Bayes optimal learning of attention-indexed models

Fabrizio Boncoraglio FABRIZIO.BONCORAGLIO@EPFL.CH
Emanuele Troiani EMANUELE.TROIANI@EPFL.CH
Vittorio Erba VITTORIO.ERBA@EPFL.CH
Lenka Zdeborová LENKA.ZDEBOROVA@EPFL.CH
Statistical Physics of Computation Laboratory,
École Polytechnique Fédérale de Lausanne (EPFL), CH-1015 Lausanne

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 modern attention architectures.

1. Introduction

The Transformer architecture [43] has transformed machine learning, achieving state-of-the-art results in natural language processing [9, 23], computer vision [16], 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* [15] 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.

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 [1–3, 5–7, 10, 14, 33, 44]. In statistical physics, similar models appear as teacher-student perceptrons [20, 21, 37, 45] or committee machines [4, 17]. 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 [11–13]. 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 $\boldsymbol{x}_a \in \mathbb{R}^d$ for $a = 1, \dots, T$. We define L *attention indices* $h^{(\ell)} \in \mathbb{R}^{T\times T}$ with components $h^{(\ell)}_{ab}$. The labels y for each input $X \in \mathbb{R}^{T\times d}$ are generated via a general output function $g : \mathbb{R}^{L\times T\times T} \to \mathbb{R}^{T\times T}$

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

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 [43], 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

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

2. Setting

We consider a dataset $\mathcal{D} = \{y^\mu\,, \boldsymbol{x}_a^\mu\}$ of n samples. Each sample consists of the embeddings of T tokens $\boldsymbol{x}_a^\mu \in \mathbb{R}^d$, taken as standard Gaussian $\boldsymbol{x}_a^\mu \sim \mathcal{N}(0, \mathbb{I}_d)$ 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 [46, 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 $P_S(S) = P_S(O^\top S O)$ 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 $\kappa_1,\ \kappa_2 = \mathcal{O}(1)$. 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 [40].

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 the weights S^* (estimation task), i.e. find the estimator $\hat{S}(\mathcal{D})$ that minimizes

$$\mathcal{E}_{\text{est}}(\hat{S}) = \mathbb{E}_{\mathcal{D},S^*} \frac{1}{d} \sum_{\ell=1}^{L} ||\hat{S}(\mathcal{D})_\ell - S^*_\ell||_F^2\,, \tag{3}$$

We will call the error achieved by the optimal estimator the BO estimation error.

The BO estimator 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

$$P(S_1, ..., S_L|\mathcal{D}) = \frac{1}{\mathcal{Z}(\mathcal{D})} \prod_{\ell=1}^{L} P_S(S_\ell) \prod_{\mu=1}^{n} \delta\left(y^\mu - g\left(h^{(1)}(S_1, \boldsymbol{x}^\mu), ..., h^{(L)}(S_L, \boldsymbol{x}^\mu)\right)\right) , \quad (4)$$

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.

3. Results for single-layer attention

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

$$y_{ab} = \sigma_\beta\left(\frac{\boldsymbol{x}_a S\, \boldsymbol{x}_b^\top - \delta_{ab}}{\sqrt{d}}\right) = \sigma_\beta\left(\frac{\frac{1}{\sqrt{rd}}\boldsymbol{x}_a W W^\top \boldsymbol{x}_b^\top - \delta_{ab}}{\sqrt{d}}\right) \quad (5)$$

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 σ_{hard} and Softmax σ_{soft} (x), both applied row-wise in (5):

$$\sigma_{\rm hard}(z_1 \dots z_T)_i = \delta(i = \arg\max_j x_j) , \quad \text{and} \quad \sigma_{soft}(z_1 \dots z_T)_i = \frac{e^{\beta z_i}}{\sum_{j=1}^{T} e^{\beta z_j}} . \quad (6)$$

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 as detailed in the Appendix. We provide an explicit solution in the $T = 2$ case in Appendix D. We plot the estimation error in Figure 1 left, for several values of the attention width ratio ρ, comparing with runs of the associated AMP Algorithm 1 in the Appendix 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. In the limit of small width our results simplify. 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 analytically

$$\bar{\alpha}_{\rm weak}^{hardmax} = \frac{1}{4\mathbb{E}_{y,\omega}\left[\sum_{a\leq b}^{T=2} g_{out}(y(\omega, V), \omega, V)_{ab}^{\otimes 2}\right]_{q=0,Q=1}} \approx 0.563 . \quad (7)$$

Softmax target. We now discuss the target function that uses a softmax non-linearity (6). 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 $\beta \in \mathbb{R}_+$.

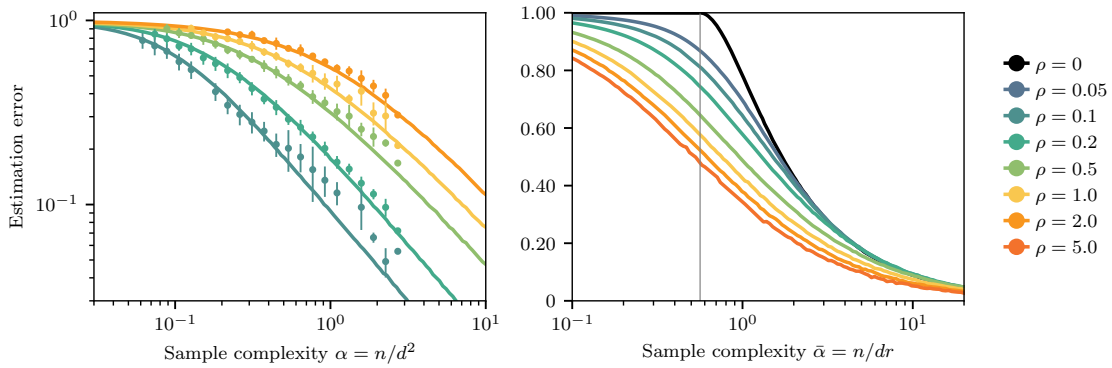

Figure 1: **(Left)** Bayes optimal-error for the single-layer attention-indexed model with $T = 2$ tokens and hardmax activation. We also plot the corresponding AMP algorithm (dots) at $d = 100$, over 16 realizations of the data and teacher weights. Error bars are computed with respect to the mean. **(Right)** Hardmax small width limit. We rescale the sample complexity to $\bar{\alpha} = \alpha/\rho$. The gray vertical line represents the weak recovery threshold of Eq.(7).

Result 1 (Bayes-optimal errors for softmax tied-attention, $T \geq 2$) *Consider the model (5) with softmax activation, $T \geq 2$ and inverse temperature $\beta \in \mathbb{R}_+$. In the high-dimensional limit $d, n \to \infty$ with $\alpha = n/d^2$ finite, the asymptotic BO estimation error is given by:*

$$\textit{MMSE} = \frac{\alpha(T^2+T-2)}{\hat{q}}\,, \qquad 1-(T^2+T-2)\alpha = \frac{4\pi^2}{3\hat{q}}\int \mu_{1/\hat{q}}(x)^3 dx \tag{8}$$

with $\mu_{1/\hat{q}} = \mu_S \boxplus \mu_{s.c.,1/\sqrt{\hat{q}}}$ the free convolution of the spectral distribution of the matrix S in Eq.(2) and the semicircle distribution with variance $\Delta = 1/\sqrt{\hat{q}}$.

