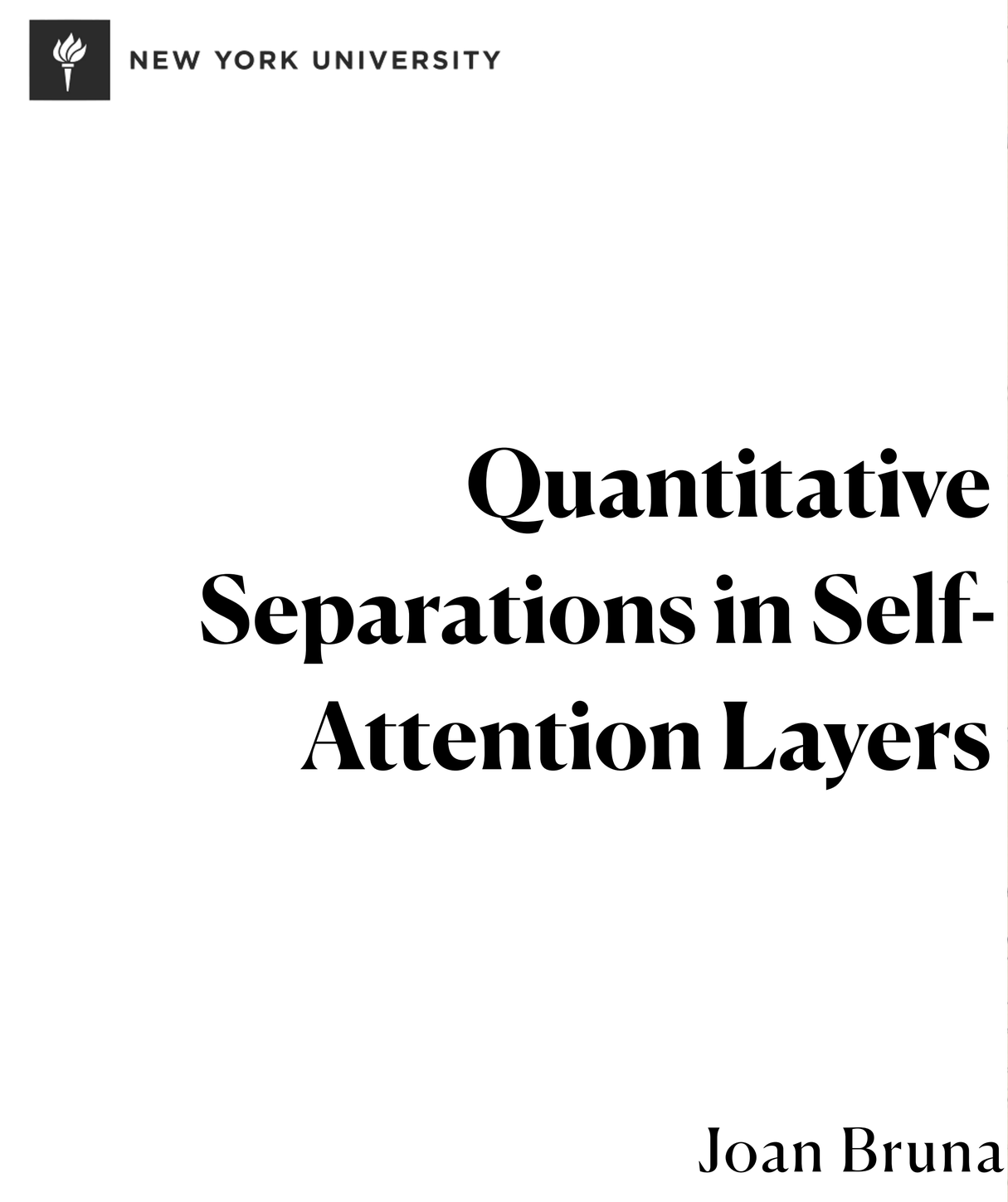
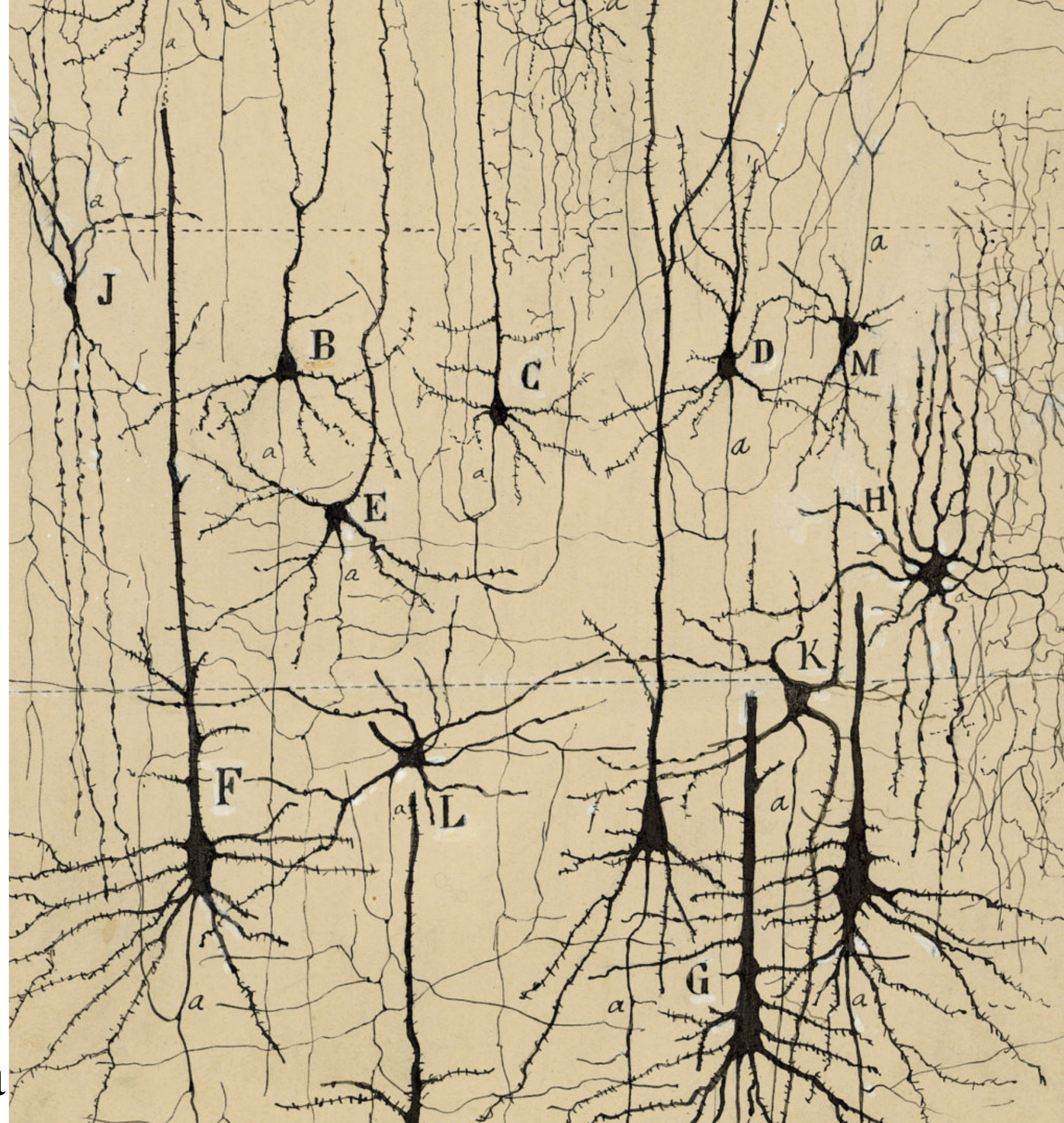


NEW YORK UNIVERSITY

Quantitative Separations in Self-Attention Layers

Joan Bruna



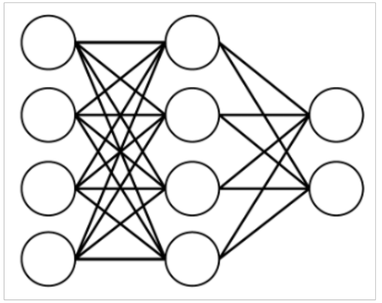
Joint Work with

Noah Amsel

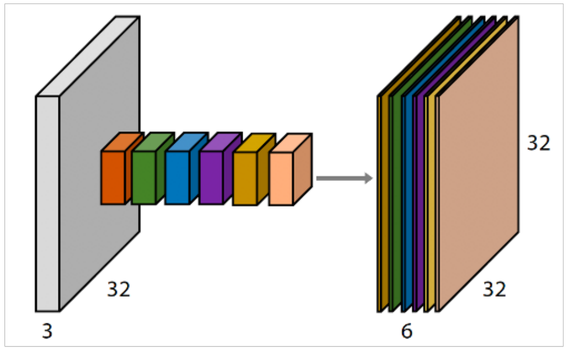
Gilad Yehudai

Aaron Zweig

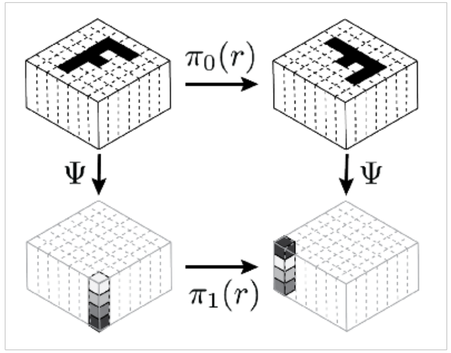
A "Zoo" of Neural Architectures



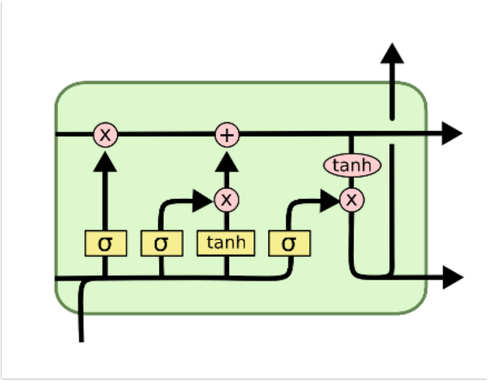
Perceptrons
Function regularity



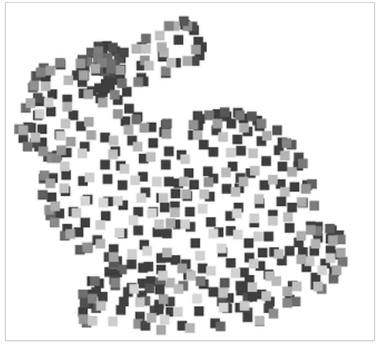
CNNs
Translation



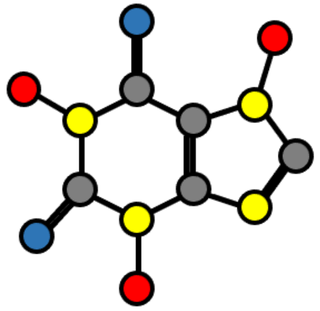
Group-CNNs
Translation+Rotation,
Global groups



LSTMs
Time warping



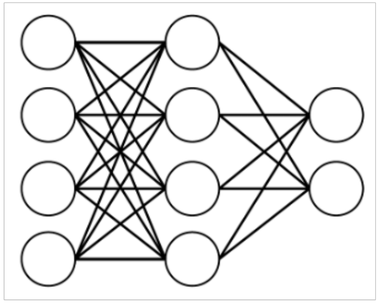
DeepSets / Transformers
Permutation



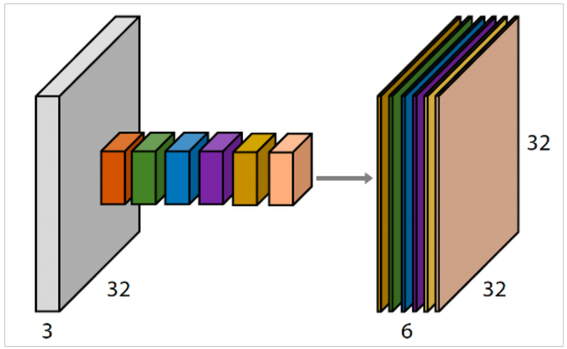
GNNs
Permutation

Intrinsic CNNs
Isometry / Gauge choice

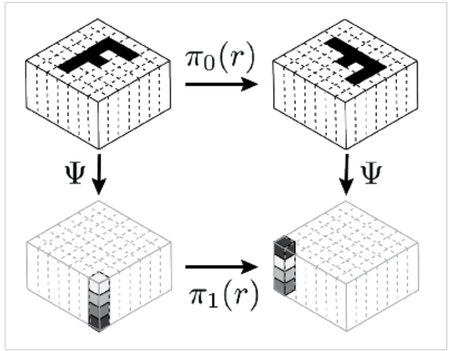
A "Zoo" of Neural Architectures



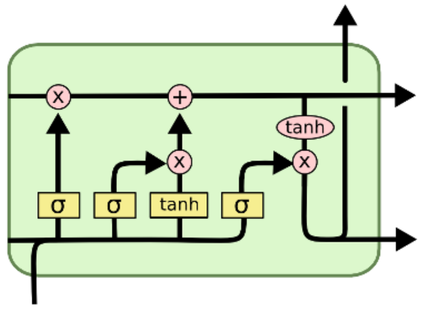
Perceptrons
Function regularity



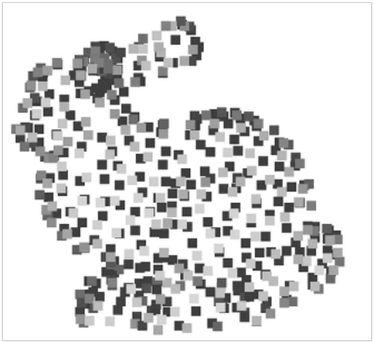
CNNs
Translation



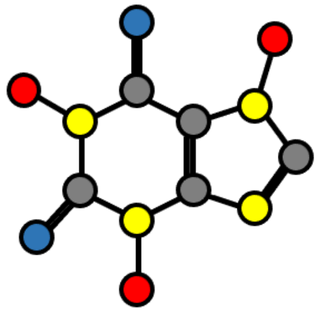
Group-CNNs
Translation+Rotation,
Global groups



LSTMs
Time warping



DeepSets / Transformers
Permutation



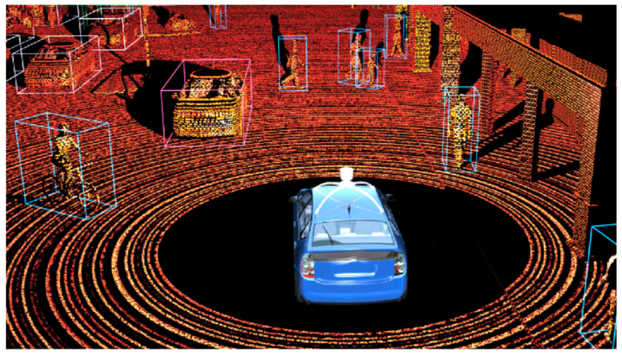
GNNs
Permutation

Intrinsic CNNs
Isometry / Gauge choice

Symmetric Functions

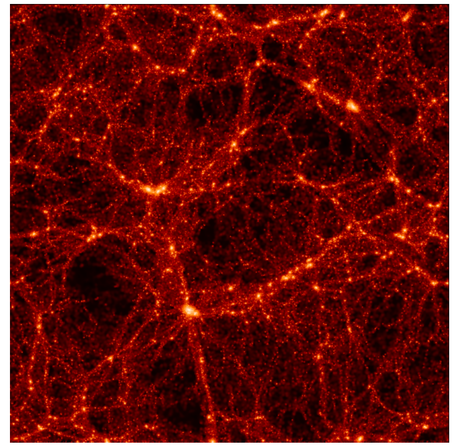
- Permutation-invariant functions: given domain Ω, consider $f : \Omega^{\otimes N} \to \mathbb{R}$ such that $f(x_{\sigma(1)}, \ldots, x_{\sigma(N)}) = f(x_1, \ldots, x_N)$, for any input X and permutation σ.
- Equivariant analog: $f : \Omega^{\otimes N} \to \mathscr{Y}^{\otimes N}$ such that $f(\sigma \,.\, X) = \sigma \,.\, f(X)$.

