Benefits of Rank in Attention Layers

Noah Amsel, Gilad Yehudai, Joan Bruna

Motivation

- How do we get the most out of transformers?
- How do we set the hyperparameters?

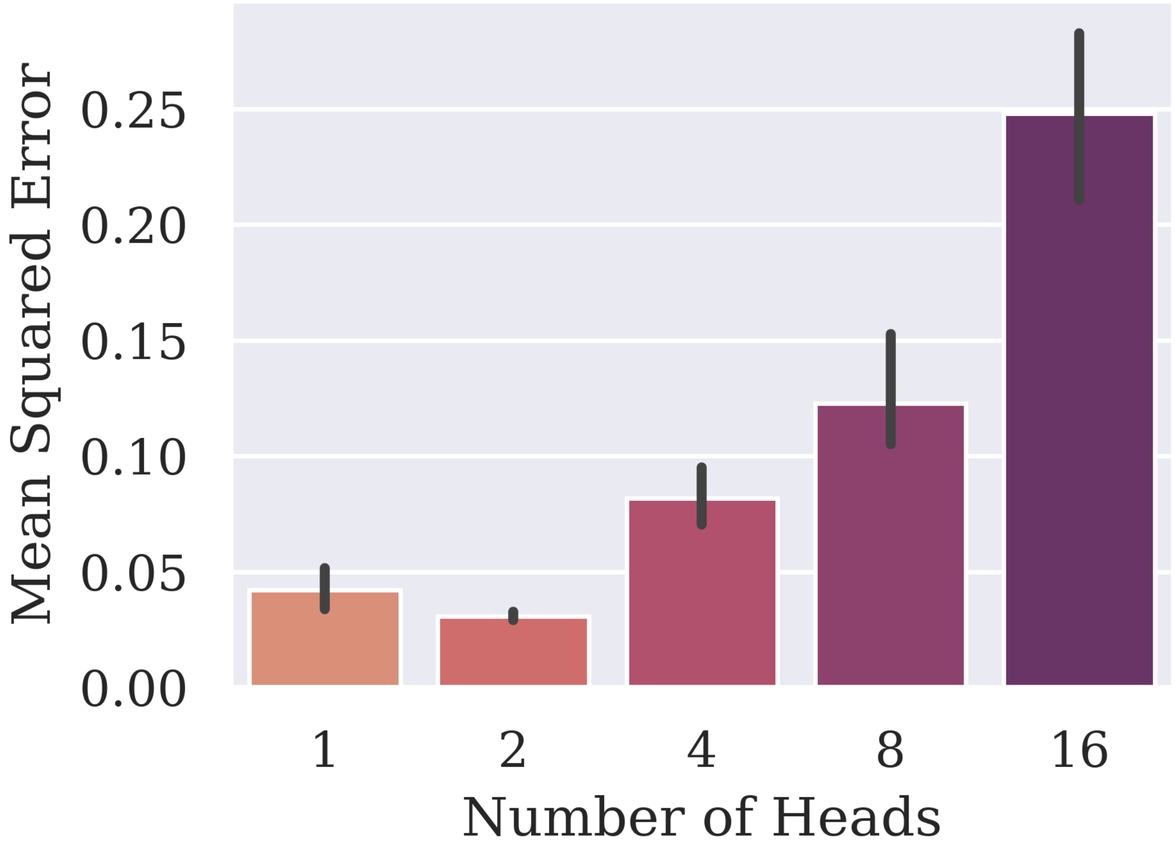
What Can Transformers Learn In-Context?
A Case Study of Simple Function Classes

Shivam Garg*
Stanford University
shivamg@cs.stanford.edu

Dimitris Tsipras*
Stanford University
tsipras@stanford.edu

Percy Liang
Stanford University
pliang@cs.stanford.edu

Gregory Valiant
Stanford University
valiant@stanford.edu



Ancient History

- "Neural Machine Translation by Jointly Learning to Align and Translate" [Bahdanau, Cho & Bengio, 2016]
  - Given query $\mathbf{q}$, keys $\{\mathbf{k}_j\}$, and values $\{\mathbf{v}_j\}$ in $\mathbb{R}^r$
  - Attention score: $s_j = \mathbf{k}_j \cdot \mathbf{q}$ $\quad\quad$ $\mathbf{s} = \mathbf{K}\mathbf{q}$
  - Attention coefficients: $\alpha_j = \dfrac{\exp(s_j)}{\sum_k \exp(s_k)}$ $\quad\quad$ $\alpha = \mathrm{sm}(\mathbf{s})$
  - Convex combo of values: $\sum_j \alpha_j \mathbf{v}_j$ $\quad\quad$ $\mathbf{V}\alpha$ $\quad\quad$ $= \mathbf{V}\ \mathrm{sm}(\mathbf{K}^\top \mathbf{q})$

Singe-head Attention

Given $\mathbf{y}$ and $\mathbf{x}_1, \ldots \mathbf{x}_N \in \mathbb{R}^d$, let

$$\mathbf{q} = \mathbf{W}^Q \mathbf{y} \qquad \mathbf{K} = \mathbf{W}^K \mathbf{X} \qquad \mathbf{V} = \mathbf{W}^V \mathbf{X}$$

where $\mathbf{W}^Q, \mathbf{W}^K, \mathbf{W}^V, \mathbf{W}^O \in \mathbb{R}^{r \times d}$.

$$\left(\mathbf{W}^O\right)^\top \mathbf{W}^V \mathbf{X} \ \mathrm{sm}\left(\left(\mathbf{W}^K \mathbf{X}\right)^\top \mathbf{W}^Q \mathbf{y}\right)$$

Multi-head Attention

Given $\mathbf{y}$ and $\mathbf{x}_1, \ldots \mathbf{x}_N \in \mathbb{R}^d$

$$\sum_{h=1}^{H} \left(\mathbf{W}_h^O\right)^\top \mathbf{W}_h^V \, \mathbf{X} \, \mathrm{sm}\left(\left(\mathbf{W}_h^K \, \mathbf{X}\right)^\top \mathbf{W}_h^Q \, \mathbf{y}\right)$$

where $\mathbf{W}_h^Q, \mathbf{W}_h^K, \mathbf{W}_h^V, \mathbf{W}_h^O \in \mathbb{R}^{r \times d}$.

- Num parameters: $4rHd$

Multi-head Attention

Given $\mathbf{y}$ and $\mathbf{x}_1, \ldots \mathbf{x}_N \in \mathbb{R}^d$

$$\sum_{h=1}^{H} \mathbf{W}_h \, \mathbf{X} \, \mathrm{sm}\left(\mathbf{X}^\top \mathbf{M}_h \mathbf{y}\right)$$

where $\mathbf{M}_h, \mathbf{W}_h \in \mathbb{R}^{d \times d}$ are rank-r.

