LEARNING WASSERSTEIN-STABLE PERMUTATION-INVARIANT FUNCTIONS: N-INDEPENDENT LATENT DIMENSION AND SAMPLE COMPLEXITY

Anonymous authors
Paper under double-blind review

ABSTRACT

Learning permutation-invariant functions over sets is fundamental to many applications, including point-cloud analysis, multi-instance learning, and data-driven stochastic optimization. DeepSets provide a standard architecture for this task by applying a shared encoder to individual elements and aggregating their representations. However, whether increasing the set size N necessarily increases the complexity of either representation or learning remains open. Existing latent-dimension bounds scale from $\mathcal{O}(N)$ for exact multiset representation to $\mathcal{O}(\log N)$ for Wasserstein-stable approximation, while the dependence of the required number of training samples on N remains open. In this paper, we show that both the latent dimension and the sample complexity can be made independent of N for Wasserstein-stable permutation-invariant functions. We first develop a Soft-Histogram Encoder and prove an N-independent latent-dimension bound for continuous DeepSets approximation. We then establish an N-independent sample-complexity bound for learning such functions from i.i.d. training sets. To improve data efficiency under limited labeling budgets arising from expensive target evaluations, we further develop permutation-aware maximin sampling (PAMS), which selects training sets to maximize their separation under optimal permutation matching. Numerical experiments on resource-allocation problems validate the predicted N-independent scaling of both the critical latent dimension and the number of training sets required to reach a fixed prediction accuracy, and show that PAMS achieves lower prediction error than random sampling under the same labeling budget. Together, these results establish that Wasserstein stability enables cardinality-independent representation and learning, while providing a practical route toward data-efficient permutation-invariant learning.

1 INTRODUCTION

Learning permutation-invariant functions over sets is fundamental to many applications, including point-cloud perception (Qi et al., 2017; Guo et al., 2019), particle-jet classification (Komiske et al., 2019), and data-driven stochastic optimization (Zhou et al., 2025; 2026). A widely used architecture for this purpose is DeepSets (Zaheer et al., 2017), which applies a shared encoder to individual set elements and aggregates their representations through a commutative pooling operation. DeepSets is known to be a universal approximator of permutation-invariant functions (Zaheer et al., 2017; Schmidt-Hieber, 2021; Tabaghi & Wang, 2024). However, these universality results do not characterize the latent dimension and sample complexity required for learning, which directly govern the computational efficiency of the model.

Existing quantitative analyses have primarily focused on the latent dimension required for representing permutation-invariant functions. For exact multiset representation, the latent dimension must grow at least linearly with the set size N (Wagstaff et al., 2022; Tabaghi & Wang, 2024; Dym & Gortler, 2025). More recent work relaxes exact representation to approximation and shows that, for Wasserstein-stable functions, the required latent dimension can be reduced to $\mathcal{O}(\log N)$ (Zhou et al., 2026). These results show that allowing approximation under suitable stability can substantially reduce the representation requirement compared with exact multiset encoding.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

Despite this progress, existing latent-dimension bounds still depend explicitly on N, and it remains unknown whether this dependence is fundamental under Wasserstein stability. Moreover, approximation guarantees alone do not characterize learnability from finite data. Even when a suitable DeepSets representation exists, it is unclear how many i.i.d. set-valued training samples are required to achieve a prescribed population risk, and whether this number must increase with N. These two issues lead to a common question: *how do representation complexity and sample complexity scale with the set cardinality N?*

In this paper, we prove that both the latent dimension and the sample complexity for learning Wasserstein-stable permutation-invariant functions can be bounded independently of the set size N. Specifically, we derive an N-independent latent-dimension guarantee for continuous DeepSets approximation and an N-independent sample-complexity bound under i.i.d. training. Our main contributions are as follows.

▷ **N-independent latent dimension.** In Sec. 4.1, we construct a Soft-Histogram Encoder to prove that every Wasserstein-stable permutation-invariant function with Lipschitz constant L_F admits a continuous DeepSets approximation with any prescribed uniform error tolerance $\tau > 0$ and latent dimension $m \leq \lceil 3\sqrt{D}L_F/\tau \rceil^D$, independently of the set size N. This establishes that the cardinality dependence in existing latent-dimension bounds is not intrinsic to Wasserstein-stable approximation. More broadly, this result theoretically justifies the common practice of using a fixed representation width across varying set sizes (Zaheer et al., 2017; Wang et al., 2016; Arandjelovic et al., 2016) for Wasserstein-stable functionals by showing that a sufficient width depends only on the input dimension, stability constant, and target accuracy, independently of the set cardinality N.

▷ **N-independent sample complexity.** In Sec. 4.2, we establish an N-independent sample-complexity guarantee for learning Wasserstein-stable permutation-invariant functions with DeepSets. For any prescribed population-risk tolerance $\varepsilon_r > 0$ and confidence level $1 - \zeta$, we show that a sufficient number of i.i.d. set-valued training samples can be bounded in terms of the encoder and decoder complexities, with no dependence on the set cardinality N. Consequently, increasing the number of elements in each input set need not increase the number of training sets required to achieve a fixed population risk.

▷ **Permutation-aware maximin sampling.** In Sec. 5, we consider training-data construction under a limited labeling budget, where each target evaluation may require solving an expensive optimization problem or running a numerical simulation. To better utilize the available labels, we adapt coverage-based maximin selection to permutation-invariant set inputs and refer to the resulting scheme as permutation-aware maximin sampling (PAMS), which selects training sets that are well separated under optimal permutation matching. This distance coincides with the Wasserstein distance between equally weighted empirical measures, aligning the sampling criterion with the Wasserstein-stable structure of the problem. Under the same labeling budget, PAMS achieves lower prediction error than random sampling. See Appendix A for related work on label-efficient sampling for permutation-invariant learning.

To our knowledge, this work provides the first cardinality-independent guarantees for both representation and learning of Wasserstein-stable permutation-invariant functions, while further improving data efficiency through permutation-aware sampling.

2 Related Work

Sum-decomposable architectures are widely used for learning permutation-invariant functions. Representative examples include DeepSets (Zaheer et al., 2017) and PointNet (Qi et al., 2017), which encode each element independently and aggregate the resulting representations through a commutative pooling operation. Their simple structure enables efficient computation and universal approximation guarantees (Zaheer et al., 2017; Schmidt-Hieber, 2021; Tabaghi & Wang, 2024). Beyond sum-decomposable models, alternative permutation-invariant architectures have been developed to explicitly capture interactions among elements, including Set Transformer (Lee et al., 2019), graph networks (Scarselli et al., 2008; Xu et al., 2018), Janossy pooling (Murphy et al., 2018), and higher-order tensor methods (Maron et al., 2019). Across these architectures, a central theoretical challenge is to characterize the latent dimension required for approximation and, in particular, how it scales with the set cardinality N.

Table 1: Comparison of latent-dimension and sample-complexity bounds for permutation-invariant learning. "Exact rep." denotes exact representation, while "Univ. approx." denotes universal approximation. "Exp. in N" and "Poly. in N" denote exponential and polynomial dependence on N, respectively, while "Indep. of N" denotes independence from the set cardinality N. The sample-complexity column reports only the dependence on the set cardinality N.

Study	Guarantee	Latent dim.	Samp. Comp.[c]	D	Cond.[d]
Zaheer et al. (2017)	Exact rep.	$\mathcal{O}(N^D)$	Exp. in N	≥ 1	–
Wagstaff et al. (2022)	Exact rep.	$\mathcal{O}(N)$[a]	Exp. in N	1	–
Dym & Gortler (2025)	Exact rep.	$2DN+1$	Exp. in N	≥ 1	–
Amir et al. (2023)	Exact rep.	$2N(D+1)+1$	Exp. in N	≥ 1	–
Tabaghi & Wang (2024)	Exact rep.	$2DN$	Exp. in N	≥ 1	Ident.
Qi et al. (2017)[b]	Univ. approx.	$\mathcal{O}_D(\varepsilon^{-D})$	Indep. of N	≥ 1	H. Cont.
Zhou et al. (2026)	Univ. approx.	$\mathcal{O}_{\varepsilon,D}(\log N)$	Poly. in N	≥ 1	W. Stable
This paper	Univ. approx.	$\mathcal{O}\Big((\sqrt{D}L_F/\tau)^D\Big)$	**Indep. of N**	≥ 1	**W. Stable**

[a] a matching lower bound of $m \geq N$ is also established in (Wagstaff et al., 2022).
[b] Under Hausdorff continuity, which is insensitive to element multiplicity.
[c] For prior work, we infer the sample-complexity scaling in N from latent-dimension bounds and standard VC-/pseudo-dimension bounds. The cited papers do not claim these sample-complexity results.
[d] "Ident." denotes identifiable multisets, "H. Cont." denotes Hausdorff continuity, and "W. Stable" denotes Wasserstein stability.

Existing work on latent dimension has largely focused on exact representation, with the goal of characterizing the minimum latent dimension required for sum-decomposable architectures to represent arbitrary multiset functions exactly. Zaheer et al. (2017) first established a sufficient bound of $m = \mathcal{O}(N^D)$. For scalar inputs, Wagstaff et al. (2022) sharpened this to $m = N$ and proved a matching lower bound. Subsequent work substantially improved the dependence on the input dimension. Dym & Gortler (2025) obtained $m = 2DN + 1$ almost surely using random projections, Amir et al. (2023) derived $m = 2N(D+1)+1$ using neural moments, and Tabaghi & Wang (2024) further reduced the bound to $m = 2DN$ while retaining a continuous decoder. These results establish that exact representation generally requires latent dimensions that scale at least linearly with the set cardinality N.

Beyond exact representation, recent work has investigated whether additional regularity of the target functional can reduce the latent dimension required for accurate approximation. Under Hausdorff continuity, Qi et al. (2017) established an N-independent bound of $m = \mathcal{O}_D(\varepsilon^{-D})$. However, the Hausdorff metric is insensitive to element multiplicity, limiting its applicability to mass-sensitive multiset functions. Wasserstein distance accounts for this information by measuring the transport of mass between empirical distributions. Under Wasserstein stability, Zhou et al. (2026) showed through a probabilistic construction that $m = \mathcal{O}_{\varepsilon,D}(\log N)$ is sufficient, reducing the dependence on the set cardinality from linear to logarithmic. Nevertheless, the latent dimension still grows with N, leaving open whether this dependence is fundamental to Wasserstein-stable approximation.

More importantly, an approximation guarantee alone does not characterize how efficiently the target function can be learned from finite data. In practice, only finitely many labeled sets are available, making the number of training samples required to achieve a prescribed population risk a central measure of learning complexity. For a generic continuous decoder on an m-dimensional latent space, worst-case approximation results imply that achieving a fixed accuracy may require a network whose size grows exponentially with m (Yarotsky, 2018). This dependence propagates to sample complexity through standard neural-network generalization bounds. Consequently, when $m = \Theta(N)$, as in exact-representation results, the resulting sample-complexity bounds can exhibit exponential dependence on N (Wagstaff et al., 2022; Dym & Gortler, 2025; Amir et al., 2023; Tabaghi & Wang, 2024). The $m = \mathcal{O}(\log N)$ bound of Zhou et al. (2026) reduces this dependence from exponential to polynomial in N, but does not eliminate the dependence on set cardinality.

In this work, we establish cardinality-independent guarantees for both representation and learning of Wasserstein-stable permutation-invariant functions. For a D-dimensional input domain and a target functional with Wasserstein-Lipschitz constant L_F, a latent dimension $m = \mathcal{O}((\sqrt{D}\,L_F/\tau)^D)$

suffices for uniform approximation accuracy τ, while the sample complexity required to achieve a prescribed population risk can be bounded independently of N. Together, these results show that increasing the set cardinality need not require either a larger latent representation or more set-valued training samples, removing the cardinality dependence in prior approximation results.

3 Problem Setting

3.1 Permutation-Invariant Functions

We consider functions $f : \mathcal{K}^N \to \mathbb{R}$ over finite sets $\mathbf{x} = (x_1, \ldots, x_N) \in \mathcal{K}^N$, where $\mathcal{K} = [0,1]^D$ is a compact domain, $D \geq 1$ is the element dimension, and $N \geq 1$ is the set cardinality.

Definition 3.1 (Permutation invariance)**.** *Let S_N denote the symmetric group of all permutations of $\{1, \ldots, N\}$. A function $f : \mathcal{K}^N \to \mathbb{R}$ is* permutation-invariant *if*

$$f(x_1, \ldots, x_N) = f(x_{\pi(1)}, \ldots, x_{\pi(N)})$$

for every $\pi \in S_N$ and every $(x_1, \ldots, x_N) \in \mathcal{K}^N$.

We consider the widely adopted DeepSets architecture (Zaheer et al., 2017) for learning permutation-invariant functions:

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

where $\phi : \mathcal{K} \to \mathbb{R}^m$ is a shared pointwise encoder, $\rho : \mathbb{R}^m \to \mathbb{R}$ is a decoder, and m is the *latent dimension*. Architecture (1) is permutation-invariant by construction. The latent dimension m governs both the expressive power and the computational cost of the architecture, and is the central quantity of interest in this paper.

3.2 Measure-Theoretic Representation

Permutation invariance implies that the target depends only on the input multiset. This motivates representing each input $\mathbf{x} \in \mathcal{K}^N$ as an empirical probability measure

$$\mu_{\mathbf{x}} \;=\; \frac{1}{N}\sum_{i=1}^{N}\delta_{x_i} \;\in\; \mathcal{P}(\mathcal{K}), \tag{2}$$

where $\mathcal{P}(\mathcal{K})$ denotes the space of Borel probability measures on $\mathcal{K}$ and δ_{x_i} is the Dirac mass at x_i. The following proposition confirms that this identification is lossless.

Proposition 3.2 (Measure-theoretic identification)**.** *A function $f : \mathcal{K}^N \to \mathbb{R}$ is permutation-invariant if and only if there exists a unique functional $F : \mathcal{P}_N(\mathcal{K}) \to \mathbb{R}$ such that $F(\mu_{\mathbf{x}}) = f(\mathbf{x})$ for all $\mathbf{x} \in \mathcal{K}^N$, where*

$$\mathcal{P}_N(\mathcal{K}) := \left\{\tfrac{1}{N}\sum_{i=1}^{N}\delta_{x_i} : x_i \in \mathcal{K}\right\}$$

denotes the space of N-point empirical measures on $\mathcal{K}$.

The proof is given in Appendix D.1. This measure-theoretic representation allows us to characterize the stability of F through perturbations of the underlying probability measure. We use the Wasserstein-1 distance to quantify such perturbations and impose Lipschitz continuity of F with respect to this metric.

Definition 3.3 (Wasserstein-1 distance)**.** *For $\mu, \nu \in \mathcal{P}(\mathcal{K})$, the Wasserstein-1 distance is*

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

where $\mathrm{Lip}_1(\mathcal{K})$ denotes the set of functions that are 1-Lipschitz with respect to the Euclidean norm on $\mathcal{K}$.

The quantity $W_1(\mu, \nu)$ measures the minimum cost of transporting mass from μ to ν, and is the natural metric for functionals that respond smoothly to perturbations of the input distribution (Villani, 2009). We impose the following regularity condition on F.