Symmetric Functions



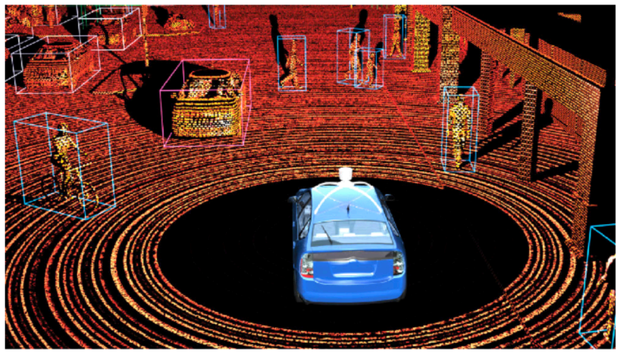
(Source: S. Grunewald, credit Qi et al.)

- Permutation-invariant functions: given domain Ω, consider $f : \Omega^{\otimes N} \to \mathbb{R}$ such that $f(x_{\sigma(1)}, \ldots, x_{\sigma(N)}) = f(x_1, \ldots, x_N)$, for any input X and permutation σ.
- Equivariant analog: $f : \Omega^{\otimes N} \to \mathcal{Y}^{\otimes N}$ such that $f(\sigma \,.\, X) = \sigma \,.\, f(X)$.
- Very broad range of application in Science (cosmology, biology)
  - Canonical model for *exchangeable* systems of particles.



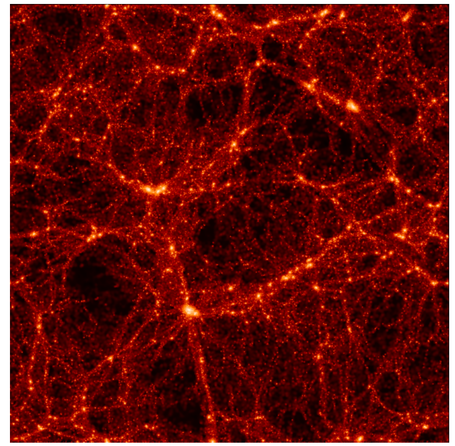
Joerg Colberg, Virgo Simulation

Symmetric Functions



(Source: S. Grunewald, credit Qi et al.)

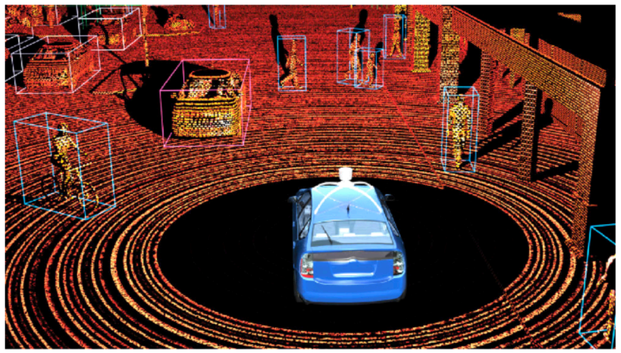
- Permutation-invariant functions: given domain Ω, consider $f : \Omega^{\otimes N} \to \mathbb{R}$ such that $f(x_{\sigma(1)}, \ldots, x_{\sigma(N)}) = f(x_1, \ldots, x_N)$, for any input X and permutation σ.
- Equivariant analog: $f : \Omega^{\otimes N} \to \mathcal{Y}^{\otimes N}$ such that $f(\sigma \,.\, X) = \sigma \,.\, f(X)$.
- Very broad range of application in Science (cosmology, biology)
  - Canonical model for *exchangeable* systems of particles.
- Universal model for structured (and discrete) inputs via *positional encodings* [Varswani et al.'17].



Joerg Colberg, Virgo Simulation

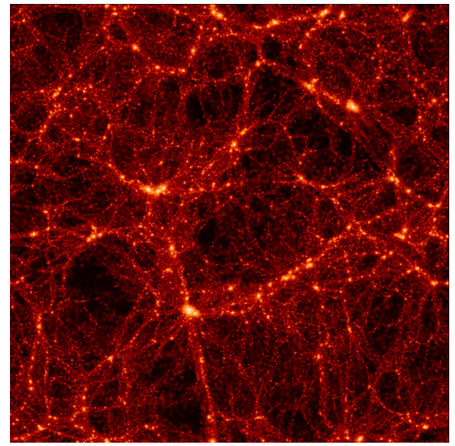
"The quick brown fox jumps over the lazy dog"

Symmetric Functions



(Source: S. Grunewald, credit Qi et al.)

- Permutation-invariant functions: given domain Ω, consider $f : \Omega^{\otimes N} \to \mathbb{R}$ such that $f(x_{\sigma(1)}, \ldots, x_{\sigma(N)}) = f(x_1, \ldots, x_N)$, for any input X and permutation σ.
- Equivariant analog: $f : \Omega^{\otimes N} \to \mathcal{Y}^{\otimes N}$ such that $f(\sigma \,.\, X) = \sigma \,.\, f(X)$.
- Very broad range of application in Science (cosmology, biology)
  - Canonical model for *exchangeable* systems of particles.
- Universal model for structured (and discrete) inputs via *positional encodings* [Varswani et al.'17].



Joerg Colberg, Virgo Simulation

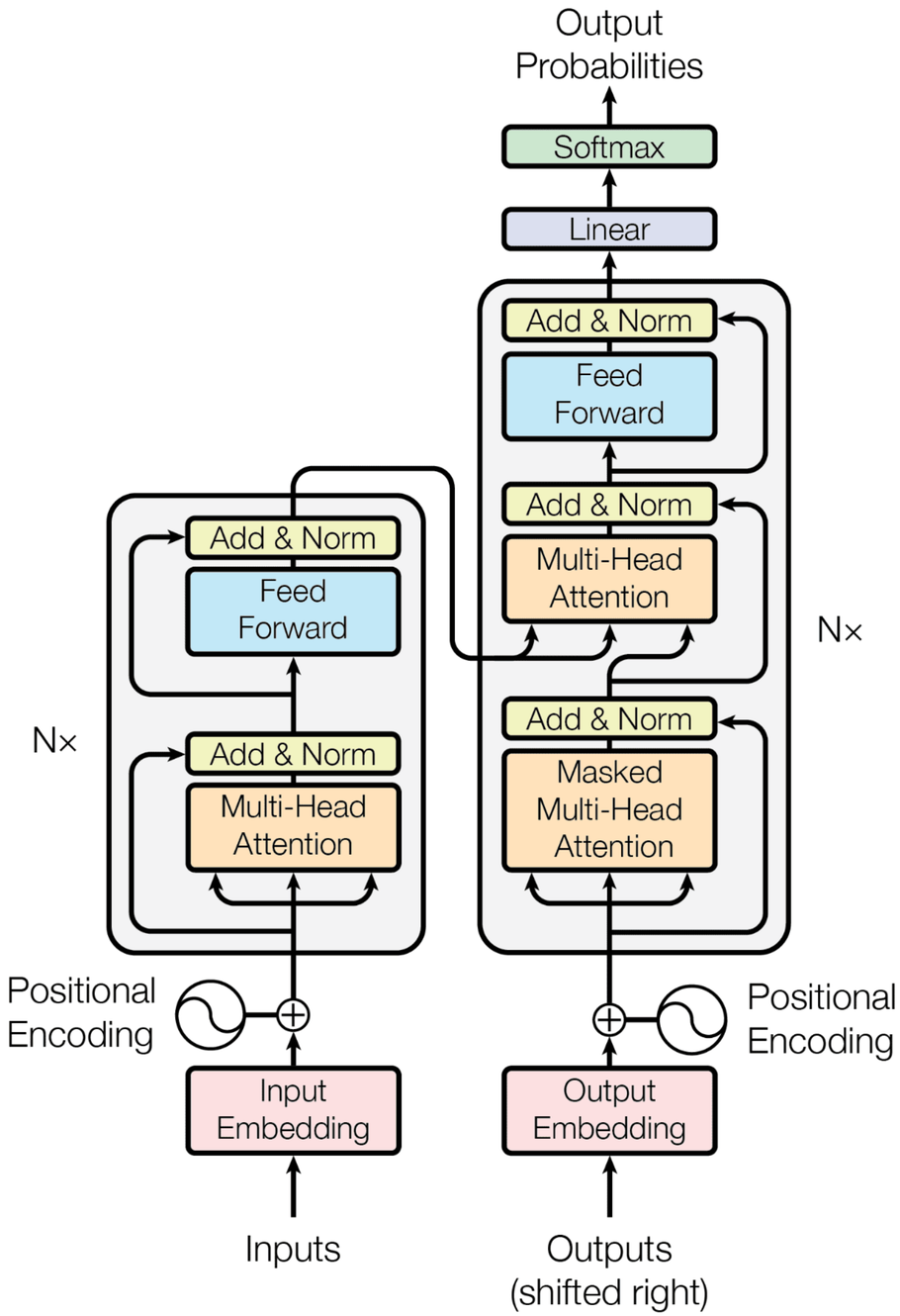
"The quick brown fox jumps over the lazy dog"

What are the canonical architectures, and what distinguishes them?

Transformers

[Varswani et al'18]

- Powers most modern large-scale NNs.
- Enables 'soft' equivariance via positional encodings.



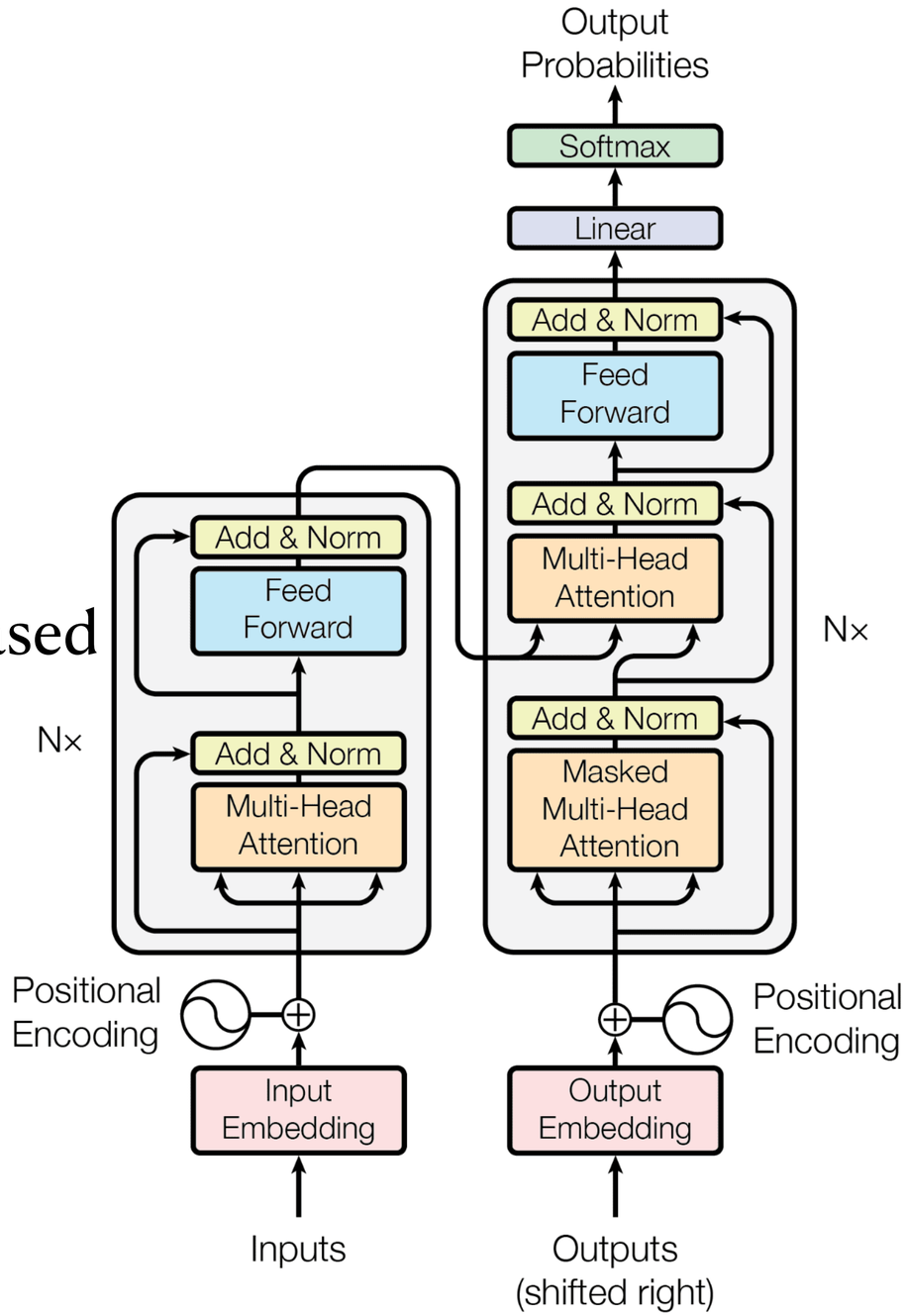
Transformers

[Varswani et al'18]

- Powers most modern large-scale NNs.
- Enables 'soft' equivariance via positional encodings.
- Key component: multiple self-attention mechanisms based on ***pairwise*** interactions:

  $S_h = \sigma(XK_hQ_h{}^\top X^\top)$, $K_h, Q_h \in \mathbb{R}^{d\times r}$,

  σ : row-wise softmax .



