) - \rho(\Psi(\mu))|}_{\text{(II) Latent Stability}}.$$

Term (I) follows from the L_F-Lipschitz continuity of F: $|F(\mu) - F(\nu)| \leq L_F \varepsilon$. Term (II) follows from the β-Hölder continuity of ρ and the Lipschitz stability of Ψ (Lemma 4.2):

$$\begin{aligned} |\rho(\Psi(\nu)) - \rho(\Psi(\mu))| &\leq L_F c_D^{-\beta} \|\Psi(\nu) - \Psi(\mu)\|_2^\beta \\ &\leq L_F c_D^{-\beta} (\frac{8}{3})^{\beta/2} \varepsilon^\beta. \end{aligned}$$

Summing the two terms and using $\varepsilon \leq \varepsilon^\beta$ for $\varepsilon \in (0, 1)$ and $\beta \in (0, 1)$:

$$|F(\mu) - \rho(\Psi(\mu))| \leq \widehat{C}_D \, L_F \, \varepsilon^\beta, \tag{15}$$

where $\widehat{C}_D := 1 + c_D^{-\beta} (\frac{8}{3})^{\beta/2}$. Taking the supremum over all $\mu \in \mathcal{P}_N(\mathcal{K})$ and applying the identification $f(\mathbf{x}) = F(\mu_{\mathbf{x}})$ yields:

$$\sup_{\mathbf{x} \in \mathcal{K}^N} |f(\mathbf{x}) - h(\mathbf{x})| \leq \widehat{C}_D \, L_F \, \varepsilon^\beta. \tag{16}$$

We complete the proof of Theorem 1.

5. Numerical Experiments

To empirically validate our theoretical predictions, we investigate how the critical latent dimension scales with set size N and feature dimension D, using geometric median estimation as a testbed. This task, defined by $y^* = \arg\min_y \sum_{i=1}^N \|x_i - y\|_2$, is Lipschitz in W_1 under standard nondegeneracy conditions, making it well-suited for our framework.

As shown in Figure 3, the empirical results support the central scaling message of our theory: stable permutation-invariant tasks do not require a latent width that grows linearly with the set size. Across the tested configurations, DeepSets maintains strong predictive performance even under substantial latent compression, while the sequence-based baseline degrades more sharply despite parameter matching. This behavior is consistent with the proposed view that, once the task is stable under Wasserstein perturbations, the effective representational burden is governed by geometric resolution rather than by the raw cardinality of the input set. Additional validation on the MNIST digit-sum task, probing the manifold hypothesis with real image data, is provided in Appendix I.

6. Conclusion

We address the dimensionality bottleneck in learning permutation-invariant functions by shifting the theoretical analysis from combinatorial injectivity to geometric stability. While exact representation of arbitrary multisets requires a latent dimension of $\mathcal{O}(N)$, we show this can be avoided for Wasserstein-stable functions. Our constructive proof builds a Wasserstein separation net with metric entropy logarithmic in N, embeds it via a multiscale RFF encoder with provable Lipschitz stability and Hölder separation guarantees, and recovers the target function via a McShane-extended Hölder decoder. Together, these yield a DeepSets architecture whose latent width is $\mathcal{O}(\varepsilon^{-(3D+4)} \log(N + \varepsilon^{-D}))$, logarithmic in N, providing a rigorous theoretical foundation for the expressivity of compact sum-decomposable architectures on stable set-function tasks. Limitations and future research directions are discussed in Appendix J.

Acknowledgments

This work is supported in part by a General Research Fund from Research Grants Council, Hong Kong (Project No. 11200124), a Collaborative Research Fund from Research Grants Council, Hong Kong (Project No. C1049-24G), an InnoHK initiative, The Government of the HKSAR, Laboratory for AI-Powered Financial Technologies, a Shenzhen-Hong Kong-Macau Science & Technology Project (Category C, Project No. SGDX20220530111203026), and a Start-up Research Grant from The Chinese University of Hong Kong, Shenzhen (Project No. UDF01004086). We thank the anonymous reviewers for their constructive feedback. In particular, one reviewer suggested a technique that may improve the ε-dependency in our bounds. Due to time constraints, we were unable to fully incorporate this technique into the present work; we plan to investigate this direction thoroughly in future work.

Impact Statement

This paper presents foundational theoretical work whose goal is to advance the mathematical understanding of machine learning architectures for permutation-invariant data. Because the contribution is theoretical and does not introduce a deployed system, dataset, or application-specific decision pipeline, we do not identify any direct or immediate negative societal consequences.

References

Amir, T., Gortler, S., Avni, I., Ravina, R., and Dym, N. Neural injective functions for multisets, measures and graphs via a finite witness theorem. *Advances in Neural Information Processing Systems*, 36:42516–42551, 2023.

Baraniuk, R. G. and Wakin, M. B. Random projections of smooth manifolds. *Foundations of computational mathematics*, 9(1):51–77, 2009.

Boyd, S., Diamond, S., Koh, K., Nystrup, P., Busseti, E., Kahn, R. N., and Speth, J. Multi-period trading via convex optimization. *Foundations and Trends in Optimization*, 3(1):1–76, 2017.

Bueno, C. and Hylton, A. On the representation power of set pooling networks. *Advances in Neural Information Processing Systems*, 34:17170–17182, 2021.

Chen, S., Zhang, H., Guo, M., Lu, Y., Wang, P., and Qu, Q. Exploring low-dimensional subspace in diffusion models for controllable image editing. *Advances in neural information processing systems*, 37:27340–27371, 2024.

Dwivedi, V. P., Luu, A. T., Laurent, T., Bengio, Y., and Bresson, X. Graph neural networks with learnable structural and positional representations. *arXiv preprint arXiv:2110.07875*, 2021.

Dym, N. and Gortler, S. J. Low-dimensional invariant embeddings for universal geometric learning. *Foundations of Computational Mathematics*, 25(2):375–415, 2025.

Folland, G. B. *Fourier analysis and its applications*, volume 4. American Mathematical Soc., 2009.

Guo, Y., Baker, K., Dall'Anese, E., Hu, Z., and Summers, T. H. Data-based distributionally robust stochastic optimal power flow—part i: Methodologies. *IEEE Transactions on Power Systems*, 34(2):1483–1492, 2018.

Guo, Y., Wang, H., Hu, Q., Liu, H., Liu, L., and Bennamoun, M. Deep learning for 3d point clouds: A survey. *IEEE transactions on pattern analysis and machine intelligence*, 43(12):4338–4364, 2020.

Hornik, K., Stinchcombe, M., and White, H. Multilayer feedforward networks are universal approximators. *Neural networks*, 2(5):359–366, 1989.

Hu, H., Gu, J., Zhang, Z., Dai, J., and Wei, Y. Relation networks for object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 3588–3597, 2018.

Hu, H., Zhang, Z., Xie, Z., and Lin, S. Local relation networks for image recognition. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 3464–3473, 2019.

Khesin, B. A. and Tabachnikov, S. L. *ARNOLD: Swimming Against the Tide: Swimming Against the Tide*, volume 86. American Mathematical Society, 2014.

Komiske, P. T., Metodiev, E. M., and Thaler, J. Energy flow networks: deep sets for particle jets. *Journal of High Energy Physics*, 2019(1):1–46, 2019.

Lee, J., Lee, Y., Kim, J., Kosiorek, A., Choi, S., and Teh, Y. W. Set transformer: A framework for attention-based permutation-invariant neural networks. In *International conference on machine learning*, pp. 3744–3753. PMLR, 2019.

Levin, E., Ma, Y., Díaz, M., and Villar, S. On transferring transferability: Towards a theory for size generalization. *Advances in Neural Information Processing Systems*, 38: 67015–67083, 2026.

Li, G., Jiao, Y., Huang, Y., Wei, Y., and Chen, Y. Transformers meet in-context learning: A universal approximation theory. *arXiv preprint arXiv:2506.05200*, 2025.

Liu, H., Hu, J. Y.-C., Song, Z., and Liu, H. Attention mechanism, max-affine partition, and universal approximation. *Advances in Neural Information Processing Systems*, 38: 120443–120511, 2026.

Liu, Z., Wang, Y., Vaidya, S., Ruehle, F., Halverson, J., Soljacic, M., Hou, T., and Tegmark, M. Kan: Kolmogorov–arnold networks. In *International conference on learning representations*, volume 2025, pp. 70367–70413, 2025.

Locatello, F., Weissenborn, D., Unterthiner, T., Mahendran, A., Heigold, G., Uszkoreit, J., Dosovitskiy, A., and Kipf, T. Object-centric learning with slot attention. *Advances in neural information processing systems*, 33:11525–11538, 2020.

Maron, H., Ben-Hamu, H., Shamir, N., and Lipman, Y. Invariant and equivariant graph networks. *arXiv preprint arXiv:1812.09902*, 2018.

Maron, H., Ben-Hamu, H., Serviansky, H., and Lipman, Y. Provably powerful graph networks. *Advances in neural information processing systems*, 32, 2019a.

Maron, H., Fetaya, E., Segol, N., and Lipman, Y. On the universality of invariant networks. In *International conference on machine learning*, pp. 4363–4371. PMLR, 2019b.

McShane, E. J. Extension of range of functions. 1934.

Mialon, G., Chen, D., Selosse, M., and Mairal, J. Graphit: Encoding graph structure in transformers. *arXiv preprint arXiv:2106.05667*, 2021.

Murphy, R. L., Srinivasan, B., Rao, V., and Ribeiro, B. Janossy pooling: Learning deep permutation-invariant functions for variable-size inputs. *arXiv preprint arXiv:1811.01900*, 2018.

O'shea, K. and Nash, R. An introduction to convolutional neural networks. *arXiv preprint arXiv:1511.08458*, 2015.

Panaretos, V. M. and Zemel, Y. *An invitation to statistics in Wasserstein space*. Springer, 2020.

Pevny, T. and Kovarík, V. Approximation capability of neural networks on spaces of probability measures and tree-structured domains. *arXiv preprint arXiv:1906.00764*, 2019.

Pope, P., Zhu, C., Abdelkader, A., Goldblum, M., and Goldstein, T. The intrinsic dimension of images and its impact on learning. *arXiv preprint arXiv:2104.08894*, 2021.

Qi, C. R., Su, H., Mo, K., and Guibas, L. J. Pointnet: Deep learning on point sets for 3d classification and segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 652–660, 2017.

Rahimi, A. and Recht, B. Random features for large-scale kernel machines. *Advances in neural information processing systems*, 20, 2007.

Scarselli, F., Gori, M., Tsoi, A. C., Hagenbuchner, M., and Monfardini, G. The graph neural network model. *IEEE transactions on neural networks*, 20(1):61–80, 2008.

Schmidt-Hieber, J. The kolmogorov–arnold representation theorem revisited. *Neural networks*, 137:119–126, 2021.

Segol, N. and Lipman, Y. On universal equivariant set networks. *arXiv preprint arXiv:1910.02421*, 2019.

Shapiro, A., Dentcheva, D., and Ruszczynski, A. *Lectures on stochastic programming: modeling and theory*. SIAM, 2021.

Sung, F., Yang, Y., Zhang, L., Xiang, T., Torr, P. H., and Hospedales, T. M. Learning to compare: Relation network for few-shot learning. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 1199–1208, 2018.

Tabaghi, P. and Wang, Y. Universal representation of permutation-invariant functions on vectors and tensors. In *International Conference on Algorithmic Learning Theory*, pp. 1134–1187. PMLR, 2024.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., and Polosukhin, I. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

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

Wagstaff, E., Fuchs, F., Engelcke, M., Posner, I., and Osborne, M. A. On the limitations of representing functions on sets. In *International conference on machine learning*, pp. 6487–6494. PMLR, 2019.

Wang, P., Yang, S., Li, S., Wang, Z., and Li, P. Polynomial width is sufficient for set representation with high-dimensional features. In *International Conference on Learning Representations*, volume 2024, pp. 52223–52252, 2024.

Wu, Z., Pan, S., Chen, F., Long, G., Zhang, C., and Yu, P. S. A comprehensive survey on graph neural networks. *IEEE transactions on neural networks and learning systems*, 32 (1):4–24, 2020.

Xu, K., Hu, W., Leskovec, J., and Jegelka, S. How powerful are graph neural networks? *arXiv preprint arXiv:1810.00826*, 2018.

Zaheer, M., Kottur, S., Ravanbakhsh, S., Poczos, B., Salakhutdinov, R. R., and Smola, A. J. Deep sets. *Advances in neural information processing systems*, 30, 2017.

Zhang, L., Tabas, D., and Zhang, B. An efficient learning-based solver for two-stage dc optimal power flow with feasibility guarantees. *arXiv preprint arXiv:2304.01409*, 2023.

Zhao, H., Jiang, L., Jia, J., Torr, P. H., and Koltun, V. Point transformer. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 16259–16268, 2021.

Zhou, M., Liang, E., Chen, M., and Low, S. H. Partially permutation-invariant neural network for solving two-stage stochastic ac-opf problem. *IEEE Transactions on Power Systems*, 2025.

Zweig, A. and Bruna, J. A functional perspective on learning symmetric functions with neural networks. In *International Conference on Machine Learning*, pp. 13023–13032. PMLR, 2021.

Zweig, A. and Bruna, J. Exponential separations in symmetric neural networks. *Advances in Neural Information Processing Systems*, 35:33134–33145, 2022.

