EXPONENTIAL SEPARATION BETWEEN LINEAR TRANSFORMERS AND JANOSSY POOLING NETWORKS FROM A FUNCTIONAL PERSPECTIVE

Anonymous authors
Paper under double-blind review

ABSTRACT

Permutation-invariant architectures, including Janossy pooling networks and linear Transformers, are ubiquitous in modern deep learning, yet their relative theoretical strengths remain insufficiently understood. Through the approximation of statistical functionals, we systematically compare these architectures and establish an exponential separation in representational efficiency favoring linear Transformers. We establish their universal approximation property (UAP) for continuous functionals of probability measures. Despite this universality, we construct a normalized hard functional involving two-stage feature extraction, exactly representable by a shallow linear Transformer with a scalar pooled representation. Nevertheless, DeepSets require a latent dimension exponential in the input dimension for fixed nontrivial L^2 accuracy under a suitable input-measure distribution. We further extend this separation to r-ary Janossy pooling networks under the uniform norm. These lower bounds hold for arbitrary continuous real-valued feature maps and readouts, regardless of their expressivity. The separation stems from a structural distinction: linear Transformers use multistage information aggregation to implement input-distribution-dependent feature maps, whereas Janossy pooling employs fixed feature maps. Our results expose an intrinsic information bottleneck of single-stage fixed-arity pooling that UAP alone does not capture.

1 INTRODUCTION

Permutation invariance is a fundamental inductive bias in diverse areas of modern deep learning, including computer vision (Ilse et al., 2018; Yang et al., 2020), scientific machine learning (Gilmer et al., 2017; Schütt et al., 2018), and geometric learning (Bronstein et al., 2021; Wang et al., 2019). It ensures the output remains unchanged under any permutation of the input elements, naturally corresponding to learning functions of unordered sets, with broad applications in particle physics (Komiske et al., 2019; Qu et al., 2022), population statistics (Zaheer et al., 2017; Radev et al., 2020), and point clouds (Qi et al., 2017; Zhao et al., 2021). More recently, in-context learning (Fang et al., 2025) and operator learning (Prasthofer et al., 2022; Chiu et al., 2025) have also been formulated within the permutation-invariance framework, with feature extraction from unordered prompt examples and sampled points, respectively. Moreover, from a more general measure-theoretic perspective, such permutation-invariant functions can be studied as statistical functionals on probability measures (Zweig & Bruna, 2021; Furuya et al., 2025).

A foundational approach to capturing permutation invariance is based on Janossy pooling (Murphy et al., 2019), with representative architectures including DeepSets (Zaheer et al., 2017) and relational networks (Santoro et al., 2017), which aggregate features extracted from individual elements and pairs, respectively. This pooling principle also underlies graph neural networks through the aggregation of neighboring node features (Morris et al., 2019; Xu et al., 2019; Corso et al., 2020). More recently, the emergence of attention mechanisms (Vaswani et al., 2017) has motivated alternative permutation-invariant architectures such as Set Transformers (Lee et al., 2019), which omit positional encodings and replace conventional pooling with attention-based aggregation.

The aforementioned architectures have achieved strong predictive performance across diverse tasks (Qi et al., 2017; Horn et al., 2020; Jaegle et al., 2021; Zhang et al., 2022), with theoretical

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.

guarantees as universal approximators of continuous permutation-invariant functions (Lee et al., 2019; Wagstaff et al., 2022). Nevertheless, empirical studies report superior performance of Transformers over conventional pooling architectures (Guo et al., 2021; Qu et al., 2022; Wu et al., 2024), suggesting inherent advantages of attention-based architectures. However, a quantitative theoretical understanding of these relative advantages remains limited. An important step toward this understanding is taken by Zweig & Bruna (2022): treating relational networks as a pairwise-interaction surrogate for attention and establishing an exponential latent-dimension separation from DeepSets in approximating permutation-invariant functions. However, the scope of this result is constrained by the overly strict assumption of complex-analytic feature maps and readouts. In view of this limitation, Zweig & Bruna (2022) poses the following open question:

Does the exponential latent-dimension lower bound in sequence length N and input dimension d persist for DeepSets under the weaker assumption of real-valued continuity?

Although the authors conjecture an affirmative answer, Wang et al. (2024) subsequently shows that DeepSets remain effective universal approximators with latent dimension $\mathrm{poly}(N, d)$ under real-valued continuity assumptions, yielding a negative answer to this question. In this work, we establish an affirmative counterpart for distribution-valued inputs: an exponential latent-dimension lower bound in d persists even for r-ary Janossy pooling networks when approximating statistical functionals under the weaker continuity assumptions. Our contributions are summarized as follows:

- We prove that Janossy pooling networks and shallow linear Transformers with a single head universally approximate continuous statistical functionals, complementing prior UAP results for deep Transformers (Furuya et al., 2025).
- We construct a normalized hard functional with two-stage feature extraction. A linear Transformer implements these stages through attention-mediated global interactions followed by nonlinear feature extraction, thereby inducing input-distribution-dependent features and exactly representing the target with latent dimension $M = 1$. In contrast, DeepSets with single-stage pooling require M exponential in d to retain sufficient information for accurate approximation.
- We further extend this exponential lower bound to r-ary Janossy pooling networks, revealing an intrinsic limitation of single-stage fixed-arity pooling: multistage information cannot be efficiently captured by merely increasing the number of interacting elements or the expressivity of the feature maps and readouts.
- We introduce spherical harmonic analysis to quantify the information captured by single-stage pooling from input distributions and employ the Borsuk–Ulam theorem to identify distinct inputs with identical pooled representations. These techniques may be of independent interest.

2 Related Work

Expressivity of invariant networks. A seminal class of permutation-invariant architectures is constructed via Janossy pooling (Murphy et al., 2019), encompassing PointNet (Qi et al., 2017), DeepSets (Zaheer et al., 2017), and relational networks (Santoro et al., 2017). Their expressivity has been extensively investigated, with particular emphasis on characterizing the latent dimensionality necessary to guarantee UAP for permutation-invariant functions (Maron et al., 2019; Wagstaff et al., 2019; 2022; Wang et al., 2024; Dym & Gortler, 2025; Syed et al., 2026). A complementary line of work employs attention mechanisms to develop invariant architectures, including SINet (Ma et al., 2018), Set Transformer (Lee et al., 2019), and AttSets (Yang et al., 2020), with related expressivity analyses (Yun et al., 2020; Alberti et al., 2023; Kajitsuka & Sato, 2024; Cheng et al., 2026; Yu et al., 2026). Further developments establish quantitative approximation rates (Jiang & Li, 2024; Takeshita & Imaizumi, 2025; Siegel et al., 2026) and characterize the expressivity of equivariant networks (Puny et al., 2023; Pacini et al., 2025; 2026).

Statistical functionals. Statistical functionals provide a classical framework for representing population quantities and analyzing the asymptotic behavior of estimators (Mises, 1947). In statistical learning theory, this perspective underlies learning from distribution-valued inputs, notably through kernel mean embeddings (Smola et al., 2007; Sriperumbudur et al., 2010) and distribution regression (Póczos et al., 2013; Szabó et al., 2016). More recently, measure-theoretic formulations have

connected this perspective to in-context learning (Furuya et al., 2025) and, through graph-measure representations, operator learning (Furuya et al., 2026). Within this framework, existing work establishes approximation and statistical learning guarantees for neural networks (Zweig & Bruna, 2021; Shi et al., 2025), as well as UAP for deep Transformers (Furuya et al., 2025; Geshkovski et al., 2026). In contrast, we establish that Janossy pooling networks and shallow linear Transformers suffice to achieve UAP for continuous statistical functionals.

Architecture separations. Expressivity gaps between neural architectures have been extensively studied (Martens et al., 2013). Foundational depth-separation results distinguish the representation power of two- and three-layer networks (Eldan & Shamir, 2016; Daniely, 2017), with extensions to broader families of radial (Safran & Shamir, 2017) and non-radial functions (Venturi et al., 2022). Complementary results establish separations in growing-depth regimes (Telgarsky, 2016) and norm-bounded infinite-width networks (Parkinson et al., 2024). For sequence inputs, task-specific expressivity separations between Transformers and recurrent or feedforward networks have also been established through sparse averaging, retrieval, and formal-language recognition (Sanford et al., 2023; Bhattamishra et al., 2024; Wen et al., 2025). For symmetric architectures, Zweig & Bruna (2022) establish an exponential separation between DeepSets and relational networks with complex-analytic activations. In a complementary direction, our statistical-functional framework yields an exponential latent-dimension separation between linear Transformers and fixed-arity Janossy pooling in a more realistic real-valued setting with arbitrary continuous feature maps and readouts.

3 Setup and main results

Notation. We represent an input sequence of fixed length N by $\boldsymbol{X} = [\boldsymbol{x}_1, \dots, \boldsymbol{x}_N] \in \mathbb{R}^{d\times N}$, with $\boldsymbol{x}_i \in \Omega \subset \mathbb{R}^d$. Throughout, Ω is compact. For $r \geq 1$, write $\Omega^{\otimes r} \coloneqq \Omega \times \cdots \times \Omega$ for its r-fold Cartesian product. We denote by $C(\Omega_1; \Omega_2)$ the space of continuous maps from Ω_1 to Ω_2, abbreviated as $C(\Omega_1)$ when $\Omega_2 = \mathbb{R}$. Let S_N denote the symmetric group on N elements, acting on $\boldsymbol{X}$ by $\tau(\boldsymbol{X}) = [\boldsymbol{x}_{\tau(1)}, \dots, \boldsymbol{x}_{\tau(N)}]$ for $\tau \in S_N$. We write $\mathbb{S}^{d-1}$ for the unit sphere in $\mathbb{R}^d$, ν_d for its uniform probability measure, and $\mathcal{P}(\Omega)$ for the space of Borel probability measures on Ω. For nonnegative functions f and g, we write $f = O(g)$ if $f \leq Cg$ for some constant $C > 0$, and $f = \Theta(g)$ if both $f = O(g)$ and $g = O(f)$. We write $f = \mathrm{poly}(N, d)$ if $f = O(p(N, d))$ for a fixed polynomial p in N and d.

3.1 Permutation invariant architectures

A function $f : \mathbb{R}^{d\times N} \to \mathbb{R}$ is *permutation invariant* if $f(\tau(\boldsymbol{X})) = f(\boldsymbol{X})$ for every $\tau \in S_N$. A popular class of architectures to capture permutation invariance can be constructed via Janossy pooling. An r-ary Janossy pooling network takes the form

$$g_r(\boldsymbol{X}) = \phi_2\left(\frac{1}{N^r}\sum_{1\leq i_1,\dots,i_r\leq N}\phi_1(\boldsymbol{x}_{i_1},\dots,\boldsymbol{x}_{i_r})\right),$$

where $\phi_1 : \mathbb{R}^{d\times r} \to \mathbb{R}^M$ and $\phi_2 : \mathbb{R}^M \to \mathbb{R}$ are continuous functions, typically parameterized by multilayer perceptrons (MLPs). The encoder ϕ_1 maps each ordered r-tuple to an M-dimensional representation, which is averaged over all tuples and subsequently mapped to the output by ϕ_2. The cases $r = 1$ and $r = 2$ recover DeepSets and relational networks, respectively. The factor N^{-r} is a normalization factor and does not affect expressivity. Suppose that $\boldsymbol{x}_i \overset{\text{i.i.d.}}{\sim} \rho$ for some $\rho \in \mathcal{P}(\Omega)$. In the infinite-sequence limit $N \to \infty$, the Janossy pooling network admits the following functional form on probability measures (De Bie et al., 2019; Shi et al., 2025):

$$g_r(\rho) = \phi_2\left(\int \phi_1(\boldsymbol{x}_1, \cdots, \boldsymbol{x}_r)\mathrm{d}\rho(\boldsymbol{x}_1)\cdots\mathrm{d}\rho(\boldsymbol{x}_r)\right).$$

Self-attention provides a complementary mechanism for constructing permutation-invariant architectures. We consider a single-layer, single-head Transformer equipped with linear attention (Katharopoulos et al., 2020; Ahn et al., 2024):

$$T(\boldsymbol{X}) = \varphi_2\left(\frac{1}{N}\sum_{i=1}^{N}\varphi_1\left(\mathrm{Attn}_i(\boldsymbol{X})\right)\right), \quad \mathrm{Attn}_i(\boldsymbol{X}) = \frac{1}{N}\sum_{j=1}^{N}\langle\mathcal{A}_Q(\boldsymbol{x}_i), \mathcal{A}_K(\boldsymbol{x}_j)\rangle\,\mathcal{A}_V(\boldsymbol{x}_j),$$

where $\mathcal{A}_Q, \mathcal{A}_K : \mathbb{R}^d \to \mathbb{R}^{d_k}$ and $\mathcal{A}_V : \mathbb{R}^d \to \mathbb{R}^{d_v}$ are affine maps, while $\varphi_1 : \mathbb{R}^{d_v} \to \mathbb{R}^M$ and $\varphi_2 : \mathbb{R}^M \to \mathbb{R}$ are continuous functions. Linear attention contextualizes each input element by aggregating value features weighted by pairwise query–key interactions. Subsequently, φ_1 extracts features from the contextualized elements, and φ_2 maps their pooled representation to the output. The factors N^{-1} normalize contextualization and feature-pooling, respectively. The corresponding probability functional as $N \to \infty$ is

$$T(\rho) = \varphi_2\left(\int \varphi_1\left(\mathrm{Attn}_\rho(\boldsymbol{x})\right) \mathrm{d}\rho(\boldsymbol{x})\right), \quad \mathrm{Attn}_\rho(\boldsymbol{x}) = \int \langle \mathcal{A}_Q(\boldsymbol{x}), \mathcal{A}_K(\boldsymbol{y})\rangle \, \mathcal{A}_V(\boldsymbol{y}) \mathrm{d}\rho(\boldsymbol{y}).$$

For fixed integers $r, M \geq 1$, let $\mathcal{G}_M^{(r)}$ and $\mathcal{T}_M$ denote the classes of probability functionals on $\mathcal{P}(\Omega)$ realized by r-ary Janossy pooling networks and linear Transformers with latent dimension M, respectively:

$$\mathcal{G}_M^{(r)} \coloneqq \left\{\rho \mapsto g_r(\rho) \;\middle|\; \phi_1 \in C(\Omega^{\otimes r}; \mathbb{R}^M),\ \phi_2 \in C(\mathbb{R}^M)\right\},$$

and

$$\mathcal{T}_M \coloneqq \left\{\rho \mapsto T(\rho) \;\middle|\; \begin{array}{ll} d_k, d_v \geq 1, & \mathcal{A}_Q, \mathcal{A}_K, \mathcal{A}_V \text{ affine}, \\ \varphi_1 \in C(\mathbb{R}^{d_v}; \mathbb{R}^M), & \varphi_2 \in C(\mathbb{R}^M) \end{array}\right\}.$$

Here, d_k, d_v range over positive integers, and the affine maps have the domains and codomains specified above. We further define $\mathcal{G}^{(r)} \coloneqq \bigcup_{M\geq 1} \mathcal{G}_M^{(r)}$ and $\mathcal{T} \coloneqq \bigcup_{M\geq 1} \mathcal{T}_M$ as the full model classes with arbitrary finite latent dimensions.

3.2 Main results

Our main results establish the UAP of both $\mathcal{G}^{(r)}$ and $\mathcal{T}$ for continuous statistical functionals (functionals of probability measures). Nevertheless, we construct a hard functional that admits uniform approximation by $\mathcal{T}_M$ with $M = 1$, whereas, for any fixed r, $\mathcal{G}_M^{(r)}$ incurs a nonvanishing approximation error even with exponentially large latent dimension.

UAP for statistical functionals. Let $C(\mathcal{P}(\Omega))$ denote the space of continuous functionals on $\mathcal{P}(\Omega)$ endowed with the weak-* topology.[1] For compact Ω, Pevny & Kovarík (2019) established the UAP of $\mathcal{G}^{(1)}$. The inclusion $\mathcal{G}^{(1)} \subseteq \mathcal{G}^{(r)}$ for $r \geq 2$ yields the following extension.

Proposition 3.1 (UAP by Janossy pooling)**.** *For every $r \geq 1$, $\mathcal{G}^{(r)}$ is uniformly dense in $C(\mathcal{P}(\Omega))$.*

For linear Transformers, the proof combines a continuous, injective transformation of probability measures realized by linear attention with a point-separating algebra generated by pooled token-wise features; the generalized Stone–Weierstrass theorem (Stone, 1948) then yields the UAP. See Appendix A for the detailed proofs.

Proposition 3.2 (UAP by linear Transformers)**.** *$\mathcal{T}$ is uniformly dense in $C(\mathcal{P}(\Omega))$.*

Remark 3.3. In fact, the UAP remains valid when the feature and readout classes are universal approximators of continuous functions on arbitrary compact sets, including commonly used shallow ReLU networks and reproducing kernel Hilbert spaces associated with universal kernels (Micchelli et al., 2006). The reason is that uniform continuity on compact sets controls approximation errors under composition, while integration against probability measures preserves uniform error bounds.

Remark 3.4. Statistical functionals provide a measure-theoretic perspective on permutation-invariant functions (Zweig & Bruna, 2021). Every continuous N-variable permutation-invariant function f admits a representation $f(\boldsymbol{X}) = \tilde{f}(\hat{\rho}_{\boldsymbol{X}})$ for some statistical functional $\tilde{f} \in C(\mathcal{P}(\Omega))$, where $\hat{\rho}_{\boldsymbol{X}} \coloneqq N^{-1}\sum_{i=1}^N \delta_{\boldsymbol{x}_i}$. Likewise, evaluating the measure-input models at $\hat{\rho}_{\boldsymbol{X}}$ exactly recovers their sequence-input counterparts. Hence, Propositions 3.1 and 3.2 recover UAPs of continuous permutation-invariant functions established for DeepSets (Zaheer et al., 2017), relational networks (Santoro et al., 2017), and one-layer linear Transformers (Alberti et al., 2023).

[1] For intuition, when Ω is compact, one may equivalently regard $\mathcal{P}(\Omega)$ as equipped with the Wasserstein distance $W_1(\mu,\nu) = \inf_\pi \int_{\Omega\times\Omega} \|\boldsymbol{x} - \boldsymbol{y}\|_1 \, \mathrm{d}\pi(\boldsymbol{x}, \boldsymbol{y})$, which induces the weak-* topology, as shown in Proposition A.1. Here, π ranges over joint probability measures with marginals μ and ν.