Transformers

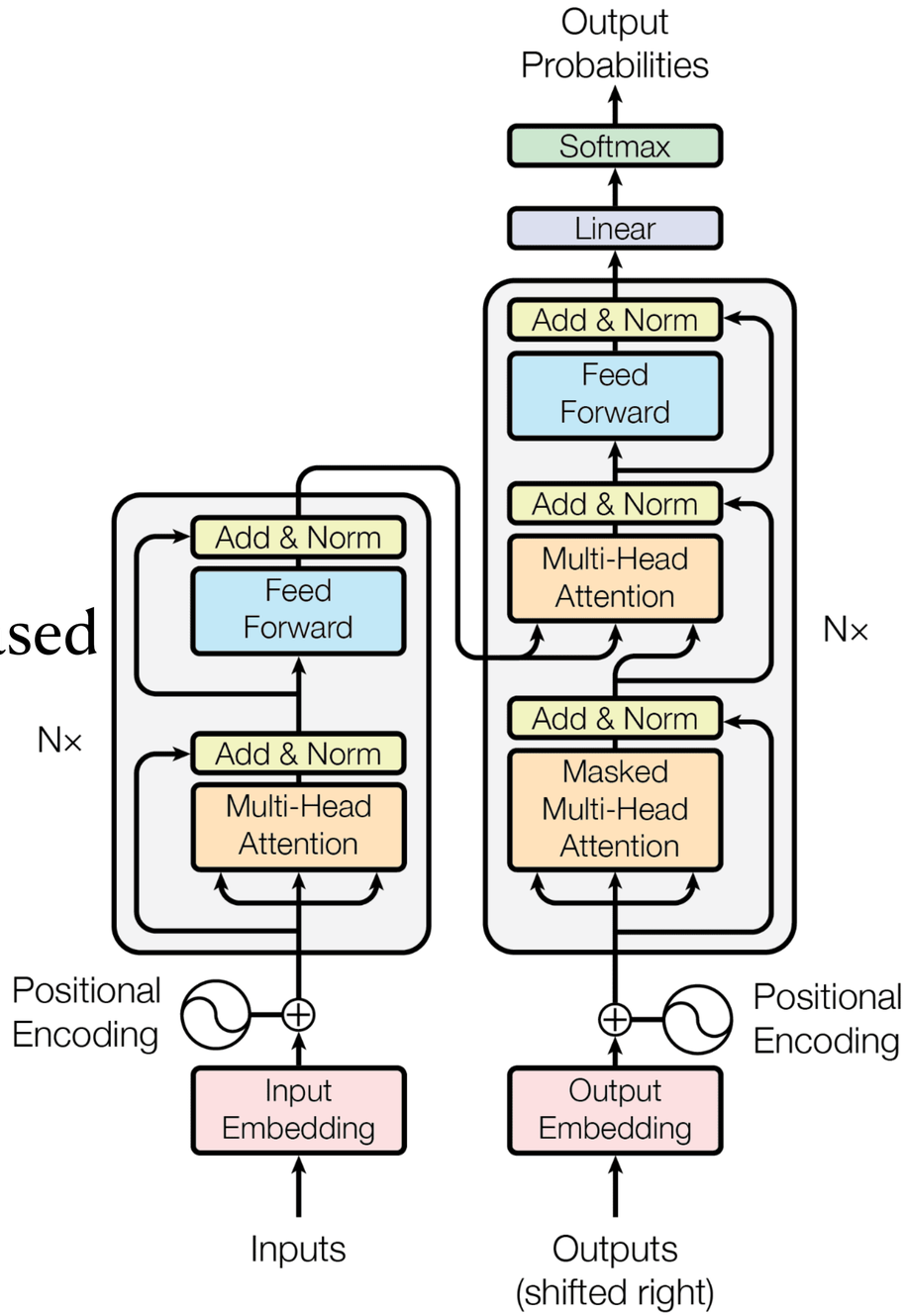
[Varswani et al'18]

- Powers most modern large-scale NNs.
- Enables 'soft' equivariance via positional encodings.
- Key component: multiple self-attention mechanisms based on ***pairwise*** interactions:

  $S_h = \sigma(XK_hQ_h{}^\top X^\top)$, $K_h, Q_h \in \mathbb{R}^{d\times r}$,

  σ : row-wise softmax .

Quantify advantage of self-attention?
What are the key parameters of the architecture?

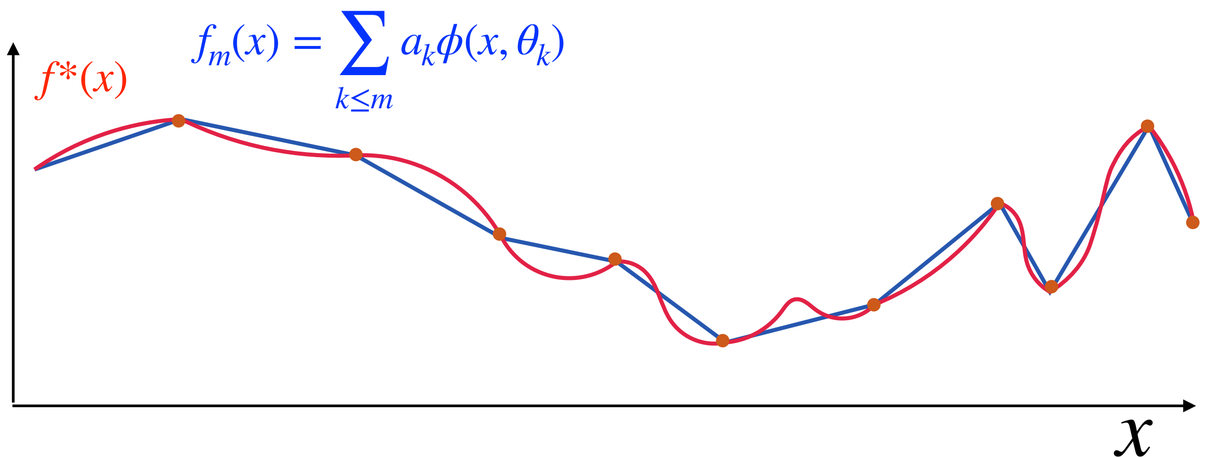


Approximation Power

- How ‘powerful’ is a given NN architecture?

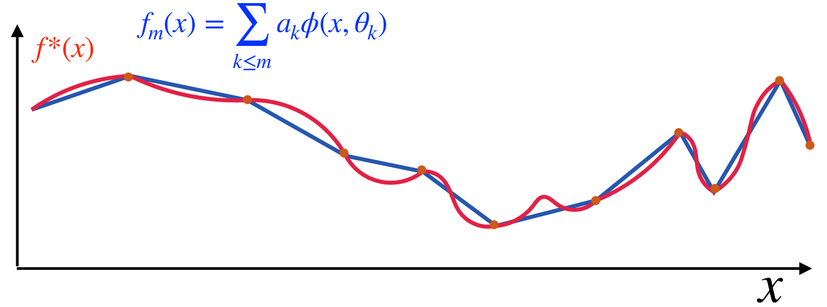
Approximation Power

- How ‘powerful’ is a given NN architecture?
- ***Qualitative*** picture $\|f^* - f_m\| \to 0$ as $m \to \infty$.
  - Universal Approximation Theorems [Cybenko, Hornik, Pinkus, ...] for MLPs.
  - [Keriven&Peyré,Maron et al,Morris et al,...] GNN expressiveness (eg WL based)



Approximation Power

- How ‘powerful’ is a given NN architecture?
- ***Qualitative*** picture $\|f^* - f_m\| \to 0$ as $m \to \infty$.
  - Universal Approximation Theorems [Cybenko, Hornik, Pinkus, ...] for MLPs.
  - [Keriven&Peyré,Maron et al,Morris et al,...] GNN expressiveness (eg WL based)
- ***Quantitative*** picture provides approximation rates, eg $\|f^* - f_m\| = \Theta(m^{-\alpha/d})$.
  - Enables separations between architectures, eg *depth* separation [Eldan & Shamir, Telgarsky, Daniely, ...]



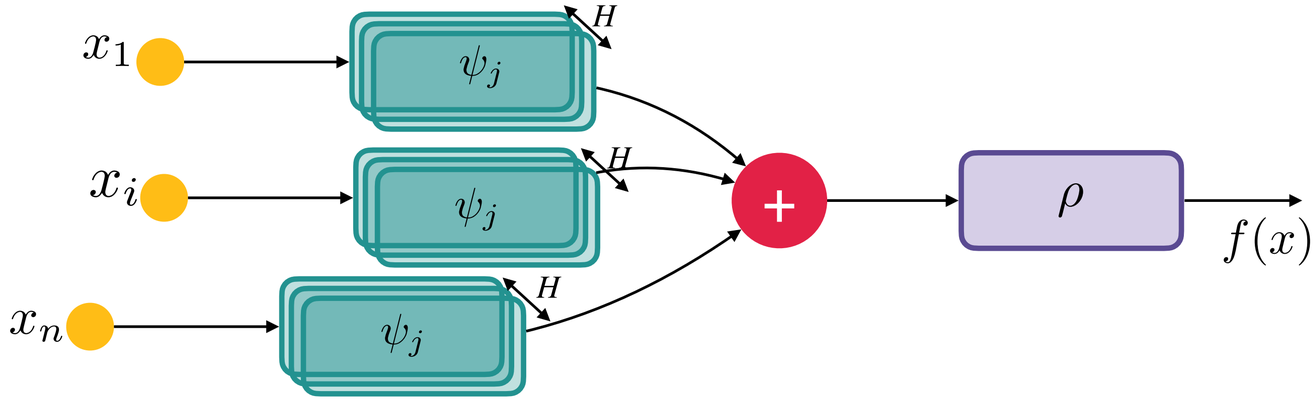
Program

1. Separation advantage of self-attention.
   joint work with Aaron Zweig

2. Rank-Width tradeoffs in transformer self-attention layers.
   joint work with Noah Amsel and Gilad Yehudai.

DeepSet

[Zaheer et al.]

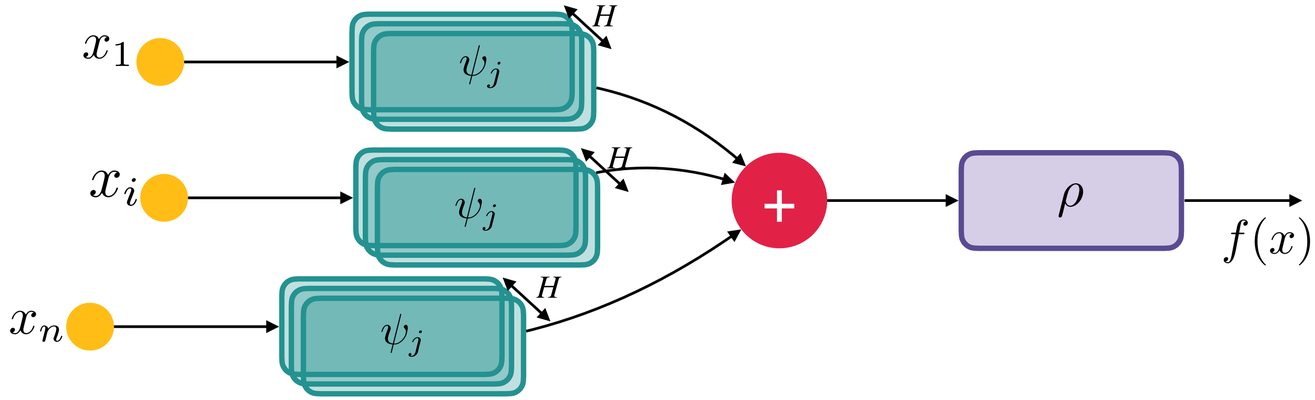


- Let $\mathrm{Sym}(H, N, d)$ be the class of all (symmetric) functions $f : (\mathbb{C}^d)^{\otimes N} \to \mathbb{C}$ of the form $f(X) = \rho(\phi_1(X), \ldots, \phi_H(X))$, with $\phi_j(X) = \sum_{i \leq N} \psi_j(x_i), j = 1, \ldots H$, where

  $\psi_j : \mathbb{C}^d \to \mathbb{C}, \rho : \mathbb{C}^H \to \mathbb{C}$ are arbitrary (smooth) maps.

DeepSet

[Zaheer et al.]



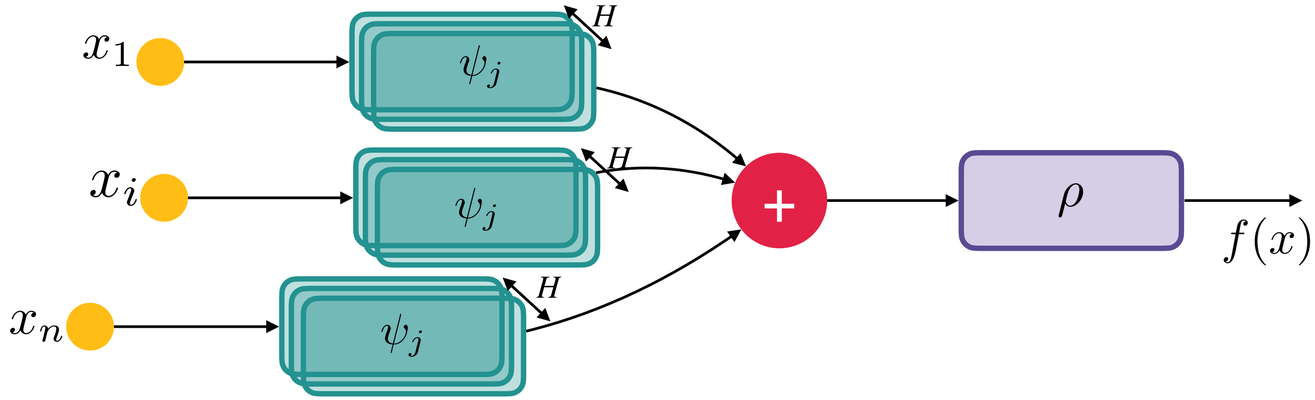
- Let Sym(H, N, d) be the class of all (symmetric) functions $f : (\mathbb{C}^d)^{\otimes N} \to \mathbb{C}$ of the form $f(X) = \rho(\phi_1(X), \ldots, \phi_H(X))$, with $\phi_j(X) = \sum_{i \leq N} \psi_j(x_i), j = 1, \ldots H$, where

  $\psi_j : \mathbb{C}^d \to \mathbb{C}, \rho : \mathbb{C}^H \to \mathbb{C}$ are arbitrary (smooth) maps.

- Q: Universal approximation properties in terms of problem parameters d, N, H ?

DeepSet

[Zaheer et al.]



- Let $\mathrm{Sym}(H, N, d)$ be the class of all (symmetric) functions $f : (\mathbb{C}^d)^{\otimes N} \to \mathbb{C}$ of the form $f(X) = \rho(\phi_1(X), \ldots, \phi_H(X))$, with $\phi_j(X) = \sum_{i \leq N} \psi_j(x_i), j = 1, \ldots H$, where

  $\psi_j : \mathbb{C}^d \to \mathbb{C}, \rho : \mathbb{C}^H \to \mathbb{C}$ are arbitrary (smooth) maps.

- Q: Universal approximation properties in terms of problem parameters d, N, H ?