Assumption 1 (Wasserstein stability)**.** *The functional $F : \mathcal{P}(\mathcal{K}) \to \mathbb{R}$ is L_F-Lipschitz with respect to W_1 for some constant $L_F > 0$:*

$$|F(\mu) - F(\nu)| \leq L_F \cdot W_1(\mu, \nu) \quad \textit{for all } \mu, \nu \in \mathcal{P}(\mathcal{K}).$$

Wasserstein stability provides a natural regularity model for set-valued learning problems whose inputs represent empirical distributions, requiring small sample perturbations to induce only small changes in the target. It is satisfied by functionals such as expectations of Lipschitz losses, optimal-transport costs to a reference distribution (Villani, 2009), and conditional value at risk for Lipschitz losses (Rockafellar & Uryasev, 2000; Prashanth & Bhat, 2022). These examples arise in stochastic optimization, distribution comparison, and risk-aware decision making, showing that the considered function class extends beyond a purely theoretical setting.

3.3 Learning Problem

We consider learning the Wasserstein-stable functional F using the DeepSets architecture in (1). Two quantities are of particular interest: the latent dimension m required for approximation and the number of training samples M required for learning.

Latent dimension. Given an approximation accuracy $\tau > 0$, we seek an encoder $\phi : \mathcal{K} \to \mathbb{R}^m$ and a decoder $\rho : \mathbb{R}^m \to \mathbb{R}$ such that

$$\sup_{\mathbf{x} \in \mathcal{K}^N} \left| F(\mu_{\mathbf{x}}) - \rho\left(\frac{1}{N} \sum_{i=1}^{N} \phi(x_i) \right) \right| \leq \tau.$$

The first question is how large the latent dimension m needs to be to achieve the prescribed approximation accuracy, and in particular whether m must grow with the set cardinality N.

Sample complexity. We further consider learning F from M i.i.d. input–output samples

$$\left\{ \left(X^{(j)}, F(\mu_{X^{(j)}}) \right) \right\}_{j=1}^{M}, \qquad X^{(j)} \sim \mathcal{D}_N,$$

where $\mathcal{D}_N$ is an unknown distribution on $\mathcal{K}^N$. We consider the ReLU DeepSets hypothesis class

$$\mathcal{H}_N = \left\{ X \mapsto \rho_\theta\left(\frac{1}{N} \sum_{i=1}^{N} \phi_\theta(x_i) \right) \right\}.$$

For any $h \in \mathcal{H}_N$, we define the population and empirical risks as

$$\mathcal{L}_N(h) = \mathbb{E}_{X \sim \mathcal{D}_N} \left[\left(h(X) - F(\mu_X) \right)^2 \right], \qquad \widehat{\mathcal{L}}_M(h) = \frac{1}{M} \sum_{j=1}^{M} \left[h(X^{(j)}) - F(\mu_{X^{(j)}}) \right]^2.$$

Let $\widehat{h}_M$ denote an empirical risk minimizer over $\mathcal{H}_N$. Given a target risk $\varepsilon_r \in (0, 1)$ and confidence level $1 - \zeta$, the second question is how large M needs to be to guarantee $\mathcal{L}_N(\widehat{h}_M) \leq \varepsilon_r$ with probability at least $1 - \zeta$, and in particular whether M must grow with N.

4 Main Results

In this section, we establish cardinality-independent guarantees for both approximation and learning. We first show that every Wasserstein-stable functional admits a continuous DeepSets approximation whose latent dimension depends only on the element dimension, stability constant, and target accuracy, but not on the set cardinality N. We then show that the number of training sets required to achieve a prescribed population risk can also be bounded independently of N.

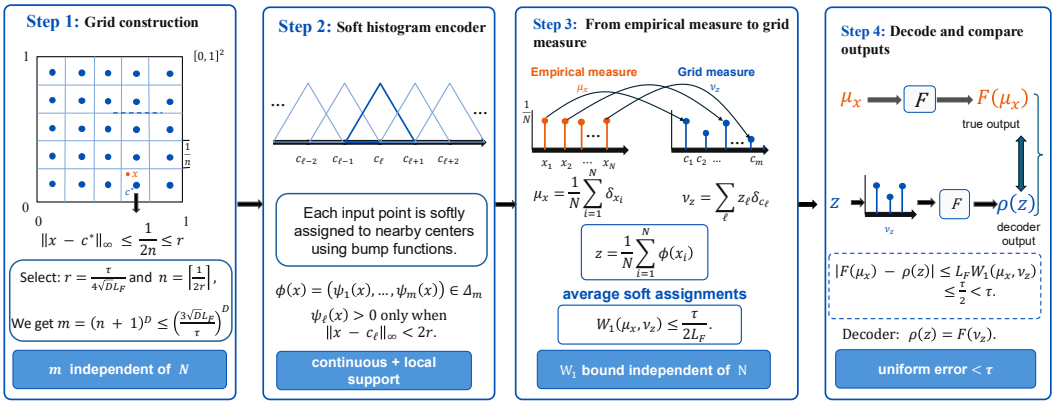

Figure 1: Illustration of the proof of Theorem 4.1.

4.1 N-Independent Latent Dimension

We begin with the approximation problem and characterize the latent dimension required by continuous DeepSets representations of Wasserstein-stable functionals.

Theorem 4.1. *Let $D \ge 1$, let $\mathcal{K} = [0,1]^D$, and let $F : \mathcal{P}(\mathcal{K}) \to \mathbb{R}$ satisfy Assumption 1 with constant $L_F > 0$. For every $\tau > 0$, there exist a latent dimension $m \le \lceil 3\sqrt{D}\, L_F/\tau \rceil^D$, a continuous encoder $\phi : \mathcal{K} \to \mathbb{R}^m$, and a continuous decoder $\rho : \mathbb{R}^m \to \mathbb{R}$ such that, simultaneously for every $N \ge 1$,*

$$\sup_{\mathbf{x} \in \mathcal{K}^N} \left| F(\mu_{\mathbf{x}}) - \rho\left(\frac{1}{N} \sum_{i=1}^N \phi(x_i) \right) \right| \le \tau,$$

The construction of Theorem 4.1 is illustrated in Figure 1. We discretize $\mathcal{K}$ with a finite grid and use a continuous soft-histogram encoder to assign inputs to nearby grid centers. Mean aggregation yields a grid-supported measure that remains uniformly close to the empirical measure in Wasserstein distance independently of N, and evaluating F on this measure gives the desired DeepSets approximation. The complete proof is provided in Appendix B.

Theorem 4.1 shows that the required representation width depends on D, L_F, and τ, but not on the set cardinality N. The same encoder–decoder pair works for all $N \ge 1$, yielding, to our knowledge, the first cardinality-independent latent-dimension guarantee for Wasserstein-stable permutation-invariant functions. This also provides a theoretical basis for fixed-width aggregation across varying set sizes (Zaheer et al., 2017; Wang et al., 2016; Arandjelovic et al., 2016).

We remark that the D-dependence in Theorem 4.1 arises from covering the ambient domain $[0,1]^D$ and therefore reflects worst-case geometric complexity. In many applications, high-dimensional data concentrate near low-dimensional structures with intrinsic dimension $d \ll D$ (Pope et al., 2021). In such cases, covering the data support rather than the ambient space can substantially reduce the effective dimension dependence.

4.2 N-Independent Sample Complexity

We next extend the cardinality-independent analysis from approximation to learning. Specifically, we show that a prescribed population risk can be achieved with a number of training sets that does not grow with the set cardinality N.

Theorem 4.2 (N-independent sample complexity)**.** *Fix a target risk $\varepsilon_r \in (0,1)$ and confidence level $1-\zeta$. Let $\mathcal{H}_N$ be the ReLU DeepSets class associated with Proposition C.1, with latent dimension m, encoder parameter count and depth (W_ϕ, L_ϕ), and decoder parameter count and depth (W_ρ, L_ρ). Here, W_ϕ counts all parameters of the vector-valued encoder $\phi : \mathcal{K} \to \mathbb{R}^m$. For every distribution $\mathcal{D}_N$ on $\mathcal{K}^N$, a sufficient number of training sets is*

$$M = \widetilde{\mathcal{O}}\left(\frac{W_\phi L_\phi + W_\rho L_\rho + \log(1/\zeta)}{\varepsilon_r^2} \right).$$

With this number of samples, any $\varepsilon_r/4$-approximate empirical risk minimizer $\widehat{h}_M \in \mathcal{H}_N$ satisfies $\mathcal{L}_N(\widehat{h}_M) \le \varepsilon_r$ with probability at least $1-\zeta$.

Theorem 4.2 shows that the cardinality independence established at the representation level extends to finite-sample learning. This follows from two properties of the DeepSets construction: the ReLU encoder and decoder can be chosen with latent dimension and architecture sizes independent of N, and mean aggregation does not amplify encoder approximation errors as N increases. Consequently, the resulting hypothesis class admits N-independent covering-number bounds, which, together with standard uniform-convergence arguments, yields Theorem 4.2. This guarantee is not dimension-free: achieving a fixed population-risk tolerance ϵ_r requires a sufficient latent dimension $m = O((\sqrt{D}L_F/\sqrt{\epsilon_r})^D)$, so the network sizes and resulting sample-complexity upper bound can inherit an exponential dependence on D. The complete ReLU realization and proof are provided in Appendix C.1 and Appendix C.2.

4.3 Discussion

We now interpret the broader significance of the preceding results by examining the mechanism behind cardinality independence, its connection to fixed-width set architectures used in practice, and the distinction between our construction and prior Wasserstein-stable approximation methods.

The role of Wasserstein stability. Permutation invariance alone does not yield an N-independent latent dimension. The key additional property is Wasserstein stability: since each element carries mass $1/N$ in the empirical measure, changing one element from x_i to x_i' changes the target by at most $\frac{L_F}{N}\|x_i - x_i'\|$. Hence, a fixed approximation accuracy τ only requires a Wasserstein resolution independent of N, allowing the same finite grid to be used across different set sizes. Increasing N changes only the masses assigned to the grid points, not the representation dimension. General permutation-invariant functions need not satisfy this $1/N$-scaled sensitivity, so existing cardinality-dependent lower bounds do not conflict with Theorem 4.1; see Appendix G.2.

Connection to fixed-size aggregation architectures. Fixed-width aggregation is widely used for sets of varying cardinalities, including point-cloud and other set-based architectures (Zaheer et al., 2017; Qi et al., 2017). Theorem 4.1 provides a theoretical justification for this practice in the Wasserstein-stable setting, showing that a width determined by D, L_F, and τ suffices independently of N. Our Soft-Histogram Encoder is also related to learnable histogram layers (Wang et al., 2016) and NetVLAD (Arandjelovic et al., 2016), which aggregate features relative to fixed bins or centers, while our analysis additionally provides an explicit bound on the required representation size. Further discussion is provided in Appendix G.1.

Relation to the prior random-feature construction. The previous $\mathcal{O}(\log N)$ result (Zhou et al., 2026) relies on random Fourier features sampled from a prescribed frequency distribution. When the target spectrum is concentrated in a limited frequency range, some sampled features may contribute little to the approximation and become redundant. The resource-allocation problem provides one example of this behavior, with most spectral power concentrated in low-frequency modes. This offers a possible explanation for why more RFF features may be required for such targets, rather than a general limitation of RFF methods. Detailed analysis is provided in Appendix G.3.

5 Permutation-Aware Maximin Sampling

The preceding analysis characterizes the number of i.i.d. training sets sufficient for learning. In practice, however, expensive target evaluations often impose a limited labeling budget. This motivates a complementary pool-based selection problem: how should a fixed number of training inputs be chosen most effectively? We address this by adapting maximin selection to permutation-invariant set inputs, using PAMS to select well-separated inputs under the permutation-aware distance d_{Q}.

Permutation-aware distance. For $\mathbf{x} = (x_1, \ldots, x_N)$ and $\mathbf{x}' = (x_1', \ldots, x_N')$ in $\mathcal{K}^N$, we define

$$d_{\mathrm{Q}}(\mathbf{x}, \mathbf{x}') := \min_{\pi \in S_N} \frac{1}{N} \sum_{i=1}^{N} \left\| x_i - x'_{\pi(i)} \right\|_2, \tag{3}$$

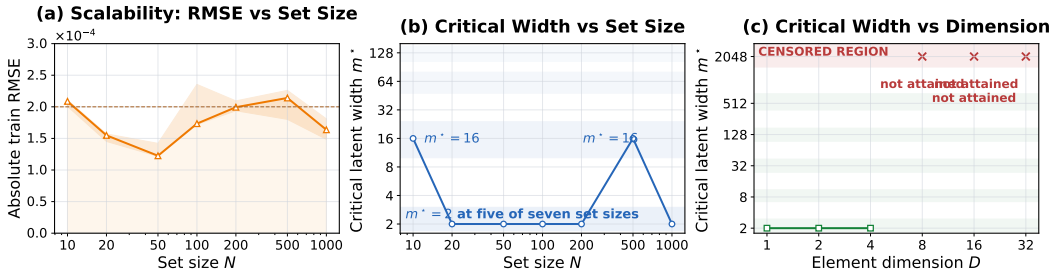

Figure 2: Latent-dimension scaling. (a) RMSE versus N at $D = 4$ and $m = 8$. (b) Critical width $m^\star$ versus N at $D = 4$. (c) $m^\star$ versus D at $N = 100$; "not attained" denotes failure to reach the target accuracy for $m \leq 2048$.

where S_N denotes the set of all permutations of N elements. The optimal matching removes the arbitrary ordering of the input elements and measures the separation between the underlying multisets. Importantly, this distance is directly aligned with the Wasserstein geometry used throughout our analysis. Recall that

$$\mu_{\mathbf{x}} = \frac{1}{N}\sum_{i=1}^{N} \delta_{x_i}, \qquad \mu_{\mathbf{x}'} = \frac{1}{N}\sum_{i=1}^{N} \delta_{x'_i}.$$

For two empirical measures with the same number of equally weighted atoms, the Wasserstein-1 distance reduces to the optimal-matching distance in (3). Hence,

$$W_1(\mu_{\mathbf{x}}, \mu_{\mathbf{x}'}) = d_{\mathrm{Q}}(\mathbf{x}, \mathbf{x}'), \qquad |F(\mu_{\mathbf{x}}) - F(\mu_{\mathbf{x}'})| \leq L_F d_{\mathrm{Q}}(\mathbf{x}, \mathbf{x}'), \tag{4}$$

where the second relation follows from Assumption 1. Thus, nearby inputs under d_{Q} have correspondingly close target values for the Wasserstein-stable function class considered in this work.

Maximin selection. PAMS selects the candidate least covered by the current training set. Let $\mathcal{U} = \{\mathbf{x}^{(1)}, \ldots, \mathbf{x}^{(K)}\} \subset \mathcal{K}^N$ denote an unlabeled candidate pool, and let $\mathcal{S}_j \subset \mathcal{U}$ denote the j inputs selected so far. Given $\mathcal{S}_j$, the next training input is selected as

$$\mathbf{x}^{(j+1)} \in \arg\max_{\mathbf{x} \in \mathcal{U} \setminus \mathcal{S}_j} \min_{\mathbf{x}' \in \mathcal{S}_j} d_{\mathrm{Q}}(\mathbf{x}, \mathbf{x}'). \tag{5}$$

The inner minimum measures each candidate's distance to the current training set, while the outer maximum selects the least-covered one. Repeating this process improves coverage under the Wasserstein-consistent metric. Appendix F shows that the resulting coverage radius directly enters a prediction-error bound and provides the complete procedure in Algorithm 1.

