QUALITY OVER QUANTITY IN ATTENTION LAYERS: WHEN ADDING MORE HEADS HURTS

Noah Amsel[1], Gilad Yehudai[1] & Joan Bruna[1,2]
[1]Courant Institute of Mathematical Sciences, New York University [2]Flatiron Institute
{noah.amsel,gy2219,jb4496}@nyu.edu

ABSTRACT

Attention-based mechanisms are widely used in machine learning, most prominently in transformers. However, hyperparameters such as the number of attention heads and the attention rank (i.e., the query/key dimension) are set nearly the same way in all realizations of this architecture, without theoretical justification. In this paper, we prove that the rank can have a dramatic effect on the representational capacity of attention. This effect persists even when the number of heads and the parameter count are very large. Specifically, we present a simple and natural target function based on nearest neighbor search that can be represented using a single full-rank attention head for any sequence length. We prove that it cannot be approximated by a low-rank attention layer even on short sequences unless the number of heads is exponential in the embedding dimension. Thus, for this target function, rank is what determines an attention layer's power. We show that, for short sequences, using multiple layers allows the target to be approximated by low-rank attention; for long sequences, we conjecture that full-rank attention is necessary regardless of depth. Finally, we present experiments with standard multilayer transformers that validate our theoretical findings. They demonstrate that, because of how all standard transformer implementations set the rank, increasing the number of attention heads can severely decrease accuracy on certain tasks.

1 INTRODUCTION

Attention-based architectures are ubiquitous in contemporary machine learning. The most prominent examples are transformers, which are invaluable tools for processing images, audio, time series, PDEs, and biological data in addition to natural language. The basic transformer architecture leaves the user free to set several hyperparameters, but few of these have been carefully studied. In fact, in the thousands of papers that use this architecture, certain hyperparameters are almost always kept the same as in the original work of Vaswani et al. (2017) that introduced transformers. (See Appendix A for a comparison.) In this paper, we study one of these hyperparameters: the attention rank.

The rank of an attention layer refers to the dimension of the query and key vectors formed by each attention head. The capacity of an attention layer is determined by its rank (r) and its number of heads (H). When the input is a sequence of embedding vectors in $\mathbb{R}^d$, its total number of parameters is of order dHr. Most transformers in the literature still use a small rank of between 64 and 128, even though the embedding dimension has grown dramatically from $d = 512$ in the original transformer (Vaswani et al., 2017) to $d = 8,192$ in LLaMA (Touvron et al., 2023a). It is not clear whether the expressive power of transformers is weakened by maintaining a fixed rank as the dimension is increased. In addition, nearly all work on transformers sets the number of heads to be $H = d/r$ (see Appendix A). In fact, this scaling is so standard that it is hard-coded into libraries like PyTorch (Paszke et al., 2019) and xFormers (Lefaudeux et al., 2022), a fact which has likely discouraged experimentation with other scalings. The original motivation for this scaling in Vaswani et al. (2017) was simply to match the parameter count of multi-head attention with that of earlier architectures, which used one full-rank attention head ($H = 1, r = d$). We know of no *a priori* reason or experimental evidence that favors this scaling over any other, as the trade-offs between the rank and the number of heads are still not well-understood. This paper raises doubts about whether the universal practice of setting $r \ll d$ and $H = d/r$ is always a wise choice.

To demonstrate how impactful these hyperparameters can be, we reproduce the influential experiment of Garg et al. (2022), which shows that transformers can be trained to perform linear regression. This experiment is widely studied by both empiricists and theoreticians as an archetype of in-context learning (Von Oswald et al., 2023; Bai et al., 2023; Ahn et al., 2023; Akyürek et al., 2023; Zhang et al., 2024). The original paper set $H = 8$. In Figure 1, we show that simply reducing H while maintaining the scaling $H = d/r$ yields significantly better accuracy without changing the number of parameters in the model. (See Appendix B.1 for details.)

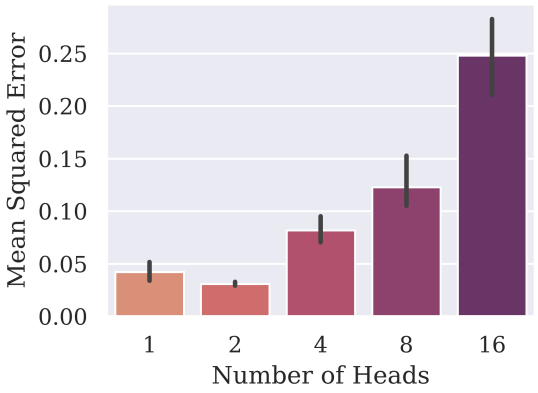

Figure 1: For the in-context learning task of Garg et al. (2022), transformers with fewer heads (that is, higher rank) perform significantly better despite having the same number of parameters. See Appendix B.1 for details.

The first step towards understanding such hyperparameter scalings in transformers is to focus on their expressive power. Our approach is analogous to a long line of work in the theory of deep learning that has analyzed the relative importance of width and depth in determining the expressive power of feedforward neural networks (FNNs). For FNNs, depth two suffices for universal approximation (Cybenko, 1989), but greater depth may be required for *efficient* approximation. That is, some functions can be efficiently represented by a three-layer FNN, but cannot be represented by a two-layer FNN unless its width is exponentially wide in the input dimension (Eldan & Shamir, 2016; Daniely, 2017; Safran & Shamir, 2017). This shows that width and depth are not exchangeable; even very large width does not compensate for the depth being too small. It is natural to ask similar questions about attention-based architectures: Does using low rank ($r \ll d$) weaken the expressive power of attention? Does using a large number of heads H compensate for this weakness? What is the most parameter-efficient way to scale up our models?

In this paper, we study precisely these trade-offs in the expressive capacity of attention layers. We present a simple target function based on nearest neighbor search which can be efficiently approximated by a single-head, full-rank attention layer (hence with a total of $O(d^2)$ parameters), but not by a low-rank attention layer. This separation in expressive power persists even if the total number of parameters satisfies $Hdr \gg d^2$. This shows that rank, rather than the total number of parameters, can be the main factor influencing an attention layer's capacity. Stacking multiple layers of low-rank attention can yield a moderately-efficient approximation of our target too, at least for short context lengths. We complement these theoretical results with experiments on fully-featured transformer architectures as implemented in PyTorch.

1.1 Our Contributions

- In Section 4, we prove a rank separation for representing the nearest neighbor function using multi-head attention. This function can be approximated to any accuracy using only a single full-rank head. Yet in the high-dimensional regime, at least $\Omega\left((d/r)^{1/\epsilon}\right)$ heads of rank r are required to achieve relative squared error ϵ. Moreover, in the high-accuracy regime (ϵ going to zero with d fixed), the required number of heads is exponential: $\Omega(\exp(d - r\log(d/r)))$.
- In Section 5, we establish an exponential separation in both regimes by combining multiple biased nearest neighbor functions. This combined target function can be approximated to any accuracy using polynomially-many full-rank heads, but $\Omega(\exp(d - r))$ rank-r heads are required to approximate it with better than a constant error.
- In Section 6, we explore ways to circumvent the weakness of low-rank attention. We show that augmenting the attention architecture and adding a second, non-linear layer can achieve

this using polynomially many heads, but unlike full-rank attention, such constructions may not scale to long sequence lengths.

- In Section 7, we support our theoretical results with experiments on standard transformer architectures with multiple layers of attention and MLPs. We show that the full rank models easily learn the target to high accuracy — even recovering our main construction — but the low rank models struggle to do so. For the well-known in-context learning task of Garg et al. (2022), larger rank likewise yields higher accuracy with fewer parameters. Users of standard transformers may not realize that setting $H = 2$ could be much worse than $H = 1$, but in some cases, it is.

2 Related Work

Theory of transformers A growing line of work has sought to provide theoretical analysis of transformers and the attention mechanism. Training dynamics, inductive biases, generalization, and in-context learning have all received significant attention. However, papers in these areas nearly always assume that full-rank attention is used (Bietti et al., 2023; Cabannes et al., 2024; Fu et al., 2023; Sanford et al., 2024a; Edelman et al., 2022; Bai et al., 2023; Zhang et al., 2024; Jelassi et al., 2024; Chen et al., 2024; Deora et al., 2023; Tian et al., 2023), even though many also assume there are multiple heads. Our work provides important context for these results, showing that full-rank models may not be good proxies for the low-rank transformers used in practice.

Expressive power of transformers Our work belongs to a body of research studying the representational capacity of transformers. Unlike other topics in transformer theory, results in this area often do apply to low-rank attention. Yun et al. (2019) prove that (exponentially deep) transformers are universal approximators even with rank one. Wei et al. (2022) and Merrill & Sabharwal (2023) show that transformers can simulate Turing machines if their size is allowed to grow with the sequence length. Kim et al. (2022) and Kajitsuka & Sato (2023) show that transformers are capable of memorizing data. Bhattamishra et al. (2024) show that transformers can efficiently implement a version of the nearest neighbor algorithm for in-context classification of points on the sphere, but their construction uses attention that is full-rank with respect to the input dimension. Our formulation of the nearest neighbor task is slightly different and can be solved with full-rank attention almost trivially (see Fact 1). Finally, an important line of work analyzes the representational capacity of transformers using classes of formal languages, finite automata, and circuits (Hahn, 2020; Liu et al., 2022; Hao et al., 2022; Merrill et al., 2022; Strobl et al., 2024), but it does not capture separations in capacity due to rank.

Limitations of low-rank attention Several other studies have investigated the role of the rank of the attention mechanism. Bhojanapalli et al. (2020) present experiments that challenge the canonical $H = d/r$ scaling. They argue that fixing d and r based on the context length N and setting H independently leads to more powerful and efficient models. They also prove that a full-rank attention head can produce any attention pattern from any input (for *some* setting of the weights), but a low-rank attention head cannot; however, this implies nothing about the representational power of low-rank attention or how it relates to the number of heads. In addition, Likhosherstov et al. (2023) show that even rank $r = \log(N)$ suffices to represent any *sparse* attention pattern. Mahdavi et al. (2024) ask how many input-output pairs a low-rank multi-head attention layer can exactly memorize. For their problem, it is not worth setting $r > N$; furthermore the memorization capacity depends on rH rather than on r or H, supporting the standard scaling. We study the more realistic and practically motivated setting of approximating a natural function over data drawn from a natural distribution. Unlike Likhosherstov et al. (2023) and Mahdavi et al. (2024), we show that high rank is sometimes essential, irrespective of H.

The paper closest to our own is Sanford et al. (2024b), which proves two separations related to rank. First, they present a function that can be well-approximated by a single attention head if and only if its rank is sufficiently large. This result prompts the following question: can using multiple heads compensate for the weakness of low-rank attention? We answer this question in the negative. Second, they present a one-dimensional function on N inputs that is impossible to represent exactly unless $rHp > N$, where p is the bits of precision. We extend this result in that our lower bounds apply (1) even for $N = 2$, (2) for infinite or finite precision (3) to function approximation over a nat-

ural distribution, not just exact representation. Additionally, our target function engenders a stronger separation: while $H \geq \Omega(1/r)$ suffices in their setting, ours requires H to grow polynomially or even exponentially in d/r to overcome the weakness of low-rank attention. However, their target functions are more closely akin to the kinds of structured reasoning tasks to which transformers are often applied. In particular, they highlight how attention is naturally suited to capturing pairwise interactions; recurrent architectures struggle to do this efficiently, while transformers struggle to capture third-order interactions.

Finally, an important reference to the dangers of low rank appears in Yang et al. (2022), which introduced the highly influential μP method for hyperparameter transfer. They find that as d and r decrease, transformers start to behave badly. Thus, they set $r > d/H$ in some cases. See Appendix D.4 and E.2.

Low rank compression and fine-tuning Much recent work in model compression (Lv et al., 2023; Hajimolahoseini et al., 2021; Ben Noach & Goldberg, 2020) and fine-tuning (Hu et al., 2022) is based on the empirical observation that the weight matrices of pretrained transformers (like those of other neural networks) can be replaced or fine-tuned by lower-dimensional proxies without sacrificing performance, and in some cases even helping it (Sharma et al., 2024). Such results contextualize our work by showing that full-rank is not *always* better than low-rank.

Depth-width trade-offs in neural networks Many previous papers have proved separations between neural networks of different depths and between neural networks and kernel methods. Several studies (Eldan & Shamir, 2016; Daniely, 2017; Safran & Shamir, 2017; Venturi et al., 2022) constructed functions that can be approximated efficiently with a 3-layer neural network, but for which 2-layer networks require the width to be exponential in the input dimension. Telgarsky (2016) and Chatziafratis et al. (2019) show depth separation for networks with constant input dimension and varying depths. Our lower bounds are also closely related technically to separation results between neural networks and kernel methods. Yehudai & Shamir (2019) prove that random features (or any other kernel method) cannot learn even a single neuron unless the number of features or magnitude of the weights is exponential in the input dimension. Kamath et al. (2020) improved on their result by removing the dependence on the magnitude of the weights. Ghorbani et al. (2021) and Misiakiewicz & Montanari (2023) study upper and lower bounds in approximating polynomials with kernel methods. They show that essentially, it is necessary and sufficient for the number of features to be exponential in the degree of the approximated polynomial. Our lower bounds are inspired by this work.

3 Setting and Notations

Attention layers A rank-r attention head is parameterized by the weight matrices $\boldsymbol{Q}, \boldsymbol{K}, \boldsymbol{V}, \boldsymbol{O} \in \mathbb{R}^{d\times r}$. (Some authors call these $\boldsymbol{W}_Q, \boldsymbol{W}_K, \boldsymbol{W}_V$, and $\boldsymbol{W}_O$.) A multi-head attention layer is simply the sum of H such attention heads. The input to a multi-head attention layer is a sequence of vectors $\boldsymbol{x}_1, \dots \boldsymbol{x}_N \in \mathbb{R}^d$ called the target points and a sequence $\boldsymbol{y}_1 \dots \boldsymbol{y}_M$ called the source points. (Note that the name "target points" is unrelated to that of the "target function" we wish to approximate.) If the columns of $\boldsymbol{X} \in \mathbb{R}^{d\times N}$ and $\boldsymbol{Y} \in \mathbb{R}^{d\times M}$ are the target and source points, respectively, then a softmax multi-head attention layer is a function of the form

$$\sum_{h=1}^{H} \boldsymbol{O}_h \boldsymbol{V}_h^\top \boldsymbol{X}\, \mathrm{sm}\left(\boldsymbol{X}^\top \boldsymbol{K}_h \boldsymbol{Q}_h^\top \boldsymbol{Y}\right) \in \mathbb{R}^{M\times d}\,, \tag{1}$$

where $\mathrm{sm}(\cdot)$ computes the softmax of each column of its input; that is, for each $\boldsymbol{y}$, it outputs a probability distribution over $[N]$ based on the scores $\boldsymbol{X}^\top \boldsymbol{K}_h \boldsymbol{Q}_h^\top \boldsymbol{y} \in \mathbb{R}^N$. A hardmax attention layer is the same, except that the hardmax function $\mathrm{hm}(\cdot)$ outputs $\boldsymbol{e}_{i^*}$, where i^* is the index of the maximum score. Note that hardmax heads are often considered to be a special case of softmax heads, since $\lim_{c\to\infty} \mathrm{sm}(\boldsymbol{X}^\top c\boldsymbol{K}_h \boldsymbol{Q}_h^\top \boldsymbol{Y}) = \mathrm{hm}(\boldsymbol{X}^\top \boldsymbol{K}_h \boldsymbol{Q}_h^\top \boldsymbol{Y})$ in pointwise convergence.

Above, we have described so-called cross-attention, which takes both source points and target points as input. The familiar self-attention layers are a special case in which the source points and target points are identical: $\boldsymbol{X} = \boldsymbol{Y}$. A given multi-head attention function can be applied to any number of source or target points, since no part of this definition depends on N or M. In addition, it is invariant to permutations of the target points and equivariant to permutations of the source points.

Generalized attention We prove our lower bounds against a class of functions that generalizes multi-head attention. Rather than computing the attention distribution as $\mathrm{sm}(\boldsymbol{X}^\top \boldsymbol{K}_h \boldsymbol{Q}_h \boldsymbol{Y})$, we allow any function depending on $\boldsymbol{y}$ and a rank-r projection of $\boldsymbol{X}$ that outputs a probability distribution over $[N]$. In addition, we replace $\boldsymbol{O}_h \boldsymbol{V}_h$ with a single matrix $\boldsymbol{V}_h \in \mathbb{R}^{d \times d}$. Thus, our model is

$$\sum_{h=1}^{H} \boldsymbol{V}_h \boldsymbol{X} \phi_h \left(\boldsymbol{K}_h^\top \boldsymbol{X}, \boldsymbol{Y} \right) , \tag{2}$$

where $\boldsymbol{K}_h \in \mathbb{R}^{d \times r}$, the function $\phi_h : \mathbb{R}^{r \times N} \times \mathbb{R}^d \to \Delta^{N-1}$ is applied column-wise to $\boldsymbol{Y}$, and Δ^{N-1} is the simplex. Note that the function ϕ_h may vary between heads. Moreover, we allow $\boldsymbol{V}_h \in \mathbb{R}^{d \times d}$ to be full-rank. Note that this class captures, beyond standard transformer architectures, the use of biases, additive positional encodings, and other encoding schemes like RoPE (Su et al., 2024) and ALiBi (Press et al., 2022) in the attention layer. We also capture architectures from early works on attention (Bahdanau et al., 2014; Xu et al., 2015), which used feedforward networks to compute the attention scores instead of the "multiplicative" or "dot product" attention scores $\boldsymbol{X}^\top \boldsymbol{K} \boldsymbol{Q} \boldsymbol{Y}$ used in transformers.

Nearest neighbor function The input to the nearest neighbor function consists of a sequence of N target points from the unit sphere $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_N \in \mathbb{S}^{d-1}$ (also denoted by $\boldsymbol{X} \in \mathbb{R}^{d \times N}$) and a source point $\boldsymbol{y} \in \mathbb{S}^{d-1}$.

The nearest neighbor function outputs the target point that is closest to the source:

$$f(\boldsymbol{x}_1, \ldots, \boldsymbol{x}_N; \boldsymbol{y}) := \operatorname*{arg\,min}_{\boldsymbol{x} \in \{\boldsymbol{x}_1, \ldots \boldsymbol{x}_N\}} \|\boldsymbol{x} - \boldsymbol{y}\|_2 . \tag{3}$$

This function is analogous to performing a semantic search, in which the goal is to retrieve the entry or word in a database or context window that most closely matches a query. This function is highly symmetric. Like multi-head attention itself, it is defined for any N and is invariant to permutations of the target points. It is also invariant to simultaneous orthogonal transformations of $\boldsymbol{X}$ and $\boldsymbol{y}$, so it has no principal directions, subspaces, or scales.

Data distribution We draw target and source points uniformly from the sphere. For our lower bounds, it is convenient to assume that the target points are orthogonal. For $N \leq d$, let $\mathcal{D}_N(\mathbb{S}^{d-1})$ denote the uniform distribution over the set of sequences $\boldsymbol{x}_1, \ldots \boldsymbol{x}_N \in \mathbb{S}^{d-1}$ for which $i \neq j \implies \boldsymbol{x}_i \perp \boldsymbol{x}_j$. Such samples can be generated by taking the first N columns of a random orthonormal matrix. Note that this is similar in essence to drawing the data points independently from the unit sphere, as isotropic random vectors in high dimension are nearly orthogonal. This distribution is invariant to orthogonal transformations of $\boldsymbol{X}$ and of $\boldsymbol{y}$.

4 Low-Rank Separation for Nearest Neighbors

In this section, we study the capacity of multi-head attention to represent the nearest-neighbor function. We show a separation in representational power based on rank. The target can be represented efficiently using full-rank attention, but under the assumptions below, approximating it using low-rank attention requires a much larger model. We begin with the upper bound using a single full-rank attention head:

Fact 1 (Full-rank Efficient Approximation, Equivariant Case)**.** *For the target function from Equation* (3)*, any $\epsilon > 0$, $N, d \in \mathbb{N}$ there exist $\boldsymbol{K}, \boldsymbol{Q}, \boldsymbol{V} \in \mathbb{R}^{d \times d}$ such that:*

$$\mathop{\mathbb{E}}_{\boldsymbol{y}, \boldsymbol{x}_1, \ldots, \boldsymbol{x}_N \sim \mathrm{Unif}(\mathbb{S}^{d-1})} \left[\left\| f(\boldsymbol{X}, \boldsymbol{y}) - \boldsymbol{V}\boldsymbol{X} \,\mathrm{sm}(\boldsymbol{X}^\top \boldsymbol{K} \boldsymbol{Q}^\top \boldsymbol{y}) \right\|^2 \right] \leq \epsilon . \tag{4}$$

The construction is straightforward. Consider for simplicity the hardmax case. Set $\boldsymbol{V} = \boldsymbol{K}\boldsymbol{Q}^\top = \boldsymbol{I}$ so that $\|\boldsymbol{x}_i - \boldsymbol{y}\|_2 = 2 - \boldsymbol{x}_i^\top \boldsymbol{K}\boldsymbol{Q}^\top \boldsymbol{y}$. Then $\mathrm{hm}(\boldsymbol{X}^\top \boldsymbol{K}\boldsymbol{Q}^\top \boldsymbol{y}) = \boldsymbol{e}_{i^*}$ where $i^* = \arg\min_{i \in [N]} \|\boldsymbol{x}_i - \boldsymbol{y}\|_2$ and $\boldsymbol{e}_i$ is the ith standard basis vector. Note that this construction using hardmax works for any input distribution on $\mathbb{S}^{d-1}$ and any number of points N, as it represents the target function exactly. The softmax case is similar; for the formal statement see appendix Appendix C.1. This construction (or one very similar to it) is easily learned by gradient descent; see Figure 5.

We now turn to the lower bound. We show that approximating the target function with rank-r heads requires the number of heads to be large unless $r \sim d$. For technical convenience, we set the number of target points to two and draw them from the distribution $\mathcal{D}_2(\mathbb{S}^{d-1})$ in which they are always orthogonal. Our main result establishes a strong quantitative separation between full-rank and low-rank self-attention layer, even when the total number of parameters is of the same order:

Theorem 2 (Low-Rank Approximation Lower Bounds, Equivariant Case)**.** *There exist universal constants c, c', C and C' such that if either of the following sets of assumptions hold:*

1. High-accuracy regime*: $r \leq d-3$, $\epsilon \leq \frac{c}{d+1}$, and*

$$H \leq C \cdot 2^{d-(r+1)\log_2(2d/r)} \,. \tag{5}$$

2. High-dimensional regime: $d \geq 5$, $\epsilon \geq \frac{c'}{d-2e^2 \cdot r}$ *and*

$$H \leq \frac{1}{2}\left(\frac{1}{2e} \cdot \frac{d}{r + C'/\epsilon}\right)^{C'/\epsilon} \,. \tag{6}$$

Then, for any choice of H rank-r generalized attention heads $\phi_h : \mathbb{R}^{r\times 2} \to \Delta^1$, $\boldsymbol{V}_h \in \mathbb{R}^{d\times d}$, $\boldsymbol{K}_h \in \mathbb{R}^{d\times r}$ the error of approximating the nearest neighbor function is bounded as follows

$$\mathop{\mathbb{E}}_{\substack{\boldsymbol{x}_1,\boldsymbol{x}_2\sim\mathcal{D}_2(\mathbb{S}^{d-1})\\ \boldsymbol{y}\sim\mathrm{Unif}(\mathbb{S}^{d-1})}} \left\| f(\boldsymbol{X};\boldsymbol{y}) - \sum_{h=1}^{H} \boldsymbol{V}_h \boldsymbol{X} \phi_h\left(\boldsymbol{K}_h^\top \boldsymbol{X}, \boldsymbol{y}\right)\right\|_2^2 \geq \epsilon \,, \tag{7}$$

where f is defined as in Equation (3)*.*

For the proof of Theorem 2, see Appendix C. Intuitively, the approximation problem becomes harder as $d \to \infty$ and as $\epsilon \to 0$. Theorem 2 combines guarantees in two different regimes. In the first regime, the desired accuracy ϵ is small. In this case, the necessary number of heads grows exponentially with $d - r$. In the second regime, the dimension d is be large. In this case, the necessary number of heads grows polynomially with d/r. Informally, both regimes show that the error is at least ϵ whenever $H \lesssim (d/r)^{1/\epsilon}$.

We emphasize that the data distribution is $\frac{1}{\sqrt{d}}$-close to the uniform product measure in Wasserstein distance, and we expect our main proof techniques to generalize to this uniform measure, as well as to other rotationally invariant distributions. Additionally, while $N = 2$ is sufficient for our goal of establishing the separation, we also believe the framework should extend to the general setting of $N > 2$, although this is out of the present scope.

Our proof uses tools from harmonic analysis on the sphere. It is reminiscent of the original depth separation work of Eldan & Shamir (2016) and Daniely (2017), which also exploited the inability of ridge functions to approximate radially-symmetric targets with substantial high-frequency energy. Due to the rotational symmetry of the target function, attention function, and data distribution, we can transform our problem to depend on a pair of points $\boldsymbol{x} = \boldsymbol{x}_1 - \boldsymbol{x}_2$ and $\boldsymbol{y}$ drawn uniformly from the sphere, rather than $\boldsymbol{x}_1, \boldsymbol{x}_2$ and $\boldsymbol{y}$. Our target is essentially given by a step function of the form $(\boldsymbol{x}, \boldsymbol{y}) \mapsto \mathrm{sgn}(\boldsymbol{x}^\top \boldsymbol{y})$, which has a slowly decaying spectrum with respect to the appropriate basis. We construct this basis using spherical harmonics, and like them, our basis functions are organized into orthogonal subspaces based on degree ℓ polynomials. Due to rotational symmetry, the energy of the target function is uniformly spread within each harmonic subspace. In contrast, each attention head is tied to a few principal directions given by the span of $\boldsymbol{K}_h$. As a result, each head is spanned by only a fraction of the basis functions in each subspace. Thus, with a limited number of heads, it is impossible to capture a substantial fraction of the energy of the target function.

We now comment on the tightness of this lower bound, focusing on the canonical setting of $r = 1$. In this case, our lower bound simplifies and strengthens slightly. For fixed ϵ and large d, the error of approximation is at least ϵ whenever $H = O\left(d^{1/(4\epsilon)}\right)$. We can construct an upper bound for our problem by considering rank-1 heads to be random features. In Appendix C.8, we argue that we can approximate our target function in the RKHS associated with the feature map $(\boldsymbol{x}_1 - \boldsymbol{x}_2, \boldsymbol{y}) \mapsto \mathrm{sgn}\left((\boldsymbol{x}_1 - \boldsymbol{x}_2)^\top \boldsymbol{k}\boldsymbol{q}^\top \boldsymbol{y}\right)$, where $\boldsymbol{k}$ and $\boldsymbol{q}$ are drawn uniformly from the unit sphere. The associated

kernel integral operator diagonalizes in the same basis of tensorized spherical harmonics used to decompose the target function above, and thus the kernel ridge regression approximation can be explicitly analysed by bounding the spectral decay of the kernel. Then, via standard arguments from random feature expansions (Bach, 2017b), one can transfer the approximation guarantees from the RKHS to the random feature model, provided that $H = \widetilde{\Omega}(d^{2/\epsilon^2})$. Thus, for $r = 1$ and fixed ϵ, the approximation lower bound of Theorem 2 captures the qualitatively correct behavior, though its precise dependence on d may not be tight.

5 Exponential Separation for Low-Rank attention

In the previous section, we proved a polynomial separation between low-rank and full-rank transformers in the constant accuracy regime (part 2 of Theorem 2). That is, to achieve error smaller than ϵ, where ϵ is a constant not depending on d or r, the number of heads must be at least poly $\left(\frac{d}{r}\right)$. In this section, we prove a stronger, exponential separation in this regime. That is, we find a target function that cannot be approximated by low-rank transformers to $O(1)$-error unless the number of heads is $\exp(\Omega(d-r))$. This new target function is defined as

$$f^*(\boldsymbol{x}_1, \ldots, \boldsymbol{x}_N, \boldsymbol{y}) = \sum_{t=1}^{2d^2+1} (-1)^t \underset{i\in\{1,\ldots,N\}}{\arg\max} \left(d \left\| \boldsymbol{x}_i - \boldsymbol{y} \right\|^2 + b_{t,i} \right) - \frac{1}{2} \sum_{j=1}^{N} \boldsymbol{x}_j \ , \tag{8}$$

where $b_{t,i} = \begin{cases} t & i = 1 \\ 0 & \text{o.w.} \end{cases}$. This function can be viewed as a sum of polynomially many nearest neighbors functions, each with a different bias term. Using this target function, we can show the following exponentially-strong separation between full-rank and low-rank attention:

Theorem 3. *The function f^* defined above satisfies the following:*

1. *For any $N \geq 2$ and $\epsilon > 0$, there exists a full-rank transformer T with $2d^2 + 2$ biased attention heads such that:*
$$\mathop{\mathbb{E}}_{\substack{\boldsymbol{x}_1,\ldots,\boldsymbol{x}_N \sim \mathcal{D}_N\left(\mathbb{S}^{d-1}\right) \\ \boldsymbol{y}\sim\mathcal{N}\left(0,\frac{1}{\sqrt{d}}I\right)}} \left[\left\| T(\boldsymbol{x}_1, \ldots, \boldsymbol{x}_N, \boldsymbol{y}) - f^*(\boldsymbol{x}_1, \ldots, \boldsymbol{x}_N, \boldsymbol{y}) \right\|^2 \right] \leq \epsilon \tag{9}$$
2. *For any rank-r transformer T with $r < d$, there exists a universal constant $c > 0$ such that if $dH \cdot \max_{h\in[H]} \left\| V_h \right\|^2 < \exp(c(d-r))$ then:*
$$\mathop{\mathbb{E}}_{\substack{\boldsymbol{x}_1,\boldsymbol{x}_2 \sim \mathcal{D}_2\left(\mathbb{S}^{d-1}\right) \\ \boldsymbol{y}\sim\mathcal{N}\left(0,\frac{1}{\sqrt{d}}I\right)}} \left[\left\| T(\boldsymbol{x}_1, \boldsymbol{x}_2, \boldsymbol{y}) - f^*(\boldsymbol{x}_1, \boldsymbol{x}_2, \boldsymbol{y}) \right\|^2 \right] > \frac{1}{40} \tag{10}$$

The full proof can be found in Appendix D. The above theorem states that the function f^* defined above can be approximated by a full-rank transformer with polynomially many heads, but rank-r transformers need exponentially-many heads to approximate it up to a constant. This result is reminiscent of the exponential separation between 2- and 3-layer networks from Eldan & Shamir (2016); Daniely (2017), in which the target function requires polynomially many neurons to be approximated by a 3-layer network, but exponentially many neurons for a 2-layer network.

Theorem 3 achieves a stronger separation than Theorem 2 but uses a more complicated target function. In addition, the full rank construction here requires attention to have a bias term inside the softmax. Another difference is that $\boldsymbol{y}$ is drawn from a Gaussian instead of the uniformly from sphere, but in sufficiently high dimension, this difference is insignificant. Note also that $\mathbb{E}\,||f^*|| = O(1)$, so our separation holds for the relative error too (see the discussion in Appendix D.2).

Remark 4 (Bound on the weights)**.** *Note that unlike Theorem 2, the bound in Theorem 3 does not apply to the number of heads per se; either H or the norm of the output weight matrix $\boldsymbol{V}_h$ must be exponential to overcome the hardness. A similar bound is found in Yehudai & Shamir (2019) which inspires our proof. In Kamath et al. (2020) the authors were able to remove the dependence on the scale of the weights by applying a more intricate analysis involving the SQ-dimension; however, in our case it is not clear how to extend their technique because of the dependence on r. We conjecture that it is still possible to remove this dependence and leave it for future work.*

The proof of the lower bound (part 2 of Theorem 3) is inspired by and extends the proof technique of Yehudai & Shamir (2019), which separates kernel methods from 2-layer networks. For the intuition behind the proof, see Appendix D.1. We note that to achieve exponential separation with constant error, we used a target that contains polynomially-many nearest neighbor functions. In fact, previous results (Hsu et al., 2021; Safran et al., 2019) imply that such a strong separation cannot be achieved with just a single nearest neighbor function. For a more in-depth discussion of this issue, see Appendix D.2.

6 Efficient Approximation Using Depth

In the previous sections, we showed that a single layer of low-rank attention fails to represent the target unless the number of heads is very large. In this section, we take up the question of whether additional layers of depth can overcome this weakness. We present a construction that approximates the target function (with slightly modified inputs) using two layers and only polynomially many rank-1 heads. However, we present constructions only for the case where the context length $N = 2$, which is also the setting of our lower bounds. We conjuncture that any construction using low-rank heads introduces an unfavorable dependence on N, a significant weakness compared to full-rank attention.

Our constructions are based on the strategy we call "majority voting", which we briefly describe here. Consider the case of $N = 2$ target points and hardmax attention. The output of each head, like the target function itself, is either $\boldsymbol{x}_1$ or $\boldsymbol{x}_2$. A random rank-1 head is weakly correlated with the target; the probability that it outputs the correct answer is $1/2 + \Omega(1/\sqrt{d})$. Thus, combining many such random heads together, their mode (the output with the most "votes") matches the target function with high probability. We use a second layer to calculate the "majority vote" of the heads in the attention layer.

Standard attention mechanisms make it difficult to count the number of votes each target point received—or even to remember what the target points $\boldsymbol{x}_1$ and $\boldsymbol{x}_2$ were—since the next layer gets only a linear combination of them with unknown coefficients. Therefore, we slightly modify the attention layer to facilitate the majority voting strategy. We concatenate labels to the vectors that allow us to count how many times $\boldsymbol{x}_1$ and $\boldsymbol{x}_2$ appear in the sum. We then use a second layer of attention to look up the full vector corresponding to the majority label. This labeling can be implemented by concatenating positional encodings to the input points. That is, instead of inputting $\boldsymbol{x}_1, \dots \boldsymbol{x}_N \in \mathbb{S}^{d-1}$ to the transformer, we now input $\begin{bmatrix}\boldsymbol{x}_1 \\ \boldsymbol{b}_1\end{bmatrix}, \dots, \begin{bmatrix}\boldsymbol{x}_N \\ \boldsymbol{b}_N\end{bmatrix}$ for $\boldsymbol{b}_i \in \mathbb{R}^e$. A linear transformation can be used to map the output of this $(d+e)$-dimensional transformer back to $\mathbb{R}^d$. Note that our target function is permutation-invariant, so the order of the points is irrelevant to the task at hand. Thus, these concatenated "positional encodings" function more like a modification to the architecture. They provide extra input dimensions that serve as scratch space in which the model can perform discrete operations like counting and indexing without corrupting the input data. Also note that, because they change the dimension of the inputs and of the transformer, these concatenated positional encodings are different from the positional encodings used in practice such as RoPE (Su et al., 2024) and ALiBi (Press et al., 2022), which are included under our framework of generalized attention.

