Bayes optimal learning of attention-indexed models

Fabrizio Boncoraglio
Statistical Physics of Computation Laboratory
EPFL, Switzerland

Emanuele Troiani
Statistical Physics of Computation Laboratory
EPFL, Switzerland

Vittorio Erba
Statistical Physics of Computation Laboratory
EPFL, Switzerland

Lenka Zdeborová
Statistical Physics of Computation Laboratory
EPFL, Switzerland

Abstract

We introduce the attention-indexed model (AIM), a theoretical framework for analyzing learning in deep attention layers. Inspired by multi-index models, AIM captures how token-level outputs emerge from layered bilinear interactions over high-dimensional embeddings. Unlike prior tractable attention models, AIM allows full-width key and query matrices, aligning more closely with practical transformers. Using tools from statistical mechanics and random matrix theory, we derive closed-form predictions for Bayes-optimal generalization error and identify sharp phase transitions as a function of sample complexity, model width, and sequence length. We propose a matching approximate message passing algorithm and show that gradient descent can reach optimal performance. AIM offers a solvable playground for understanding learning in self-attention layers, that are key components of modern architectures.

1 Introduction

The transformer architecture [1] has transformed machine learning, achieving state-of-the-art results in natural language processing [2, 3], computer vision [4], and beyond. Its core innovation—the self-attention mechanism—enables models to capture long-range dependencies between tokens. Despite their empirical success, transformers remain poorly understood theoretically, especially regarding how data structure, attention bias, and training dynamics interact in finite-sample regimes. While mechanistic interpretability has shed light on trained models, the learning process itself—what is statistically and computationally learnable from limited data—remains unexplained. A common strategy toward progress is to study simplified models in high-dimensional regimes, where the *blessing of dimensionality* [5] can yield tractable characterizations of learning. A key ingredient in this approach is a synthetic data model that captures salient aspects of real-world structure.

A common feature of natural data is that structure arises from pairwise interactions between elements. In language, for example, sentence meaning depends on both the semantics of individual words and their syntactic roles. This can be modeled by assigning each word a *semantic* embedding $\boldsymbol{x}$ and a *contextual* embedding $\boldsymbol{z}$. For instance, in "Max eats chocolate," one level of understanding identifies what each word means (e.g., Max is a person), while another captures grammatical roles (Max is the subject). These two embeddings are linked: contextual meaning emerges by comparing semantic content across words. Similar relational patterns occur in images, molecules, and graphs. The empirical success of transformers suggests that multi-layer self-attention—which performs weighted pairwise comparisons in a hierarchical way—is a natural mechanism for modeling such interactions. To analyze this, we construct a synthetic data model that reflects pairwise structure between tokens.

Theoretical understanding of fully connected neural networks has advanced significantly through the analysis of Gaussian single-index and multi-index models in the high-dimensional limit [6, 7, 8, 9,

39th Conference on Neural Information Processing Systems (NeurIPS 2025).

10, 11, 12, 13, 14, 15]. In statistical physics, similar models appear as teacher-student perceptrons [16, 17, 18, 19] or committee machines [20, 21]. These setups typically assume i.i.d. Gaussian inputs, with targets depending on a small number of random projections—"indices"—of the input. They provide a rich theoretical playground for jointly analyzing learning dynamics, generalization, and architectural biases.

Recent work has extended this framework to model key aspects of transformers, introducing the *sequence multi-index* (SMI) model [22, 23, 24]. While insightful, existing SMI models require the width of the key and query matrices to be much smaller than the token embedding dimension—a regime where only narrow attention layers can be analyzed. In contrast, practical transformers typically use key and query widths comparable to the embedding dimension. This motivates our contribution: a high-dimensional yet analyzable model where learnable matrices have extensive rank. We call this the *attention-indexed model*.

The attention-indexed model (AIM). We introduce a new class of high-dimensional functions designed to model pairwise relationships between tokens. Analogous to classical multi-index models, the *attention-indexed model* defines outputs y as nonlinear functions of high-dimensional token embeddings $\boldsymbol{x}_a^\top \in \mathbb{R}^d$ for $a = 1, \dots, T$, arranged as rows of the input sequence of tokens $X_0 \equiv X \in \mathbb{R}^{T\times d}$ for $a = 1, \dots, T$. We define L *attention indices* $h^{(\ell)} \in \mathbb{R}^{T\times T}$ with components $h_{ab}^{(\ell)}$. The labels y for each input $X \in \mathbb{R}^{T\times d}$ are generated via a general output function $g : \mathbb{R}^{L\times T\times T} \to \mathbb{R}^{T\times T}$

$$h_{ab}^{(\ell)} \equiv \frac{\boldsymbol{x}_a^\top S_\ell \, \boldsymbol{x}_b - \delta_{ab} \operatorname{Tr} S_\ell}{\sqrt{d}}, \qquad y = g\left(\{h^{(\ell)}\}_{\ell=1}^L\right). \tag{1}$$

Here each $S_\ell \in \mathbb{R}^{d\times d}$ is a learnable matrix. The diagonal mean is subtracted to avoid divergence as $d \to \infty$, ensuring the fluctuations of $h^{(\ell)}$ remain $\mathcal{O}(1)$.

While our theory applies to general rotationally invariant S_ℓ, a motivating example is when $S_\ell \simeq Q_\ell K_\ell^\top \in \mathbb{R}^{d\times d}$, as in self-attention [1], with key and query matrices $K_\ell, Q_\ell \in \mathbb{R}^{d\times r_\ell}$. We refer to r_ℓ as the *width* of the ℓth layer; it typically controls the rank of S_ℓ, though we also consider $r_\ell > d$. For analytical simplicity, we assume tied key and query, $Q_\ell = K_\ell = W_\ell$, so that

$$S_\ell = \frac{1}{\sqrt{r_\ell \, d}} W_\ell W_\ell^\top \in \mathbb{R}^{d\times d}, \quad W_\ell \in \mathbb{R}^{d\times r_\ell}. \tag{2}$$

Multi-layer attention. A key instance of the output function g arises when contextual embeddings are built via stacked attention layers. Each layer updates the representation by mixing semantic embeddings through weighted pairwise interactions. Formally, we define the multi-layer attention mechanism as:

$$y \;=\; \sigma_\beta\big(H_L(X_{L-1})\big), \qquad X_\ell \;=\; \Big[c\,\mathbb{I}_T + \sigma_\beta\big(H_\ell(X_{\ell-1})\big)\Big] X_{\ell-1} \tag{3}$$

$$H_\ell(X_{\ell-1}) \;:=\; \frac{1}{\sqrt{d}} X_{\ell-1} S_\ell X_{\ell-1}^\top - \frac{1}{\sqrt{d}} \mathbb{E}_{\mathrm{tr}}[X_{\ell-1} S_\ell X_{\ell-1}^\top] \tag{4}$$

where $X_0 \in \mathbb{R}^{T\times d}$ is the input sequence, c is the residual strength, and L is the number of layers. The function $\sigma : \mathbb{R}^{T\times T} \to \mathbb{R}^{T\times T}$ is typically the softmax with inverse temperature β; increasing β sharpens the focus on dominant pairwise interactions. Finally, the expectation in $H_\ell(X)$ is intended over the input data X_0 (empirical average).

A particularly interesting case is the limit of infinite β, where softmax becomes the *hardmax* function that selects the maximum entry in each row of the attention logits. This can be interpreted as enforcing a sparse, winner-take-all selection among candidate interactions, capturing the case where for each token only a single other token carries the relevant signal. Such mechanisms are especially relevant in symbolic, compositional, or relational reasoning, where outputs depend on identifying a unique best match per query.

This form fits into the attention-indexed framework by choosing the output function g in (1) as:

$$g(\{h^{(\ell)}\}_{\ell=1}^L) = \sigma_\beta\Big(B_c^{L-1}(h^{(1)}, \dots, h^{(L-1)})\, h^{(L)}\; B_c^{L-1}(h^{(1)}, \dots, h^{(L-1)})^\top \Big), \tag{5}$$

where the recursion is defined by:

$$B_c^0 \;=\; \mathbb{I}_T, \quad B_c^\ell \;=\; \Big[c\,\mathbb{I}_T \;+\; \sigma_\beta\Big(B_c^{\ell-1}\,h^{(\ell)}\,B_c^{\ell-1\,\top}\Big)\Big]\,B_c^{\ell-1} \quad \ell = 1,\dots,L. \tag{6}$$

In particular, with this formulation the map in (4) can be rewritten as:

$$H_\ell(X_{\ell-1}) \;=\; \frac{1}{\sqrt{d}}\,X_{\ell-1}\,S_\ell\,X_{\ell-1}^\top - \frac{\operatorname{Tr} S_\ell}{\sqrt{d}}B_c^{\ell-1}B_c^{\ell-1^\top} = B_c^{\ell-1}\,h^{(\ell)}\,B_c^{\ell-1\top} \tag{7}$$

This formulation also accommodates sequence-level outputs of the form yX_{L-1}, corresponding to the output map gB_c^{L-1}. Details of the corresponding mappings are given in Appendix B.

The existence of such mapping, from deep attention to a low-dimensional collection of attention indices, is a key contribution: it allows information-theoretic and algorithmic results derived for AIMs to transfer to multi-layer attention. In App. B we extend such mapping to multi-head and seq2seq variants. Full Transformer blocks (e.g., ViT/LLMs) interleave multi-head self-attention with token-wise MLPs, and are commonly trained in an auto-regressive fashion; our supervised setup isolates the attention learning problem, to which our results apply directly.

Our contributions. We initiate a theoretical study of learning from data generated by the attention-indexed model, without restrictions on the width of the attention matrices S_ℓ, in the high-dimensional limit. Specifically, we consider large embedding dimension d, with attention widths r_ℓ (or ranks of S_ℓ) scaling proportionally with d, while keeping sequence length T and depth L finite. The number of samples n scales quadratically with dimension to span the regime where the optimal test error changes from as bad as a random guess to zero:

$$d \to \infty\,, \quad \alpha \equiv \frac{n}{d^2} = \Theta(1)\,, \quad \rho_\ell \equiv \frac{r_\ell}{d} = \Theta(1)\,, \quad T = \Theta(1)\,, \quad L = \Theta(1)\,. \tag{8}$$

In this limit, for Gaussian i.i.d. inputs and suitable priors over S_ℓ, we analyze the Bayes-optimal estimator—that is, the posterior mean predictor. Concentration of measure allows key observables—such as the test error—to become deterministic, enabling a description in terms of low-dimensional fixed-point equations. We derive those equations using tools from statistical mechanics and random matrix theory. Beyond analytical tractability, this asymptotic setting captures relevant scalings, where model and data dimensions grow together.

Our analysis uncovers phase transitions in learnability as a function of sample complexity α, sequence length T, and attention width ρ. We show that attention mechanisms with extensive width can efficiently recover latent structure, particularly when the target model exhibits sparsity. We further study the role of the inverse temperature parameter β, contrasting the smooth regime $0 < \beta < +\infty$ with the hard assignment case $\beta = +\infty$, where the task reduces to discrete token association. We derive an approximate message passing (AMP) algorithm that matches Bayes-optimal performance in the studied setting. We also show empirically that an averaged version of gradient descent on a natural loss achieves similar optimality, aligning theory with practical training dynamics.

Further related work. The attention-indexed model (AIM) is motivated by a generative perspective, capturing how structured token-level outputs arise from layered bilinear interactions between high-dimensional embeddings—mirroring attention computations in transformers. The idea of modeling learning through such structured synthetic data dates back to the teacher–student setting in the perceptron literature [16, 17, 20], and more recently to single-index and multi-index models [6, 7, 8, 9, 10, 11, 12, 13, 14, 15].

While many theoretical studies explore simplified transformer variants, most do not rely on or benefit from the high-dimensional limit. These include works that analyze one-layer attention under finite embedding dimension [25, 26, 27, 28, 29], or study training dynamics in the linear, kernel, or random feature regimes [30, 31, 32]. Others use infinite-width approximations without access to generalization error [33, 34]. By contrast, theoretical analysis of *nonlinear* attention layers with *trainable* key and query matrices in the limit of high embedding dimension—together with sharp control of generalization—is less explored. As far as we are aware, only a few works address this regime [23, 22, 35, 36], and they all assume attention matrices of *finite* width.

Methodologically, our approach builds on techniques from high-dimensional multi-index models, particularly those developed in [21, 37], and their recent generalizations to sequence learning with multiple low-width self-attention layers [35]. The main technical challenge addressed in this paper is

extending these tools to the case where the width r of the attention matrices scales proportionally with the embedding dimension—i.e., the extensive-width regime—going beyond the key limitations of prior analyses.

To tackle this, we leverage recent results on the ellipsoid fitting problem [38, 39] and its connection to two-layer neural networks with quadratic activations and extensive width [40, 41]. Remarkably, the linear AIM model with $T = L = 1$ is mathematically equivalent to such quadratic networks, allowing us to adopt these methods. We generalize this connection to arbitrary T, L. This is enabled by a central conceptual tool, the AIM index, which disentangles the complexity of deep attention models. It allows us to split the problem into two subproblems: (i) how structure propagates across layers and tokens, and (ii) how attention matrices are learned from those structures. This separation is crucial in extending the theory to multiple layers and tokens.

Finally, we note that we focus here on the tied case $Q = K$ for clarity. The untied setting $Q \neq K$ is amenable to similar analysis following [42], and we leave its treatment for future work.

2 Setting

We consider a dataset $\mathcal{D} = \{y^\mu, X_0^\mu\}$ of n samples indexed by μ, where $X_0^\mu = \{\boldsymbol{x}_a^{\mu\top}\}_{a=1}^T \in \mathbb{R}^{T\times d}$ has rows $\boldsymbol{x}_a^{\mu\top}$. Each sample consists of the embeddings of T tokens $\boldsymbol{x}_a^{\mu\top} \in \mathbb{R}^d$, taken as standard Gaussian $\boldsymbol{x}_a^{\mu\top} \sim \mathcal{N}(0, \mathbb{I}_d)$ and of $T \times T$ matching output matrices y^μ encoding pair-wise information on the original tokens. We stress that the Gaussian assumption for the data can be relaxed in the same spirit as in [41, Assumption 2.2].

We generate y^μ using an attention-indexed model as given in (1) with matrices $\{S_\ell^*\}_{\ell=1,\dots,L}$ that are symmetric and extracted independently from a rotationally invariant ensemble $P_S(S) = P_S(O^\top SO)$ for any $d \times d$ rotation matrix O. We fix the normalizations such that $\mathbb{E}_{P_S}[\operatorname{Tr} S] = \kappa_1 d$ and $\mathbb{E}_{P_S}[\operatorname{Tr} S^2] = \kappa_2 d$ and with κ_1, $\kappa_2 = \mathcal{O}(1)$. We assume that the empirical spectral distribution of $S \sim P_S$ converges to a a distribution μ_S for $d \to +\infty$. This setting can be relaxed in several directions, allowing for different prior distributions $P_S^{(\ell)}$ for different layers, as well as considering non-symmetric matrices [43].

We consider the Bayes-optimal (BO) learning setting: the statistician knows the generative process of the dataset, i.e. the non-linearity g in (1) and the prior distribution P_S, and observes a dataset $\mathcal{D}$ but not the specific set of weights $\{S_\ell^*\}_{\ell=1,\dots,L}$ used to generate said dataset. The task is then to optimally estimate either the weights S^* (estimation task), i.e. find the estimator $\hat{S}(\mathcal{D})$ that minimizes

$$\mathcal{E}_{\text{est}}(\hat{S}) = \mathbb{E}_{\mathcal{D},S^*} \frac{1}{d} \sum_{\ell=1}^{L} ||\hat{S}(\mathcal{D})_\ell - S_\ell^*||_F^2 \,, \tag{9}$$

or the label associated to a new input sample X_{new} (generalization task), i.e. find the estimator $\hat{y}(\boldsymbol{x}, \mathcal{D})$ that minimizes

$$\mathcal{E}_{\text{gen}}(\hat{y}) = \mathbb{E}_{\mathcal{D},S^*} \mathbb{E}_{y_{\text{new}},X_{\text{new}}} ||\hat{y}(X_{\text{new}}, \mathcal{D}) - y_{\text{new}}||_F^2 \,, \tag{10}$$

where $(y_{\text{new}}, X_{\text{new}})$ is a new label-sample pair generated with the weights S^*. We will call the error achieved by the optimal estimators, respectively, the BO estimation error and BO generalization error.

Both BO estimators can be computed from the knowledge of the posterior distribution, i.e. the probability that a given set of weights S was used to generate the observed dataset

$$P(S_1, ..., S_L|\mathcal{D}) = \frac{1}{\mathcal{Z}(\mathcal{D})} \prod_{\ell=1}^{L} P_S(S_\ell) \prod_{\mu=1}^{n} \delta\left(y^\mu - g\left(h^{(1)}(S_1, \boldsymbol{x}^\mu), ..., h^{(L)}(S_L, \boldsymbol{x}^\mu)\right)\right) \,, \tag{11}$$

where the attention indices $h^{(\ell)} \in \mathbb{R}^{T\times T}$ were defined in (1) and $\mathcal{Z}(\mathcal{D})$ is a normalization factor. The BO estimator with respect to the estimation error is the mean of the posterior distribution, while BO estimator with respect to the generalization error is the mean of the predicted label under the posterior distribution, i.e.

$$\hat{S} = \mathbb{E}_{S\sim P(S|\mathcal{D})}[S] \,, \qquad \hat{y}(\boldsymbol{x}) = \mathbb{E}_{S\sim P(S|\mathcal{D})}\left[g\left(h^{(1)}(S_1, \boldsymbol{x}), ..., h^{(L)}(S_L, \boldsymbol{x})\right)\right] \,. \tag{12}$$

Under the high-dimensional limit (8), we will show that the estimation/generalization error achieved by the BO estimators are characterized through the *overlaps* q, Q defined as

$$q_{\ell k} = \frac{1}{d}\mathbb{E}_{S\sim P(S|\mathcal{D})}\left[\operatorname{Tr} S_\ell S_k^*\right] \,, \qquad Q_{\ell k} = \frac{1}{d}\mathbb{E}_{S\sim P_S}\left[\operatorname{Tr} S_\ell^* S_k^*\right] \,. \tag{13}$$

3 Statistical and computational limits for AIMs

In this Section we provide results for the information-theoretical and computational limits on attention-indexed models in a very general setting, which we then specify to the attention models in Section 4.

In order to state our results let us define, for two $L \times L$ symmetric positive-definite matrix $q, \hat{q} \in \mathbb{S}_L^+$

$$\mathcal{Z}_{\rm in}(\{Y_\ell\}_{\ell=1}^L; \hat{q}) = \int \left[\prod_{\ell=1}^L \mathrm{d}S_\ell \, P_S(S_\ell)\right] \exp\Big[-\tfrac{d}{4}\sum_{\ell,k=1}^L \hat{q}_{\ell k}\,\mathrm{Tr}(S_\ell S_k) + \tfrac{d}{2}\sum_{\ell,k=1}^L \sqrt{\hat{q}}_{\ell k}\,\mathrm{Tr}(Y_\ell S_k)\Big].$$
$$I_{\rm in}(\hat{q}) = \lim_{d\to\infty}\frac{1}{d^2}\int DY_1\dots DY_L\, \mathcal{Z}_{\rm in}(Y_1,\dots,Y_L;\hat{q})\log \mathcal{Z}_{\rm in}(Y_1,\dots,Y_L;\hat{q}) \tag{14}$$

where DY stands for integration over a $\mathrm{GOE}(d)$ (Wigner) matrix Y, and

$$\mathcal{Z}_{\rm out}(y,\omega,V) = \int \left[\prod_{a\le b}^T d^L h_{ab}\,\mathcal{N}\left(h_{ab};\omega_{ab};V_{ab}\right)\right]\delta\left(y - g(\{h_{ab}^{(\ell)}/\sqrt{2-\delta_{ab}}\}_{\ell=1}^L)\right)$$
$$I_{\rm out}(q) = \int\prod_{a,b=1}^T dy_{ab}\int \mathcal{D}\eta_1\dots\mathcal{D}\eta_L\,\mathcal{Z}_{\rm out}(y,\omega,V)\log\mathcal{Z}_{\rm out}(y,\omega,V) \tag{15}$$
$$\text{with:}\quad \omega_{ab}^{(\ell)} = \sum_{k=1}^L \sqrt{2q}_{\ell k}\,\eta_{ab}^{(k)}\,,\qquad V^{(\ell k)} = 2(Q_{\ell k} - q_{\ell k})$$

where $\mathcal{D}\eta$ stands for integration over a $L\times T\times T$ tensor symmetric in the token indices and with independent entries $\mathcal{N}(0,1)$. Moreover, let us define

$$g_{\rm out}(y,\omega,V) = \partial_\omega \ln \mathcal{Z}_{\rm out}(y,\omega,V)\,,\qquad g_{\rm in}(Y|\hat{q}) = \partial_{\hat{q}^{1/2}Y}\ln\mathcal{Z}_{\rm in}(\hat{q})\,. \tag{16}$$

Our first result provides a description of the error in the high dimensional $d\to\infty$ limit, with finite sample complexity $\alpha = n/d^2$, number of tokens T and number of layers L.

Result 3.1 (Performance of information-theoretically optimal estimators)**.** *Consider the extremization problem*

$$\inf_{\hat{q}\in\mathbb{S}_L^+}\sup_{q\in\mathbb{S}_L^+}\left\{-\frac{\mathrm{Tr}\,q\hat{q}}{4} + I_{\rm in}(\hat{q}) + \alpha I_{\rm out}(q)\right\} \tag{17}$$

where $I_{\rm in}(\hat{q})$ and $I_{\rm out}(q)$ are defined in (14) and (15), Call $(q^*,\hat{q}^*)$ the global extremizer of (17). Then, in the high dimensional limit (8), the BO estimation error is given by

$$\lim_{d\to\infty}\mathcal{E}_{\rm est}(\hat{S}_{\rm BO}) = \|Q - q^*\|_F^2\,, \tag{18}$$

while the BO generalization error is given by

$$\mathcal{E}_{\rm gen} = \mathbb{E}_{\eta,\xi}\left\|g\left(\left\{\frac{\omega_{ab}^{(\ell)}}{\sqrt{2-\delta_{ab}}}\right\}_{a\le b,\ell=1}^{T,L}\right) - g\left(\left\{\frac{h(\omega^{(\ell)},V^{(\ell k)})_{ab}}{\sqrt{2-\delta_{ab}}}\right\}_{a\le b,\ell=1}^{T,L}\right)\right\|_F^2, \tag{19}$$

where ξ a $L\times T\times T$ tensor symmetric in the token indices and with independent entries $\mathcal{N}(0,1)$ and ω, V, η are defined in (15). Finally $h(\omega^{(\ell)},V^{(\ell k)}) = \omega^{(\ell)} + \sum_{k=1}^L \sqrt{V}_{\ell k}\,\xi^{(k)}$.

These expressions provide a prediction for the information-theoretically optimal estimation and generalization errors, and as such they constitute a sharp information theoretical bound on the performance of any algorithm. We derive Result 3.1 in Appendix C using the heuristic replica method, but believe that a rigorous treatment is possible by adapting [41] to the case of multiple attention indices $L > 1$ and multiple tokens $T > 1$.

