Phase Transitions in Attention: A Bayesian Theory of Copy Head Emergence

Itay Lavie[1,2*†] **Kirsten Fischer**[3,4*] **Andrey Lekov**[5] **Frederic Van Maele**[5]
Zohar Ringel[1] **Moritz Helias**[3,6]

[1]*Racah Institute of Physics, Hebrew University of Jerusalem, Jerusalem 91904, Israel*
[2]*John A. Paulson School of Engineering and Applied Sciences, Harvard University, Cambridge, MA 02138, USA*
[3]*Institute for Advanced Simulation (IAS-6), Computational and Systems Neuroscience, Jülich Research Center, Jülich 52428, Germany*
[4]*Institute of AI for Health, Helmholtz Munich, Neuherberg, Germany*
[5]*RWTH Aachen University, Aachen, Germany*
[6]*Department of Physics, Faculty 1, RWTH Aachen University, Aachen, Germany*

Abstract

Attention is the key mechanism underlying in-context learning in transformers, and attention patterns have been observed empirically to emerge abruptly during training. We present a Bayesian theory of feature learning in attention; we then focus on how the copy subcircuit in the first layer of an induction head is learned by analyzing a single-layer softmax attention network trained on a copy task. We derive a closed-form posterior over the attention matrix and reduce it to a low-dimensional order parameter space. This reduction reveals a phase transition in the amount of training data, which we verify using both Bayesian sampling and standard training with Adam. We contrast our results with linear attention and find that softmax attention exhibits a *first-order phase transition* while in linear attention an initial *second-order phase transition* is followed by a smooth, continuous evolution toward the structured attention pattern (*crossover*). Our work provides a first-principles theoretical account of the abrupt emergence of the copy subcircuit, reminiscent of the one observed in training large language models.

1 Introduction

In-context learning (ICL) has become a cornerstone of modern large language models (LLMs) and coding agents [1], enabling models to solve and learn tasks at inference time without explicit training. Yet, while loss scaling with model size and data is well understood empirically [2, 3], the ability to predict the development of capabilities such as ICL lags far behind [4].

Whether such capabilities emerge abruptly or gradually has sparked much debate in recent years, with some works claiming some capabilities are emergent and evolve abruptly [5, 4] and others arguing that these results may be artifacts of the evaluation metric [6, 7]. The former possibility has far-reaching consequences: if capabilities are acquired without prior signal, models may cross capability thresholds that require additional safeguards [8] before these are in place. This possibility has motivated a line of work seeking progress measures that track capability acquisition continuously [9, 10, 11, 12].

ICL capabilities have been extensively studied in two tractable synthetic settings: in-context regression, where, given input-output pairs of sequences in context, the model has to predict the output for

*Equal contribution.
†Correspondence to: itaylavie@g.harvard.edu.

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

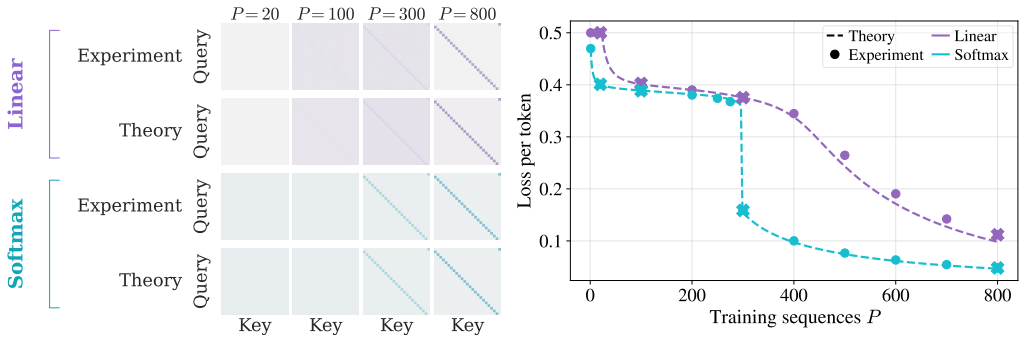

Figure 1: Copy head formation unfolds in multiple stages, with qualitatively distinct behavior for linear and softmax attention. **Left:** Attention patterns for linear (purple) and softmax (blue) attention from the trained model (top per color) and theory (bottom per color) for different amounts of training data P. (i) $P = 20$: the attention layer does not contribute in the linear case and softmax attention remains at its uniform pooling default; (ii) $P = 100$: after the second-order phase transition for linear attention, all tokens are attended to equally; (iii) $P = 300$: the copy pattern begins to appear; (iv) $P = 800$: it sharpens into the correct shift-by-one pattern. **Right:** Loss per token as a function of training set size P for linear (purple) and softmax (blue) attention. Softmax attention exhibits a sharp, discontinuous loss drop near $P \approx 300$ corresponding to a first-order phase transition. Linear attention shows a fast continuous decrease at small P (second-order transition) followed by a plateau and a slower decrease (gradual crossover). Crosses mark points on the left.

a new input [13, 14, 15, 16]

$$[x_1,\, f(x_1),\, x_2,\, f(x_2) \ldots x_L] \to f(x_L), \tag{1}$$

and in-context bigrams

$$[\ldots A,\, B \ldots A] \to B, \tag{2}$$

where the model learns context-dependent token statistics via induction heads [17, 18]. Training has been observed to evolve abruptly in both settings, for in-context regression [19, 20] as well as for in-context bigrams [21]. Remarkably, the latter was observed to be essential for a variety of ICL capabilities in LLMs [18, 22, 23] and the abrupt formation of induction heads is often termed *emergent* and associated with a *phase transition* [18, 12, 21, 24, 25, 26].

In this work, we present a Bayesian theory of feature learning in a single-head attention network and analyze how the first layer of an induction head, namely the copying subcircuit, $[\ldots A,\, B] \to A$ is learned. Marginalizing over the network weights yields a posterior over attention patterns whose dominant behavior is captured by two scalar order parameters measuring the competition between uniform and copy attention. Softmax and linear attention undergo qualitatively different transitions (first-order in the former, second-order followed by a crossover in the latter) each producing distinct loss signatures (see Fig. 1).

Our main contributions are:

- We develop a novel analytical approach to Bayesian feature learning for a single attention head: We derive a closed-form posterior over the attention matrix and reduce it to a low-dimensional order parameter space, revealing how competing attention patterns drive a phase transition in the amount of training data.
- We find that the attention activation determines the order of the transition and thus the character of the capability emergence: softmax exhibits an abrupt first-order phase transition while linear attention shows a continuous second-order transition followed by a gradual crossover, providing a first-principles account of the abrupt emergence of the copy subcircuit in softmax transformers and predicting less sharp behavior in linear transformers.
- The order parameters map directly onto measurable loss signatures, connecting the theoretical transition types to empirically observable training metrics including the unigram-to-bigram learning progression reported in prior work [21].

1.1 Related Work

Theoretical limits of transformers. A first line of work studies transformers in the lazy limit, where no feature learning occurs: [27] extended the NNGP correspondence to attention, showing single-head attention departs from Gaussianity, with μP scaling or many heads restoring it. [28] characterized the spectral bias in the lazy limit through symmetry. [29, 30] studied signal propagation in transformers and [31] the diversity of attention heads at infinite width and heads. A second line studies tractable attention models in high dimensions: [32] identify a sample-complexity-driven transition between positional and semantic learning, [33] derive Bayes-optimal and algorithmic thresholds for deep attention networks formulated as sequence multi-index models, [34, 35] extend this to extensive-width attention matrices for Bayes-optimal and empirical-risk learning, respectively. While these works focus on characterizing learning in weight space, we focus on a specific capability that is essential for ICL and characterize it by an effective posterior over interpretable attention patterns.

Bayesian feature learning in attention. Bayesian treatments of feature learning have studied Gaussian kernels that adapt to data [36, 37, 38] and mean-field theories of strong feature learning, including grokking phase transitions [39]. Closer in spirit is [40] which focused on the interplay of multiple heads while freezing the attention patterns themselves. In contrast, our analysis goes beyond Gaussian features: we find a non-Gaussian distribution for attention activations and treat the layer coupling exactly, taking the attention pattern of an individual head as the primary degree of freedom rather than freezing it.

Training dynamics of transformers on synthetic tasks. [21] showed that two-layer transformers trained on Markov sequences undergo three learning stages: (i) no in-context information used; (ii) in-context unigram statistics used; (iii) in-context bi-gram statistics used. Our theoretical results echo these empirical findings: (i) the attention is ineffective; (ii) we see a pooling like behavior where all the context is attended equally, matching unigram statistics; (iii) only the specific position is attended to, matching the bigram statistics stage (see Fig 1). [41] found similar stagewise dynamics (no context; in-context unigrams; in-context bigrams) in two-layer transformers on Omniglot. [42, 43] proved that transformers learn the causal Markov chain structure, with [42] identifying learning the first attention layer of an induction head, i.e. the copy subcircuit, as the main challenge. [44] studied the tradeoff between in-context and in-weights learning through associative memory. [45] characterize online SGD dynamics in a sequence single-index model and [46] the stagewise specialization of heads in multi-head softmax attention, complementing our equilibrium perspective.

Phase transitions and emergence of copy heads. [18] coined the term *induction head* and provided extensive empirical evidence for their abrupt emergence and causal role in ICL in LLMs across scales, with [12] analyzing this in a minimal setting. Recently, [47] gave sharp asymptotic timescales for the appearance of an induction head in a two-layer simplified linear transformer, finding two stages with tight timescales $\Theta(L)$ and $\Theta(L^2)$, corresponding to the same stages we find here. Such timescale characterization, however, cannot distinguish a true phase transition from a crossover, as both would have tight asymptotic timescales.

2 A Minimal Model for the Emergence of Copy Heads

Motivated by the emergence of induction heads in transformers [18], we focus on the copy or shift-by-one subcircuit that constitutes their first layer. To this end, we study the following copy task: inputs are sequences of i.i.d. one-hot encoded tokens $x_j^a \in \{0,1\}^{L\times V}$, where $1 \le a \le L$ with L being the sequence length, and V the vocabulary size. The target labels are $y_i^a(x) = x_i^{a-1}$ with periodic boundary condition.

We consider a simplified model of a single-layer attention network:

$$h_i^a = \sum_j W_{ij}^{\mathrm{emb}}\, x_j^a + p_i^a; \quad G_H^{ab} = \phi\!\left(\frac{\boldsymbol{h}^{a\top}\boldsymbol{W}_H^G\,\boldsymbol{h}^b}{\sqrt{d_{\mathrm{model}}}}\right)_b; \quad f_i^a(\boldsymbol{x}) = \sum_{j,H} W_{ij,H}^O/\sqrt{N_H}\sum_{b=1}^{L} G_H^{ab}\, h_j^b, \tag{3}$$

where ϕ denotes the attention activation: $\phi = \mathrm{softmax}$ for softmax attention, and $\phi(\cdot)_b = \frac{1}{L}(\cdot)_b$ for linear attention with a normalization factor $1/L$. The matrix $\boldsymbol{W}^{\mathrm{emb}}$ embeds the tokens into the model space of dimension d_{model} and p_i^a is the positional encoding. Each of the N_H heads has its own attention weights $\boldsymbol{W}_H^G$ and readout weights $\boldsymbol{W}_H^O$; the factor $1/\sqrt{N_H}$ keeps the scale of the output independent of the number of heads. We merge the standard matrices $\boldsymbol{W}^{Q\top}\boldsymbol{W}^K$ into a single matrix $\boldsymbol{W}^G$ as commonly done in theoretical works [44, 21, 42, 47]; in Appendix B we show that as long as $d_{\mathrm{model}},\ d_k \gg L$ the resulting prior over attention patterns when parameterizing with two separate matrices W^K, W^Q is the same, even though the prior over the product of weights $W^{K^\top}W^Q$ is different (see Appendix J). Numerically, both parameterizations with $d_k = d_{model}$ indeed yield the same attention patterns and losses (Fig. 5). We do not include causal masking as it is not needed for the copying task.

Linear attention has become a widely adopted setting for theoretical studies of transformers [48, 20] as it retains the token-mixing structure of softmax attention and reproduces many aspects of its optimization landscape [49, 15]. In this work, it serves as a controlled reference point: by studying both softmax and linear attention on the same task, we can isolate the effect of the attention activation on the phase transition order.

The full setup can be seen as a teacher-student problem with a teacher attention pattern given by $G^*_{ab} = \delta^{a-1,b}$ for both softmax and linear attention. The copy operation is the minimal circuit that is necessary for induction heads and other ICL capabilities [23], and our task isolates it by construction. While induction heads empirically emerge under unsupervised next-token prediction [18], we adopt a supervised setting to enable a tractable Bayesian analysis [37, 50], consistent with theoretical work using synthetic tasks to study transformers [15, 48, 20]. As we show below even this minimal subcircuit exhibits a rich phenomenology.

Scaling. We take a parametrization similar to μP scaling [51]: we downscale the attention weight scale $g_W = \bar{g}_W/\chi$, the output weight scale $g_O = \bar{g}_O/\chi$, and the observation noise $\sigma^2 \to \sigma^2/\chi$ for large $\chi \gg 1$. We also adopt the equivalent-kernel (EK) approximation, replacing the finite-dataset likelihood with the population likelihood, valid when $\sigma^2 \gg \sum_{a=1}^{L}(\boldsymbol{y}^a - \boldsymbol{f}^a(\boldsymbol{x}))^2$ [52, 50]. We first take $d_{\mathrm{model}} \to \infty$, and then take $L, \sigma^2 \to \infty$ jointly with $\tilde{\sigma}^2 := \sigma^2/L$ to maintain EK validity. Applying the χ transformation, we have $\bar{\sigma}^2 := \sigma^2\chi/L$ in total as an order one quantity. Finally, we keep χ large while remaining agnostic to its scaling with L. As shown below, the scale of P w.r.t. L varies across the problem; identifying the sample complexity of each transition is one of our main theoretical findings. We note the order of limits is important and the limits generally do not commute. Taking $d_{\mathrm{model}} \gg P, L$ ensures control of the embedding kernel fluctuations and concentrates it around its initialization value (but not the subsequent layers). Similarly, increasing σ^2, P jointly at fixed ratio improves the EK approximation, and vise versa, while changing one but not the other approaches a different point on the limiting curve. We probe the approximations experimentally in App. Q

3 Bayesian Theory of Copy Head Emergence as a Phase Transition

We develop a Bayesian framework over the attention matrix $\boldsymbol{G}$: marginalizing the network weights yields a closed-form posterior whose mode or maximum a-posteriori estimate (MAP) is the attention pattern most favored by data and prior at a given training-set size P, separately from whether an optimizer reaches it. The negative log-posterior $\mathcal{S}(\boldsymbol{G})$, which we refer to as the *action*, plays the role of an effective loss landscape on $\boldsymbol{G}$, which accounts for the number of microscopic weight configurations that produce the same attention pattern. We develop the framework for multi-layer attention in App. K and specialize here to a single-layer network.

The copy task's inherent permutation symmetry singles out two irreducible representations of the symmetric group, corresponding to two structurally distinct attention patterns: attending uniformly versus attending to the token that predicts the next one. This structure reduces the posterior over the attention matrix $\boldsymbol{G}$ to two corresponding scalar order parameters $\hat{c}_1$ and $\hat{c}_G$ from first principles; the order parameters are low-dimensional collective coordinates on which the posterior concentrates in the large-χL limit, so that the dominant attention pattern is determined by the minima of an effective log-likelihood over $(\hat{c}_1, \hat{c}_G)$. These order parameters exhibit phase transitions as a function of the training data amount P, ultimately driving the emergence of a copy subcircuit; see App. A for a primer on order parameters and phase transitions. We first derive the theory for linear attention

and characterize its phenomenology, then obtain the results for softmax attention by taking the pushforward of the resulting posterior under the softmax map.

3.1 Network Prior as a Gaussian Process over the Attention Pattern

Marginalizing the embedding $\boldsymbol{W}^{\text{emb}}$ and positional encoding $\boldsymbol{p}^a$ (full derivation in App. B) shows that the token embeddings $\boldsymbol{h}^a_{i,\alpha}$ are i.i.d. Gaussian in the model dimension i with zero mean, conditioned on their covariance (kernel). By a large-deviation argument, its empirical covariance concentrates in the $d_{\text{model}} \to \infty$ limit

$$Q^{ab}_{\alpha\beta} \coloneqq \frac{1}{d_{\text{model}}} \sum_{i=1}^{d_{\text{model}}} h^a_{i\alpha} h^b_{i\beta} \to C^{(xx)}_{ab,\alpha\beta} = g_{\text{emb}} \sum_j x^a_{j\alpha}\, x^b_{j\beta} + g_p\, \delta_{ab}. \tag{4}$$

As the merged key and query weights $\boldsymbol{W}^G_H$ are independent across heads, marginalizing them yields an i.i.d. conditional Gaussian prior on each head's attention scores $G^{ab}_{H,\alpha}$ with per-head covariance

$$C^{(G_H G_H)}_{ab,cd,\alpha\beta} = \frac{g_K g_Q}{L^2 d^2_{\text{model}}} \big(\boldsymbol{h}^a_\alpha \cdot \boldsymbol{h}^c_\beta\big)\big(\boldsymbol{h}^b_\alpha \cdot \boldsymbol{h}^d_\beta\big) \;\to\; \frac{g_K g_Q}{L^2}\, C^{(xx)}_{ac,\alpha\beta}\, C^{(xx)}_{bd,\alpha\beta}, \tag{5}$$

where we again use the concentration of $\boldsymbol{Q}$ to $\boldsymbol{C}^{(xx)}$. We get a Gaussian prior factorized over heads

$$p(\{\boldsymbol{G}_H\} \mid \boldsymbol{X}) \propto \prod_H \exp\!\left(-\frac{1}{2} \sum_{ab,cd,\alpha\beta} G^{ab}_{H,\alpha} \Big[\frac{g_W}{L^2} C^{(xx)}_{ac,\alpha\beta} C^{(xx)}_{bd,\alpha\beta}\Big]^{-1} G^{cd}_{H,\beta}\right), \quad g_W \coloneqq g_K g_Q. \tag{6}$$

Finally, integrating out the per-head output weights $\boldsymbol{W}^O_H$, which are independent across heads, the conditional distribution of the network outputs $\boldsymbol{F} = (\boldsymbol{f}^a_\alpha)_{\alpha=1,\dots,P;\, a=1,\dots,L}$ couples the heads only through the summed kernel and yields

$$p(\boldsymbol{F} \mid \{\boldsymbol{G}_H\}) \propto \int \mathcal{D}\tilde{\boldsymbol{f}}\; \exp\!\Big(-\tilde{f}^a_{i\alpha} f^a_{i\alpha} + \tfrac{g_O}{2N_H}\, \tilde{f}^a_{i\alpha} \tilde{f}^c_{i\beta} \sum_H G^{ab}_{H,\alpha} G^{cd}_{H,\beta}\, C^{(xx)}_{bd,\alpha\beta}\Big), \tag{7}$$

where repeated indices are summed over and $\tilde{\boldsymbol{f}}$ is an auxiliary field conjugate to $\boldsymbol{f}$, which arises from the Fourier representation of the Dirac δ-distribution enforcing the readout (see App. B) and is integrated over. Eq. (7) shows that, given the attention patterns, the outputs are Gaussian and independent across i, with covariance $\frac{g_O}{N_H} \sum_H G^{ab}_{H,\alpha} G^{cd}_{H,\beta} C^{(xx)}_{bd,\alpha\beta}$.

3.2 Network Posterior over the Attention Patterns

Assuming Gaussian noise on the outputs $\boldsymbol{y}_\alpha = \boldsymbol{f}_\alpha + \boldsymbol{\xi}$ with $\boldsymbol{\xi} \sim \mathcal{N}(0, \sigma^2 \boldsymbol{I})$, the posterior then is

$$p(\{\boldsymbol{G}_H\} \mid \{\boldsymbol{x}_\alpha, \boldsymbol{y}_\alpha\}_\alpha) \propto p(\{\boldsymbol{G}_H\} \mid \boldsymbol{X}) \int \mathcal{D}\boldsymbol{f}\, \exp\Big(-\frac{1}{2\sigma^2} \sum_\alpha \|\boldsymbol{y}_\alpha - \boldsymbol{f}_\alpha\|_2^2\Big)\, p(\boldsymbol{F} \mid \{\boldsymbol{G}_H\}), \tag{8}$$

where we dropped the normalization constant (full derivation in App. C). Inserting the results for the network prior and integrating out the network outputs $\boldsymbol{f}$ and their conjugate variables $\tilde{\boldsymbol{f}}$, we get

$$p(\{\boldsymbol{G}_H\} \mid \{\boldsymbol{x}_\alpha, \boldsymbol{y}_\alpha\}) \propto \exp\big(-\mathcal{S}(\{\boldsymbol{G}_H\})\big), \tag{9}$$

with the negative log-posterior or *action* being

$$\begin{aligned}\mathcal{S}(\{\boldsymbol{G}_H\}) = \frac{1}{2}\, \boldsymbol{y}^\top \Big[\tfrac{g_O}{N_H} \textstyle\sum_H \boldsymbol{G}_H \boldsymbol{C}^{(xx)} \boldsymbol{G}_H^\top + \sigma^2 \boldsymbol{I}\Big]^{-1} \boldsymbol{y} + \tfrac{1}{2} \ln\det\Big[\tfrac{g_O}{N_H} \textstyle\sum_H \boldsymbol{G}_H \boldsymbol{C}^{(xx)} \boldsymbol{G}_H^\top + \sigma^2 \boldsymbol{I}\Big] \\ + \sum_H \Big(\frac{1}{2}\, \boldsymbol{G}_H \big[\tfrac{g_W}{L^2} \boldsymbol{C}^{(xx)} \boldsymbol{C}^{(xx)}\big]^{-1} \boldsymbol{G}_H + \frac{1}{2} \ln\det\Big[\tfrac{g_W}{L^2} \boldsymbol{C}^{(xx)} \boldsymbol{C}^{(xx)}\Big]\Big).\end{aligned} \tag{10}$$

From this, we observe a competition between the network prior (last term) and the conditioning on the training labels (first two terms); the prior separates over heads, whereas the data terms couple them only through the total attention-weighted kernel $\frac{g_O}{N_H} \sum_H \boldsymbol{G}_H \boldsymbol{C}^{(xx)} \boldsymbol{G}_H^\top$.

3.3 Reduction of the Posterior to Interpretable Order Parameters

We now specialize to a single head $N_H = 1$. Empirically, the copy subcircuit that forms the first layer of an induction head localizes to individual attention heads [18, 17], so a single head is the natural granularity at which to study its emergence. Since the heads enter the prior independently and couple only through the summed kernel in the data term, a single head already contains the competition between uniform pooling and copying that is our object of interest for both linear and softmax attention; how the coupling between several heads affects the transition is left to future work.

Based on the permutation equivariance of the problem, we make the following representation theory-inspired ansatz for the attention matrix

$$\boldsymbol{G} = \hat{c}_1 \frac{\mathbf{1}_L}{L} + \hat{c}_G \frac{\boldsymbol{G}^*}{\sqrt{L}}, \tag{11}$$

where $\mathbf{1}_L$ is the $L \times L$ all-ones matrix and $G^*_{ab} = \delta^{a-1,b}$ is the *copy* or *shift-by-one* operator. Each term is weighted by a coefficient $(\hat{c}_1, \hat{c}_G)$, which we identify as the order parameters of the problem: $\hat{c}_G \gg \hat{c}_1$ implements previous token copying, while $\hat{c}_G \ll \hat{c}_1$ implements uniform pooling over the whole context, seeing only the single token empirical frequencies in context (in-context unigrams). The normalization factors give both modes unit Frobenius norm.

NNGP Prior Term. We evaluate $\mathcal{S}_{\mathrm{NNGP}} = \frac{L^2}{2g_W} \sum_{ab,cd,\alpha\beta} G^{ab}_{\alpha} \left[C^{(xx)} C^{(xx)}\right]^{-1}_{\substack{ac;bd \\ \alpha\beta}} G^{cd}_{\beta}$ by computing the summed pseudo-inverse $\Sigma_{(ab),(cd)} = \sum_{\alpha\beta} (\mathcal{M}^{\dagger})_{ab\alpha,cd\beta}$ of $\mathcal{M}_{ab\alpha,cd\beta} = C^{(xx)}_{ac,\alpha\beta} C^{(xx)}_{bd,\alpha\beta}$ (full derivation in App. D). For one-hot encoded inputs and training sets satisfying a rank condition (see App. D), the summed pseudo-inverse follows in closed form and is independent of P as $\Sigma_{(ab),(cd)} = \Pi_{ac}\, \Pi_{bd}$ with $\Pi = \boldsymbol{I} - \frac{1}{L+V} \mathbf{1}_L$.

Contracting the attention patterns with the summed pseudo-inverse gives

$$\mathcal{S}_{\mathrm{NNGP}}(\hat{c}_1, \hat{c}_G) = \frac{\chi L}{2\bar{g}_W} \Big[\frac{1}{L} \frac{V^2}{(1+V/L)^2} \hat{c}_1^2 + \frac{2}{L^{3/2}} \frac{V^2}{(1+V/L)^2} \hat{c}_1 \hat{c}_G + L \frac{(1+V/L)^2 - L^{-1} - 2VL^{-2}}{(1+V/L)^2} \hat{c}_G^2 \Big]. \tag{12}$$

Data Term We evaluate $\mathcal{S}_{\mathrm{data}} = \frac{1}{2} \boldsymbol{y}^{\top} [g_O \boldsymbol{G} \boldsymbol{C}^{(xx)} \boldsymbol{G}^{\top} + \sigma^2 \boldsymbol{I}]^{-1} \boldsymbol{y}$ by diagonalizing the operator $\boldsymbol{G} \boldsymbol{C}^{(xx)} \boldsymbol{G}^{\top}$ (full derivation in App. E). We choose a Fourier basis adapted to the copy task:

$$\begin{aligned} \left[v_{0,0}\right]^a_{\alpha} &= \frac{1}{\sqrt{PL}}, \qquad \left[v_{0,k\geq 1}\right]^a_{\alpha} = \frac{1}{\sqrt{PL^2}} \sum_b \sum_j e^{i\frac{2\pi}{V} kj}\, x^b_{j,\alpha}, \\ \left[v_{1,k\geq 1}\right]^a_{\alpha} &= \frac{1}{\sqrt{PL(1-L^{-1})}} \sum_j e^{i\frac{2\pi}{V} kj} \left(x^{a-1}_{j,\alpha} - \frac{1}{L} \sum_b x^b_{j,\alpha} \right). \end{aligned} \tag{13}$$

where $k = 1, \ldots, V-1$ and we note $\boldsymbol{v}_{1,0} = 0$. The mode $\boldsymbol{v}_{0,0}$ is uniform in both position and data; the modes $\boldsymbol{v}_{0,k\geq 1}$ are uniform in position but carry Fourier structure over the vocabulary; and the modes $\boldsymbol{v}_{1,k\geq 1}$ capture position-to-position correlations by isolating the deviation of token $a-1$ from the positional mean. We denote the data-dependent kernel without the noise terms $\sigma^2 \boldsymbol{I}$ as $\mathcal{K}^{ac}_{\alpha\beta} \coloneqq g_O \sum_{bd} G^{ab}_{\alpha} G^{cd}_{\beta} C^{(xx)}_{bd,\alpha\beta}$.

The projection of the kernel $\mathcal{K}$ onto the uniform mode $\boldsymbol{v}_{0,0}$ reduces to a scalar $\mathcal{K}_0$ given below. For each $k \geq 1$, the kernel acts on the two-dimensional space $(\boldsymbol{v}_{0,k}, \boldsymbol{v}_{1,k})$ as a 2×2 matrix. Computing the action of each of the four operator pairs $(\mathbf{1}_L, \mathbf{1}_L)$, $(\mathbf{1}_L, \boldsymbol{G}^*)$, $(\boldsymbol{G}^*, \mathbf{1}_L)$, $(\boldsymbol{G}^*, \boldsymbol{G}^*)$ on both modes and assembling with weights $\hat{c}_1^2$, $\hat{c}_1 \hat{c}_G$, $\hat{c}_G \hat{c}_1$, $\hat{c}_G^2$ respectively yields

$$\mathcal{K}_0 = g_O P \left(\frac{L}{V} + 1 \right) \left(\hat{c}_1 + \frac{\hat{c}_G}{\sqrt{L}} \right)^2; \qquad \mathcal{K}_k = \frac{g_O P}{V} \boldsymbol{u}\boldsymbol{u}^{\top}; \quad \boldsymbol{u} = \begin{pmatrix} \hat{c}_1 + \hat{c}_G/\sqrt{L} \\ \hat{c}_G \sqrt{(L-1)/L} \end{pmatrix}. \tag{14}$$

By inverting with the Sherman–Morrison formula and summing over all Fourier modes, we get

$$\mathcal{S}_{\mathrm{data}}(\hat{c}_1, \hat{c}_G) = \frac{\chi L}{2} \left[\frac{1}{L} \frac{1}{\bar{g}_O \left(\hat{c}_1 + \hat{c}_G/\sqrt{L} \right)^2 \left(1 + \frac{V}{L}\right) + \frac{\bar{\sigma}}{P}^2 V} - \frac{P}{L\bar{\sigma}^2} \frac{V-1}{V} \left(\frac{\bar{g}_O \left(\hat{c}_G + \hat{c}_1/\sqrt{L} \right)^2}{\bar{g}_O \|\boldsymbol{u}\|^2 + \frac{\bar{\sigma}^2 LV}{P}} - 1 \right) \right]. \tag{15}$$

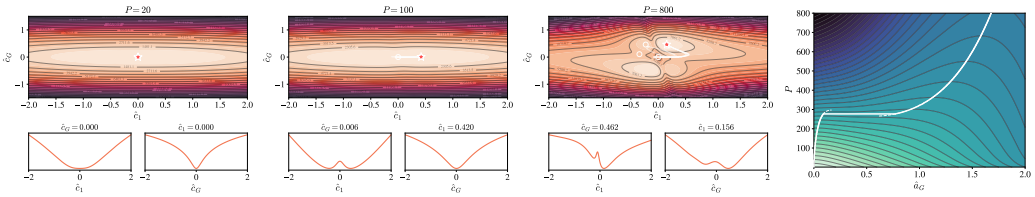

Figure 2: Different topologies of actions $\mathcal{S}^{\text{lin}}(\hat{c}_1, \hat{c}_G)$ and $\mathcal{S}^{\text{softmax}}(\hat{a}_G; P)$. **Linear attention (left):** Contour plots of action at increasing dataset size P, corresponding to the attention stages (i), (ii), (iv) in Fig 1. From left to right: (i) the action has a single minimum at the origin - the model does not use the context; (ii) after the second-order phase transition, a minimum at nonzero $\hat{c}_1$ emerges, corresponding to uniform pooling attention where the model attends to the entire context equally; (iv) finally, the landscape reorganizes and large $\hat{c}_G$ is preferred. The minimum deepens and splits, and two possible solutions are formed, but one is preferred and leads to lower loss values. Bottom row: one-dimensional sections through the global minimum (marked with a star in the contour plot). The plot reveals a clear Landau-type phase transition from (i) to (ii). Curves show the training trajectories projected onto $(\hat{c}_1, \hat{c}_G)$; these do not follow the gradient of $\mathcal{S}$ as optimization occurs in weight space. **Softmax attention (right):** Contour plot of the action $\mathcal{S}^{\text{softmax}}(\hat{a}_G)$ at varying values of P. The white solid line indicates the location of the global minimum and dashed lines indicate local minima. At $P \approx 300$ both minima become equally deep, the global minimum jumps from the local minimum at smaller $\hat{a}_G$ to the one at larger $\hat{a}_G$, leading to a first-order phase transition.

Together, Eqs. (15) and (12) reduce the high-dimensional inference problem to a closed-form two-dimensional posterior that is asymptotically exact assuming the ansatz in Eq. (11). Next, we study the behavior of the action in terms of the order parameters for both linear and softmax attention. In addition to the order parameters, we obtain the network loss from the action directly (see App. H).

Linear attention action, order parameters, and sample complexities. Pulling out a common scaling factor of χL, we see that in the limit $L, \chi \gg 1$ the action simplifies to $\mathcal{S} = (\mathcal{S}_{\text{NNGP}} + \mathcal{S}_{\text{data}}) + O(\chi^{-1})$ where the log-determinant term (see App. F) is subleading and can be dropped. We note that the leading terms are not strictly organized by powers of L or P as $\hat{c}_1, \hat{c}_G$ naturally have different L and P scales, as we show below. Crucially, the χL scaling in front of $\mathcal{S}$ means $(\hat{c}_1, \hat{c}_G)$ concentrate on deterministic values determined by a saddle point approximation.

Notably, this action has a $\mathbb{Z}_2$ symmetry $(\hat{c}_1, \hat{c}_G) \to (-\hat{c}_1, -\hat{c}_G)$, which governs its second-order phase transition. Interestingly, this action has a larger approximate $\mathbb{Z}_2 \times \mathbb{Z}_2$ symmetry that allows sign flips of $\hat{c}_1, \hat{c}_2$ independently. The presence of this approximate symmetry leads to the second plateau and the subsequent crossover; we formalize this in a field-theoretic toy model in App. N.

Expanding the action around the origin gives two sample-complexity scales. The phase transition occurs when the $\mathbb{Z}_2$ symmetry is first broken, where the curvature in the $\hat{c}_1$ direction changes sign

$$P^*_{\hat{c}_1} = \frac{L^2 V^2 \bar{\sigma}^2}{L+V} \sqrt{\frac{1}{\bar{g}_O \bar{g}_W \, (L+1)\,(L+V-1)}} \sim V^2. \tag{16}$$

The crossover is estimated from the point where the curvature along the $\hat{c}_G$ direction changes sign

$$P^*_{\hat{c}_G} = \frac{L^2 \bar{\sigma}^2}{L+V} \sqrt{\frac{V\,(L(L-1)+2(L-1)V+V^2)}{(L+1)\bar{g}_O \bar{g}_W}} \sim L^{3/2} V^{1/2}. \tag{17}$$

Fig. 2 (left panels) illustrates the topology of the action $\mathcal{S}(\hat{c}_1, \hat{c}_G)$ for different P values. At $P < P^*_{\hat{c}_1}$, the action has a single minimum at the origin. At $P = P_{\hat{c}_1} \approx 23$, a new minimum at nonzero $\hat{c}_1$ emerges continuously and the sections reveal the characteristic flat-then-tilted shape of a Landau-type second-order phase transition (see App. A); enabling the model to exploit in-context unigram statistics. Beyond this, as P increases toward $P_{\hat{c}_G} \approx 322$, the saddle at the origin becomes a maximum, driving $\hat{c}_G$ away from zero and the crossover from context pooling to copying. Critically, this reorganization is not a second phase transition: the symmetry is already broken at $P_{\hat{c}_1}$.

Softmax attention action, order parameter, and sample complexity. We now take the push-forward of the posterior under the softmax map. We use $\hat{a}_1$, $\hat{a}_G$ for the order parameters past the

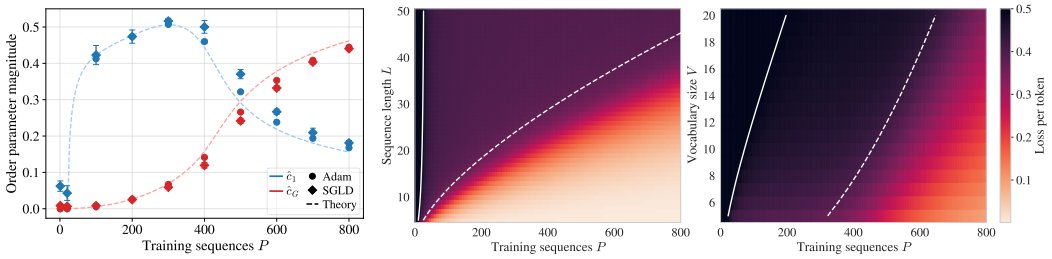

Figure 3: *Linear attention* exhibits a second-order phase transition and subsequent gradual crossover. **Left:** Scalar order parameters $\hat{c}_1, \hat{c}_G$ measured from models trained with Adam (circles) and SGLD (diamonds) compared to the theory predictions (dashed) as a function of the amount of training data or sample complexity P. At the phase transition, $\hat{c}_1$ rises sharply from zero, while $\hat{c}_G$ remains near zero; beyond $P \approx 400$, $\hat{c}_1$ declines while $\hat{c}_G$ rises, marking the crossover from uniform pooling to copy attention; theory tracks the empirically observed behavior closely. **Center and right:** Phase diagram of copy subcircuit emergence. We show the loss per token as a function of training sequences P vs. sequence length L for $V = 5$ fixed (center), and vs. vocabulary size V for fixed $L = 25$ (right). The two phases identified from the order parameters $\hat{c}_1, \hat{c}_G$ manifest as distinct loss plateaus separated by two transition regions; the phase boundaries are predicted by our theory as a function of $P^*(L, V)$ in Eqs. (16) (solid white) and (17) (dashed white).

softmax and define them as projections onto our base function; mirroring $\hat{c}_1$, $\hat{c}_G$. First, the row-wise normalization of softmax eliminates the $\hat{a}_1$ order parameter, as

$$\hat{a}_1 = 1 - \frac{\hat{a}_G}{\sqrt{L}}. \tag{18}$$

Furthermore, the softmax activation is invariant under overall shifts of the inputs, which makes $\hat{a}_1$, $\hat{a}_G$ independent of $\hat{c}_1$; therefore it will be fixed solely by the weight decay / prior term and $\hat{c}_1 = 0$ at the saddle point. Next, since the data term is insensitive to how the attention pattern is parameterized, it requires no adaptation and we can simply plug in $(\hat{a}_1, \hat{a}_G) = (\hat{c}_1, \hat{c}_G)$. Finally an explicit calculation of the mapping between $\hat{c}_G$ and $\hat{a}_G$ gives (full derivation in App. G)

$$\hat{c}_G = \frac{1}{\sqrt{L}} \log \left[\frac{1 + \frac{L-1}{\sqrt{L}} \hat{a}_G}{1 - \frac{1}{\sqrt{L}} \hat{a}_G} \right]. \tag{19}$$

Substituting Eq. (18) and $\hat{c}_1 = 0$ into the prior term Eq. (12), and replacing $(\hat{c}_1, \hat{c}_G) \to (\hat{a}_1, \hat{a}_G)$ in the data term Eq. (15) (which depends only on the post-softmax attention matrix), gives the action for softmax attention. Dropping constants and subleading terms we find

$$S_{\text{softmax}}(\hat{a}_G) = \frac{\chi L}{2} \left(\frac{1}{\bar{g}_W} \log^2 \left(\frac{1 + \sqrt{L} \hat{a}_G}{1 - \hat{a}_G / \sqrt{L}} \right) - \frac{P}{\bar{\sigma}^2 L} \frac{V-1}{V} \frac{\left(\hat{a}_G + \frac{1}{\sqrt{L}} \right)^2 \bar{g}_O}{(\hat{a}_G^2 + 1)\, \bar{g}_O + \frac{\bar{\sigma}^2 L}{P} V} + O\left(\frac{1}{L} \right) \right). \tag{20}$$

The self-consistent equation $\partial_{\hat{a}_G} S \stackrel{!}{=} 0$ has to be solved numerically for $\hat{a}_G^*$ to obtain the MAP values. Investigating the topology of the action in Fig. 2 (right), we find two competing minima, with the global minimum jumping from one to the other at a critical amount of training data $P^*_{\hat{a}_G}$, indicating a first-order phase transition. We can get an analytical approximation for the sample complexity $P^*_{\hat{a}_G}$ by solving $S(0) = S(\hat{a})$ for P and plugging in $\hat{a} = O(1)$ at large L

$$P^*_{\hat{a}_G} \approx \frac{1}{4\bar{g}_W} L \frac{V}{V-1} \bar{\sigma}^2 \ln^2(L) \left(\sqrt{1 + 4 \frac{V-1}{\ln^2(L)} \frac{\bar{g}_W}{\bar{g}_O}} + 1 \right) \sim L \ln^2(L), \tag{21}$$

constituting an improvement of $O(\sqrt{LV})$ in the sample complexity compared to linear attention.

4 Experiments

We validate our theoretical predictions against networks trained with Adam [53] and sampled with SGLD [54, 55] in Fig. 1. All networks in the main text use the merged matrix $\boldsymbol{W}^G$. With separate

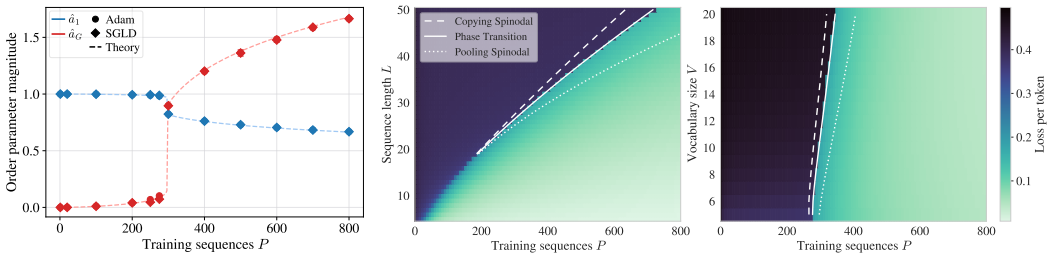

Figure 4: *Softmax attention* exhibits a first-order phase transition. **Left:** Scalar order parameters $\hat{a}_1, \hat{a}_G$ measured from models trained with Adam (circles) and SGLD (diamonds) compared to the theory predictions (dashed lines) as a function of the amount of training data or sample complexity P. At the phase transition the value of $\hat{a}_G$ jumps marking a first-order phase transition. Note that the dots corresponding to SGLD are often covered by the diamonds corresponding to Adam due to their precise agreement. **Center and right:** Phase diagram of copy subcircuit emergence. We show the loss per token as a function of training sequences P vs. sequence length L for $V = 5$ fixed (center), and vs. vocabulary size V for fixed $L = 25$ (right). The transition from a phase of uniform attention to a phase of copy attention manifests as a sharp drop in the loss; the phase boundary $P^*(L, V)$ (white) is predicted by our theory in Eq. (21). Additionally, the appearance of the copy local minimum is marked with a dashed line, and the disappearance of the pooling local minimum is marked by a dotted line

matrices $\boldsymbol{W}^Q, \boldsymbol{W}^K$ at $d_k = d_{\text{model}}$, we find the same transitions and loss curves (see App. J, Fig. 5). Theory and experiment agree quantitatively across all stages. The two attention mechanisms differ qualitatively: linear attention exhibits a second-order phase transition at $P \approx 100$, where $\hat{c}_1$ departs from zero and the model begins pooling uniformly over the context, followed by a crossover as $\hat{c}_G$ grows and the copy pattern gradually sharpens. Softmax attention instead undergoes a first-order transition at $P \approx 300$, with a discontinuous loss drop from the initial uniform pooling solution directly to a copy solution. In both cases, the pooling stage likely corresponds to learning the first-order statistics of the data while the second stage reflects the landscape topology change as discussed in Fig. 2: the uniform pooling minimum destabilizes and the copy minimum deepens.