The following theorem describes our majority voting construction using a two-layer transformer with rank-1 heads, concatenated positional encodings, and self-attention. Because we use self-attention, the source and target points are now the same. For the precise definition of this model and the proof of this theorem, see Appendix E.1 and Appendix E.4.

Theorem 5. *There exist universal constants c_1, c_2 such that for all $d > c_1$, and $\epsilon \in \left(0, \frac{1}{2}\right)$, and $H \geq c_2 \cdot \frac{d^3}{\epsilon^2}$, there exists a 2-layer, rank-1 transformer T with H heads and 2-dimensional positional encodings (as defined in Definition 33) for which*

$$\mathop{\mathbb{E}}_{\boldsymbol{x}_1, \boldsymbol{x}_2, \boldsymbol{y} \sim \mathrm{Unif}(\mathbb{S}^{d-1})} \left\| f(\boldsymbol{x}_1, \boldsymbol{x}_2; \boldsymbol{y}) - T\left([\boldsymbol{x}_1 \quad \boldsymbol{x}_2 \quad \boldsymbol{y}]\right) \right\|_2^2 \leq \epsilon \,. \tag{11}$$

One might wonder whether the concatenated positional encodings are necessary to make this construction work, especially since they break permutation invariance in order to represent a permutation invariant target. In Appendix E, we present an alternative construction (Theorem 32) that is

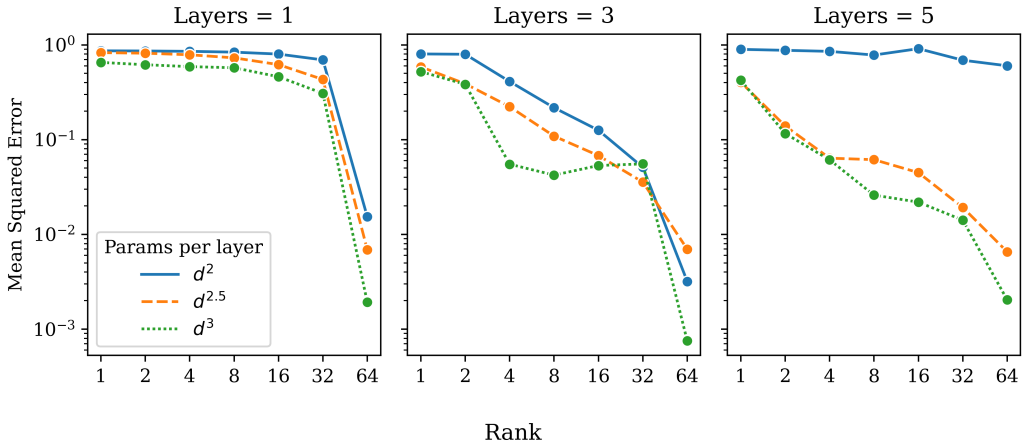

Figure 2: Standard transformers trained on the farthest neighbor function. The dimension is $d = 64$ and the number of input points is $N = 16$. Line shows best of five runs (see Appendix B.2 for the spread). Across different numbers of layers, high-rank models significantly outperform low-rank models that have the same number of parameters.

permutation invariant. However, it modifies the architecture by concatenating the outputs of the attention heads together instead of summing them, and it passes the result (which has dimension dH) to an MLP layer instead of a second attention layer.

Although these constructions assume $N = 2$ source points, it seems feasible to generalize them to larger N. However, the major drawback of such a generalization is that the size of the transformer will depend on N. Even the simple step of calculating the majority between N possible terms does not seem to be possible without at least a linear dependence on N. On the other hand, Fact 1 shows that the target function can be approximated for any N using a single full rank attention. We conjecture that such a dependence on N is necessary when using low-rank attention:

Conjecture 6. *There is no multi-layer transformer (with fixed size and weight matrices) of rank $r < d$ that approximates the target of Equation* (3) *for all N.*

That is, while it may be possible to construct a transformer that approximates the target for a given fixed N (as we do above), we conjecture that there is no such construction that is independent of N. Proving or refuting the above conjecture would have very different implications. A counterexample would mean that the the weakness of low-rank can be compensated by depth, and thus the rank does not play a decisive role in the expressive power of multi-layer transformers. A proof would show that, even in the multi-layer case, low-rank attention is fundamentally weaker than high-rank attention.

7 Experiments

In this section, we complement our theoretical results with experiments on a more realistic architectures.[1] We train fully-featured transformer encoders—which include multiple layers of self-attention, MLPs, skip connections, and normalization—on our nearest neighbor target function. Relaxing the assumptions of our theory, we draw $N > 2$ input points uniformly and i.i.d. from $\mathbb{S}^{d-1}$ without constraining them to be orthogonal, and we make no distinction between source and target points. To test our predictions, we modify the attention layers by allowing the number of heads to be different from the standard setting $H = d/r$.

Our first experiment studies the influence of rank on the performance of the transformer for various numbers of heads (H) and layers (L). We fix the dimension $d = 64$ and the number of points $N = 16$. We use three different scaling rules for the number of heads: $H = d^c/r$ for $c \in \{1, 1.5, 2\}$; in other words, we set number of parameters per attention layer to be of order $rdH = d^{c+1}$. In this

[1] Our code is available at `https://github.com/NoahAmsel/attention-formers`.

experiment, we use no positional encodings. Figure 2 plots the results, showing the best of five runs for each setting. Each line corresponds to one choice of c (that is, one choice of scaling rule for the number of heads) so the number of parameters per layer is constant along each line. When $L = 1$, the results suggest that using full-rank ($r = d = 64$) is necessary and sufficient to learn the target function accurately; even $d^2/2$ heads of rank $d/2$ fails. For $L > 1$, the trade-off between rank and accuracy is more favorable, but full-rank attention still significantly outperforms low-rank attention, even using fewer parameters. The best low-rank model we trained ($L = 5, c = 2, r = 32$) performs no better than the worst full-rank model ($L = 1, c = 1, r = 64$) despite having 80x fewer parameters in its attention layers. (Here we are excluding the models with $L = 5, c = 1$, which seem to suffer from optimization difficulties on this problem.) In short, a standard transformer with $H = 1$ (full-rank) performs much better on this task than one with even slightly larger H (lower rank).

In Appendix B.2, we present additional experiments exploring the solution learned by full rank heads (cf. Fact 1), the use of positional encodings (cf. Theorem 5), and the role of the number of points N. In each case, the empirical results support the assumptions and conclusions of our theory. We also give full details of our models and training procedure.

8 Conclusions and Limitations

In this paper, we have investigated the role of rank in attention mechanisms. We question the nearly universal practice of using low-rank attention and setting the number of heads according to $H = d/r$. We show that for a simple and natural target function inspired by semantic search, low-rank attention is fundamentally weaker than full-rank attention, even when $H \gg d/r$. We demonstrate this strict separation between the low-rank and high-rank regimes both theoretically, by proving hardness of approximation in the shallow setting, and empirically, through experiments with standard multilayer transformers. Our results suggest that using a larger rank can improve the expressivity and parameter-efficiency of attention.

That said, our theoretical analysis is inherently limited to the study of shallow transformers, and the results of Section 6 illustrate how adding depth may overcome the limitations of low-rank self-attention in some cases. However, we hope that our findings will motivate theoreticians and practitioners to more carefully consider the settings and scalings of transformer hyperparameters. In particular, they show that if theoreticians assume attention is full-rank, their analyses may fail to accurately describe the transformers used in practice. Moreover, increasing the rank or number of heads beyond their standard settings may have practical benefits that allow SOTA models to be smaller overall. Such benefits may have been obscured by overreliance on heuristics like $H = d/r$, which prevent us from isolating the effects of any one hyperparameter. Clearly, much remains to be understood about the successes and failure modes of attention-based architectures.

Our findings suggest several promising directions for future work. The basic transformer architecture of Vaswani et al. (2017) allows the user to set a variety of hyperparameters. However, transformers have been scaled almost exclusively by increasing the embedding dimension and number of layers, without significantly changing the other hyperparameters (see Table 1). While considerable prior work has studied scaling laws for the dimension and number of layers, we believe that future research should also consider the other hyperparameters and seek to understand the trade-offs, dependencies, and scaling laws that govern them. Here, we focus on the query/key rank and its relationship to the number of heads, but the depth and width of the MLPs and the value/output rank are also of interest.

Additionally, the rotational invariance of the input data distribution is instrumental in establishing our lower bounds. Given the inherently discrete nature of text-based transformers, a natural question is to understand how to generalize our techniques beyond the rotationally-invariant setting. Another direction for future work is to understand the relationship between the rank and the context length. Focusing on the $N = 2$ case suffices for us to prove rank separation, but we believe a similar result should hold at least for all $N \leq d$; Figure 7 provides preliminary experimental evidence. Understanding the $N > 2$ case may also help address a final open question: What is the relationship between rank and depth? In particular, does Conjecture 6 hold?

Acknowledgements: This work was partially supported by the Alfred P. Sloan Foundation, and awards NSF RI-1816753, NSF CAREER CIF 1845360, NSF CHS-1901091 and NSF DMS-MoDL 2134216. We thank Ohad Shamir for useful discussions while this work was being completed.

Reproducibility Statement: Assumptions of all our theoretical results are described in the main text, and complete proofs are given in Appendices C to E. Details of all experiments are given in Appendix B, and our source code is included in the supplementary material and available at `https://github.com/NoahAmsel/attention-formers`. All data used in our experiments in synthetic.

REFERENCES

Kwangjun Ahn, Xiang Cheng, Hadi Daneshmand, and Suvrit Sra. Transformers learn to implement preconditioned gradient descent for in-context learning. In *Proceedings of the 37th International Conference on Neural Information Processing Systems*, NIPS '23, Red Hook, NY, USA, 2023. Curran Associates Inc.

Ekin Akyürek, Dale Schuurmans, Jacob Andreas, Tengyu Ma, and Denny Zhou. What learning algorithm is in-context learning? investigations with linear models. In *The Eleventh International Conference on Learning Representations*, 2023. URL `https://openreview.net/forum?id=0g0X4H8yN4I`.

Francis Bach. Breaking the curse of dimensionality with convex neural networks. *The Journal of Machine Learning Research*, 18(1):629–681, 2017a.

Francis Bach. On the equivalence between kernel quadrature rules and random feature expansions. *Journal of Machine Learning Research*, 18(21):1–38, 2017b.

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. Neural machine translation by jointly learning to align and translate. *CoRR abs/1409.0473*, 2014. doi: abs/1409.0473.

Yu Bai, Fan Chen, Huan Wang, Caiming Xiong, and Song Mei. Transformers as statisticians: Provable in-context learning with in-context algorithm selection. In A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine (eds.), *Advances in Neural Information Processing Systems*, volume 36, pp. 57125–57211. Curran Associates, Inc., 2023. URL `https://proceedings.neurips.cc/paper_files/paper/2023/file/b2e63e36c57e153b9015fece2352a9f9-Paper-Conference.pdf`.

Matan Ben Noach and Yoav Goldberg. Compressing pre-trained language models by matrix decomposition. In Kam-Fai Wong, Kevin Knight, and Hua Wu (eds.), *Proceedings of the 1st Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 10th International Joint Conference on Natural Language Processing*, pp. 884–889, Suzhou, China, December 2020. Association for Computational Linguistics. URL `https://aclanthology.org/2020.aacl-main.88`.

Satwik Bhattamishra, Michael Hahn, Phil Blunsom, and Varun Kanade. Separations in the representational capabilities of transformers and recurrent architectures, 2024. URL `https://arxiv.org/abs/2406.09347`.

Srinadh Bhojanapalli, Chulhee Yun, Ankit Singh Rawat, Sashank Reddi, and Sanjiv Kumar. Low-rank bottleneck in multi-head attention models. In *International conference on machine learning*, pp. 864–873. PMLR, 2020.

Alberto Bietti, Vivien Cabannes, Diane Bouchacourt, Herve Jegou, and Leon Bottou. Birth of a transformer: A memory viewpoint. In A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine (eds.), *Advances in Neural Information Processing Systems*, volume 36, pp. 1560–1588. Curran Associates, Inc., 2023. URL `https://proceedings.neurips.cc/paper_files/paper/2023/file/0561738a239a995c8cd2ef0e50cfa4fd-Paper-Conference.pdf`.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. Language models are few-shot learners, 2020.

Vivien Cabannes, Elvis Dohmatob, and Alberto Bietti. Scaling laws for associative memories. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=Tzh6xAJSll.

Vaggos Chatziafratis, Sai Ganesh Nagarajan, Ioannis Panageas, and Xiao Wang. Depth-width tradeoffs for relu networks via sharkovsky's theorem. *arXiv preprint arXiv:1912.04378*, 2019.

Siyu Chen, Heejune Sheen, Tianhao Wang, and Zhuoran Yang. Training dynamics of multi-head softmax attention for in-context learning: Emergence, convergence, and optimality. *arXiv preprint arXiv:2402.19442*, 2024.

Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann, Parker Schuh, Kensen Shi, Sashank Tsvyashchenko, Joshua Maynez, Abhishek Rao, Parker Barnes, Yi Tay, Noam Shazeer, Vinodkumar Prabhakaran, Emily Reif, Nan Du, Ben Hutchinson, Reiner Pope, James Bradbury, Jacob Austin, Michael Isard, Guy Gur-Ari, Pengcheng Yin, Toju Duke, Anselm Levskaya, Sanjay Ghemawat, Sunipa Dev, Henryk Michalewski, Xavier Garcia, Vedant Misra, Kevin Robinson, Liam Fedus, Denny Zhou, Daphne Ippolito, David Luan, Hyeontaek Lim, Barret Zoph, Alexander Spiridonov, Ryan Sepassi, David Dohan, Shivani Agrawal, Mark Omernick, Andrew M. Dai, Thanumalayan Sankaranarayana Pillai, Marie Pellat, Aitor Lewkowycz, Erica Moreira, Rewon Child, Oleksandr Polozov, Katherine Lee, Zongwei Zhou, Xuezhi Wang, Brennan Saeta, Mark Diaz, Orhan Firat, Michele Catasta, Jason Wei, Kathy Meier-Hellstern, Douglas Eck, Jeff Dean, Slav Petrov, and Noah Fiedel. Palm: scaling language modeling with pathways. *J. Mach. Learn. Res.*, 24(1), mar 2024. ISSN 1532-4435.

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

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

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

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In Jill Burstein, Christy Doran, and Thamar Solorio (eds.), *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pp. 4171–4186, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics. doi: 10.18653/v1/N19-1423. URL https://aclanthology.org/N19-1423.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=YicbFdNTTy.

Benjamin L Edelman, Surbhi Goel, Sham Kakade, and Cyril Zhang. Inductive biases and variable creation in self-attention mechanisms. In *International Conference on Machine Learning*, pp. 5793–5831. PMLR, 2022.

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

Christopher Frye and Costas J Efthimiou. Spherical harmonics in p dimensions. *arXiv preprint arXiv:1205.3548*, 2012.

Hengyu Fu, Tianyu Guo, Yu Bai, and Song Mei. What can a single attention layer learn? a study through the random features lens. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023. URL https://openreview.net/forum?id=wX8GuzDSJR.

Shivam Garg, Dimitris Tsipras, Percy S Liang, and Gregory Valiant. What can transformers learn in-context? a case study of simple function classes. In S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh (eds.), *Advances in Neural Information Processing Systems*, volume 35, pp. 30583–30598. Curran Associates, Inc., 2022. URL https://proceedings.neurips.cc/paper_files/paper/2022/file/c529dba08a146ea8d6cf715ae8930cbe-Paper-Conference.pdf.

Behrooz Ghorbani, Song Mei, Theodor Misiakiewicz, and Andrea Montanari. Linearized two-layers neural networks in high dimension. *The Annals of Statistics*, 49(2):1029 – 1054, 2021. doi: 10.1214/20-AOS1990. URL https://doi.org/10.1214/20-AOS1990.

Yuan Gong, Yu-An Chung, and James Glass. AST: Audio Spectrogram Transformer. In *Proc. Interspeech 2021*, pp. 571–575, 2021. doi: 10.21437/Interspeech.2021-698.

Dirk Groeneveld, Iz Beltagy, Pete Walsh, Akshita Bhagia, Rodney Kinney, Oyvind Tafjord, Ananya Harsh Jha, Hamish Ivison, Ian Magnusson, Yizhong Wang, et al. Olmo: Accelerating the science of language models. *arXiv preprint arXiv:2402.00838*, 2024.

Michael Hahn. Theoretical limitations of self-attention in neural sequence models. *Transactions of the Association for Computational Linguistics*, 8:156–171, 2020.

Habib Hajimolahoseini, Mehdi Rezagholizadeh, Vahid Partovinia, Marzieh Tahaei, Omar Mohamed Awad, and Yang Liu. Compressing pre-trained language models using progressive low rank decomposition. *Advances in Neural Information Processing Systems*, 2021.

Yiding Hao, Dana Angluin, and Robert Frank. Formal language recognition by hard attention transformers: Perspectives from circuit complexity. *Transactions of the Association for Computational Linguistics*, 10:800–810, 2022.

Jordan Hoffmann, Sebastian Borgeaud, Arthur Mensch, Elena Buchatskaya, Trevor Cai, Eliza Rutherford, Diego de Las Casas, Lisa Anne Hendricks, Johannes Welbl, Aidan Clark, Thomas Hennigan, Eric Noland, Katherine Millican, George van den Driessche, Bogdan Damoc, Aurelia Guy, Simon Osindero, Karén Simonyan, Erich Elsen, Oriol Vinyals, Jack Rae, and Laurent Sifre. An empirical analysis of compute-optimal large language model training. In S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh (eds.), *Advances in Neural Information Processing Systems*, volume 35, pp. 30016–30030. Curran Associates, Inc., 2022. URL https://proceedings.neurips.cc/paper_files/paper/2022/file/c1e2faff6f588870935f114ebe04a3e5-Paper-Conference.pdf.

Daniel Hsu, Clayton H Sanford, Rocco Servedio, and Emmanouil Vasileios Vlatakis-Gkaragkounis. On the approximation power of two-layer networks of random relus. In *Conference on Learning Theory*, pp. 2423–2461. PMLR, 2021.

Edward J Hu, yelong shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=nZeVKeeFYf9.

Samy Jelassi, David Brandfonbrener, Sham M Kakade, and Eran Malach. Repeat after me: Transformers are better than state space models at copying. *arXiv preprint arXiv:2402.01032*, 2024.

Tokio Kajitsuka and Issei Sato. Are transformers with one layer self-attention using low-rank weight matrices universal approximators?, 2023.

Pritish Kamath, Omar Montasser, and Nathan Srebro. Approximate is good enough: Probabilistic variants of dimensional and margin complexity. In *Conference on Learning Theory*, pp. 2236–2262. PMLR, 2020.

Junghwan Kim, Michelle Kim, and Barzan Mozafari. Provable memorization capacity of transformers. In *The Eleventh International Conference on Learning Representations*, 2022.

Beatrice Laurent and Pascal Massart. Adaptive estimation of a quadratic functional by model selection. *Annals of statistics*, pp. 1302–1338, 2000.

Benjamin Lefaudeux, Francisco Massa, Diana Liskovich, Wenhan Xiong, Vittorio Caggiano, Sean Naren, Min Xu, Jieru Hu, Marta Tintore, Susan Zhang, Patrick Labatut, Daniel Haziza, Luca Wehrstedt, Jeremy Reizenstein, and Grigory Sizov. xformers: A modular and hackable transformer modelling library. https://github.com/facebookresearch/xformers, 2022.

Valerii Likhosherstov, Krzysztof Choromanski, and Adrian Weller. On the expressive flexibility of self-attention matrices. *Proceedings of the AAAI Conference on Artificial Intelligence*, 37(7): 8773–8781, Jun. 2023. doi: 10.1609/aaai.v37i7.26055. URL https://ojs.aaai.org/index.php/AAAI/article/view/26055.

Bingbin Liu, Jordan T Ash, Surbhi Goel, Akshay Krishnamurthy, and Cyril Zhang. Transformers learn shortcuts to automata. *arXiv preprint arXiv:2210.10749*, 2022.

Xiuqing Lv, Peng Zhang, Sunzhu Li, Guobing Gan, and Yueheng Sun. LightFormer: Light-weight transformer using SVD-based weight transfer and parameter sharing. In Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki (eds.), *Findings of the Association for Computational Linguistics: ACL 2023*, pp. 10323–10335, Toronto, Canada, July 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.findings-acl.656. URL https://aclanthology.org/2023.findings-acl.656.

Sadegh Mahdavi, Renjie Liao, and Christos Thrampoulidis. Memorization capacity of multi-head attention in transformers. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=MrR3rMxqqv.

William Merrill and Ashish Sabharwal. The expresssive power of transformers with chain of thought. *arXiv preprint arXiv:2310.07923*, 2023.

William Merrill, Ashish Sabharwal, and Noah A Smith. Saturated transformers are constant-depth threshold circuits. *Transactions of the Association for Computational Linguistics*, 10:843–856, 2022.

Theodor Misiakiewicz and Andrea Montanari. Six lectures on linearized neural networks. *arXiv preprint arXiv:2308.13431*, 2023.

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. Pytorch: An imperative style, high-performance deep learning library. *Advances in neural information processing systems*, 32, 2019.

Ofir Press, Noah Smith, and Mike Lewis. Train short, test long: Attention with linear biases enables input length extrapolation. In *International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=R8sQPpGCv0.

Alec Radford, Karthik Narasimhan, Tim Salimans, Ilya Sutskever, et al. Improving language understanding by generative pre-training, 2018.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9, 2019.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In Marina Meila and Tong Zhang (eds.), *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 8748–8763. PMLR, 18–24 Jul 2021. URL https://proceedings.mlr.press/v139/radford21a.html.

Jack W Rae, Sebastian Borgeaud, Trevor Cai, Katie Millican, Jordan Hoffmann, Francis Song, John Aslanides, Sarah Henderson, Roman Ring, Susannah Young, et al. Scaling language models: Methods, analysis & insights from training gopher. *arXiv preprint arXiv:2112.11446*, 2021.

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

Itay Safran, Ronen Eldan, and Ohad Shamir. Depth separations in neural networks: what is actually being separated? In *Conference on Learning Theory*, pp. 2664–2666. PMLR, 2019.

Clayton Sanford, Daniel Hsu, and Matus Telgarsky. Transformers, parallel computation, and logarithmic depth, 2024a.

Clayton Sanford, Daniel J Hsu, and Matus Telgarsky. Representational strengths and limitations of transformers. *Advances in Neural Information Processing Systems*, 36, 2024b.

Ohad Shamir. Distribution-specific hardness of learning neural networks. *The Journal of Machine Learning Research*, 19(1):1135–1163, 2018.

Pratyusha Sharma, Jordan T. Ash, and Dipendra Misra. The truth is in there: Improving reasoning in language models with layer-selective rank reduction. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=ozX92bu8VA.

Lena Strobl, William Merrill, Gail Weiss, David Chiang, and Dana Angluin. What Formal Languages Can Transformers Express? A Survey. *Transactions of the Association for Computational Linguistics*, 12:543–561, 05 2024. ISSN 2307-387X. doi: 10.1162/tacl_a_00663. URL https://doi.org/10.1162/tacl_a_00663.

Jianlin Su, Murtadha Ahmed, Yu Lu, Shengfeng Pan, Wen Bo, and Yunfeng Liu. Roformer: Enhanced transformer with rotary position embedding. *Neurocomputing*, 568:127063, 2024.

Matus Telgarsky. Benefits of depth in neural networks. In Vitaly Feldman, Alexander Rakhlin, and Ohad Shamir (eds.), *29th Annual Conference on Learning Theory*, volume 49 of *Proceedings of Machine Learning Research*, pp. 1517–1539, Columbia University, New York, New York, USA, 23–26 Jun 2016. PMLR. URL https://proceedings.mlr.press/v49/telgarsky16.html.

Romal Thoppilan, Daniel De Freitas, Jamie Hall, Noam Shazeer, Apoorv Kulshreshtha, Heng-Tze Cheng, Alicia Jin, Taylor Bos, Leslie Baker, Yu Du, YaGuang Li, Hongrae Lee, Huaixiu Steven Zheng, Amin Ghafouri, Marcelo Menegali, Yanping Huang, Maxim Krikun, Dmitry Lepikhin, James Qin, Dehao Chen, Yuanzhong Xu, Zhifeng Chen, Adam Roberts, Maarten Bosma, Vincent Zhao, Yanqi Zhou, Chung-Ching Chang, Igor Krivokon, Will Rusch, Marc Pickett, Pranesh Srinivasan, Laichee Man, Kathleen Meier-Hellstern, Meredith Ringel Morris, Tulsee Doshi, Renelito Delos Santos, Toju Duke, Johnny Soraker, Ben Zevenbergen, Vinodkumar Prabhakaran, Mark Diaz, Ben Hutchinson, Kristen Olson, Alejandra Molina, Erin Hoffman-John, Josh Lee, Lora Aroyo, Ravi Rajakumar, Alena Butryna, Matthew Lamm, Viktoriya Kuzmina, Joe Fenton, Aaron Cohen, Rachel Bernstein, Ray Kurzweil, Blaise Aguera-Arcas, Claire Cui, Marian Croak, Ed Chi, and Quoc Le. Lamda: Language models for dialog applications, 2022.

Yuandong Tian, Yiping Wang, Beidi Chen, and Simon S Du. Scan and snap: Understanding training dynamics and token composition in 1-layer transformer. *Advances in Neural Information Processing Systems*, 36:71911–71947, 2023.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*, 2023a.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*, 2023b.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. Attention is all you need. In I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett (eds.), *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017. URL https://proceedings.neurips.cc/paper_files/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf.

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

Roman Vershynin. *High-dimensional probability: An introduction with applications in data science*, volume 47. Cambridge university press, 2018.

Johannes Von Oswald, Eyvind Niklasson, Ettore Randazzo, João Sacramento, Alexander Mordvintsev, Andrey Zhmoginov, and Max Vladymyrov. Transformers learn in-context by gradient descent. In *Proceedings of the 40th International Conference on Machine Learning*, ICML'23. JMLR.org, 2023.

Colin Wei, Yining Chen, and Tengyu Ma. Statistically meaningful approximation: a case study on approximating turing machines with transformers. *Advances in Neural Information Processing Systems*, 35:12071–12083, 2022.

Kelvin Xu, Jimmy Ba, Ryan Kiros, Kyunghyun Cho, Aaron Courville, Ruslan Salakhudinov, Rich Zemel, and Yoshua Bengio. Show, attend and tell: Neural image caption generation with visual attention. In Francis Bach and David Blei (eds.), *Proceedings of the 32nd International Conference on Machine Learning*, volume 37 of *Proceedings of Machine Learning Research*, pp. 2048–2057, Lille, France, 07–09 Jul 2015. PMLR. URL https://proceedings.mlr.press/v37/xuc15.html.

Greg Yang, Edward J. Hu, Igor Babuschkin, Szymon Sidor, Xiaodong Liu, David Farhi, Nick Ryder, Jakub Pachocki, Weizhu Chen, and Jianfeng Gao. Tensor programs v: Tuning large neural networks via zero-shot hyperparameter transfer, 2022. URL https://arxiv.org/abs/2203.03466.

Gilad Yehudai and Ohad Shamir. On the power and limitations of random features for understanding neural networks. *Advances in Neural Information Processing Systems*, 32, 2019.

Chulhee Yun, Srinadh Bhojanapalli, Ankit Singh Rawat, Sashank J Reddi, and Sanjiv Kumar. Are transformers universal approximators of sequence-to-sequence functions? *arXiv preprint arXiv:1912.10077*, 2019.

Biao Zhang and Rico Sennrich. Root mean square layer normalization. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett (eds.), *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019. URL https://proceedings.neurips.cc/paper_files/paper/2019/file/1e8a19426224ca89e83cef47f1e7f53b-Paper.pdf.

Ruiqi Zhang, Spencer Frei, and Peter L. Bartlett. Trained transformers learn linear models in-context. *Journal of Machine Learning Research*, 25(49):1–55, 2024. URL http://jmlr.org/papers/v25/23-1042.html.

A Hyperparameters of Transformer

The transformer architecture (Vaswani et al., 2017) leaves the user free to set the following hyperparameters:

- The embedding dimension (d)
- The number of layers (L)
- The width of the MLPs (w)
- The depth of the MLPs (D)

- The rank of the $\boldsymbol{W}_Q$ and $\boldsymbol{W}_K$ matrices for each head (r)
- The rank of the $\boldsymbol{W}_V$ and $\boldsymbol{W}_O$ matrices for each head (r_2)
- The number of attention heads in each layer (H)

In this paper, we consider the dimension d to be given by the domain of the target function, rather than being a hyperparameter as in language modeling. As Table 1 shows, only d and L have been significantly changed relative to the original model. For all models of which we are aware, w lies within a factor of two from Vaswani et al. (2017), r lies within a factor of four, and D and r_2 are not changed at all. H has been scaled, but always according to the standard scaling (up to a factor of 2).

Table 1: Hyperparameter settings of popular transformer models (largest versions reported). Except for d and L, they are strikingly consistent. See text of Appendix A for notation.

Year	Model	d	L	w	D	r	r_2	H
2017	Attention is all you need (Vaswani et al., 2017)	512	6	$4d$	2	64	r	d/r
2018	GPT, GPT-2 (Radford et al., 2018; 2019)	768	12	$4d$	2	64	r	d/r
2019	Bert-Large (Devlin et al., 2019)	1,024	24	$4d$	2	64	r	d/r
2020	GPT-3 (Brown et al., 2020)	12,288	96	$4d$	2	128	r	d/r
2021	ViT-Huge (Dosovitskiy et al., 2021)	1,280	32	$4d$	2	80	r	d/r
	CLIP (text encoder) (Radford et al., 2021)	768	12	$4d$	2	64	r	d/r
	CLIP (image encoder) (Radford et al., 2021)	1,024	24	$4d$	2	64	r	d/r
	Jurassic-1	13,824	76	$4d$	2	144	r	d/r
	Gopher 280B (Rae et al., 2021)	16,384	80	$4d$	2	128	r	d/r
	AST (audio) (Gong et al., 2021)	1,024	24	$4d$	2	64	r	d/r
	LaMDA (Thoppilan et al., 2022)	8,192	64	$8d$	2	128	r	$2d/r$
2022	Chinchilla 70B (Hoffmann et al., 2022)	8,192	80	$4d$	2	128	r	d/r
2023	PaLM (Chowdhery et al., 2024)	18,432	118	$4d$	2	256	r	$2d/3r$
	LLaMA, Llama-2 (Touvron et al., 2023a;b)	8,192	80	$8d/3$	2	128	r	d/r
2024	OLMo (Groeneveld et al., 2024)	8,192	80	$8d/3$	2	128	r	d/r

B Additional Experiments and Details

B.1 In-Context Learning of Linear Functions

We reproduce a striking experiment from Garg et al. (2022) that demonstrates the ability of transformers to perform in-context learning. In this task, the prompt is a sequence of input-output pairs from a randomly selected function it has not seen before, plus one additional input point. The model is trained to predict the output corresponding to this final point. That is, given the sequence

$$\boldsymbol{x}_1, f(\boldsymbol{x}_1), \ldots, \boldsymbol{x}_N, f(\boldsymbol{x}_N), \boldsymbol{x}_{\text{query}} \tag{12}$$

the desired output is $f(\boldsymbol{x}_{\text{query}})$. In the linear version of this task, $f(\boldsymbol{x}) = \boldsymbol{w}^\top \boldsymbol{x}$. The input points $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_N$, the query point $\boldsymbol{x}_{\text{query}}$, and the target weight $\boldsymbol{w}$ are all drawn from the isotropic Gaussian distribution in 20 dimensions. The model consists of a linear embedding layer mapping from $\mathbb{R}^{20}$ to $\mathbb{R}^d$, a 12-layer transformer, and a linear unembedding layer from $\mathbb{R}^d$ to $\mathbb{R}$.

We use this experiment as a benchmark to demonstrate the importance of rank in attention-based models, specifically the consequences of using a rank that is too small for the task. We run the exact code of Garg et al. (2022), changing nothing in the data, model, or training procedure except for the embedding dimension d and the number of heads H. As is standard, the rank is given by $r = d/H$. The original paper uses a model with embedding dimension $d = 256$, which is quite large relative to the intrinsic dimension of the inputs, and $H = 8, r = 32$. We show that we can still achieve good performance with much smaller models, such as $d = 48$, but *only* if the rank is sufficiently large. Figure 3 plots the performance of these models. Figure 1 shows the same data, but focusing only on the $d = 48, N = 40$ case. Across different embedding dimensions, the high-rank models ($H = 1$ or 2) significantly outperform the low-rank ones ($H = 8$ or 16).

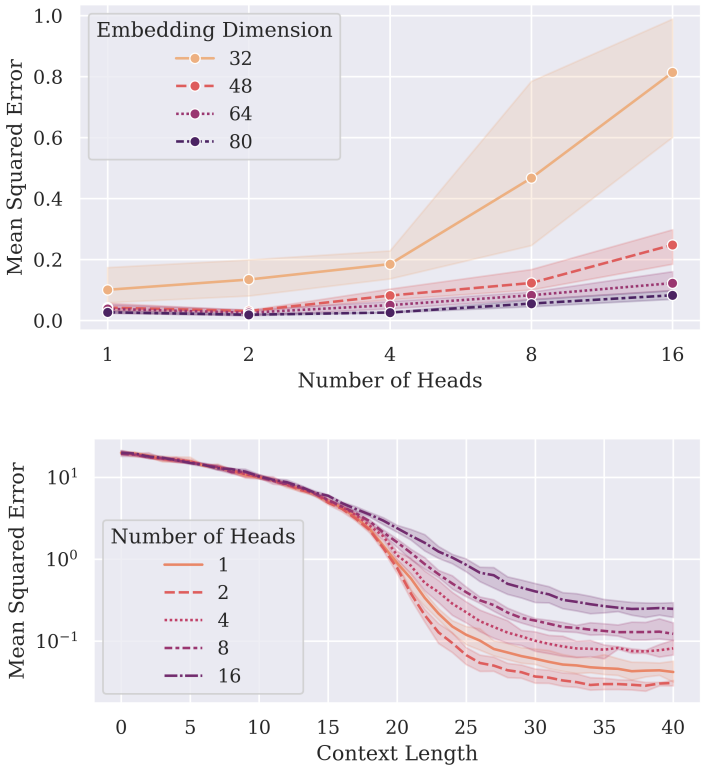

Figure 3: Performance of 12-layer transformers on in-context learning of linear functions in 20 dimensions. Top panel shows the error on a prompt of $N = 40$ input-output pairs for various embedding dimensions d. Bottom panel shows $d = 48$ given various context-lengths N. Across different choices of embedding dimension and context length, performance is significantly improved by reducing the number of attention heads H. Due to the standard scaling rule $r = d/H$, this is equivalent to increasing the attention rank.