Remark 3.5. Complementing the universality results of Furuya et al. (2025) for deep Transformers, we show that one-layer linear Transformers suffice to approximate continuous statistical functionals. Their formulation considers query-dependent targets $f(\rho, \boldsymbol{x})$, with an additional query token $\boldsymbol{x} \in \Omega$ to formulate in-context learning. Our UAP proofs extend naturally to this setting by incorporating $\boldsymbol{x}$ into the feature maps ϕ_1 and φ_1.

Detailed discussions of the preceding remarks are provided in Appendix A.

Exponential separations. We first establish an exponential separation in latent dimension between DeepSets and linear Transformers. For integers $d \geq 2$ and $q \geq 0$, define $L_{d,q} := \binom{d+q-1}{q} - \binom{d+q-3}{q-2}$, with $\binom{n}{k} = 0$ for $k < 0$, and set $q_d := 2\lfloor d/2 \rfloor$.

Theorem 3.6 (Separation between DeepSets and linear Transformers)**.** *Let $d \geq 3$. There exist a functional $f \in C(\mathcal{P}(\mathbb{S}^{d-1}))$ and a random probability measure $\rho_{\boldsymbol{A},\boldsymbol{b}}$ generated by independent random variables $\boldsymbol{A} \sim \nu_{L_{d,q_d}}$ and $\boldsymbol{b} \sim \nu_d$ with joint distribution π_d, such that $\|f(\rho_{\boldsymbol{A},\boldsymbol{b}})\|_{L^2(\pi_d)} = 1$ and:*

- *For $M \leq c_1 4^d/\sqrt{d}$, where $c_1 > 0$ is a constant independent of d,*

$$\inf_{g \in \mathcal{G}_M^{(1)}} \|g(\rho_{\boldsymbol{A},\boldsymbol{b}}) - f(\rho_{\boldsymbol{A},\boldsymbol{b}})\|_{L^2(\pi_d)} \geq \frac{1}{\sqrt{2}}.$$

- *There exists $T \in \mathcal{T}_M$ with $d_k = d, d_v = 1$ and $M = 1$ that exactly represents f, with both φ_1 and φ_2 realized by two-layer ReLU networks of width 2:*

$$f(\rho) = T(\rho), \qquad \rho \in \mathcal{P}(\mathbb{S}^{d-1}).$$

For general r-ary Janossy pooling, the same hard functional yields a slightly weaker but still exponential separation in latent dimension under the uniform norm.

Theorem 3.7 (Lower bound for general Janossy pooling)**.** *Fix $r \geq 1$ and $d \geq 3$, and let f be the same functional given in Theorem 3.6. There exists a constant $c_r > 0$, depending only on r, such that whenever $M \leq c_r 4^d/d^{r-\frac{1}{2}}$,*

$$\inf_{g \in \mathcal{G}_M^{(r)}} \sup_{\rho \in \mathcal{P}(\mathbb{S}^{d-1})} |g(\rho) - f(\rho)| \geq 1.$$

Remark 3.8. The lower bounds in Theorems 3.6 and 3.7 hold for arbitrary continuous feature maps ϕ_1 and output maps ϕ_2, encompassing MLPs of arbitrary depth and hidden width with any continuous activation. Beyond the latent dimension M, continuity is the sole constraint on these components. This generality removes additional architectural restrictions, thereby isolating the distinct aggregation mechanisms of Janossy pooling and linear Transformers as the source of the exponential separation in latent dimension. These results complement exponential separations between permutation-invariant architectures established for neural networks with complex-analytic activations (Zweig & Bruna, 2022).

Remark 3.9. Wang et al. (2024) established that DeepSets with latent dimension $M = \mathrm{poly}(N, d)$ can represent any continuous permutation-invariant function on sequences of fixed length N. However, when extended to more general continuous statistical functionals, our results reveal a fundamental limitation of the single-sum pooling mechanism, manifested by an exponential lower bound on the latent dimension in d, even for functionals exactly representable by linear Transformers with $M = 1$.

We provide proof sketches for Theorems 3.6 and 3.7 in Section 4, with complete proofs deferred to Appendix B.

4 Proof Sketches

We consider the following hard functional for $\rho \in \mathcal{P}(\mathbb{S}^{d-1})$

$$\boldsymbol{m}_\rho = \int \boldsymbol{x} \mathrm{d}\rho(\boldsymbol{x}), \qquad h(\rho) = \int |\langle \boldsymbol{x}, \boldsymbol{m}_\rho \rangle| \, \mathrm{d}\rho(\boldsymbol{x}). \tag{1}$$

The functional h is continuous on $\mathcal{P}(\mathbb{S}^{d-1})$, as shown in Proposition B.1. It measures the mean absolute alignment of the input elements with their barycenter.

Evaluating h involves two stages of feature extraction: projection onto the input barycenter, followed by an absolute-value nonlinearity. Consequently, the effective feature map $\boldsymbol{x} \mapsto |\langle \boldsymbol{x}, \boldsymbol{m}_\rho \rangle|$ adapts to the input distribution through $\boldsymbol{m}_\rho$. Linear Transformers provide a simple and efficient implementation of this two-stage information aggregation process: linear attention realizes the adaptive barycentric projection, while φ_1 applies the absolute-value nonlinearity before final pooling. Combining these operations, linear Transformers implement the desired input-distribution-dependent feature map. In contrast, r-ary Janossy pooling employs a fixed feature map ϕ_1 that cannot adapt to the barycenter direction. It must therefore retain, within an M-dimensional latent representation, sufficient information to evaluate the target across all possible barycenter directions. We employ spherical harmonic analysis to characterize the target's dependence on the barycenter direction and quantify the information that Janossy pooling must retain for accurate approximation. We begin with the requisite spherical-harmonic preliminaries.

4.1 Elementary spherical harmonic analysis

Spherical harmonics serve as Fourier modes on $\mathbb{S}^{d-1}$. A spherical harmonic of degree q is the restriction to $\mathbb{S}^{d-1}$ of a real homogeneous harmonic polynomial $p : \mathbb{R}^d \to \mathbb{R}$, satisfying

$$p(t\boldsymbol{x}) = t^q p(\boldsymbol{x}), \qquad \Delta_d p = 0,$$

where $\Delta_d = \sum_{i=1}^{d} \frac{\partial^2}{\partial x_i^2}$ is the Euclidean Laplacian. Let $\mathcal{H}_{d,q}$ denote the space of spherical harmonics of degree q, and let $\mathscr{P}_{\le s}(\mathbb{S}^{d-1})$ denote the restrictions to $\mathbb{S}^{d-1}$ of real polynomials of total degree at most s. The following lemma collects the properties used in our analysis.

Lemma 4.1 (Spherical harmonic orthogonality and the cosine transform)**.** *Let $d \ge 2$. The following properties hold.*

(i) *For every integer $q \ge 0$, $\dim(\mathcal{H}_{d,q}) = L_{d,q}$. Each $\mathcal{H}_{d,q}$ admits a real orthonormal basis $\{Y_{d,q,\ell}\}_{\ell=1}^{L_{d,q}}$ in $L^2(\nu_d)$, with*

$$\langle Y_{d,q,\ell}, Y_{d,q',\ell'} \rangle_{L^2(\nu_d)} = \delta_{q,q'}\delta_{\ell,\ell'},$$

for all admissible indices.

(ii) *Define $\boldsymbol{Y}_{d,q}(\boldsymbol{x}) \coloneqq \big(Y_{d,q,1}(\boldsymbol{x}), \cdots, Y_{d,q,L_{d,q}}(\boldsymbol{x})\big)^\top$. Then*

$$\|\boldsymbol{Y}_{d,q}(\boldsymbol{x})\|_2^2 = L_{d,q} \quad \text{for every } \boldsymbol{x} \in \mathbb{S}^{d-1}, \qquad \int_{\mathbb{S}^{d-1}} \boldsymbol{Y}_{d,q}(\boldsymbol{x})\boldsymbol{Y}_{d,q}(\boldsymbol{x})^\top \, \mathrm{d}\nu_d(\boldsymbol{x}) = \boldsymbol{I}_{L_{d,q}}.$$

(iii) *For $f \in L^2(\nu_d)$, the spherical cosine transform is defined as*

$$(\mathcal{C}f)(\boldsymbol{b}) \coloneqq \int_{\mathbb{S}^{d-1}} |\langle \boldsymbol{b}, \boldsymbol{x} \rangle| f(\boldsymbol{x}) \, \mathrm{d}\nu_d(\boldsymbol{x}), \qquad \boldsymbol{b} \in \mathbb{S}^{d-1}.$$

For every $Y \in \mathcal{H}_{d,q}$,

$$(\mathcal{C}Y)(\boldsymbol{b}) = \gamma_{d,q} Y(\boldsymbol{b}), \qquad \boldsymbol{b} \in \mathbb{S}^{d-1}, \tag{2}$$

where $\gamma_{d,q}$ is the eigenvalue depending on d and q. Moreover, $\gamma_{d,q} = 0$ for odd q, whereas $\gamma_{d,q} \neq 0$ for every even $q \ge 2$.

(iv) *The polynomial restriction space admits the orthogonal decomposition $\mathscr{P}_{\le s}(\mathbb{S}^{d-1}) = \bigoplus_{q=0}^{s} \mathcal{H}_{d,q}$, with dimension*

$$K_{d,s} \coloneqq \sum_{q=0}^{s} L_{d,q} = \binom{d+s}{s} - \binom{d+s-2}{s-2}.$$

The identity in equation 2 shows that the spherical cosine transform acts diagonally on the harmonic decomposition, scaling each degree-q component by $\gamma_{d,q}$. Since the hard functional in equation 1 is defined using a similar absolute alignment operation, spherical harmonics provide a natural basis for its analysis. These properties and further background on spherical harmonic analysis can be found in Atkinson & Han (2012); Ólafsson et al. (2012).

4.2 Proof sketch of Theorem 3.6

We first outline the lower bound for DeepSets in several steps. Complete proofs of all intermediate results are provided in Appendix B.1.

Step 1: Normalization of the hard functional. We first construct a random probability measure on $\mathcal{P}(\mathbb{S}^{d-1})$ and normalize the hard functional under the resulting distribution. Fix an even integer $q \geq 2$ and $\eta \in (0, \frac{1}{2})$. We suppress the dependence on η throughout for brevity. For $\boldsymbol{A} \in \mathbb{S}^{L_{d,q}-1}$ and $\boldsymbol{b} \in \mathbb{S}^{d-1}$, define

$$\frac{\mathrm{d}\rho_{\boldsymbol{A},\boldsymbol{b}}}{\mathrm{d}\nu_d}(\boldsymbol{x}) \coloneqq 1 + \eta \langle \boldsymbol{b}, \boldsymbol{x} \rangle + \eta H_{\boldsymbol{A}}(\boldsymbol{x}), \qquad H_{\boldsymbol{A}}(\boldsymbol{x}) \coloneqq \frac{1}{\sqrt{L_{d,q}}} \langle \boldsymbol{A}, \boldsymbol{Y}_{d,q}(\boldsymbol{x}) \rangle . \tag{3}$$

By the spherical harmonic properties in Lemma 4.1, $\rho_{\boldsymbol{A},\boldsymbol{b}} \in \mathcal{P}(\mathbb{S}^{d-1})$, as established in Proposition B.2. Let $\boldsymbol{A} \sim \nu_{L_{d,q}}$ and $\boldsymbol{b} \sim \nu_d$ be independent, with joint distribution denoted by $\pi_{d,q}$. Then $\rho_{\boldsymbol{A},\boldsymbol{b}}$ defines a measure-valued random variable over $\mathcal{P}(\mathbb{S}^{d-1})$.

We normalize the hard functional h in equation 1 as follows.

$$f_{d,q}(\rho) = c_{d,q} \left(h(\rho) - \beta_d \right), \qquad \beta_d \coloneqq \frac{\eta}{d} \frac{\Gamma(d/2)}{\sqrt{\pi}\Gamma((d+1)/2)}, \qquad c_{d,q} \coloneqq \frac{d\sqrt{L_{d,q}}}{\eta^2 \gamma_{d,q}}. \tag{4}$$

By the continuity of h, we have $f_{d,q} \in C(\mathcal{P}(\mathbb{S}^{d-1}))$. The following lemma characterizes the normalized functional under the constructed random measures.

Lemma 4.2 (Harmonic encoding)**.** *For $\rho_{\boldsymbol{A},\boldsymbol{b}}$ defined in equation 3,*

$$\boldsymbol{m}_{\rho_{\boldsymbol{A},\boldsymbol{b}}} = \frac{\eta}{d}\boldsymbol{b}, \qquad f_{d,q}(\rho_{\boldsymbol{A},\boldsymbol{b}}) = \langle \boldsymbol{A}, \boldsymbol{Y}_{d,q}(\boldsymbol{b}) \rangle .$$

Moreover, $\| f_{d,q}(\rho_{\boldsymbol{A},\boldsymbol{b}}) \|_{L^2(\pi_{d,q})} = 1$.

The degree-one perturbation $\eta\langle \boldsymbol{b}, \boldsymbol{x} \rangle$ in equation 3 stores the directional information of $\boldsymbol{b}$ in the barycenter, while the orthogonal degree-q perturbation $\eta H_{\boldsymbol{A}}(\boldsymbol{x})$ stores the $L_{d,q}$ coefficients specified by $\boldsymbol{A}$. The hard functional $f_{d,q}$ then encodes this information as the linear projection of $\boldsymbol{A}$ onto the spherical-harmonic feature vector $\boldsymbol{Y}_{d,q}(\boldsymbol{b})$ evaluated at $\boldsymbol{b}$. The hard functional used in Theorems 3.6 and 3.7 is the normalized functional f_{d,q_d} with $q_d = 2\lfloor d/2 \rfloor$.

Remark 4.3. The normalization in equation 4 gives f_{d,q_d} unit $L^2(\pi_d)$ energy, where $\pi_d = \pi_{d,q_d}$. Without it, h is exponentially close to the constant β_d, since $\| h(\rho_{\boldsymbol{A},\boldsymbol{b}}) - \beta_d \|_{L^2(\pi_d)} = |c_{d,q_d}|^{-1} = \Theta(\eta^2 4^{-d} d^{-9/4})$ by Lemma C.4 and equation 4. We therefore focus on the normalized hard functional f_{d,q_d} to obtain a nontrivial approximation problem at fixed accuracy.

Step 2: Optimal predictor with linear representation. For the constructed measures $\rho_{\boldsymbol{A},\boldsymbol{b}}$, the pooled DeepSets representation induced by ϕ_1 depends on $\boldsymbol{A}$ and $\boldsymbol{b}$ only through linear measurements, as formalized below.

Lemma 4.4 (Linear information representation)**.** *For any continuous $\phi_1 : \mathbb{S}^{d-1} \to \mathbb{R}^M$, there exist $\boldsymbol{z}_0 \in \mathbb{R}^M$, $\boldsymbol{U} \in \mathbb{R}^{M \times d}$, and $\boldsymbol{W} \in \mathbb{R}^{M \times L_{d,q}}$ such that*

$$Z(\boldsymbol{A}, \boldsymbol{b}) \coloneqq \int \phi_1(\boldsymbol{x}) \mathrm{d}\rho_{\boldsymbol{A},\boldsymbol{b}}(\boldsymbol{x}) = \boldsymbol{z}_0 + \eta \boldsymbol{U}\boldsymbol{b} + \frac{\eta}{\sqrt{L_{d,q}}} \boldsymbol{W}\boldsymbol{A}.$$

The outer map ϕ_2 forms its prediction solely from $Z(\boldsymbol{A}, \boldsymbol{b})$, which is an affine function of the barycenter direction $\boldsymbol{b}$ and the degree-q harmonic component $\boldsymbol{W}\boldsymbol{A}$. Since $(\boldsymbol{b}, \boldsymbol{W}\boldsymbol{A})$ is at least as informative as $Z(\boldsymbol{A}, \boldsymbol{b})$, it suffices to lower bound the optimal recovery error under the richer observation $(\boldsymbol{b}, \boldsymbol{W}\boldsymbol{A})$, as formalized in the following lemma.

Lemma 4.5 (Projection lemma)**.** *Let $\boldsymbol{A} \sim \nu_{L_{d,q}}$ and $\boldsymbol{b} \sim \nu_d$ be independent, with joint distribution $\pi_{d,q}$. Fix $\boldsymbol{W} \in \mathbb{R}^{M \times L_{d,q}}$, set $\boldsymbol{S} \coloneqq \mathrm{range}(\boldsymbol{W}^\top)$, and let $P_{\boldsymbol{S}}$ be the orthogonal projection onto $\boldsymbol{S}$. Then, almost surely,*

$$\mathbb{E}[\boldsymbol{A} \mid \boldsymbol{W}\boldsymbol{A}] = P_{\boldsymbol{S}}\boldsymbol{A}, \qquad \mathbb{E}[\langle \boldsymbol{A}, \boldsymbol{Y}_{d,q}(\boldsymbol{b}) \rangle \mid \boldsymbol{b}, \boldsymbol{W}\boldsymbol{A}] = \langle P_{\boldsymbol{S}}\boldsymbol{A}, \boldsymbol{Y}_{d,q}(\boldsymbol{b}) \rangle.$$

The latter is the unique mean-square optimal predictor, up to almost-sure equality, with

$$\inf_{g \in \mathcal{H}_{\boldsymbol{W}}} \mathbb{E}_{\boldsymbol{A},\boldsymbol{b}} \left| f_{d,q}(\rho_{\boldsymbol{A},\boldsymbol{b}}) - g(\boldsymbol{b}, \boldsymbol{W}\boldsymbol{A}) \right|^2 = \mathbb{E}_{\boldsymbol{A},\boldsymbol{b}} \left| \langle P_{\boldsymbol{S}^\perp}\boldsymbol{A}, \boldsymbol{Y}_{d,q}(\boldsymbol{b}) \rangle \right|^2,$$

where $\mathcal{H}_{\boldsymbol{W}}$ consists of Borel maps $g : \mathbb{S}^{d-1} \times \mathbb{R}^M \to \mathbb{R}$ satisfying $g(\boldsymbol{b}, \boldsymbol{W}\boldsymbol{A}) \in L^2(\pi_{d,q})$.

Step 3: Estimation of remaining energy. The error of the optimal predictor corresponds to the harmonic energy remaining in $\boldsymbol{S}^{\perp}$, quantified by the following lemma.

Lemma 4.6 (Energy lemma)**.** *Under the setting of Lemma 4.5,*

$$\mathbb{E}_{\boldsymbol{A},\boldsymbol{b}} \left|\langle P_{\boldsymbol{S}^{\perp}} \boldsymbol{A}, \boldsymbol{Y}_{d,q}(\boldsymbol{b})\rangle\right|^2 = 1 - \frac{\operatorname{rank}(\boldsymbol{W})}{L_{d,q}}.$$

The harmonic encoding distributes the unit target energy isotropically across the $L_{d,q}$ harmonic directions. Therefore, the prediction error is determined by the number of directions not contained in $\boldsymbol{S}$, i.e., $L_{d,q} - \operatorname{rank}(\boldsymbol{W})$.