In our next result we provide an efficient polynomial-time approximate message passing (AMP) algorithm that in the high-dimensional limit saturates this bound under the condition that there is a unique local extremizer of (17) (if multiple extermizers are present, computational-to-statistical gaps may arise [44]).

Algorithm 1: AMP

Result: The estimators $\hat{S}_\ell$
Input: Observations $y^\mu \in \mathbb{R}^{T\times T}$ and "sensing matrices"
$Z^\mu_{ij,ab} \equiv (\boldsymbol{x}^\mu_{i,a}\boldsymbol{x}^\mu_{j,b} + \boldsymbol{x}^\mu_{j,a}\boldsymbol{x}^\mu_{i,b} - 2\delta_{ij}\delta_{ab})/\sqrt{2d(1+\delta_{ab})} \in \mathbb{R}$;
Initialize $\hat{S}^{t=0}_\ell \sim P_S$ and $\hat{C}^{t=0} = 2(\kappa_2 - \kappa_1^2)\mathbb{I}_L$;
while *not converging* **do**

- *Estimation of the variance and mean of* $\mathrm{Tr}[Z^\mu_{ab}S_\ell]$;

$V^t = 2\hat{C}^t$ and $\omega^t_{\mu,\,ab} = \mathrm{Tr}[Z^\mu_{ab}\hat{S}^t] - (1-\delta_{0t})g_{\rm out}(y^\mu, \omega^{t-1}_\mu, V^{t-1})_{ab}V^t \in \mathbb{R}^L$;

- *Variance and mean of* S_ℓ *estimated from the "output" observations*;

$$\hat{q}^t_{\ell k} = \frac{4\alpha}{n}\sum_{\mu,a\le b}^{n,T,T} g_{\rm out}(y^\mu, \omega^t_\mu, V^t)^{(\ell)}_{ab}\, g_{\rm out}(y^\mu, \omega^t_\mu, V^t)^{(k)}_{ab} \quad \text{and}$$

$$R^t_{ij} = \hat{S}^t_{ij} + (\hat{q}^t)^{-1}\frac{2}{d}\sum_{\mu,a\le b}^{n,T,T} g_{\rm out}(y^\mu, \omega^t_\mu, V^t)_{ab}Z^\mu_{ij,ab} \in \mathbb{R}^L;$$

- *Update of the estimation of* S *with the "input" information*;

$\hat{S}^{t+1}_\ell = g_{\rm in}\left(R^t, \hat{q}^t\right)_\ell$ and $\hat{C}^{t+1}_{\ell k} = \frac{1}{d^2}\nabla_{R_k}\cdot g_{\rm in}\left(R^t, \hat{q}^t\right)_\ell$;

$t = t+1$;

end

Result 3.2 (State evolution of AMP)**.** *Call* $\hat{S}^t_\ell$ *the time-*t *iterate of the AMP algorithm 1. In the high dimensional limit* (8) *of large* d *and for a finite number of iterations, we have that*

$$\frac{1}{d}\mathrm{Tr}[\hat{S}^t_\ell \hat{S}^t_k] \to q^t_{\ell k}\,, \qquad \frac{1}{d}\mathrm{Tr}[\hat{S}^t_\ell S^*_k] \to q^t_{\ell k}\,, \tag{20}$$

where

$$\begin{aligned}
\hat{q}^{t+1}_{\ell k} = 4\alpha\mathbb{E}_{\xi,\eta}\sum_{a\le b}^{T} & g_{\rm out}\left(g\left(\left\{\frac{h(\omega^t, V^t)_{ab}}{\sqrt{2-\delta_{ab}}}\right\}\right), \omega^t, V^t\right)^{(\ell)}_{ab} \\
& \times g_{\rm out}\left(g\left(\left\{\frac{h(\omega^t, V^t)_{ab}}{\sqrt{2-\delta_{ab}}}\right\}\right), \omega^t, V^t\right)^{(k)}_{ab}\,, \\
q^{t+1}_{\ell k} = \lim_{d\to+\infty}\frac{1}{d}\mathbb{E}_{S,Y}\,\mathrm{Tr}\left[g_{\rm in}(Y(S,\hat{q}^{t+1}),\hat{q}^{t+1})_\ell g_{\rm in}(Y(S,\hat{q}^{t+1}),\hat{q}^{t+1})_k\right]\,,
\end{aligned} \tag{21}$$

with $V^t = 2(Q-q^t)$ *and* $\omega^t_\ell = \sum_{k=1}^L(\sqrt{2q^t})_{\ell k}\,\eta_k$*, where* (a,b) *are token and* (ℓ,k) *layer indices and with* η_ℓ *a standard Gaussian in every component.* $h(\omega^t, V^t)$, η *and* ξ *are defined as in Result 3.1. Finally we have* $Y(S,\Delta)_\ell = S_\ell + \sum_{m=1}^L\sqrt{\Delta}_{\ell m}\Xi_m$*, all* Ξ_m *are* $\mathrm{GOE}(d)$ *and all* $S_\ell \sim P_S$.

Result 3.2 (which we derive in Appendix C) is a classic statement in the theory of AMP, that here we adapt to take into account multiple attention indices and multiple tokens. We remark that even though the denoiser $g_{\rm in}$ in (16) is a complicated non-separable function, state evolution still holds [45, 46]. Additionally, we show in Appendix C that the fixed point of (21) are the extremisers of (17). The AMP algorithm 1 is thus both a tool that gives us theoretical guarantees though (21) and a practical algorithm that can be efficiently implemented and run on a machine. We discuss the implementation in Appendix C.

The prior-related elements $I_{\rm in}$ and $g_{\rm in}$ of (14) and (21) require discussion. Both of them involve integrals in dimension $\mathcal{O}(Ld^2)$ whose large d asymptotics is highly non-trivial.

This is due to the prior P_S on the weights being non-separable, and crucially to the coupling between weights of different layers caused by the non-diagonal matrix $\hat{q}$ in (14) and (21). We remark that for a generic number of layers $L > 1$, the evaluation of these two equations is equivalent to the evaluation

in the high-dimensional limit of the free entropy and of the Bayes-optimal estimator for the following L-wise matrix denoising problem

$$Y_\ell = S_\ell + \sum_{m=1}^{L} \sqrt{\Delta}_{\ell m} \Xi_m \,, \tag{22}$$

where from a heavily correlated set of L noisy observations $Y_\ell \in \mathbb{R}^{d\times d}$, $\ell = 1, \dots, L$, one needs to estimate back the L independent matrices $S_m \sim P_S$. Here $\Delta \in \mathbb{S}_L^+$ is an L-dimensional symmetric positive-definite matrix acting as a noise-to-signal ratio, and Ξ_ℓ are independent $\mathrm{GOE}(d)$ matrices acting as noise. To the best of our knowledge, for $L > 1$ and non-diagonal Δ this is still a challenging open problem (in case of diagonal Δ, the problem factorizes in L independent matrix denoising problems). In practice, one can approximate rotationally-invariant matrix denoisers by low-degree spectral polynomials, following [47]; this provides a practical route to AMP for multi-layer $\hat{q}$ when the exact HCIZ-based denoiser is unavailable [48].

For $L = 1$, the problem (22) reduces to Bayes-optimal denoising of a rotationally invariant matrix [40, 43]. We start by remarking that for $L = 1$ the quantities $Q, q, \hat{q}$ governing the BO performance are scalar.

For Result (3.1) at leading order in $d \to \infty$ we have

$$I_{\rm in}(\hat{q}) = \frac{Q}{4}\hat{q} - \frac{1}{4}\log\hat{q} - \frac{1}{2}\Sigma(\mu_{1/\hat{q}}) - \frac{1}{8} \quad \text{where} \quad \Sigma(\mu) = \mathbb{E}_{x,y\sim\mu} \log|x-y| \,, \tag{23}$$

and for the AMP algorithm Result (3.2) we have

$$g_{{\rm in},L=1}(X,\Delta) = O^\top f_{\rm RIE}(\Lambda,\Delta)O \quad \text{where} \quad f_{\rm RIE}(\Lambda,\Delta)_i = \Lambda_i - 2\Delta \int \frac{\mathrm{d}\mu_\Delta(t)}{\Lambda_i - t} \,, \tag{24}$$

where $Q = \kappa_2$, μ_Δ is the asymptotic spectral density of a matrix $X = S + \sqrt{\Delta}Z$ with $S \sim P_S$ and $Z \sim \mathrm{GOE}(d)$, and $X = O^\top \Lambda O$ is the eigen-decomposition of X (see [40] for details).

Similarly, at leading order in $d \to \infty$ we have

$$\frac{1}{d^2}\nabla_X \cdot g_{{\rm in},L=1}(X,\Delta) = \Delta - \frac{4\pi^2\Delta^2}{3}\int \mathrm{d}\mu_\Delta(t)^3 \,. \tag{25}$$

Finally, the state evolution equation for q in (21) reads

$$q = Q - \frac{1}{\hat{q}} + \frac{4\pi^2}{3\hat{q}^2}\int \mathrm{d}t\,[\mu_{1/\hat{q}}(t)]^3 \,. \tag{26}$$

4 Results for single-layer attention

We now apply our general results to the single-layer ($L = 1$) tied-attention model

$$y_{ab} = \sigma_\beta\left(\frac{\boldsymbol{x}_a^\top S\, \boldsymbol{x}_b - \delta_{ab}\,\mathrm{Tr}\,S}{\sqrt{d}}\right) = \sigma_\beta\left(\frac{\frac{1}{\sqrt{rd}}\boldsymbol{x}_a^\top W W^\top\, \boldsymbol{x}_b - \delta_{ab}\,\mathrm{Tr}\big(\frac{1}{\sqrt{rd}}WW^\top\big)}{\sqrt{d}}\right) \tag{27}$$

where we parametrized the weight matrix S as a tied-attention with extensive-width $r = \rho d$ and $W \in \mathbb{R}^{d\times r}$ has independent entries $W_{ij} \sim \mathcal{N}(0,1)$. For the activation, we consider the case of Hardmax $\sigma_{\rm hard}$ and Softmax $\sigma_{\rm soft}$, both applied row-wise in (27):

$$\sigma_{\rm hard}(z_1 \dots z_T)_i = \delta(i = \arg\max_j x_j) \,, \quad \text{and} \quad \sigma_{\rm soft}(z_1 \dots z_T)_i = \frac{e^{\beta z_i}}{\sum_{j=1}^T e^{\beta z_j}} \,. \tag{28}$$

We stress that both these tasks are well-defined only for $T \geq 2$, as the $T = 1$ the output of both activations equals 1 regardless of the input. As discussed in the introduction, the model with hardmax provides an interesting token-association task.

Hardmax target. The BO treatment of the hardmax activation for generic number of tokens T is challenging due to the complex form of the state equation for $\hat{q}$ (21). We provide an explicit solution in the $T = 2$ case.

Result 4.1 (Bayes-optimal errors for hardmax tied-attention, $T = 2$)**.** *Consider the model* (27) *with hardmax activation. In the high-dimensional limit* (8)*, the asymptotic BO estimation and generalization errors are given by* (18) *and* (19)*, where* $(q, \hat{q})$ *is the solution of* (26) *with* $Q = 1 + \rho$ *and*

$$g_{\text{out}}(y, \omega, V)_{ab} = \frac{1}{\sqrt{6(Q-q)}} \frac{\phi(k_1, k_2, c)}{\Phi(k_1, k_2, c)} \begin{pmatrix} \sqrt{2}s_1 & -(s_1+s_2) \\ -(s_1+s_2) & \sqrt{2}s_2 \end{pmatrix}_{ab}, \tag{29}$$

where $\phi(k_1, k_2, c)$ *is the p.d.f. of a bi-variate Gaussian with zero mean, variances* $1/(1-c^2)$ *and covariance* $c/(1-c^2)$*, and* $\Phi(k_1, k_2, c)$ *is its c.d.f (see Appendix A). Moreover,* $s_a = 2y_{aa} - 1$*,* $k_a = s_a(\sqrt{2}\omega_{aa} - \omega_{12})/\sqrt{6(Q-q)}$*,* $c = s_1 s_2/3$ *and* $\omega_{ab} = \sqrt{2q}\,\eta_{ab}$*.*

We detail the derivation of Result 4.1 in App. C. We plot the estimation error given in Result (4.1) in Figure 1 left, for several values of the attention width ratio ρ, comparing with runs of the associated AMP Algorithm 1 at size $d = 100$.

We observe that for all finite α the estimation error is strictly positive, and that it approaches zero as α grows with rate compatible with $\mathcal{O}(1/\alpha)$. Moreover, as soon as $\alpha > 0$, we observe that the estimation error is smaller than 1, i.e. the value achieved in the absence of data. Intuitively, the hardmax output function enforces discrete token association per row. This is akin to a multiclass classification target, which explains the lack of a finite strong-recovery threshold and the observed power-law decay at large α (Fig. 1, left).

In the limit of small width Result 4.1 simplifies. Notice that in this limit the correct sample scale is given by $\bar{\alpha} = \alpha/\rho = n/(dr)$, as the matrix to infer is not extensive-width anymore. In this limit there appears a so–called weak recovery threshold, a value of sample complexity below which the estimator reaches the same performance as if there were no data. We characterize it as follows.

Corollary 4.2 (Small width limit for hardmax activation)**.** *Consider the model* (27) *with hardmax activation and* $T = 2$*. In the high-dimensional limit* (8)*, the equation for* q *of Result 3.1 simplifies to* $q = [\max(1-t, 0)]^2$ *under the rescaling* $\alpha = \bar{\alpha}\rho$ *and* $\hat{q} = t\rho$*. In particular, the BO error is the same as that of the data-less estimator for all* $\bar{\alpha} < \bar{\alpha}_{\text{weak}}^{\text{hardmax}}$ *where*

$$\bar{\alpha}_{\text{weak}}^{\text{hardmax}} = \frac{1}{4\mathbb{E}_{y,\omega}\left[\sum_{a\le b}^{T=2} g_{\text{out}}(y(\omega, V), \omega, V)_{ab}^{\otimes 2}\right]_{q=0, Q=1}} \approx 0.563\,. \tag{30}$$

Corollary 4.2 follows by combining Result 4.2 with the small-width analysis of [40]. We remark that (30) holds for all activation (using the appropriate g_{out} function) and all values of $T \geq 2$. We plot the analytical prediction for the BO estimation error given in Corollary 4.2 in Figure 1 right.

Softmax target. We now discuss the target function that uses a softmax non-linearity (28). This choice of activation allows for an analytic treatment for any number of tokens $T \geq 2$, and any finite value of the softmax inverse temperature $\beta \in \mathbb{R}_+$.

Result 4.3 (Bayes-optimal errors for softmax tied-attention, $T \geq 2$)**.** *Consider the model* (27) *with softmax activation,* $T \geq 2$ *and inverse temperature* $\beta \in \mathbb{R}_+$*. In the high-dimensional limit* (8)*, the asymptotic BO estimation and generalization errors are given by* (18) *and* (19)*, where* $(q, \hat{q}) \in \mathbb{R}_+^2$ *is the solution of* (26)*,* $Q = 1 + \rho$ *and* $\hat{q} = \alpha(T^2 + T - 2)/(Q - q)$*.*

We plot the BO estimation error given in Result 4.3 in Figure 2 left, and observe that contrary to the hardmax case, the BO estimation error vanishes at a finite value of α (the so-called strong recovery threshold). Interestingly, the BO errors given in Result (4.3) is independent of the value of the inverse temperature β and reduces to the case of a single-token model with linear activation [40], modulo a rescaling of the sample ratio α to $2\alpha/(T^2 + T - 2)$ (notice that the rescaling is not just given by the total number unordered couples of tokens $T(T+1)/2$, as it would be in the case of a multi-token case with bijective activation, see App. C). The softmax activation is almost invertible, meaning that given the output, the input is fully determined apart for a common additive shift (acting as a noise correlated with the data), and is additionally constrained by the symmetry of the attention matrix. Result (4.3) precisely quantifies the amount of samples required to estimate this undetermined shift. More precisely, fix a given estimation error. Then, achieving this error with the BO estimator in the softmax case with $T \geq 2$ requires a factor $1 + 2/(T(T+1))$ more samples than the case of a fully bijective activation.

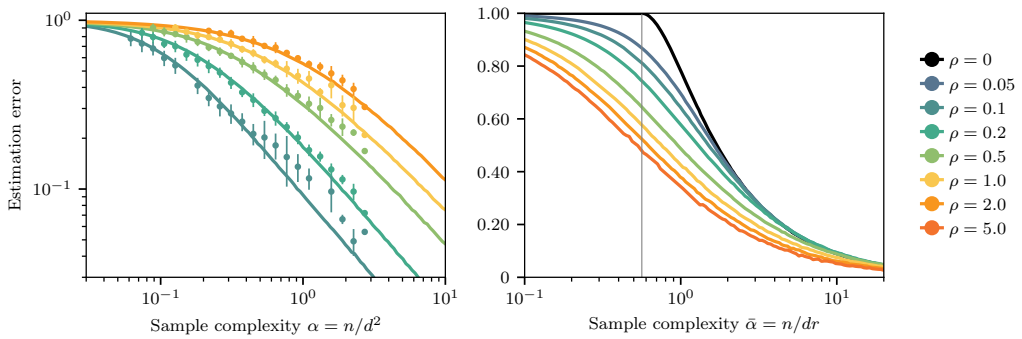

Figure 1: **(Left)** The Bayes optimal-error for the single-layer attention-indexed model with $T = 2$ tokens and hardmax activation for and several values of the width ratio ρ (Result (4.1)). The log-log scale highlights a large α power-law decay of the BO estimation error, strikingly different from the softmax behaviour (see Figure 2). We also plot the corresponding errors achieved by the AMP algorithm (dots) at $d = 100$, averaged over 16 realizations of the data and teacher weights. Error bars are computed with respect to the mean. We find a good agreement even for such a moderate size. **(Right)** Focus on the small width Bayes-optimal error case (Corollary (4.2)) of the same model. We rescale the sample complexity to $\bar{\alpha} = \alpha/\rho$, and we highlight the theoretical prediction of the weak recovery thresholds (gray vertical line).

On the other hand, we remark that the AMP algorithm for the softmax activation at $T \geq 2$ *is not* a simple rescaling of the AMP for the single-token linear-activation case given in [40]. The AMP output function g_{out} is given by

$$g_{\text{out},ab} = -\frac{\tau_{ab}}{T^2 V}\Big[\sum_{c\leq d}^{T}[\tau_{cd}^2\phi_{Tc} - \tau_{cd}\omega_{cd}] + \sum_{c\leq d}^{T-1}\tau_{cd}^2\phi_{cd}\Big] + \frac{\tau_{ab}\phi_{Ta} - \omega_{ab} + (1-\delta_{bT})\phi_{ab}\tau_{ab}}{V}\,, \quad (31)$$

where $\tau_{ab} = \sqrt{2-\delta_{ab}}$, $\phi(y)_{ab} = \beta^{-1}\log(y_{ab}/y_{aT})$ and ω, V defined in (15). Thus, AMP processes the data in a non-trivial, optimal way to perform this effective inversion of the softmax activation. We plot experiments for AMP at $d = 100$ in the $T = 2, 3$ case in Figure 2 right (purple and blue dots, to be compared with the prediction of Result (4.3) given by the black line), and observe a nice agreement. We also remark that while the BO performance is independent of the inverse temperature β, as long as it is finite, again AMP output function is not.

Thanks to the mentioned reduction, one can transfer directly several results from [40] to the case of softmax tied-attention, including an explicit prediction for the strong recovery threshold (the value of α after which the BO error is zero), the slope of the error at strong recovery, and the small-width and large-width limits (see App. D). In particular, the strong recovery threshold satisfies

$$\alpha_{\text{recovery}}^{\text{softmax}} = \frac{2}{T^2+T-2}\begin{cases}\rho - \rho^2/2 & \text{if} \quad 0<\rho<1\\ 1/2 & \text{if} \quad \rho \geq 1\end{cases}\,. \quad (32)$$

We remark again that this threshold does not coincide with the naive counting argument, which would give a factor $\frac{T(T+1)}{2}$ at denominator instead. In particular, the $2/(T^2+T-2) = \frac{T(T+1)}{2} - 1$ factor reflects the near-invertibility of the row-wise softmax under a global row-shift, modulo the symmetry constraint; cf. App. D. Finally, contrarily to hardmax (which implements a discrete winner-takes-all assignment and exhibits power-law error decay), softmax behaves like a smooth regression target with a finite strong-recovery threshold (Eq. (32)), see Fig. 1 vs Fig. 2.

We finally remark that our analysis keeps T, L finite while $d, n \to \infty$ with $n/d^2 = \alpha = \Theta(1)$. The predictions remain accurate as long as T grows much slower than d; the curves obtained after the T-dependent rescaling of α (Sec. 4) remain valid for very large T provided $T << d$.

Finally, we consider the performance of gradient descent minimizing the loss

$$\mathcal{L}(W) = \sum_{\mu=1}^{n}\sum_{a,b=1}^{T}\left(y_{ab}^{\mu} - \sigma_\beta\left(\frac{\boldsymbol{x}_a^{\mu\top} W W^\top \boldsymbol{x}_b^{\mu} - \delta_{ab}\,\mathrm{Tr}\,WW^\top}{\sqrt{r}\,d}\right)\right)^2\,, \quad (33)$$

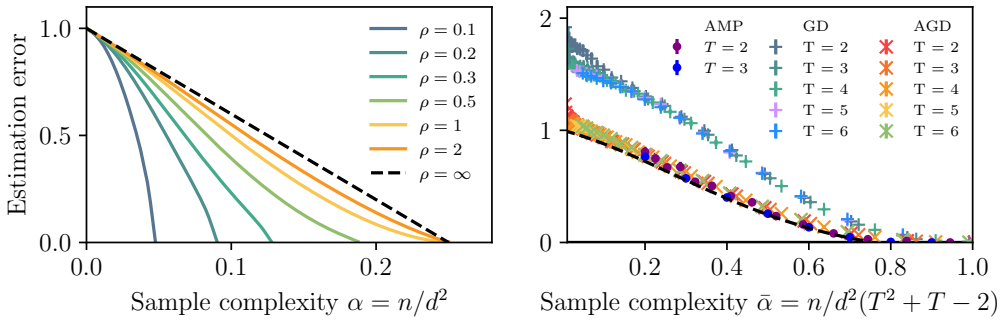

Figure 2: **(Left)** Illustration of the Bayes-optimal estimation error for the softmax tied-attention model (Result 4.3), eq. (27) for any $0 < \beta < +\infty$ and $T = 2$ tokens, and for several values of the attention width ratio $\rho = r/d$. The model reaches zero BO error at finite α depending on ρ (eq. (32)). **(Right)** We show, in black dashed lines the theoretical prediction of the BO estimation error computed for the sample complexity rescaled by the number of tokens $\alpha/(T^2 + T - 2)$ and $\rho = 0.5$. We show the performance of the corresponding AMP algorithm for $T = 2, 3$ tokens, correctly achieving the BO error. We also compare the BO performance with those of Adam GD and its averaged version AGD with $d = 100$. We average each numerical experiment (GD,AGD,AMP) over 16 realizations of the data and teacher weights. Error bars are the standard deviation on the mean.

with training set generated by Eq. (27) (we optimize using the ADAM optimizer [49]). In line with previous work [40, 42], we also consider the Averaged GD (AGD) estimator given by

$$\hat{S}_{\text{GD,avg}} = \frac{1}{M} \sum_{m=1}^{M} \frac{W_m^{\text{final}} (W_m^{\text{final}})^\top}{\sqrt{rd}} , \tag{34}$$

where we average over M initial matrices $W_m^{(0)}$, and W_m^{final} is the corresponding set of weights at convergence. We plot the results of our numerical experiments at $d = 200$ for both GD and AGD in Figure 2 right. As already observed in [40, 42], AGD reaches performances compatible with the BO estimation error, while GD has worse error. We remark that both variants seem to achieve perfect recovery at the BO threshold (32). This phenomenon, at this point well documented within this class of models, is still not understood.

