TRANSFORMERS' SPECTRAL BIAS AND THE SYMMETRIC GROUP

Itay Lavie
Racah Institute of Physics
Hebrew University of Jerusalem
Jerusalem 91904, Israel

Guy Gur-Ari
Augment Computing

Zohar Ringel
Racah Institute of Physics
Hebrew University of Jerusalem
Jerusalem 91904, Israel

ABSTRACT

We study inductive bias in transformers in the infinitely over-parameterized kernel 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; 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)).

We approach the challenge from the infinitely over-parameterized kernel 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; Jacot et al., 2018). We rely on the established NNGP (Neal, 1996; Lee et al., 2020; Naveh et al., 2021) and NTK (Jacot et al., 2018) correspondences between infinitely wide transformer NN and kernel methods (Hron et al., 2020), and understand their inductive bias through the eigenvalue decomposition of the kernel (Canatar et al., 2021b; Cohen et al., 2021; Simon et al., 2023). We characterize the inductive bias by *learnability* i.e. 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. 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 attention at the kernel 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 permutation symmetry in its principal components, suggesting that the toolbox presented can be of use on natural language datasets.

2 MODEL

We study a transformer-like NN (Vaswani et al., 2017) 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), one multi-head self-attention layer (MHA) with a non-linearity Φ (commonly chosen to be $\mathrm{softmax}$), a one hidden layer MLP with non-linearity ϕ (commonly chosen to be ReLU) and a final linear readout layer. The input to the NN is made out of $L+1$ tokens $\vec{x}^s$ indexed by an upper sequence index $s = 1, 2, ..., L+1$ with each token having an internal (vocabulary or embedding) 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 $X^s_{i,\mu}$, where we drop the sample index μ when we discuss only a single sample. One-hot encoding is used for the tokens, such that $[\vec{x}^s]_i = \delta_{i,v}$ where $v = 1, ..., N_{\mathrm{voc}}$ is the token represented by $\vec{x}^s$. For detailed model description see appendix D.

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 traceability. The HMMs have a vocabulary of size $N_{\mathrm{voc}} = 2$ and $d_{\mathrm{hidden}} = 2$ hidden states, 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 across all samples, with probability 1 to switch a hidden state at each state. For a complementary introduction to HMMs and a detailed description of the dataset used see appendix E.

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 4 and present evidence for an approximate permutation symmetry in the principal components of WikiText.

3 THEORY

Here, we derive a bound on the sample complexity of a target function, its *learnability*, as a function of the context length and the decomposition of the target to irreps of the symmetric group.

Infinitely wide NNs admit kernel limits, where Bayesian inference is described by the regression with the NNGP kernel (Lee et al., 2018) and learning with gradient flow is described by regression with the NTK (Jacot et al., 2018). For transformers, the existence of such limits was established in Hron et al. (2020), when the key's dimension (d_k) and the number of heads (N_h) go to infinity $d_k, N_h \to \infty$. We denote the kernel (NTK or NNGP) by $k(x, y)$. The kernel view allows us to study the inductive bias through the continuum limit (Canatar et al., 2021b; Cohen & Welling, 2016; Simon et al., 2023), where the kernel admits an eigenfunction decomposition and symmetries are explicitly manifested. In the continuum setting, predictions can be made using the kernel regression formula on the eigenbasis of the kernel operator ($\hat{K}$)

$$\hat{f}(X_*) = \sum_{i=1}^{\infty} \frac{\lambda_i}{\lambda_i + \delta/N} g_i \varphi_i(X_*); \qquad \begin{array}{l} \hat{K}\varphi_i(X) = \mathbb{E}_{Y\sim p_{\mathrm{train}}}\left[k(X,Y)\varphi_i(Y)\right] = \lambda\varphi_i(X) \\ g_i = \langle g(x), \varphi_i(x)\rangle_x = \mathbb{E}_{x\sim p_{\mathrm{data}}}\left[g(x)\varphi_i(x)\right] \end{array}, \tag{1}$$

where δ is the ridge or the effective ridge parameter (Canatar et al., 2021b; Cohen et al., 2021); 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 defined above. We can give equation 1 an intuitive interpretation: The architecture and dataset dictates the *learnability*. All eigenfunctions corresponding to $\lambda = 0$ will not be expressible by the NN, while eigenfunctions corresponding to $\lambda \neq 0$ will require $N \sim \sigma^2/\lambda$ samples to be learned. Accordingly, predicting the learning curves of the network is reduced to solving the eigenvalue problem for the kernel operator corresponding to the network and finding the projections of the target on the eigenbasis.

For the NN described in Sec.2, the fact that the network 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 token but the last one.

3.1 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 1 above. We say an operator like $\hat{K}$ is symmetric under the action of a group G if

$$\forall g \in G,\ \ k(\vec{x}_g, \vec{y}_g) = k(\vec{x}, \vec{y})\ \&\ p_{\text{data}}(\vec{x}_g) = p_{\text{data}}(\vec{x}), \tag{2}$$

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. 1 . A symmetry, as described in equation 2, 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 1, 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 $\varphi_j(x)$ 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 be members of such degenerate blocks. See Fig 1.