Remarkably, while our predictions are for the MAP values, we find excellent agreement with models trained using Adam: minimizing the action (10) at large χ corresponds to training with gradient flow with weight decay (see App. M for derivation and refined statement). The two procedures, however, take the limits of long training time and large χ in opposite order, and these limits do not commute when the action has several minima: the posterior concentrates on the global minimum, whereas gradient flow can remain in a local one, depending on the initialization. We therefore ran an ensemble of models per P value and report results from runs that reached the global minimum (see App. R.1 for details; for softmax all runs converge to the global minimum); SGLD instead targets the Bayesian posterior as its stationary distribution and, given sufficient mixing time, would sample basins by their posterior weights without requiring such explicit filtering.

Next, we examine the two transitions in detail through the order parameters $\hat{c}_1, \hat{c}_G$. For linear attention, Fig.3 (left panel) shows that both Adam and SGLD results track the theoretical predictions closely across different amounts of training data P. The transient deviation of SGLD near the crossover is attributable to large fluctuations as the uniform pooling minimum destabilizes and the copy minimum emerges (Fig. 2), causing the posterior to broaden and the network to transiently explore directions outside the two-dimensional order parameter subspace; we quantify this coverage in App. P. The three-phase structure is confirmed in the (P, L) and (P, V) phase diagrams (Fig. 3 center and right): the order parameter transitions manifest as distinct loss plateaus, with the two boundaries scaling with L and V as predicted by Eqs. (16),(17). The first boundary marks the onset of structured attention, the second the emergence of the copy pattern.

Softmax attention in Fig. 4 presents a strikingly different picture: rather than the three-phase structure of linear attention, $\hat{a}_G$ jumps discontinuously at a single phase boundary and the loss drops abruptly, with no initial learning phase. The boundary is predicted by Eq. (21) across the full (P, L) and (P, V) planes, with the theory tracking the empirical transition remarkably well. The sharp separation between the two phases is a direct signature of the first-order phase transition.

5 Discussion

We presented a novel Bayesian theory of feature learning in a single attention head, obtaining a first-principles theory explaining how the copy subcircuit emerges as a phase transition in the amount of training data. By reducing the posterior over attention patterns to a low-dimensional order parameter space, we obtain quantitative predictions for the phase boundary and the attention pattern evolution.

A key finding is that the transition order depends on the attention mechanism. For linear attention, an initial *second-order phase transition* to uniform pooling attention is followed by a smooth, continuous crossover toward the copy attention; this is not a second phase transition, since the symmetry is already broken at the first transition and no additional singularity can occur. For softmax attention, we instead find a *first-order phase transition*, a discontinuous onset of structured attention. That the same Bayesian framework yields qualitatively different transition orders for the two mechanisms is a demonstration of its discriminative power: it resolves mechanistic differences that coarser analyses would miss. Together, these results provide a fine-grained and quantitative account of copy subcircuit emergence.

Implications for the predictability of emergent phenomena. The phase transition order has direct consequences on whether capability emergence can be anticipated. The second-order phase transition of linear attention is preceded by diverging fluctuations [56] (we are aware of one work that utilizes these [57]), and the subsequent gradual crossover allows capabilities to be tracked before they are fully expressed. In contrast, the first-order phase transition of softmax attention is inherently unpredictable from any observable at $P < P_*$. While the pooling minimum is the global minimum, the probability of sampling away from the pooling minimum is exponentially suppressed, ensuring samples are uninformative about the approaching phase transition. At the phase transition, the pooling minimum maintains stability so no large fluctuations or other precursor signals appear, making the transition inherently unpredictable. This makes understanding which type of transition governs a given capability in a given architecture crucial for designing a monitoring strategy, and suggests fundamental limits to what can be monitored. However, several important questions remain. In particular, capabilities of practical importance may be composite, arising from many "microscopic" capabilities that emerge abruptly but collectively produce a gradual, observable buildup. It remains to be determined whether such collective signals can enable prediction despite the unpredictability of individual transitions. Finally, a further direction for investigation is how these dynamics unfold as a function of training steps rather than sample complexity.

Bayesian Posterior vs. Dynamics. In this work we focus on analyzing the emergence of a copy head as a function of sample complexity under Bayesian inference. This framework can naturally be mapped to equilibrium statistical mechanics where phase transitions are well defined. A separate question is the characterization of the optimization dependent approach to equilibrium. We show the trajectories of the loss and order parameters throughout training for all seeds and P values in Appendix R.1; we leave the theoretical study to future work.

Limitations and Outlook. To reduce the posterior to a low-dimensional order parameter space, we make several assumptions: First, the projection onto the two-dimensional subspace is an ansatz we support experimentally (see App. M) rather then derive rigorously. Second, we focus here on the single-head, single-layer setting and the supervised copy task to isolate the copy subcircuit, which is our main subject of interest. The Bayesian framework is more general and derived for N_H heads, in which case the heads separate in the prior and couple only through the summed data kernel. Further, our Bayesian framework admits a multi-layer generalization outlined in App. K. A natural next step is therefore to characterize how the copy head feeds into the induction head and, in particular, whether that changes the transition character. The phase boundary $P^*(L, V)$ derived here could further serve as a basis for scaling laws predicting copy subcircuit emergence as a function of model and data scale. Structured inputs can be incorporated into our framework, as demonstrated for Markov transitions in App. L. Extending our results beyond i.i.d. uniform inputs would be an interesting direction for future work, particularly to understand qualitative changes in the transition and shifts in the sample complexity P^*. Finally, our derivation relies on the equivalent-kernel approximation and the $d_{\text{model}} \to \infty$ limit at fixed d_{model}/d_k, suppressing finite-size corrections of the data distribution and of the Gaussian prior on the attention pattern, respectively. These approximations are accurate in the strong regularization regime and for large model dimensions, respectively, but a systematic treatment of small regularization or finite P and finite d_{model} corrections similar to [37, 38] will be an interesting direction for future work.

Acknowledgments

We thank Noa Rubin for the insightful discussion and comments on the manuscript. I.L. acknowledges the Helmholtz Information & Data Science Academy (HIDA) for providing financial support that enabled a short-term research stay at IAS-6, during which this project was initiated. Z.R. and I.L. acknowledge support from ISF Grant 2250/19. The authors gratefully acknowledge the computing time granted by the JARA Vergabegremium and provided on the JARA Partition part of the supercomputer JURECA at Forschungszentrum Jülich (computation grant JINB33). K.F. is supported by the Joachim Herz Foundation.

Societal Impact Statement

This paper works towards understanding the emergence of copy heads in transformers, thus aiming to advance mechanistic explainability of transformers. While the latter surely has societal impacts, these will be much further down the line.

Use of AI Assistance

The authors conceived the project, derived the Bayesian framework and the theory for linear attention apart from the closed form of Σ by hand, and wrote an initial implementation of the numerical experiments. AI assistants, namely Claude (Anthropic) and ChatGPT (OpenAI) in different versions over the course of the project, were used to check these derivations, which revealed minor errors such as missing prefactors, to derive the closed form of the summed pseudo-inverse Σ (App. D). They were further used to support writing and editing the text. All AI-assisted content was checked, verified, and revised by the authors, who take full responsibility for the content of the paper.

References

[1] 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 H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901. Curran Associates, Inc., 2020.

[2] 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. arXiv:2001.08361 [cs].

[3] Jordan Hoffmann, Sebastian Borgeaud, Arthur Mensch, Elena Buchatskaya, Trevor Cai, Eliza Rutherford, Diego de Las Casas, Lisa Anne Hendricks, Johannes Welbl, Aidan Clark, Thomas Hennigan, Eric Noland, Katherine Millican, George van den Driessche, Bogdan Damoc, Aurelia Guy, Simon Osindero, Karén Simonyan, Erich Elsen, Oriol Vinyals, Jack Rae, and Laurent Sifre. An empirical analysis of compute-optimal large language model training. In S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh, editors, *Advances in Neural Information Processing Systems*, volume 35, pages 30016–30030. Curran Associates, Inc., 2022.

[4] Nicholas Lourie, Michael Y. Hu, and Kyunghyun Cho. Scaling Laws Are Unreliable for Downstream Tasks: A Reality Check, October 2025. arXiv:2507.00885 [cs].

[5] Jason Wei, Yi Tay, Rishi Bommasani, Colin Raffel, Barret Zoph, Sebastian Borgeaud, Dani Yogatama, Maarten Bosma, Denny Zhou, Donald Metzler, Ed H. Chi, Tatsunori Hashimoto, Oriol Vinyals, Percy Liang, Jeff Dean, and William Fedus. Emergent Abilities of Large Language Models. *Transactions on Machine Learning Research*, June 2022.

[6] Rylan Schaeffer, Brando Miranda, and Sanmi Koyejo. Are emergent abilities of large language models a mirage? In A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine, editors, *Advances in Neural Information Processing Systems*, volume 36, pages 55565–55581. Curran Associates, Inc., 2023.

[7] Leonardo Berti, Flavio Giorgi, and Gjergji Kasneci. Emergent Abilities in Large Language Models: A Survey, March 2025. arXiv:2503.05788 [cs].

[8] Anthropic. Anthropic's responsible scaling policy. Policy document, Anthropic, September 2023.

[9] Boaz Barak, Benjamin Edelman, Surbhi Goel, Sham Kakade, Eran Malach, and Cyril Zhang. Hidden progress in deep learning: Sgd learns parities near the computational limit. In S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh, editors, *Advances in Neural Information Processing Systems*, volume 35, pages 21750–21764. Curran Associates, Inc., 2022.

[10] Kenzo Clauw, Daniele Marinazzo, and Sebastiano Stramaglia. Information-theoretic progress measures reveal grokking is an emergent phase transition. In *ICML 2024 Workshop on Mechanistic Interpretability*, 2024.

[11] Neel Nanda, Lawrence Chan, Tom Lieberum, Jess Smith, and Jacob Steinhardt. Progress measures for grokking via mechanistic interpretability. In *The Eleventh International Conference on Learning Representations*, 2023.

[12] Gautam Reddy. The mechanistic basis of data dependence and abrupt learning in an in-context classification task. In *The Twelfth International Conference on Learning Representations*, 2024.

[13] Shivam Garg, Dimitris Tsipras, Percy S Liang, and Gregory Valiant. What can transformers learn in-context? a case study of simple function classes. In S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh, editors, *Advances in Neural Information Processing Systems*, volume 35, pages 30583–30598. Curran Associates, Inc., 2022.

[14] Ekin Akyürek, Dale Schuurmans, Jacob Andreas, Tengyu Ma, and Denny Zhou. What learning algorithm is in-context learning? investigations with linear models. In *The Eleventh International Conference on Learning Representations*, 2023.

[15] 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*, pages 35151–35174. PMLR, July 2023.

[16] Yue M. Lu, Mary Letey, Jacob A. Zavatone-Veth, Anindita Maiti, and Cengiz Pehlevan. Asymptotic theory of in-context learning by linear attention. *Proceedings of the National Academy of Sciences*, 122(28):e2502599122, 2025.

[17] Nelson Elhage, Neel Nanda, Catherine Olsson, Tom Henighan, Nicholas Joseph, Ben Mann, Amanda Askell, Yuntao Bai, Anna Chen, Tom Conerly, Nova DasSarma, Dawn Drain, Deep Ganguli, Zac Hatfield-Dodds, Danny Hernandez, Andy Jones, Jackson Kernion, Liane Lovitt, Kamal Ndousse, Dario Amodei, Tom Brown, Jack Clark, Jared Kaplan, Sam McCandlish, and Christopher Olah. A mathematical framework for transformer circuits. *Transformer Circuits Thread*, 2021.

[18] 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. arXiv:2209.11895 [cs].

[19] Chen Siyu, Sheen Heejune, Wang Tianhao, and Yang Zhuoran. Training dynamics of multi-head softmax attention for in-context learning: Emergence, convergence, and optimality (extended abstract). In Shipra Agrawal and Aaron Roth, editors, *Proceedings of Thirty Seventh Conference on Learning Theory*, volume 247 of *Proceedings of Machine Learning Research*, pages 4573–4573. PMLR, 30 Jun–03 Jul 2024.

[20] Yedi Zhang, Aaditya K Singh, Peter E. Latham, and Andrew M Saxe. Training dynamics of in-context learning in linear attention. In Aarti Singh, Maryam Fazel, Daniel Hsu, Simon Lacoste-Julien, Felix Berkenkamp, Tegan Maharaj, Kiri Wagstaff, and Jerry Zhu, editors, *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pages 76047–76087. PMLR, 13–19 Jul 2025.

[21] Ezra Edelman, Nikolaos Tsilivis, Benjamin L. Edelman, eran malach, and Surbhi Goel. The evolution of statistical induction heads: In-context learning markov chains. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024.

[22] Vivien Cabannes, Charles Arnal, Wassim Bouaziz, Xingyu Alice Yang, Francois Charton, and Julia Kempe. Iteration head: A mechanistic study of chain-of-thought. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024.

[23] Johannes von Oswald, Eyvind Niklasson, Maximilian Schlegel, Seijin Kobayashi, Nicolas Zucchet, Nino Scherrer, Nolan Miller, Mark Sandler, Blaise Agüera y Arcas, Max Vladymyrov, Razvan Pascanu, and João Sacramento. Uncovering mesa-optimization algorithms in Transformers, September 2023. arXiv:2309.05858 [cs].

[24] Mingze Wang, Ruoxi Yu, Weinan E, and Lei Wu. How Transformers Get Rich: Approximation and Dynamics Analysis, January 2025. arXiv:2410.11474 [cs] version: 3.

[25] Gouki Minegishi, Hiroki Furuta, Shohei Taniguchi, Yusuke Iwasawa, and Yutaka Matsuo. Beyond induction heads: In-context meta learning induces multi-phase circuit emergence. In Aarti Singh, Maryam Fazel, Daniel Hsu, Simon Lacoste-Julien, Felix Berkenkamp, Tegan Maharaj, Kiri Wagstaff, and Jerry Zhu, editors, *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pages 44372–44395. PMLR, 13–19 Jul 2025.

[26] Tatsuya Aoyama, Ethan Gotlieb Wilcox, and Nathan Schneider. Predicting the Emergence of Induction Heads in Language Model Pretraining, February 2026. arXiv:2511.16893 [cs].

[27] Jiri Hron, Yasaman Bahri, Jascha Sohl-Dickstein, and Roman Novak. Infinite attention: NNGP and NTK for deep attention networks. In Hal Daumé III and Aarti Singh, editors, *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 4376–4386. PMLR, 13–18 Jul 2020.

[28] Itay Lavie, Guy Gur-Ari, and Zohar Ringel. Towards Understanding Inductive Bias in Transformers: A View From Infinity. In *Proceedings of the 41st International Conference on Machine Learning*, pages 26043–26069. PMLR, July 2024.

[29] Aditya Cowsik, Tamra Nebabu, Xiaoliang Qi, and Surya Ganguli. Geometric dynamics of signal propagation predict trainability of transformers. *Phys. Rev. E*, 112:055301, Nov 2025.

[30] Lorenzo Noci, Chuning Li, Mufan Li, Bobby He, Thomas Hofmann, Chris Maddison, and Dan Roy. The shaped transformer: Attention models in the infinite depth-and-width limit. In A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine, editors, *Advances in Neural Information Processing Systems*, volume 36, pages 54250–54281. Curran Associates, Inc., 2023.

[31] Blake Bordelon, Hamza Chaudhry, and Cengiz Pehlevan. Infinite limits of multi-head transformer dynamics. In A. Globerson, L. Mackey, D. Belgrave, A. Fan, U. Paquet, J. Tomczak, and C. Zhang, editors, *Advances in Neural Information Processing Systems*, volume 37, pages 35824–35878. Curran Associates, Inc., 2024.

[32] Hugo Cui, Freya Behrens, Florent Krzakala, and Lenka Zdeborová. A phase transition between positional and semantic learning in a solvable model of dot-product attention. In A. Globerson, L. Mackey, D. Belgrave, A. Fan, U. Paquet, J. Tomczak, and C. Zhang, editors, *Advances in Neural Information Processing Systems*, volume 37, pages 36342–36389. Curran Associates, Inc., 2024.

[33] Emanuele Troiani, Hugo Cui, Yatin Dandi, Florent Krzakala, and Lenka Zdeborova. Fundamental limits of learning in sequence multi-index models and deep attention networks: high-dimensional asymptotics and sharp thresholds. In Aarti Singh, Maryam Fazel, Daniel Hsu, Simon Lacoste-Julien, Felix Berkenkamp, Tegan Maharaj, Kiri Wagstaff, and Jerry Zhu, editors, *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pages 60147–60182. PMLR, 13–19 Jul 2025.

[34] Fabrizio Boncoraglio, Emanuele Troiani, Vittorio Erba, and Lenka Zdeborová. Bayes optimal learning of attention-indexed models. In D. Belgrave, C. Zhang, H. Lin, R. Pascanu, P. Koniusz, M. Ghassemi, and N. Chen, editors, *Advances in Neural Information Processing Systems*, volume 38, Main Conference, pages 105029–105074. Curran Associates, Inc., 2025.

[35] Fabrizio Boncoraglio, Vittorio Erba, Emanuele Troiani, Yizhou Xu, Florent Krzakala, and Lenka Zdeborová. Single-head attention in high dimensions: A theory of generalization, weights spectra, and scaling laws. In Tong Zhang, Miroslav Dudik, Martin Jaggi, Alekh Agarwal, Sharon Li, Dale Schuurmans, Jerry Zhu, Felix Berkenkamp, Hanze Dong, and Alberto Bietti, editors, *Proceedings of the 43rd International Conference on Machine Learning*, volume 306 of *Proceedings of Machine Learning Research*, pages 9058–9093. PMLR, 06–11 Jul 2026.

[36] 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, 2023.

[37] Kirsten Fischer, Javed Lindner, David Dahmen, Zohar Ringel, Michael Krämer, and Moritz Helias. Critical feature learning in deep neural networks. In Ruslan Salakhutdinov, Zico Kolter, Katherine Heller, Adrian Weller, Nuria Oliver, Jonathan Scarlett, and Felix Berkenkamp, editors, *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pages 13660–13690. PMLR, 21–27 Jul 2024.

[38] Noa Rubin, Kirsten Fischer, Javed Lindner, Inbar Seroussi, Zohar Ringel, Michael Krämer, and Moritz Helias. From kernels to features: A multi-scale adaptive theory of feature learning. In Aarti Singh, Maryam Fazel, Daniel Hsu, Simon Lacoste-Julien, Felix Berkenkamp, Tegan Maharaj, Kiri Wagstaff, and Jerry Zhu, editors, *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pages 52225–52257. PMLR, 13–19 Jul 2025.

[39] Noa Rubin, Inbar Seroussi, and Zohar Ringel. Grokking as a first order phase transition in two layer networks. In *The Twelfth International Conference on Learning Representations*, 2024.

[40] Lorenzo Tiberi, Francesca Mignacco, Kazuki Irie, and Haim Sompolinsky. Dissecting the interplay of attention paths in a statistical mechanics theory of transformers. In A. Globerson, L. Mackey, D. Belgrave, A. Fan, U. Paquet, J. Tomczak, and C. Zhang, editors, *Advances in Neural Information Processing Systems*, volume 37, pages 72710–72753. Curran Associates, Inc., 2024.

[41] Aaditya K Singh, Ted Moskovitz, Felix Hill, Stephanie C.Y. Chan, and Andrew M Saxe. What needs to go right for an induction head? A mechanistic study of in-context learning circuits and their formation. In Ruslan Salakhutdinov, Zico Kolter, Katherine Heller, Adrian Weller, Nuria Oliver, Jonathan Scarlett, and Felix Berkenkamp, editors, *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pages 45637–45662. PMLR, 21–27 Jul 2024.

[42] Eshaan Nichani, Alex Damian, and Jason D. Lee. How transformers learn causal structure with gradient descent. In Ruslan Salakhutdinov, Zico Kolter, Katherine Heller, Adrian Weller, Nuria Oliver, Jonathan Scarlett, and Felix Berkenkamp, editors, *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pages 38018–38070. PMLR, 21–27 Jul 2024.

[43] Siyu Chen, Heejune Sheen, Tianhao Wang, and Zhuoran Yang. Unveiling induction heads: Provable training dynamics and feature learning in transformers. In A. Globerson, L. Mackey, D. Belgrave, A. Fan, U. Paquet, J. Tomczak, and C. Zhang, editors, *Advances in Neural Information Processing Systems*, volume 37, pages 66479–66567. Curran Associates, Inc., 2024.

[44] Alberto Bietti, Vivien Cabannes, Diane Bouchacourt, Herve Jegou, and Leon Bottou. Birth of a transformer: A memory viewpoint. In A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine, editors, *Advances in Neural Information Processing Systems*, volume 36, pages 1560–1588. Curran Associates, Inc., 2023.

[45] Luca Arnaboldi, Bruno Loureiro, Ludovic Stephan, Florent Krzakala, and Lenka Zdeborová. Asymptotics of sgd in sequence-single index models and single-layer attention networks. In D. Belgrave, C. Zhang, H. Lin, R. Pascanu, P. Koniusz, M. Ghassemi, and N. Chen, editors, *Advances in Neural Information Processing Systems*, volume 38, Main Conference, pages 20611–20645. Curran Associates, Inc., 2025.

[46] M. Sagitova, O. Duranthon, and L. Zdeborová. Specialization of softmax attention heads: insights from the high-dimensional single-location model. arXiv preprint arXiv:2603.03993, 2026.

[47] Tiberiu Musat, Tiago Pimentel, Lorenzo Noci, Alessandro Stolfo, Mrinmaya Sachan, and Thomas Hofmann. On the Emergence of Induction Heads for In-Context Learning, January 2026. arXiv:2511.01033 [cs].

[48] Yuandong Tian, Yiping Wang, Beidi Chen, and Simon Du. Scan and snap: Understanding training dynamics and token composition in 1-layer transformer. In A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine, editors, *Advances in Neural Information Processing Systems*, volume 36, pages 71911–71947. Curran Associates, Inc., 2023.

[49] Kwangjun Ahn, Xiang Cheng, Minhak Song, Chulhee Yun, Ali Jadbabaie, and Suvrit Sra. Linear attention is (maybe) all you need (to understand transformer optimization). In *The Twelfth International Conference on Learning Representations*, 2024.

[50] Zohar Ringel, Noa Rubin, Edo Mor, Moritz Helias, and Inbar Seroussi. Applications of statistical field theory in deep learning. *arXiv:2502.18553*, 2025.

[51] Ge Yang, Edward Hu, Igor Babuschkin, Szymon Sidor, Xiaodong Liu, David Farhi, Nick Ryder, Jakub Pachocki, Weizhu Chen, and Jianfeng Gao. Tuning large neural networks via zero-shot hyperparameter transfer. In M. Ranzato, A. Beygelzimer, Y. Dauphin, P.S. Liang, and J. Wortman Vaughan, editors, *Advances in Neural Information Processing Systems*, volume 34, pages 17084–17097. Curran Associates, Inc., 2021.

[52] B. W. Silverman. Spline smoothing: The equivalent variable kernel method. *The Annals of Statistics*, 12(3):898–916, 1984.

[53] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *International Conference on Learning Representations (ICLR)*, 2015.

[54] Max Welling and Yee Whye Teh. Bayesian learning via stochastic gradient langevin dynamics. In *Proceedings of the 28th International Conference on International Conference on Machine Learning*, ICML'11, page 681–688, Madison, WI, USA, 2011. Omnipress.

[55] Gadi Naveh, Oded Ben David, Haim Sompolinsky, and Zohar Ringel. Predicting the outputs of finite deep neural networks trained with noisy gradients. *Phys. Rev. E*, 104:064301, Dec 2021.

[56] Mehran Kardar. *Statistical Physics of Fields*. Cambridge University Press, Cambridge, 2007.

[57] Max Hennick and Guillaume Corlouer. From density matrices to phase transitions in deep learning: Spectral early warnings and interpretability. *arXiv:2603.29805*, 2026.

[58] E. Gardner. The space of interactions in neural network models. *Journal of Physics A: Mathematical and General*, 21(1):257, January 1988.

[59] E. Gardner and B. Derrida. Optimal storage properties of neural network models. *Journal of Physics A: Mathematical and General*, 21(1):271, January 1988.

[60] Moritz Helias and David Dahmen. *Statistical Field Theory for Neural Networks*, volume 970 of *Lecture Notes in Physics*. Springer International Publishing, Cham, 2020.

[61] Yasaman Bahri, Jonathan Kadmon, Jeffrey Pennington, Sam S. Schoenholz, Jascha Sohl-Dickstein, and Surya Ganguli. Statistical Mechanics of Deep Learning. *Annual Review of Condensed Matter Physics*, 11(Volume 11, 2020):501–528, March 2020.

[62] Hugo Cui. High-dimensional learning of narrow neural networks. *Journal of Statistical Mechanics: Theory and Experiment*, 2025(2):023402, February 2025.

[63] Hugo Touchette. The large deviation approach to statistical mechanics. *Physics Reports*, 478(1):1–69, 2009.

[64] Alexander van Meegen and Haim Sompolinsky. Coding schemes in neural networks learning classification tasks. *Nature Communications*, 16:3354, 2025.

[65] Jan P. Bauer, Kirsten Fischer, Moritz Helias, and Agostina Palmigiano. A unified theory of feature learning in rnns and dnns. *arXiv:2602.15593*, 2026.

[66] Radford M. Neal. *Bayesian Learning for Neural Networks*, volume 118 of *Lecture Notes in Statistics*. Springer, 1996.

[67] Jaehoon Lee, Jascha Sohl-dickstein, Jeffrey Pennington, Roman Novak, Sam Schoenholz, and Yasaman Bahri. Deep neural networks as gaussian processes. In *International Conference on Learning Representations*, 2018.

[68] Heejune Sheen, Siyu Chen, Tianhao Wang, and Harrison H. Zhou. Implicit Regularization of Gradient Flow on One-Layer Softmax Attention, March 2024. arXiv:2403.08699 [cs] version: 1.

[69] Davoud Ataee Tarzanagh, Yingcong Li, Christos Thrampoulidis, and Samet Oymak. Transformers as Support Vector Machines, February 2024. arXiv:2308.16898 [cs, math].

[70] Hannes Risken. *Fokker-Planck Equation*, pages 63–95. Springer Berlin Heidelberg, Berlin, Heidelberg, 1996.

[71] Anders Krogh and John Hertz. A simple weight decay can improve generalization. In J. Moody, S. Hanson, and R.P. Lippmann, editors, *Advances in Neural Information Processing Systems*, volume 4. Morgan-Kaufmann, 1991.

Appendix

A A short introduction to Landau Theory and Phase Transitions

This appendix gives a self-contained statistical physics background for some of the ideas used in the text. We focus on Landau theory, order parameters, symmetries, first- and second-order phase transitions, and crossovers. Phase transitions and statistical mechanics have a long history in machine learning, beginning with work on the phase transition in the capacity of the perceptron [58, 59]. This appendix, however, does not aim to cover that rich history, some recent reviews of the field include [60, 61, 62, 50]. We discuss the most directly related work in Section 1.1.

A.1 Phases, Control Parameters, and Large-System Limits

A *phase* is a macroscopic regime characterized by qualitatively stable behavior. In this paper, the phases are regimes of the learned attention pattern: for example, a regime where the attention mechanism does not use the context, a regime where it pools uniformly over all positions, and a regime where it implements the copy pattern.

A *control parameter* is a quantity whose variation can drive the system between phases. In our setting the central control parameter is the amount of training data P, while other quantities such as sequence length L, vocabulary size V, and noise scale σ^2 also affect the phase boundaries.

A true phase transition is strictly defined only in a large-system limit. For finite systems, observables are typically smooth functions of the control parameter. Sharp nonanalytic behavior, which is required for a phase transition, appears only after taking an asymptotic limit. In our theory, the analogous large parameter is the overall scale multiplying the reduced action, in particular the χL scaling appearing in the posterior. At finite χ and finite L, transitions are softened; in the large-χL limit, the posterior concentrates around the minima of the action and sharp transitions can appear.

A.2 Macroscopic Actions and Concentration

The central mathematical structure behind Landau theory of phase transitions, which we be explained below, is a probability distribution of the form

$$p(m) \propto \exp[-\Lambda\, s(m)], \tag{22}$$

where m is a low-dimensional macroscopic variable, $s(m)$ is an action density, and Λ is a large coefficient; naturally this is intimately related to the theory of large deviations [63]. In our setting, Λ is represented by the large factor χL multiplying the reduced action.

The normalization of this distribution is

$$Z = \int dm\, \exp[-\Lambda s(m)]. \tag{23}$$

The associated free energy is

$$\mathcal{F} = -\log Z. \tag{24}$$

When Λ is large, the integral is dominated by the minimum of the action density:

$$m^\star = \arg\min_m s(m). \tag{25}$$

The saddle-point approximation then gives

$$\mathcal{F} = -\log Z \simeq \Lambda s(m^\star) + \text{subleading fluctuation terms}. \tag{26}$$

Thus the phase selected by the posterior is determined by the minimum of the reduced action, while the free energy is obtained by evaluating the corresponding marginal integral.

The large coefficient in Eq. (22) has two roles. First, it justifies a saddle-point approximation. If $m^\star$ is a global minimum, then expanding around it gives

$$s(m) = s(m^\star) + \frac{1}{2} s''(m^\star)(m - m^\star)^2 + \cdots . \tag{27}$$

Substituting this into Eq. (22) gives

$$p(m) \approx \exp[-\Lambda s(m^\star)] \exp\left[-\frac{\Lambda}{2} s''(m^\star)(m-m^\star)^2\right]. \tag{28}$$

Thus typical fluctuations around the minimum scale as

$$|m - m^\star| \sim \frac{1}{\sqrt{\Lambda}}. \tag{29}$$

As $\Lambda \to \infty$, the distribution collapses onto the minimizing value of the macroscopic variable.

Second, Eq. (22) is a sign that a successful macroscopic description has emerged. The original system may have many microscopic degrees of freedom, such as network weights and hidden activations. After marginalizing or averaging over them, the dominant behavior is governed by a small number of collective variables and a scalar action. In this paper, the collective variables are the attention order parameters, and the large factor χL makes their saddle-point values sharply defined.

In our Bayesian setting, the posterior over attention patterns has the form

$$p(\boldsymbol{G} \mid \text{data}) \propto \exp[-\mathcal{S}(\boldsymbol{G})]. \tag{30}$$

After reducing to the relevant order-parameter space, this becomes

$$p(\hat{c}_1, \hat{c}_G \mid \text{data}) \propto \exp[-\mathcal{S}(\hat{c}_1, \hat{c}_G)]. \tag{31}$$

In the scaling regime used in the main text,

$$\mathcal{S}(\hat{c}_1, \hat{c}_G) = \chi L \, S(\hat{c}_1, \hat{c}_G) + \text{subleading terms}. \tag{32}$$

Therefore the dominant learned attention pattern is determined, to leading order, by the minima of the reduced action S.

A.3 Order Parameters

An *order parameter* is a low-dimensional quantity that distinguishes phases. It is chosen to capture the collective degree of freedom that changes qualitatively across a transition.

In this paper, the order parameters are scalar projections of the attention matrix onto interpretable attention patterns. For linear attention, we have

$$\boldsymbol{G} = \hat{c}_1 \frac{\mathbf{1}_L}{L} + \hat{c}_G \frac{\boldsymbol{G}^*}{\sqrt{L}}. \tag{33}$$

The coefficient $\hat{c}_1$ measures the strength of uniform pooling over the context, while $\hat{c}_G$ measures the strength of the copy or shift-by-one pattern. In the softmax case, row normalization removes one independent direction, leaving the copy order parameter $\hat{a}_G$.

The key point is that an order parameter is not merely a summary statistic. It is the macroscopic coordinate in which the posterior concentrates. In the present work, the order parameters distinguish whether the network ignores context, pools over context, or selectively attends to the previous token.

A.4 Landau Polynomials, Minima, and Phase Transitions

Landau theory studies the reduced action as a function of the order parameter. Near a transition, the action is often expanded as a polynomial. For a single scalar order parameter m, one writes

$$S(m) = S_0 + hm + \frac{r}{2}m^2 + \frac{w}{3}m^3 + \frac{u}{4}m^4 + \cdots. \tag{34}$$

This polynomial is a local expansion of the macroscopic action after the microscopic degrees of freedom have been integrated out. The coefficients depend on the control parameter. In our setting, they depend on P and on problem parameters such as L, V, and σ^2.

The phases are determined by the minima of this polynomial action. If the global minimum is at $m = 0$, then the corresponding order parameter is absent. If the global minimum is at $m \neq 0$, then the corresponding structure is present. A phase transition occurs when the identity or location of the global minimum changes nonanalytically in the large-χL limit.

For example, suppose symmetry forbids odd powers of m, giving

$$S(m) = \frac{r}{2}m^2 + \frac{u}{4}m^4, \qquad u > 0. \tag{35}$$

When $r > 0$, the unique minimum is at

$$m^\star = 0. \tag{36}$$

When $r < 0$, the point $m = 0$ becomes unstable and two new minima appear at

$$m^\star = \pm\sqrt{\frac{-r}{u}}. \tag{37}$$

Thus a change in the sign of a polynomial coefficient changes the structure of the minima, and hence changes the phase selected by the posterior.

In our Bayesian setting, the reduced action is the negative log-posterior restricted to the relevant order-parameter subspace. The learned attention pattern is determined by the minimum of $\mathcal{S}_{\mathrm{linear}}(\hat{c}_1, \hat{c}_G)$ for linear attention, or $\mathcal{S}_{\mathrm{softmax}}(\hat{a}_G)$ for softmax attention. The phase transition is therefore a change in the global minimum of the action as the amount of training data P is varied.

The structure of the action reflects a competition between two effects. The prior term favors simple, small-norm attention patterns. The data term favors attention patterns that correctly predict the copy labels. A phase transition occurs when increasing the dataset size P makes a structured attention pattern favorable enough to overcome the prior cost.

A.5 Symmetry and Symmetry Breaking

Symmetry is central to the classification of phase transitions. A system has a symmetry if the action is unchanged under a transformation of the order parameter. For example, the polynomial in Eq. (35) is invariant under

$$m \mapsto -m. \tag{38}$$

This is a $\mathbb{Z}_2$ symmetry. When $r > 0$, the minimum is at $m = 0$, which is invariant under this symmetry. When $r < 0$, the minima are at nonzero values of m, and choosing one of them breaks the symmetry.

Symmetry also determines which polynomial terms are allowed. If the action is invariant under $m \mapsto -m$, then odd powers such as m and m^3 are forbidden. If this symmetry is absent, odd powers may appear, and the transition structure can change.

In the linear-attention theory, the reduced action has an exact diagonal sign symmetry

$$(\hat{c}_1, \hat{c}_G) \mapsto (-\hat{c}_1, -\hat{c}_G). \tag{39}$$

The second-order transition corresponds to the spontaneous selection of a nonzero ordered direction in this order-parameter space. The action also has an approximate larger symmetry that nearly allows independent sign flips of $\hat{c}_1$ and $\hat{c}_G$. This approximate structure is responsible for the later confined reorganization toward the copy mode, but because the exact symmetry has already been broken, this later reorganization is a crossover rather than a second phase transition.

In the softmax-attention case, the softmax map changes the effective geometry of the order-parameter space. Row normalization removes the uniform attention direction as an independent degree of freedom, and the resulting one-dimensional action develops two competing minima. The transition occurs when the global minimum jumps from the weak-copy solution to the strong-copy solution.

A.6 Second-Order Phase Transitions

A *second-order* or *continuous* phase transition is one in which the order parameter turns on continuously at the critical point. In the Landau polynomial of Eq. (35), minimizing $S(m)$ gives

$$m^\star = 0 \quad \text{for} \quad r > 0, \qquad |m^\star| = \sqrt{\frac{-r}{u}} \quad \text{for} \quad r < 0. \tag{40}$$

Thus $m^\star$ approaches zero continuously as $r \to 0^-$. The curvature at the origin,

$$\left.\frac{d^2 S}{dm^2}\right|_{m=0} = r, \tag{41}$$

vanishes at the transition. This vanishing curvature means that fluctuations in the order parameter become large near the critical point:

$$\langle (m - m^\star)^2 \rangle \sim \frac{1}{\Lambda\, S''(m^\star)}. \tag{42}$$

When the curvature tends to zero, the distribution becomes broad in the ordering direction before the ordered solution fully develops.

For linear attention in the main text, the first transition is of this kind. As P crosses the critical value $P^*_{\hat{c}_1}$, the minimum at the origin loses stability and the pooling order parameter $\hat{c}_1$ turns on continuously. The network begins to exploit in-context information, but it initially does so through uniform pooling rather than through a sharp copy pattern.

This distinction matters for predictability. A second-order transition has precursor signals: the action becomes increasingly flat along the unstable direction, and fluctuations in the corresponding order parameter grow. In principle, this can make the approach to the transition detectable before the ordered capability is fully expressed.

A.7 First-Order Phase Transitions

A *first-order* or *discontinuous* phase transition occurs when the global minimum of the action jumps discontinuously from one local minimum to another. Unlike a second-order transition, the original minimum need not lose local stability at the transition. Instead, two locally stable phases coexist at the critical point, and the transition occurs when their action values become equal.

A minimal Landau polynomial with two competing minima is

$$S(m) = \frac{r}{2} m^2 - \frac{w}{4} m^4 + \frac{v}{6} m^6, \qquad w > 0, \quad v > 0. \tag{43}$$

The sixth-order term stabilizes the polynomial, while the negative quartic term allows a nonzero local minimum to appear before the zero-order-parameter minimum becomes unstable. At the transition point, the two minima have equal action and the global minimizer jumps.

The exponential concentration makes this jump sharp. If two local minima are located at $m_1^\star$ and $m_2^\star$, then their posterior weights scale as

$$\frac{p(m_1^\star)}{p(m_2^\star)} \approx \exp[-\Lambda\,(S(m_1^\star) - S(m_2^\star))]\,. \tag{44}$$

For large Λ, even a small difference in action is exponentially amplified. Therefore, when the two action values cross, the dominant saddle switches abruptly.

The corresponding free energy is controlled by the lower of the two saddle values:

$$\mathcal{F} = -\log Z \simeq \Lambda \min\{S(m_1^\star), S(m_2^\star)\} + \text{subleading terms}. \tag{45}$$

At the crossing point, this saddle-point approximation becomes nonanalytic in the large-Λ limit. This nonanalyticity is the phase transition. The polynomial action itself is smooth; the nonanalyticity arises from selecting the dominant minimum.

In the softmax-attention theory, this is precisely the qualitative structure. The action $\mathcal{S}_{\mathrm{softmax}}(\hat{a}_G)$ has two competing minima: one with weak copy structure and one with strong copy structure. As P increases, the strong-copy minimum becomes deeper. At the critical sample size $P^*_{\hat{a}_G}$, the global minimizer changes from one local minimum to the other and the MAP value of $\hat{a}_G$ jumps discontinuously.

This jump in the order parameter produces an abrupt loss drop. It also means there need not be a smooth precursor in the order parameter itself. The weak-copy phase can remain locally stable until the moment the global optimum switches to the copy phase. This is the sense in which the softmax transition is more abrupt and harder to anticipate from smooth progress measures.

A.8 Crossovers

A *crossover* is a rapid but smooth change in behavior that is not a true phase transition. In a crossover, no quantity becomes nonanalytic in the large-system limit, and no symmetry is broken. Nevertheless, the system may show a pronounced change over a narrow range of the control parameter.

This is the interpretation of the later stage of linear-attention learning in the main text. The first true transition turns on the pooling order parameter $\hat{c}_1$. At larger P, the copy direction $\hat{c}_G$ becomes increasingly favorable and the attention pattern reorganizes from uniform pooling toward copying. This reorganization can be sharp in finite experiments and can produce a visible loss decrease.

A.9 Phase Diagrams

A *phase diagram* shows which phase is realized as a function of control parameters. In this paper, we plot quantities such as the loss as a function of P and L, or as a function of P and V.

The theoretical phase boundaries are obtained by analyzing the reduced action. For linear attention, the first boundary is where the curvature in the pooling direction changes sign, giving the second-order transition at $P^*_{\hat{c}_1}$. The second boundary marks the scale where the copy direction becomes favorable and the crossover begins. For softmax attention, the phase boundary is determined by equality of the two competing minima of $\mathcal{S}_{\mathrm{softmax}}(\hat{a}_G)$.

The phase diagrams in the main text should therefore be read as maps of attention behavior: below the boundary, the model is in an unstructured or weakly structured phase; above it, the learned attention pattern implements the copy subcircuit.