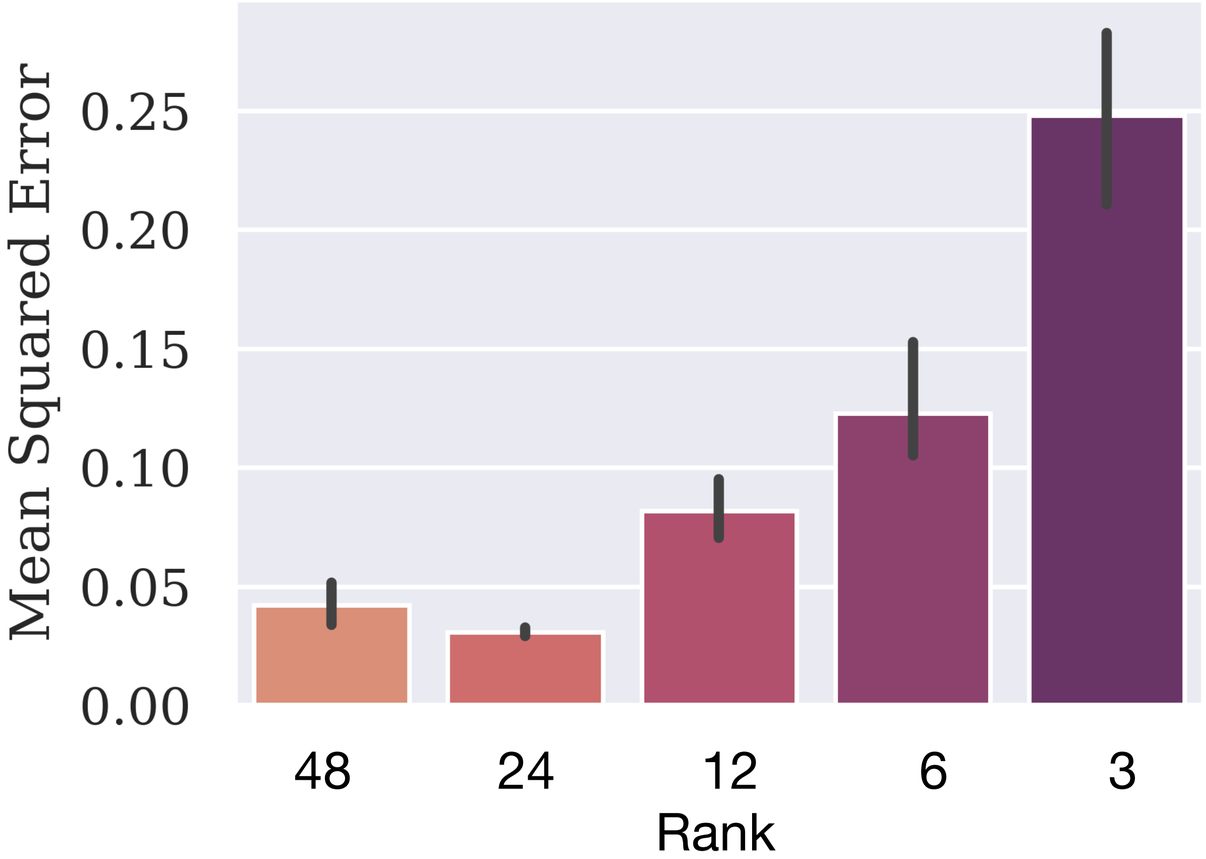
- Num parameters: $4rHd$

			num layers	MLP depth	value rank			
		hidden dim		MLP width		attn rank		num heads
Year	Model	d	L	w	D	r	r_2	H
2017	Attention is all you need [59]	512	6	$4d$	2	64	r	d/r

Num params $= 4rHd = 4d^2$

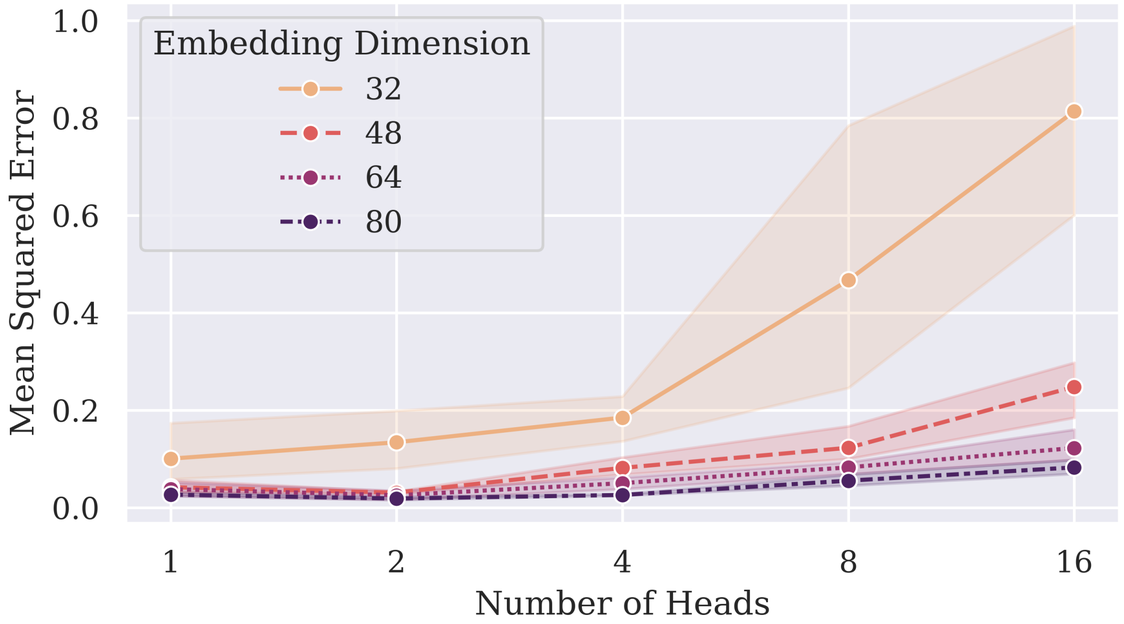
Same as $H = 1,\ r = d$

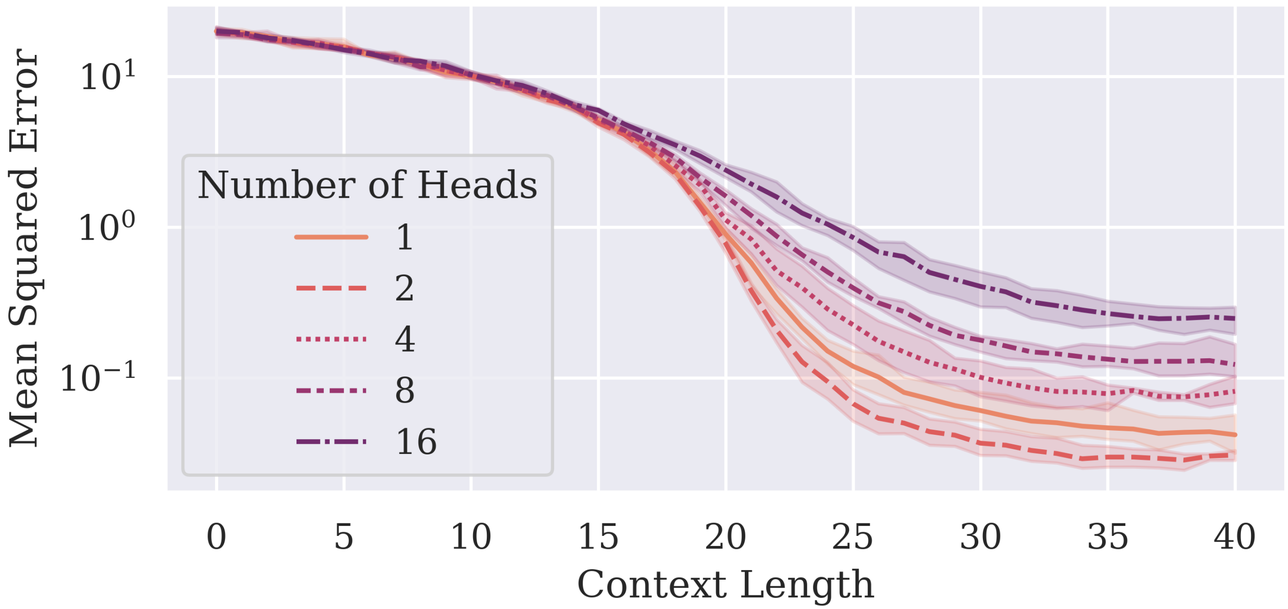
Back to Garg et al.



r = d / H

More Garg et al.