- **Fact** [Schläfli,MacMahon,Weyl,Noerther]: If $H > H^* := \binom{d+N}{N}$, then

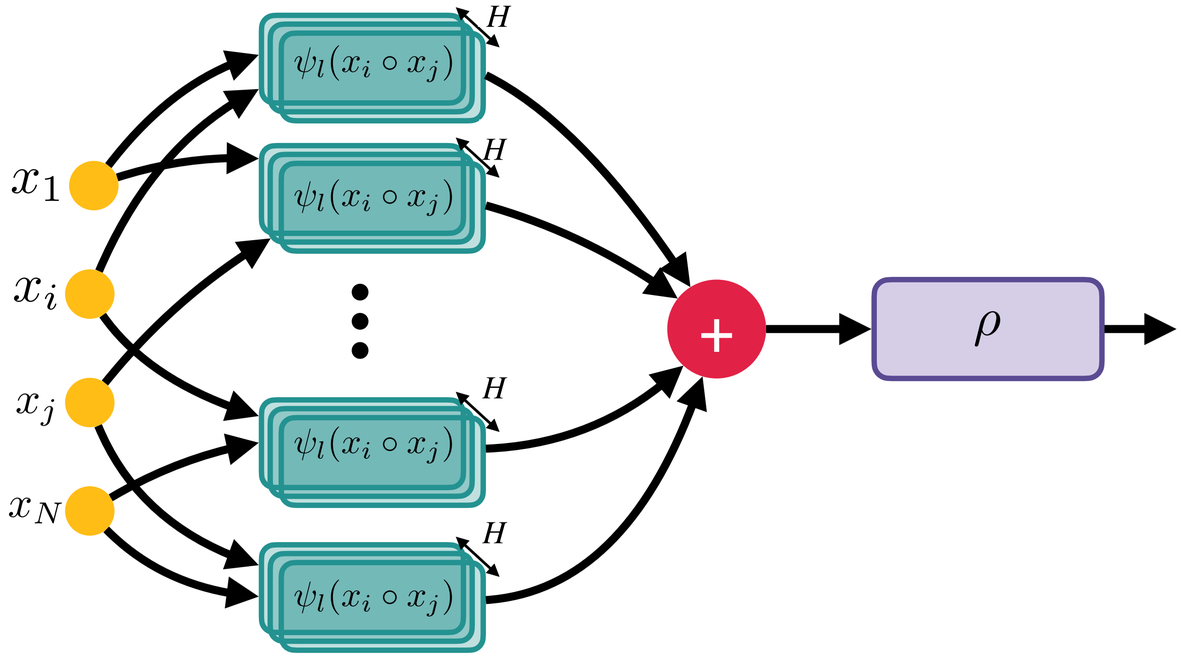
  $\mathrm{Sym}(H, N, d)$ can approximate any smooth symmetric function.

H^*: size of the basis of multisymmetric powersum polynomials.

Self-Attention Mechanism

[Badhanu et al., Varswani et al]

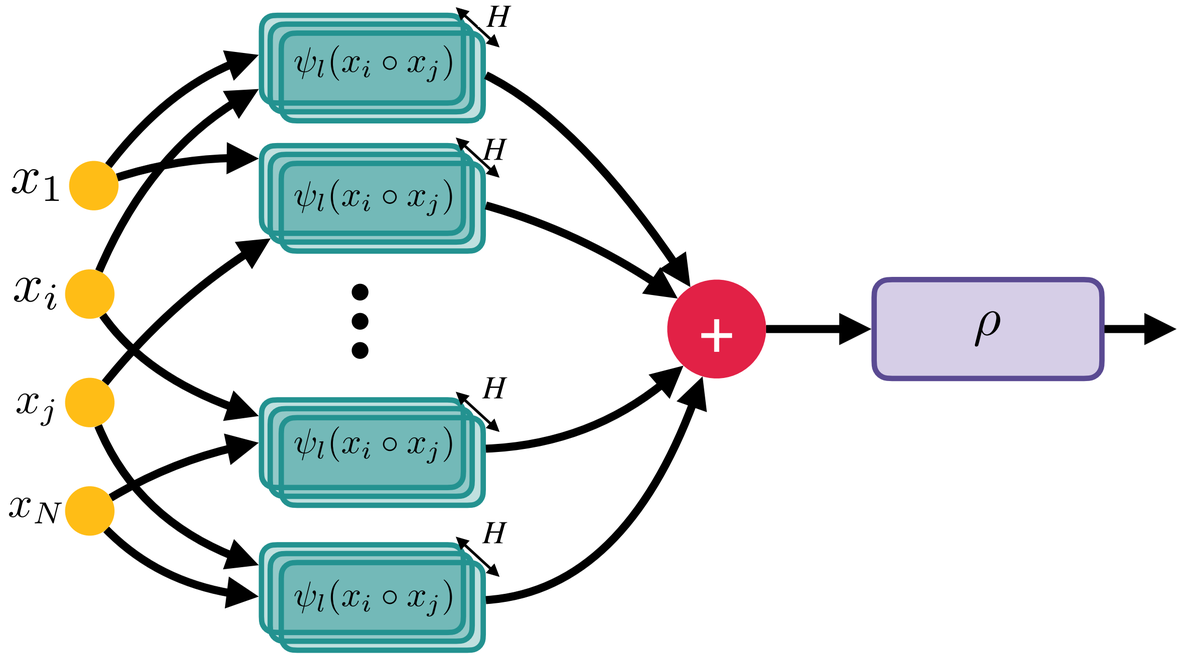
- Pairwise symmetric representation: Let $\mathrm{Sym}^2(H, N, d)$ be the class of (symmetric) functions of the form $f(X) = \rho(\phi_1(X), \ldots, \phi_H(X))$, with $\phi_j(X) = \sum_{i,i'=1}^{N} \psi_j(x_i, x_{i'})$, $j = 1, \ldots H$.
- Underlies the *Transformer* architecture.



Self-Attention Mechanism

[Badhanu et al., Varswani et al]

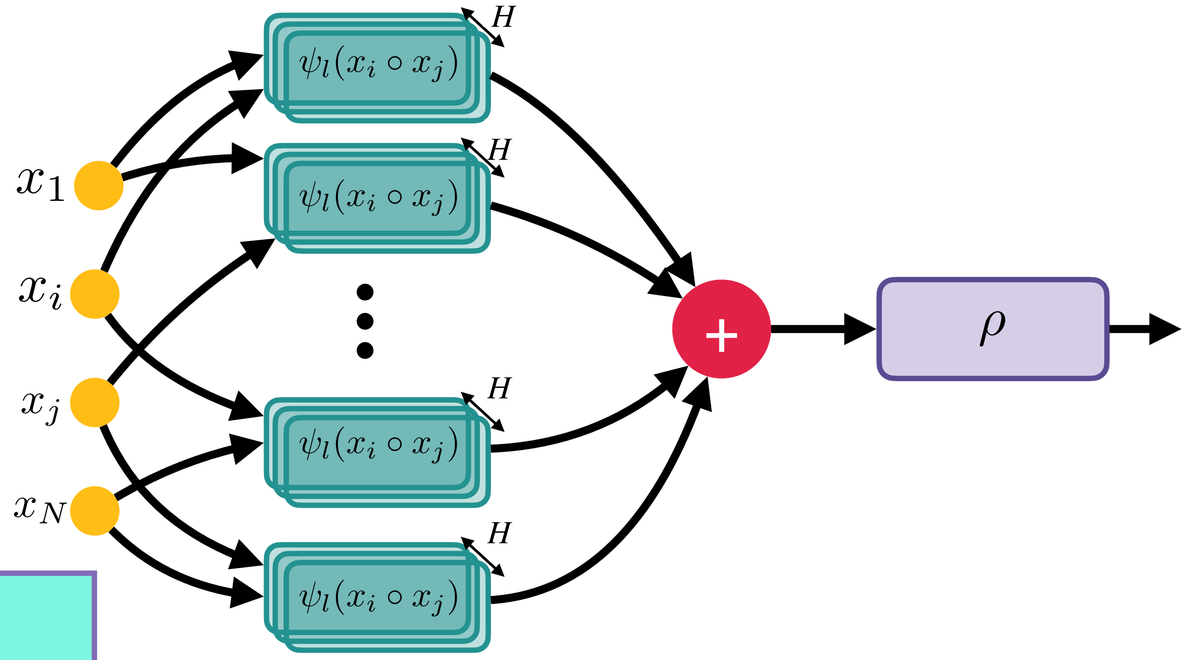
- Pairwise symmetric representation: Let $\mathrm{Sym}^2(H, N, d)$ be the class of (symmetric) functions of the form $f(X) = \rho(\phi_1(X), \ldots, \phi_H(X))$, with $\phi_j(X) = \sum_{i,i'=1}^{N} \psi_j(x_i, x_{i'})$, $j = 1, \ldots H$.
- Underlies the *Transformer* architecture.
- Clearly, $\mathrm{Sym}(H, N, d) \subseteq \mathrm{Sym}^2(H, N, d)$.



Self-Attention Mechanism

[Badhanu et al., Varswani et al]

- Pairwise symmetric representation: Let $\mathrm{Sym}^2(H, N, d)$ be the class of (symmetric) functions of the form $f(X) = \rho(\phi_1(X), \ldots, \phi_H(X))$, with $\phi_j(X) = \sum_{i,i'=1}^{N} \psi_j(x_i, x_{i'})$, $j = 1, \ldots H$.
- Underlies the *Transformer* architecture.
- Clearly, $\mathrm{Sym}(H, N, d) \subseteq \mathrm{Sym}^2(H, N, d)$.



How to quantify this gain?

Main Result

- Given a data distribution $\nu \in \mathscr{P}((\mathbb{C}^d)^{\otimes N})$, define L^2 metric $\langle f, g\rangle_\nu := \mathbb{E}_{X\sim\nu}[f(X)g^*(X)]$.
- Approximation lower bound for $\mathsf{Sym}(H, N, d)$ when $H < H^*$:

Theorem [ZB'22]: There exists a smooth distribution ν in $(\mathbb{C}^d)^{\otimes N}$ and g of unit norm $\|g\|_\nu = 1$ such that $\min_{f\in\mathsf{Sym}(H,N,d)} \|g - f\|_\nu^2 \geq \frac{1}{6}\left(1 - \frac{H}{2^{\min(n/2,d-1)}}\right)$.

Main Result

- Given a data distribution $\nu \in \mathscr{P}((\mathbb{C}^d)^{\otimes N})$, define L^2 metric $\langle f, g\rangle_\nu := \mathbb{E}_{X\sim\nu}[f(X)g^*(X)]$.
- Approximation lower bound for $\mathsf{Sym}(H, N, d)$ when $H < H^*$:

Theorem [ZB'22]: There exists a smooth distribution ν in $(\mathbb{C}^d)^{\otimes N}$ and g of unit norm $\|g\|_\nu = 1$ such that $\min_{f\in\mathsf{Sym}(H,N,d)} \|g - f\|_\nu^2 \geq \frac{1}{6}\left(1 - \frac{H}{2^{\min(n/2,d-1)}}\right)$.

- Efficient approximation using pairwise symmetric model:

Theorem [ZB'22]: For any $\epsilon > 0$, there exists $f \in \mathsf{Sym}^2(1,N, d)$ parametrised by $O(\mathrm{poly}(N, d, \epsilon^{-1}))$ neurons with weights bounded by $O(\mathrm{poly}(N, d, \epsilon^{-1}))$ such that $\|g - f\|_\infty \leq \epsilon$.

Main Result

- Given a data distribution $\nu \in \mathscr{P}((\mathbb{C}^d)^{\otimes N})$, define L^2 metric $\langle f, g\rangle_\nu := \mathbb{E}_{X\sim\nu}[f(X)g^*(X)]$.
- Approximation lower bound for $\mathsf{Sym}(H, N, d)$ when $H < H^*$:

Theorem [ZB'22]: There exists a smooth distribution ν in $(\mathbb{C}^d)^{\otimes N}$ and g of unit norm $\|g\|_\nu = 1$ such that $\min_{f\in\mathsf{Sym}(H,N,d)} \|g - f\|_\nu^2 \geq \frac{1}{6}\left(1 - \frac{H}{2^{\min(n/2,d-1)}}\right)$.

- Efficient approximation using pairwise symmetric model:

Theorem [ZB'22]: For any $\epsilon > 0$, there exists $f \in \mathsf{Sym}^2(1,N, d)$ parametrised by $O(\mathrm{poly}(N, d, \epsilon^{-1}))$ neurons with weights bounded by $O(\mathrm{poly}(N, d, \epsilon^{-1}))$ such that $\|g - f\|_\infty \leq \epsilon$.

  - *Exponential separation* between Sym and Sym^2. Arbitrary depth in Sym .
  - Function g is an appropriate symmetric polynomial of large 'rank'.

Proof Sketch

- Powersums $p_k(X) = k^{-1/2} \sum_{i=1}^{N} x_i^k$ generate an algebra of symmetric polynomials

  $p_\lambda(X) = \prod_j p_{\lambda_j}(X)$ for partition $\lambda = \{\lambda_1, \ldots, \lambda_S\}$.

- We can define measure ν where products of powersums are *orthogonal* wrt $\langle f, g \rangle_\nu$.

Proof Sketch

- Powersums $p_k(X) = k^{-1/2} \sum_{i=1}^{N} x_i^k$ generate an algebra of symmetric polynomials

  $p_\lambda(X) = \prod_j p_{\lambda_j}(X)$ for partition $\lambda = \{\lambda_1, \ldots, \lambda_S\}$.

- We can define measure ν where products of powersums are *orthogonal* wrt $\langle f, g \rangle_\nu$.
- Approximation error $\|f - g\|_\nu$ lower bounded by $\|\mathscr{P}_2(f - g)\|_\nu$, where $\mathscr{P}_2$ is the projection onto products of two powersums, for certain class of multisymmetric polynomials g.

Proof Sketch

- Powersums $p_k(X) = k^{-1/2} \sum_{i=1}^{N} x_i^k$ generate an algebra of symmetric polynomials

  $p_\lambda(X) = \prod_j p_{\lambda_j}(X)$ for partition $\lambda = \{\lambda_1, \ldots, \lambda_S\}$.

- We can define measure ν where products of powersums are *orthogonal* wrt $\langle f, g \rangle_\nu$.
- Approximation error $\|f - g\|_\nu$ lower bounded by $\|\mathscr{P}_2(f - g)\|_\nu$, where $\mathscr{P}_2$ is the projection onto products of two powersums, for certain class of multisymmetric polynomials g.
- When $f \in \mathsf{Sym}(H, N, d)$, we have $\|\mathscr{P}_2(f - g)\| \gtrsim \|F - G\|_F$ where F,G are matrix representations of f, g, and F has rank at most H.

Proof Sketch

- Powersums $p_k(X) = k^{-1/2} \sum_{i=1}^{N} x_i^k$ generate an algebra of symmetric polynomials

  $p_\lambda(X) = \prod_j p_{\lambda_j}(X)$ for partition $\lambda = \{\lambda_1, \ldots, \lambda_S\}$.

- We can define measure ν where products of powersums are *orthogonal* wrt $\langle f, g \rangle_\nu$.
- Approximation error $\|f - g\|_\nu$ lower bounded by $\|\mathscr{P}_2(f - g)\|_\nu$, where $\mathscr{P}_2$ is the projection onto products of two powersums, for certain class of multisymmetric polynomials g.
- When $f \in \mathsf{Sym}(H, N, d)$, we have $\|\mathscr{P}_2(f - g)\| \gtrsim \|F - G\|_F$ where F,G are matrix representations of f, g, and F has rank at most H.
- Target g has associated matrix representation G of rank $\exp(\Theta(\min(N, d))$.

