Towards Understanding Inductive Bias in Transformers: A View From Infinity

Itay Lavie [1] **Guy Gur-Ari** [2] **Zohar Ringel** [1]

Abstract

We study inductive bias in Transformers in the infinitely over-parameterized Gaussian process limit and argue transformers tend to be biased towards more permutation symmetric functions in sequence space. We show that the representation theory of the symmetric group can be used to give quantitative analytical predictions when the dataset is symmetric to permutations between tokens. We present a simplified transformer block and solve the model at the limit, including accurate predictions for the learning curves and network outputs. We show that in common setups, one can derive tight bounds in the form of a scaling law for the learnability as a function of the context length. Finally, we argue WikiText dataset, does indeed possess a degree of permutation symmetry.

1. Introduction

Transformers show state-of-the-art performance on a wide variety of tasks (Wolf et al., 2020; Dosovitskiy et al., 2021; Chen et al., 2020; Min et al., 2022; Brown et al., 2020) with seemingly ever-improving performance (Kaplan et al., 2020; Henighan et al., 2020). The past year has brought forth larger and more capable models than ever before (Jiang et al., 2024; OpenAI, 2023; GeminiTeam, 2023), yet our understanding of them falls behind (Goyal & Bengio, 2022; Wen et al., 2023).

Recent works have advanced us in understanding specific aspects and behaviors like grokking (Nanda et al., 2023; Rubin et al., 2023; Liu et al., 2022b;a), in-context learning (Von Oswald et al., 2023; Olsson et al., 2022), and out-of-distribution (OOD) generalization (Nam et al., 2022; Canatar et al., 2021a). However, a unified view of the inductive bias of transformers is still lacking. It has been claimed that understanding and designing networks with better inductive bias is a necessary step toward AI (Goyal & Bengio, 2022); this can also make them safer and more suitable for deployment in high-risk situations (see for example Bommasani et al. (2021) and European Union AI act (EuropeanCommission, 2021)).

We approach the challenge from the infinitely over-parameterized Gaussian process (GP) limit, where the neural network (NN) becomes more analytically tractable but still shares many qualitative and quantitative similarities with finite NNs used in real life (Lee et al., 2020; Shankar et al., 2020; Novak et al., 2018). An established correspondence between infinitely wide NNs and Bayesian inference with a GP (Neal, 1996; Lee et al., 2020; Welling & Teh, 2011; Naveh et al., 2021) allows us to identify the inductive bias of the NN in the GP limit with its Bayesian prior, thereby yielding a demystified concrete expression for the inductive bias. This approach has proven itself in the study of deep fully connected networks and convolutional neural networks (CNNs) by enabling accurate prediction of learning curves and explaining phenomena such as reduction of task complexity due to weight sharing (Novak et al., 2018; Naveh et al., 2021) and CNN's utilization of hierarchical structure in the data (Cagnetta et al., 2023).

In this work, we characterize the inductive bias by specifying how many samples will be required to learn a target function. We show that when the dataset possesses a permutation symmetry, learnability is closely tied to the irreducible representations (irreps) of the symmetric group. Namely, the more symmetric the function to permutations, as quantified below, the more learnable it is. We present an instructive example, work through it, and show we can accurately predict its outputs at the limit including performance under distributional shift. We then show that our predictions are a good approximation for wide but finite networks; these predictions again hold OOD. We present learnability bounds for richer networks, including a standard transformer block. Finally, we argue natural language (NL) does have some permutation symmetry, based on an analysis of WikiText-2 (Merity et al., 2016).

Our main contributions are:

- We give explicit analytical predictions for the outputs and generalization performance of a NN with linear

[1]Racah Institute of Physics, Hebrew University of Jerusalem, Jerusalem 91904, Israel [2]Augment Computing. Correspondence to: Itay Lavie <itay.lavie@mail.huji.ac.il>.

Proceedings of the 41st International Conference on Machine Learning, Vienna, Austria. PMLR 235, 2024. Copyright 2024 by the author(s).

attention at the GP limit, in distribution and OOD. We show how irreducible representations of the symmetric group can be built and used for to predict learnability in this case.

- We extend our results to a transformer block with standard $\mathrm{softmax}$ attention. We show experimentally the learnability bounds found based on the dimension of the relevant irreducible representations are tight.
- We analyze WikiText-2 and show evidence for an approximate permutation symmetry in its principal components, suggesting that the toolbox presented can be of use in natural language datasets.

2. Methods

This section presents the model we study and the tools used to analyze it at the GP limit. We show how to implement these tools to reveal the inductive bias in terms of the space of *expressible* functions and their *learnability* by the NN.

2.1. Model

2.1.1. Neural network architecture

We study a transformer-like NN with one transformer block, for simplicity, we do not include residual connections or layer normalization, although these can be added. The NN is made of an embedding layer with added learned positional encoding (PE) $\vec{p}$, one multi-head self-attention layer (MHA), an MLP with one hidden layer and a final linear readout layer.

The input to the NN is made out of $L+1$ tokens $\vec{x}^a$ indexed by an upper sequence index $a = 1, 2, ..., L+1$ with each token having an internal vocabulary dimension indexed by a lower index i. We group these with a Greek letter sample index $\mu = 1, 2, ..., N$ into a rank 3 tensor $X^a_{i,\mu}$, where we drop the sample index μ when we discuss only a single sample. One-hot encoding is used for the tokens, such that $[\vec{x}^a]_i = \delta_{i,v}$ where $v = 1, ..., N_{\rm voc}$ is the token represented by $\vec{x}^a$. Denoting the input by x^a_i and the output of l'th layer by $z^{(l),a}_i$ the resulting NN is

$$
\begin{aligned}
z^{(1),a}_i &= W^{\rm emb}_{ij} x^a_j + p^a_i \\
z^{(2),a}_{i,h} &= \Phi\left(\frac{Q^a_{j,h} K^b_{j,h}}{\sqrt{d_k}}\right) V^b_{i,h} \quad \text{(no } h \text{ summation)} \\
&= \Phi\left(\frac{W^Q_{lm,h} z^{(1),a}_m W^K_{ln,h} z^{(1),b}_n}{\sqrt{d_k}}\right) W^V_{ij,h} z^{(1),b}_j \qquad (1) \\
z^{(3),a}_i &= W^O_{ij,h} z^{(2),a}_{j,h} \\
z^{(4),a}_i &= \phi\left(W^{(4)}_{ij} z^{(3),a}_j + b^{(4)}_i\right) \\
z^{(5),a}_i &= W^{(5)}_{ij} z^{(4),a}_j + b^{(5)}_i \\
f^a_i(X) &= z^{(6),a}_i = W^{\rm d-emb}_{ij} z^{(5),a}_j
\end{aligned}
$$

using Einstein's summation convention, with Φ and ϕ being some activation functions[1]. The NN parameters

$$
\begin{aligned}
W^{\rm emb} &\in \mathbb{R}^{d_m \times N_{\rm voc}}, & \vec{p}^a &\in \mathbb{R}^{d_m} \\
W^Q, W^K, W^V &\in \mathbb{R}^{d_k \times d_m}, & W^O &\in \mathbb{R}^{d_m \times d_k \times N_{\rm heads}} \\
W^{(4)} &\in \mathbb{R}^{d_{ff} \times d_m}, & \vec{b}^{(4)} &\in \mathbb{R}^{d_{ff}},\ \vec{b}^{(5)} \in \mathbb{R}^{d_m} \\
W^{(5)} &\in \mathbb{R}^{d_m \times d_{ff}}, & W^{\rm d-emb} &\in \mathbb{R}^{N_{\rm voc} \times d_m}
\end{aligned} \qquad (2)
$$

are all learned. For the MHA we use $N_{\rm heads}$ heads and the same dimension $d_k = d_v = d_m/N_{\rm heads}$ for keys, queries, and values. Lastly, for the hidden layer $z^{(4)}$ we use dimension d_{ff} which is of the same order of magnitude as the model dimension $d_{ff} \sim d_m$. Notably, consecutive affine transformations can be combined together without loss of generality, but they are kept in this way to align with standard notation[2].

As an instructive example, we will use a linearized MHA[3] $\Phi(x) = \frac{1}{L+1}x$ and linear MLP $\phi(x) = x$, as this setting allows for closed-form analytical predictions at the GP limit. Note that because we remove the common $\mathrm{softmax}$ nonlinearity we add a division by the length to make sure the network's output stays $O(1)$ and does not scale with L.

2.1.2. Task, Loss Function, Initialization, and Training Protocol

The task is a pretraining task, namely, predicting the conditional probability for the next token given the context $p(\vec{x}^{L+2}|X)$. For simplicity, we limit the discussion to predicting the next token probability from a full context window of length $L+1$, and looking only at the prediction for the unknown token, meaning we define $f(X) := f^{L+1}(X)$.

Mean square error (MSE) loss with weight decay is used. The weights are initialized according to LeCun initialization, meaning the weights in each layer are i.i.d with $w \sim \mathcal{N}(0, \frac{1}{\sqrt{\text{fan-in}}})$, and the biases are initialized to zero. For the convenience of the analytical calculations, we will initialize the PE as Gaussian i.i.d entries $p^a_i \sim \mathcal{N}(0, 1/2)$ for $a \neq L+1$, for the last token we will initialize the PE to zero $p^{L+1}_i = 0$.

The training is done with uncorrected Langevin dynamics (Neal, 1993), that is, gradient descent with noise $\eta \sim N(0, \sigma^2)$ added to the gradients, as a model for stochastic gradient descent. As shown in Mingard et al. (2020); Liu et al. (2021); Mandt et al. (2018) the result is indicative

[1] A common choice would be $\Phi = \mathrm{softmax}$ acting on the b index and $\phi = \mathrm{ReLU}$

[2] Combining such affine transformations would also induce a different prior in finite-sized NNs as shown in Li & Sompolinsky (2021).

[3] similar to the one suggested by Von Oswald et al. (2023); Hron et al. (2020) and recently studied by Ahn et al. (2024).

of the outcome of SGD training. We adjust the noise and weight decay such that following training with no data the network's weights distribution is the same as at initialization, as described in Naveh et al. (2021). From a Bayesian perspective, such training protocol samples from the posterior distribution of a Bayesian NN with prior induced by the weights' initialization distribution.

2.1.3. Datasets

We use a mixture of hidden Markov models (HMMs) (Baum & Petrie, 1966) as a dataset. The mixture of HMMs is chosen for its balance between aspects of language, like long-range dependencies and sensitivity to (elementary) context (Xie et al., 2021), and analytical tractability. This setting also yields a well-defined concept of distributional shift, as the NN can be trained on a fraction of the mixture and tested on another.

A HMM is composed of two stochastic processes, h^a and x^a, where a is the time-step index. The process h^a is dubbed "hidden" while x^a is the observed process. The hidden process is Markovian, with d_{hidden} different states. The observed process depends only on the hidden state at the same time, where each of the possible N_{voc} outputs is given a different probability under each hidden state.

HMMs are conveniently described by stochastic emission and transition matrices. The i, j entry of the transition matrix $T \in \mathbb{R}^{d_{\text{hidden}} \times d_{\text{hidden}}}$ represent the transition probability from the j'th hidden state to the i'th. Similarly, the i, j entry of emission matrix $O \in \mathbb{R}^{N_{\text{voc}} \times d_{\text{hidden}}}$ represent the probability to emit the i'th output in the vocabulary when in the j'th hidden state.

Our dataset is a mixture of HMMs with $N_{\text{voc}} = 2$ and $d_{\text{hidden}} = 2$, where the emission probabilities that define the HMM p, q are themselves drawn from uniform distributions $p \sim U(p_a, p_a + w),\ q \sim U(q_a, q_a + w)$. The transition probabilities are constant and deterministic. The transition and emission probabilities for a HMM in the mixture are given in matrix form by

$$T = \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix}; \quad O = \begin{bmatrix} p & q \\ 1-p & 1-q \end{bmatrix}. \tag{3}$$

Finally, the initial hidden state, h^1, is a random variable with equal probability for each of the two possible hidden states.

As a primer for the discussion to follow, we point out that the probability distribution defined by an HMM is invariant to permutation of tokens outputted under the same hidden state. We re-examine this point in section 3 and present evidence for an approximate permutation symmetry in the principal components of WikiText.

2.2. Theory

First, we present the correspondence between an infinite transformer-like NN and a Gaussian process (GP), known as the neural network Gaussian process correspondence (NNGP) (Lee et al., 2018; Hron et al., 2020). To discuss the inductive bias in terms of functions rather than individual samples, we study the resulting GP averaged over datasets and its spectral decomposition. The null space is identified as the space of inexpressible functions, and the eigenvalues are interpreted as a measure of learnability for their corresponding eigenfunctions. Capitalizing on the permutation symmetry of the kernel eigenvalue problem, we use tools from representation theory to provide upper bounds on the scaling of eigenvalues with context length.

As shown in Hron et al. (2020) when $d_k, N_{\text{heads}} \to \infty$ the distribution of NN outputs induced by the initialization distribution converges to a GP $f(x) \sim \mathcal{GP}(0, k)$ with k being the kernel of the NN. The kernel $k(X, Y)$ is the covariance between samples given by

$$k(X, Y) = \mathbb{E}_\Theta \left[f_\Theta(X) f_\Theta(Y) \right], \tag{4}$$

where Θ stands for the NN parameters, drawn from the initialization distribution defined in 2.1.2. Following our training protocol, this correspondence carries over to the trained NN, such that the trained NN is equivalent to the result of Bayesian inference with the NNGP as prior (Naveh et al., 2021).

While exact GP inference is generally hard, we can get an insight into the learning process by looking at the continuum limit where GP regression is averaged over all possible datasets of size N. For a large N, a continuum kernel can be used in an approximation known as the equivalent kernel (EK) (Silverman, 1984; Sollich & Williams, 2004; Cohen et al., 2021). We can define the EK integral operator $\hat{K}$ associated with the continuum kernel,

$$\hat{K} u(X) = \mathbb{E}_{Y \sim p_{\text{train}}} \left[k(X, Y) u(Y) \right], \tag{5}$$

where p_{train} is the training distribution.

The inductive bias can now be understood by looking at the expression for the predictor (the output of the NN, averaged over the possible initializations) under the EK approximation on the eigenbasis of $\hat{K}$

$$\mathbb{E}_\Theta \left[f_\Theta(X_*) \right] = \sum_{i=1}^{\infty} \frac{\lambda_i}{\lambda_i + \sigma^2/N} g_i \varphi_i(X_*) \tag{6}$$

where σ^2 is the variance of the noise added to the gradients, or the observation uncertainty from a Bayesian perspective; φ_i's are the eigenfunctions; λ_i's are the corresponding eigenvalues and g_i is the projection of $g(x)$ on φ_i given by the inner product

$$\langle g(x), \varphi_i(x) \rangle_x = \mathbb{E}_{x \sim p_{\text{data}}} \left[g(x) \varphi_i(x) \right]. \tag{7}$$

We can give equation 6 an intuitive interpretation: The architecture and dataset dictate both the *learnability* and *expressibility*. All eigenfunctions corresponding to $\lambda = 0$ will not be expressible by the NN. For the expressible eigenfunctions, learnability for the eigenfunction φ_i is dictated by the quantity

$$L_i := \left| \frac{\langle f(x), \varphi_i(x) \rangle_{x \sim p}}{\langle y(x), \varphi_i(x) \rangle_{x \sim p}} \right| = \frac{\lambda_i}{\lambda_i + \sigma^2/N}. \quad (8)$$

Where the last equality is given by using the EK predictor for $f(x)$ calculated on the same distribution $p(x)$ as the inner product, recovering the result of (Simon et al., 2023) (for effective ridge $\delta = \sigma^2$). With this matching EK predictor, the learnability $0 \leq L \leq 1$ acts as a filter, passing almost nothing when $N \ll \sigma^2 \lambda^{-1}$ and passing the information from the target almost perfectly when $N \gg \sigma^2 \lambda^{-1}$. We see that learning the feature $\varphi_i(x)$ requires $N \simeq \sigma^2 \lambda^{-1}$ samples, predicting a performance improvement at that scale.

Accordingly, predicting the learning of the NN is reduced to solving the eigenvalue problem for the EK kernel corresponding to the NN

$$\hat{K} \varphi_i(X) = \mathbb{E}_{Y \sim p_{\text{train}}} [k(X, Y) \varphi_i(Y)] = \lambda \varphi_i(X), \quad (9)$$

and finding the projections of the target on the eigenbasis.