A.10 Interpretation for Attention and Emergence

The statistical-physics terminology used in this paper can be summarized as follows. The attention matrix is a high-dimensional object, but the task symmetry identifies a small number of collective directions that matter. These directions are the order parameters. The Bayesian posterior over attention patterns induces an effective action over these order parameters. Because the action is multiplied by a large factor, the posterior concentrates near the minima of this action. As the number of training examples increases, the action changes shape. A phase transition occurs when the dominant saddle, equivalently the global MAP solution in the concentrated limit, changes qualitatively.

For linear attention, the action first undergoes a second-order transition into a pooling phase. The copy component then grows through a crossover. Thus the emergence of copying is preceded by an intermediate, partially useful context-aggregation regime. For softmax attention, the nonlinearity changes the effective action so that two minima coexist. The global minimum jumps directly to the copy phase, yielding a first-order transition and an abrupt loss drop.

This distinction is the main conceptual role of Landau theory in the paper. It explains why two attention mechanisms can solve the same task but display qualitatively different emergence profiles: the mechanism determines the structure of the reduced action, whose saddle points determine the order of the transition.

B Deriving the network prior as a function of the attention

We are ultimately interested in the network posterior conditioned on the training data. Our approach follows a recent line of studies describing the network in terms of the feature kernels [37, 40, 38, 64, 65]. We start by deriving the network prior given inputs $\boldsymbol{X} = (\boldsymbol{x}^a_\alpha)_{\alpha=1,\dots,P;\, a=1,\dots,L}$ and network outputs $\boldsymbol{F} = (\boldsymbol{f}^a_\alpha)_{\alpha=1,\dots,P;\, a=1,\dots,L}$:

$$p(\boldsymbol{F}|\boldsymbol{X}) = \int \prod_{\alpha,a} p(\boldsymbol{f}^a_\alpha|\boldsymbol{x}^a_\alpha, \boldsymbol{\theta})\, p(\boldsymbol{\theta})\, d\boldsymbol{\theta}. \tag{46}$$

For fixed network parameters $\boldsymbol{\theta} \coloneqq \{\boldsymbol{W}^{\text{emb}}, \boldsymbol{p}, \{\boldsymbol{W}_H^G\}, \{\boldsymbol{W}_H^O\}\}$, $H = 1, \ldots, N_H$, the probability $p(\boldsymbol{F}|\boldsymbol{X}, \boldsymbol{\theta})$ is given by enforcing the network model

$$\boldsymbol{h}_\alpha^a = \boldsymbol{W}^{\text{emb}} \boldsymbol{x}_\alpha^a + \boldsymbol{p}^a, \quad \boldsymbol{G}_{H,\alpha}^{ab} = \frac{1}{L\sqrt{d_{\text{model}}}} \boldsymbol{h}_\alpha^{a\intercal} \boldsymbol{W}_H^G \boldsymbol{h}_\alpha^b, \quad \boldsymbol{f}_\alpha^a = \frac{1}{\sqrt{N_H}} \sum_{H,b} \boldsymbol{W}_H^O \boldsymbol{G}_{H,\alpha}^{ab} \boldsymbol{h}_\alpha^b, \tag{47}$$

with Dirac δ-distributions as

$$\begin{aligned} p(\boldsymbol{F}|\boldsymbol{X}, \boldsymbol{\theta}) = \int \mathcal{D}\boldsymbol{h}\, \mathcal{D}\boldsymbol{G} \prod_{a,\alpha} \delta(-\boldsymbol{h}_\alpha^a + \boldsymbol{W}^{\text{emb}} \boldsymbol{x}_\alpha^a + \boldsymbol{p}^a) \prod_{H,a,b,\alpha} \delta\Big(- G_{H,\alpha}^{ab} + \frac{1}{L\sqrt{d_{\text{model}}}} \boldsymbol{h}_\alpha^{a\intercal} \boldsymbol{W}_H^G \boldsymbol{h}_\alpha^b \Big) \\ \times \prod_{a,\alpha} \delta\Big(- \boldsymbol{f}_\alpha^a + \frac{1}{\sqrt{N_H}} \sum_{H,b} \boldsymbol{W}_H^O G_{H,\alpha}^{ab} \boldsymbol{h}_\alpha^b \Big). \end{aligned} \tag{48}$$

We use the Fourier representation of the Dirac δ-distribution

$$\delta(\boldsymbol{z}) = \int \exp\left(\tilde{\boldsymbol{z}}^\intercal \boldsymbol{z}\right) d\tilde{\boldsymbol{z}} \tag{49}$$

with $\tilde{\boldsymbol{z}}$ the conjugate variable to $\boldsymbol{z}$ and $\tilde{\boldsymbol{z}}^\intercal \boldsymbol{z} = \sum_{i=1}^N \tilde{z}_i z_i$, where we integrate along the imaginary axis $\int d\tilde{\boldsymbol{z}} = \prod_k \int_{i\mathbb{R}} \frac{d\tilde{z}_k}{2\pi i}$ for notational convenience, and rewrite

$$\begin{aligned} p(\boldsymbol{F}|\boldsymbol{X}, \boldsymbol{\theta}) = \int \mathcal{D}\boldsymbol{h} \int \mathcal{D}\tilde{\boldsymbol{h}} \exp\Big(- \sum_{a,\alpha} \tilde{\boldsymbol{h}}_\alpha^{a\intercal} \boldsymbol{h}_\alpha^a + \sum_{a,\alpha} \tilde{\boldsymbol{h}}_\alpha^{a\intercal} \boldsymbol{W}^{\text{emb}} \boldsymbol{x}_\alpha^a + \sum_{a,\alpha} \tilde{\boldsymbol{h}}_\alpha^{a\intercal} \boldsymbol{p}^a \Big) \\ \times \prod_H \int \mathcal{D}\boldsymbol{G}_H \int \mathcal{D}\tilde{\boldsymbol{G}}_H \exp\Big(- \sum_{ab,\alpha} \tilde{G}_{H,\alpha}^{ab} G_{H,\alpha}^{ab} + \frac{1}{L\sqrt{d_{\text{model}}}} \sum_{ab,\alpha} \tilde{G}_{H,\alpha}^{ab} \boldsymbol{h}_\alpha^{a\intercal} \boldsymbol{W}_H^G \boldsymbol{h}_\alpha^b \Big) \\ \times \int \mathcal{D}\tilde{\boldsymbol{f}} \exp\Big(- \sum_{a,\alpha} \tilde{\boldsymbol{f}}_\alpha^{a\intercal} \boldsymbol{f}_\alpha^a + \frac{1}{\sqrt{N_H}} \sum_{H,ab,\alpha} \tilde{\boldsymbol{f}}_\alpha^{a\intercal} \boldsymbol{W}_H^O G_{H,\alpha}^{ab} \boldsymbol{h}_\alpha^b \Big), \end{aligned} \tag{50}$$

where we have used the shorthands $\mathcal{D}\boldsymbol{h} = \prod_{a,\alpha} d\boldsymbol{h}_\alpha^a$, $\mathcal{D}\boldsymbol{G}_H = \prod_{ab,\alpha} dG_{H,\alpha}^{ab}$, and similarly for the conjugate fields. All weights are drawn independently, in particular independently across heads, with $W_{H,ij}^G \sim \mathcal{N}(0, g_K g_Q / d_{\text{model}})$ and $W_{ij,H}^O \sim \mathcal{N}(0, g_O / d_{\text{model}})$; the priors of $\boldsymbol{W}^{\text{emb}}$ and $\boldsymbol{p}$ are given below. The number of heads N_H is kept fixed as $d_{\text{model}} \to \infty$.

Embedding Layer

Next, we perform the marginalization over network parameters. In the following, we denote the expectation value over the statistics of network parameters W° as $\langle \ldots \rangle_{W^\circ}$. For the embedding layer, we have the following

$$p(\boldsymbol{H}|\boldsymbol{X}) = \int \mathcal{D}\tilde{\boldsymbol{h}}\; p(\boldsymbol{h}_\alpha^a | \boldsymbol{x}_\alpha^a, \boldsymbol{W}^{\text{emb}}, \boldsymbol{p})\, p(\boldsymbol{W}^{\text{emb}}, \boldsymbol{p})\, d\boldsymbol{W}^{\text{emb}} d\boldsymbol{p} \tag{51}$$

$$= \int \mathcal{D}\tilde{\boldsymbol{h}}\, \langle \exp(- \sum_{a,\alpha,i} \tilde{h}_{i,\alpha}^a h_{i,\alpha}^a + \sum_{a,\alpha,ij} \tilde{h}_{i,\alpha}^a W_{ij}^{\text{emb}} x_{j,\alpha}^a + \sum_{a,\alpha,i} \tilde{h}_{i,\alpha}^a p_i^a) \rangle_{W_{ij}^{\text{emb}}, p_i} \tag{52}$$

$$= \int \mathcal{D}\tilde{\boldsymbol{h}}\, \exp(- \sum_{a,\alpha,i} \tilde{h}_{i,\alpha}^a h_{i,\alpha}^a)\, \langle \exp(\sum_{a,\alpha,ij} \tilde{h}_{i,\alpha}^a W_{ij}^{\text{emb}} x_{j,\alpha}^a) \rangle_{W_{ij}^{\text{emb}}}\, \langle \exp(\sum_{a,\alpha,i} \tilde{h}_{i,\alpha}^a p_i^a) \rangle_{p_i} \tag{53}$$

$$= \int \mathcal{D}\tilde{\boldsymbol{h}}\, \exp(- \sum_{a,\alpha,i} \tilde{h}_{i,\alpha}^a h_{i,\alpha}^a) \exp\left(\frac{1}{2} \frac{g_{\text{emb}}}{V} \sum_{ab,\alpha\beta,ij} \tilde{h}_{i,\alpha}^a \tilde{h}_{i,\beta}^b x_{j,\alpha}^a x_{j,\beta}^b \right) \tag{54}$$

$$\times \exp\left(\frac{1}{2} g_p \sum_{ab,\alpha\beta,i} \tilde{h}_{i,\alpha}^a \delta_{ab} \tilde{h}_{i,\beta}^b \right)$$

$$= \int \mathcal{D}\tilde{\boldsymbol{h}}\, \exp(- \sum_{a,\alpha,i} \tilde{h}_{i,\alpha}^a h_{i,\alpha}^a) \exp\left(\frac{1}{2} \sum_{ab,\alpha\beta,ij} \tilde{h}_{i,\alpha}^a \tilde{h}_{i,\beta}^b \left[\frac{g_{\text{emb}}}{V} x_{j,\alpha}^a x_{j,\beta}^b + g_p \delta_{ab} \right] \right). \tag{55}$$

Integrating the conjugate variables $\tilde{\boldsymbol{h}}$, we find that the token embeddings $h^a_{i,\alpha}$ are i.i.d. Gaussian distributed in model space i with zero mean and covariance matrix

$$\boldsymbol{C}^{(xx)}_{ab,\alpha\beta} = g_{\text{emb}} \sum_j x^a_{j,\alpha} x^b_{j,\beta} + g_p \delta_{ab}. \tag{56}$$

Attention Layer

The attention scores of head H are given by

$$\boldsymbol{G}^{ab}_{H,\alpha} = \frac{1}{L\sqrt{d_{\text{model}}}} \boldsymbol{h}^{a\mathsf{T}}_\alpha \boldsymbol{W}^G_H \boldsymbol{h}^b_\alpha \,. \tag{57}$$

Since the $\boldsymbol{W}^G_H$ are independent across heads, the marginalization factorizes over heads given the embeddings, $p(\{\boldsymbol{G}_H\}|\boldsymbol{H}) = \prod_H p(\boldsymbol{G}_H|\boldsymbol{H})$, with

$$p(\boldsymbol{G}_H|\boldsymbol{H}) = \int \mathcal{D}\tilde{\boldsymbol{G}}_H \left\langle \exp\Big(- \sum_{ab,\alpha} \tilde{G}^{ab}_{H,\alpha} G^{ab}_{H,\alpha} + \sum_{ab,\alpha,ij} \frac{1}{L\sqrt{d_{\text{model}}}} \tilde{G}^{ab}_{H,\alpha} h^a_{i,\alpha} W^G_{H,ij} h^b_{j,\alpha} \Big) \right\rangle_{W^G_{H,ij}} \tag{58}$$

$$= \int \mathcal{D}\tilde{\boldsymbol{G}}_H \, \exp\Big(- \sum_{ab,\alpha} \tilde{G}^{ab}_{H,\alpha} G^{ab}_{H,\alpha} \Big) \, \exp\Big(\frac{1}{2} \frac{g_K \, g_Q}{L^2 d^2_{\text{model}}} \sum_{ab,cd,\alpha\beta} \tilde{G}^{ab}_{H,\alpha} \tilde{G}^{cd}_{H,\beta} (\boldsymbol{h}^a_\alpha \cdot \boldsymbol{h}^c_\beta)(\boldsymbol{h}^b_\alpha \cdot \boldsymbol{h}^d_\beta) \Big) \tag{59}$$

$$= \mathcal{N}\big(\boldsymbol{G}_H \mid 0, C^{(\boldsymbol{GG})}\big), \tag{60}$$

with the same covariance for every head,

$$C^{(\boldsymbol{GG})}_{ab,cd,\alpha\beta} = \frac{g_K \, g_Q}{L^2 d^2_{\text{model}}} (\boldsymbol{h}^a_\alpha \cdot \boldsymbol{h}^c_\beta)(\boldsymbol{h}^b_\alpha \cdot \boldsymbol{h}^d_\beta) \tag{61}$$

To decouple the appearing terms $(\boldsymbol{h}^a_\alpha \cdot \boldsymbol{h}^c_\beta)$ in the covariance, we introduce an auxiliary variable as

$$\boldsymbol{Q}^{ab}_{\alpha\beta} \coloneqq \frac{1}{d_{\text{model}}} \sum_i h^a_{i,\alpha} h^b_{i,\beta}. \tag{62}$$

To obtain the distribution of $\boldsymbol{Q}$, we will make use of the Gärtner-Ellis theorem [63]. For this, we first compute its cumulant-generating function $\mathcal{W}_{\boldsymbol{Q}}$ as

$$\mathcal{W}_{\boldsymbol{Q}}(\tilde{\boldsymbol{Q}}) = \ln \int \mathcal{D}\boldsymbol{h} \, \exp\Big(\tilde{Q}^{ab}_{\alpha\beta} \frac{1}{d_{\text{model}}} \sum_i h^a_{i,\alpha} h^b_{i,\beta} \Big) \, \mathcal{N}(\boldsymbol{h}|0, \boldsymbol{C}^{(xx)}) \tag{63}$$

$$= d_{\text{model}} \, \ln \int \mathcal{D}h \, \exp\Big(\tilde{Q}^{ab}_{\alpha\beta} \frac{1}{d_{\text{model}}} h^a_\alpha h^b_\beta \Big) \, \mathcal{N}(h|0, \boldsymbol{C}^{(xx)}) \tag{64}$$

$$= d_{\text{model}} \, \ln \int \mathcal{D}h \, \exp\Big(- \frac{1}{2} h^a_\alpha \Big[- \frac{2}{d_{\text{model}}} \tilde{Q}^{ab}_{\alpha\beta} + \big(C^{(xx)}_{ab,\alpha\beta}\big)^{-1} \Big] h^b_\beta \Big) \tag{65}$$

$$\times \Big(\det\big(C^{(xx)}_{ab,\alpha\beta}\big) \Big)^{-1/2}$$

$$= d_{\text{model}} \, \ln \Big(\det\big(- \frac{2}{d_{\text{model}}} \tilde{Q}^{ab}_{\alpha\beta} + \big(C^{(xx)}_{ab,\alpha\beta}\big)^{-1} \big) \Big)^{-1/2} \Big(\det\big(C^{(xx)}_{ab,\alpha\beta}\big) \Big)^{-1/2} \tag{66}$$

$$= -\frac{d_{\text{model}}}{2} \ln \det \Big(\boldsymbol{I} - \frac{2}{d_{\text{model}}} \tilde{\boldsymbol{Q}} \boldsymbol{C}^{(xx)} \Big). \tag{67}$$

Since the cumulant-generating function has a scaling form $\mathcal{W}_{\boldsymbol{Q}}(\tilde{\boldsymbol{Q}}) = d_{\text{model}} \, \lambda(\tilde{\boldsymbol{Q}}/d_{\text{model}})$ with λ being a d_{model}-independent function, we may obtain the distribution of the auxiliary variable $\boldsymbol{Q}$ by a large deviation principle as

$$\ln p(\boldsymbol{Q}|C^{(xx)}) = \sup_{\tilde{\boldsymbol{Q}}} \operatorname{Tr}\big[\boldsymbol{Q}^{\mathsf{T}} \tilde{\boldsymbol{Q}}\big] - \mathcal{W}_{\boldsymbol{Q}}(\tilde{\boldsymbol{Q}}) \tag{68}$$

$$=: \Gamma(\boldsymbol{Q}), \tag{69}$$

expressed in terms of the rate function Γ [63].

We compute the rate function via the supremum condition

$$0 \stackrel{!}{=} Q^{sz}_{\mu\nu} - \frac{d_{\text{model}}}{2}\Big[\boldsymbol{I} - \frac{2}{d_{\text{model}}}\tilde{\boldsymbol{Q}}\boldsymbol{C}^{(xx)}\Big]^{-1}_{sa,\mu\alpha} C^{(xx)}_{az,\alpha\nu}, \tag{70}$$

yielding

$$\tilde{\boldsymbol{Q}} = \frac{d_{\text{model}}}{2}\Big(\big[\boldsymbol{C}^{(xx)}\big]^{-1} - \big[\boldsymbol{Q}\big]^{-1}\Big). \tag{71}$$

This gives for the rate function

$$\begin{aligned}\Gamma(\boldsymbol{Q}) &= \operatorname{Tr}\Big[\boldsymbol{Q}\frac{d_{\text{model}}}{2}\Big(\big[\boldsymbol{C}^{(xx)}\big]^{-1} - \big[\boldsymbol{Q}\big]^{-1}\Big)\Big] \qquad (72)\\ &\quad + \frac{d_{\text{model}}}{2}\ln\det\Big(I - \frac{2}{d_{\text{model}}}\frac{d_{\text{model}}}{2}\Big(\big[\boldsymbol{C}^{(xx)}\big]^{-1} - \big[\boldsymbol{Q}\big]^{-1}\Big)\boldsymbol{C}^{(xx)}\Big)\\ &= \frac{d_{\text{model}}}{2}\operatorname{Tr}\Big(\boldsymbol{Q}\big[\boldsymbol{C}^{(xx)}\big]^{-1} - \boldsymbol{I}\Big) + \frac{d_{\text{model}}}{2}\ln\det\Big(\big[\boldsymbol{Q}\big]^{-1}\boldsymbol{C}^{(xx)}\Big). \qquad (73)\end{aligned}$$

For $d_{\text{model}} \to \infty$ at fixed N_H, the first term dominates the rate function $\Gamma(\boldsymbol{Q})$ and $\boldsymbol{Q}$ concentrates on $\boldsymbol{C}^{(xx)}$, so we may replace $\boldsymbol{Q}$ in (60) with $\boldsymbol{C}^{(xx)}$. Thus, we obtain

$$p(\{\boldsymbol{G}_H\}|\boldsymbol{X}) = \prod_H \int \mathcal{D}\tilde{\boldsymbol{G}}_H \, \exp\Big(-\sum_{ab,\alpha}\tilde{G}^{ab}_{H,\alpha}G^{ab}_{H,\alpha}\Big)\,\exp\Big(\frac{1}{2}\,\frac{g_K\,g_Q}{L^2}\sum_{ab,cd,\alpha\beta}\tilde{G}^{ab}_{H,\alpha}\tilde{G}^{cd}_{H,\beta}C^{(xx)}_{ac,\alpha\beta}C^{(xx)}_{bd,\alpha\beta}\Big) \tag{74}$$

$$= \prod_H \exp\Big(-\frac{1}{2}\sum_{ab,cd,\alpha\beta}G^{ab}_{H,\alpha}\Big[\frac{g_K\,g_Q}{L^2}C^{(xx)}_{ac,\alpha\beta}C^{(xx)}_{bd,\alpha\beta}\Big]^{-1}G^{cd}_{H,\beta} - \ln\det\Big[\frac{g_K\,g_Q}{L^2}C^{(xx)}_{ac,\alpha\beta}C^{(xx)}_{bd,\alpha\beta}\Big]^{1/2}\Big). \tag{75}$$

For notational brevity, we denote $g_W \coloneqq g_K g_Q$ in the following.

Attention Layer parameterized with W^K, W^Q

We show the derivation for a single head and drop the head index; since the key and query weights are independent across heads, the result applies to each head separately. The attention mechanism is given by

$$\boldsymbol{G}^{ab}_{\alpha} = \frac{1}{L\sqrt{d_{\text{model}}}}h^a_{i,\alpha}W^Q_{li}W^K_{lj}h^b_{j,\alpha}. \tag{76}$$

We now marginalize with respect to the W^K and W^Q,

$$p(\boldsymbol{G}|\boldsymbol{H}) = \int \mathcal{D}\tilde{\boldsymbol{G}}\,\Big\langle \exp\Big(-\sum_{ab,\alpha}\tilde{G}^{ab}_{\alpha}G^{ab}_{\alpha} + \frac{1}{L\sqrt{d_{\text{model}}}}\sum_{ab,\alpha,ijl}\tilde{G}^{ab}_{\alpha}h^a_{i,\alpha}W^Q_{li}W^K_{lj}h^b_{j,\alpha}\Big)\Big\rangle_{\boldsymbol{W}^Q,\boldsymbol{W}^K}. \tag{77}$$

Using independence in the index l, this becomes

$$p(\boldsymbol{G}|\boldsymbol{H}) = \int \mathcal{D}\tilde{\boldsymbol{G}}\,\exp\Big(-\sum_{ab,\alpha}\tilde{G}^{ab}_{\alpha}G^{ab}_{\alpha} + \mathcal{W}_{\boldsymbol{G}}(\tilde{\boldsymbol{G}})\Big), \tag{78}$$

where

$$\mathcal{W}_{\boldsymbol{G}}(\tilde{\boldsymbol{G}}) = d_k \ln\Big\langle\Big\langle \exp\Big(\frac{1}{L\sqrt{d_{\text{model}}}}\sum_{ab,\alpha,ij}\tilde{G}^{ab}_{\alpha}h^a_{i,\alpha}W^Q_{li}W^K_{lj}h^b_{j,\alpha}\Big)\Big\rangle_{\boldsymbol{W}^K}\Big\rangle_{\boldsymbol{W}^Q}. \tag{79}$$

Performing first the Gaussian integral over the row W^K_{lj} gives

$$\mathcal{W}_{\boldsymbol{G}}(\tilde{\boldsymbol{G}}) = d_k \ln\Big\langle \exp\Big(\frac{g_K}{2L^2 d^2_{\text{model}}}\sum_{ab,cd,\alpha\beta,ijk}\tilde{G}^{ab}_{\alpha}\tilde{G}^{cd}_{\beta}h^a_{i,\alpha}W^Q_{li}h^b_{j,\alpha}h^d_{j,\beta}h^c_{k,\beta}W^Q_{lk}\Big)\Big\rangle_{\boldsymbol{W}^Q}. \tag{80}$$

To decouple the appearing hidden-state contractions, we introduce an auxiliary variable as

$$Q^{ab}_{\alpha\beta} \coloneqq \frac{1}{d_{\text{model}}} \sum_i h^a_{i,\alpha} h^b_{i,\beta}. \tag{81}$$

Then

$$\mathcal{W}_{\boldsymbol{G}}(\tilde{\boldsymbol{G}}) = d_k \ln \Big\langle \exp\Big(\frac{g_K}{2L^2 d_{\text{model}}} \sum_{ab,cd,\alpha\beta,ik} \tilde{G}^{ab}_{\alpha} \tilde{G}^{cd}_{\beta} Q^{bd}_{\alpha\beta} h^a_{i,\alpha} W^Q_{li} h^c_{k,\beta} W^Q_{lk} \Big) \Big\rangle_{\boldsymbol{W}^Q}. \tag{82}$$

The remaining Gaussian integral is over the feature-indexed row vector W^Q_{li}, and therefore gives a determinant over the feature indices i, k:

$$\mathcal{W}_{\boldsymbol{G}}(\tilde{\boldsymbol{G}}) = -\frac{d_k}{2} \ln \det_{ik} \left[\delta_{ik} - \frac{g_Q g_K}{L^2 d^2_{\text{model}}} \sum_{ab,cd,\alpha\beta} \tilde{G}^{ab}_{\alpha} Q^{bd}_{\alpha\beta} \tilde{G}^{cd}_{\beta} h^a_{i,\alpha} h^c_{k,\beta} \right]. \tag{83}$$

We now rewrite this determinant in the token-sample space. Let the combined token-sample index be denoted by $(a\alpha)$, and define

$$H_{i,a\alpha} \coloneqq h^a_{i,\alpha}, \qquad \widehat{\tilde{G}}_{a\alpha,b\beta} \coloneqq \tilde{G}^{ab}_{\alpha} \delta_{\alpha\beta}, \qquad Q_{a\alpha,b\beta} \coloneqq Q^{ab}_{\alpha\beta}. \tag{84}$$

Then Eq. (83) can be written in matrix form as

$$\mathcal{W}_{\boldsymbol{G}}(\tilde{\boldsymbol{G}}) = -\frac{d_k}{2} \ln \det_{i} \left[\boldsymbol{I} - \frac{g_Q g_K}{L^2 d^2_{\text{model}}} \boldsymbol{H} \widehat{\tilde{\boldsymbol{G}}} \boldsymbol{Q} \widehat{\tilde{\boldsymbol{G}}}^{\mathsf{T}} \boldsymbol{H}^{\mathsf{T}} \right]. \tag{85}$$

The matrix inside the feature-space determinant is symmetric. We factor it as

$$\boldsymbol{H} \widehat{\tilde{\boldsymbol{G}}} \boldsymbol{Q} \widehat{\tilde{\boldsymbol{G}}}^{\mathsf{T}} \boldsymbol{H}^{\mathsf{T}} = \left(\boldsymbol{H} \widehat{\tilde{\boldsymbol{G}}} \boldsymbol{Q}^{1/2} \right) \left(\boldsymbol{H} \widehat{\tilde{\boldsymbol{G}}} \boldsymbol{Q}^{1/2} \right)^{\mathsf{T}}. \tag{86}$$

Using the Weinstein–Aronszajn identity,

$$\det(\boldsymbol{I} - \boldsymbol{A}\boldsymbol{A}^{\mathsf{T}}) = \det(\boldsymbol{I} - \boldsymbol{A}^{\mathsf{T}}\boldsymbol{A}), \tag{87}$$

together with

$$\boldsymbol{H}^{\mathsf{T}} \boldsymbol{H} = d_{\text{model}}\, \boldsymbol{Q}, \tag{88}$$

we obtain

$$\mathcal{W}_{\boldsymbol{G}}(\tilde{\boldsymbol{G}}) = -\frac{d_k}{2} \ln \det \left[\boldsymbol{I} - \frac{g_Q g_K}{L^2 d_{\text{model}}} \boldsymbol{Q}^{1/2} \widehat{\tilde{\boldsymbol{G}}}^{\mathsf{T}} \boldsymbol{Q} \widehat{\tilde{\boldsymbol{G}}} \boldsymbol{Q}^{1/2} \right]. \tag{89}$$

To obtain the distribution of $\boldsymbol{Q}$, we will make use of the Gärtner-Ellis theorem [63]. For this, we first compute its cumulant-generating function $\mathcal{W}_{\boldsymbol{Q}}$ as

$$\mathcal{W}_{\boldsymbol{Q}}(\tilde{\boldsymbol{Q}}) = \ln \int \mathcal{D}\boldsymbol{h}\, \exp\Big(\tilde{Q}^{ab}_{\alpha\beta} \frac{1}{d_{\text{model}}} \sum_i h^a_{i,\alpha} h^b_{i,\beta} \Big) \mathcal{N}(\boldsymbol{h}|0, \boldsymbol{C}^{(xx)}) \tag{90}$$

$$= d_{\text{model}} \ln \int \mathcal{D}h\, \exp\Big(\tilde{Q}^{ab}_{\alpha\beta} \frac{1}{d_{\text{model}}} h^a_{\alpha} h^b_{\beta} \Big) \mathcal{N}(h|0, \boldsymbol{C}^{(xx)}) \tag{91}$$

$$= d_{\text{model}} \ln \int \mathcal{D}h\, \exp\Big(-\frac{1}{2} h^a_{\alpha} \Big[-\frac{2}{d_{\text{model}}} \tilde{Q}^{ab}_{\alpha\beta} + \big(C^{(xx)}_{ab,\alpha\beta}\big)^{-1} \Big] h^b_{\beta} \Big) \Big(\det \boldsymbol{C}^{(xx)} \Big)^{-1/2} \tag{92}$$

$$= -\frac{d_{\text{model}}}{2} \ln \det \Big(\boldsymbol{I} - \frac{2}{d_{\text{model}}} \tilde{\boldsymbol{Q}} \boldsymbol{C}^{(xx)} \Big). \tag{93}$$

Since the cumulant-generating function has a scaling form

$$\lambda_{\boldsymbol{Q}} = \frac{1}{d_{\text{model}}} \mathcal{W}_{\boldsymbol{Q}}(d_{\text{model}} \tilde{\boldsymbol{Q}}) = -\frac{1}{2} \ln \det \Big(\boldsymbol{I} - 2\tilde{\boldsymbol{Q}} \boldsymbol{C}^{(xx)} \Big) \tag{94}$$

we may obtain the distribution of the auxiliary variable $\boldsymbol{Q}$ by a large deviation principle as

$$\ln p(\boldsymbol{Q}|C^{(xx)}) = -d_{\text{model}}\Gamma(\boldsymbol{Q}), \tag{95}$$

$$\Gamma(\boldsymbol{Q}) = \sup_{\tilde{\boldsymbol{Q}}} \left\{ \text{Tr}[\boldsymbol{Q}^{\mathsf{T}}\tilde{\boldsymbol{Q}}] - \lambda_{\boldsymbol{Q}}(\tilde{\boldsymbol{Q}}) \right\}. \tag{96}$$

We compute the rate function Γ by evaluating the supremum condition

$$0 = Q^{sz}_{\mu\nu} - \frac{d_{\text{model}}}{2}\left[\boldsymbol{I} - \frac{2}{d_{\text{model}}}\tilde{\boldsymbol{Q}}\boldsymbol{C}^{(xx)}\right]^{-1}_{sa,\mu\alpha}\frac{2}{d_{\text{model}}}C^{(xx)}_{az,\alpha\nu}, \tag{97}$$

yielding

$$\tilde{\boldsymbol{Q}} = \frac{d_{\text{model}}}{2}\left([\boldsymbol{C}^{(xx)}]^{-1} - [\boldsymbol{Q}]^{-1}\right). \tag{98}$$

This gives for the rate function

$$\Gamma(\boldsymbol{Q}) = \frac{d_{\text{model}}}{2}\text{Tr}\Big(\boldsymbol{Q}[\boldsymbol{C}^{(xx)}]^{-1} - \boldsymbol{I}\Big) + \frac{d_{\text{model}}}{2}\ln\det\left([\boldsymbol{Q}]^{-1}\boldsymbol{C}^{(xx)}\right). \tag{99}$$

For $d_{\text{model}} \to \infty$, $\boldsymbol{Q}$ concentrates on $\boldsymbol{C}^{(xx)}$, so we may replace $\boldsymbol{Q}$ in Eq. (89) with $\boldsymbol{C}^{(xx)}$. Defining

$$\boldsymbol{C}_{a\alpha,b\beta} \coloneqq C^{(xx)}_{ab,\alpha\beta}, \tag{100}$$

we obtain

$$\mathcal{W}_{\boldsymbol{G}}(\tilde{\boldsymbol{G}}) = -\frac{d_k}{2}\ln\det\left[\boldsymbol{I} - \frac{g_Q g_K}{L^2 d_{\text{model}}}\boldsymbol{C}^{1/2}\widehat{\tilde{\boldsymbol{G}}}^{\mathsf{T}}\boldsymbol{C}\widehat{\tilde{\boldsymbol{G}}}\boldsymbol{C}^{1/2}\right]. \tag{101}$$

We now evaluate the remaining integral over $\tilde{\boldsymbol{G}}$ by the same large-deviation argument. Define

$$\widehat{G}_{a\alpha,b\beta} \coloneqq G^{ab}_{\alpha}\delta_{\alpha\beta}. \tag{102}$$

We define the "whiten" matrices

$$\bar{\boldsymbol{G}} \coloneqq \boldsymbol{C}^{-1/2}\widehat{\boldsymbol{G}}\boldsymbol{C}^{-1/2}, \qquad \bar{\tilde{\boldsymbol{G}}} \coloneqq \boldsymbol{C}^{1/2}\widehat{\tilde{\boldsymbol{G}}}\boldsymbol{C}^{1/2}. \tag{103}$$

Then

$$\sum_{ab,\alpha}\tilde{G}^{ab}_{\alpha}G^{ab}_{\alpha} = \text{Tr}[\bar{\tilde{\boldsymbol{G}}}^{\mathsf{T}}\bar{\boldsymbol{G}}], \tag{104}$$

and

$$\mathcal{W}_{\boldsymbol{G}}(\bar{\tilde{\boldsymbol{G}}}) = -\frac{d_k}{2}\ln\det\left[\boldsymbol{I} - \frac{g_Q g_K}{L^2 d_{\text{model}}}\bar{\tilde{\boldsymbol{G}}}^{\mathsf{T}}\bar{\tilde{\boldsymbol{G}}}\right]. \tag{105}$$

Using $k = \frac{d_{\text{model}}}{d_k} = O(1)$, where k is often taken to be equal to the number of heads, so for our single-head attention we specialize to the case $k = 1 \Rightarrow d_{\text{model}} = d_k$ [3]. We can expand the cumulant-generating function in a power series using $\ln\det = \text{Tr}\ln$

$$\frac{d_k}{2}\text{Tr}\left[\sum_{n=1}^{\infty}\frac{1}{n}\left(\frac{g_Q g_K}{L^2 d_{\text{model}}}\bar{\tilde{\boldsymbol{G}}}^{\mathsf{T}}\bar{\tilde{\boldsymbol{G}}}\right)^n\right] = \frac{d_k}{2}\left(\frac{g_Q g_K}{L^2 d_{\text{model}}}\text{Tr}\left[\bar{\tilde{\boldsymbol{G}}}^{\mathsf{T}}\bar{\tilde{\boldsymbol{G}}}\right] + O\left(\text{Tr}\left[\left(\frac{g_Q g_K}{L^2 d_{\text{model}}}\bar{\tilde{\boldsymbol{G}}}^{\mathsf{T}}\bar{\tilde{\boldsymbol{G}}}\right)^2\right]\right)\right) \tag{106}$$

So the expansion's control parameter is $\epsilon \coloneqq \frac{g_Q g_K}{L^2 d_{\text{model}}}\left\|\bar{\tilde{\boldsymbol{G}}}\right\|^2_{op}$ and we may truncate the series at Gaussian order if it is small. We thus turn to estimate $\left\|\bar{\tilde{\boldsymbol{G}}}\right\|^2_{op}$ self-consistently. At this Gaussian order we see

$$\bar{\boldsymbol{G}} \sim \frac{g_Q g_K}{L^2} \quad \Rightarrow \quad \bar{\tilde{\boldsymbol{G}}} \sim L\sqrt{\frac{1}{g_Q g_K}}. \tag{107}$$

Since $\bar{\tilde{\boldsymbol{G}}}$ is an $L \times L$ matrix it operator norm is of order

$$\left\|\bar{\tilde{\boldsymbol{G}}}\right\|^2_{op} \sim \frac{L^3}{g_Q g_K} \tag{108}$$

[3] the result is unchanged for any $k = O(1)$; for general k the expansion parameter below becomes $\epsilon \sim L/d_k$

we find our expansion parameter ϵ is of order

$$\epsilon \sim \frac{L}{d_{\text{model}}}. \tag{109}$$

We see that under our scaling, where $d_{\text{model}} \gg L$ the expansion is consistent and we may approximate

$$\mathcal{W}_{\boldsymbol{G}}(\bar{\tilde{\boldsymbol{G}}}) \approx \frac{d_k g_Q g_K}{2L^2 d_{\text{model}}} \operatorname{Tr}\left[\bar{\tilde{\boldsymbol{G}}}^{\mathsf{T}} \bar{\tilde{\boldsymbol{G}}}\right] \tag{110}$$

which agrees with the central limit theorem intuition that when we sum a large number of modes d_{model} into a small matrix L we should get Gaussian behavior. We conclude the prior distribution for the attention pattern $\boldsymbol{G}$ is

$$p(\boldsymbol{G}|\boldsymbol{X}) = \exp\Big(-\frac{1}{2} \sum_{ab,cd,\alpha\beta} G^{ab}_{\alpha} \Big[\frac{d_k g_K\, g_Q}{d_{\text{model}} L^2} C^{(xx)}_{ac,\alpha\beta} C^{(xx)}_{bd,\alpha\beta}\Big]^{-1} G^{cd}_{\beta} - \ln\det\Big[\frac{d_k g_K\, g_Q}{d_{\text{model}} L^2} C^{(xx)}_{ac,\alpha\beta} C^{(xx)}_{bd,\alpha\beta}\Big]^{1/2}\Big). \tag{111}$$

For notational brevity, we denote $g_W \coloneqq g_K g_Q d_k / d_{\text{model}}$ in the following.

