PROMPTING A PRETRAINED TRANSFORMER CAN BE A UNIVERSAL APPROXIMATOR

Aleksandar Petrov, Philip H.S. Torr, Adel Bibi
University of Oxford
aleks@robots.ox.ac.uk

ABSTRACT

Despite the widespread adoption of prompting, prompt tuning and prefix-tuning of transformer models, our theoretical understanding of these fine-tuning methods remains limited. A key question is whether one can arbitrarily modify the behavior of pretrained model by prompting or prefix-tuning it, i.e., whether prompting and prefix-tuning a pretrained model can universally approximate sequence-to-sequence functions. This paper answers in the affirmative and demonstrates that much smaller pretrained models than previously thought can be universal approximators when prefixed. In fact, prefix-tuning a single attention head being sufficient to approximate any continuous function. Moreover, any sequence-to-sequence function can be approximated by prefixing a transformer with depth linear in the sequence length. Beyond these density-type results, we also offer bounds on the length of the prefix needed to approximate a function to a desired precision.

1 INTRODUCTION

Motivated by the success of few- and zero-shot learning (Wei et al., 2021; Kojima et al., 2022), context-based fine-tuning methods do not change the model parameters but the way the input is presented. With prompting, one fine-tunes a string of tokens (*a prompt*) which is prepended to the user input (Liu et al., 2023). One can optimize the real-valued embeddings instead (*soft prompting, prompt tuning*, Lester et al. 2021). A generalization to this approach is the optimization over the embeddings of every attention layer (*prefix-tuning*, Li and Liang 2021). As every prompt and soft prompt can be expressed as prefix-tuning (Petrov et al., 2024), we focus primarily on prefix-tuning.

While these context-based fine-tuning techniques have seen widespread adoption and can be competitive to full fine-tuning (Liu et al., 2022), our understanding of their abilities and restrictions remains limited. How much can the behavior of a model be modified without changing any model parameter? Can prefix-tuning of a pretrained transformer be a universal approximator? Given a pretrained transformer and an arbitrary target function, how long should the prefix be so that the transformer approximates this function to an arbitrary precision? These are some of the questions we aim to address in this work.

It is well-known that fully-connected neural networks can approximate any continuous function (Cybenko, 1989; Hornik et al., 1989; Barron, 1993; Telgarsky, 2015). The attention mechanism (Bahdanau et al., 2015) has also been studied in its own right. Deora et al. (2023) derived convergence and generalization guarantees for gradient-descent training of a single-layer multi-head self-attention model, and Mahdavi et al. (2023) showed that the memorization capacity increases linearly with the number of attention heads. On the other hand, it was shown that attention layers are not enough as it loses rank doubly exponentially with depth if Multi-Layer Perceptrons (MLPs) and residual connections are not present (Dong et al., 2021). However, attention layers, with a hidden size that grows only logarithmically in the sequence lengths, were shown to be good approximators for sparse attention patterns (Likhosherstov et al., 2021), except for a few tasks that require a linear scaling of the size of the hidden layers in the sequence length (Sanford et al., 2023).

Considering universal approximation using encoder-only transformers, Yun et al. (2019) showed that transformers are universal approximators of sequence-to-sequence functions by demonstrating that self-attention layers can compute contextual mappings of input sequences. Jiang and Li

(2023) demonstrated universality by instead leveraging the Kolmogorov-Albert representation theorem. Moreover, Alberti et al. (2023) provided universal approximation results for architectures with non-standard attention mechanisms. The closest to our objective is the work of Wang et al. (2023). They quantize the input and output spaces allowing them to enumerate all possible sequence-to-sequence functions. All possible functions and inputs can then be hard-coded in a transformer using the constructions by Yun et al. (2019). As this approach relies on memorization, the depth of the model depends on the desired approximation precision ϵ.

In this work, we demonstrate that prefix-tuning can be a universal approximator much more efficiently than previously assumed. In particular:

i. We show that attention heads are especially suited to model functions over hyperspheres, concretely, prefix-tuning *a single attention head* is sufficient to approximate any continuous function on the hypersphere S^m to any desired precision ϵ;
ii. We quantify the prompt length depending on ϵ and the smoothness of the target function;
iii. We demonstrate how this result can be leveraged to approximate general sequence-to-sequence functions with transformers of depth linear in the sequence length and independent of ϵ;

2 Background Material

2.1 Transformer Architecture

The sequence fed to a transformer model is split into two parts: a *prefix* sequence $P = (\boldsymbol{p}_1, \dots, \boldsymbol{p}_N)$ which is to be learnt or hand-crafted and an *input* sequence $X = (\boldsymbol{x}_1, \dots, \boldsymbol{x}_T)$, where $\boldsymbol{x}_i, \boldsymbol{p}_i \in \mathbb{R}^d$. Since we will only be interested in the output at the positions corresponding to the inputs X, we use $u(\cdot\,; P) : \mathbb{R}^{d\times(N+T)} \to \mathbb{R}^{d\times T}$ to denote the output of u at the locations corresponding to the input X when prefixed with P. Therefore, the k-th output of u is defined as:

$$[u(X;P)]_k = \frac{\sum_{i=1}^N \exp(\boldsymbol{x}_k^\top \boldsymbol{H} \boldsymbol{p}_i) \boldsymbol{W}_V \boldsymbol{p}_i + \sum_{j=1}^T \exp(\boldsymbol{x}_k^\top \boldsymbol{H} \boldsymbol{x}_j) \boldsymbol{W}_V \boldsymbol{x}_j}{\sum_{i=1}^N \exp(\boldsymbol{x}_k^\top \boldsymbol{H} \boldsymbol{p}_i) + \sum_{j=1}^T \exp(\boldsymbol{x}_k^\top \boldsymbol{H} \boldsymbol{x}_j)}, \tag{1}$$

where $\boldsymbol{W}_V, \boldsymbol{H} \in \mathbb{R}^{d\times d}$. For simplicity, we only use attention blocks with a single head.

We consider *pretrained* transformers but, in the context of this work, these are constructed rather than trained. $\boldsymbol{W}_V, \boldsymbol{W}_Q, \boldsymbol{W}_K$ along with the parameters of the MLPs are *pretrained parameters* and the prefix P is the only variable that can be modified to change the behavior of the model.

2.2 Universal Approximation

Let $\mathcal{X}$ and $\mathcal{Y}$ be normed vector spaces. Given a concept space $\mathcal{C} \subseteq \mathcal{Y}^{\mathcal{X}}$ and a hypothesis space $\mathcal{H} \subset \mathcal{Y}^{\mathcal{X}}$, we are interested in how well functions in $\mathcal{H}$ approximate functions in $\mathcal{C}$. The hypothesis classes we consider are the set of all prefixed attention heads and the set of prefixed transformers:

Definition 1 (Prefixed Attention Heads Class ($\mathcal{H}^{N,T}_{-,d}$)). This is the class of all attention heads as defined in Equation (1) of dimension d, input/output sequence of length T, prefix of length at most N, and *fixed* pretrained components $\boldsymbol{H}, \boldsymbol{W}_V \in \mathbb{R}^{d\times d}$. For simplicity, we say that $\mathcal{H}^{N,T}_{-,d}$ *is dense in* $\mathcal{C}$ to imply that there exists a pair $(\boldsymbol{H}, \boldsymbol{W}_V)$ such that $\mathcal{H}^{N,T}_{-,d}(\boldsymbol{H}, \boldsymbol{W}_V)$ is dense in $\mathcal{C}$. When considering all possible prefix lengths, we drop the N: $\mathcal{H}^{T}_{-,d} = \bigcup_{N\in\mathbb{N}} \mathcal{H}^{N,T}_{-,d}$.

Definition 2 (Prefixed Transformers Class ($\mathcal{H}^{N,T}_{\equiv,d}$)). A transformer consists of L layers with each layer l consisting of an attention head with $\boldsymbol{H}^l$ and $\boldsymbol{W}_V^l$ followed by an MLP consisting of k_l linear layers, each parameterized as $\mathcal{L}_k^l(\boldsymbol{x}) = \boldsymbol{A}^{l,k}\boldsymbol{x} + \boldsymbol{b}^{l,k}$ interspersed with non-linear activation σ. This gives rise to the following hypothesis class when prefixed: Again, we say $\mathcal{H}^{N,T}_{\equiv,d}$ *is dense in* $\mathcal{C}$, as a shorthand, to *there exists* $\{\boldsymbol{H}^l, \boldsymbol{W}_V^l, \boldsymbol{A}^{l,k}, \boldsymbol{b}^{l,k}\}_{l=1}^L$ *such that* $\mathcal{H}^{N,T}_{\equiv,d}(\{\boldsymbol{H}^l, \boldsymbol{W}_V^l, \boldsymbol{A}^{l,k}, \boldsymbol{b}^{l,k}\}_{l=1}^L)$ *is dense in* $\mathcal{C}$.

In this paper, we consider several different concept classes. For reasons that will become apparent in the following section, we focus on functions whose domain is a hypersphere $S^m = \{\boldsymbol{y} \in \mathbb{R}^{m+1} \mid \|\boldsymbol{y}\|_2 = 1\} \subset \mathbb{R}^{m+1}$.

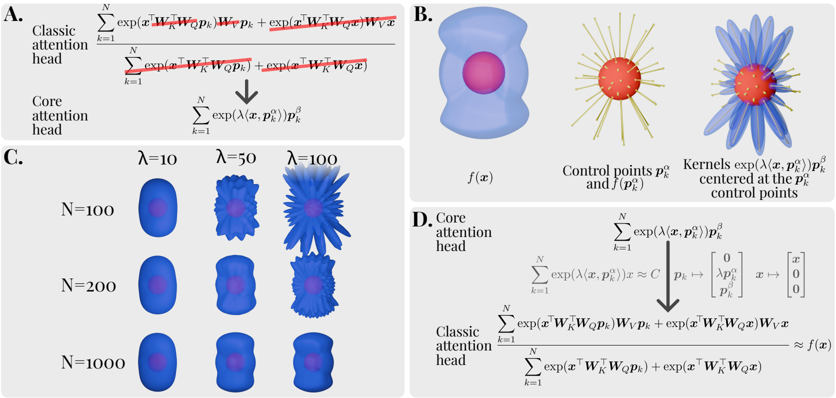

A.
Classic attention head
$\frac{\sum_{k=1}^{N} \exp(\boldsymbol{x}^\top \boldsymbol{W}_K^\top \boldsymbol{W}_Q \boldsymbol{p}_k)\boldsymbol{W}_V \boldsymbol{p}_k + \exp(\boldsymbol{x}^\top \boldsymbol{W}_K^\top \boldsymbol{W}_Q \boldsymbol{x})\boldsymbol{W}_V \boldsymbol{x}}{\sum_{k=1}^{N} \exp(\boldsymbol{x}^\top \boldsymbol{W}_K^\top \boldsymbol{W}_Q \boldsymbol{p}_k) + \exp(\boldsymbol{x}^\top \boldsymbol{W}_K^\top \boldsymbol{W}_Q \boldsymbol{x})}$
Core attention head
$\sum_{k=1}^{N} \exp(\lambda\langle \boldsymbol{x}, \boldsymbol{p}_k^\alpha\rangle)\boldsymbol{p}_k^\beta$
B.
$f(\boldsymbol{x})$
Control points $\boldsymbol{p}_k^\alpha$ and $f(\boldsymbol{p}_k^\alpha)$
Kernels $\exp(\lambda\langle \boldsymbol{x}, \boldsymbol{p}_k^\alpha\rangle)\boldsymbol{p}_k^\beta$ centered at the $\boldsymbol{p}_k^\alpha$ control points
C.
λ=10
λ=50
λ=100
N=100
N=200
N=1000
D.
Core attention head
$\sum_{k=1}^{N} \exp(\lambda\langle \boldsymbol{x}, \boldsymbol{p}_k^\alpha\rangle)\boldsymbol{p}_k^\beta$
$\sum_{k=1}^{N} \exp(\lambda\langle \boldsymbol{x}, \boldsymbol{p}_k^\alpha\rangle)x \approx C$
$\boldsymbol{p}_k \mapsto \begin{bmatrix} 0 \\ \lambda \boldsymbol{p}_k^\alpha \\ \boldsymbol{p}_k^\beta \end{bmatrix}$
$\boldsymbol{x} \mapsto \begin{bmatrix} x \\ 0 \\ 0 \end{bmatrix}$
Classic attention head
$\frac{\sum_{k=1}^{N} \exp(\boldsymbol{x}^\top \boldsymbol{W}_K^\top \boldsymbol{W}_Q \boldsymbol{p}_k)\boldsymbol{W}_V \boldsymbol{p}_k + \exp(\boldsymbol{x}^\top \boldsymbol{W}_K^\top \boldsymbol{W}_Q \boldsymbol{x})\boldsymbol{W}_V \boldsymbol{x}}{\sum_{k=1}^{N} \exp(\boldsymbol{x}^\top \boldsymbol{W}_K^\top \boldsymbol{W}_Q \boldsymbol{p}_k) + \exp(\boldsymbol{x}^\top \boldsymbol{W}_K^\top \boldsymbol{W}_Q \boldsymbol{x})} \approx f(\boldsymbol{x})$

Figure 1: **Approximating functions on the hypersphere with a single attention head. A.** We simplify the classical attention head into a *core attention head*. **B.** The $\exp(\lambda\langle \boldsymbol{x}, \boldsymbol{p}_k^\alpha\rangle)\boldsymbol{p}_k^\beta$ terms act like kernels when $\boldsymbol{x}$ is restricted to a hypersphere. We can approximate a function f by placing N control points $\boldsymbol{p}_1^\alpha, ..., \boldsymbol{p}_N^\alpha$ and centering a kernel at each of them. **C.** Increasing λ results in less smoothing, while increasing N results in more control points and hence better approximation. With large enough λ and N, we can approximate f to any desired accuracy. **D.** With the normalization term in classical attention close to a constant, and giving $\boldsymbol{x}$, $\boldsymbol{p}_k^\alpha$ and $\boldsymbol{p}_k^\beta$ orthogonal subspaces, core attention can be represented as classical attention. Hence, a classical attention head can also approximate f with arbitrary precision.

Definition 3 (Vector-valued Functions on the Hypersphere)**.** The class of vector-valued functions on the hypersphere is:

$$\mathcal{C}_{v,m} = \{f : S^m \to \mathbb{R}^{m+1} \mid [f]_i \in C(S^m), i = 1, ..., m+1\},$$

with $C(S^m) \subset \mathbb{R}^{S^m}$ the space of all continuous functions defined on S^m with bounded norm.

Transformers are typically used to learn mappings over sequences rather than individual inputs. Hence, we define several sequence-to-sequence concept classes.

Definition 4 (General Sequence-to-sequence Functions)**.** Given a fixed sequence length $T \in \mathbb{N}_{>0}$, we define the sequence-to-sequence function class as:

$$\mathcal{C}_{T,m} = \{f : (S^m)^T \to (\mathbb{R}^{m+1})^T \mid f \text{ continuous and bounded}\}.$$

3 Main Results

We briefly summarise our main results and leave the formal treatment for the appendix. First, we observe that the classical attention head can be simplified in what we refer to as a *core attention head* by dropping the terms depending only on $\boldsymbol{x}$, setting $\boldsymbol{H} = \lambda \boldsymbol{I}_d$, $\lambda > 0$, and $\boldsymbol{W}_V = \boldsymbol{I}_d$ and removing the denominator. We also allow for different values of the prefix positions when computing the attention and when computing the value:

$$h_{\circledast}(\boldsymbol{x}) = \textstyle\sum_{k=1}^{N} \exp(\lambda\langle \boldsymbol{x}, \boldsymbol{p}_k^\alpha\rangle)\boldsymbol{p}_k^\beta, \tag{2}$$

which gives rise to the hypothesis class:

$$\mathcal{H}_{\circledast,d}^{N} = \left\{ \boldsymbol{x} \mapsto \textstyle\sum_{k=1}^{N'} \exp(\lambda\langle \boldsymbol{x}, \boldsymbol{p}_k^\alpha\rangle)\boldsymbol{p}_k^\beta, \text{ where } \boldsymbol{p}_k^\alpha, \boldsymbol{p}_k^\beta \in \mathbb{R}^d, N' \leq N, \lambda > 0 \right\}.$$

As the dot product is a notion of similarity, one can interpret $h_{\circledast}$ in Equation (6) as an interpolator. The $\boldsymbol{p}_i^\alpha$ vectors act as control points, while the $\boldsymbol{p}_i^\beta$ vectors designate the output value at the location of the corresponding control point. The dot product with the input $\boldsymbol{x}$ controls how much each control point should contribute to the final result, with control points closer to $\boldsymbol{x}$ (larger dot product) contributing more. However, while functions of the type in Equation (6) cannot approximate arbitrary functions over $\mathbb{R}^d$, we show they can approximate any function defined over the hypersphere S^m.

Moreover, we give a Jackson-type result, that is, we quantify the number of terms N in the sum in Equation (6) we need in order to approximate a function to a given precision. In our prefix-tuning setting, N coresponds to the prefix length. Our result is as following:

Informal Theorem 1. *Let $f : S^m \to \mathbb{R}^{m+1}$, $m \geq 8$ satisfy some continuity conditions. Then, for any $\epsilon > 0$, there exist $\boldsymbol{p}_1^\alpha, ..., \boldsymbol{p}_N^\alpha \in S^m$ and $\boldsymbol{p}_1^\beta, ..., \boldsymbol{p}_N^\beta \in \mathbb{R}^{m+1}$ such that*

$$\sup_{\boldsymbol{x} \in S^m} \left\| f(\boldsymbol{x}) - \textstyle\sum_{k=1}^N \exp(\lambda \langle \boldsymbol{x}, \boldsymbol{p}_k^\alpha \rangle) \boldsymbol{p}_k^\beta \right\|_2 \leq \epsilon,$$

with $\lambda = \mathcal{O}(\epsilon^{-4})$ and for any $N \geq N(\epsilon) = \mathcal{O}(\epsilon^{\text{-1-3}m\text{-2}m^2})$. That is, $\mathcal{H}_{\circledast, m+1}$ is dense in $\mathcal{C}_{v,m}$ with respect to the $\| \cdot \|_2$ norm.