We plot the BO estimation error given in Result 1 in Figure 2 left. We observe that the BO estimation error vanishes at a finite value of α (the so-called strong recovery threshold). Interestingly, the BO error given in Result (1) is independent of the value of the inverse temperature β and reduces to the case of a single-token model with linear activation [30], 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. D). 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 (1) 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.

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, as the AMP output function g_{out}, given in Appendix D, is indeed very different from the one in [30]. 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

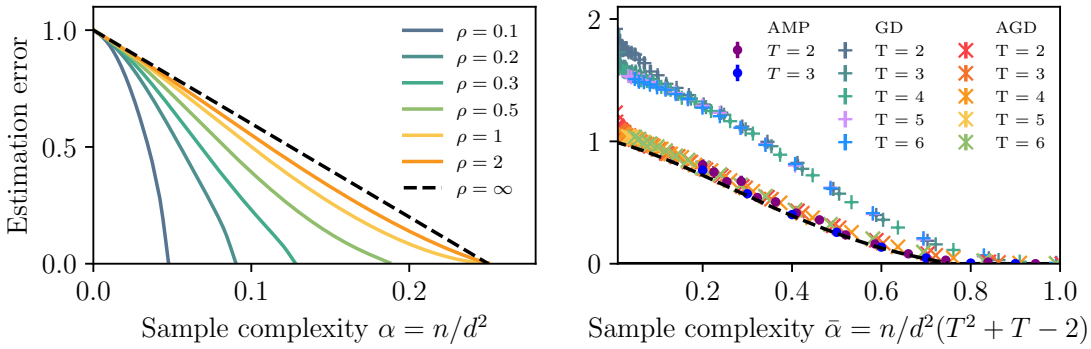

Figure 2: **(Left)** Bayes-optimal estimation error for the softmax tied-attention model (Result 1), eq. (5) for $T = 2$ tokens. **(Right)** We show, in black dashed lines the theoretical prediction of the BO estimation error computed for the rescaled sample complexity $\alpha/(T^2 + T - 2)$ and $\rho = 0.5$. We show the performance of the corresponding AMP algorithm, and 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.

and blue dots, to be compared with the prediction of Result (1) 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 [30] 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. E). In particular, the strong recovery threshold satisfies $\alpha_{recovery}^{softmax} = (1 - (max\{0, 1 - \rho\})^2)/(T^2 + T - 2)$. We remark again that this threshold does not coincide with the naive counting argument, which would give a factor $T(T + 1)$ at denominator instead.

Finally, we consider the performance of gradient descent minimizing the MSE loss with training set generated by Eq. (5) (we optimize using the ADAM optimizer [24]). In line with previous work [18, 30], we also consider the Averaged GD (AGD) estimator given by $\hat{S}_{\text{GD,avg}} = \sum_{m=1}^{M} W_m^{final}(W_m^{final})^\top / M\sqrt{rd}$, 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 [18, 30], 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. This phenomenon, at this point well documented within this class of models, is still not understood.

References

[1] 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.

[2] 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.

[3] 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.

[4] 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.

[5] 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.

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

[7] Alberto Bietti, Joan Bruna, and Loucas Pillaud-Vivien. On learning gaussian multi-index models with gradient flow. *arXiv preprint arXiv:2310.19793*, 2023.

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

[9] 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.

[10] 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.

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

[12] 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.

[13] 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. ISSN 1742-5468. doi: 10.1088/1742-5468/ada696. URL https://dx.doi.org/10.1088/1742-5468/ada696. Publisher: IOP Publishing.

[14] 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.

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

[16] 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. URL https://arxiv.org/abs/2010.11929.

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

[18] 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. *arXiv preprint arXiv:2410.18858*, 2024.

[19] 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.

[20] 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.

[21] 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.

[22] 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.

[23] Jacob Devlin Ming-Wei Chang Kenton and Lee Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding, 2019.

[24] 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. URL http://arxiv.org/abs/1412.6980.

[25] 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.

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

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

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

[29] 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.

[30] 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. URL https://proceedings.neurips.cc/paper_files/paper/2024/hash/953e742190ca02fc8f9f710052f2fead-Abstract-Conference.html.

[31] 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.

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

[33] 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.

[34] 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.

[35] 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.

[36] 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.

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

[38] 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.

[39] 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.

[40] Emanuele Troiani, Vittorio Erba, Florent Krzakala, Antoine Maillard, and Lenka Zdeborova. 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.

[41] Emanuele Troiani, Yatin Dandi, Leonardo Defilippis, Lenka Zdeborová, Bruno Loureiro, and Florent Krzakala. Fundamental limits of weak learnability in high-dimensional multi-index models. *arXiv preprint arXiv:2405.15480*, 2024.

[42] Emanuele Troiani, Hugo Cui, Yatin Dandi, Florent Krzakala, and Lenka Zdeborová. Fundamental limits of learning in sequence multi-index models and deep attention networks: High-dimensional asymptotics and sharp thresholds, 2025. URL https://arxiv.org/abs/2502.00901.

[43] A Vaswani. Attention is all you need. *Advances in Neural Information Processing Systems*, 2017.

[44] 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.

[45] 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.

[46] Yizhou Xu, Antoine Maillard, Lenka Zdeborová, and Florent Krzakala. Fundamental Limits of Matrix Sensing: Exact Asymptotics, Universality, and Applications, March 2025. URL http://arxiv.org/abs/2503.14121. arXiv:2503.14121 [stat].

[47] 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.

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

Appendix A. 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 [17, 20, 37], and more recently to single-index and multi-index models [1–3, 5–7, 10, 14, 33, 44].

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 [31, 34, 38, 39, 47], or study training dynamics in the linear, kernel, or random feature regimes [19, 22, 25]. Others use infinite-width approximations without access to generalization error [8, 36]. 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 [11, 12, 32, 42], 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 [4, 41], and their recent generalizations to sequence learning with multiple low-width self-attention layers [42]. 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 [26–28] and its connection to two-layer neural networks with quadratic activations and extensive width [30, 46]. 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 [18], and we leave its treatment for future work.

Appendix B. Notations and model description and known theory results

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 $u, v = 0 \dots n$ as the replica indices from 0 to n.

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:

$$y = g\left(\{h^{(\ell)}\}_{\ell=1}^{L}\right) \tag{9}$$

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

$$h_{ab}^{(\ell)} \equiv \frac{\boldsymbol{x}_a S_\ell\, \boldsymbol{x}_b^\top - \delta_{ab} \operatorname{Tr} S_\ell}{\sqrt{d}} \tag{10}$$

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 Marcenko-Pastur laws in the contex of random matrix theory. In particular

$$\sigma_{\mathrm{sc},\,\Delta} = \frac{\sqrt{4\Delta - x^2}}{2\pi\Delta}\mathbb{I}\{|x| \le 2\sqrt{\Delta}\}\,, \qquad \mu_{\mathrm{MP},\rho}(x) = \begin{cases} (1-\rho)\delta(x) + \rho\frac{\sqrt{(\lambda_+ - x)(x-\lambda_-)}}{2\pi x}\,, & \text{if } \rho \le 1 \\ \rho\frac{\sqrt{(\lambda_+ - x)(x-\lambda_-)}}{2\pi x}\,, & \text{if } \rho > 1 \end{cases} \tag{11}$$

Finally, we recall the following following definitions.