5 Limitations

The attention-indexed model introduced in this work and its high-dimensional analysis provide promising stepping stone to the analysis of learning in multi-layer attention. We, however, so far only analyzed the Bayes-optimal performance and the associated AMP algorithm in the single layer case, gradient descent was only explored numerically and clearly its theoretical analysis in this class of models is an interesting topic for future work. While the concept of attention-indexed model captures multi-layer attention networks, the matrix denoiser (22) needed to study the optimal performance is a challenging open problem for random matrix theory and thus the analysis of the model for more than one layer remains open. While out results rely on exact statistical physics tools, mathematically rigorous proof of the obtained results for $T \geq 2$ is an open problem. Future work should also explore the inclusion of multiple heads and MLP layers to mimic yet closer practicalities of the current transformer architectures. The structure of the input data considered in this paper is very simple, future work should also include studied of how the performance depends on the input structure.

Acknowledgments and Disclosure of Funding

We thank Antoine Maillard, Florent Krzakala, Hugo Cui and Yizhou Xu for insightful discussions related to this work. We acknowledge funding from the Swiss National Science Foundation grants SNSF SMArtNet (grant number 212049).

References

[1] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

[2] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies, volume 1 (long and short papers)*, pages 4171–4186, 2019.

[3] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901, 2020.

[4] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale, 2021.

[5] David L Donoho et al. High-dimensional data analysis: The curses and blessings of dimensionality. *AMS math challenges lecture*, 1(2000):32, 2000.

[6] Gerard Ben Arous, Reza Gheissari, and Aukosh Jagannath. Online stochastic gradient descent on non-convex losses from high-dimensional inference. *Journal of Machine Learning Research*, 22(106):1–51, 2021.

[7] Emmanuel Abbe, Enric Boix Adsera, and Theodor Misiakiewicz. The merged-staircase property: a necessary and nearly sufficient condition for sgd learning of sparse functions on two-layer neural networks. In *Conference on Learning Theory*, pages 4782–4887. PMLR, 2022.

[8] Rodrigo Veiga, Ludovic Stephan, Bruno Loureiro, Florent Krzakala, and Lenka Zdeborová. Phase diagram of stochastic gradient descent in high-dimensional two-layer neural networks. *Advances in Neural Information Processing Systems*, 35:23244–23255, 2022.

[9] Jimmy Ba, Murat A Erdogdu, Taiji Suzuki, Zhichao Wang, Denny Wu, and Greg Yang. High-dimensional asymptotics of feature learning: How one gradient step improves the representation. *Advances in Neural Information Processing Systems*, 35:37932–37946, 2022.

[10] Luca Arnaboldi, Ludovic Stephan, Florent Krzakala, and Bruno Loureiro. From high-dimensional & mean-field dynamics to dimensionless odes: A unifying approach to sgd in two-layers networks. In *The Thirty Sixth Annual Conference on Learning Theory*, pages 1199–1227. PMLR, 2023.

[11] Elizabeth Collins-Woodfin, Courtney Paquette, Elliot Paquette, and Inbar Seroussi. Hitting the high-dimensional notes: An ode for sgd learning dynamics on glms and multi-index models. *Information and Inference: A Journal of the IMA*, 13(4):iaae028, 2024.

[12] Alex Damian, Eshaan Nichani, Rong Ge, and Jason D Lee. Smoothing the landscape boosts the signal for sgd: Optimal sample complexity for learning single index models. *Advances in Neural Information Processing Systems*, 36, 2024.

[13] Alberto Bietti, Joan Bruna, and Loucas Pillaud-Vivien. On learning gaussian multi-index models with gradient flow part i: General properties and two-timescale learning. *Communications on Pure and Applied Mathematics*, 78(12):2354–2435, 2025.

[14] Behrad Moniri, Donghwan Lee, Hamed Hassani, and Edgar Dobriban. A theory of non-linear feature learning with one gradient step in two-layer neural networks. In *Proceedings of the 41st International Conference on Machine Learning*, pages 36106–36159, 2024.

[15] Raphaël Berthier, Andrea Montanari, and Kangjie Zhou. Learning time-scales in two-layers neural networks. *Foundations of Computational Mathematics*, pages 1–84, 2024.

[16] Elizabeth Gardner and Bernard Derrida. Three unfinished works on the optimal storage capacity of networks. *Journal of Physics A: Mathematical and General*, 22(12):1983, 1989.

[17] Haim Sompolinsky, Naftali Tishby, and H Sebastian Seung. Learning from examples in large neural networks. *Physical Review Letters*, 65(13):1683, 1990.

[18] T.L.H. Watkin, A. Rau, and M. Biehl. The statistical mechanics of learning a rule. *Reviews of Modern Physics*, 65(2):499–556, 1993.

[19] Sebastian Goldt, Madhu Advani, Andrew M Saxe, Florent Krzakala, and Lenka Zdeborová. Dynamics of stochastic gradient descent for two-layer neural networks in the teacher-student setup. *Advances in neural information processing systems*, 32, 2019.

[20] Andreas Engel. *Statistical mechanics of learning*. Cambridge University Press, 2001.

[21] Benjamin Aubin, Antoine Maillard, Florent Krzakala, Nicolas Macris, Lenka Zdeborová, et al. The committee machine: Computational to statistical gaps in learning a two-layers neural network. *Advances in Neural Information Processing Systems*, 31, 2018.

[22] 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. *Advances in Neural Information Processing Systems*, 37:36342–36389, 2024.

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

[24] Hugo Cui, Florent Krzakala, and Lenka Zdeborová. Bayes-optimal learning of deep random networks of extensive-width. *Journal of Statistical Mechanics: Theory and Experiment*, 2025(1):014001, January 2025. Publisher: IOP Publishing.

[25] Bingqing Song, Boran Han, Shuai Zhang, Jie Ding, and Mingyi Hong. Unraveling the gradient descent dynamics of transformers. *Advances in Neural Information Processing Systems*, 37:92317–92351, 2024.

[26] Yuandong Tian, Yiping Wang, Beidi Chen, and Simon S Du. Scan and snap: Understanding training dynamics and token composition in 1-layer transformer. *Advances in Neural Information Processing Systems*, 36:71911–71947, 2023.

[27] Ashok Vardhan Makkuva, Marco Bondaschi, Adway Girish, Alliot Nagle, Hyeji Kim, Michael Gastpar, and Chanakya Ekbote. Local to global: Learning dynamics and effect of initialization for transformers. *Advances in Neural Information Processing Systems*, 37:86243–86308, 2024.

[28] Eshaan Nichani, Alex Damian, and Jason D Lee. How transformers learn causal structure with gradient descent. In *Proceedings of the 41st International Conference on Machine Learning*, pages 38018–38070, 2024.

[29] Hongru Yang, Bhavya Kailkhura, Zhangyang Wang, and Yingbin Liang. Training dynamics of transformers to recognize word co-occurrence via gradient flow analysis. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024.

[30] Jiri Hron, Yasaman Bahri, Jascha Sohl-Dickstein, and Roman Novak. Infinite attention: Nngp and ntk for deep attention networks. In *International Conference on Machine Learning*, pages 4376–4386. PMLR, 2020.

[31] Hengyu Fu, Tianyu Guo, Yu Bai, and Song Mei. What can a single attention layer learn? a study through the random features lens. *Advances in Neural Information Processing Systems*, 36:11912–11951, 2023.

[32] Yue M Lu, Mary I Letey, Jacob A Zavatone-Veth, Anindita Maiti, and Cengiz Pehlevan. Asymptotic theory of in-context learning by linear attention. *CoRR*, 2024.

[33] Blake Bordelon, Hamza Chaudhry, and Cengiz Pehlevan. Infinite limits of multi-head transformer dynamics. *Advances in Neural Information Processing Systems*, 37:35824–35878, 2024.

[34] Lorenzo Noci, Chuning Li, Mufan Li, Bobby He, Thomas Hofmann, Chris J Maddison, and Dan Roy. The shaped transformer: Attention models in the infinite depth-and-width limit. *Advances in Neural Information Processing Systems*, 36:54250–54281, 2023.

[35] 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. *Forty-second International Conference on Machine Learning*, 2025.

[36] Pierre Marion, Raphaël Berthier, Gérard Biau, and Claire Boyer. Attention layers provably solve single-location regression. *arXiv preprint arXiv:2410.01537*, 2024.

[37] Emanuele Troiani, Yatin Dandi, Leonardo Defilippis, Lenka Zdeborova, Bruno Loureiro, and Florent Krzakala. Fundamental computational limits of weak learnability in high-dimensional multi-index models. In *International Conference on Artificial Intelligence and Statistics*, pages 2467–2475. PMLR, 2025.

[38] Antoine Maillard and Afonso S Bandeira. Exact threshold for approximate ellipsoid fitting of random points. *arXiv preprint arXiv:2310.05787*, 2023.

[39] Antoine Maillard and Dmitriy Kunisky. Fitting an ellipsoid to random points: predictions using the replica method. *IEEE Transactions on Information Theory*, 2024.

[40] Antoine Maillard, Emanuele Troiani, Simon Martin, Lenka Zdeborová, and Florent Krzakala. Bayes-optimal learning of an extensive-width neural network from quadratically many samples. *Advances in Neural Information Processing Systems*, 37:82085–82132, December 2024.

[41] Yizhou Xu, Antoine Maillard, Lenka Zdeborová, and Florent Krzakala. Fundamental limits of matrix sensing: Exact asymptotics, universality, and applications. In Nika Haghtalab and Ankur Moitra, editors, *Proceedings of Thirty Eighth Conference on Learning Theory*, volume 291 of *Proceedings of Machine Learning Research*, pages 5757–5823. PMLR, 30 Jun–04 Jul 2025.

[42] Vittorio Erba, Emanuele Troiani, Luca Biggio, Antoine Maillard, and Lenka Zdeborová. Bilinear sequence regression: A model for learning from long sequences of high-dimensional tokens. *Physical Review X*, 15(2):021092, 2025.

[43] Emanuele Troiani, Vittorio Erba, Florent Krzakala, Antoine Maillard, and Lenka Zdeborová. Optimal denoising of rotationally invariant rectangular matrices. In Bin Dong, Qianxiao Li, Lei Wang, and Zhi-Qin John Xu, editors, *Proceedings of Mathematical and Scientific Machine Learning*, volume 190 of *Proceedings of Machine Learning Research*, pages 97–112. PMLR, 15–17 Aug 2022.

[44] Lenka Zdeborová and Florent Krzakala. Statistical physics of inference: Thresholds and algorithms. *Advances in Physics*, 65(5):453–552, 2016.

[45] Raphael Berthier, Andrea Montanari, and Phan-Minh Nguyen. State evolution for approximate message passing with non-separable functions. *Information and Inference: A Journal of the IMA*, 9(1):33–79, 2020.

[46] Cédric Gerbelot and Raphaël Berthier. Graph-based approximate message passing iterations. *Information and Inference: A Journal of the IMA*, 12(4):2562–2628, 2023.

[47] Guilhem Semerjian. Matrix denoising: Bayes-optimal estimators via low-degree polynomials. *Journal of Statistical Physics*, 191(10):139, 2024.

[48] Joël Bun, Romain Allez, Jean-Philippe Bouchaud, and Marc Potters. Rotational invariant estimator for general noisy matrices. *IEEE Transactions on Information Theory*, 62(12):7475–7490, 2016.

[49] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In Yoshua Bengio and Yann LeCun, editors, *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*, 2015.

[50] Hidetoshi Nishimori. Exact results and critical properties of the ising model with competing interactions. *Journal of Physics C: Solid State Physics*, 13(21):4071, 1980.

[51] Antoine Maillard, Florent Krzakala, Marc Mézard, and Lenka Zdeborová. Perturbative construction of mean-field equations in extensive-rank matrix factorization and denoising. *Journal of Statistical Mechanics: Theory and Experiment*, 2022(8):083301, 2022.

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: The paper aims at studying Attention Index Models and its connection with self attention mechanism. We do so by providing analytical and numerical results.

   Guidelines:

   - The answer NA means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: We state our limitations in the main.

   Guidelines:

   - The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

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: We presented in this paper exact claims and conjectures based on well-known analytical methods. Specifically, we developed an application of the so-called replica method in this new context, so we can think of our results at the level of rigor of theoretical physics. A fully rigorous derivation of our results is a very technical and lengthy avenue, and is left for a more suitable mathematical venue.

Guidelines:

- The answer NA means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

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 release the code and discuss the details of the implementation in the supplementary material.

Guidelines:

- The answer NA means that the paper does not include experiments.
- If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
- If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
- Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
  (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
  (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
  (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
  (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

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: [Yes]

   Justification: We release the code and the data for reproducing the figures.

   Guidelines:

   - The answer NA means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (https://nips.cc/public/guides/CodeSubmissionPolicy) for more details.
   - While we encourage the release of code and data, we understand that this might not be possible, so "No" is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
   - The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (https://nips.cc/public/guides/CodeSubmissionPolicy) for more details.
   - The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
   - The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
   - At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
   - Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

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, etc.) necessary to understand the results?

   Answer: [Yes]

   Justification: Yes, in the supplementary material

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

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 provide error bars in each numerical experiment and explain how they were constructed.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
   - The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).

- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
- If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

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: In the supplementary material.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

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: Our work is theoretical in nature, and we respected the code of ethics.

   Guidelines:

   - The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [NA]

    Justification: Our work is theoretical on syntetic data.

    Guidelines:

    - The answer NA means that there is no societal impact of the work performed.
    - If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
    - Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.

- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

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., pretrained language models, image generators, or scraped datasets)?

    Answer: [NA]

    Justification: Our work is theoretical in nature and on synthetic data.

    Guidelines:

    - The answer NA means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

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: [NA]

    Justification: We use synthetic data.

    Guidelines:

    - The answer NA means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
    - If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
    - For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.

- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [NA]

    Justification: We use synthetic data that we generate ourself.

    Guidelines:

    - The answer NA means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

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: [NA]

    Justification: We use synthetic data that we generate ourself.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

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: [NA]

    Justification: We use synthetic data that we generate ourself.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
    - We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
    - For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

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 rigorousness, or originality of the research, declaration is not required.

Answer: [NA]

Justification: The core method development in this research does not involve LLMs as any important, original, or non-standard components.

Guidelines:

- The answer NA means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
- Please refer to our LLM policy (https://neurips.cc/Conferences/2025/LLM) for what should or should not be described.

A Notations and model description

In this appendix we first remind all the notations and settings of the attention-indexed models. We then remind mathematical concepts and definitions that are present in the main text.

Throughout this work, we use $\ell, k = 1, \dots, L$ as the layer index where L is the total number of layer matrices used, while $a, b = 1, \dots, T$ are the token index and T is the total number of tokens. Then $i, j = 1, \dots, d$ are the indices for the dimensions and d is the embedding dimension of each token, and $\mu = 1 \dots n$ is the sample index and n is the total number of samples. We will also use $u, v = 0 \dots m$ as the replica indices from 0 to m.

We list below the specifics of our model:

- $X \equiv X_0 \in \mathbb{R}^{T \times d}$: The matrix of T tokens (rows), each token of embedding dimension d.
- $S_\ell \in \mathbb{R}^{d \times d}$ symmetric matrices for $\ell = 1, \dots, L$ and extracted independently from a rotationally invariant ensemble $P_S(S) = P_S(O^\top S O)$ for any rotation matrix O. We fix the normalizations such that $\mathbb{E}_{P_S}[\operatorname{Tr} S] = \kappa_1 d$ and $\mathbb{E}_{P_S}[\operatorname{Tr} S^2] = \kappa_2 d$ and with $\kappa_1,\ \kappa_2 = \mathcal{O}(1)$. Contextually, we assume that the empirical spectral distribution of S will converge to a well defined measure μ_S. For the purpose of the analysis, we will specify our general framework to symmetric matrices of the form $S_\ell = W^\top W / \sqrt{r_\ell d}$ where $W \in \mathbb{R}^{d \times r}$ with entries $W_{ij} \sim \mathcal{N}(0, 1)$. We refer to the finite quantities $\rho_l > 0$ as the width ratios of each layer.
- We define the AIM as the following model:

$$y = g\left(\{h^{(\ell)}\}_{\ell=1}^{L}\right) \tag{35}$$

  with the generic map $g : \mathbb{R}^{L \times T \times T} \to \mathbb{R}^{T \times T}$ which depends on the quadratic preactivations

$$h_{ab}^{(\ell)} \equiv \frac{\boldsymbol{x}_a^\top S_\ell \, \boldsymbol{x}_b - \delta_{ab} \operatorname{Tr} S_\ell}{\sqrt{d}} \tag{36}$$

  In the following appendix, we will show the tight link between the generic definition of the AIM with deep attention networks.

In the rest of this appendix, we recall the definition of the semicircle and Marchenko-Pastur laws in the context of random matrix theory. In particular

$$\begin{aligned} \sigma_{\mathrm{sc},\,\Delta} &= \frac{\sqrt{4\Delta - x^2}}{2\pi\Delta} \mathbb{I}\{|x| \le 2\sqrt{\Delta}\}\,, \\ \mu_{\mathrm{MP},\rho}(x) &= \begin{cases} (1-\rho)\delta(x) + \rho \frac{\sqrt{(\lambda_+ - x)(x - \lambda_-)}}{2\pi x}\,, & \text{if } \rho \le 1 \\ \rho \frac{\sqrt{(\lambda_+ - x)(x - \lambda_-)}}{2\pi x}\,, & \text{if } \rho > 1 \end{cases} \end{aligned} \tag{37}$$

Finally, we recall the following following definitions.

- Standard normal pdf and cdf

$$\phi(z) = \frac{e^{-z^2/2}}{\sqrt{2\pi}}, \qquad \Phi(z) = \int_{-\infty}^{z} \phi(t)\, dt = \tfrac{1}{2}\big(1 + \operatorname{erf}(z/\sqrt{2})\big) \tag{38}$$

- Bivariate normal density and cdf with correlation c

$$\phi_2(u, v; c) = \frac{\exp\Big[-\frac{u^2 - 2cuv + v^2}{2(1-c^2)}\Big]}{2\pi\sqrt{1-c^2}}, \quad \Phi_2(u, v; c) = \int_{-\infty}^{u}\int_{-\infty}^{v} \phi_2(t_1, t_2; c)\, dt_2\, dt_1. \tag{39}$$

We also remark that we formally define the Dirac delta function $\delta(x) = \lim_{\sigma \to 0} \mathcal{N}(0, \sigma)(x)$ as the limit to zero variance of a centered Gaussian.

We finally define the row-wise softmax function with inverse temperature β acting on the matrix $h \in \mathbb{R}^{T \times T}$ matrix:

$$\sigma_\beta(h_{ab}) = \text{Softmax}(\beta h_{ab}) = \frac{\exp(\beta h_{ab})}{\sum_b \exp(\beta h_{ab})} \tag{40}$$

B From deep self–attention to the *attention-indexed models*

In this appendix we highlight the connection between the AIM models defined in Eq. (1) with those of two crucial architectures employed in the analysis of Large Language Models (LLMs), namely deep attention networks and their sequence-to-sequence (seq2seq) version. In particular, we show that both the deep self–attention encoder and its sequence–to–sequence (seq2seq) variant can be rewritten exactly as an attention-indexed model of the form (1).

We keep the notation of the main text and the previous appendix: tokens are indexed by $a, b \in [T]$, embeddings by $\boldsymbol{x}_a^\top \in \mathbb{R}^d$, and every layer $l \in [L]$ carries a tied key–query weight matrix[1] $S_l \in \mathbb{R}^{d\times d}$ with extensive width $r_l = \rho_l d$ and rotationally–invariant prior P_S.

Deep encoder. Let $X_0 \in \mathbb{R}^{T\times d}$ be the matrix whose rows are the token embeddings, $(X_0)_{a:} = \boldsymbol{x}_a^\top$. A deep self–attention network with a residual (skip) connection and readout strength $c \geq 0$ is given by the recursive formula:

$$X_\ell \;=\; \Big[c\,\mathbb{I}_T + \sigma_\beta\big(H_\ell(X_{\ell-1})\big)\Big]\, X_{\ell-1}, \qquad \ell = 1, \dots, L, \tag{41}$$

where $\sigma : \mathbb{R}^{T\times T} \to \mathbb{R}^{T\times T}$ is the row–wise softmax with inverse temperature $\beta > 0$ implicitly contained in the symbol $\sigma(\cdot)$. The function $H_{\ell-1}(X_{\ell-1})$ is given by:

$$H_\ell(X_{\ell-1}) \;=\; \frac{1}{\sqrt{d}}\, X_{\ell-1}\, S_\ell\, X_{\ell-1}^\top - \frac{1}{\sqrt{d}}\mathbb{E}_{\rm tr}[X_{\ell-1} S_\ell X_{\ell-1}^\top] \tag{42}$$

where the expectation is intended over the input data X_0. Define the preactivations

$$h_{ab}^{(\ell)} \;:=\; \frac{1}{\sqrt{d}}\, \boldsymbol{x}_a^\top S_\ell \boldsymbol{x}_b - \frac{1}{\sqrt{d}}\,\mathrm{Tr}\, S_\ell\, \delta_{ab}, \qquad \ell = 1, \dots, L,\ a, b \in [T], \tag{43}$$

and the sequence of token–space operators

$$B_c^0 := \mathbb{I}_T, \qquad B_c^\ell := \Big[c\,\mathbb{I}_T + \sigma_\beta\big(\, B_c^{\ell-1} h^{(\ell)} B_c^{\ell-1\top}\big)\Big] B_c^{\ell-1}, \quad \ell = 1, \dots, L, \tag{44}$$

One verifies inductively that

$$X_\ell = B_c^\ell\, X_0, \qquad \ell = 0, \dots, L, \tag{45}$$

so that every hidden representation depends on the data only through the collection $\big\{h^{(1)}, \dots, h^{(\ell)}\big\}$. In this way, we can write:

$$H_\ell(X_{\ell-1}) = \frac{1}{\sqrt{d}}\, X_{\ell-1}\, S_\ell\, X_{\ell-1}^\top - \frac{1}{\sqrt{d}}\mathbb{E}_{\rm tr}[X_{\ell-1} S_\ell X_{\ell-1}^\top] \tag{46}$$

$$= \frac{1}{\sqrt{d}}\, X_{\ell-1}\, S_\ell\, X_{\ell-1}^\top - \frac{1}{\sqrt{d}}\mathbb{E}_{\rm tr}[B_c^{\ell-1} X_0\; S_\ell\; X_0^\top B_c^{\ell-1\top}] \tag{47}$$

$$= \frac{1}{\sqrt{d}}\, X_{\ell-1}\, S_\ell\, X_{\ell-1}^\top - \frac{\mathrm{Tr} S_\ell}{\sqrt{d}} B_c^{\ell-1} B_c^{\ell-1\top} \tag{48}$$

Which is exactly the formula shown in (4). Furthermore, this map does only depend on the preactivations $h^{(\ell)}$ in the following way:

$$H_\ell(X_{\ell-1}) = \frac{1}{\sqrt{d}}\, (B_c^{\ell-1} X_0)\, S_\ell\, (B_c^{\ell-1} X_0)^\top - \frac{\mathrm{Tr} S_\ell}{\sqrt{d}} B_c^{\ell-1} B_c^{\ell-1\top} \tag{49}$$

$$= B_c^{\ell-1}\Big(\frac{1}{\sqrt{d}}\, X_0 S_\ell X_0^\top\Big) B_c^{\ell-1\top} - \frac{\mathrm{Tr} S_\ell}{\sqrt{d}} B_c^{\ell-1} B_c^{\ell-1\top} \tag{50}$$

$$= B_c^{\ell-1}\Big(h^{(\ell)} + \frac{\mathrm{Tr} S_\ell}{\sqrt{d}}\mathbb{I}_T\Big) B_c^{\ell-1\top} - \frac{\mathrm{Tr} S_\ell}{\sqrt{d}} B_c^{\ell-1} B_c^{\ell-1\top} \tag{51}$$

$$= B_c^{\ell-1}\, h^{(\ell)}\, B_c^{\ell-1^\top} \tag{52}$$

[1] For simplicity we restrict to the single–head, tied setting; extending to multi–head merely introduces an additional block index.

where we used the definition in (1). This completes our mapping in (7). In particular the *deep-attention output* is given by:

$$y \;=\; \sigma_\beta\Big(H_L(X_{L-1})\Big) \;=\; g_{\text{deep}}\Big(h^{(1)},\dots,h^{(L)}\Big) \;\in \mathbb{R}^{T\times T}, \tag{53}$$

with[2] $g_{\text{deep}}(h^{(1)},\dots,h^{(L)}) := \sigma_\beta\big(\, B_c^{L-1}(h^{(1:L-1)})\, h^{(L)}\, B_c^{L-1}(h^{(1:L-1)})^\top\big)$. Equation (53) is *exact* and has the attention-indexed model structure (1): the whole deep network collapses to a deterministic multivariate function g_{deep} of the L bilinear indices $h_{ab}^{(\ell)} \sim \{\boldsymbol{x}_a^\top S_\ell \boldsymbol{x}_b\}_{\ell,a,b}$.