6 Numerical Experiments

We carry out numerical experiments on a resource-allocation problem to evaluate the latent-dimension and sample-complexity results, as well as the sampling efficiency of PAMS. Specifically, we investigate (i) how the latent dimension required to achieve a prescribed prediction accuracy varies with the set size N and element dimension D, (ii) how the number of training sets required to achieve a prescribed prediction accuracy varies with N, and (iii) whether PAMS improves prediction accuracy over random sampling under the same labeling budget. The resource-allocation formulation and the simulation setup are described in Appendix E.

6.1 Latent-Dimension Scaling

We investigate the impact of the set size N and element dimension D on the latent dimension required by DeepSets. We define the critical latent width $m^\star$ as the smallest tested width whose median training RMSE over three random seeds is below 2×10^{-4}. As illustrated in Fig. 2(a)–(b), for fixed $D = 4$, both the prediction error and $m^\star$ remain nearly unchanged as N increases from 10 to 1000, showing no systematic dependence on the set size. In contrast, Fig. 2(c) shows that $m^\star$

Table 2: Prediction RMSE ($\times 10^{-3}$, mean $\pm$ standard deviation) for random sampling and PAMS.

Setting		$M = 32$		$M = 64$		$M = 128$	
		Random	PAMS	Random	PAMS	Random	PAMS
Set-size scaling ($D = 1$)	$N = 2$	9.9 ± 6.8	2.9 ± 0.6	4.6 ± 3.0	2.0 ± 0.6	3.4 ± 1.4	1.4 ± 0.3
	$N = 5$	10.7 ± 4.0	4.0 ± 0.7	5.3 ± 1.1	2.0 ± 0.0	2.7 ± 1.3	1.7 ± 0.3
	$N = 10$	7.2 ± 2.6	2.6 ± 1.6	4.9 ± 1.3	1.2 ± 0.2	3.2 ± 1.6	0.8 ± 0.1
Element-dimension scaling ($N = 10$)	$D = 2$	3.1 ± 0.5	3.1 ± 0.3	3.0 ± 0.8	2.1 ± 0.7	2.5 ± 1.0	1.3 ± 0.9
	$D = 4$	3.6 ± 0.7	3.2 ± 0.3	2.2 ± 0.2	2.2 ± 0.2	1.9 ± 0.1	1.8 ± 0.2
	$D = 8$	5.1 ± 0.5	4.0 ± 0.7	3.4 ± 0.4	2.9 ± 0.7	2.5 ± 0.3	2.2 ± 0.2

increases sharply with D: it equals 2 for $D = 1, 2, 4$, whereas the target accuracy is not reached for $D = 8, 16, 32$ within $m \leq 2048$. These observations corroborate Theorem 4.1: the required latent dimension need not grow with N, while it can increase substantially with the element dimension D.

6.2 Sample-Complexity Scaling

We examine how the number of training sets required to achieve a fixed prediction accuracy scales with the set size N. We define $M^\star(N)$ as the smallest tested training budget for which the test RMSE falls below $\varepsilon = 2 \times 10^{-4}$. As shown in Fig. 3, $M^\star(N)$ exhibits no systematic growth as N increases from 20 to 1000. This behavior is consistent with Theorem 4.2, indicating that the number of training sets required to achieve a fixed accuracy need not increase with the set cardinality.

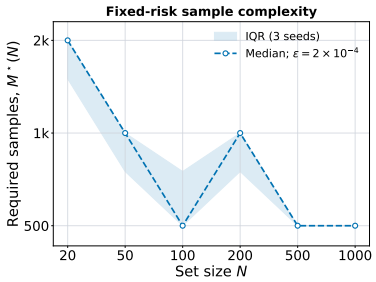

Figure 3: Sample-complexity scaling. Fixed-risk sample requirement $M^\star$ versus N at $\varepsilon = 2 \times 10^{-4}$; shading shows the interquartile range over three seeds.

6.3 Sampling Efficiency

We compare PAMS with random sampling under the same labeling budget using the same DeepSets predictor. As shown in Table 2, PAMS achieves lower prediction errors across all tested set sizes and in almost all element-dimension settings, with results reported over three random seeds. By selecting well-separated inputs under the optimal-matching distance, PAMS improves training-set coverage and consistently reduces prediction error. These results demonstrate the benefit of PAMS for data-efficient training under limited labeling budgets.

7 Conclusion

In this paper, we study the representation and learning complexity of Wasserstein-stable permutation-invariant functions with DeepSets. We show that both the latent dimension and the number of training sets required to achieve a fixed accuracy can be bounded independently of the set cardinality N. Our analysis is based on a Soft-Histogram Encoder that approximates empirical measures at a fixed Wasserstein resolution, together with an N-independent sample-complexity analysis for the resulting DeepSets class. We further develop permutation-aware maximin sampling (PAMS) to improve data efficiency under a fixed labeling budget. Numerical experiments on resource-allocation problems corroborate the predicted scaling behavior and demonstrate the benefit of PAMS over random sampling.

Several limitations remain. (i) Improving the dependence of the latent-dimension bound on D remains open. (ii) Extending cardinality-independent guarantees beyond uniformly Wasserstein-stable functionals is an important direction. (iii) PAMS targets geometric coverage but does not account for non-uniform data distributions, motivating distribution-aware extensions.

AI USE DISCLOSURE

We used generative AI tools to polish the language and improve the readability of this paper. The authors take full responsibility for the final content.

REFERENCES

Charu C Aggarwal, Xiangnan Kong, Quanquan Gu, Jiawei Han, and S Yu Philip. Active learning: A survey. In *Data classification: algorithms and applications*, pp. 599–634. Chapman and Hall/CRC, 2019.

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

Relja Arandjelovic, Petr Gronat, Akihiko Torii, Tomas Pajdla, and Josef Sivic. Netvlad: Cnn architecture for weakly supervised place recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 5297–5307, 2016.

Jordan T Ash, Chicheng Zhang, Akshay Krishnamurthy, John Langford, and Alekh Agarwal. Deep batch active learning by diverse, uncertain gradient lower bounds. *arXiv preprint arXiv:1906.03671*, 2019.

Peter L Bartlett, Nick Harvey, Christopher Liaw, and Abbas Mehrabian. Nearly-tight vc-dimension and pseudodimension bounds for piecewise linear neural networks. *Journal of Machine Learning Research*, 20(63):1–17, 2019.

Qi Chen, Kai Liu, Ruilong Yao, and Hu Ding. Sublinear time algorithms for greedy selection in high dimensions. In *Uncertainty in Artificial Intelligence*, pp. 346–356. PMLR, 2022.

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

Yonatan Geifman and Ran El-Yaniv. Deep active learning over the long tail. *arXiv preprint arXiv:1711.00941*, 2017.

Teofilo F Gonzalez. Clustering to minimize the maximum intercluster distance. *Theoretical computer science*, 38:293–306, 1985.

Yulan Guo, Hanyun Wang, Qingyong Hu, Hao Liu, Li Liu, and Mohammed Bennamoun. Deep learning for 3d point clouds: A survey. *arXiv preprint arXiv:1912.12033*, 2019.

David Haussler. Decision theoretic generalizations of the pac model for neural net and other learning applications. *Information and computation*, 100(1):78–150, 1992.

Dorit S Hochbaum and David B Shmoys. A best possible heuristic for the k-center problem. *Mathematics of operations research*, 10(2):180–184, 1985.

Patrick Kidger and Terry Lyons. Universal approximation with deep narrow networks. In *Conference on learning theory*, pp. 2306–2327. PMLR, 2020.

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

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

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

Ryan L Murphy, Balasubramaniam Srinivasan, Vinayak Rao, and Bruno Ribeiro. Janossy pooling: Learning deep permutation-invariant functions for variable-size inputs. *arXiv preprint arXiv:1811.01900*, 2018.

Phillip Pope, Chen Zhu, Ahmed Abdelkader, Micah Goldblum, and Tom Goldstein. The intrinsic dimension of images and its impact on learning. *arXiv preprint arXiv:2104.08894*, 2021.

LA Prashanth and Sanjay P Bhat. A wasserstein distance approach for concentration of empirical risk estimates. *Journal of Machine Learning Research*, 23(238):1–61, 2022.

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

R Tyrrell Rockafellar and Stanislav Uryasev. Optimization of conditional value-at-risk. *Journal of risk*, 2:21–42, 2000.

Franco Scarselli, Marco Gori, Ah Chung Tsoi, Markus Hagenbuchner, and Gabriele Monfardini. The graph neural network model. *IEEE transactions on neural networks*, 20(1):61–80, 2008.

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

Ozan Sener and Silvio Savarese. Active learning for convolutional neural networks: A core-set approach. *arXiv preprint arXiv:1708.00489*, 2017.

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

Cédric Villani. *Optimal transport: old and new*, volume 338. Springer, 2009.

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

Zhe Wang, Hongsheng Li, Wanli Ouyang, and Xiaogang Wang. Learnable histogram: Statistical context features for deep neural networks. In *European conference on computer vision*, pp. 246–262. Springer, 2016.

Keyulu Xu, Weihua Hu, Jure Leskovec, and Stefanie Jegelka. How powerful are graph neural networks? *arXiv preprint arXiv:1810.00826*, 2018.

Shuo Yang, Zhe Cao, Sheng Guo, Ruiheng Zhang, Ping Luo, Shengping Zhang, and Liqiang Nie. Mind the boundary: Coreset selection via reconstructing the decision boundary. In *Forty-first International Conference on Machine Learning*, 2024.

Dmitry Yarotsky. Optimal approximation of continuous functions by very deep relu networks. In *Conference on learning theory*, pp. 639–649. PMLR, 2018.

Manzil Zaheer, Satwik Kottur, Siamak Ravanbakhsh, Barnabas Poczos, Russ R Salakhutdinov, and Alexander Smola. Deep sets. *Advances in neural information processing systems*, 30, 2017.

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

Min Zhou, Enming Liang, and Minghua Chen. $\mathcal{O}(\log N)$ Latent Dimension Suffices for Universal Approximation of Permutation-Invariant Functions. In *Forty-third International Conference on Machine Learning*, 2026.

APPENDIX

A Label-Efficient Sampling for Permutation-Invariant Learning **12**

B Proof of the N-Independent Latent-Dimension Result **13**

B.1 Grid Construction and Covering Bound 13

B.2 Soft-Histogram Encoder 15

B.3 Approximation by a Grid-Supported Measure 16

B.4 Decoder Construction and Continuity 18

B.5 Proof of Theorem 4.1 21

C Proof of the N-Independent Sample-Complexity Result **22**

C.1 ReLU Realization 22

C.2 Proof of Theorem 4.2 24

D Auxiliary Proofs **27**

D.1 Measure-Theoretic Identification 27

E Simulation Setup **27**

E.1 Resource-Allocation Problem 27

E.2 Data Preparation 28

E.3 Network and Training Settings 28

F Permutation-Aware Maximin Sampling Algorithm **28**

F.1 Coverage-Based Prediction Error 28

G Understanding Cardinality Independence **30**

G.1 Connection to Fixed-Size Aggregation Architectures 30

G.2 The Role of Wasserstein Stability 31

G.3 Frequency-Domain Analysis 31

A LABEL-EFFICIENT SAMPLING FOR PERMUTATION-INVARIANT LEARNING

Data selection under limited labeling budgets. When target evaluations are expensive, active learning and core-set selection have been widely used to reduce the number of labeled samples required for training. Active-learning methods select informative inputs according to criteria such as model uncertainty or prediction disagreement (Aggarwal et al., 2019; Ash et al., 2019), while core-set methods construct representative subsets that preserve the geometric coverage of a larger unlabeled pool (Sener & Savarese, 2017; Yang et al., 2024). Among geometry-based approaches, maximin and farthest-point selection iteratively choose samples that are far from the current selected set, closely relating the selection problem to the k-center objective (Gonzalez, 1985; Hochbaum & Shmoys, 1985). These approaches provide effective mechanisms for reducing redundant samples under a fixed labeling budget. However, they are typically formulated for vector-valued inputs under Euclidean or feature-space distances, which do not directly account for the permutation symmetry of set-valued inputs.

Coverage-based subset selection. A closely related class of methods selects training samples by explicitly improving the geometric coverage of the candidate pool. The k-center formulation seeks a subset that minimizes the maximum distance from any candidate to its nearest selected sample (Sener & Savarese, 2017), while farthest-point or maximin selection provides a simple sequential strategy by repeatedly choosing the least-covered candidate (Geifman & El-Yaniv, 2017; Chen et al., 2022). Such coverage-based criteria have been widely used in core-set construction and batch active learning because they avoid repeatedly querying labels or retraining the prediction model during sample selection. Their effectiveness, however, depends critically on the distance used to measure similarity between candidate inputs.

Permutation-aware selection. Existing coverage-based methods rely on a prescribed distance between candidate inputs. For set-valued inputs, directly using Euclidean distance on ordered representations is inappropriate because the resulting distance depends on the arbitrary ordering of set elements. To address this issue, PAMS performs maximin selection under an optimal-matching distance that is invariant to permutations of the input elements. For equally weighted empirical measures, this distance coincides with the Wasserstein distance, making the sampling criterion consistent with the Wasserstein stability considered throughout this work. Moreover, the resulting coverage radius directly enters an upper bound on the prediction error, as shown in Appendix F.1.

B Proof of the N-Independent Latent-Dimension Result

We prove Theorem 4.1 constructively by developing the Soft-Histogram Encoder (SHE) and a corresponding decoder. The argument proceeds in four steps. First, Sec. B.1 constructs a finite grid on $\mathcal{K}$ and bounds the number of grid centers. Second, Sec. B.2 defines a continuous partition-of-unity encoder that softly assigns each input to nearby centers. Third, Sec. B.3 shows that averaging these assignments induces a grid-supported measure that approximates the empirical measure in Wasserstein distance. Fourth, Sec. B.4 constructs a continuous decoder on the latent space. These components are assembled in Sec. B.5 to establish the uniform approximation and latent-dimension bounds. The main construction concerns the non-degenerate regime $0 < \tau \le L_F\sqrt{D}/2$; the case $\tau > L_F\sqrt{D}/2$ admits a direct one-center construction and is handled separately in Sec. B.5.

B.1 Grid Construction and Covering Bound

We begin with the grid construction used in the proof of Theorem 4.1. The following lemma shows that the grid provides the required spatial resolution while keeping the number of grid points independent of the set cardinality N.