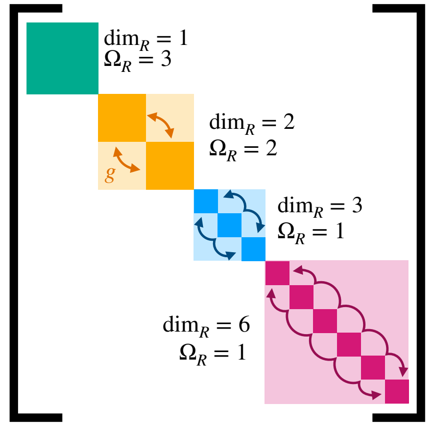

Figure 1: **(Illustration of diagonalization using symmetries)** The figure illustrates the direct sum (block) structure described in Prop. 1. Each color-shaded block represents an irrep, and each solid color represents a multiplicity block within the irrep. All elements outside the multiplicity blocks 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 2. Formally, the blocks correspond to the irreps of the group over the space of expressible functions (see Prop. 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 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 2 with $x_g = T_g x$.* **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.

Proposition 2. *Under the same conditions as Prop 1 and given the kernel is normalized, the trace is given by* $\mathbb{E}_{x\sim p_{\mathrm{data}}}[k(x,x)] \simeq 1$. *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.

Focusing back on our model, We can now state symmetry formally as symmetry under the action of the symmetric group in L symbols S_L i.e. $k(T_s X, T_{s_L} Y) = k(X, Y)$ where T_s is a representation of any element $s \in S_L$ that acts naturally on the sequence index [1]. Following Prop. 1,2 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. 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 3.1. *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 $(L-k, k)$ is $\dim_k = \frac{L!}{k!\frac{(L-k+1)!}{L-2k+1}} \sim L^k$. 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). We thus see that the sample complexity of a function in the representation $(L-k, k)$ is asymptotically bounded from below by $N \simeq \lambda_{(L-k,k)}^{-1}\sigma^2 = \Omega(L^k)$. We therefore see that the more symmetric a function is to permutations (smaller k) the more learnable it is.

4 Experimental Results

In this section, our theory is compared to numerical experiments. We start by comparing our predictions for the example of linear activation functions ($\Phi(x) = x/L,\ \phi(x) = x$) with exact Bayesian inference using the NNGP. We predict the performance OOD and show good agreement with experiments. We then present the NNGP kernel's spectrum of an NN with standard $\mathrm{softmax}$ attention and show that the scaling law bounds derived on the eigenvalues are 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.

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. Detailed analytical calculations for this case are given in appendix F.

In the center panel of Fig.2 we see the spectrum of the kernel, for a NN with $\Phi = \mathrm{softmax}$ and $\phi(x) = x$. 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.

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 G)

[1]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.

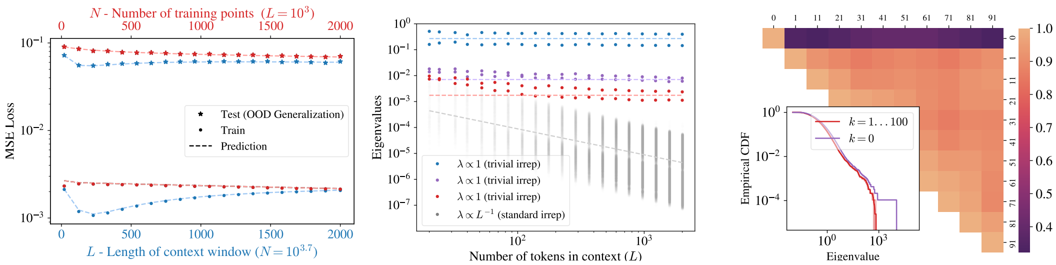

Figure 2: **Left: (theory vs. experiment)** Loss as a function of L (in blue) and N (in red) for a NN with linear attention. We find good agreement between our theoretical predictions (calculated for the train and test distributions) and exact Bayesian inference with the NNGP kernel, equivalent to inference 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: (kernel eigenvalues scaling law)** The spectrum of the empirical NNGP kernel of a NN 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 permutation symmetry in WikiText)** The triangle shows the cosine similarity 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.

$$C^{kk'}_{ij} := \mathbb{E}_{X\sim \mathrm{WikiText-2}}\left[X^a_i V^{ak} X^b_j V^{bk'}\right]; \quad V^{ak} := \exp\left(i\frac{2\pi}{L}ak\right), \begin{array}{l} a = 1, ..., L \\ k = 0, ..., L-1 \end{array}. \tag{3}$$

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 and by their spectrum. As shown in Fig. 2 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. $k = 0, k \neq 0$).

5 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.

Critically, the above results depend on a permutation symmetric dataset while some settings do have this exact symmetry[2], 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 corrections in NL to second order, as a concrete mechanism for in-context learning can already appear there; we leave this for future work.