- Standard normal pdf and cdf

$$\phi(z) = \frac{e^{-z^2/2}}{\sqrt{2\pi}}, \qquad \Phi(z) = \int_{-\infty}^{z} \phi(t)\, dt = \tfrac{1}{2}\big(1 + \operatorname{erf}(z/\sqrt{2})\big) \tag{12}$$

- Bivariate normal density and cdf with correlation c

$$\phi_2(u,v;c) = \frac{\exp\Big[-\frac{u^2 - 2cuv + v^2}{2(1-c^2)}\Big]}{2\pi\sqrt{1-c^2}}, \qquad \Phi_2(u,v;c) = \int_{-\infty}^{u}\int_{-\infty}^{v} \phi_2(t_1,t_2;c)\, dt_2\, dt_1. \tag{13}$$

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:

$$\sigma_\beta(h_{ab}) = \text{Softmax}(\beta h_{ab}) = \frac{\exp(\beta h_{ab})}{\sum_b \exp(\beta h_{ab})} \tag{14}$$

Appendix C. 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 $\boldsymbol{x}_a \in \mathbb{R}^d$, 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:

$$X_\ell \;=\; \Big[\, c\,\mathbb{I}_T \;+\; \sigma_\beta\Big(\tfrac{1}{\sqrt{d}}\, X_{\ell-1} S_\ell X_{l-1}^\top\Big)\Big]\, X_{\ell-1}, \qquad \ell = 1,\dots,L, \tag{15}$$

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 $\sigma(\cdot)$.
Define the *pre–activations*

$$h_{ab}^{(\ell)} \;:=\; \frac{1}{\sqrt{d}}\, \boldsymbol{x}_a S_\ell \boldsymbol{x}_b^\top, \qquad \ell = 1,\dots,L,\; a,b \in [T], \tag{16}$$

and the sequence of token–space operators

$$B_0 := \mathbb{I}_T, \qquad B_\ell := \Big[c\,\mathbb{I}_T + \sigma_\beta\big(\, B_{\ell-1} h^{(\ell)} B_{\ell-1}^\top\big)\Big] B_{\ell-1}, \quad \ell = 1,\dots,L, \tag{17}$$

One verifies inductively that

$$X_\ell = B_\ell\, X_0, \qquad \ell = 0,\dots,L, \tag{18}$$

so that every hidden representation depends on the data *only* through the collection $\{h^{(1)},\dots,h^{(\ell)}\}$. In particular the *deep-attention output*

$$y \;=\; \sigma_\beta\Big(\tfrac{1}{\sqrt{d}}\, X_{L-1} S_L X_{L-1}^\top\Big) \;=\; g_{\text{deep}}\Big(h^{(1)},\dots,h^{(L)}\Big) \;\in \mathbb{R}^{T\times T}, \tag{19}$$

with[2] $g_{\text{deep}}(h^{(1)},\dots,h^{(L)}) := \sigma_\beta\big(\, B_{L-1}(h^{(1:L-1)})\, h^{(L)}\, B_{L-1}^\top(h^{(1:L-1)})\big)$. Equation (19) 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 $\{\boldsymbol{x}_a S_\ell \boldsymbol{x}_b^\top\}_{\ell,a,b}$.

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

$$X_L \;=\; \sigma_\beta\Big(\tfrac{1}{\sqrt{d}}\, X_{L-1} S_L X_{L-1}^\top\Big)\, X_{L-1}, \tag{20}$$

1. For simplicity we restrict to the single–head, tied setting; extending to multi–head merely introduces an additional block index.
2. The explicit form of g_{deep} is obtained by inserting (18) with $l = L-1$.

with exactly the same algebra

$$X_L \;=\; g_{\text{seq}}\Big(h^{(1)},\dots,h^{(L)}\Big)\,X_0, \qquad g_{\text{seq}}(h^{(1:L)}) := \sigma_\beta\big(\,B_{L-1}h^{(L)}B_{L-1}^\top\big)\,B_{L-1}. \tag{21}$$

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:

$$X_1 = g_{\text{seq}}\Big(\{h^{(1)}\}_{a\le b}^{T}\Big)\,X_0, \quad g_{\text{seq}}(h^{(1)}) = \sigma_\beta\big(\,B_0h^{(1)}B_0^\top\big)\,B_0 = \sigma_\beta(h^{(1)}) = g_{deep}(h^{(1)}) \tag{22}$$

From this paragraph we can hence conclude that, as shown in equations (19) and (21), 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. 2–3 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).

Appendix D. Bayes optimal analysis of Attention-Indexed Models (AIM)

We study a model described by the general setting:

$$y_\mu \sim P_{out}\Big(\frac{\boldsymbol{x}_a^\mu S_\ell\;\boldsymbol{x}_b^{\mu\,\top} - \delta_{ab}\,\mathrm{Tr}\,S_\ell}{\sqrt{d}}\Big)_{\ell=1,\dots,L}^{a,b=1,\dots,T} \tag{23}$$

with $\boldsymbol{x}_a$ 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. 23 as y_μ outputs generated by a model of attention from data X that are processed in a bilinear way through :

$$y = g\left(\{\frac{\boldsymbol{x}_a^\mu S_\ell\;\boldsymbol{x}_b^{\mu\,\top} - \delta_{ab}\,\mathrm{Tr}\,S_\ell}{\sqrt{d}}\}_{\ell=1}^{L}\right). \tag{24}$$

or

$$y_\mu = g_{\text{deep}}(h^{(1)},\dots,h^{(L)}) = B_c^L\Big(\{\frac{\boldsymbol{x}_a^\mu S_\ell\;\boldsymbol{x}_b^{\mu\,\top} - \delta_{ab}\,\mathrm{Tr}\,S_\ell}{\sqrt{d}}\}_{a,b=1\dots T}^{\ell=1\dots L}\Big) \in \mathbb{R}^{T\times T} \tag{25}$$

and

$$P_{out}\Big(\frac{\boldsymbol{x}_a^\mu S_\ell\;\boldsymbol{x}_b^{\mu\,\top} - \delta_{ab}\,\mathrm{Tr}\,S_\ell}{\sqrt{d}}\Big)_{\ell=1,\dots,L}^{a,b=1,\dots,T} = \delta\Big(y - B_c^L(\frac{\boldsymbol{x}_a^\mu S_\ell\;\boldsymbol{x}_b^{\mu\,\top} - \delta_{ab}\,\mathrm{Tr}\,S_\ell}{\sqrt{d}})\Big) \tag{26}$$

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

We define the preactivations

$$h_{ab}^{(\ell)\,\mu} = \frac{\boldsymbol{x}_a^\mu\;S_\ell\;\boldsymbol{x}_b^{\mu\,\top} - \delta_{ab}\,\mathrm{Tr}\,S_\ell}{\sqrt{d}} \tag{27}$$

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

$$Z^\mu_{ij,ab} \equiv (\boldsymbol{x}^\mu_{i,a}\boldsymbol{x}^\mu_{j,b} + \boldsymbol{x}^\mu_{j,a}\boldsymbol{x}^\mu_{i,b} - 2\delta_{ij}\delta_{ab})/\sqrt{2d(1+\delta_{ab})} \in \mathbb{R} \tag{28}$$