We now turn to use symmetry to simplify the eigenvalue problem. The fact that the NN defined in 2.1, never explicitly acts in sequence space (that is, the weights do not carry a sequence index) and the PE is drawn i.i.d guarantees a permutation symmetry between all the tokens but the last one[4] [5].

2.2.1. Symmetry and Representation Theory

We start with an intuitive understanding of the role of symmetries and give a precise formulation later in this section. A fuller introduction and examples are given in Appendix B. For a simple example where our use of representation theory amounts to a simple discrete Fourier transform, and introduction to permutation symmetry in appendix A.

Symmetries can greatly simplify the eigenvalue problems like equation 9 above. We say an operator like $\hat{K}$ is symmetric under the action of a group G if

$$\forall g \in G, \;\; k(\vec{x}_g, \vec{y}_g) = k(\vec{x}, \vec{y}) \;\&\; p_{\text{data}}(\vec{x}_g) = p_{\text{data}}(\vec{x}), \quad (10)$$

where $\vec{x}_g$ is the result of acting with a symmetry action g on $\vec{x}$, e.g. rotating $\vec{x}$ or permuting the entries of $\vec{x}$. Such an action is formalized through a *representation* of the group, we give a precise definition in Prop. 2.1 . A symmetry, as described in equation 10, means we are allowed to act with a symmetry action $g \in G$ but our model will stay invariant to this action. In the context of the eigenvalue problem in equation 9, such an action can be viewed as mixing different eigenfunctions $\varphi_i(x)$ (say by rotating the inputs x, such that the outputs $\varphi_i(\vec{x}_g)$ overlaps with $\phi_j(x)$ for $i \neq j$) without changing the eigenvalues. This scenario implies, that all the eigenvalues of the mixed eigenfunctions must be identical, i.e. degenerate. Moreover, all eigenfunctions must either be members of such degenerate blocks, or in the simplest case be transformed back to themselves under the action of the symmetry group. See Fig 1.

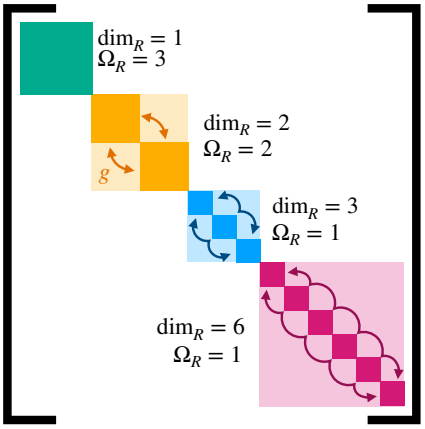

Figure 1. **(Illustration of diagonalization using symmetries)** The figure illustrates the direct sum (block) structure described in Prop. 2.1. Each color-shaded block represents an irrep, and each solid color represents a multiplicity block within the irrep. All elements outside the multiplicity blocks vanish, both between different irreps and within an irrep. The symmetry actions $g \in G$ can mix multiplicity blocks as indicated by the arrows. Since all multiplicity blocks inside an irrep are linked by the symmetry actions they are all degenerate .

If we study precisely how a symmetry group mixes the functions, we can identify the above-mentioned blocks in the space of expressible functions. The blocks would be a property of the symmetry group itself and would hold for any kernel satisfying equation 10. Formally, the blocks correspond to the irreps of the group over the space of expressible functions (see Prop. 2.1). These can be understood as the minimal spaces of functions that mix with one another. The functions in those spaces cannot be "untangled" under the symmetry, hence the name irreducible.

Proposition 2.1. *Recalling results from Tung (1985); Fulton & Harris (2004). Given linear transformations $\{T_g | g \in G\}$ which constitute a representation of G ($\forall g_1, g_2 \in G, \;\; T_{g_1 g_2} = T_{g_1} T_{g_2}$) and a model symmetric under the action of a group G, i.e. satisfying equation 10 with $x_g = T_g x$.*

[4] This is also true for the NTK (Jacot et al., 2018) and we expect similar results to hold in that setting.

[5] the last token is "signaled out" at inference time as the only one who's output is desired. For a single transformer block, it is therefore sufficient to use only the last token as a query, and it is thereby not symmetric to the other tokens.

It holds that: *The kernel operator can be decomposed into a direct sum, where each summand corresponds to an irrep of G (shaded blocks in Fig.1). For an irrep R that appears Ω_R times in $\hat{K}$ (said to have a multiplicity Ω_R), each such block consists of Ω_R different eigenvalues, each with m-fold degeneracy, equal to the dimension of the irrep ($\dim_R$).* **As a corollary**, *each irrep of multiplicity 1 gives exact eigenvectors of the kernel. For an irrep of multiplicity Ω_R, finding the spaces of the irrep allows one to diagonalize in the $\Omega_R \times \Omega_R$ (multiplicity) space for each irrep individually; these spaces are guaranteed not to mix different irreps under the kernel.*

Going back to a more intuitive level, multiplicity means different sets of functions mix in the same way, but not between themselves. To separate these sets into eigenspaces the eigenvalue problem in the $\Omega_R \times \Omega_R$ multiplicity space needs to be solved in other means, but we are guaranteed we need to solve it in only one such multiplicity block, as all blocks are guaranteed to be degenerate (one solid color square of each color in Fig 1).

Degeneracy not only allows us to simplify the problem but also to give an asymptotic upper bound on the eigenvalues. Mercer's theorem (König, 1986) guarantees $\hat{K}$ has a finite trace, which can be thought of as a fixed budget. Since all the eigenvalues are positive, they must share this fixed budget; leading to Prop. 2.2.

Proposition 2.2. *Under the same conditions as Prop 2.1 and given the kernel is normalized, the trace is given by*

$$\mathbb{E}_{x\sim p_{\text{data}}}\left[k(x,x)\right] \simeq 1. \tag{11}$$

An eigenvalue λ belonging to a space corresponding to an irrep R, is bound from above, $\lambda = O(\dim_R^{-1})$ where $\dim_R$ is the dimension of R.

Since the kernel's trace is fixed to unity, and all eigenvalues in an irrep are degenerate (equal), the eigenvalues are upper bounded by one over the number of such eigenvalues, as they have to sum up to at most 1.

We can now state formally the symmetry of our model as symmetry under the action of the symmetric group in L symbols S_L i.e. $k(T_{s_L} X, T_{s_L} Y) = k(X,Y)$ where T_{s_L} is a representation of any element $s_L \in S_L$ that acts naturally on the sequence index [6]. Following Prop. 2.1,2.2 and the symmetry manifested in the model, we are interested in the irreps of the symmetric group.

Irreps of the symmetric group S_L are uniquely labeled by partitions of L to integers, written as ordered sets from the largest part to the smallest, such that the sum of the parts is L. For example, a partition to L could be $(L-1,1)$ or $(L-4,2,1,1)$. To decompose the space of expressible functions we use the extensive literature on the representations of the symmetric group; a less formal introduction is given in Appendix B, and a formal treatment is given in Appendix C.

Since the input is one-hot encoded, every target function will be a multilinear polynomial in the input tokens; that is, fixing all other variables we will remain with a linear function of x_i^a for some particular a, i. This fact can be seen by considering each variable x_i^a can only take on values $\{0,1\}$ so $(x_i^a)^n = x_i^a$ for $0 < n \in \mathbb{Z}$. We thus wish to consider the decomposition of multilinear polynomials to irreps of the symmetric group.

Theorem 2.1. *The space of homogeneous multilinear polynomials in n variables of degree d can be* fully *decomposed into* $\min\{d+1, n-d+1\}$ unique *irreps of S_n labeled by the partitions $(n-k,k)$ for $0 \le k \le d, n-d$.*

See proof in appendix C. We can therefore expand any analytic function into polynomials and decompose them into the irreps of the symmetric group.

The dimension of the k'th irrep ($\dim_k$) of the form $(L-k,k)$ scale as $\dim_k = \frac{L!}{k!\frac{(L-k+1)!}{L-2k+1}} \sim L^k$. We can now quantitatively define a measure for symmetry to permutations: the more symmetric a function is, the less it may mix with other functions, and the smaller the dimension of the irreps it belongs to (smaller k). Going back to the definition of learnability in equation 8 we see that the number of samples required to learn a function in the representation $(L-k,k)$ is asymptotically bounded from below by

$$N \simeq \lambda_{(L-k,k)}^{-1}\sigma^2 = \Omega(L^k). \tag{12}$$

We therefore see that the more symmetric a function is to permutations (smaller k) the more learnable it is.

2.3. Example: Linear Activations

In this example, we choose $\Phi(x) = \frac{1}{L+1}x$ and linear MLP $\phi(x) = x$, as previously noted in 2.1.1 and solve the eigenvalue problem presented in the previous section. Note the linear activation functions Φ, ϕ do not imply a linear NN as the attention layer is inherently non-linear. While this example is a minimal transformer like NN, our dataset already goes beyond the landscape of complete permutation invariance and demonstrates how the tools presented above can be adapted to richer datasets where the permutation invariance is partially broken.

2.3.1. Expressibility

First, we want to identify the space of functions spanned by φ_i with $\lambda_i \neq 0$, the space of expressible functions.

Claim 1. *The space of functions expressible by the model stated in section 2.1 is spanned by the linear functions of*

[6] We note this is a symmetry of the prior distribution and this is all that is required for our theory. The posterior distribution need not have this symmetry, as is often the case with learned positional encoding.

$\{x_1^a\}_{a=1}^L$ multiplied by linear functions of x_1^{L+1}, which is a $2L+2$ dimensional space.

The kernel function corresponding to our NN is given by

$$k(X,Y) = \frac{1}{8}\vec{x}^{L+1}\cdot\vec{y}^{L+1}\frac{1}{(L+1)^2}\sum_{a,b=1}^{L+1}\left(\vec{x}^a\cdot\vec{y}^b+\delta^{a,b}\right)^2. \tag{13}$$

One-hot encoding not only implies multilinearity of the outputs, but also guarantees multilinearity of the kernel in the inner product of two vectors $(\vec{x}^a\cdot\vec{y}^b)^n = (\vec{x}^a\cdot\vec{y}^b)$ for $0 < n \in \mathbb{Z}$. In this example, it means only linear terms in the context window $a, b = 1, ..., L$ are present.

We can further restrict the model's expressibility in our case, by considering large context windows $L \gg 1$. In that case, we can approximate the kernel given in equation 13 by summing only up to L, and dropping sub-leading contributions in $\frac{1}{L}$. We show these indeed give only sub-leading corrections in appendix F. These simplifications result in

$$k(X,Y) = \underbrace{\vec{x}^{L+1}\cdot\vec{y}^{L+1}}_{\mathfrak{A}}\frac{1}{8L^2}\underbrace{\sum_{a,b=1}^{L}\left(\vec{x}^a\cdot\vec{y}^b+\frac{\vec{x}^a\cdot\vec{y}^a}{L}+\frac{1}{L}\right)}_{\mathfrak{B}}. \tag{14}$$

Finally, Since our particular model uses a vocabulary of size 2 the entries of a one-hot vector are completely determined by one another $x_2^a = 1 - x_1^a$, allowing us to write it using only the first entry. As can be seen in equation 9, the only X dependence in the l.h.s comes from the kernel $k(X,Y)$, thus for the equality to hold for every X, the eigenfunction $\varphi_i(X)$ of $\lambda_i \neq 0$ must be in the space of functions spanned by $k(X,\cdot)$. For example, if $k(X,Y)$ is linear in X only linear functions will be expressible. Based on this argument (formally given by the representer theorem (Schölkopf et al., 2001)), we may conclude the space of expressible functions is spanned by linear functions of $\{x_1^a\}_{a=1}^L$ multiplied by linear functions of x_1^{L+1}; a space of dimension $2L+2$.

2.3.2. Learnability

Moving from expressibility to learnability requires knowledge of the full spectrum of the kernel. While this problem is generally hard, we will use the representation theory tools developed above to simplify it. We would like to decompose the zeroth and first-degree multilinear polynomial to irreps.

Claim 2. *The space of expressible functions of the model described above can be decomposed into irreps as follows. A trivial representation ($\dim = 1$) of multiplicity 6 and a standard representation ($\dim = L/2-1$) of multiplicity 4.*

Starting from the largest structure, notice the kernel is a product of two terms ($\mathfrak{A}, \mathfrak{B}$ in equation D.1). The $\mathfrak{A}$ part is diagonalized in the basis

$$a(\vec{x}^{L+1}) = x_1^{L+1},\ \ b(\vec{x}^{L+1}) = (1 - x_1^{L+1}), \tag{15}$$

which leads to multiplicity 2 for all irreps involved; one belonging to the a block and one to the b block of $\mathfrak{A}$.

Moving on to the $\mathfrak{B}$ term, as expected from the general argument presented in the previous section, we find it is symmetric under the action of the permutation in the symmetric group S_L on the set of tokens in the context window $\{x^a\}_{s=1}^L$. The full S_L symmetry is not, however, presented in the probability distribution of our chosen dataset, as tokens have different emission probabilities under different hidden states. Nevertheless, a smaller symmetry is preserved, allowing permutations only within the same hidden states. Since the transition between hidden states is deterministic, we find that all odd (even) tokens belong to the same hidden state and can be permuted between themselves, giving rise to the smaller symmetry group $S_{L/2}^{\text{odd}} \times S_{L/2}^{\text{even}} := \mathcal{S}$ [7] as a symmetry of $\hat{K}$.

As discussed in the previous subsection, in this example, only polynomials up to first degree can have non-vanishing eigenvalues. First degree polynomials are decomposed to two irreps (see theorem 2.1), namely the trivial $(L/2)$ and standard representation $(L/2-1, 1)$. The trivial representation has dimension 1 with multiplicity 4: one for the even subspace and one for the odd subspace, times 2 for the multiplicity from $\mathfrak{A}$. The standard representation has dimension $L/2-1$ with multiplicity 4, broken down in the same way as before. For zeroth degree polynomials (constants) only the trivial representation exists, of multiplicity 2: again, the coming from $\mathfrak{A}$. The process of concretely writing down the functions that span the space can be simplified further using the cyclic sub-group $C_{L/2}$, this process is details in appendix D.

Using symmetries and the partition to $\mathfrak{A}, \mathfrak{B}$ we are able to reduce the eigenvalue problem to two[8] 3×3 spaces of the trivial representation, which are diagonalizable in closed form, and a diagonalized $2L-4$ dimensional space of the standard representation. We can repeat the same procedure for polynomials of any order and decompose them to irreps (see appendix B for a discussion of the method, and an example); thereby allowing us to expand the results to a wider class of NNs including non-linear and deeper NNs.

Gathering the results of this section, **Given**: (1) equation 6, together with the (2) learnable target given in equation I.8, the (3) eigendecomposition given in equations D.3, I.7, and the (4) eigendecomposition of the two 3×3 spaces spanned by the basis in equation I.1. **One can** predict accurately the output of the model described in section 2.1 with linear activation functions in the GP limit. Additionally, One can make accurate predictions for the generalization loss, even under a distributional shift.

[7] Assuming L is even for simplicity

[8] One for the a block and one form the b block

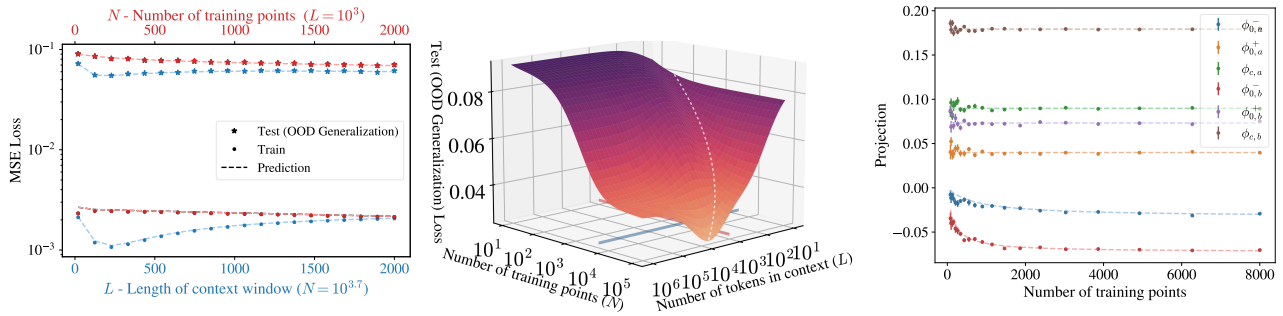

Figure 2. **Left: (theory vs. experiment)** Two sections, along constant N (in blue) and L (in red), of the MSE loss as shown in the center. We find good agreement between our theoretical predictions (calculated for the train and test distributions) and exact inference with a GP, equivalent to inference with an infinitely wide NN. Stars indicate the experimental MSE loss calculated on the test dataset, where the majority of samples are OOD w.r.t to training dataset. **Center: (learnability scaling law)** Prediction for the generalization MSE. We see the learnability threshold as a diagonal valley (marked by the dashed line) of constant N/L ratio as a consequence of having target eigenvalues of scaling $\lambda \sim L^{-1}$. This is the onset of the regime where there are enough datapoints to use the full potential of the context length. **Right: (linear MLP predictions extend to non-linear MLP)** Projections of the exact GP inference predictor with *non-linear MLP* on the basis vectors $\{\varphi\}$ specified in equation I.1, together with the theoretical prediction for the projection for a GP with *linear MLP* as a function of the number of training points. The predictions for the transformer block with linear attention and linear MLP are useful for non-linear MLP as well.