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 kernel limit. As recent works have shown (Seroussi et al., 2023; Pacelli et al., 2023), the kernel 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.

[2]For example the settings in in Power et al. (2022) and common setting in which in context learning has been studied (Von Oswald et al., 2023; Garg et al., 2022; Ahuja et al., 2023)

REFERENCES

Kabir Ahuja, Madhur Panwar, and Navin Goyal. In-Context Learning through the Bayesian Prism. June 2023. URL http://arxiv.org/abs/2306.04891. arXiv:2306.04891 [cs].

Gernot Akemann, Jinho Baik, and Philippe Di Francesco. *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.

Leonard E. Baum and Ted Petrie. 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.

Gaspard Beugnot, Julien Mairal, and Alessandro Rudi. 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.

Rishi Bommasani, Drew A. Hudson, Ehsan Adeli, Russ Altman, Simran Arora, Sydney von Arx, Michael S. Bernstein, Jeannette Bohg, Antoine Bosselut, Emma Brunskill, Erik Brynjolfsson, S. Buch, Dallas Card, Rodrigo Castellon, Niladri S. Chatterji, Annie S. Chen, Kathleen A. Creel, Jared Davis, Dora Demszky, Chris Donahue, Moussa Doumbouya, Esin Durmus, Stefano Ermon, John Etchemendy, Kawin Ethayarajh, Li Fei-Fei, Chelsea Finn, Trevor Gale, Lauren E. Gillespie, Karan Goel, Noah D. Goodman, Shelby Grossman, Neel Guha, Tatsunori Hashimoto, Peter Henderson, John Hewitt, Daniel E. Ho, Jenny Hong, Kyle Hsu, Jing Huang, Thomas F. Icard, Saahil Jain, Dan Jurafsky, Pratyusha Kalluri, Siddharth Karamcheti, Geoff Keeling, Fereshte Khani, O. Khattab, Pang Wei Koh, Mark S. Krass, Ranjay Krishna, Rohith Kuditipudi, Ananya Kumar, Faisal Ladhak, Mina Lee, Tony Lee, Jure Leskovec, Isabelle Levent, Xiang Lisa Li, Xuechen Li, Tengyu Ma, Ali Malik, Christopher D. Manning, Suvir P. Mirchandani, Eric Mitchell, Zanele Munyikwa, Suraj Nair, Avanika Narayan, Deepak Narayanan, Benjamin Newman, Allen Nie, Juan Carlos Niebles, Hamed Nilforoshan, J. F. Nyarko, Giray Ogut, Laurel Orr, Isabel Papadimitriou, Joon Sung Park, Chris Piech, Eva Portelance, Christopher Potts, Aditi Raghunathan, Robert Reich, Hongyu Ren, Frieda Rong, Yusuf H. Roohani, Camilo Ruiz, Jack Ryan, Christopher R'e, Dorsa Sadigh, Shiori Sagawa, Keshav Santhanam, Andy Shih, Krishna Parasuram Srinivasan, Alex Tamkin, Rohan Taori, Armin W. Thomas, Florian Tramèr, Rose E. Wang, William Wang, Bohan Wu, Jiajun Wu, Yuhuai Wu, Sang Michael Xie, Michihiro Yasunaga, Jiaxuan You, Matei A. Zaharia, Michael Zhang, Tianyi Zhang, Xikun Zhang, Yuhui Zhang, Lucia Zheng, Kaitlyn Zhou, and Percy Liang. On the opportunities and risks of foundation models. *ArXiv*, 2021. URL https://crfm.stanford.edu/assets/report.pdf.

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

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

Abdulkadir Canatar, Blake Bordelon, and Cengiz Pehlevan. 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.

Mark Chen, Alec Radford, Rewon Child, Jeffrey Wu, Heewoo Jun, David Luan, and Ilya Sutskever. 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.

Omry Cohen, Or Malka, and Zohar Ringel. 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.

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

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale, June 2021. URL http://arxiv.org/abs/2010.11929. arXiv:2010.11929 [cs].

William Fulton and Joe Harris. *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.

Shivam Garg, Dimitris Tsipras, Percy Liang, and Gregory Valiant. 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.

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

Anirudh Goyal and Yoshua Bengio. 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.

Tom Henighan, Jared Kaplan, Mor Katz, Mark Chen, Christopher Hesse, Jacob Jackson, Heewoo Jun, Tom B. Brown, Prafulla Dhariwal, Scott Gray, Chris Hallacy, Benjamin Mann, Alec Radford, Aditya Ramesh, Nick Ryder, Daniel M. Ziegler, John Schulman, Dario Amodei, and Sam McCandlish. Scaling Laws for Autoregressive Generative Modeling, November 2020. URL http://arxiv.org/abs/2010.14701. arXiv:2010.14701 [cs].

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

Arthur Jacot, Franck Gabriel, and Clement Hongler. Neural Tangent Kernel: Convergence and Generalization in Neural Networks. In *Advances in Neural Information Processing Systems*, volume 31. Curran Associates, Inc., 2018. URL https://proceedings.neurips.cc/paper_files/paper/2018/hash/5a4be1fa34e62bb8a6ec6b91d2462f5a-Abstract.html.