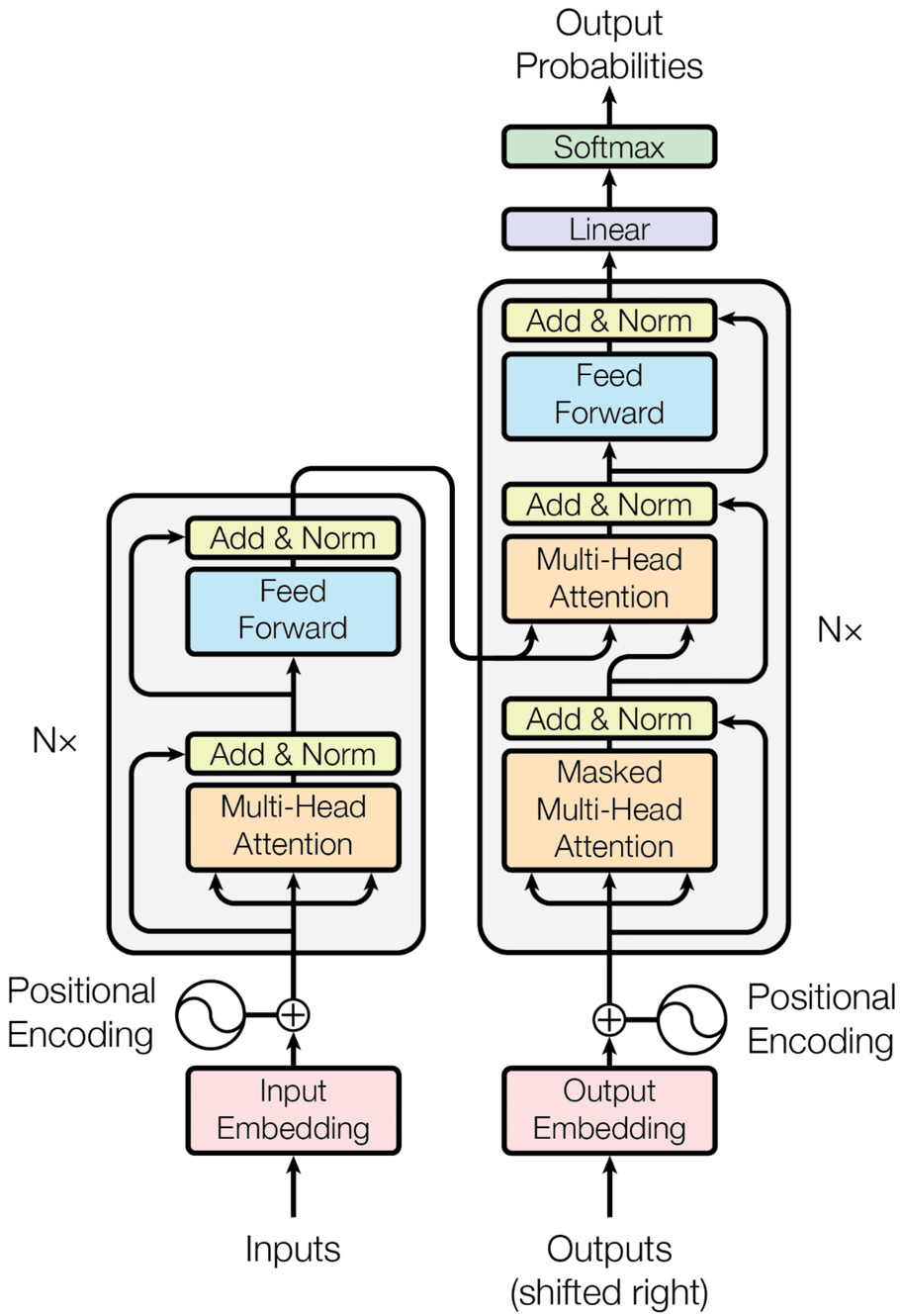
In Summary

- Exponential separation between unary (*DeepSet*) and pairwise (*Transformer*) symmetric representations.
- Leveraging algebraic structure of symmetric polynomials for analytic functions.
- **Open**: High-order extensions, e.g. [Sanford et al]? Inherent advantage of pairwise interactions model (standard model)?
- **Open**: non analytic-activations?

Transformers

[Varswani et al'18]

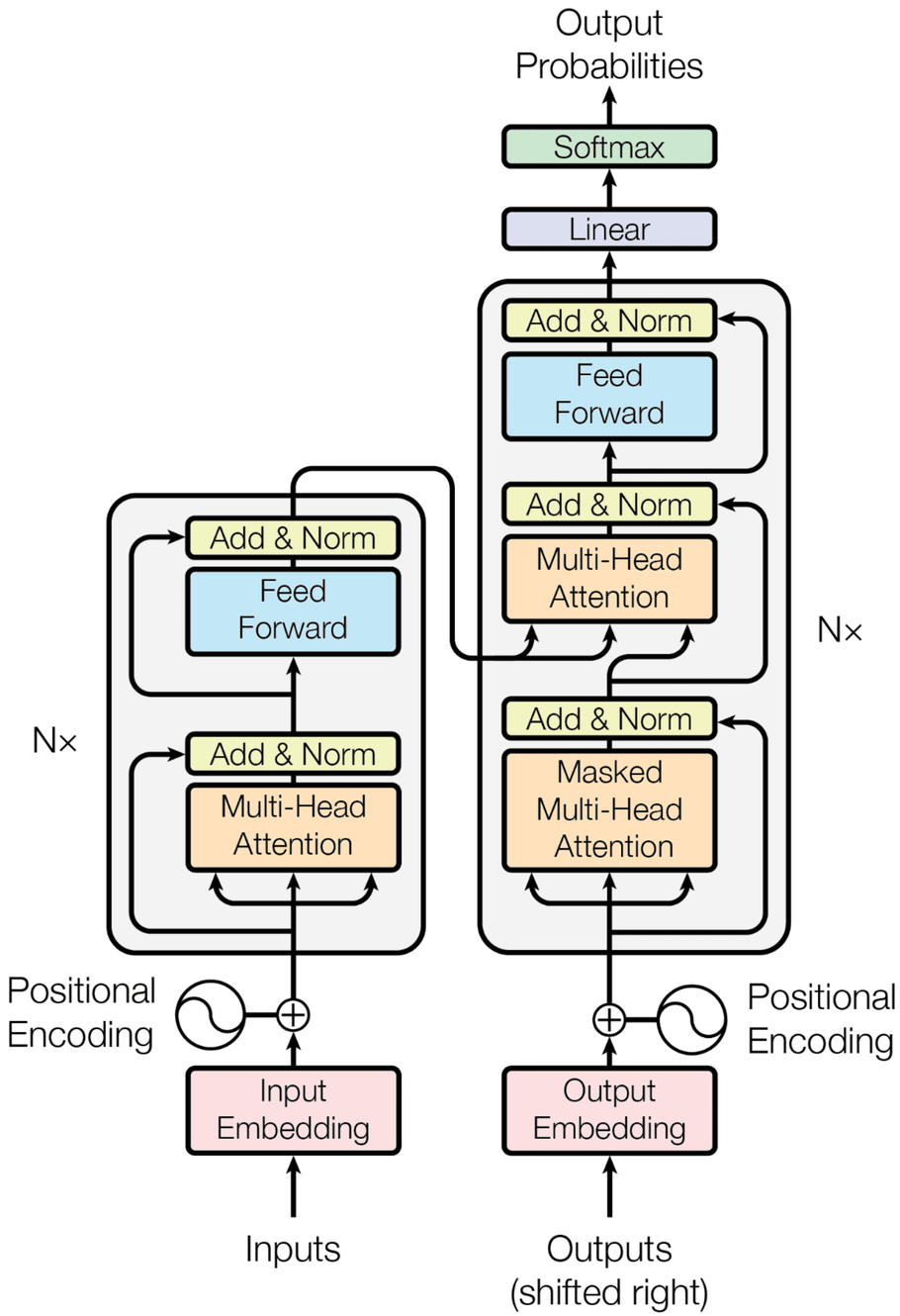
- Powers most modern large-scale NNs.
- Enables 'soft' invariance via positional encodings.



Transformers

[Varswani et al'18]

- Powers most modern large-scale NNs.
- Enables 'soft' invariance via positional encodings.
- Given inputs $X \in \mathbb{R}^{N \times d}$, each layer computes $F(X) = \Phi_\theta\Big(\sum_{h \le H} S_h X V_h\Big)$, with
  - $S_h = \sigma(X K_h Q_h^\top X^\top)$ self-attention kernel, $K_h, Q_h \in \mathbb{R}^{d \times r}$,
  - $V_h \in \mathbb{R}^{d \times d}$: 'value' feature map,
  - Φ_θ : Element-wise MLP transformation.
- Two main hyperparameters: width H and rank r.

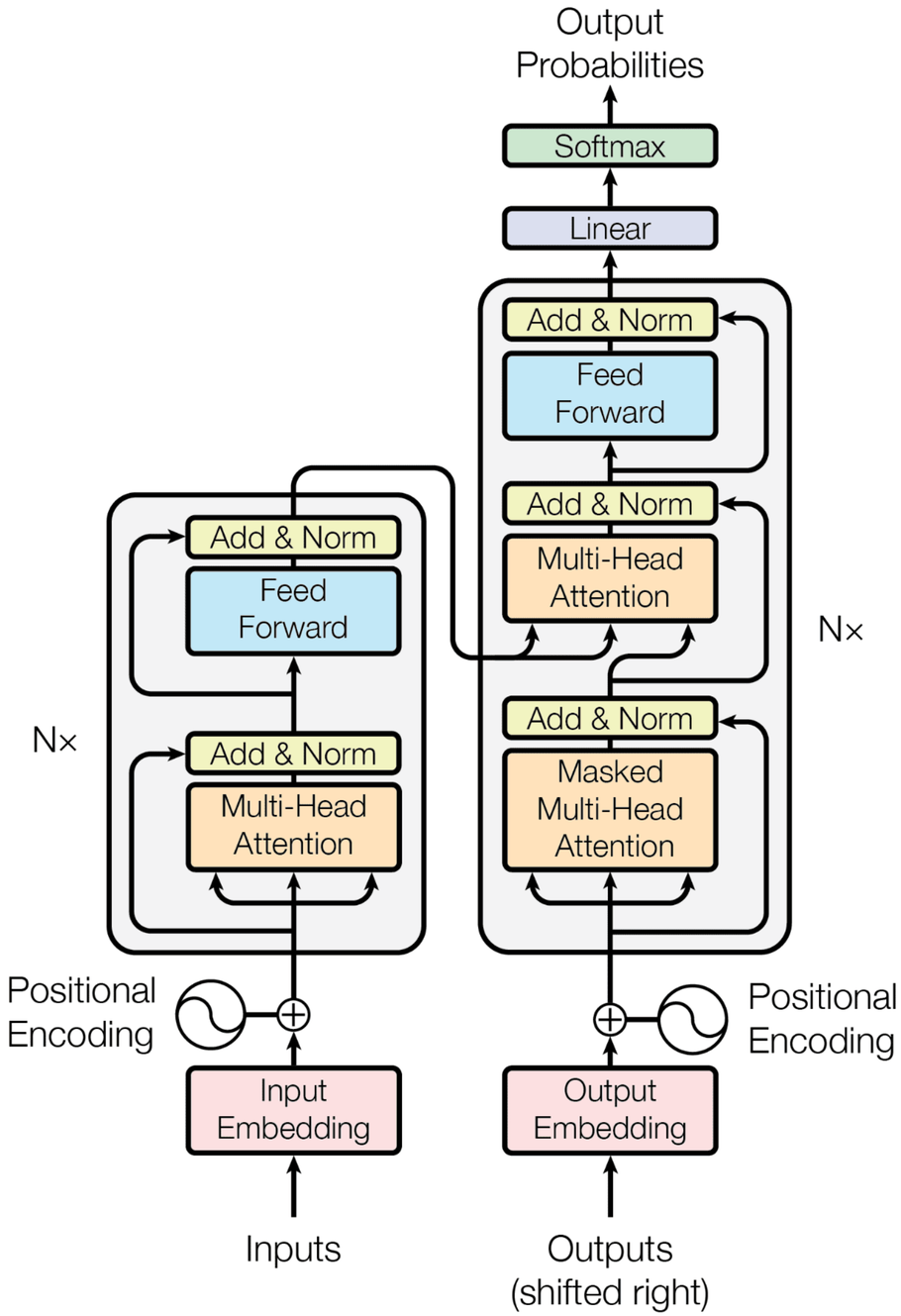


Transformers

[Varswani et al'18]

- Powers most modern large-scale NNs.
- Enables 'soft' invariance via positional encodings.
- Given inputs $X \in \mathbb{R}^{N \times d}$, each layer computes $F(X) = \Phi_\theta\Big(\sum_{h \le H} S_h X V_h\Big)$, with
  - $S_h = \sigma(X K_h Q_h^\top X^\top)$ self-attention kernel, $K_h, Q_h \in \mathbb{R}^{d \times r}$,
  - $V_h \in \mathbb{R}^{d \times d}$: 'value' feature map,
  - Φ_θ : Element-wise MLP transformation.
- Two main hyperparameters: width H and rank r.

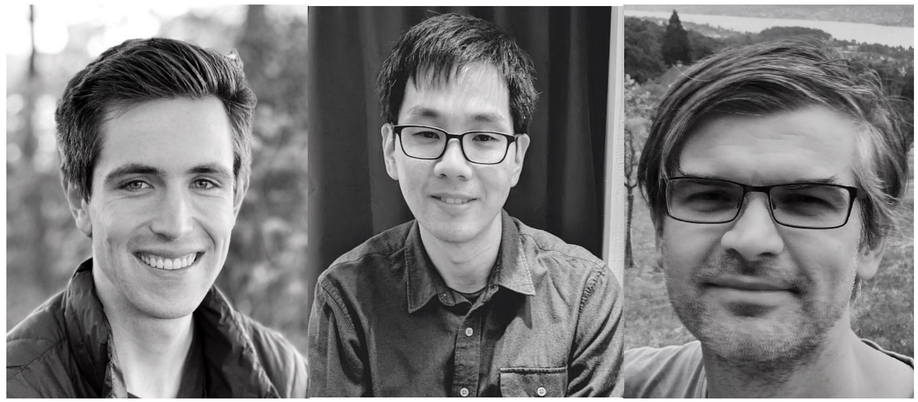
What do they capture?



Known Separations in Transformers

[Sanford, Hsu, Telgarsky, '24]

- Focus on the multi-head self-attention (MHSA) layer (no MLP): $F(X) = \sum_{h \le H} S_h X V_h$.
- There exists an equivariant function $f : \mathcal{X}^{\otimes N} \to \{0,1\}^{\otimes N}$ such that one layer of multi-head attention with p bits of precision cannot express it unless $prH \gtrsim N$.



Known Separations in Transformers

[Sanford, Hsu, Telgarsky, '24]

- Focus on the multi-head self-attention (MHSA) layer (no MLP): $F(X) = \sum_{h \le H} S_h X V_h$.
- There exists an equivariant function $f : \mathcal{X}^{\otimes N} \to \{0,1\}^{\otimes N}$ such that one layer of multi-head attention with p bits of precision cannot express it unless $prH \gtrsim N$.
  - Target function is a 'third-order' self-attention function.
  - Lower bound is based on the framework of communication complexity.