3. Experimental Results

In this section, the theory is compared to numerical experiments. We start by comparing our predictions for the example of linear activation functions with exact Bayesian inference using the NNGP prior, which is equivalent to the output of a NN train with Langevin dynamics. We predict the performance OOD and show good agreement with experiments. The N-L scaling law is shown and the kernel for erf activation function is shown to give a similar predictor per N. We then move on to real, wide but finite, NN with non-linear activation functions trained by SGD. We show nontrivial predictions such as OOD performance carryover from the linear MLP to the non-linear MLP case. We then present the kernel's spectrum of an NN with standard $\mathrm{softmax}$ attention and show that the scaling law bounds derived on the eigenvalues become tight. Lastly, we analyze WikiText-2 and show that at leading order correlations the dataset does indeed appear to be permutation symmetric to a good approximation.

3.1. Linear Activation Functions

On the left of Fig. 2 the predictions for the loss as a function of N and L are presented, together with exact Bayesian inference, showing good agreement both on train ($p \sim U(0.4, 0.4 + 10^{-1.5})$, $q \sim U(0.5, 0.5 + 10^{-1.5})$) and test ($p, q \sim U(0, 1)$) distribution loss. For details on the analytical predictions for the loss see Appendix E.

The $2d$ manifold of the analytical prediction for the (OOD generalization) loss computed on a mixture of HMMs drawn from $p, q \sim U(0, 1)$, as a function of N, L is presented in Fig. 2 center. As shown in Appendix D, only functions that belong to the trivial representation contribute to spanning the target. Four out of six corresponding eigenvalues scale as $\lambda \sim 1$ and the other two scales as $\lambda \sim L^{-1}$. Resulting in the scaling law $N \simeq 1/\lambda \propto L$ This result appears as a connected valley of minimal loss when holding N constant, i.e. a constant $N \propto L$ ratio.

Lastly, we compare the GP predictor for a network with non-linear MLP $\phi(x) = \mathrm{erf}(x)$ (in reference to equation 1) to that of the linear example solved above. To the right in Fig 2, the exact GP inference (computed for *non-linear* MLP) predictions projected on to the $\{\varphi\}$ basis vectors in equation I.1 are compared with analytical predictions for a network with *linear* MLP as a function of the number of samples in the dataset. They match very well in terms of learnability as seen by the rate at which the projection coefficients reach their asymptotic value . We see the NN with non-linear MLP follows the same trend and learns a similar predictor to the one with linear MLP.

3.2. Finite Neural Networks and Natural Language

While the example of linear activation functions can be fully understood one may wonder whether the conclusions carry over to more common scenarios, here we address these concerns. The doubts we will try to dispel are

- While there is an established correspondence between Bayesian inference with a GP and an infinite NN, does a finite NN, trained with SGD with a finite learning rate show similar results?
- Could it be that the OOD generalization is a result of the limited expressibility and or the GP limit?

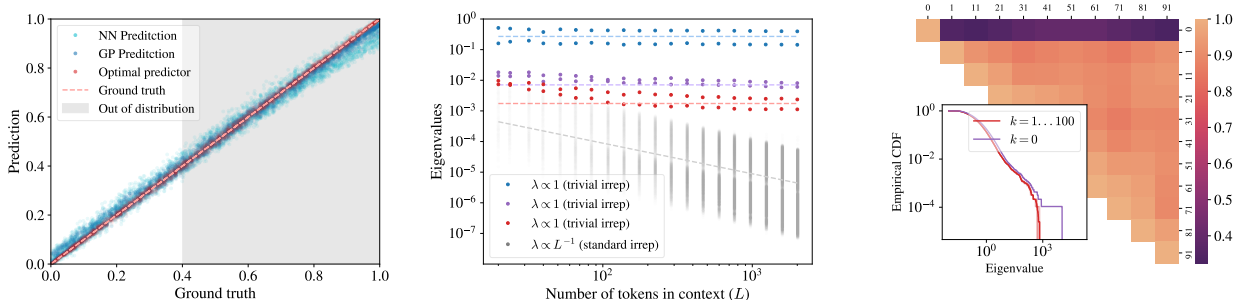

Figure 3. **Left: (Predicted OOD generalization matches experiment)** Finite NN and exact GP inference (cyan and blue respectively), optimal predictions (red) and ground truth (dashed light red) as a function of the ground truth target for the sample. We see good agreement between predictions and experiment in distribution and OOD. The performance OOD is only slightly worse than in distribution, as indicated by the spread around the ground truth. **Center: (the scaling bounds are tight for eigenvalues of softmax attention)** The spectrum of the empirical kernel of a network with $\mathrm{softmax}$ attention as a function of the context length (L). The scaling with L is bound tightly by the scaling deduced from the dimension of the corresponding irrep of the symmetric group. The light dashed lines serve only as a guide to the eye for the scaling law; they are not predictions for specific values. **Right: (evidence for approximate permutation symmetry in WikiText)** The triangle shows the cosine similarity induced by the Frobenius inner product between the linear features of WikiText C^{kk} and $C^{k'k'}$ for the k's indicated on the boundary. We see all sampled $k \neq 0$ are similar to one another but different from $k = 0$ as predicted by the irrep decomposition. The Empirical CDF plot shows the CDF for the eigenvalues of those sampled matrices. Different k's for $k \neq 0$ are almost identical. $k = 0$ has a distinct distribution.

- Are the bounds derived using the dimension of the corresponding irreps tight?
- Among other tasks, transformers are widely used in NLP, is NL permutation symmetric to any degree?

The first result involves again a non-linear MLP with erf activation function and linearized MHA. The NN is trained on $8,000$ samples drawn from the mixture $p, q \sim U(0, 0.4)$ with SGD with a mini-batch size of 50 and a learning rate of 10^{-3} for $10,000$ epochs. In figure 3 left, the output of the NN, together with exact Bayesian inference with the NNGP prior, the optimal predictor, and the ground truth, are shown as a function of the ground truth target value; each point on the plot can be thought of as correlation between the output and the true value. The performance OOD only starts to decline slightly around 0.8, matching well with the exact GP predictions despite the finite size and SGD training. It remains to be seen whether these results hold for a large transformer.

In the center panel of Fig. 3 we see the spectrum of the kernel matrix as defined in equation 4, for a NN with $\Phi = \mathrm{softmax}$ and $\phi(x) = x$. The scaling of the eigenvalues is improved such that the eigenvalues in the trivial irrep scale as L^0 and the eigenvalues in the standard irrep scale as L^{-1}, meaning, they take the maximum scaling possible based on the degeneracy of the irrep. Comparing these values to those of a fully connected network (FCN) can give a view from infinity on the advantages transformers have over FCNs. For our dataset and task, the target function belongs to the trivial representation of the symmetric groups, this means that the network with $\Phi = \mathrm{softmax}$ can learn the target in $N = \Theta(1)$ samples. On the other hand, an FCN will need $N = O(L)$ samples, as its symmetry group would be the sign-symmetric group, where all L linear functions are degenerate (Yang & Salman, 2020). Such a result extends further to settings used to study in-context learning (Von Oswald et al., 2023), as they too aggregate symmetrically over context, allowing transformers to gain significant advantage over FCNs.

Finally, we present some evidence suggesting NL does possess an approximate permutation symmetry, at least up to linear correlations. We examine the (first order) correlations in WikiText-2 at the basis of the cyclic permutation irreps (for experimental details see appendix H)

$$C_{ij}^{kk'} := \mathbb{E}_{X \sim \mathrm{WikiText-2}} \left[X_i^a V^{ak} X_j^b V^{bk'} \right]; \tag{16}$$

$$V^{ak} := \exp\left(i \frac{2\pi}{L} ak \right), \quad \begin{matrix} a = 1, ..., L \\ k = 0, ..., L-1 \end{matrix}. \tag{17}$$

If permutation symmetry were to hold, we would expect all C^{kk} correlation matrices with $k \neq 0$ to be interchangeable, as they are all part of the standard irrep. We quantify this quality by the cosine similarity induced by Frobenius inner product and by their spectrum. As shown in Fig. 3 right, there is indeed a large similarity in the standard irrep. This similarity does not exist with the trivial irrep ($k = 0$). The spectrum of the different correlation matrices inside the standard irrep is almost identical as well, as indicated by the eigenvalue CDF in the same figure. This similarity, again, does not exist between the two irreps (i.e. $k = 0, k \neq 0$).

4. Discussion

In this work, we analyzed a family of transformer-like models and showed that their inductive bias can be understood using the representation theory of the symmetric group when the dataset possesses permutation symmetry. In this setting, we derived a scaling law for the number of data samples required to learn a target as a function of the context length. We showed these predictions are generic and hold with added non-linear activation functions, as well as OOD.

Critically, the above results depend on a permutation symmetric dataset[9] while some settings do have this exact symmetry like transformers for sets (Lee et al., 2019; Kim et al., 2021) or setting studied for ICL (Power et al., 2022; Von Oswald et al., 2023), natural language does not seem to have it prima facie. We have shown that, in fact, first-order correlations in WikiText-2 seem to largely manifest this symmetry. This means that when learning linear targets or up to $O(L)$ samples, such models will be bound by the scaling laws discussed above. One such linear function (in the context tokens) that is relevant to NLP is the copying heads discussed in Olsson et al. (2022), while induction heads would be second order in the context tokens. This fact motivates examining the correlations in NL to second order, as a concrete mechanism for in-context learning can already appear there; we leave this for future work.

Notwithstanding, *some* conclusions can be carried from a symmetric dataset to an arbitrary one. Assume the target is one of the eigenfunctions of the kernel when diagonalized on a permutation symmetric measure $p(x)$, namely $y(x) = \varphi_t(x)$, we may substitute into equation 8 a predictor $f(x)$ attained by inference with the same GP on a different, possibly non-symmetric dataset $D_N = \{x_\mu, y(x_\mu)\}_{\mu=1}^N$. The learnability with the new $f(x)$ can nevertheless be bounded (Lavie & Ringel, 2024)

$$L_t \leq \frac{\lambda_t N \mathbb{E}_{x\sim D_N}[\phi_t^2(x)]}{\sigma^2}. \tag{18}$$

suggesting $N \approx \lambda_t^{-1}\sigma^2/\mathbb{E}_{x\sim D_N}[\phi_t^2(x)]$ samples will be required to learn ϕ_t from D_N when measuring learnability on $p(x)$.

Lastly, while our work accounts for the implicit inductive bias of the architecture, it does not address other sources of inductive bias, like finite learning rate (Lewkowycz et al., 2020; Beugnot et al., 2022; Mohtashami et al., 2023) and finite size corrections to the GP limit. As recent works have shown (Seroussi et al., 2023; Bordelon & Pehlevan, 2023; Segadlo et al., 2022; Li & Sompolinsky, 2021; Pacelli et al., 2023; Dyer & Gur-Ari, 2019; Aitchison et al., 2021), the GP limit is used as a starting point for more advanced methods that study finite size corrections and capture important phenomena like representation learning. Studying such corrections is left to future work.

[9]Though the solved linear example shows permutation symmetry between subsets in the sequence is also powerful

4.1. Related Works

Inductive bias. The term spectral bias was coined by Rahaman et al. (2019), referring specifically to the Fourier spectrum. The term has since been used to describe other forms of inductive bias, including those coming from the spectrum of the corresponding kernel (eigenvalue decomposition) (Bordelon et al., 2020; Cao et al., 2021; Canatar et al., 2021b). The form of inductive bias studied in this paper has been extensively studied for fully connected networks (FCNs): for data uniformly distributed on a hypersphere (Basri et al., 2019; Bietti & Bach, 2021; Azevedo & Menegatto, 2014; Scetbon & Harchaoui, 2021) and non-uniformly (Basri et al., 2020), and for Gaussian and uniform distribution on the hypercube (Yang & Salman, 2020). Bietti et al. (2021) have studied how additional symmetries above the rotational symmetry of FCNs can reduce sample complexity and Tahmasebi & Jegelka (2023) generalized these results. In the context of CNNs symmetries were used to design better inductive bias (see Cohen & Welling (2016)), and their corresponding kernels' spectra were analyzed as well (Bietti, 2021; Xiao, 2022; Cagnetta et al., 2023; Geifman et al., 2022) showing how their inductive bias is beneficial for learning a range of tasks. To the best of our knowledge, our work is the first to study inductive bias of transformers from the perspective of symmetries in the infinite width (GP) limit. Recently an empirical work has characterized some aspects of the inductive bias of transformers (Bhattamishra et al., 2023); our work complements it and can be seen as a theoretical explanation. Lastly, Fu et al. (2023) has studied random feature attention under limiting assumptions (no positional encoding, limited depth); our work approaches the challenge from a the perspective of symmetry & representation theory and its results are valid for a more general family of models.

Understanding transformers. A large body of knowledge has accumulated regarding understanding specific phenomena related to transformers like grokking (Liu et al., 2022a;b; Rubin et al., 2023; Nanda et al., 2023), in-context learning (Olsson et al., 2022; Garg et al., 2022; Akyürek et al., 2022; Von Oswald et al., 2023). Other works focused on other aspects like signal propagation (Cowsik et al., 2024; Noci et al., 2022) and OOD generalization (Nam et al., 2022). Our work focuses directly on the transformers' inductive bias and characterizes it at the GP limit.

Acknowledgements

Z.R. and I.L. acknowledge support from ISF Grant 2250/19.

Impact Statement

This paper presents work whose goal is to advance the field of Machine Learning. There are many potential societal consequences of our work, none which we feel must be specifically highlighted here.

References

Ahn, K., Cheng, X., Song, M., Yun, C., Jadbabaie, A., and Sra, S. Linear attention is (maybe) all you need (to understand transformer optimization), March 2024. URL http://arxiv.org/abs/2310.01082. arXiv:2310.01082 [cs, math].

Aitchison, L., Yang, A., and Ober, S. W. Deep kernel processes. In *Proceedings of the 38th International Conference on Machine Learning*, pp. 130–140. PMLR, July 2021. URL https://proceedings.mlr.press/v139/aitchison21a.html. ISSN: 2640-3498.

Akemann, G., Baik, J., and Di Francesco, P. *The Oxford Handbook of Random Matrix Theory*. Oxford University Press, September 2015. ISBN 978-0-19-874419-1. doi: 10.1093/oxfordhb/9780198744191.001.0001. URL https://doi.org/10.1093/oxfordhb/9780198744191.001.0001.

Akyürek, E., Schuurmans, D., Andreas, J., Ma, T., and Zhou, D. What learning algorithm is in-context learning? Investigations with linear models. September 2022. URL https://openreview.net/forum?id=0g0X4H8yN4I.

Azevedo, D. and Menegatto, V. A. Sharp estimates for eigenvalues of integral operators generated by dot product kernels on the sphere. *Journal of Approximation Theory*, 177:57–68, January 2014. ISSN 0021-9045. doi: 10.1016/j.jat.2013.10.002. URL https://www.sciencedirect.com/science/article/pii/S0021904513001652.

Basri, R., Jacobs, D., Kasten, Y., and Kritchman, S. The Convergence Rate of Neural Networks for Learned Functions of Different Frequencies. In *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019. URL https://proceedings.neurips.cc/paper_files/paper/2019/hash/5ac8bb8a7d745102a978c5f8ccdb61b8-Abstract.html.

Basri, R., Galun, M., Geifman, A., Jacobs, D., Kasten, Y., and Kritchman, S. Frequency Bias in Neural Networks for Input of Non-Uniform Density. In *Proceedings of the 37th International Conference on Machine Learning*, pp. 685–694. PMLR, November 2020. URL https://proceedings.mlr.press/v119/basri20a.html. ISSN: 2640-3498.