Step 4: Quantitative lower bound for DeepSets. Combining the preceding steps yields an explicit $L^2(\pi_{d,q})$ approximation lower bound for DeepSets with latent dimension M.

Lemma 4.7 (Lower bound for DeepSets)**.** *For every $d \geq 3$, even $q \geq 2$, and $M \geq 1$,*

$$\inf_{g \in \mathcal{G}_M^{(1)}} \|g(\rho_{\boldsymbol{A},\boldsymbol{b}}) - f_{d,q}(\rho_{\boldsymbol{A},\boldsymbol{b}})\|_{L^2(\pi_{d,q})} \geq \sqrt{\left(1 - \frac{M}{L_{d,q}}\right)_+},$$

where $(t)_+ \coloneqq \max\{t, 0\}$.

With $q_d = 2\lfloor d/2 \rfloor$ for $d \geq 3$, we have $M/L_{d,q_d} = O(M\sqrt{d}/4^d)$ by Proposition C.3. Hence, there exists a universal constant $c_1 > 0$ such that $M \leq c_1 4^d/\sqrt{d}$ implies $M/L_{d,q_d} \leq 1/2$. Lemma 4.7 then establishes the lower bound in Theorem 3.6.

Exact representation by linear Transformers. Linear attention realizes the first stage, adaptive barycentric projection, through global interactions: setting $\mathcal{A}_Q(\boldsymbol{x}) = \boldsymbol{x}$, $\mathcal{A}_K(\boldsymbol{y}) = \boldsymbol{y}$, and $\mathcal{A}_V(\boldsymbol{y}) = 1$ yields $\mathrm{Attn}_\rho(\boldsymbol{x}) = \langle \boldsymbol{x}, \boldsymbol{m}_\rho \rangle$. For the second stage, set $\varphi_1(x) = \mathrm{ReLU}(x) + \mathrm{ReLU}(-x) = |x|$ to extract the absolute-alignment feature $|\langle \boldsymbol{x}, \boldsymbol{m}_\rho \rangle|$. The readout $\varphi_2(x) = c_{d,q}\mathrm{ReLU}(x - \beta_d) - c_{d,q}\mathrm{ReLU}(\beta_d - x) = c_{d,q}(x - \beta_d)$ then performs the normalization. Both maps are realized by two-layer ReLU networks of width 2. Thus, a linear Transformer with latent dimension $M = 1$ exactly represents the normalized hard functional, namely $T(\rho) = f_{d,q}(\rho)$ for every $\rho \in \mathcal{P}(\mathbb{S}^{d-1})$.

4.3 Proof sketch of Theorem 3.7

For $r \geq 2$, the pooled representation is generally nonlinear in $(\boldsymbol{A}, \boldsymbol{b})$; consequently, neither the linear-projection argument in Lemma 4.5 nor the rank-based energy estimate in Lemma 4.6 extends directly. We instead establish the lower bound through indistinguishability: an insufficient latent dimension forces certain input distributions with separated target values to share the same pooled representation. The output map cannot distinguish these inputs, resulting in a nonvanishing uniform approximation error. We outline the proof sketches with detailed proofs completed in Appendix B.2.

Step 1: Antipodal feature differences. We compare the pooled representations of input distributions with opposite harmonic perturbations:

$$\boldsymbol{\Delta}^{(r)}(\boldsymbol{A}, \boldsymbol{b}) \coloneqq \boldsymbol{Z}^{(r)}(\boldsymbol{A}, \boldsymbol{b}) - \boldsymbol{Z}^{(r)}(-\boldsymbol{A}, \boldsymbol{b}),$$

where

$$\boldsymbol{Z}^{(r)}(\boldsymbol{A}, \boldsymbol{b}) \coloneqq \int_{(\mathbb{S}^{d-1})^r} \phi_1(\boldsymbol{x}_1, \ldots, \boldsymbol{x}_r)\, \mathrm{d}\rho_{\boldsymbol{A},\boldsymbol{b}}(\boldsymbol{x}_1) \cdots \mathrm{d}\rho_{\boldsymbol{A},\boldsymbol{b}}(\boldsymbol{x}_r).$$

The following lemma establishes a low-degree expansion in $\boldsymbol{b}$ whose coefficients are continuous and odd in $\boldsymbol{A}$.

Lemma 4.8 (Low-degree expansion with odd coefficients)**.** *Let $d \geq 3$, and let $q \geq 2$ be even. Fix $r \geq 1$ and a continuous feature map ϕ_1. For every $\boldsymbol{A} \in \mathbb{S}^{L_{d,q}-1}$ and $j = 1, \ldots, M$, $\boldsymbol{\Delta}_j^{(r)}(\boldsymbol{A}, \cdot) \in \mathscr{P}_{\leq r-1}(\mathbb{S}^{d-1})$. For any fixed $L^2(\nu_d)$-orthonormal basis $\{p_k\}_{k=1}^{K_{d,r-1}}$ of $\mathscr{P}_{\leq r-1}(\mathbb{S}^{d-1})$, each coordinate admits the expansion*

$$\boldsymbol{\Delta}_j^{(r)}(\boldsymbol{A}, \boldsymbol{b}) = \sum_{k=1}^{K_{d,r-1}} c_{j,k}(\boldsymbol{A}) p_k(\boldsymbol{b}),$$

where the coefficient maps $c_{j,k} : \mathbb{S}^{L_{d,q}-1} \to \mathbb{R}$ are continuous and odd: $c_{j,k}(-\boldsymbol{A}) = -c_{j,k}(\boldsymbol{A})$.

Step 2: Antipodal inputs with indistinguishable features. The continuity and oddness of the coefficient maps allow us to apply the Borsuk–Ulam theorem (Matoušek, 2003). When the latent dimension is below the following threshold, this yields distinct distributions with opposite harmonic perturbations but identical pooled representations. A similar Borsuk–Ulam-based method was employed by Syed et al. (2026) to characterize the minimal latent dimension required for Janossy pooling networks to achieve UAP for permutation-invariant functions on fixed-length sequences.

Lemma 4.9 (Indistinguishable pooled representations)**.** *Under the assumptions of Lemma 4.8, if* $MK_{d,r-1} < L_{d,q}$*, then there exists* $\boldsymbol{A}_* \in \mathbb{S}^{L_{d,q}-1}$ *such that*

$$\boldsymbol{Z}^{(r)}(\boldsymbol{A}_*, \boldsymbol{b}) = \boldsymbol{Z}^{(r)}(-\boldsymbol{A}_*, \boldsymbol{b}) \qquad \textit{for every } \boldsymbol{b} \in \mathbb{S}^{d-1}.$$

Step 3: Target separation and approximation lower bound. For the previously identified harmonic components $\boldsymbol{A}_*$, we select a barycenter direction $\boldsymbol{b}_*$ such that $\rho_{\boldsymbol{A}_*,\boldsymbol{b}_*}$ and $\rho_{-\boldsymbol{A}_*,\boldsymbol{b}_*}$ have well-separated target values. Together with their indistinguishable pooled representations, this separation yields a nonvanishing constant lower bound on the approximation error.

Lemma 4.10 (Target separation)**.** *Let* $d \geq 3$ *and let* $q \geq 2$ *be even. For every* $\boldsymbol{A}_* \in \mathbb{S}^{L_{d,q}-1}$*, there exists* $\boldsymbol{b}_* \in \mathbb{S}^{d-1}$ *such that*

$$|f_{d,q}(\rho_{\boldsymbol{A}_*,\boldsymbol{b}_*}) - f_{d,q}(\rho_{-\boldsymbol{A}_*,\boldsymbol{b}_*})| \geq 2.$$

Combining the indistinguishability of pooled representations in Lemma 4.9 with the functional value separation in Lemma 4.10 yields the following uniform approximation lower bound.

Lemma 4.11 (Lower bound for r-ary Janossy pooling)**.** *Let* $d \geq 3$*, let* $q \geq 2$ *be even, and let* $r, M \geq 1$ *be integers. If* $MK_{d,r-1} < L_{d,q}$*, then*

$$\inf_{g \in \mathcal{G}_M^{(r)}} \sup_{\rho \in \mathcal{P}(\mathbb{S}^{d-1})} |g(\rho) - f_{d,q}(\rho)| \geq 1.$$

With $q_d = 2\lfloor d/2 \rfloor$, Proposition C.3 gives $K_{d,r-1} = \Theta(d^{r-1})$ and $L_{d,q_d} = \Theta(4^d/\sqrt{d})$. Hence, there exists a constant $c_r > 0$, depending only on r, such that $M \leq c_r 4^d/d^{r-1/2}$ implies $MK_{d,r-1} < L_{d,q_d}$. Lemma 4.11 then establishes the lower bound in Theorem 3.7.

5 Concluding Remarks

We study the expressivity of r-ary Janossy pooling networks and linear Transformers from a functional perspective. Although both architectures possess UAP for continuous statistical functionals, we establish an exponential separation in latent dimension, favoring linear Transformers. This advantage stems from attention followed by pooling, a two-stage process of global interaction and aggregation that enables distribution-dependent feature extraction, whereas Janossy pooling networks are restricted to single-stage feature extraction using feature maps fixed across input distributions.

Complementing the separation established by Zweig & Bruna (2022) for networks with complex-analytic activations, our result holds in the more practical real-valued setting for arbitrary continuous feature maps and readouts, irrespective of depth or internal complexity. This generality stems from pooling's intrinsic information bottleneck: insufficient latent dimension leaves target-relevant distinctions between input distributions unresolved, which no subsequent readout can recover.

Looking ahead, a natural extension is to investigate how intermediate permutation-equivariant layers (Zaheer et al., 2017) affect the separation between pooling-based and attention-based architectures. Characterizing separations between shallow and deep Transformers would further elucidate the role of multistage information interaction and aggregation across stacked blocks in approximating statistical functionals. Beyond approximation, an important question is whether analogous architectural separations arise in sample complexity and optimization efficiency (Lahoti et al., 2024; Malach & Shalev-Shwartz, 2021).

Limitations. For fixed arity $r \geq 2$, our separation is established in the uniform norm; we conjecture that an analogous exponential separation holds in L^2 under a suitable distribution over input measures. The sharpness of the latent-dimension lower bound $M \geq c_r 4^d/d^{r-1/2}$ for the constructed functional at fixed approximation accuracy is also unresolved. Finally, our separation concerns a single hard functional; identifying broader classes that admit efficient approximation by linear Transformers but not by Janossy pooling networks remains an open direction.

AI USE STATEMENT

In this work, we used generative AI tools to polish the writing for clarity and readability and to assist with literature searches, with the aim of improving the coverage of related work. The proof strategies and mathematical arguments were developed by the authors, while generative AI was used to assist in checking derivations and mathematical correctness. We take responsibility for the final content of this work, including all text, claims, and artifacts produced with the aid of generative AI.

REFERENCES

Kwangjun Ahn, Xiang Cheng, Minhak Song, Chulhee Yun, Ali Jadbabaie, and Suvrit Sra. Linear attention is (maybe) all you need (to understand Transformer optimization). In *International Conference on Learning Representations*, volume 2024, pp. 16193–16205, 2024.

Silas Alberti, Niclas Dern, Laura Thesing, and Gitta Kutyniok. Sumformer: Universal approximation for efficient Transformers. In *Topological, Algebraic and Geometric Learning Workshops 2023*, pp. 72–86. PMLR, 2023.

Kendall Atkinson and Weimin Han. *Spherical harmonics and approximations on the unit sphere: An introduction. Lecture Notes in Mathematics*, volume 2044. Springer Heidelberg, 2012.

Satwik Bhattamishra, Michael Hahn, Phil Blunsom, and Varun Kanade. Separations in the representational capabilities of transformers and recurrent architectures. *Advances in Neural Information Processing Systems*, 37:36002–36045, 2024.

Michael M Bronstein, Joan Bruna, Taco Cohen, and Petar Veličković. Geometric deep learning: Grids, groups, graphs, geodesics, and gauges. *arXiv preprint arXiv:2104.13478*, 2021.

Jingpu Cheng, Ting Lin, Zuowei Shen, and Qianxiao Li. A unified framework for establishing the universal approximation of Transformer-type architectures. *Advances in Neural Information Processing Systems*, 38:129131–129163, 2026.

Shao-Ting Chiu, Aditya Nambiar, Ali Syed, Jonathan W Siegel, and Ulisses Braga-Neto. In-context multi-operator learning with DeepOSets. *arXiv preprint arXiv:2512.16074*, 2025.

Gabriele Corso, Luca Cavalleri, Dominique Beaini, Pietro Liò, and Petar Veličković. Principal neighbourhood aggregation for graph nets. *Advances in Neural Information Processing Systems*, 33:13260–13271, 2020.

Amit Daniely. Depth separation for neural networks. In *Conference on Learning Theory*, pp. 690–696. PMLR, 2017.

Gwendoline De Bie, Gabriel Peyré, and Marco Cuturi. Stochastic deep networks. In *International Conference on Machine Learning*, pp. 1556–1565. PMLR, 2019.

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

Ronen Eldan and Ohad Shamir. The power of depth for feedforward neural networks. In *Conference on Learning Theory*, pp. 907–940. PMLR, 2016.

Lizhe Fang, Yifei Wang, Khashayar Gatmiry, Lei Fang, and Yisen Wang. Rethinking invariance in in-context learning. In *International Conference on Learning Representations*, volume 2025, pp. 503–525, 2025.

Takashi Furuya, Maarten V de Hoop, and Gabriel Peyré. Transformers are universal in-context learners. In *International Conference on Learning Representations*, volume 2025, pp. 80820–80845, 2025.

Takashi Furuya, David Mis, Ivan Dokmanić, Maarten V de Hoop, and Matti Lassas. Function graph Transformers universally approximate operators between function spaces. *arXiv preprint arXiv:2605.17968*, 2026.

Borjan Geshkovski, Philippe Rigollet, and Domènec Ruiz-Balet. Measure-to-measure interpolation using transformers. *Foundations of Computational Mathematics*, pp. 1–50, 2026.

Justin Gilmer, Samuel S Schoenholz, Patrick F Riley, Oriol Vinyals, and George E Dahl. Neural message passing for quantum chemistry. In *International Conference on Machine Learning*, pp. 1263–1272. PMLR, 2017.

Ronald L. Graham, Donald E. Knuth, and Oren Patashnik. *Concrete mathematics: A foundation for computer science*. Addison-Wesley Reading, MA, 2 edition, 1994.

Meng-Hao Guo, Jun-Xiong Cai, Zheng-Ning Liu, Tai-Jiang Mu, Ralph R Martin, and Shi-Min Hu. PCT: Point cloud Transformer. *Computational Visual Media*, 7(2):187–199, 2021.

Max Horn, Michael Moor, Christian Bock, Bastian Rieck, and Karsten Borgwardt. Set functions for time series. In *International Conference on Machine Learning*, pp. 4353–4363. PMLR, 2020.

Maximilian Ilse, Jakub Tomczak, and Max Welling. Attention-based deep multiple instance learning. In *International Conference on Machine Learning*, pp. 2127–2136. Pmlr, 2018.

Andrew Jaegle, Felix Gimeno, Andy Brock, Oriol Vinyals, Andrew Zisserman, and Joao Carreira. Perceiver: General perception with iterative attention. In *International Conference on Machine Learning*, pp. 4651–4664. PMLR, 2021.

Haotian Jiang and Qianxiao Li. Approximation rate of the Transformer architecture for sequence modeling. *Advances in Neural Information Processing Systems*, 37:68926–68955, 2024.

Tokio Kajitsuka and Issei Sato. Are transformers with one layer self-attention using low-rank weight matrices universal approximators? In *International Conference on Learning Representations*, volume 2024, pp. 35177–35205, 2024.

Angelos Katharopoulos, Apoorv Vyas, Nikolaos Pappas, and François Fleuret. Transformers are RNNs: Fast autoregressive transformers with linear attention. In *International Conference on Machine Learning*, pp. 5156–5165. 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.

Aakash Sunil Lahoti, Stefani Karp, Ezra Winston, Aarti Singh, and Yuanzhi Li. Role of locality and weight sharing in image-based tasks: A sample complexity separation between CNNs, LCNs, and FCNs. In *International Conference on Learning Representations*, volume 2024, pp. 26826–26865, 2024.

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.

Chih-Yao Ma, Asim Kadav, Iain Melvin, Zsolt Kira, Ghassan AlRegib, and Hans Peter Graf. Attend and interact: Higher-order object interactions for video understanding. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 6790–6800. IEEE, 2018.

Eran Malach and Shai Shalev-Shwartz. Computational separation between convolutional and fully-connected networks. In *International Conference on Learning Representations*, 2021.

Haggai Maron, Ethan Fetaya, Nimrod Segol, and Yaron Lipman. On the universality of invariant networks. In *International Conference on Machine Learning*, pp. 4363–4371. PMLR, 2019.

James Martens, Arkadev Chattopadhya, Toni Pitassi, and Richard Zemel. On the representational efficiency of restricted Boltzmann machines. *Advances in Neural Information Processing Systems*, 26, 2013.

Jiří Matoušek. *Using the Borsuk–Ulam theorem: Lectures on topological methods in combinatorics and geometry*. Springer, 2003.

Charles A Micchelli, Yuesheng Xu, and Haizhang Zhang. Universal kernels. *Journal of Machine Learning Research*, 7(12), 2006.

R v Mises. On the asymptotic distribution of differentiable statistical functions. *The Annals of Mathematical Statistics*, 18(3):309–348, 1947.

Christopher Morris, Martin Ritzert, Matthias Fey, William L Hamilton, Jan Eric Lenssen, Gaurav Rattan, and Martin Grohe. Weisfeiler and Leman go neural: Higher-order graph neural networks. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pp. 4602–4609, 2019.

Ryan L. Murphy, Balasubramaniam Srinivasan, Vinayak A. Rao, and Bruno Ribeiro. Janossy pooling: Learning deep permutation-invariant functions for variable-size inputs. In *International Conference on Learning Representations*, 2019.

Gestur Ólafsson, Angela Pasquale, and Boris Rubin. Analytic and group-theoretic aspects of the cosine transform. *arXiv preprint arXiv:1209.1822*, 2012.

Marco Pacini, Xiaowen Dong, Bruno Lepri, and Gabriele Santin. Separation power of equivariant neural networks. In *International Conference on Learning Representations*, volume 2025, pp. 86956–86988, 2025.