Known Separations in Transformers

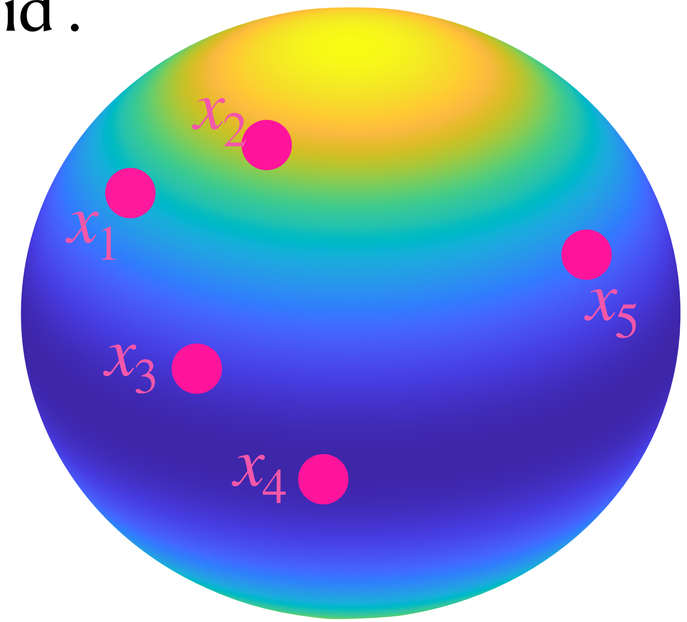
[Sanford, Hsu, Telgarsky, '24]

- Focus on the multi-head self-attention (MHSA) layer (no MLP): $F(X) = \sum_{h \le H} S_h X V_h$.
- There exists an equivariant function $f : \mathcal{X}^{\otimes N} \to \{0,1\}^{\otimes N}$ such that one layer of multi-head attention with p bits of precision cannot express it unless $prH \gtrsim N$.
  - Target function is a 'third-order' self-attention function.
  - Lower bound is based on the framework of communication complexity.

Can we disentangle role of rank and width?
Beyond fixed precision?
Quantitative approximation rates?

Metric Functions on the Sphere

- Consider inputs in d-dimensional unit-sphere $\mathcal{X} = \mathcal{S}_{d-1}$, and approximation in $L^2(\mathcal{S}_{d-1}^{\otimes N}, \tau_{d-1}^{\otimes N})$:
  $\|f - g\| := \mathbb{E}_{x_1,\ldots,x_N}\|f(x) - g(x)\|^2$, where $x_i \sim \text{Unif}(\mathcal{S}_{d-1})$ iid .

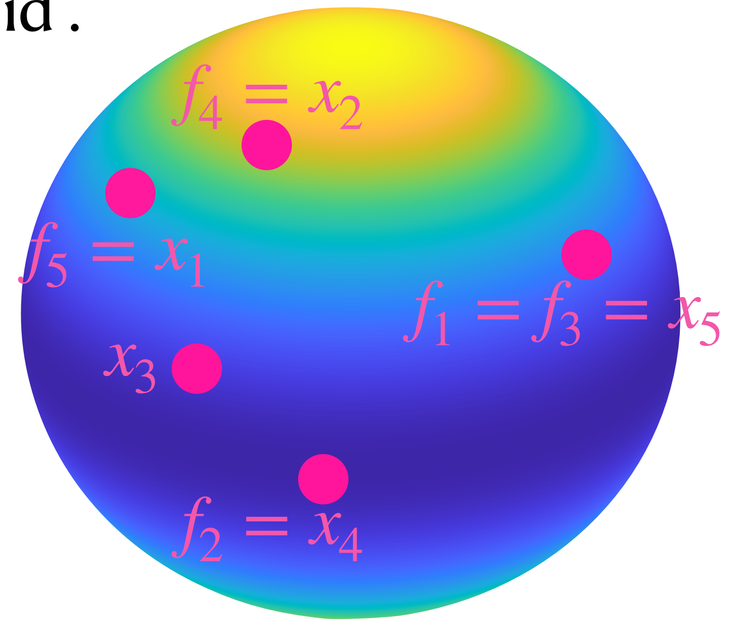


τ_{d-1} : Uniform measure on $\mathcal{S}_{d-1}$.

Metric Functions on the Sphere

- Consider inputs in d-dimensional unit-sphere $\mathcal{X} = \mathcal{S}_{d-1}$, and approximation in $L^2(\mathcal{S}_{d-1}^{\otimes N}, \tau_{d-1}^{\otimes N})$:
  $\|f - g\| := \mathbb{E}_{x_1,\ldots,x_N}\|f(x) - g(x)\|^2$, where $x_i \sim \text{Unif}(\mathcal{S}_{d-1})$ iid .

- **Furthest-neighbor** function: $f : \mathcal{S}_{d-1}^{\otimes N} \to \mathcal{S}_{d-1}^{\otimes N}$ defined as
  $f_j(x_1, \ldots, x_N) = \arg\max_i \|x_i - x_j\| = \arg\min_i x_i \cdot x_j$.



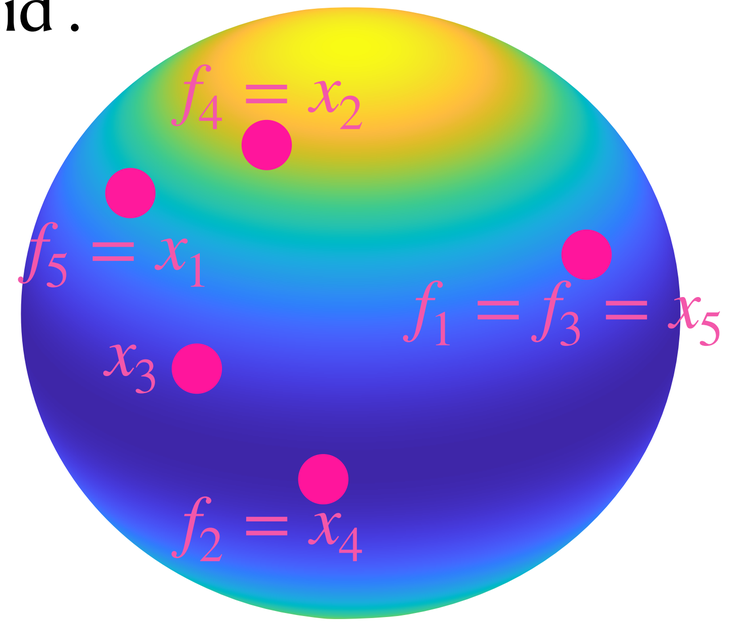
τ_{d-1} : Uniform measure on $\mathcal{S}_{d-1}$.

Metric Functions on the Sphere

- Consider inputs in d-dimensional unit-sphere $\mathcal{X} = \mathcal{S}_{d-1}$, and approximation in $L^2(\mathcal{S}_{d-1}^{\otimes N}, \tau_{d-1}^{\otimes N})$:
  $\|f - g\| := \mathbb{E}_{x_1,\ldots,x_N}\|f(x) - g(x)\|^2$, where $x_i \sim \text{Unif}(\mathcal{S}_{d-1})$ iid .

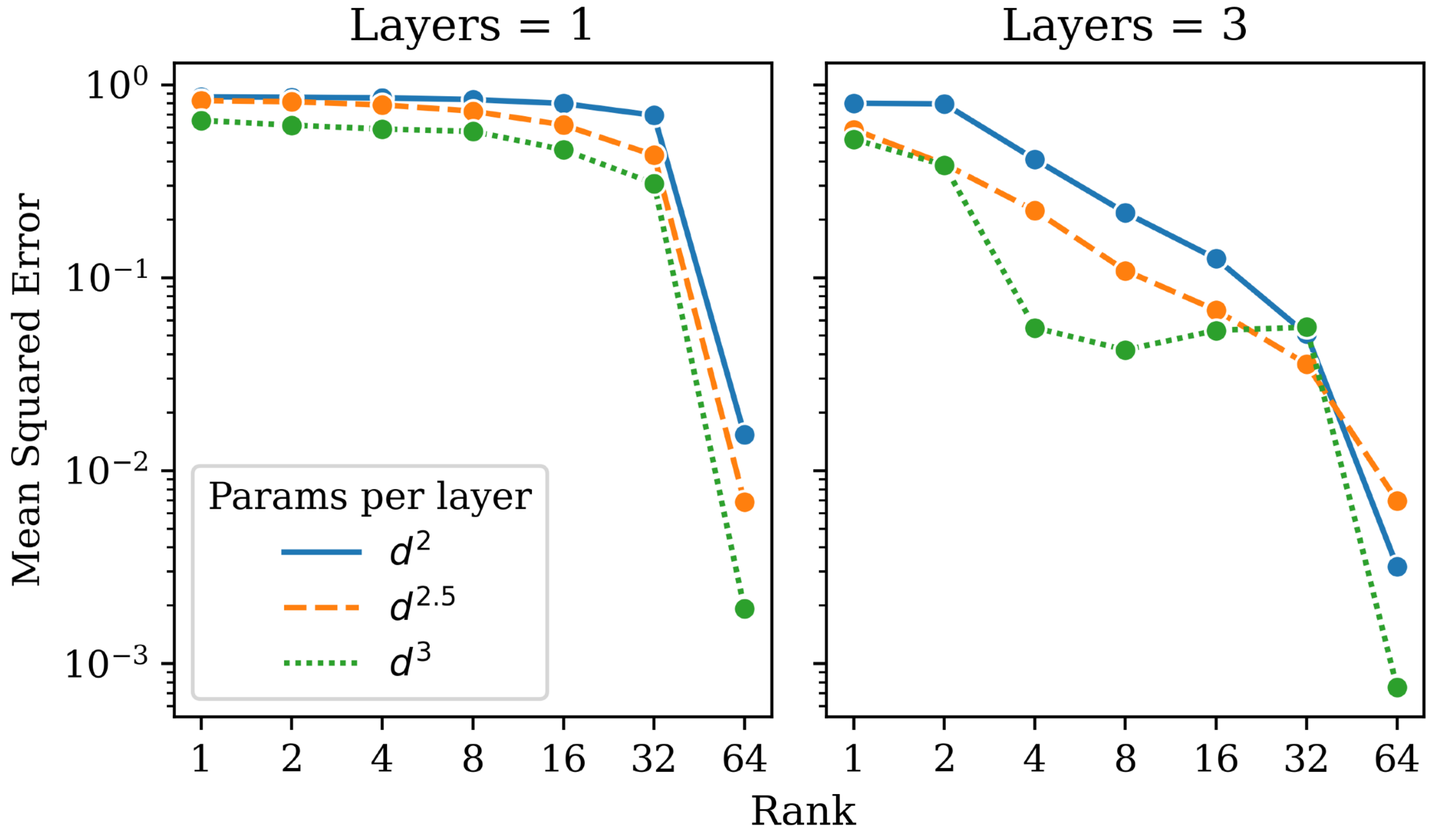
- **Furthest-neighbor** function: $f : \mathcal{S}_{d-1}^{\otimes N} \to \mathcal{S}_{d-1}^{\otimes N}$ defined as
  $f_j(x_1, \ldots, x_N) = \arg\max_i \|x_i - x_j\| = \arg\min_i x_i \cdot x_j$.

Approximation with Self-Attention?



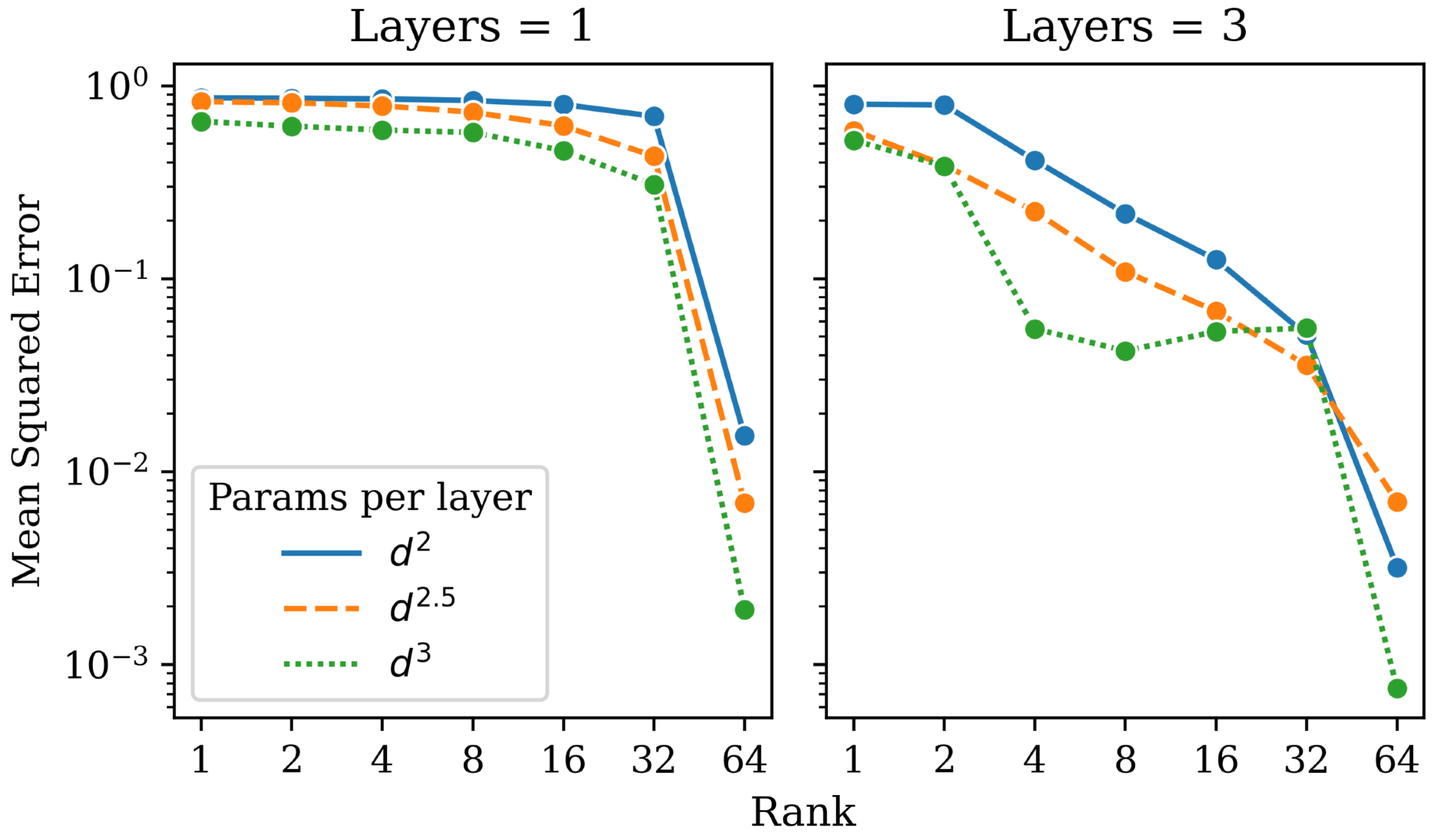
τ_{d-1} : Uniform measure on $\mathcal{S}_{d-1}$.

Metric Functions on the Sphere



- Off-the-shelf transformers, $d = 64, N = 16$. Each line has constant rH.
- Low-rank transformers don't learn.