Seq2seq variant. If the last layer keeps the token embeddings instead of collapsing them, i.e.

$$X_L \;=\; \sigma_\beta\Big(H_L(X_{L-1})\Big)\, X_{L-1}, \tag{54}$$

with exactly the same algebra

$$X_L \;=\; g_{\text{seq}}\Big(h^{(1)},\dots,h^{(L)}\Big)\, X_0, \qquad g_{\text{seq}}(h^{(1:L)}) := \sigma_\beta\big(\, B_c^{L-1} h^{(L)} B_c^{L-1\top}\big)\, B_c^{L-1}. \tag{55}$$

Thus the seq2seq readout is also an attention-indexed model: a (matrix-valued) function of the same quadratic statistics, followed by a fixed linear map X_0.

Note that in the particular case of just $L = 1$ layer the seq2seq map simplifies into:

$$X_1 = g_{\text{seq}}\Big(\{h^{(1)}\}_{a\le b}^{T}\Big)\, X_0, \quad g_{\text{seq}}(h^{(1)}) = \sigma_\beta\big(\, B_c^0 h^{(1)} B_c^{0\top}\big)\, B_c^0 = \sigma_\beta(h^{(1)}) = g_{deep}(h^{(1)}) \tag{56}$$

From this paragraph we can hence conclude that, as shown in equations (53) and (55), any L-layer tied self-attention network with extensive-width weights is information-theoretically equivalent to an attention-indexed model with L indices. Consequently all the Bayes–optimal analysis carried out in Secs. 2–4 applies verbatim to deep self-attention and to its seq2seq counterpart: learning the matrices $\{S_\ell\}$ under the deep architecture is statistically equivalent to learning them under the attention-indexed model (1).

Multi-head self-attention. Let heads $m = 1,\dots,M$ with (tied) weights $S_\ell^{(m)}$ and per-head logits $H_\ell^{(m)}(X_{\ell-1}) = \frac{1}{\sqrt{d}} X_{\ell-1} S_\ell^{(m)} X_{\ell-1}^\top - \mathbb{E}_{\text{tr}}[\frac{1}{\sqrt{d}} X_{\ell-1} S_\ell^{(m)} X_{\ell-1}^\top]$. A standard multi-head layer computes head-wise weights $\sigma_\beta(H_\ell^{(m)})$ and then aggregates (by concatenation or averaging) before a token-wise linear map; for our purposes, averaging:

$$X_\ell = \Big[c\,\mathbb{I}_T + \frac{1}{M}\sum_{m=1}^{M} \sigma_\beta\big(H_\ell^{(m)}(X_{\ell-1})\big)\Big] X_{\ell-1}. \tag{57}$$

Applying the transport identity headwise and averaging gives

$$\frac{1}{M}\sum_{m=1}^{M} H_\ell^{(m)}(X_{\ell-1}) = B_c^{\ell-1}\left(\frac{1}{M}\sum_{m=1}^{M} h^{(\ell,m)}\right) B_c^{\ell-1\top}. \tag{58}$$

If we define the quantity $\bar{h}^{(\ell)} = \frac{1}{M}\sum_{m=1}^{M} h^{(\ell,m)}$, hence, the closed recursion is

$$B_c^\ell \;=\; \Big[c\,\mathbb{I}_T + \sigma_\beta\Big(B_c^{\ell-1}\,\bar{h}^{(\ell)}\, B_c^{\ell-1\top}\Big)\Big]\, B_c^{\ell-1}\,, \quad B_c^0 = \mathbb{I}_T\,, \tag{59}$$

with output

$$y \;=\; \sigma_\beta\Big(B_c^{L-1}\,\bar{h}^{(L)}\, B_c^{L-1\top}\Big), \tag{60}$$

and the seq2seq variant

$$X_L \;=\; \sigma_\beta\Big(B_c^{L-1}\,\bar{h}^{(L)}\, B_c^{L-1\top}\Big)\, B_c^{L-1}\, X_0\,. \tag{61}$$

Finally, also this multi-head variant of the model is again an AIM. Here, the collection of indices is simply enlarged to $\{h^{(\ell,m)}\}_{\ell=1..L,\, m=1..M}$, with each $h^{(\ell,m)}$ defined as in (1). The extensive-rank regime corresponds to ranks $r_\ell^{(m)} = \rho_\ell^{(m)} d$.

[2]The explicit form of g_{deep} is obtained by inserting (45) with $l = L-1$.

C Bayes optimal analysis of attention-indexed models (AIM)

We study a model described by the general setting:

$$y_\mu \sim P_{\rm out}\Big(\frac{\boldsymbol{x}_a^{\mu\top} S_\ell\, \boldsymbol{x}_b^\mu - \delta_{ab}\,{\rm Tr}\, S_\ell}{\sqrt{d}}\Big)_{\ell=1,\dots,L}^{a,b=1,\dots,T} \tag{62}$$

with $\boldsymbol{x}_a^\top$ rows of $X \in \mathbb{R}^{T\times d}$, $S_\ell \in \mathbb{R}^{d\times d}$ symmetric and $y \in \mathbb{R}^{T\times T}$. Indices range from $\mu = 1\dots n$ samples, with $d, n >> 1$. Instead the number of tokens and layers $T, L << d$: we interpret Eq. (62) as y_μ outputs generated by a model of attention from data X that are processed in a bilinear way through:

$$y_\mu = g\left(\{\frac{\boldsymbol{x}_a^\mu S_\ell\, \boldsymbol{x}_b^{\mu\,\top} - \delta_{ab}\,{\rm Tr}\, S_\ell}{\sqrt{d}}\}_{\ell=1}^L\right). \tag{63}$$

or following Eq. (6):

$$y_\mu = g_{\rm deep}(h^{(1)},\dots,h^{(L)}) = B_c^L\Big(\{\frac{\boldsymbol{x}_a^{\mu\,\top} S_\ell\, \boldsymbol{x}_b^\mu - \delta_{ab}\,{\rm Tr}\, S_\ell}{\sqrt{d}}\}_{a,b=1\dots T}^{\ell=1\dots L}\Big) \in \mathbb{R}^{T\times T} \tag{64}$$

and

$$P_{\rm out}\Big(\frac{\boldsymbol{x}_a^{\mu\,\top} S_\ell\, \boldsymbol{x}_b^\mu - \delta_{ab}\,{\rm Tr}\, S_\ell}{\sqrt{d}}\Big)_{\ell=1,\dots,L}^{a,b=1,\dots,T} = \delta\Big(y - B_c^L(\frac{\boldsymbol{x}_a^{\mu\,\top} S_\ell\, \boldsymbol{x}_b^\mu - \delta_{ab}\,{\rm Tr}\, S_\ell}{\sqrt{d}})\Big) \tag{65}$$

In our setting, the matrices S_ℓ are symmetrical for each layer ℓ and we consider multiple layers indices $\ell = 1,\dots,L$. $\boldsymbol{x}_a^\top$ is the a-th row of X for $a, b = 1,\dots,T$. Each row $\boldsymbol{x}_a^\top \in \mathbb{R}^d$ has i.i.d. Gaussian entries, so $x_{ai}^\mu \sim \mathcal{N}(0,1)$.

We define the preactivations

$$h_{ab}^{(\ell)\,\mu} = \frac{\boldsymbol{x}_a^{\mu\,\top} S_\ell\, \boldsymbol{x}_b^\mu - \delta_{ab}\,{\rm Tr}\, S_\ell}{\sqrt{d}} \tag{66}$$

Since the matrices S_ℓ are symmetric, so are the preactivations of the model. Finally, for convenience, we rewrite the preactivations of the model in terms of the symmetrized sensing matrices

$$Z_{ij,ab}^\mu \equiv (x_{i,a}^\mu x_{j,b}^\mu + x_{j,a}^\mu x_{i,b}^\mu - 2\delta_{ij}\delta_{ab})/\sqrt{2d(1+\delta_{ab})} \in \mathbb{R} \tag{67}$$

The preactivations of the model can thus be expressed as:

$$\sqrt{2-\delta_{ab}}\, h_{ab}^{(\ell)\,\mu} = {\rm Tr}(S_\ell Z_{ab}^\mu) = \tilde{h}_{ab}^{(\ell)\,\mu} \tag{68}$$

In the rest of the analysis, we will refer to this equivalent representation of the model by considering symmetrized data $\tilde{h}_{ab}^{(\ell)\,\mu}$ that we will just recall $h_{ab}^{(\ell)\,\mu}$, while incorporating the factor $\sqrt{2-\delta_{ab}}$ in the output function part.

C.1 Replica analysis of AIM and their state evolution

Starting from the posterior distribution of the model:

$$P(S_1,...,S_L|\mathcal{D}) = \frac{1}{\mathcal{Z}(\mathcal{D})}\prod_{\ell=1}^L P_S(S_\ell)\prod_{\mu=1}^n \delta\left(y^\mu - g\left(h^{(1)}(S_1,\boldsymbol{x}^\mu),\dots,h^{(L)}(S_L,\boldsymbol{x}^\mu)\right)\right), \tag{69}$$

the replicated partition function of the model in Eq. (62) is:

$$\langle \mathcal{Z}(\mathcal{D})^m\rangle = \mathbb{E}_{y,\,X}\int\prod_{\ell=1}^L\prod_{u=0}^m {\rm d}S_\ell^u\, P_0(S_\ell^u)\prod_{\mu=1}^n\prod_{a\le b}^T P_{\rm out}\Big(y^\mu\,\Big|\,\Big\{\frac{h_{ab}^{(\ell),\mu,u}}{\sqrt{2-\delta_{ab}}}\Big\}_{ab}\Big)\,\delta\Big(h_{ab}^{(\ell),\mu,u} - {\rm Tr}(S_\ell^u Z_{ab}^\mu)\Big) \tag{70}$$

where $P_0(S_\ell^u)$ is the rotational invariant prior distribution of each S_ℓ, and $h_{ab}^{(\ell),\mu,u}$ are the replicated preactivations in terms of the symmetrized data as explained in (68). u is the replica index, we work in a Bayes optimal setting. Above, $\mu \in \{1,\dots,n\}$ enumerates data samples, $\ell \in \{1,\dots,L\}$ indexes the distinct layers, $u \in \{0,\dots,m\}$ indexes the replicas, and $a, b \in \{1,\dots,T\}$ are the token indices.

We compute the expectation with respect to the data exploiting the Gaussian-equivalence principle:

$$\mathbb{E}_X \delta\Big(h_{ab}^{(\ell),\mu,u} - \mathrm{Tr}(S_\ell^u Z_{ab}^\mu)\Big) \quad \mapsto \quad P_h\Big(\{h_{ab}^{(\ell),\mu,u}\}_{\ell,\mu,a,b,u}\Big), \tag{71}$$

where P_h is a joint Gaussian distribution with the means and covariances:

$$\mathbb{E}\big[h_{ab}^{(\ell),\mu,u} - \mathrm{Tr}(S_\ell^u Z_{ab}^\mu)\big] = 0, \quad \mathrm{Cov}_{x^\mu}\left(h_{a\le b}^{(\ell)\ u}, h_{c\le d}^{(k)\ v}\right) = \frac{1}{d}\left[2\,\delta_{ac}\delta_{bd}\right]\mathrm{Tr}\left(S_\ell^u S_k^v\right) \tag{72}$$

We introduce the order parameters measuring the S_ℓ^u–S_k^v overlaps:

$$Q_{\ell k}^{uv} \ := \ \frac{1}{d}\,\mathrm{Tr}\Big(S_\ell^u\, S_k^v\Big), \quad \text{for } \ell, k = 1, \dots, L,\ u, v = 0, \dots, m. \tag{73}$$

We enforce the definitions of the overlaps by inserting δ-functions:

$$\prod_{\ell,k=1}^{L} \prod_{u\le v=0}^{m} \delta\Big(d^2\, Q_{\ell k}^{uv} - d\ \mathrm{Tr}\big[S_\ell^u\, S_k^v\big]\Big), \tag{74}$$

and introduce the corresponding conjugate fields $\widehat{Q}_{\ell k}^{uv}$. We insert

$$\delta\Big(d^2\, Q_{\ell k}^{uv} - d\mathrm{Tr}[S_\ell^u\, S_k^v]\Big) = \int \mathrm{d}\widehat{Q}_{\ell k}^{uv}\ \exp\Big\{\, i\,\frac{\widehat{Q}_{\ell k}^{uv}}{2}\,\Big(d^2\, Q_{\ell k}^{uv} - d\ \mathrm{Tr}[S_\ell^u\, S_k^v]\Big)\Big\}. \tag{75}$$

Hence the replicated partition function can be schematically written:

$$\begin{aligned}\big\langle \mathcal{Z}(\mathcal{D})^m \big\rangle = & \int \Big(\prod_{u,\ell} \mathrm{d}S_\ell^u\ P_0(S_\ell^u)\Big) \int \Big(\prod_{u\le v,\ell,k} \mathrm{d}Q_{\ell k}^{uv}\ \mathrm{d}\hat{Q}_{\ell k}^{uv}\Big) \\ & \times \exp\Big[\frac{i}{2} \sum_{u\le v,\ell,k} \hat{Q}_{\ell k}^{uv}\big(d^2\, Q_{\ell k}^{uv}\big)\Big] \times \exp\Big[-\frac{i\,d}{2} \sum_{u\le v,\ell,k} \hat{Q}_{\ell k}^{uv}\,\mathrm{Tr}(S_\ell^u S_k^v)\Big] \\ & \times \prod_{\mu=1}^{n}\Big[\int \prod_{u,\ell} \mathrm{d}h_{ab}^{(\ell),\mu,u}\ P_h\big(h^{(\ell),\mu,u}\big) \prod_{u,\ell,a\le b} P_{\mathrm{out}}(y_{ab}^\mu \mid \{\frac{h_{ab}^{(\ell),\mu,u}}{\sqrt{2-\delta_{ab}}}\}_{ab})\Big],\end{aligned} \tag{76}$$

In a replica-symmetric (RS) scenario, we let

$$Q_{\ell k}^{uv} = \begin{cases} Q_{\ell k}, & (u = v), \\ q_{\ell k}, & (u \neq v). \end{cases} \tag{77}$$

and:

$$i\hat{Q}_{\ell k}^{uv} = \begin{cases} \hat{Q}_{\ell k}, & \text{if } u = v \\ -\hat{q}_{\ell k}, & \text{if } u \neq v \end{cases} \tag{78}$$

Hence, e.g. the exponent $\sum_{\ell,k,u,v} i\,\widehat{Q}_{\ell,k}^{uv}\, d^2\, Q_{\ell,k}^{uv}$ becomes

$$i\, d^2 \sum_{\ell,k}\Big[\frac{(m+1)}{2}\,\widehat{Q}_{\ell k}\, Q_{\ell k} - \frac{m(m+1)}{4}\,\widehat{q}_{\ell k}\, q_{\ell k}\Big]. \tag{79}$$

Likewise, $-\sum_{\ell,k,u,v} \widehat{Q}_{\ell k}^{uv}\mathrm{Tr}(S_\ell^u S_k^v)$ can be reorganized in a form that leads in the limit $m \to 0$ to typical terms $\widehat{Q}_{\ell k}^{uu} = 0$ or similar. Moreover $\widehat{Q}_{\ell k}^{uv} = -\frac{\hat{q}_{\ell k}}{2}$.

So finally the replicated partition function, hence, takes the following form:

$$\big\langle \mathcal{Z}(\mathcal{D})^m \big\rangle = \int \prod_{u\le v,\ell,k} dQ_{\ell k}^{uv} d\hat{Q}_{\ell k}^{uv} \exp\left(\frac{i}{2} d^2 \sum_{u\le v,\ell,k} \hat{Q}_{\ell k}^{uv} Q_{\ell k}^{uv}\right) I_{\mathrm{in}} I_{\mathrm{out}} \tag{80}$$

with:

$$d^2 I_{\mathrm{in}}(\hat{q}) = \int \prod_{u,\ell} dS_\ell^u\, P_0(S_\ell^u) \exp\left(-\frac{i\,d}{2} \sum_{u\le v,\ell,k} \hat{Q}_{\ell k}^{uv}\,\mathrm{Tr}\left(S_\ell^u S_k^v\right)\right), \tag{81}$$

$$I_{\mathrm{out}}(q) = \left[\int dy \int \prod_{u,\ell,a\le b} dh_{ab}^{(\ell)\ u}\, P\left(h_{ab}^{(\ell)\ u}\right) \prod_{u,\ell} P_{\mathrm{out}}\left(y|\Big\{\frac{h_{ab}^{(\ell)\ u}}{\sqrt{2-\delta_{ab}}}\Big\}_{ab}\right)\right]^n. \tag{82}$$

The free entropy per degree of freedom of the problem is defined as

$$\Phi = \lim_{d\to\infty} \frac{1}{d^2} \lim_{n\to\infty} \lim_{m\to 0} \frac{1}{m} \ln\langle Z^m \rangle \,. \tag{83}$$

After introducing $n = \alpha d^2$ data samples, the free entropy decomposes into a prior contribution and an output contribution:

$$\Phi = \text{extr}_{\{q,\hat{q}\}} \left\{ -\frac{\text{Tr}\, q\hat{q}}{4} + I_{\text{in}}(\hat{q}) + \alpha I_{\text{out}}(q) \right\} . \tag{84}$$

Thus obtaining the state equations:

$$q = 4\, \partial_{\hat{q}} I_{\text{in}}(\hat{q}) \tag{85}$$

$$\hat{q} = 4\alpha\, \partial_q I_{\text{out}}(q) \tag{86}$$

C.2 Prior Term Computation

First we compute under the RS ansatz:

$$-\frac{i\,d}{2} \sum_{u\le v=0}^{m} \hat{Q}^{uv}_{\ell k} \,\text{Tr}\,(S^u_\ell S^v_k) = -\frac{i\,d}{2} \left(\sum_{u=0}^{m} \hat{Q}_{\ell k} \,\text{Tr}\,(S^u_\ell S^u_k) + \sum_{u<v} (-\hat{q}_{\ell k}) \,\text{Tr}\,(S^u_\ell S^v) \right) \tag{87}$$

$$= -\frac{\hat{Q}_{\ell k}\, d}{2} \sum_{u=0}^{m} \text{Tr}\,(S^u_\ell S^u_k) + \frac{\hat{q}_{\ell k}\, d}{2} \sum_{u<v} \text{Tr}\,(S^u_\ell S^v_k) \tag{88}$$

$$= -\frac{d}{2} \left(\hat{Q}_{\ell k} + \frac{\hat{q}_{\ell k}}{2} \right) \sum_{u=0}^{m} \text{Tr}\,(S^u_\ell S^u_k) + \frac{\hat{q}_{\ell k}\, d}{4} \sum_{u,v=0}^{m} \text{Tr}\,(S^u_\ell S^v_k) \tag{89}$$

We remind that each S_ℓ is a rank-$\rho_\ell d$ rotationally invariant matrix of order $O(d \times d)$. The prior factor that emerges from the partition function, after decoupling the replica indices by applying a Hubbard-Stratonovich transformation, reads:

$$I_{\text{in}}(\hat{q}) = \int \prod_{\ell=1}^{L} \prod_{u=0}^{m} dS^u_\ell \, P_0(S^u_\ell) \, \exp\left\{ -\frac{i\,d}{2} \sum_{\ell,k=1}^{L} \sum_{u\le v=0}^{m} \widehat{Q}^{(u,v)}_{\ell k} \, \text{Tr}\big(S^u_\ell S^v_k\big) \right\} \tag{90}$$

$$= \int d\bar{S} P_0(\bar{S}) \exp\left\{ \sum_{\ell,k}^{L} -\frac{d}{2} \left(\hat{Q}_{\ell k} + \frac{\hat{q}_{\ell k}}{2} \right) \sum_{u=0}^{m} \text{Tr}\,(S^u_\ell S^u_k) + \frac{\hat{q}_{\ell k}\, d}{4} \sum_{u,v=0}^{m} \text{Tr}\,(S^u_\ell S^v_k) \right\} \tag{91}$$

$$= \int d\bar{S} P_0(\bar{S}) \exp\left\{ -\sum_{\ell,k}^{L} \sum_{u} \frac{\hat{q}_{\ell,k}\, d}{4} \text{Tr}(S^u_\ell S^u_k) + \sum_{\ell k}^{L} \sum_{u,v} \frac{\hat{q}_{\ell k} d}{4} \text{Tr}(S^u_\ell S^v_k) \right\} \tag{92}$$

$$= \int d\bar{S} P_0(\bar{S}) \mathcal{D}(Y) \exp\left\{ -\sum_{\ell,k}^{L} \sum_{u} \frac{\hat{q}_{\ell k}\, d}{4} \text{Tr}(S^u_\ell S^v_k) + \sum_{\ell,k}^{L} \sum_{u} \frac{\sqrt{\hat{q}_{\ell k}}\, d}{2} \text{Tr}(S^u_k Y_\ell) \right\} \tag{93}$$

$$= \int \mathcal{D}(Y) \left\{ \int d\bar{S} P_0(\bar{S}) \exp\Big\{ -\frac{d}{4} \sum_{\ell,k}^{L} \hat{q}_{\ell k} \,\text{Tr}(S_\ell S_k) + d \sum_{\ell,k}^{L} \frac{\sqrt{\hat{q}_{\ell k}}}{2} \,\text{Tr}(S_k Y_\ell) \right\}^{m+1} \tag{94}$$

where $\mathcal{D}(Y_\ell)$ are GOE(d) measures $\forall \ell \in [L]$ and $Y_\ell \in \mathbb{R}^{d\times d}$ and also $\bar{S} \in [\mathbb{R}^{d\times d}]^L$. In Eq.(78) we used the identity:

$$\mathbb{E}_{Y\sim \text{GOE}(d)} \left[e^{\frac{d}{2}\,\text{Tr}[SY]} \right] = e^{\frac{d}{4}\,\text{Tr}\left[S^2\right]}$$

Finally, taking the zero replica $m \to 0$ limit, we can write the prior contribution to the free entropy of the model as:

$$I_{\text{in}}(\hat{q}) = \lim_{d\to\infty} \frac{1}{d^2} \int DY_1 \dots DY_L \, \mathcal{Z}_{\text{in}}(Y_1, \dots, Y_L; \hat{q}) \log \mathcal{Z}_{\text{in}}(Y_1, \dots, Y_L; \hat{q})$$

$$\mathcal{Z}_{\text{in}}(\{Y_\ell\}_{\ell=1}^{L}; \hat{q}) = \int \left[\prod_{\ell=1}^{L} \mathrm{d}S_\ell \, P_S(S_\ell) \right] \times \exp\Big[-\frac{d}{4} \sum_{\ell,k=1}^{L} \hat{q}_{\ell k} \,\text{Tr}(S_\ell S_k) + \frac{d}{2} \sum_{\ell,k=1}^{L} \sqrt{\hat{q}}_{\ell k} \,\text{Tr}(Y_\ell S_k) \Big]. \tag{95}$$

The matrices $Y_\ell \in \mathbb{R}^{d\times d}$ are the auxiliary fields introduced by the Hubbard–Stratonovich transformation. Notably, they can be interpreted as "noisy measurements" of the S_ℓ matrices with coupled indices. In particular, the denoising problem which is solved by the free-entropy contribution of the prior is:

$$Y_\ell^{ij} = \sum_k \sqrt{\hat{q}_{\ell k}}\, S_k^{ij} + Z_\ell^{ij} \quad \forall i,j,\ell \tag{96}$$

with Z_ℓ GOE(d) matrices and $S_\ell \in \mathbb{R}^{d\times d}$ rotationally invariant matrices, leading to an exponential term of the form of $-\frac{1}{2}\sum_\ell \mathrm{Tr}((\sum_k \sqrt{\hat{q}}_{\ell k} S_k - Y_\ell)^2)$. Such equivalence between the matrix denoising problem in (96) and (95) is analogous to those of [42, 40].