The preactivations of the model can thus be expressed as :

$$\sqrt{2-\delta_{ab}}\; h^{(\ell)\,\mu}_{ab} = \mathrm{Tr}(S_\ell Z^\mu_{ab}) \tag{29}$$

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

D.1. Replica analysis of AIM and their state evolution

Starting from the posterior distribution of the model:

$$P(S_1,...,S_L|\mathcal{D}) = \frac{1}{\mathcal{Z}(\mathcal{D})}\prod_{\ell=1}^{L} P_S(S_\ell)\prod_{\mu=1}^{n}\delta\left(y^\mu - g\left(h^{(1)}(S_1,\boldsymbol{x}^\mu),...,h^{(L)}(S_L,\boldsymbol{x}^\mu)\right)\right), \tag{30}$$

the replicated partition function of the model in Eq. (23) is:

$$\left\langle \mathcal{Z}(\mathcal{D})^m \right\rangle = \mathbb{E}_{y,\,X}\int \prod_{\ell=1}^{L}\prod_{u=0}^{m} \mathrm{d}S^u_\ell\; P_0(S^u_\ell)\prod_{\mu=1}^{n}\prod_{a\le b}^{T} P_{\mathrm{out}}\Big(y^\mu \,\Big|\, \Big\{\frac{h^{(\ell),\mu,u}_{ab}}{\sqrt{2-\delta_{ab}}}\Big\}_{ab}\Big)\,\delta\Big(h^{(\ell),\mu,u}_{ab} - \mathrm{Tr}(S^u_\ell Z^\mu_{ab})\Big) \tag{31}$$

where $P_0(S^u_\ell)$ is the rotational invariant prior distribution of each S_ℓ, and $h^{(\ell),\mu,u}_{ab}$ are the replicated preactivations in terms of the symmetrized data as explained in (29). 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:

$$\mathbb{E}_X\delta\Big(h^{(\ell),\mu,u}_{ab} - \mathrm{Tr}(S^u_\ell Z^\mu_{ab})\Big) \quad \mapsto \quad P_h\Big(\{h^{(\ell),\mu,u}_{ab}\}_{\ell,\mu,a,b,u}\Big), \tag{32}$$

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

$$\mathbb{E}\big[h^{(\ell),\mu,u}_{ab} - \mathrm{Tr}(S^u_\ell Z^\mu_{ab})\big] = 0, \quad \mathrm{Cov}_{x^\mu}\left(h^{(\ell)\,u}_{a\le b}, h^{(k)\,v}_{c\le d}\right) = \frac{1}{d}\left[2\,\delta_{ac}\delta_{bd}\right]\mathrm{Tr}\left(S^u_\ell S^v_k\right) \tag{33}$$

We introduce the order parameters measuring the S^u_ℓ–S^v_k overlaps:

$$Q^{uv}_{\ell k} \;:=\; \frac{1}{d}\,\mathrm{Tr}\Big(S^u_\ell\, S^v_k\Big), \quad \text{for } \ell,k = 1,\dots,L,\; u,v = 0,\dots,m. \tag{34}$$

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

$$\prod_{\ell,k=1}^{L}\prod_{u\le v=0}^{m}\delta\Big(d^2\,Q^{uv}_{\ell k} - d\;\mathrm{Tr}\big[S^u_\ell\, S^v_k\big]\Big), \tag{35}$$

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

$$\delta\Big(d^2\, Q^{uv}_{\ell k} - d\mathrm{Tr}[S^u_\ell\, S^v_k]\Big) = \int \mathrm{d}\widehat{Q}^{uv}_{\ell k}\, \exp\Big\{\, i\, \frac{\widehat{Q}^{uv}_{\ell k}}{2}\, \Big(d^2\, Q^{uv}_{\ell k} - d\, \mathrm{Tr}[S^u_\ell\, S^v_k]\Big)\Big\}. \tag{36}$$

Hence the replicated partition function can be schematically written:

$$\begin{aligned}\big\langle \mathcal{Z}(\mathcal{D})^m \big\rangle = & \int \Big(\prod_{u,\ell} \mathrm{d}S^u_\ell\, P_0(S^u_\ell)\Big) \int \big(\prod_{u\le v,\ell,k} \mathrm{d}Q^{uv}_{\ell k}\, \mathrm{d}\hat{Q}^{uv}_{\ell k}\big) \\ & \times \exp\big[\frac{i}{2} \sum_{u\le v,\ell,k} \hat{Q}^{uv}_{\ell k} (d^2\, Q^{uv}_{\ell k})\big] \times \exp\Big[-\frac{i\, d}{2} \sum_{u\le v,\ell,k} \hat{Q}^{uv}_{\ell k}\, \mathrm{Tr}(S^u_\ell S^v_k)\Big] \\ & \times \prod_{\mu=1}^{n} \Big[\int \prod_{u,\ell} \mathrm{d}h^{(\ell),\mu,u}_{ab}\, P_h\big(h^{(\ell),\mu,u}\big) \prod_{u,\ell,a\le b} P_{\mathrm{out}}(y^\mu_{ab} \mid \frac{\{h^{(\ell),\mu,u}_{ab}}{\sqrt{2-\delta_{ab}}}\}_{ab})\Big],\end{aligned} \tag{37}$$

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

$$Q^{uv}_{\ell k} = \begin{cases} Q_{\ell k}, & (u = v), \\ q_{\ell k}, & (u \neq v). \end{cases} \tag{38}$$

and:

$$i\hat{Q}^{uv}_{\ell k} = \begin{cases} \hat{Q}_{\ell k}, & \text{if } u = v \\ -\hat{q}_{\ell k}, & \text{if } u \neq v \end{cases} \tag{39}$$

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

$$i\, d^2 \sum_{\ell,k} \Big[\frac{(m+1)}{2}\, \widehat{Q}_{\ell k}\, Q_{\ell k} - \frac{m(m+1)}{4}\, \widehat{q}_{\ell k}\, q_{\ell k}\Big]. \tag{40}$$

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

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

$$\big\langle \mathcal{Z}(\mathcal{D})^m \big\rangle = \int \prod_{u\le v,\ell,k} dQ^{uv}_{\ell k} d\hat{Q}^{uv}_{\ell k} \exp\left(\frac{i}{2} d^2 \sum_{u\le v,\ell,k} \hat{Q}^{uv}_{\ell k} Q^{uv}_{\ell k}\right) I_{\mathrm{in}} I_{\mathrm{out}} \tag{41}$$

with:

$$d^2 I_{\mathrm{in}}(\hat{q}) = \int \prod_{u,\ell} dS^u_\ell\, P_0(S^u_\ell) \exp\left(-\frac{i\, d}{2} \sum_{u\le v,\ell,k} \hat{Q}^{uv}_{\ell k}\, \mathrm{Tr}\,(S^u_\ell S^v_k)\right), \tag{42}$$

$$I_{\mathrm{out}}(q) = \left[\int dy \int \prod_{u,\ell,a\le b} dh^{(\ell)\,u}_{ab}\, P\left(\{h^{(\ell)\,u}_{ab}\}\right) \prod_{u,\ell} P_{\mathrm{out}}\left(y|\frac{h^{(\ell)\,u}_{ab}}{\sqrt{2-\delta_{ab}}}\right)\right]^n. \tag{43}$$