Marco Pacini, Mircea Petrache, Bruno Lepri, Shubhendu Trivedi, and Robin Walters. On universality of deep equivariant networks. In *International Conference on Learning Representations*, volume 2026, pp. 40184–40205, 2026.

Suzanna Parkinson, Greg Ongie, Rebecca Willett, Ohad Shamir, and Nathan Srebro. Depth separation in norm-bounded infinite-width neural networks. In *The Thirty Seventh Annual Conference on Learning Theory*, pp. 4082–4114. PMLR, 2024.

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

Allan Pinkus. Approximation theory of the MLP model in neural networks. *Acta Numerica*, 8: 143–195, 1999.

Barnabás Póczos, Aarti Singh, Alessandro Rinaldo, and Larry Wasserman. Distribution-free distribution regression. In *Artificial Intelligence and Statistics*, pp. 507–515. PMLR, 2013.

Michael Prasthofer, Tim De Ryck, and Siddhartha Mishra. Variable-input deep operator networks. *arXiv preprint arXiv:2205.11404*, 2022.

Omri Puny, Derek Lim, Bobak Kiani, Haggai Maron, and Yaron Lipman. Equivariant polynomials for graph neural networks. In *International Conference on Machine Learning*, pp. 28191–28222. PMLR, 2023.

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.

Huilin Qu, Congqiao Li, and Sitian Qian. Particle Transformer for jet tagging. In *International Conference on Machine Learning*, pp. 18281–18292. PMLR, 2022.

Stefan T Radev, Ulf K Mertens, Andreas Voss, Lynton Ardizzone, and Ullrich Köthe. BayesFlow: Learning complex stochastic models with invertible neural networks. *IEEE Transactions on Neural Networks and Learning Systems*, 33(4):1452–1466, 2020.

Itay Safran and Ohad Shamir. Depth-width tradeoffs in approximating natural functions with neural networks. In *International Conference on Machine Learning*, pp. 2979–2987. PMLR, 2017.

Clayton Sanford, Daniel J Hsu, and Matus Telgarsky. Representational strengths and limitations of Transformers. *Advances in Neural Information Processing Systems*, 36:36677–36707, 2023.

Adam Santoro, David Raposo, David G Barrett, Mateusz Malinowski, Razvan Pascanu, Peter Battaglia, and Timothy Lillicrap. A simple neural network module for relational reasoning. *Advances in Neural Information Processing Systems*, 30, 2017.

Kristof T Schütt, Huziel E Sauceda, P-J Kindermans, Alexandre Tkatchenko, and K-R Müller. SchNet–a deep learning architecture for molecules and materials. *The Journal of Chemical Physics*, 148(24), 2018.

Zhongjie Shi, Zhan Yu, and Ding-Xuan Zhou. Learning theory of distribution regression with neural networks. *Constructive Approximation*, 62(1):61–104, 2025.

Jonathan W Siegel, Snir Hordan, Hannah Lawrence, Ali Syed, and Nadav Dym. Quantitative approximation rates for group equivariant learning. *arXiv preprint arXiv:2602.20370*, 2026.

Alex Smola, Arthur Gretton, Le Song, and Bernhard Schölkopf. A Hilbert space embedding for distributions. In *International Conference on Algorithmic Learning Theory*, pp. 13–31. Springer, 2007.

Bharath K Sriperumbudur, Arthur Gretton, Kenji Fukumizu, Bernhard Schölkopf, and Gert RG Lanckriet. Hilbert space embeddings and metrics on probability measures. *The Journal of Machine Learning Research*, 11:1517–1561, 2010.

Marshall H Stone. The generalized Weierstrass approximation theorem. *Mathematics Magazine*, 21 (5):237–254, 1948.

Ali Syed, Aditya Nambiar, and Jonathan W Siegel. Embedding dimension lower bounds for universality of deep sets and Janossy pooling. *arXiv preprint arXiv:2605.08377*, 2026.

Zoltán Szabó, Bharath K Sriperumbudur, Barnabás Póczos, and Arthur Gretton. Learning theory for distribution regression. *Journal of Machine Learning Research*, 17(152):1–40, 2016.

Naoki Takeshita and Masaaki Imaizumi. Approximation of permutation invariant polynomials by Transformers: Efficient construction in column-size. *arXiv preprint arXiv:2502.11467*, 2025.

Matus Telgarsky. Benefits of depth in neural networks. In *Conference on Learning Theory*, pp. 1517–1539. PMLR, 2016.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in Neural Information Processing Systems*, 30, 2017.

Luca Venturi, Samy Jelassi, Tristan Ozuch, and Joan Bruna. Depth separation beyond radial functions. *Journal of Machine Learning Research*, 23(122):1–56, 2022.

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

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

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

Peihao Wang, Shenghao Yang, Shu Li, Zhangyang Wang, and Pan Li. Polynomial width is sufficient for set representation with high-dimensional features. In *International Conference on Learning Representations*, volume 2024, pp. 52223–52252, 2024.

Yue Wang, Yongbin Sun, Ziwei Liu, Sanjay E Sarma, Michael M Bronstein, and Justin M Solomon. Dynamic graph cnn for learning on point clouds. *ACM Transactions on Graphics (tog)*, 38(5): 1–12, 2019.

Kaiyue Wen, Xingyu Dang, and Kaifeng Lyu. RNNs are not transformers (yet): The key bottleneck on in-context retrieval. In *International Conference on Learning Representations*, volume 2025, pp. 48813–48856, 2025.

Haixu Wu, Huakun Luo, Haowen Wang, Jianmin Wang, and Mingsheng Long. Transolver: A fast Transformer solver for PDEs on general geometries. In *International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 53681–53705. PMLR, 2024.

Keyulu Xu, Weihua Hu, Jure Leskovec, and Stefanie Jegelka. How powerful are graph neural networks? In *International Conference on Learning Representations*, 2019.

Bo Yang, Sen Wang, Andrew Markham, and Niki Trigoni. Robust attentional aggregation of deep feature sets for multi-view 3d reconstruction. *International Journal of Computer Vision*, 128(1): 53–73, 2020.

Kôsaku Yosida. *Functional analysis*. Springer Science & Business Media, 2012.

Penghao Yu, Haotian Jiang, Zeyu Bao, Ruoxi Yu, and Qianxiao Li. The effect of attention head count on Transformer approximation. In *International Conference on Learning Representations*, volume 2026, pp. 77573–77607, 2026.

Chulhee Yun, Srinadh Bhojanapalli, Ankit Singh Rawat, Sashank J. Reddi, and Sanjiv Kumar. Are Transformers universal approximators of sequence-to-sequence functions? In *International Conference on Learning Representations*, 2020.

Manzil Zaheer, Satwik Kottur, Siamak Ravanbakhsh, Barnabas Poczos, Russ R Salakhutdinov, and Alexander Smola. Deep sets. *Advances in Neural Information Processing Systems*, 30, 2017.

Lily Zhang, Veronica Tozzo, John Higgins, and Rajesh Ranganath. Set norm and equivariant skip connections: Putting the deep in deep sets. In *International Conference on Machine Learning*, pp. 26559–26574. PMLR, 2022.

Hengshuang Zhao, Li Jiang, Jiaya Jia, Philip HS Torr, and Vladlen Koltun. Point transformer. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 16259–16268, 2021.

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

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

A UNIVERSAL APPROXIMATION OF STATISTICAL FUNCTIONALS

In this section, we provide detailed proofs of the UAP results for continuous statistical functionals.

A.1 TOPOLOGY OF $\mathcal{P}(\Omega)$

In this subsection, we recall several fundamental topological properties of $\mathcal{P}(\Omega)$. We first identify its weak-$*$ topology with the topology induced by the 1-Wasserstein distance.

Proposition A.1 (Wasserstein metrization of the weak-$*$ topology)**.** *Let $\Omega \subset \mathbb{R}^d$ be nonempty and compact. For $\rho_1, \rho_2 \in \mathcal{P}(\Omega)$, define*

$$W_1(\rho_1, \rho_2) \coloneqq \inf_{\pi \in \Pi(\rho_1, \rho_2)} \int_{\Omega \times \Omega} \|\boldsymbol{x} - \boldsymbol{y}\|_1 \, \mathrm{d}\pi(\boldsymbol{x}, \boldsymbol{y}),$$

where $\Pi(\rho_1, \rho_2)$ denotes the set of Borel probability measures on $\Omega \times \Omega$ with marginals ρ_1 and ρ_2. Then W_1 metrizes the weak-$*$ topology on $\mathcal{P}(\Omega)$. Equivalently, for any sequence $(\rho_n)_{n \geq 1} \subset \mathcal{P}(\Omega)$ and $\rho \in \mathcal{P}(\Omega)$,

$$W_1(\rho_n, \rho) \longrightarrow 0 \quad \Longleftrightarrow \quad \int_\Omega f \, \mathrm{d}\rho_n \longrightarrow \int_\Omega f \, \mathrm{d}\rho \quad \textit{for every } f \in C(\Omega).$$

The proposition follows from Villani (2009, Corollary 6.13) and provides a concrete and intuitive metric characterization of weak-$*$ convergence on $\mathcal{P}(\Omega)$.

We next establish the compactness of $\mathcal{P}(\Omega)$ in the weak-$*$ topology.

Lemma A.2 (Weak-$*$ compactness of the probability-measure space)**.** *Let $\Omega \subset \mathbb{R}^d$ be nonempty and compact. Then $\mathcal{P}(\Omega)$ is compact Hausdorff in the weak-$*$ topology.*

Proof. Equip $C(\Omega)$ with the uniform norm. By the Riesz representation theorem (Yosida, 2012, Example 6, p. 119), each $\rho \in \mathcal{P}(\Omega)$ is identified with the positive linear functional

$$\Lambda_\rho(f) := \int_\Omega f \, \mathrm{d}\rho, \qquad f \in C(\Omega),$$

satisfying $\Lambda_\rho(\mathbf{1}) = 1$. For every $f \in C(\Omega)$,

$$|\Lambda_\rho(f)| \leq \Lambda_\rho(|f|) \leq \|f\|_\infty.$$

Thus $\|\Lambda_\rho\| \leq 1$, while $\Lambda_\rho(\mathbf{1}) = \|\mathbf{1}\|_\infty = 1$ gives the reverse inequality. Hence $\|\Lambda_\rho\| = 1$.

Under this identification, $\mathcal{P}(\Omega)$ consists precisely of the elements of the closed unit ball of $C(\Omega)^*$ satisfying

$$\Lambda(\mathbf{1}) = 1, \qquad \Lambda(f) \geq 0 \quad \text{for every } f \in C(\Omega) \text{ with } f \geq 0.$$

Each condition defines a weak-$*$ closed set, since $\Lambda \mapsto \Lambda(f)$ is weak-$*$ continuous for every fixed $f \in C(\Omega)$. Their intersection is therefore weak-$*$ closed.

The Banach–Alaoglu theorem (Yosida, 2012, Theorem 1, p. 137) ensures that the closed unit ball of $C(\Omega)^*$ is weak-$*$ compact. As a closed subset of this ball, $\mathcal{P}(\Omega)$ is also weak-$*$ compact. Finally, the weak-$*$ topology on $C(\Omega)^*$ is Hausdorff, and this property is inherited by $\mathcal{P}(\Omega)$. □

This compactness result provides the natural setting for studying the UAP of continuous statistical functionals on $\mathcal{P}(\Omega)$. In particular, continuous functionals are bounded on $\mathcal{P}(\Omega)$, while the compact Hausdorff structure enables the application of the Stone–Weierstrass theorem in the proof below.

A.2 PROOF OF PROPOSITION 3.2

In this subsection, we prove Proposition 3.2 by constructing an injective transformation of probability measures with a fixed choice of linear attention, and then applying the generalized Stone–Weierstrass theorem to an algebra of pooled features. Throughout, $\Omega \subset \mathbb{R}^d$ is nonempty and compact. Taking $d_k = d_v = d$ and $\mathcal{A}_Q = \mathcal{A}_K = \mathcal{A}_V = \mathrm{Id}_{\mathbb{R}^d}$ gives

$$\mathrm{Attn}_\rho(\boldsymbol{x}) = \int_\Omega \langle \boldsymbol{x}, \boldsymbol{y} \rangle \boldsymbol{y} \, \mathrm{d}\rho(\boldsymbol{y}) = \Sigma_\rho \boldsymbol{x}, \qquad \Sigma_\rho \coloneqq \int_\Omega \boldsymbol{x}\boldsymbol{x}^\top \, \mathrm{d}\rho(\boldsymbol{x}). \tag{5}$$

Let $\Upsilon(\rho)$ denote the distribution of these attention outputs. Explicitly,

$$\Upsilon(\rho) \coloneqq (\Sigma_\rho)_{\#}\rho, \qquad \Upsilon(\rho)(E) = \rho\big(\{\boldsymbol{x} \in \Omega : \Sigma_\rho \boldsymbol{x} \in E\}\big)$$

for every Borel set $E \subseteq \mathbb{R}^d$. Thus, if $X \sim \rho$, then $\Sigma_\rho X$ has distribution $\Upsilon(\rho)$. We identify measures on Ω with their extensions to $\mathbb{R}^d$ assigning zero mass outside Ω.

Lemma A.3 (Injectivity of the attention-induced transformation)**.** *The map $\Upsilon : \mathcal{P}(\Omega) \to \mathcal{P}(\mathbb{R}^d)$ defined by $\Upsilon(\rho) = (\Sigma_\rho)_{\#}\rho$ is injective.*

Proof. Fix $\rho \in \mathcal{P}(\Omega)$, let $X \sim \rho$, and set $Z = \Sigma_\rho X$ and $\nu = \Upsilon(\rho)$. Compactness of Ω ensures that all second moments below are finite. Since Σ_ρ is symmetric positive semidefinite,

$$\int_{\mathbb{R}^d} \boldsymbol{z}\boldsymbol{z}^\top \,\mathrm{d}\nu(\boldsymbol{z}) = \mathbb{E}[ZZ^\top] = \Sigma_\rho \mathbb{E}[XX^\top]\Sigma_\rho = \Sigma_\rho^3. \tag{6}$$

The unique positive semidefinite cube root of the left-hand side therefore determines Σ_ρ from ν.

To recover ρ even when Σ_ρ is singular, observe that for every $\boldsymbol{v} \in \ker \Sigma_\rho$,

$$0 = \boldsymbol{v}^\top \Sigma_\rho \boldsymbol{v} = \int_\Omega |\boldsymbol{v}^\top \boldsymbol{x}|^2 \,\mathrm{d}\rho(\boldsymbol{x}).$$

Hence $\boldsymbol{v}^\top X = 0$ almost surely. Applying this identity to a finite orthonormal basis of $\ker \Sigma_\rho$ yields

$$X \in (\ker \Sigma_\rho)^\perp = \mathrm{range}(\Sigma_\rho) \qquad \text{almost surely.}$$

Let $\Sigma_\rho^\dagger$ denote the Moore–Penrose pseudoinverse. Since $\Sigma_\rho^\dagger \Sigma_\rho$ is the orthogonal projection onto $\mathrm{range}(\Sigma_\rho)$,

$$\Sigma_\rho^\dagger Z = X \qquad \text{almost surely,} \qquad \rho = (\Sigma_\rho^\dagger)_{\#}\nu. \tag{7}$$

Both Σ_ρ and its pseudoinverse are determined by ν. Thus ν uniquely determines ρ, proving injectivity. □

The lemma shows that the fixed attention transformation preserves all information in the input measure. Its injectivity will provide the point-separation property required by the Stone–Weierstrass theorem.

Proof of Proposition 3.2. By Proposition A.1 and Lemma A.2, $\mathcal{P}(\Omega)$ is compact and metrizable in the weak-$*$ topology. We construct an algebra $\mathscr{A} \subseteq \mathcal{T} \subseteq C(\mathcal{P}(\Omega))$ and verify the two conditions of Theorem C.6: algebraic closure, and inclusion of constants together with separation of points. Throughout, $\|\cdot\|_{\mathrm{op}}$ denotes the Euclidean operator norm.

Continuity of the model class. Fix $T \in \mathcal{T}$ and its constituent maps. Writing

$$\mathbf{D}_\rho \coloneqq \int_\Omega \mathcal{A}_V(\boldsymbol{y})\mathcal{A}_K(\boldsymbol{y})^\top \,\mathrm{d}\rho(\boldsymbol{y}), \qquad \mathrm{Attn}_\rho(\boldsymbol{x}) = \mathbf{D}_\rho \mathcal{A}_Q(\boldsymbol{x}),$$

weak-$*$ convergence $\rho_n \rightharpoonup \rho$ implies $\mathbf{D}_{\rho_n} \to \mathbf{D}_\rho$ in operator norm. Boundedness of $\mathcal{A}_Q$ on Ω then gives $\mathrm{Attn}_{\rho_n} \to \mathrm{Attn}_\rho$ uniformly on Ω. Moreover, all attention outputs of this fixed model lie in a common compact ball, since its affine maps are bounded on Ω. Uniform continuity of φ_1 on that ball implies

$$\boldsymbol{h}_n \coloneqq \varphi_1 \circ \mathrm{Attn}_{\rho_n} \longrightarrow \boldsymbol{h} \coloneqq \varphi_1 \circ \mathrm{Attn}_\rho \qquad \text{uniformly on } \Omega.$$

Consequently,

$$\left\| \int_\Omega \boldsymbol{h}_n \,\mathrm{d}\rho_n - \int_\Omega \boldsymbol{h} \,\mathrm{d}\rho \right\|_\infty \le \sup_{\boldsymbol{x}\in\Omega} \|\boldsymbol{h}_n(\boldsymbol{x}) - \boldsymbol{h}(\boldsymbol{x})\|_\infty + \left\| \int_\Omega \boldsymbol{h} \,\mathrm{d}(\rho_n - \rho) \right\|_\infty \longrightarrow 0.$$

The second term vanishes coordinatewise by weak-$*$ convergence. Continuity of φ_2 therefore gives $T(\rho_n) \to T(\rho)$. Since $\mathcal{P}(\Omega)$ is metrizable, this proves $\mathcal{T} \subseteq C(\mathcal{P}(\Omega))$.