Baum, L. E. and Petrie, T. Statistical Inference for Probabilistic Functions of Finite State Markov Chains. *The Annals of Mathematical Statistics*, 37(6):1554–1563, 1966. ISSN 0003-4851. URL https://www.jstor.org/stable/2238772. Publisher: Institute of Mathematical Statistics.

Beugnot, G., Mairal, J., and Rudi, A. On the Benefits of Large Learning Rates for Kernel Methods. In *Proceedings of Thirty Fifth Conference on Learning Theory*, pp. 254–282. PMLR, June 2022. URL https://proceedings.mlr.press/v178/beugnot22a.html. ISSN: 2640-3498.

Bhattamishra, S., Patel, A., Kanade, V., and Blunsom, P. Simplicity Bias in Transformers and their Ability to Learn Sparse Boolean Functions. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 5767–5791, Toronto, Canada, 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.acl-long.317. URL https://aclanthology.org/2023.acl-long.317.

Bietti, A. Approximation and Learning with Deep Convolutional Models: a Kernel Perspective. October 2021. URL https://openreview.net/forum?id=lrocYB-0ST2.

Bietti, A. and Bach, F. Deep Equals Shallow for ReLU Networks in Kernel Regimes, August 2021. URL http://arxiv.org/abs/2009.14397. arXiv:2009.14397 [cs, stat].

Bietti, A., Venturi, L., and Bruna, J. On the Sample Complexity of Learning under Invariance and Geometric Stability, November 2021. URL http://arxiv.org/abs/2106.07148. arXiv:2106.07148 [cs, stat].

Bommasani, R., Hudson, D. A., Adeli, E., Altman, R., Arora, S., von Arx, S., Bernstein, M. S., Bohg, J., Bosselut, A., Brunskill, E., Brynjolfsson, E., Buch, S., Card, D., Castellon, R., Chatterji, N. S., Chen, A. S., Creel, K. A., Davis, J., Demszky, D., Donahue, C., Doumbouya, M., Durmus, E., Ermon, S., Etchemendy, J., Ethayarajh, K., Fei-Fei, L., Finn, C., Gale, T., Gillespie, L. E., Goel, K., Goodman, N. D., Grossman, S., Guha, N., Hashimoto, T., Henderson, P., Hewitt, J., Ho, D. E., Hong, J., Hsu, K., Huang, J., Icard, T. F., Jain, S., Jurafsky, D., Kalluri,

P., Karamcheti, S., Keeling, G., Khani, F., Khattab, O., Koh, P. W., Krass, M. S., Krishna, R., Kuditipudi, R., Kumar, A., Ladhak, F., Lee, M., Lee, T., Leskovec, J., Levent, I., Li, X. L., Li, X., Ma, T., Malik, A., Manning, C. D., Mirchandani, S. P., Mitchell, E., Munyikwa, Z., Nair, S., Narayan, A., Narayanan, D., Newman, B., Nie, A., Niebles, J. C., Nilforoshan, H., Nyarko, J. F., Ogut, G., Orr, L., Papadimitriou, I., Park, J. S., Piech, C., Portelance, E., Potts, C., Raghunathan, A., Reich, R., Ren, H., Rong, F., Roohani, Y. H., Ruiz, C., Ryan, J., R'e, C., Sadigh, D., Sagawa, S., Santhanam, K., Shih, A., Srinivasan, K. P., Tamkin, A., Taori, R., Thomas, A. W., Tramèr, F., Wang, R. E., Wang, W., Wu, B., Wu, J., Wu, Y., Xie, S. M., Yasunaga, M., You, J., Zaharia, M. A., Zhang, M., Zhang, T., Zhang, X., Zhang, Y., Zheng, L., Zhou, K., and Liang, P. On the opportunities and risks of foundation models. *ArXiv*, 2021. URL https://crfm.stanford.edu/assets/report.pdf.

Bordelon, B. and Pehlevan, C. Self-consistent dynamical field theory of kernel evolution in wide neural networks*. *Journal of Statistical Mechanics: Theory and Experiment*, 2023(11):114009, November 2023. ISSN 1742-5468. doi: 10.1088/1742-5468/ad01b0. URL https://dx.doi.org/10.1088/1742-5468/ad01b0. Publisher: IOP Publishing.

Bordelon, B., Canatar, A., and Pehlevan, C. Spectrum Dependent Learning Curves in Kernel Regression and Wide Neural Networks. In *Proceedings of the 37th International Conference on Machine Learning*, pp. 1024–1034. PMLR, November 2020. URL https://proceedings.mlr.press/v119/bordelon20a.html. ISSN: 2640-3498.

Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D., Wu, J., Winter, C., Hesse, C., Chen, M., Sigler, E., Litwin, M., Gray, S., Chess, B., Clark, J., Berner, C., McCandlish, S., Radford, A., Sutskever, I., and Amodei, D. Language Models are Few-Shot Learners. In *Advances in Neural Information Processing Systems*, volume 33, pp. 1877–1901. Curran Associates, Inc., 2020. URL https://papers.nips.cc/paper/2020/hash/1457c0d6bfcb4967418bfb8ac142f64a-Abstract.html.

Cagnetta, F., Favero, A., and Wyart, M. What Can Be Learnt With Wide Convolutional Neural Networks? In *Proceedings of the 40th International Conference on Machine Learning*, pp. 3347–3379. PMLR, July 2023. URL https://proceedings.mlr.press/v202/cagnetta23a.html. ISSN: 2640-3498.

Canatar, A., Bordelon, B., and Pehlevan, C. Out-of-Distribution Generalization in Kernel Regression. In *Advances in Neural Information Processing Systems*, November 2021a. URL https://openreview.net/forum?id=-h6Ldc0MO-.

Canatar, A., Bordelon, B., and Pehlevan, C. Spectral bias and task-model alignment explain generalization in kernel regression and infinitely wide neural networks. *Nature Communications*, 12(1):2914, May 2021b. ISSN 2041-1723. doi: 10.1038/s41467-021-23103-1. URL https://www.nature.com/articles/s41467-021-23103-1. Number: 1 Publisher: Nature Publishing Group.

Cao, Y., Fang, Z., Wu, Y., Zhou, D.-X., and Gu, Q. Towards Understanding the Spectral Bias of Deep Learning. volume 3, pp. 2205–2211, August 2021. doi: 10.24963/ijcai.2021/304. URL https://www.ijcai.org/proceedings/2021/304. ISSN: 1045-0823.

Chen, M., Radford, A., Child, R., Wu, J., Jun, H., Luan, D., and Sutskever, I. Generative Pretraining From Pixels. In *Proceedings of the 37th International Conference on Machine Learning*, pp. 1691–1703. PMLR, November 2020. URL https://proceedings.mlr.press/v119/chen20s.html. ISSN: 2640-3498.

Cohen, O., Malka, O., and Ringel, Z. Learning curves for overparametrized deep neural networks: A field theory perspective. *Physical Review Research*, 3(2):023034, April 2021. doi: 10.1103/PhysRevResearch.3.023034. URL https://link.aps.org/doi/10.1103/PhysRevResearch.3.023034. Publisher: American Physical Society.

Cohen, T. S. and Welling, M. Group Equivariant Convolutional Networks, June 2016. URL http://arxiv.org/abs/1602.07576. arXiv:1602.07576 [cs, stat].

Cowsik, A., Nebabu, T., Qi, X.-L., and Ganguli, S. Geometric Dynamics of Signal Propagation Predict Trainability of Transformers, March 2024. URL http://arxiv.org/abs/2403.02579. arXiv:2403.02579 [cond-mat].

Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., and Houlsby, N. An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale, June 2021. URL http://arxiv.org/abs/2010.11929. arXiv:2010.11929 [cs].

Dyer, E. and Gur-Ari, G. Asymptotics of Wide Networks from Feynman Diagrams. In *International Conference on Learning Representations*, September

2019. URL https://openreview.net/forum?id=S1gFvANKDS.

EuropeanCommission. Proposal for a REGULATION OF THE EUROPEAN PARLIAMENT AND OF THE COUNCIL LAYING DOWN HARMONISED RULES ON ARTIFICIAL INTELLIGENCE (ARTIFICIAL INTELLIGENCE ACT) AND AMENDING CERTAIN UNION LEGISLATIVE ACTS, 2021. URL https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=celex%3A52021PC0206.

Fu, H., Guo, T., Bai, Y., and Mei, S. What can a Single Attention Layer Learn? A Study Through the Random Features Lens. *Advances in Neural Information Processing Systems*, 36:11912–11951, December 2023.

Fulton, W. and Harris, J. *Representation Theory: A First Course*, volume 129 of *Graduate Texts in Mathematics*. Springer, New York, NY, 2004. ISBN 978-3-540-00539-1 978-1-4612-0979-9. doi: 10.1007/978-1-4612-0979-9. URL http://link.springer.com/10.1007/978-1-4612-0979-9.

Garg, S., Tsipras, D., Liang, P., and Valiant, G. What Can Transformers Learn In-Context? A Case Study of Simple Function Classes. In *Advances in Neural Information Processing Systems*, May 2022. URL https://openreview.net/forum?id=flNZJ2eOet.

Geifman, A., Galun, M., Jacobs, D., and Ronen, B. On the Spectral Bias of Convolutional Neural Tangent and Gaussian Process Kernels. *Advances in Neural Information Processing Systems*, 35:11253–11265, December 2022. URL https://proceedings.neurips.cc/paper_files/paper/2022/hash/48fd58527b29c5c0ef2cae43065636e6-Abstract-Conference.html.

GeminiTeam. Gemini: A Family of Highly Capable Multimodal Models, December 2023. URL http://arxiv.org/abs/2312.11805. arXiv:2312.11805 [cs].

Goyal, A. and Bengio, Y. Inductive biases for deep learning of higher-level cognition. *Proceedings of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 478(2266):20210068, October 2022. doi: 10.1098/rspa.2021.0068. URL https://royalsocietypublishing.org/doi/full/10.1098/rspa.2021.0068. Publisher: Royal Society.

Henighan, T., Kaplan, J., Katz, M., Chen, M., Hesse, C., Jackson, J., Jun, H., Brown, T. B., Dhariwal, P., Gray, S., Hallacy, C., Mann, B., Radford, A., Ramesh, A., Ryder, N., Ziegler, D. M., Schulman, J., Amodei, D., and McCandlish, S. Scaling Laws for Autoregressive Generative Modeling, November 2020. URL http://arxiv.org/abs/2010.14701. arXiv:2010.14701 [cs].

Hron, J., Bahri, Y., Sohl-Dickstein, J., and Novak, R. Infinite attention: NNGP and NTK for deep attention networks, June 2020. URL http://arxiv.org/abs/2006.10540. arXiv:2006.10540 [cs, stat].

Jacot, A., Gabriel, F., and Hongler, C. Neural Tangent Kernel: Convergence and Generalization in Neural Networks. In *Advances in Neural Information Processing Systems*, volume 31. Curran Associates, Inc., 2018.

Jiang, A. Q., Sablayrolles, A., Roux, A., Mensch, A., Savary, B., Bamford, C., Chaplot, D. S., Casas, D. d. l., Hanna, E. B., Bressand, F., Lengyel, G., Bour, G., Lample, G., Lavaud, L. R., Saulnier, L., Lachaux, M.-A., Stock, P., Subramanian, S., Yang, S., Antoniak, S., Scao, T. L., Gervet, T., Lavril, T., Wang, T., Lacroix, T., and Sayed, W. E. Mixtral of Experts, January 2024. URL http://arxiv.org/abs/2401.04088. arXiv:2401.04088 [cs].

Kaplan, J., McCandlish, S., Henighan, T., Brown, T. B., Chess, B., Child, R., Gray, S., Radford, A., Wu, J., and Amodei, D. Scaling Laws for Neural Language Models, January 2020. URL http://arxiv.org/abs/2001.08361. arXiv:2001.08361 [cs, stat].

Kim, J., Oh, S., and Hong, S. Transformers Generalize DeepSets and Can be Extended to Graphs & Hypergraphs. In *Advances in Neural Information Processing Systems*, volume 34, pp. 28016–28028. Curran Associates, Inc., 2021.

König, H. *Eigenvalue Distribution of Compact Operators*, volume 16 of *Operator Theory: Advances and Applications*. Birkhäuser, Basel, 1986. ISBN 978-3-0348-6280-6 978-3-0348-6278-3. doi: 10.1007/978-3-0348-6278-3. URL http://link.springer.com/10.1007/978-3-0348-6278-3.

Lavie, I. and Ringel, Z. Symmetric kernels with non-symmetric data: A data-agnostic learnability bound. 2024.

Lee, J., Bahri, Y., Novak, R., Schoenholz, S. S., Pennington, J., and Sohl-Dickstein, J. Deep Neural Networks as Gaussian Processes. February 2018. URL https://openreview.net/forum?id=B1EA-M-0Z.

Lee, J., Lee, Y., Kim, J., Kosiorek, A., Choi, S., and Teh, Y. W. Set Transformer: A Framework for Attention-based Permutation-Invariant Neural Networks. In *Proceedings of the 36th International Conference on Machine Learning*, pp. 3744–3753. PMLR, May 2019.

URL https://proceedings.mlr.press/v97/lee19d.html. ISSN: 2640-3498.

Lee, J., Schoenholz, S. S., Pennington, J., Adlam, B., Xiao, L., Novak, R., and Sohl-Dickstein, J. Finite Versus Infinite Neural Networks: an Empirical Study, September 2020. URL http://arxiv.org/abs/2007.15801. arXiv:2007.15801 [cs, stat].

Lewkowycz, A., Bahri, Y., Dyer, E., Sohl-Dickstein, J., and Gur-Ari, G. The large learning rate phase of deep learning: the catapult mechanism, March 2020. URL http://arxiv.org/abs/2003.02218. arXiv:2003.02218 [cs, stat].

Li, Q. and Sompolinsky, H. Statistical Mechanics of Deep Linear Neural Networks: The Backpropagating Kernel Renormalization. *Physical Review X*, 11(3): 031059, September 2021. doi: 10.1103/PhysRevX.11.031059. URL https://link.aps.org/doi/10.1103/PhysRevX.11.031059. Publisher: American Physical Society.

Liu, K., Ziyin, L., and Ueda, M. Noise and Fluctuation of Finite Learning Rate Stochastic Gradient Descent, June 2021. URL http://arxiv.org/abs/2012.03636. arXiv:2012.03636 [cs, stat].

Liu, Z., Kitouni, O., Nolte, N. S., Michaud, E., Tegmark, M., and Williams, M. Towards Understanding Grokking: An Effective Theory of Representation Learning. *Advances in Neural Information Processing Systems*, 35:34651–34663, December 2022a. URL https://proceedings.neurips.cc/paper_files/paper/2022/hash/dfc310e81992d2e4cedc09ac47eff13e-Abstract-Conference.html.

Liu, Z., Michaud, E. J., and Tegmark, M. Omnigrok: Grokking Beyond Algorithmic Data. In *The Eleventh International Conference on Learning Representations*, September 2022b. URL https://openreview.net/forum?id=zDiHoIWa0q1.

Mandt, S., Hoffman, M. D., and Blei, D. M. Stochastic Gradient Descent as Approximate Bayesian Inference, January 2018. URL http://arxiv.org/abs/1704.04289. arXiv:1704.04289 [cs, stat].

Merity, S., Xiong, C., Bradbury, J., and Socher, R. Pointer Sentinel Mixture Models. In *International Conference on Learning Representations*, November 2016. URL https://openreview.net/forum?id=Byj72udxe.

Min, E., Chen, R., Bian, Y., Xu, T., Zhao, K., Huang, W., Zhao, P., Huang, J., Ananiadou, S., and Rong, Y. Transformer for Graphs: An Overview from Architecture Perspective, February 2022. URL http://arxiv.org/abs/2202.08455. arXiv:2202.08455 [cs].

Mingard, C., Valle-Pérez, G., Skalse, J., and Louis, A. A. Is SGD a Bayesian sampler? Well, almost, October 2020. URL http://arxiv.org/abs/2006.15191. arXiv:2006.15191 [cs, stat].

Mohtashami, A., Jaggi, M., and Stich, S. Special Properties of Gradient Descent with Large Learning Rates, February 2023. URL http://arxiv.org/abs/2205.15142. arXiv:2205.15142 [cs, math].

Nam, A. J., Abdool, M., Maxfield, T., and McClelland, J. L. Achieving and Understanding Out-of-Distribution Generalization in Systematic Reasoning in Small-Scale Transformers, December 2022. URL http://arxiv.org/abs/2210.03275. arXiv:2210.03275 [cs].