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

$$\Phi = \lim_{d\to\infty} \frac{1}{d^2} \lim_{n\to\infty} \lim_{m\to 0} \frac{1}{m} \ln\langle Z^m \rangle \,. \tag{44}$$

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

$$\Phi = \text{extr}_{\{q,\hat{q}\}} \left\{ -\frac{\text{Tr}\, q\hat{q}}{4} + I_{\text{in}}(\hat{q}) + \alpha I_{out}(q) \right\} . \tag{45}$$

Thus obtaining the state equations:

$$q = 4\, \partial_{\hat{q}} I_{\text{in}}(\hat{q}) \tag{46}$$

$$\hat{q} = 4\alpha\, \partial_q I_{out}(q) \tag{47}$$

D.2. Prior Term Computation

First we compute under the RS ansatz:

$$-\frac{i\,d}{2} \sum_{u\le v=0}^{m} \hat{Q}^{uv}_{\ell k} \,\text{Tr}\,(S^u_\ell S^v_k) = -\frac{i\,d}{2} \left(\sum_{u=0}^{m} \hat{Q}_{\ell k} \,\text{Tr}\,(S^u_\ell S^u_k) + \sum_{u<v} (-\hat{q}_{\ell k}) \,\text{Tr}\,(S^u_\ell S^v) \right) \tag{48}$$

$$= -\frac{\hat{Q}_{\ell k}\,d}{2} \sum_{u=0}^{m} \text{Tr}\,(S^u_\ell S^u_k) + \frac{\hat{q}_{\ell k}\,d}{2} \sum_{u<v} \text{Tr}\,(S^u_\ell S^v_k) \tag{49}$$

$$= -\frac{d}{2} \left(\hat{Q}_{\ell k} + \frac{\hat{q}_{\ell k}}{2} \right) \sum_{u=0}^{m} \text{Tr}\,(S^u_\ell S^u_k) + \frac{\hat{q}_{\ell k}\,d}{4} \sum_{u,v=0}^{m} \text{Tr}\,(S^u_\ell S^v_k) \tag{50}$$

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:

$$I_{\text{in}}(\hat{q}) = \int \prod_{\ell=1}^{L} \prod_{u=0}^{m} dS^u_\ell \, P_0(S^u_\ell) \, \exp\left\{ -\frac{i\,d}{2} \sum_{\ell,k=1}^{L} \sum_{u\le v=0}^{m} \widehat{Q}^{(u,v)}_{\ell k} \,\text{Tr}\big(S^u_\ell S^v_k\big) \right\} \tag{51}$$

$$= \int d\bar{S} P_0(\bar{S}) \exp\left\{ \sum_{\ell,k}^{L} -\frac{d}{2} \left(\hat{Q}_{\ell k} + \frac{\hat{q}_{\ell k}}{2} \right) \sum_{u=0}^{m} \text{Tr}\,(S^u_\ell S^u_k) + \frac{\hat{q}_{\ell k}\,d}{4} \sum_{u,v=0}^{m} \text{Tr}\,(S^u_\ell S^v_k) \right\} \tag{52}$$

$$= \int d\bar{S} P_0(\bar{S}) \exp\left\{ -\sum_{\ell,k}^{L} \sum_{u} \frac{\hat{q}_{\ell,k}\,d}{4} \,\text{Tr}(S^u_\ell S^u_k) + \sum_{\ell k}^{L} \sum_{u,v} \frac{\hat{q}_{\ell k} d}{4} \,\text{Tr}(S^u_\ell S^v_k) \right\} \tag{53}$$

$$= \int d\bar{S} P_0(\bar{S}) \mathcal{D}(Y) \exp\left\{ -\sum_{\ell,k}^{L} \sum_{u} \frac{\hat{q}_{\ell k}\,d}{4} \,\text{Tr}(S^u_\ell S^v_k) + \sum_{\ell,k}^{L} \sum_{u} \frac{\sqrt{\hat{q}_{\ell k}}\,d}{2} \,\text{Tr}(S^u_k Y_\ell) \right\} \tag{54}$$

$$= \int \mathcal{D}(Y) \left\{ \int d\bar{S} P_0(\bar{S}) \exp\Big\{ -\frac{d}{4} \sum_{\ell,k}^{L} \hat{q}_{\ell k} \,\text{Tr}(S_\ell S_k) + d \sum_{\ell,k}^{L} \frac{\sqrt{\hat{q}_{\ell k}}}{2} \,\text{Tr}(S_k Y_\ell) \right\}^{m+1} \tag{55}$$

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.(73) we used the identity:

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

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

$$\begin{aligned} I_{\text{in}}(\hat{q}) &= \lim_{d\to\infty} \frac{1}{d^2} \int DY_1 \dots DY_L \, \mathcal{Z}_{\text{in}}(Y_1,\dots,Y_L;\hat{q}) \log \mathcal{Z}_{\text{in}}(Y_1,\dots,Y_L;\hat{q}) \\ \mathcal{Z}_{\text{in}}(\{Y_\ell\}_{\ell=1}^L;\hat{q}) &= \int \left[\prod_{\ell=1}^{L} \mathrm{d}S_\ell \, P_S(S_\ell) \right] \\ &\quad \times \exp\Big[-\frac{d}{4} \sum_{\ell,k=1}^{L} \hat{q}_{\ell k} \,\mathrm{Tr}(S_\ell S_k) + \frac{d}{2} \sum_{\ell,k=1}^{L} \sqrt{\hat{q}}_{\ell k} \,\mathrm{Tr}(Y_\ell S_k) \Big]. \end{aligned} \tag{56}$$

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 :

$$Y_\ell^{ij} = \sum_k \sqrt{\hat{q}_{\ell k}}\, S_k^{ij} + Z_\ell^{ij} \quad \forall i,j,\ell \tag{57}$$

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 (57) and (56) is analogous to those of [18, 30].

D.3. Output Channel Computation

Starting from the replicated partition function in Eq.(41), we can see that the output channel contribution to the free entropy of the model, factorized with respect to the data, is given by:

$$I_{\text{out}}(q) = \left[\int dy \int \prod_{u,\ell,a\le b} dh_{ab}^{(\ell)\,u} \, P\left(\{h_{ab}^{(\ell)\,u}\} \right) \prod_{u,\ell} P_{\text{out}}\left(y | \frac{h_{ab}^{(\ell)\,u}}{\sqrt{2-\delta_{ab}}} \right) \right]^n . \tag{58}$$

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

$$\mathbb{E}[h_{ab}^{(\ell)}] = 0 \qquad \mathrm{Cov}_{x^\mu}\left(h_{a\le b}^{(\ell)\,u}, h_{c\le d}^{(k)\,v} \right) = \frac{1}{d} \left[2\,\delta_{ac}\delta_{bd} \right] \mathrm{Tr}\left(S_\ell^u S_k^v \right) = \left[2\,\delta_{ac}\delta_{bd} \right] Q_{\ell k}^{uv} \tag{59}$$

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:

$$-\frac{1}{2} \sum_{u,v=0}^{n} \sum_{a\le b, c\le d} \sum_{\ell,k} \left(h_{ab}^{(\ell)\,u} \right) \left(\Sigma_h^{-1} \right)_{ab,cd}^{uv,\ell k} \left(h_{cd}^{(k)\,v} \right) \tag{60}$$