Construction of the feature algebra. We now fix the attention maps in equation 5. Set

$$R \coloneqq \max\left\{1, \max_{\boldsymbol{x}\in\Omega} \|\boldsymbol{x}\|_2\right\}, \qquad \mathbb{B}^d(R^3) \coloneqq \{\boldsymbol{z} \in \mathbb{R}^d : \|\boldsymbol{z}\|_2 \le R^3\}.$$

For every $\rho \in \mathcal{P}(\Omega)$,

$$\|\Sigma_\rho\|_{\mathrm{op}} \le \int_\Omega \|\boldsymbol{x}\|_2^2 \,\mathrm{d}\rho(\boldsymbol{x}) \le R^2, \qquad \sup_{\boldsymbol{x}\in\Omega} \|\Sigma_\rho \boldsymbol{x}\|_2 \le R^3.$$

Thus every transformed measure $\Upsilon(\rho)$ is supported in the same compact ball $\mathbb{B}^d(R^3)$. For $\psi \in C(\mathbb{B}^d(R^3))$, define

$$J_\psi(\rho) \coloneqq \int_\Omega \psi(\Sigma_\rho \boldsymbol{x}) \,\mathrm{d}\rho(\boldsymbol{x}) = \int_{\mathbb{B}^d(R^3)} \psi(\boldsymbol{z}) \,\mathrm{d}\Upsilon(\rho)(\boldsymbol{z}). \tag{8}$$

To realize these features with maps defined on $\mathbb{R}^d$, use the continuous projection

$$\mathrm{Proj}_{\mathbb{B}^d(R^3)}(\boldsymbol{z}) \coloneqq \frac{\boldsymbol{z}}{\max\{1, \|\boldsymbol{z}\|_2/R^3\}}, \qquad \bar{\psi} \coloneqq \psi \circ \mathrm{Proj}_{\mathbb{B}^d(R^3)}.$$

Then $\bar{\psi} \in C(\mathbb{R}^d)$ agrees with ψ on $\mathbb{B}^d(R^3)$. Taking $\varphi_1 = \bar{\psi}$ and $\varphi_2 = \mathrm{Id}_{\mathbb{R}}$ realizes J_ψ as an element of $\mathcal{T}_1$. Hence $J_\psi \in C(\mathcal{P}(\Omega))$. Since these integrals determine the weak-$*$ topology on $\mathcal{P}(\mathbb{B}^d(R^3))$, they also establish the continuity of Υ.

Let $\mathscr{A}$ consist of all finite polynomial combinations of these features:

$$\mathscr{A} \coloneqq \left\{ \rho \mapsto P\big(J_{\psi_1}(\rho), \dots, J_{\psi_M}(\rho)\big) \;\middle|\; \begin{array}{l} M \ge 1, \quad \psi_j \in C(\mathbb{B}^d(R^3)), \\ P : \mathbb{R}^M \to \mathbb{R} \text{ a real polynomial} \end{array} \right\}.$$

Every element is realized by the same fixed attention head with $\varphi_1 = (\bar{\psi}_1, \dots, \bar{\psi}_M)^\top$ and $\varphi_2 = P$. Therefore,

$$\mathscr{A} \subseteq \mathcal{T} \subseteq C(\mathcal{P}(\Omega)).$$

Condition (i): Algebraic closure. For $G_1, G_2 \in \mathscr{A}$, concatenate their finite lists of pooled features. Both functionals are polynomials in the resulting common feature vector, and so are their sum, product, and scalar multiples. Thus, for every $\alpha \in \mathbb{R}$,

$$G_1 + G_2 \in \mathscr{A}, \qquad G_1 G_2 \in \mathscr{A}, \qquad \alpha G_1 \in \mathscr{A}.$$

Condition (ii): Constants and separation of points. Constant polynomials show that $\mathscr{A}$ contains every constant functional. For distinct $\rho, \rho' \in \mathcal{P}(\Omega)$, Lemma A.3 gives $\Upsilon(\rho) \ne \Upsilon(\rho')$. By uniqueness in the Riesz representation theorem, there exists $\psi \in C(\mathbb{B}^d(R^3))$ such that

$$\int_{\mathbb{B}^d(R^3)} \psi \,\mathrm{d}\Upsilon(\rho) \ne \int_{\mathbb{B}^d(R^3)} \psi \,\mathrm{d}\Upsilon(\rho').$$

Equivalently, $J_\psi(\rho) \ne J_\psi(\rho')$. Since $J_\psi \in \mathscr{A}$, the algebra separates points of $\mathcal{P}(\Omega)$.

Both conditions of Theorem C.6 are therefore satisfied on the compact Hausdorff space $\mathcal{P}(\Omega)$. Consequently, $\mathscr{A}$ is uniformly dense in $C(\mathcal{P}(\Omega))$, and the same holds for $\mathcal{T}$. Explicitly, for every $F \in C(\mathcal{P}(\Omega))$ and $\varepsilon > 0$, there exist a finite M, $\psi_1, \dots, \psi_M \in C(\mathbb{B}^d(R^3))$, and a real polynomial $P : \mathbb{R}^M \to \mathbb{R}$ such that

$$\sup_{\rho\in\mathcal{P}(\Omega)} \big|F(\rho) - P\big(\boldsymbol{u}(\rho)\big)\big| < \varepsilon, \qquad \boldsymbol{u}(\rho) \coloneqq \int_\Omega \boldsymbol{\psi}(\Sigma_\rho \boldsymbol{x}) \,\mathrm{d}\rho(\boldsymbol{x}), \tag{9}$$

where $\boldsymbol{\psi} = (\psi_1, \dots, \psi_M)^\top$. These choices are independent of the input measure ρ, and no additional attention heads or layers are introduced. □

Parameterization of feature and readout maps. We now justify Remark 3.3 for neural-network parameterizations of the feature and readout maps. The argument applies to continuous multilayer perceptron (MLP) classes admitting uniform approximation of continuous functions on compact sets. For example, shallow neural networks with continuous nonpolynomial activations satisfy the requirement (Pinkus, 1999).

Fix $F \in C(\mathcal{P}(\Omega))$ and $\varepsilon > 0$. By equation 9, there exist a finite M, $\boldsymbol{\psi} = (\psi_1, \dots, \psi_M)^\top$ with $\psi_j \in C(\mathbb{B}^d(R^3))$, and a polynomial $P : \mathbb{R}^M \to \mathbb{R}$ such that

$$\sup_{\rho \in \mathcal{P}(\Omega)} \big|F(\rho) - P\big(\boldsymbol{u}(\rho)\big)\big| < \frac{\varepsilon}{3}, \qquad \boldsymbol{u}(\rho) = \int_\Omega \boldsymbol{\psi}(\Sigma_\rho \boldsymbol{x}) \, \mathrm{d}\rho(\boldsymbol{x}).$$

Retain the fixed attention maps in equation 5, so that $\mathrm{Attn}_\rho(\boldsymbol{x}) = \Sigma_\rho \boldsymbol{x} \in \mathbb{B}^d(R^3)$ for every ρ and $\boldsymbol{x} \in \Omega$. To control the readout approximation, define

$$a_\psi \coloneqq 1 + \sup_{\boldsymbol{z} \in \mathbb{B}^d(R^3)} \|\boldsymbol{\psi}(\boldsymbol{z})\|_\infty, \qquad \mathbb{B}^M_\infty(a_\psi) \coloneqq [-a_\psi, a_\psi]^M.$$

Since ρ has total mass one, $\|\boldsymbol{u}(\rho)\|_\infty \le a_\psi - 1$ for every ρ. Uniform continuity of P on the compact set $\mathbb{B}^M_\infty(a_\psi)$ provides $\delta \in (0, 1)$ such that

$$\sup_{\substack{\boldsymbol{v}, \boldsymbol{w} \in \mathbb{B}^M_\infty(a_\psi) \\ \|\boldsymbol{v} - \boldsymbol{w}\|_\infty \le \delta}} |P(\boldsymbol{v}) - P(\boldsymbol{w})| < \frac{\varepsilon}{3}.$$

Applying the MLP approximation theorem to each coordinate of $\boldsymbol{\psi}$ and to P yields two-layer networks $\varphi_1 : \mathbb{R}^d \to \mathbb{R}^M$ and $\varphi_2 : \mathbb{R}^M \to \mathbb{R}$ satisfying

$$\sup_{\boldsymbol{z} \in \mathbb{B}^d(R^3)} \|\varphi_1(\boldsymbol{z}) - \boldsymbol{\psi}(\boldsymbol{z})\|_\infty < \delta, \qquad \sup_{\boldsymbol{v} \in \mathbb{B}^M_\infty(a_\psi)} |\varphi_2(\boldsymbol{v}) - P(\boldsymbol{v})| < \frac{\varepsilon}{3}.$$

The coordinatewise approximants for φ_1 can be combined into a single finite hidden layer.

Define the approximating pooled representation by

$$\widehat{\boldsymbol{u}}(\rho) \coloneqq \int_\Omega \varphi_1(\Sigma_\rho \boldsymbol{x}) \, \mathrm{d}\rho(\boldsymbol{x}).$$

Since $\Upsilon(\rho)$ is a probability measure supported on $\mathbb{B}^d(R^3)$, integration preserves the uniform error bound:

$$\sup_{\rho \in \mathcal{P}(\Omega)} \|\widehat{\boldsymbol{u}}(\rho) - \boldsymbol{u}(\rho)\|_\infty \le \sup_{\boldsymbol{z} \in \mathbb{B}^d(R^3)} \|\varphi_1(\boldsymbol{z}) - \boldsymbol{\psi}(\boldsymbol{z})\|_\infty < \delta.$$

In particular, $\|\widehat{\boldsymbol{u}}(\rho)\|_\infty \le a_\psi - 1 + \delta < a_\psi$. Thus both pooled representations belong to $\mathbb{B}^M_\infty(a_\psi)$, uniformly over ρ. The resulting Transformer $T(\rho) = \varphi_2(\widehat{\boldsymbol{u}}(\rho))$ therefore satisfies

$$\begin{aligned} \sup_{\rho \in \mathcal{P}(\Omega)} |F(\rho) - T(\rho)| \le{}& \sup_{\rho \in \mathcal{P}(\Omega)} |F(\rho) - P(\boldsymbol{u}(\rho))| + \sup_{\rho \in \mathcal{P}(\Omega)} |P(\boldsymbol{u}(\rho)) - P(\widehat{\boldsymbol{u}}(\rho))| \\ &+ \sup_{\rho \in \mathcal{P}(\Omega)} |P(\widehat{\boldsymbol{u}}(\rho)) - \varphi_2(\widehat{\boldsymbol{u}}(\rho))| < \varepsilon. \end{aligned}$$

The finite MLP widths, parameters, and pooled dimension may depend on F and ε, but not on ρ; the attention maps remain fixed. Hence the UAP persists under these neural-network parameterizations.

The same argument applies to r-ary Janossy pooling networks for every fixed $r \ge 1$, by approximating ϕ_1 on the compact set $\Omega^{\otimes r}$ and ϕ_2 on a compact neighborhood of the pooled feature range, since $\rho^{\otimes r}$ is again a probability measure.

Connection to fixed-length sequences. We justify Remark 3.4. Fix $N \ge 1$ and identify $\boldsymbol{X} = [\boldsymbol{x}_1, \dots, \boldsymbol{x}_N]$ with an element of $\Omega^{\otimes N}$. Define the empirical-measure map and its image by

$$e_N : \Omega^{\otimes N} \to \mathcal{P}(\Omega), \qquad e_N(\boldsymbol{X}) = \hat{\rho}_{\boldsymbol{X}} \coloneqq \frac{1}{N} \sum_{i=1}^N \delta_{\boldsymbol{x}_i}, \qquad E_N \coloneqq e_N(\Omega^{\otimes N}).$$

For $\boldsymbol{X}, \boldsymbol{Y} \in \Omega^{\otimes N}$, coupling corresponding entries gives

$$W_1(\hat{\rho}_{\boldsymbol{X}}, \hat{\rho}_{\boldsymbol{Y}}) \leq \frac{1}{N} \sum_{i=1}^{N} \|\boldsymbol{x}_i - \boldsymbol{y}_i\|_1.$$

Thus e_N is continuous by Proposition A.1. Since $\Omega^{\otimes N}$ is compact, E_N is compact and therefore closed in the Hausdorff space $\mathcal{P}(\Omega)$.

Two sequences of length N have the same empirical measure if and only if they differ by a permutation, with multiplicities preserved. Consequently, every continuous permutation-invariant function $f_N : \Omega^{\otimes N} \to \mathbb{R}$ factors uniquely as $f_N = g_N \circ e_N$ for a function $g_N : E_N \to \mathbb{R}$. The continuous surjection $e_N : \Omega^{\otimes N} \to E_N$ is a quotient map, since its domain is compact and its codomain is Hausdorff. Hence g_N is continuous. Moreover, $\mathcal{P}(\Omega)$ is metrizable and thus normal, so the Tietze extension theorem provides $\tilde{f}_N \in C(\mathcal{P}(\Omega))$ satisfying $\tilde{f}_N|_{E_N} = g_N$. In particular,

$$f_N(\boldsymbol{X}) = \tilde{f}_N(\hat{\rho}_{\boldsymbol{X}}), \qquad \boldsymbol{X} \in \Omega^{\otimes N}.$$

The measure-input architectures likewise recover their sequence-input counterparts upon evaluation at empirical measures. For linear attention,

$$\mathrm{Attn}_{\hat{\rho}_{\boldsymbol{X}}}(\boldsymbol{x}_i) = \frac{1}{N} \sum_{j=1}^{N} \langle \mathcal{A}_Q(\boldsymbol{x}_i), \mathcal{A}_K(\boldsymbol{x}_j) \rangle \mathcal{A}_V(\boldsymbol{x}_j) = \mathrm{Attn}_i(\boldsymbol{X}),$$

and therefore $T(\hat{\rho}_{\boldsymbol{X}}) = T(\boldsymbol{X})$. Similarly, integration against $\hat{\rho}_{\boldsymbol{X}}^{\otimes r}$ averages over all ordered r-tuples, including repeated indices, giving

$$g_r(\hat{\rho}_{\boldsymbol{X}}) = \phi_2 \left(\frac{1}{N^r} \sum_{1 \leq i_1, \ldots, i_r \leq N} \phi_1(\boldsymbol{x}_{i_1}, \ldots, \boldsymbol{x}_{i_r}) \right) = g_r(\boldsymbol{X}).$$

For any $\varepsilon > 0$, applying Proposition 3.2 to $\tilde{f}_N$ yields $T \in \mathcal{T}$ such that

$$\sup_{\boldsymbol{X} \in \Omega^{\otimes N}} |f_N(\boldsymbol{X}) - T(\boldsymbol{X})| \leq \sup_{\rho \in \mathcal{P}(\Omega)} |\tilde{f}_N(\rho) - T(\rho)| < \varepsilon.$$

Proposition 3.1 gives the same conclusion for Janossy pooling networks of every fixed arity $r \geq 1$. Thus the statistical-functional UAP results recover universal approximation of continuous permutation-invariant functions on fixed-length sequences. The approximating models may depend on N, f_N, and ε.

Query-dependent targets. We justify Remark 3.5 by supplying the query $\boldsymbol{x} \in \Omega$ directly to the feature maps. The extended Transformer takes the form

$$T(\rho, \boldsymbol{x}) = \varphi_2 \left(\int_{\Omega} \varphi_1(\mathrm{Attn}_\rho(\boldsymbol{y}), \boldsymbol{x}) \, \mathrm{d}\rho(\boldsymbol{y}) \right),$$

where $\varphi_1 \in C(\mathbb{R}^{d_v} \times \mathbb{R}^d; \mathbb{R}^M)$ and $\varphi_2 \in C(\mathbb{R}^M)$. The continuity argument in the proof of Proposition 3.2 extends to joint continuity in $(\rho, \boldsymbol{x})$ for every such model.

For the fixed attention maps in equation 5, consider the pooled features

$$J_\psi(\rho, \boldsymbol{x}) \coloneqq \int_{\Omega} \psi(\Sigma_\rho \boldsymbol{y}, \boldsymbol{x}) \, \mathrm{d}\rho(\boldsymbol{y}), \qquad \psi \in C(\mathbb{B}^d(R^3) \times \Omega).$$

These features are jointly continuous by the same two-term estimate used above, together with uniform continuity of ψ on $\mathbb{B}^d(R^3) \times \bar{\Omega}$ and uniform convergence of the attention outputs. They also separate points of $\mathcal{P}(\Omega) \times \Omega$. Indeed, distinct measures are distinguished by query-independent features, using Lemma A.3 and uniqueness in the Riesz representation theorem. Distinct queries are distinguished by a coordinate function $\psi(\boldsymbol{z}, \boldsymbol{x}) = x_k$, since $J_\psi(\rho, \boldsymbol{x}) = x_k$.

Finite polynomial combinations of these features therefore form a subalgebra containing the constants and separating points. By the Tietze extension theorem, each $\psi \in C(\mathbb{B}^d(R^3) \times \Omega)$ admits a

continuous extension $\bar{\psi}$ to $\mathbb{R}^{2d}$. Consequently, every such polynomial combination is realized by the extended Transformer with $\varphi_1 = (\bar{\psi}_1, \dots, \bar{\psi}_M)^\top$ and a polynomial readout φ_2. Since $\mathcal{P}(\Omega) \times \Omega$ is compact Hausdorff, Theorem C.6 yields uniform density of the extended Transformer class in $C(\mathcal{P}(\Omega) \times \Omega)$.

For query-dependent Janossy pooling, the same argument uses

$$I_\chi(\rho, \boldsymbol{x}) \coloneqq \int_\Omega \chi(\boldsymbol{y}, \boldsymbol{x}) \, \mathrm{d}\rho(\boldsymbol{y}), \qquad \chi \in C(\Omega \times \Omega).$$

These features are jointly continuous and separate distinct measures and queries in the same manner, without the attention-induced transformation Υ. For any fixed $r \geq 1$, a finite collection of these features is realized by choosing

$$\phi_1(\boldsymbol{y}_1, \dots, \boldsymbol{y}_r, \boldsymbol{x}) = \big(\bar{\chi}_1(\boldsymbol{y}_1, \boldsymbol{x}), \dots, \bar{\chi}_M(\boldsymbol{y}_1, \boldsymbol{x})\big)^\top,$$

where each $\bar{\chi}_j$ is a continuous extension of χ_j to $\mathbb{R}^{2d}$. Integration over the remaining $r-1$ variables contributes a factor of one, so the pooled representation equals $(I_{\chi_1}(\rho, \boldsymbol{x}), \dots, I_{\chi_M}(\rho, \boldsymbol{x}))^\top$. Taking ϕ_2 to be a polynomial realizes the corresponding feature algebra. Theorem C.6 therefore also yields UAP on $\mathcal{P}(\Omega) \times \Omega$ for query-dependent Janossy pooling networks of every fixed arity $r \geq 1$.

B Exponential Separations

In this section, we provide detailed proofs of the main separation results for linear Transformers and r-ary Janossy pooling networks in approximating statistical functionals.