Now, we can bring back all the components of the classical attention head (Equation (1)) that we removed and combine the two parts of the prefix $\boldsymbol{p}_k^\alpha$ and $\boldsymbol{p}_k^\beta$. We do this by increasing the hidden dimension of our attention head to $3(m{+}1)$ while preserving the asymptotic behaviour for λ and N:

Informal Theorem 2. *Let $f : S^m \to \mathbb{R}^{m+1}$, $m \geq 8$ satisfy some continuity conditions. Then, there exists a matrix $\Pi \in \mathbb{R}^{3(m+1) \times (m+1)}$ such that for any $0 < \epsilon$ there exists a prefix for an attention head $h \in \mathcal{H}_{-,3(m+1)}^{N,1}$ such that:*

$$\sup_{\boldsymbol{x} \in S^m} \| f(\boldsymbol{x}) - (\Pi^T \circ h \circ \Pi)(\boldsymbol{x}) \|_2 \leq \epsilon, \tag{3}$$

for any $N \geq N(\lambda, {}^{\epsilon}/\sqrt{m+1})$. That is, $\Pi^{-1} \circ \mathcal{H}_{-,3(m+1)}^1 \circ \Pi$ is dense in $\mathcal{C}_{v,m}$.

Appendix B has a detailed development of these result with the formal proofs in Appendix E, while Figure 1 illustrates the idea behind the results.

Still, one typically uses the transformer architecture for operations over sequences rather than over single inputs (the case with $T \geq 1$). By using a variant of the Kolmogorov-Arnold theorem (Kolmogorov, 1957) due to Schmidt-Hieber (2021), we can extend the above results to the $T \geq 1$ case:

Informal Theorem 3. *Let $f \in \mathcal{C}_{T,m}$ satisfy some continuity conditions. Then, there exists a transformer with $T + 2$ layers, such that for any $0 < \epsilon$ there exists a $h \in \mathcal{H}_{\equiv,d}^T$ such that*

$$\sup_{\{\boldsymbol{x}_i\}_{i=1}^T} \max_{1 \leq k \leq T} \left\| \left[f(\{\boldsymbol{x}_i\}) - (\Pi^{-1} \circ h^T \circ \Pi)(\{\boldsymbol{x}_i\}) \right]_k \right\|_2 \leq \epsilon.$$

This result is discussed in detail in Appendix C.

4 Discussion and conclusions

Just like us, Wang et al. (2023) show that prefix-tuning can be a universal approximator. This approach has several limitations: i) the model has exponential depth $\mathcal{O}(T\epsilon^{-m})$; ii) reducing the approximation error ϵ requires increasing the model depth; iii) the prefix length is fixed, hence a constant function and a highly non-smooth function would have equal prefix lengths, and iv) it effectively has memorized all possible functions and inputs, explaining the exponential size of their constructions. In contrast, we show that memorization is not needed: attention heads are naturally suited for universal approximation. Informal Theorem 3 showed that $T + 2$ layers are enough, we require shorter prefixes for more smooth functions and reducing the approximation error ϵ can be done by increasing the prefix length, without modifying the pretrained model.

While this work focused on prefix-tuning, the results can extend to prompting. Observe that prefix tuning can be reduced to soft prompting by using an appropriate attention mechanism and position embeddings. Hence, if a function $f \in \mathcal{C}_{T,m}$ requires N prefixes to be approximated to precision ϵ with prefix-tuning, it would require $\mathcal{O}(TN)$ soft tokens to be approximated with soft prompting. A soft token can be encoded with a sequence of hard tokens, hence f could be approximated with $\mathcal{O}(\log_V(\epsilon^{-1})mTN)$ hard tokens, with V the vocabulary size. Therefore, our universal approximation results may translate to prompting. This raises concerns as to whether it is at all possible to prevent a transformer model to exhibit undesirable behaviors (Zou et al., 2023; Wolf et al., 2023; Chao et al., 2023). Still, our results require specific form of the attention and value matrices and, hence, it is not clear whether these risk translate to real-world models.

References

Silas Alberti, Niclas Dern, Laura Thesing, and Gitta Kutyniok. 2023. Sumformer: Universal approximation for efficient transformers. In *Proceedings of 2nd Annual Workshop on Topology, Algebra, and Geometry in Machine Learning (TAG-ML)*.

Donald E Amos. 1974. Computation of modified Bessel functions and their ratios. *Mathematics of Computation*, 28(125):239–251.

Kendall Atkinson and Weimin Han. 2012. *Spherical Harmonics and Approximations on the Unit Sphere: An Introduction*.

Yogesh J Bagul and Satish K Panchal. 2018. Certain inequalities of Kober and Lazarević type. *Research Group in Mathematical Inequalities and Applications Research Report Collection*, 21(8).

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2015. Neural machine translation by jointly learning to align and translate. In *International Conference on Learning Representations*.

Alex Barnett. 2021. Lower bounds on the modified Bessel function of the first kind. Mathematics Stack Exchange.

Andrew R Barron. 1993. Universal approximation bounds for superpositions of a sigmoidal function. *IEEE Transactions on Information Theory*, 39(3):930–945.

Shaked Brody, Uri Alon, and Eran Yahav. 2023. On the expressivity role of LayerNorm in transformers' attention. *arXiv preprint arXiv:2305.02582*.

Patrick Chao, Alexander Robey, Edgar Dobriban, Hamed Hassani, George J. Pappas, and Eric Wong. 2023. Jailbreaking black box large language models in twenty queries. *arXiv preprint arXiv:2310.08419*.

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

Feng Dai and Yuan Xu. 2013. *Approximation Theory and Harmonic Analysis on Spheres and Balls*.

David Demeter, Gregory Kimmel, and Doug Downey. 2020. Stolen probability: A structural weakness of neural language models. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*.

Puneesh Deora, Rouzbeh Ghaderi, Hossein Taheri, and Christos Thrampoulidis. 2023. On the optimization and generalization of multi-head attention. *arXiv preprint arXiv:2310.12680*.

Yihe Dong, Jean-Baptiste Cordonnier, and Andreas Loukas. 2021. Attention is not all you need: Pure attention loses rank doubly exponentially with depth. In *International Conference on Machine Learning*.

Ricardo Estrada. 2014. On radial functions and distributions and their Fourier transforms. *Journal of Fourier Analysis and Applications*, 20(2):301–320.

Uriel Feige and Gideon Schechtman. 2002. On the optimality of the random hyperplane rounding technique for MAX CUT. *Random Structures & Algorithms*, 20(3):403–440.

Paul Funk. 1915. Beiträge zur Theorie der Kugelfunktionen. *Mathematische Annalen*, 77:136–152.

Federico Girosi and Tomaso Poggio. 1989. Representation properties of networks: Kolmogorov's theorem is irrelevant. *Neural Computation*, 1(4):465–469.

E Hecke. 1917. Über orthogonal-invariante Integralgleichungen. *Mathematische Annalen*, 78:398–404.

Kurt Hornik, Maxwell Stinchcombe, and Halbert White. 1989. Multilayer feedforward networks are universal approximators. *Neural networks*, 2(5):359–366.

Haotian Jiang and Qianxiao Li. 2023. Approximation theory of transformer networks for sequence modeling. *arXiv preprint arXiv:2305.18475*.

Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. 2022. Large language models are zero-shot reasoners. *Advances in Neural Information Processing Systems*.

Andrei Nikolaevich Kolmogorov. 1957. On the representation of continuous functions of many variables by superposition of continuous functions of one variable and addition. In *Doklady Akademii Nauk*, volume 114, pages 953–956. Russian Academy of Sciences.

Brian Lester, Rami Al-Rfou, and Noah Constant. 2021. The power of scale for parameter-efficient prompt tuning. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*.

Shengqiao Li. 2010. Concise formulas for the area and volume of a hyperspherical cap. *Asian Journal of Mathematics & Statistics*, 4(1):66–70.

Xiang Lisa Li and Percy Liang. 2021. Prefix-Tuning: Optimizing continuous prompts for generation. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*.

Valerii Likhosherstov, Krzysztof Choromanski, and Adrian Weller. 2021. On the expressive power of self-attention matrices. *arXiv preprint arXiv:2106.03764*.

Pengfei Liu, Weizhe Yuan, Jinlan Fu, Zhengbao Jiang, Hiroaki Hayashi, and Graham Neubig. 2023. Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing. *ACM Computing Surveys*.

Xiao Liu, Kaixuan Ji, Yicheng Fu, Weng Tam, Zhengxiao Du, Zhilin Yang, and Jie Tang. 2022. P-Tuning: Prompt tuning can be comparable to fine-tuning across scales and tasks. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*.

Sadegh Mahdavi, Renjie Liao, and Christos Thrampoulidis. 2023. Memorization capacity of multi-head attention in transformers. *arXiv preprint arXiv:2306.02010*.

Valdir Antônio Menegatto. 1997. Approximation by spherical convolution. *Numerical Functional Analysis and Optimization*, 18(9-10):995–1012.

Tin Lok James Ng and Kwok-Kun Kwong. 2022. Universal approximation on the hypersphere. *Communications in Statistics – Theory and Methods*, 51(24):8694–8704.

Aleksandar Petrov, Philip HS Torr, and Adel Bibi. 2024. When do prompting and prefix-tuning work? A theory of capabilities and limitations. In *International Conference on Learning Representations*.

David L Ragozin. 1971. Constructive polynomial approximation on spheres and projective spaces. *Transactions of the American Mathematical Society*, 162:157–170.

C. A. Rogers. 1963. Covering a sphere with spheres. *Mathematika*, 10(2):157–164.

Clayton Sanford, Daniel Hsu, and Matus Telgarsky. 2023. Representational strengths and limitations of transformers. *arXiv preprint arXiv:2306.02896*.

Johannes Schmidt-Hieber. 2021. The Kolmogorov–Arnold representation theorem revisited. *Neural Networks*, 137:119–126.

Matus Telgarsky. 2015. Representation benefits of deep feedforward networks. *arXiv preprint arXiv:1509.08101*.

Yihan Wang, Jatin Chauhan, Wei Wang, and Cho-Jui Hsieh. 2023. Universality and limitations of prompt tuning. In *Advances in Neural Information Processing Systems*.

Jason Wei, Maarten Bosma, Vincent Zhao, Kelvin Guu, Adams Wei Yu, Brian Lester, Nan Du, Andrew M Dai, and Quoc V Le. 2021. Finetuned language models are zero-shot learners. In *International Conference on Learning Representations*.

Yotam Wolf, Noam Wies, Oshri Avnery, Yoav Levine, and Amnon Shashua. 2023. Fundamental limitations of alignment in large language models. *arXiv preprint arXiv:2304.11082*.

Chulhee Yun, Srinadh Bhojanapalli, Ankit Singh Rawat, Sashank Reddi, and Sanjiv Kumar. 2019. Are transformers universal approximators of sequence-to-sequence functions? In *International Conference on Learning Representations*.

Andy Zou, Zifan Wang, J Zico Kolter, and Matt Fredrikson. 2023. Universal and transferable adversarial attacks on aligned language models. *arXiv preprint arXiv:2307.15043*.

A Some further definitions

Definition A.1 (Universal Approximation (Density-Type))**.** We say that $\mathcal{H}$ is a universal approximator for $\mathcal{C}$ over a compact set $S \subseteq \mathcal{X}$ if for every $f \in \mathcal{C}$ and every $\epsilon > 0$ there exists an $h \in \mathcal{H}$ such that $\sup_{x\in S} \|f(x) - h(x)\| \leq \epsilon$. One typically says that $\mathcal{H}$ *is dense in* $\mathcal{C}$.

Lemma A.1 (Transitivity)**.** *If $\mathcal{A}$ is dense in $\mathcal{B}$ and $\mathcal{B}$ is dense in $\mathcal{C}$, then $\mathcal{A}$ is dense in $\mathcal{C}$.*

Definition A.2 (Approximation Rate (Jackson-Type))**.** Fix a hypothesis space $\mathcal{H}$. Let $\{\mathcal{H}^N : N \in \mathbb{N}_+\}$ be a collection of subsets of $\mathcal{H}$ such that $\mathcal{H}^N \subset \mathcal{H}^{N+1}$ and $\bigcup_{N\in\mathbb{N}_+} \mathcal{H}^N = \mathcal{H}$. Here, N is a measure of the complexity of the approximation candidates, and $\mathcal{H}^N$ is the subset of hypotheses with complexity at most N. Then, the approximation rate estimate for $\mathcal{C}$ over a compact $S \subseteq \mathcal{X}$ is a bound $Z_{\mathcal{H}}$:

$$N \geq Z_{\mathcal{H}}(f, \epsilon) \implies \inf_{h\in\mathcal{H}^N} \sup_{x\in S} \|f(x) - h(x)\| \leq \epsilon, \forall f{\in}\mathcal{C}.$$

$Z_{\mathcal{H}}$ gives a lower bound on the hypothesis complexity necessary to reach the target precision ϵ and typically depends on the smoothness of f.

Lemma A.2. *A Jackson bound for $\{\mathcal{H}^N \mid N \in \mathbb{N}_+\}$ with finite $Z_{\mathcal{H}}$ for all $f \in \mathcal{C}$,$\epsilon > 0$ immediately implies that $\bigcup_{N\in\mathbb{N}_+} \mathcal{H}^N = \mathcal{H}$ is dense in $\mathcal{C}$. Hence, Jackson bounds (Definition A.2) are stronger than density results (Definition A.1).*

Definition A.3 (Scalar Functions on the Hypersphere)**.** Define $C(S^m) \subset \mathbb{R}^{S^m}$ to be the space of all continuous functions defined on S^m with bounded norm, i.e.,

$$\|f\|_\infty = \sup_{\boldsymbol{x}\in S^m} |f(\boldsymbol{x})| < \infty, \; f \in C(S^m). \tag{4}$$

This is the concept class $\mathcal{C}_{s,m} = C(S^m) \subset \mathbb{R}^{S^m}$.

Definition A.4 (Element-wise functions)**.** Element-wise functions operate over sequences of inputs but apply the exact same function independently to all inputs:

$$\mathcal{C}_{\|,T,m}{=}\left\{ f \in \mathcal{C}_{T,m} \;\middle|\; \begin{array}{l} \text{there exists } g \in \mathcal{C}_{v,m}, \text{ such that} \\ f(\boldsymbol{x}_1, ..., \boldsymbol{x}_T){=}(g(\boldsymbol{x}_1), ..., g(\boldsymbol{x}_T)) \\ \text{for all } (\boldsymbol{x}_1, ..., \boldsymbol{x}_T) \in (S^m)^T \end{array} \right\}.$$

B Universal Approximation with a Single Attention Head

In this section, we will restrict ourselves to the setting when the input sequence is of length T=1, i.e., X=$(\boldsymbol{x})$. General sequence-to-sequence functions will be discussed in Appendix C. We will show that a single attention head can approximate any continuous function on the hypersphere, or that $\mathcal{H}^1_{-,m+1}$ is dense in $\mathcal{C}_{s,m}$. To do this, we first simplify the classical attention head in Equation (1), resulting in what we call a *core attention head*. Then, we show that each of the terms in the core attention act as a kernel, meaning that it can approximate any function in $\mathcal{C}_s$. Finally, we show that any core attention head can be approximated by a classical attention head, hence, $\mathcal{H}^1_{-,m+1}$ is indeed dense in $\mathcal{C}_s$. The complete pipeline is illustrated in Figure 1.