Readout Layer

Finally, we look at the network output and marginalize over the readout weights, which are independent across heads,

$$p(\boldsymbol{F}|\{\boldsymbol{G}_H\}) = \int \mathcal{D}\tilde{\boldsymbol{f}} \left\langle \exp\Big(-\sum_{a,\alpha,i} \tilde{f}^a_{i,\alpha} f^a_{i,\alpha} + \frac{1}{\sqrt{N_H}} \sum_H \sum_{ab,\alpha,ij} \tilde{f}^a_{i,\alpha} W^O_{ij,H} G^{ab}_{H,\alpha} h^b_{j,\alpha} \Big) \right\rangle_{\{W^O_{ij,H}\}} \tag{112}$$

$$= \int \mathcal{D}\tilde{\boldsymbol{f}}\, \exp\Big(-\sum_{a,\alpha,i} \tilde{f}^a_{i,\alpha} f^a_{i,\alpha} \Big)\, \exp\Big(\frac{1}{2} \frac{g_O}{N_H d_{\text{model}}} \sum_H \sum_{abcd,\alpha\beta,i} \tilde{f}^a_{i,\alpha} \tilde{f}^c_{i,\beta}\, G^{ab}_{H,\alpha} G^{cd}_{H,\beta} \sum_j h^b_{j,\alpha} h^d_{j,\beta} \Big) \tag{113}$$

$$= \int \mathcal{D}\tilde{\boldsymbol{f}}\, \exp\Big(-\sum_{a,\alpha,i} \tilde{f}^a_{i,\alpha} f^a_{i,\alpha} \Big)\, \exp\Big(\frac{1}{2} \frac{g_O}{N_H} \sum_H \sum_{abcd,\alpha\beta,i} \tilde{f}^a_{i,\alpha} \tilde{f}^c_{i,\beta}\, G^{ab}_{H,\alpha} G^{cd}_{H,\beta}\, Q^{bd}_{\alpha\beta} \Big), \tag{114}$$

where we have identified the last sum term as the auxiliary variable $\boldsymbol{Q}$ defined in (62). For $d_{\text{model}} \gg 1$, $\boldsymbol{Q}$ concentrates and we may replace by $\boldsymbol{C}^{(xx)}$ as in (75), yielding

$$p(\boldsymbol{F}|\{\boldsymbol{G}_H\}) = \int \mathcal{D}\tilde{\boldsymbol{f}}\, \exp\Big(-\sum_{a,\alpha,i} \tilde{f}^a_{i,\alpha} f^a_{i,\alpha} \Big)\, \exp\Big(\frac{1}{2} \frac{g_O}{N_H} \sum_H \sum_{abcd,\alpha\beta,i} \tilde{f}^a_{i,\alpha} \tilde{f}^c_{i,\beta}\, G^{ab}_{H,\alpha} G^{cd}_{H,\beta}\, C^{(xx)}_{bd,\alpha\beta} \Big). \tag{115}$$

C Deriving the Network Posterior

We are interested in the posterior over the attention patterns $\{\boldsymbol{G}_H\}$. For the posterior, we assume Gaussian noise on the network outputs $\boldsymbol{y}_\alpha = \boldsymbol{f}_\alpha + \boldsymbol{\xi}_\alpha$ with $\boldsymbol{\xi}_\alpha \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, \sigma^2 \boldsymbol{I})$. Then, we obtain

$$p(\{\boldsymbol{G}_H\}|\{\boldsymbol{x}_\alpha, \boldsymbol{y}_\alpha\}_\alpha) \propto p(\{\boldsymbol{G}_H\}|\boldsymbol{X}) \int \mathcal{D}\boldsymbol{f}\, p(\{\boldsymbol{y}_\alpha\}|\boldsymbol{F})\, p(\boldsymbol{F}|\{\boldsymbol{G}_H\}) \tag{116}$$

$$\propto p(\{\boldsymbol{G}_H\}|\boldsymbol{X}) \int \mathcal{D}\boldsymbol{f}\, \exp\Big(-\frac{1}{2\sigma^2} \sum_\alpha \|\boldsymbol{y}_\alpha - \boldsymbol{f}_\alpha\|_2^2 \Big)\, p(\boldsymbol{F}|\{\boldsymbol{G}_H\}) \tag{117}$$

$$= \int \mathcal{D}\boldsymbol{f} \int \mathcal{D}\tilde{\boldsymbol{f}}\, \exp\Big(-\frac{1}{2\sigma^2} \sum_\alpha \|\boldsymbol{y}_\alpha - \boldsymbol{f}_\alpha\|_2^2 - \tilde{\boldsymbol{f}}_\alpha^{\mathsf{T}} \boldsymbol{f}_\alpha \Big) \tag{118}$$

$$\times \exp\Big(\frac{1}{2} \frac{g_O}{N_H} \sum_H \sum_{abcd,\alpha\beta,i} \tilde{f}^a_{i,\alpha} \tilde{f}^c_{i,\beta}\, G^{ab}_{H,\alpha} G^{cd}_{H,\beta}\, C^{(xx)}_{bd,\alpha\beta} \Big)\, p(\{\boldsymbol{G}_H\}|\boldsymbol{X}), \tag{119}$$

where we drop the normalization constant throughout since we are ultimately interested in the most likely values of the attention in the posterior. Integrating out the network outputs $\boldsymbol{f}$ in the first term yields

$$\int \mathcal{D}\boldsymbol{f}\, \exp\Big(-\frac{1}{2\sigma^2}\sum_\alpha \|\boldsymbol{y}_\alpha - \boldsymbol{f}_\alpha\|_2^2 - \tilde{\boldsymbol{f}}_\alpha^\intercal \boldsymbol{f}_\alpha\Big) \propto \exp\Big(\sum_\alpha \frac{\sigma^2}{2}\tilde{\boldsymbol{f}}_\alpha^\intercal \tilde{\boldsymbol{f}}_\alpha - \tilde{\boldsymbol{f}}_\alpha^\intercal \boldsymbol{y}_\alpha\Big). \tag{120}$$

Overall, this gives

$$p(\{\boldsymbol{G}_H\}|\{\boldsymbol{x}_\alpha, \boldsymbol{y}_\alpha\}_\alpha) \propto \int \mathcal{D}\tilde{\boldsymbol{f}}\, \exp\Big(-\sum_\alpha \tilde{\boldsymbol{f}}_\alpha^\intercal \boldsymbol{y}_\alpha\Big) \tag{121}$$

$$\times \exp\left(\frac{1}{2}\sum_{abcd,\alpha\beta,i} \tilde{f}^a_{i,\alpha}\Big[\frac{g_O}{N_H}\sum_H G^{ab}_{H,\alpha}G^{cd}_{H,\beta}\, C^{(xx)}_{bd,\alpha\beta} + \sigma^2\boldsymbol{I}\Big]\tilde{f}^c_{i,\beta}\right) p(\{\boldsymbol{G}_H\}|\boldsymbol{X}) \tag{122}$$

$$= \exp\Bigg(-\frac{1}{2}\sum_{abcd,\alpha\beta,i} y^a_{i,\alpha}\Big[\frac{g_O}{N_H}\sum_H G^{ab}_{H,\alpha}G^{cd}_{H,\beta}\, C^{(xx)}_{bd,\alpha\beta} + \sigma^2\boldsymbol{I}\Big]^{-1} y^c_{i,\beta}$$
$$-\frac{1}{2}\ln\det\Big[\frac{g_O}{N_H}\sum_H G^{ab}_{H,\alpha}G^{cd}_{H,\beta}\, C^{(xx)}_{bd,\alpha\beta} + \sigma^2\boldsymbol{I}\Big]\Bigg) p(\{\boldsymbol{G}_H\}|\boldsymbol{X}). \tag{123}$$

Since we are interested in the dominant attention behavior of the posterior, we may focus on the negative exponent or action $\mathcal{S}(\{\boldsymbol{G}_H\}) \coloneqq -\ln p(\{\boldsymbol{G}_H\}|\{\boldsymbol{x}_\alpha, \boldsymbol{y}_\alpha\}_\alpha)$. Inserting the result for $p(\{\boldsymbol{G}_H\}|\boldsymbol{X})$ from (75), we have

$$\begin{aligned}\mathcal{S}(\{\boldsymbol{G}_H\}) = &\frac{1}{2}\sum_{abcd,\alpha\beta,i} y^a_{i,\alpha}\Big[\frac{g_O}{N_H}\sum_H G^{ab}_{H,\alpha}G^{cd}_{H,\beta}\, C^{(xx)}_{bd,\alpha\beta} + \sigma^2\boldsymbol{I}\Big]^{-1} y^c_{i,\beta}\\ &+\frac{1}{2}\ln\det\Big[\frac{g_O}{N_H}\sum_H G^{ab}_{H,\alpha}G^{cd}_{H,\beta}\, C^{(xx)}_{bd,\alpha\beta} + \sigma^2\boldsymbol{I}\Big]\\ &+\sum_H\Big(\frac{1}{2}\sum_{ab,cd,\alpha\beta} G^{ab}_{H,\alpha}\Big[\frac{g_W}{L^2}C^{(xx)}_{ac,\alpha\beta}C^{(xx)}_{bd,\alpha\beta}\Big]^{-1}G^{cd}_{H,\beta}\\ &+\frac{1}{2}\ln\det\Big[\frac{g_W}{L^2}C^{(xx)}_{ac,\alpha\beta}C^{(xx)}_{bd,\alpha\beta}\Big]\Big).\end{aligned} \tag{124}$$

The first two terms in the log probability are the likelihood of the labels given the data, which couples the heads only through the summed kernel. The last term is the prior term, which separates over heads and is for each head identical to the one found in NNGP theory [66, 67].

We now look at the scaling of the different terms. We downscale the output weights, the attention weights and the noise by a large factor χ, similar to taking μP scaling (so-called "rich regime") [51], such that the transformation is

$$g_W \to \frac{g_W}{\chi}, \qquad g_O \to \frac{g_O}{\chi}, \qquad \sigma^2 \to \frac{\sigma^2}{\chi}. \tag{125}$$

Intuitively, the scaling with χ suppresses fluctuations in directions that are irrelevant to performing the task. Consequently, we may neglect the term $\ln\det$, which is a direct result of the Gaussian fluctuations of $\tilde{\boldsymbol{f}}$. This scaling also increases the overall scale of the action and thus the sharpness of the probability landscape in $\{\boldsymbol{G}_H\}$.

Explicitly, we have with these scalings

$$\begin{aligned}\mathcal{S}(\{\boldsymbol{G}_H\}) = &\chi\frac{1}{2}\sum_{abcd,\alpha\beta,i} y^a_{i,\alpha}\Big[\frac{g_O}{N_H}\sum_H G^{ab}_{H,\alpha}G^{cd}_{H,\beta}\, C^{(xx)}_{bd,\alpha\beta} + \sigma^2\boldsymbol{I}\Big]^{-1} y^c_{i,\beta}\\ &+\frac{1}{2}\ln\det\Big[\frac{g_O}{N_H}\sum_H G^{ab}_{H,\alpha}G^{cd}_{H,\beta}\, C^{(xx)}_{bd,\alpha\beta} + \sigma^2\boldsymbol{I}\Big]\\ &+L^2\chi\frac{1}{2}\sum_H\sum_{ab,cd,\alpha\beta} G^{ab}_{H,\alpha}\big[g_W C^{(xx)}_{ac,\alpha\beta}C^{(xx)}_{bd,\alpha\beta}\big]^{-1}G^{cd}_{H,\beta}\\ &+\frac{N_H}{2}\ln\det\big[g_W C^{(xx)}_{ac,\alpha\beta}C^{(xx)}_{bd,\alpha\beta}\big] + const.(\{\boldsymbol{G}_H\})\end{aligned} \tag{126}$$

D Projecting the NNGP term onto the order parameters

We first look at the term

$$\mathcal{S}_{\mathrm{NNGP}}(G) \;=\; \sum_{ab,\,cd,\,\alpha\beta} G^{ab}\,\big[\frac{2g_W}{L^2}\mathcal{M}_{ab\alpha,\,cd\beta}\big]^{\dagger}\,G^{cd}\,, \tag{127}$$

where $\mathcal{M}_{ab\alpha,\,cd\beta} = C^{(xx)}_{ac,\alpha\beta}\,C^{(xx)}_{bd,\alpha\beta}$ is the NNGP kernel built from the input kernel (4) with $g_{\mathrm{emb}} = g_p = 1$ as in our experiments, $C^{(xx)}_{ac,\alpha\beta} = \sum_j x^a_{j\alpha}\,x^c_{j\beta} + \delta^{ac}$. We will project this operator onto the attention patterns $(\mathbf{1}_L, \boldsymbol{G}^*)$ and in the end transform to the normalized attention patterns with coefficients $(\hat{c}_1, \hat{c}_G)$ as in (11) of the main text. We use the pseudo-inverse denoted by $\dagger$ here to account for zero eigenvalues of the matrix.

We restrict to attention patterns that do not depend on the sample index, $G^{ab}_\alpha = G^{ab}$, as in the ansatz (11). The quadratic form (127) then involves only the summed pseudo-inverse

$$\Sigma_{(ab),(cd)} \;=\; \sum_{\alpha,\beta=1}^{P} \big(\mathcal{M}^{\dagger}\big)_{ab\alpha,\,cd\beta}\,. \tag{128}$$

Each data sample α is a sequence $\sigma_\alpha\colon \{1,\dots,L\} \to \{1,\dots,V\}$, one-hot encoded as $x^a_{j\alpha} = \delta_{j,\sigma_\alpha(a)}$. The data overlap matrix $D^{(\alpha\beta)}_{ac} = \delta_{\sigma_\alpha(a),\sigma_\beta(c)}$ gives the input kernel as $C^{(\alpha\beta)}_{ac} = D^{(\alpha\beta)}_{ac} + \delta_{ac}$. Introducing the embedding matrix $\Phi_\alpha = [\mathrm{Id}_L \mid X_\alpha]$, which extends each data vector $(X_\alpha)_{aj} = \delta_{j,\sigma_\alpha(a)}$ by the identity matrix Id_L to represent the positional encoding, we have $C^{(\alpha\beta)} = \Phi_\alpha\Phi_\beta^{\mathsf{T}}$, and the (α,β)-block of the forward operator takes the Kronecker form

$$\mathcal{S}_{\alpha\beta} \;=\; C^{(\alpha\beta)} \otimes C^{(\alpha\beta)} \;=\; (\Phi_\alpha \otimes \Phi_\alpha)\,(\Phi_\beta \otimes \Phi_\beta)^{\mathsf{T}}\,. \tag{129}$$

Setting $\Psi_\alpha = \Phi_\alpha \otimes \Phi_\alpha$ and stacking over samples gives $\mathcal{S} = \Psi\Psi^{\mathsf{T}}$. Since $\Phi_\alpha E = \mathrm{Id}_L$ for $E = \begin{pmatrix}\mathrm{Id}_L\\ 0\end{pmatrix}$, every sample-independent pattern satisfies $G = \Phi_\alpha(EGE^{\mathsf{T}})\Phi_\alpha^{\mathsf{T}}$ for all α and thus lies in $\mathrm{range}(\Psi) = \mathrm{range}(\mathcal{S})$, the support of the Gaussian prior, on which the pseudo-inverse in (127) gives its negative log-density up to a $\boldsymbol{G}$-independent constant.

Closed form for Σ

We now determine Σ in closed form. Regarding notation, we write $\boldsymbol{u}_n = (1,\dots,1)^{\mathsf{T}} \in \mathbb{R}^n$ for the all-ones vector, so that $\boldsymbol{J}_L = \boldsymbol{u}_L\boldsymbol{u}_L^{\mathsf{T}}$, and we use the Kronecker product $(A\otimes B)_{(ab),(ij)} = A_{ai}B_{bj}$ throughout. Besides $\mathcal{S} = \Psi\Psi^{\mathsf{T}}$, which is $\mathcal{M}$ written as an $L^2P \times L^2P$ matrix, we define

$$\mathcal{R} \;:=\; \Psi^{\mathsf{T}}\Psi \;=\; \sum_\alpha \Gamma_\alpha \otimes \Gamma_\alpha\,, \qquad \Gamma_\alpha := \Phi_\alpha^{\mathsf{T}}\Phi_\alpha\,, \qquad \mathcal{B} \;:=\; \sum_\alpha \Psi_\alpha^{\mathsf{T}} \;=\; \sum_\alpha \Phi_\alpha^{\mathsf{T}} \otimes \Phi_\alpha^{\mathsf{T}}\,, \tag{130}$$

of sizes $(L+V)^2 \times (L+V)^2$ and $(L+V)^2 \times L^2$, respectively. Note that $\mathcal{R}$ is symmetric and positive semi-definite.

We start by rewriting Σ as a matrix product. Let $\Psi = U\Delta\tilde{U}^{\mathsf{T}}$ be the compact singular value decomposition, with Δ the diagonal matrix of the nonzero singular values. Then $\mathcal{M}^\dagger = U\Delta^{-2}U^{\mathsf{T}}$ and $(\mathcal{R}^\dagger)^2 = \tilde{U}\Delta^{-4}\tilde{U}^{\mathsf{T}}$, and therefore

$$\Psi\,(\mathcal{R}^\dagger)^2\,\Psi^{\mathsf{T}} \;=\; U\Delta\tilde{U}^{\mathsf{T}}\,\tilde{U}\Delta^{-4}\tilde{U}^{\mathsf{T}}\,\tilde{U}\Delta U^{\mathsf{T}} \;=\; U\Delta^{-2}U^{\mathsf{T}} \;=\; \mathcal{M}^\dagger\,. \tag{131}$$

The (α,β)-block of $\mathcal{S}^\dagger$ is thus $\Psi_\alpha(\mathcal{R}^\dagger)^2\Psi_\beta^{\mathsf{T}}$, and summing over sample pairs gives

$$\Sigma \;=\; \sum_{\alpha\beta}\Psi_\alpha(\mathcal{R}^\dagger)^2\Psi_\beta^{\mathsf{T}} \;=\; \mathcal{B}^{\mathsf{T}}(\mathcal{R}^\dagger)^2\mathcal{B} \;=\; X^{\mathsf{T}}X\,, \qquad X := \mathcal{R}^\dagger\mathcal{B}\,, \tag{132}$$

where we used $\sum_\alpha \Psi_\alpha = \mathcal{B}^{\mathsf{T}}$ and the symmetry of $\mathcal{R}^\dagger$. The pseudo-inverse of the large $L^2P \times L^2P$ matrix has been eliminated, and the problem is reduced to identifying the single matrix $X \in \mathbb{R}^{(L+V)^2\times L^2}$.

Instead of computing $\mathcal{R}^\dagger$ directly, we use the following elementary criterion, which allows us to verify a candidate. If a matrix Y satisfies

$$\text{(a)} \quad \mathcal{R}\,Y = \mathcal{B}\,, \qquad\qquad \text{(b)} \quad \text{every column of } Y \text{ lies in } \mathrm{range}(\mathcal{R})\,, \tag{133}$$

then $Y = X$. Indeed, for symmetric $\mathcal{R}$ the product $\mathcal{R}^\dagger\mathcal{R}$ is the orthogonal projector onto $\mathrm{range}(\mathcal{R})$, so that $X = \mathcal{R}^\dagger\mathcal{B} = \mathcal{R}^\dagger\mathcal{R}\,Y = Y$ by (a) and (b).

Next, we determine the kernel of $\mathcal{R}$. One-hot encoding implies $X_\alpha \boldsymbol{u}_V = \boldsymbol{u}_L$, since each row of X_α contains exactly one nonzero entry. Hence the single, data-independent vector

$$w_0 := \begin{pmatrix} \boldsymbol{u}_L \\ -\boldsymbol{u}_V \end{pmatrix} \in \mathbb{R}^{L+V} \qquad \text{satisfies} \qquad \Phi_\alpha w_0 = \boldsymbol{u}_L - X_\alpha \boldsymbol{u}_V = 0 \quad \text{for every } \alpha\,. \tag{134}$$

This degeneracy makes $\mathcal{R}$ singular, which is why its pseudo-inverse enters X: for one-hot inputs the sum of the positional coordinates and the sum of the token coordinates of an embedded token coincide, so the two directions cannot be distinguished by any sample. From (134) we get $\Gamma_\alpha w_0 = \Phi_\alpha^\mathsf{T}(\Phi_\alpha w_0) = 0$ and hence, for any $v \in \mathbb{R}^{L+V}$, $\mathcal{R}\,(w_0 \otimes v) = \sum_\alpha (\Gamma_\alpha w_0) \otimes (\Gamma_\alpha v) = 0$ and likewise $\mathcal{R}\,(v \otimes w_0) = 0$. The subspace

$$\mathcal{N} := \mathrm{span}\big\{\, w_0 \otimes v,\; v \otimes w_0 \;:\; v \in \mathbb{R}^{L+V} \,\big\}\,, \qquad \dim \mathcal{N} = 2(L+V) - 1\,, \tag{135}$$

therefore obeys $\mathcal{N} \subseteq \ker \mathcal{R}$ for any training set; the dimension is $2(L+V)$ generators minus one for the single overlap $w_0 \otimes w_0$. Hence $\mathrm{rank}\,\mathcal{S} = \mathrm{rank}\,\mathcal{R} \le (L+V-1)^2$ for any training set, so that $\mathcal{S}$ is singular as soon as $L^2 P > (L+V-1)^2$. Since $\mathcal{R}$ is symmetric, $\mathrm{range}(\mathcal{R}) = (\ker \mathcal{R})^\perp \subseteq \mathcal{N}^\perp$. We assume that this inclusion is tight, which we refer to as the rank condition,

$$\ker \mathcal{R} = \mathcal{N}\,, \qquad \text{equivalently} \qquad \mathrm{rank}\,\mathcal{R} = (L{+}V{-}1)^2\,, \qquad \text{equivalently} \qquad \mathrm{range}(\mathcal{R}) = \mathcal{N}^\perp\,, \tag{136}$$

i.e. w_0 is the *only* degeneracy of the training set.

Now we can construct X. We look for a solution of $\mathcal{R}\,Y = \mathcal{B}$ of the form $Y = \phi \otimes \phi$ with $\phi \in \mathbb{R}^{(L+V)\times L}$, written in position and token blocks as $\phi = (\phi_\mathrm{p}; \phi_\mathrm{t})$ with $\phi_\mathrm{p} \in \mathbb{R}^{L\times L}$ and $\phi_\mathrm{t} \in \mathbb{R}^{V\times L}$. Two conditions on ϕ are sufficient, one for each part of the criterion (133).

Condition (i) is $\Phi_\alpha \phi = \mathrm{Id}_L$ for all α, i.e. ϕ is a right inverse of every embedding simultaneously. It implies $\Gamma_\alpha \phi = \Phi_\alpha^\mathsf{T}(\Phi_\alpha \phi) = \Phi_\alpha^\mathsf{T}$ and therefore, by the mixed-product rule and (130),

$$\mathcal{R}\,(\phi \otimes \phi) = \sum_\alpha (\Gamma_\alpha \phi) \otimes (\Gamma_\alpha \phi) = \sum_\alpha \Phi_\alpha^\mathsf{T} \otimes \Phi_\alpha^\mathsf{T} = \mathcal{B}\,, \tag{137}$$

which is part (a) of (133). Condition (ii) is $\phi^\mathsf{T} w_0 = 0$. The columns of $\phi \otimes \phi$ are the vectors $\phi_{\cdot a} \otimes \phi_{\cdot b}$, and for any v, $(w_0 \otimes v)^\mathsf{T}(\phi_{\cdot a} \otimes \phi_{\cdot b}) = (w_0^\mathsf{T}\phi_{\cdot a})(v^\mathsf{T}\phi_{\cdot b}) = 0$ and likewise $(v \otimes w_0)^\mathsf{T}(\phi_{\cdot a} \otimes \phi_{\cdot b}) = 0$. Every column of $\phi \otimes \phi$ therefore lies in $\mathcal{N}^\perp$, which equals $\mathrm{range}(\mathcal{R})$ by (136); this is part (b).

Both conditions are met by the following construction. Writing $\Phi_\alpha \phi = \phi_\mathrm{p} + X_\alpha \phi_\mathrm{t}$, condition (i) requires $X_\alpha \phi_\mathrm{t}$ to be independent of the data. Since row a of $X_\alpha \phi_\mathrm{t}$ is row $\sigma_\alpha(a)$ of ϕ_t, this is achieved by giving ϕ_t identical rows, $\phi_\mathrm{t} = \boldsymbol{u}_V m^\mathsf{T}$ with $m \in \mathbb{R}^L$, so that $X_\alpha \phi_\mathrm{t} = (X_\alpha \boldsymbol{u}_V) m^\mathsf{T} = \boldsymbol{u}_L m^\mathsf{T}$, and condition (i) then fixes $\phi_\mathrm{p} = \mathrm{Id}_L - \boldsymbol{u}_L m^\mathsf{T}$. The vector m is fixed by condition (ii): using $\phi_\mathrm{p}^\mathsf{T}\boldsymbol{u}_L = \boldsymbol{u}_L - Lm$ and $\phi_\mathrm{t}^\mathsf{T}\boldsymbol{u}_V = Vm$,

$$\phi^\mathsf{T} w_0 = \phi_\mathrm{p}^\mathsf{T}\boldsymbol{u}_L - \phi_\mathrm{t}^\mathsf{T}\boldsymbol{u}_V = \boldsymbol{u}_L - (L+V)\,m \;\overset{!}{=}\; 0 \qquad \Longrightarrow \qquad m = \frac{\boldsymbol{u}_L}{L+V}\,. \tag{138}$$

With $\Pi := \mathrm{Id}_L - \frac{1}{L+V}\boldsymbol{J}_L$ this gives the explicit matrix

$$\phi = \begin{pmatrix} \Pi \\ \frac{1}{L+V}\,\boldsymbol{u}_V \boldsymbol{u}_L^\mathsf{T} \end{pmatrix}\,, \qquad \text{and hence} \qquad X = \mathcal{R}^\dagger\mathcal{B} = \phi \otimes \phi\,, \tag{139}$$

by the criterion (133). Interpreted directly, ϕ distributes each position uniformly over the token coordinates and removes the corresponding mean from the positional coordinates; among all common right inverses of the embeddings it is the one orthogonal to the degenerate direction w_0, and $\|w_0\|^2 = L+V$ is the origin of the factor $(L+V)^{-1}$.

Inserting (139) into (132) and using $(A \otimes B)^{\mathsf{T}}(C \otimes D) = (A^{\mathsf{T}}C) \otimes (B^{\mathsf{T}}D)$,

$$\Sigma = X^{\mathsf{T}}X = (\phi \otimes \phi)^{\mathsf{T}}(\phi \otimes \phi) = (\phi^{\mathsf{T}}\phi) \otimes (\phi^{\mathsf{T}}\phi)\,, \tag{140}$$

so that the Kronecker structure of the answer is automatic and only the $L \times L$ matrix $\phi^{\mathsf{T}}\phi$ remains to be evaluated. Abbreviating $s := (L+V)^{-1}$ and using $\boldsymbol{J}_L^2 = L\,\boldsymbol{J}_L$ together with $\boldsymbol{u}_V^{\mathsf{T}}\boldsymbol{u}_V = V$,

$$\phi^{\mathsf{T}}\phi = \Pi^2 + s^2 V \boldsymbol{J}_L = \mathrm{Id}_L - s\big[\,2 - sL - sV\,\big]\boldsymbol{J}_L = \mathrm{Id}_L - s\big[\,2 - s(L+V)\,\big]\boldsymbol{J}_L = \mathrm{Id}_L - s\,\boldsymbol{J}_L = \Pi\,, \tag{141}$$

where the last step used $s(L+V) = 1$. Combining (140) and (141),

$$\Sigma_{(ab),(cd)} \;=\; \delta_{ac}\,\delta_{bd} \;-\; \frac{\delta_{ac}}{L+V} \;-\; \frac{\delta_{bd}}{L+V} \;+\; \frac{1}{(L+V)^2} \;=\; \Pi_{ac}\,\Pi_{bd}\,, \tag{142}$$

with $\Pi = \mathrm{Id}_L - \frac{1}{L+V}\boldsymbol{J}_L$. In particular Σ depends on the training set only through the rank condition (136), and is otherwise independent of P and of the particular sequences drawn.

A sufficient condition for the rank condition (136). Because $\mathcal{R} = \Psi^{\mathsf{T}}\Psi$ we have $\ker\mathcal{R} = \ker\Psi$, and identifying a vector in $\mathbb{R}^{(L+V)^2}$ with a matrix γ of size $(L+V)\times(L+V)$, the blocks of $\Psi\gamma$ are $\Phi_\alpha\gamma\Phi_\alpha^{\mathsf{T}}$. Hence

$$\ker\mathcal{R} = \big\{\,\gamma \;:\; \Phi_\alpha\gamma\Phi_\alpha^{\mathsf{T}} = 0 \;\text{ for all } \alpha\,\big\}\,, \tag{143}$$

an explicit linear condition with no pseudo-inverse involved. Writing γ in position and token blocks as $\gamma = \left(\begin{smallmatrix} A & B \\ C & D \end{smallmatrix}\right)$ and using $\Phi_\alpha = [\mathrm{Id}_L \mid X_\alpha]$,

$$\big(\Phi_\alpha\gamma\Phi_\alpha^{\mathsf{T}}\big)_{ac} = A_{ac} + B_{a\,\sigma_\alpha(c)} + C_{\sigma_\alpha(a)\,c} + D_{\sigma_\alpha(a)\,\sigma_\alpha(c)}\,. \tag{144}$$

Suppose the training set is such that for every pair of positions $a \neq c$ and every pair of tokens (j,k) there exists a sample α with $\sigma_\alpha(a) = j$ and $\sigma_\alpha(c) = k$. Fixing $a \neq c$, (144) then gives $A_{ac} + B_{ak} + C_{jc} + D_{jk} = 0$ for all j,k. Taking the difference of this relation at j and at j' shows that $D_{jk} - D_{j'k}$ is independent of k, whence $D_{jk} = d_k + \varepsilon_j$; the same difference then forces $C_{jc} - C_{j'c}$ to be independent of c, whence $C_{jc} = r_c - \varepsilon_j$. Taking instead the difference at k and k' gives, in the same way, $B_{ak} = \mu_a - d_k$. Substituting these three forms back yields $A_{ac} = -\mu_a - r_c$, first for $a \neq c$ and then, from the $a = c$ relations, on the diagonal as well. The resulting general solution,

$$A = -\mu\boldsymbol{u}_L^{\mathsf{T}} - \boldsymbol{u}_L r^{\mathsf{T}}, \quad B = \mu\boldsymbol{u}_V^{\mathsf{T}} - \boldsymbol{u}_L d^{\mathsf{T}}, \quad C = -\varepsilon\boldsymbol{u}_L^{\mathsf{T}} + \boldsymbol{u}_V r^{\mathsf{T}}, \quad D = \varepsilon\boldsymbol{u}_V^{\mathsf{T}} + \boldsymbol{u}_V d^{\mathsf{T}}, \tag{145}$$

is exactly $\gamma = w_0 v^{\mathsf{T}} + w\,w_0^{\mathsf{T}}$ with $v = (-r,-d)$ and $w = (-\mu,-\varepsilon)$, i.e. $\ker\mathcal{R} = \mathcal{N}$, which is (136).

This condition on the training set is sufficient but not necessary, and in practice the rank may be attained for considerably smaller P than it requires; rank counting alone already shows that $P \geq \lceil (L+V-1)^2/L^2 \rceil$ is necessary, since each sample contributes at most L^2 to the rank. This bound is at most 4 for $V \leq L+1$ and equals 2 for $L = 25$, $V = 5$. For the training sets used in our experiments that satisfies this bound, we verified numerically that (142) applies.

Determining the projections onto the attention patterns

Using the expression for (142), we compute the bilinear form $\frac{2g_W}{L^2}\mathcal{S}_{\mathrm{NNGP}}(G,G') = \sum_{abcd} G_{ab}\,\Sigma_{(ab),(cd)}\,G'_{cd}$ explicitly

$$\begin{aligned}
\frac{2g_W}{L^2}\mathcal{S}_{\mathrm{NNGP}}(G,G') &= \sum_{abcd} G_{ab}\Big[\delta_{ac}\delta_{bd} - \frac{\delta_{ac}}{L+V} - \frac{\delta_{bd}}{L+V} + \frac{1}{(L+V)^2}\Big]G'_{cd} \\
&= \sum_{ab} G_{ab}G'_{ab} - \frac{1}{L+V}\sum_{abd} G_{ab}\,G'_{ad} - \frac{1}{L+V}\sum_{abc} G_{ab}\,G'_{cb} \qquad (146)\\
&\quad + \frac{1}{(L+V)^2}\Big(\sum_{ab} G_{ab}\Big)\Big(\sum_{ab} G'_{ab}\Big) \\
&= \sum_{ab} G_{ab}G'_{ab} - \frac{\sum_a r_a r'_a + \sum_b c_b c'_b}{L+V} + \frac{ss'}{(L+V)^2}\,, \qquad (147)
\end{aligned}$$

where $r_a = \sum_b G_{ab}$ and $r'_a = \sum_b G'_{ab}$ are the row sums, $c_b = \sum_a G_{ab}$ and $c'_b = \sum_a G'_{ab}$ are the column sums, and $s = \sum_{ab} G_{ab}$ and $s' = \sum_{ab} G'_{ab}$ are the total sums.

We use this expression to compute the projections onto the basis $\{\mathbf{1}_L,\ G^*\}$, yielding

$$\frac{1}{(L+V)^2}\begin{pmatrix} L^2V^2 & LV^2 \\ LV^2 & L\big[(L+V)^2-(L+2V)\big] \end{pmatrix}. \tag{148}$$

We project onto the attention patterns normalized with respect to the Frobenius norm $\frac{1}{L}\,\mathbf{1}_L$, $\frac{1}{\sqrt{L}}\,G^*$, so making the ansatz $G = \hat{c}_1 \frac{1}{L}\,\mathbf{1}_L + \hat{c}_G \frac{1}{\sqrt{L}}\,G^*$, we overall get

$$\frac{1}{(L+V)^2}\Big[V^2\,\hat{c}_1^2 \;+\; 2\frac{V^2}{\sqrt{L}}\,\hat{c}_1\,\hat{c}_G \;+\; \big((L+V)^2 - L - 2V\big)\,\hat{c}_G^2\Big]. \tag{149}$$

Multiplying the above result by $L^2/(2g_W)$ and setting $g_W = \bar{g}_W/\chi$ yields Eq. (12) of the main text.

E Projecting the Data Term onto the Order Parameters

We now look at the data term

$$\boldsymbol{y}^\mathsf{T}\Big[g_O \boldsymbol{G}\boldsymbol{C}^{(xx)}\boldsymbol{G}^\top + \sigma^2\boldsymbol{I}\Big]^{-1}\boldsymbol{y} \tag{150}$$

and make the ansatz $\boldsymbol{G} = c_1\mathbf{1}_L + c_G\boldsymbol{G}^*$. In the end, we transform to the normalized operators basis with coefficients $(\hat{c}_1, \hat{c}_G)$. Due to the contraction with the target values $\boldsymbol{y}$, we will use its Fourier modes as a base. For the copy task, the targets are the inputs shifted by one, so we choose

$$\big[\boldsymbol{v}_{0,0}\big]^a_\alpha = \frac{1}{\sqrt{PL}}\frac{1}{L}\sum_b\sum_j x^b_{j,\alpha} = \frac{1}{\sqrt{PL}}, \tag{151}$$

$$\big[\boldsymbol{v}_{0,k\geq 1}\big]^a_\alpha = \frac{1}{\sqrt{PL^2}}\sum_b\sum_j e^{i\frac{2\pi}{V}kj}x^b_{j,\alpha}, \tag{152}$$

$$\big[\boldsymbol{v}_{1,k}\big]^a_\alpha = \frac{1}{\sqrt{PL(1-L^{-1})}}\sum_j e^{i\frac{2\pi}{V}kj}\left(x^{a-1}_{j,\alpha} - \frac{1}{L}\sum_b x^b_{j,\alpha}\right). \tag{153}$$

Note that $\boldsymbol{v}_{1,0} = 0$ and that the entries of $\big[\boldsymbol{v}_{0,k}\big]^a$ are independent of a. The different normalization constants for $\big[\boldsymbol{v}_{0,0}\big]^a_\alpha$ and $\big[\boldsymbol{v}_{0,k\geq 1}\big]^a_\alpha$ result from the application of the equivalent kernel approximation. For $k \geq 1$, we will use the Fourier orthogonality relation $\sum_j e^{i\frac{2\pi}{V}kj} = 0$ and the decomposition identity

$$\sum_j e^{i\frac{2\pi}{V}kj}x^{a-1}_{j,\alpha} = \sqrt{PL(1-L^{-1})}\,[\boldsymbol{v}_{1,k}]^a_\alpha + \sqrt{P}\,[\boldsymbol{v}_{0,k}]^a_\alpha. \tag{154}$$

Throughout, we will use the equivalent kernel approximation replacing $\sum_\alpha \circ \approx P\boldsymbol{E}_x[\circ]$ and the one-hot expectations:

$$\boldsymbol{E}[x^a_{i,\beta}x^a_{j,\beta}] = \frac{\delta_{ij}}{V}, \qquad \boldsymbol{E}[x^a_{i,\beta}x^b_{j,\beta}] = \frac{1}{V^2}\ \ (a\neq b), \qquad \boldsymbol{E}[x^a_{j,\beta}] = \frac{1}{V}. \tag{155}$$

Decomposition of labels

We start by computing the spanning coefficients of y

$$\boldsymbol{y}_i^\mathsf{T}\boldsymbol{v}_{0,0} = \frac{1}{\sqrt{PL}}\sum_{\alpha,a} x_{i\alpha}^{a-1} \approx \sqrt{PL}\frac{1}{V}, \tag{156}$$

$$\boldsymbol{y}_i^\mathsf{T}\boldsymbol{v}_{0,k\geq 1} = \frac{1}{\sqrt{PL^2}}\sum_{\alpha,a} x_{i,\alpha}^{a-1}\sum_j e^{i\frac{2\pi}{V}kj}\sum_b x_{j,\alpha}^b \tag{157}$$

$$= \frac{1}{\sqrt{P}}\sum_{\alpha,a}\sum_j e^{i\frac{2\pi}{V}kj}\left(\frac{1}{L}\sum_b \delta^{a-1,b}x_{i,\alpha}^{a-1}x_{j,\alpha}^{a-1} + \left(1-\delta^{a-1,b}\right)x_{i,\alpha}^{a-1}x_{j,\alpha}^b\right) \tag{158}$$

$$\approx \frac{1}{\sqrt{P}}P\sum_a\sum_j e^{i\frac{2\pi}{V}kj}\frac{1}{L}\left(\frac{\delta_{ij}}{V} + (L-1)\frac{1}{V^2}\right) = \sqrt{P}e^{i\frac{2\pi}{V}ki}\frac{1}{V}, \tag{159}$$

$$\boldsymbol{y}_i^\mathsf{T}\boldsymbol{v}_{1,k\geq 1} = \frac{1}{\sqrt{PL(1-L^{-1})}}\sum_{\alpha,a}\sum_j e^{i\frac{2\pi}{V}kj}x_{i,\alpha}^{a-1}\left(x_{j,\alpha}^{a-1} - \frac{1}{L}\sum_b x_{j,\alpha}^b\right) \tag{160}$$

$$= \frac{1}{\sqrt{PL(1-L^{-1})}}\sum_{\alpha,a}\sum_j e^{i\frac{2\pi}{V}kj}\Bigg(x_{i,\alpha}^{a-1}x_{j,\alpha}^{a-1} \tag{161}$$

$$- \frac{1}{L}\sum_b\left(\delta^{a-1,b}x_{i,\alpha}^{a-1}x_{j,\alpha}^{a-1} + \left(1-\delta^{a-1,b}\right)x_{i,\alpha}^{a-1}x_{j,\alpha}^b\right)\Bigg) \tag{162}$$

$$\approx \frac{1}{\sqrt{PL(1-L^{-1})}}P\sum_a\sum_j e^{i\frac{2\pi}{V}kj}\left(\frac{\delta_{ij}}{V} - \frac{1}{L}\left(\frac{\delta_{ij}}{V} + (L-1)\frac{1}{V^2}\right)\right) \tag{163}$$

$$= \sqrt{\frac{PL}{1-L^{-1}}}e^{i\frac{2\pi}{V}ki}\frac{1}{V}\left(1-\frac{1}{L}\right). \tag{164}$$

Decomposition of operator-matrix-contraction

We look at the term $\boldsymbol{G}\boldsymbol{C}^{(xx)}\boldsymbol{G}^\top$ and make the ansatz $\boldsymbol{G} = c_1\mathbf{1}_L + c_G\boldsymbol{G}_*$. Inserting this ansatz, we have

$$\sum_{bd} G^{ab}G^{cd}C_{bd,\alpha\beta}^{(xx)} = c_1^2\sum_{bd}C_{bd} + c_1c_G\sum_b C_{b,c-1} + c_Gc_1\sum_d C_{a-1,d} + c_G^2C_{a-1,c-1}, \tag{165}$$

corresponding to the four operator pairs $(\mathbf{1}_L,\mathbf{1}_L)$, $(\mathbf{1}_L,\boldsymbol{G}^*)$, $(\boldsymbol{G}^*,\mathbf{1}_L)$, $(\boldsymbol{G}^*,\boldsymbol{G}^*)$. We compute the action of each on the Fourier basis vectors $\boldsymbol{v}_{0,0}$, $\boldsymbol{v}_{0,k\geq 1}$, $\boldsymbol{v}_{1,k\geq 1}$ from the right, i.e. we evaluate $\sum_{\beta,c}\sum_{bd}M_L^{ab}C_{bd,\alpha\beta}^{(xx)}M_R^{cd}[\boldsymbol{v}]_\beta^c$.

We first look at the term corresponding to $(\mathbf{1}_L,\mathbf{1}_L)$, so the application of $\sum_{bd}C_{bd,\alpha\beta}$ to the Fourier basis.

For $\boldsymbol{v}_{0,0}$, we have

$$\sum_{\beta,c}\sum_{bd}C_{bd,\alpha\beta}^{(xx)}\left[\boldsymbol{v}_{0,0}\right]_\beta^c = \sum_{\beta,c}\sum_{bd}\left(\sum_i x_{i,\alpha}^b x_{i,\beta}^d + \delta^{bd}\right)\frac{1}{\sqrt{PL}} \tag{166}$$

$$\approx \frac{P}{\sqrt{PL}}\sum_c\sum_{bd}\left(\sum_i x_{i,\alpha}^b\frac{1}{V} + \delta^{bd}\right) \tag{167}$$

$$= \frac{PL^2}{\sqrt{PL}}\left(\frac{L}{V}+1\right) = PL^2\left(\frac{L}{V}+1\right)\left[\boldsymbol{v}_{0,0}\right]_\alpha^a. \tag{168}$$

For $\boldsymbol{v}_{0,k\geq 1}$, we get

$$\sum_{\beta,c}\sum_{bd}C_{bd,\alpha\beta}^{(xx)}\left[\boldsymbol{v}_{0,k\geq 1}\right]_\beta^c \tag{169}$$

$$= \sum_{\beta,c}\sum_{bd}\Big(\sum_i x^b_{i,\alpha}x^d_{i,\beta}+\delta^{bd}\Big)\frac{1}{\sqrt{PL^2}}\sum_a\sum_j e^{i\frac{2\pi}{V}kj}x^a_{j,\beta} \tag{170}$$

$$= \frac{L}{\sqrt{PL^2}}\sum_\beta\Big(\sum_{bd}\sum_i x^b_{i,\alpha}x^d_{i,\beta}\sum_a\sum_j e^{i\frac{2\pi}{V}kj}x^a_{j,\beta}+L\sum_a\sum_j e^{i\frac{2\pi}{V}kj}x^a_{j,\beta}\Big) \tag{171}$$

$$= \frac{1}{\sqrt{P}}\sum_\beta\Big(\sum_{bda}\sum_{ij} e^{i\frac{2\pi}{V}kj}x^b_{i,\alpha}\big(\delta^{ad}x^a_{i,\beta}x^a_{j,\beta}+\big(1-\delta^{ad}\big)x^d_{i,\beta}x^a_{j,\beta}\big)+L\sum_a\sum_j e^{i\frac{2\pi}{V}kj}x^a_{j,\beta}\Big) \tag{172}$$

$$\approx \frac{1}{\sqrt{P}}P\Big(\sum_b\sum_{ij}e^{i\frac{2\pi}{V}kj}x^b_{i,\alpha}L\Big(\frac{\delta_{ij}}{V}+\frac{L-1}{V^2}\Big)+L\sum_a\sum_j e^{i\frac{2\pi}{V}kj}\frac{1}{V}\Big) \tag{173}$$

$$= \sqrt{P}\Big(\sum_b\sum_j e^{i\frac{2\pi}{V}kj}x^b_{j,\alpha}L\frac{1}{V}\Big)=\frac{PL^2}{V}\big[\boldsymbol{v}_{0,k\geq 1}\big]^a_\alpha, \tag{174}$$

where the $1/V^2$ and $1/V$ terms vanish by Fourier orthogonality for $k\geq 1$.