Lemma B.1 (Grid covering). *Suppose $L_F > 0$ and $0 < \tau \le L_F\sqrt{D}/2$. Define*

$$r = \frac{\tau}{4\sqrt{D}\,L_F}, \qquad n = \left\lceil \frac{1}{2r} \right\rceil,$$

and let

$$\mathcal{C} = \left\{ \left(\frac{j_1}{n}, \dots, \frac{j_D}{n} \right) : (j_1, \dots, j_D) \in \{0, \dots, n\}^D \right\} = \{c_1, \dots, c_m\}$$

be the uniform grid on $\mathcal{K} = [0,1]^D$. Then every $x \in \mathcal{K}$ has a grid center $c_\ell \in \mathcal{C}$ satisfying

$$\|x - c_\ell\|_\infty \le \frac{1}{2n} \le r,$$

and the number of grid centers satisfies

$$m \le \left(\frac{3\sqrt{D}\,L_F}{\tau} \right)^D.$$

Proof. We prove the covering property first and then bound the number of grid centers. To express the grid resolution in terms of the prescribed tolerance τ, substitute the definition of r into that of n to obtain

$$n = \left\lceil \frac{2\sqrt{D}\,L_F}{\tau} \right\rceil.$$

Covering property. Fix any point $x = (x_1, \ldots, x_D) \in \mathcal{K}$. Because $x_k \in [0, 1]$, we may choose, for each coordinate k, an integer $j_k \in \{0, \ldots, n\}$ nearest to nx_k. Every real number is within distance $1/2$ of a nearest integer, so

$$|nx_k - j_k| \leq \frac{1}{2}.$$

Dividing by n gives

$$\left| x_k - \frac{j_k}{n} \right| \leq \frac{1}{2n}, \qquad k = 1, \ldots, D.$$

Thus, the selected indices define the grid center

$$c_\ell = \left(\frac{j_1}{n}, \ldots, \frac{j_D}{n} \right),$$

which belongs to $\mathcal{C}$ by its definition. The coordinatewise bound therefore yields

$$\|x - c_\ell\|_\infty = \max_{1 \leq k \leq D} \left| x_k - \frac{j_k}{n} \right| \leq \frac{1}{2n}.$$

Moreover, by the definition of n,

$$n = \left\lceil \frac{1}{2r} \right\rceil \geq \frac{1}{2r},$$

and therefore

$$\frac{1}{2n} \leq r.$$

Combining the two inequalities above, for every $x \in \mathcal{K}$ there exists a grid center $c_\ell \in \mathcal{C}$ such that

$$\|x - c_\ell\|_\infty \leq \frac{1}{2n} \leq r.$$

Number of grid centers. It remains to control the size of $\mathcal{C}$. Each coordinate index j_k takes one of $n + 1$ values independently, so

$$m = (n + 1)^D.$$

Using $\lceil a \rceil \leq a + 1$, we obtain

$$n + 1 = \left\lceil \frac{2\sqrt{D}\, L_F}{\tau} \right\rceil + 1 \leq \frac{2\sqrt{D}\, L_F}{\tau} + 2.$$

To simplify this upper bound, the assumed inequality

$$\tau \leq \frac{L_F \sqrt{D}}{2},$$

implies

$$\frac{\sqrt{D}\, L_F}{\tau} \geq 2,$$

so the additive constant can be absorbed as follows:

$$\frac{2\sqrt{D}\, L_F}{\tau} + 2 \leq \frac{3\sqrt{D}\, L_F}{\tau}.$$

Combining this estimate with $m = (n + 1)^D$ gives

$$m = (n + 1)^D \leq \left(\frac{3\sqrt{D}\, L_F}{\tau} \right)^D.$$

This completes the proof. □

B.2 SOFT-HISTOGRAM ENCODER

The grid constructed above provides finitely many representatives of points in $\mathcal{K}$. We now use this grid to build a continuous finite-dimensional representation of each point. A hard assignment to the nearest grid center would generally be discontinuous, so instead we assign locally supported soft weights to nearby centers and normalize them to form a partition of unity. This construction is the pointwise Soft-Histogram Encoder used later to represent an entire input set.

Definition B.2 (Soft-Histogram Encoder)**.** *For each $x \in \mathcal{K}$ and each grid center c_ℓ, define*

$$\widetilde{\psi}_\ell(x) = \max\left\{0, 1 - \frac{\|x - c_\ell\|_\infty}{2r}\right\}, \qquad \psi_\ell(x) = \frac{\widetilde{\psi}_\ell(x)}{\sum_{q=1}^m \widetilde{\psi}_q(x)}.$$

The pointwise Soft-Histogram Encoder is

$$\phi(x) = \big(\psi_1(x), \dots, \psi_m(x)\big).$$

We now show that the denominator in Definition B.2 is uniformly positive, so the normalized weights are well-defined and continuous. We also show that $\psi_\ell(x) > 0$ exactly when c_ℓ lies within ℓ^∞-distance $2r$ of x. Hence the encoder assigns weight only to nearby grid centers, a property used later to bound the approximation error.

Lemma B.3 (Encoder continuity and locality)**.** *Suppose $L_F > 0$ and $0 < \tau \le L_F\sqrt{D}/2$, and consider the grid from Lemma B.1 and the encoder from Definition B.2. Define the normalizing factor*

$$S(x) = \sum_{q=1}^m \widetilde{\psi}_q(x).$$

Then $S(x) \ge 1/2$ for every $x \in \mathcal{K}$. Hence the normalized weights $\psi_1, \dots, \psi_m$ are well-defined, nonnegative, and continuous on $\mathcal{K}$, and they form a partition of unity:

$$\sum_{\ell=1}^m \psi_\ell(x) = 1.$$

Consequently, $\phi : \mathcal{K} \to \Sigma_m$ is continuous. In addition, each weight is supported near its corresponding grid center: for every $x \in \mathcal{K}$ and $\ell \in \{1, \dots, m\}$,

$$\psi_\ell(x) > 0 \quad \Longleftrightarrow \quad \|x - c_\ell\|_\infty < 2r.$$

Proof. We first recall from Lemma B.1 the quantities used in the encoder construction:

$$\mathcal{K} = [0,1]^D, \qquad r = \frac{\tau}{4\sqrt{D}\, L_F} > 0.$$

For each grid center c_ℓ, the weight function is

$$\widetilde{\psi}_\ell(x) = \max\left\{0, 1 - \frac{\|x - c_\ell\|_\infty}{2r}\right\}, \tag{6}$$

and its normalizing factor is

$$S(x) = \sum_{q=1}^m \widetilde{\psi}_q(x).$$

The normalized weights are therefore given by

$$\psi_\ell(x) = \frac{\widetilde{\psi}_\ell(x)}{S(x)}, \tag{7}$$

and the pointwise encoder is

$$\phi(x) = \big(\psi_1(x), \dots, \psi_m(x)\big). \tag{8}$$

We will show that the encoder takes values in the probability simplex

$$\Sigma_m = \left\{u \in \mathbb{R}^m : u_\ell \ge 0 \text{ for all } \ell,\ \sum_{\ell=1}^m u_\ell = 1\right\}.$$

The proof follows the dependencies in this construction. We first establish a uniform positive lower bound for $S(x)$. This ensures that the denominator in (7) never vanishes and hence that the normalized weights are well-defined. We then prove that ϕ is continuous and simplex-valued, and finally verify the locality property.

Step 1: the normalization is well-defined. Fix $x \in \mathcal{K}$. By Lemma B.1, there exists a center $c_{q^\star}$ such that

$$\|x - c_{q^\star}\|_\infty \le \frac{1}{2n} \le r.$$

Using (6) and $r > 0$, we obtain

$$\widetilde{\psi}_{q^\star}(x) \ge 1 - \frac{\|x - c_{q^\star}\|_\infty}{2r} \ge 1 - \frac{r}{2r} = \frac{1}{2}.$$

Since every $\widetilde{\psi}_q(x)$ is nonnegative,

$$S(x) = \sum_{q=1}^{m} \widetilde{\psi}_q(x) \ge \widetilde{\psi}_{q^\star}(x) \ge \frac{1}{2}.$$

As x was arbitrary, $S(x) \ge 1/2$ on $\mathcal{K}$. Hence the denominator in (7) never vanishes, and every normalized weight is well-defined.

Step 2: the encoder is continuous and simplex-valued. For each ℓ, the map $x \mapsto \|x - c_\ell\|_\infty$ is continuous. Equation (6) then shows that $\widetilde{\psi}_\ell$ is continuous, because it is the maximum of two continuous functions. Hence S, being a finite sum of the bumps, is also continuous. Step 1 gives $S(x) \ge 1/2$, so dividing by S preserves continuity. Therefore every ψ_ℓ is continuous.

Moreover, $\widetilde{\psi}_\ell(x) \ge 0$ and $S(x) > 0$, so $\psi_\ell(x) \ge 0$. Summing the normalized weights gives

$$\sum_{\ell=1}^{m} \psi_\ell(x) = \frac{1}{S(x)} \sum_{\ell=1}^{m} \widetilde{\psi}_\ell(x) = \frac{S(x)}{S(x)} = 1.$$

Thus, $\phi(x) \in \Sigma_m$ for every $x \in \mathcal{K}$. Since all of its coordinate functions are continuous, $\phi : \mathcal{K} \to \Sigma_m$ is continuous.

Step 3: positive weights are local. Fix $\ell \in \{1, \dots, m\}$ and $x \in \mathcal{K}$. Since $S(x) > 0$,

$$\psi_\ell(x) > 0 \quad \Longleftrightarrow \quad \widetilde{\psi}_\ell(x) > 0.$$

By (6), the latter condition is equivalent to

$$1 - \frac{\|x - c_\ell\|_\infty}{2r} > 0 \quad \Longleftrightarrow \quad \|x - c_\ell\|_\infty < 2r,$$

where the final equivalence uses $r > 0$. This proves the locality statement and completes the proof. □

Figure 4 illustrates the construction in one dimension.

B.3 Approximation by a Grid-Supported Measure

To compare the averaged encoder representation with the original input set, we interpret its coordinates as probability masses assigned to the corresponding grid centers. Accordingly, for any $z = (z_1, \dots, z_m) \in \Sigma_m$, define the grid-supported measure

$$\nu_z = \sum_{\ell=1}^{m} z_\ell \, \delta_{c_\ell}. \tag{9}$$

Here, δ_{c_ℓ} denotes the Dirac mass at the grid center c_ℓ. Because the coordinates of z are nonnegative and sum to one and every c_ℓ lies in $\mathcal{K}$, we have $\nu_z \in \mathcal{P}(\mathcal{K})$. We now show that, when z is the averaged encoder output, ν_z approximates the empirical measure $\mu_{\mathbf{x}}$ in Wasserstein distance.

Lemma B.4 (Grid-measure approximation)**.** *Suppose $L_F > 0$ and $0 < \tau \le L_F\sqrt{D}/2$. For every $N \ge 1$ and every $\mathbf{x} = (x_1, \dots, x_N) \in \mathcal{K}^N$, let*

$$z = \frac{1}{N} \sum_{i=1}^{N} \phi(x_i) \in \Sigma_m.$$

Then the grid-supported measure ν_z satisfies

$$W_1(\mu_{\mathbf{x}}, \nu_z) \le 2\sqrt{D}\, r = \frac{\tau}{2L_F}.$$

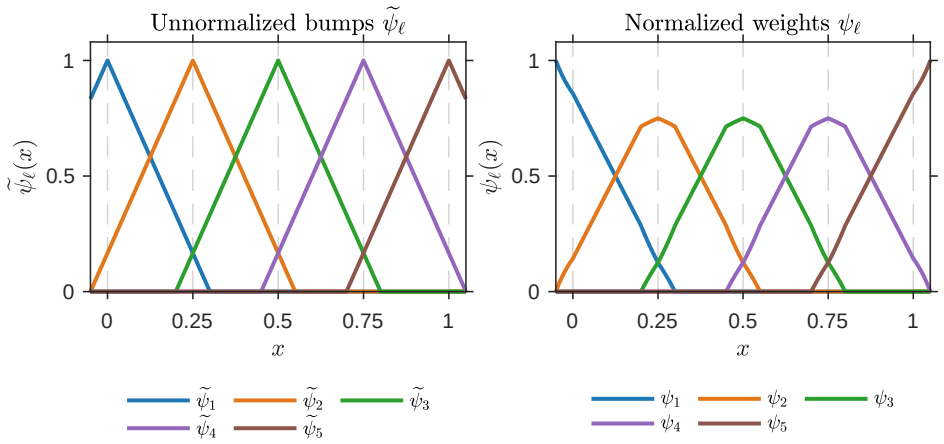

Figure 4: Soft-histogram encoder in one dimension. **Left:** each triangular bump $\widetilde{\psi}_\ell$ equals 1 at its grid center c_ℓ and decreases to 0 at distance $2r$. **Right:** normalization produces nonnegative weights that sum to one. A weight $\psi_\ell(x)$ is positive only when $|x - c_\ell| < 2r$.

Proof. We prove the claim by constructing an explicit coupling between $\mu_{\mathbf{x}}$ and ν_z and bounding its transport cost. From the definition of z and the coordinate representation of ϕ, we have

$$z_\ell = \frac{1}{N} \sum_{i=1}^{N} \psi_\ell(x_i). \tag{10}$$

By the Kantorovich formulation, any coupling Π with marginals $\mu_{\mathbf{x}}$ and ν_z gives the upper bound

$$W_1(\mu_{\mathbf{x}}, \nu_z) \le \int \|x - c\|_2 \, d\Pi.$$

It therefore suffices to construct such a coupling with cost at most $\tau/(2L_F)$.

Step 1: define the transport plan. Each input point x_i carries mass $1/N$ in $\mu_{\mathbf{x}}$. We split this mass among the grid centers, sending the fraction $\psi_\ell(x_i)$ to c_ℓ. The resulting transport plan is

$$\Pi := \frac{1}{N} \sum_{i=1}^{N} \sum_{\ell=1}^{m} \psi_\ell(x_i) \, \delta_{(x_i, c_\ell)}.$$

Here, $\delta_{(x_i,c_\ell)}$ records transport from the source x_i to the destination c_ℓ. Since $\psi_\ell(x_i) \ge 0$ and $\sum_\ell \psi_\ell(x_i) = 1$ (Lemma B.3), Π is a probability measure.

Step 2: verify that Π is a coupling. For Π to couple $\mu_{\mathbf{x}}$ and ν_z, its first marginal must be $\mu_{\mathbf{x}}$ and its second marginal must be ν_z. We verify these two requirements in turn. For each indexed input x_i, the total mass transported from x_i is $N^{-1} \sum_\ell \psi_\ell(x_i) = 1/N$. Hence the first marginal is $N^{-1} \sum_i \delta_{x_i} = \mu_{\mathbf{x}}$. For each center c_ℓ, the total mass received is $N^{-1} \sum_i \psi_\ell(x_i) = z_\ell$ by (10). Hence the second marginal is $\sum_\ell z_\ell \, \delta_{c_\ell} = \nu_z$.