Albert Q. Jiang, Alexandre Sablayrolles, Antoine Roux, Arthur Mensch, Blanche Savary, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Emma Bou Hanna, Florian Bressand, Gianna Lengyel, Guillaume Bour, Guillaume Lample, Lélio Renard Lavaud, Lucile Saulnier, Marie-Anne Lachaux, Pierre Stock, Sandeep Subramanian, Sophia Yang, Szymon Antoniak, Teven Le Scao, Théophile Gervet, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. Mixtral of Experts, January 2024. URL http://arxiv.org/abs/2401.04088. arXiv:2401.04088 [cs].

Jared Kaplan, Sam McCandlish, Tom Henighan, Tom B. Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeffrey Wu, and Dario Amodei. Scaling Laws for Neural Language Models, January 2020. URL http://arxiv.org/abs/2001.08361. arXiv:2001.08361 [cs, stat].

Hermann König. *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.

Jaehoon Lee, Yasaman Bahri, Roman Novak, Samuel S. Schoenholz, Jeffrey Pennington, and Jascha Sohl-Dickstein. Deep Neural Networks as Gaussian Processes, March 2018. URL http://arxiv.org/abs/1711.00165. arXiv:1711.00165 [cs, stat].

Jaehoon Lee, Samuel S. Schoenholz, Jeffrey Pennington, Ben Adlam, Lechao Xiao, Roman Novak, and Jascha Sohl-Dickstein. Finite Versus Infinite Neural Networks: an Empirical Study, September 2020. URL http://arxiv.org/abs/2007.15801. arXiv:2007.15801 [cs, stat].

Aitor Lewkowycz, Yasaman Bahri, Ethan Dyer, Jascha Sohl-Dickstein, and Guy Gur-Ari. 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].

Qianyi Li and Haim Sompolinsky. 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.

Ziming Liu, Ouail Kitouni, Niklas S. Nolte, Eric Michaud, Max Tegmark, and Mike Williams. 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.

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

Stephen Merity, Caiming Xiong, James Bradbury, and Richard Socher. Pointer Sentinel Mixture Models. In *International Conference on Learning Representations*, November 2016. URL https://openreview.net/forum?id=Byj72udxe.

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

Andrew J. Nam, Mustafa Abdool, Trevor Maxfield, and James L. McClelland. 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].

Neel Nanda, Lawrence Chan, Tom Lieberum, Jess Smith, and Jacob Steinhardt. Progress measures for grokking via mechanistic interpretability, October 2023. URL http://arxiv.org/abs/2301.05217. arXiv:2301.05217 [cs].

Gadi Naveh, Oded Ben David, Haim Sompolinsky, and Zohar Ringel. 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.

Radford M. Neal. Priors for Infinite Networks. In Radford M. Neal (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.

Catherine Olsson, Nelson Elhage, Neel Nanda, Nicholas Joseph, Nova DasSarma, Tom Henighan, Ben Mann, Amanda Askell, Yuntao Bai, Anna Chen, Tom Conerly, Dawn Drain, Deep Ganguli, Zac Hatfield-Dodds, Danny Hernandez, Scott Johnston, Andy Jones, Jackson Kernion, Liane Lovitt, Kamal Ndousse, Dario Amodei, Tom Brown, Jack Clark, Jared Kaplan, Sam McCandlish, and Chris Olah. 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].

R. Pacelli, S. Ariosto, M. Pastore, F. Ginelli, M. Gherardi, and P. Rotondo. 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.

Marc Potters and Jean-Philippe Bouchaud. *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.

Alethea Power, Yuri Burda, Harri Edwards, Igor Babuschkin, and Vedant Misra. Grokking: Generalization Beyond Overfitting on Small Algorithmic Datasets, January 2022. URL http://arxiv.org/abs/2201.02177. arXiv:2201.02177 [cs].

Noa Rubin, Inbar Seroussi, and Zohar Ringel. 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].

Bruce E. Sagan. *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.

Inbar Seroussi, Gadi Naveh, and Zohar Ringel. 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.

James B. Simon, Madeline Dickens, Dhruva Karkada, and Michael Deweese. 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.

Peter Sollich and Christopher Williams. Using the Equivalent Kernel to Understand Gaussian Process Regression. In *Advances in Neural Information Processing Systems*, volume 17. MIT Press, 2004. URL https://proceedings.neurips.cc/paper/2004/hash/d89a66c7c80a29b1bdbab0f2a1a94af8-Abstract.html.

Wu-Ki Tung. *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.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is All you Need. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017. URL https://papers.nips.cc/paper_files/paper/2017/hash/3f5ee243547dee91fbd053c1c4a845aa-Abstract.html.

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

Kaiyue Wen, Yuchen Li, Bingbin Liu, and Andrej Risteski. (Un)interpretability of Transformers: a case study with Dyck grammars. June 2023. URL https://openreview.net/forum?id=kaILSVAspn#all.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander M. Rush. HuggingFace's Transformers: State-of-the-art Natural Language Processing, July 2020. URL http://arxiv.org/abs/1910.03771. arXiv:1910.03771 [cs].