Appendices

A. Set Function Universality

Standard sum-decomposable frameworks for fixed-size sets approach permutation-invariant learning by treating inputs as discrete multisets embedded in Euclidean space. Theoretical analyses in this regime heavily prioritize exact representation and combinatorial injectivity. For scalar inputs, distinguishing all possible multisets fundamentally requires a latent width of $m \approx N$ (Zaheer et al., 2017; Wagstaff et al., 2019). In high-dimensional settings ($D > 1$), achieving exact representation or separating orbits demands a latent capacity that scales at least linearly as $\mathcal{O}(ND)$ or polynomially (Wang et al., 2024; Tabaghi & Wang, 2024; Dym & Gortler, 2025). Collectively, these results establish a strict "linear barrier": by mathematically insisting on the separation of all theoretically distinct configurations, the network width must scale at least linearly with the set size, imposing a prohibitive computational bottleneck for massive-scale applications.

To circumvent the limitations imposed by combinatorial injectivity, a parallel line of work models sets of arbitrary cardinality as empirical probability measures. In this formulation, studies such as Pevny & Kovarík (2019); Zweig & Bruna (2021); Bueno & Hylton (2021) analyze the expressivity of normalized, mean-aggregated architectures over spaces of measures. Using the Stone-Weierstrass theorem, these works establish universal approximation for continuous functionals on probability measures under topologies induced by metrics such as the Wasserstein distance. This measure-theoretic perspective also enables a principled analysis of downstream properties, including generalization across varying set sizes (Levin et al., 2026). However, these results are primarily qualitative: although they guarantee the existence of universal approximators for Wasserstein-continuous functionals, they do not yield explicit quantitative bounds on the minimum latent dimension m.

In this work, we bridge the gap between these two paradigms by applying the measure-theoretic stability perspective to the fixed-cardinality regime. By relaxing the strict requirement of exact combinatorial injectivity, we derive a rigorous, constructive quantitative capacity bound. Specifically, we prove that a latent dimension logarithmic in N for fixed accuracy and confidence is sufficient for the uniform approximation of Wasserstein-stable permutation-invariant functions. This result avoids the linear-scaling bottleneck for this stable function class, while remaining consistent with existing lower bounds for exact representation of arbitrary multisets.

B. Proof of Proposition 3.1

We first restate the result regarding the equivalence between permutation-invariant functions on sets and functionals on empirical measures.

Proposition 3.1. A function $f : (\mathcal{K})^N \to \mathbb{R}$ is permutation-invariant if and only if there exists a unique function $F : \mathcal{P}_N(\mathcal{K}) \to \mathbb{R}$ such that $f(\mathbf{x}) = F(\mu_{\mathbf{x}})$, where $\mathcal{P}_N(\mathcal{K})$ denotes the set of empirical probability measures of the form $N^{-1}\sum_{i=1}^{N} \delta_{x_i}$, equivalently measures with at most N support points and masses in multiples of $1/N$.

Proof. We establish the equivalence by proving the implication in both directions.

($\Leftarrow$) Sufficiency. Assume there exists a functional $F : \mathcal{P}_N(\mathcal{K}) \to \mathbb{R}$ such that $f(\mathbf{x}) = F(\mu_{\mathbf{x}})$. Let $\mathbf{x} = (x_1, \ldots, x_N) \in (\mathcal{K})^N$ be an arbitrary input tuple and let $\pi \in \mathfrak{S}_N$ be any permutation. The permuted input is denoted by $\mathbf{x}^\pi = (x_{\pi(1)}, \ldots, x_{\pi(N)})$. By definition, the empirical measure associated with the permuted input is:

$$\mu_{\mathbf{x}^\pi} = \frac{1}{N}\sum_{i=1}^{N} \delta_{x_{\pi(i)}}. \tag{17}$$

Since the summation operation is commutative and associative, the ordering of terms does not alter the resulting measure. Consequently, $\mu_{\mathbf{x}^\pi} = \mu_{\mathbf{x}}$. It follows that:

$$f(\mathbf{x}^\pi) = F(\mu_{\mathbf{x}^\pi}) = F(\mu_{\mathbf{x}}) = f(\mathbf{x}). \tag{18}$$

As this equality holds for any $\mathbf{x}$ and any $\pi \in \mathfrak{S}_N$, we conclude that f is permutation-invariant.

($\Rightarrow$) Necessity. Conversely, assume f is permutation-invariant. We construct the functional F as follows. For any measure $\mu \in \mathcal{P}_N(\mathcal{K})$, the set of pre-images $\mathcal{K}_\mu := \{\mathbf{z} \in (\mathcal{K})^N : \mu_{\mathbf{z}} = \mu\}$ is non-empty by definition. We define $F(\mu) := f(\mathbf{x})$ for any representative $\mathbf{x} \in \mathcal{K}_\mu$.

To show F is well-defined, we must verify that the assignment is independent of the choice of representative. Let $\mathbf{x}, \mathbf{y} \in \mathcal{K}_\mu$. The equality of empirical measures $\mu_{\mathbf{x}} = \mu_{\mathbf{y}}$ implies that $\mathbf{x}$ and $\mathbf{y}$ define the same multiset. Consequently, $\mathbf{y}$ is a permutation of $\mathbf{x}$; that is, there exists $\pi \in \mathfrak{S}_N$ such that $\mathbf{y} = \mathbf{x}^\pi$. Invoking the permutation-invariance of f, we obtain:

$$f(\mathbf{y}) = f(\mathbf{x}^\pi) = f(\mathbf{x}). \tag{19}$$

Since f is constant over $\mathcal{K}_\mu$, the value $F(\mu)$ is unique.

□

C. Proof of Proposition 3.2

We restate the result for completeness.

Proposition 3.2. Consider a stochastic optimization problem where the goal is to find a decision $y \in \mathcal{Y}$ that minimizes the expected loss under an uncertain distribution μ:

$$y^*(\mu) := \arg\min_{y \in \mathcal{Y}} \mathbb{E}_{\xi \sim \mu}[\ell(y, \xi)], \tag{20}$$

where $\mathcal{Y} \subset \mathbb{R}^d$ is closed and convex, and ξ represents the uncertain parameter drawn from distribution μ. The unique optimal solution $y^*(\mu)$ exists when $\ell(y, \xi)$ is α-strongly convex in y. Suppose further that $\ell(y, \xi)$ is differentiable with integrable gradient $\nabla_y \ell$, and has L_g-Lipschitz continuous gradients in ξ uniformly over $y \in \mathcal{Y}$. Then the solution mapping $y^* : \mu \mapsto y^*(\mu)$ is Lipschitz continuous in the Wasserstein-1 metric:

$$\|y^*(\mu) - y^*(\nu)\| \leq \frac{L_g}{\alpha} W_1(\mu, \nu). \tag{21}$$

Proof. We establish the Lipschitz continuity of the solution mapping $y^* : \mu \mapsto y^*_\mu$ with respect to the Wasserstein-1 metric. Let $\mu, \nu \in \mathcal{P}_N(\mathcal{K})$ be two empirical measures, and let y^*_μ and y^*_ν denote their respective optimal solutions. Since $Q(y, \mu) = \mathbb{E}_{\xi \sim \mu}[\ell(y, \xi)]$ is α-strongly convex in y, these minimizers are unique.

Step 1: First-order optimality. For the constrained minimization over the convex set $\mathcal{Y}$, the optimal solution satisfies the variational inequality:

$$\langle \nabla_y Q(y^*_\mu, \mu), y - y^*_\mu \rangle \geq 0, \quad \forall y \in \mathcal{Y}. \tag{22}$$

Applying this to $y^*_\nu \in \mathcal{Y}$ gives $\langle \nabla_y Q(y^*_\mu, \mu), y^*_\nu - y^*_\mu \rangle \geq 0$. Symmetrically, optimality of y^*_ν for ν gives $\langle \nabla_y Q(y^*_\nu, \nu), y^*_\mu - y^*_\nu \rangle \geq 0$. Summing these two inequalities:

$$\langle \nabla_y Q(y^*_\nu, \nu) - \nabla_y Q(y^*_\mu, \mu), y^*_\mu - y^*_\nu \rangle \geq 0. \tag{23}$$

Step 2: Strong convexity bound. Rearranging to isolate the variation due to the measure:

$$\langle \nabla_y Q(y^*_\nu, \mu) - \nabla_y Q(y^*_\mu, \mu), y^*_\nu - y^*_\mu \rangle \leq \langle \nabla_y Q(y^*_\nu, \mu) - \nabla_y Q(y^*_\nu, \nu), y^*_\nu - y^*_\mu \rangle. \tag{24}$$

By α-strong convexity of $Q(\cdot, \mu)$:

$$\alpha \|y^*_\nu - y^*_\mu\|^2 \leq \langle \nabla_y Q(y^*_\nu, \mu) - \nabla_y Q(y^*_\mu, \mu), y^*_\nu - y^*_\mu \rangle. \tag{25}$$

Combining with (24) and applying Cauchy–Schwarz:

$$\alpha \|y^*_\nu - y^*_\mu\|^2 \leq \|\nabla_y Q(y^*_\nu, \mu) - \nabla_y Q(y^*_\nu, \nu)\| \cdot \|y^*_\nu - y^*_\mu\|. \tag{26}$$

Dividing by $\|y^*_\nu - y^*_\mu\|$ (the claim is trivial if $y^*_\nu = y^*_\mu$):

$$\|y^*_\nu - y^*_\mu\| \leq \frac{1}{\alpha} \|\nabla_y Q(y^*_\nu, \mu) - \nabla_y Q(y^*_\nu, \nu)\|. \tag{27}$$

Step 3: Wasserstein bound via dual-norm argument. Define $g(\xi) := \nabla_y \ell(y^*_\nu, \xi) \in \mathbb{R}^d$. Then:

$$\|\nabla_y Q(y^*_\nu, \mu) - \nabla_y Q(y^*_\nu, \nu)\| = \|\mathbb{E}_{\xi \sim \mu}[g(\xi)] - \mathbb{E}_{\xi \sim \nu}[g(\xi)]\|. \tag{28}$$

By the variational characterization of the Euclidean norm:

$$\|\mathbb{E}_\mu[g(\xi)] - \mathbb{E}_\nu[g(\xi)]\| = \sup_{\|w\|_2 \le 1} \left(\mathbb{E}_\mu[\langle w, g(\xi)\rangle] - \mathbb{E}_\nu[\langle w, g(\xi)\rangle]\right), \tag{29}$$

where $w \in \mathbb{R}^d$. For each fixed unit vector w with $\|w\|_2 \le 1$, the scalar function $\xi \mapsto \langle w, g(\xi)\rangle$ is L_g-Lipschitz in ξ uniformly over all unit w, since by Cauchy–Schwarz:

$$|\langle w, g(\xi)\rangle - \langle w, g(\xi')\rangle| \le \|w\|_2 \cdot \|g(\xi) - g(\xi')\|_2 \le L_g \|\xi - \xi'\|. \tag{30}$$

Applying scalar Kantorovich–Rubinstein duality to each term in the supremum:

$$\sup_{\|w\|_2 \le 1} \left(\mathbb{E}_\mu[\langle w, g(\xi)\rangle] - \mathbb{E}_\nu[\langle w, g(\xi)\rangle]\right) \le L_g W_1(\mu, \nu). \tag{31}$$

Therefore:

$$\|\mathbb{E}_\mu[g(\xi)] - \mathbb{E}_\nu[g(\xi)]\| \le L_g W_1(\mu, \nu). \tag{32}$$

Step 4: Combining the bounds. Substituting back into the bound from Step 2:

$$\|y^*(\mu) - y^*(\nu)\| = \|y^*_\mu - y^*_\nu\| \le \frac{L_g}{\alpha} W_1(\mu, \nu). \tag{33}$$

Hence y^* is (L_g/α)-Lipschitz continuous with respect to W_1, completing the proof. □

D. Proof of Proposition 3.3

We restate the result for completeness.

Proposition 3.3. Let P be a probability distribution on $\mathbb{R}^d$ with finite second moment, and let $\zeta \sim \mathcal{N}(0, \sigma^2 I_d)$ be independent Gaussian noise. Define the perturbed distribution $Q = P * \mathcal{N}(0, \sigma^2 I_d)$ as the law of $Y = X + \zeta$ where $X \sim P$. Then:

$$W_1(P, Q) \le \sigma\sqrt{d}. \tag{34}$$

Proof. We proceed by constructing an explicit coupling between P and Q and bounding the expected transport cost.

Step 1: Constructing an explicit coupling. Let $X \sim P$ and let $\zeta \sim \mathcal{N}(0, \sigma^2 I_d)$ be independent of X. Define $Y = X + \zeta$, so that $Y \sim Q$ by definition. The pair $(X, Y) = (X, X + \zeta)$ constitutes a valid coupling of (P, Q), since the marginal distributions of X and Y are P and Q respectively.