Theory papers assume full-rank

- Practitioners use low r and $H = d/r$
- Theoreticians assume $r = d$ and $H \gg 1$
  - Is this a good proxy for real-world transformers?

Related Work

Depth Separation for MLPs

- **Q:** 2-layer NNs are universal approximators. So why are deep NNs better?
- **A:** Deep NNs are more parameter-efficient
  - Pay for some depth, save a *lot* of width
- **Thm:** There exists a function which
  - depth 3 MLP can represent with width d^5
  - depth 2 MLP needs width e^{cd} to approximate
- [Eldan & Shamir ‘16] [Telgarsky ‘16] [Daniely ‘17] [Chatziafratis+ ’19] …

Separations in Expressive Power

Construct a function showing that you can…

MLP:

…pay a bit for depth, save a *lot* of width

Multi-head attention layer:

…pay a bit for rank, save a *lot* of heads

Another rank separation result

- “Representational Strengths and Limitations of Transformers” [Sanford, Hsu, Telgarsky, 2023]
- **Thm:** A single layer of multi-head self-attention cannot compute Match3 unless $rHp > \tilde{\Omega}(N)$
  - Proved using communication complexity (p is precision)
- **Q:** Is low r a limitation, or just low rH?
- We extend hardness guarantee to…
  - Prohibitively large H
  - $p = \infty$
  - ϵ-approximation

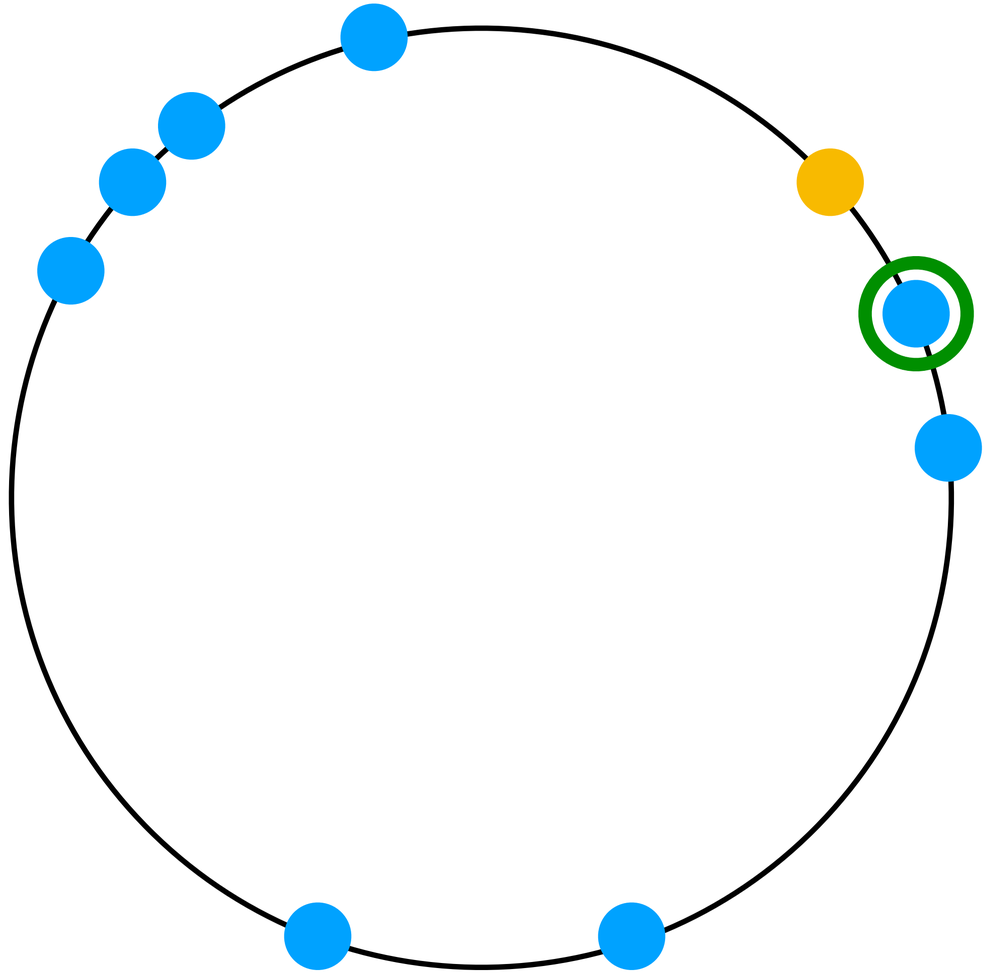
Our Rank Separation

Nearest Neighbor Function

target points $\mathbf{x}_i$

source point $\mathbf{y}$

output $f\,(\mathbf{x}_1, \ldots, \mathbf{x}_N; \mathbf{y})$



Target Function: Nearest Neighbor

For $\mathbf{y}, \mathbf{x}_1, \ldots, \mathbf{x}_N \in \mathbb{S}^{d-1}$

$$f(\mathbf{x}_1, \ldots, \mathbf{x}_N; \mathbf{y}) := \operatorname*{argmin}_{\mathbf{x} \in \{\mathbf{x}_1, \ldots, \mathbf{x}_N\}} \|\mathbf{x} - \mathbf{y}\|_2$$

$$= \operatorname*{argmax}_{\mathbf{x} \in \{\mathbf{x}_1, \ldots, \mathbf{x}_N\}} \mathbf{x}^\top \mathbf{y}$$

$$= \mathbf{X}\ \mathrm{hm}(\mathbf{X}^\top \mathbf{y})$$

$$= \underbrace{\mathbf{I}_d}\ \mathbf{X}\ \mathrm{sm}\left(\mathbf{X}^\top \underbrace{\left(10^{10} \cdot \mathbf{I}_d\right)}\ \mathbf{y}\right)$$

full rank

Upper Bound

For $\mathbf{y}, \mathbf{x}_1, \ldots, \mathbf{x}_N \in \mathbb{S}^{d-1}$, the nearest neighbor function can be exactly represented by a single full-rank hardmax attention head.