To illuminate the approximation abilities of the attention head mechanism we relax it a bit. That is, we allow for different values of the prefix positions when computing the attention (the exp terms in Equation (1)) and when computing the value (the right multiplication with $\boldsymbol{W}_V$). We will also

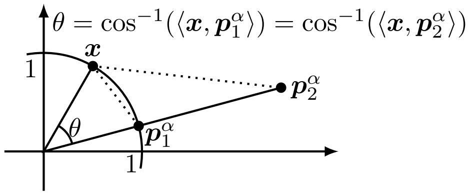

$\theta = \cos^{-1}(\langle \boldsymbol{x}, \boldsymbol{p}_1^\alpha \rangle) = \cos^{-1}(\langle \boldsymbol{x}, \boldsymbol{p}_2^\alpha \rangle)$
$\boldsymbol{x}$
1
$\boldsymbol{p}_2^\alpha$
θ
$\boldsymbol{p}_1^\alpha$
1

Figure 2: **The dot product is a measure of closeness over the hypersphere.** We want large dot product for points with lower distances. That is not the case for general $\boldsymbol{p}_1^\alpha, \boldsymbol{p}_2^\alpha \in \mathbb{R}^{m+1}$: above we show larger dot product for points which are further away, i.e., $\langle \boldsymbol{x}, \boldsymbol{p}_1^\alpha \rangle < \langle \boldsymbol{x}, \boldsymbol{p}_2^\alpha \rangle$ despite $\|\boldsymbol{x} - \boldsymbol{p}_1^\alpha\|_2 < \|\boldsymbol{x} - \boldsymbol{p}_1^\alpha\|_2$. However, if we restrict $\boldsymbol{x}$, $\boldsymbol{p}_i^\alpha$, and $\boldsymbol{p}_j^\alpha$ to the hypersphere S^m, then the dot product measures the cosine between $\boldsymbol{x}$ and $\boldsymbol{p}_i$ which is truly a measure of closeness: $\langle \boldsymbol{x}, \boldsymbol{p}_i^\alpha \rangle < \langle \boldsymbol{x}, \boldsymbol{p}_j^\alpha \rangle \iff \|\boldsymbol{x} - \boldsymbol{p}_i^\alpha\|_2 > \|\boldsymbol{x} - \boldsymbol{p}_j^\alpha\|_2$.

drop the terms depending only on $\boldsymbol{x}$, set $\boldsymbol{H} = \lambda \boldsymbol{I}_d$, $\lambda > 0$, and $\boldsymbol{W}_V = \boldsymbol{I}_d$. We refer to this relaxed version as a *split attention head* with its corresponding hypothesis class:

$$h_{\rfloor\lceil}(\boldsymbol{x}) = \frac{\sum_{k=1}^{N'} \exp(\lambda \langle \boldsymbol{x}, \boldsymbol{p}_k^\alpha \rangle) \boldsymbol{p}_k^\beta}{\sum_{k=1}^{N'} \exp(\lambda \langle \boldsymbol{x}, \boldsymbol{p}_k^\alpha \rangle)}. \tag{5}$$

Definition B.1 (Split Attention Head Class)**.**

$$\mathcal{H}_{\rfloor\lceil, d}^N = \left\{ h_{\rfloor\lceil} \text{ as in (5)}, \boldsymbol{p}_k^\alpha, \boldsymbol{p}_k^\beta \in \mathbb{R}^d, N' \leq N, \lambda > 0 \right\}.$$

We will later show that a split head can be represented by a classical attention head. For now, let us simplify a bit further: we drop the denominator, resulting in a *core attention head*:

$$h_{\circledast}(\boldsymbol{x}) = \textstyle\sum_{k=1}^{N} \exp(\lambda \langle \boldsymbol{x}, \boldsymbol{p}_k^\alpha \rangle) \boldsymbol{p}_k^\beta, \tag{6}$$

which gives rise to the hypothesis class:

$$\mathcal{H}_{\circledast, d}^N = \left\{ \begin{array}{c} \boldsymbol{x} \mapsto \sum_{k=1}^{N'} \exp(\lambda \langle \boldsymbol{x}, \boldsymbol{p}_k^\alpha \rangle) \boldsymbol{p}_k^\beta, \text{ where} \\ \boldsymbol{p}_k^\alpha, \boldsymbol{p}_k^\beta \in \mathbb{R}^d, N' \leq N, \lambda > 0 \end{array} \right\}.$$

We also have their scalar-valued counterparts:

$$h_{\odot}(\boldsymbol{x}) = \textstyle\sum_{k=1}^{N} \exp(\lambda \langle \boldsymbol{x}, \boldsymbol{p}_k^\alpha \rangle) p_k^\beta, \tag{7}$$

$$\mathcal{H}_{\odot, d}^N = \left\{ \begin{array}{c} \boldsymbol{x} \mapsto \sum_{k=1}^{N'} \exp(\lambda \langle \boldsymbol{x}, \boldsymbol{p}_k^\alpha \rangle) p_k^\beta, \text{ where} \\ \boldsymbol{p}_k^\alpha \in \mathbb{R}^d, p_k^\beta \in \mathbb{R}, N' \leq N, \lambda > 0 \end{array} \right\}.$$

As the dot product is a notion of similarity, one can interpret $h_{\circledast}$ in Equation (6) as an interpolator. The $\boldsymbol{p}_i^\alpha$ vectors act as control points, while the $\boldsymbol{p}_i^\beta$ vectors designate the output value at the location of the corresponding control point. The dot product with the input $\boldsymbol{x}$ controls how much each control point should contribute to the final result, with control points closer to $\boldsymbol{x}$ (larger dot product) contributing more.

Unfortunately, it is not generally true that higher dot product means smaller distance, hence the above interpretation fails in $\mathbb{R}^{m+1}$. To see this, consider two control points $\boldsymbol{p}_1^\alpha, \boldsymbol{p}_2^\alpha \in \mathbb{R}^{m+1}$ such that $\boldsymbol{p}_2^\alpha = t\boldsymbol{p}_1^\alpha$, with $t > 1$. Then for $\boldsymbol{x} = \boldsymbol{p}_1^\alpha$ we would have $\langle \boldsymbol{x}, \boldsymbol{p}_1^\alpha \rangle = \|\boldsymbol{p}_1^\alpha\|_2^2 < \langle \boldsymbol{x}, \boldsymbol{p}_2^\alpha \rangle = t\|\boldsymbol{p}_1^\alpha\|_2^2$; the dot product is smaller for $\boldsymbol{p}_1^\alpha$, the control point that is closer to $\boldsymbol{x}$, than for the much further away $\boldsymbol{p}_2^\alpha$ (see Figure 2). Therefore, the further away control point has a larger contribution than the closer point, which is at odds with the interpolation behaviour we desire. In general, the contribution of control points with larger norms will "dominate" the one of points with smaller norms. This has been observed for the attention mechanism in general by Demeter et al. (2020).

Fortunately, the domination of larger norm control points $\boldsymbol{p}_i^\alpha$ is not an issue if all control points have the same norm. In particular, if $\boldsymbol{x}$ and $\boldsymbol{p}_i^\alpha$ lie on the unit hypersphere $S^m = \{\boldsymbol{y} \in \mathbb{R}^{m+1} \mid \|\boldsymbol{y}\|_2 = 1\}$ then $\langle \boldsymbol{x}, \boldsymbol{p}_i^\alpha \rangle = \cos(\angle(\boldsymbol{x}, \boldsymbol{p}_i^\alpha))$ and it has the desired property that the closer $\boldsymbol{x}$ is to $\boldsymbol{p}_i^\alpha$, the higher their dot product. By doing this, we restrict $h_{\circledast}$ to be a function from the hypersphere S^m to $\mathbb{R}^{m+1}$. While this might seem artificial, modern transformer architectures do operate over hyperspheres as LayerNorm projects activations from $\mathbb{R}^{m+2}$ to S^m (Brody et al., 2023).

The central result of this section is that the functions of the form of Equation (7) can approximate any continuous function defined on the hypersphere, i.e., $\mathcal{H}_{\odot,m+1} = \bigcup_{N=1}^{\infty} \mathcal{H}_{\odot,m+1}^{N}$ is dense in $\mathcal{C}_{s,m}$ (Definition A.3) and $\mathcal{H}_{\circledast,m+1} = \bigcup_{N=1}^{\infty} \mathcal{H}_{\circledast,m+1}^{N}$ is dense in $\mathcal{C}_{v,m}$ (Definition 3). Furthermore, we offer a Jackson-type approximation rate result which gives us a bound on the necessary prefix length N to achieve a desired approximation quality.

Theorem B.1 (Jackson-type Bound for Universal Approximation on the Hypersphere)**.** *Let $f \in C(S^m)$ be a continuous function on S^m, $m \geq 8$ with modulus of continuity*

$$\omega(f;t){=}\sup\{|f(\boldsymbol{x}){-}f(\boldsymbol{y})| \mid \boldsymbol{x},\boldsymbol{y}{\in}S^m, \cos^{\text{-}1}(\langle\boldsymbol{x},\boldsymbol{y}\rangle) \leq t\} \leq Lt,$$

for some $L > 0$. Then, for any $\epsilon > 0$, there exist $\boldsymbol{p}_1^{\alpha}, \ldots, \boldsymbol{p}_N^{\alpha} \in S^m$ and $p_1^{\beta}, \ldots, p_N^{\beta} \in \mathbb{R}$ such that

$$\sup_{x \in S^m} \left| f(\boldsymbol{x}) - \sum_{k=1}^{N} \exp(\lambda\langle\boldsymbol{x}, \boldsymbol{p}_k^{\alpha}\rangle) p_k^{\beta} \right| \leq \epsilon,$$

where $\lambda = \Lambda(\epsilon/2)$ with

$$\Lambda(\sigma) = \frac{(8LC_R{+}m\sigma{+}\sigma)\left(1\text{-}\frac{\sigma^2}{8LC_HC_R{+}2\sigma C_H}\right)^{\frac{\sigma}{4LC_R{+}\sigma}}}{\sigma\left(1\text{-}\left(1\text{-}\frac{\sigma^2}{8LC_HC_R{+}2\sigma C_H}\right)^{\frac{2\sigma}{4LC_R{+}\sigma}}\right)} = \mathcal{O}\left(\frac{L^3C_H}{\sigma^4}\right), \tag{8}$$

and any $N \geq N(\lambda, \epsilon)$ with

$$N(\lambda,\epsilon){=}\Phi(m)\left(\frac{3\pi(L{+}\|f\|_{\infty})c_{m+1}(\lambda)\exp(\lambda)}{\epsilon}\right)^{m+1}{=}\mathcal{O}(\epsilon^{\text{-}1\text{-}3m\text{-}2m^2}), \tag{9}$$

with C_H being a constant depending on the smoothness of f (formally defined in the proof), C_R being a constant not depending on f or ϵ, $\Phi(m) = \mathcal{O}(m \log m)$ being a function that depends only on the dimension m and c_{m+1} being a normalization function.

Corollary B.1. *$\mathcal{H}_{\odot,m+1}$ is dense in $\mathcal{C}_{s,m}$*

Proof. Theorem B.1 holds for all $\epsilon > 0$ and Lemma A.2. □

Theorem B.1 is a Jackson-type result as Equation (9) gives the number N of control points needed to approximate f with accuracy ϵ. This corresponds to the length of the prefix sequence. Moreover, the smoother the target f is, i.e., the smaller L, C_H, the shorter the prefix length N. Thus, our construction uses only as much prefix positions as necessary.

The proof of Theorem B.1 follows closely (Ng and Kwong, 2022). While they only provide a density result, we offer a Jackson-type bound which is non-trivial and may be of an independent interest. The idea behind the proof is as following. We first approximate f with its convolution with a kernel having the form of the terms in Equation (7):

$$(f * K_{\lambda}^{\text{vMF}})(\boldsymbol{x}){=}\int_{S_m} c_{m+1}(\lambda)\exp(\lambda\langle\boldsymbol{x},\boldsymbol{y}\rangle)f(\boldsymbol{y})\, dw_m(\boldsymbol{y}). \tag{10}$$

The larger the λ is, the closer $f * K_{\lambda}^{\text{vMF}}$ is to f and hence the smaller the approximation error (Menegatto, 1997). $\Lambda(\epsilon/2)$ gives the smallest value for λ such that this error is $\epsilon/2$. Equation (10) can then be approximated with sums: we partition S^m into N sets $V_1, ..., V_N$ small enough that f does not vary too much within each set. Each control point $\boldsymbol{p}_k^{\alpha}$ is placed in its corresponding V_k. Then, $\exp(\lambda\langle\boldsymbol{x},\boldsymbol{y}\rangle)f(\boldsymbol{y})$ can be approximated with $\exp(\lambda\langle\boldsymbol{x},\boldsymbol{p}_k^{\alpha}\rangle)f(\boldsymbol{p}_k^{\alpha})$ when $\boldsymbol{y}$ is in the k-th set V_k. Hence, Equation (10) can be approximated with $\sum_{k=1}^{N} \exp(\lambda\langle\boldsymbol{x},\boldsymbol{p}_k^{\alpha}\rangle)\; Cf(\boldsymbol{p}_k^{\alpha})$ for some suitable constant C. By increasing N we can reduce the error of approximating the convolution with the sum. Equation (9) gives us the minimum N needed so that this error is $\epsilon/2$. Hence, we have error of at most $\epsilon/2$ from approximating f with the convolution and $\epsilon/2$ from approximating the convolution with the sum, resulting in our overall error being bounded by ϵ. The full proof is in Appendix E and is illustrated in Figure 4. The theorem can be extended to vector-valued functions in $\mathcal{C}_{v,m}$ with a multiplicative factor $1/\sqrt{m+1}$:

Corollary B.2. *Let $f : S^m \to \mathbb{R}^{m+1}$, $m \geq 8$ be such that each component f_i satisfies the conditions in Theorem B.1. Define $\|f\|_\infty = \max_{1 \leq i \leq m+1} \|f_i\|_\infty$. Then, for any $\epsilon > 0$, there exist $\boldsymbol{p}_1^\alpha, ..., \boldsymbol{p}_N^\alpha \in S^m$ and $\boldsymbol{p}_1^\beta, ..., \boldsymbol{p}_N^\beta \in \mathbb{R}^{m+1}$ such that*

$$\sup_{\boldsymbol{x} \in S^m} \left\| f(\boldsymbol{x}) - \sum_{k=1}^{N} \exp(\lambda \langle \boldsymbol{x}, \boldsymbol{p}_k^\alpha \rangle) \boldsymbol{p}_k^\beta \right\|_2 \leq \epsilon,$$

with $\lambda = \Lambda({}^{\epsilon}/_{2\sqrt{m+1}})$ for any $N \geq N(\lambda, {}^{\epsilon}/_{\sqrt{m+1}})$. That is, $\mathcal{H}_{\circledast,m+1}$ is dense in $\mathcal{C}_{v,m}$ with respect to the $\|\cdot\|_2$ norm.

Thanks to Theorem B.1 and Corollary B.2, we know that functions in $\mathcal{C}_{v,m}$ can be approximated by core attention (Equation (6)). We only have to demonstrate that a core attention head can be represented as a classical attention head (Equation (1)). We do this by reversing the simplifications we made when constructing the core attention head.

Let us start by bringing the normalization term back, resulting in $\mathcal{H}^N_{\Vert,d}$, the split attention head hypothesis (Definition B.1). Intuitively, $d(\boldsymbol{x}) = \sum_{k=1}^{n} \exp(\lambda \langle \boldsymbol{x}, \boldsymbol{p}_k^\alpha \rangle)$ is almost constant when the $\boldsymbol{p}_k^\alpha$ are uniformly distributed over the sphere as the distribution of distances from $\boldsymbol{x}$ to $\boldsymbol{p}_k^\alpha$ will be similar, regardless of where $\boldsymbol{x}$ lies. We can bound how far $d(\boldsymbol{x})$ is from being a constant and adjust the approximation error to account for it. Appendix F has the full proof.

Theorem B.2. *Let $f : S^m \to \mathbb{R}^{m+1}$, $m \geq 8$ be such that each component f_i satisfies the conditions in Theorem B.1. Then, for any $0 < \epsilon < 2\|f\|_\infty$, there exist $\boldsymbol{p}_1^\alpha, ..., \boldsymbol{p}_N^\alpha \in S^m$ such that*

$$\sup_{\boldsymbol{x} \in S^m} \left\| f(\boldsymbol{x}) - \frac{\sum_{k=1}^{N} \exp(\lambda \langle \boldsymbol{x}, \boldsymbol{p}_k^\alpha \rangle) \boldsymbol{p}_k^\beta}{\sum_{k=1}^{N} \exp(\lambda \langle \boldsymbol{x}, \boldsymbol{p}_k^\alpha \rangle)} \right\|_2 \leq \epsilon,$$

with

$$\lambda = \Lambda\left(\frac{\epsilon(\|f\|_\infty + L)}{\sqrt{m+1}(3\|f\|_\infty + 2L) - {}^{\epsilon}/_{2}} \right)$$
$$\boldsymbol{p}_k^\beta = f(\boldsymbol{p}_k^\alpha), \ \forall k = 1, \ldots, N,$$

for any $N \geq N(\lambda, {}^{\epsilon}/_{\sqrt{m+1}})$. That is, $\mathcal{H}_{\Vert,m+1}$ is dense in $\mathcal{C}_{v,m}$ with respect to the $\|\cdot\|_2$ norm.

An interesting observation is that adding the normalization term has not affected the asymptotic behavior of λ and hence also of the prefix length N. Furthermore, notice how the value $\boldsymbol{p}_i^\beta$ at the control point $\boldsymbol{p}_i^\alpha$ is simply $f(\boldsymbol{p}_k^\alpha)$, the target function evaluated at this control point.

We ultimately care about the ability of the classical attention head (Definition 1) to approximate functions in $\mathcal{C}_{v,m}$ by prefixing. Hence, we need to bring back the terms depending only on the input $\boldsymbol{x}$, combine the α and β parts of the prefix into a single vector and bring back the $\boldsymbol{H}$ and $\boldsymbol{W}_V$ matrices. One can do this by considering an attention head with a hidden dimension $3(m+1)$ allowing us to place $\boldsymbol{x}, \boldsymbol{p}_k^\alpha$ and $\boldsymbol{p}_k^\beta$ in different subspaces of the embedding space. To do this, define a pair of embedding and projection operations:

$$\Pi : S^m \to \mathbb{R}^{3(m+1)} \qquad \Pi^{-1} : \mathbb{R}^{3(m+1)} \to \mathbb{R}^{m+1}$$
$$\boldsymbol{x} \mapsto \begin{bmatrix} \boldsymbol{I}_{m+1} \\ \boldsymbol{0}_{m+1} \\ \boldsymbol{0}_{m+1} \end{bmatrix} \boldsymbol{x} \qquad \boldsymbol{x} \mapsto \begin{bmatrix} \boldsymbol{I}_{m+1} \\ \boldsymbol{0}_{m+1} \\ \boldsymbol{0}_{m+1} \end{bmatrix}^\top \boldsymbol{x}.$$

Lemma B.1. *$\Pi^{-1} \circ \mathcal{H}^1_{-,3(m+1)} \circ \Pi$ is dense in $\mathcal{H}_{\Vert,m+1}$, with the composition applied to each function in the class.*

The proof is in Appendix G. Lemma B.1 shows that every split attention head can be *exactly* represented as 3 times bigger classical attention head. Note that our choice for $\boldsymbol{H}$ and $\boldsymbol{W}_V$ is not unique. Equivalent constructions are available by multiplying each component by an invertible matrix, effectively changing the basis of the system. Finally, the embedding and projection operations can be represented as MLPs and hence can be embedded in a transformer architecture. Now, we can provide the final result of this section, namely that the standard attention head of a transformer can approximate any vector-valued function on the hypersphere:

Theorem B.3. *Let $f : S^m \to \mathbb{R}^{m+1}$, $m \geq 8$ be such that each component f_i satisfies the conditions in Theorem B.1. Then, for any $0 < \epsilon \leq 2\|f\|_\infty$, there exists an attention head $h \in \mathcal{H}^{N,1}_{-,3(m+1)}$ such that*

$$\sup\nolimits_{\boldsymbol{x}\in S^m} \|f(\boldsymbol{x}) - (\Pi^{-1} \circ h \circ \Pi)(\boldsymbol{x})\|_2 \leq \epsilon, \tag{11}$$

for any $N \geq N(\lambda, \epsilon/\sqrt{m+1})$. That is, $\Pi^{-1} \circ \mathcal{H}^1_{-,3(m+1)} \circ \Pi$ is dense in $\mathcal{C}_{v,m}$ with respect to the $\|\cdot\|_2$ norm.