Step 2: Bounding the Wasserstein-1 distance. By the primal definition of the Wasserstein-1 distance as an infimum over all couplings:

$$W_1(P, Q) \le \mathbb{E}\left[\|X - Y\|_2\right] = \mathbb{E}\left[\|X - (X + \zeta)\|_2\right] = \mathbb{E}\left[\|\zeta\|_2\right]. \tag{35}$$

Step 3: Computing $\mathbb{E}[\|\zeta\|_2]$. Since $\zeta \sim \mathcal{N}(0, \sigma^2 I_d)$, we write $\zeta = \sigma u$ where $u \sim \mathcal{N}(0, I_d)$. Therefore:

$$\mathbb{E}\left[\|\zeta\|_2\right] = \sigma \cdot \mathbb{E}\left[\|u\|_2\right]. \tag{36}$$

Applying Jensen's inequality to the concave function $t \mapsto \sqrt{t}$:

$$\mathbb{E}\left[\|u\|_2\right] \le \sqrt{\mathbb{E}\left[\|u\|_2^2\right]}. \tag{37}$$

Since $u \sim \mathcal{N}(0, I_d)$, we have $\|u\|_2^2 = \sum_{i=1}^d u_i^2$ where $u_i \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, 1)$. Therefore:

$$\mathbb{E}\left[\|u\|_2^2\right] = \sum_{i=1}^d \mathbb{E}[u_i^2] = d. \tag{38}$$

Step 4: Combining the bounds. Substituting back:

$$W_1(P, Q) \le \mathbb{E}\left[\|\zeta\|_2\right] = \sigma \cdot \mathbb{E}\left[\|u\|_2\right] \le \sigma \cdot \sqrt{\mathbb{E}\left[\|u\|_2^2\right]} = \sigma\sqrt{d}. \tag{39}$$

This completes the proof. □

E. Proof of Lemma 4.1

We restate the result for completeness.

Lemma 4.1. Let $\mathcal{K} = [0,1]^D$. For any $\varepsilon \in (0,1)$, let $Q_\varepsilon \subset \mathcal{P}_N(\mathcal{K})$ be a maximal ε-separated subset with respect to W_1. Then Q_ε is an ε-cover of $\mathcal{P}_N(\mathcal{K})$, and its log-cardinality satisfies

$$\log |Q_\varepsilon| \leq \mathcal{O}\big(\varepsilon^{-D} \log(N + \varepsilon^{-D})\big)\,, \tag{40}$$

where the implicit constant depends only on D.

Proof. We proceed in three steps: constructing the net, bounding its cardinality via a grid cover, and evaluating the resulting combinatorial bound.

Step 1: Construction of Q_ε. Let $Q_\varepsilon \subset \mathcal{P}_N(\mathcal{K})$ be a maximal ε-separated subset with respect to W_1, meaning $W_1(\nu_j, \nu_k) \geq \varepsilon$ for all distinct $\nu_j, \nu_k \in Q_\varepsilon$, and no further point can be added while maintaining this separation. Such a set exists by a standard greedy argument: repeatedly add points at W_1-distance $> \varepsilon$ from all previously added points until no such point remains. Maximality implies the covering property: if some $\mu \in \mathcal{P}_N(\mathcal{K})$ were not within ε of any point in Q_ε, we could add μ while maintaining separation, contradicting maximality. Hence Q_ε is simultaneously an ε-separated set and an ε-cover of $\mathcal{P}_N(\mathcal{K})$.

Step 2: Bounding the cardinality via a grid cover. Since Q_ε is ε-separated, any two distinct centers $\nu_j, \nu_k \in Q_\varepsilon$ satisfy $W_1(\nu_j, \nu_k) \geq \varepsilon$, so the open W_1-balls $\{B_{W_1}(\nu, \varepsilon/2)\}_{\nu \in Q_\varepsilon}$ are pairwise disjoint. Therefore:

$$|Q_\varepsilon| \leq \mathcal{N}(\mathcal{P}_N(\mathcal{K}), W_1, \varepsilon/2), \tag{41}$$

where $\mathcal{N}$ denotes the $\varepsilon/2$-covering number. To bound this, partition $\mathcal{K} = [0,1]^D$ into $M = \lceil 2/\varepsilon \rceil^D \asymp \varepsilon^{-D}$ cells of side length $\varepsilon/2$ with centroids $g_1, \ldots, g_M$. For any empirical measure $\mu = \frac{1}{N}\sum_i \delta_{x_i}$, project each atom x_i to its nearest centroid $g_{j(i)}$, yielding a quantized measure ν. Since each atom moves at most the cell side length $\varepsilon/2$ under projection, the W_1 distance satisfies:

$$W_1(\mu, \nu) \leq \frac{1}{N}\sum_{i=1}^{N} \|x_i - g_{j(i)}\| \leq \frac{\varepsilon}{2}. \tag{42}$$

The collection of all such quantized measures forms a valid $\varepsilon/2$-cover of $\mathcal{P}_N(\mathcal{K})$, consistent with the packing bound above. Since Q_ε is also an ε-cover by Step 1, all subsequent approximation arguments use cover radius ε, not $\varepsilon/2$.

Step 3: Evaluating the combinatorial bound. Each quantized measure is uniquely determined by allocating N atoms across M bins, i.e., a nonnegative integer solution $(k_1, \ldots, k_M)$ to $\sum_j k_j = N$. By the stars and bars theorem, the number of such allocations is $\binom{N+M-1}{M-1}$. Applying the standard inequality $\binom{a}{b} \leq (ea/b)^b$ with $a = N + M - 1$ and $b = M - 1$:

$$\mathcal{N}(\mathcal{P}_N(\mathcal{K}), W_1, \varepsilon/2) \leq \binom{N+M-1}{M-1} \leq \left(\frac{e(N+M)}{M}\right)^M. \tag{43}$$

Taking logarithms and substituting $M = \lceil 2/\varepsilon \rceil^D \asymp \varepsilon^{-D}$:

$$\log |Q_\varepsilon| \leq M \log \frac{e(N+M)}{M} \leq M\left(1 + \log \frac{N+M}{M}\right) \leq M \log(N+M) = \mathcal{O}\big(\varepsilon^{-D}\log(N+\varepsilon^{-D})\big)\,. \tag{44}$$

This completes the proof. □

F. Proof of Lemma 4.2

Lemma 4.2.

The multiscale encoder Ψ is Lipschitz with respect to W_1: for any $\mu, \nu \in \mathcal{P}_N(\mathcal{K})$,

$$\|\Psi(\mu) - \Psi(\nu)\|_2 \leq \sqrt{8/3} \cdot W_1(\mu, \nu), \tag{45}$$

Proof. The gradient of $\varphi_{s,r}(x) = \sqrt{2}\cos(a_{s,r}^\top x + b_{s,r})$ satisfies

$$\|\nabla_x \varphi_{s,r}(x)\| = \sqrt{2}|\sin(a_{s,r}^\top x + b_{s,r})| \cdot \|a_{s,r}\| \le \sqrt{2} \cdot 2^s.$$

So $\varphi_{s,r}$ is $\sqrt{2} \cdot 2^s$-Lipschitz. By Kantorovich–Rubinstein:

$$\left| \int \varphi_{s,r}\, d(\mu - \nu) \right| \le \sqrt{2} \cdot 2^s \cdot W_1(\mu, \nu).$$

Since blocks are in orthogonal subspaces:

$$\|\Psi(\mu) - \Psi(\nu)\|_2^2 = \sum_{s=0}^{S} \frac{w_s^2}{m_s} \sum_{r=1}^{m_s} \Big(\int \varphi_{s,r}\, d(\mu - \nu) \Big)^2 \le 2W_1(\mu, \nu)^2 \sum_{s=0}^{S} w_s^2 \cdot 2^{2s}.$$

Substituting $w_s = 2^{-2s}$ gives $w_s^2 \cdot 2^{2s} = 2^{-2s}$. Bounding by the infinite series:

$$\sum_{s=0}^{S} 2^{-2s} \le \sum_{s=0}^{\infty} 2^{-2s} = \frac{1}{1 - 2^{-2}} =: 4/3.$$

Taking square roots gives the result. □ □

G. Proof of Lemma 4.3

We first restate the Lemma for completeness.

Lemma 4.3.

Let Q_ε be the maximal ε-separated net from Lemma 4.1. With probability at least $1 - \delta$, the multiscale encoder Ψ satisfies the following Hölder separation bound simultaneously for all distinct pairs $\nu_j, \nu_k \in Q_\varepsilon$:

$$\|\Psi(\nu_j) - \Psi(\nu_k)\|_2 \ge c_D \cdot W_1(\nu_j, \nu_k)^\alpha, \tag{46}$$

Here $\alpha = (D + 6)/2$ and the constant $c_D > 0$ depends only on the domain and ambient dimension. This guarantee holds provided

$$m \ge C_D \left(\varepsilon^{-(3D+4)} \log(N + \varepsilon^{-D}) + \varepsilon^{-(2D+4)} \log \frac{1}{\delta} \right),$$

which simplifies to $m = \mathcal{O}(C_D \varepsilon^{-(3D+4)} \log(N + \varepsilon^{-D}))$ for fixed confidence level δ.

Constant Summary for Appendix G. The proof below introduces several constants from the Fourier-energy, concentration, and dimension-count arguments. Table 2 summarizes only these constants; variables such as r, λ, S, s^*, m_s, and Z_ℓ are defined directly in the text when used.

Table 2. Constants used in the lower-separation proof. Here $\mathcal{K}_1 := \{x \in \mathbb{R}^D : \mathrm{dist}(x, \mathcal{K}) \le 1\}$, q is the smooth spatial cutoff, $g = q\overline{f}$, and h is the smoothing kernel.

Constant	Definition	Role
M_D	$(\mathrm{diam}(\mathcal{K}) + 1)\lvert\mathcal{K}_1\rvert^{1/2}$; for $\mathcal{K} = [0,1]^D$, $M_D \le (\sqrt{D} + 1)3^{D/2}$.	Uniform L^2 bound for the compactly supported witness g.
L_q	$\|q\|_{\mathrm{Lip}}$.	Lipschitz constant of the spatial cutoff q.
L_D	$1 + L_q(\mathrm{diam}(\mathcal{K}) + 1)$.	Lipschitz bound for $g = q\overline{f}$; controls the low-pass smoothing error.
A_0	$\int_{\mathbb{R}} \lvert\eta(t)\rvert\, dt$.	One-dimensional L^1 constant of the inverse Fourier transform η.

Continued on next page

Constant	Definition	Role
A_1	$\int_{\mathbb{R}} \lvert t\rvert \, \lvert\eta(t)\rvert \, dt$.	One-dimensional first-moment constant of η.
A_h	$\int_{\mathbb{R}^D} \lvert h(x)\rvert \lVert x\rVert_2 \, dx \le D^{3/2} A_1 A_0^{D-1}$.	First absolute moment of the smoothing kernel.
$c_{0,D}$	$\max\{1, \operatorname{diam}(\mathcal{K}), 8L_D A_h\}$.	Sets the smoothing scale and the matched Fourier radius; the harmless $\operatorname{diam}(\mathcal{K})$ enlargement covers the dyadic case $s^* = 0$.
$c_{1,D}$	$\frac{(2\pi)^D}{16 M_D^2 \lvert B(0,1)\rvert c_{0,D}^D}$.	Gives $\frac{1}{\lvert B(0,c_{0,D}/r)\rvert} \int_{B(0,c_{0,D}/r)} \lvert\widehat{\lambda}(\omega)\rvert^2 \, d\omega \ge c_{1,D} r^{D+2}$.
$c_{2,D}$	$2^{-D} c_{1,D}$.	Accounts for comparing the Fourier-energy ball with the sampled dyadic ball.
$c_{3,D}$	$\sqrt{c_{2,D}/2}$.	Block-norm lower-bound constant after Hoeffding.
$c_{4,D}$	$\frac{128}{c_{2,D}^2 c_{0,D}^{2D+4}}$.	Sufficient block-size prefactor for one pair.
$c_{5,D}$	$\frac{c_{3,D}}{(2c_{0,D})^2}$.	Full-embedding separation constant for the current $w_s = 2^{-2s}$ proof.
c_D	$c_{5,D}$.	Final lower-separation constant appearing in Lemma 4.3.
$c_{6,D}$	$\frac{2^{2D+4}}{2^{2D+4}-1}$.	Geometric-series constant for $\sum_{s=0}^{S} 2^{s(2D+4)}$.
$c_{7,D}$	$2c_{4,D}c_{6,D}$.	Intermediate dimension-count constant.
$c_{8,D}$	$(2c_{0,D})^{2D+4}$.	Converts $2^{S(2D+4)}$ into $\varepsilon^{-(2D+4)}$.
$c_{9,D}$	$c_{7,D}c_{8,D}$.	Final dimension-count prefactor.
C_D	$c_{9,D}$.	Final width constant appearing in Lemma 4.3.