"$r = d,\ H = 1$ suffices"

Lower Bound (Informal)

If $H \lesssim (d/r)^{1/\epsilon}$, then

$$\mathop{\mathbb{E}}_{\mathbf{x}_1 \perp \mathbf{x}_2,\ \mathbf{y} \sim \mathcal{U}(\mathbb{S}^{d-1})} \left\| f(\mathbf{x}_1, \mathbf{x}_2; \mathbf{y}) - \sum_{h=1}^{H} \mathrm{attn}_h(\mathbf{x}_1, \mathbf{x}_2; \mathbf{y}) \right\|_2^2 \geq \epsilon$$

Generalizing Attention

Standard:

$$\sum_{h=1}^{H} \left(\mathbf{W}_h^O\right)^\top \mathbf{W}_h^V \, \mathbf{X} \, \mathrm{sm}\left(\left(\mathbf{W}_h^K \, \mathbf{X}\right)^\top \mathbf{W}_h^Q \, \mathbf{y}\right)$$

where $\mathbf{W}_h^Q, \mathbf{W}_h^K, \mathbf{W}_h^V, \mathbf{W}_h^O \in \mathbb{R}^{r \times d}$.

Generalized:

$$\sum_{h=1}^{H} \mathbf{V}_h \, \mathbf{X} \, \phi_h\left(\mathbf{K}_h^\top \mathbf{X}, \, \mathbf{y}\right)$$

where $\mathbf{V}_h \in \mathbb{R}^{d \times d}, \quad \mathbf{K}_h \in \mathbb{R}^{d \times r}, \quad \phi_h : \mathbb{R}^{r \times N} \times \mathbb{R}^d \to \Delta^N$.

Proof Sketch

Reduction to Scalar Function on $\mathbb{S}^{d-1} \times \mathbb{S}^{d-1}$

$$\mathop{\mathbb{E}}_{\mathbf{x}_1 \perp \mathbf{x}_2,\ \mathbf{y} \sim \mathcal{U}(\mathbb{S}^{d-1})} \left\| f(\mathbf{x}_1, \mathbf{x}_2; \mathbf{y}) - \sum_{h=1}^{H} \mathrm{attn}_h(\mathbf{x}_1, \mathbf{x}_2; \mathbf{y}) \right\|_2^2$$

$$\geq \frac{1}{2} \mathop{\mathbb{E}}_{\mathbf{x},\ \mathbf{y} \sim \mathcal{U}(\mathbb{S}^{d-1})} \left(\mathrm{sgn}\left(\mathbf{x}^\top \mathbf{y}\right) - \sum_{h=1}^{H} g_h(\mathbf{x}, \mathbf{y}) \right)^2$$

(follows by projecting onto $\mathbf{x}_1 - \mathbf{x}_2$)

where $g_h(\mathbf{x}, \mathbf{y}) = \tilde{\phi}_h\left(\mathbf{K}_h^\top \mathbf{x}, \mathbf{y}\right)$

Spherical Harmonics

An orthonormal basis for $L^2(\mathbb{S}^{d-1} \to \mathbb{R})$

Frequency / Degree:

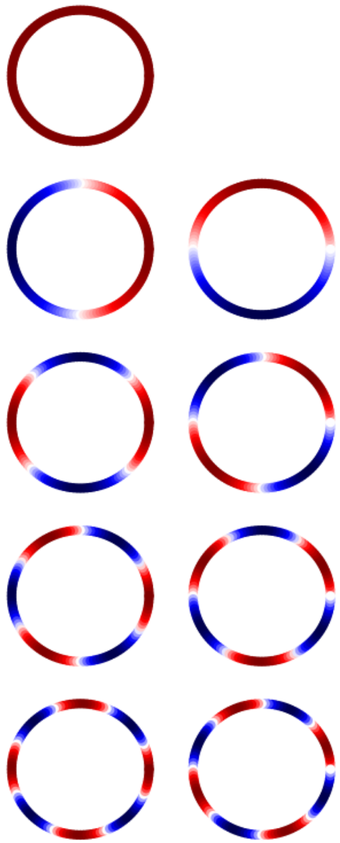
0

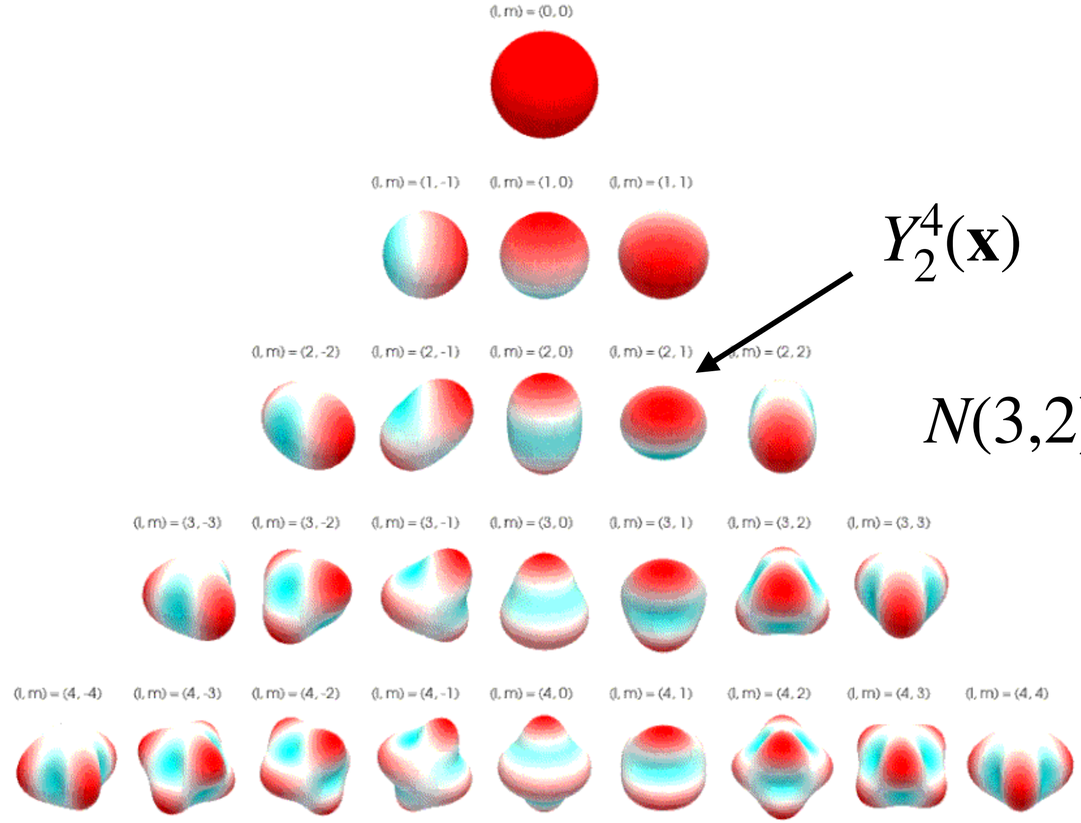
1

2

3

4





$Y_2^4(\mathbf{x})$

$N(3,2) = 5$

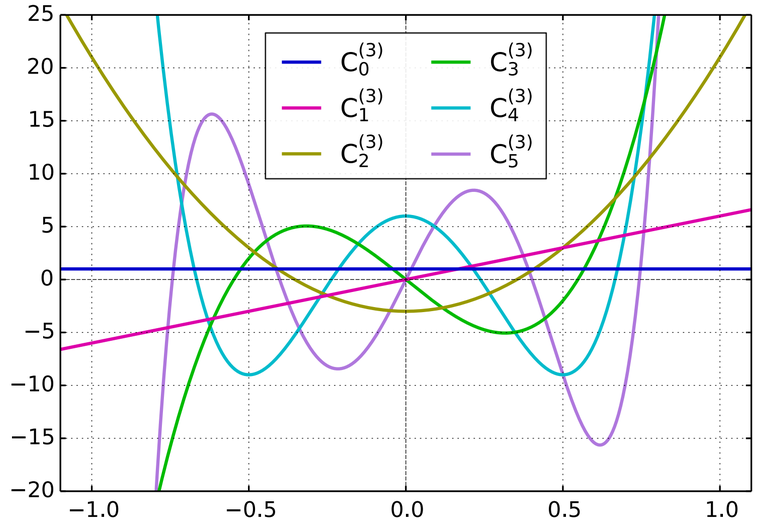
basis elements of degree ℓ in on $\mathbb{S}^{d-1}$ is $N(d, \ell)$