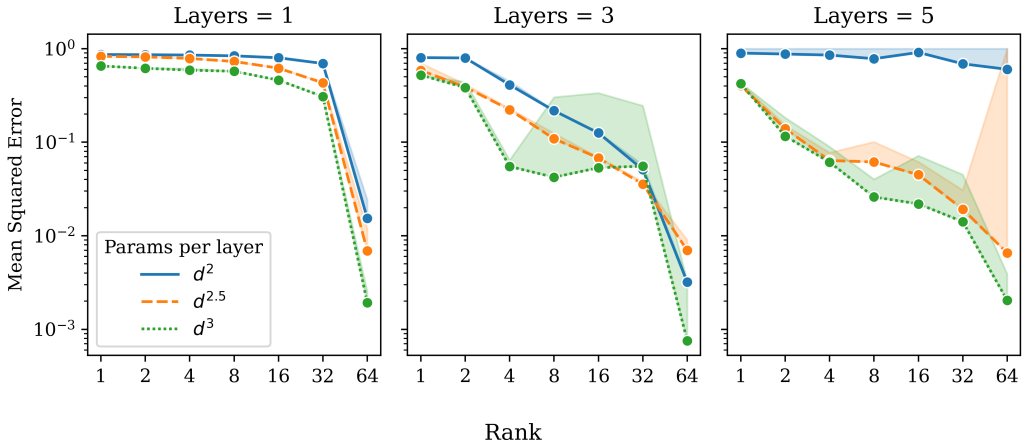

Figure 4: Copy of Figure 2 with the addition of a shaded region showing the range over five runs. With a few exceptions, these ranges are quite narrow.

B.2 Nearest Neighbor Task

In this section, we provide details about the models and training procedure used for the experiments in Section 7 and present additional experiments on our nearest neighbor target function.

Model and training details We use the PyTorch implementation of transformer encoders (Paszke et al., 2019) with two modifications. First, we generalize the standard scaling $H = d/r$, allowing H to be any multiple of d/r. (In particular, we try $H = d^{1.5}/r$ and $H = d^2/r$.) Second, we replace the layer normalization with RMSNorm (Zhang & Sennrich, 2019), a standard choice in modern transformers (Touvron et al., 2023a; Chowdhery et al., 2024) that is also better suited to our target function. We train with biases, but preliminary experiments showed that these make little difference.[2] We run each experiment on a single Nvidia GPU (usually a V100) for no more than a few hours.

Since we are using self-attention, there is no distinction between the source and target points. The N input points are drawn uniformly and i.i.d. from $\mathbb{S}^{d-1}$, and they are not constrained to be orthogonal. We change our target function accordingly. For each input point, the target now outputs whichever of the other points is *farthest* from it. We output the farthest instead of the nearest point because otherwise, each point would map to itself. The loss function is the average mean squared error over the N points. We do not use any attention mask. In particular, we allow points to attend to themselves. Our dataset is synthetic, so we train and test on a stream of freshly generated samples that never repeat. We train on 10^5 batches of size 256 each. For all experiments, we use AdamW with the same learning rate of 0.01 and a learning rate schedule of cosine annealing with a linear warm-up.

Statistical consistency of main experiment Our focus in this paper is on representational capacity rather than learnability or generalization. However, for completeness, we reproduce Figure 2 with the addition of a shaded region showing the range over five runs (Figure 4). In most cases the results are strikingly consistent between runs. Optimization difficulties were observed for the regimes $L = 3$, params $= d^3$ and $L = 5$, params $= d^2$. In the first case, we mostly overcame these by performing three additional runs each (so eight instead of five) for the settings $r = 16$ and $r = 32$.

Full-rank solution In the full-rank case the transformer learns the target, but what representation has it learned? Figure 5 suggests that, in some cases, it is very nearly the construction of Fact 1.

[2]Note that biases in the key, query, and value transformations have a different role from additive positional encodings. These biases differ between heads but are constant across tokens; in contrast, the positional encodings differ between tokens but not heads. The biases implemented by PyTorch are also slightly different from those studied in Section 5.

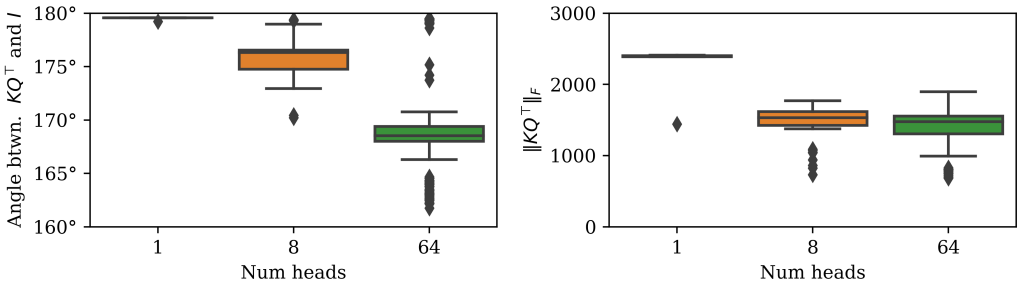

Figure 5: Properties of learned $\boldsymbol{KQ}^\top$ matrices for full-rank models with one layer. Boxplots show distribution over heads from five runs, each on a model which has between 1 and 64 full-rank heads. Left panel plots Frobenius angle with the identity: $\arccos\left(\langle \boldsymbol{KQ}^\top, \boldsymbol{I}\rangle_F / (\|\boldsymbol{KQ}^\top\|_F \|\boldsymbol{I}\|_F)\right)$. Results show that $\boldsymbol{KQ}^\top$ nearly equals $-c\boldsymbol{I}$ for $c > 1000$ in all cases.

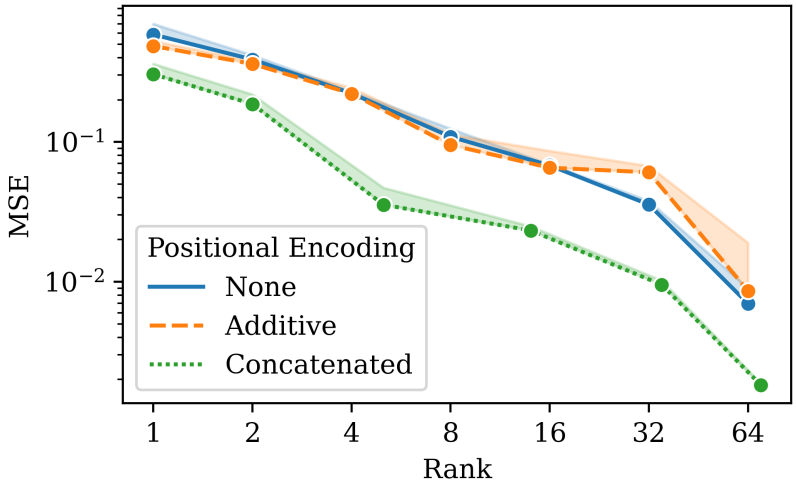

Figure 6: Standard 3-layer transformers with positional encodings on the the farthest neighbor task ($d = 64$, $H = 512/r$, $N = 16$). Line shows best of five runs; shaded region shows range over five runs. Concatenating 6-dimensional learned absolute positional encodings to the input points improves performance somewhat, but does not ameliorate the low-rank bottleneck. Additive positional encodings do not improve performance because they do not increase the embedding dimension.

Recall that in Fact 1, we use a hardmax attention head with $\boldsymbol{K}_h\boldsymbol{Q}_h^\top = \boldsymbol{I}$. In our experiments however, we use the farthest neighbor target function and softmax heads, so the corresponding construction is $\boldsymbol{K}_h\boldsymbol{Q}_h^\top = -c\boldsymbol{I}$ for $c \gg 1$. The first panel shows the median Frobenius angle between the matrices $\boldsymbol{K}_h\boldsymbol{Q}_h^\top$ and $\boldsymbol{I}$ learned by the full-rank, single layer models in the previous experiment. This shows that $\boldsymbol{K}_h\boldsymbol{Q}_h^\top$ very nearly equals $-\boldsymbol{I}$ up to a constant factor. Moreover, as the second panel shows, the norm of this matrix is large, which causes the softmax to act like a hardmax. Results are similar for three layer networks with a single full-rank head, but when $L > 1$ and $H > 1$, it seems the network learns some other, less interpretable strategy to represent the target.

Positional encodings Since our target function is permutation-invariant, no positional information exists in the data. However, in Section 6, we showed that concatenated positional encodings can help low-rank attention succeed when $L > 1$ by giving the model extra dimensions of scratch space. The positional encoding schemes used in practice, like additive encodings (Vaswani et al., 2017), RoPE (Su et al., 2024) and ALiBi (Press et al., 2022), cannot be used in this way, being versions of the generalized attention heads studied in this paper. In Figure 6, we experiment with positional

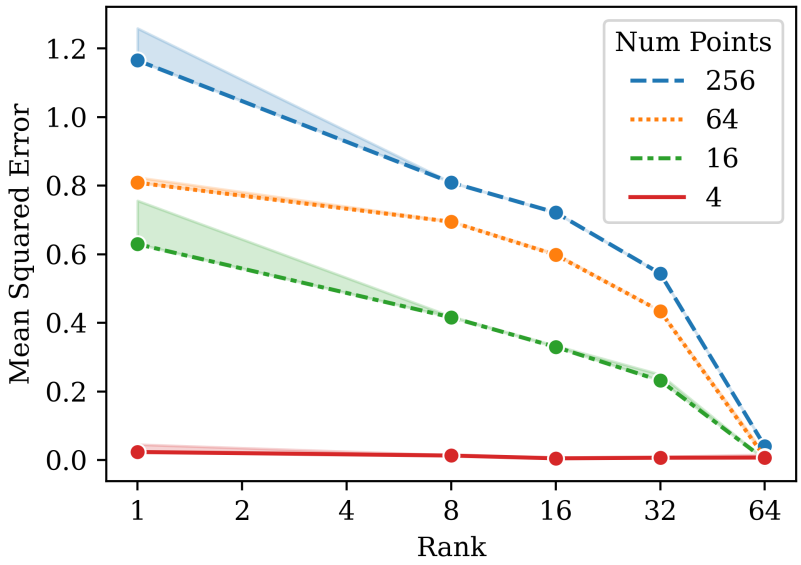

Figure 7: Effect of the number of points (N) on the difficulty of learning the farthest neighbor function. Full-rank attention learns an accurate representation across many Ns, but the performance of low-rank attention degrades as N grows. Dimension is 64. All models have two layers with $H = d^2/r$ heads each. Line shows best of five runs; shaded region shows range over five runs.

encodings. As expected, additive encodings fail to improve performance at all. Concatenated 6-dimensional positional encodings help a little bit, since the transformer is now larger ($d = 70$ instead of 64) and has more parameters. However, unlike in Theorem 5, these encodings do not mitigate the low-rank bottleneck in practice, even for 3-layer transformer. Thus, the low-rank construction of Theorem 5 is primarily of theoretical interest.

Role of N In Figure 7, we explore how the number of input points N affects the difficulty of learning the target function. We fix $d = 64$, $H = d^2/r$, and the number of layer $L = 2$. The results show that, as predicted by Fact 1, the full-rank heads learn the target accurately across a range of N. However, the low-rank heads suffer declining accuracy as N grows. This accords with Conjecture 6, which predicts that low-rank transformers of a fixed size fail to accurately represent the target for sufficiently large N.

C Proofs from Section 4

In this section, we prove the upper bound Fact 1, the lower bound Theorem 2 and some important properties relating to the approximation of the target by random heads.

We begin with the proof of Fact 1 in Appendix C.1. In Appendix C.2, we review the basics of spherical harmonics and describe the corresponding family of ultraspherical orthogonal polynomials on the interval. In Appendix C.3, we construct a basis for functions of pairs of points on the sphere that we will use to analyze the target and the attention mechanism. In Appendix C.4, we show how to expand the target function in this basis, proving the critical properties of slow spectral decay and rotational invariance between basis elements of the same degree. In Appendix C.5, we expand a single attention head in this basis, showing that the number of basis elements with which it is correlated is limited by the rank of the attention head. In Appendix C.6, we use these results to obtain a lower bound on the error of approximation that depends only on certain universal constants related to the spherical harmonics, particularly the number of spherical harmonics of a given degree and the coefficients of the ultraspherical expansion of the sign function. In Appendix C.7, we analyze this expression to derive a bound on the necessary number of heads that depends only on the

dimension d, the rank r, and the error level ϵ. Finally, in Appendix C.8, we analyze a construction that approximates the target function using random rank-1 heads.

C.1 Proof of Fact 1

Let $\epsilon > 0$. We set $\boldsymbol{V} = \boldsymbol{I}$, $\boldsymbol{K}\boldsymbol{Q}^\top = \alpha \boldsymbol{I}$ for $\alpha > 0$ to be chosen later. Since $\boldsymbol{x}_i, \boldsymbol{y} \sim \mathrm{Unif}(\mathbb{S}^{d-1})$, for every $i \in \{1, \dots, N\}$, there exists $\delta > 0$ (which depends on ϵ) such that for the set:

$$A_\delta := \{(\boldsymbol{x}_1, \dots, \boldsymbol{x}_N, \boldsymbol{y}) \in (\mathbb{S}^{d-1})^{N+1} : \forall i \neq j,\ |(\boldsymbol{x}_i - \boldsymbol{x}_j)^\top \boldsymbol{y}| > \delta\}\,, \tag{13}$$

we have that $\Pr((\boldsymbol{x}_1, , \dots, \boldsymbol{x}_N, \boldsymbol{y}) \notin A_\delta) \leq \frac{\epsilon}{2}$. Note that:

$$\boldsymbol{X}\,\mathrm{sm}(\alpha \boldsymbol{X}^\top \boldsymbol{y}) \underset{\alpha \to \infty}{\longrightarrow} \arg\max_{\boldsymbol{x}_i} (\boldsymbol{x}_i^\top \boldsymbol{y}) = \arg\max_{\boldsymbol{x}_i} \|\boldsymbol{x}_i - \boldsymbol{y}\|^2\,, \tag{14}$$

where the convergence is uniform on A_δ, and the equality follows since all the vectors are from the unit sphere. In particular, there exists $\alpha > 0$ such that:

$$\sup_{(\boldsymbol{x}_1, \dots, \boldsymbol{x}_N, \boldsymbol{y}) \in A_\delta} \left\| \boldsymbol{X}\,\mathrm{sm}(\alpha \boldsymbol{X}^\top \boldsymbol{y}) - \arg\max_{\boldsymbol{x}_i} \|\boldsymbol{x}_i - \boldsymbol{y}\|^2 \right\|^2 \leq \frac{\epsilon}{2}\,. \tag{15}$$

Combining both bounds and taking expectation over the vectors finishes the proof.

C.2 Spherical Harmonics

We begin by reviewing some basic results from the theory of spherical harmonics. Let $\tau(\cdot)$ denote the uniform distribution over $\mathbb{S}^{d-1}$ and define the inner product $\langle \cdot, \cdot \rangle_\tau$ over $L^2(\mathbb{S}^{d-1})$ as follows

$$\langle f, g \rangle_\tau := \int_{\mathbb{S}^{d-1}} f(\boldsymbol{x}) g(\boldsymbol{x}) d\tau(\boldsymbol{x}) \tag{16}$$

A polynomial $H : \mathbb{R}^d \to \mathbb{R}$ is called harmonic and degree-ℓ homogeneous if

$$\nabla^2 H = 0, \qquad H(a\boldsymbol{x}) = a^\ell H(\boldsymbol{x}) \tag{17}$$

A spherical harmonic of degree ℓ is the restriction of a harmonic homogeneous polynomial to the sphere $\mathbb{S}^{d-1}$. That is, a function $Y : \mathbb{S}^{d-1} \to \mathbb{R}$ is a spherical harmonic of degree ℓ if and only if the $\mathbb{R}^d \to \mathbb{R}$ function defined by

$$\boldsymbol{x} \mapsto \|\boldsymbol{x}\|^\ell Y\left(\frac{\boldsymbol{x}}{\|\boldsymbol{x}\|}\right) \tag{18}$$

is a harmonic homogeneous polynomial of degree ℓ. The set of spherical harmonics of degree ℓ on $\mathbb{S}^{d-1}$ form a function space $\mathcal{F}_\ell \subset L^2(\mathbb{S}^{d-1})$. These subspaces have the following dimensions (Theorem 4.4 of Frye & Efthimiou (2012)):

$$N(d, \ell) := \dim \mathcal{F}_\ell = \frac{2\ell + d - 2}{\ell} \binom{\ell + d - 3}{\ell - 1}\,. \tag{19}$$

The reason spherical harmonics are so useful is that the $\mathcal{F}_\ell$ are linearly independent, and their direct sum is $L^2(\mathbb{S}^{d-1})$. That is, if $\{Y_\ell^j\}_{j=1}^{N(d,\ell)}$ is an orthonormal basis of $\mathcal{F}_\ell$, then $\cup_{\ell=0}^\infty \{Y_\ell^j\}_{j=1}^{N(d,\ell)}$ is an orthonormal basis of $L^2(\mathbb{S}^{d-1})$ with respect to $\langle \cdot, \cdot \rangle_\tau$.

For a unit vector $\boldsymbol{e}$, let u_d denote the distribution of $\boldsymbol{x}^\top \boldsymbol{e}$ when $\boldsymbol{x} \sim \tau$. Then for $t \in [-1, 1]$,

$$u_d(t) := \frac{A_{d-2}}{A_{d-1}} \cdot (1 - t^2)^{\frac{d-3}{2}} \tag{20}$$

where A_{d-1} is the surface area of $\mathbb{S}^{d-1}$ (see Lemma 4.17 of Frye & Efthimiou (2012)). Define the following inner product over functions mapping $[-1, 1] \to \mathbb{R}$:

$$\langle f, g \rangle_{u_d} := \int_{-1}^{1} f(t) g(t) u_d(t) dt \tag{21}$$

The ultraspherical polynomials $P_\ell : [-1, 1] \to \mathbb{R}$ for $\ell \in \mathbb{N}_{\geq 0}$ are defined by the following properties:

1. P_ℓ has degree ℓ
2. $\ell \neq \ell' \iff \langle P_\ell, P_{\ell'} \rangle_{u_d} = 0$
3. $P_\ell(1) = 1$

These polynomials form an orthogonal basis for $L^2([-1,1], u_d)$, which includes all bounded functions on $[-1,1]$. Moreover, they are intimately connected to the spherical harmonics. We exploit three such connections. First (Equation 4.30 of Frye & Efthimiou (2012))

$$\|P_\ell\|_{u_d}^2 = \frac{1}{N(d,\ell)} \tag{22}$$

Second, the addition formula states that each ultraspherical polynomial can be expressed in terms of the spherical harmonics of the same degree and vice versa (Theorem 4.11[3] of Frye & Efthimiou (2012))

$$P_\ell(\boldsymbol{x}^\top \boldsymbol{y}) = \frac{1}{N(d,\ell)} \sum_{j=1}^{N(d,\ell)} Y_\ell^j(\boldsymbol{x}) Y_\ell^j(\boldsymbol{y}) \tag{23}$$

Finally, the Hecke-Funk formula (Theorem 4.24 of Frye & Efthimiou (2012)) gives the relationship between the ultraspherical expansion of $t \mapsto f(t)$ and the spherical harmonic expansion of $\boldsymbol{y} \mapsto f(\boldsymbol{x}^\top \boldsymbol{y})$. For any degree-$\ell$ spherical harmonic Y_ℓ,

$$\left\langle f\big(\langle \boldsymbol{x}, \cdot \rangle\big), Y_\ell \right\rangle_\tau := \int_{\mathbb{S}^{d-1}} f(\boldsymbol{x}^\top \boldsymbol{y}) Y_\ell(\boldsymbol{y}) d\tau(\boldsymbol{y}) = Y_\ell(\boldsymbol{x}) \langle f, P_\ell \rangle_{u_d} \tag{24}$$

We will make use of the ultraspherical expansion of two particular functions:

Definition 7. *Let $\{\alpha_\ell\}$ be the ultraspherical series for* arcsin *and let $\{\eta_\ell\}$ be the ultraspherical series for* sign*. That is,*

$$\arcsin(t) = \sum_{\ell=0}^{\infty} \alpha_\ell \frac{P_\ell(t)}{\|P_\ell\|_{u_d}} \tag{25}$$

$$\operatorname{sign}(t) = \sum_{\ell=0}^{\infty} \eta_\ell \frac{P_\ell(t)}{\|P_\ell\|_{u_d}} \qquad \forall t \in [-1,1] \tag{26}$$

C.3 Orthonormal Basis for Target and Attention Heads

The goal of this section is to define the orthonormal basis that we will use to analyze the (surrogate) target and attention functions. We define the input space for these functions as follows: $\mathcal{X} = \mathbb{S}^{d-1} \times \mathbb{S}^{d-1}$. We denote elements of this set by $(\boldsymbol{x}, \boldsymbol{y})$ or z for short. For any two functions, define their tensorization by

$$(f \otimes g)(z) = f(\boldsymbol{x}) f(\boldsymbol{y}) \tag{27}$$

We let $\bar{\tau} = \tau \otimes \tau$ be the uniform measure on $\mathcal{X}$. We also define a feature space $\Omega = \mathbb{S}^{d-1} \times \mathbb{S}^{d-1}$ and denote elements of this space by $(\boldsymbol{q}, \boldsymbol{k})$ or ω. Of course, $\Omega = \mathcal{X}$, but since they are used in different contexts, we use separate notation for readability.

We define the feature mapping that we will use to analyze the surrogate target and attention functions:

Definition 8. *Define the "rank-1 head" function $\rho : \mathcal{X} \times \Omega \to \{\pm 1\}$ by*

$$\rho(z, \omega) := \operatorname{sign}\left(\boldsymbol{x}^\top \boldsymbol{k} \boldsymbol{q}^\top \boldsymbol{y}\right) \tag{28}$$

and the feature map linear operator $\mathcal{T} : L^1(\Omega) \to L^2(\mathcal{X})$ by

$$(\mathcal{T}u)(z) := \int_\Omega \rho(z, \omega) u(\omega) d\bar{\tau}(\omega) \tag{29}$$

[3]Note that Frye & Efthimiou (2012) has an extra factor of A_{d-1} in the theorem statement. This is because they use a different normalization for the spherical harmonics.

The intuition is as follows. For a fixed value of $\omega = (\boldsymbol{k}, \boldsymbol{q})$, the function $\rho(\cdot, \omega)$ acts like a hardmax attention head with rank 1. More precisely, if $\boldsymbol{x} = \boldsymbol{x}_1 - \boldsymbol{x}_2$ and $\boldsymbol{V} = \boldsymbol{I}$, then $\rho(z, \omega)$ is the output of the head applied to the source $\boldsymbol{y}$ and targets $\boldsymbol{x}_1$ and $\boldsymbol{x}_2$, projected onto $\boldsymbol{x}$. Furthermore, $\mathcal{T}u$ is a weighted linear combination of all possible rank-1 hardmax heads.

We will construct a basis using functions of the form $\mathcal{T}(Y \otimes Y')$ for spherical harmonics Y and Y'. The rationale for choosing this basis is as follows. $\mathcal{T}$ defines a positive semidefinite operator $\mathcal{T}^*\mathcal{T} : L^1(\Omega) \to L^2(\Omega)$, which is described by the following formula:

$$(\mathcal{T}^*\mathcal{T}u)(\omega) = \int_\Omega \mathop{\mathbb{E}}_{z\sim\bar{\tau}}[\rho(z,\omega)\rho(z,\omega')] \cdot u(\omega')d\bar{\tau}(\omega') \tag{30}$$

Functions of the form $Y \otimes Y'$ will turn out to be eigenfunctions of this operator. To see why, we must first analyze the kernel $\mathbb{E}_{z\sim\bar{\tau}}[\rho(z,\omega)\rho(z,\omega')]$, which we do in the following lemma.

Lemma 9.

$$\mathop{\mathbb{E}}_{z\sim\bar{\tau}}[\rho(z,\omega)\rho(z,\omega')] = \frac{4}{\pi^2}\arcsin(\boldsymbol{q}^\top\boldsymbol{q}')\arcsin(\boldsymbol{k}^\top\boldsymbol{k}') \tag{31}$$

Proof. To begin, we compute a closely related property – the probability that the signs are equal:

$$\Pr_{\boldsymbol{z}\sim\bar{\tau}}[\rho(\boldsymbol{z},\omega) = \rho(\boldsymbol{z},\omega')] = \Pr_{\boldsymbol{z}\sim\bar{\tau}}[\langle\boldsymbol{x},\boldsymbol{k}\rangle\langle\boldsymbol{q},\boldsymbol{y}\rangle\langle\boldsymbol{x},\boldsymbol{k}'\rangle\langle\boldsymbol{q}',\boldsymbol{y}\rangle > 0] \tag{32}$$

$$\tag{33}$$

Let θ be the angle between $\boldsymbol{q}$ and $\boldsymbol{q}'$ and let ϕ be the angle between $\boldsymbol{k}$ and $\boldsymbol{k}'$. We have

$$\Pr_{\boldsymbol{y}}[\langle\boldsymbol{y},\boldsymbol{q}\rangle\langle\boldsymbol{y},\boldsymbol{q}'\rangle \geq 0] = 1 - \frac{\theta}{\pi} \tag{34}$$

$$\Pr_{\boldsymbol{x}}[\langle\boldsymbol{x},\boldsymbol{k}\rangle\langle\boldsymbol{x},\boldsymbol{k}'\rangle \geq 0] = 1 - \frac{\phi}{\pi} \tag{35}$$

$$\Pr_{\boldsymbol{x},\boldsymbol{y}}[\langle\boldsymbol{y},\boldsymbol{q}\rangle\langle\boldsymbol{y},\boldsymbol{q}'\rangle \geq 0 \wedge \langle\boldsymbol{x},\boldsymbol{k}\rangle\langle\boldsymbol{x},\boldsymbol{k}'\rangle \geq 0] = \left(1-\frac{\theta}{\pi}\right)\left(1-\frac{\phi}{\pi}\right) \tag{36}$$

$$\Pr_{\boldsymbol{x},\boldsymbol{y}}[\langle\boldsymbol{y},\boldsymbol{q}\rangle\langle\boldsymbol{y},\boldsymbol{q}'\rangle \leq 0 \wedge \langle\boldsymbol{x},\boldsymbol{k}\rangle\langle\boldsymbol{x},\boldsymbol{k}'\rangle \leq 0] = \frac{\theta}{\pi}\frac{\phi}{\pi} \tag{37}$$

$$\Pr_{\boldsymbol{x},\boldsymbol{y}}[\langle\boldsymbol{x},\boldsymbol{k}\rangle\langle\boldsymbol{x},\boldsymbol{k}'\rangle\langle\boldsymbol{y},\boldsymbol{q}\rangle\langle\boldsymbol{y},\boldsymbol{q}'\rangle \geq 0] = \left(1-\frac{\theta}{\pi}\right)\left(1-\frac{\phi}{\pi}\right) + \frac{\theta}{\pi}\frac{\phi}{\pi} \tag{38}$$

A bit of algebra now shows

$$\Pr_{\boldsymbol{z}\sim\bar{\tau}}[\rho(\boldsymbol{z},\omega) = \rho(\boldsymbol{z},\omega')] = \left(1-\frac{\theta}{\pi}\right)\left(1-\frac{\phi}{\pi}\right) + \frac{\theta}{\pi}\frac{\phi}{\pi} \tag{39}$$

$$= \frac{1}{2} + \frac{2}{\pi^2}\left(\frac{\pi}{2}-\theta\right)\left(\frac{\pi}{2}-\phi\right) \tag{40}$$

By definition, $\theta = \arccos(\langle\boldsymbol{q},\boldsymbol{q}'\rangle)$ and $\phi = \arccos(\langle\boldsymbol{k},\boldsymbol{k}'\rangle)$. Using the identity $\arcsin(z) = \pi/2 - \arccos(z)$, we obtain

$$\Pr_{\boldsymbol{z}\sim\bar{\tau}}[\rho(\boldsymbol{z},\omega) = \rho(\boldsymbol{z},\omega')] = \frac{1}{2} + \frac{2}{\pi^2}\arcsin(\boldsymbol{q}^\top\boldsymbol{q}')\arcsin(\boldsymbol{k}^\top\boldsymbol{k}') \tag{41}$$

Finally,

$$\mathop{\mathbb{E}}_{\boldsymbol{z}\sim\bar{\tau}}[\rho(\boldsymbol{z},\omega)\rho(\boldsymbol{z},\omega')] = \Pr_{\boldsymbol{z}\sim\bar{\tau}}[\rho(\boldsymbol{z},\omega) = \rho(\boldsymbol{z},\omega')] - \Pr_{\boldsymbol{z}\sim\bar{\tau}}[\rho(\boldsymbol{z},\omega) \neq \rho(\boldsymbol{z},\omega')] \tag{42}$$

$$= 2\Pr_{\boldsymbol{z}\sim\bar{\tau}}[\rho(\boldsymbol{z},\omega) = \rho(\boldsymbol{z},\omega')] - 1 \tag{43}$$

$$= \frac{4}{\pi^2}\arcsin(\boldsymbol{q}^\top\boldsymbol{q}')\arcsin(\boldsymbol{k}^\top\boldsymbol{k}') \tag{44}$$

□

The above lemma gives us a handy expression for $\mathcal{T}^*\mathcal{T}$ that allows to show the following:

Lemma 10. *Let Y, Y' be spherical harmonics of degrees ℓ and ℓ', respectively. Then $Y \otimes Y'$ is an eigenfunction of the operator $\mathcal{T}^*\mathcal{T}$:*

$$\mathcal{T}^*\mathcal{T}(Y \otimes Y') = \frac{4}{\pi^2} \frac{\alpha_\ell \alpha_{\ell'}}{\sqrt{N(d,\ell)N(d,\ell')}} \cdot Y \otimes Y' \tag{45}$$

Proof. It is easily seen that

$$(\mathcal{T}^* f)(\cdot) = \int_{\mathcal{X}} \rho(z, \cdot) f(z) d\bar{\tau}(z) \tag{46}$$

and thus, substituting and changing the order of integration

$$[\mathcal{T}^*\mathcal{T}(Y \otimes Y')](\omega) = \int_\Omega \mathop{\mathbb{E}}_{z \sim \bar{\tau}} [\rho(z,\omega)\rho(z,\omega')] \cdot (Y \otimes Y')(\omega') d\bar{\tau}(\omega') \tag{47}$$

Applying Lemma 9 and expanding $d\bar{\tau}(\omega)$ and $Y \otimes Y'$,

$$= \frac{4}{\pi^2} \int_\Omega \arcsin(\boldsymbol{q}^\top \boldsymbol{q}') \arcsin(\boldsymbol{k}^\top \boldsymbol{k}') \cdot (Y \otimes Y')(\omega') d\bar{\tau}(\omega') \tag{48}$$

$$= \frac{4}{\pi^2} \int_{\mathbb{S}^{d-1}} \arcsin(\boldsymbol{q}^\top \boldsymbol{q}') Y(\boldsymbol{q}') d\tau(\boldsymbol{q}') \cdot \int_{\mathbb{S}^{d-1}} \arcsin(\boldsymbol{k}^\top \boldsymbol{k}') Y'(\boldsymbol{k}') d\tau(\boldsymbol{k}') \tag{49}$$

Applying the Hecke-Funke formula (Equation (24)) to the first integral,

$$\int_{\mathbb{S}^{d-1}} \arcsin(\boldsymbol{q}^\top \boldsymbol{q}') Y(\boldsymbol{q}') d\tau(\boldsymbol{q}') = Y(\boldsymbol{q}) \langle \arcsin, P_\ell \rangle_{u_d} \tag{50}$$

$$= Y(\boldsymbol{q}) \left\langle \arcsin, \frac{P_\ell}{\|P_\ell\|_{u_d}} \right\rangle_{u_d} \cdot \|P_\ell\|_{u_d} \tag{51}$$

$$= Y(\boldsymbol{q}) \frac{\alpha_\ell}{\sqrt{N(d,\ell)}} \tag{52}$$

By the same logic, the second integral equals $Y'(\boldsymbol{k}') \cdot \alpha_{\ell'} / \sqrt{N(d,\ell)}$. Combining these proves the lemma. □

The previous lemma immediately implies that the functions $\mathcal{T}(Y \otimes Y')$ form an orthogonal basis:

Lemma 11. *Let B be a set of orthonormal spherical harmonics. Then the elements of $\{\mathcal{T}(Y \otimes Y') \mid Y, Y' \in B\}$ are also orthogonal. Furthermore, if Y and Y' have degrees ℓ and ℓ', then*

$$\|\mathcal{T}(Y \otimes Y')\|_{\bar{\tau}}^2 = \frac{4}{\pi^2} \frac{\alpha_\ell \alpha_{\ell'}}{\sqrt{N(d,\ell)N(d,\ell')}} \tag{53}$$

Proof. Let $Y_i, Y_j, Y_{i'}, Y_{j'} \in B$. Let Y_i' have degree ℓ and Y_j' have degree ℓ'. Then

$$\langle \mathcal{T}(Y_i \otimes Y_j), \mathcal{T}(Y_{i'} \otimes Y_{j'}) \rangle = \langle Y_i \otimes Y_j, \mathcal{T}^*\mathcal{T}(Y_{i'} \otimes Y_{j'}) \rangle \tag{54}$$

$$= \langle Y_i \otimes Y_j, Y_{i'} \otimes Y_{j'} \rangle \cdot \frac{4}{\pi^2} \frac{\alpha_\ell \alpha_{\ell'}}{\sqrt{N(d,\ell)N(d,\ell')}} \tag{55}$$

But $\langle Y_i \otimes Y_j, Y_{i'} \otimes Y_{j'} \rangle$ is one if $Y_i = Y_{i'}$ and $Y_j = Y_{j'}$, and zero otherwise. □

C.4 Expansion of the Target Function

We define a surrogate target function that will turn out to be the relevant one for our analysis.