Proof. We first define the multi-scale random Fourier embedding used in the proof. For each scale index $s = 0, 1, \ldots, S$, define the frequency domain

$$\Omega_s := B(0, 2^s) = \{\omega \in \mathbb{R}^D : \lVert\omega\rVert_2 \le 2^s\}.$$

For each scale $s = 0, \ldots, S$, we sample m_s, which will be chosen later, independent frequency-phase pairs

$$(a_{s,\ell}, b_{s,\ell}), \qquad \ell = 1, \ldots, m_s,$$

where

$$a_{s,\ell} \sim \operatorname{Unif}(\Omega_s), \qquad b_{s,\ell} \sim \operatorname{Unif}[0, 2\pi].$$

Define the corresponding random Fourier features, for $s = 0, \ldots, S$ and $\ell = 1, \ldots, m_s$, by

$$\varphi_{s,\ell}(x) = \sqrt{2}\cos(a_{s,\ell}^\top x + b_{s,\ell}).$$

For a probability measure ν defined over the domain $\mathcal{K}$, we define the scale-s block embedding by

$$\Phi_s(\nu) = \left(\int \varphi_{s,1} \, d\nu, \ldots, \int \varphi_{s,m_s} \, d\nu\right) := \left(\int_{\mathcal{K}} \varphi_{s,1}(x) \, d\nu(x), \ldots, \int_{\mathcal{K}} \varphi_{s,m_s}(x) \, d\nu(x)\right) \in \mathbb{R}^{m_s}.$$

Finally, define the full multi-scale embedding by

$$\Psi(\nu) := \left(\frac{w_0}{\sqrt{m_0}}\Phi_0(\nu), \frac{w_1}{\sqrt{m_1}}\Phi_1(\nu), \ldots, \frac{w_S}{\sqrt{m_S}}\Phi_S(\nu)\right),$$

where, for $s = 0, \ldots, S$,

$$w_s = 2^{-2s}.$$

Before we proceed, we recall the notation. The index $s \in \{0, \ldots, S\}$ is called the scale index. Scale s corresponds to the frequency radius 2^s, because all frequency vectors in this scale are sampled from $\Omega_s = B(0, 2^s)$. Thus, larger s allows larger frequency vectors and therefore more rapidly oscillating cosine features.

The collection of random Fourier features

$$\{\varphi_{s,1}, \ldots, \varphi_{s,m_s}\}$$

is called the s-th frequency block, or simply block s. Each $\varphi_{s,\ell}$ is a function on $\mathbb{R}^D$, and in this proof it is evaluated on the compact domain $\mathcal{K}$, where the input measures are supported. Thus, block s is the group of m_s random Fourier features associated with the same frequency domain Ω_s.

For a probability measure ν supported on $\mathcal{K}$, the vector

$$\Phi_s(\nu) = \left(\int \varphi_{s,1}\, d\nu, \ldots, \int \varphi_{s,m_s}\, d\nu \right)$$

is called the scale-s block embedding vector of ν. It records the averages of the m_s features in block s under the measure ν.

The full embedding $\Psi(\nu)$ is obtained by concatenating the weighted and normalized block embeddings. Here, concatenation means that we place the vectors

$$\frac{w_s}{\sqrt{m_s}}\Phi_s(\nu), \qquad s = 0, \ldots, S,$$

one after another to form one long vector. The factor $1/\sqrt{m_s}$ normalizes the block, while the weight $w_s = 2^{-2s}$ downweights higher-frequency blocks.

The intuition behind this construction, which will be made precise in the proof below, is that different Wasserstein separations are detected at different frequency resolutions. A pair of measures with distance r is matched with a scale whose frequency radius is of order $1/r$. The multi-scale embedding includes all scales $s = 0, \ldots, S$, so for every separated pair there is an appropriate block that can detect it. The scale weights $w_s = 2^{-2s}$ ensure that the sum over all scales is well controlled.

Because a multi-scale embedding $\Psi(\nu)$ is a concatenation of block vectors, the squared Euclidean distance between two multi-scale embedding is the sum of the squared Euclidean distances of the corresponding block vectors. Hence, for any two probability measures ν and ν',

$$\|\Psi(\nu) - \Psi(\nu')\|_2^2 = \sum_{s=0}^{S} w_s^2 \frac{1}{m_s} \|\Phi_s(\nu) - \Phi_s(\nu')\|_2^2.$$

Step 1: Fix one pair and choose the appropriate frequency block.

Recall that Q_ε is the finite set constructed in Lemma 4.1. By definition, Q_ε is ε-separated under the Wasserstein distance W_1. That is, for any two distinct elements $\nu_j, \nu_k \in Q_\varepsilon$,

$$W_1(\nu_j, \nu_k) \geq \varepsilon.$$

Fix two distinct measures $\nu_j, \nu_k \in Q_\varepsilon$, and define

$$r := W_1(\nu_j, \nu_k), \qquad \lambda := \nu_j - \nu_k.$$

Then

$$r \geq \varepsilon.$$

Let $c_{0,D} > 1$ be the constant from Lemma H.1, and we have $\frac{c_{0,D}}{\varepsilon} \geq 1$. We choose S as the smallest integer satisfying

$$2^S \geq \frac{c_{0,D}}{\varepsilon}.$$

For the fixed pair (ν_j, ν_k) with a Wasserstein distance r, define s^* as the first block whose frequency radius is at least $c_{0,D}/r$:

$$s^* := \min\left\{ s \in \{0, 1, \ldots, S\} : 2^s \geq \frac{c_{0,D}}{r} \right\}.$$

Observing $r \geq \varepsilon$ and $2^S \geq c_{0,D}/\varepsilon$, both s^* and S are non-negative integers, and the definition of s^*, we have

$$\frac{c_{0,D}}{r} \leq \min\{2^{s^*}, \frac{c_{0,D}}{\varepsilon}\} \leq 2^S.$$

Hence s^* is a valid scale index and satisfies $s^* \leq S$. By the definition of s^* as the first scale with $2^{s^*} \geq c_{0,D}/r$, we also have

$$\frac{c_{0,D}}{r} \leq 2^{s^*} < \frac{2c_{0,D}}{r}.$$

We have therefore identified a valid block s^* whose frequency radius is comparable to $1/r$. We also note that for any valid $r \geq \varepsilon$, the corresponding s^* lies in $[0, S]$. This observation will be useful in **Step 5** later. In the next step, we show that this selected block has a nontrivial expected signal for separating the fixed pair (ν_j, ν_k).

Step 2: Show that the identified frequency block has nontrivial expected signal.

We now study the frequency block s^*, i.e., $\{\varphi_{s^*,1}, \ldots, \varphi_{s^*,m_{s^*}}\}$.

By definition of the block embedding,

$$\Phi_{s^*}(\nu) = \left(\int \varphi_{s^*,1}\, d\nu, \ldots, \int \varphi_{s^*,m_{s^*}}\, d\nu \right).$$

Thus, using $\lambda = \nu_j - \nu_k$, we have

$$\Phi_{s^*}(\nu_j) - \Phi_{s^*}(\nu_k) = \left(\int \varphi_{s^*,1}\, d\lambda, \ldots, \int \varphi_{s^*,m_{s^*}}\, d\lambda \right),$$

where

$$\int \varphi_{s^*,\ell}\, d\lambda := \int \varphi_{s^*,\ell}\, d\nu_j - \int \varphi_{s^*,\ell}\, d\nu_k.$$

We call $\int \varphi_{s^*,\ell}\, d\lambda$ the signal of the ℓ-th feature for the fixed pair (ν_j, ν_k), because it measures the difference between the average value of this feature under the two measures.

Therefore,

$$\|\Phi_{s^*}(\nu_j) - \Phi_{s^*}(\nu_k)\|_2^2 = \sum_{\ell=1}^{m_{s^*}} \left(\int \varphi_{s^*,\ell}\, d\lambda \right)^2 = \sum_{\ell=1}^{m_{s^*}} Z_\ell.$$

where

$$Z_\ell := \left(\int \varphi_{s^*,\ell}\, d\lambda \right)^2, \quad \ell = 1, \ldots, m_{s^*}.$$

Dividing both sides by m_{s^*} gives

$$\frac{1}{m_{s^*}} \|\Phi_{s^*}(\nu_j) - \Phi_{s^*}(\nu_k)\|_2^2 = \frac{1}{m_{s^*}} \sum_{\ell=1}^{m_{s^*}} Z_\ell.$$

We compute the expectation of one Z_ℓ. Temporarily write

$$a := a_{s^*,\ell}, \qquad b := b_{s^*,\ell}.$$

For a fixed frequency a, define

$$P(a) := \int \cos(a^\top x)\, d\lambda(x), \qquad Q(a) := \int \sin(a^\top x)\, d\lambda(x).$$

Using

$$\cos(a^\top x + b) = \cos(a^\top x) \cos b - \sin(a^\top x) \sin b,$$

we get

$$\int \cos(a^\top x + b)\, d\lambda(x) = P(a) \cos b - Q(a) \sin b.$$

Since the feature also includes a constant $\sqrt{2}$,

$$Z_\ell = 2\big(P(a) \cos b - Q(a) \sin b\big)^2.$$

Averaging over the random phase b, and using

$$\mathbb{E}_b[\cos^2 b] = \frac{1}{2}, \qquad \mathbb{E}_b[\sin^2 b] = \frac{1}{2}, \qquad \mathbb{E}_b[\cos b \sin b] = 0,$$

we obtain

$$\mathbb{E}_b[Z_\ell] = P(a)^2 + Q(a)^2.$$

Now define the Fourier transform of λ by

$$\widehat{\lambda}(a) := \int e^{ia^\top x}\, d\lambda(x).$$

Since

$$e^{ia^\top x} = \cos(a^\top x) + i\sin(a^\top x),$$

we have

$$\widehat{\lambda}(a) = P(a) + iQ(a).$$

Therefore,

$$|\widehat{\lambda}(a)|^2 = P(a)^2 + Q(a)^2.$$

Thus,

$$\mathbb{E}_b[Z_\ell] = |\widehat{\lambda}(a)|^2.$$

Finally, we average over a. Since a is uniformly distributed on the ball $B(0, 2^{s^*})$, its density is

$$\frac{1}{|B(0, 2^{s^*})|},$$

where $|B(0, 2^{s^*})|$ denotes the D-dimensional Lebesgue volume of the ball $B(0, 2^{s^*})$. Therefore,

$$\mathbb{E}[Z_\ell] = \frac{1}{|B(0, 2^{s^*})|} \int_{B(0,2^{s^*})} |\widehat{\lambda}(\omega)|^2\, d\omega.$$

From **Step 1**, we have

$$\frac{c_{0,D}}{r} \le 2^{s^*} < \frac{2c_{0,D}}{r}.$$

Therefore,

$$B(0, c_{0,D}/r) \subseteq B(0, 2^{s^*}),$$

Since $2^{s^*} < 2c_{0,D}/r$ and the volume of a D-dimensional ball scales as the D-th power of its radius, we have

$$|B(0, 2^{s^*})| \le |B(0, 2c_{0,D}/r)| = 2^D |B(0, c_{0,D}/r)|.$$

Therefore,

$$\begin{aligned}\mathbb{E}[Z_\ell] &= \frac{1}{|B(0, 2^{s^*})|} \int_{B(0,2^{s^*})} |\widehat{\lambda}(\omega)|^2\, d\omega \\ &\ge \frac{1}{2^D |B(0, c_{0,D}/r)|} \int_{B(0,c_{0,D}/r)} |\widehat{\lambda}(\omega)|^2\, d\omega.\end{aligned}$$

Now applying Lemma H.1, we obtain

$$\mathbb{E}[Z_\ell] \ge 2^{-D} c_{1,D} r^{D+2}, \quad \ell = 1, \dots, m_{s^*}.$$

where $c_{1,D}$ is a constant from Lemma H.1. Define

$$c_{2,D} := 2^{-D} c_{1,D}.$$

Then we arrive at

$$\mathbb{E}[Z_\ell] \ge c_{2,D} r^{D+2}, \quad \ell = 1, \dots, m_{s^*}.$$

By now, we have shown that each random feature in the identified block s^* has a nontrivial expected squared signal:

$$\mathbb{E}[Z_\ell] \ge c_{2,D} r^{D+2}, \qquad \ell = 1, \dots, m_{s^*}.$$

Next, we show that the empirical average of these squared differences over the whole block is also large with high probability.

Step 3: Bound the value of empirical observations.

We first show that each Z_ℓ is bounded. Recall that $Z_\ell = \left(\int \varphi_{s^*,\ell}\, d\lambda\right)^2,\ \ \ell = 1, \ldots, m_{s^*}$, and

$$\varphi_{s^*,\ell}(x) = \sqrt{2}\cos(a_{s^*,\ell}^\top x + b_{s^*,\ell}),$$

we have for every x

$$|\varphi_{s^*,\ell}(x)| \le \sqrt{2}.$$

Also recall that

$$\int \varphi_{s^*,\ell}\, d\lambda = \int \varphi_{s^*,\ell}\, d\nu_j - \int \varphi_{s^*,\ell}\, d\nu_k.$$

Since ν_j and ν_k are probability measures, each integral on the right-hand side is bounded in absolute value by $\sqrt{2}$. Hence

$$\left|\int \varphi_{s^*,\ell}\, d\lambda\right| \le 2\sqrt{2}.$$

Squaring gives

$$0 \le Z_\ell \le 8.$$

The random variables $Z_1, \ldots, Z_{m_{s^*}}$ are independent because the corresponding random features are sampled independently. We now apply Hoeffding's inequality.[6] In our case, $Y_\ell = Z_\ell$, $m = m_{s^*}$, and $B = 8$. For any $t > 0$,

$$\mathbb{P}\left[\frac{1}{m_{s^*}}\sum_{\ell=1}^{m_{s^*}} Z_\ell \le \mathbb{E}[Z_\ell] - t\right] \le \exp\left(-\frac{2m_{s^*}t^2}{8^2}\right).$$

From **Step 2**, we have

$$\mathbb{E}[Z_\ell] \ge c_{2,D} r^{D+2},\ \ \ell = 1, \ldots, m_{s^*}.$$

Choose

$$t = \frac{1}{2} c_{2,D} r^{D+2}.$$

Then Hoeffding's inequality implies

$$\mathbb{P}\left[\frac{1}{m_{s^*}}\sum_{\ell=1}^{m_{s^*}} Z_\ell \le \left(1 - \frac{1}{2}\right) c_{2,D} r^{D+2}\right] \le \mathbb{P}\left[\frac{1}{m_{s^*}}\sum_{\ell=1}^{m_{s^*}} Z_\ell \le \mathbb{E}[Z_\ell] - \frac{1}{2} c_{2,D} r^{D+2}\right] \le \exp\left(-\frac{m_{s^*} c_{2,D}^2 r^{2D+4}}{128}\right).$$