Spherical harmonic expansion of rank-1 functions

- Spherical harmonic expansion: $f = \sum_{\ell=0}^{\infty} \sum_{i=1}^{N(d,\ell)} \langle f, Y_\ell^i \rangle \cdot Y_\ell^i$
  - where $\langle f, g \rangle = \mathop{\mathbb{E}}_{\|\mathbf{x}\|=1} \left[f(\mathbf{x}) g(\mathbf{x}) \right]$
- Rank-1 function: $f(\mathbf{x}) = \psi(\mathbf{x}^\mathsf{T} \mathbf{a})$ for $\psi : \mathbb{R} \to \mathbb{R}$
- Theorem (Hecke-Funk):

$$\langle f, Y_\ell^i \rangle = Y_\ell^i(\mathbf{a}) \cdot \langle \psi, P_\ell \rangle$$

  - where P_ℓ is the ℓ^{th} Gegenbauer polynomial



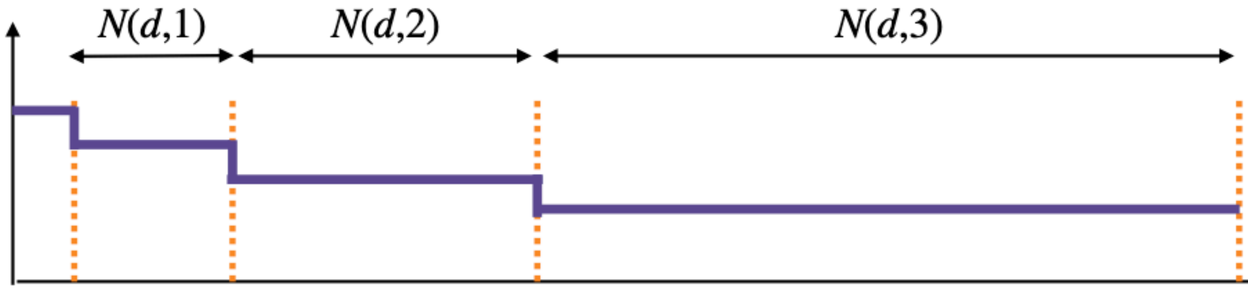
Representing $\operatorname{sgn}\left(\mathbf{x}^\top \mathbf{y}\right)$ with spherical harmonics

- $\left\{Y_\ell^i \otimes Y_{\ell'}^{i'}\right\}_{\ell,\ell',i,i'}$ is orthonormal basis for $L^2\left(\left(\mathbb{S}^{d-1} \times \mathbb{S}^{d-1}\right) \to \mathbb{R}\right)$
- By Hecke-Funk

$$\left\langle \operatorname{sgn}(\mathbf{x}^\top \mathbf{y}), Y_\ell^i(\mathbf{x}) Y_{\ell'}^{i'}(\mathbf{y}) \right\rangle = \langle \operatorname{sgn}, P_\ell \rangle \cdot \langle Y_\ell^i, Y_{\ell'}^{i'} \rangle = \eta_\ell \cdot \begin{cases} 1 & \ell = \ell', i = i' \\ 0 & \text{o.w.} \end{cases}$$

$$\implies \operatorname{sgn}(\mathbf{x}^\top \mathbf{y}) = \sum_{\ell=0}^{\infty} \sum_{i=1}^{N(d,\ell)} \eta_\ell Y_\ell^i(\mathbf{x}) Y_\ell^i(\mathbf{y})$$

where $\eta_\ell \sim \ell^{-1}$



Expansion of the head functions

- $\phi_h\left(\mathbf{K}_h^\top \mathbf{x}, \mathbf{y}\right)$ only cares about an r-dimensional projection of $\mathbf{x}$

  $\Rightarrow$ ortho to many spherical harmonics, e.g. $\phi_h\left(\begin{bmatrix} x_1 \\ x_2 \end{bmatrix}, \mathbf{y}\right) \perp x_3^5$

- **Lemma:** $\phi_h\left(\mathbf{K}_h^\top \mathbf{x}, \mathbf{y}\right)$ is orthogonal to $Y_\ell \otimes Y_\ell$ for all Y_ℓ in ℓ^{th} harmonic except for a subspace of dimension $M(d, \ell) \leq \binom{r+\ell}{\ell}$

  - out of total dimension $N(d, \ell)$
  - All H heads are spanned by $\leq H \cdot M(d, \ell)$ harmonics

Combining

$$\mathop{\mathbb{E}}_{\mathbf{x},\,\mathbf{y}\sim\mathcal{U}(\mathbb{S}^{d-1})}\left(\operatorname{sgn}\left(\mathbf{x}^{\top}\mathbf{y}\right)-\sum_{h=1}^{H}g_h(\mathbf{x},\mathbf{y})\right)^2$$

$$=\left\|\sum_{\ell=0}^{\infty}\left[\sum_{i=1}^{N(d,\ell)}\eta_\ell Y_\ell^i\otimes Y_\ell^i-\sum_{h=1}^{H}\sum_{i=1}^{M(d,\ell)}?_{h,\ell,i}Y_\ell^i\otimes Y_\ell^i\right]\right\|^2$$

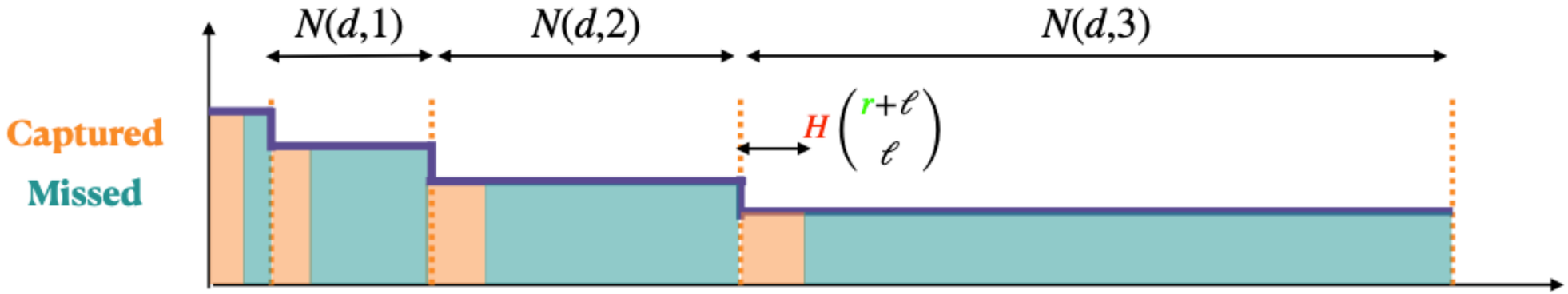
$$\geq\left\|\sum_{\ell=0}^{\infty}\sum_{i=H\cdot M(d,\ell)}^{N(d,\ell)}\eta_\ell\cdot Y_\ell^i\otimes Y_\ell^i\right\|^2=\sum_{\ell=0}^{\infty}\eta_\ell^2\left(N(d,\ell)-H\cdot M(d,\ell)\right)$$