Finally, for $\boldsymbol{v}_{1,k\geq 1}$ it is

$$\sum_{\beta,c}\sum_{bd}C^{(xx)}_{bd,\alpha\beta}\big[\boldsymbol{v}_{1,k\geq 1}\big]^c_\beta \tag{175}$$

$$= \sum_{\beta,c}\sum_{bd}\Big(\sum_i x^b_{i,\alpha}x^d_{i,\beta}+\delta^{bd}\Big)\frac{1}{\sqrt{PL(1-L^{-1})}}\sum_j e^{i\frac{2\pi}{V}kj}\Big(x^{c-1}_{j,\beta}-\frac{1}{L}\sum_a x^a_{j,\beta}\Big) \tag{176}$$

$$= \frac{1}{\sqrt{PL(1-L^{-1})}}\sum_{\beta,c}\sum_j e^{i\frac{2\pi}{V}kj}\Big(\sum_i\Big(\sum_{bd}x^b_{i,\alpha}x^d_{i,\beta}x^{c-1}_{j,\beta}-\frac{1}{L}\sum_{abd}x^b_{i,\alpha}x^d_{i,\beta}x^a_{j,\beta}\Big)$$
$$+L\Big(x^{c-1}_{j,\beta}-\frac{1}{L}\sum_a x^a_{j,\beta}\Big)\Big) \tag{177}$$

$$\approx \frac{P}{\sqrt{PL(1-L^{-1})}}\sum_{c}\sum_j e^{i\frac{2\pi}{V}kj}\Big(\sum_i\Big(\sum_{bd}x^b_{i,\alpha}\Big(\delta^{d,c-1}\frac{\delta_{ij}}{V}+\Big(1-\delta^{d,c-1}\Big)\frac{1}{V^2}\Big)$$
$$-\frac{1}{L}\sum_{abd}x^b_{i,\alpha}\Big(\delta^{ad}\frac{\delta_{ij}}{V}+\big(1-\delta^{ad}\big)\frac{1}{V^2}\Big)\Big)+L\Big(\frac{1}{V}-\frac{1}{L}\sum_a\frac{1}{V}\Big)\Big) \tag{178}$$

$$= \frac{P}{\sqrt{PL(1-L^{-1})}}\sum_c\sum_{ij}e^{i\frac{2\pi}{V}kj}$$
$$\Big(\sum_b x^b_{i,\alpha}\Big(\frac{\delta_{ij}}{V}+(L-1)\frac{1}{V^2}\Big)-\sum_b x^b_{i,\alpha}\Big(\frac{\delta_{ij}}{V}+(L-1)\frac{1}{V^2}\Big)\Big) \tag{179}$$

$$= 0. \tag{180}$$

Next, we consider the term from $(\mathbf{1}_L, \boldsymbol{G}^*)$ corresponding to $\sum_b C_{b,c-1,\alpha\beta}$.

For $\boldsymbol{v}_{0,0}$, we get

$$\sum_{\beta,c}\sum_{bd}C^{(xx)}_{bd,\alpha\beta}\delta^{c-1,d}\big[\boldsymbol{v}_{0,0}\big]^c_\beta=\sum_{\beta,c}\sum_b\Big(\sum_i x^b_{i,\alpha}x^{c-1}_{i,\beta}+\delta^{b,c-1}\Big)\frac{1}{\sqrt{PL}} \tag{181}$$

$$\approx \frac{P}{\sqrt{PL}}\sum_{bc}\Big(\sum_i x^b_{i,\alpha}\frac{1}{V}+\delta^{b,c-1}\Big) \tag{182}$$

$$= \frac{PL}{\sqrt{PL}}\Big(\frac{L}{V}+1\Big)=PL\Big(\frac{L}{V}+1\Big)\big[\boldsymbol{v}_{0,0}\big]^a_\alpha. \tag{183}$$

For $\boldsymbol{v}_{0,k\geq 1}$, it is

$$\sum_{\beta,c}\sum_{bd} C^{(xx)}_{bd,\alpha\beta}\delta^{c-1,d}\big[\boldsymbol{v}_{0,k\geq 1}\big]^c_\beta \tag{184}$$

$$= \sum_{\beta,c}\sum_{b}\Big(\sum_i x^b_{i,\alpha}x^{c-1}_{i,\beta} + \delta^{b,c-1}\Big)\frac{1}{\sqrt{P}}\frac{1}{L}\sum_a\sum_j e^{i\frac{2\pi}{V}kj}x^a_{j,\beta} \tag{185}$$

$$= \frac{1}{\sqrt{P}}\frac{1}{L}\sum_\beta\Big(\sum_{bd}\sum_i x^b_{i,\alpha}x^d_{i,\beta}\sum_a\sum_j e^{i\frac{2\pi}{V}kj}x^a_{j,\beta} + L\sum_a\sum_j e^{i\frac{2\pi}{V}kj}x^a_{j,\beta}\Big) \tag{186}$$

$$= \frac{1}{\sqrt{P}}\frac{1}{L}\sum_\beta\Big(\sum_{bda}\sum_{ij} e^{i\frac{2\pi}{V}kj}x^b_{i,\alpha}\Big(\delta^{ad}x^a_{i,\beta}x^a_{j,\beta} + \big(1-\delta^{ad}\big)x^d_{i,\beta}x^a_{j,\beta}\Big) + L\sum_a\sum_j e^{i\frac{2\pi}{V}kj}x^a_{j,\beta}\Big) \tag{187}$$

$$\approx \frac{1}{\sqrt{P}}\frac{1}{L}P\Big(\sum_b\sum_{ij} e^{i\frac{2\pi}{V}kj}x^b_{i,\alpha}L\Big(\frac{\delta_{ij}}{V}+\frac{L-1}{V^2}\Big) + L\sum_a\sum_j e^{i\frac{2\pi}{V}kj}\frac{1}{V}\Big) \tag{188}$$

$$= \frac{1}{\sqrt{P}}\frac{1}{L}LP\Big(\sum_b\sum_j e^{i\frac{2\pi}{V}kj}x^b_{j,\alpha}\frac{1}{V}\Big) = \frac{PL}{V}\big[\boldsymbol{v}_{0,k\geq 1}\big]^a_\alpha. \tag{189}$$

Applied to $\boldsymbol{v}_{1,k\geq 1}$, we have

$$\sum_{\beta,c}\sum_{bd} C^{(xx)}_{bd,\alpha\beta}\delta^{c-1,d}\big[\boldsymbol{v}_{1,k\geq 1}\big]^c_\beta \tag{190}$$

$$= \sum_{\beta,c}\sum_{b}\Big(\sum_i x^b_{i,\alpha}x^{c-1}_{i,\beta} + \delta^{b,c-1}\Big)\frac{1}{\sqrt{PL(1-L^{-1})}}\sum_j e^{i\frac{2\pi}{V}kj}\Big(x^{c-1}_{j,\beta} - \frac{1}{L}\sum_a x^a_{j,\beta}\Big) \tag{191}$$

$$= \frac{1}{\sqrt{PL(1-L^{-1})}}\sum_{\beta,c}\sum_j e^{i\frac{2\pi}{V}kj}\Big(\sum_i\Big(\sum_b x^b_{i,\alpha}x^c_{i,\beta}x^c_{j,\beta} - \frac{1}{L}\sum_{ab}x^b_{i,\alpha}x^c_{i,\beta}x^a_{j,\beta}\Big) + \Big(x^c_{j,\beta} - \frac{1}{L}\sum_a x^a_{j,\beta}\Big)\Big), \tag{192}$$

where we relabelled $c-1\to c$. Applying the equivalent kernel:

$$\approx \frac{P}{\sqrt{PL(1-L^{-1})}}\sum_c\sum_j e^{i\frac{2\pi}{V}kj}\Big(\sum_i\Big(\sum_b x^b_{i,\alpha}\frac{\delta_{ij}}{V} - \frac{1}{L}\sum_{ab}x^b_{i,\alpha}\Big(\delta^{ac}\frac{\delta_{ij}}{V} + \Big(1-\delta^{ac}\Big)\frac{1}{V^2}\Big)\Big) + \Big(\frac{1}{V} - \frac{1}{L}\sum_a\frac{1}{V}\Big)\Big) \tag{193}$$

$$= \frac{P}{\sqrt{PL(1-L^{-1})}}\sum_c\sum_{ij} e^{i\frac{2\pi}{V}kj}\Big(\sum_b x^b_{i,\alpha}\frac{\delta_{ij}}{V} - \frac{1}{L}\sum_b x^b_{i,\alpha}\Big(\frac{\delta_{ij}}{V} + (L-1)\frac{1}{V^2}\Big)\Big) \tag{194}$$

$$= \frac{P}{\sqrt{PL(1-L^{-1})}}\frac{L}{V}\Big(1-\frac{1}{L}\Big)\sum_j e^{i\frac{2\pi}{V}kj}\sum_b x^b_{j,\alpha} \tag{195}$$

$$= \frac{PL}{V}\sqrt{L\Big(1-\frac{1}{L}\Big)}\Big(\frac{1}{\sqrt{P}}\frac{1}{L}\sum_b\sum_j e^{i\frac{2\pi}{V}kj}x^b_{j,\alpha}\Big) = \frac{PL\sqrt{L-1}}{V}\big[\boldsymbol{v}_{0,k\geq 1}\big]^a_\alpha. \tag{196}$$

For the terms switched in order $(\boldsymbol{G}^*, \boldsymbol{1}_L)$, we have $\delta^{a-1,b}\sum_d C_{bd,\alpha\beta}$.

For $\boldsymbol{v}_{0,0}$, this yields

$$\sum_{\beta,c}\sum_{bd}\delta^{a-1,b}C^{(xx)}_{bd,\alpha\beta}\big[\boldsymbol{v}_{0,0}\big]^c_\beta = \sum_{\beta,c}\sum_d\Big(\sum_i x^{a-1}_{i,\alpha}x^d_{i,\beta} + \delta^{a-1,d}\Big)\frac{1}{\sqrt{PL}} \tag{197}$$

$$\approx \frac{P}{\sqrt{PL}} \sum_c \sum_d \Big(\sum_i x_{i,\alpha}^{a-1} \frac{1}{V} + \delta^{a-1,d} \Big) \tag{198}$$

$$= \frac{PL}{\sqrt{PL}} \Big(\frac{L}{V} + 1 \Big) = PL \Big(\frac{L}{V} + 1 \Big) \big[\boldsymbol{v}_{0,0} \big]_\alpha^a. \tag{199}$$

For $\boldsymbol{v}_{0,k\geq 1}$, it is

$$\sum_{\beta,c} \sum_{bd} \delta^{a-1,b} C_{bd,\alpha\beta}^{(xx)} \big[\boldsymbol{v}_{0,k\geq 1} \big]_\beta^c \tag{200}$$

$$= \sum_{\beta,c} \sum_d \Big(\sum_i x_{i,\alpha}^{a-1} x_{i,\beta}^d + \delta^{a-1,d} \Big) \frac{1}{\sqrt{P}} \frac{1}{L} \sum_{a'} \sum_j e^{i\frac{2\pi}{V} kj} x_{j,\beta}^{a'} \tag{201}$$

$$= \frac{1}{\sqrt{P}} \sum_\beta \Big(\sum_d \sum_i x_{i,\alpha}^{a-1} x_{i,\beta}^d \sum_{a'} \sum_j e^{i\frac{2\pi}{V} kj} x_{j,\beta}^{a'} + \sum_{a'} \sum_j e^{i\frac{2\pi}{V} kj} x_{j,\beta}^{a'} \Big) \tag{202}$$

$$= \frac{1}{\sqrt{P}} \sum_\beta \Big(\sum_{da'} \sum_{ij} e^{i\frac{2\pi}{V} kj} x_{i,\alpha}^{a-1} \big(\delta^{da'} x_{i,\beta}^{a'} x_{j,\beta}^{a'} + \big(1 - \delta^{da'} \big) x_{i,\beta}^d x_{j,\beta}^{a'} \big) + \sum_{a'} \sum_j e^{i\frac{2\pi}{V} kj} x_{j,\beta}^{a'} \Big) \tag{203}$$

$$\approx \frac{P}{\sqrt{P}} \Big(\sum_{da'} \sum_{ij} e^{i\frac{2\pi}{V} kj} x_{i,\alpha}^{a-1} \Big(\delta^{da'} \frac{\delta_{ij}}{V} + \Big(1 - \delta^{da'} \Big) \frac{1}{V^2} \Big) + \sum_{a'} \sum_j e^{i\frac{2\pi}{V} kj} \frac{1}{V} \Big) \tag{204}$$

$$= \sqrt{P} \frac{L}{V} \sum_j e^{i\frac{2\pi}{V} kj} x_{j,\alpha}^{a-1}. \tag{205}$$

Now we decompose using (154), inserting $x_{j,\alpha}^{a-1} = (x_{j,\alpha}^{a-1} - \frac{1}{L} \sum_{a'} x_{j,\alpha}^{a'}) + \frac{1}{L} \sum_{a'} x_{j,\alpha}^{a'}$:

$$= \frac{PL}{V} \sqrt{L(1-L^{-1})} \big[\boldsymbol{v}_{1,k\geq 1} \big]_\alpha^a + \frac{PL}{V} \big[\boldsymbol{v}_{0,k\geq 1} \big]_\alpha^a. \tag{206}$$

For $\boldsymbol{v}_{1,k\geq 1}$, we have

$$\sum_{\beta,c} \sum_{bd} \delta^{a-1,b} C_{bd,\alpha\beta}^{(xx)} \big[\boldsymbol{v}_{1,k\geq 1} \big]_\beta^c \tag{207}$$

$$= \sum_{\beta,c} \sum_d \Big(\sum_i x_{i,\alpha}^{a-1} x_{i,\beta}^d + \delta^{a-1,d} \Big) \frac{1}{\sqrt{PL(1-L^{-1})}} \sum_j e^{i\frac{2\pi}{V} kj} \Big(x_{j,\beta}^{c-1} - \frac{1}{L} \sum_{a'} x_{j,\beta}^{a'} \Big) \tag{208}$$

$$= \frac{1}{\sqrt{PL(1-L^{-1})}} \sum_{\beta,c} \sum_j e^{i\frac{2\pi}{V} kj} \Big(\sum_i \Big(\sum_d x_{i,\alpha}^{a-1} x_{i,\beta}^d x_{j,\beta}^{c-1} - \frac{1}{L} \sum_{da'} x_{i,\alpha}^{a-1} x_{i,\beta}^d x_{j,\beta}^{a'} \Big) + \Big(x_{j,\beta}^{c-1} - \frac{1}{L} \sum_{a'} x_{j,\beta}^{a'} \Big) \Big) \tag{209}$$

$$\approx \frac{P}{\sqrt{PL(1-L^{-1})}} \sum_c \sum_j e^{i\frac{2\pi}{V} kj} \Big(\sum_i \Big(x_{i,\alpha}^{a-1} \Big(\frac{\delta_{ij}}{V} + \Big(L - 1 \Big) \frac{1}{V^2} \Big) - \frac{1}{L} \sum_{a'} x_{i,\alpha}^{a-1} \Big(\frac{\delta_{ij}}{V} + \Big(L - 1 \Big) \frac{1}{V^2} \Big) \Big) + \Big(\frac{1}{V} - \frac{1}{L} \sum_{a'} \frac{1}{V} \Big) \Big) \tag{210}$$

$$= 0, \tag{211}$$

since the two terms in the parentheses cancel identically after summing over c.

Finally, for $(\boldsymbol{G}^*, \boldsymbol{G}^*)$ we determine how $C_{a-1,c-1,\alpha\beta}$ acts onto the Fourier basis.

For $\boldsymbol{v}_{0,0}$, we get

$$\sum_{\beta,c} \sum_{bd} \delta^{a-1,b} C_{bd,\alpha\beta}^{(xx)} \delta^{c-1,d} \big[\boldsymbol{v}_{0,0} \big]_\beta^c = \sum_{\beta,c} \Big(\sum_i x_{i,\alpha}^{a-1} x_{i,\beta}^{c-1} + \delta^{a-1,c-1} \Big) \frac{1}{\sqrt{PL}} \tag{212}$$

$$\approx \frac{P}{\sqrt{PL}} \sum_{c} \left(\sum_{i} x_{i,\alpha}^{a-1} \frac{1}{V} + \delta^{a-1,c-1} \right) \tag{213}$$

$$= \frac{P}{\sqrt{PL}} \Big(\frac{L}{V} + 1 \Big) = P \Big(\frac{L}{V} + 1 \Big) \big[\boldsymbol{v}_{0,0} \big]_{\alpha}^{a}. \tag{214}$$

For $\boldsymbol{v}_{0,k\geq 1}$, we write

$$\sum_{\beta,c} \sum_{bd} \delta^{a-1,b} C_{bd,\alpha\beta}^{(xx)} \delta^{c-1,d} \big[\boldsymbol{v}_{0,k\geq 1} \big]_{\beta}^{c} \tag{215}$$

$$= \sum_{\beta,c} \left(\sum_{i} x_{i,\alpha}^{a-1} x_{i,\beta}^{c-1} + \delta^{a-1,c-1} \right) \frac{1}{\sqrt{PL^2}} \sum_{a'} \sum_{j} e^{i\frac{2\pi}{V} kj} x_{j,\beta}^{a'} \tag{216}$$

$$= \frac{1}{\sqrt{PL^2}} \sum_{\beta} \left(\sum_{c} \sum_{i} x_{i,\alpha}^{a-1} x_{i,\beta}^{c-1} \sum_{a'} \sum_{j} e^{i\frac{2\pi}{V} kj} x_{j,\beta}^{a'} + \sum_{a'} \sum_{j} e^{i\frac{2\pi}{V} kj} x_{j,\beta}^{a'} \right) \tag{217}$$

$$= \frac{1}{\sqrt{PL^2}} \sum_{\beta} \Bigg(\sum_{ca'} \sum_{ij} e^{i\frac{2\pi}{V} kj} x_{i,\alpha}^{a-1} \big(\delta^{c-1,a'} x_{i,\beta}^{a'} x_{j,\beta}^{a'} + \big(1 - \delta^{c-1,a'}\big) x_{i,\beta}^{c-1} x_{j,\beta}^{a'} \big) + \sum_{a'} \sum_{j} e^{i\frac{2\pi}{V} kj} x_{j,\beta}^{a'} \Bigg) \tag{218}$$

$$\approx \frac{P}{\sqrt{PL^2}} \left(\sum_{ca'} \sum_{ij} e^{i\frac{2\pi}{V} kj} x_{i,\alpha}^{a-1} \left(\delta^{c-1,a'} \frac{\delta_{ij}}{V} + \left(1 - \delta^{c-1,a'}\right) \frac{1}{V^2} \right) + \sum_{a'} \sum_{j} e^{i\frac{2\pi}{V} kj} \frac{1}{V} \right) \tag{219}$$

$$= \frac{P}{\sqrt{PL^2}} \frac{L}{V} \sum_{j} e^{i\frac{2\pi}{V} kj} x_{j,\alpha}^{a-1} = \frac{\sqrt{P}}{V} \sum_{j} e^{i\frac{2\pi}{V} kj} x_{j,\alpha}^{a-1}. \tag{220}$$

Decomposing via (154):

$$= \frac{P}{V} \sqrt{L(1 - L^{-1})} \big[\boldsymbol{v}_{1,k\geq 1} \big]_{\alpha}^{a} + \frac{P}{V} \big[\boldsymbol{v}_{0,k\geq 1} \big]_{\alpha}^{a}. \tag{221}$$

For $\boldsymbol{v}_{1,k\geq 1}$, it is

$$\sum_{\beta,c} \sum_{bd} \delta^{a-1,b} C_{bd,\alpha\beta}^{(xx)} \delta^{c-1,d} \big[\boldsymbol{v}_{1,k\geq 1} \big]_{\beta}^{c} \tag{222}$$

$$= \sum_{\beta,c} \left(\sum_{i} x_{i,\alpha}^{a-1} x_{i,\beta}^{c-1} + \delta^{a-1,c-1} \right) \frac{1}{\sqrt{PL(1 - L^{-1})}} \sum_{j} e^{i\frac{2\pi}{V} kj} \left(x_{j,\beta}^{c-1} - \frac{1}{L} \sum_{a'} x_{j,\beta}^{a'} \right) \tag{223}$$

$$= \frac{1}{\sqrt{PL(1 - L^{-1})}} \sum_{\beta,c} \sum_{j} e^{i\frac{2\pi}{V} kj} \left(\sum_{i} \left(x_{i,\alpha}^{a-1} x_{i,\beta}^{c-1} x_{j,\beta}^{c-1} - \frac{1}{L} \sum_{a'} x_{i,\alpha}^{a-1} x_{i,\beta}^{c-1} x_{j,\beta}^{a'} \right) + \left(x_{j,\beta}^{c-1} - \frac{1}{L} \sum_{a'} x_{j,\beta}^{a'} \right) \right). \tag{224}$$

Applying the equivalent kernel approximation:

$$\approx \frac{P}{\sqrt{PL(1 - L^{-1})}} \sum_{c} \sum_{j} e^{i\frac{2\pi}{V} kj} \left(\sum_{i} \left(x_{i,\alpha}^{a-1} \frac{\delta_{ij}}{V} - \frac{1}{L} \sum_{a'} x_{i,\alpha}^{a-1} \left(\delta^{c-1,a'} \frac{\delta_{ij}}{V} + \left(1 - \delta^{c-1,a'}\right) \frac{1}{V^2} \right) \right) + \left(\frac{1}{V} - \frac{1}{L} \sum_{a'} \frac{1}{V} \right) \right) \tag{225}$$

$$= \frac{P}{\sqrt{PL(1 - L^{-1})}} \sum_{c} \sum_{ij} e^{i\frac{2\pi}{V} kj} \left(x_{i,\alpha}^{a-1} \frac{\delta_{ij}}{V} - \frac{1}{L} x_{i,\alpha}^{a-1} \left(\frac{\delta_{ij}}{V} + (L-1) \frac{1}{V^2} \right) \right) \tag{226}$$

$$= \frac{P}{\sqrt{PL(1-\frac{1}{L})}} \frac{L}{V}\Big(1-\frac{1}{L}\Big)\sum_j e^{i\frac{2\pi}{V}kj} x_{j,\alpha}^{a-1}. \tag{227}$$

Now decomposing $x_{j,\alpha}^{a-1}$ via (154):

$$= \frac{P}{\sqrt{PL(1-\frac{1}{L})}} \frac{L}{V}\Big(1-\frac{1}{L}\Big)\Big(\sqrt{PL(1-L^{-1})}\,[\boldsymbol{v}_{1,k}]_\alpha^a + \sqrt{P}\,[\boldsymbol{v}_{0,k}]_\alpha^a\Big) \tag{228}$$

$$= \frac{P(L-1)}{V}\big[\boldsymbol{v}_{1,k\geq 1}\big]_\alpha^a + \frac{P\sqrt{L-1}}{V}\big[\boldsymbol{v}_{0,k\geq 1}\big]_\alpha^a. \tag{229}$$

Explicit expression for the data term

Collecting the results from all previous derivations, we have:

Term	$\boldsymbol{v}_{0,0}$	$\boldsymbol{v}_{0,k\geq 1}$	$\boldsymbol{v}_{1,k\geq 1}$
$(\mathbf{1}_L, \mathbf{1}_L)$	$PL^2(\frac{L}{V}+1)\ \boldsymbol{v}_{0,0}$	$\frac{PL^2}{V}\ \boldsymbol{v}_{0,k}$	0
$(\mathbf{1}_L, \boldsymbol{G}^*)$	$PL(\frac{L}{V}+1)\ \boldsymbol{v}_{0,0}$	$\frac{PL}{V}\ \boldsymbol{v}_{0,k}$	$\frac{PL\sqrt{L-1}}{V}\ \boldsymbol{v}_{0,k}$
$(\boldsymbol{G}^*, \mathbf{1}_L)$	$PL(\frac{L}{V}+1)\ \boldsymbol{v}_{0,0}$	$\frac{PL}{V}\ \boldsymbol{v}_{0,k} + \frac{PL\sqrt{L-1}}{V}\ \boldsymbol{v}_{1,k}$	0
$(\boldsymbol{G}^*, \boldsymbol{G}^*)$	$P(\frac{L}{V}+1)\ \boldsymbol{v}_{0,0}$	$\frac{P}{V}\ \boldsymbol{v}_{0,k} + \frac{P\sqrt{L-1}}{V}\ \boldsymbol{v}_{1,k}$	$\frac{P\sqrt{L-1}}{V}\ \boldsymbol{v}_{0,k} + \frac{P(L-1)}{V}\ \boldsymbol{v}_{1,k}$

We now assemble the full kernel $g_O \sum_{bd} G^{ab}G^{cd}C^{(xx)}_{bd,\alpha\beta}$ in the Fourier basis. Since $G = c_1\mathbf{1}_L + c_G\boldsymbol{G}^*$, we have

$$G^{ab}G^{cd} = c_1^2 + c_1c_G\delta^{c-1,d} + c_Gc_1\delta^{a-1,b} + c_G^2\delta^{a-1,b}\delta^{c-1,d}, \tag{230}$$

so the kernel is a weighted sum of the four terms computed in the previous section.

For $k=0$, only $\boldsymbol{v}_{0,0}$ contributes since $\boldsymbol{v}_{1,0}=0$. Reading off the $\boldsymbol{v}_{0,0}\to\boldsymbol{v}_{0,0}$ eigenvalue from each term in the summary table, we get

$$\mathcal{K}_0 = g_O\Big[c_1^2\cdot PL^2\Big(\frac{L}{V}+1\Big) + c_1c_G\cdot PL\Big(\frac{L}{V}+1\Big) + c_Gc_1\cdot PL\Big(\frac{L}{V}+1\Big) + c_G^2\cdot P\Big(\frac{L}{V}+1\Big)\Big] \tag{231}$$

$$= g_O P\Big(\frac{L}{V}+1\Big)\big(c_1L + c_G\big)^2. \tag{232}$$

Switching to normalized operators with $\hat{c}_1 = \frac{c_1}{L}$ and $\hat{c}_G = \frac{c_G}{\sqrt{L}}$, we get

$$\mathcal{K}_0 = g_O P\Big(\frac{L}{V}+1\Big)\Big(\hat{c}_1 + \frac{\hat{c}_G}{\sqrt{L}}\Big)^2. \tag{233}$$

For $k\geq 1$, the kernel acts on the two-dimensional space $(\boldsymbol{v}_{0,k},\ \boldsymbol{v}_{1,k})$. We can read off the 2×2 matrix representation of each term from the summary table.

$(\mathbf{1}_L, \mathbf{1}_L)$**:** $\boldsymbol{v}_{0,k}\to\frac{PL^2}{V}\boldsymbol{v}_{0,k}$, $\boldsymbol{v}_{1,k}\to 0$, so

$$\frac{P}{V}\begin{pmatrix} L^2 & 0 \\ 0 & 0\end{pmatrix}. \tag{234}$$

$(\mathbf{1}_L, \boldsymbol{G}^*)$**:** $\boldsymbol{v}_{0,k}\to\frac{PL}{V}\boldsymbol{v}_{0,k}$, $\boldsymbol{v}_{1,k}\to\frac{PL\sqrt{L-1}}{V}\boldsymbol{v}_{0,k}$, so

$$\frac{P}{V}\begin{pmatrix} L & L\sqrt{L-1} \\ 0 & 0\end{pmatrix}. \tag{235}$$

$(\boldsymbol{G}^*, \mathbf{1}_L)$**:** $\boldsymbol{v}_{0,k}\to\frac{PL}{V}\boldsymbol{v}_{0,k} + \frac{PL\sqrt{L-1}}{V}\boldsymbol{v}_{1,k}$, $\boldsymbol{v}_{1,k}\to 0$, so

$$\frac{P}{V}\begin{pmatrix} L & 0 \\ L\sqrt{L-1} & 0\end{pmatrix}. \tag{236}$$

$(\boldsymbol{G}^*, \boldsymbol{G}^*)$**:** $\boldsymbol{v}_{0,k} \to \frac{P}{V}\boldsymbol{v}_{0,k} + \frac{P\sqrt{L-1}}{V}\boldsymbol{v}_{1,k}$, $\boldsymbol{v}_{1,k} \to \frac{P\sqrt{L-1}}{V}\boldsymbol{v}_{0,k} + \frac{P(L-1)}{V}\boldsymbol{v}_{1,k}$, so

$$\frac{P}{V}\begin{pmatrix} 1 & \sqrt{L-1} \\ \sqrt{L-1} & L-1 \end{pmatrix}. \tag{237}$$

The full kernel matrix for $k \geq 1$ is

$$\begin{aligned} \mathcal{K}_k = g_O \Bigg[& c_1^2 \frac{P}{V}\begin{pmatrix} L^2 & 0 \\ 0 & 0 \end{pmatrix} + c_1 c_G \frac{P}{V}\begin{pmatrix} L & L\sqrt{L-1} \\ 0 & 0 \end{pmatrix} \\ & + c_G c_1 \frac{P}{V}\begin{pmatrix} L & 0 \\ L\sqrt{L-1} & 0 \end{pmatrix} + c_G^2 \frac{P}{V}\begin{pmatrix} 1 & \sqrt{L-1} \\ \sqrt{L-1} & L-1 \end{pmatrix} \Bigg] \end{aligned} \tag{238}$$

$$= \frac{g_O P}{V}\begin{pmatrix} c_1^2 L^2 + 2c_1 c_G L + c_G^2 & c_1 c_G L\sqrt{L-1} + c_G^2\sqrt{L-1} \\ c_G c_1 L\sqrt{L-1} + c_G^2\sqrt{L-1} & c_G^2(L-1) \end{pmatrix}. \tag{239}$$

Altogether, we obtain

$$\mathcal{K}_k = \frac{g_O P}{V}\begin{pmatrix} (c_1 L + c_G)^2 & (c_1 L + c_G) c_G \sqrt{L-1} \\ (c_1 L + c_G) c_G \sqrt{L-1} & c_G^2 (L-1) \end{pmatrix} = \frac{g_O P}{V}\, \boldsymbol{u}\,\boldsymbol{u}^\mathsf{T}, \tag{240}$$

where

$$\boldsymbol{u} = \begin{pmatrix} c_1 L + c_G \\ c_G\sqrt{L-1} \end{pmatrix} = \begin{pmatrix} \hat{c}_1 + \hat{c}_G/\sqrt{L} \\ \hat{c}_G\sqrt{(L-1)/L} \end{pmatrix}. \tag{241}$$

We again replaced the bare operators by normalized operators $c_1 \mathbf{1}_L \mapsto \hat{c}_1 \frac{1}{L}\mathbf{1}_L$ and $c_G \boldsymbol{G}^* \mapsto \hat{c}_G \frac{1}{\sqrt{L}}\boldsymbol{G}^*$. The final result is a rank-1 matrix, which will allow closed-form inversion via Sherman–Morrison.

Inversion of kernel using Sherman-Morrison

We want to express

$$S_{\text{data}} = \frac{1}{2}\sum_{ac,\alpha\beta,i} y^a_{i,\alpha}\big[\mathcal{K} + \sigma^2 \boldsymbol{I}\big]^{-1} y^c_{i,\beta}, \tag{242}$$

which in the Fourier basis becomes a sum over modes.

First, we look at the contributions of modes with $k = 0$. Since $\boldsymbol{v}_{1,0} = 0$, only the scalar mode $\boldsymbol{v}_{0,0}$ contributes with

$$S_{\text{data},0} = \frac{1}{2}\sum_i \frac{|\boldsymbol{y}_i^\mathsf{T}\boldsymbol{v}_{0,0}|^2}{\mathcal{K}_0 + \sigma^2} = \frac{1}{2}\cdot\frac{V \cdot PL/V^2}{\mathcal{K}_0 + \sigma^2} = \frac{1}{2}\cdot\frac{PL/V}{\mathcal{K}_0 + \sigma^2}, \tag{243}$$

where we used $|\boldsymbol{y}_i^\mathsf{T}\boldsymbol{v}_{0,0}|^2 = PL/V^2$ (independent of i) and $\sum_i 1 = V$. Substituting $\mathcal{K}_0$:

$$S_{\text{data},0} = \frac{PL/V}{2\Big[g_O P\big(\hat{c}_1 + \hat{c}_G/\sqrt{L}\big)^2(L/V+1) + \sigma^2\Big]}. \tag{244}$$

For modes with $k \geq 1$, the matrix to invert is $\mathcal{K}_k + \sigma^2\boldsymbol{I} = \frac{g_O P}{V}\boldsymbol{u}\boldsymbol{u}^\mathsf{T} + \sigma^2\boldsymbol{I}$. Since this is a rank-1 update of $\sigma^2\boldsymbol{I}$, the Sherman–Morrison formula gives

$$\big(\sigma^2\boldsymbol{I} + A\boldsymbol{u}\boldsymbol{u}^\mathsf{T}\big)^{-1} = \frac{1}{\sigma^2}\boldsymbol{I} - \frac{A}{\sigma^2(\sigma^2 + A|\boldsymbol{u}|^2)}\,\boldsymbol{u}\boldsymbol{u}^\mathsf{T}, \tag{245}$$

where $A = g_O P/V$. We now compute $|\boldsymbol{u}|^2$:

$$|\boldsymbol{u}|^2 = (c_1 L + c_G)^2 + c_G^2(L-1) \tag{246}$$

$$= \Big(\hat{c}_1 + \frac{\hat{c}_G}{\sqrt{L}}\Big)^2 + \hat{c}_G^2\frac{L-1}{L}. \tag{247}$$

The quadratic form $\boldsymbol{y}^\dagger_{i,k}(\mathcal{K}_k + \sigma^2\boldsymbol{I})^{-1}\boldsymbol{y}_{i,k}$ becomes

$$\boldsymbol{y}^\dagger_{i,k}(\mathcal{K}_k + \sigma^2\boldsymbol{I})^{-1}\boldsymbol{y}_{i,k} = \frac{1}{\sigma^2}|\boldsymbol{y}_{i,k}|^2 - \frac{A}{\sigma^2(\sigma^2 + A|\boldsymbol{u}|^2)}|\boldsymbol{u}^\mathsf{T}\boldsymbol{y}_{i,k}|^2. \tag{248}$$

The first term gives

$$|\boldsymbol{y}_{i,k}|^2 = \frac{P}{V^2}(1 + L - 1) = \frac{PL}{V^2}. \tag{249}$$

Next, the inner product $\boldsymbol{u}^\mathsf{T}\boldsymbol{y}_{i,k}$ yields

$$\boldsymbol{u}^\mathsf{T}\boldsymbol{y}_{i,k} = \frac{\sqrt{P}}{V} e^{i\frac{2\pi}{V}ki}\Big[(c_1 L + c_G)\cdot 1 + c_G\sqrt{L-1}\cdot\sqrt{L-1}\Big] \tag{250}$$

$$= \frac{\sqrt{P}}{V} e^{i\frac{2\pi}{V}ki}\big[c_1 L + c_G + c_G(L-1)\big] \tag{251}$$

$$= \frac{\sqrt{P}}{V} e^{i\frac{2\pi}{V}ki}\cdot L(c_1 + c_G). \tag{252}$$

In normalized coordinates, $L(c_1 + c_G) = L(\hat{c}_1/L + \hat{c}_G/\sqrt{L}) = \hat{c}_1 + \sqrt{L}\,\hat{c}_G$, so

$$|\boldsymbol{u}^\mathsf{T}\boldsymbol{y}_{i,k}|^2 = \frac{P}{V^2}(\hat{c}_1 + \sqrt{L}\,\hat{c}_G)^2. \tag{253}$$

Substituting into the quadratic form yields

$$\boldsymbol{y}_{i,k}^\dagger(\mathcal{K}_k + \sigma^2\boldsymbol{I})^{-1}\boldsymbol{y}_{i,k} = \frac{PL}{V^2\sigma^2} - \frac{(g_O P/V)\cdot P(\hat{c}_1 + \sqrt{L}\,\hat{c}_G)^2}{V^2\sigma^2(\sigma^2 + (g_O P/V)|\boldsymbol{u}|^2)}. \tag{254}$$

This is independent of i since $|e^{i2\pi ki/V}|^2 = 1$. Summing over $i = 0, \dots, V-1$ gives a factor of V, and summing over $k = 1, \dots, V-1$ gives $V-1$ identical terms (using that $\mathcal{K}_k$ is independent of k for $k \geq 1$):

$$S_{\mathrm{data},k\geq 1} = \frac{1}{2}\sum_{k=1}^{V-1}\sum_{i=0}^{V-1}\boldsymbol{y}_{i,k}^\dagger(\mathcal{K}_k + \sigma^2\boldsymbol{I})^{-1}\boldsymbol{y}_{i,k} \tag{255}$$

$$= \frac{(V-1)V}{2}\left(\frac{PL}{V^2\sigma^2} - \frac{g_O P^2(\hat{c}_1 + \sqrt{L}\,\hat{c}_G)^2}{V^3\sigma^2(\sigma^2 + \frac{g_O P}{V}|\boldsymbol{u}|^2)}\right) \tag{256}$$

$$= \frac{V-1}{2}\left(\frac{PL}{V\sigma^2} - \frac{g_O P^2(\hat{c}_1 + \sqrt{L}\,\hat{c}_G)^2}{V^2\sigma^2(\sigma^2 + \frac{g_O P}{V}|\boldsymbol{u}|^2)}\right). \tag{257}$$

Combining all terms, we get

$$S_{\mathrm{data}}(\hat{c}_1, \hat{c}_G) = \frac{PL/V}{2\big[g_O P(\hat{c}_1 + \hat{c}_G/\sqrt{L})^2(L/V+1) + \sigma^2\big]} + \frac{V-1}{2}\left(\frac{PL}{V\sigma^2} - \frac{g_O P^2(\hat{c}_1 + \sqrt{L}\,\hat{c}_G)^2}{V^2\sigma^2\big(\sigma^2 + \frac{g_O P}{V}|\boldsymbol{u}|^2\big)}\right), \tag{258}$$

where $|\boldsymbol{u}|^2 = (\hat{c}_1 + \hat{c}_G/\sqrt{L})^2 + \hat{c}_G^2(L-1)/L$.

F Fluctuation Term

In this section, we derive an expression for

$$\ln\det\Big[g_O\sum_{b,d} G^{ab}G^{cd}\,C^{(xx)}_{bd,\alpha\beta} + \sigma^2\boldsymbol{I}\Big], \tag{259}$$

in the basis $\boldsymbol{G} = c_1\mathbf{1}_L + c_G\boldsymbol{G}^*$ and then transform to the normalized basis $(\hat{c}_1, \hat{c}_G)$ in the end. We denote the matrix inside the determinant as $\mathcal{K} + \sigma^2\boldsymbol{I}$.

Identifying all eigenspaces and eigenvalue structure

The Fourier basis $\{\boldsymbol{v}_{0,0},\ \boldsymbol{v}_{0,k\geq 1},\ \boldsymbol{v}_{1,k\geq 1}\}$ used in the previous section spans a $(2V-1)$-dimensional subspace of $\mathbb{R}^{LP}$. For the fluctuation term, we must carefully check what happens on the orthogonal complement, because the δ^{bd} term in $C^{(xx)}_{bd,\alpha\beta} = \sum_i x^b_{i,\alpha}x^d_{i,\beta} + \delta^{bd}$ can produce non-trivial eigenvalues there.

Consider vectors of the form

$$[w_m]^c_\beta = \frac{1}{\sqrt{P}} f_m(c), \qquad m = 1, \dots, L-1, \tag{260}$$

that are constant in the data index β while being traceless in the sequences index $\sum_c f_m(c) = 0$. These give $L-1$ independent directions and they are orthogonal to all Fourier basis vectors:

- $\boldsymbol{v}_{0,0}$: $\sum_{c,\beta}[w_m]^c_\beta[\boldsymbol{v}_{0,0}]^c_\beta \propto \sum_c f_m(c) = 0$.
- $\boldsymbol{v}_{0,k\geq 1}$: $\sum_{c,\beta}[w_m]^c_\beta[\boldsymbol{v}_{0,k}]^c_\beta \propto \sum_c f_m(c) \sum_\beta \sum_j e^{i\frac{2\pi}{V}kj} x^a_{j,\beta} \propto 0 \sum_\beta \sum_j e^{i\frac{2\pi}{V}kj} x^a_{j,\beta} = 0$.
- $\boldsymbol{v}_{1,k\geq 1}$: $\sum_{c,\beta}[w_m]^c_\beta[\boldsymbol{v}_{1,k}]^c_\beta \propto \sum_\beta \sum_j e^{i\frac{2\pi}{V}kj} \sum_c f_m(c)(x^{c-1}_{j,\beta} - \frac{1}{L}\sum_a x^a_{j,\beta})$. Under the equivalent kernel, $\sum_\beta x^c_{j,\beta} \approx P/V$, so $\sum_\beta (x^{c-1}_{j,\beta} - \frac{1}{L}\sum_a x^a_{j,\beta}) \approx 0$.

We now compute $[\mathcal{K}\, w_m]^a_\alpha = g_O \sum_{c,\beta}\sum_{bd} G^{ab}G^{cd}C^{(xx)}_{bd,\alpha\beta}[w_m]^c_\beta$, separating the two parts of $C^{(xx)}$.

For the data term $\sum_i x^b_{i,\alpha}x^d_{i,\beta}$, we get

$$g_O \sum_{c,\beta}\sum_{bd} G^{ab}G^{cd}\Big(\sum_i x^b_{i,\alpha}x^d_{i,\beta}\Big)\frac{f_m(c)}{\sqrt{P}} \tag{261}$$

$$= \frac{g_O}{\sqrt{P}} \sum_{bd} G^{ab}\Big(\sum_i x^b_{i,\alpha}\sum_\beta x^d_{i,\beta}\Big)\sum_c G^{cd}f_m(c). \tag{262}$$

Under the equivalent kernel, $\sum_\beta x^d_{i,\beta} \approx P/V$, so the factor $\sum_i x^b_{i,\alpha} \cdot P/V = P x^b_{\cdot,\alpha}/V$ is independent of d. Then

$$\sum_d\sum_c G^{cd}f_m(c) = \sum_c f_m(c)\sum_d G^{cd} = \sum_c f_m(c)\big(c_1 L + c_G\big) = (c_1 L + c_G)\sum_c f_m(c) = 0, \tag{263}$$

since $\sum_d G^{cd} = c_1 L + c_G$ is independent of c and $\sum_c f_m(c) = 0$. So the data part vanishes.