B.1 Separations between Deepsets and linear Transformers

In this subsection, we provide a detailed proof of the exponential separation between DeepSets and linear Transformers under the $L^2(\pi_d)$ metric, as established in Theorem 3.6. We first verify the continuity of the hard functional.

Proposition B.1 (Continuity of the hard functional)**.** *For $\rho \in \mathcal{P}(\mathbb{S}^{d-1})$, define*

$$\boldsymbol{m}_\rho = \int \boldsymbol{x} \, \mathrm{d}\rho(\boldsymbol{x}), \qquad h(\rho) = \int |\langle \boldsymbol{x}, \boldsymbol{m}_\rho \rangle| \, \mathrm{d}\rho(\boldsymbol{x}).$$

Then $h \in C(\mathcal{P}(\mathbb{S}^{d-1}))$.

Proof. Since $\mathbb{S}^{d-1}$ is compact, the weak-* topology on $\mathcal{P}(\mathbb{S}^{d-1})$ is metrized by the 1-Wasserstein distance (Villani, 2009, Corollary 6.13), defined by

$$W_1(\rho_1, \rho_2) \coloneqq \inf_{\pi \in \Pi(\rho_1, \rho_2)} \int_{\mathbb{S}^{d-1} \times \mathbb{S}^{d-1}} \|\boldsymbol{x} - \boldsymbol{y}\|_1 \, \mathrm{d}\pi(\boldsymbol{x}, \boldsymbol{y}),$$

where $\Pi(\rho_1, \rho_2)$ denotes the set of couplings of ρ_1 and ρ_2. It therefore suffices to prove continuity with respect to W_1.

Fix $\rho_1, \rho_2 \in \mathcal{P}(\mathbb{S}^{d-1})$ and $\pi \in \Pi(\rho_1, \rho_2)$, and write $\boldsymbol{m}_i = \boldsymbol{m}_{\rho_i}$ for $i \in \{1, 2\}$. By the marginal constraints of π,

$$\begin{aligned} |h(\rho_1) - h(\rho_2)| &= \left| \int |\langle \boldsymbol{x}, \boldsymbol{m}_1 \rangle| \, \mathrm{d}\pi(\boldsymbol{x}, \boldsymbol{y}) - \int |\langle \boldsymbol{y}, \boldsymbol{m}_2 \rangle| \, \mathrm{d}\pi(\boldsymbol{x}, \boldsymbol{y}) \right| \\ &\leq \int \big| |\langle \boldsymbol{x}, \boldsymbol{m}_1 \rangle| - |\langle \boldsymbol{y}, \boldsymbol{m}_2 \rangle| \big| \, \mathrm{d}\pi(\boldsymbol{x}, \boldsymbol{y}). \end{aligned}$$

For every $(\boldsymbol{x}, \boldsymbol{y}) \in \mathbb{S}^{d-1} \times \mathbb{S}^{d-1}$, the reverse triangle inequality and Hölder's inequality yield

$$\begin{aligned} \big| |\langle \boldsymbol{x}, \boldsymbol{m}_1 \rangle| - |\langle \boldsymbol{y}, \boldsymbol{m}_2 \rangle| \big| &\leq |\langle \boldsymbol{x}, \boldsymbol{m}_1 \rangle - \langle \boldsymbol{y}, \boldsymbol{m}_2 \rangle| \\ &= |\langle \boldsymbol{x} - \boldsymbol{y}, \boldsymbol{m}_1 \rangle + \langle \boldsymbol{y}, \boldsymbol{m}_1 - \boldsymbol{m}_2 \rangle| \\ &\leq \|\boldsymbol{x} - \boldsymbol{y}\|_1 \|\boldsymbol{m}_1\|_\infty + \|\boldsymbol{y}\|_\infty \|\boldsymbol{m}_1 - \boldsymbol{m}_2\|_1. \end{aligned}$$

Since $\|\boldsymbol{y}\|_\infty \le 1$ and $\|\boldsymbol{m}_1\|_\infty \le 1$, it follows that

$$||\langle \boldsymbol{x}, \boldsymbol{m}_1\rangle| - |\langle \boldsymbol{y}, \boldsymbol{m}_2\rangle|| \le \|\boldsymbol{x}-\boldsymbol{y}\|_1 + \|\boldsymbol{m}_1 - \boldsymbol{m}_2\|_1.$$

Consequently,

$$|h(\rho_1) - h(\rho_2)| \le \int \|\boldsymbol{x}-\boldsymbol{y}\|_1 \,\mathrm{d}\pi(\boldsymbol{x},\boldsymbol{y}) + \|\boldsymbol{m}_1 - \boldsymbol{m}_2\|_1.$$

Moreover, again using the marginal constraints of π,

$$\|\boldsymbol{m}_1 - \boldsymbol{m}_2\|_1 = \left\| \int (\boldsymbol{x}-\boldsymbol{y}) \,\mathrm{d}\pi(\boldsymbol{x},\boldsymbol{y}) \right\|_1 \le \int \|\boldsymbol{x}-\boldsymbol{y}\|_1 \,\mathrm{d}\pi(\boldsymbol{x},\boldsymbol{y}).$$

Therefore,

$$|h(\rho_1) - h(\rho_2)| \le 2 \int \|\boldsymbol{x}-\boldsymbol{y}\|_1 \,\mathrm{d}\pi(\boldsymbol{x},\boldsymbol{y}).$$

Taking the infimum over $\pi \in \Pi(\rho_1,\rho_2)$ gives

$$|h(\rho_1) - h(\rho_2)| \le 2W_1(\rho_1,\rho_2).$$

Hence, h is W_1-Lipschitz and therefore weak-* continuous. □

We next verify that harmonic perturbations of the uniform measure define valid probability measures.

Proposition B.2 (Validity of the measure construction)**.** *Let $d \ge 3$, let $q \ge 2$ be even, and let $\eta \in (0, \frac{1}{2})$. For $\boldsymbol{A} \in \mathbb{S}^{L_{d,q}-1}$ and $\boldsymbol{b} \in \mathbb{S}^{d-1}$, define $\rho_{\boldsymbol{A},\boldsymbol{b}}$ through*

$$\frac{\mathrm{d}\rho_{\boldsymbol{A},\boldsymbol{b}}}{\mathrm{d}\nu_d}(\boldsymbol{x}) = 1 + \eta\,\langle \boldsymbol{b}, \boldsymbol{x}\rangle + \eta H_{\boldsymbol{A}}(\boldsymbol{x}), \quad H_{\boldsymbol{A}}(\boldsymbol{x}) \coloneqq \frac{1}{\sqrt{L_{d,q}}}\,\langle \boldsymbol{A}, \boldsymbol{Y}_{d,q}(\boldsymbol{x})\rangle\,.$$

Then $\rho_{\boldsymbol{A},\boldsymbol{b}} \in \mathcal{P}(\mathbb{S}^{d-1})$.

Proof. By Lemma 4.1 (ii) and the Cauchy–Schwarz inequality,

$$|H_{\boldsymbol{A}}(\boldsymbol{x})| \le \frac{\|\boldsymbol{A}\|_2 \|\boldsymbol{Y}_{d,q}(\boldsymbol{x})\|_2}{\sqrt{L_{d,q}}} = 1.$$

Since $|\langle \boldsymbol{b}, \boldsymbol{x}\rangle| \le 1$, the density satisfies

$$1 + \eta\langle \boldsymbol{b}, \boldsymbol{x}\rangle + \eta H_{\boldsymbol{A}}(\boldsymbol{x}) \ge 1 - 2\eta > 0.$$

Moreover, $\boldsymbol{x} \mapsto \langle \boldsymbol{b}, \boldsymbol{x}\rangle$ belongs to $\mathcal{H}_{d,1}$, while $H_{\boldsymbol{A}} \in \mathcal{H}_{d,q}$ by Lemma 4.1 (i). Both are orthogonal to the constant functions in $\mathcal{H}_{d,0}$, again by Lemma 4.1(i), and hence

$$\int_{\mathbb{S}^{d-1}} \langle \boldsymbol{b}, \boldsymbol{x}\rangle \,\mathrm{d}\nu_d(\boldsymbol{x}) = \int_{\mathbb{S}^{d-1}} H_{\boldsymbol{A}}(\boldsymbol{x}) \,\mathrm{d}\nu_d(\boldsymbol{x}) = 0.$$

Therefore,

$$\int_{\mathbb{S}^{d-1}} \frac{\mathrm{d}\rho_{\boldsymbol{A},\boldsymbol{b}}}{\mathrm{d}\nu_d}(\boldsymbol{x}) \,\mathrm{d}\nu_d(\boldsymbol{x}) = 1.$$

Thus, $\rho_{\boldsymbol{A},\boldsymbol{b}}$ has a nonnegative density satisfying the probability normalization condition, and hence $\rho_{\boldsymbol{A},\boldsymbol{b}} \in \mathcal{P}(\mathbb{S}^{d-1})$. □

On this measure family, we normalize the hard functional and derive the harmonic encoding representation.

Proof of Lemma 4.2. Since each coordinate function belongs to $\mathcal{H}_{d,1}$ and $H_{\boldsymbol{A}} \in \mathcal{H}_{d,q}$ with $q \ge 2$, orthogonality to constants and between distinct harmonic degrees in Lemma 4.1 (i) gives

$$\int_{\mathbb{S}^{d-1}} \boldsymbol{x} \,\mathrm{d}\nu_d(\boldsymbol{x}) = \boldsymbol{0}, \qquad \int_{\mathbb{S}^{d-1}} \boldsymbol{x} H_{\boldsymbol{A}}(\boldsymbol{x}) \,\mathrm{d}\nu_d(\boldsymbol{x}) = \boldsymbol{0}.$$

Moreover, rotational invariance implies $\int_{\mathbb{S}^{d-1}} \boldsymbol{x}\boldsymbol{x}^\top \,\mathrm{d}\nu_d(\boldsymbol{x}) = \boldsymbol{I}_d/d$. Consequently,

$$\begin{aligned}\boldsymbol{m}_{\rho_{\boldsymbol{A},\boldsymbol{b}}} &= \int_{\mathbb{S}^{d-1}} \boldsymbol{x}\left(1+\eta\langle \boldsymbol{b},\boldsymbol{x}\rangle + \eta H_{\boldsymbol{A}}(\boldsymbol{x})\right)\,\mathrm{d}\nu_d(\boldsymbol{x}) \\ &= \eta\left(\int_{\mathbb{S}^{d-1}} \boldsymbol{x}\boldsymbol{x}^\top \,\mathrm{d}\nu_d(\boldsymbol{x})\right)\boldsymbol{b} = \frac{\eta}{d}\boldsymbol{b}.\end{aligned}$$

To evaluate h, we invoke the rotational invariance of ν_d to obtain, for any $\boldsymbol{b}\in\mathbb{S}^{d-1}$,

$$\int_{\mathbb{S}^{d-1}} |\langle \boldsymbol{b},\boldsymbol{x}\rangle|\,\mathrm{d}\nu_d(\boldsymbol{x}) = \int_{\mathbb{S}^{d-1}} |\langle \mathbf{e}_1,\boldsymbol{x}\rangle|\,\mathrm{d}\nu_d(\boldsymbol{x}) = \int_{\mathbb{S}^{d-1}} |x_1|\,\mathrm{d}\nu_d(\boldsymbol{x}).$$

Since $x_1^2 \sim \mathrm{Beta}(\frac{1}{2},\frac{d-1}{2})$ for $\boldsymbol{x}\sim\nu_d$, it follows that

$$\int_{\mathbb{S}^{d-1}} |x_1|\,\mathrm{d}\nu_d(\boldsymbol{x}) = \frac{\Gamma(d/2)}{\sqrt{\pi}\Gamma((d-1)/2)}\int_{-1}^{1}|t|(1-t^2)^{(d-3)/2}\,\mathrm{d}t = \frac{\Gamma(d/2)}{\sqrt{\pi}\Gamma((d+1)/2)} = \frac{d\beta_d}{\eta}.$$

The antipodal symmetry of ν_d also gives

$$\int_{\mathbb{S}^{d-1}} |\langle \boldsymbol{b},\boldsymbol{x}\rangle|\langle \boldsymbol{b},\boldsymbol{x}\rangle\,\mathrm{d}\nu_d(\boldsymbol{x}) = 0.$$

Since $H_{\boldsymbol{A}}\in\mathcal{H}_{d,q}$, the cosine-transform identity in Lemma 4.1 (iii) implies

$$\int_{\mathbb{S}^{d-1}} |\langle \boldsymbol{b},\boldsymbol{x}\rangle| H_{\boldsymbol{A}}(\boldsymbol{x})\,\mathrm{d}\nu_d(\boldsymbol{x}) = \gamma_{d,q}H_{\boldsymbol{A}}(\boldsymbol{b}) = \frac{\gamma_{d,q}}{\sqrt{L_{d,q}}}\langle \boldsymbol{A},\boldsymbol{Y}_{d,q}(\boldsymbol{b})\rangle.$$

Combining these identities with $\boldsymbol{m}_{\rho_{\boldsymbol{A},\boldsymbol{b}}} = \eta\boldsymbol{b}/d$, we obtain

$$\begin{aligned}h(\rho_{\boldsymbol{A},\boldsymbol{b}}) &= \frac{\eta}{d}\int_{\mathbb{S}^{d-1}} |\langle \boldsymbol{b},\boldsymbol{x}\rangle|\left(1+\eta\langle \boldsymbol{b},\boldsymbol{x}\rangle+\eta H_{\boldsymbol{A}}(\boldsymbol{x})\right)\,\mathrm{d}\nu_d(\boldsymbol{x}) \\ &= \beta_d + \frac{\eta^2\gamma_{d,q}}{d\sqrt{L_{d,q}}}\langle \boldsymbol{A},\boldsymbol{Y}_{d,q}(\boldsymbol{b})\rangle.\end{aligned}$$

For even $q\geq 2$, Lemma 4.1 (iii) ensures that $\gamma_{d,q}\neq 0$. Thus, the definition of $c_{d,q}$ yields

$$f_{d,q}(\rho_{\boldsymbol{A},\boldsymbol{b}}) = c_{d,q}\left(h(\rho_{\boldsymbol{A},\boldsymbol{b}})-\beta_d\right) = \langle \boldsymbol{A},\boldsymbol{Y}_{d,q}(\boldsymbol{b})\rangle.$$

Finally, independence of $\boldsymbol{A}$ and $\boldsymbol{b}$, together with the orthonormality identity in Lemma 4.1 (ii), gives

$$\begin{aligned}\|f_{d,q}(\rho_{\boldsymbol{A},\boldsymbol{b}})\|^2_{L^2(\pi_{d,q})} &= \mathbb{E}_{\boldsymbol{A}}\left[\boldsymbol{A}^\top\left(\int_{\mathbb{S}^{d-1}}\boldsymbol{Y}_{d,q}(\boldsymbol{b})\boldsymbol{Y}_{d,q}(\boldsymbol{b})^\top\,\mathrm{d}\nu_d(\boldsymbol{b})\right)\boldsymbol{A}\right] \\ &= \mathbb{E}_{\boldsymbol{A}}\left[\|\boldsymbol{A}\|_2^2\right] = 1,\end{aligned}$$

completing the proof. □

We now characterize the information retained by the pooled DeepSets representation.

Proof of Lemma 4.4. Define

$$\boldsymbol{z}_0 \coloneqq \int_{\mathbb{S}^{d-1}}\phi_1(\boldsymbol{x})\,\mathrm{d}\nu_d(\boldsymbol{x}), \qquad \boldsymbol{U}\coloneqq\int_{\mathbb{S}^{d-1}}\phi_1(\boldsymbol{x})\boldsymbol{x}^\top\,\mathrm{d}\nu_d(\boldsymbol{x}),$$

and

$$\boldsymbol{W}\coloneqq\int_{\mathbb{S}^{d-1}}\phi_1(\boldsymbol{x})\boldsymbol{Y}_{d,q}(\boldsymbol{x})^\top\,\mathrm{d}\nu_d(\boldsymbol{x}).$$

These quantities are well-defined by continuity on the compact sphere, have the stated dimensions, and are independent of $\boldsymbol{A}$ and $\boldsymbol{b}$. Substituting the density of $\rho_{\boldsymbol{A},\boldsymbol{b}}$ and using linearity of integration yield

$$\begin{aligned}Z(\boldsymbol{A},\boldsymbol{b}) &= \int_{\mathbb{S}^{d-1}}\phi_1(\boldsymbol{x})\left(1+\eta\boldsymbol{x}^\top\boldsymbol{b}+\frac{\eta}{\sqrt{L_{d,q}}}\boldsymbol{Y}_{d,q}(\boldsymbol{x})^\top\boldsymbol{A}\right)\,\mathrm{d}\nu_d(\boldsymbol{x}) \\ &= \boldsymbol{z}_0+\eta\boldsymbol{U}\boldsymbol{b}+\frac{\eta}{\sqrt{L_{d,q}}}\boldsymbol{W}\boldsymbol{A},\end{aligned}$$

which proves the claim. □

To analyze optimal recovery, we recall the mean-square optimality of conditional expectation.

Lemma B.3 (Conditional expectation as the optimal predictor)**.** *Let Y be a square-integrable real-valued random variable and $\mathcal{F}$ a sub-σ-algebra. Denote by $L^2(\mathcal{F})$ the space of square-integrable $\mathcal{F}$-measurable random variables. For every $H \in L^2(\mathcal{F})$,*

$$\mathbb{E}|Y - H|^2 = \mathbb{E}|Y - \mathbb{E}[Y \mid \mathcal{F}]|^2 + \mathbb{E}|\mathbb{E}[Y \mid \mathcal{F}] - H|^2. \tag{10}$$

Consequently,

$$\inf_{H \in L^2(\mathcal{F})} \mathbb{E}|Y - H|^2 = \mathbb{E}|Y - \mathbb{E}[Y \mid \mathcal{F}]|^2, \tag{11}$$

with the unique minimizer $H = \mathbb{E}[Y \mid \mathcal{F}]$, up to almost-sure equality.

Proof. Set $Y_{\mathcal{F}} := \mathbb{E}[Y \mid \mathcal{F}]$. Conditional Jensen's inequality gives $\mathbb{E}|Y_{\mathcal{F}}|^2 \le \mathbb{E}|Y|^2$. By the defining property of conditional expectation,

$$\mathbb{E}[(Y - Y_{\mathcal{F}})Z] = 0$$

for every bounded $\mathcal{F}$-measurable Z. Truncation and the Cauchy–Schwarz inequality extend this identity to every $Z \in L^2(\mathcal{F})$. Taking $Z = Y_{\mathcal{F}} - H$ and expanding $Y - H = (Y - Y_{\mathcal{F}}) + (Y_{\mathcal{F}} - H)$ yields equation 10. Its last term is nonnegative and vanishes precisely when $H = Y_{\mathcal{F}}$ almost surely, proving equation 11 and uniqueness. □

This characterization reduces optimal prediction from $(\boldsymbol{b}, \boldsymbol{W}\boldsymbol{A})$ to a conditional-expectation calculation.