Nanda, N., Chan, L., Lieberum, T., Smith, J., and Steinhardt, J. Progress measures for grokking via mechanistic interpretability, October 2023. URL http://arxiv.org/abs/2301.05217. arXiv:2301.05217 [cs].

Naveh, G., Ben David, O., Sompolinsky, H., and Ringel, Z. Predicting the outputs of finite deep neural networks trained with noisy gradients. *Physical Review E*, 104(6): 064301, December 2021. doi: 10.1103/PhysRevE.104.064301. URL https://link.aps.org/doi/10.1103/PhysRevE.104.064301. Publisher: American Physical Society.

Neal, R. M. Probabilistic Inference Using Markov Chain Monte Carlo Methods, 1993.

Neal, R. M. Priors for Infinite Networks. In Neal, R. M. (ed.), *Bayesian Learning for Neural Networks*, Lecture Notes in Statistics, pp. 29–53. Springer, New York, NY, 1996. ISBN 978-1-4612-0745-0. doi: 10.1007/978-1-4612-0745-0_2. URL https://doi.org/10.1007/978-1-4612-0745-0_2.

Noci, L., Anagnostidis, S., Biggio, L., Orvieto, A., Singh, S. P., and Lucchi, A. Signal Propagation in Transformers: Theoretical Perspectives and the Role of Rank Collapse. *Advances in Neural Information Processing Systems*, 35: 27198–27211, December 2022.

Novak, R., Xiao, L., Bahri, Y., Lee, J., Yang, G., Hron, J., Abolafia, D. A., Pennington, J., and Sohl-dickstein, J. Bayesian Deep Convolutional Networks with Many Channels are Gaussian Processes. September 2018. URL https://openreview.net/forum?id=B1g30j0qF7.

Olsson, C., Elhage, N., Nanda, N., Joseph, N., DasSarma, N., Henighan, T., Mann, B., Askell, A., Bai, Y., Chen, A., Conerly, T., Drain, D., Ganguli, D., Hatfield-Dodds, Z., Hernandez, D., Johnston, S., Jones, A., Kernion, J., Lovitt, L., Ndousse, K., Amodei, D., Brown, T., Clark, J., Kaplan, J., McCandlish, S., and Olah, C. In-context Learning and Induction Heads, September 2022. URL http://arxiv.org/abs/2209.11895. arXiv:2209.11895 [cs].

OpenAI. GPT-4 Technical Report, March 2023. URL http://arxiv.org/abs/2303.08774. arXiv:2303.08774 [cs].

Pacelli, R., Ariosto, S., Pastore, M., Ginelli, F., Gherardi, M., and Rotondo, P. A statistical mechanics framework for Bayesian deep neural networks beyond the infinite-width limit. *Nature Machine Intelligence*, 5(12):1497–1507, December 2023. ISSN 2522-5839. doi: 10.1038/s42256-023-00767-6. URL https://www.nature.com/articles/s42256-023-00767-6. Number: 12 Publisher: Nature Publishing Group.

Potters, M. and Bouchaud, J.-P. *A First Course in Random Matrix Theory: for Physicists, Engineers and Data Scientists*. Cambridge University Press, Cambridge, 2020. ISBN 978-1-108-48808-2. doi: 10.1017/9781108768900. URL https://www.cambridge.org/core/books/first-course-in-random-matrix-theory/2292A554A9BB9E2A4697C35BCE920304.

Power, A., Burda, Y., Edwards, H., Babuschkin, I., and Misra, V. Grokking: Generalization Beyond Overfitting on Small Algorithmic Datasets, January 2022. URL http://arxiv.org/abs/2201.02177. arXiv:2201.02177 [cs].

Rahaman, N., Baratin, A., Arpit, D., Draxler, F., Lin, M., Hamprecht, F., Bengio, Y., and Courville, A. On the Spectral Bias of Neural Networks. In *Proceedings of the 36th International Conference on Machine Learning*, pp. 5301–5310. PMLR, May 2019. URL https://proceedings.mlr.press/v97/rahaman19a.html. ISSN: 2640-3498.

Rubin, N., Seroussi, I., and Ringel, Z. Droplets of Good Representations: Grokking as a First Order Phase Transition in Two Layer Networks, October 2023. URL http://arxiv.org/abs/2310.03789. arXiv:2310.03789 [cond-mat, stat].

Sagan, B. E. *The Symmetric Group*, volume 203 of *Graduate Texts in Mathematics*. Springer, New York, NY, 2001. ISBN 978-1-4419-2869-6 978-1-4757-6804-6. doi: 10.1007/978-1-4757-6804-6. URL http://link.springer.com/10.1007/978-1-4757-6804-6.

Scetbon, M. and Harchaoui, Z. A Spectral Analysis of Dot-product Kernels. In *Proceedings of The 24th International Conference on Artificial Intelligence and Statistics*, pp. 3394–3402. PMLR, March 2021. URL https://proceedings.mlr.press/v130/scetbon21b.html. ISSN: 2640-3498.

Schölkopf, B., Herbrich, R., and Smola, A. J. A Generalized Representer Theorem. In Helmbold, D. and Williamson, B. (eds.), *Computational Learning Theory*, pp. 416–426, Berlin, Heidelberg, 2001. Springer. ISBN 978-3-540-44581-4. doi: 10.1007/3-540-44581-1_27.

Segadlo, K., Epping, B., Meegen, A. v., Dahmen, D., Krämer, M., and Helias, M. Unified field theoretical approach to deep and recurrent neuronal networks. *Journal of Statistical Mechanics: Theory and Experiment*, 2022(10):103401, October 2022. ISSN 1742-5468. doi: 10.1088/1742-5468/ac8e57. URL https://dx.doi.org/10.1088/1742-5468/ac8e57. Publisher: IOP Publishing and SISSA.

Seroussi, I., Naveh, G., and Ringel, Z. Separation of scales and a thermodynamic description of feature learning in some CNNs. *Nature Communications*, 14(1):908, February 2023. ISSN 2041-1723. doi: 10.1038/s41467-023-36361-y. URL https://www.nature.com/articles/s41467-023-36361-y. Number: 1 Publisher: Nature Publishing Group.

Shankar, V., Fang, A., Guo, W., Fridovich-Keil, S., Schmidt, L., Ragan-Kelley, J., and Recht, B. Neural Kernels Without Tangents, March 2020. URL http://arxiv.org/abs/2003.02237. arXiv:2003.02237 [cs, stat].

Silverman, B. W. Spline Smoothing: The Equivalent Variable Kernel Method. *The Annals of Statistics*, 12(3):898–916, 1984. ISSN 0090-5364. URL https://www.jstor.org/stable/2240968. Publisher: Institute of Mathematical Statistics.

Simon, J. B., Dickens, M., Karkada, D., and Deweese, M. The Eigenlearning Framework: A Conservation Law Perspective on Kernel Ridge Regression and Wide Neural Networks. *Transactions on Machine Learning Research*, February 2023. ISSN 2835-8856. URL https://openreview.net/forum?id=FDbQGCAViI.

Sollich, P. and Williams, C. Using the Equivalent Kernel to Understand Gaussian Process Regression. In *Advances in Neural Information Processing Systems*, volume 17. MIT Press, 2004.

Tahmasebi, B. and Jegelka, S. The Exact Sample Complexity Gain from Invariances for Kernel Regression. *Advances in Neural Information Processing Systems*, 36: 55616–55646, December 2023.

Tung, W.-K. *Group Theory in Physics: An Introduction to Symmetry Principles, Group Representations, and Special Functions in Classical and Quantum Physics*. WORLD SCIENTIFIC, August 1985. ISBN 978-9971-966-57-7 978-981-238-498-0. doi: 10.1142/0097. URL http://www.worldscientific.com/worldscibooks/10.1142/0097.

Von Oswald, J., Niklasson, E., Randazzo, E., Sacramento, J., Mordvintsev, A., Zhmoginov, A., and Vladymyrov, M. Transformers Learn In-Context by Gradient Descent. In *Proceedings of the 40th International Conference on Machine Learning*, pp. 35151–35174. PMLR, July 2023. URL https://proceedings.mlr.press/v202/von-oswald23a.html. ISSN: 2640-3498.

Welling, M. and Teh, Y. W. Bayesian learning via stochastic gradient langevin dynamics. In *Proceedings of the 28th International Conference on International Conference on Machine Learning*, ICML'11, pp. 681–688, Madison, WI, USA, June 2011. Omnipress. ISBN 978-1-4503-0619-5.

Wen, K., Li, Y., Liu, B., and Risteski, A. Transformers are uninterpretable with myopic methods: a case study with bounded Dyck grammars, December 2023. URL http://arxiv.org/abs/2312.01429. arXiv:2312.01429 [cs, stat].

Wolf, T., Debut, L., Sanh, V., Chaumond, J., Delangue, C., Moi, A., Cistac, P., Rault, T., Louf, R., Funtowicz, M., Davison, J., Shleifer, S., von Platen, P., Ma, C., Jernite, Y., Plu, J., Xu, C., Scao, T. L., Gugger, S., Drame, M., Lhoest, Q., and Rush, A. M. HuggingFace's Transformers: State-of-the-art Natural Language Processing, July 2020. URL http://arxiv.org/abs/1910.03771. arXiv:1910.03771 [cs].

Xiao, L. Eigenspace Restructuring: A Principle of Space and Frequency in Neural Networks. In *Proceedings of Thirty Fifth Conference on Learning Theory*, pp. 4888–4944. PMLR, June 2022. URL https://proceedings.mlr.press/v178/xiao22a.html. ISSN: 2640-3498.

Xie, S. M., Raghunathan, A., Liang, P., and Ma, T. An Explanation of In-context Learning as Implicit Bayesian Inference. In *International Conference on Learning Representations*, October 2021. URL https://openreview.net/forum?id=RdJVFCHjUMI.

Yang, G. and Salman, H. A Fine-Grained Spectral Perspective on Neural Networks, April 2020. URL http://arxiv.org/abs/1907.10599. arXiv:1907.10599 [cs, stat].

A. Introduction to Key Concepts in Representation Theory for Eigenvalue Problems

Symmetries can greatly simplify the above eigenvalue problem. Let G be a symmetry group, we say the eigenvalue problem possesses this symmetry provided

$$\forall g \in G,\ k(T_g\, x, T_g\, y) = k(x, y) \qquad p_{\text{data}}(T_g\, x) = p_{\text{data}}(x) \tag{A.1}$$

where the linear transformations (T_g) are some faithful representation of G (i.e. $T_g T_{g'} = T_{gg'}$ and $T_g T_{g'} = Id$ iff gg' is the identity element of G).

As a concrete example and to make contact with the terminology in the main text, consider the case where $x \in \mathrm{R}^2$ which we express in polar coordinates $x = (r_x \cos(\theta_x), r_x \sin(\theta_x))$, and $p(x)$ effectively discretizes θ and fixes r (i.e. $p(x) = \delta(r_x - 1)N^{-1}\sum_{j=1}^{N}\delta(\theta_x - 2\pi j/N)$). Let $K(x, y) = ||x - y||$, $G = Z_N$ given by the rotation of x in units of $2\pi/N$, and T_g's given by the corresponding 2×2 rotation matrices on x.

Next we utilize G to find the spectrum of K w.r.t. $p(x)$. To this end, we consider the space on which $\hat{K}$ acts— the vector space of functions of x ($f(x)$) with the distance induced by $p(x)$. This space is N dimensional and spanned by $[f(x_1), ..., f(x_N)] \equiv \vec{f}$. The linear action of T_g on x induces a linear action on function space (equivalently on $\vec{f}$) via $\hat{T}_g \cdot f(x) = f(T_g x)$. Symmetry under G, as defined above, implies that $\hat{T}_g$'s all commute with $\hat{K}$. Consequently eigenspaces of $\hat{K}$ are invariant under all $\hat{T}_g$'s.

The above guides us to look for the minimal vector spaces which are invariant under all $\hat{T}_g$'s. These are known as irreducible representations (irreps). The group Z_N is known to have N distinct irreducible representations of dimension 1 labelled by $k \in \{2\pi/N, 4\pi/N, ..., 2\pi\}$. The corresponding invariant spaces are simply the discrete Fourier mode vectors $\vec{v}_k = [e^{2\pi i k/N}, e^{4\pi i k/N}, ..., 1]$. It can be checked that all $\hat{T}_g$'s leave each of these spaces/vectors invariant. This implies $\hat{K}$ is diagonal on the $\vec{v}_k$ basis. Allowing more complicated radial dependence, say by taking $p(x)$ with $\delta(r - 1)$ replaced by $\frac{1}{2}[\delta(r - 1) + \delta(r - 2)]$, the resulting blocks of $\hat{K}$ associated with each irrep would be 2×2. Equivalently stated each block would contain the irrep at multiplicity 2. Furthermore, for non-abelian G (e.g. augmenting Z_N with reflections), irreps of dimension larger than 1 generally appear.

B. A Gentle Introduction to The Use of Symmetry in Kernel Learning and The Symmetric Group

Spectral properties of kernels with respect to the data measure, provide a detailed description of the implicit bias of infinitely wide neural networks. However, diagonalizing a generic kernel operator on a generic measure is challenging. For fully connected networks and rotation symmetric datasets, this difficulty is largely lifted. In fact for uniform data on the hypersphere closed-form expressions for the spectrum and eigenfunctions exist (Cohen et al., 2021; Canatar et al., 2021b), the latter being hyperspherical harmonics. These results follow directly from studying the representation theory of the orthogonal group acting on multivariate polynomials.

For transformer models like the ones introduced above, the analog task is to find representations of the symmetric group acting on multivariate polynomials. Below we provide several concrete examples of such representations, flesh out their implications on spectral bias, and provide a road map for deriving higher representations.

As a starting point consider a kernel $K(x, y)$ where $x, y \in \mathcal{R}^d$ and some generic dataset measure $p(x)$. Let S_d denote the symmetric group (the group of all possible permutations) on $1, 2, ..., d$ where an element $\sigma \in S_d$ acts on x as $[\sigma x]_i = x_{\sigma(i)}$ (i.e. the natural action). Assuming $K(x, y) = K(\sigma x, \sigma y)$ and $p(x) = p(\sigma x)$ we wish to solve the following eigenvalue problem

$$\int dy p(y) K(x, y) \varphi_\lambda(y) = \lambda \varphi_\lambda(x) \tag{B.1}$$

to simplify the problem, let us assume that $k(x, y)$ contains powers of x and y only up to some finite degree q. In that case, any $\varphi_\lambda(x)$ with non-zero λ must be at most a q'th order multivariate polynomial.

To proceed with finding the spectrum and eigenfunctions, we first address the question of what are the irreducible representations of the symmetric group acting on finite degree polynomials. Irreducible representations (irreps) of the symmetric group are labelled by partitions of d which we denote by $(d_1, d_2, ..., d_k)$ such that $d_1 \geq d_2 \geq ... \geq d_k$ and

$\sum_k d_k = d$. These partitions are in one-to-one correspondence with Young Diagrams wherein one simply draws a row of boxes of length d_1, followed by a left aligned row of boxes of length d_2 etc...

Conveniently, there is a direct way of constructing an irrep from its Young diagram (see Fulton & Harris (2004)). As shown in theorem 2.1, particularly relevant here are Young diagrams of the form $(n-k,k)$. Considering those, the first step is finding all standard Young Tableaux associated with the Young diagram. Standard Young Tableaux are assignments of integers between $1..d$, with no repetitions, to the boxes of the Young Diagram such that all columns and rows are of increasing order. For instance, for the case $(d-2,2)$ and $d=6$ these would be

$$\begin{array}{|c|c|c|c|}\hline 1&3&5&6\\\hline 2&4\\\cline{1-2}\end{array}\quad\begin{array}{|c|c|c|c|}\hline 1&3&4&6\\\hline 2&5\\\cline{1-2}\end{array}\quad\begin{array}{|c|c|c|c|}\hline 1&3&4&5\\\hline 2&6\\\cline{1-2}\end{array}\quad\begin{array}{|c|c|c|c|}\hline 1&2&5&6\\\hline 3&4\\\cline{1-2}\end{array}\quad\begin{array}{|c|c|c|c|}\hline 1&2&4&6\\\hline 3&5\\\cline{1-2}\end{array}$$
$$\begin{array}{|c|c|c|c|}\hline 1&2&4&5\\\hline 3&6\\\cline{1-2}\end{array}\quad\begin{array}{|c|c|c|c|}\hline 1&2&3&6\\\hline 4&5\\\cline{1-2}\end{array}\quad\begin{array}{|c|c|c|c|}\hline 1&2&3&5\\\hline 4&6\\\cline{1-2}\end{array}\quad\begin{array}{|c|c|c|c|}\hline 1&2&3&4\\\hline 5&6\\\cline{1-2}\end{array} \tag{B.2}$$

An important observation here, true for any $(d-k,k)$, is that the lower row completely determines the upper one. Indeed the upper row must consist of all integers besides those in the lower row, arranged in strictly increasing order. We may thus denote such tableaux by their set of lower row integers $i_1, i_2, .., i_k$ (although some combinations may be disallowed). We next associated a monomial of the form $x_{i_1}x_{i_2}...x_{i_k}$ with each such standard Young Tableaux[10] [11]. To proceed with the construction we further consider the group of column permutations $C \subset S_d$ wherein we only allow switching of pairs along columns. We then construct the following polynomial element from the monomial

$$M^1(x)_{i_1..i_k} = \sum_{\sigma\in C} \text{sign}(\sigma) x_{\sigma i_1}..x_{\sigma i_k} \tag{B.3}$$

it then follows (see appendix C) that these k'th degree polynomials span the irreps $(d-k,k)$, where the action of S_d amounts to its natural action on the indices x. Notably this basis is typically not an orthonormal one. Furthermore, the same representation may appear with any power of x_i, namely $M^{(m)} = \sum_{\sigma\in C} \text{sign}(\sigma) x^m_{\sigma i_1}..x^m_{\sigma i_k}$, $m \in \mathbb{N}$, however for discrete measures some of these may collapse onto one another or to the trivial representation. For instance if $x_i \in \{+1,-1\}$,$M^{(2m)}$ is just a constant and $M^{(2m+1)} = M^{(1)}$.