Step 3: bound the transport cost. By the locality property of Lemma B.3, $\psi_\ell(x_i) > 0$ only when $\|x_i - c_\ell\|_\infty < 2r$. For any such pair, $\|x_i - c_\ell\|_2 \le \sqrt{D} \, \|x_i - c_\ell\|_\infty < 2r\sqrt{D}$. Since terms with $\psi_\ell(x_i) = 0$ contribute nothing:

$$\int \|x - c\|_2 \, d\Pi = \sum_{i,\ell} \frac{1}{N} \psi_\ell(x_i) \|x_i - c_\ell\|_2 \le 2r\sqrt{D} \sum_i \frac{1}{N} \underbrace{\sum_\ell \psi_\ell(x_i)}_{=1} = 2r\sqrt{D}.$$

Step 4: substitute r and conclude.

$$W_1(\mu_{\mathbf{x}}, \nu_z) \le 2r\sqrt{D} = 2 \cdot \frac{\tau}{4\sqrt{D}\, L_F} \cdot \sqrt{D} = \frac{\tau}{2L_F}.$$

The resulting bound is independent of N and holds uniformly over all set sizes. □

B.4 Decoder Construction and Continuity

We next convert the averaged encoder representation z into an output. By Lemma B.4, its associated measure ν_z approximates the input measure $\mu_{\mathbf{x}}$. We therefore first define a decoder on the set of valid encoder outputs by

$$g : \Sigma_m \to \mathbb{R}, \qquad g(z) := F(\nu_z).$$

The theorem requires a decoder whose domain is the entire latent space $\mathbb{R}^m$, whereas g is defined only on Σ_m. To extend g, for each $u \in \mathbb{R}^m$ define its Euclidean projection onto Σ_m by

$$\Pi_{\Sigma_m}(u) := \arg\min_{v \in \Sigma_m} \|u - v\|_2. \tag{11}$$

Because Σ_m is nonempty, closed, and convex, this minimizer exists and is unique. Thus, $\Pi_{\Sigma_m}(u)$ is the point in Σ_m closest to u in Euclidean distance. We define the decoder on $\mathbb{R}^m$ by

$$\rho(u) := g(\Pi_{\Sigma_m}(u)) = F\big(\nu_{\Pi_{\Sigma_m}(u)}\big).$$

If $z \in \Sigma_m$, then $\Pi_{\Sigma_m}(z) = z$ and hence $\rho(z) = g(z) = F(\nu_z)$. The following lemma verifies that this extension is continuous.

Lemma B.5 (Continuous decoder extension)**.** *Under Assumption 1, the decoder $\rho : \mathbb{R}^m \to \mathbb{R}$ defined above is continuous and satisfies $\rho(z) = F(\nu_z)$ for every $z \in \Sigma_m$. Moreover,*

$$|\rho(z) - \rho(z')| \le \frac{L_F\sqrt{D}}{2}\|z - z'\|_1, \qquad z, z' \in \Sigma_m.$$

Proof. Recall that each $z \in \Sigma_m$ specifies the grid-supported probability measure

$$\nu_z = \sum_{\ell=1}^{m} z_\ell \, \delta_{c_\ell} \in \mathcal{P}(\mathcal{K}),$$

so $g(z) = F(\nu_z)$ is well-defined. We first establish the Lipschitz bound for g on Σ_m by controlling the Wasserstein distance between ν_z and $\nu_{z'}$. We then use the fact that projecting two points onto Σ_m cannot increase the distance between them to show that the extended decoder ρ is continuous on all of $\mathbb{R}^m$.

Step 1: quantify the unmatched mass. For any $z, z' \in \Sigma_m$, at the grid center c_ℓ, the measure ν_z places mass z_ℓ, whereas $\nu_{z'}$ places mass z'_ℓ. When moving the mass of ν_z to obtain $\nu_{z'}$, we do not need to move the portion that is already present at the same center in both measures. The largest portion that can stay at c_ℓ is

$$s_\ell = \min\{z_\ell, z'_\ell\}.$$

After leaving this amount at c_ℓ, define

$$p_\ell = z_\ell - s_\ell, \qquad q_\ell = z'_\ell - s_\ell.$$

Here p_ℓ is the excess mass of ν_z at c_ℓ, which must be moved away from that center. Similarly, q_ℓ is the additional mass that must be moved to c_ℓ in order to obtain $\nu_{z'}$. For a fixed ℓ, only one of these quantities can be positive. Indeed, if $z_\ell \ge z'_\ell$, then $s_\ell = z'_\ell$, so

$$p_\ell = z_\ell - z'_\ell, \qquad q_\ell = 0.$$

If $z_\ell < z'_\ell$, then $s_\ell = z_\ell$, so

$$p_\ell = 0, \qquad q_\ell = z'_\ell - z_\ell.$$

Therefore, in either case,

$$p_\ell + q_\ell = |z_\ell - z'_\ell|.$$

We next calculate the total excess mass that must leave its original grid centers. Because both z and z' belong to Σ_m, their coordinates sum to one: $\sum_{\ell=1}^m z_\ell = \sum_{\ell=1}^m z'_\ell = 1$. Therefore, using $p_\ell = z_\ell - s_\ell$, we have

$$\sum_{\ell=1}^{m} p_\ell = \sum_{\ell=1}^{m} (z_\ell - s_\ell) = 1 - \sum_{\ell=1}^{m} s_\ell.$$

Similarly, using $q_\ell = z'_\ell - s_\ell$, we obtain

$$\sum_{\ell=1}^{m} q_\ell = \sum_{\ell=1}^{m} (z'_\ell - s_\ell) = 1 - \sum_{\ell=1}^{m} s_\ell.$$

It follows that

$$\sum_{\ell=1}^{m} p_\ell = \sum_{\ell=1}^{m} q_\ell =: \alpha.$$

Using the identity above, we obtain

$$2\alpha = \sum_{\ell=1}^{m} p_\ell + \sum_{\ell=1}^{m} q_\ell = \sum_{\ell=1}^{m} (p_\ell + q_\ell) = \sum_{\ell=1}^{m} |z_\ell - z'_\ell| = \|z - z'\|_1.$$

Consequently,

$$\alpha = \frac{1}{2}\|z - z'\|_1.$$

Thus, α is the total mass that must be transported between distinct grid centers.

Step 2: construct a coupling and bound its cost. Our goal is to bound $|g(z) - g(z')|$. Since $g(z) = F(\nu_z)$, Assumption 1 gives

$$|g(z) - g(z')| \le L_F W_1(\nu_z, \nu_{z'}).$$

It therefore remains to bound $W_1(\nu_z, \nu_{z'})$. For this purpose, it is enough to exhibit one coupling of the two measures and bound its transport cost. If $\alpha = 0$, then no mass needs to be moved and $z = z'$, so the desired bound holds immediately. We therefore assume $\alpha > 0$. We leave the mass s_ℓ at each c_ℓ and distribute the excess mass p_ℓ among the grid centers that require the additional mass q_k. This transport plan is given by

$$\gamma = \sum_{\ell=1}^{m} s_\ell \, \delta_{(c_\ell, c_\ell)} + \frac{1}{\alpha} \sum_{\ell=1}^{m} \sum_{k=1}^{m} p_\ell q_k \, \delta_{(c_\ell, c_k)}.$$

We now verify that γ is a coupling of ν_z and $\nu_{z'}$. To do so, we must show that its first marginal is ν_z and its second marginal is $\nu_{z'}$.

We verify the two marginals by integration. Let $h : \mathcal{K} \to \mathbb{R}$ be an arbitrary continuous function. We begin with the integral that determines the first marginal:

$$\int_{\mathcal{K}\times\mathcal{K}} h(x) \, d\gamma(x, y).$$

For a Dirac measure at (c_ℓ, c_k),

$$\int_{\mathcal{K}\times\mathcal{K}} h(x) \, d\delta_{(c_\ell, c_k)}(x, y) = h(c_\ell).$$

This holds because the Dirac measure places all its mass at (c_ℓ, c_k) and zero mass elsewhere, so the integrand is evaluated only at $x = c_\ell$. Substituting the definition of γ therefore gives

$$\begin{aligned}\int_{\mathcal{K}\times\mathcal{K}} h(x) \, d\gamma(x, y) &= \sum_{\ell=1}^{m} s_\ell h(c_\ell) + \frac{1}{\alpha} \sum_{\ell=1}^{m} p_\ell h(c_\ell) \underbrace{\left(\sum_{k=1}^{m} q_k \right)}_{=\alpha} \\ &= \sum_{\ell=1}^{m} \underbrace{(s_\ell + p_\ell)}_{=z_\ell} h(c_\ell) = \sum_{\ell=1}^{m} z_\ell h(c_\ell) = \int_{\mathcal{K}} h \, d\nu_z.\end{aligned}$$

Thus, the first marginal of γ is ν_z.

For the second marginal, we instead integrate $h(y)$. Since a Dirac measure at (c_ℓ, c_k) evaluates this integrand at $y = c_k$, the same calculation gives

$$\int_{\mathcal{K}\times\mathcal{K}} h(y)\, d\gamma(x,y) = \sum_{k=1}^{m} s_k h(c_k) + \frac{1}{\alpha}\sum_{k=1}^{m} q_k h(c_k) \underbrace{\left(\sum_{\ell=1}^{m} p_\ell\right)}_{=\alpha}$$
$$= \sum_{k=1}^{m} \underbrace{(s_k + q_k)}_{=z'_k} h(c_k) = \sum_{k=1}^{m} z'_k h(c_k) = \int_{\mathcal{K}} h\, d\nu_{z'}.$$

Thus, the second marginal of γ is $\nu_{z'}$. Therefore, γ is a coupling of ν_z and $\nu_{z'}$.

For any coupling, its transport cost bounds the difference between the expectations of every 1-Lipschitz function under the two marginals. Taking the supremum over such functions shows that this cost upper-bounds W_1. For γ, the diagonal part has zero cost. Therefore,

$$\int_{\mathcal{K}\times\mathcal{K}} \|x-y\|_2\, d\gamma(x,y) = \sum_{\ell=1}^{m} s_\ell \|c_\ell - c_\ell\|_2 + \frac{1}{\alpha}\sum_{\ell=1}^{m}\sum_{k=1}^{m} p_\ell q_k \|c_\ell - c_k\|_2$$
$$= \frac{1}{\alpha}\sum_{\ell=1}^{m}\sum_{k=1}^{m} p_\ell q_k \|c_\ell - c_k\|_2.$$

Since $W_1(\nu_z, \nu_{z'})$ is no greater than the cost of this particular coupling and $\operatorname{diam}(\mathcal{K}) = \sqrt{D}$, we obtain

$$W_1(\nu_z, \nu_{z'}) \le \frac{1}{\alpha}\sum_{\ell=1}^{m}\sum_{k=1}^{m} p_\ell q_k \|c_\ell - c_k\|_2 \le \frac{\sqrt{D}}{\alpha}\left(\sum_{\ell=1}^{m} p_\ell\right)\left(\sum_{k=1}^{m} q_k\right)$$
$$= \alpha\sqrt{D} = \frac{\sqrt{D}}{2}\|z - z'\|_1.$$

By Assumption 1, we have

$$|g(z) - g(z')| = |F(\nu_z) - F(\nu_{z'})| \le L_F W_1(\nu_z, \nu_{z'}) \le \frac{L_F\sqrt{D}}{2}\|z - z'\|_1.$$

Hence g is Lipschitz on Σ_m with respect to the ℓ^1 norm.

Step 3: extend the decoder to the full latent space. To obtain a decoder on all of $\mathbb{R}^m$ that agrees with g on Σ_m, we compose g with the Euclidean projection onto Σ_m. The simplex is a nonempty closed convex subset of $\mathbb{R}^m$, so its Euclidean projection is nonexpansive:

$$\|\Pi_{\Sigma_m}(u) - \Pi_{\Sigma_m}(v)\|_2 \le \|u - v\|_2, \qquad u, v \in \mathbb{R}^m.$$

Define

$$\rho(u) = g(\Pi_{\Sigma_m}(u)) = F\left(\nu_{\Pi_{\Sigma_m}(u)}\right), \qquad u \in \mathbb{R}^m.$$

It remains to verify that this extension is continuous on $\mathbb{R}^m$, as required by the theorem. The Lipschitz bound from Step 2, the inequality $\|w\|_1 \le \sqrt{m}\|w\|_2$, and the nonexpansiveness of the projection give, for all $u, v \in \mathbb{R}^m$,

$$|\rho(u) - \rho(v)| \le \frac{L_F\sqrt{D}}{2}\|\Pi_{\Sigma_m}(u) - \Pi_{\Sigma_m}(v)\|_1$$
$$\le \frac{L_F\sqrt{Dm}}{2}\|u - v\|_2.$$

Thus, ρ is continuous on $\mathbb{R}^m$. Finally, the projection leaves every $z \in \Sigma_m$ unchanged, and therefore

$$\rho(z) = g(z) = F(\nu_z), \qquad z \in \Sigma_m. \tag{12}$$

Hence ρ extends $z \mapsto F(\nu_z)$, and the estimate from Step 2 is the claimed Lipschitz bound on Σ_m. □

B.5 Proof of Theorem 4.1

We now assemble the preceding lemmas and verify the continuity, uniform approximation, and dimension claims in Theorem 4.1.

Proof. We consider separately the degenerate and non-degenerate regimes.

Degenerate (large-tolerance) regime. Suppose $\tau > L_F\sqrt{D}/2$. Let $c_1 = (1/2, \dots, 1/2) \in \mathcal{K}$, take $m = 1$, and define

$$\phi(x) = 1, \qquad \rho(u) = F(\delta_{c_1}).$$

Both maps are continuous. For any $\mathbf{x} = (x_1, \dots, x_N) \in \mathcal{K}^N$, the coupling that sends every x_i to c_1 gives

$$W_1(\mu_{\mathbf{x}}, \delta_{c_1}) \le \frac{1}{N}\sum_{i=1}^{N} \|x_i - c_1\|_2 \le \frac{\sqrt{D}}{2}.$$

The first inequality follows from the coupling that transports each atom x_i to c_1, while the second follows from $\|x_i - c_1\|_2 \le \sqrt{D}/2$ for every $x_i \in \mathcal{K}$. Therefore, Assumption 1 yields

$$\left| F(\mu_{\mathbf{x}}) - \rho\left(\frac{1}{N}\sum_{i=1}^{N}\phi(x_i)\right)\right| \le \frac{L_F\sqrt{D}}{2} < \tau.$$

Moreover,

$$m = 1 \le \left\lceil \frac{3\sqrt{D}\, L_F}{\tau} \right\rceil^D,$$

because the quantity inside the ceiling is positive. Hence the theorem holds in this regime.

Non-degenerate regime. Suppose $0 < \tau \le L_F\sqrt{D}/2$.

Step 1: assemble the encoder and decoder. Let $\{c_\ell\}_{\ell=1}^m$ be the grid from Lemma B.1, let ϕ be the Soft-Histogram Encoder from Definition B.2, and let ρ be the decoder from Lemma B.5. By Lemma B.3, $\phi : \mathcal{K} \to \Sigma_m \subset \mathbb{R}^m$ is continuous. By Lemma B.5, $\rho : \mathbb{R}^m \to \mathbb{R}$ is continuous.

Step 2: verify the uniform approximation bound. For any $\mathbf{x} = (x_1, \dots, x_N) \in \mathcal{K}^N$, define

$$z = \frac{1}{N}\sum_{i=1}^{N}\phi(x_i), \qquad \nu_z = \sum_{\ell=1}^{m} z_\ell\, \delta_{c_\ell}.$$

Every $\phi(x_i)$ lies in the convex set Σ_m, so $z \in \Sigma_m$. Consequently, (12), Assumption 1, and Lemma B.4 give

$$\begin{aligned}\left| F(\mu_{\mathbf{x}}) - \rho\left(\frac{1}{N}\sum_{i=1}^{N}\phi(x_i)\right)\right| &= |F(\mu_{\mathbf{x}}) - F(\nu_z)| \\ &\le L_F W_1(\mu_{\mathbf{x}}, \nu_z) \\ &\le \frac{\tau}{2} < \tau.\end{aligned}$$

Because this estimate holds for every $\mathbf{x} \in \mathcal{K}^N$, taking the supremum over $\mathcal{K}^N$ proves the required uniform approximation bound.