Before we proceed, from the last paragraph in **Step 1**, we have

$$2^{s^*} \ge \frac{c_{0,D}}{r},$$

Consequently, raising the powers of both sides by $2D + 4$ and re-arranging terms, we have

$$c_{0,D}^{-(2D+4)} 2^{s^*(2D+4)} \ge r^{-(2D+4)}.$$

Now for any $\delta_{\text{pair}} \in (0, 1)$, if m_{s^*} satisfies the following requirement

$$m_{s^*} \ge \frac{128}{c_{2,D}^2}\left(c_{0,D}^{-(2D+4)} 2^{s^*(2D+4)}\right) \log\frac{1}{\delta_{\text{pair}}} \ge \frac{128}{c_{2,D}^2} r^{-(2D+4)} \log\frac{1}{\delta_{\text{pair}}},$$

[6] We use the following standard form of Hoeffding's inequality: if $Y_1, \ldots, Y_m$ are independent random variables satisfying $0 \le Y_\ell \le B$, then for any $t > 0$,

$$\mathbb{P}\left[\frac{1}{m}\sum_{\ell=1}^{m} Y_\ell \le \mathbb{E}[Y_\ell] - t\right] \le \exp\left(-\frac{2mt^2}{B^2}\right).$$

then the probability

$$\mathbb{P}\left[\frac{1}{m_{s^*}}\sum_{\ell=1}^{m_{s^*}} Z_\ell \le \frac{1}{2}c_{2,D}r^{D+2}\right] \le \delta_{\text{pair}}.$$

That is, with probability at least $1-\delta_{\text{pair}}$,

$$8 \ge \frac{1}{m_{s^*}}\sum_{\ell=1}^{m_{s^*}} Z_\ell \ge \frac{1}{2}c_{2,D}r^{D+2}.$$

(As a sanity check, by the definition of $c_{2,D}$ in **Step 2**, the right-hand-side value $\frac{1}{2}c_{2,D}r^{D+2}$ is indeed no larger than than the left-hand-side value 8.) Thus,

$$\frac{1}{m_{s^*}}\|\Phi_{s^*}(\nu_j)-\Phi_{s^*}(\nu_k)\|_2^2 = \sum_{\ell=1}^{m_{s^*}} Z_\ell \ge \frac{1}{2}c_{2,D}r^{D+2},$$

with probability at least $1-\delta_{\text{pair}}$. Taking square roots, we now arrive at the observation that if

$$m_{s^*} \ge c_{4,D}2^{s^*(2D+4)}\log\frac{1}{\delta_{\text{pair}}},$$

where

$$c_{4,D} := \frac{128}{c_{2,D}^2 c_{0,D}^{2D+4}}.$$

then with probability at least $1-\delta_{\text{pair}}$,

$$\frac{1}{\sqrt{m_{s^*}}}\|\Phi_{s^*}(\nu_j)-\Phi_{s^*}(\nu_k)\|_2 \ge c_{3,D}r^{(D+2)/2},$$

where

$$c_{3,D} := \sqrt{\frac{c_{2,D}}{2}}.$$

By now, we have shown that the identified block s^* separates the fixed pair (ν_j,ν_k) after normalization by $\sqrt{m_{s^*}}$, provided that block s^* contains enough random features. Next, we pass from this single-block lower bound to a lower bound for the full multi-scale embedding Ψ.

Step 4: Pass from the selected block to the full embedding.

From **Step 3**, with probability at least $1-\delta_{\text{pair}}$,

$$\frac{1}{\sqrt{m_{s^*}}}\|\Phi_{s^*}(\nu_j)-\Phi_{s^*}(\nu_k)\|_2 \ge c_{3,D}r^{(D+2)/2},$$

where $c_{3,D} := \sqrt{\frac{c_{2,D}}{2}}$ and $c_{2,D}$ is defined in Lemma H.1. The full embedding contains all blocks, and its squared distance is

$$\|\Psi(\nu_j)-\Psi(\nu_k)\|_2^2 = \sum_{s=0}^{S} w_s^2 \frac{1}{m_s}\|\Phi_s(\nu_j)-\Phi_s(\nu_k)\|_2^2.$$

All terms in the summation are nonnegative. Hence, the contribution from the selected block s^* alone gives a lower bound:

$$\|\Psi(\nu_j)-\Psi(\nu_k)\|_2 \ge w_{s^*}\frac{1}{\sqrt{m_{s^*}}}\|\Phi_{s^*}(\nu_j)-\Phi_{s^*}(\nu_k)\|_2.$$

Using the lower bound from **Step 3** for the selected block s^*,

$$\|\Psi(\nu_j)-\Psi(\nu_k)\|_2 \ge w_{s^*}c_{3,D}r^{(D+2)/2}.$$

By the definition of w_s, i.e., $w_s = 2^{-2s}$, and $2^{s^*} < \frac{2c_{0,D}}{r}$ from **Step 1**, we have

$$w_{s^*} = 2^{-2s^*} > \left(\frac{r}{2c_{0,D}}\right)^2.$$

Substituting this into the full embedding lower bound gives

$$\|\Psi(\nu_j) - \Psi(\nu_k)\|_2 \geq c_{3,D} \left(\frac{r}{2c_{0,D}}\right)^2 r^{(D+2)/2}.$$

Combining the powers of r,

$$r^2 r^{(D+2)/2} = r^{(D+6)/2}.$$

Hence

$$\|\Psi(\nu_j) - \Psi(\nu_k)\|_2 \geq c_{5,D} r^{(D+6)/2},$$

where

$$c_{5,D} := \frac{c_{3,D}}{(2c_{0,D})^2}.$$

Since $r = W_1(\nu_j, \nu_k)$, this becomes

$$\|\Psi(\nu_j) - \Psi(\nu_k)\|_2 \geq c_{5,D} W_1(\nu_j, \nu_k)^{(D+6)/2}.$$

By now, we have lifted the normalized separation from the selected block s^* to the full multi-scale embedding Ψ. For the fixed pair (ν_j, ν_k), we have shown that, with probability at least $1 - \delta_{\text{pair}}$,

$$\|\Psi(\nu_j) - \Psi(\nu_k)\|_2 \geq c_{5,D} W_1(\nu_j, \nu_k)^{(D+6)/2}.$$

Next, we remove the dependence on the fixed pair and make this lower bound hold simultaneously for all distinct pairs in Q_ε.

Step 5: Make the bound hold for all pairs in Q_ε.

The argument above was for one fixed pair. We now make the result hold simultaneously for all distinct pairs in Q_ε.

Since Q_ε contains $|Q_\varepsilon|$ measures, the number of ordered pairs is at most

$$|Q_\varepsilon|^2.$$

For one fixed pair, the failure probability is at most δ_{pair}. By the union bound, the probability that at least one pair fails is at most

$$|Q_\varepsilon|^2 \delta_{\text{pair}}.$$

Now for any $\delta \in (0, 1)$, choose

$$\delta_{\text{pair}} = \frac{\delta}{|Q_\varepsilon|^2}.$$

Then

$$|Q_\varepsilon|^2 \delta_{\text{pair}} = \delta.$$

Thus, with probability at least $1 - \delta$, no pair fails.

It remains to choose the number of features in each block. From Step 3, for a pair whose selected block is s^*, it is enough to require

$$m_{s^*} \geq c_{4,D} 2^{s^*(2D+4)} \log \frac{1}{\delta_{\text{pair}}}.$$

Substituting

$$\delta_{\text{pair}} = \frac{\delta}{|Q_\varepsilon|^2},$$

we obtain

$$\log \frac{1}{\delta_{\text{pair}}} = 2 \log |Q_\varepsilon| + \log \frac{1}{\delta}.$$

The selected block s^* depends on the pair (ν_j, ν_k), because s^* is chosen according to the distance

$$r = W_1(\nu_j, \nu_k).$$

The selected block s^* depends on the pair (ν_j, ν_k), because s^* is chosen according to the distance

$$r = W_1(\nu_j, \nu_k).$$

Different pairs may therefore select different blocks. For a pair whose selected block is s^*, the sufficient condition above becomes

$$m_{s^*} \geq c_{4,D} 2^{s^*(2D+4)} \left(2 \log |Q_\varepsilon| + \log \frac{1}{\delta} \right).$$

Since the embedding must be fixed before we know which pair is being considered, we impose this requirement on every possible block. Thus, for each $s = 0, \ldots, S$, we choose

$$m_s = \left\lceil c_{4,D} 2^{s(2D+4)} \left(2 \log |Q_\varepsilon| + \log \frac{1}{\delta} + 1 \right) \right\rceil, \qquad s = 0, \ldots, S.$$

The additional $+1$ is included only to handle the ceiling function in the dimension count. It ensures that the quantity inside the ceiling is bounded below by a positive constant, so that we can use $\lceil x \rceil \leq 2x$ without introducing a separate additive $(S+1)$ term.

With this choice, whichever block s^* is selected for any pair, the required feature condition is satisfied. Indeed, by Step 1, every valid pair (ν_j, ν_k) with $r = W_1(\nu_j, \nu_k) \geq \varepsilon$ selects a block

$$s^* \in \{0, \ldots, S\}$$

satisfying

$$\frac{c_{0,D}}{r} \leq 2^{s^*} < \frac{2c_{0,D}}{r}.$$

Since we have chosen m_s to satisfy the required lower bound for every $s = 0, \ldots, S$, the selected block s^* also satisfies the Step 3 requirement. Therefore, with probability at least $1 - \delta$, for all distinct $\nu_j, \nu_k \in Q_\varepsilon$,

$$\|\Psi(\nu_j) - \Psi(\nu_k)\|_2 \geq c_{5,D} W_1(\nu_j, \nu_k)^{(D+6)/2}.$$

By now, we have upgraded the fixed-pair result to a uniform result over the finite set Q_ε. With the above choice of block sizes m_s, the required feature condition is satisfied for whichever block s^* is selected by each pair. Therefore, with probability at least $1 - \delta$, every distinct pair $\nu_j, \nu_k \in Q_\varepsilon$ satisfies

$$\|\Psi(\nu_j) - \Psi(\nu_k)\|_2 \geq c_{5,D} W_1(\nu_j, \nu_k)^{(D+6)/2}.$$

It remains only to count the total number of random features, i.e., the total number of scalars, used by the multi-scale embedding.

For convenience, we refer $c_D := c_{5,D}$ in Lemma 4.3, Sec. 4. This gives:

$$\|\Psi(\nu_j) - \Psi(\nu_k)\|_2 \geq c_D W_1(\nu_j, \nu_k)^{(D+6)/2}.$$

Step 6: Count the total embedding dimension.

The total embedding dimension is

$$m = \sum_{s=0}^{S} m_s.$$

Using the definition of m_s, and the elementary inequality $\lceil x \rceil \leq 2x$ for $x \geq 1$, we obtain

$$m \leq 2c_{4,D} \left(2 \log |Q_\varepsilon| + \log \frac{1}{\delta} + 1\right) \sum_{s=0}^{S} 2^{s(2D+4)}.$$

Let $q := 2^{2D+4}$. Then

$$\sum_{s=0}^{S} 2^{s(2D+4)} = \sum_{s=0}^{S} q^s = \frac{q^{S+1} - 1}{q - 1} \leq \frac{q}{q-1} q^S.$$

Therefore, with

$$c_{6,D} := \frac{2^{2D+4}}{2^{2D+4} - 1},$$

we have

$$\sum_{s=0}^{S} 2^{s(2D+4)} \leq c_{6,D} 2^{S(2D+4)}.$$

Using this particular bound, we obtain

$$\begin{aligned} m &\leq 2c_{4,D} \left(2 \log |Q_\varepsilon| + \log \frac{1}{\delta} + 1\right) \sum_{s=0}^{S} 2^{s(2D+4)} \\ &\leq 2c_{4,D} c_{6,D} 2^{S(2D+4)} \left(2 \log |Q_\varepsilon| + \log \frac{1}{\delta} + 1\right). \end{aligned}$$

Define

$$c_{7,D} := 2c_{4,D} c_{6,D}.$$

Then

$$m \leq c_{7,D} 2^{S(2D+4)} \left(2 \log |Q_\varepsilon| + \log \frac{1}{\delta} + 1\right).$$

Since by its definition S is the smallest integer satisfying

$$2^S \geq \frac{c_{0,D}}{\varepsilon},$$

we also have

$$2^{S-1} < \frac{c_{0,D}}{\varepsilon}.$$

Hence

$$2^S < \frac{2c_{0,D}}{\varepsilon}.$$

Raising both sides to the power $2D + 4$, we obtain

$$2^{S(2D+4)} \leq (2c_{0,D})^{2D+4} \varepsilon^{-(2D+4)}.$$

Defining

$$c_{8,D} := (2c_{0,D})^{2D+4},$$

we get

$$2^{S(2D+4)} \leq c_{8,D} \varepsilon^{-(2D+4)}.$$

Substituting

$$2^{S(2D+4)} \leq c_{8,D} \varepsilon^{-(2D+4)}$$

into the previous bound gives

$$m \leq c_{9,D} \varepsilon^{-(2D+4)} \left(2 \log |Q_\varepsilon| + \log \frac{1}{\delta} + 1\right),$$

where $c_{9,D} := c_{7,D}c_{8,D}$, which is defined in Table 2.