One notable example of a $(d-k,k)$ representation is the standard representation $(d-1,1)$ equivalent to the natural action on

$$\text{Span}\{x_i - x_0\}_{i=1}^d \tag{B.4}$$

this representation is also equivalent to considering the discrete Fourier modes

$$\varphi_k(x) = \sum_j e^{i2\pi kj/d} x_j \quad k \in \{1,2,...,d-1\} \tag{B.5}$$

but omitting $\varphi_{k=0}(x)$ (the trivial representation). The different k numbers, via $e^{2\pi ik/d}$, can also be understood as one-dimensional-irreps of the cyclic group ($Z_n \subset S$).

Another relevant irrep is the trivial one, corresponding to symmetric (multivariate) polynomials. These are spanned by the Schur polynomials which are again in one-to-one correspondence with Young Diagrams, via however a different association than the one above. Up to an order of, say order 3, these are spanned by 1, $\sum_i x_i$,$\sum_{i=j} x_i x_j$, $\sum_{i\neq j} x_i x_j$, $\sum_{i=j=k} x_i x_j x_k$, $\sum_{i\neq j=k} x_i x_j x_k$, $\sum_{i\neq j\neq k} x_i x_j x_k$.

Another low dimensional representation is the sign representation of the symmetric group, associated with alternating polynomials (polynomials which are anti-symmetric with respect to exchanging any two variables). All such polynomials are of degree higher than that of the Vandermonde polynomial ($\pi_{1\leq i<j\leq d,n-d}(x_i - x_j)$), thus having a degree higher than $n-1+n-2+...+0 = n(n-1)/2$. Due to their high order they would not appear for any $q < d$. We conjecture that these would be exponentially suppressed in d for any NNGP or NTK kernel.

[10] Similar to the construction of Specht modules from Young tabloids(Fulton & Harris, 2004).

[11] In the next appendix, where we prove theorem 2.1 we take a different approach for the construction of the irreps of the Symmetric group. Here we effectively directly associate monomials with Young Tabloids, while in the next appendix, we use the Young symmetrizers as projectors to irrep spaces without the need for such a less formal, yet more intuitive, association between tabloids and monomials.

The above irreps and their associations with polynomials, facilitate the construction of low order polynomial representations. For instance, let us assume that $x_i \in \{+1, -1\}$ and consider all possible polynomials up to second order. These are spanned by three trivial representations (i.e. (d) partition/Young-Diagram)

$$1, \sum_{i=1}^{d} x_i, \sum_{1 \leq i < j \leq d, n-d} x_i x_j \tag{B.6}$$

two $d-1$ dimension standard representations ($(d-1, 1)$)

$$\varphi_k(x) \quad k \in \{1..d-1\} \tag{B.7}$$
$$\left(\sum_{i=1}^{d} x_i\right) \varphi_k(x) \quad k \in \{1..d-1\}$$

and one $(d-1)(d-2)/2 - 1$ dimension ($(d-2, 2)$) representation spanned by

$$\varphi_{ij}(x) = x_i x_j - x_0 x_j - x_i x_b + x_0 x_b \quad b = \min[\{k\}_{k=1}^{d} \backslash \{i, j\}], 1 < i < j \neq 3 \tag{B.8}$$

Given a measure ($p(x)$) which respects the symmetry, any two polynomials associated with distinct representation would be orthogonal. However, their normalization and the orthogonality relations within the same representations would vary based on the measure.

Turning to the spectrum, it then follows from standard representation theory arguments that a kernel with $q = 2$ has 6 generally distinct eigenvalues. Three generally-non-degenerate eigenvalues are associated with linear combinations of the 3 trivial representations. Two, generally distinct sets, of $d-1$-degenerate eigenvalues associated with the two linear combinations of the two standard representations. Last, one $(d-1)(d-2)/2 - 1$ degenerate eigenvalue associated with the $(d-2, 2)$ representations.

Finally, we note that the eigenfunctions associated with the two standard representations can mix in a limited manner. Following the assignment of k numbers (or equivalently eigenvalues with respect to the subgroup of S consisting of cyclic permutations of the indices), each basis element we used is also an irrep of the cyclic group. Hence two different values of k cannot be mixed. In addition, other elements in the permutation group are capable of shifting between these k values, hence the linear combinations are constant as a function of k. As the eigenfunctions associated with one of the $d-1$-degenerate eigenvalue can be written as $a\varphi_k + b(\sum_i x_i)\varphi_k$ where a, b are k independent coefficients. The corresponding eigenfunction associated with the other $d-1$-degenerate eigenvalues is simply the orthogonal one.

C. Decomposition of multilinear polynomials to irreps of the symmetric group

Definition C.1 (Partition)**.** A partition of n is an ordered set of positive integers $\lambda = (\lambda_1, \lambda_2, ..., \lambda_m)$ such that $\{\lambda_i\}_{i=1}^{m} \subset \mathbb{N}$, $\sum_{i=1}^{m} \lambda_i = n$ and $\lambda_1 \geq \lambda_2 \geq ... \geq \lambda_m \geq 1$.

Theorem C.2. *Irreps of the symmetric group of n symbols S_n are uniquely labeled by partitions of n (Fulton & Harris, 2004)*

Definition C.3 (Young Diagram)**.** A Young diagram Θ_λ of a partition λ of n is a diagram where one draws a row of λ_i boxes for each element in lambda starting with λ_1, with each subsequent element below it. For example given the partition $\lambda = (3, 2, 1)$ the Young diagram is

$$\Theta_\lambda = \begin{array}{|c|c|c|}\hline & & \\ \hline & \\ \cline{1-2} \\ \cline{1-1}\end{array} \tag{C.1}$$

Definition C.4 (Young Tableau)**.** A Young Tableau Θ_λ^p associated with a Young diagram Θ_λ with n boxes is a filling where each box is filled with an integer$1, ..., n$ with no repetitions (definition vary, here we follow (Sagan, 2001)). For example some of the Young Tableaux associated with the Young diagram from the previous example are:

$$\Theta_\lambda^C = \begin{array}{|c|c|c|}\hline 1 & 2 & 3 \\ \hline 4 & 5 \\ \cline{1-2} 6 \\ \cline{1-1}\end{array}, \ \Theta_\lambda^a = \begin{array}{|c|c|c|}\hline 2 & 5 & 3 \\ \hline 6 & 1 \\ \cline{1-2} 4 \\ \cline{1-1}\end{array}, \ \Theta_\lambda^b = \begin{array}{|c|c|c|}\hline 1 & 4 & 5 \\ \hline 2 & 6 \\ \cline{1-2} 3 \\ \cline{1-1}\end{array} \tag{C.2}$$

Definition C.5 (Standard Young Tableau)**.** A standard Young tableau is a Young tableau where the rows and columns increase to the right and to the bottom respectively (Again definitions vary, here we follow (Sagan, 2001)). For example, Θ^b_λ and Θ^C_λ in equation C.2 are standard Young tableaux but Θ^a_λ is not.

Definition C.6 (Canonical Young Tableau)**.** A canonical Young tableau Θ^C_λ is a standard Young tableau where the numbers $1, ..., \lambda_1$ appear in the first row, the numbers $\lambda_1 + 1, ..., \lambda_2$ appear is the second row and so on. For example, the Θ^C_λ in equation C.2 is the canonical Young tableau.

Definition C.7 (Rows and columns subgroups)**.** Given a Young tableau Θ^p_λ of partition λ and assignment p, we define the rows subgroup $\mathcal{R}^p_\lambda$ which leave invariant the (unordered) sets of numbers appearing in the same row of Θ^p_λ. Similarly, we define columns subgroup $\mathcal{C}^p_\lambda$ which leave invariant the (unordered) sets of numbers appearing in the same column of Θ^p_λ.

Definition C.8 (Permutation action on the multilinear polynomials)**.** Let $\mathcal{T}$ be linear representations of the the symmetric group S_n on the multilinear polynomials, such that the permutation acts naturally on the variables indices. E.g. let $\sigma \in S_n$ be a permutation, and let $P(x_1, ..., x_n) = x_1 x_2$ be a multilinear polynomial, then $\mathcal{T}(s)P = x_{\sigma(1)} x_{\sigma(2)}$.

Define the groups of row and column actions on the multilinear polynomials

$$R^p_\lambda = \{\mathcal{T}(\sigma) | \sigma \in \mathcal{R}^p_\lambda\}, \quad C^p_\lambda = \{\mathcal{T}(\sigma) | \sigma \in \mathcal{C}^p_\lambda\} \tag{C.3}$$

Definition C.9 (Row symmetrizer, column anti-symmetrizer and young symmetrizer)**.** Define the row symmetrizer, column anti-symmetrizer and Young symmetrizer linear operators:

$$\hat{R}^p_\lambda = \sum_{r \in R^p_\lambda} r \tag{C.4}$$

$$\hat{C}^p_\lambda = \sum_{c \in C^p_\lambda} \text{sign}(c) c \tag{C.5}$$

$$\hat{Y}^p_\lambda = \hat{C}^p_\lambda \hat{R}^p_\lambda \tag{C.6}$$

Theorem C.10. *Young symmetrizers associated with standard Young tableaux are projectors to irrep spaces of the symmetric group (Fulton & Harris, 2004)*

Lemma C.11. *If there exists a transposition $t^* \in C^p_\lambda$ that leaves a monomial M unchanged, M vanishes under the action of the column anti-symmetrizer -*

$$\exists t^* \in C^p_\lambda \text{ s.t. } t^* M = M \to \hat{C}^p_\lambda M = 0.$$

Proof. Let Θ^p_λ be a standard Young tableau of a partition λ. Let $\hat{C}^p_\lambda$ be the column anti-symmetrizer associated with Θ^p_λ. Let $M(x_1, x_2, ..., x_n)$ be a multilinear monomial in the variables $x_1, x_2, ..., x_n$. Let $t^* \in C^p_\lambda$ be a transposition such that $t^* M = M$. A transposition is an involution, that means, it is a bijection from the group to itself and $t^* t^* = e$, where e is the identity element. Right multiplication with t^* maps any element $c_i \in C^p_\lambda$ from the column group to $c_j = c_i t^*$ such that

$$\text{sign}(c_i) c_i M = \text{sign}(c_i t^* t^*) c_i t^* t^* M = \text{sign}(c_j t^*) c_j t^* M = -\text{sign}(c_j) c_j M. \tag{C.7}$$

We have constructed a unique pairing between each $c_i \in C^p_\lambda$ and $c_j \in C^p_\lambda$ such that $c_i \neq c_j$ and $\text{sign}(c_i) c_i M = -\text{sign}(c_j) c_j M$ that is

$$\forall c_i \in C^p_\lambda \ \exists! c_j \in C^p_\lambda \text{ s.t. } c_i \neq c_j \wedge \text{sign}(c_i) c_i M = -\text{sign}(c_j) c_j M.$$

That means the terms in the sum cancel in pairs $\hat{C}^p_\lambda M = \sum_{c \in C^p_\lambda} \text{sign}(c) c M = 0$. ■

Lemma C.12. *All multilinear monomials in n variables, vanish when acted upon with a column anti-symmetrizer that corresponds to a Young tableau with more than 2 rows*

Proof. Let $M(x_1, x_2, ..., x_n)$ be a multilinear monomial in the variables $x_1, x_2, ..., x_n$. Let Θ^p_λ be a standard Young tableau of a partition λ that has more than 2 rows. The first column in Θ^p_λ gives raise to at least 3 transpositions $(ab), (bc), (ac)$. Since each variable must either appear in $M(x_1, x_2, ..., x_n)$ to a single power or zeroth power, out of the 3 variables x_a, x_b, x_c at least two must appear to the same power. Because the product of our variables is not ordered, at least one of the 3 transpositions leaves $M(x_1, x_2, ..., x_n)$ unchanged. Applying lemma C.11, $M(x_1, x_2, ..., x_n)$ must vanish under the action. ■

Lemma C.13. *All multilinear monomials of degree d in n variables, vanish when acted upon with a column anti-symmetrizer associated with a partition $(n-k,k)$ for $k > \min\{d, n-d\}$.*

Proof. for $k > \{d, n-d\}$ there exists a column transposition $(ab) \in C^p_\lambda$ where both x_a, x_b appear in the monomial to zeroth power, therefore the transposition (ab) leaves it unchanged. Applying lemma C.11, the monomial must vanish under the action. ■

Remark C.14. The multilinear monomial can be thought of as picking specific boxes in the Young tableau, one can then permute inside the rows, writing down the numbers that appear in the chosen boxes as the indices in the monomial. Finally one can act with the column permutations, while adding their signs, on the monomials found by the rows actions. Summing up all terms gives the result of acting with the Young symmetrizer on the monomial. The necessary conditions above for $\hat{C}^p_\lambda M \neq 0$ translate to being able to pick d boxes such that at most one box is picked in every column, and no column has more than one box unpicked in it.

Lemma C.15. *There exists a multilinear monomial of degree d in n variables, that does not vanish when acted upon with a Young symmetrizer associated with a partition $(n-k,k)$ for every k such that $0 \leq k \leq d, n-d$.*

Proof. Let $M = \prod_{i=1}^d x_i$ be a multilinear monomial of degree d in n variables. Let $\Theta^C_{(n-k,k)}$ be the canonical Young tableau associated with the partition $(n-k,k)$ for $0 \leq k \leq d, n-d$,

$$\Theta^C_{(n-k,k)} = \begin{array}{|c|c|ccc|c|} \hline 1 & 2 & \dots & k & \dots & n-k \\ \hline n-k+1 & n-k+2 & \dots & n & & \\ \cline{1-4} \end{array} \,. \tag{C.8}$$

We now verify $\hat{Y}^C_{(n-k,k)} M \neq 0$:

The row symmetrizer sums positive elements, therefore the sum cannot vanish

$$P = \hat{R}^C_{(n-k,k)} M = \sum_{r \in R^C_{(n-k,k)}} rM \neq 0. \tag{C.9}$$

Since $\{x_i\}_{i=1}^n$ are independent variables all elements in the sum above are linearly independent (up to identical elements). We may conclude it is sufficient to show a single element doesn't vanish to prove $\hat{C}^C_{(n-k,k)} P$ doesn't vanish, since $\hat{C}^C_{(n-k,k)}$ includes the trivial element. In particular, we will show that for $r = e$ the summand $rM = M$ does not vanish under the action of the column symmetrizer.

The column symmetrizer $\hat{C}^C_{(n-k,k)}$ is a sum of closed, independent, column transpositions and their products. All non-trivial transpositions, when acting on M specifically, create linearly independent elements, therefore the sum of such transpositions acting on M cannot vanish.

We may conclude $\hat{C}^C_{(n-k,k)} P$ includes at least one non vanishing term (that is M) and therefore $\hat{Y}^C_{(n-k,k)} M \neq 0$. ■

Definition C.16 (Hook Length)**.** The hook length $h_\lambda(i,j)$ of a box, where i (j) denotes the row (column) of the box in the Young diagram Θ_λ, is the number of boxes to the right of the i,j'th box in the i'th row, plus the number of boxes below the box in the j'th column plus one.