For the identity term δ^{bd} part, we get

$$g_O \sum_{c,\beta}\sum_{bd} G^{ab}G^{cd}\delta^{bd}\frac{f_m(c)}{\sqrt{P}} = \frac{g_O}{\sqrt{P}} \underbrace{\sum_\beta 1}_{=P} \sum_b G^{ab}\sum_c G^{cb}f_m(c). \tag{264}$$

We compute $\sum_b G^{ab}G^{cb}$ explicitly. Since $G^{ab} = c_1 + c_G\delta^{a-1,b}$:

$$\sum_b G^{ab}G^{cb} = \sum_b (c_1 + c_G\delta^{a-1,b})(c_1 + c_G\delta^{c-1,b}) \tag{265}$$

$$= c_1^2 L + c_1 c_G + c_G c_1 + c_G^2\delta^{a-1,c-1} \tag{266}$$

$$= c_1^2 L + 2c_1 c_G + c_G^2\delta_{ac}. \tag{267}$$

The first two terms are independent of c, so contracting with $f_m(c)$ with $\sum_c f_m(c) = 0$ kills them:

$$\sum_c (c_1^2 L + 2c_1 c_G + c_G^2\delta_{ac})f_m(c) = c_G^2 f_m(a). \tag{268}$$

Therefore

$$[\mathcal{K}\, w_m]^a_\alpha = g_O P\, c_G^2 \cdot \frac{f_m(a)}{\sqrt{P}} = g_O P\, c_G^2\, [w_m]^a_\alpha, \tag{269}$$

with eigenvalue $\lambda_w = g_O P c_G^2 = \frac{g_O P \hat{c}_G^2}{L}$.

Deriving the full log-determinant

Taking together the results from the previous section, we have that the LP-dimensional space decomposes as:

Subspace	Multiplicity	Eigenvalue of $\mathcal{K} + \sigma^2 \boldsymbol{I}$
$\boldsymbol{v}_{0,0}$	1	$\mathcal{K}_0 + \sigma^2$
$(\boldsymbol{v}_{0,k}, \boldsymbol{v}_{1,k})$, non-trivial direction	$V-1$	$\sigma^2 + \frac{g_O P}{V}\lvert\boldsymbol{u}\rvert^2$
$(\boldsymbol{v}_{0,k}, \boldsymbol{v}_{1,k})$, trivial direction	$V-1$	σ^2
$\boldsymbol{w}_m$	$L-1$	$g_O P c_G^2 + \sigma^2$
Null space	$LP - L - 2V + 2$	σ^2

We now switch to the normalized coordinates $c_1 \mapsto \frac{\hat{c}_1}{L}$ and $c_G \mapsto \frac{\hat{c}_G}{\sqrt{L}}$.

For $\boldsymbol{v}_{0,0}$, we get a scalar:

$$\ln(\mathcal{K}_0 + \sigma^2) = \ln\left[g_O P\Big(\hat{c}_1 + \frac{\hat{c}_G}{\sqrt{L}}\Big)^2\Big(\frac{L}{V} + 1\Big) + \sigma^2\right]. \tag{270}$$

Next, we determine the $\boldsymbol{v}_{i,k\geq 1}$ component. Each 2×2 block has the form $\sigma^2 \boldsymbol{I}_2 + A\boldsymbol{u}\boldsymbol{u}^\mathsf{T}$ where $A = g_O P/V$. By the matrix determinant lemma:

$$\det(\sigma^2 \boldsymbol{I}_2 + A\boldsymbol{u}\boldsymbol{u}^\mathsf{T}) = \det(\sigma^2 \boldsymbol{I}_2) \cdot \big(1 + A\,\boldsymbol{u}^\mathsf{T}(\sigma^2 \boldsymbol{I}_2)^{-1}\boldsymbol{u}\big) = \sigma^4\Big(1 + \frac{A}{\sigma^2}|\boldsymbol{u}|^2\Big). \tag{271}$$

Taking the logarithm and summing over $k = 1, \ldots, V-1$ (the kernel is independent of k), we obtain

$$\sum_{k=1}^{V-1} \ln\det\Big[\frac{g_O P}{V}\boldsymbol{u}\boldsymbol{u}^\mathsf{T} + \sigma^2 \boldsymbol{I}_2\Big] = (V-1)\Big[2\ln\sigma^2 + \ln\Big(1 + \frac{g_O P}{V\sigma^2}|\boldsymbol{u}|^2\Big)\Big]. \tag{272}$$

For $\boldsymbol{w}_m$, each of the $(L-1$ directions contributes the same term, yielding:

$$(L-1)\ln\Big(\frac{g_O P \hat{c}_G^2}{L} + \sigma^2\Big). \tag{273}$$

From the remaining $LP - L - 2V + 2$ directions of the null-space, each contributes $\ln\sigma^2$. Combining all terms, we get

$$\begin{aligned}\ln\det[\mathcal{K} + \sigma^2 \boldsymbol{I}] = (LP - L)\ln\sigma^2 &+ \ln\left[g_O P\Big(\hat{c}_1 + \frac{\hat{c}_G}{\sqrt{L}}\Big)^2\Big(\frac{L}{V} + 1\Big) + \sigma^2\right] \\ &+ (V-1)\ln\left(\sigma^2 + \frac{g_O P}{V}|\boldsymbol{u}|^2\right) + (L-1)\ln\left(\frac{g_O P \hat{c}_G^2}{L} + \sigma^2\right),\end{aligned} \tag{274}$$

where $|\boldsymbol{u}|^2 = (\hat{c}_1 + \hat{c}_G/\sqrt{L})^2 + \hat{c}_G^2(L-1)/L$.

The term $(LP - L)\ln\sigma^2$ is independent of $(\hat{c}_1, \hat{c}_G)$ and can be dropped. The term $(L - 1)\ln(g_O P \hat{c}_G^2/L + \sigma^2)$ arises from the δ^{bd} part of $C^{(xx)}$ acting on directions that are constant in the data index α but traceless in the copy index a.

G Mapping the Posterior Through the Softmax

We derive here the one-dimensional posterior for softmax attention theory, with $\mathrm{softmax}$ explicitly defined as

$$\mathrm{softmax}^a(\boldsymbol{x}; \beta) = \frac{e^{\beta x^a}}{\sum_{a=1}^{L} e^{\beta x^a}} \tag{275}$$

with β the softmax inverse temperature, which we take to be 1 in the main text. The starting point is the posterior over the pre-softmax attention scores from linear attention restricted to the two-dimensional order-parameter ansatz

$$\boldsymbol{G} = \hat{c}_1 \frac{\mathbf{1}_L}{L} + \hat{c}_G \frac{\boldsymbol{G}^*}{\sqrt{L}}, \qquad G^*_{ab} = \delta^{a-1,b}. \tag{276}$$

For softmax attention, the actual attention matrix is the row-wise softmax of the score matrix. With the normalization used in the main text, this means

$$\boldsymbol{A} = \mathrm{softmax}(L\boldsymbol{G}; \beta). \tag{277}$$

The posterior over $\boldsymbol{A}$ is therefore the pushforward of the posterior over $\boldsymbol{G}$,

$$p(\boldsymbol{A}) = \int d\hat{c}_1 \, d\hat{c}_G \, \delta(\boldsymbol{A} - \mathrm{softmax}(\beta L \boldsymbol{G}(\hat{c}_1, \hat{c}_G))) \, p(\hat{c}_1, \hat{c}_G) \tag{278}$$

Eliminating the uniform mode. The softmax is invariant under adding a constant to every entry in a row. In the ansatz above, the $\hat{c}_1$ direction shifts all entries in each row by the same amount, and therefore drops out of the softmax. Explicitly,

$$\mathrm{softmax}(\beta L G) = \frac{e^{\beta\sqrt{L}\hat{c}_G + \beta\hat{c}_1}}{(L-1)e^{\beta\hat{c}_1} + e^{\beta\sqrt{L}\hat{c}_G + \beta\hat{c}_1}}, \tag{279}$$

so that the common factor $e^{\beta\hat{c}_1}$ cancels between numerator and denominator. Thus the softmax attention pattern depends only on $\hat{c}_G$. Thus, each row can be taken to have one copy entry, where $G_{ab} = \hat{c}_G/\sqrt{L}$, and $L-1$ non-copy entries, where $G_{ab} = 0$, hence

$$A_G(\hat{c}_G) = \frac{e^{\beta\sqrt{L}\hat{c}_G}}{L - 1 + e^{\beta\sqrt{L}\hat{c}_G}}, \qquad A_1(\hat{c}_G) = \frac{1}{L - 1 + e^{\beta\sqrt{L}\hat{c}_G}}, \tag{280}$$

where A_G denotes the attention weight on the copy position and A_1 the weight on each of the other $L-1$ positions. These variables obey the row-normalization constraint

$$A_G + (L-1)A_1 = 1. \tag{281}$$

Consequently, the pushforward posterior is supported on a one-dimensional manifold in the space of row-normalized attention matrices.

Let

$$p(\hat{c}_G) = \int d\hat{c}_1 \, p(\hat{c}_1, \hat{c}_G) \tag{282}$$

be the marginal posterior over the softmax-relevant score direction. Then the induced distribution over the copy entry is

$$p(A_G) = \int d\hat{c}_G \, \delta\left(A_G - \frac{e^{\beta\sqrt{L}\hat{c}_G}}{L - 1 + e^{\beta\sqrt{L}\hat{c}_G}}\right) p(\hat{c}_G) \tag{283}$$

and,

$$p(A_G, A_1) = \delta\left(A_1 - \frac{1 - A_G}{L-1}\right) p(A_G). \tag{284}$$

Inverse map. The map in Eq. (280) is monotone in $\hat{c}_G$, so we can invert it explicitly. Starting from

$$A_G = \frac{e^{\beta\sqrt{L}\hat{c}_G}}{L - 1 + e^{\beta\sqrt{L}\hat{c}_G}}, \tag{285}$$

we obtain

$$A_G(L-1) = e^{\beta\sqrt{L}\hat{c}_G}(1 - A_G), \tag{286}$$

and therefore

$$\hat{c}_G(A_G) = \frac{1}{\beta\sqrt{L}} \log\left[\frac{(L-1)A_G}{1 - A_G}\right]. \tag{287}$$

The corresponding Jacobian is

$$\frac{d\hat{c}_G}{dA_G} = \frac{1}{\beta\sqrt{L}\, A_G(1 - A_G)}, \tag{288}$$

so the one-dimensional pushforward density can be written as

$$p(A_G) = \frac{p(\hat{c}_G(A_G))}{\beta\sqrt{L}\, A_G(1 - A_G)}. \tag{289}$$

Softmax order parameter. In the main text we express the post-softmax attention matrix in the same structural basis as the linear-attention ansatz

$$\boldsymbol{A} = \hat{a}_1 \frac{\mathbf{1}_L}{L} + \hat{a}_G \frac{\boldsymbol{G}^*}{\sqrt{L}}. \tag{290}$$

Since $\boldsymbol{A}$ is row-normalized, the two coefficients are not independent:

$$\hat{a}_1 + \frac{\hat{a}_G}{\sqrt{L}} = 1, \qquad \hat{a}_1 = 1 - \frac{\hat{a}_G}{\sqrt{L}}. \tag{291}$$

The entries of the attention matrix are therefore

$$A_G = \frac{\hat{a}_1}{L} + \frac{\hat{a}_G}{\sqrt{L}} = \frac{1 + \frac{L-1}{\sqrt{L}}\hat{a}_G}{L}, \qquad A_1 = \frac{\hat{a}_1}{L} = \frac{1 - \frac{1}{\sqrt{L}}\hat{a}_G}{L}. \tag{292}$$

Substituting Eq. (292) into Eq. (287) gives the softmax inverse map used in the main text:

$$\hat{c}_G(\hat{a}_G) = \frac{1}{\beta\sqrt{L}} \log\left[\frac{1 + \frac{L-1}{\sqrt{L}}\hat{a}_G}{1 - \frac{1}{\sqrt{L}}\hat{a}_G}\right]. \tag{293}$$

The allowed range is

$$-\frac{\sqrt{L}}{L-1} < \hat{a}_G < \sqrt{L}, \tag{294}$$

which is exactly the condition that both A_G and A_1 are positive.

For completeness, the Jacobian in the $\hat{a}_G$ coordinate is

$$\frac{d\hat{c}_G}{d\hat{a}_G} = \frac{1}{\left(1 + \frac{L-1}{\sqrt{L}}\hat{a}_G\right)\left(1 - \frac{1}{\sqrt{L}}\hat{a}_G\right)}. \tag{295}$$

Thus the exact one-dimensional pushforward density contains a Jacobian contribution

$$p(\hat{a}_G) \propto \exp[-\mathcal{S}_{\text{softmax}}(\hat{a}_G)]\,, \tag{296}$$

$$\mathcal{S}_{\text{softmax}}(\hat{a}_G) = \mathcal{S}_{\text{prior}}(\hat{c}_G(\hat{a}_G)) + \mathcal{S}_{\text{data}}(\hat{a}_1(\hat{a}_G), \hat{a}_G) - \log\left|\frac{d\hat{c}_G}{d\hat{a}_G}\right|. \tag{297}$$

The Jacobian term is subleading in the large-χL saddle-point limit because the prior and data terms scale as χL, while the Jacobian contributes only an additive logarithmic correction. We therefore drop it in the saddle-point analysis in the main text.

Softmax action. We now substitute the map above into the reduced action. The prior term is obtained from Eq. (12) by setting $\hat{c}_1 = 0$, since $\hat{c}_1$ is a uniform mode that gets absorbed by the softmax and is fixed to zero by the prior at the saddle point. Using Eq. (293), this gives

$$\mathcal{S}_{\text{prior}}^{\text{softmax}}(\hat{a}_G) = \frac{\chi L}{2} \frac{(L+V)^2 - L - 2V}{\beta^2 g_W (L+V)^2} \log^2\left[\frac{1 + \frac{L-1}{\sqrt{L}}\hat{a}_G}{1 - \frac{1}{\sqrt{L}}\hat{a}_G}\right]. \tag{298}$$

The data term depends only on the post-softmax attention matrix itself. Therefore, in Eq. (15) we substitute

$$\hat{c}_1 \mapsto \hat{a}_1 = 1 - \frac{\hat{a}_G}{\sqrt{L}}, \qquad \hat{c}_G \mapsto \hat{a}_G. \tag{299}$$

Under this substitution,

$$\hat{a}_1 + \frac{\hat{a}_G}{\sqrt{L}} = 1, \tag{300}$$

so the uniform-mode contribution is independent of $\hat{a}_G$ and can be dropped when determining the saddle point. The nontrivial copy-sensitive part gives

$$\mathcal{S}_{\text{data}}^{\text{softmax}}(\hat{a}_G) = -\frac{\chi L}{2} \frac{P}{\sigma^2} \frac{V-1}{V} \frac{\left(1 + \sqrt{L}\,\hat{a}_G - \frac{\hat{a}_G}{\sqrt{L}}\right)^2 g_O}{\left(1 + \frac{L-1}{L}\hat{a}_G^2\right) g_O + \frac{\sigma^2}{P} V}, \tag{301}$$

up to terms independent of $\hat{a}_G$.

Combining Eqs. (298) and (301), and dropping constants and the subleading Jacobian term, yields

$$\mathcal{S}_{\text{softmax}}(\hat{a}_G) = \frac{\chi L}{2}\left[\frac{(L+V)^2 - L - 2V}{\beta^2 g_W (L+V)^2}\log^2\left(\frac{1+\frac{L-1}{\sqrt{L}}\hat{a}_G}{1-\frac{1}{\sqrt{L}}\hat{a}_G}\right) - \frac{P}{\sigma^2}\frac{V-1}{V}\frac{\left(1+\sqrt{L}\,\hat{a}_G - \frac{\hat{a}_G}{\sqrt{L}}\right)^2 g_O}{\left(1+\frac{L-1}{L}\hat{a}_G^2\right)g_O + \frac{\sigma^2}{P}V}\right]. \tag{302}$$

Finally, writing $\sigma^2 = \bar{\sigma}^2 L$ and keeping the leading large-L terms gives the expression in the main text:

$$\mathcal{S}_{\text{softmax}}(\hat{a}_G) = \frac{\chi L}{2}\left[\frac{1}{\beta^2 g_W}\log^2\left(\frac{1+\sqrt{L}\,\hat{a}_G}{1-\hat{a}_G/\sqrt{L}}\right) - \frac{P}{\bar{\sigma}^2 L}\frac{V-1}{V}\frac{\left(\hat{a}_G + \frac{1}{\sqrt{L}}\right)^2 g_O}{(\hat{a}_G^2+1)g_O + \frac{\bar{\sigma}^2 L}{P}V} + O\left(\frac{1}{L}\right)\right] \tag{303}$$

In the main text we take $\beta = 1$ for simplicity, but the β dependence can be recovered by $g_W \to g_W\beta^2$.

H From the Reduced Action to the Training Loss

We start from the marginal likelihood of the labels before integrating out the network outputs:

$$Z(\sigma^2) \;\coloneqq\; p(\{\boldsymbol{y}_\alpha\} \mid \{\boldsymbol{x}_\alpha\}_\alpha) \propto \int \mathcal{D}\boldsymbol{G}\int \mathcal{D}\boldsymbol{f}\; p(\boldsymbol{G}\mid\boldsymbol{X})\,p(\boldsymbol{F}\mid\boldsymbol{G})\,\exp\Big(-\frac{1}{2\sigma^2}\sum_{\alpha=1}^{P}\|\boldsymbol{y}_\alpha - \boldsymbol{f}_\alpha\|_2^2\Big). \tag{304}$$

Differentiating (304) with respect to σ^2 gives

$$\begin{aligned}\frac{\partial Z}{\partial\sigma^2} &= \int \mathcal{D}\boldsymbol{G}\int \mathcal{D}\boldsymbol{f}\; p(\boldsymbol{G}\mid\boldsymbol{X})\,p(\boldsymbol{F}\mid\boldsymbol{G})\,\frac{\partial}{\partial\sigma^2}\exp\Big(-\frac{1}{2\sigma^2}\sum_{\alpha=1}^{P}\|\boldsymbol{y}_\alpha - \boldsymbol{f}_\alpha\|_2^2\Big)\\ &= \frac{1}{2\sigma^4}\int \mathcal{D}\boldsymbol{G}\int \mathcal{D}\boldsymbol{f}\;\Big(\sum_{\alpha=1}^{P}\|\boldsymbol{y}_\alpha - \boldsymbol{f}_\alpha\|_2^2\Big)\,p(\boldsymbol{G}\mid\boldsymbol{X})\,p(\boldsymbol{F}\mid\boldsymbol{G})\,\exp\Big(-\frac{1}{2\sigma^2}\sum_{\alpha=1}^{P}\|\boldsymbol{y}_\alpha - \boldsymbol{f}_\alpha\|_2^2\Big).\end{aligned} \tag{305}$$

Dividing by $Z(\sigma^2)$, we obtain

$$\sigma^4\frac{\partial}{\partial\sigma^2}\ln Z(\sigma^2) = \frac{1}{2}\left\langle\sum_{\alpha=1}^{P}\|\boldsymbol{y}_\alpha - \boldsymbol{f}_\alpha\|_2^2\right\rangle, \tag{306}$$

where $\langle\cdots\rangle$ denotes the posterior average with respect to the measure in (304). Thus, differentiating the log-marginal likelihood with respect to σ^2 brings down the quadratic data-fit term.

Relation to the reduced action. After integrating out $\boldsymbol{F}$ and projecting onto the two-dimensional ansatz

$$\boldsymbol{G} = \hat{c}_1\,\frac{\mathbf{1}_L}{L} + \hat{c}_G\,\frac{\boldsymbol{G}^*}{\sqrt{L}}, \tag{307}$$

the same marginal likelihood takes the form

$$Z(\sigma^2) \propto \int d\hat{c}_1\,d\hat{c}_G\; e^{-\mathcal{S}(\hat{c}_1,\hat{c}_G;\sigma^2)}. \tag{308}$$

Differentiating (308) yields

$$\frac{\partial}{\partial\sigma^2}\ln Z(\sigma^2) = \frac{\int d\hat{c}_1\,d\hat{c}_G\,\big(\partial_{\sigma^2}\mathcal{S}(\hat{c}_1,\hat{c}_G;\sigma^2)\big)e^{-\mathcal{S}(\hat{c}_1,\hat{c}_G;\sigma^2)}}{\int d\hat{c}_1\,d\hat{c}_G\;e^{-\mathcal{S}(\hat{c}_1,\hat{c}_G;\sigma^2)}} = \Big\langle\partial_{\sigma^2}\mathcal{S}\Big\rangle_{\hat{c}_1,\hat{c}_G}, \tag{309}$$

where the average is now taken with respect to the reduced posterior over $(\hat{c}_1, \hat{c}_G)$.

In the large-χ limit used in the main text, the reduced posterior is sharply peaked and the integral (308) is dominated by its saddle point. Writing

$$(\hat{c}_1^\star, \hat{c}_G^\star) = \arg\min_{\hat{c}_1, \hat{c}_G} \mathcal{S}(\hat{c}_1, \hat{c}_G; \sigma^2), \tag{310}$$

we have, to leading order in the saddle-point approximation,

$$\ln Z(\sigma^2) \simeq \mathcal{S}(\hat{c}_1^\star, \hat{c}_G^\star; \sigma^2), \qquad \frac{\partial}{\partial \sigma^2} \ln Z(\sigma^2) \simeq \frac{d}{d\sigma^2} \mathcal{S}(\hat{c}_1^\star, \hat{c}_G^\star; \sigma^2). \tag{311}$$

Using the chain rule,

$$\frac{d}{d\sigma^2} \mathcal{S}(\hat{c}_1^\star, \hat{c}_G^\star; \sigma^2) = \frac{\partial \mathcal{S}}{\partial \sigma^2} + \frac{\partial \mathcal{S}}{\partial \hat{c}_1} \frac{d\hat{c}_1^\star}{d\sigma^2} + \frac{\partial \mathcal{S}}{\partial \hat{c}_G} \frac{d\hat{c}_G^\star}{d\sigma^2}. \tag{312}$$

At the saddle point, however,

$$\left.\frac{\partial \mathcal{S}}{\partial \hat{c}_1}\right|_{(\hat{c}_1^\star, \hat{c}_G^\star)} = 0, \qquad \left.\frac{\partial \mathcal{S}}{\partial \hat{c}_G}\right|_{(\hat{c}_1^\star, \hat{c}_G^\star)} = 0, \tag{313}$$

so the implicit derivatives of the order parameters drop out and

$$\frac{d}{d\sigma^2} \mathcal{S}(\hat{c}_1^\star, \hat{c}_G^\star; \sigma^2) = \left.\frac{\partial \mathcal{S}(\hat{c}_1, \hat{c}_G; \sigma^2)}{\partial \sigma^2}\right|_{(\hat{c}_1, \hat{c}_G) = (\hat{c}_1^\star, \hat{c}_G^\star)}. \tag{314}$$

Combining (306), (311), and (314), we obtain the relation

$$\frac{1}{2} \left\langle \sum_{\alpha=1}^{P} \|\boldsymbol{y}_\alpha - \boldsymbol{f}_\alpha\|_2^2 \right\rangle \simeq \sigma^4 \left.\frac{\partial \mathcal{S}(\hat{c}_1, \hat{c}_G; \sigma^2)}{\partial \sigma^2}\right|_{(\hat{c}_1, \hat{c}_G) = (\hat{c}_1^\star, \hat{c}_G^\star)}. \tag{315}$$

Under the equivalent-kernel (EK) approximation ($\sum_\alpha \circ \approx P\, \mathbb{E}_x[\circ]$), which is valid when $\sigma^2 \gg \sum_a (\boldsymbol{y}^a - \boldsymbol{f}^a(\boldsymbol{x}))^2$, the generalization gap is small [52, 50] and we find

$$\frac{1}{2} \left\langle \mathbb{E}_{\boldsymbol{x}} \|\boldsymbol{y}(\boldsymbol{x}) - \boldsymbol{f}(\boldsymbol{x})\|_2^2 \right\rangle \simeq \frac{\sigma^4}{P} \left.\frac{\partial \mathcal{S}(\hat{c}_1, \hat{c}_G; \sigma^2)}{\partial \sigma^2}\right|_{(\hat{c}_1, \hat{c}_G) = (\hat{c}_1^\star, \hat{c}_G^\star)}. \tag{316}$$

I Spinodals of the Softmax Action

In addition to the binodal (phase transition) curve $P_{\hat{a}_G}$ characterized in Eq. 21, there are two other interesting events: the first spinodal when the copy minimum appears but is not yet the global minimum, and the second spinodal when the pooling minimum is annihilated by the maximum and the metastable branch disappears. In this appendix, we estimate their sample complexity.

I.1 Pooling State Loss of Stability

Starting from Eq.20 we want to solve for $\frac{\partial}{\partial \hat{a}_G} S \stackrel{!}{=} 0$ and simultaneously $\frac{\partial^2}{\partial \hat{a}_G^2} S \stackrel{!}{=} 0$ when $\hat{a}_G$ itself is small (that is, on the pooling branch). In this phase $\hat{a}_G \sim 1/\sqrt{L}$ so we define $x = \sqrt{L}\hat{a}_G = O(1)$, we also assume $L \gg 1$ throughout. In terms of x, the action is, to leading order in L

$$S_{\text{softmax}}(x) \approx \frac{\chi L}{2} \left(\frac{1}{\bar{g}_W} \log^2(1+x) - \frac{P}{\bar{\sigma}^2 L^2} \frac{V-1}{V} \frac{(x+1)^2 \bar{g}_O}{\bar{g}_O + \frac{\bar{\sigma}^2 L}{P} V} \right). \tag{317}$$

Defining $z = 1 + x$, which has the same minimum, the first-order derivative equation reads

$$\partial_z : \quad \frac{1}{\bar{g}_W} 2 \frac{\log(z)}{z} = 2 \frac{P}{\bar{\sigma}^2 L^2} \frac{V-1}{V} \frac{z \bar{g}_O}{\bar{g}_O + \frac{\bar{\sigma}^2 L}{P} V} \tag{318}$$

which can be simplified to

$$\frac{\log(z)}{z^2} = \underbrace{\frac{P}{\bar{\sigma}^2 L^2} \frac{V-1}{V} \frac{\bar{g}_O \bar{g}_W}{\bar{g}_O + \frac{\bar{\sigma}^2 L}{P} V}}_{=:Q}, \tag{319}$$

where we also defined Q. Moving on to the second-order derivative equation, we have

$$\partial_z^2: \quad \frac{1}{z^2} - \frac{\log(z)}{z^2} = Q \tag{320}$$

Comparing the two eliminates Q, which lets us solve for the location of the spinodal, while eliminating the dependence on P. The solution for z is

$$z_* = \sqrt{e} \tag{321}$$

with e the mathematical constant. Plugging this solution for z into the first derivative equation, we find that the solution for Q is

$$Q_* = \frac{1}{2e} \tag{322}$$

Isolating P from within Q, we find the metastable branch ends at

$$P_{\text{anni}} = L^2 \frac{V}{V-1} \frac{1}{4e\bar{g}_W} \bar{\sigma}^2 \left(1 + \sqrt{1 + 8e\left(\frac{V-1}{L}\right)\frac{\bar{g}_W}{\bar{g}_O}}\right) \approx L^2 \frac{V}{V-1} \frac{1}{2e\bar{g}_W} \bar{\sigma}^2 \sim L^2 \tag{323}$$

A factor of $L/\ln^2(L)$ after the phase transition.

I.2 Copy Behavior Appears as a Metastable State

Here we look for when the copy minimum first appears, at first as a metastable state. In this case we expect $\hat{a}_G = O(1)$ so to leading order in L we have

$$S_{\text{softmax}}(\hat{a}_G) \approx \frac{\chi L}{2} \left(\frac{1}{\bar{g}_W} \log^2\left(\sqrt{L}\hat{a}_G\right) - \frac{P}{\bar{\sigma}^2 L} \frac{V-1}{V} \frac{\hat{a}_G^2 \bar{g}_O}{(\hat{a}_G^2+1)\bar{g}_O + \frac{\bar{\sigma}^2 L}{P} V} \right) \tag{324}$$

The first order derivative equation $\partial_{\hat{a}_G} S \stackrel{!}{=} 0$ reads

$$\frac{\log\left(\sqrt{L}\hat{a}_G\right)}{\hat{a}_G} = \frac{P\bar{g}_W \bar{g}_O}{\bar{\sigma}^2 L} \frac{V-1}{V} \left(\frac{\hat{a}_G}{(\hat{a}_G^2+1)\bar{g}_O + \frac{\bar{\sigma}^2 L}{P} V} - \frac{\hat{a}_G^3}{\left(\hat{a}_G^2 + \bar{g}_O + \frac{\bar{\sigma}^2 L}{P} V\right)^2} \right), \tag{325}$$

using the definition $A := \bar{g}_O + \frac{\bar{\sigma}^2 L}{P} V$ we can write it as

$$\frac{\log\left(\sqrt{L}\hat{a}_G\right)}{\hat{a}_G} = \frac{P\bar{g}_W}{\bar{\sigma}^2 L} \frac{V-1}{V} \frac{\bar{g}_O A \hat{a}_G}{\left(\hat{a}_G^2 \bar{g}_O + A\right)^2}. \tag{326}$$

The second order derivative equation $\partial_{\hat{a}_G}^2 S \stackrel{!}{=} 0$ reads

$$\frac{1}{\hat{a}_G^2}\left(1 - \log\left(\sqrt{L}\hat{a}_G\right)\right) = \frac{P\bar{g}_W}{\bar{\sigma}^2 L} \frac{V-1}{V} \frac{\bar{g}_O A\left(A - 3\hat{a}_G^2 \bar{g}_O\right)}{\left(\hat{a}_G^2 \bar{g}_O + A\right)^3} \tag{327}$$

dividing the two equation eliminates P and let us solve for $\hat{a}_G$

$$\hat{a}_G^2 \bar{g}_O = \left(\frac{2\log\left(\sqrt{L}\hat{a}_G\right) - 1}{2\log\left(\sqrt{L}\hat{a}_G\right) + 1} \right) A \tag{328}$$

estimating this quantity for very large L that satisfy $\ln(L) \gg 1$ we can approximate $\frac{2\log\left(\sqrt{L}\hat{a}_G\right)-1}{2\log\left(\sqrt{L}\hat{a}_G\right)+1} \approx 1$ which yields

$$\hat{a}_G^* = \sqrt{A/\bar{g}_O}. \tag{329}$$

Plugging this result into the first-order derivative equation, we get an equation for P_{start}; solving it, we find

$$P_{\text{creat}} \approx \frac{2\bar{\sigma}^2 L V \ln L}{\bar{g}_W (V-1)} \sim L \ln L, \tag{330}$$

which is $\ln(L)$ smaller than $P_{\hat{a}_G}$.

We note this result, in particular, is a rough estimate for the metastability creation event, and it does not capture the correct scaling with V; it only captures the asymptotic scaling with L. The theory's prediction is given by the spinodal conditions $\partial_{\hat{a}_G} S \stackrel{!}{=} 0$, $\partial_{\hat{a}_G}^2 S \stackrel{!}{=} 0$, which result in two scalar transcendental equations; the results in this appendix are a mere estimation of the value.

J Probability Distribution of Gaussian Matrix Product $W^{Q\mathsf{T}}W^K$

The probability distribution of the matrix product $\boldsymbol{W^G} = \boldsymbol{W}^{Q\mathsf{T}}\boldsymbol{W}^K$ in the attention layer is given by

$$p(\boldsymbol{W^G}) = \int \mathcal{D}\boldsymbol{W}^Q \int \mathcal{D}\boldsymbol{W}^K \prod_{nm} \delta\big(W^{\boldsymbol{G}}_{nm} - \sum_l W^Q_{ln} W^K_{lm}\big)\, p(\boldsymbol{W}^K)\, p(\boldsymbol{W}^Q) \tag{331}$$

$$= \int \mathcal{D}\tilde{\boldsymbol{W}}^{\boldsymbol{G}} \int \mathcal{D}\boldsymbol{W}^Q \int \mathcal{D}\boldsymbol{W}^K \exp\big(\sum_{nm} \tilde{W}^{\boldsymbol{G}}_{nm}\big(W^{\boldsymbol{G}}_{nm} - \sum_l W^Q_{ln} W^K_{lm}\big)\big)\, p(\boldsymbol{W}^K)\, p(\boldsymbol{W}^Q) \tag{332}$$

$$= \int \mathcal{D}\tilde{\boldsymbol{W}}^{\boldsymbol{G}} \exp\Big(\sum_{nm} \tilde{W}^{\boldsymbol{G}}_{nm} W^{\boldsymbol{G}}_{nm} - \ln\Big\langle \exp\Big(\sum_{nml} \tilde{W}^{\boldsymbol{G}}_{nm} W^Q_{ln} W^K_{lm}\Big)\Big\rangle_{\boldsymbol{W}^Q,\boldsymbol{W}^K}\Big), \tag{333}$$

where we rewrite the δ-function using its Fourier representation as in previous sections.

We identify the second term as the cumulant-generating function of $\boldsymbol{W^G}$ and calculate it to obtain its statistics

$$\mathcal{W}(\tilde{\boldsymbol{W}}^{\boldsymbol{G}}) = \ln\Big\langle \exp\Big(\sum_{nml} \tilde{W}^{\boldsymbol{G}}_{nm} W^Q_{ln} W^K_{lm}\Big)\Big\rangle_{\boldsymbol{W}^Q,\boldsymbol{W}^K} \tag{334}$$

$$= \ln \prod_l \Big\langle\Big\langle \exp\Big(\sum_{nm} \tilde{W}^{\boldsymbol{G}}_{nm} W^Q_{ln} W^K_{lm}\Big)\Big\rangle_{\boldsymbol{W}^K}\Big\rangle_{\boldsymbol{W}^Q} \tag{335}$$

$$= \ln \prod_l \Big\langle \exp\Big(\frac{1}{2}\frac{g_K}{d_{\text{model}}} \sum_{nmk} \tilde{W}^{\boldsymbol{G}}_{nm} \tilde{W}^{\boldsymbol{G}}_{km} W^Q_{ln} W^Q_{lk}\Big)\Big\rangle_{\boldsymbol{W}^Q} \tag{336}$$

$$= d_k \ln \Big\langle \exp\Big(\frac{1}{2}\frac{g_K}{d_{\text{model}}} \sum_{nmk} \tilde{W}^{\boldsymbol{G}}_{nm} \tilde{W}^{\boldsymbol{G}}_{km} W^Q_{n} W^Q_{k}\Big)\Big\rangle_{\boldsymbol{W}^Q}. \tag{337}$$

In the last line, we used the independence in l to pull out a factor d_k. Explicitly inserting the prior of $\boldsymbol{W}^Q$, we get

$$\mathcal{W}(\tilde{\boldsymbol{W}}^{\boldsymbol{G}}) = d_k \ln \exp\Big(-\frac{1}{2}\sum_{nm} W^Q_n \Big[-\frac{g_K}{d_{\text{model}}}\sum_m \tilde{W}^{\boldsymbol{G}}_{nm}\tilde{W}^{\boldsymbol{G}}_{nm} + \frac{d_{\text{model}}}{g_Q}\delta_{nm}\Big] W^Q_m\Big) \times \det\Big[\Big(\frac{d_{\text{model}}}{g_Q}\boldsymbol{I}\Big)^{1/2}\Big] \tag{338}$$

$$= d_k \ln\Big(\det\Big[\sum_{nm}\frac{d_{\text{model}}}{g_Q}\delta_{nm} - \frac{g_K}{d_{\text{model}}}\sum_{nm}\tilde{W}^{\boldsymbol{G}}_{nm}\tilde{W}^{\boldsymbol{G}}_{nm}\Big]^{-1/2} \det\Big[\Big(\frac{d_{\text{model}}}{g_Q}\boldsymbol{I}\Big)^{1/2}\Big]\Big) \tag{339}$$

$$= d_k \ln\Big(\det\Big[\sum_{nm}\delta_{nm} - \frac{g_K g_Q}{d^2_{\text{model}}}\sum_{nm}\tilde{W}^{\boldsymbol{G}}_{nm}\tilde{W}^{\boldsymbol{G}}_{nm}\Big]^{-1/2}\Big) \tag{340}$$

$$= -\frac{d_k}{2}\ln\det\Big[\boldsymbol{I} - \frac{g_K g_Q}{d^2_{\text{model}}}\tilde{\boldsymbol{W}}^{\boldsymbol{G}}\tilde{\boldsymbol{W}}^{\boldsymbol{G}\mathsf{T}}\Big] \tag{341}$$

Thus, up to normalization constants and prefactors,

$$I(\boldsymbol{W^G}) := -\ln p(\boldsymbol{W^G}) \tag{342}$$

$$= \sup_{\tilde{\boldsymbol{W}}^{\boldsymbol{G}}}\Big\{\mathrm{Tr}\big[\tilde{\boldsymbol{W}}^{\boldsymbol{G}\mathsf{T}}\boldsymbol{W^G}\big] - \mathcal{W}(\tilde{\boldsymbol{W}}^{\boldsymbol{G}})\Big\} \tag{343}$$

$$= \sup_{\tilde{\boldsymbol{W}}^{\boldsymbol{G}}}\Big\{\mathrm{Tr}\big[\tilde{\boldsymbol{W}}^{\boldsymbol{G}\mathsf{T}}\boldsymbol{W^G}\big] + \frac{d_k}{2}\ln\det\Big[\boldsymbol{I} - \frac{g_K g_Q}{d^2_{\text{model}}}\tilde{\boldsymbol{W}}^{\boldsymbol{G}}\tilde{\boldsymbol{W}}^{\boldsymbol{G}\mathsf{T}}\Big]\Big\}. \tag{344}$$

Since the second term depends only on the singular values of $\tilde{\boldsymbol{W}}^{\boldsymbol{G}}$, the saddle point aligns the singular vectors of $\tilde{\boldsymbol{W}}^{\boldsymbol{G}}$ with those of $\boldsymbol{W^G}$. Writing

$$\boldsymbol{W^G} = \boldsymbol{U}\mathrm{diag}(\sigma_i)\boldsymbol{V}^{\mathsf{T}}, \qquad \tilde{\boldsymbol{W}}^{\boldsymbol{G}} = \boldsymbol{U}\mathrm{diag}(\tilde{\sigma}_i)\boldsymbol{V}^{\mathsf{T}}, \tag{345}$$

we obtain

$$I(\boldsymbol{W}^{\boldsymbol{G}}) = \sup_{\tilde{\sigma}_i} \sum_i \left[\tilde{\sigma}_i \sigma_i + \frac{d_k}{2} \ln \left(1 - \frac{g_K g_Q}{d_{\text{model}}^2} \tilde{\sigma}_i^2 \right) \right]. \tag{346}$$

The saddle point satisfies

$$0 = \frac{\partial}{\partial \tilde{\sigma}_j} \left(\sum_i \tilde{\sigma}_i \sigma_i + \frac{d_k}{2} \sum_i \ln \left[1 - \frac{g_K g_Q}{d_{\text{model}}^2} \tilde{\sigma}_i^2 \right] \right) \tag{347}$$

$$= \sigma_j - d_k \frac{g_K g_Q}{d_{\text{model}}^2} \frac{\tilde{\sigma}_j}{1 - \frac{g_K g_Q}{d_{\text{model}}^2} \tilde{\sigma}_j^2}. \tag{348}$$

Equivalently,

$$\sigma_j \tilde{\sigma}_j^2 + d_k \tilde{\sigma}_j - \sigma_j \frac{d_{\text{model}}^2}{g_K g_Q} = 0. \tag{349}$$

The maximizing branch is

$$\tilde{\sigma}_j^\star = \frac{-d_k + \sqrt{d_k^2 + 4\sigma_j^2 \frac{d_{\text{model}}^2}{g_K g_Q}}}{2\sigma_j} = \sqrt{\frac{d_k^2}{4\sigma_j^2} + \frac{d_{\text{model}}^2}{g_K g_Q}} - \frac{d_k}{2\sigma_j}. \tag{350}$$

Substituting this saddle point back into the Legendre transform gives

$$I(\boldsymbol{W}^{\boldsymbol{G}}) = \frac{d_k}{2} \sum_i \left[\sqrt{1 + \frac{4 d_{\text{model}}^2 \sigma_i^2}{d_k^2 g_K g_Q}} - 1 + \ln \left(\frac{2}{1 + \sqrt{1 + \frac{4 d_{\text{model}}^2 \sigma_i^2}{d_k^2 g_K g_Q}}} \right) \right]. \tag{351}$$

At small singular values, the exact rate function expands as

$$I(\boldsymbol{W}^{\boldsymbol{G}}) = \frac{d_{\text{model}}^2}{2 d_k g_K g_Q} \sum_i \sigma_i^2 + O\left(\sum_i \sigma_i^4 \right) \approx \frac{d_{\text{model}}^2}{2 d_k g_K g_Q} \|\boldsymbol{W}^{\boldsymbol{G}}\|_F^2 \tag{352}$$

Thus, near the origin, the effective regularization induced by the Gaussian product prior is standard ℓ_2 weight decay / Gaussian prior on the merged matrix $\boldsymbol{W}^{\boldsymbol{G}}$.

At large singular values, the leading term is linear in the singular values,

$$I(\boldsymbol{W}^{\boldsymbol{G}}) = \frac{d_{\text{model}}}{\sqrt{g_K g_Q}} \sum_i |\sigma_i| + O(\ln \sigma_i) \approx \frac{d_{\text{model}}}{\sqrt{g_K g_Q}} \|\boldsymbol{W}^{\boldsymbol{G}}\|_*. \tag{353}$$

Therefore, for general d_k, the exact product prior interpolates between Frobenius-norm regularization at small singular values and nuclear-norm regularization in the tails.