C.3 Output Channel Computation

Starting from the replicated partition function in Eq. (80), we can see that the output channel contribution to the free entropy of the model, factorized with respect to the data, is given by:

$$I_{\text{out}}(q) = \left[\int dy \int \prod_{u,\ell,a\le b} dh_{ab}^{(\ell)\,u}\, P\left(\{h_{ab}^{(\ell)\,u}\}_{ab}\right) \prod_{u,\ell} P_{\text{out}}\left(y \;\middle|\; \left\{\frac{h_{ab}^{(\ell)\,u}}{\sqrt{2-\delta_{ab}}}\right\}_{ab}\right)\right]^n, \tag{97}$$

where we consider only the upper triangular token indices $a \le b$.
$P_h\left(\{h_{ab}^{(\ell),u}\}_{ab}\right)$ is a multivariate Gaussian distribution with means and covariance:

$$\mathbb{E}[h_{ab}^{(\ell)}] = 0, \qquad \mathrm{Cov}_{x^\mu}\left(h_{a\le b}^{(\ell)\,u}, h_{c\le d}^{(k)\,v}\right) = \frac{1}{d}\left[2\,\delta_{ac}\delta_{bd}\right]\mathrm{Tr}\left(S_\ell^u S_k^v\right) = \left[2\,\delta_{ac}\delta_{bd}\right] Q_{\ell k}^{uv}. \tag{98}$$

Under the RS ansatz and in the limit $m \to 0$, we can decouple the replicas through another Hubbard-Stratonovich transformation. The exponent involving $h_{ab}^{(\ell)\,u}$ becomes:

$$-\frac{1}{2}\sum_{u,v=0}^{m}\sum_{a\le b,c\le d}\sum_{\ell,k}\left(h_{ab}^{(\ell)\,u}\right)\left(\Sigma_h^{-1}\right)_{ab,cd}^{uv,\ell k}\left(h_{cd}^{(k)\,v}\right). \tag{99}$$

Substituting back, the output term becomes:

$$I_{\text{out}}(q) = \left[\int dy \int \prod_{u,\ell,a\le b} dh_{ab}^{(\ell)\,u} \exp\left(-\frac{1}{2}\sum_{u,v=0}^{m}\sum_{a,b,c,d}\sum_{\ell,k} h_{ab}^{(\ell)\,u}\left(\Sigma_h^{-1}\right)_{a\le b,c\le d}^{uv,\ell k} h_{cd}^{(k)\,v}\right)\prod_{u,\ell} P_{\text{out}}\left(y \;\middle|\; \left\{\frac{h_{ab}^{(\ell)\,u}}{\sqrt{2-\delta_{ab}}}\right\}_{ab}\right)\right]^n. \tag{100}$$

For a fixed channel ℓ and for each token pair (a,b) with $a \le b$, the covariance in the replica space is given by

$$(\Sigma_h)_{a\le b,c\le d}^{uv,\ell k} = \left[2\,\delta_{ac}\delta_{bd}\right] Q_{\ell k}^{uv} = \left[2\,\delta_{ac}\delta_{bd}\right]\Big[(Q_{\ell k} - q_{\ell k})\,\delta_{uv} + q_{\ell k}\Big]. \tag{101}$$

Because of the Kronecker structure $\delta_{ac}\delta_{bd}$, the Gaussian law over all $\{h_{ab}^{(\ell)u}\}_{a\le b,\ell,u}$ factorizes over token pairs (a,b). Hence it is sufficient to treat one fixed pair (a,b) and then take the product over $a \le b$. For notational clarity in the next steps, we temporarily fix a pair (a,b) and write $h^{(\ell)u} \equiv h_{ab}^{(\ell)u}$. For this pair, the covariance across replicas and layers reads

$$\mathbb{E}\Big[h^{(\ell)u}h^{(k)v}\Big] = 2\Big[(Q_{\ell k} - q_{\ell k})\,\delta_{uv} + q_{\ell k}\Big], \qquad u,v = 0,\dots,m,\;\; \ell,k = 1,\dots,L. \tag{102}$$

We now construct explicitly a family of Gaussian random variables with covariance (102). Introduce a shared Gaussian vector $\omega = (\omega^{(1)},\dots,\omega^{(L)})$ with mean zero and covariance

$$\mathbb{E}\Big[\omega^{(\ell)}\omega^{(k)}\Big] = 2q_{\ell k}. \tag{103}$$

For each replica $u = 0,\dots,m$, an independent Gaussian vector $\xi^u = (\xi^{(1)u},\dots,\xi^{(L)u})$ with mean zero and covariance

$$\mathbb{E}\Big[\xi^{(\ell)u}\xi^{(k)v}\Big] = 2(Q_{\ell k} - q_{\ell k})\,\delta_{uv}. \tag{104}$$

ω is independent of all $\{\xi^u\}_{u=0}^m$. Define for each replica u and layer ℓ:

$$h^{(\ell)u} := \omega^{(\ell)} + \xi^{(\ell)u}. \tag{105}$$

Then, for all ℓ, k and u, v,

$$\begin{aligned}\mathbb{E}\Big[h^{(\ell)u}h^{(k)v}\Big] &= \mathbb{E}\Big[(\omega^{(\ell)}+\xi^{(\ell)u})(\omega^{(k)}+\xi^{(k)v})\Big] \\ &= \mathbb{E}\Big[\omega^{(\ell)}\omega^{(k)}\Big] + \mathbb{E}\Big[\xi^{(\ell)u}\xi^{(k)v}\Big] \\ &= 2q_{\ell k} + 2(Q_{\ell k}-q_{\ell k})\delta_{uv} = 2\Big[(Q_{\ell k}-q_{\ell k})\delta_{uv} + q_{\ell k}\Big], \end{aligned} \tag{106}$$

which is exactly (102). Therefore the law of $\{h^{(\ell)u}\}_{\ell,u}$ is exactly the RS Gaussian law.

From (105), conditional on ω the replicas are independent and

$$\{h^{(\ell)u}\}_{\ell=1}^{L} \,\Big|\, \omega \;\sim\; \mathcal{N}\Big(\{\omega^{(\ell)}\}_{\ell=1}^{L},\, V\Big), \qquad V^{(\ell k)} := 2(Q_{\ell k}-q_{\ell k}). \tag{107}$$

Equivalently, for each u,

$$p\Big(\{h^{(\ell)u}\}_{\ell} \mid \omega\Big) = \frac{1}{\sqrt{\det(2\pi V)}}\,\exp\left(-\frac{1}{2}\sum_{\ell=1}^{L}\sum_{k=1}^{L}\big(h^{(\ell)u}-\omega^{(\ell)}\big)\,[V^{-1}]_{\ell k}\,\big(h^{(k)u}-\omega^{(k)}\big)\right). \tag{108}$$

Moreover $\omega \sim \mathcal{N}(0, 2q)$, i.e.

$$p(\omega) = \frac{1}{\sqrt{\det(2\pi\, 2q)}}\,\exp\left(-\frac{1}{2}\sum_{\ell=1}^{L}\sum_{k=1}^{L}\omega^{(\ell)}\,[(2q)^{-1}]_{\ell k}\,\omega^{(k)}\right). \tag{109}$$

By marginalization over ω, the joint density of $\{h^{u}\}_{u=0}^{m}$ is therefore:

$$p\Big(\{h^{(\ell)u}\}_{\ell,u}\Big) = \int d^L\omega\; p(\omega)\;\prod_{u=0}^{m} p\Big(\{h^{(\ell)u}\}_{\ell} \mid \omega\Big). \tag{110}$$

We now express $\omega \sim \mathcal{N}(0, 2q)$ through standard normals. Introduce i.i.d. auxiliary variables $\eta^{(1)}, \ldots, \eta^{(L)} \sim \mathcal{N}(0, 1)$ and define

$$\omega^{(\ell)} = \sum_{r=1}^{L}(\sqrt{2q})_{\ell r}\,\eta^{(r)}, \tag{111}$$

where $(\sqrt{2q})$ is any matrix square root such that, for all ℓ, k,

$$\sum_{r=1}^{L}(\sqrt{2q})_{\ell r}(\sqrt{2q})_{kr} = (2q)_{\ell k}. \tag{112}$$

Then ω has covariance (103) since

$$\mathbb{E}[\omega^{(\ell)}\omega^{(k)}] = \sum_{r,s}(\sqrt{2q})_{\ell r}(\sqrt{2q})_{ks}\,\mathbb{E}[\eta^{(r)}\eta^{(s)}] = \sum_{r}(\sqrt{2q})_{\ell r}(\sqrt{2q})_{kr} = (2q)_{\ell k}. \tag{113}$$

Consequently, for any function $F(\omega)$,

$$\int d^L\omega\;\mathcal{N}(\omega; 0, 2q)\,F(\omega) = \int\prod_{r=1}^{L}\frac{d\eta^{(r)}}{\sqrt{2\pi}}e^{-(\eta^{(r)})^2/2}\;F\left(\left\{\sum_{r}(\sqrt{2q})_{\ell r}\eta^{(r)}\right\}_{\ell=1}^{L}\right). \tag{114}$$

Restoring the token indices, for each (a, b) we introduce independent $\eta_{ab}^{(r)} \sim \mathcal{N}(0, 1)$ and define

$$\omega_{ab}^{(\ell)} = \sum_{k=1}^{L}(\sqrt{2q})_{\ell k}\,\eta_{ab}^{(k)}, \qquad V^{(\ell k)} = 2(Q_{\ell k}-q_{\ell k}). \tag{115}$$

Using (107)–(114), the Gaussian exponent for each replica u and each token pair (a, b) is therefore exactly

$$-\frac{1}{2}\sum_{\ell=1}^{L}\sum_{k=1}^{L}\big(h_{ab}^{(\ell)\,u}-\omega_{ab}^{(\ell)}\big)\,[V^{-1}]_{\ell k}\,\big(h_{ab}^{(k)\,u}-\omega_{ab}^{(k)}\big) \;-\frac{1}{2}\sum_{r=1}^{L}(\eta_{ab}^{(r)})^2, \tag{116}$$

which is the desired form with ω and V identified as in (115).

Define the auxiliary measure

$$\mathcal{D}\eta = \prod_{a\le b}\prod_{\ell=1}^{L} \frac{d\eta_{ab}^{(\ell)}}{\sqrt{2\pi}}\, \exp\Big[-\frac{(\eta_{ab}^{(\ell)})^2}{2}\Big]. \tag{117}$$

For fixed η, the replicas are independent and identically distributed through the conditional Gaussian $\mathcal{N}(h_{ab};\omega_{ab},V)$ on layers ℓ. Hence, introducing the single-replica output partition function

$$\mathcal{Z}_{\rm out}(y,\omega,V) = \int \prod_{a\le b} d^L h_{ab}\, \mathcal{N}(h_{ab};\omega_{ab},V)\; P_{\rm out}\left(y \;\middle|\; \Big\{\frac{h_{ab}^{(\ell)}}{\sqrt{2-\delta_{ab}}}\Big\}_{a\le b,\ \ell=1}^{L}\right), \tag{118}$$

we can write

$$I_{\rm out}(q) = \left[\int dy \int \mathcal{D}\eta\, \Big(\mathcal{Z}_{\rm out}(y,\omega,V)\Big)^{m+1}\right]^{n}, \tag{119}$$

where ω and V are the functions of (q,Q) defined in (115). Expanding Eq. (119) for small number of replicas $m\to 0$ and using $A^{m+1} = A\, e^{m\ln A} = A\,(1+m\ln A + o(m))$, the contribution to the free entropy is governed by

$$I_{\rm out}(q) = \int dy \int \mathcal{D}\eta\; \mathcal{Z}_{\rm out}(y,\omega,V)\; \ln \mathcal{Z}_{\rm out}(y,\omega,V), \tag{120}$$

We aim to compute $\partial_q I_{\rm out}(q)$ to finally reach the state equation in Eq. (86), namely:

$$\widehat{q}_{\ell k} = 4\alpha \int dy \int \mathcal{D}\eta\, \left[1+\ln \mathcal{Z}_{\rm out}(y,\omega,V)\right]\, \frac{\partial \mathcal{Z}_{\rm out}(y,\omega,V)}{\partial q_{\ell k}}. \tag{121}$$

It is convenient to rewrite the η-integral as an integral over ω itself. For each token pair (a,b) we have $\omega_{ab}\sim\mathcal{N}(0,2q)$ with independent draws across $a\le b$. Thus we may equivalently write

$$I_{\rm out}(q) = \int dy \int \Big[\prod_{a\le b} d^L\omega_{ab}\, \mathcal{N}(\omega_{ab};0,2q)\Big]\; \mathcal{Z}_{\rm out}(y,\omega,V)\; \ln \mathcal{Z}_{\rm out}(y,\omega,V), \tag{122}$$

where $\omega = \{\omega_{ab}^{(\ell)}\}_{a\le b,\ell}$. Now define for each (a,b) and layer ℓ:

$$(g_{\rm out}(y,\omega,V))_{ab}^{(\ell)} := \frac{\partial}{\partial \omega_{ab}^{(\ell)}} \ln \mathcal{Z}_{\rm out}(y,\omega,V). \tag{123}$$

In (122), q appears in: (i) in the Gaussian measure $\omega_{ab}\sim\mathcal{N}(0,2q)$, and (ii) in $V = 2(Q-q)$ inside $\mathcal{Z}_{\rm out}$. Let

$$F(\omega,q) := \int dy\; \mathcal{Z}_{\rm out}(y,\omega,V)\; \ln \mathcal{Z}_{\rm out}(y,\omega,V).$$

Then

$$I_{\rm out}(q) = \int \Big[\prod_{a\le b} d^L\omega_{ab}\, \mathcal{N}(\omega_{ab};0,2q)\Big]\; F(\omega,q). \tag{124}$$

Now fix a single pair (a,b), under $\omega_{ab}\sim\mathcal{N}(0,\Sigma)$ with $\Sigma = 2q$, we get the Gaussian identity, for any smooth $G(\omega_{ab})$:

$$\frac{\partial}{\partial \Sigma_{\ell k}} \mathbb{E}_{\omega_{ab}\sim\mathcal{N}(0,\Sigma)}[G(\omega_{ab})] = \frac{1}{2}\, \mathbb{E}_{\omega_{ab}}\left[\frac{\partial^2 G(\omega_{ab})}{\partial \omega_{ab}^{(\ell)}\, \partial \omega_{ab}^{(k)}}\right]. \tag{125}$$

Since $\Sigma = 2q$, we have $\partial/\partial q_{\ell k} = 2\,\partial/\partial\Sigma_{\ell k}$, hence

$$\frac{\partial}{\partial q_{\ell k}} \mathbb{E}_{\omega_{ab}\sim\mathcal{N}(0,2q)}[G(\omega_{ab})] = \mathbb{E}_{\omega_{ab}}\left[\frac{\partial^2 G(\omega_{ab})}{\partial \omega_{ab}^{(\ell)}\, \partial \omega_{ab}^{(k)}}\right]. \tag{126}$$

Applying (126) to (124) (and summing over independent pairs) yields

$$\frac{\partial I_{\rm out}(q)}{\partial q_{\ell k}} = \int \Big[\prod_{a\le b} d^L\omega_{ab}\, \mathcal{N}(\omega_{ab};0,2q)\Big]\; \left(\sum_{a\le b} \frac{\partial^2 F(\omega,q)}{\partial \omega_{ab}^{(\ell)}\, \partial \omega_{ab}^{(k)}} \;+\; \frac{\partial F(\omega,q)}{\partial q_{\ell k}}\Big|_{\omega}\right), \tag{127}$$

where $\partial F/\partial q|_\omega$ is the explicit derivative through $V = 2(Q - q)$ at fixed ω.

Since $V^{(rs)} = 2(Q_{rs} - q_{rs})$, we have $\partial V^{(rs)}/\partial q_{\ell k} = -2\delta_{r\ell}\delta_{sk}$, hence

$$\frac{\partial F(\omega, q)}{\partial q_{\ell k}}\Big|_\omega = -2\,\frac{\partial F(\omega, q)}{\partial V^{(\ell k)}}. \tag{128}$$

Plugging (128) into (127) gives

$$\frac{\partial I_{\text{out}}(q)}{\partial q_{\ell k}} = \int \Big[\prod_{a\le b} d^L\omega_{ab}\,\mathcal{N}(\omega_{ab}; 0, 2q)\Big] \left(\sum_{a\le b} \frac{\partial^2 F}{\partial \omega_{ab}^{(\ell)}\,\partial \omega_{ab}^{(k)}} \;-\; 2\,\frac{\partial F}{\partial V^{(\ell k)}}\right). \tag{129}$$

Now fix ω, V and define for brevity

$$Z(y) := \mathcal{Z}_{\text{out}}(y, \omega, V), \qquad g_{ab}^{(\ell)}(y) := (g_{\text{out}}(y, \omega, V))_{ab}^{(\ell)}. \tag{130}$$

Then

$$F(\omega, q) = \int dy\; Z(y)\; \ln Z(y). \tag{131}$$

We compute the two derivatives in (129) exactly. Since Z is an integral of a Gaussian density $\mathcal{N}(h_{ab}; \omega_{ab}, V)$, then by definition:

$$\frac{\partial Z(y)}{\partial \omega_{ab}^{(\ell)}} = Z(y)\, g_{ab}^{(\ell)}(y), \qquad \frac{\partial^2 Z(y)}{\partial \omega_{ab}^{(\ell)} \partial \omega_{ab}^{(k)}} = Z(y)\, g_{ab}^{(\ell)}(y)\, g_{ab}^{(k)}(y) + Z(y)\,\frac{\partial g_{ab}^{(\ell)}(y)}{\partial \omega_{ab}^{(k)}}. \tag{132}$$

Therefore,

$$\frac{\partial F}{\partial \omega_{ab}^{(k)}} = \int dy\; (1 + \ln Z(y))\,\frac{\partial Z(y)}{\partial \omega_{ab}^{(k)}} = \int dy\; (1 + \ln Z(y))\, Z(y)\, g_{ab}^{(k)}(y), \tag{133}$$

$$\frac{\partial^2 F}{\partial \omega_{ab}^{(\ell)} \partial \omega_{ab}^{(k)}} = \int dy\, \Big[Z(y)\, g_{ab}^{(\ell)}(y)\, g_{ab}^{(k)}(y) + (1 + \ln Z(y))\Big(Z(y)\, g_{ab}^{(\ell)}(y)\, g_{ab}^{(k)}(y) + Z(y)\,\partial_{\omega_{ab}^{(k)}} g_{ab}^{(\ell)}(y)\Big)\Big]. \tag{134}$$

For a Gaussian integral, the derivative of $\ln Z$ with respect to V satisfies the identity

$$\frac{\partial \ln Z(y)}{\partial V^{(\ell k)}} = \frac{1}{2}\left(g_{ab}^{(\ell)}(y)\, g_{ab}^{(k)}(y) + \frac{\partial g_{ab}^{(\ell)}(y)}{\partial \omega_{ab}^{(k)}}\right), \tag{135}$$

Using $\partial_{V^{(\ell k)}} Z = Z\,\partial_{V^{(\ell k)}} \ln Z$ and (135), we obtain

$$\frac{\partial F}{\partial V^{(\ell k)}} = \int dy\; (1 + \ln Z(y))\,\frac{\partial Z(y)}{\partial V^{(\ell k)}} = \frac{1}{2}\int dy\; (1 + \ln Z(y))\, Z(y)\; \Big(g_{ab}^{(\ell)}(y)\, g_{ab}^{(k)}(y) + \partial_{\omega_{ab}^{(k)}} g_{ab}^{(\ell)}(y)\Big). \tag{136}$$

Subtracting $2\,\partial F/\partial V^{(\ell k)}$ from $\partial^2 F/(\partial \omega_{ab}^{(\ell)} \partial \omega_{ab}^{(k)})$ using (134) and (136), all terms proportional to $(1 + \ln Z)Z(\cdots)$ cancel exactly, yielding

$$\frac{\partial^2 F}{\partial \omega_{ab}^{(\ell)} \partial \omega_{ab}^{(k)}} - 2\,\frac{\partial F}{\partial V^{(\ell k)}} = \int dy\; Z(y)\, g_{ab}^{(\ell)}(y)\, g_{ab}^{(k)}(y). \tag{137}$$

Plugging (137) into (129) and summing over $a \le b$ gives

$$\frac{\partial I_{\text{out}}(q)}{\partial q_{\ell k}} = \int \Big[\prod_{a\le b} d^L\omega_{ab}\,\mathcal{N}(\omega_{ab}; 0, 2q)\Big] \sum_{a\le b} \int dy\; \mathcal{Z}_{\text{out}}(y, \omega, V)\, (g_{\text{out}}(y, \omega, V))_{ab}^{(\ell)} (g_{\text{out}}(y, \omega, V))_{ab}^{(k)}. \tag{138}$$