Proof. The density result follows directly from Theorem B.2 and Lemma B.1 and transitivity (Lemma A.1). The Jackson bound is the same as in Theorem B.2 as transforming the split attention head to a classical attention head is exact and does not contribute further error. □

Therefore, we have shown that a single attention head with a hidden state $3(m+1)$ can approximate any continuous function $f : C(S^m) \to \mathbb{R}^{m+1}$ to an arbitrary accuracy. This is for *fixed* pre-trained components, that is, $\boldsymbol{H}$ and $\boldsymbol{W}_V$ are as given in the proof of Lemma B.1 and depend neither on the input $\boldsymbol{x}$ nor on the target function f. Therefore, the behavior of the attention head is fully controlled by the prefix. This is a Jackson-type result, with the length N of the prefix given in Theorem B.2. To the best of our knowledge, Theorem B.3 is the first bound on the necessary prefix length to achieve a desired accuracy of function approximation using an attention head. Most critically, Theorem B.3 demonstrates that attention heads are more expressive than commonly thought. A *single* attention head with a very simple structure can be a universal approximator.

C Universal Approximation of Sequence-to-Sequence Functions

The previous section showed how we can approximate any continuous $f : S^m \to \mathbb{R}^{m+1}$ with a single attention head. Still, one typically uses the transformer architecture for operations over sequences rather than over single inputs (the case with $T \geq 1$). We will now show how we can leverage Theorem B.3 to model general sequence-to-sequence functions. First, we show the simpler case of functions that apply the exact same mapping to all inputs. We then show how to model general sequence-to-sequence functions using a variant of the Kolmogorov–Arnold theorem.

Element-wise functions Theorem B.3 can be extended to element-wise functions where the exact same function is applied to each element in the input sequence, i.e., the concept class $\mathcal{C}_{\|,T,m}$ from Definition A.4. If $f \in \mathcal{C}_{\|,T,m}$, then there exists a $g \in \mathcal{C}_{v,m}$ such that $f(\boldsymbol{x}_1, \ldots, \boldsymbol{x}_T) = (g(\boldsymbol{x}_1), \ldots, g(\boldsymbol{x}_T))$. By Theorem B.3, there exists a prefix $\boldsymbol{p}_1, ..., \boldsymbol{p}_N$ that approximates g. As the construction in Lemma B.1 prevents interactions between two different inputs $\boldsymbol{x}_i$ and $\boldsymbol{x}_j$, an attention head $h^T \in \mathcal{H}^{N,T}_{-,3(m+1)}$ for a T-long input (Equation (1)) with the exact same prefix $\boldsymbol{p}_1, \ldots, \boldsymbol{p}_N$ approximates f:

Corollary C.1. $\Pi^{-1} \circ \mathcal{H}^T_{-,3(m+1)} \circ \Pi$ *is dense in $\mathcal{C}_{\|,T,m}$ with respect to the $\|\cdot\|_2$ norm applied element-wise. That is, for every $\epsilon > 0$, there exists $h^T \in \mathcal{H}^{N,T}_{-,3(m+1)}$ such that:*

$$\sup_{\{\boldsymbol{x}_i\}\in(S^m)^T} \max_{1\leq k\leq T} \left\| \left[f(\{\boldsymbol{x}_i\}) - (\Pi^{-1} \circ h^T \circ \Pi)(\{\boldsymbol{x}_i\}) \right]_k \right\|_2 \leq \epsilon,$$

with Π and Π^{-1} applied element-wise, $[\cdot]_k$ selecting the k-th element, and approximate rate bound on N as in Theorem B.3.

General sequence-to-sequence functions Ultimately, we are interested in modeling arbitrary functions from sequences of inputs $(\boldsymbol{x}_1, \ldots, \boldsymbol{x}_T)$ to sequences of outputs $(\boldsymbol{y}_1, \ldots, \boldsymbol{y}_T)$, that is, the $\mathcal{C}_{T,m}$. We will use a version of the Kolmogorov–Arnold representation Theorem. The theorem is typically defined over functions over the unit hypercube $[0,1]^m$. As there exists a homeomorphism between $[0,1]^m$ and a subset of S^m (Lemma G.1), for simplicity, we will ignore this technical detail. Our construction requires only $T + 2$ attention layers, each with a single head.

The original Kolmogorov-Arnold representation theorem (Kolmogorov, 1957) identifies every continuous function $f : [0,1]^d \to \mathbb{R}$ with univariate functions g_q, $\psi_{p,q}$ such that:

$$f(x_1, \ldots, x_d) = \sum\nolimits_{q=0}^{2d} g_q \left(\sum\nolimits_{p=1}^{d} \psi_{p,q}(x_p) \right).$$

In other words, multivariate functions can be represented as sums and compositions of univariate functions. As transformers are good at summing and attention heads are good at approximating functions, they can approximate functions of this form. However, g_q and $\psi_{p,q}$ are generally not well-behaved (Girosi and Poggio, 1989), so we will use the construction by Schmidt-Hieber (2021) instead.

Lemma C.1 (Theorem 2 in (Schmidt-Hieber, 2021))**.** *For a fixed d, there exists a monotone functions $\psi : [0,1] \to C$ (the Cantor set) such that for any function $f : [0,1]^d \to \mathbb{R}$, we can find a function $g : C \to \mathbb{R}$ such that*

i. $f(x_1, \dots, x_d) = g\left(3 \sum_{p=1}^{d} 3^{-p}\, \psi(x_p)\right),$ (12)

ii. if f is continuous, then g is also continuous,

iii. if $|f(\boldsymbol{x}) - f(\boldsymbol{y})| \leq Q|\boldsymbol{x} - \boldsymbol{y}|_\infty$, for all $\boldsymbol{x}, \boldsymbol{y} \in [0,1]^d$ and some Q, then $|g(x) - g(y)| \leq 2Q, \forall x, y \in C$.

In comparison with the original Kolmogorov–Arnold theorem, we need a single inner function ψ which does not depend on the target function f and only one outer function g. Furthermore, both ψ and g are Lipschitz. Hence, we can approximate them with our results from Appendix B.

We need to modify Lemma C.1 a bit to make it fit the sequence-to-sequence setting. First, flatten a sequence of T $(m+1)$-dimensional vectors into a single vector in $[0,1]^{(m+1)T}$. Second, define $\Psi_d : [0,1]^d \to \mathbb{R}^d$ to be the element-wise application of ψ: $\Psi_d(\{x_i\}_{i=1}^d) = \{\psi(x_i)\}_{i=1}^d$. We can also define $G_i : C \to \mathbb{R}^{m+1},\ i = 1, \dots, T$ and extend Equation (12) for our setting:

$$\begin{aligned} f(\boldsymbol{x}_1, ..., \boldsymbol{x}_T) = (G_1(R), \dots, G_T(R)),\ \text{with} & \\ R = 3 \sum_{i=1}^{T} 3^{-(i-1)(m+1)} \sum_{p=1}^{m+1} 3^{-p} \psi(\boldsymbol{x}_{i,p}) & \qquad (13) \\ = 3 \sum_{i=1}^{T} 3^{-(i-1)(m+1)} \begin{bmatrix} 3^{-1} \\ \vdots \\ 3^{-(m+1)} \end{bmatrix}^{\top} \Psi_{m+1}(\boldsymbol{x}_i). & \end{aligned}$$

Equation (13) can now be represented with a transformer with $T + 2$ attention layers. Ψ_{m+1} is applied element-wise, hence, all $\Psi_{m+1}(\boldsymbol{x}_i)$ can be computed in parallel with a single attention head (Corollary C.1). The dot product with the $\begin{bmatrix} 3^{-1} & \cdots & 3^{-(m+1)} \end{bmatrix}$ vector can be computed using a single MLP. The product with the $3^{-(i-1)(m+1)}$ scalar is a bit more challenging as it depends on the position in the sequence. However, if we concatenate position encodings to the input, another MLP can use them to compute this factor and the multiplication. The outer sum over the T inputs and the multiplication by 3 can be achieved with a single attention head. Hence, using only 2 attention layers, we have compressed the whole sequence in a single scalar R.[1]

The only thing left is to apply $G_1, \dots, G_T$ to R to compute each of the T outputs. As each one of these is Lipschitz, we can approximate each with a single attention head using Theorem B.3. Each G_i is different and would need its own set of prefixes, requiring T attention heads arranged in T attention layers. Using the positional encodings, each layer can compute the output for its corresponding position and pass the input unmodified for the other positions. The overall prefix size would be the longest of the prefixes necessary to approximate $\Psi_{m+1}, G_1, \dots, G_T$.

Hence, we have constructed an architecture that can approximate any sequence-to-sequence function $f \in \mathcal{C}_{T,m}$ with only T+2 attention layers. Thus, $\mathcal{H}^T_{\equiv,d}$ is dense in $\mathcal{C}_{T,m}$.

D Background on Analysis on the Sphere

As mentioned in the main text, the investigation of the properties of attention heads naturally leads to analysing functions over the hypersphere. To this end, our results require some basic facts about the

[1] Yun et al. (2019) use a similar approach but use discretization to enumerate all possible sequences and require $\mathcal{O}(\epsilon^{-m})$ attention layers. In our continuous setting, R is computed with 2 layers.

analysis on the hypersphere. We will review them in this appendix. For a comprehensive reference, we recommend (Atkinson and Han, 2012) and (Dai and Xu, 2013).

Define $\mathbb{P}_k(\mathbb{R}^{m+1})$ to be the space of polynomials of degree at most k. The restriction of a polynomial $p \in \mathbb{P}_k(\mathbb{R}^{m+1})$ to the unit hypersphere $S^m = \{\boldsymbol{x} \in \mathbb{R}^{m+1} \mid \|\boldsymbol{x}\|_2 = 1\}$ is called a *spherical polynomial*. We can thus define the space of spherical polynomials:

$$\mathbb{P}_k(S^m) = \{p|_{S^m} \text{ for } p \in \mathbb{P}_k(\mathbb{R}^{m+1})\}.$$

Define by $\mathbb{H}_k(\mathbb{R}^{m+1})$ the space of polynomials of degree k that are homogeneous:

$$\mathbb{H}_k(\mathbb{R}^{m+1}) = \left\{(x_1, \ldots, x_{m+1}) \mapsto x_1^{\alpha_1} \times \cdots \times x_{m+1}^{\alpha_{m+1}} \;\middle|\; \sum_{i=1}^{m+1} \alpha_i = k\right\}.$$

Its restriction to the sphere $\mathbb{H}_k(S^m)$ is defined analogously to $\mathbb{P}_k(S^m)$. Finally, we can define the space $\mathbb{Y}(\mathbb{R}^{m+1})$ of harmonic homogeneous polynomials:

$$\mathbb{Y}_k(\mathbb{R}^{m+1}) = \left\{p \in \mathbb{H}_k(\mathbb{R}^{m+1}) \;\middle|\; \frac{\partial^2}{\partial x^2}\, p(\boldsymbol{x}) = 0,\ \forall \boldsymbol{x} \in \mathbb{R}^{m+1}\right\}.$$

$\mathbb{Y}_k(S^m)$ which is the restriction of $\mathbb{Y}_k(\mathbb{R}^{m+1})$ to S^m is the set of *spherical harmonics* of degree k. Spherical harmonics are the higher-dimensional extension of Fourier series.

Notably, even though

$$\mathbb{Y}_k(S^m) \subset \mathbb{H}_k(S^m) \subset \mathbb{P}_k(S^m),$$

the restriction of any polynomial on S^m is a sum of spherical harmonics:

$$\mathbb{P}_k(S^m) = \mathbb{Y}_0(S^m) \oplus \cdots \oplus \mathbb{Y}_k(S^m),$$

with $\oplus$ being the direct sum (Atkinson and Han, 2012, Corollary 2.19).

We define $C(S^m)$ to be the space of all continuous functions defined on S^m with the uniform norm

$$\|f\|_\infty = \sup_{\boldsymbol{x} \in S^m} |f(\boldsymbol{x})|,\ f \in C(S^m). \tag{14}$$

Similarly, $\mathcal{L}_p(S^m), 1 \leq p < \infty$ is the space of all functions defined on S^m which are integrable with respect to the standard surface measure dw_m. The norm in this space is:

$$\|f\|_p = \left(\frac{1}{w_m} \int_{S^m} |f(\boldsymbol{x})|^p \, dw_m(\boldsymbol{x})\right)^{1/p},\ f \in \mathcal{L}_p(S^m), \tag{15}$$

with the surface area being

$$w_m = \int_{S^m} dw_m = \frac{2\pi^{(m+1)/2}}{\Gamma(^{(m+1)}/_2)}. \tag{16}$$

We will use V_m to denote any of these two spaces and $\|\cdot\|_m$ the corresponding norm.

A key property of spherical harmonics is that sums of spherical harmonics can uniformly approximate the functions in $C(S^m)$. In other words, the span of $\bigcup_{k=0}^\infty \mathbb{Y}_k(S^m)$ is dense in $C(S^m)$ with respect to the uniform norm $\|\cdot\|_\infty$. Hence, any $f \in C(S^m)$ can be expressed as a series of spherical harmonics:

$$f(\boldsymbol{x}) = \sum_{k=0}^\infty Y_k^m(\boldsymbol{x}),\ \text{with } Y_k^m \in \mathbb{Y}_k(S^m),\ \forall k.$$

We will also make a heavy use of the concept of spherical convolutions. Define the space of kernels $\mathcal{L}^{1,m}$ to consist of all measurable functions K on $[-1, 1]$ with norm

$$\|K\|_{1,m} = \frac{w_{m-1}}{w_m} \int_{-1}^1 |K(t)|(1-t^2)^{(m-2)/2} \, dt \ < \infty.$$

Definition D.1 (Spherical convolution)**.** The spherical convolution $K * f$ of a kernel K in $\mathcal{L}^{1,m}$ with a function $f \in V_m$ is defined by:

$$(K * f)(\boldsymbol{x}) = \frac{1}{w_m} \int_{S^m} K(\langle \boldsymbol{x}, \boldsymbol{y} \rangle) f(\boldsymbol{y}) \, dw_m(\boldsymbol{y}),\ \boldsymbol{x} \in S^m.$$

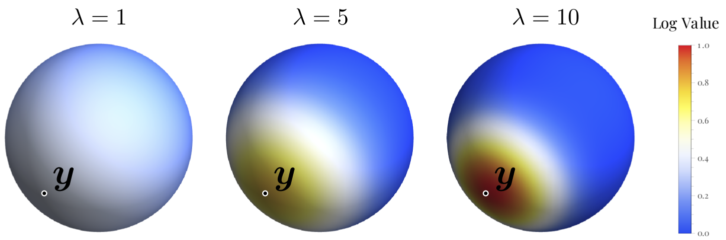


Figure 3: Plots of the von Mises-Fisher kernel $K_\lambda^{\text{vMF}}(\langle \boldsymbol{x}, \boldsymbol{y} \rangle)$ for $\lambda = 1, 5, 10$ and fixed $\boldsymbol{y}$ in three dimensions ($m = 2$). The larger λ is, the more concentrated the kernel is around $\boldsymbol{y}$.

Spherical convolutions map functions $f \in V_m$ to functions in V_m. Furthermore, the spherical harmonics are eigenfunctions of the function generated by a kernel in $\mathcal{L}^{1,m}$:

Lemma D.1 (Funk and Hecke's formula (Funk, 1915; Hecke, 1917; Estrada, 2014))**.**

$$K * Y_k^m = a_k^m(K) Y_k^m, \textit{ when } K \in \mathcal{L}^{1,m},\ Y_k^m \in \mathbb{Y}_k(S^m), k = 0, 1, \dots,$$

where $a_k^m(K)$ are the coefficients in the series expansion in terms of Gegenbauer polynomials associated with the kernel K:

$$a_k^m(K) = \frac{w_{m-1}}{w_m} \int_{-1}^{1} K(t) \frac{Q_k^{(m-1)/2}(t)}{Q_k^{(m-1)/2}(1)} (1 - t^2)^{(m-2)/2}\, dt, k = 0, 1, \dots . \tag{17}$$

Here, $Q_k^{(m-1)/2}$ is the Gegenbauer polynomial of degree k.

Note also that with a change of variables we have:

$$\int_{S^m} K(\langle \boldsymbol{x}, \boldsymbol{y} \rangle)\, dw_m(\boldsymbol{y}) = w_{m-1} \int_{-1}^{1} K(t)(1 - t^2)^{(m-2)/2}\, dt. \tag{18}$$

Ideally, we would like a kernel that acts as an identity for the convolution operation. In this case, we would have $\|K * f - f\|_\infty = 0, f \in \mathcal{C}(S^m)$ which would be rather convenient. However, there is no such kernel in the spherical setting (Menegatto, 1997). The next best thing is to construct a sequence of kernels $\{K_n\} \in \mathcal{L}^{1,m}$ such that $\|K_n * f - f\|_m \to 0$ as $n \to \infty$ for all $f \in V_m$. This sequence of kernels is called an *approximate identity*. The specific such sequence of kernels we will use is based on the von Mises-Fisher distribution as this gives us the $\exp(\langle \boldsymbol{x}, \boldsymbol{y} \rangle)$ form that we also observe in the transformer attention mechanism.