The Gaussian approximation is valid when the typical singular values of $\boldsymbol{W}^{\boldsymbol{G}}$ are small compared to the crossover scale $\sigma_* = d_k \sqrt{g_K g_Q}/(2 d_{\text{model}})$ at which corrections to Eq. (352) become large. Since the entries of $\boldsymbol{W}^{\boldsymbol{G}}$ have variance $d_k g_K g_Q / d_{\text{model}}^2$ and $\boldsymbol{W}^{\boldsymbol{G}}$ is a $d_{\text{model}} \times d_{\text{model}}$ matrix, its typical singular values scale as $\sigma_{\text{typ}} \sim \sqrt{d_k g_K g_Q / d_{\text{model}}}$. The ratio to the crossover scale is

$$\frac{\sigma_{\text{typ}}}{\sigma_*} \sim 2 \sqrt{\frac{d_{\text{model}}}{d_k}}, \tag{354}$$

which is independent of $g_K g_Q$ – rescaling the variances shifts both scales identically, so they cancel. The Gaussian approximation is therefore self-consistently valid only when $d_k \gg d_{\text{model}}$, and breaks down when $d_k \approx d_{\text{model}}$, where typical singular values sit right at the crossover between the ℓ_2 and nuclear-norm regimes.

This nuclear-norm behavior aligns with recent work showing that separate key and query matrices change the implicit regularization of attention. For one-layer softmax attention, [68] showed that gradient flow with separately trained key and query matrices implicitly minimizes the nuclear norm of their product. Similarly, [69] showed that the separate $(\boldsymbol{W}^K, \boldsymbol{W}^Q)$ parametrization induces a nuclear-norm SVM over the merged matrix $\boldsymbol{W} = \boldsymbol{W}^Q \boldsymbol{W}^{K\top}$, rather than the Frobenius-norm

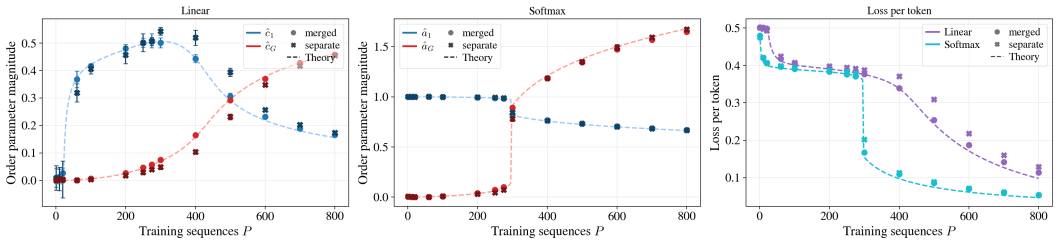

Figure 5: *Separate key and query matrices* reproduce the phase transitions of the merged model. **Left:** Scalar order parameters $\hat{c}_1, \hat{c}_G$ for linear attention, measured from models trained with a merged matrix $\boldsymbol{W}^{\boldsymbol{G}}$ (circles) and with separate matrices $\boldsymbol{W}^Q, \boldsymbol{W}^K$ with $d_k = d_{\text{model}}$ (crosses), compared to the theory predictions (dashed lines) against the number of training samples P. Both parametrizations show the second-order phase transition followed by the crossover. The agreement with the theory in the unmerged case is slightly worsened at the crossover. **Center:** Same for softmax attention. For both parametrizations, $\hat{a}_G$ jumps at the same $P \approx 300$, marking the first-order phase transition. **Right:** Loss per token for linear (purple) and softmax (cyan) attention.

objective obtained by optimizing $\boldsymbol{W}$ directly. Relatedly, [20] analyzed the in-context regression setting and compared the saddle-point structure of the dynamics with merged and with separate key and query matrices.

Nevertheless, many theoretical works replace the key-query product by a single merged matrix [44, 21, 42, 47]. Our results above clarify that this replacement is benign when $d_k \gg d_{\text{model}}$ and pinpoint when the effective regularization crosses over from ℓ_2 weight decay to nuclear norm. Remarkably, at the level of attention patterns rather than weights, this approximation holds when $d_{\text{model}}, d_k \gg L$ as we show in Appendix B.

We confirm this numerically by training the network of Eq. (3) with separate matrices $\boldsymbol{W}^Q, \boldsymbol{W}^K \in \mathbb{R}^{d_k \times d_{\text{model}}}$ with Gaussian i.i.d. entries in place of the merged $\boldsymbol{W}^{\boldsymbol{G}}$, with variances chosen such that the second moments of $\boldsymbol{W}^{Q\intercal}\boldsymbol{W}^K$ and $\boldsymbol{W}^{\boldsymbol{G}}$ coincide, $d_k = d_{\text{model}} = 10000$, and all other settings as in Appendix R. Since $d_k = d_{\text{model}}$ (the case treated in Appendix B), the prior on $\boldsymbol{W}^{Q\intercal}\boldsymbol{W}^K$ is non-Gaussian (see above), whereas $d_k, d_{\text{model}} \gg L$ ensures a Gaussian prior on the attention pattern. The order parameters are extracted from the trained attention pattern exactly as for the merged model and compared to the unchanged theory.

Figure 5 shows that both parametrizations display the same phenomenology: the second-order transition followed by the crossover for linear attention, the first-order transition at $P \approx 300$ for softmax attention, and the corresponding loss signatures. This confirms that the relevant condition for our theory is $d_k, d_{\text{model}} \gg L$ rather than $d_k \gg d_{\text{model}}$: the non-Gaussianity of the weight prior does not propagate to the attention pattern. The only systematic deviation is a slightly delayed growth of $\hat{c}_G$ with separate key and query matrices at the onset of the crossover.

K Extension to multiple attention layers

To extend the theory to linear transformers with M layers, each with a single attention head, we generalize the single layer case as

$$h_i^{(\ell+1)a}(x) = G^{(\ell)ab}\, B_{ij}^{(\ell)} h_j^{(\ell)b}, \quad 1 \le \ell \le M-1\,, \tag{355}$$

$$G^{(\ell)ab} = \frac{1}{L\sqrt{d}}\, h_i^{(\ell)a}\, W_{ij}^{(\ell)}\, h_j^{(\ell)b}\,, \tag{356}$$

together with the convention $h^{(0)} = x$ and $h^{(M+1)} = f$. We here again combine the key and the query matrix into a single matrix W with prior $W_{ij}^{(\ell)} \overset{\text{i.i.d.}}{\sim} \mathcal{N}\big(0, g_W\, d^{-1}\big)$ and denote the output matrix W^O of each layer as $B_{ij}^{(\ell)} \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, g_B\, d^{-1})$ here. We also use the shorthand $d = d_{\text{model}}$ in the following.

As the priors on $W^{(\ell)}$ and $B^{(\ell)}$ are independent across different layers ℓ, we may consider one layer at a time. Performing the analogous steps as in App. B, we get by enforcing the definition of the attention layer (356)

$$\begin{aligned}
&\int d\tilde{G}^{(\ell)} \Big\langle \exp\big(-\mathrm{tr}\tilde{G}^{(\ell)\mathrm{T}} G^{(\ell)} + \tilde{G}^{(\ell)ab}_{\alpha} \frac{1}{L\sqrt{d}}\, h^{(\ell)a}_{\alpha i}\, W^{(\ell)}_{ij}\, h^{(\ell)b}_{\alpha j}\big) \Big\rangle_{W^{(\ell)}_{ij} \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, g_W\, 1/d)} \\
&= \int d\tilde{G}^{(\ell)}\, \exp\big(-\mathrm{tr}\tilde{G}^{(\ell)\mathrm{T}} G^{(\ell)} + \tilde{G}^{(\ell)ab}_{\alpha} \tilde{G}^{(\ell)a'b'}_{\beta}\, \frac{g_W}{2L^2}\, C^{(\ell)aa'}_{\alpha\beta}\, C^{(\ell)bb'}_{\alpha\beta}\big)\,,
\end{aligned}$$

where we defined

$$C^{(\ell)ab}_{\alpha\beta} = d^{-1}\, h^{(\ell)a}_{\alpha i} h^{(\ell)b}_{\beta i}\,.$$

Enforcing the equation (355) which projects the activity from one layer to the next, we likewise have

$$\begin{aligned}
&\int d\tilde{h}^{(\ell+1)} \Big\langle \exp\big(-\mathrm{tr}\tilde{h}^{(\ell+1)\mathrm{T}} h^{(\ell+1)} + \tilde{h}^{(\ell+1)a}_{\alpha i} G^{(\ell)ab}_{\alpha}\, B^{(\ell)}_{ij} h^{(\ell)b}_{\alpha j}\big) \Big\rangle_{B^{(\ell)}_{ij} \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, g_B\, d^{-1})} \\
&= \int d\tilde{h}^{(\ell+1)} \exp\big(-\mathrm{tr}\tilde{h}^{(\ell+1)\mathrm{T}} h^{(\ell+1)} + \frac{1}{2} g_B\, \tilde{h}^{(\ell+1)a}_{\alpha i} \tilde{h}^{(\ell+1)a'}_{\beta i} G^{(\ell)ab}_{\alpha} G^{(\ell)a'b'}_{\beta}\, C^{(\ell)bb'}_{\alpha\beta}\big)\,,
\end{aligned}$$

where again the attention-weighted input kernel appears in each layer ℓ

$$\mathcal{K}^{(\ell)aa'}_{\alpha\beta} := G^{(\ell)ab}_{\alpha} G^{(\ell)a'b'}_{\beta}\, C^{(\ell)bb'}_{\alpha\beta}\,.$$

The appearance of $\sum_i \tilde{h}^{(\ell+1)a}_{\alpha i} \tilde{h}^{(\ell+1)a'}_{\beta i}$ shows that, given $G^{(\ell)}$ and $C^{(\ell)}$, the distribution of the pre-activations $h_{\alpha i}$ factorizes over i, so that the kernel C is expected to concentrate also at intermediate layers. Its definition is enforced by

$$C^{(0)ab}_{\alpha\beta} = x^{a}_{\alpha i} x^{b}_{\beta i} + g_p\, \delta_{ab},,$$

$$\begin{aligned}
&p(C^{(\ell)}|C^{(\ell-1)}, G^{(\ell-1)}) \\
&= \int d\tilde{C}^{(\ell)} \Big\langle \exp\big(-\mathrm{tr}\tilde{C}^{(\ell)\mathrm{T}} C^{(\ell)} + \tilde{C}^{(\ell)ab}_{\alpha\beta}\, d^{-1}\, h^{(\ell)a}_{\alpha i} h^{(\ell)b}_{\beta i}\big) \Big\rangle_{h^{(\ell)a}_{\alpha i} \overset{\text{i.i.d. in } i}{\sim} \mathcal{N}\left(0, g_B \mathcal{K}^{(\ell-1)ab}_{\alpha\beta}\right)} \\
&= \int d\tilde{C}^{(\ell)} \exp\Big(-\mathrm{tr}\tilde{C}^{(\ell)\mathrm{T}} C^{(\ell)} + d\, \ln \Big\langle \exp\big(\tilde{C}^{(\ell)ab}_{\alpha\beta}\, d^{-1}\, h^{(\ell)a}_{\alpha} h^{(\ell)b}_{\beta}\big) \Big\rangle_{h^{(\ell)a}_{\alpha} \sim \mathcal{N}\left(0, g_B \mathcal{K}^{(\ell-1)ab}_{\alpha\beta}\right)} \Big)\,,
\end{aligned}$$

where we exploited the independence over i in the last step. The appearance of the scaling form $d\, \ln\langle \exp\big(d^{-1}\, \tilde{C} \ldots\big)\rangle$ of the cumulant-generating function shows that C concentrates. In the limit of large d, we may thus approximate its distribution with the rate function

$$-\ln p(C^{(\ell)}|C^{(\ell-1)})/d \simeq \sup_{\tilde{C}^{(\ell)}} \mathrm{tr}\tilde{C}^{(\ell)\mathrm{T}} C^{(\ell)} - d\, W_C\big(\tilde{C}^{(\ell)}|\mathcal{K}^{(\ell-1)}\big)\,, \tag{357}$$

$$W_C(\tilde{C}) := \ln \Big\langle \exp\big(\tilde{C}^{(\ell)ab}_{\alpha\beta}\, d^{-1}\, h^{(\ell)a}_{\alpha} h^{(\ell)b}_{\beta}\big) \Big\rangle_{h^{(\ell)a}_{\alpha} \sim \mathcal{N}\left(0, g_B \mathcal{K}^{(\ell-1)ab}_{\alpha\beta}\right)}\,. \tag{358}$$

Evaluating the supremum condition yields

$$0 = C^{(\ell)ab}_{\alpha\beta} - \langle h^{(\ell)a}_{\alpha} h^{(\ell)b}_{\beta} \rangle\,,$$

which shows that the second order statistics of h is fixed by $C^{(\ell)}$, the argument of the left hand side of (357). This implies that the quadratic form A in the exponent $-\frac{1}{2} h^{\mathrm{T}} A h$ appearing in (358) satisfies

$$\big[C^{(\ell)}\big]^{-1} = A = -\,2\, d^{-1}\, \tilde{C}^{(\ell)ab}_{\alpha\beta} + \big[g_B \mathcal{K}^{(\ell-1)}\big]^{-1}\,, \tag{359}$$

so that

$$\tilde{C}^{(\ell)ab}_{\alpha\beta} = \frac{d}{2} \left(\big[g_B \mathcal{K}^{(\ell-1)}\big]^{-1} - \big[C^{(\ell)}\big]^{-1}\right).$$

Likewise it implies that the Gaussian integral over h in W_C (358) yields $\frac{1}{2}\ln\det\left(C^{(\ell)}\right)$ and has the normalization constant $-\frac{1}{2}\ln\det\left(g_B\mathcal{K}^{(\ell-1)}\right)$, so that together we get

$$-\ln p(C^{(\ell)}|C^{(\ell-1)})/d \simeq \frac{1}{2}\mathrm{tr}\left[g_B\mathcal{K}^{(\ell-1)}\right]^{-1}C^{(\ell)} - \frac{1}{2}\ln\det\left(C^{(\ell)}\right) + \frac{1}{2}\ln\det\left(g_B\mathcal{K}^{(\ell-1)}\right) + \text{const.}\,, \tag{360}$$

$$\equiv \mathrm{KL}\big(\mathcal{N}(0,C^{(\ell)})\|\mathcal{N}(0,g_B\mathcal{K}^{(\ell-1)})\big)\,, \tag{361}$$

where we dropped the constant term $-\frac{1}{2}\mathrm{tr}\big[{C^{(\ell)}}^{-1}C^{(\ell)}\big] = -\frac{PL}{2}$. The expression (360) is that of a KL divergence between two Gaussian distributions with covariance matrices $g_B\mathcal{K}^{(\ell-1)}$ and $C^{(\ell)}$. Since the KL divergence is minimal for $C^{(\ell)} = g_B\mathcal{K}^{(\ell-1)}$, this shows that this term tries to keep these two kernels close; the point where the distance vanishes is the neural network Gaussian process solution. So together with the likelihood of the labels as $\mathcal{N}(y|f,\sigma^2)$, we have

$$-\ln\, p(y|C^{(0)}) + \text{const.} \simeq \inf_G \inf_C \tag{362}$$

$$d\sum_{\ell=1}^{M-1}\frac{1}{2}C^{(\ell)}\left[g_B\mathcal{K}^{(\ell-1)}\right]^{-1} - \frac{1}{2}\ln\det\left(C^{(\ell)}\right) + \frac{1}{2}\ln\det\left(g_B\mathcal{K}^{(\ell-1)}\right)$$

$$+\sup_{\tilde G}\sum_{\ell=1}^{M-1}\mathrm{tr}\tilde G^{(\ell)\mathrm{T}}G^{(\ell)} - \tilde G^{(\ell)ab}_{\alpha}\tilde G^{(\ell)a'b'}_{\beta}\,\frac{g_W}{2L^2}\,C^{(\ell)aa'}_{\alpha\beta}\,C^{(\ell)bb'}_{\alpha\beta} \tag{363}$$

$$+\sup_{\tilde f}\left\{\tilde f^{\mathrm{T}}y - \frac{\sigma^2}{2}\sum_i \tilde f^a_{\alpha i}\tilde f^a_{\alpha i} + \frac{g_B}{2}\sum_i \tilde f^a_{\alpha i}\tilde f^b_{\beta i}\,\mathcal{K}^{(M)\,ab}_{\alpha\beta}\right\} \tag{364}$$

A more compact form is achieved by performing the Gaussian integrals over $\tilde G$ and $\tilde f$ instead of taking their supremum; this yields a slightly better approximation, since a normalization determinant is contained and yields

$$-\ln\, p(y|C^{(0)}) = \inf_G \inf_C \tag{365}$$

$$d\sum_{\ell=1}^{M-1}\mathrm{KL}\big(\mathcal{N}(0,C^{(\ell)})\|\mathcal{N}(0,g_B\mathcal{K}^{(\ell-1)})\big)$$

$$-\sum_{\ell=0}^{M-1}\ln\mathcal{N}(G^{(\ell)}|0,g_W L^{-2}\,C^{(\ell)}C^{(\ell)})$$

$$-\ln\mathcal{N}(y|0,g_B\mathcal{K}^{(M)}+\sigma^2\mathrm{I})$$

$$\left[C^{(\ell)}C^{(\ell)}\right]_{(\alpha ab),(\beta a'b')} := C^{(\ell)aa'}_{\alpha\beta}\,C^{(\ell)bb'}_{\alpha\beta}\,,$$

$$\mathcal{K}^{(\ell)aa'}_{\alpha\beta} := G^{(\ell)ab}_{\alpha}G^{(\ell)a'b'}_{\beta}\,C^{(\ell)bb'}_{\alpha\beta}\,.$$

This expression shows the different tendencies of the network:

- The first line aims to minimize the KL-divergence between the attention-weighted Gaussian kernel matrix $g_B\mathcal{K}^{(\ell-1)}$ of the previous layer and the Gaussian kernel matrix of the next layer $C^{(\ell)}$. The KL-divergence measures the information loss, when approximating the Gaussian with $C^{(\ell)}$ by the Gaussian with $g_B K^{(\ell-1)}$.
- The second line is minimal, if the attention matrix G is most likely under the Gaussian with covariance $g_W L^{-2}C^{(\ell)}C^{(\ell)}$; the structure of the indices shows that the attention patterns G^{ab}_{α} and $G^{a'b'}_{\beta}$ tend to be correlated if the sequences α and β of the previous layer are similar at positions a and a' and at positions b and b'. The sign of the resulting correlation depends on the product of the signs of the correlations in each factor $C^{(\ell)}$. This term shows the tendency of the attention patterns to focus attention on positions with similar entries.

- The third line is minimal, if the target y is most likely under the Gaussian with covariance $\mathcal{K}^{(M)} + \sigma^2 \mathrm{I}$.

The state after learning is a compromise between these three forces.

L Extension to Markov transition statistics

The framework of the main text and the preceding appendices is not tied to the copy task. Since the prior over attention patterns depends only on the inputs, the task enters solely through the data term of the action (10). As an example beyond copying, we consider labels generated by a Markov chain.

Inputs are drawn as in Sec. 2, while the label y^a_α at position a is generated from the preceding token x^{a-1}_α by a transition matrix $T \in \mathbb{R}^{V\times V}$,

$$\mathbb{E}\big[y^a_{\alpha i} \,\big|\, x_\alpha\big] = \sum_{j=1}^{V} T_{ij}\, x^{a-1}_{\alpha j}\,, \tag{366}$$

with periodic boundary conditions, where T_{ij} is the probability of the label token i given $x^{a-1}_\alpha = j$. We take T to be doubly stochastic, $\sum_i T_{ij} = \sum_j T_{ij} = 1$, so that the stationary distribution is uniform, and decompose

$$T = \frac{J_V}{V} + \tilde{T}\,, \qquad \sum_i \tilde{T}_{ij} = \sum_j \tilde{T}_{ij} = 0\,, \qquad \|T\|_F^2 = 1 + \|\tilde{T}\|_F^2\,. \tag{367}$$

The copy task corresponds to $T = \mathrm{Id}_V$, for which $\|\tilde{T}\|_F^2 = V - 1$, whereas for $T = J_V/V$ the labels are independent of the context and $\tilde{T} = 0$. Since T acts on token indices only, the labels transform under permutations of positions exactly as in the copy task, so the symmetry argument of App. E and the ansatz (11) carry over unchanged.

For simplicity, we only consider the label following the last token, with attention on the preceding position as in the copy pattern. The network output is then a linear readout of the last token, $f_{\alpha i} = \sum_j B_{ij}\, x^L_{\alpha j}$, and the prior $B_{ij} \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, g_B)$ induces

$$f_{\cdot i} \overset{\text{i.i.d. in } i}{\sim} \mathcal{N}\big(0,\, g_B\, C\big)\,, \qquad C_{\alpha\beta} = \sum_{j=1}^{V} x^L_{\alpha j}\, x^L_{\beta j}\,. \tag{368}$$

With Gaussian observation noise of variance σ^2, the log-likelihood of the labels is

$$\begin{aligned}
\ln p(y^{L+1}) &= \ln \int \mathcal{D}f\, \exp\Big(-\frac{1}{2\sigma^2}\sum_{\alpha,i}\big(y^{L+1}_{\alpha i} - f_{\alpha i}\big)^2\Big)\, \mathcal{N}\big(f \,\big|\, 0, g_B C\big) \qquad (369)\\
&= \ln \int \mathcal{D}f\, \mathcal{D}\tilde{f}\, \exp\Big(-\frac{1}{2\sigma^2}\sum_{\alpha,i}\big(y^{L+1}_{\alpha i} - f_{\alpha i}\big)^2 - \sum_{\alpha,i}\tilde{f}_{\alpha i} f_{\alpha i} + \frac{g_B}{2}\sum_{\alpha,\beta,i}\tilde{f}_{\alpha i}\, C_{\alpha\beta}\, \tilde{f}_{\beta i}\Big)\\
&= \ln \int \mathcal{D}\tilde{f}\, \exp\big(S(\tilde{f})\big) + \text{const.}\,,
\end{aligned}$$

where $\tilde{f}$ is integrated along the imaginary axis and the action reads

$$S(\tilde{f}) = -\sum_{\alpha,i}\tilde{f}_{\alpha i}\, y^{L+1}_{\alpha i} + \frac{\sigma^2}{2}\sum_{\alpha,i}\tilde{f}^2_{\alpha i} + \frac{g_B}{2}\sum_{\alpha,\beta,i}\tilde{f}_{\alpha i}\, C_{\alpha\beta}\, \tilde{f}_{\beta i}\,. \tag{370}$$

In the EK approximation, we replace the action by its average over the data. As usual, the conjugate field then becomes a function of the input, $\tilde{f}_\alpha \to \tilde{f}(x^L_\alpha)$, so that samples are distinguished only by their last token. For discrete tokens, we implement this by sorting the samples into the blocks

$$\Sigma(v) = \big\{\alpha \in \{1,\dots,P\} \,\big|\, x^L_\alpha = v\big\}\,, \qquad 1 \le v \le V\,, \tag{371}$$

and ordering them block by block, which leaves the likelihood unchanged. In the sorted order, $x^L_{\cdot j}$ is the indicator of the j-th block, and the inputs enter only through the block sizes $|\Sigma(v)|$, whose relative fluctuations are of order $\sqrt{V/P}$. Replacing them by their mean P/V, we obtain

$$\langle x^L_{\cdot j}\rangle_{x^L} = e_j \otimes \mathbf{1}\,, \qquad \langle C\rangle_{x^L} = \mathrm{Id}_V \otimes \mathbf{1}\mathbf{1}^{\mathrm{T}}\,, \tag{372}$$

with $\mathbf{1} \in \mathbb{R}^{P/V}$ the all-ones vector and $e_j \in \mathbb{R}^V$ the j-th unit vector.

In contrast to the average in the original order, $\langle x^L_{\alpha j}\rangle = 1/V$ and $\langle C_{\alpha\beta}\rangle = \delta_{\alpha\beta} + (1-\delta_{\alpha\beta})/V$, (372) contains no factor $1/V$. The two are related by a uniformly random permutation, which puts a given sample into block j with probability $1/V$ and two distinct samples into the same block with probability $(P/V-1)/(P-1) \simeq 1/V$. This permutation leaves the exact likelihood unchanged, but not its data average at fixed $\tilde{f}$, because the sorting depends on the inputs. Averaging in the original order would retain only the marginal label distribution, $\langle y^{L+1}_{\alpha i}\rangle = \sum_j T_{ij}/V = 1/V$, and lose its dependence on the preceding token, so that only the uniform part J_V/V of T would remain in the action.

Averaging (366) over the labels at fixed sorted inputs, we find that the mean labels are constant within each block,

$$\big\langle y^{L+1}_{\cdot i}\big\rangle_{x^L,\, y^{L+1}} = \sum_{j=1}^{V} T_{ij}\,\big(e_j \otimes \mathbf{1}\big) = T^{\mathrm{T}}_{i\cdot} \otimes \mathbf{1}\,, \tag{373}$$

The averaged action is thus invariant under permutations of the samples within each block. Since its saddle point is unique for $\sigma^2 > 0$, the conjugate field is indeed constant within each block,

$$\tilde{f}_{\cdot i} = \hat{\tilde{f}}_{\cdot i} \otimes \mathbf{1}\,. \tag{374}$$

We insert (372)–(374) into (370) and use the mixed-product rule $(A\otimes B)(A'\otimes B') = AA'\otimes BB'$. Each term then splits into a contraction over the token index and a factor $\mathbf{1}^{\mathrm{T}}\mathbf{1} = P/V$ from the sum over the samples within a block,

$$\begin{aligned}
-\sum_i \tilde{f}^{\mathrm{T}}_{\cdot i}\big\langle y^{L+1}_{\cdot i}\big\rangle_{x^L,\, y^{L+1}} &= -\sum_i \big(\hat{\tilde{f}}_{\cdot i}\otimes\mathbf{1}\big)^{\mathrm{T}}\big(T^{\mathrm{T}}_{i\cdot}\otimes\mathbf{1}\big) \\
&= -\sum_i \hat{\tilde{f}}^{\mathrm{T}}_{\cdot i} T^{\mathrm{T}}_{i\cdot}\;\mathbf{1}^{\mathrm{T}}\mathbf{1} = -\frac{P}{V}\sum_{v,i}\hat{\tilde{f}}_{vi}\,T_{iv}\,, \\
\frac{\sigma^2}{2}\sum_i \tilde{f}^{\mathrm{T}}_{\cdot i}\tilde{f}_{\cdot i} &= \frac{\sigma^2}{2}\sum_i \big(\hat{\tilde{f}}_{\cdot i}\otimes\mathbf{1}\big)^{\mathrm{T}}\big(\hat{\tilde{f}}_{\cdot i}\otimes\mathbf{1}\big) \\
&= \frac{\sigma^2}{2}\sum_i \hat{\tilde{f}}^{\mathrm{T}}_{\cdot i}\hat{\tilde{f}}_{\cdot i}\;\mathbf{1}^{\mathrm{T}}\mathbf{1} = \frac{\sigma^2}{2}\,\frac{P}{V}\sum_{v,i}\hat{\tilde{f}}^2_{vi}\,, \\
\frac{g_B}{2}\sum_i \tilde{f}^{\mathrm{T}}_{\cdot i}\langle C\rangle_{x^L}\,\tilde{f}_{\cdot i} &= \frac{g_B}{2}\sum_i \big(\hat{\tilde{f}}_{\cdot i}\otimes\mathbf{1}\big)^{\mathrm{T}}\big(\mathrm{Id}_V\otimes\mathbf{1}\mathbf{1}^{\mathrm{T}}\big)\big(\hat{\tilde{f}}_{\cdot i}\otimes\mathbf{1}\big) \\
&= \frac{g_B}{2}\sum_i \hat{\tilde{f}}^{\mathrm{T}}_{\cdot i}\hat{\tilde{f}}_{\cdot i}\;\big(\mathbf{1}^{\mathrm{T}}\mathbf{1}\big)^2 = \frac{g_B}{2}\Big(\frac{P}{V}\Big)^2\sum_{v,i}\hat{\tilde{f}}^2_{vi}\,.
\end{aligned}$$

Collecting the three terms, the averaged action reads

$$\big\langle S(\hat{\tilde{f}})\big\rangle_{x^L,\, y^{L+1}} = \frac{P}{V}\Big\{-\sum_{v,i=1}^{V}\hat{\tilde{f}}_{vi}\,T_{iv} + \frac{1}{2}\Big(g_B\frac{P}{V}+\sigma^2\Big)\sum_{v,i=1}^{V}\hat{\tilde{f}}^2_{vi}\Big\}\,. \tag{375}$$

The entries $\hat{\tilde{f}}_{vi}$ decouple, each sourced by T_{iv}, and the saddle point is $\hat{\tilde{f}}^*_{vi} = T_{iv}/(g_B P/V + \sigma^2)$. The corresponding mean prediction for a sample with last token v is $\bar{f}_{vi} = g_B(P/V)\,\hat{\tilde{f}}^*_{vi}$, which approaches T_{iv} for $g_B P/V \gg \sigma^2$, so the network learns column v of the transition matrix. The data term is

$$S_{\mathrm{data}} = -\big\langle S(\hat{\tilde{f}}^*)\big\rangle_{x^L,\, y^{L+1}} = \frac{P}{2V}\,\frac{1+\|\tilde{T}\|_F^2}{g_B P/V+\sigma^2}\,. \tag{376}$$

Hence T enters only through $\|T\|_F^2 = 1 + \|\tilde{T}\|_F^2$. Since each column of T is a probability vector, $\|\tilde{T}\|_F^2$ is maximal, $V-1$, for deterministic transitions such as the copy task, and vanishes for context-independent labels. Including the positional contribution g_p adds a constant to all entries of C. This only affects the mode $\sum_v \hat{\tilde{f}}_{vi}$, which is sourced by J_V/V, and leaves the contribution of $\tilde{T}$ unchanged.

M Stochastic Gradient Langevin Dynamics

To numerically validate theoretical predictions against real networks, we draw samples from the posterior over networks conditioned on training data $X = (x_\alpha)_{\alpha=1,\dots,P}, Y = (y_\alpha)_{\alpha=1,\dots,P}$. We assume Gaussian priors on the network weights at initialization as $W_{ij}^{\text{emb}} \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, g_{\text{emb}})$, $p_i^a \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, g_p/d_{\text{model}})$, $W_{ij}^Q \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, g_Q/d_{\text{model}})$, $W_{ij}^K \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, g_K/d_{\text{model}})$, $W_{ij}^O \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, g_O/d_{\text{model}})$ with d_{model} being the model dimension. We set $g_{\text{emb}} = g_P = 1$. Further, we assume Gaussian readout noise $\boldsymbol{y}_\alpha = \boldsymbol{f}_\alpha + \boldsymbol{\xi}$ with $\boldsymbol{\xi} \sim \mathcal{N}(0, \sigma^2 \boldsymbol{I})$.

Sampling from the Bayesian posteriors then corresponds to training the network via stochastic gradient Langevin dynamics (SGLD) [54]. Following [55], the parameters Θ are evolved according to the stochastic differential equation

$$\begin{aligned} \partial_t \Theta(t) &= -\gamma \Theta(t) - \nabla_\Theta \mathcal{L}(\Theta(t); Y) + \sqrt{2T}\zeta(t), \\ \langle \zeta_i(t)\zeta_j(s)\rangle &= \delta_{ij}\delta(t-s), \end{aligned} \tag{377}$$

where the squared error loss is $\mathcal{L}(\Theta; Y) = \sum_{\alpha=1}^P (f_\alpha(\Theta) - y_\alpha)^2$ and $f_\alpha(\Theta)$ denotes the network output for input α. At large times t, this dynamics converges to the equilibrium distribution over Θ, given by

$$\lim_{t\to\infty} p\left(\Theta(t)\right) \sim \exp\left(-\frac{\gamma}{2T}\|\Theta\|^2 - \frac{1}{T}\mathcal{L}(\Theta; Y)\right). \tag{378}$$

This stationary distribution follows from the Fokker-Planck equation [70] governing the evolution of the density of Θ. The corresponding marginal density over network outputs takes the form

$$\begin{aligned} p(Y|X) &\propto \int d\Theta \, \exp\big(-\frac{\gamma}{2T}\,\|\Theta\|^2 - \frac{1}{T}\,\|f - Y\|^2\big) \\ &\propto \Big\langle \exp\big(-\frac{1}{T}\,\|f - Y\|^2\big)\Big\rangle_{\Theta_k \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, T/\gamma)} \\ &\propto \mathcal{N}(Y|f, T/2)\,\langle \delta\big[f - f(\Theta)\big]\rangle_{\Theta_k \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0,T/\gamma)}, \end{aligned} \tag{379}$$

which, upon identifying $p(f|X) \equiv \langle \delta\big[f - f(\Theta)\big]\rangle_{\Theta_k \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0,T/\gamma)}$, coincides with the network prior under the identifications $\sigma^2 = T/2$ for the regularization noise and $T/\gamma = g/d$ for the variance of parameter Θ_k. In practice, realizing distinct variances per parameter requires assigning a separate weight decay γ to each parameter. The time-discretized form of (377) reads

$$\begin{aligned} \Theta_t &= \Theta_{t-1} - \eta\left(\gamma\Theta_{t-1} + \nabla_\Theta \mathcal{L}(\Theta_{t-1}; Y)\right) + \sqrt{2T\eta}\,\zeta_t, \\ \langle \zeta_t \zeta_s\rangle &= \delta_{ts}, \end{aligned} \tag{380}$$

with ζ_t drawn from a standard normal distribution and η a finite step size serving as the learning rate. Faithful discretization of the continuous-time dynamics in (377) requires η to remain small. In this form, SGLD is equivalent to full-batch gradient descent augmented with i.i.d. standard normal noise and weight decay regularization [71]. The parameter σ^2, which enters the main text as the regularization term on the diagonal of the output kernel, controls the relative weight of the prior against the likelihood. Large σ^2 implies large $T = 2\sigma^2$, amplifying the stochastic noise in SGLD and thereby upweighting the parameter prior. In contrast, small σ^2 suppresses the noise and directs the optimization towards minimizing the loss of training.

The rescaling by the factor χ used in the main text in Section 3.2 for large χ corresponds to training with gradient flow. To see this, note that the prior variance of the parameters θ is given by $g = T/\gamma$. Rescaling the regularization noise $\sigma^2 = T/2$ as T/χ and simultaneously rescaling g/χ therefore downscales the noise in the gradient update (377) but maintains the same weight decay γ. In the limit

of large χ the update equation (377) thus tends to gradient flow without noise but with regularization by weight decay.

A subtlety arises when taking the limit of Langevin noise to zero, which concerns the stationary distribution, because the limit $\lim_{\chi\to\infty}$ and the limit $\lim_{t\to\infty}$ do not necessarily commute: The stationary state reached with gradient flow corresponds to $\lim_{t\to\infty}\lim_{\chi\to\infty}$, while considering the (stationary) Bayesian posterior distribution in the limit $\chi\to\infty$, corresponds to $\lim_{\chi\to\infty}\lim_{t\to\infty}$. The two limits may not commute in cases in which ergodicity is broken in gradient flow; that is, if the training dynamics gets stuck within a local minimum. In the Bayesian setting this cannot happen as long as $T>0$, or correspondingly, $\chi<\infty$. The reason why training with Adam and our theory obtained from the limit $\lim_{\chi\to\infty}\lim_{t\to\infty}$ still agree relies on our training procedure, which filters training runs that got stuck at sub-optimal minima, as described in App. R.1. Thus we avoid this form of ergodicity breaking in the noiseless training procedure, which is equivalent to having an infinitesimal noise in the Bayesian posterior.

N Field-Theoretic Toy Model of the Avoided Second Phase Transition

It is useful to formulate a minimal Landau theory that captures the same qualitative phenomenon as in the main text: one genuine symmetry-breaking transition, followed by a second lower scale at which the subleading structure changes rapidly without an additional thermodynamic singularity.

Consider two real scalar fields ϕ_1,ϕ_2 with Euclidean action

$$S_{\text{toy}}[\phi_1,\phi_2]=\frac{r_1(T)}{2}\phi_1^2+\frac{r_2(T)}{2}\phi_2^2-J\,\phi_1\phi_2+\frac{u_1}{4}\phi_1^4+\frac{u_2}{4}\phi_2^4, \tag{381}$$

with $u_1,u_2>0$ for stability. The bilinear coupling J breaks the independent sign-flip symmetry $\mathbb{Z}_2\times\mathbb{Z}_2$ down to the diagonal $\mathbb{Z}_2$ symmetry $(\phi_1,\phi_2)\mapsto(-\phi_1,-\phi_2)$. This symmetry already suggests that for $J\neq 0$ there is generically only *one* true symmetry-breaking transition.

To make contact with a control parameter such as temperature, we take

$$r_i(T)=a_i(T-T_i^0),\qquad a_i>0. \tag{382}$$

If the fields were decoupled, each would become critical at its own bare scale T_i^0. For the analogy to the main text we take $T_1^0>T_2^0$, so that ϕ_1 is the mode that would order first upon cooling, while ϕ_2 becomes soft only at a lower scale.

One true phase transition. Around the symmetric saddle $\phi_1=\phi_2=0$, the quadratic form is

$$\mathcal{M}(T)=\begin{pmatrix} r_1(T) & -J\\ -J & r_2(T)\end{pmatrix}, \tag{383}$$

with eigenvalues

$$m_\pm^2(T)=\frac{r_1(T)+r_2(T)}{2}\pm\frac{1}{2}\sqrt{(r_1(T)-r_2(T))^2+4J^2}. \tag{384}$$

The disordered phase loses stability when the lower eigenvalue m_-^2 vanishes, equivalently when

$$r_1(T_c)\,r_2(T_c)=J^2. \tag{385}$$

Thus the transition is controlled by the mixed normal mode rather than by either field separately. For weak mixing and well-separated bare scales, one finds

$$T_c\simeq T_1^0+\frac{J^2}{a_1\,r_2(T_1^0)}, \tag{386}$$

so the true transition occurs close to the higher bare scale T_1^0.

The mean-field equations in the ordered phase are

$$r_1(T)\,\bar\phi_1-J\,\bar\phi_2+u_1\bar\phi_1^3=0, \tag{387}$$

$$r_2(T)\,\bar\phi_2-J\,\bar\phi_1+u_2\bar\phi_2^3=0. \tag{388}$$

Because $J \neq 0$, once one component condenses the other is immediately induced. Near the transition, where $\bar{\phi}_2$ is still small, Eq. (388) gives

$$\bar{\phi}_2 \simeq \frac{J\,\bar{\phi}_1}{r_2(T)}. \tag{389}$$

So the lower-scale field is not waiting for an independent symmetry-breaking transition; it is sourced explicitly by the condensate of the first field.

A second lower scale without a second singularity. Even though there is only one exact transition, there is generally still a second physically important scale. As the control parameter is varied deeper into the ordered phase, the effective mass $r_2(T)$ in (389) can become small near the lower bare scale T_2^0. When that happens, $\bar{\phi}_2$ grows rapidly and the composition of the ordered state reorganizes sharply.

The fluctuations around the ordered state are controlled by the Hessian

$$H(\bar{\phi}_1, \bar{\phi}_2) = \begin{pmatrix} r_1 + 3u_1\bar{\phi}_1^2 & -J \\ -J & r_2 + 3u_2\bar{\phi}_2^2 \end{pmatrix}. \tag{390}$$

At the true critical point, the smaller eigenvalue of H vanishes. The other eigenvalue, however, stays strictly positive for generic $J \neq 0$. This means that the second mode can become *soft* but not truly critical. Correspondingly, the associated susceptibility and correlation length,

$$\chi_{\rm soft} \sim \lambda_{\rm soft}^{-1}, \qquad \xi_{\rm soft} \sim \lambda_{\rm soft}^{-1/2}, \tag{391}$$

can become large but remain finite when the second eigenvalue $\lambda_{\rm soft}$ develops a pronounced minimum.

This is the standard avoided-critical-point scenario: for $J = 0$, the lower-scale field could undergo its own independent transition, but any nonzero J rounds this into a crossover. Thermodynamically there is only one nonanalyticity, yet dynamically and phenomenologically the second scale can look highly transition-like, with large fluctuations, long relaxation times, and a rapid change of the order-parameter composition.

Relation to the main text. This toy model provides a useful analogy for the behavior of the two attention order parameters $(\hat{c}_1, \hat{c}_G)$ in the main text. The first instability corresponds to a genuine phase transition into a state with nontrivial attention structure. The later sharp growth of the copy component is analogous to the lower-scale field ϕ_2: it is strongly enhanced when its effective mass becomes small, but because it is already linearly coupled to the ordered phase, this enhancement is a crossover rather than a second symmetry-breaking transition. In this sense, the later reorganization of the attention pattern is naturally interpreted as an *avoided* second transition.

Finally, we note that if the bilinear coupling is set to zero, $J = 0$, the symmetry is enlarged back to $\mathbb{Z}_2 \times \mathbb{Z}_2$, and two distinct transitions are generically possible when the bare scales T_1^0 and T_2^0 are well separated. The absence of such an additional singularity for $J \neq 0$ is precisely the point of the present toy model.