Definition 12. *The surrogate target function $\tilde{f} : \mathcal{X} \to \mathbb{R}$ is*

$$\tilde{f}(z) := \operatorname{sign}(\boldsymbol{x}^\top \boldsymbol{y}) \tag{56}$$

After a change of variables $(\boldsymbol{x}, \boldsymbol{w}) = (\boldsymbol{x}_1 - \boldsymbol{x}_2, \boldsymbol{x}_1 + \boldsymbol{x}_2)$, our original target function reduces simply to $\tilde{f}(z)\boldsymbol{x} + \boldsymbol{w}$. We now wish to expand $\tilde{f}$ in the basis $\{\mathcal{T}(Y \otimes Y')\}$. We will first need the following lemma, which describes the correlation of a rank-1 head with the surrogate target function.

Lemma 13. *Fix $\omega = (\boldsymbol{q}, \boldsymbol{k}) \in \Omega$. Then*

$$\left\langle \tilde{f}, \rho(\cdot, \omega) \right\rangle_{\bar{\tau}} = \sum_{\ell=0}^{\infty} c_\ell P_\ell(\boldsymbol{q}^\top \boldsymbol{k}) \tag{57}$$

where

$$c_\ell = \frac{2}{\pi} \eta_\ell \alpha_\ell \tag{58}$$

Proof. By definition,

$$\left\langle \tilde{f}, \rho(\cdot, \omega) \right\rangle_{\bar{\tau}} = \mathop{\mathbb{E}}_{\boldsymbol{x}, \boldsymbol{y} \sim \tau} \left[\operatorname{sign}(\boldsymbol{x}^\top \boldsymbol{y}) \operatorname{sign}(\boldsymbol{x}^\top \boldsymbol{k} \boldsymbol{q}^\top \boldsymbol{y}) \right] \tag{59}$$

Let τ_+ denote the uniform measure on the hemisphere $\{\boldsymbol{x} \in \mathbb{S}^{d-1} \mid \boldsymbol{x}^\top \boldsymbol{k} \geq 0\}$, and τ_- the uniform measure on the opposite hemisphere. Then we can decompose the expectation as follows:

$$\mathop{\mathbb{E}}_{\boldsymbol{x}, \boldsymbol{y} \sim \tau} [\operatorname{sign}(\boldsymbol{x}^\top \boldsymbol{y}) \operatorname{sign}(\boldsymbol{x}^\top \boldsymbol{k} \boldsymbol{q}^\top \boldsymbol{y})] = \frac{1}{2} \mathop{\mathbb{E}}_{\substack{\boldsymbol{x} \sim \tau_+ \\ \boldsymbol{y} \sim \tau}} [\operatorname{sign}(\boldsymbol{x}^\top \boldsymbol{y}) \operatorname{sign}(\boldsymbol{q}^\top \boldsymbol{y})] \tag{60}$$

$$- \frac{1}{2} \mathop{\mathbb{E}}_{\substack{\boldsymbol{x} \sim \tau_- \\ \boldsymbol{y} \sim \tau}} [\operatorname{sign}(\boldsymbol{x}^\top \boldsymbol{y}) \operatorname{sign}(\boldsymbol{q}^\top \boldsymbol{y})] \tag{61}$$

Given any fixed unit vectors $\boldsymbol{x}$, $\boldsymbol{q}$ we have that

$$\Pr_{\boldsymbol{y}}[\operatorname{sign}(\boldsymbol{x}^\top \boldsymbol{y}) = \operatorname{sign}(\boldsymbol{q}^\top \boldsymbol{y})] = 1 - \frac{\arccos(\boldsymbol{x}^\top \boldsymbol{q})}{\pi} \tag{62}$$

Therefore,

$$\mathop{\mathbb{E}}_{\boldsymbol{y}}[\operatorname{sign}(\boldsymbol{x}^\top \boldsymbol{y}) \operatorname{sign}(\boldsymbol{q}^\top \boldsymbol{y})] = \Pr_{\boldsymbol{y}}[\operatorname{sign}(\boldsymbol{x}^\top \boldsymbol{y}) = \operatorname{sign}(\boldsymbol{q}^\top \boldsymbol{y})] - \Pr_{\boldsymbol{y}}[\operatorname{sign}(\boldsymbol{x}^\top \boldsymbol{y}) \neq \operatorname{sign}(\boldsymbol{q}^\top \boldsymbol{y})] \tag{63}$$

$$= 2 \Pr_{\boldsymbol{y}}[\operatorname{sign}(\boldsymbol{x}^\top \boldsymbol{y}) = \operatorname{sign}(\boldsymbol{q}^\top \boldsymbol{y})] - 1 \tag{64}$$

$$= 1 - \frac{2 \arccos(\boldsymbol{x}^\top \boldsymbol{q})}{\pi} \tag{65}$$

Plugging this into the expression above,

$$= \frac{1}{2} \mathop{\mathbb{E}}_{\boldsymbol{x} \sim \tau_+} \left[1 - \frac{2 \arccos(\boldsymbol{x}^\top \boldsymbol{q})}{\pi} \right] - \frac{1}{2} \mathop{\mathbb{E}}_{\boldsymbol{x} \sim \tau_-} \left[1 - \frac{2 \arccos(\boldsymbol{x}^\top \boldsymbol{q})}{\pi} \right] \tag{66}$$

$$= -\frac{2}{\pi} \left(\frac{1}{2} \mathop{\mathbb{E}}_{\boldsymbol{x} \sim \tau_+} \left[\arccos(\boldsymbol{x}^\top \boldsymbol{q}) \right] - \frac{1}{2} \mathop{\mathbb{E}}_{\boldsymbol{x} \sim \tau_-} \left[\arccos(\boldsymbol{x}^\top \boldsymbol{q}) \right] \right) \tag{67}$$

$$= -\frac{2}{\pi} \left(\frac{1}{2} \mathop{\mathbb{E}}_{\boldsymbol{x} \sim \tau_+} \left[\operatorname{sign}(\boldsymbol{x}^\top \boldsymbol{k}) \arccos(\boldsymbol{x}^\top \boldsymbol{q}) \right] + \frac{1}{2} \mathop{\mathbb{E}}_{\boldsymbol{x} \sim \tau_-} \left[\operatorname{sign}(\boldsymbol{x}^\top \boldsymbol{k}) \arccos(\boldsymbol{x}^\top \boldsymbol{q}) \right] \right) \tag{68}$$

$$= -\frac{2}{\pi} \left(\mathop{\mathbb{E}}_{\boldsymbol{x} \sim \tau} \left[\operatorname{sign}(\boldsymbol{x}^\top \boldsymbol{k}) \arccos(\boldsymbol{x}^\top \boldsymbol{q}) \right] \right) \tag{69}$$

$$\tag{70}$$

Using the identity $\arccos(t) = \frac{\pi}{2} - \arcsin(t)$ and the fact that $\mathbb{E}_{\boldsymbol{x}}[\operatorname{sign}(\boldsymbol{x}^\top \boldsymbol{k})] = 0$,

$$= \frac{2}{\pi} \mathop{\mathbb{E}}_{\boldsymbol{x} \sim \tau} \left[\operatorname{sign}(\boldsymbol{x}^\top \boldsymbol{k}) \arcsin(\boldsymbol{x}^\top \boldsymbol{q}) \right] \tag{71}$$

$$= \frac{2}{\pi} \left\langle \operatorname{sign}(\langle \cdot, \boldsymbol{k} \rangle), \arcsin(\langle \cdot, \boldsymbol{q} \rangle) \right\rangle_{\tau} \tag{72}$$

We now expand $\operatorname{sign}(\langle \cdot, \boldsymbol{k} \rangle)$ and $\arcsin(\langle \cdot, \boldsymbol{q} \rangle)$ in a basis of spherical harmonics. By Hecke-Funk,

$$\left\langle \operatorname{sign}(\langle \cdot, \boldsymbol{k} \rangle), Y_\ell^j \right\rangle_{\tau} = Y_\ell^j(\boldsymbol{k}) \left\langle \operatorname{sign}, P_\ell \right\rangle_{u_d} = Y_\ell^j(\boldsymbol{k}) \eta_\ell \|P_\ell\|_{u_d} \tag{73}$$

$$\left\langle \arcsin(\langle \cdot, \boldsymbol{q} \rangle), Y_\ell^j \right\rangle_{\tau} = Y_\ell^j(\boldsymbol{q}) \left\langle \arcsin, P_\ell \right\rangle_{u_d} = Y_\ell^j(\boldsymbol{q}) \alpha_\ell \|P_\ell\|_{u_d} \tag{74}$$

$$\tag{75}$$

Thus, writing the inner product in the basis of spherical harmnoics,

$$\frac{2}{\pi}\Big\langle \operatorname{sign}(\langle\cdot,\boldsymbol{k}\rangle), \arcsin(\langle\cdot,\boldsymbol{q}\rangle)\Big\rangle_{\tau} = \frac{2}{\pi}\sum_{\ell=0}^{\infty}\sum_{j=1}^{N(d,\ell)}\left(Y_\ell^j(\boldsymbol{k})\eta_\ell\|P_\ell\|_{u_d}\right)\left(Y_\ell^j(\boldsymbol{q})\alpha_\ell\|P_\ell\|_{u_d}\right) \tag{76}$$

$$= \frac{2}{\pi}\sum_{\ell=0}^{\infty}\left(\eta_\ell\alpha_\ell\|P_\ell\|_{u_d}^2\sum_{j=1}^{N(d,\ell)}Y_\ell^j(\boldsymbol{k})Y_\ell^j(\boldsymbol{q})\right) \tag{77}$$

Applying the addition formula (Equation (22)),

$$= \frac{2}{\pi}\sum_{\ell=0}^{\infty}\eta_\ell\alpha_\ell\|P_\ell\|_{u_d}^2 N(d,\ell)P_\ell(\boldsymbol{k}^\top\boldsymbol{q}) \tag{78}$$

$$= \sum_{\ell=0}^{\infty}\frac{2}{\pi}\eta_\ell\alpha_\ell P_\ell(\boldsymbol{k}^\top\boldsymbol{q}) \tag{79}$$

$$\tag{80}$$

□

We now expand our surrogate target function $\tilde{f}$ in our basis $\{\mathcal{T}(Y\otimes Y')\}$. The following lemma shows that $\tilde{f}$ is orthogonal to any basis element for which $Y \neq Y'$, and that the coefficient of $\mathcal{T}(Y\otimes Y')$ only depends only on the degree of Y. That is, the energy of $\tilde{f}$ is evenly spread across all elements of $\{\mathcal{T}(Y_\ell\otimes Y_\ell) \mid Y_\ell\in\mathcal{F}_\ell\}$.

Lemma 14. *Let Y, Y' be spherical harmonics of odd degree. Let ℓ be the degree of Y. Then*

$$\left\langle \tilde{f}, \frac{\mathcal{T}(Y\otimes Y')}{\|\mathcal{T}(Y\otimes Y')\|_{\bar{\tau}}}\right\rangle_{\bar{\tau}} = \frac{\eta_\ell}{\sqrt{N(d,\ell)}}\delta_{Y,Y'} \tag{81}$$

where $\delta_{Y,Y'} = \mathbf{1}[Y = Y']$. That is, if the basis element is built from two identical spherical harmonics of degree ℓ, then its correlation with the target function depends only on ℓ; otherwise it is zero.

Proof. Expanding, switching the order of the integrals, and applying Lemma 13,

$$\left\langle \tilde{f}, \mathcal{T}(Y\otimes Y')\right\rangle_{\bar{\tau}} = \int_{\mathcal{X}}\int_{\Omega}\tilde{f}(\boldsymbol{z})\rho(\boldsymbol{z},\omega)(Y\otimes Y')(\omega)d\bar{\tau}(\omega)d\bar{\tau}(\boldsymbol{z}) \tag{82}$$

$$= \int_{\Omega}\left\langle \tilde{f}, \rho(\cdot,\omega)\right\rangle_{\bar{\tau}}(Y\otimes Y')(\omega)d\bar{\tau}(\omega) \tag{83}$$

$$= \sum_{\ell'=0}^{\infty}c_{\ell'}\int_{\Omega}P_{\ell'}(\boldsymbol{q}^\top\boldsymbol{k})(Y\otimes Y')(\omega)d\bar{\tau}(\omega) \tag{84}$$

Expanding the integral over Ω and applying Hecke-Funk (Equation (24)),

$$= \sum_{\ell'=0}^{\infty}c_{\ell'}\int_{\mathbb{S}^{d-1}}\int_{\mathbb{S}^{d-1}}P_{\ell'}(\boldsymbol{q}^\top\boldsymbol{k})Y'(\boldsymbol{k})Y(\boldsymbol{q})d\tau(\boldsymbol{k})d\tau(\boldsymbol{q}) \tag{85}$$

$$= \sum_{\ell'=0}^{\infty}c_{\ell'}\int_{\mathbb{S}^{d-1}}\Big(Y'(\boldsymbol{q})\left\langle P_{\ell'}, P_{\ell'}\right\rangle_{u_d}\Big)Y(\boldsymbol{q})d\tau(\boldsymbol{q}) \tag{86}$$

$$= \sum_{\ell'=0}^{\infty}c_{\ell'}\|P_{\ell'}\|_{u_d}^2\left\langle Y, Y'\right\rangle_{\tau} \tag{87}$$

$$= \frac{c_\ell}{N(d,\ell)} \tag{88}$$

Finally, applying the formula for c_ℓ from Lemma 13 and the formula for $\|\mathcal{T}(Y \otimes Y')\|_\tau$ from Lemma 11,

$$\left\langle \tilde{f}, \frac{\mathcal{T}(Y \otimes Y)}{\|\mathcal{T}(Y \otimes Y)\|_{\bar{\tau}}} \right\rangle_{\bar{\tau}} = \frac{c_\ell}{N(d,\ell)} \cdot \frac{1}{\|\mathcal{T}(Y \otimes Y)\|_{\bar{\tau}}} = \frac{\frac{2}{\pi}\eta_\ell \alpha_\ell}{N(d,\ell)} \cdot \frac{1}{\sqrt{\frac{4}{\pi^2}\alpha_{\ell(i)}^2 / N(d,\ell)}} = \frac{\eta_\ell}{\sqrt{N(d,\ell)}} \tag{89}$$

□

Up to now, we have constructed a basis without showing that its span includes our target function. Lemma 25 (in Appendix C.8) verifies that, in fact, $\tilde{f}$ lies in this span. This lemma is not needed for the proof of Theorem 2, but is used in the kernel approximation of Appendix C.8. It also shows that this step of the proof is tight. We do not lose anything by lower bounding the error only on the part of $\tilde{f}$ that lies in the span of our basis functions.

C.5 Expansion of the Head Functions

In this section, we expand the low-rank attention head function in our basis $\{\mathcal{T}(Y \otimes Y')\}$. Unlike the target function, the energy of an attention head is not spread out, but concentrated on a few basis elements in each harmonic. We first need the following lemma, which we will use to bound the number of these special basis elements.

Lemma 15. *Let $\mathcal{A}_\ell$ be the span of the harmonics of degree ℓ on $\mathbb{S}^{d-1}$ that are zero after marginalizing onto the first r coordinates. Then*

$$\dim(\mathcal{F}_\ell / \mathcal{A}_\ell) := M(r,\ell) \le \binom{r+\ell}{\ell} \tag{90}$$

where $\mathcal{F}_\ell / \mathcal{A}_\ell$ is the orthogonal complement of $\mathcal{A}_\ell$ in $\mathcal{F}_\ell$. Furthermore, $M(1,\ell) = 1$.

Proof. Let $\mathcal{L} : \mathcal{F}_\ell \to L^2(B_r)$ be the linear operator which marginalizes a degree ℓ spherical harmonic function on the first r coordinates. (Here, B_r is the unit r-ball.) That is,

$$(\mathcal{L}f)(\boldsymbol{x}) := \mathop{\mathbb{E}}_{\boldsymbol{y} \sim \mathbb{S}^{d-r-1}} f\left(\begin{bmatrix} \boldsymbol{x} \\ \boldsymbol{y}\sqrt{1-\|\boldsymbol{x}\|^2} \end{bmatrix}\right) \tag{91}$$

By definition, $\mathcal{A}_\ell$ is the null space of $\mathcal{L}$. We will show below that the range of $\mathcal{L}$ contains only polynomials of the first r coordinates of degree at most ℓ. The dimension of the space of polynomials in dimension r of degree at most ℓ is $\binom{r+\ell}{\ell}$. Thus, by the rank-nullity theorem,

$$\dim(\mathcal{F}_\ell) \le \dim(\mathcal{A}_\ell) + \binom{r+\ell}{\ell} \tag{92}$$

and therefore

$$\dim(\mathcal{F}_\ell / \mathcal{A}_\ell) = \dim(\mathcal{F}_\ell) - \dim(\mathcal{A}_\ell) \le \binom{r+\ell}{\ell} \tag{93}$$

We will now show that the range of $\mathcal{L}$ contains only polynomials in the first r coordinates of degree at most ℓ. Each spherical harmonic is the restriction to $\mathbb{S}^{d-1}$ of a harmonic homogeneous polynomial on $\mathbb{R}^d$, so it suffices to show that $\mathcal{L}$ maps monomials of degree exactly ℓ in $\mathbb{R}^d$ to polynomials of degree at most ℓ in the first r coordinates. Let

$$Y\left(\begin{bmatrix} \boldsymbol{x} \\ \boldsymbol{y} \end{bmatrix}\right) := x_1^{p_1} \cdots x_r^{p_r} y_{r+1}^{p_{r+1}} \cdots y_d^{p_d} = \left(\prod_{i=1}^{r} x_i^{p_i}\right)\left(\prod_{i=r+1}^{d} y_i^{p_i}\right) \tag{94}$$

be one such monomial. If any of $p_{r+1}, \ldots, p_d$ is odd, then $L[Y] = 0$. If all are even, then

$$L[Y](\boldsymbol{x}) = \left(\prod_{i=1}^{r} x_i^{p_i}\right)\left(\mathop{\mathbb{E}}_{\boldsymbol{y} \sim \mathbb{S}^{d-r-1}} \prod_{i=r+1}^{d} \left(y_i\sqrt{1-\|\boldsymbol{x}\|^2}\right)^{p_i}\right) \tag{95}$$

$$= \left(\prod_{i=1}^{r} x_i^{p_i}\right)\left(\prod_{i=r+1}^{d} \left(1-\|\boldsymbol{x}\|^2\right)^{p_i/2}\right)\left(\mathop{\mathbb{E}}_{\boldsymbol{y} \sim \mathbb{S}^{d-r-1}} \prod_{i=r+1}^{d} y_i^{p_i}\right) \tag{96}$$

is a polynomial in $\boldsymbol{x}$ whose highest degree term has degree $(\sum_{i=1}^{r} p_i) + \left(\sum_{i=r+1}^{d} p_i\right)$, which equals the degree of the original monomial.

For the special case of $r = 1$, it suffices to show that $\mathcal{L}$ has rank one, or equivalently that its nullspace has dimension $N(d,\ell) - 1$. Let $Y_1 = P_\ell(\langle \hat{\boldsymbol{e}}_1, \cdot \rangle)$, where $\hat{\boldsymbol{e}}_1 \in \mathbb{R}^d$ is the first standard basis vector. By Theorem 4.10 of Frye & Efthimiou (2012), Y_1 is a spherical harmonic of degree ℓ. Complete an orthonormal basis $\{Y_1, \dots Y_{N(d,\ell)}\}$ of $\mathcal{F}_\ell$. Our goal is to show that $\mathcal{L}Y_j = 0$ for all $j \in \{2, \dots N(d,\ell)\}$ (with equality in the weak sense).

To do this, it suffices to show that $\langle P_\ell, \mathcal{L}Y_j \rangle = 0$ for all ℓ:

$$\langle P_\ell, \mathcal{L}Y_j \rangle = \mathop{\mathbb{E}}_{x \sim u_d} [P_\ell(x)(\mathcal{L}Y_j)(x)] \tag{97}$$

$$= \mathop{\mathbb{E}}_{x \sim u_d} \left[P_\ell(x) \mathop{\mathbb{E}}_{\boldsymbol{y} \in \mathbb{S}^{d-2}} Y_j \left(\begin{bmatrix} x \\ \boldsymbol{y}\sqrt{1-|x|^2} \end{bmatrix} \right) \right] \tag{98}$$

$$= \mathop{\mathbb{E}}_{\boldsymbol{z} \sim \tau} [P_\ell(x) Y_j(\boldsymbol{z})] \tag{99}$$

where $\boldsymbol{z} := \begin{bmatrix} x \\ \boldsymbol{y}\sqrt{1-|x|^2} \end{bmatrix} \in \mathbb{S}^{d-1}$. But by definition, $P_\ell(x) = Y_1 \left(\begin{bmatrix} x \\ \boldsymbol{y}\sqrt{1-|x|^2} \end{bmatrix} \right)$ for all $\boldsymbol{y} \in \mathbb{S}^{d-2}$. Continuing from above,

$$= \mathop{\mathbb{E}}_{\boldsymbol{z} \sim \tau} [Y_1(\boldsymbol{z}) Y_j(\boldsymbol{z})] = \langle Y_1, Y_j \rangle_\tau = 0 \tag{100}$$

for all $j \neq 1$. □

Lemma 16. *Let $\boldsymbol{X}$ be a square matrix. Let $\mathcal{D}$ be the uniform distribution over orthogonal matrices. Then,*

$$\mathop{\mathbb{E}}_{\boldsymbol{Q} \sim \mathcal{D}} [\boldsymbol{Q}^\top \boldsymbol{X} \boldsymbol{Q}] = \operatorname{tr}(\boldsymbol{X}) \cdot \boldsymbol{I} \tag{101}$$

Proof. Let q_{ki} denote the entry in the kth row and ith column of $\boldsymbol{Q}$. Then the (i,j) entry of the expectation is

$$\mathop{\mathbb{E}}_{\boldsymbol{Q} \sim \mathcal{D}} [\boldsymbol{Q}^\top \boldsymbol{X} \boldsymbol{Q}]_{ij} = \sum_k \sum_\ell x_{k\ell} \mathop{\mathbb{E}}_{Q} [q_{ki} q_{\ell j}] \tag{102}$$

So long as $(k,i) \neq (\ell,j)$, then conditional distribution of $q_{\ell j}$ given q_{ki} is symmetric, since negating the ℓth row (or jth column) of $\boldsymbol{Q}$ would produce another orthonormal matrix. Thus, if $(k,i) \neq (\ell,j)$, then the expectation is zero. The only non-zero terms are

$$\mathop{\mathbb{E}}_{\boldsymbol{Q} \sim \mathcal{D}} [\boldsymbol{Q}^\top \boldsymbol{X} \boldsymbol{Q}]_{ii} = \sum_k x_{kk} \mathop{\mathbb{E}}_{Q} [q_{ki}^2] \tag{103}$$

Since the marginal distribution of each row (or column) is uniform on the unit sphere, the variance of each entry is 1. □

Lemma 17. *Define $M(r,\ell)$ as in Lemma 15. Assume the rank $r < d$ and consider the functions $g_h(z) = \boldsymbol{x}^\top \boldsymbol{V}_h \boldsymbol{x} \cdot \tilde{\phi}_h(\boldsymbol{K}_h^\top \boldsymbol{x}, \boldsymbol{y})$ for $\tilde{\phi}_h : \mathbb{R}^r \times \mathbb{S}^{d-1} \to \mathbb{R}$ and $\boldsymbol{K}_h \in \mathbb{R}^{d \times r}$ for $h \in [H]$. Then there exists a subspace $\mathcal{A}_\ell \subseteq \mathcal{F}_\ell$ of dimension at least $N(d,\ell) - H \cdot M(r,\ell)$ such that $\mathcal{T}(Y_\ell \otimes Y_\ell)$ is orthogonal to g_h for any $Y_\ell \in \mathcal{A}_\ell$ and any $h \in H$.*

Proof. The first part of the proof gives a construction for $\mathcal{A}_\ell$. Fix $\boldsymbol{y}$, $\boldsymbol{q}$ and h and define

$$h_{\boldsymbol{K}}(\boldsymbol{k}) := \mathop{\mathbb{E}}_{\boldsymbol{x} \sim \tau} [\rho(\boldsymbol{z}, \omega) g_h(\boldsymbol{x})] = \mathop{\mathbb{E}}_{\boldsymbol{x} \sim \tau} \left[\operatorname{sign}(\boldsymbol{x}^\top \boldsymbol{k} \boldsymbol{q}^\top \boldsymbol{y}) \boldsymbol{x}^\top \boldsymbol{V} \boldsymbol{x} \cdot \tilde{\phi}_h(\boldsymbol{K}^\top \boldsymbol{x}, \boldsymbol{y}) \right] \tag{104}$$

Define $\overline{\boldsymbol{K}} = [\boldsymbol{K} \quad \boldsymbol{k}]$. As a first step, we show that this function only depends on a particular projection of $\boldsymbol{V}$, not on $\boldsymbol{V}$ itself. Choose a basis such that the column span of $\overline{\boldsymbol{K}}$ is $\operatorname{span}(\{\boldsymbol{e}_1, \dots, \boldsymbol{e}_{r'}\}$, where $1 \leq r' \leq \min(r+1, d)$. Then we can rewrite $\boldsymbol{V} = \begin{bmatrix} \boldsymbol{A} & \boldsymbol{B} \\ \boldsymbol{C} & \boldsymbol{D} \end{bmatrix}$ where $\boldsymbol{A} \in \mathbb{R}^{r' \times r'}$. The distribution of $\boldsymbol{x}$ is isotropic and independent of $\boldsymbol{y}$. Therefore, we can rotate it without affecting the expectation. In fact, we can draw a random orthogonal matrix from any distribution, and

$\mathbb{E}_{\boldsymbol{x},\boldsymbol{Q}}[f(\boldsymbol{Q}\boldsymbol{x})]$ will equal $\mathbb{E}_{\boldsymbol{x}}[f(\boldsymbol{x})]$. We draw random orthogonal matrices that fix the column span of $\overline{\boldsymbol{K}}$, that is, matrices of the form $\boldsymbol{Q} = \begin{bmatrix} \boldsymbol{I} & \cdot \\ \cdot & \tilde{\boldsymbol{Q}} \end{bmatrix}$, where $\tilde{\boldsymbol{Q}} \in \mathbb{R}^{(d-r')\times(d-r')}$ is a uniformly distributed orthogonal matrix. Then,

$$h_{\boldsymbol{K}}(\boldsymbol{k}) = \mathop{\mathbb{E}}_{\boldsymbol{x},\boldsymbol{Q}}\left[\operatorname{sign}(\boldsymbol{x}^\top\boldsymbol{Q}^\top\boldsymbol{k}\boldsymbol{q}^\top\boldsymbol{y})\boldsymbol{x}^\top\boldsymbol{Q}^\top\boldsymbol{V}_h\boldsymbol{Q}\boldsymbol{x}\cdot\tilde{\phi}_h(\boldsymbol{K}^\top\boldsymbol{Q}\boldsymbol{x},\boldsymbol{y})\right] \tag{105}$$

$$= \mathop{\mathbb{E}}_{\boldsymbol{x},\tilde{\boldsymbol{Q}}}\left[\operatorname{sign}(\boldsymbol{x}^\top\boldsymbol{k}\boldsymbol{q}^\top\boldsymbol{y})\boldsymbol{x}^\top\begin{bmatrix}\boldsymbol{A} & \boldsymbol{B}\tilde{\boldsymbol{Q}} \\ \tilde{\boldsymbol{Q}}^\top\boldsymbol{C} & \tilde{\boldsymbol{Q}}^\top\boldsymbol{D}\tilde{\boldsymbol{Q}}\end{bmatrix}\boldsymbol{x}\cdot\tilde{\phi}_h(\boldsymbol{K}^\top\boldsymbol{x},\boldsymbol{y})\right] \tag{106}$$

Moving the expectation over $\tilde{\boldsymbol{Q}}$ inside, the off-diagonal blocks are both 0. Applying Lemma 16, the bottom right block becomes $\operatorname{tr}(\boldsymbol{D})\cdot\boldsymbol{I}$. Thus, letting $\boldsymbol{A}' = \boldsymbol{A} - \operatorname{tr}(\boldsymbol{D})\cdot\boldsymbol{I}$,

$$\mathop{\mathbb{E}}_{\tilde{\boldsymbol{Q}}}\begin{bmatrix}\boldsymbol{A} & \boldsymbol{B}\tilde{\boldsymbol{Q}} \\ \tilde{\boldsymbol{Q}}^\top\boldsymbol{C} & \tilde{\boldsymbol{Q}}^\top\boldsymbol{D}\tilde{\boldsymbol{Q}}\end{bmatrix} = \operatorname{tr}(\boldsymbol{D})\cdot\boldsymbol{I} + \boldsymbol{U}\boldsymbol{A}'\boldsymbol{U}^\top \tag{107}$$

where $\boldsymbol{U} = \begin{bmatrix}\boldsymbol{I} \\ \cdot\end{bmatrix}$ is defined to be the column span of $\overline{\boldsymbol{K}}$. In all,

$$h_{\boldsymbol{K}}(\boldsymbol{k}) = \mathop{\mathbb{E}}_{\boldsymbol{x}}\left[\operatorname{sign}(\boldsymbol{x}^\top\boldsymbol{k}\boldsymbol{q}^\top\boldsymbol{y})\boldsymbol{x}^\top\left(\operatorname{tr}(\boldsymbol{D})\cdot\boldsymbol{I} + \boldsymbol{U}\boldsymbol{A}'\boldsymbol{U}^\top\right)\boldsymbol{x}\cdot\tilde{\phi}_h(\boldsymbol{K}^\top\boldsymbol{x},\boldsymbol{y})\right] \tag{108}$$

Now that we have reduced $\boldsymbol{V}$, we can more clearly see the implications of the rotational invariance of the distribution of $\boldsymbol{x}$. Let $\boldsymbol{O}$ be an arbitrary orthonormal matrix. Then

$$h_{\boldsymbol{K}}(\boldsymbol{k}) = \mathop{\mathbb{E}}_{\boldsymbol{x}\sim\tau}\left[\operatorname{sign}(\boldsymbol{x}^\top\boldsymbol{O}^\top\boldsymbol{k}\boldsymbol{q}^\top\boldsymbol{y})\boldsymbol{x}^\top\boldsymbol{O}^\top\left(\operatorname{tr}(\boldsymbol{D})\cdot\boldsymbol{I} + \boldsymbol{U}\boldsymbol{A}'\boldsymbol{U}^\top\right)\boldsymbol{O}\boldsymbol{x}\cdot\tilde{\phi}_h(\boldsymbol{K}^\top\boldsymbol{O}\boldsymbol{x},\boldsymbol{y})\right] \tag{109}$$

$$= \mathop{\mathbb{E}}_{\boldsymbol{x}\sim\tau}\left[\operatorname{sign}(\boldsymbol{x}^\top\boldsymbol{O}^\top\boldsymbol{k}\boldsymbol{q}^\top\boldsymbol{y})\boldsymbol{x}^\top\left(\operatorname{tr}(\boldsymbol{D})\cdot\boldsymbol{I} + \boldsymbol{O}^\top\boldsymbol{U}\boldsymbol{A}'\boldsymbol{U}^\top\boldsymbol{O}\right)\boldsymbol{x}\cdot\tilde{\phi}_h(\boldsymbol{K}^\top\boldsymbol{O}\boldsymbol{x},\boldsymbol{y})\right] \tag{110}$$

$$= h_{\boldsymbol{O}^\top\boldsymbol{K}}(\boldsymbol{O}^\top\boldsymbol{k}) \tag{111}$$

where the last step follows because $\boldsymbol{O}^\top\boldsymbol{U}$ is precisely the column span of $\boldsymbol{O}^\top\boldsymbol{K}$. Thus by Weyl's fundamental theorem of invariant functions, there exists $\tilde{h}:\mathbb{R}^r\to\mathbb{R}$ such that

$$h_{\boldsymbol{K}}(\boldsymbol{k}) = \tilde{h}(\boldsymbol{K}^\top\boldsymbol{k}) \tag{112}$$

Let $\tau_{\boldsymbol{K}}$ denote the marginal distribution of τ on the column space of $\boldsymbol{K}$ and let $\tau_{\boldsymbol{K}^\perp}$ denote its marginal distribution on the orthogonal complement of the column space of $\boldsymbol{K}$. Then the random vector $\boldsymbol{v} + \boldsymbol{v}^\perp\sqrt{1-\|\boldsymbol{v}\|}$, where $\boldsymbol{v}\sim\tau_{\boldsymbol{K}}$ and $\boldsymbol{v}^\perp\sim\tau_{\boldsymbol{K}^\perp}$ is distributed uniformly on the sphere. Let Y be a spherical harmonic that is zero after marginalizing the onto the column space of $\boldsymbol{K}$. (For example, if $\boldsymbol{K}^\top = \begin{bmatrix}\tilde{\boldsymbol{K}}^\top & \boldsymbol{0}_{r\times d-r}\end{bmatrix}$, then marginalizing onto the column space means taking the average of the function over the final $d-r$ coordinates.) Then

$$\langle h_{\boldsymbol{K}}, Y\rangle = \int_{\mathbb{S}^{d-1}} h_{\boldsymbol{K}}(\boldsymbol{k})Y(\boldsymbol{k})d\tau(\boldsymbol{k}) \tag{113}$$

$$= \int\int h_{\boldsymbol{K}}(\boldsymbol{v}+\boldsymbol{v}^\perp\sqrt{1-\|\boldsymbol{v}\|})Y(\boldsymbol{v}+\boldsymbol{v}^\perp\sqrt{1-\|\boldsymbol{v}\|})d\tau_{\boldsymbol{K}^\perp}(\boldsymbol{v}^\perp)d\tau_{\boldsymbol{K}}(\boldsymbol{v}) \tag{114}$$

$$= \int \tilde{h}_{\boldsymbol{K}}(\boldsymbol{v})\left(\int Y(\boldsymbol{v}+\boldsymbol{v}^\perp\sqrt{1-\|\boldsymbol{v}\|})d\tau_{\boldsymbol{K}^\perp}(\boldsymbol{v}^\perp)\right)d\tau_{\boldsymbol{K}}(\boldsymbol{v}) \tag{115}$$

$$= 0 \tag{116}$$

Let $\mathcal{A}_\ell^h \subset \mathcal{F}_\ell$ be the space of spherical harmonics of degree ℓ that have this marginalization property with respect to $\boldsymbol{K}_h$. Let $\mathcal{A}_\ell = \cap_h \mathcal{A}_\ell^h$. Recall that $N(d,\ell)$ is the dimension of $\mathcal{F}_\ell$, and $M(r,\ell)$ is the dimension of the orthogonal complement of $\mathcal{A}_\ell^h$ in $\mathcal{F}_\ell$, denoted $\mathcal{F}_\ell/A_\ell^h$. Thus,

$$\dim(\mathcal{A}_\ell) = \dim(\mathcal{F}_\ell) - \dim(\mathcal{F}_\ell/\mathcal{A}_\ell) = N(d,l) - \dim(\oplus_h(\mathcal{F}_\ell/\mathcal{A}_\ell^h)) \geq N(d,\ell) - H\cdot M(r,\ell) \tag{117}$$