Step 3: verify the latent dimension. Finally, Lemma B.1 gives

$$m \le \left(\frac{3\sqrt{D}\, L_F}{\tau}\right)^D \le \left\lceil \frac{3\sqrt{D}\, L_F}{\tau} \right\rceil^D.$$

Together with Steps 1 and 2, this proves Theorem 4.1. □

C Proof of the N-Independent Sample-Complexity Result

In this part, we prove that the number of i.i.d. training samples required to learn a Wasserstein-stable target functional with ReLU DeepSets is independent of the input set size N. We first realize the continuous DeepSets approximation by finite ReLU networks whose sizes are independent of N. We then bound the covering number of the resulting hypothesis class and combine this bound with uniform convergence to obtain the claimed result.

C.1 ReLU Realization

We begin by converting the continuous encoder–decoder construction from Theorem 4.1 into a finite ReLU DeepSets model. For any prescribed approximation accuracy, we construct such a model whose latent dimension and encoder and decoder architectures are independent of the input set size N. The following proposition formalizes this realization.

Proposition C.1 (N-Independent ReLU DeepSets Approximation)**.** *For every target accuracy $\tau > 0$, there exist a latent dimension m and ReLU networks*

$$\phi_\theta : \mathcal{K} \to \mathbb{R}^m, \qquad \rho_\theta : \mathbb{R}^m \to \mathbb{R},$$

such that the vector-valued encoder ϕ_θ has W_ϕ parameters in total and depth L_ϕ, while the decoder ρ_θ has W_ρ parameters in total and depth L_ρ. The quantities m, (W_ϕ, L_ϕ), and (W_ρ, L_ρ) are independent of N, and, simultaneously for every $N \geq 1$,

$$\sup_{\mathbf{x}\in\mathcal{K}^N} \left| F(\mu_{\mathbf{x}}) - \rho_\theta\left(\frac{1}{N}\sum_{i=1}^{N} \phi_\theta(x_i)\right)\right| \leq \tau.$$

Proof. The proof proceeds in three steps. First, we apply Theorem 4.1 with target accuracy $\tau/3$ to obtain a continuous DeepSets approximation with an N-independent latent dimension. Second, we approximate the encoder by a finite ReLU network and show that mean aggregation does not amplify the approximation error. Third, we approximate the decoder on the resulting latent region and combine the three error bounds.

Step 1: obtain the continuous approximation and define the decoder domain. We first obtain the continuous approximation to be realized by ReLU networks and identify a compact region containing the latent inputs to the decoder. Apply Theorem 4.1 with tolerance $\tau/3$. There exist a latent dimension m, a continuous encoder, written coordinatewise as

$$\phi : \mathcal{K} \to \mathbb{R}^m, \qquad \phi(x) = \big(\phi_1(x), \dots, \phi_m(x)\big),$$

and a continuous decoder

$$\rho : \mathbb{R}^m \to \mathbb{R}$$

such that, for every $N \geq 1$,

$$\sup_{\mathbf{x}\in\mathcal{K}^N} \left| F(\mu_{\mathbf{x}}) - \rho\left(\frac{1}{N}\sum_{i=1}^{N} \phi(x_i)\right)\right| \leq \frac{\tau}{3}. \tag{13}$$

The encoder constructed in Theorem 4.1 is simplex-valued. In the degenerate regime, $m = 1$ and $\phi(x) = 1$; in the non-degenerate regime, its coordinates are the normalized weights from Definition B.2, which are nonnegative and sum to one by Lemma B.3. Therefore,

$$\phi(x) \in \Sigma_m := \left\{ z \in \mathbb{R}^m : z_\ell \geq 0,\ \sum_{\ell=1}^{m} z_\ell = 1 \right\}.$$

Hence, for every $\mathbf{x} = (x_1, \dots, x_N) \in \mathcal{K}^N$,

$$z := \frac{1}{N}\sum_{i=1}^{N} \phi(x_i) \in \Sigma_m, \tag{14}$$

because Σ_m is convex.

In Step 2, we replace ϕ by a ReLU encoder ϕ_θ and evaluate the decoder at $z_\theta = N^{-1}\sum_{i=1}^{N}\phi_\theta(x_i)$. A general ReLU approximation need not preserve the nonnegativity and unit-sum constraints defining Σ_m; consequently, z_θ need not belong to Σ_m. To obtain a common domain containing both z and z_θ, define the closed unit neighborhood

$$\mathcal{Z} := \{u \in \mathbb{R}^m : \operatorname{dist}(u, \Sigma_m) \le 1\}.$$

Since Σ_m is compact, so is $\mathcal{Z}$. This gives a compact domain on which to control ρ uniformly for both the exact and approximated latent representations.

Step 2: construct the ReLU encoder and control the pooling error. The encoder ϕ obtained in Step 1 is continuous but is not yet represented by a neural network. Our goal in this step is to replace it by a finite ReLU encoder ϕ_θ, while keeping the decoder ρ fixed, so that the resulting change in the decoded output is at most $\tau/3$, uniformly over all set sizes N. To determine how accurately ϕ must be approximated, we use the uniform continuity of ρ on $\mathcal{Z}$. Specifically, there exists $\delta_\phi \in (0, 1]$ such that

$$u, v \in \mathcal{Z}, \qquad \|u - v\|_2 \le \delta_\phi \quad \Longrightarrow \quad |\rho(u) - \rho(v)| \le \frac{\tau}{3}. \tag{15}$$

Each coordinate ϕ_ℓ is continuous on the compact set $\mathcal{K}$. Therefore, the ReLU universal approximation theorem (Kidger & Lyons, 2020) provides a finite ReLU network $\phi_{\theta,\ell}$ for each $\ell = 1, \dots, m$ such that

$$\sup_{x\in\mathcal{K}} |\phi_{\theta,\ell}(x) - \phi_\ell(x)| \le \frac{\delta_\phi}{\sqrt{m}}.$$

Stacking these networks defines the ReLU encoder

$$\phi_\theta := \big(\phi_{\theta,1}, \dots, \phi_{\theta,m}\big) : \mathcal{K} \to \mathbb{R}^m.$$

The coordinatewise bounds imply

$$\sup_{x\in\mathcal{K}} \|\phi_\theta(x) - \phi(x)\|_2 \le \delta_\phi. \tag{16}$$

For any $\mathbf{x} = (x_1, \dots, x_N) \in \mathcal{K}^N$, let

$$z_\theta := \frac{1}{N}\sum_{i=1}^{N}\phi_\theta(x_i).$$

By the triangle inequality and (16),

$$\|z_\theta - z\|_2 = \left\|\frac{1}{N}\sum_{i=1}^{N}\big(\phi_\theta(x_i) - \phi(x_i)\big)\right\|_2 \le \frac{1}{N}\sum_{i=1}^{N}\|\phi_\theta(x_i) - \phi(x_i)\|_2 \le \delta_\phi. \tag{17}$$

Because $z \in \Sigma_m$, (17) and $\delta_\phi \le 1$ imply $\operatorname{dist}(z_\theta, \Sigma_m) \le \|z_\theta - z\|_2 \le \delta_\phi \le 1$. Thus, both z and z_θ lie in $\mathcal{Z}$, and their distance is at most δ_ϕ. Applying (15) therefore gives

$$|\rho(z_\theta) - \rho(z)| \le \frac{\tau}{3}. \tag{18}$$

Step 3: construct the ReLU decoder and combine the error bounds. The decoder ρ obtained in Step 1 is continuous but is not yet represented by a neural network. Our goal in this step is to replace it by a finite ReLU decoder ρ_θ while introducing an additional error of at most $\tau/3$. Since ρ is continuous on the compact set $\mathcal{Z}$, the ReLU universal approximation theorem provides a finite ReLU network

$$\rho_\theta : \mathbb{R}^m \to \mathbb{R}$$

such that

$$\sup_{u\in\mathcal{Z}} |\rho_\theta(u) - \rho(u)| \le \frac{\tau}{3}. \tag{19}$$

Step 2 shows that $z_\theta \in \mathcal{Z}$, so (19) implies

$$|\rho_\theta(z_\theta) - \rho(z_\theta)| \le \frac{\tau}{3}. \tag{20}$$

Fix any $N \ge 1$ and $\mathbf{x} \in \mathcal{K}^N$. Combining (13), (18), and (20) gives

$$\begin{aligned} |F(\mu_{\mathbf{x}}) - \rho_\theta(z_\theta)| &\le |F(\mu_{\mathbf{x}}) - \rho(z)| + |\rho(z) - \rho(z_\theta)| + |\rho(z_\theta) - \rho_\theta(z_\theta)| \\ &\le \frac{\tau}{3} + \frac{\tau}{3} + \frac{\tau}{3} = \tau. \end{aligned} \tag{21}$$

Taking the supremum over $\mathbf{x} \in \mathcal{K}^N$ gives

$$\sup_{\mathbf{x}\in\mathcal{K}^N} \left| F(\mu_{\mathbf{x}}) - \rho_\theta\left(\frac{1}{N}\sum_{i=1}^{N} \phi_\theta(x_i)\right)\right| \le \tau.$$

Finally, the ReLU universal approximation theorem guarantees a finite network for any fixed continuous target, compact approximation domain, and positive approximation accuracy (Kidger & Lyons, 2020). In our construction, Theorem 4.1 chooses m, ϕ, and ρ independently of N. Hence the approximation domains $\mathcal{K}$ and $\mathcal{Z}$, the encoder accuracies $\delta_\phi/\sqrt{m}$, and the decoder accuracy $\tau/3$ are all independent of N. Applying the theorem to these fixed objects therefore yields finite encoder and decoder networks that can be chosen once and used for every N. Their total parameter counts and depths, (W_ϕ, L_ϕ) and (W_ρ, L_ρ), are consequently independent of N. □

C.2 Proof of Theorem 4.2

We first restate the result.

Theorem 4.2 (N-independent sample complexity). Fix a target risk $\varepsilon_r \in (0, 1)$ and confidence level $1 - \zeta$. Let $\mathcal{H}_N$ be the ReLU DeepSets class associated with Proposition C.1, with latent dimension m. Let (W_ϕ, L_ϕ) denote the total parameter count and depth of the vector-valued encoder $\phi : \mathcal{K} \to \mathbb{R}^m$, and let (W_ρ, L_ρ) denote those of the decoder. For every distribution $\mathcal{D}_N$ on $\mathcal{K}^N$, the sample size

$$M = \widetilde{\mathcal{O}}\left(\frac{W_\phi L_\phi + W_\rho L_\rho + \log(1/\zeta)}{\varepsilon_r^2}\right)$$

is sufficient. With probability at least $1 - \zeta$, every $\varepsilon_r/4$-approximate empirical risk minimizer $\widehat{h}_M \in \mathcal{H}_N$ satisfies

$$\mathcal{L}_N(\widehat{h}_M) \le \varepsilon_r.$$

Proof. Our goal is to show that the learned predictor has population risk at most ε_r. We divide this risk budget into four terms, each of size at most $\varepsilon_r/4$. The argument proceeds as follows.

First, we use Proposition C.1 to fix one DeepSets architecture that works for every set size N. Its hypothesis class contains a reference predictor h whose population risk is at most $\varepsilon_r/4$; this is the approximation term.

Second, we measure the size of the hypothesis class by constructing an empirical L^1 cover. We first cover the encoder outputs at all elements in the sample, average these outputs within each set, and then cover the decoder on the resulting pooled representations. Averaging introduces a factor $1/N$ that cancels the sum over the N elements. Consequently, the logarithm of the covering number depends on the encoder and decoder network sizes, but not on N.

Third, boundedness converts this predictor cover into a uniform generalization bound: with sufficiently many training samples, the empirical and population risks differ by at most $\varepsilon_r/4$ simultaneously for every predictor in the class.

Finally, approximate empirical optimality says that the empirical risk of the learned predictor $\widehat{h}_M$ is at most the empirical risk of the reference predictor h, plus $\varepsilon_r/4$. Applying the uniform generalization bound once to $\widehat{h}_M$ and once to h gives

$$\mathcal{L}_N(\widehat{h}_M) \le \underbrace{\mathcal{L}_N(h)}_{\text{approximation}} + \underbrace{\frac{\varepsilon_r}{4}}_{\text{optimization}} + \underbrace{2\frac{\varepsilon_r}{4}}_{\text{generalization}} \le \varepsilon_r.$$

The remainder of the proof establishes these four claims in detail.

Step 1: select a class containing an accurate predictor. We first ensure that the hypothesis class contains a predictor with population risk at most $\varepsilon_r/4$. Apply Proposition C.1 with approximation accuracy

$$\tau = \frac{\sqrt{\varepsilon_r}}{2}.$$

Use the resulting N-independent ReLU architecture to define $\mathcal{H}_N$. By the proposition, this class contains a predictor h satisfying

$$\sup_{X \in \mathcal{K}^N} |h(X) - F(\mu_X)| \le \frac{\sqrt{\varepsilon_r}}{2}.$$

Therefore,

$$\mathcal{L}_N(h) = \mathbb{E}_{X \sim \mathcal{D}_N}\left[(h(X) - F(\mu_X))^2\right] \le \frac{\varepsilon_r}{4}. \tag{22}$$

Thus, the approximation error contributes at most one quarter of the target risk.

Step 2: construct an N-independent cover of the DeepSets class. We use standard covering bounds for ReLU networks with bounded parameters. Choose $B < \infty$ large enough to contain all weights and biases of the reference predictor h, and define $\mathcal{H}_N$ by allowing the parameters of the same architecture to vary in $[-B, B]$. Thus, h remains in $\mathcal{H}_N$.

This parameter bound has two consequences needed below. First, the encoder outputs are uniformly bounded because the input domain $\mathcal{K}$ is compact. Second, every decoder is Λ_ρ-Lipschitz for one finite constant Λ_ρ, because its layer weights are uniformly bounded. Both constants are independent of N.

For completeness, only one property specific to DeepSets is needed. For any two encoders ϕ and $\widetilde{\phi}$ and any set $X = (x_1, \dots, x_N)$,

$$\left\| \frac{1}{N}\sum_{i=1}^N \phi(x_i) - \frac{1}{N}\sum_{i=1}^N \widetilde{\phi}(x_i) \right\|_1 \le \frac{1}{N}\sum_{i=1}^N \|\phi(x_i) - \widetilde{\phi}(x_i)\|_1.$$

Thus, mean pooling does not enlarge the average encoder-covering error. After composition with a Λ_ρ-Lipschitz decoder, an encoder cover of radius $u/[2(1+\Lambda_\rho)]$ and a decoder cover of radius $u/2$ together form an u-cover of the full DeepSets class.

Standard pseudo-dimension and covering-number bounds for bounded ReLU networks (Bartlett et al., 2019; Haussler, 1992) yield N-independent covers for the encoder and decoder classes. Since mean aggregation does not enlarge the encoder covering radius, these component-wise covers combine to give an N-independent cover of the full DeepSets class.

$$\log \mathcal{N}\big(u, \mathcal{H}_N, L^1\big) = \widetilde{\mathcal{O}}(W_\phi L_\phi + W_\rho L_\rho)\,, \tag{23}$$

where the suppressed factors are logarithmic in $1/u$ and in fixed architecture-dependent bounds. Crucially, no factor depends on N.