Combining

- Very roughly,

$$\sum_{\ell=0}^{\infty} \eta_\ell^2 \left(N(d,\ell) - H \cdot M(d,\ell) \right) \gtrsim \sum_{\ell} \ell^{-2} \left(d^\ell - H r^\ell \right)$$

- So unless $H > d^p / r^p$ we are left with $\sum_{\ell > p} l^{-2} = p^{-1}$ error



[End Proof Sketch]

Lower Bound

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*

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

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

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

think of $H \lesssim (d/r)^{1/\epsilon}$

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

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

Alternative Lower Bound

There exists a target function f^* such that unless $H \cdot \max_h \|\mathbf{V}_h\| \lesssim c^{d-r}$, the error of approximation is bounded as follows:

$$\mathop{\mathbb{E}}_{\substack{\boldsymbol{x}_1,\boldsymbol{x}_2 \sim \mathcal{D}_2\left(\mathbb{S}^{d-1}\right) \\ \boldsymbol{y} \sim \mathcal{N}\left(0, \frac{1}{\sqrt{d}} I\right)}} \left[\left\|T(\boldsymbol{x}_1, \boldsymbol{x}_2, \boldsymbol{y}) - f^*(\boldsymbol{x}_1, \boldsymbol{x}_2, \boldsymbol{y})\right\|^2\right] > \frac{1}{40}$$

- Differences: bias in upper bound, dependence on $\|\mathbf{V}_h\|$, proof techniques

What about more layers?

- Low-rank attention layers are weaker, even for large H
- **Q:** Is a low-rank *transformer* weaker? … idk
- Construction: *modified* rank-1 transformer works for $N = 2$
- Conjecture: low-rank transformer fails for large N, unlike full-rank one

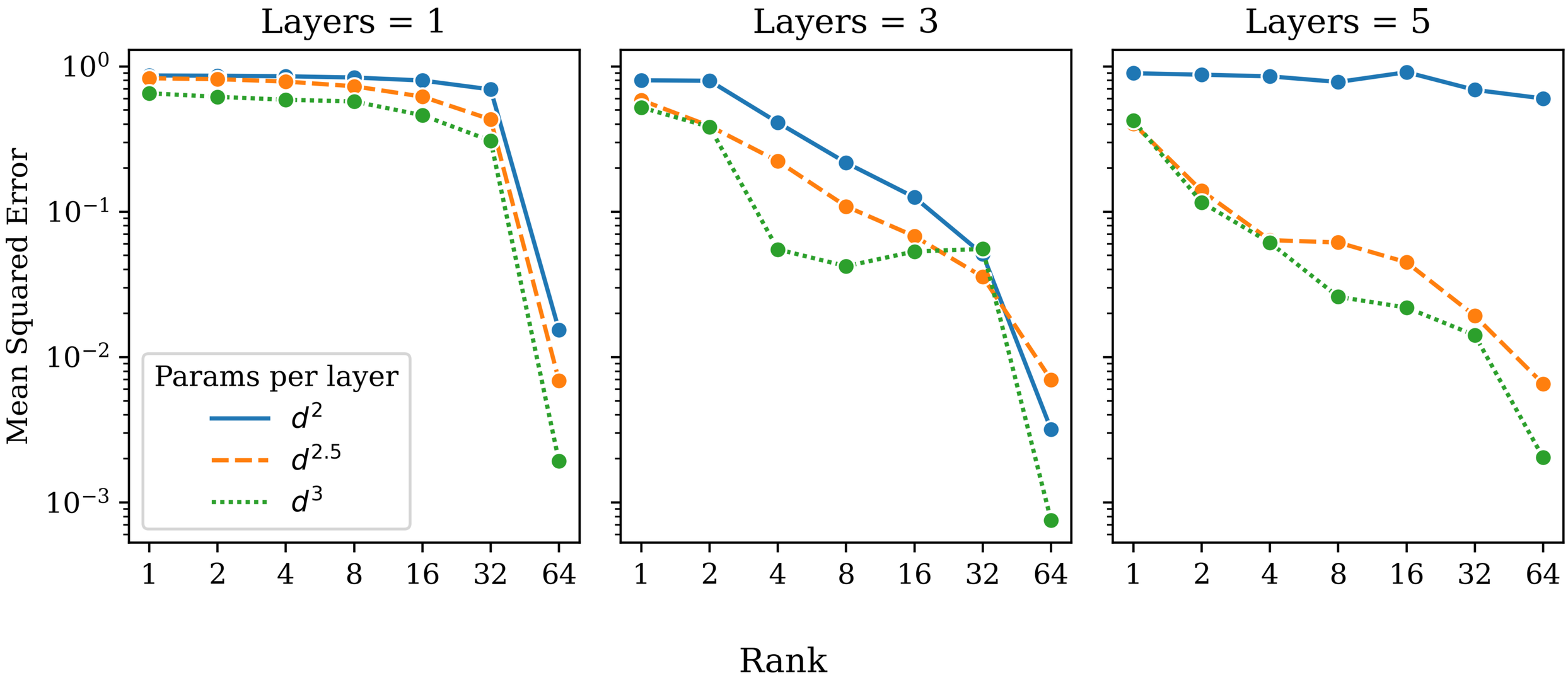
Modified Attention Construction: $L = 2$, $N = 2$

- Random rank-1 head $\mathrm{hm}\left(\mathbf{X}^\top \mathbf{q}\mathbf{q}^\top \mathbf{y}\right)$ guesses correctly w.p. $\sim \frac{1}{2} + \frac{1}{\sqrt{d}}$
- Majority vote of many such heads is correct with high probability
- To tally the votes, need extra “index” and “scratchpad” dimensions
  - Input: $\begin{bmatrix} \mathbf{x}_1 \\ 1 \\ 0 \end{bmatrix}, \begin{bmatrix} \mathbf{x}_2 \\ -1 \\ 0 \end{bmatrix}, \begin{bmatrix} \mathbf{y} \\ 0 \\ 0 \end{bmatrix}$
- Attn layer 1: each head votes. Sum the votes in index dimension. Save in $\mathbf{y}$’s scratchpad dimension
- Attn layer 2: look up the target $\mathbf{x}_1$ or $\mathbf{x}_2$ whose sign matches the tally