Metric Functions on the Sphere



- Off-the-shelf transformers, $d = 64, N = 16$. Each line has constant rH.
- Low-rank transformers don't learn. Why?

Full Rank Efficient Approximation

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

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

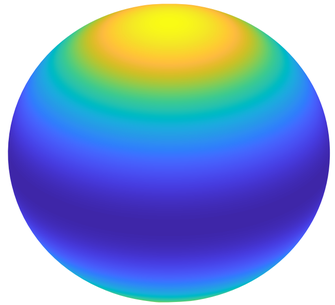
$$= \operatorname*{argmin}_{\mathbf{x} \in \{\mathbf{x}_1, \ldots, \mathbf{x}_N\}} \mathbf{x}^\top \mathbf{y}$$

Full Rank Efficient Approximation

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

$$
\begin{aligned}
f(\mathbf{x}_1, \ldots, \mathbf{x}_N; \mathbf{y}) &:= \operatorname*{argmax}_{\mathbf{x} \in \{\mathbf{x}_1, \ldots, \mathbf{x}_N\}} \|\mathbf{x} - \mathbf{y}\|_2 \\
&= \operatorname*{argmin}_{\mathbf{x} \in \{\mathbf{x}_1, \ldots, \mathbf{x}_N\}} \mathbf{x}^\top \mathbf{y} \\
&= \mathbf{X}\ \mathrm{hm}(-\mathbf{X}^\top \mathbf{y}) \\
&= \mathbf{I}_d\ \mathbf{X}\ \mathrm{sm}\left(\mathbf{X}^\top \left(-10^{10} \cdot \mathbf{I}_d\right) \mathbf{y}\right)
\end{aligned}
$$

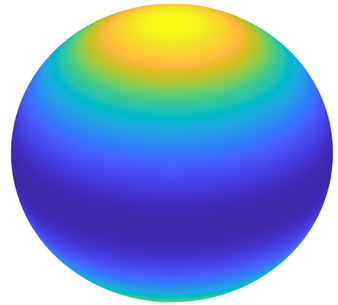
Our Rank Separation Result



τ_{d-1} : uniform measure of $\mathcal{S}_{d-1}$.

- **Theorem** [ABY'24]: Fix $N = 3$, and let f be the furthest-neighbor function. Then:
  1. f can be expressed with a single full-rank head,
  2. For each $\epsilon > 0$, if $H \lesssim (d/r)^{1/\epsilon}$, then any MHSA layer g incurs a (relative) error $\|f - g\|^2 \geq \epsilon$.

Our Rank Separation Result



τ_{d-1} : uniform measure of $\mathcal{S}_{d-1}$.

- **Theorem** [ABY'24]: Fix $N = 3$, and let f be the furthest-neighbor function. Then:
  1. f can be expressed with a single full-rank head,
  2. For each $\epsilon > 0$, if $H \lesssim (d/r)^{1/\epsilon}$, then any MHSA layer g incurs a (relative) error $\|f - g\|^2 \geq \epsilon$.
- **Consequence**: Low-Rank separation: if $r = O_d(1)$, then $rH \gtrsim d^{1/\epsilon}$ parameters are needed , while if $r = \Theta(d)$, then $rH \simeq d$ suffice.
- In the high-precision regime $\epsilon \to 0$, approximation rate becomes exponential, $H \gtrsim 2^{d - r\log(d/r)}$.

Reduction to half-space approximation

- Projecting $f(x_1, x_2, y) - g(x_1, x_2, y)$ onto $z := \dfrac{x_1 - x_2}{\|x_1 - x_2\|}$ roughly leads to

$$\mathbb{E}_{x_1,x_2,y}\|f(x_1, x_2, y) - g(x_1, x_2, y)\|^2 \gtrsim \frac{1}{2}\mathbb{E}_{z,y}\left|\operatorname{sign}(z^\top y) - \sum_{h \le H} g_h(z, y)\right|^2,$$

  where $g_h(z, y) = \phi_h(K_h^\top z, y)$ only depends on a rank-r projection of z.

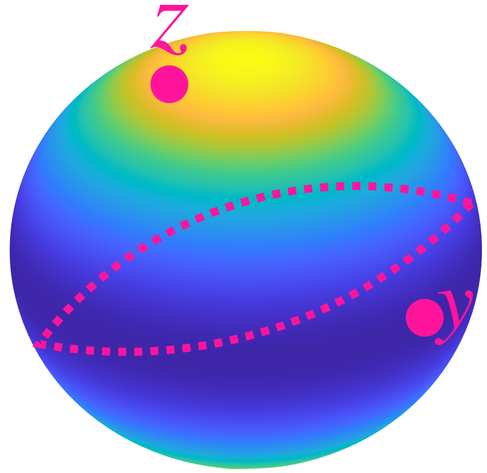
Reduction to half-space approximation

- Projecting $f(x_1, x_2, y) - g(x_1, x_2, y)$ onto $z := \dfrac{x_1 - x_2}{\|x_1 - x_2\|}$ roughly leads to

$$\mathbb{E}_{x_1,x_2,y}\|f(x_1, x_2, y) - g(x_1, x_2, y)\|^2 \gtrsim \frac{1}{2}\mathbb{E}_{z,y}\left|\operatorname{sign}(z^\top y) - \sum_{h \le H} g_h(z, y)\right|^2,$$

  where $g_h(z, y) = \phi_h(K_h^\top z, y)$ only depends on a rank-r projection of z.

- Thus, question is reduced to approximating indicator function $I : (z, y) \mapsto \operatorname{sign}(z^\top y)$ in $L^2(\mathcal{S}_{d-1}^{\otimes 2}; \tau_{d-1}^{\otimes 2})$ using low-rank neurons.



Reduction to half-space approximation

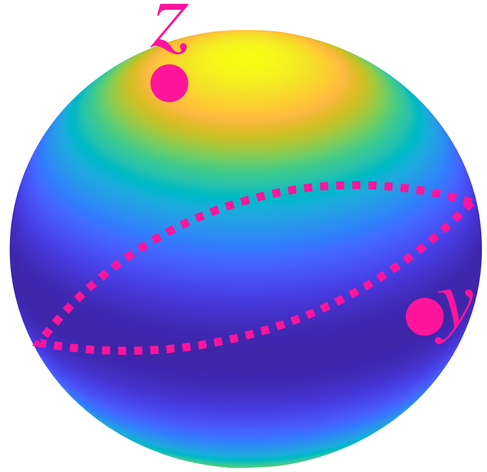
- Projecting $f(x_1, x_2, y) - g(x_1, x_2, y)$ onto $z := \dfrac{x_1 - x_2}{\|x_1 - x_2\|}$ roughly leads to

$$\mathbb{E}_{x_1,x_2,y}\|f(x_1, x_2, y) - g(x_1, x_2, y)\|^2 \gtrsim \frac{1}{2}\mathbb{E}_{z,y}\left|\operatorname{sign}(z^\top y) - \sum_{h \le H} g_h(z, y)\right|^2,$$

  where $g_h(z, y) = \phi_h(K_h^\top z, y)$ only depends on a rank-r projection of z.

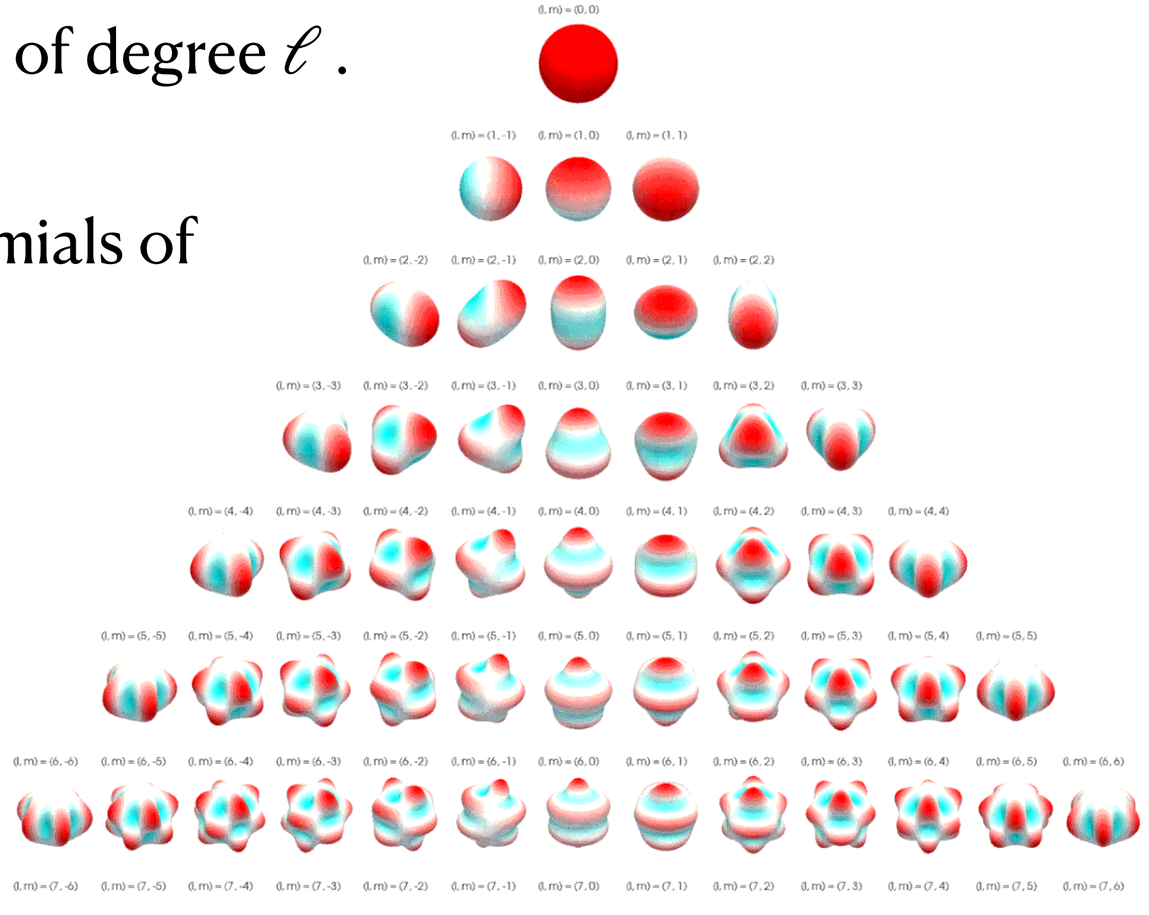
- Thus, question is reduced to approximating indicator function $I : (z, y) \mapsto \operatorname{sign}(z^\top y)$ in $L^2(\mathcal{S}_{d-1}^{\otimes 2}; \tau_{d-1}^{\otimes 2})$ using low-rank neurons.

- **Key idea:** find orthonormal basis of $L^2(\mathcal{S}_{d-1}^{\otimes 2}; \tau_{d-1}^{\otimes 2})$ such that:
  - Target function has its energy *spread* across basis elements,
  - Approximant has its energy *concentrated* on a few elements.



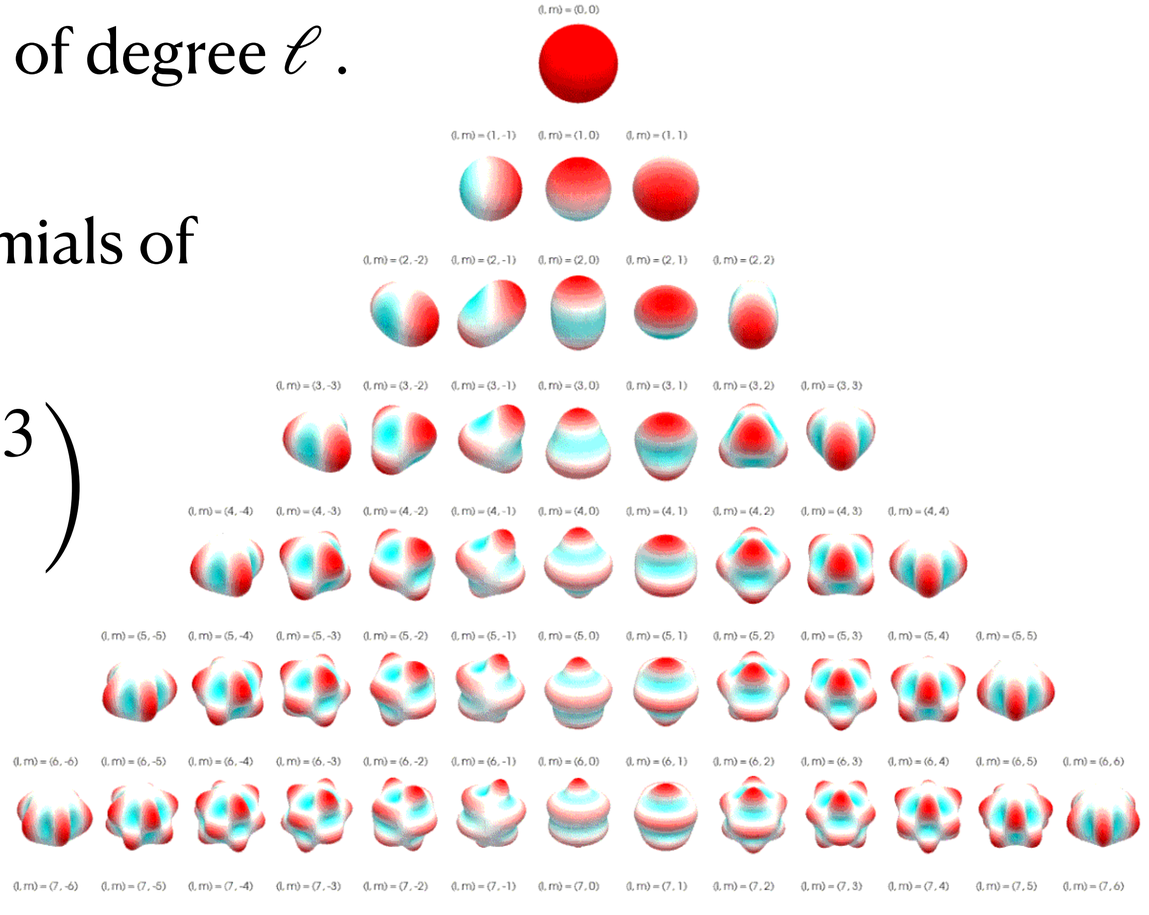
Spherical Harmonics

- Canonical harmonic expansion for $f \in L^2(\mathcal{S}_{d-1}, \tau_{d-1})$:
  $f = \sum_{\ell \geq 0} \mathsf{P}_{V_\ell} f$, with $V_\ell =$ Spherical Harmonics of degree ℓ .
- V_ℓ spanned by *homogeneous*, *harmonic* polynomials of degree ℓ in d variables.