It remains to show that for all $Y\in\mathcal{A}_\ell$, $\mathcal{T}(Y_\ell\otimes Y_\ell)$ is orthogonal to g_h.

$$\langle\mathcal{T}(Y\otimes Y), g_h\rangle_{\bar{\tau}} = \int_\Omega \mathop{\mathbb{E}}_{\boldsymbol{z}}[\rho(\boldsymbol{z},\omega)g_h(\boldsymbol{z})]Y(\boldsymbol{k})Y(\boldsymbol{q})d\tau(\boldsymbol{k})d\tau(\boldsymbol{q}) \tag{118}$$

$$= \int_{\mathbb{S}^{d-1}} \mathop{\mathbb{E}}_{\boldsymbol{y}}\left(\int_{\mathbb{S}^{d-1}} \mathop{\mathbb{E}}_{\boldsymbol{x}}[\rho(\boldsymbol{x},\boldsymbol{y},\omega)g_h(z)]Y(\boldsymbol{k})d\tau(\boldsymbol{k})\right)Y(\boldsymbol{q})\tau(\boldsymbol{q}) \tag{119}$$

But for any fixed $\boldsymbol{y}$ and $\boldsymbol{q}$,

$$\int_{\mathbb{S}^{d-1}} \mathop{\mathbb{E}}_{\boldsymbol{x}}[\rho(\boldsymbol{x}, \boldsymbol{y}, \omega) g_h(z)] Y(\boldsymbol{k}) d\tau(\boldsymbol{k}) = \langle h_{\boldsymbol{K}_h}, Y \rangle = 0 \tag{120}$$

by the calculation above, where the final step follows because $Y \in \mathcal{A}_\ell \subset \mathcal{A}_\ell^h$. □

C.6 Proof of Theorem 2

Theorem 2 (Low-Rank Approximation Lower Bounds, Equivariant Case)**.** *There exist universal constants c, c', C and C' such that if either of the following sets of assumptions hold:*

1. High-accuracy regime*: $r \leq d - 3$, $\epsilon \leq \frac{c}{d+1}$, and*

$$H \leq C \cdot 2^{d-(r+1)\log_2(2d/r)} \,. \tag{5}$$

2. High-dimensional regime: $d \geq 5$, $\epsilon \geq \frac{c'}{d - 2e^2 \cdot r}$ *and*

$$H \leq \frac{1}{2} \left(\frac{1}{2e} \cdot \frac{d}{r + C'/\epsilon} \right)^{C'/\epsilon} \,. \tag{6}$$

Then, for any choice of H rank-r generalized attention heads $\phi_h : \mathbb{R}^{r \times 2} \to \Delta^1$, $\boldsymbol{V}_h \in \mathbb{R}^{d \times d}$, $\boldsymbol{K}_h \in \mathbb{R}^{d \times r}$ the error of approximating the nearest neighbor function is bounded as follows

$$\mathop{\mathbb{E}}_{\substack{\boldsymbol{x}_1, \boldsymbol{x}_2 \sim \mathcal{D}_2(\mathbb{S}^{d-1}) \\ \boldsymbol{y} \sim \mathrm{Unif}(\mathbb{S}^{d-1})}} \left\| f(\boldsymbol{X}; \boldsymbol{y}) - \sum_{h=1}^{H} \boldsymbol{V}_h \boldsymbol{X} \phi_h \left(\boldsymbol{K}_h^\top \boldsymbol{X}, \boldsymbol{y} \right) \right\|_2^2 \geq \epsilon \,, \tag{7}$$

where f is defined as in Equation (3)*.*

Proof. We lower bound the error by projecting it onto the unit vector $(\boldsymbol{x}_1 - \boldsymbol{x}_2)/(\sqrt{2})$. For convenience, we define a basis

$$\boldsymbol{x} = \frac{\boldsymbol{x}_1 - \boldsymbol{x}_2}{\sqrt{2}} \qquad \boldsymbol{w} = \frac{\boldsymbol{x}_1 + \boldsymbol{x}_2}{\sqrt{2}} \tag{121}$$

The joint distribution of $\boldsymbol{x}$ and $\boldsymbol{w}$ is the same as that of $\boldsymbol{x}_1$ and $\boldsymbol{x}_2$. They are each uniformly distributed on the sphere, and they are always orthogonal. The projection of the target function onto $\boldsymbol{x}$ yields the surrogate target function of Definition 12:

$$\left\langle \frac{\boldsymbol{x}_1 - \boldsymbol{x}_2}{\sqrt{2}}, f(\boldsymbol{X}; y) \right\rangle = \frac{1}{\sqrt{2}} \operatorname{sign}\left(\langle \boldsymbol{x}_1 - \boldsymbol{x}_2, \boldsymbol{y} \rangle \right) =: \frac{1}{\sqrt{2}} \tilde{f}(\boldsymbol{x}, \boldsymbol{y}) \tag{122}$$

Let the attention weights produced by a softmax head be t_1 and $t_2 = 1 - t_1$. Then the output of the head before multiplication with $\boldsymbol{V}$ is

$$t\boldsymbol{x}_1 + (1-t)\boldsymbol{x}_2 = \frac{t_1 - t_2}{\sqrt{2}} \boldsymbol{x} + \frac{1}{\sqrt{2}} \boldsymbol{w} \tag{123}$$

Letting $\tilde{\phi}(\boldsymbol{K}^\top \boldsymbol{x}, \boldsymbol{y}) = (t_1 - t_2)/\sqrt{2}$, the inner product of the head with $\boldsymbol{x}$ is

$$\boldsymbol{x}^\top \boldsymbol{V} \boldsymbol{x} \cdot \tilde{\phi}(\boldsymbol{K}^\top \boldsymbol{x}, \boldsymbol{y}) + \boldsymbol{x}^\top \boldsymbol{V} \boldsymbol{w} \tag{124}$$

Notice that, since the conditional distribution of $\boldsymbol{w}$ given $\boldsymbol{x}$ is symmetric, the correlation of the second term above with the surrogate target is zero:

$$\mathop{\mathbb{E}}_{\boldsymbol{x}_1, \boldsymbol{x}_2 \sim \mathcal{D}_2(\mathbb{S}^{d-1})} \left[\tilde{f}(\boldsymbol{x}, \boldsymbol{y}) \cdot \boldsymbol{x}^\top \boldsymbol{V} \boldsymbol{w} \right] = 0 \tag{125}$$

Thus, we have the following lower bound:

$$\mathop{\mathbb{E}}_{\substack{\boldsymbol{x}_1,\boldsymbol{x}_2\sim\mathcal{D}_2(\mathbb{S}^{d-1})\\ \boldsymbol{y}\sim\mathrm{Unif}(\mathbb{S}^{d-1})}}\left\|f(\boldsymbol{X};\boldsymbol{y})-\sum_{h=1}^{H}\boldsymbol{V}_h\boldsymbol{X}\phi_h\left(\boldsymbol{K}_h^\top(\boldsymbol{x}_1-\boldsymbol{x}_2),\boldsymbol{y}\right)\right\|^2 \tag{126}$$

$$\geq \mathop{\mathbb{E}}_{\substack{\boldsymbol{x}_1,\boldsymbol{x}_2\sim\mathcal{D}_2(\mathbb{S}^{d-1})\\ \boldsymbol{y}\sim\mathrm{Unif}(\mathbb{S}^{d-1})}}\left\langle \boldsymbol{x},\quad f(\boldsymbol{X};\boldsymbol{y})-\sum_{h=1}^{H}\boldsymbol{V}_h\boldsymbol{X}\phi_h\left(\boldsymbol{K}_h^\top(\boldsymbol{x}_1-\boldsymbol{x}_2),\boldsymbol{y}\right)\right\rangle^2 \tag{127}$$

$$= \mathop{\mathbb{E}}_{\substack{\boldsymbol{x},\boldsymbol{w}\sim\mathcal{D}_2(\mathbb{S}^{d-1})\\ \boldsymbol{y}\sim\mathrm{Unif}(\mathbb{S}^{d-1})}}\frac{1}{2}\left(\tilde{f}(\boldsymbol{x},\boldsymbol{y})-\sum_{h=1}^{H}\boldsymbol{x}^\top\boldsymbol{V}_h\boldsymbol{x}\cdot\tilde{\phi}_h(\boldsymbol{K}_h^\top\boldsymbol{x},\boldsymbol{y}))-\sum_{h=1}^{H}\boldsymbol{x}^\top\boldsymbol{V}_h\boldsymbol{w}\right)^2 \tag{128}$$

$$\geq \mathop{\mathbb{E}}_{\boldsymbol{x},\boldsymbol{y}\sim\mathrm{Unif}(\mathbb{S}^{d-1})}\frac{1}{2}\left(\tilde{f}(\boldsymbol{x},\boldsymbol{y})-\sum_{h=1}^{H}\boldsymbol{x}^\top\boldsymbol{V}_h\boldsymbol{x}\cdot\tilde{\phi}_h(\boldsymbol{K}_h^\top\boldsymbol{x},\boldsymbol{y}))\right)^2 \tag{129}$$

$$= \frac{1}{2}\left\|\tilde{f}-\sum_{h=1}^{H}g_h\right\|_{\bar{\tau}}^2 \tag{130}$$

where $g_h(z)=\boldsymbol{x}^\top\boldsymbol{V}_h\boldsymbol{x}\cdot\tilde{\phi}_h(\boldsymbol{K}_h^\top\boldsymbol{x},\boldsymbol{y})$. Construct the space $\mathcal{A}_\ell\subseteq\mathcal{F}_\ell$ according to Lemma 17, and let $\{Y_\ell^i\}_{i=1}^{\dim\mathcal{A}_\ell}$ be an orthonormal basis of $\mathcal{A}_\ell$. Then each element in the following set is orthogonal to each $g_h(z)$:

$$\left\{\frac{\mathcal{T}(Y_\ell^i\otimes Y_\ell^i)}{\|\mathcal{T}(Y_\ell^i\otimes Y_\ell^i)\|_{\bar{\tau}}}\right\}_{i=1}^{\dim(\mathcal{A}_\ell)} \tag{131}$$

Furthermore, by Lemma 11, this set is orthonormal. Thus

$$\left\|\tilde{f}-\sum_{h=1}^{H}g_h\right\|_{\bar{\tau}}^2 \geq \sum_{\ell\text{ odd}}\sum_{i=1}^{\dim(\mathcal{A}_\ell)}\left\langle\tilde{f}-\sum_{h=1}^{H}g_h,\frac{\mathcal{T}(Y_\ell^i\otimes Y_\ell^i)}{\|\mathcal{T}(Y_\ell^i\otimes Y_\ell^i)\|_{\bar{\tau}}}\right\rangle^2 \tag{132}$$

$$= \sum_{\ell\text{ odd}}\sum_{i=1}^{\dim(\mathcal{A}_\ell)}\left\langle\tilde{f},\frac{\mathcal{T}(Y_\ell^i\otimes Y_\ell^i)}{\|\mathcal{T}(Y_\ell^i\otimes Y_\ell^i)\|_{\bar{\tau}}}\right\rangle^2 \tag{133}$$

$$= \sum_{\ell\text{ odd}}\dim(\mathcal{A}_\ell)\frac{\eta_\ell^2}{N(d,\ell)} \tag{134}$$

where the final step follows from Lemma 14. By the construction of $\mathcal{A}_\ell$ (Lemma 17),

$$\dim(\mathcal{A}_\ell)\geq N(d,\ell)-H\cdot M(r,\ell) \tag{135}$$

and thus

$$\geq\sum_{\ell\text{ odd}}\left(1-H\cdot\frac{M(r,\ell)}{N(d,\ell)}\right)\eta_\ell^2 \tag{136}$$

Appealing either to Lemma 22 or to Lemma 23 finishes the proof. □

C.7 Asymptotics

Lemma 18. *Let $m>\ell$ and ℓ odd. Then*

$$\int_0^1\left(\frac{d}{dt}\right)^\ell(1-t^2)^m dt=(-1)^{1+(\ell-1)/2}\binom{m}{\frac{\ell-1}{2}}(\ell-1)!\,. \tag{137}$$

Proof. We have

$$\int_0^1 \left(\frac{d}{dt}\right)^\ell (1-t^2)^m dt = -\left(\frac{d}{dt}\right)^{\ell-1} (1-t^2)^m\Big|_{t=0} \tag{138}$$

$$= -\left(\frac{d}{dt}\right)^{\ell-1} \sum_{k=0}^{m} \binom{m}{k} (-1)^k t^{2k}\Big|_{t=0} \tag{139}$$

$$= (-1)^{1+(\ell-1)/2} \binom{m}{\frac{\ell-1}{2}} (\ell-1)! \,. \tag{140}$$

□

Lemma 19. *Define η_ℓ as in Definition 7. For odd ℓ, $\eta_\ell^2 \sim \sqrt{\frac{d}{\ell^3(\ell+d)}}$.*

Proof. From the definition, we have

$$\eta_{l,d} = 2\frac{\sqrt{N(d,l)}A_{d-2}}{A_{d-1}} \int_0^1 P_{l,d}(t)(1-t^2)^{(d-3)/2} dt \,. \tag{141}$$

From the Rodrigues formula for $P_{l,d}$ (Frye & Efthimiou, 2012, Proposition 4.19), we have

$$\eta_{l,d} = 2\frac{\sqrt{N(d,l)}A_{d-2}}{A_{d-1}} \frac{(-1)^l}{2^l (l+(d-3)/2)_l} \int_0^1 \left(\frac{d}{dt}\right)^l (1-t^2)^{l+(d-3)/2} dt \,. \tag{142}$$

Now, using Lemma 18, we obtain

$$\eta_{l,d} = 2\frac{\sqrt{N(d,l)}A_{d-2}}{A_{d-1}} \frac{(-1)^l}{2^l (l+(d-3)/2)_l} (-1)^{1+(l-1)/2} \binom{l+(d-3)/2}{\frac{l-1}{2}} (l-1)! \,, \tag{143}$$

and thus, using $\frac{A_{d-2}}{A_{d-1}} \sim C'\sqrt{d}$, we have

$$|\eta_{l,d}| \sim C\sqrt{d}\sqrt{N(d,l)} 2^{-l} \frac{(l-1)!((d-3)/2)!}{(l+(d-3)/2)!} \binom{l+(d-3)/2}{\frac{l-1}{2}} \,. \tag{144}$$

$$= C\frac{\sqrt{d}}{l}\sqrt{N(d,l)} 2^{-l} \frac{\binom{l+(d-3)/2}{\frac{l-1}{2}}}{\binom{l+(d-3)/2}{l}} \tag{145}$$

$$= C\frac{\sqrt{d}}{l}\sqrt{N(d,l)} 2^{-l} \frac{l!\left(\frac{d-3}{2}\right)!}{\left(\frac{l-1}{2}\right)!\left(\frac{d+l-2}{2}\right)!} \,. \tag{146}$$

Using Stirling's approximation, we obtain

$$N(d,l) \sim \frac{l+d}{l}\left(\frac{l+d}{ld}\right)^{1/2} \frac{(l+d)^{(l+d-3)}}{l^{(l-1)}d^{(d-2)}} \tag{147}$$

$$\sim (l+d)^{l+d-3/2} l^{-l-1/2} d^{-d+3/2} \,, \tag{148}$$

as well as

$$\frac{l!\left(\frac{d-3}{2}\right)!}{\left(\frac{l-1}{2}\right)!\left(\frac{d+l-2}{2}\right)!} \sim \sqrt{\frac{ld}{l(d+l)}} l^{(l+1)/2} d^{(d-3)/2} (d+l)^{(-d-l+2)/2} 2^l \,, \tag{149}$$

$$\sim (l+d)^{(-d-l+1)/2} l^{(l+1)/2} d^{(d-2)/2} 2^l \,, \tag{150}$$

leading to

$$|\eta_{l,d}| \sim (l+d)^{(-d-l+1+l+d)/2-3/4} l^{-1-l/2-1/4+l/2+1/2} d^{1/2-d/2+3/4+d/2-1} \tag{151}$$

$$\sim (l+d)^{-1/4} l^{-3/4} d^{1/4} \,, \tag{152}$$

as claimed. □

Lemma 19 shows that η_ℓ^2 decays slowly with ℓ. Using $\sqrt{\frac{d}{\ell^3(\ell+d)}} \geq 1/\ell^2$ and including by a fudge factor c that is slightly smaller than 1, we get a form that is better suited to the proof of our lower bounds:

Corollary 20. *There exists a universal constant c'' such that $\eta_\ell^2 \geq c''/\ell^2$ for all sufficiently large d and ℓ (say, for all $d, \ell > 4$).*

Lemma 21 (Decay of α_ℓ)**.** *For ℓ odd, we have $\alpha_\ell = \frac{1}{4}\eta_\ell^2/\sqrt{N(d,\ell)}$.*

Proof. We start from $\arcsin = \pi/2 - \arccos$ and the kernel representation (Bach, 2017a, Section 3.1)

$$\frac{1}{2\pi}(\pi - \arccos(x \cdot y)) = \mathbb{E}_{\theta \in \mathbb{S}^{d-1}} \left[\mathbf{1}[x \cdot \theta > 0]\mathbf{1}[y \cdot \theta > 0]\right] . \tag{153}$$

Now, from the Hecke-Funk formula, we have, up to zeroth-harmonic terms, the following correspondence between the Gegenbauer expansion of $\arcsin$ and that of of sign, given precisely by η_ℓ. Fix any $\boldsymbol{x} \in \mathbb{S}^{d-1}$. Then

$$\begin{aligned} P_\ell(1)\langle \arcsin, P_\ell \rangle &= \int \arcsin(\boldsymbol{x} \cdot \boldsymbol{y}) P_\ell(\boldsymbol{x} \cdot \boldsymbol{y}) \tau(d\boldsymbol{y}) \\ &= \frac{1}{4} \int \int \operatorname{sign}(\boldsymbol{x} \cdot \theta) \operatorname{sign}(\boldsymbol{y} \cdot \theta) P_\ell(\boldsymbol{x} \cdot \boldsymbol{y}) \tau(d\boldsymbol{y}) \tau(d\theta) \\ &= \frac{1}{4} \langle \operatorname{sign}, P_\ell \rangle \int \operatorname{sign}(\boldsymbol{x} \cdot \theta) P_\ell(\boldsymbol{x} \cdot \theta) \tau(d\theta) \\ &= \frac{1}{4} P_\ell(1) \langle \operatorname{sign}, P_\ell \rangle^2 . \end{aligned} \tag{154}$$

Since $\langle \arcsin, P_\ell \rangle = \alpha_\ell \|P_\ell\|$ and $\langle \operatorname{sign}, P_\ell \rangle = \eta_\ell \|P_\ell\|$, so $\alpha_\ell = \frac{1}{4}\eta_\ell^2 \|P_\ell\| = \frac{1}{4}\eta_\ell^2/\sqrt{N(d,\ell)}$.

□

Lemma 22. *There are universal constants c and C such that the following hold: Assume $r \leq d-3$, $\epsilon \leq \frac{c}{d+1}$, and $H \leq C \cdot 2^{d-(r+1)\log_2(2d/r)}$. Then*

$$\sum_{\ell \text{ odd}} \left(1 - H \cdot \frac{M(r,\ell)}{N(d,\ell)}\right) \eta_\ell^2 \geq \epsilon \tag{155}$$

Proof.

$$N(d,\ell) = \frac{2\ell + d - 2}{\ell} \binom{\ell + d - 3}{\ell - 1} \tag{156}$$

Applying Stirling's approximation,

$$N(d,\ell) \gtrsim \frac{\ell + d - 3}{\ell} \frac{(\ell + d - 3)^{\ell + d - 2.5}}{(\ell - 1)^{\ell - 0.5}(d-2)^{d-1.5}} \tag{157}$$

$$\geq \frac{(\ell + d - 3)^{\ell + d - 1.5}}{\ell^{\ell + 0.5}(d-2)^{d-1.5}} \tag{158}$$

Meanwhile, Lemma 15 and Stirling's approximation give

$$M(r,\ell) \leq \binom{r + \ell}{\ell} \lesssim \frac{(r+\ell)^{r+\ell+0.5}}{r^{r+0.5}\ell^{\ell+0.5}} \tag{159}$$

By assumption, $r \leq d - 3$, so

$$\frac{M(r,\ell)}{N(d,\ell)} \lesssim \left(\frac{r + \ell}{\ell + d - 3}\right)^{r+\ell+0.5} \frac{(d-2)^{d-1.5}}{r^{r+0.5}(\ell + d - 3)^{d-r-2}} \tag{160}$$

$$\leq \frac{(d-2)^{d-1.5}}{r^{r+0.5}(\ell + d - 3)^{d-r-2}} \tag{161}$$

The above expression is decreasing in ℓ. Thus for all $\ell \geq \mu d + 1$,

$$\frac{M(r,\ell)}{N(d,\ell)} \lesssim \frac{(d-2)^{d-1.5}}{r^{r+0.5}((1+\mu)(d-2))^{d-r-2}} \tag{162}$$

$$\leq \left(\frac{d}{r}\right)^{r+0.5} \frac{1}{(1+\mu)^{d-r-2}} \tag{163}$$

$$= (1+\mu)^{-d+r+2+(r+0.5)\log_{1+\mu}(d/r)} \tag{164}$$

By assumption, $c/\epsilon \geq d+1$, so the above holds with $\mu = 1$ for all $\ell \geq c/\epsilon$:

$$\frac{M(r,\ell)}{N(d,\ell)} \lesssim 2^{-d+(r+1)\log_2(2d/r)} \tag{165}$$

Also by assumption, $H \leq C \cdot 2^{d-(r+1)\log_2(2d/r)}$. Setting C appropriately, $\left(1 - H \cdot \frac{M(r,\ell)}{N(d,\ell)}\right) \geq \frac{1}{2}$ for all $\ell \geq c/\epsilon$ Finally, applying Corollary 20,

$$\sum_{\ell \text{ odd}} \left(1 - H \cdot \frac{M(r,\ell)}{N(d,\ell)}\right) \eta_\ell^2 \geq \sum_{\substack{\ell \geq c/\epsilon \\ \ell \text{ odd}}} \frac{1}{2} \cdot \frac{c''}{\ell^2} \tag{166}$$

$$\geq \frac{c''}{4} \sum_{\ell \geq c/\epsilon} \frac{1}{\ell^2} \tag{167}$$

$$\geq \frac{c''}{4} \cdot \frac{\epsilon}{c} \tag{168}$$

Setting $c = c''/4$ completes the proof. □

Lemma 23. *There is a universal constant c such that the following holds. If $d \geq 5$,*

$$\frac{2c}{\epsilon} < \frac{d}{2e^2} - r\,, \tag{169}$$

and

$$H \leq \frac{1}{2}\left(\frac{1}{2e} \cdot \frac{d}{r + \frac{c}{\epsilon}}\right)^{\frac{c}{\epsilon}}\,, \tag{170}$$

then

$$\sum_{\ell \text{ odd}} \left(1 - H \cdot \frac{M(r,\ell)}{N(d,\ell)}\right) \eta_\ell^2 \geq \epsilon \tag{171}$$

$$\tag{172}$$

Proof. Recall the formula for $N(d,\ell)$ from Equation (19). Lower bounding, for $\ell \geq 1$ and $d \geq 5$,

$$N(d,\ell) = \frac{2\ell + d - 2}{\ell}\binom{\ell + d - 3}{\ell - 1} \tag{173}$$

$$\geq \frac{\ell + d - 3}{\ell}\left(\frac{\ell + d - 3}{\ell - 1}\right)^{\ell-1} \geq \left(\frac{\ell + d - 3}{\ell}\right)^{\ell} \tag{174}$$

$$\geq \left(\frac{d+\ell}{2\ell}\right)^{\ell} \tag{175}$$

$$\tag{176}$$

Meanwhile, Lemma 15 gives

$$M(r,\ell) \leq \binom{r+\ell}{\ell} \leq \left(\frac{e(r+\ell)}{\ell}\right)^{\ell} \tag{177}$$

Thus

$$\frac{M(r,\ell)}{N(d,\ell)} \leq \left(2e \cdot \frac{r+\ell}{d+\ell}\right)^{\ell} \leq \left(2e \cdot \frac{r+\ell}{d}\right)^{\ell} \tag{178}$$

The above is a decreasing function of ℓ for all $\ell < \frac{d}{2e^2} - r$. Assume that $\frac{2c}{\epsilon} < \frac{d}{2e^2} - r$. Then the following holds for all $\ell \in \left[\frac{c}{\epsilon}, \frac{2c}{\epsilon}\right]$:

$$\frac{M(r,\ell)}{N(d,\ell)} \leq \left(2e \cdot \frac{r + \frac{c}{\epsilon}}{d}\right)^{\frac{c}{\epsilon}} \tag{179}$$

Assume $H \leq \frac{1}{2}\left(\frac{1}{2e} \cdot \frac{d}{r+\frac{c}{\epsilon}}\right)^{\frac{c}{\epsilon}}$. Then for all $\ell \in \left[\frac{c}{\epsilon}, \frac{2c}{\epsilon}\right]$:

$$1 - H \cdot \frac{M(r,\ell)}{N(d,\ell)} \geq \frac{1}{2} \tag{180}$$

Finally, applying Corollary 20,

$$\sum_{\ell \text{ odd}} \left(1 - H \cdot \frac{M(r,\ell)}{N(d,\ell)}\right) \eta_\ell^2 \geq \frac{1}{2} \sum_{\ell \text{ odd}} \frac{c''}{\ell^2} \geq \frac{c''}{4} \sum_{\ell=c/\epsilon}^{2c/\epsilon} \frac{1}{\ell^2} \geq \frac{c''}{4} \cdot \frac{\epsilon}{2c} \tag{181}$$

Setting $c = c''/8$ completes the proof. □

C.8 Kernel Ridge Regression and Random Feature Approximation

In this section, we analyze a simple approximation of the nearest neighbor function by standard rank-1 attention heads. We show that $O(\epsilon^{-4} d^{2/\epsilon})$ heads suffice to achieve a squared approximation error of ϵ, nearly matching the lower bound of Theorem 2. First, we reduce this problem to approximating the surrogate target function $\tilde{f}$ by rank-1 hardmax heads. Then we approximate $\tilde{f}$ in the RKHS generated by rank-1 hardmax attention heads (that is, generated by the feature map $\mathcal{T}$). Finally, we appeal to standard arguments to conclude that we can approximate $\tilde{f}$ by a finite linear combination of random rank-1 hardmax heads.

Recall that a standard rank-1 attention layer has the form $\sum_h \boldsymbol{o}_h \boldsymbol{v}_h^\top \boldsymbol{X} \mathrm{sm}\left(\boldsymbol{X}^\top \boldsymbol{k}_h \boldsymbol{q}_h^\top \boldsymbol{y}\right)$ for $\boldsymbol{q}_h, \boldsymbol{k}_h, \boldsymbol{v}_h, \boldsymbol{o}_h \in \mathbb{R}^d$. For simplicity, in this section we use rank-1 heads without a value/output transform, that is $\sum_h \alpha_h \boldsymbol{X} \mathrm{sm}\left(\boldsymbol{X}^\top \boldsymbol{k}_h \boldsymbol{q}_h^\top \boldsymbol{y}\right)$ for $\alpha \in \mathbb{R}$. Any such head can be constructed out of d standard rank-1 heads by setting $\boldsymbol{v}_h = \boldsymbol{e}_i, \boldsymbol{v}_o = \alpha \boldsymbol{e}_i$ for $i \in [d]$, so this simplification does not meaningfully change our result.

Lemma 24. *For any $u \in L^1(\Omega)$, there exists a rank-1 attention layer that approximates the nearest neighbor function f up to expected squared error $\frac{1}{2}\|\tilde{f} - \mathcal{T}u\|_\tau^2$, where $\mathcal{T}$ is defined as in Definition 8 and $\tilde{f}$ is the surrogate target function of Definition 12.*

Proof. As in the proof of Theorem 2, define

$$\boldsymbol{x} = \frac{\boldsymbol{x}_1 - \boldsymbol{x}_2}{\sqrt{2}}, \qquad \boldsymbol{w} = \frac{\boldsymbol{x}_1 + \boldsymbol{x}_2}{\sqrt{2}}. \tag{182}$$

We can rewrite the target function in terms of the surrogate target function as follows:

$$f(\boldsymbol{x}_1, \boldsymbol{x}_2; \boldsymbol{y}) = \frac{\boldsymbol{x}}{\sqrt{2}} \tilde{f}(\boldsymbol{x}, \boldsymbol{y}) + \frac{\boldsymbol{w}}{\sqrt{2}}. \tag{183}$$

Likewise, we can write a rank-1 hardmax attention head as

$$\boldsymbol{X} \mathrm{hm}\left(\boldsymbol{X}^\top \boldsymbol{k} \boldsymbol{q}^\top \boldsymbol{y}\right) = \frac{\boldsymbol{x}}{\sqrt{2}} \rho(\boldsymbol{x}, \boldsymbol{y}; \boldsymbol{q}, \boldsymbol{k}) + \frac{\boldsymbol{w}}{\sqrt{2}},$$

where $\rho(\boldsymbol{x}, \boldsymbol{y}; \boldsymbol{q}, \boldsymbol{k}) := \mathrm{sgn}(\boldsymbol{x}^\top \boldsymbol{k} \boldsymbol{q}^\top \boldsymbol{y})$ is defined as in Equation (28). An "averaging head" is an attention head that always returns the average of the target points, regardless of the source point. It can be implemented by a rank-1 softmax head by setting $\boldsymbol{q} = \boldsymbol{k} = \boldsymbol{0}$:

$$\boldsymbol{X} \mathrm{sm}\left(\boldsymbol{X}^\top \boldsymbol{0} \boldsymbol{y}\right) = \frac{\boldsymbol{x}_1 + \boldsymbol{x}_2}{2} = \frac{\boldsymbol{w}}{\sqrt{2}}.$$

We construct our approximation to f by taking a linear combination of hardmax heads with coefficients given by u plus a single averaging head with coefficient $1 - \int_\Omega u(\boldsymbol{q}, \boldsymbol{k}) d\bar{\tau}(\boldsymbol{q}, \boldsymbol{k})$:

$$(\boldsymbol{X}, \boldsymbol{y}) \mapsto \int_\Omega u(\boldsymbol{q}, \boldsymbol{k})\ \boldsymbol{X}\ \mathrm{hm}\left(\boldsymbol{X}^\top \boldsymbol{k} \boldsymbol{q}^\top \boldsymbol{y}\right) d\bar{\tau}(\boldsymbol{q}, \boldsymbol{k}) + \left(1 - \int_\Omega u(\boldsymbol{q}, \boldsymbol{k}) d\bar{\tau}(\boldsymbol{q}, \boldsymbol{k})\right) \frac{\boldsymbol{x}_1 + \boldsymbol{x}_2}{2} \,. \tag{184}$$

To analyze its error, we use the Pythagorean theorem. Due to the averaging head, the projection of the error onto $\boldsymbol{w}$ is zero. What remains is the projection of the error onto $\boldsymbol{x}$:

$$\mathop{\mathbb{E}}_{\substack{\boldsymbol{x}_1, \boldsymbol{x}_2 \sim \mathcal{D}_2(\mathbb{S}^{d-1}) \\ \boldsymbol{y} \sim \mathrm{Unif}(\mathbb{S}^{d-1})}} \left\| f(\boldsymbol{X}; \boldsymbol{y}) - \left[\int_\Omega u(\boldsymbol{q}, \boldsymbol{k})\ \boldsymbol{X}\ \mathrm{hm}\left(\boldsymbol{X}^\top \boldsymbol{k} \boldsymbol{q}^\top \boldsymbol{y}\right) d\bar{\tau}(\boldsymbol{q}, \boldsymbol{k}) + \left(1 - \int_\Omega u(\boldsymbol{q}, \boldsymbol{k}) d\bar{\tau}(\boldsymbol{q}, \boldsymbol{k})\right) \frac{\boldsymbol{w}}{\sqrt{2}} \right] \right\|^2 \tag{185}$$

$$= \mathop{\mathbb{E}}_{\boldsymbol{x}, \boldsymbol{y} \sim \mathrm{Unif}(\mathbb{S}^{d-1})} \frac{1}{2} \left(\tilde{\boldsymbol{x}}^\top f(\boldsymbol{X}; \boldsymbol{y}) - \int_\Omega u(\boldsymbol{q}, \boldsymbol{k})\ \boldsymbol{x}^\top \boldsymbol{X}\ \mathrm{hm}\left(\boldsymbol{X}^\top \boldsymbol{k} \boldsymbol{q}^\top \boldsymbol{y}\right) d\bar{\tau}(\boldsymbol{q}, \boldsymbol{k}) \right)^2 =: \frac{1}{2} \left\| \tilde{f} - \mathcal{T} u \right\|_{\bar{\tau}}^2 \tag{186}$$

$$= \mathop{\mathbb{E}}_{\boldsymbol{x}, \boldsymbol{y} \sim \mathrm{Unif}(\mathbb{S}^{d-1})} \frac{1}{2} \left(\tilde{f}(\boldsymbol{x}, \boldsymbol{y}) - \int_\Omega \rho(\boldsymbol{x}, \boldsymbol{y}; \boldsymbol{q}, \boldsymbol{k}) u(\boldsymbol{q}, \boldsymbol{k}) d\bar{\tau}(\boldsymbol{q}, \boldsymbol{k}) \right)^2 =: \frac{1}{2} \left\| \tilde{f} - \mathcal{T} u \right\|_{\bar{\tau}}^2 \,. \tag{187}$$

□

By the above lemma, our task is to find a finitely supported signed measure u for which $\tilde{f} \approx \mathcal{T} u$. We next show that it is possible to exactly represent $\tilde{f}$ using a measure that is not finitely supported.