Step 3: derive a uniform generalization bound. Step 2 shows that, on any fixed sample, the prediction functions in $\mathcal{H}_N$ admit a finite empirical L^1 cover. We now use this result to construct a cover of the corresponding squared-loss functions.

We first bound both the targets and the predictions. Fix any reference measure $\nu_0 \in \mathcal{P}(\mathcal{K})$. By Assumption 1,

$$|F(\mu_X)| \le |F(\nu_0)| + L_F W_1(\mu_X, \nu_0) \le |F(\nu_0)| + L_F \operatorname{diam}(\mathcal{K})$$

for every $X \in \mathcal{K}^N$. The right-hand side is a fixed finite constant; in particular, it does not grow with N. Dividing the targets and predictions by the maximum of this constant and 1, we may work with $|F(\mu_X)| \le 1$. This fixed rescaling affects only the constants hidden in $\widetilde{\mathcal{O}}$.

We then clip every prediction $g(X)$ to $[-1, 1]$. Since the target $F(\mu_X)$ already belongs to this interval, clipping can only move the prediction closer to the target and therefore cannot increase the squared loss. Hence, in the remainder of the proof, both targets and predictions lie in $[-1, 1]$.

For $g \in \mathcal{H}_N$, define the normalized squared loss

$$\ell_g(X) := \frac{1}{4}\big(g(X) - F(\mu_X)\big)^2.$$

The factor $1/4$ ensures that $0 \le \ell_g \le 1$. If two predictors are close, then their losses are equally close. Indeed, the difference-of-squares identity gives

$$\begin{aligned} |\ell_g(X) - \ell_{\widetilde{g}}(X)| &= \frac{1}{4}|g(X) - \widetilde{g}(X)||g(X) + \widetilde{g}(X) - 2F(\mu_X)| \\ &\le \frac{1}{4}|g(X) - \widetilde{g}(X)|\big(|g(X)| + |\widetilde{g}(X)| + 2|F(\mu_X)|\big) \\ &\le \frac{1}{4}|g(X) - \widetilde{g}(X)|\,(1 + 1 + 2) \\ &= |g(X) - \widetilde{g}(X)|, \end{aligned} \tag{24}$$

where we used $g(X), \widetilde{g}(X), F(\mu_X) \in [-1, 1]$. Hence every u-cover of the predictors is also an u-cover of the normalized losses.

A standard uniform-convergence theorem for bounded loss classes (Haussler, 1992), together with the covering bound from Step 2 and rescaling from ℓ_g to the original squared loss, then yields

$$\sup_{g \in \mathcal{H}_N} |\mathcal{L}_N(g) - \widehat{\mathcal{L}}_M(g)| = \widetilde{\mathcal{O}}\left(\sqrt{\frac{W_\phi L_\phi + W_\rho L_\rho + \log(1/\zeta)}{M}}\right).$$

Consequently, if

$$M = \widetilde{\mathcal{O}}\left(\frac{W_\phi L_\phi + W_\rho L_\rho + \log(1/\zeta)}{\varepsilon_r^2}\right). \tag{25}$$

, then, with probability at least $1 - \zeta$, the random training sample satisfies the uniform bound

$$\sup_{g \in \mathcal{H}_N} |\mathcal{L}_N(g) - \widehat{\mathcal{L}}_M(g)| \le \frac{\varepsilon_r}{4}, \tag{26}$$

simultaneously for every predictor $g \in \mathcal{H}_N$.

Step 4: transfer the guarantee to the empirical risk minimizer. Our aim is to bound $\mathcal{L}_N(\widehat{h}_M)$, whereas Step 1 bounds only $\mathcal{L}_N(h)$. We connect these two quantities in three comparisons:

$$\mathcal{L}_N(\widehat{h}_M) \longrightarrow \widehat{\mathcal{L}}_M(\widehat{h}_M) \longrightarrow \widehat{\mathcal{L}}_M(h) \longrightarrow \mathcal{L}_N(h).$$

The first and third comparisons follow from uniform convergence in Step 3. The middle comparison follows from approximate empirical risk minimization, because h is one of the predictors in $\mathcal{H}_N$.

More precisely, because $\widehat{h}_M$ is an $\varepsilon_r/4$-approximate empirical risk minimizer and $h \in \mathcal{H}_N$,

$$\widehat{\mathcal{L}}_M(\widehat{h}_M) \le \widehat{\mathcal{L}}_M(h) + \frac{\varepsilon_r}{4}. \tag{27}$$

Whenever the uniform bound (26) holds, it can be applied first to $\widehat{h}_M$ and then to h:

$$\mathcal{L}_N(\widehat{h}_M) \le \widehat{\mathcal{L}}_M(\widehat{h}_M) + \frac{\varepsilon_r}{4}, \qquad \widehat{\mathcal{L}}_M(h) \le \mathcal{L}_N(h) + \frac{\varepsilon_r}{4}.$$

Combining these two inequalities with (27) gives

$$\mathcal{L}_N(\widehat{h}_M) \le \mathcal{L}_N(h) + \underbrace{\frac{\varepsilon_r}{4}}_{\text{ERM error}} + \underbrace{2\frac{\varepsilon_r}{4}}_{\text{generalization errors}} \le \frac{\varepsilon_r}{4} + \frac{3\varepsilon_r}{4} = \varepsilon_r, \tag{28}$$

where Step 1 gives $\mathcal{L}_N(h) \le \varepsilon_r/4$. Thus, $\mathcal{L}_N(\widehat{h}_M) \le \varepsilon_r$ with probability at least $1 - \zeta$.

Finally, the sample bound (25) depends on m and the encoder and decoder architecture sizes, all of which are independent of N by Proposition C.1. Therefore, the sufficient number of training sets is independent of the input set size N. □

D Auxiliary Proofs

D.1 Measure-Theoretic Identification

For convenience, we restate the proposition.

Proposition 3.2 (Measure-theoretic identification). *A function $f : \mathcal{K}^N \to \mathbb{R}$ is permutation-invariant if and only if there exists a unique functional $F : \mathcal{P}_N(\mathcal{K}) \to \mathbb{R}$ such that $f(\mathbf{x}) = F(\mu_{\mathbf{x}})$ for all $\mathbf{x} \in \mathcal{K}^N$.*

Proof. We prove the two directions separately.

($\Leftarrow$) **Sufficiency.** Suppose $F : \mathcal{P}_N(\mathcal{K}) \to \mathbb{R}$ satisfies $f(\mathbf{x}) = F(\mu_{\mathbf{x}})$ for all $\mathbf{x} \in \mathcal{K}^N$. For any $\pi \in S_N$, the empirical measure of the permuted input $\mathbf{x}^\pi = (x_{\pi(1)}, \dots, x_{\pi(N)})$ satisfies

$$\mu_{\mathbf{x}^\pi} = \frac{1}{N}\sum_{i=1}^{N} \delta_{x_{\pi(i)}} = \frac{1}{N}\sum_{i=1}^{N} \delta_{x_i} = \mu_{\mathbf{x}},$$

where the second equality holds because π is a bijection. Therefore $f(\mathbf{x}^\pi) = F(\mu_{\mathbf{x}^\pi}) = F(\mu_{\mathbf{x}}) = f(\mathbf{x})$.

($\Rightarrow$) **Existence.** Suppose f is permutation-invariant. For each $\mu \in \mathcal{P}_N(\mathcal{K})$, choose any $\mathbf{x} \in \mathcal{K}^N$ with $\mu_{\mathbf{x}} = \mu$ and set $F(\mu) := f(\mathbf{x})$.

Well-definedness. Suppose $\mathbf{x}$ and $\mathbf{y}$ satisfy $\mu_{\mathbf{x}} = \mu_{\mathbf{y}}$. Since equality of empirical measures implies that every value occurs with the same multiplicity in $\mathbf{x}$ and $\mathbf{y}$, the two tuples differ only by a permutation. Hence $\mathbf{y} = \mathbf{x}^\pi$ for some $\pi \in S_N$, and by permutation invariance $f(\mathbf{y}) = f(\mathbf{x})$. So $F(\mu)$ does not depend on the choice of representative.

Identity. By construction $F(\mu_{\mathbf{x}}) = f(\mathbf{x})$ for all $\mathbf{x}$.

Uniqueness. Any F' satisfying $f(\mathbf{x}) = F'(\mu_{\mathbf{x}})$ for all $\mathbf{x}$ must agree with F on every $\mu \in \mathcal{P}_N(\mathcal{K})$, since every such μ arises as the empirical measure of some $\mathbf{x}$. □

E Simulation Setup

This appendix provides the detailed formulation of the resource-allocation benchmark and the experimental settings used in Sec. 6.

E.1 Resource-Allocation Problem

To evaluate the dependence on set cardinality and element dimension in a controlled setting, we use the resource-allocation problem as the benchmark. Each input element represents the features of one user, while the target is the optimal value of a shared allocation problem across all users. By varying N and D, this formulation allows us to study the scaling behavior predicted by our theory while preserving the same underlying task structure.

Given the input set $\mathbf{x} = \{x_1, \dots, x_N\}$, the target $F_D(\mathbf{x})$ is defined as the optimal objective value

$$F_D(\mathbf{x}) = \max_{\mathbf{p}} \frac{1}{N}\sum_{i=1}^{N}\left[\frac{1}{D}\sum_{d=1}^{D}\log(1 + x_{i,d}p_i)\right] - \frac{\gamma}{2N}\sum_{i=1}^{N} p_i^2$$

subject to

$$p_i \geq 0, \qquad i = 1, \dots, N, \qquad \sum_{i=1}^{N} p_i \leq N.$$

Here, p_i denotes the resource allocated to user i, and throughout all experiments we set $\gamma = 0.05$, which controls the quadratic regularization on the allocation. Since the objective and constraints are shared across all users, F_D is invariant to permutations of the input elements.

Table 3: Settings for the latent-dimension and sample-complexity experiments.

Setting	Latent-dim. scaling	Sample-complexity scaling
Hidden width	128	256
Latent width	$m \in \{2, 4, 8, 16, 32, 64, 128, 256, 512, 1024, 2048\}$	$m = 128$
Learning rate	1.5×10^{-3}	10^{-3}
Training updates	20,000	40,000
Training sets	10,000 per (N, D)	$M \in \{500, 1000, 2000, 4000, 8000, 16000, 32000\}$
Set size	$N \in \{10, 20, 50, 100, 200, 500, 1000\}$	$N \in \{20, 50, 100, 200, 500, 1000\}$
Element dimension	$D \in \{1, 2, 4, 8, 16, 32\}$	$D = 4$

E.2 Data Preparation

We consider a resource-allocation problem with N homogeneous users, where each user is characterized by a feature vector $x_i \in [0,1]^D$. Each input set is sampled according to $x_i \overset{\text{i.i.d.}}{\sim} \mathrm{Unif}([0,1]^D)$, and its label is the optimal objective value of the resource-allocation problem defined in Sec. E.1. Since all users share the same objective and constraints, the target is permutation-invariant with respect to the input elements. We solve the optimization problem for every sampled input set and use DeepSets with mean aggregation in all experiments.

E.3 Network and Training Settings

We use DeepSets with ReLU encoder and decoder networks, mean aggregation, Adam optimization, batch size 128, MSE loss on normalized targets, and gradient clipping with maximum norm 1. For each configuration, we use 4096 validation sets and 8192 test sets, with $x_i \overset{\text{i.i.d.}}{\sim} \mathrm{Unif}([0,1]^D)$. The experiment-specific settings are summarized in Table 3.

For the latent-dimension experiment, the critical width $m^\star$ is the smallest tested m whose median training RMSE over three random seeds is below 2×10^{-4}. If no tested width satisfies this criterion, the configuration is reported as not attained.

For the sample-complexity experiment, the empirical sample requirement $M^\star$ is the smallest tested training budget for which the test RMSE is below $\varepsilon = 2 \times 10^{-4}$. Configurations that do not reach this threshold within the tested budget are reported as not attained.

F Permutation-Aware Maximin Sampling Algorithm

We provide the detailed procedure for PAMS introduced in Sec. 5 and analyze how its geometric coverage affects prediction error. Given an unlabeled candidate pool $\mathcal{U}$ and a training-data budget M, PAMS first selects an initial input and then iteratively expands the training set. At each iteration, for every remaining candidate, we compute its distance to the nearest selected input under the permutation-invariant metric d_{Q}, and select the candidate with the largest minimum distance. Repeating this procedure yields a training set that is progressively spread over the permutation quotient space. We further show in Sec. F.1 that the resulting coverage radius explicitly controls an upper bound on the prediction error. The complete procedure is summarized in Algorithm 1.

F.1 Coverage-Based Prediction Error

We further analyze how the geometric coverage induced by PAMS affects the prediction error. Recall that the permutation-aware distance d_Q coincides with the Wasserstein distance between the corresponding empirical measures, as shown in Eq. (4). Hence, under Wasserstein stability, inputs that are close under d_Q also have similar target values. This motivates us to characterize how well the selected training inputs cover the candidate pool.

Algorithm 1 Permutation-Aware Maximin Sampling

Require: Unlabeled candidate pool $\mathcal{U}$ and training-data budget M
Ensure: Selected training set $\mathcal{S}_M$
1: Select an initial input $\mathbf{x}^{(1)} \in \mathcal{U}$
2: $\mathcal{S}_1 \leftarrow \{\mathbf{x}^{(1)}\}$
3: **for** $j = 1, \ldots, M-1$ **do**
4: **for** each $\mathbf{x} \in \mathcal{U} \setminus \mathcal{S}_j$ **do**
5: $r(\mathbf{x}) \leftarrow \min_{\mathbf{x}' \in \mathcal{S}_j} d_{\mathrm{Q}}(\mathbf{x}, \mathbf{x}')$
6: **end for**
7: $\mathbf{x}^{(j+1)} \in \arg\max_{\mathbf{x} \in \mathcal{U} \setminus \mathcal{S}_j} r(\mathbf{x})$
8: $\mathcal{S}_{j+1} \leftarrow \mathcal{S}_j \cup \{\mathbf{x}^{(j+1)}\}$
9: **end for**
10: **return** $\mathcal{S}_M$

For a selected training set $S \subseteq \mathcal{U}$, we define its coverage radius over the candidate pool $\mathcal{U}$ as

$$r(S,\mathcal{U}) := \max_{\mathbf{x} \in \mathcal{U}} \min_{\mathbf{x}' \in S} d_Q(\mathbf{x}, \mathbf{x}').$$

The coverage radius measures the largest distance from any candidate input to its nearest selected training input. A smaller $r(S,\mathcal{U})$ therefore indicates better coverage of the candidate pool under the permutation-aware distance. In particular, for every $\mathbf{x} \in \mathcal{U}$, there exists an $\mathbf{x}' \in S$ such that $d_Q(\mathbf{x}, \mathbf{x}') \le r(S,\mathcal{U})$.

We next establish the connection between the coverage radius and the prediction error.