Sang Michael Xie, Aditi Raghunathan, Percy Liang, and Tengyu Ma. 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.

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

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

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 $\vec{v}_k = [e^{2\pi ik/N}, e^{4\pi ik/N}, ..., 1]$. 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 3.1, 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

$$\begin{array}{ccccc} \begin{ytableau} 1 & 3 & 5 & 6 \\ 2 & 4 \end{ytableau} & \begin{ytableau} 1 & 3 & 4 & 6 \\ 2 & 5 \end{ytableau} & \begin{ytableau} 1 & 3 & 4 & 5 \\ 2 & 6 \end{ytableau} & \begin{ytableau} 1 & 2 & 5 & 6 \\ 3 & 4 \end{ytableau} & \begin{ytableau} 1 & 2 & 4 & 6 \\ 3 & 5 \end{ytableau} \\ \begin{ytableau} 1 & 2 & 4 & 5 \\ 3 & 6 \end{ytableau} & \begin{ytableau} 1 & 2 & 3 & 6 \\ 4 & 5 \end{ytableau} & \begin{ytableau} 1 & 2 & 3 & 5 \\ 4 & 6 \end{ytableau} & \begin{ytableau} 1 & 2 & 3 & 4 \\ 5 & 6 \end{ytableau} & \end{array} \tag{B.2}$$

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[3] [4]. 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

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

[4] In the next appendix, where we prove theorem 3.1 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.

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.

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 ($(d-1,1)$)

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

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

Definition 2 (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

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

Definition 3 (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:

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

Definition 4 (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_λ^b and Θ_λ^C in equation C.2 are standard Young tableaux but Θ_λ^a is not.

Definition 5 (Canonical Young Tableau)**.** A canonical Young tableau Θ_λ^C 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_λ^C in equation C.2 is the canonical Young tableau.

Definition 6 (Rows and columns subgroups)**.** Given a Young tableau Θ_λ^p of partition λ and assignment p, we define the rows subgroup $\mathcal{R}_\lambda^p$ which leave invariant the (unordered) sets of numbers appearing in the same row of Θ_λ^p. Similarly, we define columns subgroup $\mathcal{C}_\lambda^p$ which leave invariant the (unordered) sets of numbers appearing in the same column of Θ_λ^p.

Definition 7 (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

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

Definition 8 (Row symmetrizer, column anti-symmetrizer and young symmetrizer)**.** Define the row symmetrizer, column anti-symmetrizer and Young symmetrizer linear operators:

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

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

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

Theorem 2. *Young symmetrizers associated with standard Young tableaux are projectors to irrep spaces of the symmetric group (Fulton & Harris, 2004)*

Lemma 1. *If there exists a transposition $t^* \in C_\lambda^p$ that leaves a monomial M unchanged, M vanishes under the action of the column anti-symmetrizer -*

$$\exists t^* \in C_\lambda^p \text{ s.t. } t^* M = M \rightarrow \hat{C}_\lambda^p M = 0.$$

Proof. Let Θ_λ^p be a standard Young tableau of a partition λ. Let $\hat{C}_\lambda^p$ be the column anti-symmetrizer associated with Θ_λ^p. Let $M(x_1, x_2, ..., x_n)$ be a multilinear monomial in the variables $x_1, x_2, ..., x_n$. Let $t^* \in C_\lambda^p$ 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

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

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 $\mathrm{sign}(c_i)c_i M = -\mathrm{sign}(c_j)c_j M$ that is

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

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

Lemma 2. *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 1, $M(x_1, x_2, ..., x_n)$ must vanish under the action. ■

Lemma 3. *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 1, the monomial must vanish under the action. ■

Remark. 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 4. *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 $0 \leq k \leq d, n-d$,

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

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 9 (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 5. *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 3.1. *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 2,3 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 4 and theorem 2 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 5)

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

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

D Model details

D.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 $\vec{x}^s$ indexed by an upper sequence index $s = 1, 2, ..., L+1$ with each token having an internal (vocabulary or embedding) 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 $X^s_{i,\mu}$, where we drop the sample index μ when we discuss only a single sample. One-hot encoding is used for the tokens, such that $[\vec{x}^s]_i = \delta_{i,v}$ where $v = 1, ..., N_{\text{voc}}$ is the token represented by $\vec{x}^s$. Denoting the input by x^a_i and the output of l'th layer by $z^{(l),a}_i$ the resulting NN is