Spherical Harmonics

- Canonical harmonic expansion for $f \in L^2(\mathcal{S}_{d-1}, \tau_{d-1})$:
  $f = \sum_{\ell \geq 0} \mathsf{P}_{V_\ell} f$, with $V_\ell =$ Spherical Harmonics of degree ℓ .
- V_ℓ spanned by *homogeneous, harmonic* polynomials of degree ℓ in d variables.
- $\dim(V_\ell) := N(d, \ell) = \dfrac{2\ell + d - 2}{\ell} \begin{pmatrix} \ell + d - 3 \\ \ell - 1 \end{pmatrix}$



Hecke-Funk Representation

- Expansion of rank-1 functions given by the **Hecke-Funk representation** formula:

If $\phi_\theta(x) = \phi(x^\top \theta)$, and Y_ℓ is a spherical harmonic of degree ℓ, then $\langle \phi_\theta, Y_\ell \rangle_{\tau_d} = Y_\ell(\theta) \langle \phi, P_{\ell,d} \rangle_{u_d}$, where

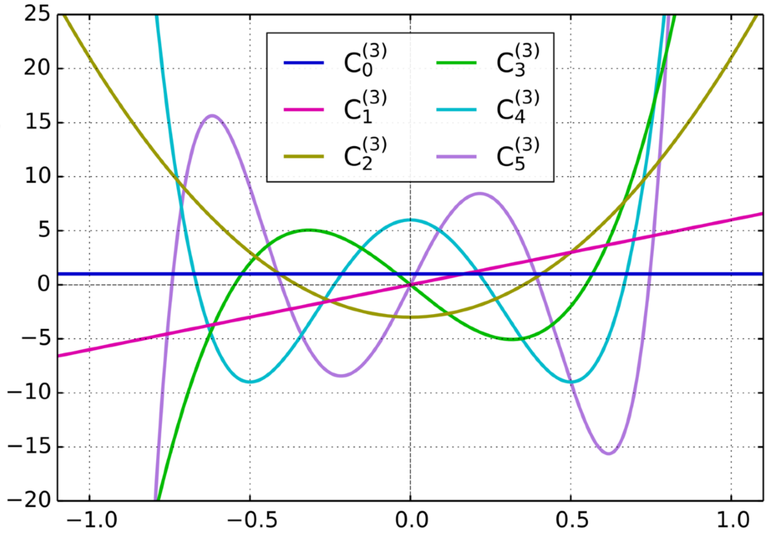
  - $P_{\ell,d}(t)$: Gegenbauer Polynomial of degree ℓ,
  - $u_d = (\theta^\top \cdot)_\# \tau_d \propto (1-t^2)^{(d-2)/3}$ marginal distribution of τ_d.

Hecke-Funk Representation

- Expansion of rank-1 functions given by the **Hecke-Funk representation** formula:

If $\phi_\theta(x) = \phi(x^\top \theta)$, and Y_ℓ is a spherical harmonic of degree ℓ, then $\langle \phi_\theta, Y_\ell \rangle_{\tau_d} = Y_\ell(\theta) \langle \phi, P_{\ell,d} \rangle_{u_d}$, where

  - $P_{\ell,d}(t)$: Gegenbauer Polynomial of degree ℓ,
  - $u_d = (\theta^\top \cdot)_\# \tau_d \propto (1-t^2)^{(d-2)/3}$ marginal distribution of τ_d.
  - $\{P_{\ell,d}\}_\ell$ forms an orthonormal basis of $L^2([-1,1], u_d)$.



Spectral decomposition of the target

- Given *any* basis of spherical harmonics $\{Y_{\ell,k}\}_{\ell\geq 0, k\leq N(d,\ell)}$, consider the *tensorised* basis $\{Y_{\ell,k} \otimes Y_{\ell',k'}\}_{\ell,\ell',k,k'}$ of $L^2(\mathcal{S}^{\otimes 2}_{d-1})$.

Spectral decomposition of the target

- Given *any* basis of spherical harmonics $\{Y_{\ell,k}\}_{\ell\geq 0, k\leq N(d,\ell)}$, consider the *tensorised* basis $\{Y_{\ell,k} \otimes Y_{\ell',k'}\}_{\ell,\ell',k,k'}$ of $L^2(\mathcal{S}^{\otimes 2}_{d-1})$.
- By H-F, $\langle I, Y_{\ell,k} \otimes Y_{\ell',k'}\rangle = \int \text{sign}(z^\top y) Y_{\ell,k}(z) Y_{\ell',k'}(y) d\tau_d(z) d\tau_d(y)$

$$= \langle \text{sign}, P_{\ell'}\rangle_{u_d} \int Y_{\ell,k}(z) Y_{\ell',k'}(z) d\tau_d(z) = \langle \text{sign}, P_{\ell'}\rangle_{u_d} N(d,\ell)^{-1/2} \delta_{\ell=\ell',k=k'}$$

Spectral decomposition of the target

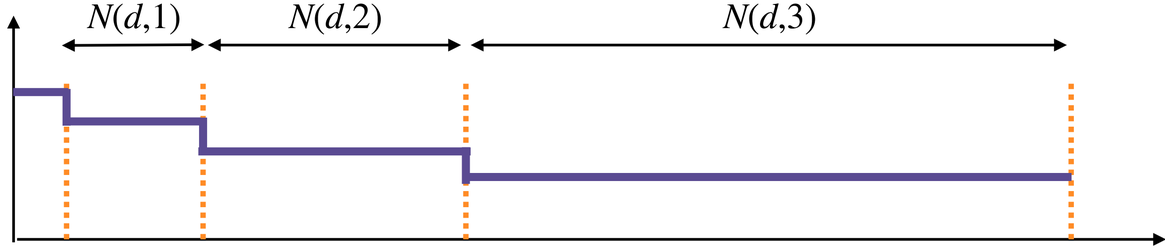
- Given *any* basis of spherical harmonics $\{Y_{\ell,k}\}_{\ell\geq 0, k\leq N(d,\ell)}$, consider the *tensorised* basis $\{Y_{\ell,k} \otimes Y_{\ell',k'}\}_{\ell,\ell',k,k'}$ of $L^2(\mathcal{S}^{\otimes 2}_{d-1})$.
- By H-F, $\langle I, Y_{\ell,k} \otimes Y_{\ell',k'}\rangle = \int \text{sign}(z^\top y) Y_{\ell,k}(z) Y_{\ell',k'}(y) d\tau_d(z) d\tau_d(y)$

$$= \langle \text{sign}, P_{\ell'}\rangle_{u_d} \int Y_{\ell,k}(z) Y_{\ell',k'}(z) d\tau_d(z) = \langle \text{sign}, P_{\ell'}\rangle_{u_d} N(d,\ell)^{-1/2} \delta_{\ell=\ell',k=k'}$$

- Gegenbauer expansion of sign: $\langle \text{sign}, P_\ell\rangle \sim \ell^{-1}$

Spectral decomposition of the target

- Given *any* basis of spherical harmonics $\{Y_{\ell,k}\}_{\ell\geq 0, k\leq N(d,\ell)}$, consider the *tensorised* basis $\{Y_{\ell,k}\otimes Y_{\ell',k'}\}_{\ell,\ell',k,k'}$ of $L^2(\mathcal{S}^{\otimes 2}_{d-1})$.
- By H-F, $\langle I, Y_{\ell,k}\otimes Y_{\ell',k'}\rangle = \int \mathrm{sign}(z^\top y) Y_{\ell,k}(z) Y_{\ell',k'}(y) d\tau_d(z) d\tau_d(y)$
$= \langle \mathrm{sign}, P_{\ell'}\rangle_{u_d} \int Y_{\ell,k}(z) Y_{\ell',k'}(z) d\tau_d(z) = \langle \mathrm{sign}, P_{\ell'}\rangle_{u_d} N(d,\ell)^{-1/2} \delta_{\ell=\ell',k=k'}$.
- Gegenbauer expansion of sign: $\langle \mathrm{sign}, P_\ell\rangle \sim \ell^{-1}$
- **Consequence**: target has energy uniformly spread in harmonics $\{Y_{\ell,k}\otimes Y_{\ell,k}\}_{\ell,k}$, with slow decay $\sim \ell^{-2}$.
- Rot.Inv $I(Uz, Uy) = I(z,y)\ \forall z, y, U\in\mathcal{O}_d$.

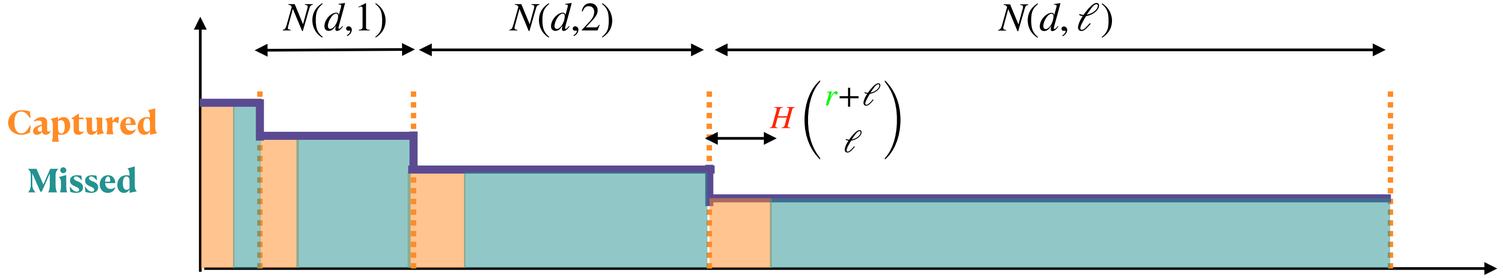


Spectral decomposition of self-attention heads

- Each self-attention head is of the form $\phi_h(K_h^\top z, y)$, with K_h of rank r .
- **Lemma**: For each ℓ and each K_h, the set $\{(z, y) \mapsto \phi_h(K_h^\top z, y)\}_{\phi_h}$ is orthogonal to at least $N(d, \ell) - \binom{r+\ell}{\ell}$ spherical harmonics $Y_{\ell,k} \otimes Y_{\ell,k}$.

Spectral decomposition of self-attention heads

- Each self-attention head is of the form $\phi_h(K_h^\top z, y)$, with K_h of rank r .
- **Lemma**: For each ℓ and each K_h, the set $\{(z,y) \mapsto \phi_h(K_h^\top z, y)\}_{\phi_h}$ is orthogonal to at least $N(d,\ell) - \binom{r+\ell}{\ell}$ spherical harmonics $Y_{\ell,k} \otimes Y_{\ell,k}$.
- **Consequence**: Approximant $\sum_{h \le H} \phi_h(K_h^\top z, y)$ is orthogonal to at least $N(d,\ell) - H\binom{r+\ell}{\ell}$ spherical harmonics.



Putting it all together

- We have $\mathbb{E}_{x_1,x_2,y}\|f(x_1,x_2,y)-g(x_1,x_2,y)\|^2 \gtrsim \mathbb{E}_{z,y}\left|\operatorname{sign}(z^\top y)-\sum_{h\le H} g_h(z,y)\right|^2$

$$= \sum_{\ell} \sum_{k\le N(d,\ell)} \langle I - \sum_{h\le H} g_h, Y_{\ell,k}\rangle^2$$

Putting it all together

- We have $\mathbb{E}_{x_1,x_2,y}\|f(x_1,x_2,y)-g(x_1,x_2,y)\|^2 \gtrsim \mathbb{E}_{z,y}\left|\operatorname{sign}(z^\top y)-\sum_{h\le H} g_h(z,y)\right|^2$

$$
\begin{aligned}
&= \sum_{\ell} \sum_{k\le N(d,\ell)} \langle I - \sum_{h\le H} g_h, Y_{\ell,k}\rangle^2 \\
&\ge \sum_{\ell} \langle \operatorname{sign}, P_\ell\rangle^2 \left(N(d,\ell) - H\binom{r+\ell}{\ell}\right) \\
&\gtrsim \sum_{\ell} \ell^{-2}\left(d^\ell - H r^\ell\right)
\end{aligned}
$$

Comments

- Extension to $N > 3$ should be doable using the same arguments.
- Approximation lower bound is nearly tight via Random Feature Approximation in a rotation-invariant RKHS — no special low-rank 'features' to be learnt, due to rotational invariance!
- Geometric arguments are similar to [Daniely'17] in the context of depth-separation.
- Extension to biased (non-equivariant) targets preserves exponential separation via positional encodings.

Comments

- Extension to $N > 3$ should be doable using the same arguments.
- Approximation lower bound is nearly tight via Random Feature Approximation in a rotation-invariant RKHS — no special low-rank 'features' to be learnt, due to rotational invariance!
- Geometric arguments are similar to [Daniely'17] in the context of depth-separation.
- Extension to biased (non-equivariant) targets preserves exponential separation via positional encodings.
- Role of MLP / extra layers?

Majority Voting

- A *random* rank-1 head hm $\left(X^\top qq^\top y\right), q \sim \tau_{d-1}$, guesses correctly with probability $1/2 + O(1/\sqrt{d})$.

Majority Voting

- A *random* rank-1 head hm $\left(X^\top qq^\top y\right), q \sim \tau_{d-1}$, guesses correctly with probability $1/2 + O(1/\sqrt{d})$.
- By Hoeffding's inequality, the majority vote of $H \gtrsim d^2$ heads thus guesses correctly with probability greater than $1 - \exp(\Theta(d))$.

Majority Voting

- A *random* rank-1 head hm $\left(X^\top q q^\top y\right), q \sim \tau_{d-1}$, guesses correctly with probability $1/2 + O(1/\sqrt{d})$.
- By Hoeffding's inequality, the majority vote of $H \gtrsim d^2$ heads thus guesses correctly with probability greater than $1 - \exp(\Theta(d))$.
- To tally the votes, need extra "index" and "scratchpad" dimensions: $\begin{bmatrix} \mathbf{x}_1 \\ 1 \\ 0 \end{bmatrix}, \begin{bmatrix} \mathbf{x}_2 \\ -1 \\ 0 \end{bmatrix}, \begin{bmatrix} \mathbf{y} \\ 0 \\ 0 \end{bmatrix}$

Majority Voting

- A *random* rank-1 head hm $\left(X^\top q q^\top y\right), q \sim \tau_{d-1}$, guesses correctly with probability $1/2 + O(1/\sqrt{d})$.
- By Hoeffding's inequality, the majority vote of $H \gtrsim d^2$ heads thus guesses correctly with probability greater than $1 - \exp(-\Theta(d))$.
- To tally the votes, need extra "index" and "scratchpad" dimensions: $\begin{bmatrix} \mathbf{x}_1 \\ 1 \\ 0 \end{bmatrix}, \begin{bmatrix} \mathbf{x}_2 \\ -1 \\ 0 \end{bmatrix}, \begin{bmatrix} \mathbf{y} \\ 0 \\ 0 \end{bmatrix}$
- 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.

Conclusions / Future Work

- Self-Attention layer: pairwise interactions provably more powerful than unary structure. “Canonical” model for interactions? (physics).
- Role of (Rank, Heads) in Self-Attention layers: *not* captured by ‘parameter count’ rH.
- Metric functions have rotational symmetry, hard to approximate with low-rank structure.
- For fixed N, may compensate with additional layers. Dependency in N necessary?
- Towards language applications: replace $\mathcal{S}_{d-1}$ by discrete metric space. Does lower bound also hold?

Thanks!



References:

- *Exponential Separations in Symmetric Neural Networks, Zweig, Bruna, NeurIPS'22.*
- *Benefits of rank in Attention Layers, Amsel, Bruna, Yehudai, preprint'24.*