Experiments

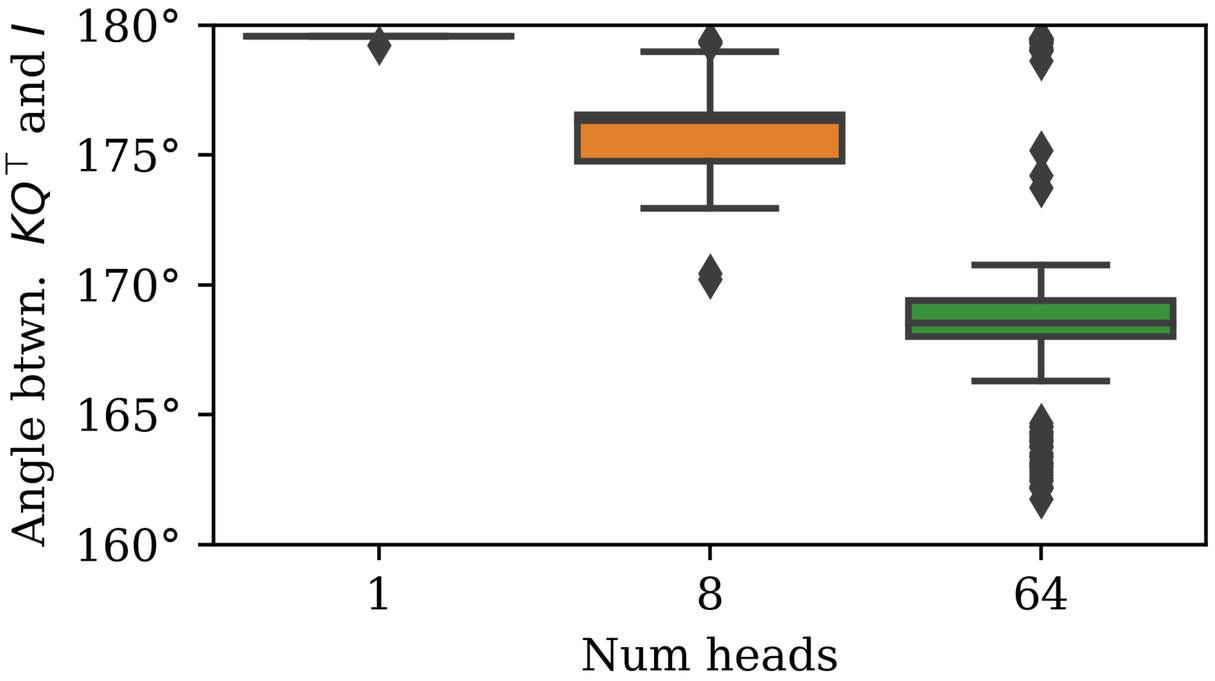
Experimental Setup

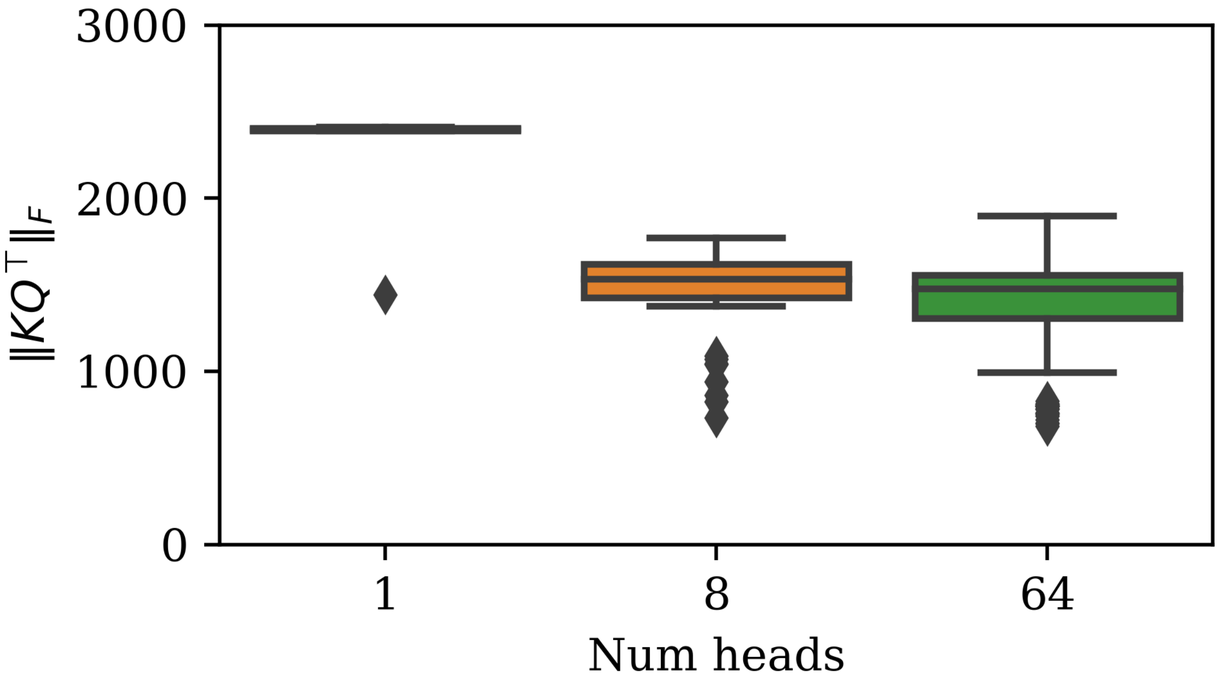
- Off-the-shelf multilayer transformers from PyTorch (but $H = d^c/r$, RMSNorm)
- "Farthest neighbor" with self-attention $\mathbf{x}_1, \ldots \mathbf{x}_N \sim (\mathbb{S}^{d-1})$
- No positional encodings (yes biases)
- Fix $d = 64, N = 16$
- Best of 5 runs


Layers = 1
Layers = 3
Layers = 5
Mean Squared Error
10^0
10^{-1}
10^{-2}
10^{-3}
Params per layer
d^2
$d^{2.5}$
d^3
1
2
4
8
16
32
64
Rank