Therefore the output-channel state equation is

$$\widehat{q}_{\ell k} = 4\alpha\; \mathbb{E}_{\omega, y}\left[\sum_{a\le b} (g_{\text{out}}(y, \omega, V))_{ab}^{(\ell)}\; (g_{\text{out}}(y, \omega, V))_{ab}^{(k)}\right], \tag{139}$$

which is equivalent to

$$\widehat{q}_{\ell k} = 4\alpha\; \mathbb{E}_{\eta, y}\left[\sum_{a\le b} (g_{\text{out}}(y, \omega, V))_{ab}^{(\ell)} (g_{\text{out}}(y, \omega, V))_{ab}^{(k)}\right], \tag{140}$$
$$\text{for} \quad \ell = 1, \dots, L \quad a \le b = 1, \dots, T.$$

The expectation $\mathbb{E}_{(\eta,y)}$ is taken over the joint measure

$$\prod_{\ell=1}^{L} \mathcal{D}\eta^{(\ell)} \tag{141}$$

and the output y is drawn from the channel density

$$P_{\text{out}}\Big(y \mid \{h_{ab}^{(\ell)}\}_\ell\Big), \quad h_{ab}^{(\ell)} \sim \mathcal{N}\Big(\omega_{ab}^{(\ell)}, V_{ab}^{(\ell k)}\Big), \tag{142}$$

with:

$$P_{\text{out}}\Big(y \;\Big|\; \{h_{ab}^{(\ell)}\}_\ell\Big) = \delta(\{y_{ab} - g(\{h_{ab}^{(\ell)}\}_{\forall \ell})_{ab}\}_{\forall ab}), \tag{143}$$

or particularly, for the deep attention case:

$$P_{\text{out}}\Big(y \;\Big|\; \{h_{ab}^{(\ell)}\}_\ell\Big) = \delta(\{y_{ab} - B_c^L(\{h_{ab}^{(\ell)}\}_{\forall \ell})_{ab}\}_{\forall ab}). \tag{144}$$

C.3.1 Recap of the state equations

In this section we summarize the findings of the previous appendices. We performed the Bayes optimal analysis of the attention-indexed models (AIM) defined in Eq. (1): we found that the problem can be split in two components, the former involving the (extensive width) rotationally invariant prior channel and the latter involving the output channel part of the model. Through a replica analysis, we found that the prior channel is described by the following function:

$$\begin{aligned}\mathcal{Z}_{\text{in}}(\{Y_\ell\}_{\ell=1}^L; \hat{q}) = \int \left[\prod_{\ell=1}^{L} \mathrm{d}S_\ell \, P_S(S_\ell)\right] \\ \times \exp\Big[-\tfrac{d}{4}\sum_{\ell,k=1}^{L} \hat{q}_{\ell k} \operatorname{Tr}(S_\ell S_k) + \tfrac{d}{2}\sum_{\ell,k=1}^{L} \sqrt{\hat{q}}_{\ell k} \operatorname{Tr}(Y_\ell S_k)\Big].\end{aligned} \tag{145}$$

The denoising function associated to the prior channel assumes the form:

$$g_{\text{in}}(Y|\hat{q}) = \partial_{\hat{q}^{1/2}Y} \ln \mathcal{Z}_{\text{in}}(Y, \hat{q}) \tag{146}$$

The free entropy contribution coming from the prior channel is:

$$I_{\text{in}}(\hat{q}) = \lim_{d\to\infty} \frac{1}{d^2} \int DY_1 \dots DY_L \, \mathcal{Z}_{\text{in}}(Y_1, \dots, Y_L; \hat{q}) \log \mathcal{Z}_{\text{in}}(Y_1, \dots, Y_L; \hat{q}) \tag{147}$$

where DY stands for integration over a $\mathrm{GOE}(d)$ (Wigner) matrix Y.

Notably, this problem is equivalently mapped to the same posterior distribution of the following matrix denoising problem:

$$Y(S, \Delta)_\ell = S_\ell + \sum_{m=1}^{L} \sqrt{\Delta}_{\ell m} \Xi_m, \quad \Xi_m \sim \mathrm{GOE}(d) \quad S_\ell \sim P_S \tag{148}$$

On the other hand, the output channel of the model is described by the function:

$$\begin{aligned}\mathcal{Z}_{\text{out}}(y, \omega, V) = \int \left[\prod_{a\le b}^{T} d^L h_{ab} \, \mathcal{N}\left(h_{ab}; \omega_{ab}; V_{ab}\right)\right] \delta\left(y - g(\{h_{ab}^{(\ell)}/\sqrt{2-\delta_{ab}}\}_{\ell=1}^L)\right) \\ \text{with:} \quad \omega_{ab}^{(\ell)} = \sum_{k=1}^{L} \sqrt{2q}_{\ell k} \, \eta_{ab}^{(k)}, \qquad V^{(\ell k)} = 2(Q_{\ell k} - q_{\ell k})\end{aligned} \tag{149}$$

The denoising function associated to the output channel takes the form:

$$g_{\text{out}}(y, \omega, V) = \partial_\omega \ln \mathcal{Z}_{\text{out}}(y, \omega, V) \tag{150}$$

The free entropy contribution coming from the output channel is:

$$I_{\text{out}}(q) = \int \prod_{a,b=1}^{T} dy_{ab} \int \mathcal{D}\eta_1 \dots \mathcal{D}\eta_L \, \mathcal{Z}_{\text{out}}(y, \omega, V) \log \mathcal{Z}_{\text{out}}(y, \omega, V) \tag{151}$$

where $\mathcal{D}\eta$ stands for integration over a $L \times T \times T$ tensor symmetric in the token indices and with independent entries $\mathcal{N}(0,1)$.

To conclude this section, in the multi-layer setting described by the AIM framework in Eq. (1), we found the following state equations:

$$\begin{aligned}\hat{q}_{\ell k} &= 4\alpha \mathbb{E}_{\xi,\eta} \sum_{a\le b}^{L} g_{\text{out}}\left(g\left(\left\{\frac{h(\omega,V)_{ab}}{\sqrt{2-\delta_{ab}}}\right\}\right),\omega,V\right)^{(\ell)}_{ab} \\ &\quad \times g_{\text{out}}\left(g\left(\left\{\frac{h(\omega,V)_{ab}}{\sqrt{2-\delta_{ab}}}\right\}\right),\omega,V\right)^{(k)}_{ab}, \\ q_{\ell k} &= \lim_{d\to+\infty}\frac{1}{d}\mathbb{E}_{S,Y}\,\text{Tr}\left[g_{\text{in}}(Y(S,\hat{q}),\hat{q})_\ell g_{\text{in}}(Y(S,\hat{q}),\hat{q})_k\right],\end{aligned} \tag{152}$$

where $\mathbb{E}_{\eta,\xi}$ is intended as the average over $L \times T \times T$ symmetric in the token indices and Gaussians with zero mean and unit variance. Moreover, the average $\mathbb{E}_{S,Y}$ is with respect to respect to Y as given in Eq. (148) and $S \sim P_S$. Finally:

$$[h(\omega,V)_{ab}]^{(\ell)} = \omega^{(\ell)}_{ab} + \sum_k \sqrt{V}^{(\ell k)}\xi^{(k)}_{ab} \tag{153}$$

C.4 The fixed point of AMP is described by the state equations

We start by defining a new variable $\omega^*_{\mu,\,ab}$ such that $y_\mu = g(\{\omega^*_{\mu,\,ab}/\sqrt{2-\delta_{ab}}\}_{a\le b})$, where we can assume that $\mathbb{E}[(\omega^*_{\mu,\,ab})_\ell\,(\omega^*_{\mu,\,ab})_k] = 2Q^t_{\ell k}$. Our first step is to define the quantities m^t and q^t on the iterates of AMP

$$m^t_{\ell k} = \text{Tr}[\hat{S}^t_\ell S^*_k]/d\,, \qquad q^t_{\ell k} = \text{Tr}[\hat{S}^t_\ell \hat{S}^t_k]/d\,. \tag{154}$$

We now claim, in analogy with [44, 35, 40, 42] that for every sample μ and every couple of tokens $a \le b$ the variables $\omega^t_{\mu,\,ab}$ at each time converge to independent centered Gaussian variables with the following covariances

$$\mathbb{E}[(\omega^t_{\mu,\,ab})_\ell\,(\omega^t_{\mu,\,ab})_k] = 2q^t_{\ell k}\,, \qquad \mathbb{E}[(\omega^*_{\mu,\,ab})_\ell\,(\omega^t_{\mu,\,ab})_k] = 2m^t_{\ell k}\,, \tag{155}$$

By Nishimori's identities [50] we can assume $m^t = q^t$. The first equation of (21) is now immediately recovered (modulo the substitution $V \to 2(Q-q^t)$ which will come after)

$$\hat{q}^t_{\ell k} \approx 4\alpha \mathbb{E}_{y,\,\omega^t} \sum_{a\le b}^{T}\left[g_{\text{out}}(y,\omega^t,V^t)^{(\ell)}_{ab} g_{\text{out}}(y,\omega^t,V^t)^{(k)}_{ab}\right] \tag{156}$$

where

$$y = g\left(\left\{\frac{\omega^*_{ab}}{\sqrt{2-\delta_{ab}}}\right\}^{T}_{a\le b}\right), \qquad \begin{pmatrix}\omega^t_{ab}\\ \omega^*_{ab}\end{pmatrix} \sim \mathcal{N}\left(0, \begin{pmatrix}2q^t & 2m^t\\ 2m^t & 2Q\end{pmatrix}\right) \tag{157}$$

Again as in [44, 35, 40, 42] we will have that in distribution

$$R^t_{ij} = S^*_{ij} + (\hat{q}^t)^{-1}\Xi^t_{ij} \tag{158}$$

We are ready to close the circle: going back to the definition of q^t we write

$$q^t_{\ell k} = \mathbb{E}_{R^t}\,\text{Tr}\left[g_{\text{in}}\left(R^t,\hat{q}^t\right)_\ell g_{\text{in}}\left(R^t,\hat{q}^t\right)_k\right]/d\,, \tag{159}$$

which is exactly the second equation in (21). Notice how the expectation is taken over the random variable R^t_{ij} in (158). The last step is to notice that

$$\hat{C}^t_{\ell k} = \text{Tr}[(\hat{S}^t_\ell - S^*_\ell)(\hat{S}^t_k - S^*_k)]/d^2 = Q - q^t \tag{160}$$

such that $V^t = 2(Q-q^t)$.

D The case of $L = 1$ layer

In this Appendix we restrict the theoretical results derived for an arbitrary number of layers to the particular case of $L = 1$ layer. In this particular case, the order parameters q and $\hat{q}$ become scalar quantities. Moreover, in the following analysis we specialize to the extensive-rank choice:

$$S = \frac{1}{\sqrt{rd}} WW^\top \in \mathbb{R}^{d\times d} \quad W \in \mathbb{R}^{d\times r} \quad (W)_{ij} \sim \mathcal{N}(0,1) \tag{161}$$

with rank ratio $\rho = r/d = O(1)$. Thus, the spectral distribution of the symmetric matrix S is that of the Marcenko-Pastur law for Wishart matrices described in App. (A).

D.1 Prior channel state equation

Starting from Eq. (95) for $L = 1$ layer, we get::

$$\begin{aligned} I_{\rm in}(\hat{q}) &= \lim_{d\to\infty} \frac{1}{d^2} \int DY\, \mathcal{Z}_{\rm in}(Y;\hat{q}) \log \mathcal{Z}_{\rm in}(Y;\hat{q}) \\ \mathcal{Z}_{\rm in}(Y;\hat{q}) &= \int dS\, P_0(S) \exp\left(-\frac{d}{2}\left(\hat{Q} + \frac{\hat{q}}{2}\right) \mathrm{Tr}\left(S^T S\right) + \frac{\sqrt{\hat{q}}d}{2} \mathrm{Tr}\left(Y^T S\right)\right). \end{aligned} \tag{162}$$

Again, at the 0-replica order $\hat{Q} = 0$ and integrating over Y:

$$\begin{aligned} &\int \mathcal{D}Y\, \mathcal{Z}_{\rm in}(Y;\hat{q}) \qquad (163)\\ &= \int \mathcal{D}Y \int dS\, P_0(S) \exp\left(-\frac{\hat{q}d}{4} \mathrm{Tr}(S^\top S) + \frac{\sqrt{\hat{q}}d}{2} \mathrm{Tr}\left(Y^\top S\right) - \frac{1}{4} \mathrm{Tr}\left(Y^\top Y\right)\right) \\ &= \int dS\, P_0(S) \exp\left(\frac{\hat{q}d}{4} \mathrm{Tr}\left(S^\top S\right)\right) \exp\left(-\frac{\hat{q}d}{4} \mathrm{Tr}\left(S^\top S\right)\right) \\ &= \int dS\, P_0(S) = 1 \qquad (164) \end{aligned}$$

Now, note that the exponent in $\mathcal{Z}_{\rm in}(Y;\hat{q})$ can be rearranged as:

$$-\tfrac{\hat{q}d}{4} \mathrm{Tr}(S^T S) + \frac{\sqrt{\hat{q}}d}{2} \mathrm{Tr}(S^T Y) = -\frac{d}{4} \mathrm{Tr}\Big(\hat{q}\, S^T S - 2\sqrt{\hat{q}}\, S^T Y\Big). \tag{165}$$

Observe

$$\mathrm{Tr}\Big[(\sqrt{\hat{q}}\, S - Y)^T (\sqrt{\hat{q}}\, S - Y)\Big] = \hat{q}\, \mathrm{Tr}(S^T S) \;-\; 2\sqrt{\hat{q}}\, \mathrm{Tr}(S^T Y) \;+\; \mathrm{Tr}(Y^T Y). \tag{166}$$

Hence

$$-\tfrac{\hat{q}}{4} \mathrm{Tr}(S^T S) \;+\; \frac{\sqrt{\hat{q}}}{2} \mathrm{Tr}(Y^T S) = -\tfrac{1}{4} \mathrm{Tr}\big[(\sqrt{\hat{q}}\, S - Y)^2\big] \;+\; \tfrac{1}{4} \mathrm{Tr}(Y^T Y). \tag{167}$$

Therefore:

$$I_0(Y) = \exp\Big[+\tfrac{1}{4} \mathrm{Tr}(Y^T Y)\Big] \;\times\; \int \mathrm{d}S\, P_0(S)\, \exp\Big[-\tfrac{1}{4} \mathrm{Tr}\big(\sqrt{\hat{q}}\, S - Y\big)^2\Big]. \tag{168}$$

Ignoring the factor $\exp(\frac{1}{4} \mathrm{Tr}(Y^T Y))$ that is independent of S, we see that

$$\int \mathrm{d}S\, P_0(S)\, \exp\Big[-\tfrac{d}{4} \mathrm{Tr}\big(\sqrt{\hat{q}}\, S - Y\big)^2\Big] \tag{169}$$

which plays the role of a posterior density for S given $Y = \sqrt{\hat{q}}\, S + Z$ with Z a GOE(d) noise.

In the large-d limit, let us parametrize S by its eigenvalues:

$$S = U\, \Lambda\, U^T \tag{170}$$

where $\Lambda = \mathrm{diag}(\lambda_1, \ldots, \lambda_d)$. Then

$$\mathrm{d}S \;=\; \Big[\prod_{i=1}^{d} \mathrm{d}\lambda_i\Big]\, \big|\Delta(\{\lambda_i\})\big|\, \mathrm{d}U \quad \text{with} \quad \Delta(\{\lambda_i\}) \;=\; \prod_{1\le i<j\le d} \big|\lambda_i - \lambda_j\big|, \tag{171}$$

Then the exponent

$$\mathrm{Tr}\Big[-\tfrac{1}{4}\,(\sqrt{\hat{q}}\,S - Y)^2\Big] \tag{172}$$

becomes

$$-\tfrac{1}{4}\,\mathrm{Tr}\Big(\sqrt{\hat{q}}\,U\,\Lambda\,U^T - Y\Big)^2. \tag{173}$$

We can factor out the integral over $U \in \mathcal{O}(d)$ and for d large:

$$\int_{\mathcal{O}(d)} \exp\Big(\tfrac{\hat{q}\,d}{2}\,\mathrm{Tr}[\Lambda\,U^T\,Y\,U]\Big)\,\mathcal{D}U \;\;\approx\;\; \exp\Big[\tfrac{d^2}{2}\,I_{\mathrm{HCIZ}}\big(\hat{q};\;\mu_\Lambda,\;\mu_Y\big)\Big], \tag{174}$$

where I_{HCIZ} is an explicit functional in the limit $d \to \infty$ of dimension $2/d^2$ times the log of that integral, and μ_Λ is the limiting spectral distribution of $\Lambda/\sqrt{d}$.

The prior contribution of the free entropy is given by

$$\Phi_{\mathrm{prior}}(\hat{q}) \;=\; \lim_{d\to\infty}\,\frac{1}{d^2}\,\mathbb{E}\big[\ln I_0(Y)\big], \tag{175}$$

or more explicitly:

$$\Phi_{\mathrm{prior}}(\hat{q}) = \lim_{d\to\infty}\,\frac{1}{d^2}\,\mathbb{E}\big[\ln \int P_0(S)\, e^{-\frac{d}{4}\,\mathrm{Tr}(\sqrt{\hat{q}}\,S - Y)^2}\,\mathrm{d}S\big]. \tag{176}$$

This term can be explicitly computed and mapped to a matrix estimation problem. i.e. a denoising problem as follows:

$$\Phi_{\mathrm{prior}}(\hat{q}) = \lim_{d\to\infty}\,\frac{1}{d^2}\,\mathbb{E}_Y \ln I_0(Y) \;=\; -\frac{\hat{q}\,Q}{4} + \frac{1}{2} I_{\mathrm{HCIZ}}\Big(\hat{q}; \mu_0, \mu_0 \boxplus \sigma_{\mathrm{sc},1/\sqrt{\hat{q}}}\Big) + \text{const}, \tag{177}$$

where $Q = 1 + \rho$.Then, one has the relation from [51]:

$$-\frac{1}{2}\Sigma\big(\mu_{\hat{q}}\big) + \frac{1}{4\hat{q}}\mathbb{E}_{\mu_{\hat{q}}}[X^2] - \frac{1}{2} I_{\mathrm{HCIZ}}\Big(\hat{q}; \mu_0, \mu_{\hat{q}}\Big) - \frac{3}{8} + \frac{1}{4}\ln\hat{q} + \frac{1}{4\hat{q}}\mathbb{E}_{\mu_0}[X^2] = 0, \tag{178}$$

where we have defined $\mu_{\hat{q}} = \mu_0 \boxplus \sigma_{\mathrm{sc},1/\sqrt{\hat{q}}}$ and $\Sigma(\mu)$ is the noncommutative entropy:

$$\Sigma(\mu) = \int \mu(dx)\mu(dy)\ln|x-y|.$$

In our normalization (with $Q = 1 + \rho$), rearranging yields

$$\frac{1}{2} I_{\mathrm{HCIZ}}(\hat{q}; \mu_0, \mu_{\hat{q}}) = \; -\frac{1}{2}\Sigma(\mu_{\hat{q}}) + \frac{1}{4}\Big[2Q\hat{q} + 1\Big] - \frac{3}{8} - \frac{1}{4}\ln\hat{q}. \tag{179}$$

Plugging back into the free entropy, we obtain

$$\Phi_{\mathrm{prior}}(\hat{q}) = \; -\frac{\hat{q}\,Q}{4} \;+\; \left[-\frac{1}{2}\Sigma(\mu_{\hat{q}}) + \frac{1}{4}(2Q\hat{q}+1) - \frac{3}{8} - \frac{1}{4}\ln\hat{q}\right] + \text{const}. \tag{180}$$

Taking the derivative with respect to $\hat{q}$ yields the "prior state" equation. In fact, differentiating we obtain

$$\frac{\partial\Phi}{\partial\hat{q}} = -\frac{q}{4} + \frac{Q}{4} - \frac{1}{4\hat{q}} - \frac{1}{2}\frac{\partial}{\partial\hat{q}}\Sigma\big(\mu_{\hat{q}}\big) = 0. \tag{181}$$

Using the derivative:

$$\frac{\partial}{\partial\hat{q}}\Sigma\big(\mu_{\hat{q}}\big) = -\frac{2\pi^2}{3\hat{q}^2}\int \mu_{\hat{q}}(x)^3\,dx, \tag{182}$$

this condition becomes

$$-\frac{q}{4} + \frac{Q}{4} - \frac{1}{4\hat{q}} + \frac{\pi^2}{3\hat{q}^2}\int \mu_Y(x)^3\,dx = 0. \tag{183}$$

which is exactly our desired state equation.

To sum up, in the problem

$$Y \;=\; \sqrt{\hat{q}}\, S \;+\; Z, \quad Z \sim \mathrm{GOE}(d), \tag{184}$$

the law of Y is asymptotically $\mu_S \;\boxplus\; \sigma_{\mathrm{sc},\,1/\sqrt{\hat{q}}}$. we finally get:

$$q = Q - \frac{1}{\hat{q}} + \frac{4\,\pi^2}{3\,\hat{q}^{\,2}} \int \big[\mu_Y(x)\big]^3 \,\mathrm{d}x, \tag{185}$$

with $\mu_Y = \mu_S \boxplus \sigma_{\mathrm{sc},\,1/\sqrt{\hat{q}}}$.

For the computation of μ_Y, we recall that if $\mu_Y = \mu_S \boxplus \sigma_{\mathrm{sc},\alpha}$, we can write

$$\mathcal{R}_{\mu_Y}(z) \;=\; \mathcal{R}_{\mu_S}(z) \;+\; \mathcal{R}_{\sigma_{\mathrm{sc},\,\alpha}}(z). \tag{186}$$

For the semicircle of radius α, we have $\mathcal{R}_{\sigma_{\mathrm{sc},\,\alpha}}(z) = \alpha^2\, z$. For μ_S (Marchenko–Pastur distribution with parameter ρ), we have

$$\mathcal{R}_{\mu_{MP,\rho}}(z) = \frac{\rho}{\sqrt{\rho} - z}. \tag{187}$$

In our case $\alpha = 1/\sqrt{\hat{q}}$, then

$$\mathcal{R}_{\mu_Y}(z) = \frac{\rho}{\sqrt{\rho} - z} \;+\; \alpha^2\, z. \tag{188}$$

From $\mathcal{R}_{\mu_Y}(z) = g^{-1}_{\mu_Y}(-z) - \frac{1}{z}$, one obtains an equation for $g_{\mu_Y}(z)$, with:

$$g_{\mu_Y}(z) = \int \frac{\mu_Y(\mathrm{d}x)}{x - z}. \tag{189}$$

So, using the identity $z = \frac{1}{x} + R_{\mu_Y}(x)$, where $x = g_{\mu_Y}(z)$. So we get

$$z \;=\; \frac{1}{x} \;+\; \frac{\rho}{\sqrt{\rho} - x} \;+\; \alpha^2\, x. \tag{190}$$

Hence the final polynomial in x is:

$$(\frac{1}{\sqrt{\rho}}\alpha^2)x^3 - (\frac{z}{\sqrt{\rho}} + \alpha^2)x^2 + (z + \frac{1}{\sqrt{\rho}} - \sqrt{\rho})x - 1 = 0 \iff x = g_{\mu_Y}(z). \tag{191}$$

We look for the solution of this equation with largest imaginary part. Moreover, we compute the discriminant of this third order equation in order to correctly quantify the edges of the spectral density we want to numerically compute.