$$\begin{aligned}
z_i^{(1),a} &= W_{ij}^{\text{emb}} x_j^a + p_i^a \\
z_{i,h}^{(2),a} &= \Phi\left(\frac{Q_{j,h}^a K_{j,h}^b}{\sqrt{d_k}}\right) V_{i,h}^b = \Phi\left(\frac{W_{lm,h}^Q z_m^{(1),a} W_{ln,h}^K z_n^{(1),b}}{\sqrt{d_k}}\right) W_{ij,h}^V z_j^{(1),b} \quad \text{(no } h \text{ summation)} \\
z_i^{(3),a} &= W_{ij,h}^O z_{j,h}^{(2),a} \\
z_i^{(4),a} &= \phi\left(W_{ij}^{(4)} z_j^{(3),a} + b_i^{(4)}\right) \\
z_i^{(5),a} &= W_{ij}^{(5)} z_j^{(4),a} + b_i^{(5)} \\
f_i^a(X) = z_i^{(6),a} &= W_{ij}^{\text{d-emb}} z_j^{(5),a}
\end{aligned} \tag{D.1}$$

using Einstein's summation convention, with Φ and ϕ being some activation functions[5]. The NN parameters

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

are all learned. For the MHA we use N_{heads} heads and the same dimension $d_k = d_v = d_{\text{model}}/N_{\text{heads}}$ 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 $d_{ff} \sim d_{\text{model}}$. Notably, consecutive affine transformations can be combined together without loss of generality, but they are kept in this way to align with standard notation[6].

As an instructive example, we will use a linearized MHA[7] $\Phi(x) = \frac{1}{L}x$ and linear MLP $\phi(x) = x$, as this setting allows for closed-form analytical predictions at the kernel limit. Note that because we remove the common softmax non-linearity we add a division by the length to make sure the network's output stays $O(1)$ and does not scale with L.

D.2 Task, Loss Function, and Initialization

The task is a pretraining task, namely, predicting the conditional probability distribution for the next token given the context $p(\vec{x}^{L+2}|X)$. For simplicity, we limit the discussion to inference-time-like

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

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

[7]similar to the one suggested by Von Oswald et al. (2023) and Hron et al. (2020)

output, i.e. when 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 $w \sim \mathcal{N}(0, \frac{1}{\sqrt{\text{fan}-\text{in}}})$, 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 $p_i^a \sim \mathcal{N}(0, 1/2)$ for $a \neq L+1$, for the last token we will initialize the PE to zero $p_i^{L+1} = 0$.

E Dataset and Hidden Markov Models

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^s and x^s, where s is the time-step index. The process h^s is dubbed "hidden" while x^s 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

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

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

F Linear Activations Example

In this example, we choose $\Phi(x) = \frac{1}{L+1}x$ and linear MLP $\phi(x) = x$, as previously noted in D.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.

F.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 is spanned by the linear functions of $\{x_1^s\}_{s=1}^{L}$ 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

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

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 F.1 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 I. Finally, the kernel can be simplified to a scalar expression. 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

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

As can be seen in equation 1, 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 $k(X, \cdot)$, i.e. it must be a linear combination of the functions $\{k(X, A)\}_A$ for some values of A. For example, if $k(X, Y)$ is linear in X only linear functions will be expressible. Based on this argument, 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}, which is a space of dimension $2L + 2$.

F.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 tools developed above to simplify it.

Claim 2. *For the model described above, the spectrum of the kernel operator is composed of four leading eigenvalue $\lambda_{0,*}, \lambda_{1,*} \sim 1$ belonging to the trivial irrep, two sub leading eigenvalues $\lambda_{2,*} \sim L^{-1}$ (again belonging to the trivial irrep) and four sets of size $L/2 - 1$ belonging to the standard irrep. All the eigenvalues in each of the four sets are exactly degenerate $\lambda_{k,*}^{\text{even}}, \lambda_{k,*}^{\text{odd}} \sim L^{-2}$, where $* = \{a, b\}$ and $k = 1, ..., L/2-1$. Furthermore, the exact eigenvectors corresponding to $\lambda_{k,*}^{\text{even}}, \lambda_{k,*}^{\text{odd}}$ are given in closed from by equation K.1.*

Starting from the largest structure, notice the kernel is a product of two terms ($\mathfrak{A}, \mathfrak{B}$ in equation F.2). The $\mathfrak{A}$ part is diagonalized in the basis

$$a(\vec{x}^{L+1}) = x_1^{L+1}, \;\; b(\vec{x}^{L+1}) = (1 - x_1^{L+1}), \tag{F.3}$$

which leaves us with a large block structure; we should expect to find two copies of each eigenvector, one belonging to the a block and one to the b block.

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 $\{x^s\}_{s=1}^L$. The full S_L symmetry is not, however, presented in the probability distribution of our chosen dataset[8], 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 $S_{L/2}^{\text{odd}} \times S_{L/2}^{\text{even}} := \mathcal{S}$ [9] as a symmetry of $\hat{K}$.

[8] Therefore it is not a symmetry of the operator $\hat{K}$.

[9] Assuming L is even for simplicity

As discussed in the previous subsection, in our case only polynomials up to first degree can have non-vanishing eigenvalues. First degree polynomials are decomposed to two irreps (see theorem 3.1), namely the trivial $(L/2)$ and standard representation $(L/2-1,1)$. The trivial representation, has dimension 1 with multiplicity 2 [10], and the standard representation, has dimension $L/2-1$ with multiplicity 2 [11]. For zeroth degree polynomials (constants) only the trivial representation exists, of multiplicity 1. Such a process can be done to an arbitrary polynomial degree as explained in appendix B.