Lemma C.17. *The dimension of an irrep associated with a partition $(n-k,k)$ is $\dim_\lambda = \frac{n!}{k! \frac{(n-k+1)!}{n-2k+1}}$.*

Proof. using the hook length formula (Fulton & Harris, 2004)

$$\dim_\lambda = \frac{n!}{\prod_{i,j \in \lambda} h_\lambda(i,j)}.$$

The product in the denominator equals

$$\prod_{i,j\in\lambda} h_\lambda(i,j) = \underbrace{(n-2k)!}_{\text{upper row with nothing below}} \underbrace{k!}_{\text{lower row}} \underbrace{\frac{(n-k+1)!}{(n-2k+1)!}}_{\text{upper row with boxes below}} = k!\frac{(n-k+1)!}{n-2k+1} = \binom{n+1}{k}\frac{(n+1)!}{n-2k+1}. \tag{C.10}$$

Resulting in

$$\dim_\lambda = \frac{n!}{k!\frac{(n-k+1)!}{n-2k+1}} \sim n^k.$$

■

Theorem 2.1. *The space of homogeneous multilinear polynomials in n variables of degree d can be* fully *decomposed into* $\min\{d+1, n-d+1\}$ unique *irreps of S_n labeled by the partitions $(n-k,k)$ for $0 \le k \le d, n-d$.*

Proof. Let Θ^p_λ be a standard Young tableau of a partition λ. Let $\hat{R}^p_\lambda, \hat{C}^p_\lambda, \hat{Y}^p_\lambda$ be the row symmetrizer, column anti-symmetrizer and Young symmetrizer (respectively) of the Θ^p_λ.

Let $\{M^d_n\}$ be the set of all multilinear monomials in n variables of degree d.

$\{M^d_n\}$ is a basis for the space of multilinear polynomials in n variables of degree d. That means $\mathrm{Span}\{M^d_n\}$ is the space of multilinear polynomials in n variables of degree d.

$\mathrm{Span}\{M^d_n\}$ is closed under the action of $\hat{R}^p_\lambda$. Therefore, if $\forall M \in \{M^d_n\},\ \hat{C}^p_\lambda M = 0$, then $\forall P \in \mathrm{Span}\{M^d_n\},\ \hat{Y}^p_\lambda P = 0$.

Using lemmas C.12,C.13 we see that all $P \in \mathrm{Span}\{M^d_n\}$ vanish under the action of the Young symmetrizers associated with a Young diagram with more than 2 rows or more than $\min\{d, n-d\}$ boxes on the second row.

Based on lemma C.15 and theorem C.10 each of the irreps $(n-k,k)$ $0 \le k \le d, n-d$ appears at least once in the decomposition of $\mathrm{Span}\{M^d_n\}$ into irreps of the symmetric group.

$\mathrm{Span}\{M^d_n\}$ is $\binom{n}{d}$ dimensional.

Summing the dimension of the irreps (lemma C.17)

$$\sum_{k=0}^{\min\{d,n-d\}} \frac{n!}{k!\frac{(n-k+1)!}{n-2k+1}} = \binom{n}{d}$$

Since the sum of dimensions of irreps equals the dimension of the space each irrep appears only once. ■

D. Linear example

In section 2.3.1 we have found the kernel can be written in be simplified down to a scalar form

$$k(X,Y) = \underbrace{\frac{1}{8}\left(x_1^{L+1}y_1^{L+1} + \left(1-x_1^{L+1}\right)\left(1-y_1^{L+1}\right)\right)}_{\mathfrak{A}} \cdot \underbrace{\begin{bmatrix} \frac{1}{L^2}\sum_{a,b=1}^{L}\left(x_1^a y_1^b + (1-x_1^a)\left(1-y_1^b\right)\right) \\ +\frac{1}{L^2}\sum_{a=1}^{L}\left(x_1^a y_1^a + (1-x_1^a)\left(1-y_1^a\right)\right) + \frac{1}{L} \end{bmatrix}}_{\mathfrak{B}}, \tag{D.1}$$

revealing the space of expressible function to be the zeroth and first-degree polynomials of the tokens in the context (part $\mathfrak{B}$) window multiplied by the last token(part $\mathfrak{A}$). We followed up by applying the Thm. 2.1, to reveal they are composed of standard and trivial irreps.

Turning to the space of the standard irrep, it can be further decomposed to one-dimensional irreps of the cyclic subgroup known as the Fourier modes, thereby acquiring eigenvectors of $\mathfrak{B}$. Putting these together with the eigenvectors of $\mathfrak{A}$ $a(\vec{x}^{L+1}),\ b(\vec{x}^{L+1})$ we find $2(L-2)$ eigenvectors of the kernel (given explicitly in equation D.3).

The eigenvalues are all independent of $k \in \{1, 2, ...(L/2-1)\}$ since all the k modes belong to the same irrep, and only differ by $O(1)$ factor from one another based on the difference between odd and even and the a, b subspaces

$$\lambda_{k,a}^{\text{odd}}, \lambda_{k,a}^{\text{even}}, \lambda_{k,b}^{\text{odd}}, \lambda_{k,b}^{\text{even}} \propto \frac{1}{L^2} \tag{D.2}$$

full expressions are given by

$$\left\{ \begin{pmatrix} \varphi_{k,a}^{\text{odd}}(X) \\ \varphi_{k,b}^{\text{odd}}(X) \end{pmatrix} = \begin{pmatrix} \frac{x^{L+1}}{Z_{k,a}^{\text{odd}}} \\ \frac{1-x^{L+1}}{Z_{k,b}^{\text{odd}}} \end{pmatrix} \sum_{s=1}^{L/2} e^{i\frac{\pi k}{L/2}s} x_1^{2s-1}, \quad \begin{pmatrix} \varphi_{k,a}^{\text{even}}(X) \\ \varphi_{k,b}^{\text{even}}(X) \end{pmatrix} = \begin{pmatrix} \frac{x^{L+1}}{Z_{k,a}^{\text{even}}} \\ \frac{1-x^{L+1}}{Z_{k,b}^{even}} \end{pmatrix} \sum_{s=1}^{L/2} e^{i\frac{\pi k}{L/2}s} x_1^{2s} \right\}_{k=1}^{L/2-1}, \tag{D.3}$$

with $Z_{k,a}^{\text{odd}}, Z_{k,b}^{\text{odd}}, Z_{k,a}^{even}, Z_{k,b}^{even}$ being appropriate normalization constants.

Following the same procedure we find the trivial representation is spanned by

$$\tilde{\varphi}_0^{\text{odd}}(X) = \sum_{s=1}^{L/2} x_1^{2s-1}; \;\; \tilde{\varphi}_0^{\text{even}}(X) = \sum_{s=1}^{L/2} x_1^{2s}; \;\; \tilde{\varphi}_c(X) = 1. \tag{D.4}$$

By a Gram–Schmidt like-process, we find a good basis for the space of permutation invariant functions $\varphi_{c,*}, \varphi_{0,*}^{+}, \varphi_{0,*}^{-}$ with $* = \{a, b\}$; the definitions are given in equation I.1. The diagonalization in the multiplicity spaces of the trivial irrep can now be carried out numerically or analytically in closed form as it can be written as two 3×3 matrices.

D.1. Learnable target

So far, the whole process has been task-independent, the last component required to predict the output of the NN is the projections of the target onto the eigenvectors, which depend on the target function and the training distribution. Since the task requires estimating a parameter not accessible to the network, the projections can never span the true target function, instead even as $N \to \infty$ the network will learn a different function which we dub the *learnable target* given by $\sum_i g_i \varphi_i(x)$. We denote the projections by $g_*^-, g_*^+, g_{c,*}, g_{k,*}^{\text{odd}}, g_{k,*}^{\text{even}}$ for $\varphi_{0,*}^-, \varphi_{0,*}^+, \varphi_{c,*}, \varphi_{k,*}^{\text{odd}}, \varphi_{k,*}^{\text{even}}$ respectively, where $* = \{a, b\}$. This projections depend on the parameters of the training distribution p_a, q_a, w, L. Keeping only leading orders of $w, \frac{1}{L}$ we find $g_{k,*}^{\text{odd}}, g_{k,*}^{\text{even}}$ vanish for all k, and $g_{c,*}$ are constants w.r.t w, L while

$$g_*^+ = \frac{Lw^2\eta_*^+}{\sqrt{L^2w^2\rho_*^+ + L\sigma_*^+}}, \quad g_*^- = \frac{Lw^2\eta_*^- + \nu_*^-}{\sqrt{L^2w^4\rho_*^- + Lw^2\sigma_*^- + \xi_*^-}}, \tag{D.5}$$

the definitions of $\eta_*^\star$, $\nu_*^\star$, $\rho_*^\star$, $\sigma_*^\star$, $\xi_*^\star$, where $* = \{a, b\}$ and $\star = \{+, -\}$, are detailed in appendix I.

E. Out of distribution predictions under EK approximation

We would like to compute the mean square loss (MSE) on an arbitrary data distribution p_{data} when training on a data distribution p_{train} for which we know the EK predictor.

Under EK approximation MSE loss can be computed by

$$\begin{aligned} &\mathbb{E}_{X\sim p_{\text{data}}}\mathbb{E}_\Theta\left[\left(f_\Theta(X) - g(X)\right)^2\right] = \mathbb{E}_{X\sim p_{\text{data}}}\mathbb{E}_\Theta\left[\left(f_\Theta(X) - g(X)\right)^2\right] = \\ &= \mathbb{E}_{X\sim p_{\text{data}}}\left[\mathbb{E}_\Theta\left[f_\Theta^2(X)\right] - 2\mathbb{E}_\Theta\left[f_\Theta(X)\right]g(X) + g^2(X)\right] \approx \mathbb{E}_{X\sim p_{\text{data}}}\left[\mathbb{E}_\Theta\left[f_\Theta(X)\right]^2 - 2\mathbb{E}_\Theta\left[f_\Theta(X)\right]g(X) + g^2(X)\right] = \\ &= \mathbb{E}_{X\sim p_{\text{data}}}\left[\left[\sum_i \frac{\lambda_i}{\lambda_i + \sigma^2/N} g_i\varphi_i(x)\right]^2 - 2\sum_i \frac{\lambda_i}{\lambda_i + \sigma^2/N} g_i\varphi_i(x)\, g(X) + g^2(X)\right], \end{aligned} \tag{E.1}$$

Where the approximation on the second line is dropping the EK variance

$$\mathbb{E}_\Theta\left[f_\Theta(X)\right]^2 = \mathbb{E}_\Theta\left[f_\Theta(X)\right]^2 + \text{Var}\left[f_\Theta(X)\right] \approx \mathbb{E}_\Theta\left[f_\Theta(X)\right]^2. \tag{E.2}$$

One can in fact calculate this quantity easily within the GP framework but we found the approximation to be good enough as is and chose to drop it for simplicity. This result is exact when using an ensemble, or when using the EK predictor itself as the predictor.

By the spectral theorem, are guaranteed $\phi_i(x)$ to be orthogonal w.r.t. an inner product defined with p_{train} as a weighting function as defined in equation 7. Therefore, choosing $p_{\text{data}} = p_{\text{train}}$ simplifies equation E.2 to

$$= \sum_i \left(\frac{\lambda_i}{\lambda_i + \sigma^2/N} \right)^2 g_i^2 - 2 \sum_i \frac{\lambda_i}{\lambda_i + \sigma^2/N} g_i^2 + \langle g, g \rangle_{X \sim p_{\text{train}}}, \tag{E.3}$$

as $\langle \varphi_i, \varphi_j \rangle_{p_{\text{train}}} = \delta_{ij}$, with δ_{ij} being the Kronecker delta.

Now, if we wish to compute the loss under distributional shift all we have to do is take the expectation value in equation E.2 w.r.t. a new distribution $p_{\text{data}} = p_{\text{test}} \neq p_{\text{train}}$

$$\sum_i \sum_j \frac{\lambda_i}{\lambda_i + \sigma^2/N} \frac{\mu_j}{\mu_j + \sigma^2/N} g_i g_j \langle \varphi_i, \varphi_j \rangle_{X \sim p_{\text{test}}} - 2 \sum_i \frac{\lambda_i}{\lambda_i + \sigma^2/N} g_i \langle \varphi_i, g \rangle_{X \sim p_{\text{test}}} + \langle g, g \rangle_{X \sim p_{\text{test}}} . \tag{E.4}$$

Notably, the eigenfunctions that were orthonormal under the inner product induced by the training distribution are no longer necessarily orthonormal under the test distribution, and the object $G_{ij} := \langle \varphi_i, \varphi_j \rangle_{X \sim p_{\text{test}}}$ can be identified as their Gram matrix. Nevertheless, the elements of the Gram matrix can be evaluated as sums or integrals, analytically or numerically. In our case, we evaluated the elements of G_{ij} for the full parameter space of p_a, q_a, w analytically.

F. Sub-leading corrections from x^{L+1}

The terms left out during the approximation are

$$\begin{aligned} k^{(1)}(X,Y) = & \frac{1}{8L^2} \left(x^{L+1} y^{L+1} + \left(1 - x^{L+1}\right)\left(1 - y^{L+1}\right) \right) \cdot ... \\ & ... \cdot \left[\sum_{a=1}^{L} x^{L+1} y^a + \sum_{a=1}^{L} \left(1 - x^{L+1}\right)\left(1 - y^a\right) + \sum_{a=1}^{L} x^a y^{L+1} + \sum_{a=1}^{L} \left(1 - x^a\right)\left(1 - y^{L+1}\right) + ... \right. \\ & \left. ... + 3 x^{L+1} y^{L+1} + 3 \left(1 - x^{L+1}\right)\left(1 - y^{L+1}\right) + 1 \right] \end{aligned} \tag{F.1}$$

All the vectors $\varphi_{k,a}^{\text{odd}}(X),\ \varphi_{k,a}^{\text{even}}(X)\ , \varphi_{k,b}^{\text{odd}}(X),\ \varphi_{k,b}^{\text{even}}(X)$ in the standard representation get no corrections at all as their matrix elements with all basis vectors vanish.

Moving on to the two 3×3 blocks of the trivial representation, $\varphi_{0,a}^{+}, \varphi_{0,a}^{-}$ ($\varphi_{0,b}^{+}, \varphi_{0,b}^{-}$) can only have non-vanishing matrix elements with the $\varphi_{c,a}$ ($\varphi_{c,b}$). These terms are at most $O\left(\frac{1}{L^3}\right)$; furthermore, they are second-order corrections in the eigenvalue perturbation and are therefore sub-leading.

Last $\varphi_{c,a}$ ($\varphi_{c,b}$) can get corrections to the diagonal term, but they will be at most $O\left(\frac{1}{L}\right)$ while the leading term is $O(1)$.

G. Large structure decomposition and non-linearities

One can write the kernel of the network when applying non-linearities in the form:

$$k(X,Y) = \sum_\alpha k_\alpha^{L+1}(x^{L+1}, y^{L+1}) k_\alpha^L(\{x^s\}_{s=1}^L, \{y^s\}_{s=1}^L). \tag{G.1}$$

for some $\left\{k_\alpha^{L+1}, k_\alpha^L\right\}_\alpha$. Since all k_α^L possess the permutation symmetry they will be diagonalized in the same basis as the symmetry operator. Suppose $\varphi_j^L\left(\{x^s\}_{s=1}^L\right)$ is a non-degenerate eigenfunction of the symmetry operator, we have that $\hat{K}_\alpha^L \varphi_j = \lambda_{\alpha,j}^L \varphi_j$ simplifying the kernel eigenvalue problem to

$$\hat{K}\left(\varphi_i^{L+1} \varphi_j^L\right) = \lambda_{ij}\left(\varphi_i^{L+1} \varphi_j^L\right), \tag{G.2}$$

where $\{\varphi_j^L\}_{j=1}^n$ are known, forming blocks of size n. Note that this is not a simple tensor product structure $\lambda_{ij} \neq \lambda_i^{L+1} \lambda_j^L$ as x^{L+1} is not independent of $\{x^s\}_{s=1}^L$.