Substituting back, the output term becomes:

$$I_{out}(q) = \left[\int dy \int \prod_{u,\ell,a\le b} dh_{ab}^{(\ell)\,u} \exp\left(-\frac{1}{2} \sum_{u,v=0}^{m} \sum_{a,b,c,d} \sum_{\ell,k} h_{ab}^{(\ell)\,u} \left(\Sigma_h^{-1}\right)_{a\le b,c\le d}^{uv,pq} h_{cd}^{(k)\,v} \right) \prod_{u,\ell} P_{\rm out}\left(y | \left\{ \frac{h_{ab}^{(\ell)\,u}}{\sqrt{2-\delta_{ab}}} \right\}_{ab} \right) \right]^n \quad (61)$$

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

$$(\Sigma_h)_{a\le b,c\le d}^{uv,\ell k} \;=\; [2\,\delta_{ac}\delta_{bd}]\; Q_{\ell k}^{uv} \;=\; [2\,\delta_{ac}\delta_{bd}]\; \Big[(Q_{\ell k} - q_{\ell k})\,\delta_{uv} + q_{\ell k} \Big], \quad (62)$$

The inverse of the covariance matrix is given by:

$$\left[(\Sigma^{-1})_{ab}\right]^{u\ell,\,vq} = \frac{1}{2} \left[\left[(Q-q)^{-1}\right]_{\ell k} \delta_{uv} \;-\; \sum_{s=1}^{L} \sum_{t=1}^{L} \left[(Q-q)^{-1}\right]_{\ell s} q_{st} \left[Q^{-1}\right]_{tk} \right] \quad (63)$$

Introducing $M_\ell := \sum_{u=0}^{m} h_{ab}^{(\ell)\,u}$, we can rewrite the Gaussian exponent in Eq.(61) as:

$$-\frac{1}{4} \Big[\sum_{u,\ell,k} h_{ab}^{(\ell)\,u} \left[(Q-q)^{-1}\right]_{\ell k} h_{ab}^{(k)\,u} - \sum_{\ell,k} M_\ell \underbrace{\left(\sum_{s,t} \left[(Q-q)^{-1}\right]_{\ell s} q_{st} \left[Q^{-1}\right]_{tk} \right)}_{=:C_{\ell k}} M_k \Big]. \quad (64)$$

We now introduce L auxiliary Gaussian variables for this token pair, $\eta_{ab}^{(\ell)} \sim \mathcal{N}(0,1)$ ($\ell = 1,\dots,L$), via

$$\exp\Big(\tfrac{1}{4} \sum_{\ell,k} M_{ab}^{(\ell)} C_{\ell k} M_{ab}^{(k)} \Big) = \int \prod_{s=1}^{L} \frac{d\eta_{ab}^{(s)}}{\sqrt{2\pi}} \exp\Big(-\tfrac{1}{2} \sum_{s} (\eta_{ab}^{(s)})^2 + \tfrac{1}{2} \sum_{\ell,k} \eta_{ab}^{(\ell)} \sqrt{C}_{\ell k}\, M_{ab}^{(k)} \Big), \quad (65)$$

where the square root is intended over the spectrum of the matrix. After some algebra, we can manipulate the Gaussian exponent in the following way:

$$-\frac{1}{4} \sum_{\ell,k} \Big(h_{ab}^{(\ell)\,u} - \sum_{s=1}^{L} \sqrt{2q_{\ell s}}\, \eta_{ab}^{(s)} \Big) \left[(Q-q)^{-1}\right]_{\ell k} \Big(h_{ab}^{(k)\,u} - \sum_{t=1}^{L} \sqrt{2q_{kt}}\, \eta_{ab}^{(t)} \Big) - \frac{1}{2} \sum_{r} (\eta_{ab}^{(r)})^2. \quad (66)$$

We finally recognize:

$$-\frac{1}{2} \sum_{\ell=1}^{L} \sum_{k=1}^{L} \big(h_{ab}^{(\ell)\,u} - \omega_{ab}^{(\ell)} \big)\, [V_{ab}^{-1}]_{\ell k}\, \big(h_{ab}^{(k)\,u} - \omega_{ab}^{(k)} \big) \;-\; \frac{1}{2} \sum_{r} (\eta_{ab}^{(r)})^2 \quad (67)$$

with

$$\omega_{ab}^{(\ell)} = \sum_{k=1}^{L} \sqrt{2q}_{\ell k}\, \eta_{ab}^{(k)}\,, \qquad V^{(\ell k)} = 2(Q_{\ell k} - q_{\ell k}) \quad (68)$$

Hence we finally recognize a Gaussian term of the form:

$$\int \prod_{a\leq b}^{T} d^L h_{ab}\, \mathcal{N}\left(h_{ab}; \omega_{ab}; V_{ab}\right) P_{out}\left(y \mid h^u_{ab}\right) \tag{69}$$

Since replicas are decoupled through η, we can finally write out output channel term as:

$$I_{\text{out}}(q) = \left[\int dy \int \prod_{a\leq b,\ell} \frac{d\eta^{(\ell)}_{ab}}{\sqrt{2\pi}} e^{-\frac{(\eta^{(\ell)}_{ab})^2}{2}} \left(\int \prod_{a\leq b}^{T} d^L h_{ab} \mathcal{N}\left(h_{ab}; \omega_{ab}, V_{ab}\right) P_{\text{out}}\left(y \mid \{h^{(\ell)}_{ab}\}^L_{\ell=1}\right)\right)\right]^{n(m+1)} \tag{70}$$

Which, as for the prior term, we can write as :

$$I_{\text{out}}(q) = \left[\int dy \int \mathcal{D}\eta\left(\mathcal{Z}_{\text{out}}(y, \omega, V)\right)\right]^{n(m+1)} \tag{71}$$

and:

$$\mathcal{Z}_{\text{out}}(y, \omega, V) = \int \prod_{a\leq b} d^L h_{ab} \mathcal{N}\left(h_{ab}; \omega_{ab}, V_{ab}\right) P_{\text{out}}\left(y \mid \{\frac{h^{(\ell)}_{ab}}{\sqrt{2-\delta_{ab}}}\}^L_{\ell=1}\right) \tag{72}$$

Where we defined the measure over the auxiliary variables as

$$\mathcal{D}\eta = \prod_{\ell=1}^{L} \mathcal{D}\eta^{(\ell)} \ = \ \prod_{\ell=1}^{L} \prod_{a\leq b} \frac{d\eta^{(\ell)}_{ab}}{\sqrt{2\pi}} \ \exp\Big[-\frac{(\eta^{(\ell)}_{ab})^2}{2}\Big]\,. \tag{73}$$

Expanding Eq.(71) for small number of replicas m we get:

$$I_{\text{out}}(q) = \int dy \int \mathcal{D}\eta\, \mathcal{Z}_{\text{out}}(y, \omega, V)\ \ln \mathcal{Z}_{\text{out}}(y, \omega, V)\,, \tag{74}$$

Recalling the state equations found in Eq.(61) we get:

$$\hat{q} = 4\alpha \int D\eta \int dy\ \left(\frac{\partial \mathcal{Z}_{\text{out}}(y, \omega, V)}{\partial q}\left(1 + \ln \mathcal{Z}_{\text{out}}(y, \omega, V)\right)\right) \tag{75}$$