For convenience, we refer $C_D := c_{9,D}$ in Lemma 4.3, Sec. 4.

By Lemma 4.1,

$$\log|Q_\varepsilon| = \mathcal{O}\big(\varepsilon^{-D}\log(N+\varepsilon^{-D})\big)\,,$$

where the hidden constant depends only on D. Therefore,

$$\begin{aligned} m &\le C_D\varepsilon^{-(2D+4)}\left(2\log|Q_\varepsilon| + \log\frac{1}{\delta} + 1\right) \\ &= \mathcal{O}\left(C_D\left(\varepsilon^{-(3D+4)}\log(N+\varepsilon^{-D}) + \varepsilon^{-(2D+4)}\left(1+\log\frac{1}{\delta}\right)\right)\right), \end{aligned}$$

Thus, the total embedding dimension satisfies

$$m = \mathcal{O}\left(C_D\left(\varepsilon^{-(3D+4)}\log(N+\varepsilon^{-D}) + \varepsilon^{-(2D+4)}\left(1+\log\frac{1}{\delta}\right)\right)\right).$$

This is the desired dimension bound. Combining this dimension estimate with the uniform separation result from **Step 5** completes the proof. □

H. Proof of Lemma H.1

H.1. Low-frequency Fourier energy

Lemma H.1 (Low-frequency Fourier energy)**.** *Let ν, ν' be two distinct probability measures supported on a compact set $\mathcal{K} \subset \mathbb{R}^D$. Let*

$$r := W_1(\nu,\nu') > 0, \qquad \lambda := \nu - \nu'.$$

For $\omega \in \mathbb{R}^D$, define the Fourier transform of λ by

$$\widehat{\lambda}(\omega) := \int_{\mathcal{K}} e^{i\omega^\top x}\, d\lambda(x).$$

For $R > 0$, define

$$B(0,R) := \{\omega \in \mathbb{R}^D : \|\omega\|_2 \le R\}.$$

Here $\|\cdot\|_2$ is the Euclidean norm, and $|B(0,R)|$ denotes the D-dimensional Lebesgue volume of $B(0,R)$.

Then there exist constants $c_{0,D} > 0$ and $c_{1,D} > 0$, depending only on D and $\mathcal{K}$, such that

$$\frac{1}{|B(0,c_{0,D}/r)|}\int_{B(0,c_{0,D}/r)} |\widehat{\lambda}(\omega)|^2\, d\omega \ge c_{1,D}r^{D+2}.$$

Proof. By Lemma H.2, there exists a function $u : \mathbb{R}^D \to \mathbb{R}$ such that:

$$\left|\int_{\mathbb{R}^D} u\, d\lambda\right| \ge \frac{r}{4},$$

and its Fourier transform ($\mathbb{R}^D \to \mathbb{R}$)

$$\widehat{u}(\omega) := \int_{\mathbb{R}^D} u(x)e^{-i\omega^\top x}\, dx.$$

satisfies that u only contains frequencies inside $B(0, c_{0,D}/r)$, i.e.,

$$\widehat{u}(\omega) = 0 \quad \text{whenever } \|\omega\|_2 > \frac{c_{0,D}}{r},$$

and

$$\int_{\mathbb{R}^D} |\widehat{u}(\omega)|^2\, d\omega \le (2\pi)^D M_D^2,$$

where $C_{0,D}$ and M_D are constants defined in Lemma H.2. By Fourier inversion,

$$u(x) = \frac{1}{(2\pi)^D} \int_{\mathbb{R}^D} \widehat{u}(\omega) e^{i\omega^\top x}\, d\omega.$$

Since $\widehat{u}(\omega) = 0$ whenever $\|\omega\|_2 > c_{0,D}/r$, the integral over $\mathbb{R}^D$ is equal to the integral over the low-frequency ball. We have

$$u(x) = \frac{1}{(2\pi)^D} \int_{B(0,c_{0,D}/r)} \widehat{u}(\omega) e^{i\omega^\top x}\, d\omega.$$

Integrating both sides with respect to λ, and applying Fubini's theorem,[7] we obtain

$$\begin{aligned}
\int_{\mathbb{R}^D} u\, d\lambda &= \int \left[\frac{1}{(2\pi)^D} \int_{B(0,c_{0,D}/r)} \widehat{u}(\omega) e^{i\omega^\top x}\, d\omega \right] d\lambda(x) \\
&= \frac{1}{(2\pi)^D} \int_{B(0,c_{0,D}/r)} \widehat{u}(\omega) \left(\int_{\mathbb{R}^D} e^{i\omega^\top x}\, d\lambda(x) \right) d\omega \\
\text{(as } \lambda\text{'s support is } \mathcal{K}) &= \frac{1}{(2\pi)^D} \int_{B(0,c_{0,D}/r)} \widehat{u}(\omega) \left(\int_{\mathcal{K}} e^{i\omega^\top x}\, d\lambda(x) \right) d\omega \\
&= \frac{1}{(2\pi)^D} \int_{B(0,c_{0,D}/r)} \widehat{u}(\omega) \widehat{\lambda}(\omega)\, d\omega.
\end{aligned}$$

Applying Cauchy–Schwarz to the two functions $\widehat{u}$ and $\widehat{\lambda}$ on the frequency ball $B(0, c_{0,D}/r)$, we get [8]

$$\left| \int_{\mathbb{R}^D} u\, d\lambda \right|^2 \le \frac{1}{(2\pi)^{2D}} \left(\int_{B(0,c_{0,D}/r)} |\widehat{u}(\omega)|^2\, d\omega \right) \left(\int_{B(0,c_{0,D}/r)} |\widehat{\lambda}(\omega)|^2\, d\omega \right).$$

Using

$$\left| \int_{\mathbb{R}^D} u\, d\lambda \right| \ge \frac{r}{4},$$

we have

$$\left| \int_{\mathbb{R}^D} u\, d\lambda \right|^2 \ge \frac{r^2}{16}.$$

Therefore,

$$\frac{r^2}{16} \le \frac{1}{(2\pi)^{2D}} \left(\int_{B(0,c_{0,D}/r)} |\widehat{u}(\omega)|^2\, d\omega \right) \left(\int_{B(0,c_{0,D}/r)} |\widehat{\lambda}(\omega)|^2\, d\omega \right).$$

[7] We use Fubini's theorem in the following form: if

$$\iint |F(x,\omega)|\, d\omega\, d\mu(x) < \infty,$$

then the order of integration can be exchanged. Here $F(x,\omega) = \widehat{u}(\omega) e^{i\omega^\top x}$ on $B(0, c_{0,D}/r)$. Since $\lambda = \nu - \nu'$, it is enough to apply the theorem separately to the probability measures ν and ν'. For $\mu \in \{\nu, \nu'\}$,

$$\int_{\mathcal{K}} \int_{B(0,c_{0,D}/r)} |\widehat{u}(\omega) e^{i\omega^\top x}|\, d\omega\, d\mu(x) = \int_{B(0,c_{0,D}/r)} |\widehat{u}(\omega)|\, d\omega \le |B(0, c_{0,D}/r)|^{1/2} \left(\int_{\mathbb{R}^D} |\widehat{u}(\omega)|^2\, d\omega \right)^{1/2} < \infty,$$

where the last inequality uses $\int_{\mathbb{R}^D} |\widehat{u}(\omega)|^2\, d\omega \le (2\pi)^D M_D^2$. Thus Fubini's theorem applies.

[8] The Cauchy–Schwarz inequality for integrals states that, for square-integrable functions A and B on a domain Ω,

$$\left| \int_\Omega A(\omega) B(\omega)\, d\omega \right|^2 \le \left(\int_\Omega |A(\omega)|^2\, d\omega \right) \left(\int_\Omega |B(\omega)|^2\, d\omega \right).$$

Here we take $\Omega = B(0, c_{0,D}/r)$, $A(\omega) = \widehat{u}(\omega)$, and $B(\omega) = \widehat{\lambda}(\omega)$.

Using the upper bound from the condition that $\mu(\omega)$ satisfies

$$\int_{B(0,c_{0,D}/r)} |\widehat{u}(\omega)|^2 \, d\omega \le (2\pi)^D M_D^2,$$

we obtain

$$\frac{r^2}{16} \le \frac{M_D^2}{(2\pi)^D} \int_{B(0,c_{0,D}/r)} |\widehat{\lambda}(\omega)|^2 \, d\omega.$$

Rearranging terms gives

$$\int_{B(0,c_{0,D}/r)} |\widehat{\lambda}(\omega)|^2 \, d\omega \ge \frac{(2\pi)^D}{16 M_D^2} r^2.$$

Finally, since the volume of a D-dimensional Euclidean ball satisfies

$$|B(0,R)| = |B(0,1)| R^D,$$

we have

$$|B(0,c_{0,D}/r)| = |B(0,1)| \left(\frac{c_{0,D}}{r}\right)^D.$$

Therefore,

$$\begin{aligned}\frac{1}{|B(0,c_{0,D}/r)|} \int_{B(0,c_{0,D}/r)} |\widehat{\lambda}(\omega)|^2 \, d\omega &\ge \frac{(2\pi)^D}{16 M_D^2} \frac{r^2}{|B(0,1)|(c_{0,D}/r)^D} \\ &= \frac{(2\pi)^D}{16 M_D^2} \frac{r^{D+2}}{|B(0,1)| c_{0,D}^D}.\end{aligned}$$

Define

$$c_{1,D} := \frac{(2\pi)^D}{16 M_D^2} \frac{1}{|B(0,1)| c_{0,D}^D}.$$

Then

$$\frac{1}{|B(0,c_{0,D}/r)|} \int_{B(0,c_{0,D}/r)} |\widehat{\lambda}(\omega)|^2 \, d\omega \ge c_{1,D} r^{D+2}.$$

This proves the lemma. □

H.2. Low-frequency separating function

Lemma H.2 (Low-frequency separating function)**.** *Let ν, ν' be two distinct probability measures supported on a compact set $\mathcal{K} \subset \mathbb{R}^D$. Let*

$$r := W_1(\nu, \nu') > 0, \qquad \lambda := \nu - \nu'.$$

Then there exist constants $c_{0,D} > 0$ and $M_D > 0$, depending only on D and $\mathcal{K}$, and there exists a function $u : \mathbb{R}^D \to \mathbb{R}$ such that:

$$\left| \int_{\mathbb{R}^D} u \, d\lambda \right| \ge \frac{r}{4},$$

and its Fourier transform ($\mathbb{R}^D \to \mathbb{R}$)

$$\widehat{u}(\omega) := \int_{\mathbb{R}^D} u(x) e^{-i\omega^\top x} \, dx.$$

satisfies that u only contains frequencies inside $B(0, c_{0,D}/r)$, i.e.,

$$\widehat{u}(\omega) = 0 \quad \textit{whenever } \|\omega\|_2 > \frac{c_{0,D}}{r},$$

and

$$\int_{\mathbb{R}^D} |\widehat{u}(\omega)|^2 \, d\omega \le (2\pi)^D M_D^2.$$

Proof. **Step 1: Convert the Wasserstein separation into an integral separation.**

By the definition of W_1,

$$W_1(\nu,\nu') = \sup_{\|f\|_{\mathrm{Lip}}\le 1} \left| \int_{\mathcal{K}} f\, d\nu - \int_{\mathcal{K}} f\, d\nu' \right|.$$

Since $r = W_1(\nu,\nu') > 0$ and $r/2 < r$, there must exist a 1-Lipschitz function $f^* : \mathcal{K} \to \mathbb{R}$ such that

$$\left| \int_{\mathcal{K}} f^*\, d\nu - \int_{\mathcal{K}} f^*\, d\nu' \right| \ge \frac{r}{2}.$$

Otherwise, $W_1(\nu,\nu')$ cannot take value greater than $r/2$. Since $\lambda = \nu - \nu'$, this is equivalently

$$\left| \int_{\mathcal{K}} f^*\, d\lambda \right| \ge \frac{r}{2}.$$

We now subtract a constant from f^* to make it bounded on $\mathcal{K}$. Choose any point $x_0 \in \mathcal{K}$, and define

$$\widetilde{f}(x) := f^*(x) - f^*(x_0), \qquad x \in \mathcal{K}.$$

Then $\widetilde{f}(x_0) = 0$ and $\widetilde{f}$ is still 1-Lipschitz. Moreover, since

$$\int_{\mathcal{K}} f^*(x_0)\, d\lambda = f^*(x_0) \left(\int_{\mathcal{K}} 1\, d\nu - \int_{\mathcal{K}} 1\, d\nu' \right) = 0,$$

subtracting the constant does not change the integral against λ:

$$\int_{\mathcal{K}} \widetilde{f}\, d\lambda = \int_{\mathcal{K}} (f^*(\cdot) - f^*(x_0)) d\lambda = \int_{\mathcal{K}} f^*\, d\lambda.$$

Hence

$$\left| \int_{\mathcal{K}} \widetilde{f}\, d\lambda \right| = \left| \int_{\mathcal{K}} f^*\, d\lambda \right| \ge \frac{r}{2}.$$

Since $\widetilde{f}(x_0) = 0$ and $\widetilde{f}$ is 1-Lipschitz, for every $x \in \mathcal{K}$,

$$|\widetilde{f}(x)| = |\widetilde{f}(x) - \widetilde{f}(x_0)| \le \|x - x_0\|_2 \le \mathrm{diam}(\mathcal{K}).$$

Thus, in this **Step 1**, for any valid r that is the W_1 distance of two distinct measures in $\mathcal{K}$, we have identified a 1-Lipschitz function $\widetilde{f} : \mathcal{K} \to \mathbb{R}$ such that

$$\left| \int \widetilde{f}\, d\lambda \right| \ge \frac{r}{2},$$

and

$$|\widetilde{f}(x)| \le \mathrm{diam}(\mathcal{K}), \qquad x \in \mathcal{K}.$$

Step 2: Extend the function and make it compactly supported.