Definition D.2 (von Mises-Fisher kernels, (Ng and Kwong, 2022))**.** We define the sequence of von Mises-Fisher kernels as:

$$K_\lambda^{\text{vMF}}(t) = c_{m+1}(\lambda) \exp(\lambda t),\ t \in [-1, 1],$$

where

$$c_{m+1}(\lambda) = \frac{w_m \lambda^{\frac{m+1}{2} - 1}}{(2\pi)^{\frac{m+1}{2}} I_{\frac{m+1}{2} - 1}(\lambda)},$$

with I_v being the modified Bessel function at order v.

Note that a von Mises-Fisher kernel can also be expressed in terms of points on S^m. In particular, for a fixed $\boldsymbol{y} \in S^m$ we have $K_\lambda^{\text{vMF}}(\langle \boldsymbol{x}, \boldsymbol{y} \rangle), \boldsymbol{x} \in S^m$. The parameter λ is a "peakiness" parameter: the large λ is, the closer $K_\lambda^{\text{vMF}}(\langle \boldsymbol{x}, \boldsymbol{y} \rangle)$ approximates the delta function centered at $\boldsymbol{y}$, as can be seen in Figure 3. It is easy to check that $\|K_\lambda^{\text{vMF}}\|_{1,m} = 1,\ \forall \lambda > 1, m > 1$ and hence the sequence is in $\mathcal{L}^{1,m}$, meaning they are valid kernels. Ng and Kwong (2022, Lemma 4.2) show that $\{K_\lambda^{\text{vMF}}\}$ is indeed an approximate identity, i.e., $\|K_\lambda^{\text{vMF}} * f - f\|_m \to 0$ as $\lambda \to \infty$ for all $f \in V_m$.[2] As we want a Jackson-type result however, we will need to upper bound the error $\|K_\lambda^{\text{vMF}} * f - f\|_m$ as a function of λ, that is a non-asymptotic result on the quality of the approximation by spherical convolutions with K_λ^{vMF}. We do that in Lemma E.5.

[2] The w_m term in the normalization constant $c_{m+1}(\lambda)$ is not in (Ng and Kwong, 2022). However, without it K_λ^{vMF} are not an approximate identity.

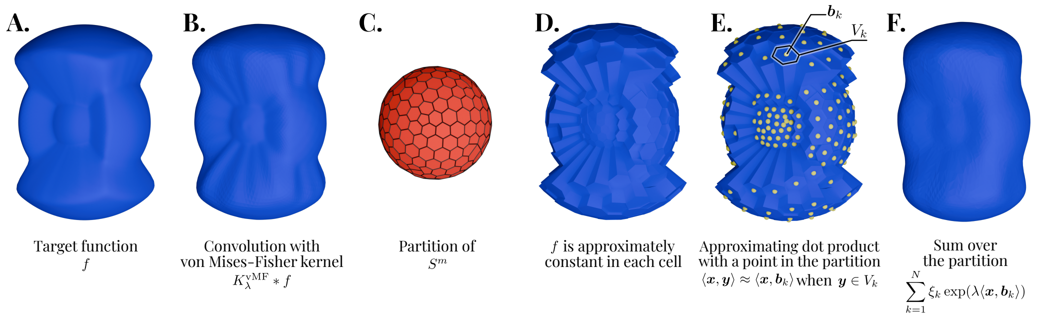

A.
B.
C.
D.
E.
F.
$\boldsymbol{b}_k$
V_k
Target function
f
Convolution with
von Mises-Fisher kernel
$K_\lambda^{\text{vMF}} * f$
Partition of
S^m
f is approximately
constant in each cell
Approximating dot product
with a point in the partition
$\langle \boldsymbol{x}, \boldsymbol{y} \rangle \approx \langle \boldsymbol{x}, \boldsymbol{b}_k \rangle$ when $\boldsymbol{y} \in V_k$
Sum over
the partition
$\sum_{k=1}^{N} \xi_k \exp(\lambda \langle \boldsymbol{x}, \boldsymbol{b}_k \rangle)$

Figure 4: **Intuition behind the proof of our Jackson-type bound for universal approximation on the hypersphere.** **A.** We want to approximate a function f over the hypersphere S^m. This illustration is in three-dimensional space, so $m = 2$. **B.** In order to get the $\exp(\lambda\langle \cdot, \boldsymbol{y}\rangle)$ form that we want, we convolve f with the $K_\lambda^{\text{vMF}}(t) = c_{m+1}(\lambda)\exp(\lambda t)$ kernel. **C.** We partition S^m into N cells $V_1,\ldots,V_N$. **D.** Our choice of N is such that f does not vary too much in each cell and hence can be approximated by a function that is constant in each V_k. **E.** As each cell is small, the dot product of $\boldsymbol{x}$ with any point in the cell V_k can be approximated by the dot product of $\boldsymbol{x}$ with a fixed point $\boldsymbol{b}_k \in V_k$. **F.** This allows us to approximate the integral in the convolution K_λ^{vMF} with a finite sum.

E A Jackson-type Bound for Universal Approximation on the Unit Hypersphere

The overarching goal in this section is to provide a Jackson-type (Definition A.2) bound for approximating functions $f : S^m \to \mathbb{R}^{m+1}$ on the hypersphere $S^m = \{\boldsymbol{x} \in \mathbb{R}^{m+1} \mid \|\boldsymbol{x}\|_2 = 1\}$ by functions of the form

$$h(\boldsymbol{x}) = \sum_{k=1}^{N} \boldsymbol{\xi}_k \exp(\lambda\langle \boldsymbol{x}, \boldsymbol{b}_k\rangle) \tag{19}$$

To this end, we will leverage results from approximation on the hypersphere using spherical convolutions by Menegatto (1997) and recent results on the universal approximation on the hypersphere by Ng and Kwong (2022). While these two works inspire the general proof strategy, they only offer uniform convergence (i.e., density-type results, Definition A.1). Instead, we offer a non-asymptotic analysis and develop the first approximation rate results on the sphere for functions of the form of Equation (19), i.e., Jackson-type results (Definition A.2).

The high-level idea of the proof is to split the goal into approximating f with the convolution $f * K_\lambda^{\text{vMF}}$ and approximating the convolution $f * K_\lambda^{\text{vMF}}$ with a sum of terms that have the $\exp(t\langle \boldsymbol{x}, \boldsymbol{b}_k\rangle)$ structure resembling the kernel K_λ^{vMF} (Definition D.2):

$$\sup_{\boldsymbol{x}\in S^m} \left\| f(\boldsymbol{x}) - \sum_{k=1}^{N} \xi_k \exp(\lambda\langle \boldsymbol{x}, \boldsymbol{b}_k\rangle) \right\| \leq \underbrace{\left\| f - f * K_\lambda^{\text{vMF}} \right\|_\infty}_{\text{Equation (21) / Lemma E.5}} + \underbrace{\left\| f * K_\lambda^{\text{vMF}} - \sum_{k=1}^{N} \xi_k \exp(\lambda\langle \cdot, \boldsymbol{b}_k\rangle) \right\|_\infty}_{\text{Lemma E.7}}. \tag{20}$$

This is also illustrated in Figure 4.

Let's focus on the first term in Equation (20). It can be further decomposed into three terms by introducing $W_q \in \mathbb{P}_q(S^m)$, the best approximation of f with a spherical polynomial of degree q:

$$\|K_\lambda^{\text{vMF}} * f - f\|_\infty \leq \underbrace{\|K_\lambda^{\text{vMF}} * f - K_\lambda^{\text{vMF}} * W_q\|_\infty}_{\text{Lemma E.2}} + \underbrace{\|K_\lambda^{\text{vMF}} * W_q - W_q\|_\infty}_{\text{Lemma E.4}} + \underbrace{\|W_q - f\|_\infty}_{\text{Lemma E.1}}. \tag{21}$$

There are a number of Jackson-type results for how well finite sums of spherical polynomials approximate functions $f \in V_m$ (the last term in Equation (21)). In particular they are interested in bounding

$$\min_{W_q\in\mathbb{P}_q(S^m)} \|f - W_q\|_p,\ 1 \leq p \leq \infty. \tag{22}$$

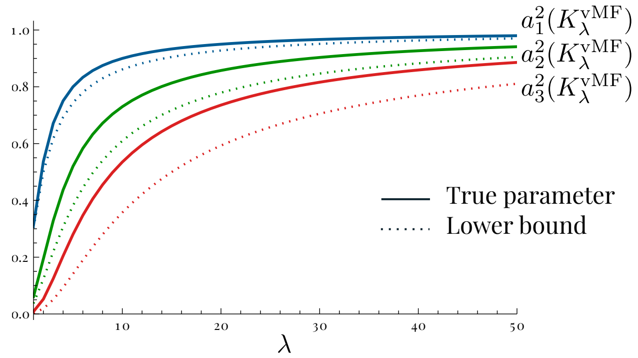


Figure 5: The coefficients a_k^m for the von Mises-Fisher kernels K_λ^{vMF} for $m = 2$ and $k \in \{1, 2, 3\}$ as well as the lower bound from Lemma E.3.

We will use a simple bound by Ragozin (1971):

Lemma E.1 (Ragozin bound)**.** *For $f \in C(S^m)$ and $q \in \mathbb{N}_{>0}$ it holds that:*

$$\min_{W_q \in \mathbb{P}_q(S^m)} \|f - W_q\|_\infty \leq C_R\, \omega\left(f; \frac{1}{q}\right), \tag{23}$$

for some constant C_R that does not depend on f or q and ω being the first modulus of continuity of f defined as:

$$\omega(f; t) = \sup\{|f(\boldsymbol{x}) - f(\boldsymbol{y})| \mid \boldsymbol{x}, \boldsymbol{y} \in S^m, \cos^{-1}(\boldsymbol{x}^\top \boldsymbol{y}) \leq t\}.$$

We recommend Atkinson and Han (2012, Chapter 4) and Dai and Xu (2013, Chapter 4) for an overview of the various bounds proposed for Equation (22) depending on the continuity properties of f and its derivatives. In particular, the above bound could be improved with a term $1/n^k$ if f has k continuous derivates (Ragozin, 1971).

We can upper-bound the first term in Equation (21) by recalling that the norm of the kernel K_λ^{vMF} is 1:

Lemma E.2.

$$\|K_\lambda^{vMF} * f - K_\lambda^{vMF} * W_q\|_m \leq \|f - W_q\|_m.$$

Hence:

$$\|K_\lambda^{vMF} * f - K_\lambda^{vMF} * W_q\|_\infty \leq \|f - W_q\|_\infty \leq C_R\, \omega\left(f; \frac{1}{q}\right).$$

Proof. Convolution is linear so $\|K_\lambda^{\mathrm{vMF}} * f - K_\lambda^{\mathrm{vMF}} * W_q\|_m = \|K_\lambda^{\mathrm{vMF}} * (f - W_q)\|_m$. Using the Hölder inequality (Dai and Xu, 2013, Theorem 2.1.2) we get $\|K_\lambda^{\mathrm{vMF}} * f - K_\lambda^{\mathrm{vMF}} * W_q\|_m \leq \|K_\lambda^{\mathrm{vMF}}\|_{1,m} \|f - W_q\|_m$. As $\|K_\lambda^{\mathrm{vMF}}\|_{1,m} = 1$ for all $\lambda > 0, m > 1$, we obtain the inequality in the lemma. For the uniform norm, we also use the Ragozin bound from Lemma E.1. □

Only the second term in Equation (21) is left. However, before we tackle it, we will need a helper lemma that bounds the eigenvalues of the von Mises-Fisher kernel (Equation (17)):

Lemma E.3 (Bounds on the eigenvalues $a_k^m(K_\lambda^{\mathrm{vMF}})$)**.** *The eigenvalues a_k^m, as defined in Equation (17), for the sequence of von Mises-Fisher kernels (Definition D.2) are bounded from below and above as:*

$$0 < \left(\frac{\lambda}{\left(\frac{m-1}{2} + k\right) + \sqrt{\lambda^2 + \left(\frac{m-1}{2} + k\right)^2}} \right)^k \leq a_k^m(K_\lambda^{vMF}) \leq 1$$

Proof. We have

$$\begin{aligned}
a_k^m(K_\lambda^{\text{vMF}}) &= \frac{w_{m-1}}{w_m}\int_{-1}^{1} K_\lambda^{\text{vMF}}(t)\frac{Q_k^{(m-1)/2}(t)}{Q_k^{(m-1)/2}(1)}(1-t^2)^{(m-2)/2}dt \\
&= \frac{w_{m-1}}{w_m}\int_{-1}^{1} c_{m+1}(\lambda)\exp(\lambda t)\frac{Q_k^{(m-1)/2}(t)}{Q_k^{(m-1)/2}(1)}(1-t^2)^{(m-2)/2}dt \\
&\overset{\star}{=} \frac{I_{\frac{m-1}{2}+k}(\lambda)}{I_{\frac{m-1}{2}}(\lambda)},
\end{aligned}$$

with $\star$ solved using Mathematica. From here we can see that $a_0^m(K_\lambda^{\text{vMF}}) = 1$ for all $m > 1, \lambda > 1$. Furthermore, for $v > 1$ and $\lambda > 0$ the modified Bessel function of the first kind $I_v(\lambda)$ is monotonically decreasing as v increases. Therefore, $a_k^m(K_\lambda^{\text{vMF}}) \leq a_0^m(K_\lambda^{\text{vMF}}) = 1$, which gives us the upper bound in the lemma.

For the lower bound, we will use the following bound on the ratio of modified Bessel functions by Amos (1974, Eq. 9):

$$\frac{I_{v+1}(x)}{I_v(x)} \geq \frac{x}{(v+1)+\sqrt{x^2+(v+1)^2}},\ v \geq 0, x \geq 0.$$

As mentioned above, $0 \leq \frac{I_{v+1}(x)}{I_v(x)} \leq 1$. Furthermore, these ratios are decreasing as v increases, i.e., $\frac{I_{v+2}(x)}{I_{v+1}(x)} \leq \frac{I_{v+1}(x)}{I_v(x)}$ for all $v \geq 0$ and $x \geq 0$ (Amos, 1974, Eq. 10). Combining these facts gives us:

$$\frac{I_{v+k}(x)}{I_v(x)} \geq \left(\frac{I_{v+k}(x)}{I_{v+k-1}(x)}\right)^k = \left(\frac{x}{(v+k)+\sqrt{x^2+(v+k)^2}}\right)^k. \tag{24}$$

We can now give the lower bound for $a_k^m(K_n^{\text{vMF}})$ using Equation (24) :

$$a_k^m(K_\lambda^{\text{vMF}}) = \frac{I_{\frac{m-1}{2}+k}(\lambda)}{I_{\frac{m-1}{2}}(\lambda)} \geq \left(\frac{\lambda}{\left(\frac{m-1}{2}+k\right)+\sqrt{\lambda^2+\left(\frac{m-1}{2}+k\right)^2}}\right)^k.$$

The lower bound for $m = 2$ is plotted in Figure 5. □

We can now provide a bound for the second term in Equation (21):

Lemma E.4. *Take an $f \in C(S^m)$. Furthermore, assume that there exists a constant $C_H \geq 0$ that upper-bounds the norms of the spherical harmonics of any best polynomial approximation W_q of f:*

$$\textit{for all } q \geq 1, W_q = \sum_{k=0}^{q} Y_k^m,\ Y_k^m \in \mathbb{Y}(S^m),\ \|Y_k^m\|_\infty \leq C_H,\ \forall k = 0,\dots,q,\ \textit{where } W_q = \arg\min_{h\in\mathbb{P}_q(S^m)} \|f-h\|_\infty.$$

Then

$$\|K_\lambda^{vMF} * W_q - W_q\|_\infty \leq C_H\, q\left(1-\left(\frac{\lambda}{\left(\frac{m-1}{2}+q\right)+\sqrt{\lambda^2+\left(\frac{m-1}{2}+q\right)^2}}\right)^q\right).$$

Proof. Using that W_q is a spherical polynomial of degree q and hence can be expressed as a sum of spherical harmonics $W_q = \sum_{k=0}^q Y_k^m$, we get:

$$\begin{aligned}
\|K_\lambda^{\text{vMF}} * W_q - W_q\|_\infty &= \left\|K_\lambda^{\text{vMF}} * \sum_{k=0}^{q} Y_k^m(x) - \sum_{k=0}^{q} Y_k^m(x)\right\|_\infty \\
&= \left\|\sum_{k=0}^{q}\left(K_\lambda^{\text{vMF}} * Y_k^m(x) - Y_k^m(x)\right)\right\|_\infty
\end{aligned}$$

$$
\begin{aligned}
&\leq \sum_{k=0}^{q} \left\| \left(K_\lambda^{\text{vMF}} * Y_k^m(x) - Y_k^m(x) \right) \right\|_\infty && \text{(Triangle inequality)} \\
&= \sum_{k=0}^{q} \left\| \left(a_k^m(K_\lambda^{\text{vMF}}) Y_k^m(x) - Y_k^m(x) \right) \right\|_\infty && \text{(from Lemma D.1)} \\
&= \sum_{k=0}^{q} \left\| \left(a_k^m(K_\lambda^{\text{vMF}}) - 1 \right) Y_k^m(x) \right\|_\infty \\
&= \sum_{k=0}^{q} \left| a_k^m(K_\lambda^{\text{vMF}}) - 1 \right| \left\| Y_k^m(x) \right\|_\infty \\
&\leq \sum_{k=0}^{q} \left| a_k^m(K_\lambda^{\text{vMF}}) - 1 \right| C_H \\
&\leq C_H \sum_{k=0}^{q} \left(1 - \left(\frac{\lambda}{\left(\frac{m-1}{2} + k\right) + \sqrt{\lambda^2 + \left(\frac{m-1}{2} + k\right)^2}} \right)^k \right) && \text{(from Lemma E.3)} \\
&\leq C_H \sum_{k=0}^{q} \left(1 - \underbrace{\left(\frac{\lambda}{\left(\frac{m-1}{2} + q\right) + \sqrt{\lambda^2 + \left(\frac{m-1}{2} + q\right)^2}} \right)}_{B}{}^k \right) && \text{(as } k \leq q\text{)} \\
&= C_H \left(q + 1 - \sum_{k=0}^{q} B^k \right) \\
&\leq C_H \left(q + 1 - (1 + qB^q) \right) && \text{(using that } 0 < B < 1\text{)} \\
&\leq C_H \, q \left(1 - B^q \right).
\end{aligned}
$$

□

We can finally combine Lemmas E.1, E.2 and E.4 in order to provide an upper bound to Equation (21):

Lemma E.5 (Bound on $\|K_\lambda^{\text{vMF}} * f - f\|_\infty$)**.** *Take an $f \in V_m$ with modulus of continuity $\omega(f;t) \leq Lt$. As in Lemma E.4, assume that there exists a constant $C_H \geq 0$ that upper-bounds the norms of the spherical harmonics of any best polynomial approximation W_q of f:*

$$
\textit{for all } q \geq 1, W_q = \sum_{k=0}^{q} Y_k^m,\ Y_k^m \in \mathbb{Y}(S^m),\ \|Y_k^m\|_\infty \leq C_H,\ \forall k = 0, \ldots, q,\ \textit{where } W_q = \arg \min_{h \in \mathbb{P}_q(S^m)} \|f - h\|_\infty .
$$

If $\lambda \geq \Lambda(\epsilon)$, where

$$
\Lambda(\epsilon) = \frac{(8LC_R + m\epsilon + \epsilon)\left(1 - \frac{\epsilon^2}{8LC_H C_R + 2\epsilon C_H}\right)^{\frac{\epsilon}{4LC_R + \epsilon}}}{\epsilon \left(1 - \left(1 - \frac{\epsilon^2}{8LC_H C_R + 2\epsilon C_H}\right)^{\frac{2\epsilon}{4LC_R + \epsilon}}\right)} = \mathcal{O}\left(\frac{L^3 C_H C_R^3}{\epsilon^4}\right). \tag{25}
$$

then $\|f - K_\lambda^{\text{vMF}} * f\|_\infty \leq \epsilon$.