Proof of Lemma 4.5. Throughout the proof, $\sigma(X)$ denotes the σ-algebra generated by a random variable X, i.e., the information contained in the observation of X. Since $\ker(\boldsymbol{W}) = \boldsymbol{S}^{\perp}$, we have $\boldsymbol{W} = \boldsymbol{W}P_{\boldsymbol{S}}$, and hence $\sigma(\boldsymbol{W}\boldsymbol{A}) \subseteq \sigma(P_{\boldsymbol{S}}\boldsymbol{A})$. The restriction $\boldsymbol{W}|_{\boldsymbol{S}} : \boldsymbol{S} \to \mathrm{range}(\boldsymbol{W})$ is injective because $\boldsymbol{S} \cap \ker(\boldsymbol{W}) = \{0\}$, and surjective because $\boldsymbol{W} = \boldsymbol{W}P_{\boldsymbol{S}}$. Therefore,

$$P_{\boldsymbol{S}}\boldsymbol{A} = (\boldsymbol{W}|_{\boldsymbol{S}})^{-1}\boldsymbol{W}\boldsymbol{A}.$$

Since both spaces are finite-dimensional, $(\boldsymbol{W}|_{\boldsymbol{S}})^{-1}$ is continuous, and hence measurable, by the bounded inverse theorem (Yosida, 2012). It follows that $\sigma(P_{\boldsymbol{S}}\boldsymbol{A}) \subseteq \sigma(\boldsymbol{W}\boldsymbol{A})$. Consequently, $\sigma(P_{\boldsymbol{S}}\boldsymbol{A}) = \sigma(\boldsymbol{W}\boldsymbol{A})$, so the projection $P_{\boldsymbol{S}}\boldsymbol{A}$ is completely determined by the observation $\boldsymbol{W}\boldsymbol{A}$.

To identify the conditional mean of the remaining component, consider the orthogonal reflection $R := P_{\boldsymbol{S}} - P_{\boldsymbol{S}^{\perp}}$. It fixes $\boldsymbol{S}$ and reverses $\boldsymbol{S}^{\perp}$, so $\boldsymbol{W}R = \boldsymbol{W}$ and $P_{\boldsymbol{S}^{\perp}}R = -P_{\boldsymbol{S}^{\perp}}$. Orthogonal invariance of the uniform spherical distribution implies that $R\boldsymbol{A}$ and $\boldsymbol{A}$ have the same distribution. Hence, for every bounded Borel function $\zeta : \mathbb{R}^M \to \mathbb{R}$,

$$\mathbb{E}[P_{\boldsymbol{S}^{\perp}}\boldsymbol{A}\,\zeta(\boldsymbol{W}\boldsymbol{A})] = \mathbb{E}[P_{\boldsymbol{S}^{\perp}}R\boldsymbol{A}\,\zeta(\boldsymbol{W}R\boldsymbol{A})] = -\mathbb{E}[P_{\boldsymbol{S}^{\perp}}\boldsymbol{A}\,\zeta(\boldsymbol{W}\boldsymbol{A})]\,.$$

Thus, the expectation vanishes. By the defining characterization of conditional expectation,

$$\mathbb{E}[P_{\boldsymbol{S}^{\perp}}\boldsymbol{A} \mid \boldsymbol{W}\boldsymbol{A}] = 0.$$

Combining this with the $\sigma(\boldsymbol{W}\boldsymbol{A})$-measurability of $P_{\boldsymbol{S}}\boldsymbol{A}$ yields

$$\mathbb{E}[\boldsymbol{A} \mid \boldsymbol{W}\boldsymbol{A}] = \mathbb{E}[P_{\boldsymbol{S}}\boldsymbol{A} + P_{\boldsymbol{S}^{\perp}}\boldsymbol{A} \mid \boldsymbol{W}\boldsymbol{A}] = P_{\boldsymbol{S}}\boldsymbol{A} \qquad \text{almost surely.}$$

Now set

$$Y := f_{d,q}(\rho_{\boldsymbol{A},\boldsymbol{b}}) = \langle \boldsymbol{A}, \boldsymbol{Y}_{d,q}(\boldsymbol{b}) \rangle, \qquad \mathcal{F} := \sigma(\boldsymbol{b}, \boldsymbol{W}\boldsymbol{A}),$$

where $\mathcal{F}$ is the σ-algebra generated by the observable pair $(\boldsymbol{b}, \boldsymbol{W}\boldsymbol{A})$, and therefore represents all information available to a predictor based on these observations. Since $\boldsymbol{b}$ is independent of $(\boldsymbol{A}, \boldsymbol{W}\boldsymbol{A})$,

$$\mathbb{E}[\boldsymbol{A} \mid \boldsymbol{b}, \boldsymbol{W}\boldsymbol{A}] = \mathbb{E}[\boldsymbol{A} \mid \boldsymbol{W}\boldsymbol{A}] = P_{\boldsymbol{S}}\boldsymbol{A}.$$

Since $\boldsymbol{Y}_{d,q}(\boldsymbol{b})$ is bounded and $\mathcal{F}$-measurable,

$$\widehat{Y} := \mathbb{E}[Y \mid \mathcal{F}] = \langle \mathbb{E}[\boldsymbol{A} \mid \boldsymbol{b}, \boldsymbol{W}\boldsymbol{A}], \boldsymbol{Y}_{d,q}(\boldsymbol{b}) \rangle = \langle P_{\boldsymbol{S}}\boldsymbol{A}, \boldsymbol{Y}_{d,q}(\boldsymbol{b}) \rangle.$$

Every $g \in \mathcal{H}_{\boldsymbol{W}}$ defines a predictor $g(\boldsymbol{b}, \boldsymbol{W}\boldsymbol{A}) \in L^2(\mathcal{F})$. Moreover, $\widehat{Y}$ is attained by an admissible predictor: define

$$g_*(\boldsymbol{b}, \boldsymbol{z}) := \left\langle (\boldsymbol{W}|_{\boldsymbol{S}})^{-1}\boldsymbol{z}, \boldsymbol{Y}_{d,q}(\boldsymbol{b}) \right\rangle \quad \text{for } \boldsymbol{z} \in \operatorname{range}(\boldsymbol{W}),$$

and extend it by zero outside this closed subspace. Then g_* is Borel measurable and $g_*(\boldsymbol{b}, \boldsymbol{W}\boldsymbol{A}) = \widehat{Y} \in L^2(\pi_{d,q})$. Applying Lemma B.3 therefore gives

$$\inf_{g \in \mathcal{H}_{\boldsymbol{W}}} \mathbb{E}_{\boldsymbol{A},\boldsymbol{b}} \left| f_{d,q}(\rho_{\boldsymbol{A},\boldsymbol{b}}) - g(\boldsymbol{b}, \boldsymbol{W}\boldsymbol{A}) \right|^2 = \mathbb{E}|Y - \widehat{Y}|^2 = \mathbb{E}_{\boldsymbol{A},\boldsymbol{b}} \left| \langle P_{\boldsymbol{S}^\perp}\boldsymbol{A}, \boldsymbol{Y}_{d,q}(\boldsymbol{b}) \rangle \right|^2.$$

The same lemma establishes uniqueness of the optimal predictor up to $\pi_{d,q}$-almost-sure equality. □

We next evaluate the harmonic energy remaining after this projection.

Proof of Lemma 4.6. Since $\boldsymbol{A}$ and $\boldsymbol{b}$ are independent, conditioning on $\boldsymbol{A}$ and applying Lemma 4.1(ii) yield

$$\begin{aligned}\mathbb{E}_{\boldsymbol{b}} \left[\left| \langle P_{\boldsymbol{S}^\perp}\boldsymbol{A}, \boldsymbol{Y}_{d,q}(\boldsymbol{b}) \rangle \right|^2 \,\middle|\, \boldsymbol{A} \right] &= (P_{\boldsymbol{S}^\perp}\boldsymbol{A})^\top \left(\int_{\mathbb{S}^{d-1}} \boldsymbol{Y}_{d,q}(\boldsymbol{b}) \boldsymbol{Y}_{d,q}(\boldsymbol{b})^\top \, \mathrm{d}\nu_d(\boldsymbol{b}) \right) P_{\boldsymbol{S}^\perp}\boldsymbol{A} \\ &= \|P_{\boldsymbol{S}^\perp}\boldsymbol{A}\|_2^2.\end{aligned}$$

Therefore,

$$\mathbb{E}_{\boldsymbol{A},\boldsymbol{b}} \left| \langle P_{\boldsymbol{S}^\perp}\boldsymbol{A}, \boldsymbol{Y}_{d,q}(\boldsymbol{b}) \rangle \right|^2 = \mathbb{E}_{\boldsymbol{A}} \|P_{\boldsymbol{S}^\perp}\boldsymbol{A}\|_2^2.$$

Since $\boldsymbol{A} \sim \nu_{L_{d,q}}$ is uniformly distributed on $\mathbb{S}^{L_{d,q}-1}$, rotational invariance implies

$$\mathbb{E}_{\boldsymbol{A}}[\boldsymbol{A}\boldsymbol{A}^\top] = \frac{1}{L_{d,q}} \boldsymbol{I}_{L_{d,q}}.$$

Hence,

$$\mathbb{E}_{\boldsymbol{A}} \|P_{\boldsymbol{S}^\perp}\boldsymbol{A}\|_2^2 = \mathbb{E}_{\boldsymbol{A}} \left[\boldsymbol{A}^\top P_{\boldsymbol{S}^\perp}\boldsymbol{A} \right] = \operatorname{tr}\left(P_{\boldsymbol{S}^\perp} \mathbb{E}_{\boldsymbol{A}}[\boldsymbol{A}\boldsymbol{A}^\top] \right) = \frac{\operatorname{tr}(P_{\boldsymbol{S}^\perp})}{L_{d,q}} = \frac{\dim(\boldsymbol{S}^\perp)}{L_{d,q}}.$$

Finally, since $\boldsymbol{S} = \operatorname{range}(\boldsymbol{W}^\top)$,

$$\dim(\boldsymbol{S}) = \operatorname{rank}(\boldsymbol{W}), \qquad \dim(\boldsymbol{S}^\perp) = L_{d,q} - \operatorname{rank}(\boldsymbol{W}).$$

Consequently,

$$\mathbb{E}_{\boldsymbol{A},\boldsymbol{b}} \left| \langle P_{\boldsymbol{S}^\perp}\boldsymbol{A}, \boldsymbol{Y}_{d,q}(\boldsymbol{b}) \rangle \right|^2 = 1 - \frac{\operatorname{rank}(\boldsymbol{W})}{L_{d,q}},$$

which proves the claim. □

The representation and recovery results together yield the quantitative DeepSets lower bound.

Proof of Lemma 4.7. Fix $g \in \mathcal{G}_M^{(1)}$. By Lemma 4.4, its prediction depends on $\boldsymbol{b}$ and $\boldsymbol{W}\boldsymbol{A}$ only through the affine combination

$$Z(\boldsymbol{A}, \boldsymbol{b}) = \boldsymbol{z}_0 + \eta \boldsymbol{U}\boldsymbol{b} + \frac{\eta}{\sqrt{L_{d,q}}} \boldsymbol{W}\boldsymbol{A}.$$

Consequently, its mean-square error cannot be smaller than that of the optimal predictor based on $(\boldsymbol{b}, \boldsymbol{W}\boldsymbol{A})$ characterized in Lemma 4.5. Applying Lemma 4.6 then yields

$$\|g(\rho_{\boldsymbol{A},\boldsymbol{b}}) - f_{d,q}(\rho_{\boldsymbol{A},\boldsymbol{b}})\|_{L^2(\pi_{d,q})}^2 \geq \mathbb{E}_{\boldsymbol{A},\boldsymbol{b}} |\langle P_{\boldsymbol{S}^\perp}\boldsymbol{A}, \boldsymbol{Y}_{d,q}(\boldsymbol{b}) \rangle|^2 = 1 - \frac{\operatorname{rank}(\boldsymbol{W})}{L_{d,q}} \geq \left(1 - \frac{M}{L_{d,q}} \right)_+,$$

where $\boldsymbol{S} := \operatorname{range}(\boldsymbol{W}^\top)$ and $\operatorname{rank}(\boldsymbol{W}) \leq \min\{M, L_{d,q}\}$. Taking square roots and the infimum over $g \in \mathcal{G}_M^{(1)}$ completes the proof. □

Combining this bound with the dimension estimates and an exact attention construction proves Theorem 3.6.

Proof of Theorem 3.6. Fix $d \geq 3$ and $\eta \in (0, \frac{1}{2})$, and set $q_d = 2\lfloor d/2 \rfloor$, $f = f_{d,q_d}$, and $\pi_d = \pi_{d,q_d}$.

Lower bound. Proposition B.2 ensures that $\rho_{\boldsymbol{A},\boldsymbol{b}} \in \mathcal{P}(\mathbb{S}^{d-1})$. By Proposition B.1 and Lemma 4.2,

$$f \in C(\mathcal{P}(\mathbb{S}^{d-1})), \qquad \|f(\rho_{\boldsymbol{A},\boldsymbol{b}})\|_{L^2(\pi_d)} = 1.$$

Let $\xi > 0$ be the universal lower-bound constant in Proposition C.3, so that $L_{d,q_d} \geq \xi 4^d/\sqrt{d}$, and choose $c_1 := \xi/2$. Whenever $M \leq c_1 4^d/\sqrt{d}$, we have

$$\frac{M}{L_{d,q_d}} \leq \frac{c_1 4^d/\sqrt{d}}{\xi 4^d/\sqrt{d}} = \frac{1}{2}.$$

Applying Lemma 4.7 with $q = q_d$ therefore gives

$$\inf_{g \in \mathcal{G}_M^{(1)}} \|g(\rho_{\boldsymbol{A},\boldsymbol{b}}) - f(\rho_{\boldsymbol{A},\boldsymbol{b}})\|_{L^2(\pi_d)} \geq \sqrt{1 - \frac{M}{L_{d,q_d}}} \geq \frac{1}{\sqrt{2}}.$$

Upper bound. Choose query–key dimension $d_k = d$, value dimension $d_v = 1$, and affine maps

$$\mathcal{A}_Q(\boldsymbol{x}) = \boldsymbol{x}, \qquad \mathcal{A}_K(\boldsymbol{y}) = \boldsymbol{y}, \qquad \mathcal{A}_V(\boldsymbol{y}) = 1.$$

The constant value map is affine, with zero linear part and bias 1. For every $\rho \in \mathcal{P}(\mathbb{S}^{d-1})$ and $\boldsymbol{x} \in \mathbb{S}^{d-1}$, these choices yield

$$\begin{aligned}\mathrm{Attn}_\rho(\boldsymbol{x}) &= \int_{\mathbb{S}^{d-1}} \langle \mathcal{A}_Q(\boldsymbol{x}), \mathcal{A}_K(\boldsymbol{y})\rangle \mathcal{A}_V(\boldsymbol{y})\, \mathrm{d}\rho(\boldsymbol{y}) \\ &= \int_{\mathbb{S}^{d-1}} \langle \boldsymbol{x}, \boldsymbol{y}\rangle \, \mathrm{d}\rho(\boldsymbol{y}) = \langle \boldsymbol{x}, \boldsymbol{m}_\rho\rangle.\end{aligned}$$

Define the scalar feature and output maps by

$$\begin{aligned}&\varphi_1(t) \coloneqq \mathrm{ReLU}(t) + \mathrm{ReLU}(-t) = |t|, \\ &\varphi_2(t) \coloneqq c_{d,q_d}\,\mathrm{ReLU}(t - \beta_d) - c_{d,q_d}\,\mathrm{ReLU}(\beta_d - t) = c_{d,q_d}(t - \beta_d).\end{aligned}$$

Both maps are realized by two-layer ReLU networks of hidden width 2. Since φ_1 has scalar output, the resulting Transformer has latent dimension $M = 1$. For every $\rho \in \mathcal{P}(\mathbb{S}^{d-1})$,

$$\begin{aligned}T(\rho) &= \varphi_2\left(\int_{\mathbb{S}^{d-1}} \varphi_1(\mathrm{Attn}_\rho(\boldsymbol{x}))\, \mathrm{d}\rho(\boldsymbol{x})\right) \\ &= c_{d,q_d}\left(\int_{\mathbb{S}^{d-1}} |\langle \boldsymbol{x}, \boldsymbol{m}_\rho\rangle|\, \mathrm{d}\rho(\boldsymbol{x}) - \beta_d\right) \\ &= c_{d,q_d}\big(h(\rho) - \beta_d\big) = f_{d,q_d}(\rho) = f(\rho).\end{aligned}$$

Thus $T \in \mathcal{T}_1$ exactly represents f on the entire probability-measure space, completing the proof. □

B.2 Lower bounds for networks with Janossy pooling

In this subsection, we give a detailed proof of the uniform approximation lower bound for general r-ary Janossy pooling networks.