The function $\widetilde{f}$ is only defined on $\mathcal{K}$. We extend it to all of $\mathbb{R}^D$ by the McShane extension (McShane, 1934):

$$\overline{f}(x) := \inf_{y\in\mathcal{K}} \left\{ \widetilde{f}(y) + \|x - y\|_2 \right\}, \qquad x \in \mathbb{R}^D.$$

Since

$$\overline{f}(x) = \widetilde{f}(x), \qquad x \in \mathcal{K}, \qquad \overline{f}(x) = \|x - y\|_2, \qquad x \in \mathbb{R}^D \setminus \mathcal{K},$$

by the analysis of $\widetilde{f}$ and the triangle inequality of ℓ_2 norm, we have that $\overline{f}$ is 1-Lipschitz, i.e.,

$$\|\overline{f}\|_{\mathrm{Lip}} \le 1 \quad \text{on } \mathbb{R}^D.$$

Next, choose a fixed smooth cutoff function $q : \mathbb{R}^D \to [0, 1]$, where smooth means infinitely differentiable, such that

$$q(x) = 1 \quad \text{for } x \in \mathcal{K},$$

and

$$q(x) = 0 \quad \text{whenever } \operatorname{dist}(x, \mathcal{K}) \geq 1.$$

On the transition region

$$0 < \operatorname{dist}(x, \mathcal{K}) < 1,$$

we only require q to vary smoothly between 1 and 0, while remaining in the interval $[0, 1]$.

Such a cutoff exists by the standard smooth cutoff lemma: if a compact set is contained in an open set, then there exists a smooth function that equals 1 on the compact set and vanishes outside the open set. Here we apply this result with the compact set $\mathcal{K}$ and the open neighborhood

$$\{x \in \mathbb{R}^D : \operatorname{dist}(x, \mathcal{K}) < 1\}.$$

Moreover, since q is smooth and supported inside the bounded set

$$\mathcal{K}_1 := \{x \in \mathbb{R}^D : \operatorname{dist}(x, \mathcal{K}) \leq 1\},$$

its Lipschitz constant is finite. We denote

$$L_q := \|q\|_{\mathrm{Lip}} < \infty.$$

Define

$$g(x) := q(x)\overline{f}(x), x \in \mathbb{R}^D.$$

Because $q = 1$ on $\mathcal{K}$, and because λ is supported on $\mathcal{K}$, we have

$$\int_{\mathbb{R}^D} g \, d\lambda = \int_{\mathbb{R}^D} \widetilde{f} \, d\lambda.$$

Therefore, by the property of $\widetilde{f}$, we have

$$\left| \int_{\mathbb{R}^D} g \, d\lambda \right| = \left| \int_{\mathbb{R}^D} \widetilde{f} \, d\lambda \right| \geq \frac{r}{2}.$$

We now record two bounds on g. Define

$$\mathcal{K}_1 := \{x \in \mathbb{R}^D : \operatorname{dist}(x, \mathcal{K}) \leq 1\}.$$

Recall that the cutoff function q satisfies

$$q(x) = 0 \quad \text{whenever } \operatorname{dist}(x, \mathcal{K}) \geq 1.$$

Since $g(x) = q(x)\overline{f}(x)$, we have $g(x) = 0$ outside $\mathcal{K}_1$. Thus, g is supported on $\mathcal{K}_1$, meaning that g vanishes outside $\mathcal{K}_1$.

Since $\mathcal{K}$ is compact, $\mathcal{K}_1$ is bounded and has finite Lebesgue volume. For any $x \in \mathcal{K}_1$, by the definition of $\mathcal{K}_1$, there exists $y \in \mathcal{K}$ such that

$$\|x - y\|_2 \leq 1.$$

Since $\overline{f}$ is 1-Lipschitz over $\mathbb{R}^D$, we have

$$|\overline{f}(x)| \leq |\overline{f}(y)| + \|x - y\|_2.$$

Because $y \in \mathcal{K}$ and $\overline{f} = \widetilde{f}$ on $\mathcal{K}$, by the property of $\widetilde{f}$ characterized in **Step 1**, we have

$$|\overline{f}(y)| = |\widetilde{f}(y)| \leq \operatorname{diam}(\mathcal{K}).$$

Therefore,

$$|\overline{f}(x)| \leq \operatorname{diam}(\mathcal{K}) + 1.$$

By the definition of q, we have $0 \le q \le 1$. This gives

$$|g(x)| = |q(x)\overline{f}(x)| \le \operatorname{diam}(\mathcal{K}) + 1, \qquad x \in \mathcal{K}_1.$$

Recall that the $L^2(\mathbb{R}^D)$-norm is defined by

$$\|g\|^2_{L^2(\mathbb{R}^D)} := \int_{\mathbb{R}^D} |g(x)|^2 \, dx.$$

Since g is supported on $\mathcal{K}_1$, and $|g(x)| \le \operatorname{diam}(\mathcal{K}) + 1$ on $\mathcal{K}_1$, we have

$$\begin{aligned}\|g\|^2_{L^2(\mathbb{R}^D)} &= \int_{\mathcal{K}_1} |g(x)|^2 \, dx \\ &\le \big(\operatorname{diam}(\mathcal{K}) + 1\big)^2 |\mathcal{K}_1|.\end{aligned}$$

Taking square roots gives

$$\|g\|_{L^2(\mathbb{R}^D)} \le \big(\operatorname{diam}(\mathcal{K}) + 1\big)|\mathcal{K}_1|^{1/2}.$$

Define

$$M_D := \big(\operatorname{diam}(\mathcal{K}) + 1\big)|\mathcal{K}_1|^{1/2}.$$

Then

$$\|g\|_{L^2(\mathbb{R}^D)} \le M_D.$$

Since $\mathcal{K} = [0,1]^D$, we have

$$\operatorname{diam}(\mathcal{K}) = \sqrt{D}.$$

Moreover,

$$\mathcal{K}_1 = \{x \in \mathbb{R}^D : \operatorname{dist}(x, \mathcal{K}) \le 1\} \subseteq [-1, 2]^D,$$

and hence

$$|\mathcal{K}_1| \le 3^D.$$

Therefore,

$$M_D \le (\sqrt{D} + 1)3^{D/2}.$$

We also record a Lipschitz bound for g. Since $g(x) := q(x)\overline{f}(x), x \in \mathbb{R}^D$, $0 \le q(x) \le 1$, $\overline{f}$ is 1-Lipschitz, and $|\overline{f}| \le \operatorname{diam}(\mathcal{K}) + 1$, for any $x, x' \in \mathbb{R}^D$,

$$\begin{aligned}|g(x) - g(x')| &= |q(x)\overline{f}(x) - q(x)\overline{f}(x') + q(x)\overline{f}(x') - q(x')\overline{f}(x')| \\ &\le |q(x)|\,|\overline{f}(x) - \overline{f}(x')| + |\overline{f}(x')|\,|q(x) - q(x')| \\ &\le \|x - x'\|_2 + \|q\|_{\mathrm{Lip}}\big(\operatorname{diam}(\mathcal{K}) + 1\big)\|x - x'\|_2.\end{aligned}$$

Thus, considering $\mathcal{K} = [0,1]^D$ and defining

$$L_D := 1 + \|q\|_{\mathrm{Lip}}\big(\operatorname{diam}(\mathcal{K}) + 1\big) = 1 + \|q\|_{\mathrm{Lip}}(\sqrt{D} + 1),$$

we have

$$\|g\|_{\mathrm{Lip}} \le L_D.$$

Thus, **Step 2** gives a compactly supported function $g : \mathbb{R}^D \to \mathbb{R}$ satisfying

$$\left|\int_{\mathbb{R}^D} g \, d\lambda\right| \ge \frac{r}{2}, \qquad \|g\|_{L^2(\mathbb{R}^D)} \le M_D, \qquad \|g\|_{\mathrm{Lip}} \le L_D.$$

Step 3: Regularize g to obtain a low-frequency separator.

At the end of Step 2, we have constructed a function g such that

$$\left|\int_{\mathbb{R}^D} g\, d\lambda\right| \geq \frac{r}{2}.$$

Thus, g separates the two measures. However, this separation may use all frequencies, including very high frequencies. Since our goal is to prove a lower bound on $\widehat{\lambda}$ inside the low-frequency ball $B(0, c_{0,D}/r)$, we need a separating function whose Fourier transform is supported inside this ball. We obtain such a function by convolving g with a smoothing kernel.

Choose the one-dimensional bump function

$$\beta(t) := \begin{cases} \exp\left(-\dfrac{1}{1-t^2}\right), & |t| < 1, \\ 0, & |t| \geq 1. \end{cases}$$

This is a standard smooth bump function on $\mathbb{R}$. Define

$$\theta(t) := \frac{\beta(t)}{\beta(0)}.$$

Then $\theta \in C_c^\infty(\mathbb{R})$, meaning that θ is infinitely differentiable and compactly supported. Moreover,

$$0 \leq \theta \leq 1, \qquad \theta(0) = 1, \qquad \theta(t) = 0 \quad \text{whenever } |t| \geq 1.$$

We define the D-dimensional frequency cutoff by

$$\widehat{h}(\omega) := \prod_{i=1}^{D} \theta(\sqrt{D}\,\omega_i), \qquad \omega = (\omega_1, \ldots, \omega_D) \in \mathbb{R}^D.$$

Then $\widehat{h} : \mathbb{R}^D \to [0,1]$ is smooth and satisfies

$$\widehat{h}(0) = 1.$$

Also, if $\widehat{h}(\omega) \neq 0$, then

$$|\omega_i| \leq \frac{1}{\sqrt{D}}, \qquad i = 1, \ldots, D.$$

Hence

$$\|\omega\|_2^2 = \sum_{i=1}^{D} \omega_i^2 \leq D \cdot \frac{1}{D} = 1.$$

Therefore,

$$\widehat{h}(\omega) = 0 \quad \text{whenever } \|\omega\|_2 > 1.$$

Let h be the inverse Fourier transform of $\widehat{h}$, namely

$$h(x) := \frac{1}{(2\pi)^D} \int_{\mathbb{R}^D} \widehat{h}(\omega) e^{i\omega^\top x}\, d\omega.$$

Since $\widehat{h}(0) = 1$, our Fourier convention gives

$$\int_{\mathbb{R}^D} h(x)\, dx = 1.$$

We now bound the first moment of h. Let η be the one-dimensional inverse Fourier transform of θ, namely

$$\eta(t) := \frac{1}{2\pi} \int_{\mathbb{R}} \theta(\xi) e^{i\xi t}\, d\xi.$$

Since $\theta \in C_c^\infty(\mathbb{R})$, its inverse Fourier transform η is a Schwartz function; in particular, η decays fast enough at infinity to be integrable and to have a finite first moment (Folland, 2009). Hence the constants

$$A_0 := \int_{\mathbb{R}} |\eta(t)|, dt, \qquad A_1 := \int_{\mathbb{R}} |t|, |\eta(t)|, dt$$

are finite.

By the product form of $\widehat{h}$ and the Fourier scaling rule, the inverse Fourier transform h satisfies

$$h(x) = D^{-D/2} \prod_{i=1}^{D} \eta(x_i/\sqrt{D}).$$

Therefore,

$$\begin{aligned} A_h &:= \int_{\mathbb{R}^D} |h(x)| \|x\|_2 \, dx \\ &= \int_{\mathbb{R}^D} D^{-D/2} \prod_{i=1}^{D} |\eta(x_i/\sqrt{D})| \|x\|_2 \, dx. \end{aligned}$$

Using the change of variables $x_i = \sqrt{D}\, y_i$, we have $dx = D^{D/2} dy$ and $\|x\|_2 = \sqrt{D}\|y\|_2$. Hence

$$A_h = \sqrt{D} \int_{\mathbb{R}^D} \prod_{i=1}^{D} |\eta(y_i)| \|y\|_2 \, dy.$$

Since

$$\|y\|_2 \le |y_1| + \cdots + |y_D|,$$

we obtain

$$\begin{aligned} A_h &\le \sqrt{D} \sum_{i=1}^{D} \int_{\mathbb{R}^D} |y_i| \prod_{j=1}^{D} |\eta(y_j)| \, dy \\ &= \sqrt{D} \sum_{i=1}^{D} A_1 A_0^{D-1} \\ &= D^{3/2} A_1 A_0^{D-1}. \end{aligned}$$

Thus,

$$A_h \le D^{3/2} A_1 A_0^{D-1} < \infty.$$

Define

$$c_{0,D} := \max\{1, 8L_D A_h\}, \qquad \sigma := \frac{r}{c_{0,D}}.$$

Now define a rescaled smoothing function $\psi : \mathbb{R}^D \to [0, \sigma^{-D}]$:

$$\psi(x) := \sigma^{-D} h(x/\sigma),$$

and set

$$u := g * \psi = \int_{\mathbb{R}^D} \psi(y) g(x - y) \, dy.$$

We prove below that u has the three desired properties.