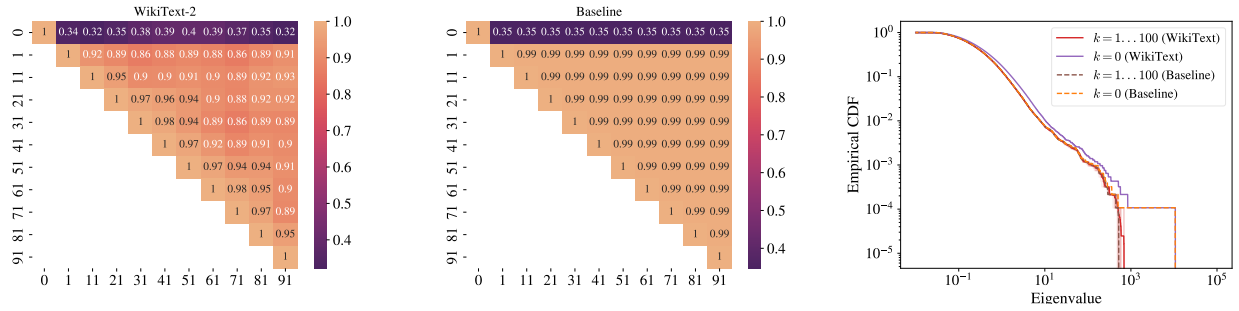

Figure 4. **Left: (Similarity measure between C^{kk} and $C^{k'k'}$ on WikiText)** the cosine similarity induced by the Frobenius inner product between the linear features of WikiText C^{kk} and $C^{k'k'}$ for the k's indicated on the boundary. We see all sampled $k \neq 0$ are similar to one another but different from $k = 0$ as predicted by the irrep decomposition. **Center: (Similarity measure between C^{kk} and $C^{k'k'}$ on a permutation symmetric baseline dataset)** The baseline dataset is created by sampling words from WikiText with frequencies as in WikiText, but with to sequential order. The underlying distribution of baseline is therefore completely permutation symmetric in sequence space. We display the same quantity as the figure on the left, this time calculated on the baseline dataset. We see all sampled $k \neq 0$ are very similar, measurably more so then the same features of WikiText. Yet, comparing the differences between the datasets to the similarity gap of $\simeq 0.6$ between $k = 0$ and all $k' \neq 0$, the results for WikiText and the baseline dataset are remarkably similar, suggesting an approximate permutation symmetry. **Right: (Comparing the spectra of C^{kk} between WikiText and the permutation symmetric baseline)** The similarity between ks with $k \neq 0$ is again seen in the spectra. One notably difference between the baseline and WikiText is the spectrum of C^{00} which differs along almost all the scale of eigenvalues, showing the principle components do capture information about sequence dependence, information that does not exist in the baseline.

H. WikiText-2 Symmetry Experiment Details

Here we give some of the details about the WikiText-2 symmetry experiment. We started with tokenizing and trimming: each sample was tokenized and trimmed to $L = 101$ tokens. We removed any sample that was shorter than 101 tokens, leaving us with about $10,000$ samples.

If the dataset is permutation invariant, Ideally, one would now want to perform principal component analysis (PCA) and find a set of generically N_{voc} different states, each with degeneracy $L - 1$ for $k = 1, ..., L - 1$ belonging to the standard irrep, and another set of generically non-degenerate N_{voc} different states, for $k = 0$ belonging to the trivial irrep. The PCA matrix would be

$$C^{ab}_{ij} := \mathbb{E}_{X \sim \text{WikiText}-2} \left[X^a_i X^b_j \right], \tag{H.1}$$

where a, i and b, j can be understood as some "flattened" super index of a $(L \cdot N_{\text{voc}}) \times (L \cdot N_{\text{voc}})$ dimensional matrix.

Moving on to Fourier space

$$\tilde{C}^{kk'}_{ij} := \mathbb{E}_{X \sim \text{WikiText}-2} \left[X^a_i V^{ak} X^b_j V^{bk'} \right]; \tag{H.2}$$

$$V^{ak} := \exp\left(i \frac{2\pi}{L} ak \right), \quad \begin{array}{l} a = 1, ..., L \\ k = 0, ..., L - 1 \end{array}. \tag{H.3}$$

One would then expect to find a block diagonal matrix where $\tilde{C}^{kk'}_{ij} = 0$ for $k \neq k'$ and $\tilde{C}^{kk}_{ij} = \tilde{C}^{k'k'}_{ij}$ for $k, k' \in \{1, ..., L - 1\}$.

However, since the number of samples $N < L \cdot N_{\text{voc}}, N_{\text{voc}}$ one cannot expect to find a block diagonal structure. Both the ranks of the matrix $\tilde{C}$ and the block $\tilde{C}^{k,k'}$ are determined by N, such that $\operatorname{rank} \tilde{C} = \operatorname{rank} \tilde{C}^{k,k} = N$, so the off-block-diagonal elements must not vanish to make the equality possible. A well-studied similar setting is that of the Wishart ensemble in random matrix theory (Potters & Bouchaud, 2020; Akemann et al., 2015). Even with $N < L \cdot N_{\text{voc}}$ we may still expect $\tilde{C}^{kk}_{ij} = \tilde{C}^{k'k'}_{ij}$ for $k, k' \in \{1, ..., L - 1\}$, but we would have to consider the noise due to the finite sampling.

To measure whether $\tilde{C}^{kk}_{ij} = \tilde{C}^{k'k'}_{ij}$ for $k, k' \in \{1, ..., L - 1\}$ we present in the main text the cosine similarity induced by the Frobenius inner product and compare the spectrum's empirical cumulative distribution function (ECDF). Here we give a more detailed plot of the similarity, where the cosine similarity value is written on top of each corresponding square

It is important to noted our results support the hypothesis WikiText-2 has an *approximate* permutation symmetry, but not an exact one. We create a baseline dataset where the word frequencies are identical to those in WikiText-2, but are drawn uniformly over the sequence, i.e. the underlying distribution is exactly permutation symmetric. One can see it shows greater similarity between $k, k' \neq 0$ blocks than the one found on WikiText-2. We can also see the information contain in WikiText-2 principal components goes beyond the frequency of words, merely by the fact the results are not identical. We also compare the empirical eigenvalue CDF, showing agreement on the spectra of $\tilde{C}^{kk}$ for $k \neq 0$ but a significant difference for $k = 0$.

In principle, using this method, one can look at correlations up to an arbitrary order, e.g. the third-order correlator would be

$$\mathcal{C}^{abc}_{ijk} := \mathbb{E}_{X\sim\text{WikiText}-2}\left[X^a_i X^b_j X^c_k\right]. \tag{H.4}$$

I. Full expressions of quantities in the main text

Here we provide the full expressions for some of the quantities defined in the main text and in Appendix D.

The basis chosen for the trivial representation is

$$\begin{aligned}
\begin{pmatrix}\varphi_{c,a}\\ \varphi_{c,b}\end{pmatrix}(X) &= \begin{pmatrix}\frac{1}{Z_{c,a}}\\ \frac{1}{Z_{c,b}}\end{pmatrix}\begin{pmatrix}x_1^{L+1}\\ 1-x_1^{L+1}\end{pmatrix} \\
\begin{pmatrix}\varphi^+_{0,a}\\ \varphi^+_{0,b}\end{pmatrix}(X) &= \begin{pmatrix}\frac{1}{Z^+_{0,a}}\\ \frac{1}{Z^+_{0,b}}\end{pmatrix}\begin{pmatrix}x_1^{L+1}\\ 1-x_1^{L+1}\end{pmatrix}\frac{1}{L} \quad \left[\sum_{s=1}^{L} x^s - \begin{pmatrix}\frac{c_a^{\text{odd}}+c_a^{\text{even}}}{2}\\ \frac{c_b^{\text{odd}}+c_b^{\text{even}}}{2}\end{pmatrix}\right] \\
\begin{pmatrix}\varphi^-_{0,a}\\ \varphi^-_{0,b}\end{pmatrix}(X) &= \begin{pmatrix}\frac{1}{Z^-_{0,a}}\\ \frac{1}{Z^-_{0,b}}\end{pmatrix}\begin{pmatrix}x_1^{L+1}\\ 1-x_1^{L+1}\end{pmatrix}\frac{1}{L} \quad \left[\begin{pmatrix}\alpha_a\\ \alpha_b\end{pmatrix}\left(\sum_{s=1}^{L/2} x^{2s-1} - \begin{pmatrix}c_a^{\text{odd}}\\ c_b^{\text{odd}}\end{pmatrix}\right)\cdots\right. \\
&\qquad \left.\cdots - \begin{pmatrix}\beta_a\\ \beta_b\end{pmatrix}\left(\sum_{s=1}^{L/2} x^{2s} - \begin{pmatrix}c_a^{\text{even}}\\ c_b^{\text{even}}\end{pmatrix}\right)\right]
\end{aligned} \tag{I.1}$$

with

$$\begin{aligned}
\alpha_a =& \frac{-24p_aq_a(p_a+q_a-2)(p_a+q_a) - 12w\left({p_a}^3 + {p_a}^2(7q_a-2) + p_aq_a(7q_a-8) + (q_a-2){q_a}^2\right) + ...}{48(p_a+q_a+w)} \\
& \frac{... + 2w^2\left((L-16){p_a}^2 + q_a((L-16)q_a+18) + p_a(18-44q_a)\right) + 2w^3 + ...}{48(p_a+q_a+w)} \\
& \frac{... + ((L-14)p_a + (L-14)q_a + 6) + (L-8)w^4}{48(p_a+q_a+w)}
\end{aligned} \tag{I.2}$$

$$\begin{aligned}
\beta_a =& \frac{-36(p_a+q_a)\left((p_a-1){p_a}^2 + (q_a-1){q_a}^2\right) - 18w\left(5{p_a}^3 + {p_a}^2(3q_a-4) + p_aq_a(3q_a-4) + {q_a}^2(5q_a-4)\right) + ...}{72(p_a+q_a+w)} \\
& \frac{... + 6w^2\left(p_a((L-12)q_a+10) - 15{p_a}^2 + 5q_a(2-3q_a)\right) + 3w^3((L-18)p_a + (L-18)q_a + 8) + (L-18)w^4}{72(p_a+q_a+w)}
\end{aligned} \tag{I.3}$$

$$\begin{aligned}
\alpha_b =& -\frac{-24(p_a-1)(q_a-1)(p_a+q_a-2)(p_a+q_a) + ...}{48(p_a+q_a+w-2)} \\
& \frac{... - 12w\left({p_a}^3 + {p_a}^2(7q_a-8) + p_a(q_a-2)(7q_a-6) + (q_a-6)(q_a-2)q_a - 4\right) + ...}{48(p_a+q_a+w-2)} \\
& \frac{... + 2w^2\left(L((p_a-2)p_a + (q_a-2)q_a + 2) - 2\left(8{p_a}^2 + p_a(22q_a-29) + q_a(8q_a-29) + 20\right)\right) + ...}{48(p_a+q_a+w-2)} \\
& \frac{... + 2w^3(L(p_a+q_a-2) - 2(7p_a+7q_a-11)) + (L-8)w^4}{48(p_a+q_a+w-2)}
\end{aligned} \tag{I.4}$$

$$\begin{aligned}\beta_b =& \frac{36(p_a+q_a-2)\left(p_a(p_a-1)^2+(q_a-1)^2q\right)+...}{72(p_a+q_a+w-2)}\\ &\frac{...+18w\left(5{p_a}^3+{p_a}^2(3q_a-14)+p_a(q_a(3q_a-8)+12)+q_a(q_a(5q_a-14)+12)-4\right)+...}{72(p_a+q_a+w-2)}\\ &\frac{...+6w^2\left(L(p_a(-q_a)+p_a+q_a-1)+15{p_a}^2+4p_a(3q_a-8)+q_a(15q_a-32)+22\right)+...}{72(p_a+q_a+w-2)}\\ &\frac{...-3w^3(L(p_a+q_a-2)-2(9p_a+9q_a-14))-\left((L-18)w^4\right)}{72(p_a+q_a+w-2)}\end{aligned} \tag{I.5}$$

$$\begin{aligned}c_a^{\text{odd}} &= \frac{3\left({p_a}^2+{q_a}^2\right)+3w(p_a+q_a)+2w^2}{3(p_a+q_a+w)}\\ c_a^{\text{even}} &= \frac{(2p_a+w)(2q_a+w)}{2(p_a+q_a+w)}\\ c_b^{\text{odd}} &= \frac{3w(p_a+q_a-1)+3(p_a-1)p_a+3(q_a-1)q_a+2w^2}{3(p_a+q_a+w-2)}\\ c_b^{\text{even}} &= \frac{2p_a(2q_a+w-1)+2q_a(w-1)+(w-2)w}{2(p_a+q_a+w-2)}\end{aligned} \tag{I.6}$$

$$\begin{aligned}\lambda_{k,a}^{\text{odd}} &= \frac{1}{8L^2}\left[2\left((1-p_a)\,p_a^2+(1-q_a)\,q_a^2\right)+O(w)\right],\\ \lambda_{k,a}^{\text{even}} &= \frac{1}{8L^2}\left[2p_aq_a\left(1-p_a+1-q_a\right)+O(w)\right],\\ \lambda_{k,b}^{\text{odd}} &= \frac{1}{8L^2}\left[2\left(p_a\left(1-p_a\right)^2+q_a\left(1-q_a\right)^2\right)+O(w)\right],\\ \lambda_{k,b}^{\text{even}} &= \frac{1}{8L^2}\left[2\left(1-p_a\right)\left(1-q_a\right)\left(p_a+q_a\right)+O(w)\right]\end{aligned} \tag{I.7}$$

To leading order in $\frac{1}{L}, w$, the spanning coefficients of the learnable target are given by

$$\begin{aligned}g_{k,*}^{\text{odd}} &= 0, & g_{k,*}^{\text{even}} &= 0\\ g_*^+ &= \frac{Lw^2\eta_*^+}{\sqrt{L^2w^2\rho_*^+ + L\sigma_*^+}}, & g_*^- &= \frac{Lw^2\eta_*^- + \nu_*^-}{\sqrt{L^2w^4\rho_*^- + Lw^2\sigma_*^- + \xi_*^-}},\\ g_{c,a} &= \frac{p_aq_a}{\sqrt{\frac{p_a+q_a}{2}}}, & g_{c,b} &= \frac{q_a+p_a-2p_aq_a}{\sqrt{2\left(1-p_a+1-q_a\right)}};\end{aligned} \tag{I.8}$$

with

$$\begin{aligned}\eta_{0,a}^+ &= 2\left({p_a}^2+{q_a}^2\right)\\ \rho_{0,a}^+ &= 48(p_a+q_a)^3\\ \sigma_{0,a}^+ &= -576(p_a+q_a)^3((p_a-1)p_a+(q_a-1)q_a)\end{aligned} \tag{I.9}$$

$$\begin{aligned}\eta_{0,b}^+ &= 2(p_a-2)p_a+2(q_a-2)q_a+4\\ \rho_{0,b}^+ &= 2(p_a-2)p_a+2(q_a-2)q_a+4\\ \sigma_{0,b}^+ &= 576(p_a+q_a-2)^3((p_a-1)p_a+(q_a-1)q_a)\end{aligned} \tag{I.10}$$

$$\begin{aligned}\eta_{0,a}^- &= -72p_aq_a(p_a-q_a)^2(p_a+q_a)\\ \nu_{0,a}^- &= 864p_aq_a(p_a-q_a)^2(p_a+q_a)((p_a-1)p_a+(q_a-1)q_a)\\ \rho_{0,a}^- &= 10368p_aq_a(p_a-q_a)^2(p_a+q_a)^3\\ \sigma_{0,a}^- &= -248832p_aq_a(p_a-q_a)^2(p_a+q_a)^3((p_a-1)p_a+(q_a-1)q_a)\\ \xi_{0,a}^- &= 1492992p_aq_a(p_a-q_a)^2(p_a+q_a)^3((p_a-1)p_a+(q_a-1)q_a)^2\end{aligned} \tag{I.11}$$

$$
\begin{aligned}
\eta_{0,a}^{-} &= -72(p_a-1)(q_a-1)(p_a-q_a)^2(p_a+q_a-2) \\
\nu_{0,a}^{-} &= 864(p_a-1)(q_a-1)(p_a-q_a)^2(p_a+q_a-2)((p_a-1)p_a+(q_a-1)q_a) \\
\rho_{0,a}^{-} &= -10368(p_a-1)(q_a-1)(p_a-q_a)^2(p_a+q_a-2)^3 \\
\sigma_{0,a}^{-} &= 248832(p_a-1)(q_a-1)(p_a-q_a)^2(p_a+q_a-2)^3((p_a-1)p_a+(q_a-1)q_a) \\
\xi_{0,a}^{-} &= -1492992(p_a-1)(q_a-1)(p_a-q_a)^2(p_a+q_a-2)^3((p_a-1)p_a+(q_a-1)q_a)^2
\end{aligned}
\tag{I.12}
$$