Now, since we know that

$$\frac{\partial \omega^{(r)}_{ab}}{\partial \eta^{(k)}_{ab}} = (\sqrt{2q})_{rk} \tag{76}$$

for each token pair (a, b), we define the denoising function:

$$\left(g_{\text{out}}(y, \omega, V)\right)^{(\ell)}_{ab} = \frac{\partial}{\partial \omega^{(\ell)}_{ab}} \ln \mathcal{Z}_{\text{out}}(y, \omega, V) = \frac{\partial \mathcal{Z}_{\text{out}}(y, \omega, V)}{\partial \omega^{(\ell)}_{ab}} \mathcal{Z}_{\text{out}}(y, \omega, V)^{-1}\,. \tag{77}$$

So, using the chain rule and the definition (77):

$$\partial_{\eta^{(k)}_{ab}} \mathcal{Z}_{\text{out}} = \sum_{r=1}^{L} \frac{\partial \mathcal{Z}_{\text{out}}}{\partial \omega^{(r)}_{ab}} \frac{\partial \omega^{(r)}_{ab}}{\partial \eta^{(k)}_{ab}} = Z_{\text{out}}(y, \omega, V) \sum_{r=1}^{L} g^{(r)}_{ab}(y, \omega, V)\ \ (\sqrt{2q})_{rk}.$$

We can write the derivative explicitly:

$$\frac{\partial \mathcal{Z}_{\rm out}}{\partial q_{\ell k}} = \sum_{a\le b}\sum_{r=1}^{L} \frac{\partial \mathcal{Z}_{\rm out}}{\partial \omega_{ab}^{(r)}} \frac{\partial \omega_{ab}^{(r)}}{\partial q_{\ell k}} \;+\; \frac{\partial \mathcal{Z}_{\rm out}}{\partial V_{\ell k}} \frac{\partial V_{\ell k}}{\partial q_{\ell k}}. \tag{78}$$

Regarding the first term, one can show that:

$$\frac{\partial \omega_{ab}^{(r)}}{\partial q_{\ell k}} = (\sqrt{2q})^{-1}_{\ell k}\, \omega_{ab}^{(r)} \;+\; \text{(sym. in } \ell, k). \tag{79}$$

Multiplying by $\partial_{\omega_{ab}^{(r)}} \mathcal{Z}_{\rm out} = g_{ab}^{(r)}\, \mathcal{Z}_{\rm out}$ and using $\omega_{ab}^{(r)} = \sum_s (\sqrt{2q})_{rs} \eta_{ab}^{(s)}$ gives

$$\frac{1}{2}\sum_{a\le b}\left[\eta_{ab}^{(k)}\, g_{ab}^{(\ell)} + \eta_{ab}^{(\ell)}\, g_{ab}^{(k)}\right] \mathcal{Z}_{\rm out} \tag{80}$$

Regarding the second piece contribution, we first notice $V_{\ell k} = 2(Q_{\ell k} - q_{\ell k})$, giving us $\partial_{q_{\ell k}} V_{\ell k} = -2$. A straightforward but lengthy calculation shows that the term coming from $\partial \mathcal{Z}_{\rm out}/\partial V_{\ell k}$ exactly cancels the $1/2$ factor above. Finally, combining the two pieces yields

$$\frac{\partial \mathcal{Z}_{\rm out}}{\partial q_{\ell k}} = \sum_{a\le b}\Big[\eta_{ab}^{(k)}\, g_{ab}^{(\ell)} + \eta_{ab}^{(\ell)}\, g_{ab}^{(k)}\Big] \mathcal{Z}_{\rm out} \tag{81}$$

which, by the definition of the denoising function and the means $\omega_{ab}^{(\ell)}$ can be manipulated into:

$$\frac{\partial \mathcal{Z}_{\rm out}}{\partial q_{\ell k}} = \frac{1}{2\, q_{\ell k}} \sum_{a\le b}\Big[\eta_{ab}^{(k)}\, \partial_{\eta_{ab}^{(\ell)}} \mathcal{Z}_{\rm out} + \eta_{ab}^{(\ell)}\, \partial_{\eta_{ab}^{(k)}} \mathcal{Z}_{\rm out}\Big]. \tag{82}$$

Now, recalling that:

$$\widehat{q}_{\ell k} = 4\alpha \int \mathcal{D}\eta\, dy\, \left[1 + \ln \mathcal{Z}_{\rm out}\right]\, \partial_{q_{\ell k}} \mathcal{Z}_{\rm out}. \tag{83}$$

Inserting (82) leads:

$$\widehat{q}_{\ell k} = \frac{2\alpha}{q_{\ell k}} \sum_{a\le b} \int \mathcal{D}\eta\, dy\, \left[1 + \ln \mathcal{Z}_{\rm out}\right]\left[\eta_{ab}^{(k)}\, \partial_{\eta_{ab}^{(\ell)}} \mathcal{Z}_{\rm out} + \eta_{ab}^{(\ell)}\, \partial_{\eta_{ab}^{(k)}} \mathcal{Z}_{\rm out}\right]. \tag{84}$$

Using $\int \mathcal{D}\eta\, \eta F = \int \mathcal{D}\eta\, \partial_\eta F$ on both terms, i.e.

$$\int \mathcal{D}\eta\, \eta_{ab}^{(k)} \partial_{\eta_{ab}^{(\ell)}} \mathcal{Z}_{\rm out} = \int \mathcal{D}\eta\, \partial_{\eta_{ab}^{(k)}} \big(\partial_{\eta_{ab}^{(\ell)}} \mathcal{Z}_{\rm out}\big). \tag{85}$$

and exploiting the symmetry: $(\ell \leftrightarrow k)$, we finally arrive to

$$\widehat{q}_{\ell k} = \frac{4\alpha}{q_{\ell k}} \sum_{a\le b} \int \mathcal{D}\eta\, dy\, \partial_{\eta_{ab}^{(\ell)}} \mathcal{Z}_{\rm out}\, \partial_{\eta_{ab}^{(k)}} \ln \mathcal{Z}_{\rm out}. \tag{86}$$

which, because $\partial_\eta \ln \mathcal{Z}_{\rm out} = (\partial_\eta \mathcal{Z}_{\rm out})/\mathcal{Z}_{\rm out}$, finally becomes

$$\widehat{q}_{\ell k} = \frac{4\alpha}{q_{\ell k}} \sum_{a\le b} \int \mathcal{D}\eta\, dy\, \frac{\big(\partial_{\eta_{ab}^{(\ell)}} \mathcal{Z}_{\rm out}\big)\, \big(\partial_{\eta_{ab}^{(k)}} \mathcal{Z}_{\rm out}\big)}{\mathcal{Z}_{\rm out}}. \tag{87}$$

Recalling the definition of the output denoising function, we can express the output channel state equation of our model as:

$$\widehat{q}_{\ell k} = 4\alpha\, \mathbb{E}_{\eta,y}\left[\sum_{a\le b}(g_{out}(y,\omega,V))^{(\ell)}_{ab}(g_{out}(y,\omega,V))^{(k)}_{ab}\right], \quad \text{for} \quad \ell = 1,\dots,L \quad a \le b = 1,\dots,T \tag{88}$$