Proposition F.1 (Prediction error under geometric coverage)**.** *Let $S \subseteq \mathcal{U}$ be a labeled training set, and let h be a predictor that is L_h-Lipschitz with respect to d_Q, i.e.,*

$$|h(\mathbf{x}) - h(\mathbf{x}')| \le L_h d_Q(\mathbf{x}, \mathbf{x}'), \qquad \mathbf{x}, \mathbf{x}' \in \mathcal{U}.$$

Then its worst-case prediction error over the candidate pool satisfies

$$\max_{\mathbf{x} \in \mathcal{U}} |h(\mathbf{x}) - F(\mu_{\mathbf{x}})| \le \max_{\mathbf{x}' \in S} |h(\mathbf{x}') - F(\mu_{\mathbf{x}'})| + (L_h + L_F) r(S,\mathcal{U}).$$

Proposition F.1 shows that the prediction error over the candidate pool is bounded by two factors. The first term measures the prediction error over the selected training inputs and therefore reflects how accurately the predictor fits the labeled data. The second term measures the additional error associated with extending predictions from the selected inputs to the remaining candidates, and is directly controlled by the coverage radius $r(S,\mathcal{U})$.

This bound provides a theoretical justification for the PAMS selection criterion. At each iteration, PAMS selects the candidate with the largest distance to the current training set, i.e., an input in the currently least-covered region. Adding this input directly targets the worst-covered part of the candidate pool and progressively reduces the coverage radius. Consequently, for predictors with comparable fitting errors on the selected inputs, improved coverage under d_Q leads to a tighter worst-case prediction-error bound.

Proof. For any $\mathbf{x} \in \mathcal{U}$, let

$$\mathbf{x}_S \in \arg\min_{\mathbf{x}' \in S} d_Q(\mathbf{x}, \mathbf{x}')$$

denote a nearest selected training input. We decompose the prediction error as

$$\begin{aligned} |h(\mathbf{x}) - F(\mu_{\mathbf{x}})| \le |h(\mathbf{x}) - h(\mathbf{x}_S)| + |h(\mathbf{x}_S) - F(\mu_{\mathbf{x}_S})| \\ + |F(\mu_{\mathbf{x}_S}) - F(\mu_{\mathbf{x}})|. \end{aligned}$$

Since h is L_h-Lipschitz with respect to d_Q,

$$|h(\mathbf{x}) - h(\mathbf{x}_S)| \le L_h d_Q(\mathbf{x}, \mathbf{x}_S).$$

Moreover, by Eq. (4),

$$|F(\mu_{\mathbf{x}_S}) - F(\mu_{\mathbf{x}})| \le L_F d_Q(\mathbf{x}, \mathbf{x}_S).$$

Combining the above inequalities yields

$$|h(\mathbf{x}) - F(\mu_{\mathbf{x}})| \leq |h(\mathbf{x}_S) - F(\mu_{\mathbf{x}_S})| + (L_h + L_F) d_Q(\mathbf{x}, \mathbf{x}_S).$$

By the definition of the coverage radius,

$$d_Q(\mathbf{x}, \mathbf{x}_S) = \min_{\mathbf{x}' \in S} d_Q(\mathbf{x}, \mathbf{x}') \leq r(S, \mathcal{U}).$$

Hence,

$$|h(\mathbf{x}) - F(\mu_{\mathbf{x}})| \leq \max_{\mathbf{x}' \in S} |h(\mathbf{x}') - F(\mu_{\mathbf{x}'})| + (L_h + L_F) r(S, \mathcal{U}).$$

Taking the maximum over $\mathbf{x} \in \mathcal{U}$ gives

$$\max_{\mathbf{x} \in \mathcal{U}} |h(\mathbf{x}) - F(\mu_{\mathbf{x}})| \leq \max_{\mathbf{x}' \in S} |h(\mathbf{x}') - F(\mu_{\mathbf{x}'})| + (L_h + L_F) r(S, \mathcal{U}),$$

which completes the proof. □

G Understanding Cardinality Independence

In this section, we examine the mechanisms underlying cardinality independence from three complementary perspectives. We connect our result to fixed-size aggregation architectures, clarify the role of Wasserstein stability, and analyze the frequency content of the experimental target.

G.1 Connection to Fixed-Size Aggregation Architectures

Fixed-size aggregation is widely used in set-based learning to process inputs with different cardinalities. For example, DeepSets maps each input element to an m-dimensional feature and then aggregates these features through a symmetric pooling operation (Zaheer et al., 2017). Once m is fixed, the same architecture can therefore process sets with different numbers of elements without changing its representation width. However, this architectural property alone does not imply that a fixed latent dimension is sufficient for accurate approximation as the set size N increases. For general continuous permutation-invariant functions, existing results show that the required latent dimension can still grow with N (Wagstaff et al., 2022; Tabaghi & Wang, 2024).

Theorem 4.1 shows that this cardinality dependence can be removed under Wasserstein stability. Specifically, for an L_F-Wasserstein-stable functional on $[0,1]^D$, a latent dimension determined by D, L_F, and the target accuracy τ is sufficient independently of N. Hence, for the function class considered in this work, a fixed-width representation is not only compatible with variable-size inputs, but is also theoretically sufficient to achieve any prescribed approximation accuracy. This result provides a theoretical justification for using a common representation width across different set sizes in the Wasserstein-stable setting.

We further note that the proposed Soft-Histogram Encoder is structurally related to several practical fixed-size aggregation mechanisms. Learnable histogram layers (Wang et al., 2016) softly assign local features to a fixed collection of bins and aggregate the resulting memberships into histogram statistics. NetVLAD (Arandjelovic et al., 2016) similarly assigns local descriptors to a prescribed set of cluster centers and aggregates assignment-weighted residuals. In both cases, the representation size is determined by the number of bins or centers rather than by the number of input elements.

Our construction follows a related principle. Each input element is softly assigned to a fixed collection of spatial grid centers, and averaging these assignments over the input set yields the masses of a grid-supported probability measure. The key distinction is that the number of grid centers in our construction is not introduced only as an architectural hyperparameter. Instead, it is chosen according to the Wasserstein resolution required for the target accuracy. Theorem 4.1 then gives an explicit sufficient bound on this representation size in terms of D, L_F, and τ. Thus, while the Soft-Histogram Encoder shares the fixed-reference aggregation principle used by learnable histograms and NetVLAD, our analysis further characterizes when such a fixed-size representation is sufficient and how its size can be selected to guarantee a prescribed approximation accuracy.

G.2 The Role of Wasserstein Stability

In this subsection, we examine the role of Wasserstein stability in the cardinality-independent result. Theorem 4.1 shows that the latent dimension can be bounded independently of the set cardinality N, while existing lower bounds for general continuous permutation-invariant functions scale with N. To understand this difference, we first study how Wasserstein stability controls the effect of perturbing the elements of an input set, and then compare this behavior with the function family used in the existing lower-bound construction.

For an empirical measure $\mu_{\mathbf{x}} = \frac{1}{N}\sum_{i=1}^{N}\delta_{x_i}$, each element contributes a mass of $1/N$. Consider two inputs that differ only in the i-th element, x_i and x_i'. Their empirical measures satisfy

$$W_1(\mu_{\mathbf{x}}, \mu_{\mathbf{x}'}) \le \frac{1}{N}\|x_i - x_i'\|.$$

Under Assumption 1, the corresponding variation of the target functional is therefore bounded by

$$|F(\mu_{\mathbf{x}}) - F(\mu_{\mathbf{x}'})| \le \frac{L_F}{N}\|x_i - x_i'\|.$$

Hence, when L_F is independent of N, the approximation problem is governed by a fixed Wasserstein resolution. To achieve output error at most τ, it is sufficient to approximate the empirical measure within Wasserstein distance of order τ/L_F. Such an approximation can be constructed using a fixed spatial discretization of $\mathcal{K}$: each input element transfers its mass to nearby grid points, while the aggregated masses form a probability measure supported on the same finite grid. The number of grid points is determined by the element dimension D and the required Wasserstein resolution, but not by the number of input elements. Increasing N therefore changes only the masses assigned to these grid points, rather than the dimension of the representation. This is the mechanism that allows the latent dimension to remain independent of N.

The above mechanism does not extend to general continuous permutation-invariant functions. To see this, consider the lower-bound construction in (Wagstaff et al., 2022). For scalar inputs ($D = 1$), they consider the alternating order-statistic functional

$$g_N(\mathbf{x}) = \sum_{i=1}^{N}(-1)^{i+1}x_{(i)},$$

where $x_{(1)} \ge \cdots \ge x_{(N)}$. Using the one-dimensional identity

$$W_1(\mu_{\mathbf{x}}, \mu_{\mathbf{y}}) = \frac{1}{N}\sum_{i=1}^{N}|x_{(i)} - y_{(i)}|,$$

we obtain

$$|g_N(\mathbf{x}) - g_N(\mathbf{y})| \le N\, W_1(\mu_{\mathbf{x}}, \mu_{\mathbf{y}}).$$

Moreover, the bound is attained for $\mathbf{x} = (0, \ldots, 0)$ and $\mathbf{y} = (\delta, 0, \ldots, 0)$, and hence $\mathrm{Lip}_{W_1}(g_N) = N$. Thus, the Wasserstein sensitivity of the lower-bound family grows with the set cardinality. To maintain a fixed output accuracy, the required Wasserstein resolution must therefore decrease as $1/N$, and the spatial discretization must become correspondingly finer as N increases. This is precisely the behavior excluded by Assumption 1, where L_F is uniform in N. The existing $\Omega(N)$ lower bound and Theorem 4.1 are therefore consistent: the former applies to general permutation-invariant functions with cardinality-dependent sensitivity, whereas the latter applies to uniformly Wasserstein-stable functionals.

G.3 Frequency-Domain Analysis

We analyze the frequency content of the resource-allocation target using one-dimensional slices of its input space. For each slice, we estimate the empirical Fourier spectrum and determine the number of low-frequency modes required to capture most of its spectral power. We then relate this spectral concentration to the random Fourier feature construction of (Zhou et al., 2026). Because RFF frequencies are sampled from a prescribed distribution rather than selected according to the target spectrum, some features may fall outside the target's dominant frequency range. Such features contribute less effectively to the approximation, potentially increasing the latent dimension required to achieve a given accuracy.

Experimental setting. We consider the scalar-input instance of the resource-allocation problem introduced in Sec. 6, with $D = 1$, $N = 10$, and $\gamma = 0.05$. The user features are independently sampled from $\mathrm{Unif}[0, 1]$. For each of $R = 100$ independently sampled configurations

$$c^{(r)} = \big(x_2^{(r)}, \ldots, x_N^{(r)}\big), \qquad r = 1, \ldots, R,$$

we fix $N - 1$ user features and vary the remaining one over $[0, 1]$, yielding the one-dimensional target slice

$$g_r(t) := F_1\big(t, x_2^{(r)}, \ldots, x_N^{(r)}\big), \qquad t \in [0, 1].$$

By permutation invariance, the choice of the varying element does not affect the analysis. Each slice is evaluated at $t_j = j/T$, $j = 0, \ldots, T - 1$, with $T = 1024$, so that the sampled grid lies in $[0, 1)$ and avoids duplicating the endpoint in the discrete Fourier transform. We use random seed 2023.

Target frequency spectrum. For each slice, we subtract its discrete mean and compute the discrete Fourier transform

$$\widehat{g}_r(k) = \sum_{j=0}^{T-1} \left(g_r(t_j) - \frac{1}{T} \sum_{q=0}^{T-1} g_r(t_q) \right) \exp\!\left(-\frac{2\pi \mathrm{i} jk}{T} \right).$$

We exclude the zero-frequency component and normalize the spectrum to compare slices with different amplitudes. Specifically, we define

$$a_r(k) := \frac{|\widehat{g}_r(k)|}{\sum_{\ell=1}^{\lfloor T/2 \rfloor} |\widehat{g}_r(\ell)|}, \qquad p_r(k) := \frac{|\widehat{g}_r(k)|^2}{\sum_{\ell=1}^{\lfloor T/2 \rfloor} |\widehat{g}_r(\ell)|^2}.$$

To quantify the concentration of spectral power, we define

$$K_q := \min \left\{ K : \sum_{k=1}^{K} \frac{1}{R} \sum_{r=1}^{R} p_r(k) \geq q \right\}.$$

Thus, K_q is the smallest number of nonzero Fourier modes accounting for a fraction q of the mean normalized spectral power across the sampled configurations. For each individual configuration, we also define

$$K_q^{(r)} := \min \left\{ K : \sum_{k=1}^{K} p_r(k) \geq q \right\},$$

which measures the corresponding spectral cutoff for the r-th slice. Figure 5 reports the median normalized magnitude $a_r(k)$ across the $R = 100$ configurations and highlights the frequency range corresponding to the 95% spectral-power cutoff.

Target frequency concentration. As shown in Figure 5, the resource-allocation target is dominated by low-frequency components. The first 5, 9, and 45 nonzero Fourier modes account for 90%, 95%, and 99% of the mean normalized spectral power, respectively. This concentration is consistent across different configurations, with the configuration-wise $K_{0.95}^{(r)}$ having median 9 and interquartile range $[9, 9]$ across configurations. Repeating the analysis with $T = 2048$ over ten independently sampled configurations gives the same spectral cutoffs.

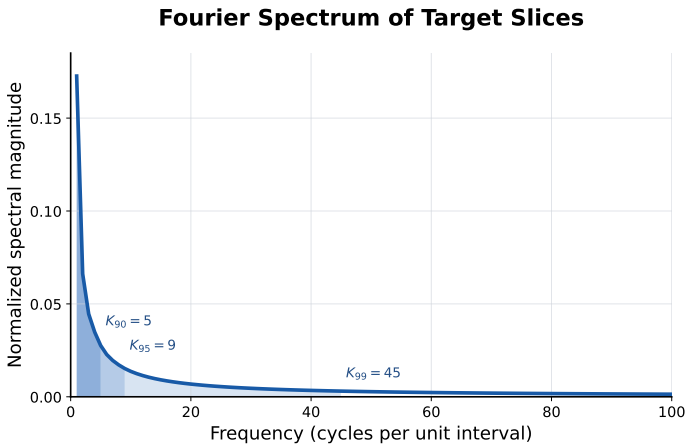

Figure 5: Empirical Fourier spectrum of one-dimensional slices of the resource-allocation target for $D = 1$ and $N = 10$. The solid curve shows the median normalized spectral magnitude across 100 independently sampled configurations. The shaded frequency ranges mark the spectral-power cutoffs $K_{0.90} = 5$, $K_{0.95} = 9$, and $K_{0.99} = 45$: the first K_q nonzero Fourier modes account for at least a fraction q of the mean normalized spectral power. The concentration of spectral power in the lowest modes indicates that the target slices are predominantly low-frequency.

The observed spectral concentration offers a possible explanation for why the RFF representation in (Zhou et al., 2026) can be redundant for this target. Its frequencies are sampled from a fixed distribution without adapting to the spectrum of the target. When most of the target's spectral power is concentrated in a narrow low-frequency range, samples outside that range may contribute comparatively little to reducing the approximation error. Increasing the number of RFF features may therefore add redundant coordinates rather than yield a proportional improvement in approximation accuracy. This interpretation is specific to the target studied here and is not intended as a general limitation of RFF representations.