Lemma 25. *The surrogate target function $\tilde{f}$ lies in the span of $\{\mathcal{T}(Y_\ell^i \otimes Y_\ell^i)\}$. Furthermore, $\tilde{f} = \mathcal{T} u$ where $u : \Omega \to \mathbb{R}$ is defined as follows:*

$$u(\omega) = \frac{\pi}{2} \sum_{\ell\ odd} \frac{\eta_\ell}{\alpha_\ell} N(d, \ell) \cdot P_\ell(\boldsymbol{q}^\top \boldsymbol{k}) \,. \tag{188}$$

Proof. For each odd ℓ, let $\{Y_\ell^i\}_{i=1}^{N(d,\ell)}$ be an orthonormal basis for $\mathcal{F}_\ell$. Applying Lemma 14, the norm of the projection of $\tilde{f}$ onto the span of $\{\mathcal{T}(Y_\ell^i \otimes Y_\ell^i)\}$ is

$$\sum_{i=1}^{N(d,\ell)} \left\langle \tilde{f}, \frac{\mathcal{T}(Y_\ell^i \otimes Y_\ell^i)}{\|\mathcal{T}(Y_\ell^i \otimes Y_\ell^i)\|_{\bar{\tau}}} \right\rangle_{\bar{\tau}}^2 = \sum_{i=1}^{N(d,\ell)} \frac{\eta_\ell^2}{N(d, \ell)} = \eta_\ell^2 \,. \tag{189}$$

Summing across all (odd) degrees, the energy equals that of $\tilde{f}$ itself.

$$\sum_{\ell=0}^{\infty} \eta_{2\ell+1}^2 = \|\operatorname{sign}\|_{\bar{\tau}}^2 = 1 = \|\tilde{f}\|_{\bar{\tau}}^2 \,. \tag{190}$$

Thus, the projection of $\tilde{f}$ onto this basis equals $\tilde{f}$. In addition, this implies that $\tilde{f}$ is in the range of $\mathcal{T}$:

$$\tilde{f} = \sum_{\ell\ \mathrm{odd}} \sum_{i=1}^{N(d,\ell)} \left\langle \tilde{f}, \frac{\mathcal{T}(Y_\ell^i \otimes Y_\ell^i)}{\|\mathcal{T}(Y_\ell^i \otimes Y_\ell^i)\|_{\bar{\tau}}} \right\rangle_{\bar{\tau}} \frac{\mathcal{T}(Y_\ell^i \otimes Y_\ell^i)}{\|\mathcal{T}(Y_\ell^i \otimes Y_\ell^i)\|_{\bar{\tau}}} \tag{191}$$

$$= \sum_{\ell\ \mathrm{odd}} \sum_{i=1}^{N(d,\ell)} \frac{\eta_\ell}{\sqrt{N(d,\ell)}} \cdot \frac{1}{\frac{2}{\pi} \alpha_\ell \sqrt{N(d,\ell)}} \mathcal{T}(Y_\ell^i \otimes Y_\ell^i) \tag{192}$$

$$= \mathcal{T}\left(\frac{\pi}{2} \sum_{\ell\ \mathrm{odd}} \frac{\eta_\ell}{\alpha_\ell} \sum_{i=1}^{N(d,\ell)} (Y_\ell^i \otimes Y_\ell^i) \right) \tag{193}$$

$$= \mathcal{T}(u) \,, \tag{194}$$

where, by the addition formula,

$$u(\omega) = \frac{\pi}{2} \sum_{\ell\ \mathrm{odd}} \frac{\eta_\ell}{\alpha_\ell} N(d, \ell) \cdot P_\ell(\boldsymbol{q}^\top \boldsymbol{k}) \,. \tag{195}$$

□

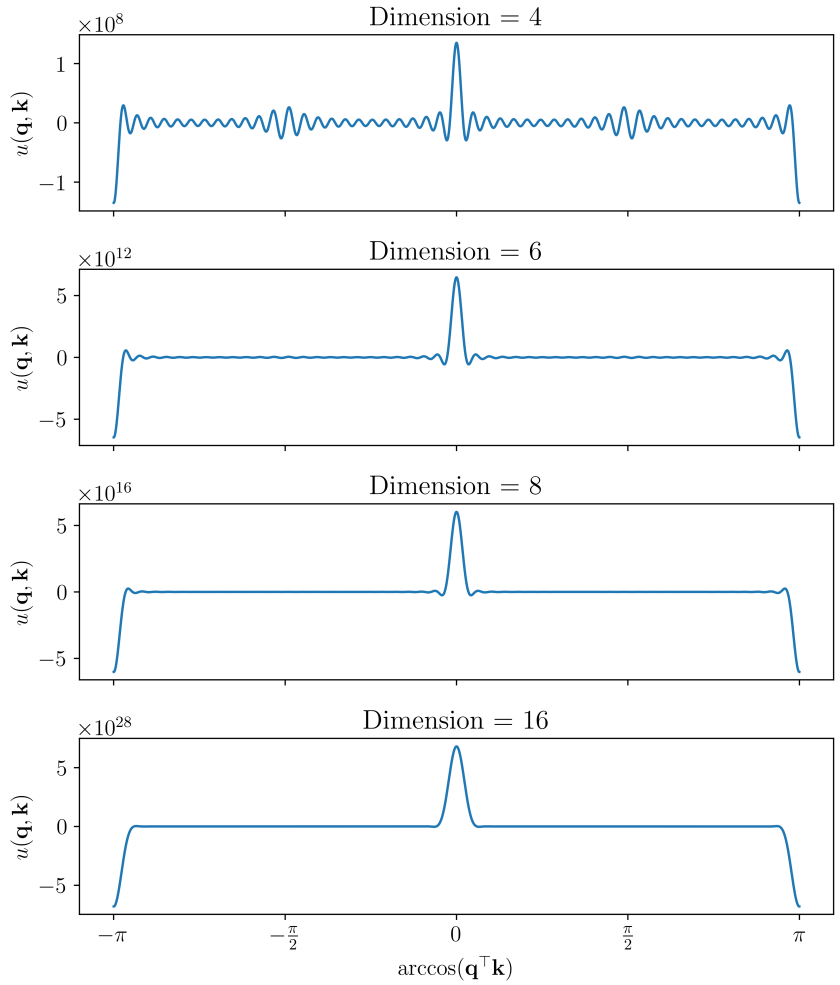

Figure 8: Approximation to $u(\cdot)$ of Equation (188) for several dimensions, using degree-50 ultraspherical expansion. Heads with $\angle(\boldsymbol{q}, \boldsymbol{k}) = \theta$ are equivalent to those with angles $\theta \pm \pi$ up to a sign flip. For large dimension, the distribution over $\angle(\boldsymbol{q}, \boldsymbol{k})$ induced by u approaches a Gaussian with mean 0.

Thus, it is possible to exactly represent the surrogate target with an infinite number of rank-1 heads, each weighted according to $u(\cdot)d\bar{\tau}(\cdot)$. See Figure 8 for an illustration of this function. We can think of $u(\cdot)d\bar{\tau}(\cdot)$ as a signed measure over rank-1 heads that depends only on $\angle(\boldsymbol{q}, \boldsymbol{k})$. Notice that the hardmax head function ρ is odd in each of its arguments $\boldsymbol{q}$ and $\boldsymbol{k}$. Since $u(\cdot)$ is also an odd function, we get the same results by restricting this measure to $[-\frac{\pi}{2}, \frac{\pi}{2}]$. Figure 8 shows that for large d, the (restricted) measure $u(\cdot)$ approaches a Gaussian distribution centered at angle 0.

We have now shown how to represent $\tilde{f}$ using $\mathcal{T}$. This representation gives us a great deal of insight into the structure of $\tilde{f}$ for the following reason. Implicit in the discussion above is the reproducing kernel Hilbert structure induced by the map $\mathcal{T}$, as the following lemma shows:

Lemma 26. *Let $\mathcal{H} \subseteq L^2(\mathcal{X})$ be the image of $\mathcal{T}$. Then $\mathcal{H}$ is a reproducing kernel Hilbert space with norm:*

$$\|f\|_{\mathcal{H}} = \inf\{\|u\|_{\bar{\tau}} : u \in \mathcal{G}, f = \mathcal{T}u\} \tag{196}$$

and kernel:

$$(z, z') \mapsto \mathop{\mathbb{E}}_{\omega \sim \bar{\tau}} [\rho(z, \omega)\rho(z', \omega)] \ . \tag{197}$$

The proof is given in Bach (2017a), Appendix A. Also note that kernel of this RKHS directly corresponds to the operator $\mathcal{T}\mathcal{T}^*$ by the following formula:

$$\left(\mathcal{T}\mathcal{T}^* f\right)(z) = \int_{\mathcal{X}} \mathop{\mathbb{E}}_{\omega\sim\bar{\tau}} \left[\rho(z,\omega)\rho(z',\omega)\right] f(z') d\bar{\tau}(z') \,. \tag{198}$$

If our target function $\tilde{f}$ were an element of this Hilbert space, we would immediately be able to approximate it using random features. Unfortunately, $\tilde{f} \notin \mathcal{H}$ because

$$\|\tilde{f}\|_{\mathcal{H}} = \|u\|_{\bar{\tau}} = \sum_{\ell \text{ odd}} \left(\frac{\eta_\ell}{\alpha_\ell}\right)^2 N(d,\ell) = \infty \,. \tag{199}$$

However, we can approximate f by an element of $\mathcal{H}$ obtained from solving a ridge regression problem. For any $\lambda > 0$, let $\tilde{f}_\lambda$ be the solution to the following optimization problem:

$$\min_{\hat{f}\in\mathcal{H}} \|\tilde{f} - \hat{f}\|_{\bar{\tau}}^2 + \lambda \|\hat{f}\|_{\mathcal{H}}^2 \,. \tag{200}$$

By tuning λ, we can find an function that accurately approximates $\tilde{f}$ and that is smooth enough to be approximated using random features. The following lemma constructs this $\tilde{f}_\lambda$. Though we obtained this construction by solving Equation (200), for brevity we do not prove that it is the solution since it is not necessary for our construction.

Lemma 27. *For any regularization parameter $\lambda > 0$, there exists a function $\tilde{f}_\lambda \in \mathcal{H}$ for which*

$$\|\tilde{f} - \tilde{f}_\lambda\|_{\bar{\tau}}^2 \leq \sum_{\ell \text{ odd}} \eta_\ell^2 \left(\frac{\lambda N(d,\ell)}{(\frac{2}{\pi}\alpha_\ell)^2 + \lambda N(d,\ell)}\right)^2 \,. \tag{201}$$

Proof. Define

$$\tilde{f}_\lambda := \mathcal{T} g_\lambda \tag{202}$$

$$g_\lambda := \sum_{\ell \text{ odd}} \sum_{i=1}^{N(d,\ell)} \gamma_\ell \cdot (Y_\ell^i \otimes Y_\ell^i) \tag{203}$$

$$\gamma_\ell := \frac{\frac{2}{\pi}\alpha_\ell \eta_\ell}{(\frac{2}{\pi}\alpha_\ell)^2 + \lambda N(d,\ell)} \,. \tag{204}$$

Then by Lemma 26

$$\|\tilde{f}_\lambda\|_{\mathcal{H}}^2 \leq \|g_\lambda\|_{\bar{\tau}}^2 = \sum_{\ell \text{ odd}} N(d,\ell)\gamma_\ell^2 \tag{205}$$

$$\leq \sum_{\ell=1}^{\ell_\lambda} N(d,\ell)\gamma_\ell^2 + \sum_{\ell>\ell_\lambda} N(d,\ell)\eta_\ell^2 \left(\frac{\frac{2}{\pi}\alpha_\ell}{\lambda N(d,\ell)}\right)^2 \tag{206}$$

$$\leq \sum_{\ell=1}^{\ell_\lambda} N(d,\ell)\gamma_\ell^2 + \frac{1}{\lambda^2} \tag{207}$$

$$< \infty \,. \tag{208}$$

Thus $\tilde{f} \in \mathcal{H}$. Furthermore, by the representation $\tilde{f} = \mathcal{T}u$ of Lemma 25

$$\|\tilde{f} - \tilde{f}_\lambda\|_{\bar{\tau}}^2 = \|\mathcal{T}(u - g_\lambda)\|_{\bar{\tau}}^2 = \left\| \sum_{\ell \text{ odd}} \sum_{i=1}^{N(d,\ell)} \left(\frac{\pi}{2}\frac{\eta_\ell}{\alpha_\ell} - \gamma_\ell\right) \cdot \mathcal{T}(Y_\ell^i \otimes Y_\ell^i) \right\|_{\bar{\tau}}^2 \,. \tag{209}$$

By Lemma 11, this is equal to

$$= \sum_{\ell \text{ odd}} \sum_{i=1}^{N(d,\ell)} \left(\frac{\pi}{2} \frac{\eta_\ell}{\alpha_\ell} - \gamma_\ell \right)^2 \cdot \left\| \mathcal{T}(Y_\ell^i \otimes Y_\ell^i) \right\|_{\bar{\tau}}^2 \tag{210}$$

$$= \sum_{\ell \text{ odd}} N(d,\ell) \left(\frac{\pi}{2} \frac{\eta_\ell}{\alpha_\ell} - \gamma_\ell \right)^2 \frac{4}{\pi^2} \frac{\alpha_\ell^2}{N(d,\ell)} \tag{211}$$

$$= \sum_{\ell \text{ odd}} \left(\eta_\ell - \frac{2}{\pi} \alpha_\ell \gamma_\ell \right)^2 \tag{212}$$

$$= \sum_{\ell \text{ odd}} \eta_\ell^2 \left(\frac{\lambda N(d,\ell)}{(\frac{2}{\pi}\alpha_\ell)^2 + \lambda N(d,\ell)} \right)^2 . \tag{213}$$

□

We now derive an informal expression for the kernel ridge regression approximation using a tuned regularization and describe its implications for random feature approximation in the high-dimensional regime. From Lemma 19 and Lemma 21, we have $\eta_\ell^2 \lesssim \ell^{-3/2}$ and $\alpha_\ell^2 \sim \eta_\ell^4 / N(d,\ell)$. By Lemma 27, for the kernel ridge regression approximation $\tilde{f}_\lambda$ to attain squared error ϵ, we should set λ so that $\lambda N(d,\ell^*) \simeq \alpha_{\ell^*}^2$, where $\ell^* \sim 1/\epsilon^2$. This roughly ensures that only degrees $\ell \gtrsim \ell^*$ are kept, while $\ell \lesssim \ell^*$ are shrunk, and hence

$$\|\tilde{f} - \tilde{f}_\lambda\| \lesssim \sum_{\substack{\ell \gtrsim \ell^* \\ \ell \text{ odd}}} \eta_\ell^2 \lesssim \frac{1}{2} \sum_{\ell \gtrsim \epsilon^{-2}} \ell^{-3/2} \sim \epsilon \, . \tag{214}$$

We thus obtain $\lambda \sim \alpha_{\ell^*}^2 / N(d,\ell^*) \sim \epsilon^6 N(d,\epsilon^{-2})^{-2}$.

Now that we have a sufficiently accurate kernel ridge regression approximation $\tilde{f}_\lambda \in \mathcal{H}$, we can approximate it using random features. The key quantity controlling the number of random features needed is the *degrees of freedom* of the kernel integral operator, defined as $D(\lambda) := \operatorname{tr}\left[\mathcal{T}\mathcal{T}^*(\mathcal{T}\mathcal{T}^* + \lambda \boldsymbol{I})^{-1}\right]$. The eigenvalues of $\mathcal{T}\mathcal{T}^*$ are the same as those of $\mathcal{T}^*\mathcal{T}$. By Lemma 10, these are $\left\{ \frac{4}{\pi^2} \frac{\alpha_\ell \alpha_{\ell'}}{\sqrt{N(d,\ell)N(d,\ell')}} \mid \ell, \ell' \geq 0 \right\}$, with the (ℓ,ℓ')-th eigenvalue having multiplicity $N(d,\ell)N(d,\ell')$. Hence

$$D(\lambda) = \sum_{\ell,\ell'} N(d,\ell)N(d,\ell') \cdot \frac{\frac{\alpha_\ell \alpha_{\ell'}}{\sqrt{N(d,\ell)N(d,\ell')}}}{\frac{\alpha_\ell \alpha_{\ell'}}{\sqrt{N(d,\ell)N(d,\ell')}} + \lambda} \leq \sum_{\ell,\ell'} N(d,\ell)N(d,\ell') \cdot \frac{\frac{\alpha_\ell \alpha_{\ell'}}{\sqrt{N(d,\ell)N(d,\ell')}}}{\lambda} . \tag{215}$$

By Lemma 21, $\frac{\alpha_\ell \alpha_{\ell'}}{\sqrt{N(d,\ell)N(d,\ell')}} \sim \frac{\eta_\ell^2 \eta_{\ell'}^2}{N(d,\ell)N(d,\ell')}$, so

$$D(\lambda) \sim \frac{1}{\lambda} \sum_{\ell,\ell'} \eta_\ell^2 \eta_{\ell'}^2 = \frac{1}{\lambda} \left(\sum_\ell \eta_\ell^2 \right)^2 = \frac{1}{\lambda} \sim \frac{N(d,\epsilon^{-2})^2}{\epsilon^6} \lesssim \frac{1}{\epsilon^6} \cdot (ed\epsilon^2)^{2/\epsilon^2} \tag{216}$$

In the high-dimensional regime (where ϵ is fixed and d goes to infinity), $D(\lambda) = \Theta\left(d^{2/\epsilon^2}\right)$.

By standard arguments about random feature expansions (Bach, 2017b), if the number of random features H is of the order $H \gtrsim D(\lambda)\log(D(\lambda)) = \widetilde{\Theta}\left(d^{2/\epsilon^2}\right)$, then with high probability the random features achieve the same approximation accuracy ϵ as the associated kernel ridge regression solution $\tilde{f}_\lambda$. It is likely that a better rate can be obtained by drawing the random features from a problem-specific distribution instead of uniformly at random. Observe that the condition required by our lower bound in the rank-1 case has the same form, though a somewhat weaker dependence on d. It is $H \leq \frac{1}{2} N(d, \frac{1}{4\epsilon})$ or $H = O\left(d^{\frac{1}{4\epsilon}}\right)$ for sufficiently large d.[4]

[4]To see this from Equation (136), recall that $M(1,\ell) = 1$, follow the final steps of Lemma 22, and use the fact that we can replace c'' by 1 for large d.

D PROOFS FROM SECTION 5

D.1 PROOF INTUITION

Part (i) of Theorem 3 is easy to prove. We approximate each summand of the target (Equation (8)) by a single attention head with a large enough temperature. We note that this construction doesn't depend on the input distribution, and can be readily generalized to any other continuous distribution. The error parameter ϵ affects the necessary temperature (i.e., the magnitude of the weights), but not the number of heads.

The proof of part (ii) of Theorem 3 is more intricate. The crux of the proof is to construct a linear combination of threshold functions that behaves like a periodic function with high frequency. Our proof is inspired and extends the proof method of Yehudai & Shamir (2019) for separation between kernel methods and 2-layer neural networks. Note that each summand t of the target in Equation (8) can be written as

$$\arg\max_{\boldsymbol{x}_i} \langle \boldsymbol{x}_i, \boldsymbol{y} \rangle + b_{t,i} = \mathbb{1}(\langle \boldsymbol{x}_1 - \boldsymbol{x}_2, \boldsymbol{y} \rangle + b_t^* > 0)\boldsymbol{x}_1 + \mathbb{1}(\langle \boldsymbol{x}_1 - \boldsymbol{x}_2, \boldsymbol{y} \rangle + b_t^* < 0)\boldsymbol{x}_2 \ , \quad (217)$$

where $b_t^* = b_{t,1} - b_{t,2}$. Let $a := 2d^2 + 1$, and recall that $b_{t,i} = t$ for $i = 1$ and 0 for $i \neq 1$. Rewriting the target function using the above substitution, we can show that it is periodic in the interval $[-a, a]$, considered as a function of $\langle \boldsymbol{x}_1 - \boldsymbol{x}_2, \boldsymbol{y} \rangle$. Denote this function by ψ_a. We extend a result from Shamir (2018) to show that:

$$\mathop{\mathbb{E}}_{\boldsymbol{x}_1, \boldsymbol{x}_2, \boldsymbol{y}} [|\psi_a(\langle \boldsymbol{x}_1 - \boldsymbol{x}_2, \boldsymbol{y} \rangle) \cdot g(\boldsymbol{K}\boldsymbol{x}_1, \boldsymbol{K}\boldsymbol{x}_2, \boldsymbol{y})|] \leq \|g\| \cdot \exp(-\Omega(d - r)) \ . \quad (218)$$

for any function g with bounded outputs if $\boldsymbol{K}$ has rank r. Finally, a rank-r transformer with H heads can be written as a linear combination of such functions with bounded outputs, where the weights of the linear combination are $\boldsymbol{V}_h$.

We draw $\boldsymbol{y}$ from a Gaussian instead of uniformly from the unit sphere in order to use the result from Shamir (2018). There, the bound on the correlation is expressed in the Fourier domain. Since the Fourier transform of a Gaussian function is also a Gaussian function, the calculations become significantly easier.

D.2 FROM POLYNOMIAL TO EXPONENTIAL SEPARATION

As can be seen from Theorem 3, to get an exponential separation between low- and high-rank transformers, we used a more complex function than the nearest neighbor as used in Theorem 2. It is natural to ask whether this more complex function is necessary to achieve this strong separation.

We claim that it is not possible to achieve exponential separation using a target that is just one nearest neighbor function, even with a bias term. Fix some $r < d$, and consider some target function f that is hard to approximate using rank-r attention heads. There are now three important parameters that play a role in the separation result: (1) The desired approximation error ϵ; (2) The number of heads H required to achieve this desired error; and (3) The Lipschitz constant L of the function f.

In Hsu et al. (2021) it is proven that it is both necessary and sufficient to approximate any target function by polynomially many random ReLU neurons if both $\epsilon = O(1)$ and $L = O(1)$, that is, if they are both independent of d. However, if either $\epsilon = \frac{1}{\text{poly}(d)}$ or $L = \text{poly}(d)$, then exponentially many random ReLU neurons are required. Our results resemble approximation by random neurons (with ReLU or other activations), where the "effective" dimension is $d - r$. Using this analogy, Theorem 2 covers two cases: (1) $\epsilon = O(1)$ and $L = O(1)$, and thus the separation is polynomial; and (2) $\epsilon = O(1/d)$ and $L = O(1)$ where the separation is exponential.

Thus, to achieve exponential separation for $\epsilon = O(1)$ we need to construct a hard to approximate target function with Lipschitz constant $L = \text{poly}(d)$. One method is to "cheat" by multiplying an $L = O(1)$ target function by a large number that depends on d. However, this will not result in a small *relative* error; effectively, by multiplying the target by some constant, we also multiplied ϵ by the same constant. In Theorem 3, we construct a target function f^* with $\|f^*\| \leq O(1)$[5],

[5]This can be seen by equation 224 and Lemma 28, where the target for $N = 2$ can be written as a sum of two periodic functions, both bounded in $[-1, 1]$.

meaning that our hardness result applies to the relative error too. However, this target function is highly oscillatory, resulting in a Lipschitz constant of $L = \text{poly}(d)$, and thus we get exponential separation even for a relative error of $\epsilon = O(1)$. We believe that the result from Hsu et al. (2021) can be extended beyond random ReLU neurons. If so, then achieving an exponential separation with a "simple" function (i.e. Lipschitz constant independent of d) would be impossible.

We also note that a similar impossibility result is known for separation between 2- and 3-layer neural networks. Safran et al. (2019) shows that such an exponential separation cannot be achieved if both $\epsilon = O(1)$ and $L = O(1)$.

D.3 Proof of item (i) in Theorem 3

The proof is similar to the proof of Fact 1, where it is applied to each head separately. Namely, let $\epsilon > 0$, and for each $t = 1, \ldots, 2d^2+1$, set $\boldsymbol{V}_t = \boldsymbol{I}$, $\boldsymbol{K}\boldsymbol{Q}^\top = \alpha \boldsymbol{I}$ for $\alpha > 0$ to be chosen later. There exists $\delta > 0$ (which depends on ϵ) such that for the set:

$$A_{\delta,t} = \{(\boldsymbol{x}_1, \ldots, \boldsymbol{x}_N, \boldsymbol{y}) : \forall i \neq j,\ |(\boldsymbol{x}_i - \boldsymbol{x}_j)^\top \boldsymbol{y} > \delta\} \tag{219}$$

we have that $\Pr(((\boldsymbol{x}_1, \ldots, \boldsymbol{x}_N, \boldsymbol{y}) \notin A_{\delta,t}) < \frac{\epsilon}{4d^2+2}$. Again, note that in $A_{\delta,t}$, softmax converges to hardmax uniformly as $\alpha \to \infty$. In particular, there exists $\alpha > 0$ such that:

$$\sup_{(\boldsymbol{x}_1,\ldots,\boldsymbol{x}_N,\boldsymbol{y}) \in A_{\delta,t}} \left\| \boldsymbol{X} \operatorname{sm}\left(\alpha \left(\boldsymbol{X}^\top \boldsymbol{y} + \boldsymbol{b}_t\right)\right) - \left(\arg\max_i \|\boldsymbol{x}_i - \boldsymbol{y}\|^2 + b_{t,i}\right) \right\|^2 \leq \frac{\epsilon}{4d^2+2} \,. \tag{220}$$

We define an additional head with $\boldsymbol{V} = -\frac{N}{2}\boldsymbol{I}$, $\boldsymbol{K}\boldsymbol{Q}^\top = 0$. Since the attention matrix is zero, for any target vector $\boldsymbol{y}$ this head will attend to uniformly to the source vectors. Hence, the output of this head will be $-\frac{1}{2}\sum_{i=1}^n \boldsymbol{x}_i$. Summing over all the above heads approximates the target in expectation up to an error of ϵ , which finishes the proof.

D.4 Proof of item (ii) in Theorem 3

For algebraic simplicity, we first rescale the problem by a factor of $\sqrt{d}$ (see Lemma 31). Define the rescaled target function:

$$f^{**}(\boldsymbol{x}_1, \ldots, \boldsymbol{x}_N, \boldsymbol{y}) = \frac{1}{\sqrt{d}} \sum_{t=1}^{2d^2+1} (-1)^t \arg\max_{i \in \{1,\ldots,N\}} \left(\|\boldsymbol{x}_i - \boldsymbol{y}\|^2 + b_{t,i} \right) - \frac{1}{2\sqrt{d}} \sum_{i=1}^N \boldsymbol{x}_i \,, \tag{221}$$

We also draw $\boldsymbol{x}_1, \boldsymbol{x}_2$ and $\boldsymbol{y}$ from scaled versions of the uniform and Gaussian distributions, so that we now wish to lower bound the following:

$$\mathop{\mathbb{E}}_{\substack{\boldsymbol{x}_1,\boldsymbol{x}_2 \sim \mathcal{D}_2\left(\sqrt{d}\mathbb{S}^{d-1}\right) \\ \boldsymbol{y} \sim \mathcal{N}(0,\boldsymbol{I})}} \left[\|T(\boldsymbol{x}_1, \boldsymbol{x}_2, \boldsymbol{y}) - f^{**}(\boldsymbol{x}_1, \boldsymbol{x}_2, \boldsymbol{y})\|^2 \right] \tag{222}$$

In the following proof, when taking norms or inner products of functions, we always consider them to be in expectation over $\mathcal{N}(0, \boldsymbol{I})$, i.e., in expectation over $\boldsymbol{y}$. Whenever we consider the expectation over $\boldsymbol{x}_1$ and $\boldsymbol{x}_2$, we explicitly write the expectation symbol. To normalize the expectation over $\boldsymbol{y}$, we introduce the constant $c_d := \left(\frac{1}{\sqrt{2\pi}}\right)^d$.

We will first construct a periodic function on the real line using a linear combination of thresholds. Let $a \in \mathbb{N}_{>2}$ and denote $H_z(x) = \mathbb{1}(x + z \geq 0)$. We define the following functions:

$$\psi_a^+(x) = \sum_{n=1}^{2a} H_{a-n}(x) \cdot (-1)^n - \frac{1}{2}, \quad \psi_a^-(x) = \sum_{n=1}^{2a} H_{-(a-n)}(x) \cdot (-1)^n - \frac{1}{2}. \tag{223}$$

Note that since $N = 2$ and $\|\boldsymbol{x}_1\| = \|\boldsymbol{x}_2\|$, we can write the target f^{**} in the following sense:

$$f^{**}(\boldsymbol{x}_1, \boldsymbol{x}_2, \boldsymbol{y}) = \frac{1}{\sqrt{d}} \psi_a^+(\langle \boldsymbol{x}_1 - \boldsymbol{x}_2, \boldsymbol{y} \rangle) \boldsymbol{x}_1 + \frac{1}{\sqrt{d}} \psi_a^-(\langle \boldsymbol{x}_2 - \boldsymbol{x}_1, \boldsymbol{y} \rangle) \boldsymbol{x}_2 \,. \tag{224}$$

For ease of notation, we will denote $\psi_a := \psi_a^+$, and we will prove the following for ψ_a. However, note that similar proofs can be shown for ψ_a^- in the exact same manner, thus we omit them for brevity. First, ψ_a has the following properties:

Lemma 28. *The function $\psi_a(x)$ defined in Equation (223) satisfies that:*

1. *It is a periodic function in the interval $[-a, a]$, and odd if a is an odd number.*
2. *For every $\boldsymbol{w}$ with $\|\boldsymbol{w}\| \geq d$, if $a > \|\boldsymbol{w}\|$ then $\left\|\psi_a(\langle \boldsymbol{w}, \cdot\rangle)^2\right\|^2 \geq \frac{1}{40}$*

Proof. Let $x_0 \in [-a, a-2]$. There is $n_0 \in \{1, \ldots, 2a\}$ such that $\lceil x_0 \rceil, \lceil x_0+2 \rceil \in [a-n_0, a-n_0+2]$. For every $n < n_0$ or $n > n_0 + 2$ we have that $H_{a-n}(x_0) = H_{a-n}(x_0 + 2)$, since the bump in the threshold is either left of x_0 or right of $x_0 + 2$. We also have that $H_{a-n_0}(x_0) + H_{a-n_0+1}(x_0) = H_{a-n_0}(x_0+2) + H_{a-n_0+1}(x_0+2) = 1$. Hence $\psi_a(x_0) = \psi_a(x_0+2)$, which means it is a periodic function with a period of 2.

If a is an odd number, then for every $x_0 \in [-1, 0]$ we have $\psi_a(x_0) = -\frac{1}{2}$ and for every $x_0 \in [0, 1]$ we have $\psi_a(x_0) = \frac{1}{2}$. Since it is periodic with a period of 2, it is odd in the interval $[-a, a]$.

For the second item, since $\boldsymbol{x}$ has a spherically symmetric distribution, we can assume w.l.o.g that $\boldsymbol{w} = \|\boldsymbol{w}\| \boldsymbol{e}_1$. We now have that:

$$\|\psi_a(\langle \boldsymbol{w}, \cdot\rangle)\|^2 = c_d \int_{\boldsymbol{x} \in \mathbb{R}^d} |\psi_a(\langle \boldsymbol{w}, \boldsymbol{x}\rangle)|^2 e^{-\frac{\|\boldsymbol{x}\|^2}{2}} d\boldsymbol{x} \tag{225}$$

$$= c_d \int_{-\infty}^{\infty} |\psi_a(\|\boldsymbol{w}\| x_1)|^2 e^{-\frac{x_1^2}{2}} dx_1 \cdot \int_{-\infty}^{\infty} e^{-\frac{x_2^2}{2}} dx_2 \cdots \int_{-\infty}^{\infty} e^{-\frac{x_d^2}{2}} dx_d \tag{226}$$

$$= \frac{1}{\|\boldsymbol{w}\| \sqrt{2\pi}} \int_{-\infty}^{\infty} |\psi_a(z)|^2 e^{-\frac{z^2}{2\|\boldsymbol{w}\|^2}} dz \tag{227}$$

$$\geq \frac{1}{\|\boldsymbol{w}\| e\sqrt{2\pi}} \int_{-\sqrt{2}\|\boldsymbol{w}\|}^{\sqrt{2}\|\boldsymbol{w}\|} |\psi_a(z)|^2 dz \tag{228}$$

where we used that if $z \leq \sqrt{2}\|\boldsymbol{w}\|$ then $e^{-\frac{z^2}{2\|\boldsymbol{w}\|^2}} \leq e^{-1}$. Since $a > \|\boldsymbol{w}\|$, then in the interval $\left[-\sqrt{2}\|\boldsymbol{w}\|, \sqrt{2}\|\boldsymbol{w}\|\right]$ there are at least $\lfloor \|\boldsymbol{w}\| \rfloor$ intervals of the form $[n, n+2]$ for $n \in \{-a, ..., a-2\}$ where $\int_n^{n+2} |\psi_a(z)|^2 \geq \frac{1}{4}$. In total, we can bound the norm by:

$$\|\psi_a(\langle \boldsymbol{w}, \cdot\rangle)\|^2 \geq \frac{1}{4e\sqrt{2\pi}} \geq \frac{1}{40} \tag{229}$$

□

Next, we now show that the correlation of this function with any other function that depends only on $w_1, \ldots, w_r$ is small:

Theorem 29. *Let $g(w_1, \ldots, w_r, \boldsymbol{y})$ be some function that depends on the first r coordinates of $\boldsymbol{w}$ with $\sup_{\boldsymbol{x}} |g(\boldsymbol{x})| \leq 1$, and take $a = 2d^2 + 1$. Then, we have that:*

$$\mathop{\mathbb{E}}_{\boldsymbol{w} \sim \mathcal{U}(\sqrt{d}\mathbb{S}^{d-1})} \left[\mathop{\mathbb{E}}_{\boldsymbol{y} \sim \mathcal{N}(0, \boldsymbol{I})} \left[|\psi_a(\langle \boldsymbol{w}, \boldsymbol{y}\rangle) \cdot g(w_1, \ldots, w_r, \boldsymbol{y})| \right] \right] \leq \exp(-c(d-r)) \tag{230}$$

for some universal constant $c > 0$.