Proof. As we want to upper-bound Equation (21) with ϵ, we will split our ϵ budget over the three terms.

For the first and the third terms, using the Ragozin bound from Lemma E.1 we have:

$$
\|f - W_q\|_\infty \leq C_R \, \omega\left(f; \frac{1}{q}\right) \leq \frac{C_R L}{q},\ q \geq 1. \tag{26}
$$

We want to select an integer q large enough so that $\|f - W_q\|_\infty \leq {}^{\epsilon}\!/_4$. That is $q = \left\lceil \frac{4C_R L}{\epsilon} \right\rceil$. This will be how we bound the first and last terms in Equation (21).

Let's focus on the second term. Pick W_q to be the best approximation from the Ragozin bound. From Lemma E.4 we have that

$$\|K_\lambda^{\text{vMF}} * W_q - W_q\|_\infty \leq C_H\, q \left(1 - B^q\right),$$

where

$$B = \frac{\lambda}{\left(\frac{m-1}{2} + q\right) + \sqrt{\lambda^2 + \left(\frac{m-1}{2} + q\right)^2}}.$$

The error budget we want to allocate for the $\|K_\lambda^{\text{vMF}} * W_q - W_q\|_\infty$ term is ${}^{\epsilon}\!/_2$. Hence:

$$B \geq \left(1 - \frac{\epsilon}{2C_H q}\right)^{1/q} = D \implies \|K_n^{\text{vMF}} * W_q - W_q\|_\infty \leq \frac{\epsilon}{2}. \tag{27}$$

We just need to find the minimum value for λ such that Equation (27) holds. We have:

$$B = \frac{\lambda}{\underbrace{\left(\frac{m-1}{2} + q\right)}_{E} + \sqrt{\lambda^2 + \left(\frac{m-1}{2} + q\right)^2}} = \frac{\lambda}{E + \sqrt{\lambda^2 + E^2}}. \tag{28}$$

Then, combining Equations (27) and (28) we get:

$$\begin{aligned} \frac{\lambda}{E + \sqrt{\lambda^2 + E^2}} &\geq D \\ \lambda &\geq \frac{2DE}{1 - D^2}. \end{aligned}$$

Finally, replacing D and E with the expressions in Equations (27) and (28), upper-bounding $q = \left\lceil \frac{4C_R L}{\epsilon} \right\rceil$ as $q \geq \frac{4C_R L}{\epsilon} + 1$ and simplifying the expression we get our final bound for λ. If $\lambda \geq \Lambda(\epsilon)$, with

$$\Lambda(\epsilon) = \frac{\left(8LC_R + m\epsilon + \epsilon\right)\left(1 - \frac{\epsilon^2}{8LC_H C_R + 2\epsilon C_H}\right)^{\frac{\epsilon}{4LC_R + \epsilon}}}{\epsilon\left(1 - \left(1 - \frac{\epsilon^2}{8LC_H C_R + 2\epsilon C_H}\right)^{\frac{2\epsilon}{4LC_R + \epsilon}}\right)},$$

then $\|K_\lambda^{\text{vMF}} * W_q - W_q\|_\infty \leq {}^{\epsilon}\!/_2$.

Hence, for any $\lambda \geq \Lambda(\epsilon)$ we have:

$$\begin{aligned} \|K_\lambda^{\text{vMF}} * f - f\|_\infty &\leq \|K_\lambda^{\text{vMF}} * f - K_\lambda^{\text{vMF}} * W_q\|_\infty + \|K_\lambda^{\text{vMF}} * W_q - W_q\|_\infty + \|W_q - f\|_\infty \\ &\leq \|f - W_q\|_\infty + \frac{\epsilon}{2} + \|W_q - f\|_\infty \\ &\leq \frac{\epsilon}{4} + \frac{\epsilon}{2} + \frac{\epsilon}{4} \\ &= \epsilon. \end{aligned}$$

This concludes our bound on Equation (21).

Finally, to give the asymptotic behavior of $\Lambda(\epsilon)$ as $\epsilon \to 0$ we observe that the Taylor series expansion of Λ around $\epsilon = 0$ is:

$$\Lambda(\epsilon) = \frac{128L^3 C_H C_R^3}{\epsilon^4} + \frac{16L^2(m-1)C_H C_R^2}{\epsilon^3} + \mathcal{O}\left(\frac{1}{\epsilon^2}\right),$$

hence:

$$\Lambda(\epsilon) = \mathcal{O}\left(\frac{L^3 C_H C_R^3}{\epsilon^4}\right).$$

□

Lemma E.5 is our bound on Equation (21) which is also the first term of Equation (20). Recall that bounding Equation (20) is our ultimate goal. Hence, we are halfway done with our proof. Let's focus now on the second term in Equation (20), that is, how well we can approximate the convolution of f with the von Mises-Fisher kernel using a finite sum:

$$\left\| f * K_\lambda^{\text{vMF}} - \sum_{k=1}^{N} \xi_k \exp(\lambda \langle \cdot, \boldsymbol{b}_k \rangle) \right\|_\infty .$$

The basic idea behind bounding this term is that we can partition the hypersphere S^m into N sets ($\{V_1, \ldots, V_N\}$), each small enough so that, for a fixed $\boldsymbol{x} \in S^m$, the $K(\langle \boldsymbol{x}, \boldsymbol{y} \rangle) f(\boldsymbol{y})$ term in the convolution

$$(K * f)(\boldsymbol{x}) = \frac{1}{w_m} \int_{S^m} K(\langle \boldsymbol{x}, \boldsymbol{y} \rangle) f(\boldsymbol{y}) \, dw_m(\boldsymbol{y})$$

is almost the same for all values $\boldsymbol{y} \in V_k$ in that element of the partition. Hence we can approximate the integral over the partition with estimate over a single point $\boldsymbol{b}_k$:

$$\int_{V_k} K(\langle \boldsymbol{x}, \boldsymbol{y} \rangle) f(\boldsymbol{y}) \, dw_m(\boldsymbol{y}) \approx |V_k| K(\langle \boldsymbol{x}, \boldsymbol{b}_k \rangle) f(\boldsymbol{b}_k).$$

The rest of this section will make this formal.

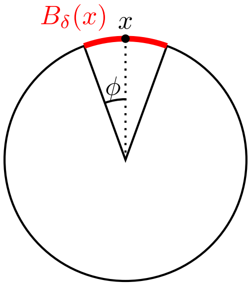
First, in order to construct our partition $\{V_1, \ldots, V_N\}$ of S^m we will first construct a cover of S^m. Then, our partition will be such that each element V_k is a subset of the corresponding element of the cover of S^m. In this way, we can control the maximum size of the elements of the cover.

Lemma E.6. *Consider a cover $\{B_\delta(\boldsymbol{b}_1), \ldots, B_\delta(\boldsymbol{b}_{N_\delta^m})\}$ of S^m by N_δ^m hyperspherical caps $B_\delta(\boldsymbol{x}) = \{\boldsymbol{y} \in S^m \mid \langle \boldsymbol{x}, \boldsymbol{y} \rangle \geq 1 - \delta\}$ for $0 < \delta < 1$. By cover we mean that $\bigcup_{i=1}^{N_\delta^m} B_\delta(\boldsymbol{b}_i) = S^m$, with N_δ^m being the smallest number of hyperspherical caps to cover S^m (its covering number). Then, for $m \geq 8$ we have:*

$$\frac{2}{I_{(\delta(2-\delta))}\left(\frac{m}{2}, \frac{1}{2}\right)} \leq N_\delta^m < \frac{\Phi(m)}{(\delta(2-\delta))^{\frac{m+1}{2}}} < \frac{\Phi(m)}{\delta^{m+1}},$$

with I being the regularized incomplete beta function and $\Phi(m) = \mathcal{O}(m \log m)$ being a function that depends only on the dimension m.

Proof. Define $\phi = \cos^{-1}(1 - \delta)$:



Naturally, the area of the caps needs to be at least as much as the area of the hypersphere for the set of caps to be a cover. This gives us our lower bound. The area of a cap with colatitude angle ϕ as above is (Li, 2010):

$$w_m^\phi = \frac{1}{2} w_m I_{\sin^2 \phi} \left(\frac{m}{2}, \frac{1}{2} \right).$$

As $\sin \phi = \sin\left(\cos^{-1}(1 - \delta)\right) = \sqrt{1 - (1 - \delta)^2} = \sqrt{\delta(2 - \delta)}$, we have our lower bound:

$$N_\delta^m \geq \frac{w_m}{w_m^\phi} = \frac{2}{I_{\sin^2 \phi}\left(\frac{m}{2}, \frac{1}{2}\right)} = \frac{2}{I_{(\delta(2-\delta))}\left(\frac{m}{2}, \frac{1}{2}\right)}$$

For the upper bound, we can use the observation that if a unit *ball* is covered with *balls* of radius r, then the unit *sphere* is also covered with *caps* of radius r. From (Rogers, 1963, intermediate result from the proof of theorem 3) we have that for $m \geq 8$ and $1/r \geq m+1$, a unit ball can be covered by less than

$$e\left((m+1)\log(m+1) + (m+1)\log\log(m+1) + 5(m+1)\right)\frac{1}{r^{m+1}} = \Phi(m)\frac{1}{r^{m+1}}$$

balls of radius r. For high m, this is a pretty good approximation since most of the volume of the hypersphere lies near its surface. Our caps B_δ can fit inside balls of radius $r = \sin\phi$. Hence, we have the upper bound:

$$N_\delta^m < \frac{\Phi(m)}{\sin^{m+1}\phi} = \frac{\Phi(m)}{(\delta(2-\delta))^{\frac{m+1}{2}}}.$$

□

Now we can use a partition resulting from this covering in order to bound the error between the integral and its Riemannian sum approximation:

Lemma E.7 (Approximation via Riemann sums)**.** *Let $g(x,y) : S^m \times S^m \to \mathbb{R}$ with $m \geq 8$ be a continuous function with modulus of continuity for both arguments $\omega(f(\cdot;\boldsymbol{y}),t) \leq Lt,\ \forall \boldsymbol{y} \in S^m$ and $\omega(f(\boldsymbol{x},\cdot);t) \leq Lt,\ \forall \boldsymbol{x} \in S^m$. Take any $0 < \delta < 1$. Then, there exists a partition $\{V_1,\dots,V_{N_\delta^m}\}$ of S^m into $N_\delta^m = \lceil \Phi(m)/\delta^{m+1} \rceil$ subsets, as well as $\boldsymbol{b}_1,\dots,\boldsymbol{b}_{N_\delta^m} \in S^m$ such that:*

$$\max_{\boldsymbol{x}\in S^m} \frac{1}{w_m}\left|\int_{S^m} g(\boldsymbol{x},\boldsymbol{y})\, dw_m(\boldsymbol{y}) - \sum_{k=1}^{N_\delta^m} g(\boldsymbol{x},\boldsymbol{b}_k)\, w_m(V_k)\right| \leq 3L\cos^{-1}(1-\delta).$$

Here, $\Phi(m) = \mathcal{O}(m\log m)$ is a function that depends only on the dimension m.

Proof. This proof is a non-asymptotic version of the proof of Lemma 4.3 from Ng and Kwong (2022). First, we can use Lemma E.6 to construct a covering $\{B_\delta(\boldsymbol{b}_1),\dots,B_\delta(\boldsymbol{b}_{N_\delta^m})\}$ of S^m. If we have a covering of S^m it is trivial to construct a partition of it $\{V_1,\dots,V_{N_\delta^m}\}$, $\bigcup_{k=1}^{N_\delta^m} V_k = S^m$, $V_i \cap V_j = \emptyset, i \neq j$ such that $V_k \subseteq B_\delta(\boldsymbol{b}_k), \forall k$. This partition can also be selected to be such that all elements of it have the same measure $w_m(V_1) = w_m(V_i), \forall i$ (Feige and Schechtman, 2002, Lemma 21). While this is not necessary for this proof, we will use this equal measure partition in Lemma F.2 and Theorem F.1.

We can then use the triangle inequality to split the term we want to bound in three separate terms:

$$\begin{aligned}\left|\int_{S^m} g(\boldsymbol{x},\boldsymbol{y})\, dw_m(\boldsymbol{y}) - \sum_{k=1}^{N_\delta^m} g(\boldsymbol{x},\boldsymbol{b}_k)\, w_m(V_k)\right| &\leq \left|\int_{S^m} g(\boldsymbol{x},\boldsymbol{y})\, dw_m(\boldsymbol{y}) - \int_{S^m} g(\boldsymbol{b}_\star,\boldsymbol{y})\, dw_m(\boldsymbol{y})\right| \\ &+ \left|\int_{S^m} g(\boldsymbol{b}_\star,\boldsymbol{y})\, dw_m(\boldsymbol{y}) - \sum_{k=1}^{N_\delta^m} g(\boldsymbol{b}_\star,\boldsymbol{b}_k)\, w_m(V_k)\right| \\ &+ \left|\sum_{k=1}^{N_\delta^m} g(\boldsymbol{b}_\star,\boldsymbol{b}_k)\, w_m(V_k) - \sum_{k=1}^{N_\delta^m} g(\boldsymbol{x},\boldsymbol{b}_k)\, w_m(V_k)\right|,\end{aligned} \tag{29}$$

where $\boldsymbol{b}_\star$ is the center of one of the caps whose corresponding partition contains $\boldsymbol{x}$, i.e., $\boldsymbol{b}_\star = \boldsymbol{b}_i \iff \boldsymbol{x} \in V_i$. Due to $\{V_1,\dots,V_{N_\delta^m}\}$ being a partition, $\boldsymbol{b}_\star$ is well-defined as $\boldsymbol{x}$ is in exactly one of the elements of the partition.