Recalling $\alpha^2 = 1/\hat{q}$, the imaginary part of x yields μ_Y (Stieltjes–Perron inversion), i.e.

$$\mu_Y(x_0) \;=\; \lim_{\epsilon \to 0^+} \frac{1}{\pi}\, \mathrm{Im}\, g_{\mu_Y}(x_0 - i\,\epsilon). \tag{192}$$

D.2 Small/Large width limit of the prior channel

We recall that the state equations are of the form

$$\begin{aligned} Q - q &= \frac{1}{\hat{q}} - \frac{4\pi^2}{3\hat{q}^2} \int dx\, \mu_{1/\hat{q}}(x)^3 \\ \hat{q} &= 2\alpha F(Q - q, q) \end{aligned} \tag{193}$$

where $\mu_{1/\hat{q}}$ is the spectral distribution of $S_* + \frac{1}{\sqrt{\hat{q}}} Z$ and $Q = d^{-1}\,\mathrm{Tr}(S_*^2)$. In our examples, S^* is $\sqrt{\rho}$ times a standard Wishart, with $Q = 1 + \rho$.

D.2.1 Small width limit

We follow [40, Section E.1.1]. Call $t = \rho/\hat{q}$, and $\bar{\alpha} = \alpha/\rho$. Call ν the distribution of $\sqrt{\rho}(S_* + \frac{1}{\sqrt{\hat{q}}}Z) = \sqrt{\rho}S_* + \sqrt{t}Z$, i.e.

$$\nu(y) = \rho^{-1/2}\mu_{1/\hat{q}}(\rho^{-1/2}y)\,. \tag{194}$$

Notice that this is precisely the ν defined in [40, Eq. 56]. Then we have

$$\begin{aligned} Q - q &= \frac{1}{\hat{q}} - \frac{4\pi^2}{3\hat{q}^2}\int dx\,\mu_{1/\hat{q}}(x)^3 \\ &= \frac{t}{\rho} - \frac{4\pi^2 t^2}{3\rho^2}\rho\int dy\,[\rho^{-1/2}\mu_{1/\hat{q}}(\rho^{-1/2}y)]^3 \\ &= \frac{t}{\rho} - \frac{4\pi^2 t^2}{3\rho}\int dy\,\nu(y)^3 \\ &= \frac{t}{\rho}\left[1 - \frac{4\pi^2 t}{3}\int dy\,\nu(y)^3\right] \\ &\approx \begin{cases} t(2-t) & \text{if} \quad t \le 1 \\ 1 & \text{if} \quad t > 1 \end{cases} \end{aligned} \tag{195}$$

where we used [40, Eq. 57 and following] to take the limit of small κ at leading order. Thus, the equations can be recast to

$$\begin{aligned} Q - q &= \begin{cases} t(2-t) & \text{if} \quad t \le 1 \\ 1 & \text{if} \quad t > 1 \end{cases} \\ t &= \frac{1}{2\bar{\alpha}F(Q-q)}\,. \end{aligned} \tag{196}$$

In particular, we have a weak recovery threshold. Indeed, as long as

$$\bar{\alpha} < \frac{1}{2F(1)} \tag{197}$$

we have that $Q - q = 1$, i.e. the same error as the average from the prior (BO estimator with no data).

D.2.2 Large width limit

Recall that $Q = 1 + \rho$ and $q \in [\rho, 1+\rho]$, so that $Q - q \in [0, 1]$ even in the $\rho \to \infty$ limit. Then we have

$$\begin{aligned} Q - q &= \frac{1}{\hat{q}} - \frac{4\pi^2}{3\hat{q}^2}\int dx\,\mu_{1/\hat{q}}(x)^3 \\ &= \frac{1}{\hat{q}}\left[1 - \frac{4\pi^2}{3\hat{q}\rho}\int dy\,[\sqrt{\rho}\mu_{1/\hat{q}}(\sqrt{\rho}y)]^3\right] \\ &= \frac{1}{\hat{q}}\left[1 - \frac{4\pi^2}{3\hat{q}\rho}\int dy\,\mu_{1/\rho\hat{q}}(y)^3\right] \\ &\approx \frac{1}{\hat{q}}\left[1 - \frac{1}{1+\hat{q}}\right] \\ &\approx \frac{1}{1+\hat{q}}\,, \end{aligned} \tag{198}$$

where we used [40, Section E.2].

D.3 Output channel state equation

For $L = 1$ layers we obtain the state equation for the output channel contribution:

$$\hat{q} = 4\,\alpha\,\mathbb{E}_{(\eta,y)}\Big[\sum_{a\le b}(g_{\text{out}}(y,\omega,V))^2_{ab}\Big] \tag{199}$$

Where we remind $\eta_{ab} \sim \mathcal{N}(0,1)$ with $a \leq b = 1, \dots, T$ and $\omega_{ab} = \sqrt{2q}\ \eta_{ab}$, $V = 2(Q-q)$, $Q = 1+\rho$. The denoising function is given by:

$$(g_{\text{out}}(y,\omega,V))_{ab} = \partial_{\omega_{ab}} \ln \mathcal{Z}_{\text{out}}(y,\omega,V) \tag{200}$$

and:

$$\mathcal{Z}_{\text{out}}(y,\omega,V) = \int \prod_{a\leq b} dh_{ab} \mathcal{N}(h_{ab},\omega_{ab},V)\, \delta\big(y - f(h)\big) \tag{201}$$

where $f(h)$ depends on the precise choice of the model. In particular, in the following sections we consider the three cases dealt in the main text. In the following, we first consider a linear output channel for a generic number of tokens T. This simple case serves as a baseline for the more interesting case of the softmax channel, namely the self-attention layer for an arbitrary number of tokens. We also consider the hardmax variant of the model treated in the main text for $T = 2$ tokens.

D.4 Linear output channel for generic number of tokens

We consider $P_{\text{out}}(y_{ab} \mid h_{ab}) = \delta(y_{ab} - \frac{h_{ab}}{\sqrt{2-\delta_{ab}}})$. Then

$$\mathcal{Z}_{\text{out}}(y,\omega,V) = \int \Big[\prod_{a\leq b} \mathcal{N}\big(h_{ab};\omega_{ab},V\big)\Big] \prod_{a\leq b} \delta\big[y_{ab} - \frac{h_{ab}}{\sqrt{2-\delta_{ab}}}\big]\, \mathrm{d}h. \tag{202}$$

Enforcing $h_{ab} = \sqrt{2-\delta_{ab}}y_{ab}$, this gives directly:

$$\mathcal{Z}_{\text{out}}(y,\omega,V) = \prod_{a\leq b}\Big[\frac{1}{\sqrt{2\pi\, V}} \exp\Big(-\frac{(\sqrt{2-\delta_{ab}}y_{ab}-\omega_{ab})^2}{2\,V}\Big)\Big]. \tag{203}$$

Hence

$$\ln \mathcal{Z}_{\text{out}}(y,\omega,V) = \sum_{a\leq b}\Big[-\tfrac{1}{2}\ln(2\pi\, V)\ -\ \frac{(\sqrt{2-\delta_{ab}}y_{ab}-\omega_{ab})^2}{2\,V}\Big]. \tag{204}$$

We recall that $\omega_{ab}(\eta)$ depends linearly on η_{ab}, e.g.:

$$\omega_{ab} = \sqrt{2q}\ \eta_{ab}, \quad V_{ab} = 2(Q-q), \quad Q\ =\ 1+\rho. \tag{205}$$

Then

$$(g_{\text{out}}(y,\omega,V))_{ab} = \frac{\partial}{\partial\omega_{ab}} \ln \mathcal{Z}_{\text{out}}(y,\omega,V) = -\frac{\partial}{\partial\omega_{ab}}\Big[\frac{(\sqrt{2-\delta_{ab}}y_{ab}-\omega_{ab}(\eta))^2}{2\,V}\Big] = +\frac{(\sqrt{2-\delta_{ab}}y_{ab}-\omega_{ab})}{V}. \tag{206}$$

Thus

$$\sum_{a\leq b}(g_{\text{out}}(y,\omega,V))^2_{ab} = \sum_{a\leq b}\Big([\,\sqrt{2-\delta_{ab}}y_{ab} - \omega_{ab}(\eta)\,]\frac{1}{2(Q-q)}\Big)^2. \tag{207}$$

We can compute the expectation:

$$\begin{aligned}&\mathbb{E}\Big[\sum_{a\leq b}(g_{\text{out}}(y,\omega,V))^2_{ab}\Big]\\ &=\ \int\Big(\prod_{a\leq b}\mathrm{d}\eta_{ab}\,\frac{e^{-\eta^2_{ab}/2}}{\sqrt{2\pi}}\Big)\ \int\Big(\prod_{a\leq b}\mathrm{d}y_{ab}\Big)\,\mathcal{Z}_{\text{out}}(y,\omega,V)\ \sum_{a\leq b}(g_{\text{out}}(y,\omega,V))^2_{ab}.\end{aligned} \tag{208}$$

We can simply use:

$$\int \mathrm{d}y_{ab}\,\frac{1}{\sqrt{2\pi\, V}}\exp\Big(-\frac{(\sqrt{2-\delta_{ab}}y_{ab}-\omega_{ab})^2}{2\,V}\Big)\,\big[\sqrt{2-\delta_{ab}}y_{ab}-\omega_{ab}\big]^2 = V. \tag{209}$$

Therefore,

$$\int\Big(\prod_{a\leq b}\mathrm{d}y_{ab}\Big)\,\mathcal{Z}_{\text{out}}(y,\omega,V)\sum_{a\leq b}(g_{\text{out}}(y,\omega,V))^2_{ab}\ =\ \sum_{a\leq b}\Big[\big(\frac{1}{2(Q-q)}\big)^2\,V\Big]. \tag{210}$$

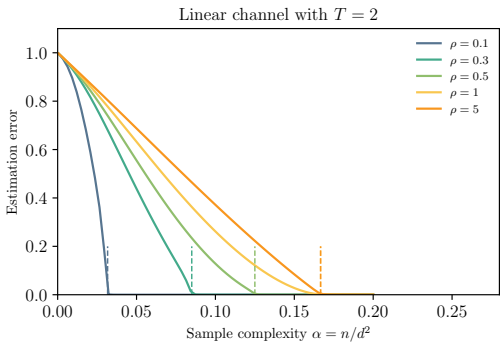

Figure 3: Illustration of the Bayes-optimal error for the linear output channel baseline in Eq. (202), for $T = 2$ tokens and several values of the width ratio $\rho = r/d$. The model reaches zero BO error at finite α. The recovery threshold matches perfectly the one find by the simple counting argument in (215), plotted in short vertical lines.

Thus each term becomes

$$\left(\frac{1}{2(Q-q)}\right)^2 2(Q-q) \;=\; \frac{1}{2(Q-q)}. \tag{211}$$

Hence the entire sum is

$$\sum_{a\le b}\frac{1}{2(Q-q)} = \frac{T(T+1)}{4}\frac{1}{Q-q}. \tag{212}$$

Notice that this result does not depend on η. Consequently, the outer integral over η becomes 1.Hence we arrive to the final form of the linear output channel state equation:

$$\hat{q} = 4\,\alpha\,\mathbb{E}_{(\eta,y)}\Big[\sum_{a\le b}(g_{\rm out}(y,\omega,V))^2_{ab}\Big] = 4\alpha\sum_{a\le b}\Big[\mathbb{E}_{(\eta,y)}g_{\rm out}(y,\omega,V))^2_{ab}\Big] = 4\alpha\frac{T(T+1)}{4(Q-q)} \tag{213}$$

which finally simplifies into the output channel state equation:

$$\hat{q} = \frac{T(T+1)\;\alpha}{Q-q} \tag{214}$$

As an example, in Fig. 3 we show the fixed point solution for the state evolution equations for the linear output channel. The prior equation (26) remains unchanged, while we use Eq. (213) for simulating the linear output channel results. We also show in vertical dashed lines the recovery threshold found by the simple counting problem:

$$\frac{T(T+1)}{2}\alpha_{\rm count} = \rho - \frac{\rho^2}{2} \tag{215}$$

The linear output channel matches perfectly the counting recovery threshold, unlike in the softmax case shown in Eq. (32).

D.5 Softmax output channel for generic number of tokens

We compute the quantity:

$$\mathcal{Z}_{\rm out}(y,\omega,V) = \int\prod_{a\le b}dh_{ab}\,\frac{1}{\sqrt{2\pi V_{ab}}}e^{-\frac{(h_{ab}-\omega_{ab})^2}{2V_{ab}}}\prod_{a\le b}\delta(y_{ab}-{\rm Softmax}\{\frac{\beta}{\sqrt{2-\delta_{ab}}}h_{ab}\}). \tag{216}$$

where we remind the factor $\sqrt{2-\delta_{ab}}$ is present due to the symmetrization of the problem (i.e. multiply and divide by $\sqrt{2-\delta_{ab}}$), allowing a much simpler treatment of the BO analysis in change of this slight modification of the output channel.

From now on, we define the quantity $\tau_{ab} = \sqrt{2 - \delta_{ab}}$. We thus aim to compute the quantity:

$$\mathcal{Z}_{\text{out}}(y, \omega, V) = \int \prod_{a \leq b} dh_{ab} \delta(y - \sigma(\frac{h_{ab}}{\tau_{ab}}) \prod_{a \leq b} \mathcal{N}(h_{ab}, \omega_{ab}, V_{ab}) \tag{217}$$

We introduce the variable $z_{ab} = h_{ab}/\tau_{ab}$ and exploit $dh\mathcal{N}(h, \mu, \sigma) = dz\mathcal{N}(z, \mu/\tau, \sigma/\tau^2)$, we get:

$$\begin{aligned} \mathcal{Z}_{\text{out}}(y, \omega, V) &= \int \prod_{a \leq b} dz_{ab}\, \delta\big(y - \sigma(z)\big) \prod_{a \leq b} \mathcal{N}\big(z_{ab}, \tfrac{\omega_{ab}}{\tau_{ab}}, \tfrac{V_{ab}}{\tau_{ab}^2}\big) = \\ &= \int \prod_{a \leq b < T} dt_{ab}\, \mathcal{N}\big(t_{ab}, \tfrac{\omega_{ab}}{\tau_{ab}} - s_a, \tfrac{V_{ab}}{\tau_{ab}^2}\big) \prod_{a=1}^{T} ds_a\, \mathcal{N}\big(s_a, \tfrac{\omega_{aT}}{\tau_{aT}}, \tfrac{V_{aT}}{\tau_{aT}^2}\big) \end{aligned} \tag{218}$$

where in the last equality we introduced the inverse mapping of the row-wise softmax function, defined in Eq. (40). In particular, we introduce:

$$e^{\beta t_{ab}} = \frac{e^{\beta z_{ab}}}{e^{\beta z_{aT}}} = \frac{e^{\beta z_{ab}}}{\sum_{b=1}^{T} e^{\beta z_{ab}}} \Big(\frac{e^{\beta z_{aT}}}{\sum_{b=1}^{T} e^{\beta z_{ab}}} \Big)^{-1} = \frac{y_{ab}}{y_{aT}} \quad \forall a \leq b < T \tag{219}$$

which leads to:

$$t_{ab} = \frac{1}{\beta} \log(\frac{y_{ab}}{y_{aT}}) = \phi_{ab}(y) \quad \forall a \leq b < T \tag{220}$$

while for $b = T$:

$$\frac{y_{Ta}}{y_{TT}} = \frac{e^{\beta z_{Ta}}}{e^{\beta z_{TT}}} = \frac{e^{\beta z_{aT}}}{e^{\beta z_{TT}}} = e^{\beta(s_a - s_{TT})} \to s_a = s_{TT} + \phi_{Ta}(y) \quad \forall a < T \tag{221}$$

having introduced the change of variables:

$$z_{ab} \to t_{ab} = z_{ab} - z_{aT} \to z_{ab} = t_{ab} + s_a = \phi_{ab} + \phi_{Ta} + s_{TT} \quad \forall a \leq b < T \tag{222}$$

and

$$z_{aT} \to s_a = z_{aT} \to z_{aT} = s_a = s_{TT} + \phi_{Ta} \quad \forall a < T \tag{223}$$

Having this mapping clear and introducing the short-hand notation $\tilde{\omega} = \omega/\tau$ and $\tilde{V} = V/\tau^2$, $s_{TT} = x$, we can see that we can reduce the computation of Eq. (218) to that of one simple scalar integral in the variable $x = s_T$, namely:

$$\begin{aligned} \mathcal{Z}_{\text{out}}(y, \omega, V) &= \int dx \mathcal{N}(x, \tilde{\omega}_{TT}, \tilde{V}_{TT}) \prod_{a=1}^{T-1} \mathcal{N}(x + \phi_{Ta}(y), \tilde{\omega}_{aT}, \tilde{V}_{aT}) \\ &\quad \times \prod_{a \leq b < T} \mathcal{N}(\phi_{ab}(y) + \phi_{Ta}(y) + x, \tilde{\omega}_{ab}, \tilde{V}_{ab}) \\ &= \int dx \exp\Big\{ -\frac{1}{2} \Big[\sum_{a=1}^{T} \frac{(x + \phi_{Ta} - \tilde{\omega}_{aT})^2)}{\tilde{V}_{aT}} + \sum_{a \leq b < T} \frac{(\phi_{ab} + \phi_{Ta} + x - \tilde{\omega}_{ab})^2}{\tilde{V}_{ab}} \Big] \Big\} \end{aligned} \tag{224}$$

We thus obtain a simple gaussian integral whose exponential is of the form:

$$\begin{aligned} &-\frac{1}{2} \Big[x^2 \Big(\sum_{a \leq b} \tilde{V}_{ab}^{-1} \Big) + 2x \Big(\sum_{a=1}^{T} \frac{\phi_{Ta} - \tilde{\omega}_{aT}}{\tilde{V}_{aT}} + \sum_{a \leq b < T} \frac{\phi_{ab} + \phi_{Ta} - \tilde{\omega}_{aT}}{\tilde{V}_{ab}} \Big) \\ &\quad + \Big(\sum_{a=1}^{T} \frac{(\phi_{Ta} - \tilde{\omega}_{aT})^2}{\tilde{V}_{aT}} + \sum_{a \leq b < T} \frac{(\phi_{ab} + \phi_{Ta} - \tilde{\omega}_{ab})^2}{\tilde{V}_{ab}} \Big) \Big] \end{aligned} \tag{225}$$

Having computed this simple gaussian integral, we can hence compute the quantity of interest:

$$\begin{aligned} \log \mathcal{Z} &= \frac{1}{2\tilde{V}} \Big[\sum_{a=1}^{T} \frac{\phi_{Ta} - \tilde{\omega}_{aT}}{\tilde{V}_{aT}} + \sum_{a \leq b < T} \frac{\phi_{ab} + \phi_{Ta} - \tilde{\omega}_{aT}}{\tilde{V}_{ab}} \Big]^2 \\ &\quad - \frac{1}{2} \Big[\sum_{a=1}^{T} \frac{(\phi_{Ta} - \tilde{\omega}_{aT})^2}{\tilde{V}_{aT}} + \sum_{a \leq b < T} \frac{(\phi_{ab} - \phi_{Ta} - \tilde{\omega}_{ab})^2}{\tilde{V}_{ab}} \Big] + \text{cost} \end{aligned} \tag{226}$$

with $\tilde{V} = \sum_{a\leq b} \tilde{V}_{ab}^{-1}$ and again $\phi_{ab}(y) = \frac{1}{\beta}\log\frac{y_{ab}}{y_{aT}}$, $\tilde{\omega}_{ab} = \frac{\omega_{ab}}{\sqrt{2-\delta_{ab}}}$,$\tilde{V}_{ab} = \frac{Vab}{2-\delta_{ab}}$,$\omega_a = \sqrt{2q}\eta_{ab}$, $V_{ab} = V = 2(Q-q)$, $h \sim \mathcal{N}(\tilde{\omega}, \tilde{V})$, $y = \sigma(h)$. The constant term contains those terms independent from ω, as we are finally interested in the denoising function , which is the derivative:

$$g_{\text{out}}(y,\omega,V)_{ab} = \partial_{\omega_{ab}} \log \mathcal{Z}_{\text{out}}(y,\omega,V) \tag{227}$$

We thus compute the denoising function deriving with quantity $\log \mathcal{Z}_{\text{out}}(y,\omega,V)$ with respect to $\tilde{\omega}$, thus computing $\tau_{ij}\partial_{\omega_{ij}} \log \mathcal{Z}_{\text{out}}(y,\omega,V)$ for $i \leq j < T$ and for $j = T$. We also consider that $\sum_{a\leq b}^{T} \tilde{V}_{ab}^{-1} = \frac{1}{V}\sum_{a\leq b}^{T}(2-\delta_{ab}) = \frac{T^2}{V}$, $V = 2(Q-q)$.

Finally, we obtain the final form of the denoising function of the softmax output channel in Eq. (27) for an arbitrary number of tokens, substituting back the original V and ω:

$$V(g_{\text{out}})_{ij} = -\frac{\tau_{ij}}{T^2}\Big[\sum_{a\leq b}^{T}\tau_{ab}^2\phi_{Ta} - \sum_{a\leq b}^{T}\tau_{ab}\omega_{ab} + \sum_{a\leq b}^{T-1}\tau_{ab}^2\phi_{ab}\Big] + \tau_{ij}\phi_{Ti} - \omega_{ij} + \delta(j<T)\phi_{ij}\tau_{ij} \tag{228}$$

which is exactly the same form appeared in the main text in Eq. (31).