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 K.1).

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

$$\lambda_{k,a}^{\text{odd}}, \lambda_{k,a}^{\text{even}}, \lambda_{k,b}^{\text{odd}}, \lambda_{k,b}^{\text{even}} \propto \frac{1}{L^2} \tag{F.4}$$

full expressions are given in equation K.8.

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

$$\tilde{\varphi}_0^{\text{odd}}(X) = \sum_{s=1}^{L/2} x_1^{2s-1};\ \ \tilde{\varphi}_0^{\text{even}}(X) = \sum_{s=1}^{L/2} x_1^{2s};\ \ \tilde{\varphi}_c(X) = 1. \tag{F.5}$$

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 K.2. 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.

Using symmetries and the partition to $\mathfrak{A}, \mathfrak{B}$ we were able to reduce the eigenvalue problem to two[12] 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.

F.3 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

$$g_*^+ = \frac{Lw^2\eta_*^+}{\sqrt{L^2w^2\rho_*^+ + L\sigma_*^+}}, \quad g_*^- = \frac{Lw^2\eta_*^- + \nu_*^-}{\sqrt{L^2w^4\rho_*^- + Lw^2\sigma_*^- + \xi_*^-}}, \tag{F.6}$$

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

Gathering the results of this section, ***Given***: (1) equation 1, together with the (2) learnable target given in equation K.9, the (3) eigendecomposition given in equations K.1, K.8, and the (4) eigendecomposition of the two 3×3 spaces spanned by the basis in equation K.2. ***One can*** predict accurately the output of the model described in section 2 with linear activation functions in the GP limit. Additionally, One can make accurate predictions for the generalization loss, even under a distributional shift.

[10] one for the even subspace and one for the odd subspace

[11] again broken down to $L/2-1$ from the even and odd subspaces

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

G 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

$$C_{ij}^{ab} := \mathbb{E}_{X\sim\text{WikiText}-2}\left[X_i^a X_j^b\right], \tag{G.1}$$

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

$$\tilde{C}_{ij}^{kk'} := \mathbb{E}_{X\sim\text{WikiText}-2}\left[X_i^a V^{ak} X_j^b V^{bk'}\right]; \tag{G.2}$$

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

One would then expect to find a block diagonal matrix where $\tilde{C}_{ij}^{kk'} = 0$ for $k \neq k'$ and $\tilde{C}_{ij}^{kk} = \tilde{C}_{ij}^{k'k'}$ 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 $\tilde{C}_{ij}^{kk} = \tilde{C}_{ij}^{k'k'}$ for $k, k' \in \{1, ..., L-1\}$, but we would have to consider the noise due to the finite sampling.

To measure whether $\tilde{C}_{ij}^{kk} = \tilde{C}_{ij}^{k'k'}$ 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).

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

$$\mathcal{C}_{ijj}^{abc} := \mathbb{E}_{X\sim\text{WikiText}-2}\left[X_i^a X_j^b X_k^c\right]. \tag{G.4}$$

H Out of distribution predictions under Equivalent Kernel approximation

Under Equivalent Kernel (EK) approximation Sollich & Williams (2004); Cohen et al. (2021) MSE loss can be computed by

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

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

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

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.

Now if we wish to compute the loss under distributional shift all we have to do is take the expectation value w.r.t. a new distribution

$$\begin{aligned}&\mathbb{E}_{X\sim p_{\mathrm{test}}}\mathbb{E}_{\Theta}\left[\left(f_{\Theta}\left(X\right)-g\left(X\right)\right)^{2}\right]\approx\\&\approx\sum_{i}\sum_{j}\frac{\lambda_{i}}{\lambda_{i}+\sigma^{2}/N}\frac{\mu_{j}}{\mu_{j}+\sigma^{2}/N}g_{i}g_{j}\left\langle\varphi_{i},\varphi_{j}\right\rangle_{X\sim p_{\mathrm{test}}}-2\sum_{i}\frac{\lambda_{i}}{\lambda_{i}+\sigma^{2}/N}g_{i}\left\langle\varphi_{i},g\right\rangle_{X\sim p_{\mathrm{test}}}+\left\langle g,g\right\rangle_{X\sim p_{\mathrm{test}}}.\end{aligned} \tag{H.3}$$

Notably, the eigenfunctions that were orthonormal under the inner product induced by the training distribution are no longer necessarily orthonormal under the test distribution.

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

The terms left out during the approximation are

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

All the vectors $\varphi_{k,a}^{\mathrm{odd}}(X),\ \varphi_{k,a}^{\mathrm{even}}(X)\ ,\varphi_{k,b}^{\mathrm{odd}}(X),\ \varphi_{k,b}^{\mathrm{even}}(X)$ 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 $O\left(1\right)$.