Step 3(a): u only uses frequencies up to $B(0, c_{0,D}/r)$.

Recall that both ψ and g are defined upon $\mathbb{R}^D$, and

$$\psi(x) := \sigma^{-D} h(x/\sigma), \qquad u := g * \psi.$$

Since convolution becomes multiplication after taking the Fourier transform,

$$\widehat{u}(\omega) = \widehat{g}(\omega)\widehat{\psi}(\omega).$$

Because ψ is the rescaled version of h, the Fourier scaling rule gives

$$\widehat{\psi}(\omega) = \widehat{h}(\sigma\omega).$$

Therefore,

$$\widehat{u}(\omega) = \widehat{g}(\omega)\widehat{h}(\sigma\omega).$$

By construction, the frequency cutoff $\widehat{h}$ satisfies

$$\widehat{h}(\zeta) = 0 \quad \text{whenever } \|\zeta\|_2 > 1.$$

Taking $\zeta = \sigma\omega$, we get

$$\widehat{h}(\sigma\omega) = 0 \quad \text{whenever } \|\sigma\omega\|_2 > 1.$$

Thus,

$$\widehat{u}(\omega) = 0 \quad \text{whenever } \|\sigma\omega\|_2 > 1.$$

Equivalently,

$$\widehat{u}(\omega) = 0 \quad \text{whenever } \|\omega\|_2 > \frac{1}{\sigma}.$$

Since $\sigma = r/c_{0,D}$, we have

$$\frac{1}{\sigma} = \frac{c_{0,D}}{r}.$$

Hence

$$\widehat{u}(\omega) = 0 \quad \text{whenever } \|\omega\|_2 > \frac{c_{0,D}}{r}.$$

Step 3(b): u still separates the two measures.

We now show that replacing g by its smoothed version u does not destroy the separation. Recall that

$$u = g * \psi, \qquad \psi(x) = \sigma^{-D} h(x/\sigma).$$

Since $\widehat{h}(0) = 1$ and consequently

$$\int_{\mathbb{R}^D} h(x)\,dx = 1,$$

a change of variables $y = \sigma z$ gives

$$\int_{\mathbb{R}^D} \psi(y)\,dy = \int_{\mathbb{R}^D} \sigma^{-D} h(y/\sigma)\,dy = \int_{\mathbb{R}^D} h(z)\,dz = 1.$$

By definition of $u(\cdot)$ and the above observation,

$$u(x) = \int_{\mathbb{R}^D} \psi(y) g(x-y)\,dy, \qquad g(x) = g(x)\int_{\mathbb{R}^D} \psi(y)\,dy.$$

, we subtract the two identities to obtain

$$u(x) - g(x) = \int \psi(y)\big(g(x-y) - g(x)\big)\,dy.$$

Taking absolute values and using the triangle inequality,

$$|u(x) - g(x)| \le \int |\psi(y)|\,|g(x-y) - g(x)|\,dy.$$

Since g is L_D-Lipschitz over $\mathbb{R}^D$,

$$|g(x-y)-g(x)| \le L_D\|y\|_2.$$

Hence, for every $x \in \mathbb{R}^D$,

$$|u(x)-g(x)| \le L_D \int |\psi(y)|\|y\|_2\, dy.$$

We now compute the last integral. Since

$$\psi(y) = \sigma^{-D} h(y/\sigma),$$

and using the change of variables $y = \sigma z$, so that $dy = \sigma^D dz$ and $\|y\|_2 = \sigma\|z\|_2$, we obtain

$$\begin{aligned}\int |\psi(y)|\|y\|_2\, dy &= \int \sigma^{-D}|h(y/\sigma)|\|y\|_2\, dy \\ &= \int \sigma^{-D}|h(z)|\sigma\|z\|_2\sigma^D\, dz \\ &= \sigma \int |h(z)|\|z\|_2\, dz \\ &= \sigma A_h.\end{aligned}$$

Therefore, for every $x \in \mathbb{R}^D$,

$$|u(x)-g(x)| \le L_D A_h \sigma.$$

Recall that

$$\|u-g\|_\infty := \sup_{x\in\mathbb{R}^D} |u(x)-g(x)|.$$

Thus,

$$\|u-g\|_\infty \le L_D A_h \sigma.$$

Since

$$\sigma = \frac{r}{c_{0,D}},$$

we have

$$L_D A_h \sigma = \frac{L_D A_h}{c_{0,D}} r.$$

By the definition

$$c_{0,D} := \max\{1, 8L_D A_h\},$$

we have $c_{0,D} \ge 8L_D A_h$, and therefore

$$\frac{L_D A_h}{c_{0,D}} \le \frac{1}{8}.$$

Hence

$$\|u-g\|_\infty \le \frac{r}{8}.$$

It remains to transfer this pointwise closeness into separation of the two measures. Since $\lambda = \nu - \nu'$,

$$\int_{\mathbb{R}^D} (u-g)\, d\lambda = \int_{\mathbb{R}^D} (u-g)\, d\nu - \int_{\mathbb{R}^D} (u-g)\, d\nu'.$$

Therefore,

$$\begin{aligned}\left|\int_{\mathbb{R}^D} (u-g)\, d\lambda\right| &\le \left|\int_{\mathbb{R}^D} (u-g)\, d\nu\right| + \left|\int_{\mathbb{R}^D} (u-g)\, d\nu'\right| \\ &\le \|u-g\|_\infty + \|u-g\|_\infty \\ &= 2\|u-g\|_\infty \\ &\le \frac{r}{4}.\end{aligned}$$

Using the separation from Step 2,

$$\left|\int_{\mathbb{R}^D} g\,d\lambda\right| \geq \frac{r}{2},$$

we finally obtain

$$\begin{aligned}\left|\int_{\mathbb{R}^D} u\,d\lambda\right| &\geq \left|\int_{\mathbb{R}^D} g\,d\lambda\right| - \left|\int_{\mathbb{R}^D} (u-g)\,d\lambda\right| \\ &\geq \frac{r}{2} - \frac{r}{4} = \frac{r}{4}.\end{aligned}$$

Thus, u still separates the two measures.

Step 3(c): u has controlled Fourier L^2-norm.

We finally show that the Fourier transform of u has a controlled L^2-norm. Recall that

$$u = g * \psi, \qquad \psi(x) = \sigma^{-D} h(x/\sigma).$$

By the convolution rule for Fourier transforms,

$$\widehat{u}(\omega) = \widehat{g}(\omega)\widehat{\psi}(\omega).$$

By the scaling rule for the rescaled kernel ψ,

$$\widehat{\psi}(\omega) = \widehat{h}(\sigma\omega).$$

Hence

$$\widehat{u}(\omega) = \widehat{g}(\omega)\widehat{h}(\sigma\omega).$$

Since $0 \leq \widehat{h}(\omega) \leq 1$ for every $\omega \in \mathbb{R}^D$, we also have

$$|\widehat{h}(\sigma\omega)| \leq 1, \qquad \omega \in \mathbb{R}^D.$$

Therefore,

$$|\widehat{u}(\omega)| = |\widehat{g}(\omega)|\,|\widehat{h}(\sigma\omega)| \leq |\widehat{g}(\omega)|.$$

Squaring both sides and integrating over $\mathbb{R}^D$ gives

$$\int_{\mathbb{R}^D} |\widehat{u}(\omega)|^2\,d\omega \leq \int_{\mathbb{R}^D} |\widehat{g}(\omega)|^2\,d\omega.$$

By Plancherel's theorem under our Fourier convention (Folland, 2009),

$$\int_{\mathbb{R}^D} |\widehat{g}(\omega)|^2\,d\omega = (2\pi)^D \int_{\mathbb{R}^D} |g(x)|^2\,dx = (2\pi)^D \|g\|^2_{L^2(\mathbb{R}^D)}.$$

By the L^2-bound from **Step 2**,

$$\|g\|_{L^2(\mathbb{R}^D)} \leq M_D.$$

Combining the previous three displays, we obtain

$$\int_{\mathbb{R}^D} |\widehat{u}(\omega)|^2\,d\omega \leq (2\pi)^D M_D^2.$$

Combining Steps 3(a), 3(b), and 3(c), we have constructed a function u satisfying

$$\widehat{u}(\omega) = 0 \quad \text{whenever } \|\omega\|_2 > \frac{c_{0,D}}{r},$$

$$\left|\int u\,d\lambda\right| \geq \frac{r}{4},$$

and

$$\int_{\mathbb{R}^D} |\widehat{u}(\omega)|^2\,d\omega \leq (2\pi)^D M_D^2.$$

This proves the lemma. □

I. Empirical Validation on the MNIST Digit-Sum Task

We validate our framework on the MNIST Digit-Sum task: given an unordered set of $N = 10$ digit images, the goal is to predict the scalar sum of their labels. Our pipeline proceeds in two stages. In the first stage, a CNN Autoencoder is trained to compress each 28×28 digit image into a compact $D = 16$ dimensional latent vector. Once trained, the encoder is frozen and used purely as a fixed feature extractor, reducing the effective input dimension from the ambient pixel space to a low-dimensional latent representation. This directly operationalizes the manifold relief argument of Section 4.1, where the intrinsic dimension d of the data, rather than the ambient dimension D, governs the required latent capacity. In the second stage, both a DeepSets model and a parameter-matched LSTM baseline are trained on sets of $N = 10$ latent vectors to predict the digit-sum label. We sweep the latent dimension across $m \in \{4, 8, 16, 32, 64\}$ and measure performance via exact integer matching of the rounded scalar prediction against the ground-truth sum.

Results are shown in Figure 4. DeepSets achieves strong accuracy even at small latent dimensions, with performance plateauing well before the linear barrier $m = N$ predicted by classical bounds. This saturation is consistent with our theoretical analysis: once m is sufficient to resolve the metric entropy of the underlying latent distribution, additional width yields diminishing returns. The parameter-matched LSTM, by contrast, fails to reliably approximate the digit-sum across all tested configurations.

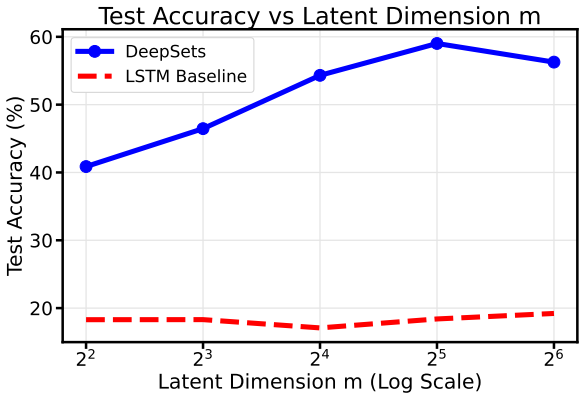

Figure 4. Empirical evaluation on the MNIST Image-Digit Sum task across varying latent dimensions m. Images are first mapped into a 16-dimensional latent representation using a frozen Autoencoder, which serves as a low-dimensional proxy for the underlying manifold structure. We compare standard DeepSets against a parameter-matched LSTM sequence baseline. The results show that DeepSets maintains strong test accuracy even under substantial latent compression, whereas the LSTM baseline fails to reliably approximate the sum on unseen sets. This empirical result is consistent with our theoretical analysis that permutation-invariant functions can be learned effectively by DeepSets with a low latent dimension.

J. Limitations and Future Work

Limitations. Our analysis introduces two primary theoretical constraints. First, the constructive bound has a worst-case polynomial dependence on the accuracy parameter whose exponent grows with the ambient dimension D. While practical data often resides on low-dimensional intrinsic manifolds ($d \ll D$), mathematically tightening this ambient dependency remains open. Second, our guarantee relies on the global Wasserstein Lipschitz condition, restricting scope to stable functions. Functions lacking this geometric stability fall outside our current bounds.

Future Work. We identify the following directions for future research.

(i) Lower Bounds. Determining whether the $\mathcal{O}(\varepsilon^{-(3D+4)} \log N)$ capacity is necessary, or whether further dimensionality reduction is possible, remains an important open question.

(ii) Extension to Attention Mechanisms. Investigating whether our logarithmic-in-N bound extends to the feature dimensions within Self-Attention layers could theoretically justify why Transformers perform well with compressed representations.

(iii) Uniformity Across Cardinalities. Developing a unified theory that guarantees bounded approximation error uniformly across all set sizes, rather than a fixed N, would represent a significant theoretical refinement.

(iv) Generalization and Learnability. It remains open whether the logarithmic-in-N width translates into improved generalization bounds or lower sample complexity.

(v) Beyond Wasserstein Stability. Relaxing the global Lipschitz assumption by exploiting local regularity or adopting weaker conditions such as Hölder continuity would broaden the scope of our framework.

(vi) Extension to Graph Neural Networks. Investigating whether the logarithmic sufficiency bound extends to Graph Neural Networks, where elements are connected by edges, would provide a unified view of scalability across set and graph learning.

(vii) Alternative Group Invariances. Quantifying how enforcing a smaller invariance group beyond the full permutation group affects the required latent dimension remains an open question.

(viii) Extension to Arbitrary Set Sizes. Leveraging continuous covering numbers and geometric regularity (Panaretos & Zemel, 2020) to establish size-independent capacity bounds would generalize our current N-dependent analysis to sets of arbitrary cardinality.