Proof. For a vector $\boldsymbol{v}$ denote by $\bar{\boldsymbol{v}}$ its last $d - r$ coordinates. Using the law of total expectation, we can rewrite the expectation in the following way:

$$\begin{aligned}
&\mathop{\mathbb{E}}_{\boldsymbol{w} \sim \mathcal{U}(\sqrt{d}\mathbb{S}^{d-1})} \left[\mathop{\mathbb{E}}_{\boldsymbol{y} \sim \mathcal{N}(0, \boldsymbol{I})} \left[|\psi_a(\langle \boldsymbol{w}, \boldsymbol{y}\rangle) \cdot g(w_1, \ldots, w_r, \boldsymbol{y})| \right] \right] \\
&= \mathop{\mathbb{E}}_{w_1, \ldots, w_r} \left[\mathop{\mathbb{E}}_{\bar{\boldsymbol{w}}} \mathop{\mathbb{E}}_{y_1, \ldots, y_r} \left[\mathop{\mathbb{E}}_{\bar{\boldsymbol{y}}} \left[\left| \psi_a \left(\sum_{i=1}^{r} w_i y_i + \langle \bar{\boldsymbol{w}}, \bar{\boldsymbol{y}} \rangle \right) \cdot g(w_1, \ldots, w_r, \boldsymbol{y}) \right| \middle| y_1, \ldots, y_r \right] \mid w_1, \ldots, w_r \right] \right] \\
&= \mathop{\mathbb{E}}_{w_1, \ldots, w_r} \mathop{\mathbb{E}}_{y_1, \ldots, y_r} \mathop{\mathbb{E}}_{\bar{\boldsymbol{w}}} \mathop{\mathbb{E}}_{\bar{\boldsymbol{y}}} \left[\left| \psi_a \left(\sum_{i=1}^{r} w_i y_i + \langle \bar{\boldsymbol{w}}, \bar{\boldsymbol{y}} \rangle \right) \cdot g(w_1, \ldots, w_r, \boldsymbol{y}) \right| \mid y_1, \ldots, y_r, w_1, \ldots, w_r \right] .
\end{aligned} \tag{231}$$

Namely, we consider the expectation conditioned on drawing the first r coordinates of both $\boldsymbol{w}$ and $\boldsymbol{y}$. Note that we could change the order of expectations since all the expectations are bounded and finite.

Let $\tilde{\psi}$ be a continuation of ψ_a from $[-a,a]$ to $\mathbb{R}$ such that it is periodic. Fix $w_1,\ldots,w_r,y_1,\ldots,y_r$ and denote by $s := \sum_{i=1}^{r} w_i y_i$ and $\|\bar{\boldsymbol{w}}\| = 2\rho$. Using Claim 30 we have that:

$$\mathop{\mathbb{E}}_{\bar{\boldsymbol{w}}\sim\mathcal{U}(2\rho\mathbb{S}^{d-r-1})}\left[\left|\left\langle g(\cdot),\tilde{\psi}(s+\langle\bar{\boldsymbol{w}},\cdot\rangle)\right\rangle\right|\right] \le c_1\cdot\left(\exp(-c_2(d-r))+\sum_{n=1}^{\infty}\exp(-n\rho^2)\right) . \quad (232)$$

Note that in the above equation, g is independent of $\bar{\boldsymbol{w}}$ (although it does depend on $w_1,\ldots,w_r$), and also that $\|g\|\le 1$ since $\sup_{\boldsymbol{x}}|g(\boldsymbol{x})|\le 1$ (recall that the norm is w.r.t a Gaussian measure).

We now have that:

$$\begin{aligned}&\mathop{\mathbb{E}}_{\bar{\boldsymbol{w}}\sim\mathcal{U}(2\rho\mathbb{S}^{d-r-1})}\left[\left|\langle g(\cdot),\psi_a(s+\langle\bar{\boldsymbol{w}},\cdot\rangle)\rangle\right|\right]\\ &\le \mathop{\mathbb{E}}_{\bar{\boldsymbol{w}}\sim\mathcal{U}(2\rho\mathbb{S}^{d-r-1})}\left[\left|\left\langle g(\cdot),\tilde{\psi}(s+\langle\bar{\boldsymbol{w}},\cdot\rangle)\right\rangle\right|\right]+\mathop{\mathbb{E}}_{\bar{\boldsymbol{w}}\sim\mathcal{U}(2\rho\mathbb{S}^{d-r-1})}\left[\left|\left\langle g(\cdot),\psi_a(s+\langle\bar{\boldsymbol{w}},\cdot\rangle)-\tilde{\psi}(s+\langle\bar{\boldsymbol{w}},\cdot\rangle)\right\rangle\right|\right]\end{aligned} \quad (233)$$

The first term in Equation (233) can be bounded by $c_1\cdot\left(\exp(-c_2(d-r))+\sum_{n=1}^{\infty}\exp(-n\rho^2)\right)$ by Equation (232). For the second term, by Cauchy-Schwartz we have that:

$$\mathop{\mathbb{E}}_{\bar{\boldsymbol{w}}}\left[\left|\left\langle g(\cdot),\psi_a(s+\langle\bar{\boldsymbol{w}},\cdot\rangle)-\tilde{\psi}(s+\langle\bar{\boldsymbol{w}},\cdot\rangle)\right\rangle\right|\right] \quad (234)$$

$$\le \|g\|\cdot\mathop{\mathbb{E}}_{\bar{\boldsymbol{w}}}\left[\left\|\psi_a(s+\langle\bar{\boldsymbol{w}},\cdot\rangle)-\tilde{\psi}(s+\langle\bar{\boldsymbol{w}},\cdot\rangle)\right\|\right] \quad (235)$$

$$\le \mathop{\mathbb{E}}_{\bar{\boldsymbol{w}}}[\Pr(|s+\langle\bar{\boldsymbol{w}},\bar{\boldsymbol{y}}\rangle|>a)] \quad (236)$$

where we used that $\|g\|\le 1$ and it is independent of $\bar{\boldsymbol{w}}$, and that $\psi_a(z)=\tilde{\psi}(z)$ for every $|z|\le a$. We have that $s+\langle\bar{\boldsymbol{w}},\bar{\boldsymbol{y}}\rangle=\langle\boldsymbol{w},\boldsymbol{y}\rangle$, and $\langle\boldsymbol{w},\boldsymbol{y}\rangle\sim\mathcal{N}(0,d)$ for every $\boldsymbol{w}$ of norm $\sqrt{d}$. In particular, for $a\ge 2d^2$ there is some constant c_3 such that $\Pr(|s+\langle\bar{\boldsymbol{w}},\bar{\boldsymbol{y}}\rangle|>a)\le\exp(-c_3 d)$. Combining the above we have that:

$$\mathop{\mathbb{E}}_{\bar{\boldsymbol{w}}\sim\mathcal{U}(2\rho\mathbb{S}^{d-r-1})}\left[\left|\langle g(\cdot),\psi_a(s+\langle\bar{\boldsymbol{w}},\cdot\rangle)\rangle\right|\right]\le c_1\cdot\left(\exp(-c_2(d-r))+\sum_{n=1}^{\infty}\exp(-n\rho^2)\right)+\exp(-c_3 d) . \quad (237)$$

We now go back to Equation (231) and consider the conditional probability over $y_1,\ldots,y_r$ and $w_1,\ldots,w_r$. Note that when taking expectation over $y_1,\ldots,y_r$ we either have that $|\langle\boldsymbol{w},\boldsymbol{y}\rangle|\le a$ which happens w.p $>1-\exp(-c_3 d)$ or $|\langle\boldsymbol{w},\boldsymbol{y}\rangle|\ge a$ in which case, since , $|g(z)|,|\psi_a(z)|\le 1$ for every $z\in\mathbb{R}$ also their product is bounded by 1.

Finally, we consider the expectation over $w_1,\ldots,w_r$. We need to show that with high probability, $\rho=\frac{1}{2}\cdot\|\bar{\boldsymbol{w}}\|$ is large. Instead, we will consider the probability over $w_{r+1},\ldots,w_d$ (note that since $\|\boldsymbol{w}\|=\sqrt{d}$, if we lower bound $\|\bar{\boldsymbol{w}}\|$ it will also upper bound $\sqrt{\sum_{i=1}^{r}w_i^2}$). Since $\boldsymbol{w}$ is sampled uniformly from $\mathcal{U}(\sqrt{d}\mathbb{S}^{d-1})$, we can instead consider sampling z_i from $\mathcal{N}(0,1)$ and setting $(\boldsymbol{w})_i=\sqrt{d}\cdot\frac{z_i}{\|\boldsymbol{z}\|}$. By standard concentration bound on the norm of Gaussian random variables (see Section 3.1 in Vershynin (2018)) there is some constant c_4 such that $\Pr(\|\bar{\boldsymbol{w}}\|^2\notin[0.9(d-r),1.1(d-r)])\le\exp(-c_4(d-r))$. Also, $\sum_{i=r+1}^{d}z_i^2$ has a χ^2 distribution with $d-r$ degrees of freedom. From Lemma 1 in Laurent & Massart (2000) we have that $\Pr\left(\sum_{i=r+1}^{d}w_i^2\ge\frac{1}{2}\cdot(d-r)\right)\le\exp(-c_5(d-r))$ for some constant c_5. Together, there is some constant c_6 such that $\Pr(\|\bar{\boldsymbol{w}}\|^2\ge\frac{1}{6}(d-r))\le\exp(-c_6(d-r))$.

Note that if $\rho>c'\sqrt{d-r}$ then $\sum_{i=1}^{\infty}\exp(-n\rho^2)\le\exp(-c'(d-r))$. Combining all the above and changing the constant terms appropriately, there is some universal constant $c>0$ such that:

$$\mathop{\mathbb{E}}_{\boldsymbol{w}\sim\mathcal{U}(\sqrt{d}\mathbb{S}^{d-1})}\left[\mathop{\mathbb{E}}_{\boldsymbol{y}\sim\mathcal{N}(0,\boldsymbol{I})}\left[|\psi_a(\langle\boldsymbol{w},\boldsymbol{y}\rangle)\cdot g(w_1,\ldots,w_r,\boldsymbol{y})|\right]\right]\le\exp(-c(d-r)) \quad (238)$$

□

Claim 30. *For any $f \in L^2(\mathcal{N}(0, \boldsymbol{I}_d))$, odd periodic function $\psi : \mathbb{R} \to \mathbb{R}$ and $s \in \mathbb{R}$, if $d > c'$ we have that:*

$$\mathop{\mathbb{E}}_{\boldsymbol{w} \sim \mathcal{U}(2\alpha\mathbb{S}^{d-1})} [|\langle f(\cdot), \psi(s + \langle \boldsymbol{w}, \cdot \rangle)\rangle|] \leq c_1 \|f\| \cdot \left(\exp(-c_2 d) + \sum_{n=1}^{\infty} \exp(-n\alpha^2) \right) , \tag{239}$$

here $c', c_1, c_2 > 0$ are some universal constants.

Proof. The proof is similar to the proof of Claim 1 from Yehudai & Shamir (2019) (which is directly derived from Lemma 5 in Shamir (2018)), except for two changes:

1. Here we have an absolute value over the inner product, instead of a square as in Claim 1.
2. We consider a translation of ψ, namely our periodic function is $\psi(s + \cdot)$ for a fixed s.

For the first item, this is a direct application of Jensen's lemma:

$$\mathop{\mathbb{E}}_{\boldsymbol{w} \sim \mathcal{U}(2\alpha\mathbb{S}^{d-1})} \left[\sqrt{|\langle f(\cdot), \psi(s + \langle \boldsymbol{w}, \cdot \rangle)\rangle|^2} \right] \leq \sqrt{\mathop{\mathbb{E}}_{\boldsymbol{w} \sim \mathcal{U}(2\alpha\mathbb{S}^{d-1})} \left[|\langle f(\cdot), \psi(s + \langle \boldsymbol{w}, \cdot \rangle)\rangle|^2 \right]} , \tag{240}$$

where now we can apply Claim 1 from Yehudai & Shamir (2019). For the second item, note that $\psi(s + \cdot)$ is also a periodic function, and Lemma 5 from Shamir (2018) applies to it in the same way as it does on $\psi(\cdot)$. □

We are now ready to prove the main theorem:

Proof of Theorem 3 part (ii). Let T be some transformer with H heads, and output matrix $\boldsymbol{V}_h$ for each head $h \in [H]$. Then T can be written as $T(\boldsymbol{x}_1, \boldsymbol{x}_2, \boldsymbol{y}) = \sum_{h=1}^{H} \boldsymbol{V}_h \boldsymbol{X} \phi_h(\boldsymbol{K}_h \boldsymbol{X}, \boldsymbol{y})$, where $\phi_h : \mathbb{R}^{r \times 2} \to \Delta^1$ is some generalized attention function on the simplex, and $\boldsymbol{K}_h \in \mathbb{R}^{d \times r}$. We have that:

$$\mathop{\mathbb{E}}_{\boldsymbol{x}_1, \boldsymbol{x}_2, \boldsymbol{y}} \left[\|f^{**}(\boldsymbol{x}_1, \boldsymbol{x}_2, \boldsymbol{y}) - T(\boldsymbol{x}_1, \boldsymbol{x}_2, \boldsymbol{y})\|^2 \right] \tag{241}$$

$$= \mathop{\mathbb{E}}_{\boldsymbol{x}_1, \boldsymbol{x}_2, \boldsymbol{y}} \left[\left\| \psi_a^+(\langle \boldsymbol{x}_1 - \boldsymbol{x}_2, \boldsymbol{y} \rangle) \frac{\boldsymbol{x}_1}{d} + \psi_a^-(\langle \boldsymbol{x}_2 - \boldsymbol{x}_1, \boldsymbol{y} \rangle) \frac{\boldsymbol{x}_2}{d} - T(\boldsymbol{x}_1, \boldsymbol{x}_2, \boldsymbol{y}) \right\|^2 \right] \tag{242}$$

$$\geq \mathop{\mathbb{E}}_{\boldsymbol{x}_1, \boldsymbol{x}_2, \boldsymbol{y}} \left[\left\| \psi_a^+(\langle \boldsymbol{x}_1 - \boldsymbol{x}_2, \boldsymbol{y} \rangle) \frac{\boldsymbol{x}_1}{d} \right\|^2 + \left\| \psi_a^-(\langle \boldsymbol{x}_2 - \boldsymbol{x}_1, \boldsymbol{y} \rangle) \frac{\boldsymbol{x}_2}{d} \right\|^2 \right. \tag{243}$$

$$\left. - \left| \left\langle \psi_a^+(\langle \boldsymbol{x}_1 - \boldsymbol{x}_2, \boldsymbol{y} \rangle) \frac{\boldsymbol{x}_1}{d}, T(\boldsymbol{x}_1, \boldsymbol{x}_2, \boldsymbol{y}) \right\rangle \right| - \left| \left\langle \psi_a^-(\langle \boldsymbol{x}_2 - \boldsymbol{x}_1, \boldsymbol{y} \rangle) \frac{\boldsymbol{x}_2}{d}, T(\boldsymbol{x}_1, \boldsymbol{x}_2, \boldsymbol{y}) \right\rangle \right| \right] \tag{244}$$

We will split the expectation in Equation (244), using the linearity of expectation and bound each term separately. By Lemma 28 item (ii), and since $\|\boldsymbol{x}_i\|^2 = d^2$ for $i = 1, 2$ we have that:

$$\mathop{\mathbb{E}}_{\boldsymbol{x}_1, \boldsymbol{x}_2, \boldsymbol{y}} \left[\left\| \psi_a^+(\langle \boldsymbol{x}_1 - \boldsymbol{x}_2, \boldsymbol{y} \rangle) \frac{\boldsymbol{x}_1}{d} \right\|^2 \right] , \quad \mathop{\mathbb{E}}_{\boldsymbol{x}_1, \boldsymbol{x}_2, \boldsymbol{y}} \left[\left\| \psi_a^-(\langle \boldsymbol{x}_2 - \boldsymbol{x}_1, \boldsymbol{y} \rangle) \frac{\boldsymbol{x}_2}{d} \right\|^2 \right] \geq \frac{1}{40} . \tag{245}$$

For the two last terms in Equation (244) we have:

$$\mathop{\mathbb{E}}_{\boldsymbol{x}_1,\boldsymbol{x}_2,\boldsymbol{y}}\left[\left|\left\langle \psi_a^+(\langle \boldsymbol{x}_1-\boldsymbol{x}_2,\boldsymbol{y}\rangle)\frac{\boldsymbol{x}_1}{d}, T(\boldsymbol{x}_1,\boldsymbol{x}_2,\boldsymbol{y})\right\rangle\right|\right] \tag{246}$$

$$=\mathop{\mathbb{E}}_{\boldsymbol{x}_1,\boldsymbol{x}_2,\boldsymbol{y}}\left[\left|\left\langle \psi_a^+(\langle \boldsymbol{x}_1-\boldsymbol{x}_2,\boldsymbol{y}\rangle)\frac{\boldsymbol{x}_1}{d}, \sum_{h=1}^{H}\boldsymbol{V}_h\boldsymbol{X}\phi_h(\boldsymbol{K}_h\boldsymbol{X},\boldsymbol{y})\right\rangle\right|\right] \tag{247}$$

$$\leq\sum_{h=1}^{H}\mathop{\mathbb{E}}_{\boldsymbol{x}_1,\boldsymbol{x}_2,\boldsymbol{y}}\left[\left|\left\langle \psi_a^+(\langle \boldsymbol{x}_1-\boldsymbol{x}_2,\boldsymbol{y}\rangle)\frac{\boldsymbol{x}_1}{d}, \boldsymbol{V}_h\boldsymbol{X}\phi_h(\boldsymbol{K}_h\boldsymbol{X},\boldsymbol{y})\right\rangle\right|\right] \tag{248}$$

$$\leq\sum_{h=1}^{H}\mathop{\mathbb{E}}_{\boldsymbol{x}_1,\boldsymbol{x}_2,\boldsymbol{y}}\left[\left|\left\langle \frac{\boldsymbol{x}_1}{d}, \boldsymbol{V}_h\boldsymbol{X}\right\rangle\right|\cdot\left\|\psi_a^+(\langle \boldsymbol{x}_1-\boldsymbol{x}_2,\boldsymbol{y}\rangle\cdot\phi_h(\boldsymbol{K}_h\boldsymbol{X},\boldsymbol{y})\right\|\right]\,, \tag{249}$$

recall that the output of ϕ_h is in $\mathbb{R}^2$ (since $N=2$). We can bound $\left|\left\langle \frac{\boldsymbol{x}_1}{d}, \boldsymbol{V}_h\boldsymbol{X}\right\rangle\right| \leq d\cdot\max_h\left\|\boldsymbol{V}_h\right\|^2$ independent of $\boldsymbol{y}$ since $\|\boldsymbol{x}_1\|=\|\boldsymbol{x}_2\|=d$. By Theorem 29 we can bound:

$$\mathop{\mathbb{E}}_{\boldsymbol{x}_1,\boldsymbol{x}_2,\boldsymbol{y}}\left[\left\|\psi_a^+(\langle \boldsymbol{x}_1-\boldsymbol{x}_2,\boldsymbol{y}\rangle\cdot\phi_h(\boldsymbol{K}_h\boldsymbol{X},\boldsymbol{y})\right\|\right]\leq\exp(-c(d-r))\,, \tag{250}$$

for some universal constant $c>0$. In a similar way we can bound the correlation w.r.t ψ_a^-. Combining all the above, and plugging back to Equation (244) we get that there is some universal constant $c'>0$ such that:

$$\mathop{\mathbb{E}}_{\boldsymbol{x}_1,\boldsymbol{x}_2,\boldsymbol{y}}\left[\left\|f^{**}(\boldsymbol{x}_1,\boldsymbol{x}_2,\boldsymbol{y})-T(\boldsymbol{x}_1,\boldsymbol{x}_2,\boldsymbol{y})\right\|^2\right]\geq\frac{2}{40}-2dH\max_h\left\|\boldsymbol{V}_h\right\|^2\exp(-c'(d-r))\,. \tag{251}$$

In particular, if $2dH\max_h\left\|\boldsymbol{V}_h\right\|^2\exp(-c'(d-r))\leq\frac{1}{40}$, then:

$$\mathop{\mathbb{E}}_{\boldsymbol{x}_1,\boldsymbol{x}_2,\boldsymbol{y}}\left[\left\|f^{**}(\boldsymbol{x}_1,\boldsymbol{x}_2,\boldsymbol{y})-T(\boldsymbol{x}_1,\boldsymbol{x}_2,\boldsymbol{y})\right\|^2\right]\geq\frac{1}{40}\,. \tag{252}$$

Rearranging the terms, we get that the above inequality happens if $dH\max_h\left\|\boldsymbol{V}_h\right\|^2\leq\frac{1}{80}\exp(-c'(d-r))$. Changing the constant c' and rescaling everything by a factor of $\sqrt{d}$ (according to Lemma 31) finishes the proof. □

The following lemma justifies our rescaling the target function and input distributions by a factor of $\sqrt{d}$.

Lemma 31. *Define f^* as in Equation* (8) *and f^{**} as in Equation* (221)*. For each attention layer T, there is an attention layer T' for which*

$$\mathop{\mathbb{E}}_{\substack{\boldsymbol{x}_1,\boldsymbol{x}_2\sim\mathcal{D}_2(\sqrt{d}\mathbb{S}^{d-1})\\ \boldsymbol{y}\sim\mathcal{N}(0,\boldsymbol{I})}}\left[\left\|T(\boldsymbol{x}_1,\boldsymbol{x}_2,\boldsymbol{y})-f^*(\boldsymbol{x}_1,\boldsymbol{x}_2,\boldsymbol{y})\right\|^2\right] \tag{253}$$

$$=\mathop{\mathbb{E}}_{\substack{\boldsymbol{x}_1,\boldsymbol{x}_2\sim\mathcal{D}_2(\mathbb{S}^{d-1})\\ \boldsymbol{y}\sim\mathcal{N}(0,\frac{1}{d}\boldsymbol{I})}}\left[\left\|T'(\boldsymbol{x}_1,\boldsymbol{x}_2,\boldsymbol{y})-f^{**}(\boldsymbol{x}_1,\boldsymbol{x}_2,\boldsymbol{y})\right\|^2\right] \tag{254}$$

Proof. Let

$$T(\boldsymbol{X},\boldsymbol{y}):=\sum_{h=1}^{H}\boldsymbol{V}_h\boldsymbol{X}\phi_h\left(\boldsymbol{K}_h^\top\boldsymbol{X},\boldsymbol{y}\right) \tag{255}$$

Then construct the following attention layer:

$$T'(\boldsymbol{X},\boldsymbol{y}):=\sum_{h=1}^{H}(\boldsymbol{V}_h/\sqrt{d})\boldsymbol{X}\phi_h'\left(\boldsymbol{K}_h^\top\boldsymbol{X},\boldsymbol{y}\right) \tag{256}$$

where $\phi'_h(\boldsymbol{A}, \boldsymbol{Y}) := \phi_h(\boldsymbol{A}/\sqrt{d}, \boldsymbol{y}/\sqrt{d})$. For instance, if $\phi_h(\boldsymbol{A}, \boldsymbol{y}) = \mathrm{sm}(\boldsymbol{A}^\top \boldsymbol{Q}_h \boldsymbol{y})$, then $\phi'_h(\boldsymbol{A}, \boldsymbol{y}) = \mathrm{sm}(\boldsymbol{A}^\top (\boldsymbol{Q}_h/d)\boldsymbol{y})$. Clearly, $T(\boldsymbol{X}/\sqrt{d}, \boldsymbol{y}/\sqrt{d}) = T'(\boldsymbol{X}, \boldsymbol{y})$. Likewise, by construction, $f^{**}(\boldsymbol{x}_1, \dots \boldsymbol{x}_N, \boldsymbol{y}) = f^*(\boldsymbol{x}_1/\sqrt{d}, \dots \boldsymbol{x}_N/\sqrt{d}, \boldsymbol{y}/\sqrt{d})$. Thus,

$$
\begin{aligned}
&\mathop{\mathbb{E}}_{\substack{\boldsymbol{x}_1, \boldsymbol{x}_2 \sim \mathcal{D}_2(\mathbb{S}^{d-1}) \\ \boldsymbol{y} \sim \mathcal{N}(0, \frac{1}{d}\boldsymbol{I})}} \left[\left\| T(\boldsymbol{x}_1, \boldsymbol{x}_2, \boldsymbol{y}) - f^*(\boldsymbol{x}_1, \boldsymbol{x}_2, \boldsymbol{y}) \right\|^2 \right] \\
&= \mathop{\mathbb{E}}_{\substack{\boldsymbol{x}_1, \boldsymbol{x}_2 \sim \mathcal{D}_2(\sqrt{d}\mathbb{S}^{d-1}) \\ \boldsymbol{y} \sim \mathcal{N}(0, \boldsymbol{I})}} \left[\left\| T\left(\boldsymbol{x}_1/\sqrt{d}, \boldsymbol{x}_2/\sqrt{d}, \boldsymbol{y}/\sqrt{d}\right) - f^*\left(\boldsymbol{x}_1/\sqrt{d}, \boldsymbol{x}_2/\sqrt{d}, \boldsymbol{y}/\sqrt{d}\right) \right\|^2 \right] \\
&= \mathop{\mathbb{E}}_{\substack{\boldsymbol{x}_1, \boldsymbol{x}_2 \sim \mathcal{D}_2(\sqrt{d}\mathbb{S}^{d-1}) \\ \boldsymbol{y} \sim \mathcal{N}(0, \boldsymbol{I})}} \left[\left\| T'(\boldsymbol{x}_1, \boldsymbol{x}_2, \boldsymbol{y}) - f^{**}(\boldsymbol{x}_1, \boldsymbol{x}_2, \boldsymbol{y}) \right\|^2 \right]
\end{aligned}
$$

□

E Proofs from Section 6 and an Additional Construction

In Section 6, we present a construction (Theorem 5) that uses concatenated positional encodings to facilitate the majority voting strategy. This construction has the strange property that it breaks the permutation invariance of standard attention layers in order to approximate a function that is permutation invariant. It also increases the dimension of the transformer. This begs the question of whether these properties are necessary to allow low-rank attention to represent the target. Below, we presenting an alternative construction that does not have these properties. Instead, it modifies the attention mechanism by concatenating the outputs of the heads together rather than summing them. It then passes the concatenated outputs to an MLP layer that computes the mode.

Theorem 32 (Majority Voting Approximation Upper Bound)**.** *There exist universal constants $c_1, c_2, c_3, c_4 > 0$ such that for all $d > c_1$, $\epsilon \in \left(0, \frac{1}{2}\right)$, and $H \geq c_2 \cdot \frac{d^3}{\epsilon^2}$, there exist vectors $\boldsymbol{q}_1, \dots, \boldsymbol{q}_H$ and a 4-layer feedforward network $g : \mathbb{R}^{dH} \to \mathbb{R}^d$ of width $c_3 d^2 H$ such that*

$$
\mathop{\mathbb{E}}_{\boldsymbol{x}_1, \boldsymbol{x}_2, \boldsymbol{y} \sim \mathrm{Unif}(\mathbb{S}^{d-1})} \left\| f(\boldsymbol{x}_1, \boldsymbol{x}_2; \boldsymbol{y}) - g\left(\begin{bmatrix} \boldsymbol{X}\, \mathrm{sm}(\boldsymbol{X}^\top \boldsymbol{q}_1 \boldsymbol{q}_1^\top \boldsymbol{y}) \\ \vdots \\ \boldsymbol{X}\, \mathrm{sm}(\boldsymbol{X}^\top \boldsymbol{q}_H \boldsymbol{q}_H^\top \boldsymbol{y}) \end{bmatrix} \right) \right\|_2^2 \leq \epsilon + \exp(-c_4 d)\,. \tag{257}
$$

This construction shows that using a constant-depth MLP to combine the heads can overcome the weakness of low rank attention. The full proof can be found in Appendix E.3. The idea behind the construction of the MLP $g(\cdot)$ is to perform an inner product between the outputs of the heads, allowing us to compare which one of the outputs $\boldsymbol{x}_1$ or $\boldsymbol{x}_2$ received more votes. The inner products can be approximated by a ReLU network, as long as the input vectors are not too close to each other, which happens with exponentially large probability. This is the cause of the extra exponentially small term in the loss.

E.1 Two-layer transformers with masked self attention

In Theorem 5, we use the following definition of a multi-layer transformer with self-attention for our majority voting construction. Self-attention simply means that the source and target points are the same, in contrast to the cross-attention used in the rest of this paper. We modify the attention mechanism by adding a self-excluding mask so that each input point cannot attend to itself (see below, where we form $\tilde{\boldsymbol{X}}_i$ by deleting the ith column of $\boldsymbol{X}$). Following standard practice, we also use a skip connection. We do not need a MLP or normalization layer, though our construction can easily be extended to include them.

Definition 33. *A rank-r **self-masked transformer layer** with H heads is a function $T : \mathbb{R}^{d\times N} \to \mathbb{R}^{d\times N}$ parameterized by rank-k attention heads $\{(\boldsymbol{M}_h, \boldsymbol{V}_h)\}_{h=1}^{H}$ and defined as follows:*

$$\tilde{\boldsymbol{X}}_i := \begin{bmatrix} | & & | & | & & | \\ \boldsymbol{x}_1 & \cdots & \boldsymbol{x}_{i-1} & \boldsymbol{x}_{i+1} & \cdots & \boldsymbol{x}_N \\ | & & | & | & & | \end{bmatrix} \tag{258}$$

$$T_i(\boldsymbol{X}) := \boldsymbol{x}_i + \sum_{h=1}^{H} \boldsymbol{V}_h \tilde{\boldsymbol{X}}_i \,\mathrm{sm}\left(\tilde{\boldsymbol{X}}_i{}^{\top} \boldsymbol{M}_h \boldsymbol{x}_i\right) \tag{259}$$

$$\tag{260}$$

Here, T_i denotes the ith output (or ith column of the output) $[T_1(\boldsymbol{X}) \quad \cdots \quad T_N(\boldsymbol{X})]$.

*A **2-layer, rank-r transformer with concatenated positional encodings** is a function $T : \mathbb{R}^{d\times N} \to \mathbb{R}^{d\times N}$ parameterized by a positional encoding matrix $\boldsymbol{E} = \mathbb{R}^{d_e\times N}$ and two $(d + d_e)$-dimensional self-masked transformer layers, $T^{(1)}$ and $T^{(2)}$, and an output-layer matrix $\boldsymbol{A} \in \mathbb{R}^{d\times(d+d_e)}$ and defined as follows:*

$$T(\boldsymbol{X}) = \boldsymbol{A} \cdot T_N^{(2)}\left(T^{(1)}\left(\begin{bmatrix}\boldsymbol{X}\\ \boldsymbol{E}\end{bmatrix}\right)\right) . \tag{261}$$

Note that in the above definition, we consider the Nth output point of the transformer to be the output of the model as a whole.

E.2 Lemmas

To prove Theorems 5 and 32 we will need several lemmas.

The first shows that for a fixed set of inputs, drawing a rank-1 head randomly will have the same output as the target f with probability slightly larger than $\frac{1}{2}$. This lemma justifies our majority voting strategy.

Lemma 34. *Fix $\boldsymbol{x}_1, \boldsymbol{x}_2, \boldsymbol{y} \in \mathbb{S}^{d-1}$ with $|\langle \boldsymbol{x}_1 - \boldsymbol{x}_2, \boldsymbol{y}\rangle| \geq a$ for some $a > 0$. Then for $d > c_1$ we have that:*

$$\Pr_{\mathbf{q}\sim\mathcal{U}(\mathbb{S}^{d-1})}\left(\arg\max_i \langle \boldsymbol{x}_i, \mathbf{q}\rangle \cdot \langle \boldsymbol{y}, \mathbf{q}\rangle = \arg\max_i \langle \boldsymbol{x}_i, \boldsymbol{y}\rangle\right) \geq \frac{1}{2} + c_2 \cdot \frac{a}{\sqrt{d}} \tag{262}$$

for some universal constants $c_1, c_2 > 0$.