Observe also that the modulus of continuity gives us a Lipschitz-like bound, i.e., if $\langle \boldsymbol{x},\boldsymbol{y}\rangle \geq 1-\delta$ for $\boldsymbol{x},\boldsymbol{y} \in S^m$ and $\omega(f;t) \leq Lt$, then

$$|f(\boldsymbol{x}) - f(\boldsymbol{y})| \leq \omega(f;\cos^{-1}(\langle \boldsymbol{x},\boldsymbol{y}\rangle)) \leq \omega(f;\cos^{-1}(1-\delta)) \leq L\cos^{-1}(1-\delta). \tag{30}$$

Let's start with the first term in Equation (29). Using the fact that we selected $\boldsymbol{b}_\star$ to be such that $\langle \boldsymbol{b}_\star, \boldsymbol{x} \rangle \geq 1 - \delta$ and Equation (30), we have:

$$\begin{aligned}
\left| \int_{S^m} g(\boldsymbol{x}, \boldsymbol{y}) \, dw_m(\boldsymbol{y}) - \int_{S^m} g(\boldsymbol{b}_\star, \boldsymbol{y}) \, dw_m(\boldsymbol{y}) \right| &= \left| \int_{S^m} (g(\boldsymbol{x}, \boldsymbol{y}) - g(\boldsymbol{b}_\star, \boldsymbol{y})) \, dw_m(\boldsymbol{y}) \right| \\
&\leq \int_{S^m} |g(\boldsymbol{x}, \boldsymbol{y}) - g(\boldsymbol{b}_\star, \boldsymbol{y})| \, dw_m(\boldsymbol{y}) \\
&\leq \int_{S^m} L \cos^{-1}(1 - \delta) \, dw_m(\boldsymbol{y}) \\
&= L \cos^{-1}(1 - \delta) \int_{S^m} dw_m(\boldsymbol{y}) \\
&= L \cos^{-1}(1 - \delta) \, w_m.
\end{aligned}$$

We can similarly upper-bound the second term of Equation (29) using also the fact that $\{V_k\}$ is a partition of S^m:

$$\begin{aligned}
\left| \int_{S^m} g(\boldsymbol{b}_\star, \boldsymbol{y}) \, dw_m(\boldsymbol{y}) - \sum_{k=1}^{N_\delta^m} g(\boldsymbol{b}_\star, \boldsymbol{b}_k) \, w_m(V_k) \right| &= \left| \sum_{k=1}^{N_\delta^m} \int_{V_k} g(\boldsymbol{b}_\star, \boldsymbol{y}) \, dm_w(\boldsymbol{y}) - \sum_{k=1}^{N_\delta^m} g(\boldsymbol{b}_\star, \boldsymbol{b}_k) \, w_m(V_k) \right| \\
&= \left| \sum_{k=1}^{N_\delta^m} \int_{V_k} (g(\boldsymbol{b}_\star, \boldsymbol{y}) - g(\boldsymbol{b}_\star, \boldsymbol{b}_k)) \, dm_w(\boldsymbol{y}) \right| \\
&\leq \sum_{k=1}^{N_\delta^m} \int_{V_k} |g(\boldsymbol{b}_\star, \boldsymbol{y}) - g(\boldsymbol{b}_\star, \boldsymbol{b}_k)| \, dm_w(\boldsymbol{y}) \\
&\leq \sum_{k=1}^{N_\delta^m} \int_{V_k} L \cos^{-1}(1 - \delta) \, dm_w(\boldsymbol{y}) \\
&= L \cos^{-1}(1 - \delta) \sum_{k=1}^{N_\delta^m} \int_{V_k} dm_w(\boldsymbol{y}) \\
&= L \cos^{-1}(1 - \delta) w_m.
\end{aligned}$$

And analogously, for the third term we get:

$$\begin{aligned}
\left| \sum_{k=1}^{N_\delta^m} g(\boldsymbol{b}_\star, \boldsymbol{b}_k) \, w_m(V_k) - \sum_{k=1}^{N_\delta^m} g(\boldsymbol{x}, \boldsymbol{b}_k) \, w_m(V_k) \right| &= \left| \sum_{k=1}^{N_\delta^m} (g(\boldsymbol{b}_\star, \boldsymbol{b}_k) - g(\boldsymbol{x}, \boldsymbol{b}_k)) \, w_m(V_k) \right| \\
&\leq \sum_{k=1}^{N_\delta^m} |g(\boldsymbol{b}_\star, \boldsymbol{b}_k) - g(\boldsymbol{x}, \boldsymbol{b}_k)| \, w_m(V_k) \\
&\leq L \cos^{-1}(1 - \delta) \sum_{k=1}^{N_\delta^m} w_m(V_k) \\
&= L \cos^{-1}(1 - \delta) w_m.
\end{aligned}$$

Finally, observing that the above bounds do not depend on the choice of $\boldsymbol{x} \in S^m$ and combining the three results we obtain our desired bound. □

By observing that we can set $g(\boldsymbol{x}, \boldsymbol{y}) = K_\lambda^{\text{vMF}}(\langle \boldsymbol{x}, \boldsymbol{y} \rangle) f(\boldsymbol{y})$, it becomes clear how Lemma E.7 can be used to bound the second term in Equation (20). For that we will also need to know what is the modulus of continuity of the von Mises-Fisher kernels K_λ^{vMF}.

Lemma E.8 (Modulus of continuity of K_λ^{vMF})**.** *The von Mises-Fisher kernels K_λ^{vMF} have modulus of continuity $\omega(K_\lambda^{vMF}; t) \leq c_{m+1}(\lambda) \exp(\lambda)$.*

Proof. Recall that $K_\lambda^{\text{vMF}}(t)$ is defined on $t \in [-1, 1]$. $K_\lambda^{\text{vMF}}(t)$ and its derivative are both monotonically increasing in t. Hence:

$$\begin{aligned}
\omega(K_\lambda^{\text{vMF}}; t) &= \sup\left\{ |K_\lambda^{\text{vMF}}(\langle \boldsymbol{z}, \boldsymbol{x}\rangle) - K_\lambda^{\text{vMF}}(\langle \boldsymbol{z}, \boldsymbol{y}\rangle) \mid \boldsymbol{x}, \boldsymbol{y} \in S^m, \cos^{-1}(\langle \boldsymbol{x}, \boldsymbol{y}\rangle) \le t \right\} \\
&= K_\lambda^{\text{vMF}}(1) - K_\lambda^{\text{vMF}}(\cos t) \\
&= c_{m+1}(\lambda)\left(\exp(\lambda) - \exp(\lambda \cos t)\right) \\
&\le c_{m+1}(\lambda) \exp(\lambda).
\end{aligned}$$

□

Our final result, a bound on Equation (20), combines Lemma E.5 and Lemma E.7, each bounding one of the two terms in Equation (20).

Theorem E.1 (Jackson-type bound for universal approximation on the hypersphere, Theorem B.1 in the main text)**.** *Let $f \in C(S^m)$ be a continuous function on S^m with modulus of continuity $\omega(f;t) \le Lt$ for some $L \in \mathbb{R}_{>0}$ and $m \ge 8$. Assume that there exists a constant $C_H \ge 0$ that upper-bounds the norms of the spherical harmonics of any best polynomial approximation W_q of f:*

$$\textit{for all } q \ge 1, W_q = \sum_{k=0}^{q} Y_k^m,\ Y_k^m \in \mathbb{Y}(S^m),\ \|Y_k^m\|_\infty \le C_H,\ \forall k = 0, \ldots, q,\ \textit{where } W_q = \arg\min_{h \in \mathbb{P}_q(S^m)} \|f - h\|_\infty.$$

Then, for any $\epsilon > 0$, there exist $\xi_1, \ldots, \xi_N \in \mathbb{R}$ and $\boldsymbol{b}_1, \ldots, \boldsymbol{b}_N \in S^m$ such that

$$\sup_{\boldsymbol{x} \in S^m} \left| f(\boldsymbol{x}) - \sum_{k=1}^{N} \xi_k \exp(\lambda \langle \boldsymbol{x}, \boldsymbol{b}_k\rangle) \right| \le \epsilon,$$

where $\lambda = \Lambda(\epsilon/2)$ (Equation (25)) and for any N such that

$$N \ge N(\lambda, \epsilon) = \Phi(m)\left(\frac{3\pi(L + \|f\|_\infty) c_{m+1}(\lambda)\exp(\lambda)}{\epsilon}\right)^{m+1} = \mathcal{O}(\epsilon^{-1-3m-2m^2}). \tag{31}$$

Proof. Recall the decomposition in Equation (20). We will split our error budget ϵ in half. Hence, we first select λ such that approximating f with its convolution with K_λ^{vMF} results in an error at most $\lambda/2$. Then, using this λ, we find how finely we need to partition S^m in order to be able to approximate the convolution with a sum up to an error $\epsilon/2$.

Let's select how "peaky" we need the kernel K_λ^{vMF} to be, that is, how big should λ be. From Lemma E.5 we have that if $\lambda = \Lambda(\epsilon/2)$, then we would have $\|f - f * K_\lambda^{\text{vMF}}\|_\infty \le \epsilon/2$.

Now, for the second term in Equation (20), consider Lemma E.7 with $g(\boldsymbol{x}, \boldsymbol{y}) = K_\lambda^{\text{vMF}}(\langle \boldsymbol{x}, \boldsymbol{y}\rangle) f(\boldsymbol{y})$. From Lemma E.8 we have that the modulus of continuity of K_λ^{vMF} is $\omega(K_\lambda^{\text{vMF}}; t) \le t\, c_{m+1}(\lambda)\exp(\lambda)$. Hence, we have modulus of continuity for $g(\boldsymbol{x}, \boldsymbol{y})$ being bounded as:

$$\begin{aligned}
\omega(g;t) &\le \|K_\lambda^{\text{vMF}}\|_\infty \omega(f;t) + \|f\|_\infty \omega(K_\lambda^{\text{vMF}}; t) \\
&\le K_\lambda^{\text{vMF}}(1)\ Lt + \|f\|_\infty c_{m+1}(\lambda)\exp(\lambda)\ t \\
&= c_{m+1}(\lambda)\exp(\lambda)\ Lt + \|f\|_\infty c_{m+1}(\lambda)\exp(\lambda)\ t \\
&= c_{m+1}(\lambda)\exp(\lambda)\,(L + \|f\|_\infty)\, t.
\end{aligned}$$

Take

$$\delta = \left(\frac{2\epsilon}{6\pi(L + \|f\|_\infty) c_{m+1}(\lambda)\exp(\lambda)}\right)^2.$$

Then, by Lemma E.7, there exists a partition $\{V_1, \ldots, V_N\}$ of S^m and $\boldsymbol{b}_1, \ldots, \boldsymbol{b}_N \in S^m$ for N as in the lemma such that:

$$\max_{\boldsymbol{x} \in S^m} \left| \frac{1}{w_m}\int_{S^m} K_\lambda^{\text{vMF}}(\langle \boldsymbol{x}, \boldsymbol{y}\rangle) f(\boldsymbol{y})\, dw_m(\boldsymbol{y}) - \frac{1}{w_m}\sum_{k=1}^{N} K_\lambda^{\text{vMF}}(\langle \boldsymbol{x}, \boldsymbol{b}_k\rangle) f(\boldsymbol{b}_k)\, w_m(V_k) \right| \le 3(L + \|f\|_\infty) c_{m+1} \exp(\lambda)\cos^{-1}(1 - \delta).$$

As $(2x/\pi)^2 < 1 - \cos(x)$ (Bagul and Panchal, 2018, Theorem 1), we have:

$$\delta = \left(\frac{2\epsilon}{6\pi(L + \|f\|_\infty) c_{m+1}(\lambda)\exp(\lambda)}\right)^2 < 1 - \cos\left(\frac{\epsilon}{6(L + \|f\|_\infty) c_{m+1}(\lambda)\exp(\lambda)}\right). \tag{32}$$

Hence:

$$\begin{aligned}
&\max_{\boldsymbol{x}\in S^m} \left| \frac{1}{w_m} \int_{S^m} K_\lambda^{\text{vMF}}(\langle \boldsymbol{x}, \boldsymbol{y}\rangle) f(\boldsymbol{y}) \, dw_m(\boldsymbol{y}) - \frac{1}{w_m} \sum_{k=1}^{N} K_\lambda^{\text{vMF}}(\langle \boldsymbol{x}, \boldsymbol{b}_k\rangle) f(\boldsymbol{b}_k) \, w_m(V_k) \right| \\
&\le 3(L + \|f\|_\infty) c_{m+1} \exp(\lambda) \cos^{-1}(1-\delta) \\
&< 3(L + \|f\|_\infty) c_{m+1} \exp(\lambda) \cos^{-1}\left(\cos \left(\frac{\epsilon}{6(L + \|f\|_\infty) c_{m+1}(\lambda) \exp(\lambda)} \right)\right) \\
&= {}^{\epsilon}\!/_{2}.
\end{aligned}$$

Combining the two results we have:

$$\begin{aligned}
\left\| f - \frac{1}{w_m} \sum_{k=1}^{N} K_\lambda^{\text{vMF}}(\langle \boldsymbol{x}, \boldsymbol{b}_k\rangle) f(\boldsymbol{y}) \, w_m(V_k) \right\|_\infty &\le \|f - f * K_\lambda^{\text{vMF}}\|_\infty + \left\| f * K_\lambda^{\text{vMF}} - \frac{1}{w_m} \sum_{k=1}^{N} K_\lambda^{\text{vMF}}(\langle \boldsymbol{x}, \boldsymbol{b}_k\rangle) f(\boldsymbol{y}) \, w_m(V_k) \right\|_\infty \\
&\le {}^{\epsilon}\!/_{2} + {}^{\epsilon}\!/_{2} \\
&= \epsilon.
\end{aligned}$$

Now, the only thing left is to show that this expression can be expressed in the form of Equation (19).

$$\begin{aligned}
\frac{1}{w_m} \sum_{k=1}^{N_\delta^m} K_\lambda^{\text{vMF}}(\langle \boldsymbol{x}, \boldsymbol{b}_k\rangle) \, f(\boldsymbol{b}_k) \, w_m(V_k) &= \sum_{k=1}^{N_\delta^m} \frac{1}{w_m} c_{m+1}(\lambda) \, \exp(\lambda \langle \boldsymbol{x}, \boldsymbol{b}_k\rangle) \, f(\boldsymbol{b}_k) \, w_m(V_k) \\
&= \sum_{k=1}^{N_\delta^m} \xi_k \exp(\lambda \langle \boldsymbol{x}, \boldsymbol{b}_k\rangle),
\end{aligned}$$

with

$$\xi_k = c_{m+1}(\lambda) f(\boldsymbol{b}_k) \frac{w_m(V_k)}{w_m}. \tag{33}$$

If we have chosen a partition of equal measure this further simplifies to

$$\xi_k = \frac{c_{m+1}(\lambda)}{N} f(\boldsymbol{b}_k).$$

Hence, for this choice of Λ, N and $\boldsymbol{b}_k$ and ξ_k constructed as above, we indeed have

$$\sup_{\boldsymbol{x}\in S^m} \left| f(\boldsymbol{x}) - \sum_{k=1}^{N} \xi_k \exp(\lambda \langle \boldsymbol{x}, \boldsymbol{b}_k\rangle) \right| \le \epsilon.$$

Finally, let's study the asymptotic growth of N as $\epsilon \to 0$. We have:

$$N(\lambda, \epsilon) = \Phi(m) \left(\frac{3\pi (L + \|f\|_\infty) c_{m+1}(\lambda) \exp(\lambda)}{\epsilon} \right)^{m+1}.$$

$\Phi(m)$ is constant in ϵ so we can ignore it. Expanding c_{m+1} and dropping the terms that do not depend on ϵ gives us:

$$\mathcal{O}\left(\frac{\lambda^{\frac{m+1}{2}-1} \exp(\lambda)}{\epsilon \, I_{\frac{m+1}{2}-1}(\lambda)} \right)^{m+1}. \tag{34}$$

The asymptotics of the modified Bessel function of the first kind are difficult to analyse. However, as we care about an upper bound, we can simplify the expression by lower-bounding $I_\nu(\lambda)$ using Equation (24) and that $I_0(\lambda) \ge {}^{C\exp(\lambda)}\!/_{\sqrt{\lambda}}$ for $\lambda > {}^{1}\!/_{2}$ (Barnett, 2021):

$$I_\nu(\lambda) \ge C \left(\frac{\sqrt{\nu^2 + \lambda^2} - \nu}{\lambda} \right)^{\nu+1} \frac{\exp\left(\sqrt{\nu^2 + \lambda^2}\right)}{\sqrt{\lambda}},$$

for some constant C. Plugging this in Equation (34) and replacing λ with its asymptotic growth ϵ^{-4} gives us:

$$\mathcal{O}\left(\left(\frac{2^{\frac{m+1}{2}} e^{\frac{1}{\epsilon^4} - \frac{1}{2}\sqrt{(m-1)^2 + \frac{4}{\epsilon^8}}} \left(\frac{1}{\epsilon^4}\right)^{m/2} \left(\epsilon^4 \left(\sqrt{(m-1)^2 + \frac{4}{\epsilon^8}} - m + 1 \right)\right)^{-\frac{1}{2}(m+1)}}{\epsilon} \right)^{m+1} \right)$$

$$=\mathcal{O}(\epsilon^{-1-3m-2m^2}),$$

with the last simplification coming from taking the Taylor series expansion at for $\epsilon \to 0$. □

We can easily extend Theorem E.1 to vector-valued functions:

Corollary E.1 (Corollary B.2 in the main text)**.** *Let $f : S^m \to \mathbb{R}^{m+1}$, $m \geq 8$ be such that each component f_i is in $C(S^m), i = 1, \dots, m+1$ and satisfies the conditions in Theorem E.1. Furthermore, define $\|f\|_\infty = \max_{1 \leq i \leq m+1} \|f_i\|_\infty$. Then, for any $\epsilon > 0$, there exist $\boldsymbol{\xi}_1, \dots, \boldsymbol{\xi}_N \in \mathbb{R}^{m+1}$ and $\boldsymbol{b}_1, \dots, \boldsymbol{b}_N \in S^m$ such that*

$$\sup_{\boldsymbol{x} \in S^m} \left\| f(\boldsymbol{x}) - \sum_{k=1}^{N} \boldsymbol{\xi}_k \exp(\lambda \langle \boldsymbol{x}, \boldsymbol{b}_k \rangle) \right\|_2 \leq \epsilon,$$

with $\lambda = \Lambda(\epsilon/2\sqrt{m+1})$ for any $N \geq N(\lambda, \epsilon/\sqrt{m+1})$.