O Symmetry Motivation for the Ansatz

In this section we show that under the assumptions of: (1) sample independent attention pattern $\boldsymbol{G}_\alpha = \boldsymbol{G}$, and (2) no spontaneous symmetry breaking in the sequence permutation symmetry the ansatz is exact (no pattern outside its space can appear).

under Assumption 1 we find in equation 142 the following prior

$$C^{GG}_{(ab),(cd)} = \delta_{ac}\delta_{bd} - \frac{\delta_{ac}}{L+V} - \frac{\delta_{bd}}{L+V} + \frac{1}{(L+V)^2}$$

Interestingly this prior has an enhanced symmetry $S_L^Q \times S_L^K$ compare to the sample dependent symmetry $S_L^{\rm diag}$. However, the full action is

$$\mathcal{S}(G) = \frac{1}{2} y^\top \left[g_O G^\top C^{(xx)} G + \sigma^2 I \right]^{-1} y + \frac{1}{2} \ln\det \left[g_O G^\top C^{(xx)} G + \sigma^2 I \right]$$

$$+\frac{1}{2}G\left(\delta_{ac}\delta_{bd}-\frac{\delta_{ac}}{L+V}-\frac{\delta_{bd}}{L+V}+\frac{1}{(L+V)^2}\right)G$$

and the likelihood term breaks this symmetry. We can try to apply the same transformation and see what constraints the transformation must satisfy to keep the action invariant. Transforming $G \to P_k G P_q$

$$\mathcal{S}(G)=\frac{1}{2}y^\top\left[g_O\left(P_kGP_q\right)^\top C^{(xx)}P_kGP_q+\sigma^2 I\right]^{-1}y+\frac{1}{2}\ln\det\left[g_O\left(P_kGP_q\right)^\top C^{(xx)}\left(P_kGP_q\right)+\sigma^2 I\right]$$

and simplifying

$$\mathcal{S}(G)=\frac{1}{2}y^\top P_q^\top\left[g_O G^\top P_k^\top C^{(xx)}P_kG+\sigma^2 I\right]^{-1}P_qy+\frac{1}{2}\ln\det\left[g_O G^\top P_k^\top C^{(xx)}P_lG+\sigma^2 I\right].$$

Using the input S_L symmetry $x \to Px, y \to G_*Px = G_*PG_*^{-1}y$ to move P_k to y, we then have

$$\mathcal{S}(G)=\frac{1}{2}y^\top G_*^{-1}P_k^\top G_*P_q^\top\left[g_O G^\top C^{(xx)}G+\sigma^2 I\right]^{-1}P_qG_*P_kG_*^{-1}y+\frac{1}{2}\ln\det\left[g_O G^\top C^{(xx)}G+\sigma^2 I\right]$$

so we require $P_qG_*P_kG_*^{-1}=I$ which give the constraint $P_q = G_*P_k^\top G_*^{-1}$. The remaining symmetry is thus

$$G \to \left(G_*P_k^\top G_*^{-1}\right)GP_k$$

i.e. instead of S_L^{diag} or $S_L^Q \times S_L^K$ we have the twisted $S_L^{G_*}$.

Assumption 2 implies that only the trivial irrep of this group can ever appear, so we move to identifying this space. We require $G_{\text{trivial}} = \left(G_*P_k^\top G_*^{-1}\right)G_{\text{trivial}}P_k$ which means

$$G_*^\top G_{\text{trivial}}P_k^\top = P_k^\top G_*^\top G_{\text{trivial}}$$

therefor $G_*^\top G_{\text{trivial}}$ commutes with any permutation, so $G_*^\top G_{\text{trivial}} \in \text{span}\{I, J\}$ (with J the all ones matrix). That means $G_{\text{trivial}} \in \text{span}\{G_*, G_*J\} = \text{span}\{G_*, J\}$ completing the derivation.

P Attention Pattern Coverage by Order Parameters

Fig.6 shows the fraction of the full attention pattern that is captured by the patterns in the $(\hat{c}_1, \hat{c}_G)$ and $(\hat{a}_1, \hat{a}_G)$ subspace for linear and softmax attention. We quantified the coverage by cosine similarity and norm ratio between the full empirical attention pattern and the attention pattern projected down to the two-dimensional subspace.

For linear attention (Fig. 6 left panel), the two-dimensional order parameter space faithfully captures the degrees of freedom used by the network during training. The agreement for Adam means that the competition between pooling and copy attention accounts for essentially all of the learned attention structure, with no significant orthogonal components. SGLD samples show more variance than Adam in the crossover region $300 \leq P \leq 600$, where the coverage plot reveals leakage to attention patterns outside the two-dimensional subspace but remains above 80% for all P values after the attention pattern departs from zero ($P \gtrsim 20$). We attribute this effect to the large fluctuations expected near the point where the pooling minimum destabilizes and the copy minimum emerges (Fig. 2): as the saddle structure dissolves and the copy minimum deepens, the posterior broadens and the network transiently explores directions outside out 2d space. Away from this region, coverage recovers, confirming that the two-parameter reduction is accurate throughout the ordered phase.

For softmax attention (Fig. 6 right panel), we see perfect agreement for both Adam and SGLD for the full range of P.

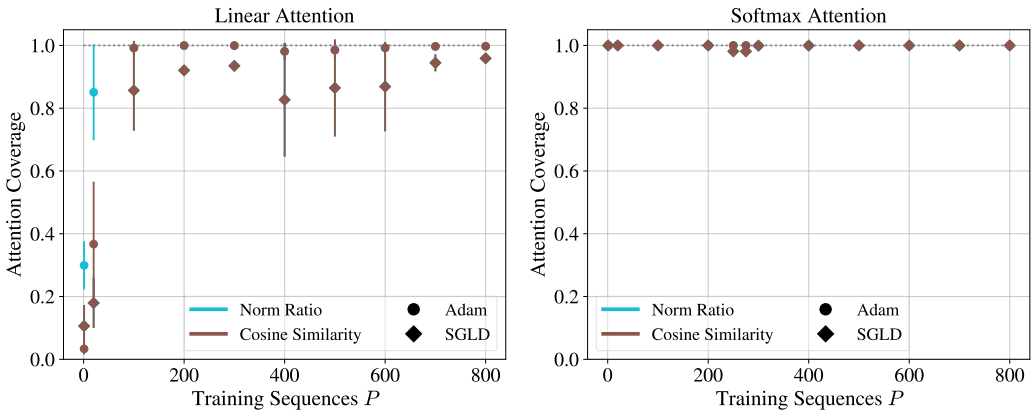

Figure 6: **Attention-pattern coverage by order-parameter projections.** We quantify how much of the empirical attention pattern lies in the two-dimensional subspace spanned by the pooling and copy order-parameter directions, using the projected norm ratio and cosine similarity. **Left:** *Linear attention* is nearly fully described by this subspace for Adam across training set sizes P; throughout much of the panel, Adam markers are hidden behind SGLD markers, and the norm-ratio points are hidden behind the cosine-similarity points. SGLD shows the only visible loss of coverage, with larger fluctuations in the crossover region $300 \leq P \leq 600$, consistent with transient attention components outside the order-parameter plane. **Right:** *Softmax attention* shows essentially perfect coverage for both Adam and SGLD over the full range of P, with all metrics and optimizers nearly overlapping at one.

Q Additional Figures

Q.1 Varying σ^2

In this section, we decrease σ^2 to probe the validity of the EK approximation in this system for small observation uncertainty/ridge. Figs. 7,8 show the order parameters as a function of sample complexity for different values of σ^2 for linear and softmax attention, respectively. Similarly, Figs. 9,10 show the test loss as a function of sample complexity for different values of σ^2 for linear and softmax attention. The agreement between theory and experiments for linear attention degrades as σ^2 decreases, with $\sigma^2 = 1/30$ showing meaningful deviations, and $\sigma^2 = 1/300$ showing very large deviations. For softmax, the agreement is maintained for all values of σ^2.

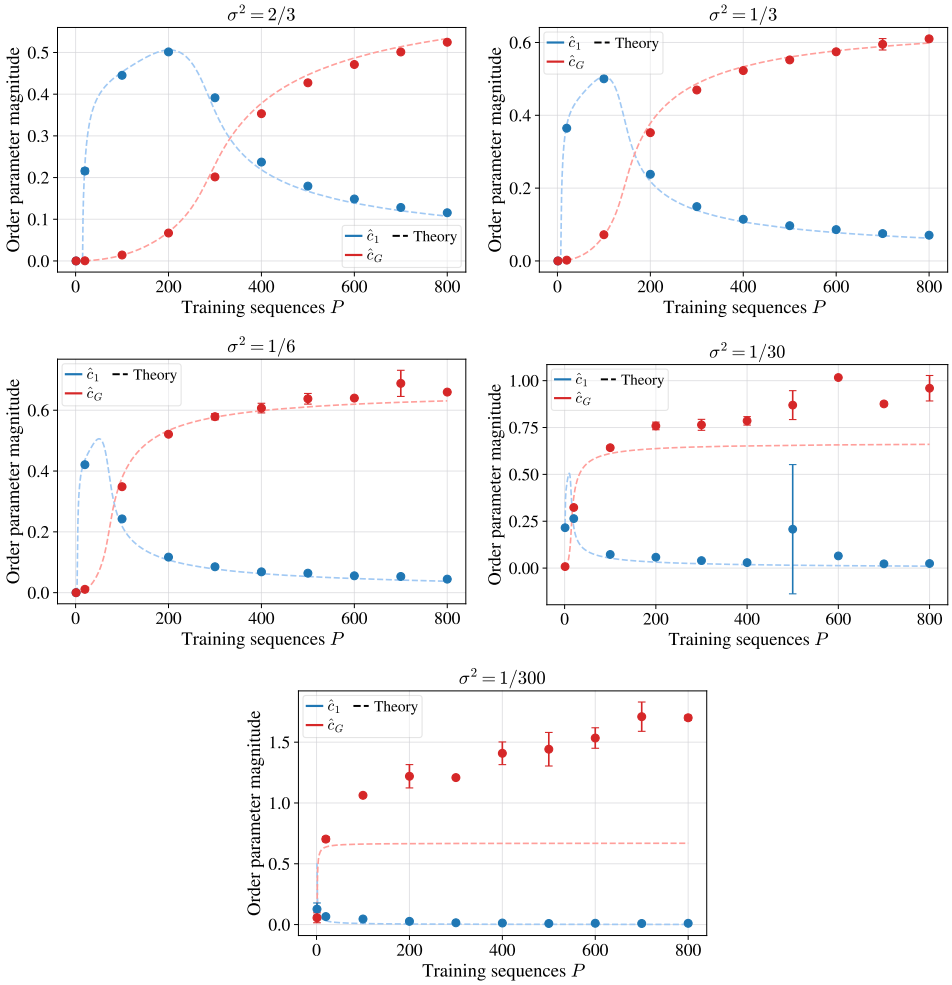

Figure 7: The order parameter for *linear attention* trained with Adam for various values of σ^2; the main text uses $\sigma^2 = 1$. We find good agreement with the theory for values down to $\sigma^2 = 1/30$, with $\sigma^2 = 1/300$ showing significant deviations from the theory.

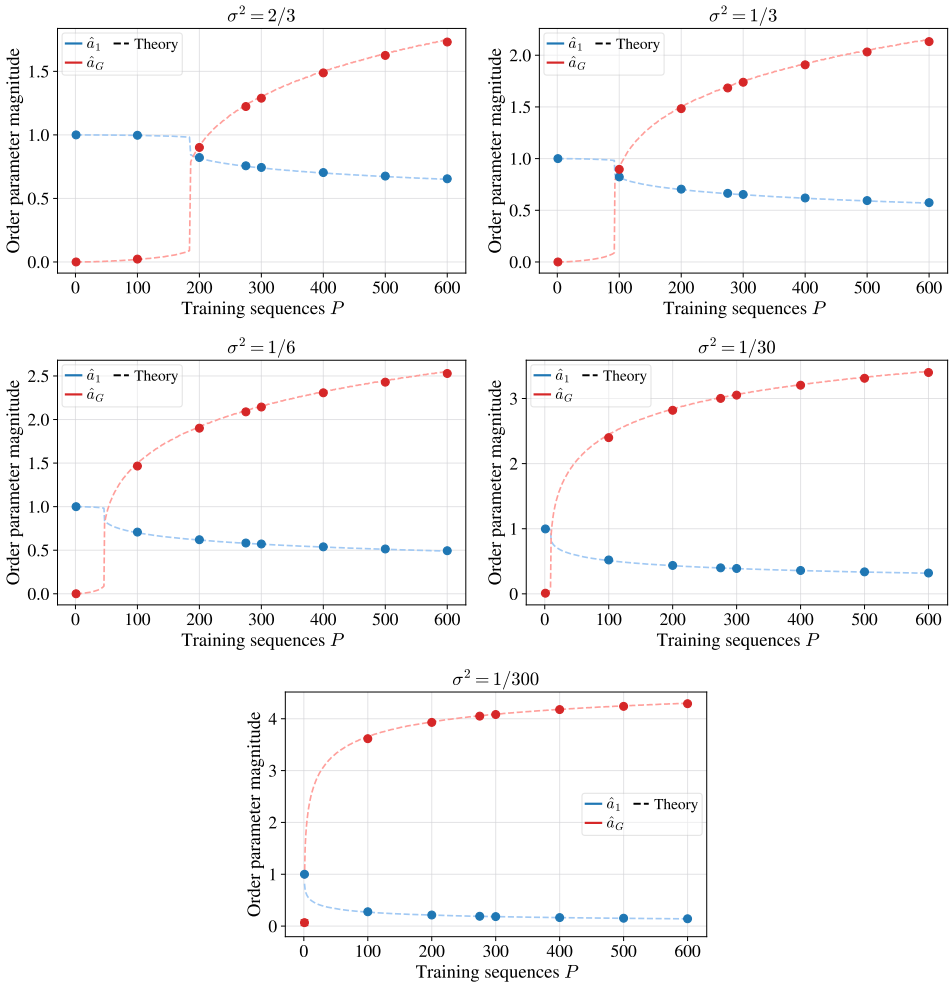

Figure 8: The order parameter for *softmax attention* trained with Adam for various values of σ^2; the main text uses $\sigma^2 = 1$. We find good agreement with the theory throughout.

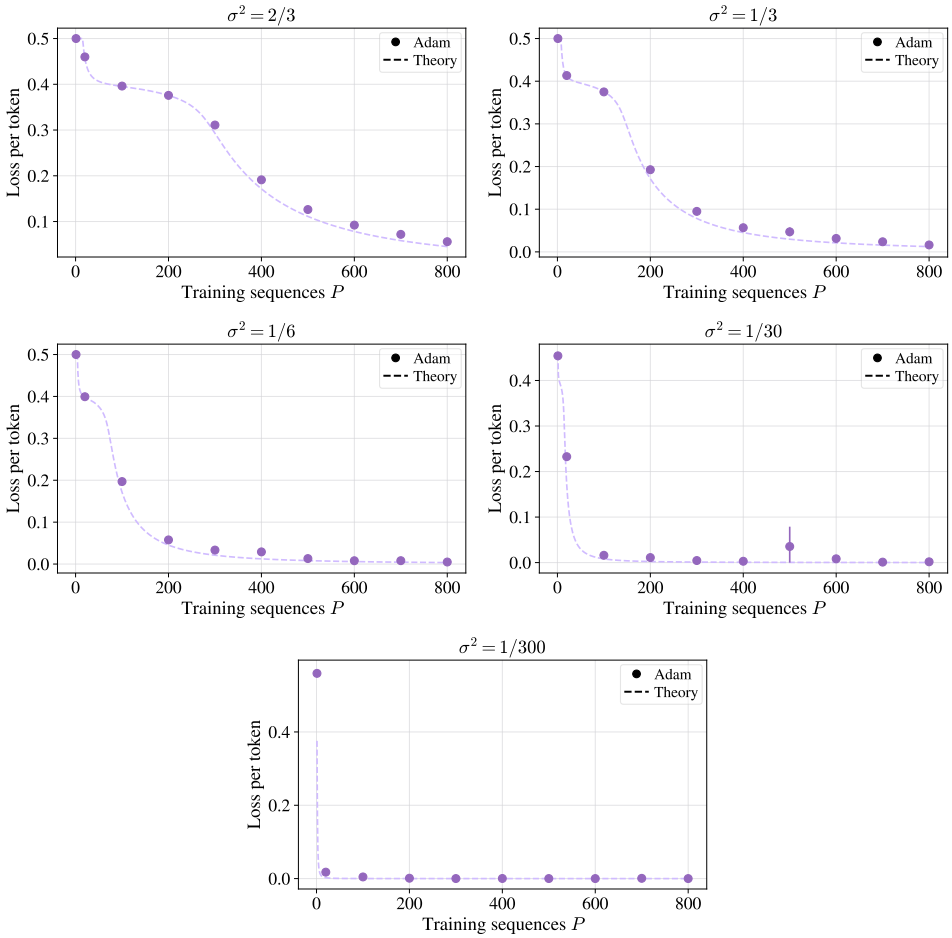

Figure 9: The test loss for *linear attention* trained with Adam for various values of σ^2; the main text uses $\sigma^2 = 1$.

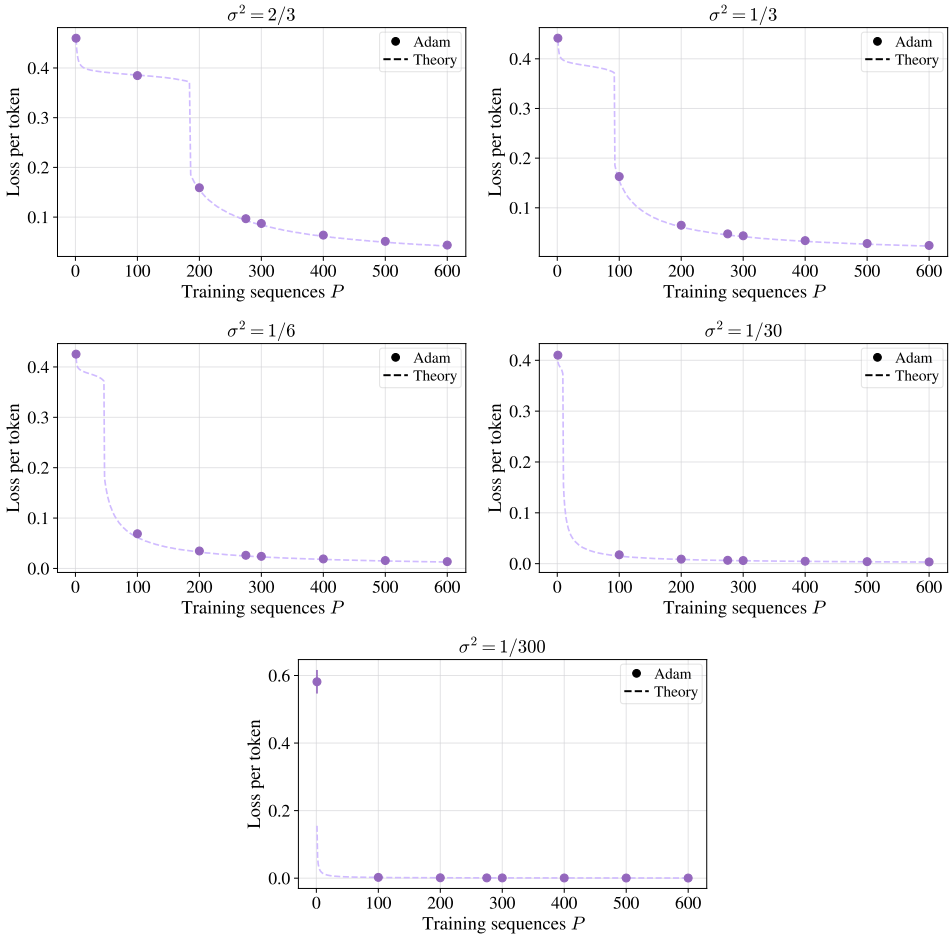

Figure 10: The test loss for *softmax attention* trained with Adam for various values of σ^2; the main text uses $\sigma^2 = 1$. We find good agreement with the theory for values down to $\sigma^2 = 1/30$, with $\sigma^2 = 1/300$ showing deviations from the theory.

Q.2 Results Under Gradient Descent

In this section, we show the results when replacing Adam with vanilla gradient descent. We find similar results, reinforcing the robustness to optimizer choice. The results for linear attention are given in Fig. 11 and for softmax in Fig. 12

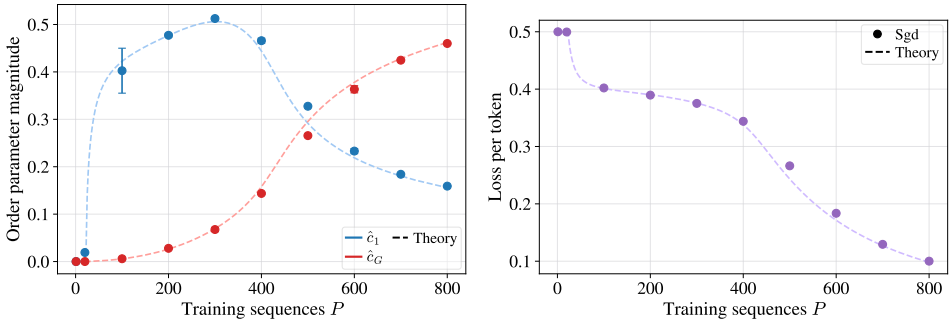

Figure 11: The order parameters (left) and test loss (right) for *linear attention* trained with GD, analogous to Fig. 3.

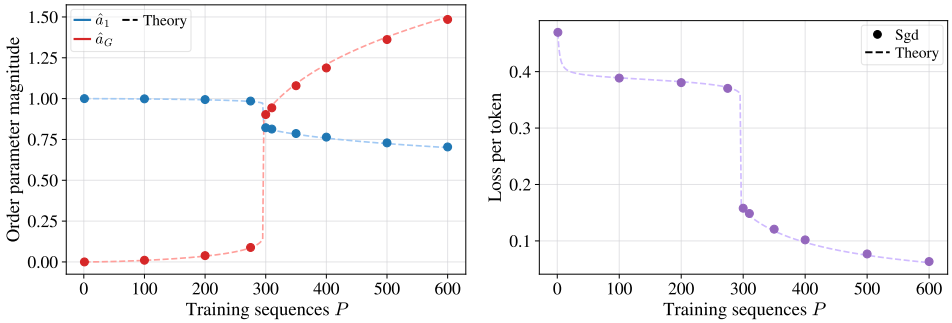

Figure 12: The order parameters (left) and test loss (right) for *linear attention* trained with GD, analogous to Fig. 4.

Q.3 Smaller Model Dimension

In Fig. 13, we show results for a model that has $\times 10$ smaller model dimension than the main text, such that $d_{\text{model}} = 1,000$ with the goal of probing the approximation $d_{\text{model}} \gg P, L$. The results appear largely the same with no notable deviations even though in this case $P \approx d_{\text{model}}$, indicating a degree of robustness. Nevertheless, we expect that pushing the ratio further such that $P \gtrsim d_{\text{model}}$ can change the results by moving the embedding kernel away from its NNGP value and softening the attention prior.

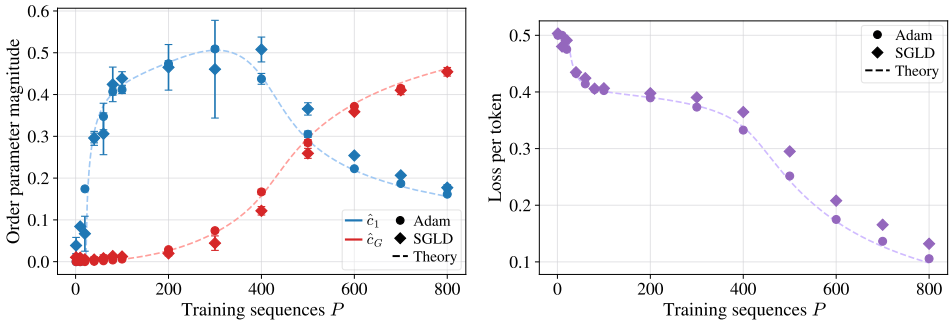

Figure 13: The order parameters (left) and test loss (right) for *linear attention* with $d_{\text{model}} = 1,000$, instead of $d_{\text{model}} = 10,000$ in Fig. 3.

R Details of Numerical Experiments

Experimental details. Unless stated otherwise, all numerical experiments use the single-layer, single-head linear attention model introduced in Section 2 with merged attention and readout matrices. Training data consist of length-L sequences of one-hot encoded tokens drawn uniformly from a vocabulary of size V, with shift-by-one copy labels. Across experiments, we vary only the number of training sequences

$$P \in \{1, 10, 20, 60, 100, 250, 275, 200, 300, 400, 500, 600, 700, 800\},$$

and linear/softmax attention while keeping all other settings fixed.

For each value of P, we train an ensemble of 10 models. Optimization is performed in two stages. We first run an Adam warmup phase for 2,000 epochs with learning rate 10^{-5}. For softmax we use learning rate warmup from 10^{-7} to 10^{-5} over 10% of the Adam epochs. Early stopping is used with patience 100 and minimum improvement threshold 0. We then continue training with Langevin updates using learning rate 5×10^{-10}, noise level $\sigma^2 = 30$, and batch size 1,000, for at most 15,000 epochs in total. Unless noted otherwise, reported results are averaged over the ensemble at fixed P.

Each run of the full ensemble required roughly 5 hours on an H100, such that in total the experiments required 100 H100 hours.

Hyperparameter	Value
Sequence length L	25
Vocabulary size V	5
Training set size P	$\{1, 10, 20, 60, 100, 200, 250, 275, 300, 400, 500, 600, 700, 800\}$
Test set size	100
Ensemble size	10
Architecture	single-layer linear/softmax attention
Number of heads	1
Model dimension d_{model}	10,000
χ	30
Warmup optimizer	Adam
Warmup epochs	2,000
Warmup learning rate	10^{-5}
Second-stage optimizer	Langevin / SGLD
Second-stage learning rate	5×10^{-10}
Batch size	1,000
Maximum number of epochs	15,000
Noise level σ^2 (post scaling with χ, L)	1
Early stopping patience	100
Early stopping minimum improvement	0

Table 1: Hyperparameters used in the numerical experiments. The only swept quantity is the number of training sequences P; all other settings are kept fixed across runs.

R.1 Training Trajectories and Filtering

To expedite the posterior sampling, we warm start all our SGLD [54] runs from the Adam early stopping minimum. We note again that the equilibrium distribution sampled by SGLD is independent of the initial conditions. Since optimizing with Adam is a more common practice, we also show the results at the end of the Adam training stage; in these cases, we filter out training runs that reached a local minimum (which is not the global minimum). Importantly, filtering was performed only on the basis of convergence to the global loss minimum, and not on the resulting order parameters or their agreement with theory. Figs. 14 and 15 show the trajectories of all runs using Adam as a function of the epoch (equiv. step, batch). For each setting, we trained an ensemble of 10 networks. Figures 16, 15 show the corresponding time evolution of the order parameters.

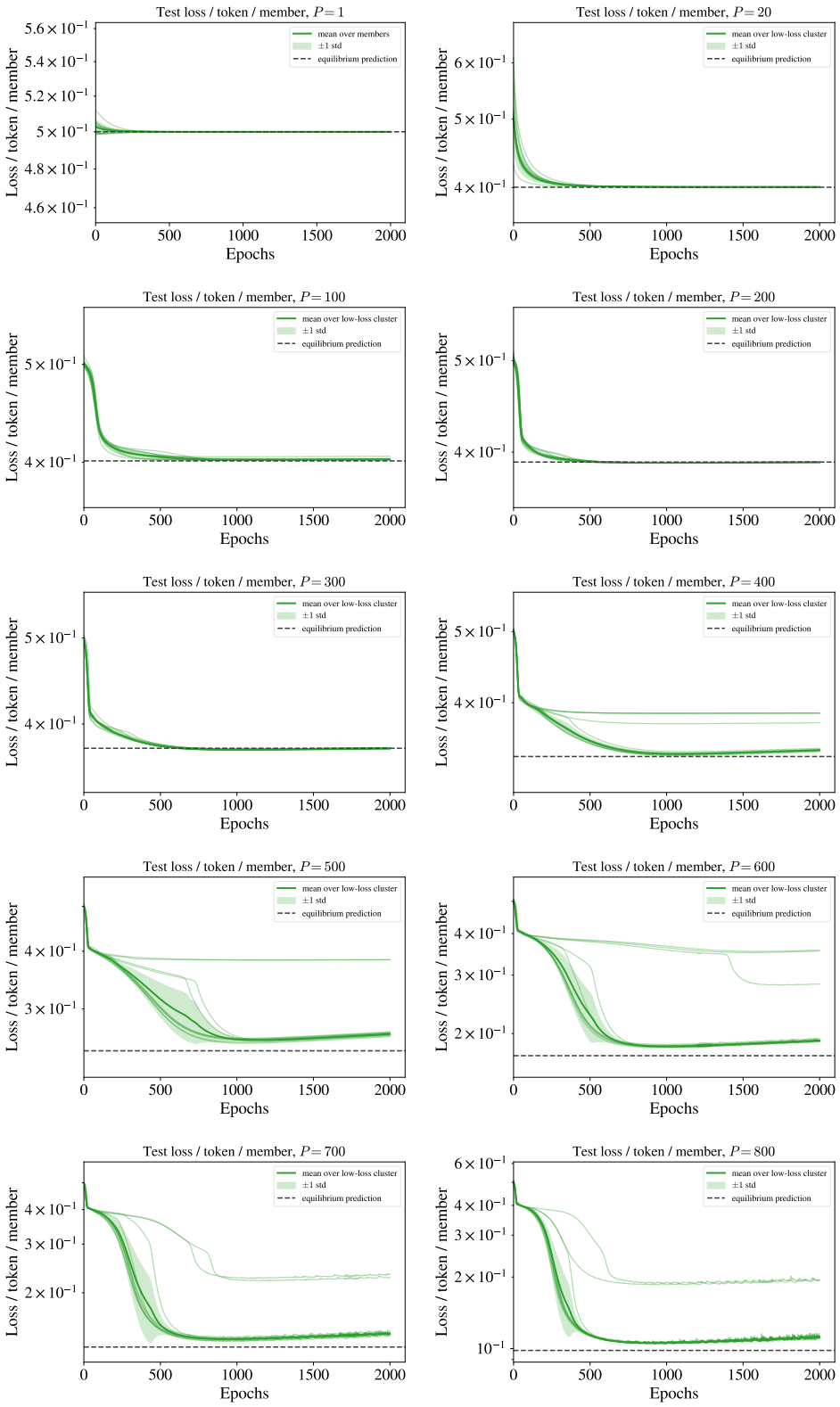

Figure 14: Test loss for *linear attention* trained with Adam. We show the test loss as a function of the number of epochs for different values of P on models trained with Adam. We clearly see two clusters with one reaching a lower loss value. Since the Bayesian posterior, and hence SGLD in principle, is insensitive to the initial conditions and would, in principle, reach the global minimum given sufficient time, we include only converged runs in the cluster with lower loss in the results in the main text.

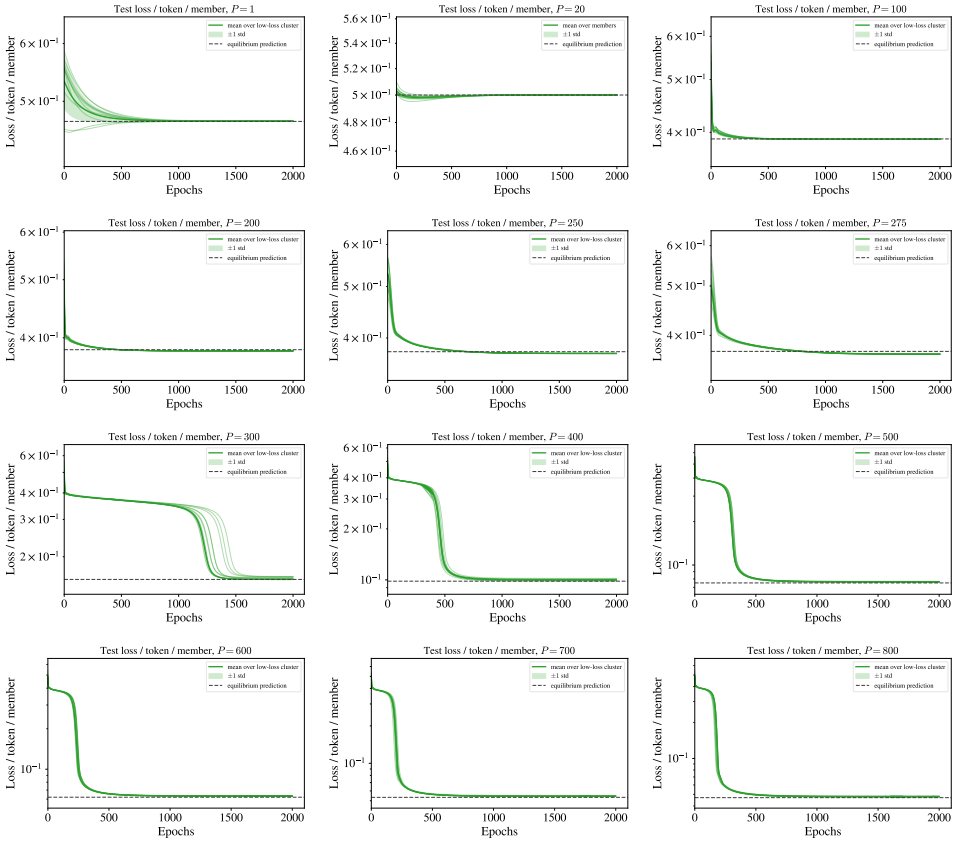

Figure 15: Test loss for *softmax attention* trained with Adam. We show the test loss as a function of the number of epochs for different values of P on models trained with Adam. With softmax Adam converged to the global minimum in all of our seeds.

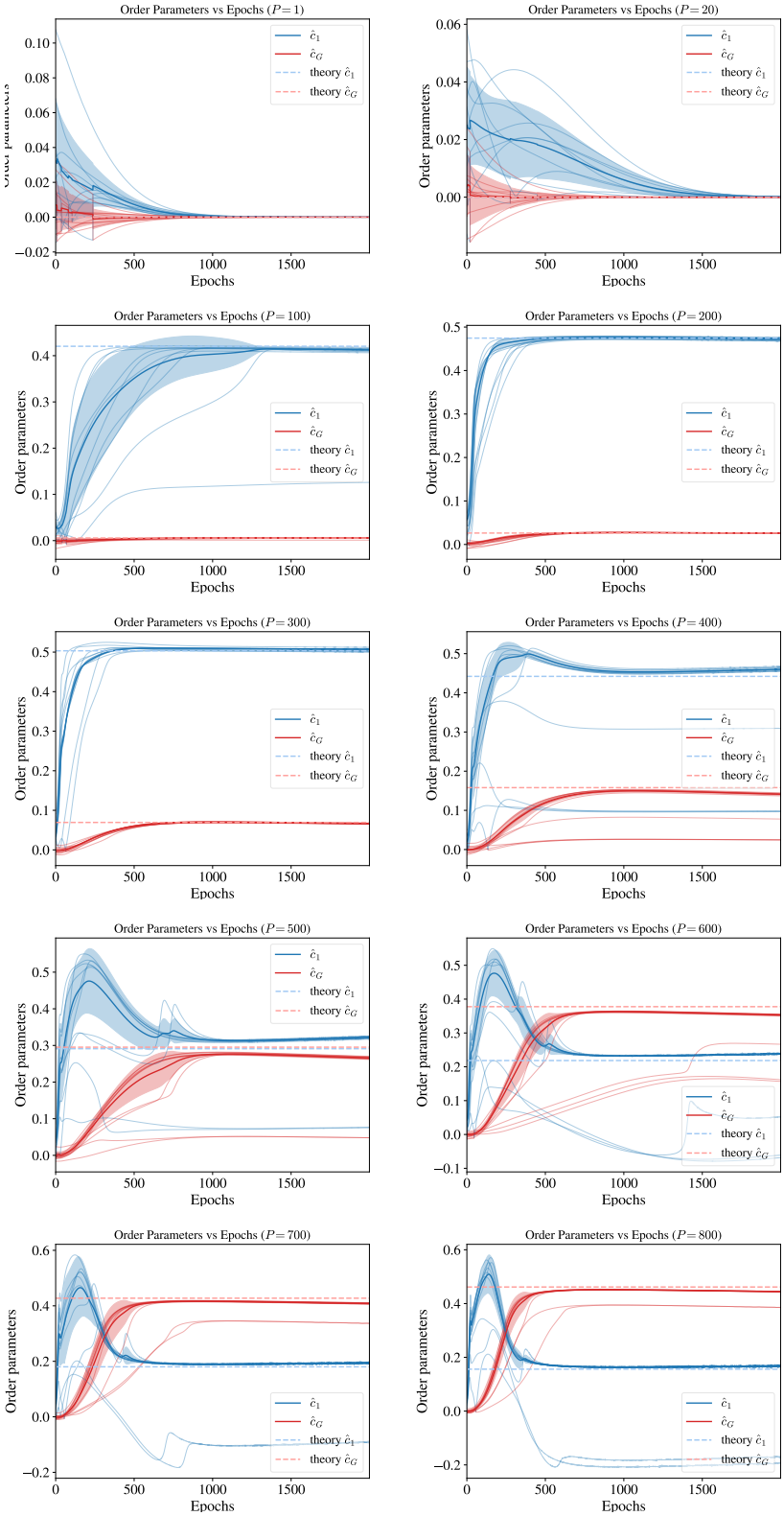

Figure 16: Order parameters for *linear attention* trained with Adam for the same runs as in Fig. 14.

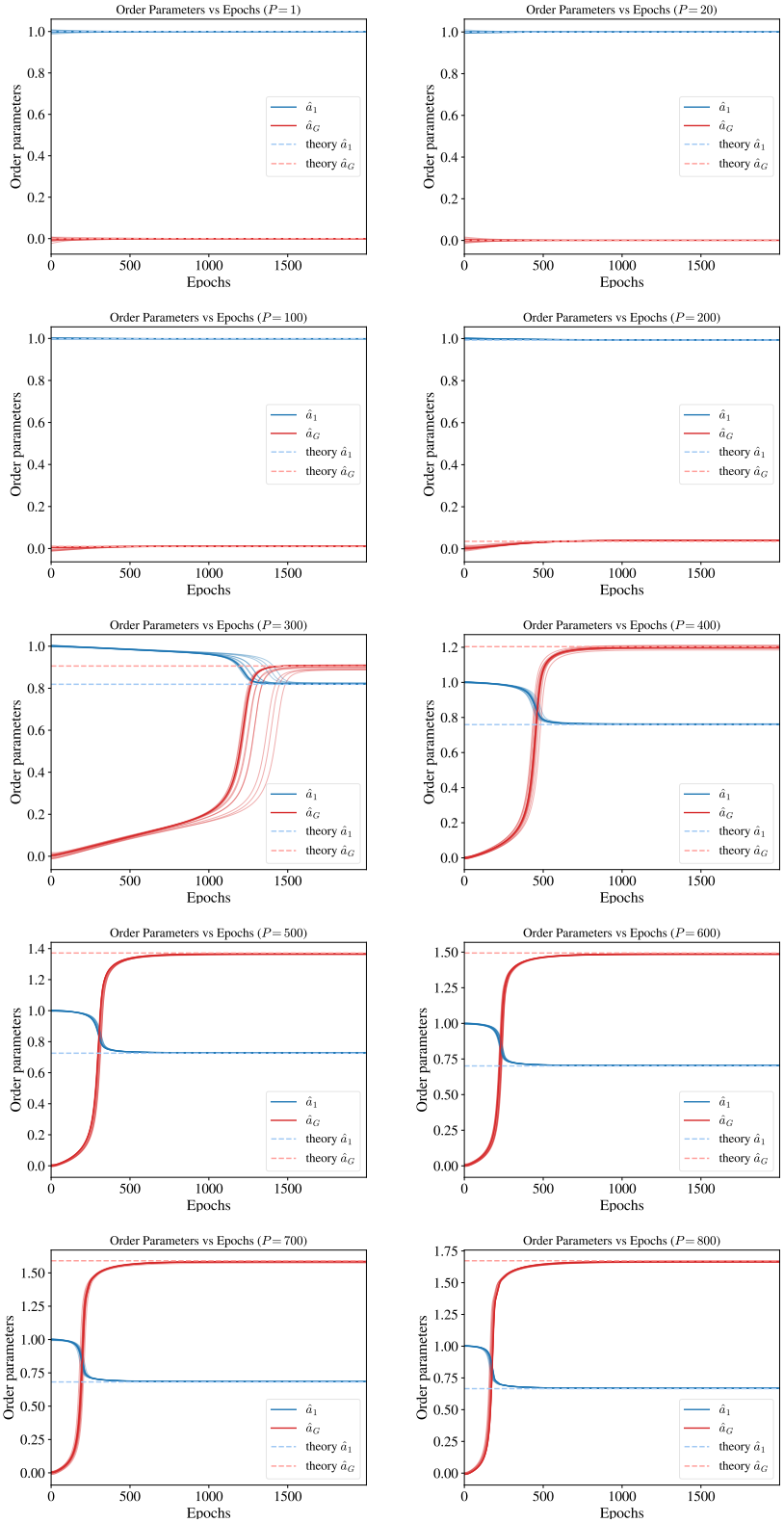

Figure 17: Order parameters for *softmax attention* trained with Adam for the same runs as in Fig. 15.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: We reflect the paper's contributions and scope in both abstract and introduction.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: We have a paragraph dedicated to limitations of the work in the discussion section.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [Yes]

   Justification: Yes.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: We include all information to reproduce the main experimental results of the paper in the Appendix Section on Langevin Stochastic Gradient Descent.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [No]

   Justification: We will publish the code and include a reference in the paper for the final version.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

   Answer: [Yes]

   Justification: We include all relevant details in the Section on Langevin Stochastic Gradient Descent.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [Yes]

   Justification: We report error bars over multiple network initializations, corresponding to sampling from the network ensemble defined by the Gaussian weight prior.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

Justification: We report them in the experimental details.

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: We adhere to the NeurIPS Code of Ethics.

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

    Justification: We include a paragraph in the discussion section.

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

    Answer: [N/A]

    Justification: The paper poses no such risks.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [N/A]

    Justification: The paper does not use existing assets.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [N/A]

    Justification: The paper does not introduce new assets.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [N/A]

    Justification: The paper does not perform crowdsourcing experiments oe research with human subjects.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [N/A]

    Justification: The paper does not involve crowdsourcing nor research with human subjects.

16. **Declaration of LLM usage**

    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

Answer: [N/A]

Justification: While we study transformers from the theoretical perspective, the paper does not include LLMs as part of the core method.