Proof of Lemma 4.8. Define $\ell_{\boldsymbol{b}}(\boldsymbol{x}) := 1 + \eta\langle \boldsymbol{b}, \boldsymbol{x}\rangle$, so that

$$\frac{\mathrm{d}\rho_{\boldsymbol{A},\boldsymbol{b}}}{\mathrm{d}\nu_d}(\boldsymbol{x}) = \ell_{\boldsymbol{b}}(\boldsymbol{x}) + \eta H_{\boldsymbol{A}}(\boldsymbol{x}).$$

Substituting this density into the pooled representation gives

$$\boldsymbol{Z}^{(r)}(\boldsymbol{A}, \boldsymbol{b}) = \int_{(\mathbb{S}^{d-1})^r} \phi_1(\boldsymbol{x}_1, \ldots, \boldsymbol{x}_r) \prod_{i=1}^{r} \big(\ell_{\boldsymbol{b}}(\boldsymbol{x}_i) + \eta H_{\boldsymbol{A}}(\boldsymbol{x}_i)\big)\, \mathrm{d}\nu_d^{\otimes r}.$$

Since $H_{-\boldsymbol{A}} = -H_{\boldsymbol{A}}$, subtracting the expression with $\boldsymbol{A}$ replaced by $-\boldsymbol{A}$ cancels all terms containing an even number of harmonic factors. Writing $[r] := \{1, \ldots, r\}$, we obtain

$$\boldsymbol{\Delta}^{(r)}(\boldsymbol{A}, \boldsymbol{b}) = 2 \sum_{\substack{I \subseteq [r] \\ |I| \text{ odd}}} \eta^{|I|} \int_{(\mathbb{S}^{d-1})^r} \phi_1(\boldsymbol{x}_1, \ldots, \boldsymbol{x}_r) \times \prod_{i \in I} H_{\boldsymbol{A}}(\boldsymbol{x}_i) \prod_{i \in [r]\setminus I} \ell_{\boldsymbol{b}}(\boldsymbol{x}_i)\, \mathrm{d}\nu_d^{\otimes r}.$$

Every surviving term contains at least one harmonic factor. As $\ell_{\boldsymbol{b}}(\boldsymbol{x}_i)$ is affine in $\boldsymbol{b}$, the corresponding product has total degree at most $r - |I| \le r - 1$ in $\boldsymbol{b}$. Integration changes only its coefficients, not its degree. Thus, for each fixed $\boldsymbol{A}$ and every $j = 1, \dots, M$,

$$\boldsymbol{\Delta}_j^{(r)}(\boldsymbol{A}, \cdot) \in \mathscr{P}_{\le r-1}(\mathbb{S}^{d-1}).$$

For the fixed orthonormal basis $\{p_k\}_{k=1}^{K_{d,r-1}}$, define

$$c_{j,k}(\boldsymbol{A}) := \int_{\mathbb{S}^{d-1}} \boldsymbol{\Delta}_j^{(r)}(\boldsymbol{A}, \boldsymbol{b}) p_k(\boldsymbol{b}) \, \mathrm{d}\nu_d(\boldsymbol{b}).$$

The preceding membership and orthonormality yield

$$\boldsymbol{\Delta}_j^{(r)}(\boldsymbol{A}, \boldsymbol{b}) = \sum_{k=1}^{K_{d,r-1}} c_{j,k}(\boldsymbol{A}) p_k(\boldsymbol{b}).$$

This equality holds in $L^2(\nu_d)$ and hence ν_d-almost everywhere. Since both sides are continuous in $\boldsymbol{b}$ and ν_d has full support on $\mathbb{S}^{d-1}$, it holds pointwise for every $\boldsymbol{b} \in \mathbb{S}^{d-1}$.

Finally, $H_{\boldsymbol{A}}(\boldsymbol{x})$ is linear in $\boldsymbol{A}$, while $\ell_{\boldsymbol{b}}(\boldsymbol{x})$ is affine in $\boldsymbol{b}$. The finite product expansion therefore shows that $\boldsymbol{\Delta}_j^{(r)}(\boldsymbol{A}, \boldsymbol{b})$ is polynomial in $(\boldsymbol{A}, \boldsymbol{b})$. Integration against $p_k(\boldsymbol{b})$ with respect to ν_d preserves polynomial dependence on $\boldsymbol{A}$. Consequently, each $c_{j,k}(\boldsymbol{A})$ is polynomial in $\boldsymbol{A}$ and therefore continuous. Moreover, the definition of the antipodal difference gives $\boldsymbol{\Delta}^{(r)}(-\boldsymbol{A}, \boldsymbol{b}) = -\boldsymbol{\Delta}^{(r)}(\boldsymbol{A}, \boldsymbol{b})$, and hence

$$c_{j,k}(-\boldsymbol{A}) = \int_{\mathbb{S}^{d-1}} \boldsymbol{\Delta}_j^{(r)}(-\boldsymbol{A}, \boldsymbol{b}) p_k(\boldsymbol{b}) \, \mathrm{d}\nu_d(\boldsymbol{b}) = -c_{j,k}(\boldsymbol{A}).$$

Thus every coefficient map is continuous and odd. □

We use these coefficient maps to construct antipodal inputs with identical pooled representations.

Proof of Lemma 4.9. Fix the feature map ϕ_1 and an $L^2(\nu_d)$-orthonormal basis $\{p_k\}_{k=1}^{K_{d,r-1}}$ of $\mathscr{P}_{\le r-1}(\mathbb{S}^{d-1})$. Collect the coefficients from Lemma 4.8 into the map

$$\Pi : \mathbb{S}^{L_{d,q}-1} \to \mathbb{R}^{MK_{d,r-1}}, \qquad \Pi(\boldsymbol{A}) := \big(c_{j,k}(\boldsymbol{A})\big)_{1 \le j \le M,\, 1 \le k \le K_{d,r-1}}.$$

By Lemma 4.8, Π is continuous and odd. Since $MK_{d,r-1} < L_{d,q}$, we have $MK_{d,r-1} \le L_{d,q} - 1$. Lemma C.5 therefore gives $\boldsymbol{A}_* \in \mathbb{S}^{L_{d,q}-1}$ such that $\Pi(\boldsymbol{A}_*) = 0$.

Thus all coefficients $c_{j,k}(\boldsymbol{A}_*)$ vanish simultaneously. The expansion in Lemma 4.8 yields, for every $j = 1, \dots, M$ and $\boldsymbol{b} \in \mathbb{S}^{d-1}$,

$$\boldsymbol{\Delta}_j^{(r)}(\boldsymbol{A}_*, \boldsymbol{b}) = \sum_{k=1}^{K_{d,r-1}} c_{j,k}(\boldsymbol{A}_*) p_k(\boldsymbol{b}) = 0.$$

By the definition of $\boldsymbol{\Delta}^{(r)}$, this implies

$$\boldsymbol{Z}^{(r)}(\boldsymbol{A}_*, \boldsymbol{b}) = \boldsymbol{Z}^{(r)}(-\boldsymbol{A}_*, \boldsymbol{b}) \qquad \text{for every } \boldsymbol{b} \in \mathbb{S}^{d-1}.$$

□

We next select a barycenter direction separating the target values of these indistinguishable inputs.

Proof of Lemma 4.10. Fix $\boldsymbol{A}_* \in \mathbb{S}^{L_{d,q}-1}$. By orthonormality of the spherical harmonics,

$$\begin{aligned}\int_{\mathbb{S}^{d-1}} |\langle \boldsymbol{A}_*, \boldsymbol{Y}_{d,q}(\boldsymbol{b}) \rangle|^2 \, \mathrm{d}\nu_d(\boldsymbol{b}) &= \boldsymbol{A}_*^\top \left(\int_{\mathbb{S}^{d-1}} \boldsymbol{Y}_{d,q}(\boldsymbol{b}) \boldsymbol{Y}_{d,q}(\boldsymbol{b})^\top \, \mathrm{d}\nu_d(\boldsymbol{b}) \right) \boldsymbol{A}_* \\ &= \|\boldsymbol{A}_*\|_2^2 = 1.\end{aligned}$$

Since ν_d is a probability measure and $\boldsymbol{b} \mapsto |\langle \boldsymbol{A}_*, \boldsymbol{Y}_{d,q}(\boldsymbol{b})\rangle|$ is continuous on the compact sphere, there exists $\boldsymbol{b}_* \in \mathbb{S}^{d-1}$ such that $|\langle \boldsymbol{A}_*, \boldsymbol{Y}_{d,q}(\boldsymbol{b}_*)\rangle| \geq 1$. Lemma 4.2 then gives

$$f_{d,q}(\rho_{\pm \boldsymbol{A}_*, \boldsymbol{b}_*}) = \pm \langle \boldsymbol{A}_*, \boldsymbol{Y}_{d,q}(\boldsymbol{b}_*)\rangle.$$

Consequently,

$$|f_{d,q}(\rho_{\boldsymbol{A}_*, \boldsymbol{b}_*}) - f_{d,q}(\rho_{-\boldsymbol{A}_*, \boldsymbol{b}_*})| = 2|\langle \boldsymbol{A}_*, \boldsymbol{Y}_{d,q}(\boldsymbol{b}_*)\rangle| \geq 2,$$

which proves the claim. □

The preceding two lemmas together yield the uniform approximation lower bound.

Proof of Lemma 4.11. Fix $g \in \mathcal{G}_M^{(r)}$. For $r \geq 2$, Lemma 4.9 provides $\boldsymbol{A}_* \in \mathbb{S}^{L_{d,q}-1}$ whose antipodal inputs have identical pooled representations for every $\boldsymbol{b}$. For $r = 1$, the same conclusion follows from Lemma 4.4: since $K_{d,0} = 1$ and $M < L_{d,q}$, we may choose a unit vector $\boldsymbol{A}_* \in \ker(\boldsymbol{W})$. In either case, applying the common output map gives

$$g(\rho_{\boldsymbol{A}_*, \boldsymbol{b}}) = g(\rho_{-\boldsymbol{A}_*, \boldsymbol{b}}) \qquad \text{for every } \boldsymbol{b} \in \mathbb{S}^{d-1}.$$

Choose $\boldsymbol{b}_*$ as in Lemma 4.10, and write $\rho_\pm := \rho_{\pm \boldsymbol{A}_*, \boldsymbol{b}_*}$. Since $g(\rho_+) = g(\rho_-)$, the triangle inequality yields

$$\begin{aligned} 2 &\leq |f_{d,q}(\rho_+) - f_{d,q}(\rho_-)| \\ &\leq |f_{d,q}(\rho_+) - g(\rho_+)| + |g(\rho_-) - f_{d,q}(\rho_-)| \\ &\leq 2 \sup_{\rho \in \mathcal{P}(\mathbb{S}^{d-1})} |g(\rho) - f_{d,q}(\rho)|. \end{aligned}$$

Dividing by 2 and taking the infimum over $g \in \mathcal{G}_M^{(r)}$ proves the claim. □

The dimension estimates then give the exponential lower bound in Theorem 3.7.

Proof of Theorem 3.7. Fix $r \geq 1$ and $d \geq 3$, and set $q_d := 2\lfloor d/2 \rfloor$ and $f := f_{d,q_d}$, as in Theorem 3.6. By Proposition C.3,

$$L_{d,q_d} \geq \xi \frac{4^d}{\sqrt{d}}, \qquad K_{d,r-1} \leq \Xi_{r-1} d^{r-1},$$

where $\xi > 0$ is universal and $\Xi_{r-1} > 0$ depends only on r. Choose $c_r := \xi/(2\Xi_{r-1})$. Whenever $M \leq c_r 4^d / d^{r-1/2}$, we obtain

$$M K_{d,r-1} \leq c_r \Xi_{r-1} \frac{4^d}{\sqrt{d}} = \frac{\xi}{2} \frac{4^d}{\sqrt{d}} \leq \frac{1}{2} L_{d,q_d} < L_{d,q_d}.$$

Since $q_d \geq 2$ is even, Lemma 4.11 applied with $q = q_d$ yields

$$\inf_{g \in \mathcal{G}_M^{(r)}} \sup_{\rho \in \mathcal{P}(\mathbb{S}^{d-1})} |g(\rho) - f(\rho)| \geq 1,$$

which proves the theorem. □

C Auxiliary Tools

In this section, we collect several auxiliary mathematical results used throughout the paper.

Lemma C.1 (Central-binomial estimate (Graham et al., 1994)). *As $n \to \infty$,*

$$\binom{2n}{n} = \frac{4^n}{\sqrt{\pi n}} \left(1 - \frac{1}{8n} + O(n^{-2})\right).$$

In particular, there exist universal constants $a, b > 0$ such that

$$a \frac{4^n}{\sqrt{n}} \leq \binom{2n}{n} \leq b \frac{4^n}{\sqrt{n}}, \qquad n \geq 1.$$

Lemma C.2 (Pascal's identity (Graham et al., 1994, Eq. (5.8))). *For integers $n \geq 1$ and $0 \leq k \leq n$,*

$$\binom{n}{k} = \binom{n-1}{k} + \binom{n-1}{k-1},$$

where $\binom{m}{j} = 0$ whenever $j < 0$ or $j > m$.

Proposition C.3 (Dimension scales). *Set $q_d := 2\lfloor d/2 \rfloor$. There exist universal constants $\xi, \Xi > 0$ such that, for every integer $d \geq 3$,*

$$\xi \frac{4^d}{\sqrt{d}} \leq L_{d,q_d} \leq \Xi \frac{4^d}{\sqrt{d}}. \tag{12}$$

Moreover, for each integer $s \geq 0$, there exist constants $\xi_s, \Xi_s > 0$, depending only on s, such that

$$\xi_s d^s \leq K_{d,s} \leq \Xi_s d^s, \qquad d \geq 3. \tag{13}$$

Proof. By definition, $q_d = d$ for even d and $q_d = d - 1$ for odd d. Substituting these values into the definition of $L_{d,q}$ yields

$$L_{d,q_d} = \begin{cases} \left(\frac{3}{2} - \frac{1}{d}\right)\binom{2d-2}{d-1}, & d \text{ even}, \\ \frac{3d-4}{2(2d-3)}\binom{2d-2}{d-1}, & d \text{ odd}. \end{cases}$$

Both prefactors lie between $3/4$ and $3/2$ for $d \geq 3$. By Lemma C.1, there exist universal constants $a, b > 0$ such that

$$a\frac{4^{d-1}}{\sqrt{d-1}} \leq \binom{2d-2}{d-1} \leq b\frac{4^{d-1}}{\sqrt{d-1}}, \qquad d \geq 3.$$

Combining these inequalities with $1 \leq d/(d-1) \leq 3/2$ gives

$$\xi \frac{4^d}{\sqrt{d}} \leq L_{d,q_d} \leq \Xi \frac{4^d}{\sqrt{d}}$$

for universal constants $\xi, \Xi > 0$.

For the second assertion, $K_{d,0} = 1$, so we may take $\xi_0 = \Xi_0 = 1$. For $s \geq 1$, Lemma C.2 gives

$$K_{d,s} = \binom{d+s}{s} - \binom{d+s-2}{s-2} = \binom{d+s-1}{s} + \binom{d+s-2}{s-1},$$

where binomial coefficients with negative lower indices are zero. Hence,

$$K_{d,s} \geq \binom{d+s-1}{s} = \frac{1}{s!}\prod_{j=0}^{s-1}(d+j) \geq \frac{d^s}{s!}.$$

Conversely,

$$K_{d,s} \leq \binom{d+s}{s} \leq \frac{(d+s)^s}{s!} \leq \frac{(s+1)^s}{s!}\, d^s.$$

The claimed bounds therefore hold with $\xi_s = 1/s!$ and $\Xi_s = (s+1)^s/s!$ for $s \geq 1$. □

Lemma C.4 (Scale of the normalization factor c_{d,q_d}). *Let $d \geq 3$, $\eta \in (0, \frac{1}{2})$, and $q_d = 2\lfloor d/2 \rfloor$. The normalization factor defined in equation 4 satisfies*

$$|c_{d,q_d}| = \Theta\left(\eta^{-2} 4^d d^{9/4}\right), \tag{14}$$

where the implicit constants are independent of d and η.

Proof. For the cosine transform defined in Lemma 4.1(iii), the eigenvalue at every even degree $q \geq 2$ is

$$\gamma_{d,q} = (-1)^{q/2-1}\frac{\Gamma(d/2)\Gamma((q-1)/2)}{2\pi\Gamma((d+q+1)/2)}. \tag{15}$$

This follows from Ólafsson et al. (2012, Eqs. (2.1) and (2.6)) with $n = d$ and $\lambda = 1$, after removing their additional normalization factor $-2\pi/\Gamma(d/2)$.

Since $q_d = d$ for even d and $q_d = d - 1$ for odd d, the Gamma duplication formula

$$\Gamma(z)\Gamma(z + \tfrac{1}{2}) = 2^{1-2z}\sqrt{\pi}\,\Gamma(2z)$$

gives

$$|\gamma_{d,q_d}| = \begin{cases} \dfrac{2^{1-d}}{\sqrt{\pi}}\,\dfrac{\Gamma(d-1)}{\Gamma(d+\frac{1}{2})}, & d \text{ even}, \\ \dfrac{2^{-d}}{\sqrt{\pi}}\,\dfrac{\Gamma((d-2)/2)}{\Gamma((d+1)/2)}, & d \text{ odd}. \end{cases} \tag{16}$$

For fixed real numbers a and b, Stirling's formula yields

$$\frac{\Gamma(z+a)}{\Gamma(z+b)} = z^{a-b}\big(1 + O(z^{-1})\big), \qquad z \to \infty.$$

Applying this estimate to both cases in equation 16 gives

$$|\gamma_{d,q_d}| = \Theta\Big(2^{-d} d^{-3/2}\Big).$$

Moreover, Proposition C.3 implies $\sqrt{L_{d,q_d}} = \Theta(2^d d^{-1/4})$. Substituting these estimates into equation 4, we obtain

$$|c_{d,q_d}| = \frac{d\sqrt{L_{d,q_d}}}{\eta^2 |\gamma_{d,q_d}|} = \Theta\left(\frac{d\,2^d d^{-1/4}}{\eta^2\, 2^{-d} d^{-3/2}}\right) = \Theta\Big(\eta^{-2} 4^d d^{9/4}\Big),$$

which proves the claim. □

Lemma C.5 (Borsuk–Ulam theorem (Matoušek, 2003, Theorem 2.1.1))**.** *Let $1 \le m \le n$ be integers. Every continuous map $F : \mathbb{S}^n \to \mathbb{R}^m$ admits a point $u_* \in \mathbb{S}^n$ satisfying $F(u_*) = F(-u_*)$. In particular, if F is odd, then $F(u_*) = 0$.*

Proof. For $m = n$, the first assertion is the standard Borsuk–Ulam theorem. For $m < n$, apply the same theorem to the continuous map $\widetilde{F} : \mathbb{S}^n \to \mathbb{R}^n$ obtained by appending $n - m$ zero coordinates to F. If F is odd, then $F(u_*) = F(-u_*) = -F(u_*)$, which implies $F(u_*) = 0$. □

Theorem C.6 (Generalized Stone–Weierstrass theorem (Stone, 1948))**.** *Let $\mathcal{X}$ be a compact Hausdorff space, and suppose that $\mathscr{A} \subseteq C(\mathcal{X}; \mathbb{R})$ satisfies the following conditions:*

- *For all $f, g \in \mathscr{A}$ and $\alpha \in \mathbb{R}$, $f + g \in \mathscr{A}$, $fg \in \mathscr{A}$, and $\alpha f \in \mathscr{A}$.*
- *$\mathscr{A}$ contains all constant functions and separates points: for every pair of distinct points $x, y \in \mathcal{X}$, there exists $h \in \mathscr{A}$ such that $h(x) \neq h(y)$.*

Then $\mathscr{A}$ is uniformly dense in $C(\mathcal{X}; \mathbb{R})$. Equivalently, for every $f \in C(\mathcal{X}; \mathbb{R})$ and $\varepsilon > 0$, there exists $h \in \mathscr{A}$ such that

$$\sup_{x \in \mathcal{X}} |f(x) - h(x)| < \varepsilon.$$