Proof. The proof is the same as for Theorem E.1. As the concentration parameter λ of the kernels K_λ^{vMF} depends only on the smoothness properties of the individual components and these are assumed to be the same, the same kernel choice can be used for all components f_i. Furthermore, the choice of partition is independent of the function to be approximated and depends only on the concentration parameter of the kernel. Hence, we can also use the same partition for all components f_i. We only need to take into account that:

$$\|x - y\|_2 = \sqrt{\sum_{i=1}^{m+1} |x_i - y_i|^2} \leq \sqrt{(m+1)\epsilon^2} = \sqrt{m+1}\epsilon, \; \forall x, y \in \mathbb{R}^{m+1}, |x_i - y_i| \leq \epsilon, i = 1, \dots, m+1.$$

which results in the factor of $\sqrt{m+1}$. □

F A Jackson-type Bound for Approximation with a Split Attention Head

Lemma F.1. *Let $a, b : \mathbb{R}^d \to \mathbb{R}$, $c, d : \mathbb{R} \to \mathbb{R}$, $c(x), d(x) \neq 0, \forall x \in \mathbb{R}$ and $\epsilon_1, \epsilon_2 \geq 0$ be such that:*

$$\sup_{\boldsymbol{y} \in \mathbb{R}^d} \|a(\boldsymbol{y}) - b(\boldsymbol{y})\|_2 \leq \epsilon_1$$
$$\sup_{x \in \mathbb{R}} |c(x) - d(x)| = |c - d|_\infty \leq \epsilon_2.$$

Then for all $x \in \mathbb{R}$ and $\boldsymbol{y} \in \mathbb{R}^d$:

$$\left\| \frac{a(\boldsymbol{y})}{c(x)} - \frac{b(\boldsymbol{y})}{d(x)} \right\|_2 \leq \frac{\epsilon_1 |c|_\infty + \epsilon_2 \sup_{\boldsymbol{y} \in \mathbb{R}^d} \|a(\boldsymbol{y})\|_2}{|c(x)\, d(x)|}.$$

Proof. For a fixed $x \in \mathbb{R}$ and $\boldsymbol{y} \in \mathbb{R}^d$, using the triangle inequality gives us

$$\begin{aligned}
\left\| \frac{a(\boldsymbol{y})}{c(x)} - \frac{b(\boldsymbol{y})}{d(x)} \right\|_2 &= \left\| \frac{a(\boldsymbol{y})\, d(x) - b(\boldsymbol{y})\, c(x)}{c(x)\, d(x)} \right\|_2 \\
&= \frac{\|a(\boldsymbol{y})\, d(x) - b(\boldsymbol{y})\, c(x)\|_2}{|c(x)\, d(x)|} \\
&= \frac{\|a(\boldsymbol{y})\, d(x) - a(\boldsymbol{y})\, c(x) + a(\boldsymbol{y})\, c(x) - b(\boldsymbol{y})\, c(x)\|_2}{|c(x)\, d(x)|} \\
&\leq \frac{\|a(\boldsymbol{y})\, (d(x) - c(x))\|_2 + \|c(x)\, (a(\boldsymbol{y}) - b(\boldsymbol{y}))\|_2}{|c(x)\, d(x)|} \\
&\leq \frac{\epsilon_2 \|a(\boldsymbol{y})\|_2 + \epsilon_1 |c(x)|}{|c(x)\, d(x)|} \\
&\leq \frac{\epsilon_1 |c|_\infty + \epsilon_2 \sup_{\boldsymbol{y} \in \mathbb{R}^d} \|a(\boldsymbol{y})\|_2}{|c(x)\, d(x)|}.
\end{aligned}$$

And as this holds for all $x \in \mathbb{R}$ and $\boldsymbol{y} \in \mathbb{R}^d$, the inequality in the lemma follows. □

Lemma F.2. *Let $f : S^m \to \mathbb{R}^{m+1}$, $m \geq 8$ satisfy the requirements in Corollary E.1. Then, given an $\epsilon > 0$ and taking λ, N, and $\boldsymbol{b}_1, \ldots, \boldsymbol{b}_N \in S^m$ as prescribed by the lemma, we have that $\sum_{k=1}^N \exp(\lambda\langle \boldsymbol{x}, \boldsymbol{b}_k\rangle)$ is close to being a constant:*

$$\sup_{\boldsymbol{x} \in S^m} \left| 1 - \frac{c_{m+1}(\lambda)}{N} \sum_{k=1}^N \exp(\lambda\langle \boldsymbol{x}, \boldsymbol{b}_k\rangle) \right| \leq \frac{\epsilon}{2\sqrt{m+1}(L + \|f\|_\infty)}. \tag{35}$$

Proof. We can use Lemma E.7 by taking $g(\boldsymbol{x}, \boldsymbol{y}) = K_\lambda^{\text{vMF}}(\langle \boldsymbol{x}, \boldsymbol{y}\rangle)$. From Lemma E.8 we have that the modulus of continuity of K_λ^{vMF} is $\omega(K_\lambda^{\text{vMF}}; t) \leq t \; c_{m+1}(\lambda)\exp(\lambda)$. Observe that using Equation (18) we have

$$\int_{S^m} g(x,y) dw_m(y) = \int_{S^m} K_\lambda^{\text{vMF}}(\langle x, y\rangle) dw_m(y) = w_{m-1} \int_{-1}^{1} K_\lambda^{\text{vMF}}(t)(1-t^2)^{(m-2)/2} dt = w_m.$$

The value for δ has to be selected as in Corollary E.1 (Equation (32)):

$$\delta = \left(\frac{2\epsilon}{6\pi\sqrt{m+1}(L + \|f\|_\infty) c_{m+1}(\lambda)\exp(\lambda)} \right)^2 \; < \; 1 - \cos\left(\frac{\epsilon}{6\sqrt{m+1}(L + \|f\|_\infty) c_{m+1}(\lambda)\exp(\lambda)} \right).$$

Now, using the same partition from Lemma E.7, and recalling that we constructed it such that each element of the partition has the same measure $w_m(V_1) = w_m(V_i), \forall i$, we have:

$$\begin{aligned}
&\max_{\boldsymbol{x} \in S^m} \left| \frac{1}{w_m} \int_{S^m} g(\boldsymbol{x}, \boldsymbol{y}) dw_m(\boldsymbol{y}) - \frac{1}{w_m} \sum_{k=1}^N K_\lambda^{\text{vMF}}(\langle \boldsymbol{x}, \boldsymbol{b}_k\rangle) w_m(V_k) \right| \\
&= \max_{\boldsymbol{x} \in S^m} \left| 1 - \frac{1}{w_m} \sum_{k=1}^N c_{m+1}(\lambda)\exp(\lambda\langle \boldsymbol{x}, \boldsymbol{b}_k\rangle) w_m(V_k) \right| \\
&= \max_{\boldsymbol{x} \in S^m} \left| 1 - \frac{c_{m+1}(\lambda)}{N} \sum_{k=1}^N \exp(\lambda\langle \boldsymbol{x}, \boldsymbol{b}_k\rangle) \right| \\
&\leq 3c_{m+1}\exp(\lambda)\cos^{-1}(1-\delta) \\
&< 3c_{m+1}\exp(\lambda)\cos^{-1}\left(\cos\left(\frac{\epsilon}{6\sqrt{m+1}(L + \|f\|_\infty) c_{m+1}(\lambda)\exp(\lambda)} \right) \right) \\
&= \frac{3c_{m+1}\exp(\lambda)\epsilon}{6\sqrt{m+1}(L + \|f\|_\infty) c_{m+1}(\lambda)\exp(\lambda)} \\
&= \frac{\epsilon}{2\sqrt{m+1}(L + \|f\|_\infty)}.
\end{aligned}$$

□

Theorem F.1. *Let $f : S^m \to \mathbb{R}^{m+1}$, $m \geq 8$ satisfies the conditions in Corollary E.1. Define $\|f\|_\infty = \max_{1 \leq i \leq m+1} \|f_i\|_\infty$. For any $0 < \epsilon \leq 2\|f\|_\infty$, there exist $\boldsymbol{b}_1, \ldots, \boldsymbol{b}_N \in S^m$ such that f can be uniformly approximated to an error at most ϵ:*

$$\sup_{\boldsymbol{x} \in S^m} \left\| f(\boldsymbol{x}) - \frac{\sum_{k=1}^N \boldsymbol{\xi}_k \exp(\lambda\langle \boldsymbol{x}, \boldsymbol{b}_k\rangle)}{\sum_{k=1}^N \exp(\lambda\langle \boldsymbol{x}, \boldsymbol{b}_k\rangle)} \right\|_2 \leq \epsilon,$$

with:

$$\begin{aligned}
&\lambda = \Lambda\left(\frac{2\epsilon(\|f\|_\infty + L)}{\sqrt{m+1}(3\|f\|_\infty + 2L) - \epsilon} \right) \text{ from Equation (25)}, \\
&N \geq \Phi(m) \left(\frac{3\pi(L + \|f\|_\infty)\sqrt{m+1}\; c_{m+1}(\lambda)\exp(\lambda)}{\epsilon} \right)^{m+1}, \\
&\boldsymbol{\xi}_k = f(\boldsymbol{b}_k), \; \forall k = 1, \ldots, N.
\end{aligned}$$

Proof. From Corollary E.1 we know that $\sum_{k=1}^{N} \boldsymbol{\xi}_k \exp(\lambda\langle \boldsymbol{x}, \boldsymbol{b}_k\rangle)$ approximates $f(\boldsymbol{x})$ and from Lemma F.2 we know that $\frac{c_{m+1}(\lambda)}{N}\sum_{k=1}^{N} \exp(\lambda\langle \boldsymbol{x}, \boldsymbol{b}_k\rangle)$ approximates 1. Using Lemma F.1 we can combine the two results to bound how well

$$\frac{\sum_{k=1}^{N} \boldsymbol{\xi}_k \exp(\lambda\langle \boldsymbol{x}, \boldsymbol{b}_k\rangle)}{\frac{c_{m+1}(\lambda)}{N}\sum_{k=1}^{N} \exp(\lambda\langle \boldsymbol{x}, \boldsymbol{b}_k\rangle)}$$

approximates $f(\boldsymbol{x})/1 = f(\boldsymbol{x})$. The fact that $\frac{c_{m+1}(\lambda)}{N}\sum_{k=1}^{N} \exp(\lambda\langle \boldsymbol{x}, \boldsymbol{b}_k\rangle)$ is not identically 1 means that we will need to increase the precision of approximating the numerator by reducing ϵ in order to account for the additional error coming from the denominator. In particular, we have

$$\begin{aligned}
&\sup_{\boldsymbol{x}\in S^m} \left\| f(\boldsymbol{x}) - \sum_{k=1}^{N} \boldsymbol{\xi}_k \exp(\lambda\langle \boldsymbol{x}, \boldsymbol{b}_k\rangle) \right\|_2 && \leq \epsilon', \\
&\sup_{\boldsymbol{x}\in S^m} \left| 1 - \frac{c_{m+1}(\lambda)}{N}\sum_{k=1}^{N} \exp(\lambda\langle \boldsymbol{x}, \boldsymbol{b}_k\rangle) \right| && \leq \frac{\epsilon'}{2\sqrt{m+1}(L+\|f\|_\infty)}, \\
&\sup_{\boldsymbol{x}\in S^m} \|f(\boldsymbol{x})\|_2 && \leq \sqrt{m+1}\|f\|_\infty, \\
&\sup_{\boldsymbol{x}\in S^m} \left| \frac{c_{m+1}(\lambda)}{N}\sum_{k=1}^{N} \exp(\lambda\langle \boldsymbol{x}, \boldsymbol{b}_k\rangle) \right| && \leq 1 + \frac{\epsilon'}{2\sqrt{m+1}(L+\|f\|_\infty)}.
\end{aligned}$$

Hence, applying Lemma F.1 gives us:

$$\begin{aligned}
\left\| f(\boldsymbol{x}) - \frac{\sum_{k=1}^{N} \boldsymbol{\xi}_k \exp(\lambda\langle \boldsymbol{x}, \boldsymbol{b}_k\rangle)}{\frac{c_{m+1}(\lambda)}{N}\sum_{k=1}^{N} \exp(\lambda\langle \boldsymbol{x}, \boldsymbol{b}_k\rangle)} \right\|_2 &= \left\| \frac{f(\boldsymbol{x})}{1} - \frac{\sum_{k=1}^{N} \boldsymbol{\xi}_k \exp(\lambda\langle \boldsymbol{x}, \boldsymbol{b}_k\rangle)}{\frac{c_{m+1}(\lambda)}{N}\sum_{k=1}^{N} \exp(\lambda\langle \boldsymbol{x}, \boldsymbol{b}_k\rangle)} \right\|_2 \\
&\leq \frac{\epsilon' + \frac{\epsilon'}{2\sqrt{m+1}(L+\|f\|_\infty)}\sqrt{m+1}\|f\|_\infty}{1 + \frac{\epsilon'}{2\sqrt{m+1}(L+\|f\|_\infty)}} \\
&= \frac{\epsilon'\sqrt{m+1}(3\|f\|_\infty + 2L)}{2\sqrt{m+1}(\|f\|_\infty + L) + \epsilon'}
\end{aligned}$$

Therefore, if we want this error to be upper bounded by ϵ, we need to select

$$\epsilon' \leq \frac{2\epsilon\sqrt{m+1}(\|f\|_\infty + L)}{\sqrt{m+1}(3\|f\|_\infty + 2L) - \epsilon}.$$

The denominator must be larger than 0 but that is always the case for $\epsilon < 2\|f\|_\infty$, which is the largest non-trivial value one can select for ϵ. From Corollary E.1 (Equation (25)) that can be achieved by selecting

$$\lambda = \Lambda\left(\frac{2\epsilon(\|f\|_\infty + L)}{\sqrt{m+1}(3\|f\|_\infty + 2L) - \epsilon}\right)$$

and

$$N \geq \Phi(m)\left(\frac{3\pi(L+\|f\|_\infty)\sqrt{m+1}\, c_{m+1}(\lambda)\exp(\lambda)}{\epsilon}\right)^{m+1}.$$

Finally, observe that the ${}^{c_{m+1}(\lambda)}/_{N}$ factor can be folded in the $\boldsymbol{\xi}_k$ terms (Equation (33)):

$$\boldsymbol{\xi}_k = \frac{N}{c_{m+1}(\lambda)} f(\boldsymbol{b}_k)\, c_{m+1}(\lambda)\frac{w_m(V_k)}{w_m} = f(\boldsymbol{b}_k),$$

with $\boldsymbol{\xi}_k$ nicely reducing to be the evaluation of f at the corresponding control point $\boldsymbol{b}_k$. □

G Additional Results

Lemma G.1. *Let $f : [0,1]^m \to \mathbb{R}$ be a continuous function such that each component f_i is L-Lipschitz with respect to the $\|\cdot\|_2$ norm, $i = 1, \ldots, m$. Define the stereographic projection and its inverse:*

$$\Sigma_m : \bar{S}^m \to \mathbb{R}^m$$

$$(x_1, \dots, x_{m+1}) \mapsto \left(\frac{x_1}{1 - x_{m+1}}, \dots, \frac{x_m}{1 - x_{m+1}} \right)$$

$$\begin{aligned} \Sigma_m^{-1} &: \mathbb{R}^m \to S^m \\ (y_1, \dots, y_m) &\mapsto \left(\frac{2y_1}{\sum_{i=1}^m y_i^2 + 1}, \dots, \frac{2y_m}{\sum_{i=1}^m y_i^2 + 1}, \frac{\sum_{i=1}^m y_i^2 - 1}{\sum_{i=1}^m y_i^2 + 1} \right) \end{aligned}$$

with $\bar{S}^m$ the part of S^m that gets mapped to $[0,1]^m$, i.e., $\bar{S}^m = \Sigma_m^{-1}([0,1]^m)$. Σ_m and Σ_m^{-1} are continuous and inverses of each other and there exist L_m^{Σ} and $L_m^{\Sigma^{-1}}$ such that $\omega(\Sigma_m; t) \leq L_m^{\Sigma} t$ and $\omega(\Sigma_m^{-1}; t) \leq L_m^{\Sigma^{-1}} t$. Furthermore, $\Sigma_m \circ \mathcal{H}^1_{H,3(m+1)} \circ \Sigma_m^{-1}$ is dense in the set of continuous functions $[0,1]^m \to \mathbb{R}^m$.

Lemma G.2 (Lemma B.1 in the main text)**.** *$\Pi^{-1} \circ \mathcal{H}^1_{-,3(m+1)} \circ \Pi$ is dense in $\mathcal{H}_{\downharpoonleft\upharpoonright,m+1}$, with the composition applied to each function in the class.*

Proof. We can prove something stronger. For all $f \in \mathcal{H}_{\downharpoonleft\upharpoonright,m+1}$ there exists a $g \in \mathcal{H}^1_{-,3(m+1)}$ such that $f = \Pi^{-1} \circ g \circ \Pi$. If $f \in \mathcal{H}_{\downharpoonleft\upharpoonright,m+1}$, then

$$f(\boldsymbol{x}) = \frac{\sum_{k=1}^N \boldsymbol{p}_k^\beta \exp(\lambda \langle \boldsymbol{x}, \boldsymbol{p}_k^\alpha \rangle)}{\sum_{k=1}^N \exp(\lambda \langle \boldsymbol{x}, \boldsymbol{p}_k^\alpha \rangle)}, \ \forall \boldsymbol{x} \in S^m$$

for some N, λ, $\boldsymbol{p}_i^\alpha$, $\boldsymbol{p}_i^\beta$. Define:

$$\begin{aligned} \boldsymbol{p}_k &= \begin{bmatrix} \mathbf{0} \\ \lambda \boldsymbol{p}_k^\alpha \\ \boldsymbol{p}_k^\beta \end{bmatrix} \in \mathbb{R}^{3(m+1)}, \\ \boldsymbol{H} &= \begin{bmatrix} M\boldsymbol{I} & \boldsymbol{I} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{0} \end{bmatrix}, \boldsymbol{W}_V = \begin{bmatrix} \mathbf{0} & \mathbf{0} & \boldsymbol{I} \\ \mathbf{0} & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{0} \end{bmatrix} \in \mathbb{R}^{3(m+1) \times 3(m+1)}, \end{aligned}$$

With M a negative constant tending to $-\infty$. Then:

$$g(\boldsymbol{x}) = \frac{\sum_{i=1}^N \exp(\boldsymbol{x}^\top \boldsymbol{H} \boldsymbol{p}_i) \boldsymbol{W}_V \boldsymbol{p}_i + \exp(\boldsymbol{x}^\top \boldsymbol{H} \boldsymbol{x}) \boldsymbol{W}_V \boldsymbol{x}}{\sum_{i=1}^N \exp(\boldsymbol{x}^\top \boldsymbol{H} \boldsymbol{p}_i) + \exp(\boldsymbol{x}^\top \boldsymbol{H} \boldsymbol{x})},$$

is in $\mathcal{H}^1_{-,3(m+1)}$ and $f = \Pi^{-1} \circ g \circ \Pi$. As this holds for all $f \in \mathcal{H}_{\downharpoonleft\upharpoonright,m+1}$, it follows that $\mathcal{H}_{\downharpoonleft\upharpoonright,m+1} \subset \Pi^{-1} \circ \mathcal{H}^1_{-,3(m+1)} \circ \Pi$. Hence, $\Pi^{-1} \circ \mathcal{H}^1_{-,3(m+1)} \circ \Pi$ is dense in $\mathcal{H}_{\downharpoonleft\upharpoonright,m+1}$. □