We now complete this appendix by computing the quantity $\mathbb{E}_{\eta,y}\sum_{a\leq b}(g_{\text{out}})_{ab}^2$. To do so, we exploit the following relations:

$$\phi_{ab} = h_{ab} - h_{aT} \quad h \sim \mathcal{N}(\tilde{\omega}, \tilde{V}) \rightarrow \tau_{ab}h_{ab} = \sqrt{2q}\,\eta_{ab} + \sqrt{V}\,\xi_{ab} \quad a \leq b \leq T \tag{229}$$

with $\eta_{ab}, \xi_{ab} \sim \mathcal{N}(0,1)$ and

$$\phi_{Ta} = h_{Ta} - h_{TT} = h_{aT} - h_{TT} \tag{230}$$

We thus substitute these relationships inside Eq. (228) and finally compute $\mathbb{E}_{\eta,\xi}\sum_{a\leq b}(g_{\text{out}})_{ab}^2$. After a long but simple algebraic calculation, it is possible to show that the denoiser function reduces to simply:

$$\begin{aligned}
V(g_{\text{out}})_{ij} = &= \tau_{ij}\sqrt{V}\xi_{TT} - \frac{\tau_{ij}}{T^2}\sum_{a\leq b}^{T^2}\tau_{ab}\sqrt{V}\xi_{ab} + \frac{\tau_{ij}}{\tau_{iT}}\sqrt{V}\xi_{iT} \\
&\quad - \frac{\tau_{ij}}{\tau_{TT}}\sqrt{V}\xi_{TT} + \delta(j<T)\sqrt{V}\xi_{ij} - \delta(j<T)\frac{\tau_{ij}}{\tau_{iT}}\sqrt{V}\xi_{iT} \\
&= -\frac{\tau_{ij}}{T^2}\sum_{a\leq b}^{T^2}\tau_{ab}\sqrt{V}\xi_{ab} + \sqrt{V}\xi_{iT}\delta(j=T) + \delta(j<T)\sqrt{V}\xi_{ij} \\
&= -\frac{\tau_{ij}}{T^2}\sum_{a\leq b}^{T^2}\tau_{ab}\sqrt{V}\xi_{ab} + \sqrt{V}\xi_{ij}
\end{aligned} \tag{231}$$

which finally gives:

$$\begin{aligned}
\mathbb{E}_{\eta,\xi}V\sum_{i\leq j}^{T}(g_{\text{out}})_{ij}^2 &= \frac{T(T+1)}{2} - \frac{2}{T^2}\sum_{i\leq j}^{T}\sum_{a\leq b}^{T}\tau_{ij}\tau_{ab}\mathbb{E}\xi_{ij}\xi_{ab} \\
&\quad + \frac{1}{T^4}\sum_{i\leq j}^{T}\sum_{a\leq b}^{T}\sum_{c\leq d}^{T}\tau_{ij}^2\tau_{ab}\tau_{cd}\mathbb{E}\xi_{ab}\xi_{cd} \\
&= \frac{T(T+1)}{2} - \frac{2}{T^2}\sum_{i\leq j}^{T}\tau_{ij}^2 + \frac{1}{T^4}\sum_{i\leq j}^{T}\sum_{a\leq b}^{T}\tau_{ij}^2\tau_{ab}^2 \\
&= \frac{T^2+T-2}{2}
\end{aligned} \tag{232}$$

Hence, we can finally conclude that the output channel state equation we obtain for a self-attention layer with an arbitrary number of tokens is:

$$\hat{q} = 4\alpha\mathbb{E}_{\eta,\xi}\sum_{i\leq j}^{T}(g_{\text{out}})_{ij}^2 = \frac{4\alpha(T^2+T-2)}{2V} = \frac{\alpha(T^2+T-2)}{Q-q} \tag{233}$$

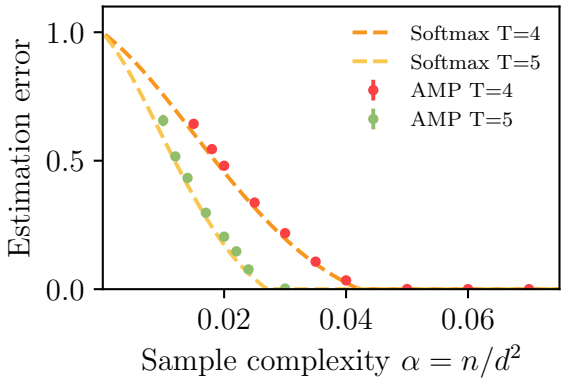

Figure 4: Comparison between the fixed points solutions of the state equations for a softmax output channel in Eq. (26) and Eq. (233) for $T = 4, 5$ tokens. We compare the theoretical solution with their corresponding AMP algorithm run over 16 different realizations and with $d = 120$. The error bars in the AMP dots are computed with respect to the mean value.

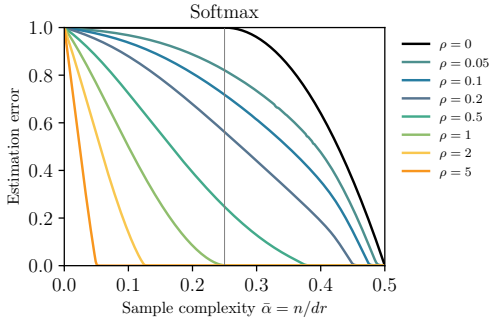

Figure 5: Low width limit of the self-attention model for $L = 1$ layer and $T = 2$ tokens in Eq. (27). We rescale the sample ratio as $\bar{\alpha} = n/dr$ and we plot several values of the width ratio $\rho = r/d$. We correctly predict the weak recovery threshold in Eq. (197).

which is the result presented in the main text in Eq. (32). We highlight this final result holds for any value of the softmax inverse temperature $0 < \beta < +\infty$. In Fig. 4 we show for completeness the state equations for the softmax output channel described by (216) for $T = 4, 5$ tokens and its corresponding AMP run for 16 different realizations and with $d = 120$. The error bars in the AMP dots are computed with respect to the mean value. We find a good agreement also in this case. In Fig. 5 we plot the low width behavior of the self-attention model for $L = 1$ layer and $T = 2$ tokens in Eq. (27), for which we recover the simple output channel state equation in Eq. (233), thus giving:

$$F(Q - q, q) = \frac{T^2 + T - 2}{2(Q - q)} \tag{234}$$

Regarding the large width result in Eq. (198) in this softmax output channel case, we get the equation:

$$\hat{q} = \frac{\alpha(T^2 - T + 2)}{Q - q} = \alpha(T^2 - T + 2)(1 + \hat{q}) \tag{235}$$

so we get

$$\hat{q} = \frac{\alpha(T^2 - T + 2)}{1 - \alpha(T^2 - T + 2)} \tag{236}$$

which gives the large ρ result:

$$\text{MMSE} = \frac{1}{1 + \hat{q}} = 1 - \alpha(T^2 - T + 2)\,. \tag{237}$$

D.6 Hardmax output channel for 2 tokens

We now discuss the hardmax output channel case, in the special case of $T = 2$ tokens. Following Eq. (28) in the main text, we need to compute the quantity:

$$\mathcal{Z}_{\rm out}(y,\omega,V) = \int \prod_{a\le b} dh_{ab}\, \frac{1}{\sqrt{2\pi V_{ab}}} e^{-\frac{(h_{ab}-\omega_{ab})^2}{2V_{ab}}} \prod_{a\le b} \delta(y_{ab} - \sigma_{\rm hard}(\{\frac{1}{\sqrt{2-\delta_{ab}}} h_{ab}\}_a)_b. \quad (238)$$

with:

$$\sigma_{\rm hard}(z_1 \dots z_T)_i = \delta(i = \arg\max_j x_j) \quad (239)$$

In this setting, in particular when $T = 2$, the output label of e.g. y_{11} becomes

$$y_{11} = \Theta(h_{11} - h_{12}) \quad (240)$$

and similarly for the other labels. Here $\Theta(u)$ is the Heaviside function:

$$\Theta(u) = \begin{cases} 1, & u > 0, \\ 0, & u < 0. \end{cases} \quad (241)$$

For the computation of the quantity in Eq. (238), it is convenient to make a change of variables by introducing the differences:

$$u = h_{11} - \frac{h_{12}}{\sqrt{2}}, \quad v = h_{22} - \frac{h_{12}}{\sqrt{2}}. \quad (242)$$

We can thus rewrite in the case of $T = 2$ tokens:

$$\begin{aligned} I_{\rm out}(\eta, y) = \int dh_{12}\, \mathcal{N}(h_{12}; \omega_{12}, V_{12}) \left\{ \int_{u\in\mathcal{R}(y_{11})} du\, \mathcal{N}(u + \frac{h_{12}}{\sqrt{2}}; \omega_{11}, V_{11}) \right\} \\ \times \left\{ \int_{v\in\mathcal{R}(y_{22})} dv\, \mathcal{N}(v + \frac{h_{12}}{\sqrt{2}}; \omega_{22}, V_{22}) \right\}, \end{aligned} \quad (243)$$

where the integration ranges are defined by the hard–threshold:

$$\mathcal{R}(y_{11}) = \begin{cases} \{u > 0\}, & \text{if } y_{11} = 1, \\ \{u < 0\}, & \text{if } y_{11} = 0, \end{cases} \qquad \mathcal{R}(y_{22}) = \begin{cases} \{v > 0\}, & \text{if } y_{22} = 1, \\ \{v < 0\}, & \text{if } y_{22} = 0. \end{cases}$$

Now, by shifting the Gaussian factors we have

$$\mathcal{N}(u + h_{12}; \omega_{11}, V_{11}) = \mathcal{N}(u;\, \omega_{11} - \frac{h_{12}}{\sqrt{2}}, V_{11}), \quad (244)$$

and similarly for the v–integral. Thus, the expression becomes

$$\mathcal{Z}_{\rm out}(y,\omega,V) = \int dh_{12}\, \mathcal{N}(h_{12}; \omega_{12}, V_{12})\; F_{11}(\frac{h_{12}}{\sqrt{2}}; \omega)\; F_{22}(\frac{h_{12}}{\sqrt{2}}; \omega)\,, \quad (245)$$

with

$$F_{11}(\frac{h_{12}}{\sqrt{2}}; \omega) = \int_{u\in\mathcal{R}(y_{11})} du\, \mathcal{N}(u;\, \omega_{11} - \frac{h_{12}}{\sqrt{2}}, V_{11}) = \Phi\Big(s_{11} \frac{\omega_{11} - \frac{h_{12}}{\sqrt{2}}}{\sqrt{V_{11}}}\Big), \quad (246)$$

$$F_{22}(\frac{h_{12}}{\sqrt{2}}; \omega) = \int_{v\in\mathcal{R}(y_{22})} dv\, \mathcal{N}(v;\, \omega_{22} - \frac{h_{12}}{\sqrt{2}}, V_{22}) = \Phi\Big(s_{22} \frac{\omega_{22} - \frac{h_{12}}{\sqrt{2}}}{\sqrt{V_{22}}}\Big), \quad (247)$$

where $\Phi(z)$ is the standard Gaussian CDF and

$$s_{11} = 2y_{11} - 1 = \begin{cases} +1, & y_{11} = 1, \\ -1, & y_{11} = 0, \end{cases} \qquad s_{22} = 2y_{22} - 1 = \begin{cases} +1, & y_{22} = 1, \\ -1, & y_{22} = 0. \end{cases}$$

Thus, in the hard–threshold limit the output channel integral is given by:

$$\mathcal{Z}_{\rm out}(y,\omega,V) = \int_{-\infty}^{+\infty} dh_{12}\, \mathcal{N}(h_{12}; \omega_{12}, V_{12})\; \Phi\Big(s_{11} \frac{\omega_{11} - \frac{h_{12}}{\sqrt{2}}}{\sqrt{V_{11}}}\Big)\, \Phi\Big(s_{22} \frac{\omega_{22} - \frac{h_{12}}{\sqrt{2}}}{\sqrt{V_{22}}}\Big) \quad (248)$$

We can further manipulate this expression.

Writing $h_{12} = \omega_{12} + \sqrt{V_{12}}\, Z$ with $Z \sim \mathcal{N}(0,1)$; then, using independence,

$$\mathcal{Z}_{\rm out}(y,\omega,V) = \mathbb{E}_Z\Big[\Phi\big(u_1 - \lambda_1 Z\big)\,\Phi\big(u_2 - \lambda_2 Z\big)\Big], \tag{249}$$

where

$$u_1 = s_{11}\,\frac{\sqrt{2}\omega_{11} - \omega_{12}}{\sqrt{2V_{11}}}, \quad u_2 = s_{22}\,\frac{\sqrt{2}\omega_{22} - \omega_{12}}{\sqrt{2V_{22}}}, \tag{250}$$

$$\lambda_1 = s_{11}\sqrt{\frac{V_{12}}{2V_{11}}}, \qquad \lambda_2 = s_{22}\sqrt{\frac{V_{12}}{2V_{22}}}. \tag{251}$$

A classical identity for jointly Gaussian variables gives

$$\mathbb{E}_Z\big[\Phi(a+bZ)\,\Phi(c+dZ)\big] = \Phi_2\Big(\tfrac{a}{\sqrt{1+b^2}},\, \tfrac{c}{\sqrt{1+d^2}};\, \tfrac{bd}{\sqrt{(1+b^2)(1+d^2)}}\Big). \tag{252}$$

Where Φ_2 is the cdf of the bivariate normal density defined in Appendix (A). Applying this relation to our model yields:

$$\mathcal{Z}_{\rm out}(y,\omega,V) = \Phi_2\Big(\kappa_1,\, \kappa_2;\, c\Big) \tag{253}$$

with the compact parameters

$$\kappa_1 = s_{11}\,\frac{\sqrt{2}\omega_{11} - \omega_{12}}{\sqrt{2V_{11} + V_{12}}}, \qquad \kappa_2 = s_{22}\,\frac{\sqrt{2}\omega_{22} - \omega_{12}}{\sqrt{2V_{22} + V_{12}}}, \tag{254}$$

$$c = s_{11}s_{22}\,\frac{V_{12}}{\sqrt{(2V_{11}+V_{12})(2V_{22}+V_{12})}} = s_{11}s_{22}\,\frac{1}{3} \quad (V_{11} = V_{22} = V_{12}). \tag{255}$$

We can hence compute denoising function:

$$(g_{\rm out}(y,\omega,V))_{ab} = \frac{\partial}{\partial \omega_{ab}} \ln \mathcal{Z}_{\rm out}(y,\omega,V). \tag{256}$$

Because V_{ab} is ω-independent, the chain rule gives

$$\frac{\partial}{\partial \omega_{11}}\Phi_2(\kappa_1,\kappa_2;\rho_{12}) = \frac{\partial \kappa_1}{\partial \omega_{11}}\,\phi_2(\kappa_1,\kappa_2;\rho_{12}), \quad \frac{\partial \kappa_1}{\partial \omega_{11}} = \frac{\sqrt{2}s_{11}}{\sqrt{2V_{11}+V_{12}}}. \tag{257}$$

The four independent derivatives are therefore

$$g_{\rm out}(y,\omega,V)_{11} = \frac{\sqrt{2}s_{11}}{\sqrt{2V_{11}+V_{12}}}\,\frac{\phi_2(\kappa_1,\kappa_2;\rho_{12})}{\Phi_2(\kappa_1,\kappa_2;\rho_{12})} \tag{258}$$

$$g_{\rm out}(y,\omega,V)_{22} = \frac{\sqrt{2}s_{22}}{\sqrt{2V_{22}+V_{12}}}\,\frac{\phi_2(\kappa_2,\kappa_1;\rho_{12})}{\Phi_2(\kappa_1,\kappa_2;\rho_{12})} \tag{259}$$

$$g_{\rm out}(y,\omega,V)_{12} = -\Big(\frac{s_{11}}{\sqrt{2V_{11}+V_{12}}} + \frac{s_{22}}{\sqrt{2V_{22}+V_{12}}}\Big)\,\frac{\phi_2(\kappa_1,\kappa_2;\rho_{12})}{\Phi_2(\kappa_1,\kappa_2;\rho_{12})} \tag{260}$$

This expression can be compactly rewritten as:

$$g_{\rm out}(y,\omega,V)_{ab} = \frac{1}{\sqrt{6(Q-q)}}\,\frac{\phi(k_1,k_2,c)}{\Phi(k_1,k_2,c)}\begin{pmatrix}\sqrt{2}s_1 & -(s_1+s_2)\\ -(s_1+s_2) & \sqrt{2}s_2\end{pmatrix}_{ab}, \tag{261}$$

where $\phi(k_1,k_2,c)$ is the p.d.f. of a bi-variate Gaussian with zero mean, variances $1/(1-c^2)$ and covariance $c/(1-c^2)$, and $\Phi(k_1,k_2,c)$ is its c.d.f (see Appendix A). Moreover, $s_a = 2y_{aa} - 1$, $k_a = s_a(\sqrt{2}\omega_{aa} - \omega_{12})/\sqrt{6(Q-q)}$, $c = s_1 s_2/3$ and $\omega_{ab} = \sqrt{2q}\,\eta_{ab}$. This is precisely the result shown in the main text in Eq. (29).

We finally compute the state equation corresponding to the output channel, namely:

$$\hat{q} = 4\alpha \mathbb{E}_{\eta,y} \sum_{a\le b} g_{\rm out}(y,\omega,V)^2_{ab} \tag{262}$$

where $\eta_{ab} \sim \mathcal{N}(0,1)$ for $a \le b = 1,\dots,T$ and $y \sim \mathcal{Z}_{\rm out}(y,\omega,V)$.

D.7 Generalization error and sequence-to-sequence version of the model

In this section we draw some consideration on the generalization error presented in Eq. (10), in the setting of a self-attention layer as in (3) and its sequence-to-sequence version as in (54).

In the main text, we showed the expression of the Bayes-optimal estimation error. In the case of one layer of self-attention this reads:

$$E_{est} = \frac{1}{d}\|S^* - \hat{S}\|_F^2 = Q - q \tag{263}$$

Regarding the generalization error, we instead aim to compute and plot the different quantity shown in Eq. (10), namely:

$$\mathcal{E}_{\text{gen}}(\hat{y}) = \mathbb{E}_{\mathcal{D},S^*}\mathbb{E}_{y_{\text{new}},\boldsymbol{x}_{\text{new}}}||\hat{y}(\boldsymbol{x}_{\text{new}},\mathcal{D}) - y_{\text{new}}||_F^2\,, \tag{264}$$

with:

$$\hat{y}_{\mathcal{D}}^{\text{BO}}\left(\mathbf{x}_{\text{test}}\right) := \mathbb{E}\left[y_{\text{test}} \mid \mathbf{x}_{\text{test}}, \mathcal{D}\right] = \int \mathbb{E}_{\mathbf{z}}\left[f_{\mathbf{S}}\left(\mathbf{x}_{\text{test}}\right)\right] \mathbb{P}(\mathbf{S} \mid \mathcal{D}) \mathrm{d}\mathbf{S}$$

Recalling the fact that, for one layer of self-attention, we simply have the relation $y = \sigma_\beta(h) = \sigma_\beta(\{h_{ab}/\sqrt{2-\delta_{ab}}\}_{ab})$, we can introduce the change of variables $h_{ab} = \frac{x_a^\top S x_b - \delta_{ab}\,\mathrm{Tr}\,S}{\sqrt{d}}$ and get the expression:

$$\mathcal{E}_{\text{gen}} = \sum_{a,b}\mathbb{E}_{x_{ab}}\int dh_{ab}d\hat{h}_{ab}\|\sigma(\frac{h_{ab}}{\sqrt{2-\delta_{ab}}}) - \sigma(\frac{\hat{h}_{ab}}{\sqrt{2-\delta_{ab}}})\|^2\delta(h_{ab} - \frac{x_a^\top S x_b - \delta_{ab}\,\mathrm{Tr}\,S}{\sqrt{d}})\delta(\hat{h}_{ab} - \frac{x_a^\top \hat{S} x_b - \delta_{ab}\,\mathrm{Tr}\,\hat{S}}{\sqrt{d}}) \tag{265}$$

We now exploit the fact that, as we know, the preactivations concentrate to:

$$\mathbb{E}_{x_{ab}}\delta(h_{ab} - \frac{x_a^\top S x_b - \delta_{ab}\,\mathrm{Tr}\,S}{\sqrt{d}})\delta(\hat{h}_{ab} - \frac{x_a^\top \hat{S} x_b - \delta_{ab}\,\mathrm{Tr}\,\hat{S}}{\sqrt{d}}) = \mathcal{N}\Big(\begin{pmatrix}h_{ab}\\ \hat{h}_{ab}\end{pmatrix}, \begin{pmatrix}0\\0\end{pmatrix}, \begin{pmatrix}q & q\\ q & Q^*\end{pmatrix}2\Big) = P(h_{ab},\hat{h}_{ab}) \tag{266}$$

Then, the overall generalization error is given by

$$\mathcal{E}_{\text{gen}} = \sum_{a,b=1}^{T}\mathbb{E}_{(h_{ab},\hat{h}_{ab})\sim P(h_{ab},\hat{h}_{ab})}\Big[\sigma(\{\frac{h_{ab}}{\sqrt{2-\delta_{ab}}}\}_{ab}) - \sigma(\{\frac{\hat{h}_{ab}}{\sqrt{2-\delta_{ab}}}\}_{ab})\Big]^2 \tag{267}$$

which exactly matches the result presented in the main text in Eq. (19), where the extension to the multi-layer setting is trivial.

Now we slightly modify our model of a self-attention layer by considering its sequence-to-sequence (seq2seq) version $y = \sigma_\beta(\{\frac{h_{ab}}{\sqrt{2-\delta_{ab}}}\}_{ab})x \in \mathbb{R}^{T\times d}$. In particular we aim to compute and plot the generalization error of Eq. (10) in this new setting.

To do so, we define $y = Ax$ and $\hat{y} = \hat{A}x$ with $A = \sigma_\beta(h)$ and $\hat{A} = \sigma_\beta(\hat{h})$ where we leave the factor $\sqrt{2-\delta_{ab}}$ implicit. We exploit the concentration of our input data, in order to compute the Frobenius norm of $y - \hat{y} = (A - \hat{A})x$. Recalling the fact that the input data are iid with $x_{ai}^\mu \sim \mathcal{N}(0,1)$, we use the fact that

$$\sum_{i=1}^{d} x_{t'i}x_{t''i} \approx \delta_{tt'} \tag{268}$$

with high probability when d is large. Hence:

$$||(A-\hat{A})x||_F^2 = \sum_{t,i}[\sum_{t'}(A_{tt'} - \hat{A}_{tt'})x_{t'i}]^2 = \sum_{t,i}\sum_{t',t''}(A_{tt'} - \hat{A}_{tt})(A_{tt''} - \hat{A}_{tt''})x_{t',i}x_{t'',i} \tag{269}$$

but using the concentration property of x we finally get:

$$||(A-\hat{A})x||_F^2 = \sum_{t,i}\sum_{t',t''}(A_{tt'} - \hat{A}_{tt'})(A_{tt''} - \hat{A}_{tt''})x_{t',i}x_{t'',i} = \sum_{t,t'}(A_{tt''} - \hat{A}_{tt'})^2 = ||A-\hat{A}||_F^2 \tag{270}$$

We hence have shown that in the case of $L = 1$ layer, the sequence-to-sequence version of the model shows the same identical state evolution with respect to a single self-attention layer.

D.8 Details on the numerical implementation

The code used to produce all the figures and the experiments is available at https://github.com/SPOC-group/ExtensiveRankAttention. Our gradient descent experiments are done in PyTorch 1.12.1 by minimizing the following loss using Adam

$$\mathcal{L}(W) = \sum_{\mu=1}^{n} \sum_{a,b=1}^{T} \left(y_{ab}^{\mu} - \sigma_\beta \left(\frac{\boldsymbol{x}_a^{\mu\top} W W^\top \boldsymbol{x}_b^{\mu} - \delta_{ab} \operatorname{Tr} W W^\top}{\sqrt{r}\, d} \right) \right)^2 , \tag{271}$$

In our implementation we sample both the input data and the weights of the target as standard Gaussian. Notice that we appropriately adjusted the loss to be consistent with the main test. We choose a learning rate 0.1 and keep the other hyperparameters at their default parameters and initializing the weights as a standard Gaussian.

When running the averaged version of the algorithm we run the optimization procedure 32 times for a fixed experiment, and average the matrix $S = WW^\top/\sqrt{rd}$ at the end of training.

Regarding the state equations in the two $L = 1$ cases of softmax and hardmax output channel: in the former case, we simply find the fixed points iterations of the state equations in Eq. (26) and Eq. (233). In the latter case, finally, we compute the expectation in Eq. (262) with Monte-Carlo over $n_{\text{samples}} = 20000$ samples. In particular, to allow for more stable results, we iterate the state equations for $T = 150$ iterations and we compute the mean overlap over the last 30 iterations of the state equations.