J Large structure decomposition and non-linearities

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

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

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

$$\hat{K}\left(\varphi_{i}^{L+1}\varphi_{j}^{L}\right)=\lambda_{ij}\left(\varphi_{i}^{L+1}\varphi_{j}^{L}\right), \tag{J.2}$$

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 $\{x^{s}\}_{s=1}^{L}$.

K FULL EXPRESSIONS OF QUANTITIES IN THE MAIN TEXT

Here we provide the full expressions for some of the quantities defined in the main text. The eigenvectors of the kernel that belong to the standard irrep are given by

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

The basis chosen for the trivial representation is

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

with

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

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

$$\begin{aligned}
\alpha_b =& -\frac{-24(p_a-1)(q_a-1)(p_a+q_a-2)(p_a+q_a) + \dots}{48(p_a+q_a+w-2)} \\
& \frac{\dots - 12w\left({p_a}^3 + {p_a}^2(7q_a-8) + p_a(q_a-2)(7q_a-6) + (q_a-6)(q_a-2)q_a - 4\right) + \dots}{48(p_a+q_a+w-2)} \\
& \frac{\dots + 2w^2\left(L((p_a-2)p_a + (q_a-2)q_a + 2) - 2\left(8{p_a}^2 + p_a(22q_a-29) + q_a(8q_a-29) + 20\right)\right) + \dots}{48(p_a+q_a+w-2)} \\
& \frac{\dots + 2w^3(L(p_a+q_a-2) - 2(7p_a+7q_a-11)) + (L-8)w^4}{48(p_a+q_a+w-2)}
\end{aligned} \tag{K.5}$$

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

$$\begin{aligned} c_a^{\mathrm{odd}} &= \frac{3\left({p_a}^2 + {q_a}^2\right) + 3w(p_a+q_a) + 2w^2}{3(p_a+q_a+w)} \\ c_a^{\mathrm{even}} &= \frac{(2p_a+w)(2q_a+w)}{2(p_a+q_a+w)} \\ c_b^{\mathrm{odd}} &= \frac{3w(p_a+q_a-1) + 3(p_a-1)p_a + 3(q_a-1)q_a + 2w^2}{3(p_a+q_a+w-2)} \\ c_b^{\mathrm{even}} &= \frac{2p_a(2q_a+w-1) + 2q_a(w-1) + (w-2)w}{2(p_a+q_a+w-2)} \end{aligned} \tag{K.7}$$

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

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

$$\begin{aligned} g_{k,*}^{\mathrm{odd}} &= 0, & g_{k,*}^{\mathrm{even}} &= 0 \\ g_*^{+} &= \frac{Lw^2\eta_*^{+}}{\sqrt{L^2w^2\rho_*^{+} + L\sigma_*^{+}}}, & g_*^{-} &= \frac{Lw^2\eta_*^{-} + \nu_*^{-}}{\sqrt{L^2w^4\rho_*^{-} + Lw^2\sigma_*^{-} + \xi_*^{-}}}, \\ g_{c,a} &= \frac{p_a q_a}{\sqrt{\frac{p_a+q_a}{2}}}, & g_{c,b} &= \frac{q_a + p_a - 2p_a q_a}{\sqrt{2\left(1-p_a+1-q_a\right)}}; \end{aligned} \tag{K.9}$$

with

$$\begin{aligned} \eta_{0,a}^{+} &= 2\left({p_a}^2 + {q_a}^2\right) \\ \rho_{0,a}^{+} &= 48(p_a+q_a)^3 \\ \sigma_{0,a}^{+} &= -576(p_a+q_a)^3((p_a-1)p_a + (q_a-1)q_a) \end{aligned} \tag{K.10}$$

$$\begin{aligned} \eta_{0,b}^{+} &= 2(p_a-2)p_a + 2(q_a-2)q_a + 4 \\ \rho_{0,b}^{+} &= 2(p_a-2)p_a + 2(q_a-2)q_a + 4 \\ \sigma_{0,b}^{+} &= 576(p_a+q_a-2)^3((p_a-1)p_a + (q_a-1)q_a) \end{aligned} \tag{K.11}$$

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

$$
\begin{aligned}
\eta^-_{0,a} &= -72(p_a-1)(q_a-1)(p_a-q_a)^2(p_a+q_a-2)\\
\nu^-_{0,a} &= 864(p_a-1)(q_a-1)(p_a-q_a)^2(p_a+q_a-2)((p_a-1)p_a+(q_a-1)q_a)\\
\rho^-_{0,a} &= -10368(p_a-1)(q_a-1)(p_a-q_a)^2(p_a+q_a-2)^3\\
\sigma^-_{0,a} &= 248832(p_a-1)(q_a-1)(p_a-q_a)^2(p_a+q_a-2)^3((p_a-1)p_a+(q_a-1)q_a)\\
\xi^-_{0,a} &= -1492992(p_a-1)(q_a-1)(p_a-q_a)^2(p_a+q_a-2)^3((p_a-1)p_a+(q_a-1)q_a)^2
\end{aligned}
\tag{K.13}
$$