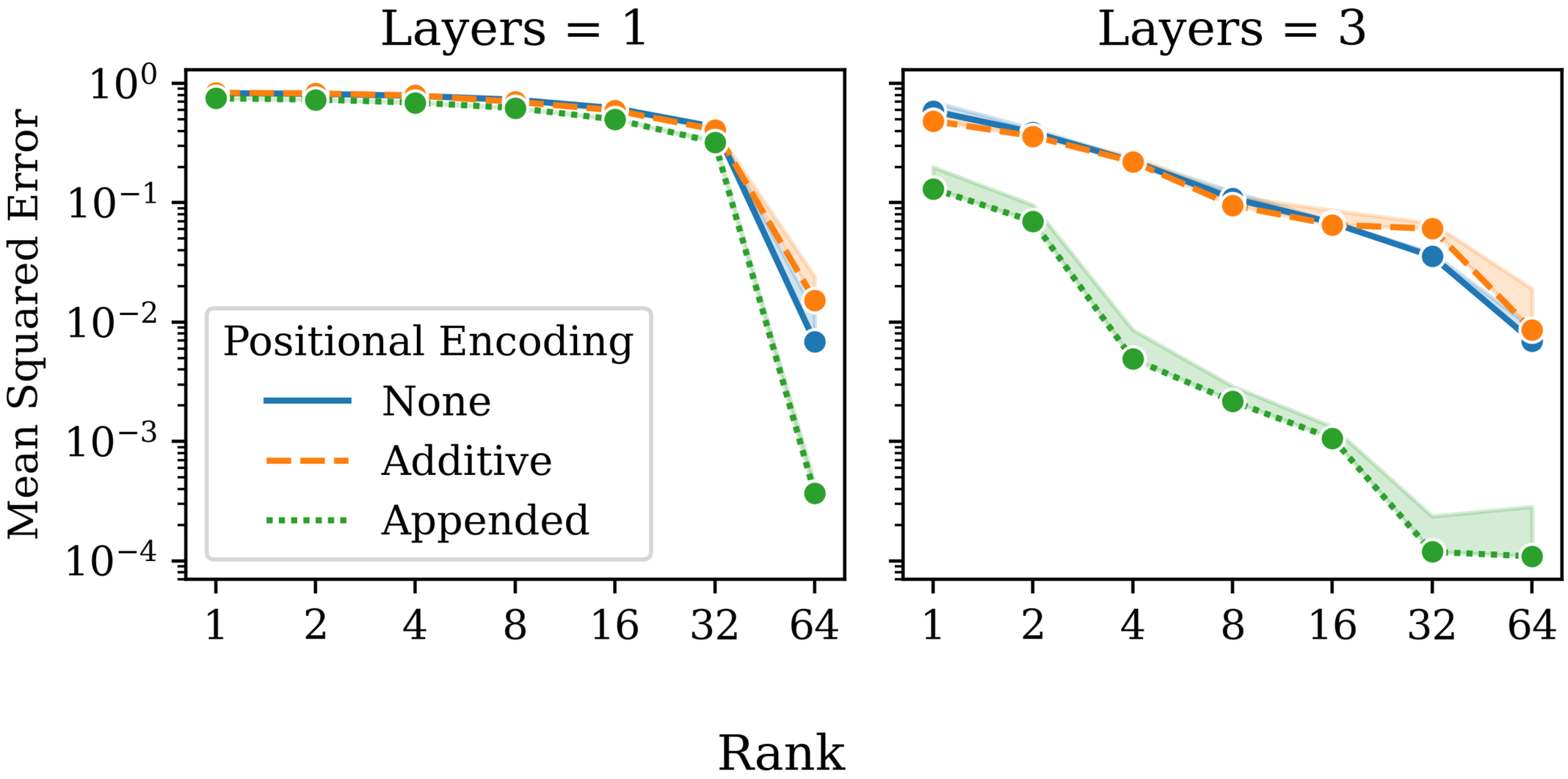
What does full-rank attention learn?

- Expected: $\mathbf{I}_d\ \mathbf{X}\ \mathrm{sm}\left(\mathbf{X}^\top\left(-10^{10}\cdot\mathbf{I}_d\right)\mathbf{y}\right)$

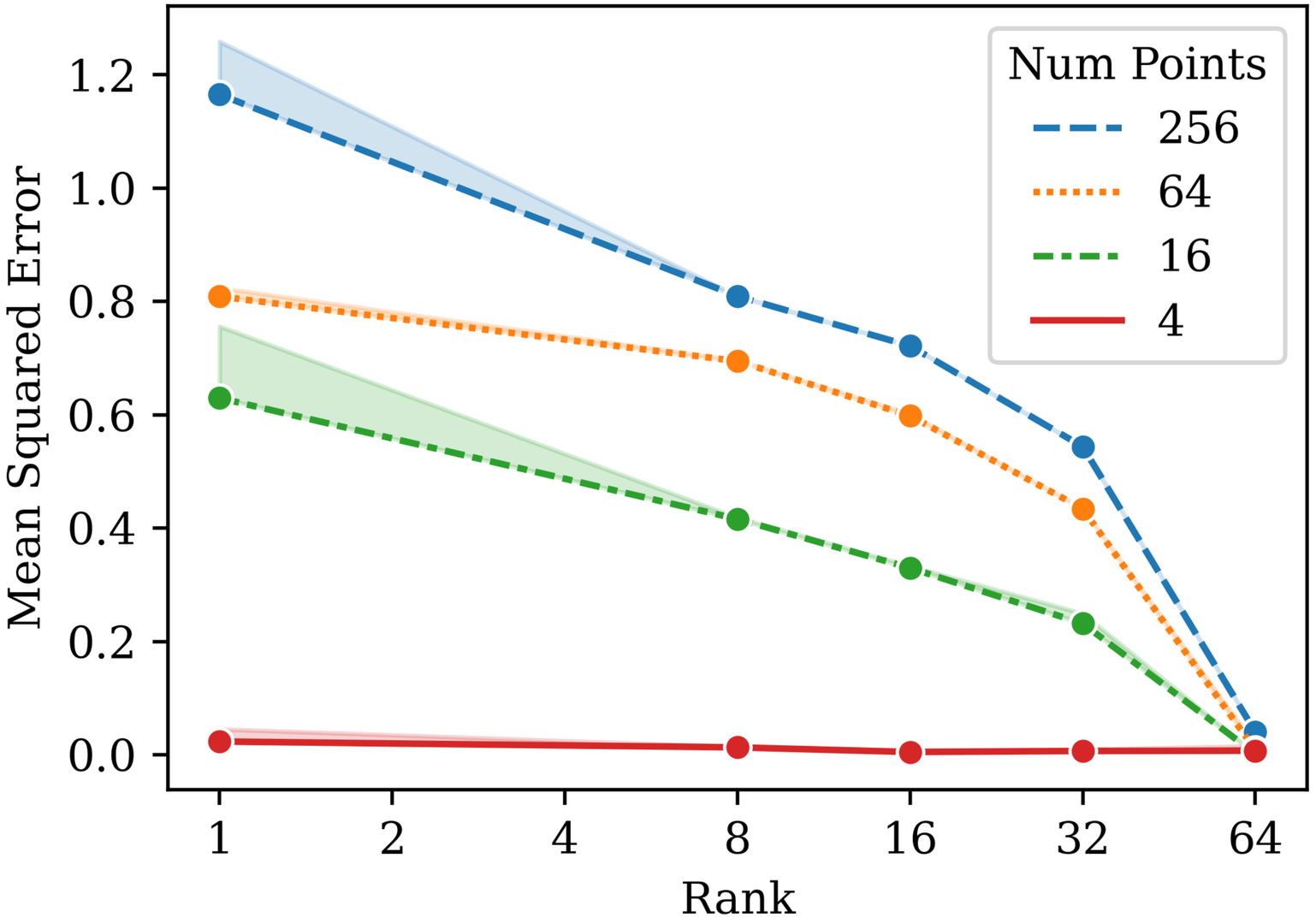




Modifying Attention



Role of N


Num Points
256
64
16
4
Mean Squared Error
1.2
1.0
0.8
0.6
0.4
0.2
0.0
1
2
4
8
16
32
64
Rank

Conclusion

Low-rank attention is fundamentally weaker than full-rank attention, even for $H \gg d/r$

Open / in progress

- Is the lower bound tight? (pretty much: just use random rank-1 heads)
- Can we prove hardness for $L > 1$? Is our conjecture true?
- Other tradeoffs in transformer hyperparameters besides r and H
- Are there hard problems that look more like text (not isotropic?)