Proof. In the proof, all probabilities are for $\mathbf{q} \sim \mathcal{U}(\mathbb{S}^{d-1})$, thus we omit this notation. Denote $\boldsymbol{w} := \boldsymbol{x}_1 - \boldsymbol{x}_2$, and assume w.l.o.g that $\langle \boldsymbol{w}, \boldsymbol{y}\rangle > 0$, the other direction is similar. We can write:

$$\begin{aligned}
&\Pr\left(\arg\max_i \langle \boldsymbol{x}_i, \mathbf{q}\rangle \cdot \langle \boldsymbol{y}, \mathbf{q}\rangle = \arg\max_i \langle \boldsymbol{x}_i, \boldsymbol{y}\rangle\right)\\
=&\Pr\left(\mathrm{sgn}(\langle \boldsymbol{w}, \boldsymbol{y}\rangle) = \mathrm{sgn}(\langle \boldsymbol{w}, \mathbf{q}\rangle \cdot \langle \boldsymbol{y}, \mathbf{q}\rangle)\right)\\
=&\Pr\left(\langle \boldsymbol{w}, \mathbf{q}\rangle \cdot \langle \boldsymbol{y}, \mathbf{q}\rangle > 0\right) .
\end{aligned}$$

Since the above probability is rotation invariant w.r.t $\mathbf{q}$, we can assume w.l.o.g that $\boldsymbol{w} = \boldsymbol{e}_1$. Hence we can write $\boldsymbol{y} = \begin{pmatrix}\tilde{a}\\ \bar{\boldsymbol{y}}\end{pmatrix}$, where $\bar{\boldsymbol{y}} \in \mathbb{R}^{d-1}$ and $\tilde{a} = \langle \boldsymbol{w}, \boldsymbol{y}\rangle$. Thus, the above probability is equal to:

$$\Pr\left(q_1(\tilde{a}q_1 + \langle \bar{\mathbf{q}}, \bar{\boldsymbol{y}}\rangle) > 0\right) \tag{263}$$

$$=\frac{1}{2}\Pr\left(q_1(\tilde{a}q_1 + \langle \bar{\mathbf{q}}, \bar{\boldsymbol{y}}\rangle) > 0 | q_1 > 0\right) + \frac{1}{2}\Pr\left(q_1(\tilde{a}q_1 + \langle \bar{\mathbf{q}}, \bar{\boldsymbol{y}}\rangle) > 0 | q_1 < 0\right) \tag{264}$$

$$=\frac{1}{2}\Pr\left(\tilde{a}q_1 + \langle \bar{\mathbf{q}}, \bar{\boldsymbol{y}}\rangle > 0 | q_1 > 0\right) + \frac{1}{2}\Pr\left(\tilde{a}q_1 + \langle \bar{\mathbf{q}}, \bar{\boldsymbol{y}}\rangle < 0 | q_1 < 0\right) \tag{265}$$

$$=\Pr\left(\tilde{a}q_1 + \langle \bar{\mathbf{q}}, \bar{\boldsymbol{y}}\rangle > 0 | q_1 > 0\right) \tag{266}$$

where the last equality is by the symmetry of the distribution of $\mathbf{q}$. Note that if $\langle \bar{\mathbf{q}}, \bar{\boldsymbol{y}} \rangle > 0$ which happens w.p $\frac{1}{2}$, then the term inside the above probability is positive. Hence, we can write:

$$\begin{aligned}
&\Pr\left(\tilde{a}q_1 + \langle \bar{\mathbf{q}}, \bar{\boldsymbol{y}} \rangle > 0 | q_1 > 0\right) \\
=&\frac{1}{2} + \frac{1}{2} \cdot \Pr\left(\tilde{a}q_1 + \langle \bar{\mathbf{q}}, \bar{\boldsymbol{y}} \rangle > 0 | q_1 > 0, \langle \bar{\mathbf{q}}, \bar{\boldsymbol{y}} \rangle < 0\right) \\
\geq&\frac{1}{2} + \frac{1}{2} \cdot \Pr\left(\tilde{a}q_1 \geq \frac{2\tilde{a}}{\sqrt{d}} | q_1 > 0\right) \cdot \Pr\left(|\langle \bar{\mathbf{q}}, \bar{\boldsymbol{y}} \rangle| \leq \frac{\tilde{a}}{\sqrt{d}} | \langle \bar{\mathbf{q}}, \bar{\boldsymbol{y}} \rangle < 0\right) \qquad (267)
\end{aligned}$$

We will now lower bound each probability separately. First, note that if we sample $\boldsymbol{u} \sim \mathcal{N}\left(0, \frac{1}{d}I\right)$, then $\frac{u_1}{\|\boldsymbol{u}\|}$ has the same distribution as q_1. By the concentration of the norm of Gaussian random variables (see Vershynin (2018) Section 3.1), there is a constant $c_1 > 0$ such that w.p $> 1 - \exp(-c_1 d)$ we have $\|\boldsymbol{u}\| \in [0.9, 1.1]$. There is also a constant $c_2 \in \left(0, \frac{1}{2}\right)$ such that $\Pr\left(u_1 > \frac{3}{\sqrt{d}}\right) > c_2$. This bounds the first probability term in Equation (267). For the second term, note that $\|\bar{\boldsymbol{y}}\| \leq \|\boldsymbol{y}\| = 1$. By the same reasoning as above we can write:

$$\begin{aligned}
&\Pr\left(|\langle \bar{\mathbf{q}}, \bar{\boldsymbol{y}} \rangle| \leq \frac{\tilde{a}}{\sqrt{d}} | \langle \bar{\mathbf{q}}, \bar{\boldsymbol{y}} \rangle < 0\right) \geq \Pr\left(|\langle \bar{\mathbf{q}}, \bar{\boldsymbol{y}} \rangle| \leq \frac{a}{\sqrt{d}} | \langle \bar{\mathbf{q}}, \bar{\boldsymbol{y}} \rangle < 0\right) \qquad (268) \\
=&\Pr_{\boldsymbol{u} \sim \mathcal{N}\left(0, \frac{1}{d}I\right)}\left(\left|\frac{u_2}{\|\boldsymbol{u}\|}\right| \leq \frac{a}{\sqrt{d}}\right) \geq (1 - \exp(-c_1 d)) \cdot \Pr_{u_2 \sim \mathcal{N}\left(0, \frac{1}{d}\right)}\left(|u_2| \leq \frac{a \cdot 0.9}{\sqrt{d}}\right) \qquad (269)
\end{aligned}$$

The above probability is bounded by $\text{erf}\left(\frac{a \cdot 0.9}{\sqrt{d}}\right) \geq \frac{a \cdot 0.9}{\sqrt{d}}$, where this inequality is since $\text{erf}(z) > z$ for $z \in \left[0, \frac{1}{2}\right]$. In total, we can bound this probability by

$$\Pr\left(|\langle \bar{\mathbf{q}}, \bar{\boldsymbol{y}} \rangle| \leq \frac{a}{\sqrt{d}} | \langle \bar{\mathbf{q}}, \bar{\boldsymbol{y}} \rangle < 0\right) \geq (1 - \exp(-c_1 d)) \cdot \frac{a \cdot 0.9}{\sqrt{d}} \,. \qquad (270)$$

We take $d > \tilde{c}$ so that $\exp(-c_1 d) \leq \frac{1}{2}$, Combining the two bounds, and changing the universal constant finishes the proof. □

The following lemma shows that a random draw of inputs will satisfy a certain condition which allows the use of the previous lemma.

Lemma 35. *Let $\epsilon > 0$, then:*

$$Pr_{\boldsymbol{x}_1, \boldsymbol{x}_2, \boldsymbol{y} \sim \text{Unif}(\mathbb{S}^{d-1})}\left(|\langle \boldsymbol{x}_1 - \boldsymbol{x}_2, \boldsymbol{y} \rangle| \leq \epsilon\right) \leq (1 - \exp(-c_1 d)) \cdot 2\epsilon\sqrt{d} \,, \qquad (271)$$

where $c_1 > 0$ is some universal constant.

Proof. By the symmetry of the distribution, we can assume w.l.o.g that $\boldsymbol{y} = \boldsymbol{e}_1$. Also, note that for $\boldsymbol{u}, \boldsymbol{v} \sim \mathcal{N}\left(0, \frac{1}{d}I\right)$, we can view the distribution of $(\boldsymbol{x}_1)_1$ and $(\boldsymbol{x}_2)_1$ as $\frac{u_1}{\|\boldsymbol{u}\|}$ and $\frac{v_1}{\|\boldsymbol{v}\|}$. Combining the above, we get that:

$$\Pr_{\boldsymbol{x}_1, \boldsymbol{x}_2, \boldsymbol{y} \sim \text{Unif}(\mathbb{S}^{d-1})}\left(|\langle \boldsymbol{x}_1 - \boldsymbol{x}_2, \boldsymbol{y} \rangle| \leq \epsilon\right) = \Pr_{\boldsymbol{u}, \boldsymbol{v} \sim \mathcal{N}\left(0, \frac{1}{d}I\right)}\left(\left|\frac{u_1}{\|\boldsymbol{u}\|} - \frac{v_1}{\|\boldsymbol{v}\|}\right| \leq \epsilon\right) \,. \qquad (272)$$

By the concentration of the norm of normal random vectors (see Vershynin (2018) section 3.1) we have w.p $> 1 - \exp(-c_1 d)$ that $\|\boldsymbol{u}\|, \|\boldsymbol{v}\| \leq 1.1$ for some universal constant $c_1 > 0$. Also $z := u_1 - v_1 \sim \mathcal{N}\left(0, \frac{2}{d}\right)$. Hence, the above probability can be upper bounded by $\Pr_{z \sim \mathcal{N}\left(0, \frac{2}{d}\right)}\left(|z| < 1.1\epsilon\right) \leq \text{erf}\left(\epsilon\sqrt{d}\right)$. Note that $\text{erf}(x) \leq 2x$ for every $x > 0$, hence the above probability can be bounded by $(1 - \exp(-c_1 d)) \cdot 2\epsilon\sqrt{d}$ □

The following lemma shows a construction of the majority function over H input vectors. This construction uses an approximation of the inner product of two inputs using a ReLU network.

Lemma 36. *Let $\boldsymbol{v}_1, \dots \boldsymbol{v}_H \in \{\boldsymbol{x}_+, \boldsymbol{x}_-\} \subset \mathbb{R}^d$, where $\langle \boldsymbol{x}_-, \boldsymbol{x}_+ \rangle \leq 0.1$. Let $\boldsymbol{v}^*$ be the mode of $\boldsymbol{v}_1, \dots \boldsymbol{v}_H$. Then there exists a 4-layer feedforward network $g : \mathbb{R}^{d(H+2)} \to \mathbb{R}^d$ with width $c \cdot d^2 H$ for some universal constant $c > 0$ and weights bounded by 2 such that*

$$g\left(\begin{bmatrix} \boldsymbol{v}_1 \\ \vdots \\ \boldsymbol{v}_H \\ \boldsymbol{x}_+ \\ \boldsymbol{x}_- \end{bmatrix}\right) = g\left(\begin{bmatrix} \boldsymbol{v}_1 \\ \vdots \\ \boldsymbol{v}_H \\ \boldsymbol{x}_- \\ \boldsymbol{x}_+ \end{bmatrix}\right) = \boldsymbol{v}^* \tag{273}$$

Proof. Let $\boldsymbol{x}$ be the finally d coordinate of $\boldsymbol{v} := [\boldsymbol{v}_1 \quad \cdots \quad \boldsymbol{v}_H \quad \boldsymbol{x}_- \quad \boldsymbol{x}_+]^\top \in \mathbb{R}^{d(H+2)}$, and let $\hat{\boldsymbol{x}}$ be the second to last block of d coordinates of $\boldsymbol{v}$. Note that either $\boldsymbol{x} = \boldsymbol{x}_+$ and $\hat{\boldsymbol{x}} = \boldsymbol{x}_-$ or the other way around. We construct a network that calculates the inner product between $\boldsymbol{x}$ and each $\boldsymbol{v}_i$ up to accuracy of $\frac{1}{10H}$. By Lemma 37 there is such a 2-layer network $M_1 : \mathbb{R}^{d(H+2)} \to \mathbb{R}^{2d+1}$ with width cd^2H for some universal constant $c > 0$ and weights bounded by 2. We add $2d$ more neurons which act as two identity matrices to keep the last $2d$ coordinates of $\boldsymbol{v}$. We add an additional output layer to M_1 which sums all the outputs of the inner products.

We now construct another network $M_2 : \mathbb{R}^{2d+1} \to \mathbb{R}^d$ which either output $\boldsymbol{x}$ if the sums of the inner product is larger than $0.2 \cdot H$ or $\hat{\boldsymbol{x}}$ otherwise. Note that by our assumption that $\langle \boldsymbol{x}_1, \boldsymbol{x}_2 \rangle \leq 0.1$, M_2 will output the mode of the $\boldsymbol{v}_i$'s. This is because M_1 calculates inner products up to an error of $\frac{1}{10H}$, summing over H such inner products returns the exact sum plus an error which is bounded by $\frac{1}{10}$. Composing M_1 and M_2 provides an MLP which will output either $\boldsymbol{x}_+$ or $\boldsymbol{x}_-$ depending on who is the mode.

The total width of the network is $c_3 d^2 H$, since we calculate inner products up to an error of $\frac{1}{10H}$, and the depth of the network is 4. □

We next show that shallow neural networks can approximately compute the inner product of two vectors.

Lemma 37. *Let $\epsilon > 0$. There exists a 2-layer network $N : \left(\mathbb{S}^{d-1}\right)^2 \to \mathbb{R}$ with width $\frac{cd^2}{\epsilon}$ and weights bounded by 2 that calculates $\langle \boldsymbol{x}, \boldsymbol{x}' \rangle$ up to accuracy ϵ. Here $c > 0$ is some universal constant.*

Proof. By Lemma 6 in Daniely (2017) there exists a depth 2 network $N_{\text{square}} : \mathbb{R} \to \mathbb{R}$ that calculates $\frac{x^2}{2}$ in $[-2, 2]$ with an error of $\frac{\epsilon}{d}$, width of at most $\frac{32d}{\epsilon}$ and weights bounded by 2. For each coordinate $i \in [d]$ we compose the linear function $(\boldsymbol{x})_i + (\boldsymbol{x}')_i$ with N_{square} to get a depth 2 network that calculates $\frac{((\boldsymbol{x})_i + (\boldsymbol{x}')_i)^2}{2}$ up to an error of $\frac{\epsilon}{d}$. Summing over these networks for every index i and subtracting 1 results in a network that calculates $\langle \boldsymbol{x}, \boldsymbol{x}' \rangle$ with an error of ϵ and width $\frac{32d^2}{\epsilon}$ □

Finally, the following lemma shows that if we draw random rank-1 attention heads, taking their "majority vote" will approximate the target function f. The rate of approximation depends on the number of sampled heads and on the input dimension.

Lemma 38. *Let $M : \left(\mathbb{R}^d\right)^H \to \mathbb{R}^d$ be the majority function over H vectors in $\mathbb{R}^d$. Namely, given a set of H vectors, M outputs the vector which appears the most times in the set, and breaks ties randomly. For a vector $\boldsymbol{q}_h$ define $g_h(\boldsymbol{x}_1, \boldsymbol{x}_2; \boldsymbol{y}) = \arg\max_{\boldsymbol{x}_i} \langle \boldsymbol{x}_i, \boldsymbol{q}_h \rangle \cdot \langle \boldsymbol{y}, \boldsymbol{q}_h \rangle$. There exist universal constants $c_1, c_2 > 0$ such that if $H > \frac{c_1 d^3}{\epsilon^2}$, then with probability at least $1 - \exp(c_2 d)$ over samples $\boldsymbol{q}_1, \dots, \boldsymbol{q}_H \sim \text{Unif}(\mathbb{S}^{d-1})$, we have that:*

$$\mathop{\mathbb{E}}_{\boldsymbol{x}_1, \boldsymbol{x}_2; \boldsymbol{y} \sim \text{Unif}(\mathbb{S}^{d-1})} \left[\left\| f(\boldsymbol{x}_1, \boldsymbol{x}_2; \boldsymbol{y}) - M\left(\{ g(\boldsymbol{x}_1, \boldsymbol{x}_2; \boldsymbol{y} \}_{h=1}^H \right) \right\|^2 \right] \leq \epsilon \,, \tag{274}$$

Here, f is defined as in Equation (3)*.*

Proof. Fix $\boldsymbol{x}_1, \boldsymbol{x}_2, \boldsymbol{y}$ with $|\langle \boldsymbol{x}_1 - \boldsymbol{x}_2, \boldsymbol{y} \rangle| \geq \epsilon$. Denote by A_h the event over sampling $\boldsymbol{q} \sim \text{Unif}(\mathbb{S}^{d-1})$ which output 1 if $\arg\max_i \langle \boldsymbol{x}_i, \mathbf{q}_h \rangle \cdot \langle \boldsymbol{y}, \mathbf{q}_h \rangle = \arg\max_i \langle \boldsymbol{x}_i, \boldsymbol{y} \rangle$ and 0 otherwise.

By Lemma 34 we have that $\Pr(A_h = 1) \geq \frac{1}{2} + c_2 \cdot \frac{\epsilon}{\sqrt{d}}$ if $d > c_1$ for some universal constants $c_1, c_2 > 0$. Note that the events $\{A_h\}_{h=1}^H$ are independent when $\boldsymbol{x}_1, \boldsymbol{x}_2, \boldsymbol{y}$ are fixed. Hence, we can use Hoeffding's inequality:

$$\Pr_{q_1,\ldots,q_H}\left(\left|\frac{1}{H}\sum_{h=1}^{H} A_h - \left(\frac{1}{2} + c_2 \cdot \frac{\epsilon}{\sqrt{d}}\right)\right| \geq t\right) \leq 2\exp(-2Ht^2)\,. \tag{275}$$

By setting $t = \frac{c2\epsilon}{\sqrt{d}}$ and $H \geq \frac{d^2}{\epsilon^2}$ we get that:

$$\Pr\left(\frac{1}{H}\sum_{h=1}^{H} A_h < \frac{1}{2}\right) \leq 2\exp(-2c_2 d)\,. \tag{276}$$

From now on, we condition on the event that $\boldsymbol{q}_1, \ldots, \boldsymbol{q}_H$ are sampled such that $\frac{1}{H}\sum_{h=1}^{H} A_h \geq \frac{1}{2}$, which happens w.p $> 1 - 2\exp(-2c_2 d)$. Note that if this event happens, then the majority of the functions $g_h(\boldsymbol{x}_1, \boldsymbol{x}_2, \boldsymbol{y})$ will output the same vector as $f(\boldsymbol{x}_1, \boldsymbol{x}_2, \boldsymbol{y})$.

By Lemma 35 we have that $\Pr\left(|\left\langle \boldsymbol{x}_1 - \boldsymbol{x}_2, \boldsymbol{y}\right\rangle| \leq \epsilon\right) \leq (1 - \exp(-c_3 d)) \cdot 2\epsilon\sqrt{d}$ for some universal constant $c_3 > 0$. Hence, we get that:

$$\mathop{\mathbb{E}}_{\boldsymbol{x}_1,\boldsymbol{x}_2,\boldsymbol{y}\sim\mathrm{Unif}(\mathbb{S}^{d-1})}\left[\left\|f(\boldsymbol{x}_1,\boldsymbol{x}_2,\boldsymbol{y}) - M\left(\{g(\boldsymbol{x}_1,\boldsymbol{x}_2,\boldsymbol{y}\}_{h=1}^H\right)\right\|^2\right] \tag{277}$$

$$= \Pr\left(|\left\langle \boldsymbol{x}_1 - \boldsymbol{x}_2, \boldsymbol{y}\right\rangle| \leq \epsilon\right) \cdot \mathbb{E}\left[\left\|f(\boldsymbol{x}_1,\boldsymbol{x}_2,\boldsymbol{y}) - M\left(\{g(\boldsymbol{x}_1,\boldsymbol{x}_2,\boldsymbol{y}\}_{h=1}^H\right)\right\|^2 \middle| |\left\langle \boldsymbol{x}_1 - \boldsymbol{x}_2, \boldsymbol{y}\right\rangle| \leq \epsilon\right] + \tag{278}$$

$$+\Pr\left(|\left\langle \boldsymbol{x}_1 - \boldsymbol{x}_2, \boldsymbol{y}\right\rangle| \geq \epsilon\right) \cdot \mathbb{E}\left[\left\|f(\boldsymbol{x}_1,\boldsymbol{x}_2,\boldsymbol{y}) - M\left(\{g(\boldsymbol{x}_1,\boldsymbol{x}_2,\boldsymbol{y}\}_{h=1}^H\right)\right\|^2 \middle| |\left\langle \boldsymbol{x}_1 - \boldsymbol{x}_2, \boldsymbol{y}\right\rangle| \geq \epsilon\right] \tag{279}$$

$$\leq (1 - \exp(-c_3 d)) \cdot 2\epsilon\sqrt{d} \cdot 1 + 1 \cdot \exp(-2c_2 d) \leq c \cdot \epsilon\sqrt{d} \tag{280}$$

where we choose d large enough such that $1 - \exp(-c_3 d) \geq \frac{1}{2}$ and $\exp(-2c_2 d) \leq \frac{1}{2}$ and changed the constant $c > 0$ accordingly. Replacing ϵ with $\tilde{\epsilon} = \frac{\epsilon}{c\sqrt{d}}$ finishes the proof.

□

E.3 Proof of Theorem 32

Proof. By Lemma 38 there exist $\boldsymbol{q}_1, \ldots, \boldsymbol{q}_{H-2} \in \mathbb{S}^{d-1}$ such that if $H \geq \frac{c_2 d^3}{\epsilon^2}$:

$$\mathop{\mathbb{E}}_{\boldsymbol{x}_1,\boldsymbol{x}_2,\boldsymbol{y}\sim\mathrm{Unif}(\mathbb{S}^{d-1})}\left[\left\|f(\boldsymbol{x}_1,\boldsymbol{x}_2,\boldsymbol{y}) - M\left(\{g(\boldsymbol{x}_1,\boldsymbol{x}_2,\boldsymbol{y}\}_{h=1}^H\right)\right\|^2\right] \leq \epsilon \tag{281}$$

where $g_h(\boldsymbol{x}_1, \boldsymbol{x}_2, \boldsymbol{y}) = \arg\max_{\boldsymbol{x}_i} \left\langle \boldsymbol{x}_i, \boldsymbol{q}_h\right\rangle \cdot \left\langle \boldsymbol{x}, \boldsymbol{q}_h\right\rangle$ and M is the majority function. We can take $H - 2$ instead of H by increasing the constant by a factor of at most 2.

We define $\boldsymbol{M}_i = \alpha \boldsymbol{q}_i \boldsymbol{q}_i^\top$ for $i = 1, \ldots, H-2$ for some $\alpha > 0$ which will be defined later. We also pick some $\boldsymbol{q}_0 \in \mathbb{S}^{d-1}$ and define $\boldsymbol{M}_{H-1} = \alpha \boldsymbol{q}_0 \boldsymbol{q}_0^\top$ and $\boldsymbol{M}_H = -\alpha \boldsymbol{q}_0 \boldsymbol{q}_0^\top$. Note that if $\boldsymbol{q}_0 \notin \{\boldsymbol{x}_1, \boldsymbol{x}_2, \boldsymbol{y}\}$ and $\arg\max_{\boldsymbol{x}_i} \boldsymbol{x}_i \boldsymbol{M}_{H-1} \boldsymbol{y} = \boldsymbol{x}_1$ then $\arg\max_{\boldsymbol{x}_i} \boldsymbol{x}_i \boldsymbol{M}_H \boldsymbol{y} = \boldsymbol{x}_2$ and vice versa.

Let $g : \mathbb{R}^{dH} \to \mathbb{R}^d$ be the 4-layer network with width $c_1 d^2 H$ as defined in Lemma 36 which simulates the majority. Denote by $\boldsymbol{v} := \begin{bmatrix} \boldsymbol{X}\,\mathrm{sm}(\boldsymbol{X}^\top \boldsymbol{M}_1 \boldsymbol{y}) \\ \vdots \\ \boldsymbol{X}\,\mathrm{sm}(\boldsymbol{X}^\top \boldsymbol{M}_H \boldsymbol{y}) \end{bmatrix}$ and by $\boldsymbol{v}_{\max} = \begin{bmatrix} \arg\max_{\boldsymbol{x}_i}(\boldsymbol{x}_i^\top \boldsymbol{M}_1 \boldsymbol{y}) \\ \vdots \\ \arg\max_{\boldsymbol{x}_i}(\boldsymbol{x}_i^\top \boldsymbol{M}_H \boldsymbol{y}) \end{bmatrix}$.

We have that:

$$\mathop{\mathbb{E}}_{\boldsymbol{x}_1,\boldsymbol{x}_2,\boldsymbol{y}\sim\mathrm{Unif}(\mathbb{S}^{d-1})}\left[\left\|f(\boldsymbol{x}_1,\boldsymbol{x}_2;\boldsymbol{y}) - g\left(\boldsymbol{v}\right)\right\|^2\right]$$

$$\leq \mathop{\mathbb{E}}_{\boldsymbol{x}_1,\boldsymbol{x}_2,\boldsymbol{y}\sim\mathrm{Unif}(\mathbb{S}^{d-1})}\left[\left\|f(\boldsymbol{x}_1,\boldsymbol{x}_2;\boldsymbol{y}) - g\left(\boldsymbol{v}_{\max}\right)\right\|^2\right] + \mathop{\mathbb{E}}_{\boldsymbol{x}_1,\boldsymbol{x}_2,\boldsymbol{y}\sim\mathrm{Unif}(\mathbb{S}^{d-1})}\left[\left\|g\left(\boldsymbol{v}_{\max}\right) - g\left(\boldsymbol{v}\right)\right\|^2\right]\,. \tag{282}$$

We will bound each term separately. For the first term in Equation (282) we can write:

$$\mathop{\mathbb{E}}_{\boldsymbol{x}_1,\boldsymbol{x}_2,\boldsymbol{y}\sim\mathrm{Unif}(\mathbb{S}^{d-1})}\left[\left\|f(\boldsymbol{x}_1,\boldsymbol{x}_2;\boldsymbol{y})-g\left(\boldsymbol{v}_{\max}\right)\right\|^2\right] \tag{283}$$

$$=\mathbb{E}\left[\left\|f(\boldsymbol{x}_1,\boldsymbol{x}_2;\boldsymbol{y})-g\left(\boldsymbol{v}_{\max}\right)\right\|^2 \mid \langle\boldsymbol{x}_1,\boldsymbol{x}_2\rangle\leq 0.1\right]\cdot\Pr(\langle\boldsymbol{x}_1,\boldsymbol{x}_2\rangle\leq 0.1)+ \tag{284}$$

$$+\mathbb{E}\left[\left\|f(\boldsymbol{x}_1,\boldsymbol{x}_2;\boldsymbol{y})-g\left(\boldsymbol{v}_{\max}\right)\right\|^2 \mid \langle\boldsymbol{x}_1,\boldsymbol{x}_2\rangle> 0.1\right]\cdot\Pr(\langle\boldsymbol{x}_1,\boldsymbol{x}_2\rangle> 0.1)\,. \tag{285}$$

By Lemma 38 the first term is bounded by ϵ. For the second term, note that $\left\|f(\boldsymbol{x}_1,\boldsymbol{x}_2;\boldsymbol{y})-g\left(\boldsymbol{v}_{\max}\right)\right\|^2\leq 2$ since the output of each function is a unit vector. Also, by standard concentration of random vectors on the unit sphere (see Section 3 in Vershynin (2018)), there is a universal constant $c_3>0$ such that $\Pr(\langle\boldsymbol{x}_1,\boldsymbol{x}_2\rangle>0.1)\leq\exp(-c_3 d)$. Hence, we can bound $\mathbb{E}\left[\left\|f(\boldsymbol{x}_1,\boldsymbol{x}_2;\boldsymbol{y})-g\left(\boldsymbol{v}_{\max}\right)\right\|^2\right]\leq\epsilon+2\exp(-c_3 d)$.

We will bound the second term in Equation (282) uniformly for any $\boldsymbol{x}_1,\boldsymbol{x}_2,\boldsymbol{y}$. Note that g is a ReLU neural network with 4 layers, width $c_1 d^2 H$ and weights bounded by 2. Hence, we can bound its Lipschitz constant by the multiplication of the Frobenius norm of its weights matrices, which is bounded by $\left(4(c_1 d^2 H)\right)^4\big)$. Hence:

$$\left\|g\left(\boldsymbol{v}_{\max}\right)-g\left(\boldsymbol{v}\right)\right\|^2\leq\left(4(c_1 d^2 H)\right)^4\big)\cdot\left\|\boldsymbol{v}_{\max}-\boldsymbol{v}\right\|^2 \tag{286}$$

$$\leq\left(4(c_1 d^2 H)\right)^4\big)\,H\cdot\max_h\left\|\boldsymbol{X}\,\mathrm{sm}(\boldsymbol{X}^\top\boldsymbol{M}_h\boldsymbol{y})-\arg\max_{\boldsymbol{x}_i}(\boldsymbol{x}_i^\top\boldsymbol{M}_h\boldsymbol{y})\right\|^2. \tag{287}$$

There is $\delta>0$ which depends on ϵ such that for the set:

$$A_\delta:=\left\{\boldsymbol{x}_1,\boldsymbol{x}_2,\boldsymbol{y}\in\mathbb{S}^{d-1}:\forall\boldsymbol{q}_h,\ (\boldsymbol{x}_1-\boldsymbol{x}_2)^\top\boldsymbol{q}_h\boldsymbol{q}_h^\top\boldsymbol{y}>\delta\right\}, \tag{288}$$

we have that $\Pr((\boldsymbol{x}_1,\boldsymbol{x}_2,\boldsymbol{y})\notin A_\delta)\leq\frac{\epsilon}{(4(c_1 d^2 H))^4)2H}$. Note that $\boldsymbol{X}\,\mathrm{sm}(\alpha\boldsymbol{X}^\top\boldsymbol{q}_h\boldsymbol{q}_h^\top\boldsymbol{y})\underset{\alpha\to\infty}{\longrightarrow}\arg\max_{\boldsymbol{x}_i}(\boldsymbol{x}_i^\top\boldsymbol{q}_h\boldsymbol{q}_h^\top\boldsymbol{y})$ uniformly on A_δ for every $\boldsymbol{q}_h$. Hence, we can find $\alpha>0$ large enough such that:

$$\sup_{\boldsymbol{x}_1,\boldsymbol{x}_2,\boldsymbol{y}\in\mathbb{S}^{d-1}}\max_h\left\|\boldsymbol{X}\,\mathrm{sm}(\boldsymbol{X}^\top\boldsymbol{M}_h\boldsymbol{y})-\arg\max_{\boldsymbol{x}_i}(\boldsymbol{x}_i^\top\boldsymbol{M}_h\boldsymbol{y})\right\|^2\leq\frac{\epsilon}{(4(c_1 d^2 H))^4)\,H}\,. \tag{289}$$

This bounds $\mathbb{E}_{\boldsymbol{x}_1,\boldsymbol{x}_2,\boldsymbol{y}\sim\mathrm{Unif}(\mathbb{S}^{d-1})}\left[\left\|g\left(\boldsymbol{v}_{\max}\right)-g\left(\boldsymbol{v}\right)\right\|^2\right]\leq\epsilon$.

Combining both bounds from Equation (282) we have:

$$\mathop{\mathbb{E}}_{\boldsymbol{x}_1,\boldsymbol{x}_2,\boldsymbol{y}\sim\mathrm{Unif}(\mathbb{S}^{d-1})}\left[\left\|f(\boldsymbol{x}_1,\boldsymbol{x}_2;\boldsymbol{y})-g\left(\boldsymbol{v}\right)\right\|^2\right]\leq\epsilon+\exp(-c_3 d) \tag{290}$$

where we changed the constant c_3 accordingly. □

E.4 Proof of Theorem 5

Proof. We first define the construction. Let $\boldsymbol{q}_1,\ldots,\boldsymbol{q}_H$ be such that the conclusions of Lemma 38 are satisfied (e.g. by drawing them uniformly from the unit sphere). Let $\boldsymbol{E}=\begin{bmatrix}1&-1&0\\0&0&0\end{bmatrix}$. We call the second dimension of the positional encodings the "scratch space". We construct the heads of the first layer as follows: For each h, let

$$\boldsymbol{M}_h^{(1)}=\alpha\begin{bmatrix}\boldsymbol{q}_h\\0\\0\end{bmatrix}\begin{bmatrix}\boldsymbol{q}_h^\top&0&0\end{bmatrix}\qquad\boldsymbol{V}_h^{(1)}=\begin{bmatrix}\boldsymbol{0}\\0\\1\end{bmatrix}\begin{bmatrix}\boldsymbol{0}^\top&1&0\end{bmatrix} \tag{291}$$

The number of heads in the first layer is H. The weights of the second layer of the transformer are defined as:

$$\boldsymbol{M}_i^{(2)}=\begin{bmatrix}\boldsymbol{0}\\1\\0\end{bmatrix}\begin{bmatrix}\boldsymbol{0}^\top&0&1\end{bmatrix}\qquad\boldsymbol{V}_i^{(2)}=\beta\begin{bmatrix}\boldsymbol{e}_i\\0\\0\end{bmatrix}\begin{bmatrix}\boldsymbol{e}_i^\top&0&0\end{bmatrix} \tag{292}$$

for the standard basis vectors $\boldsymbol{e}_i$, and $\beta > 0$ will be defined later. The number of heads in the second layer is d. Finally, we set the output layer as $\boldsymbol{A} = \frac{1}{a}[\boldsymbol{I}_d \quad 0 \quad 0]$.

We will now prove the correctness of the construction. For the following argument, assume that each head uses hardmax instead of softmax. Note that by a similar argument used in the proof of Theorem 32, this incurs an extra loss of ϵ for any $\epsilon > 0$ at the cost of increasing α.

When the first layer is applied to the input $\boldsymbol{y}$, the scratch space of the output of each head is 1 if $\boldsymbol{x}_1^\top \boldsymbol{q}_h \boldsymbol{q}_h^\top \boldsymbol{y} > \boldsymbol{x}_2^\top \boldsymbol{q}_h \boldsymbol{q}_h^\top \boldsymbol{y}$ and -1 otherwise. Let $s_y, s_{\boldsymbol{x}_1}, s_{\boldsymbol{x}_1}$ be the sum of the scratch spaces of all the H heads (we will in fact only use $s_{\boldsymbol{y}}$). Note that $s_{\boldsymbol{y}} > 0$ if the majority of the heads outputted $\boldsymbol{x}_1$ and $s_{\boldsymbol{y}} < 0$ if the majority outputted for $\boldsymbol{x}_2$. All other dimensions of the output are 0. Thus, after the skip connection, the output of the first layer is

$$T^{(1)}\left(\begin{bmatrix} \boldsymbol{x}_1 & \boldsymbol{x}_2 & \boldsymbol{y} \\ 1 & -1 & 0 \\ 0 & 0 & 0 \end{bmatrix}\right) = \begin{bmatrix} \boldsymbol{x}_1 & \boldsymbol{x}_2 & \boldsymbol{y} \\ 1 & -1 & 0 \\ s_{\boldsymbol{x}_1} & s_{\boldsymbol{x}_2} & s_{\boldsymbol{y}} \end{bmatrix}. \tag{293}$$

For the second layer of attention, note that each head attends to $\boldsymbol{x}_1$ if $s_{\boldsymbol{y}} > 0$ and to $\boldsymbol{x}_2$ otherwise. By summing d such heads, where each head corresponds to some standard basis vector, the output of the second layer is

$$T^{(2)}\left(T^{(1)}\left(\begin{bmatrix} \boldsymbol{x}_1 & \boldsymbol{x}_2 & \boldsymbol{y} \\ 1 & -1 & 0 \\ 0 & 0 & 0 \end{bmatrix}\right)\right) = \begin{bmatrix} \boldsymbol{y} \\ 0 \\ s_{\boldsymbol{y}} \end{bmatrix} + \beta \begin{bmatrix} \boldsymbol{x}_1 \\ 1 \\ s_{\boldsymbol{x}_1} \end{bmatrix} \tag{294}$$

if $s_{\boldsymbol{y}} > 0$, and the same with $\boldsymbol{x}_2$ if $s_{\boldsymbol{y}} > 0$. Finally, the after the output layer, the output of the entire transformer is $\frac{1}{\beta}\boldsymbol{y} + \boldsymbol{x}_1$ if $s_{\boldsymbol{y}} > 0$, or $\frac{1}{\beta}\boldsymbol{y} + \boldsymbol{x}_2$ otherwise.

By taking first take $\beta > \frac{1}{\epsilon}$, we get that the output of the transformer is the same as the output of the majority of the rank-1 attention heads of the first layer of the transformer, up to an extra error of ϵ. By Lemma 38 and taking the number of heads H to be large enough, we get that the majority of the heads in the first layer approximates the target up to an error of ϵ. Scaling ϵ appropriately finishes the proof.

□