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

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

and the output y is drawn from the channel density

$$P_{out}\Big(y \mid \{h^{(\ell)}_{ab}\}_\ell\Big), \quad h^{(\ell)}_{ab} \sim \mathcal{N}\Big(\omega^{(\ell)}_{ab},\, V^{(\ell k)}_{ab}\Big),$$

with:

$$P_{out}\Big(y \;\Big|\; \{h^{(\ell)}_{ab}\}_\ell\Big) = \delta(\{y_{ab} - g(\{h^{(\ell)}_{ab}\}_{\forall\ell})_{ab}\}_{\forall ab}) \tag{89}$$

or particularly, for the deep attention case:

$$P_{out}\Big(y \;\Big|\; \{h^{(\ell)}_{ab}\}_\ell\Big) = \delta(\{y_{ab} - B^L_c(\{h^{(\ell)}_{ab}\}_{\forall\ell})_{ab}\}_{\forall ab}) \tag{90}$$

D.4. 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:

$$\begin{aligned}\mathcal{Z}_{\text{in}}(\{Y_\ell\}_{\ell=1}^L;\hat{q}) = \int \left[\prod_{\ell=1}^{L} \mathrm{d}S_\ell\, P_S(S_\ell)\right] \\ \times \exp\Big[-\tfrac{d}{4}\sum_{\ell,k=1}^{L}\hat{q}_{\ell k}\,\mathrm{Tr}(S_\ell S_k) + \tfrac{d}{2}\sum_{\ell,k=1}^{L}\sqrt{\hat{q}}_{\ell k}\,\mathrm{Tr}(Y_\ell S_k)\Big].\end{aligned} \tag{91}$$

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

$$g_{in}(Y|\hat{q}) = \partial_{\hat{q}^{1/2}Y} \ln \mathcal{Z}_{\text{in}}(Y,\hat{q}) \tag{92}$$

The free entropy contribution coming from the prior channel is:

$$I_{\text{in}}(\hat{q}) = \lim_{d\to\infty}\frac{1}{d^2}\int DY_1\dots DY_L\, \mathcal{Z}_{\text{in}}(Y_1,\dots,Y_L;\hat{q})\log \mathcal{Z}_{\text{in}}(Y_1,\dots,Y_L;\hat{q}) \tag{93}$$

where DY stands for integration over a G$OE(d)$ (Wigner) matrix Y.

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

$$Y(S,\Delta)_\ell = S_\ell + \sum_{m=1}^{L} \sqrt{\Delta}_{\ell m}\Xi_m, \quad \Xi_m \sim \mathrm{G}OE(d) \quad S_\ell \sim P_S \tag{94}$$

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

$$\begin{aligned} \mathcal{Z}_{\text{out}}(y,\omega,V) &= \int \left[\prod_{a\le b}^{T} d^L h_{ab}\, \mathcal{N}\left(h_{ab};\omega_{ab};V_{ab}\right)\right] \delta\left(y - g(\{h_{ab}^{(\ell)}/\sqrt{2-\delta_{ab}}\}_{\ell=1}^{L})\right) \\ \text{with:} \quad \omega_{ab}^{(\ell)} &= \sum_{k=1}^{L} \sqrt{2q}_{\ell k}\, \eta_{ab}^{(k)}\,, \qquad V^{(\ell k)} = 2(Q_{\ell k} - q_{\ell k}) \end{aligned} \tag{95}$$

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

$$g_{out}(y,\omega,V) = \partial_\omega \ln \mathcal{Z}_{\text{out}}(y,\omega,V) \tag{96}$$

The free entropy contribution coming from the output channel is:

$$I_{\text{out}}(q) = \int \prod_{a,b=1}^{T} dy_{ab} \int \mathcal{D}\eta_1 \dots \mathcal{D}\eta_L\, \mathcal{Z}_{\text{out}}(y,\omega,V) \log \mathcal{Z}_{\text{out}}(y,\omega,V) \tag{97}$$

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:

$$\begin{aligned} \hat{q}_{\ell k} &= 4\alpha \mathbb{E}_{\xi,\eta} \sum_{a\le b}^{L} g_{\text{out}}\left(g\left(\left\{\frac{h(\omega,V)_{ab}}{\sqrt{2-\delta_{ab}}}\right\}\right),\omega,V\right)_{ab}^{(\ell)} \\ &\quad \times g_{\text{out}}\left(g\left(\left\{\frac{h(\omega,V)_{ab}}{\sqrt{2-\delta_{ab}}}\right\}\right),\omega,V\right)_{ab}^{(k)}, \\ q_{\ell k} &= \lim_{d\to+\infty} \frac{1}{d}\mathbb{E}_{S,Y}\, \mathrm{Tr}\left[g_{\text{in}}(Y(S,\hat{q}),\hat{q})_\ell g_{\text{in}}(Y(S,\hat{q}),\hat{q})_k\right], \end{aligned} \tag{98}$$

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. (94) and $S \sim P_S$. Finally:

$$[h(\omega,V)_{ab}]^{(\ell)} = \omega_{ab}^{(\ell)} + \sum_k \sqrt{V}^{(\ell k)} \xi_{ab}^{(k)} \tag{99}$$

Algorithm 1: AMP

Input: Observations $y^\mu \in \mathbb{R}^{T\times T}$ and "sensing matrices"

$$Z^\mu_{ij,ab} = \frac{\boldsymbol{x}^\mu_{i,a}\boldsymbol{x}^\mu_{j,b} + \boldsymbol{x}^\mu_{j,a}\boldsymbol{x}^\mu_{i,b} - 2\delta_{ij}\delta_{ab}}{\sqrt{2d(1+\delta_{ab})}} \in \mathbb{R}$$

Result: The estimators $\hat{S}_\ell$

Initialize $\hat{S}^0_\ell \sim P_S$ Initialize $\hat{C}^0 = 2(\kappa_2 - \kappa_1^2)\,\mathbb{I}_L$ $t \leftarrow 0$

while *not converged* **do**

`// Estimate variance and mean of` $\mathrm{Tr}[Z^\mu_{ab}S_\ell]$

$V^t \leftarrow 2\hat{C}^t$ $\omega^t_{\mu,ab} \leftarrow \mathrm{Tr}[Z^\mu_{ab}\hat{S}^t] - (1-\delta_{0t})\, g_{\rm out}(y^\mu, \omega^{t-1}_\mu, V^{t-1})_{ab}\cdot V^t$

`// Estimate variance and mean of` S_ℓ `from "output" observations`

$\hat{q}^t_{\ell k} \leftarrow \frac{4\alpha}{n}\sum_{\mu=1}^n \sum_{a\le b}^T g_{\rm out}(y^\mu, \omega^t_\mu, V^t)^{(\ell)}_{ab}\cdot g_{\rm out}(y^\mu, \omega^t_\mu, V^t)^{(k)}_{ab}$, $R^t_{ij} \leftarrow \hat{S}^t_{ij} + (\hat{q}^{\,t})^{-1}\cdot \frac{2}{d}\sum_{\mu=1}^n\sum_{a\le b}^T g_{\rm out}(y^\mu,\omega^t_\mu,V^t)_{ab}\cdot Z^\mu_{ij,ab}$

`// Update estimator with "input" information`

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

$t \leftarrow t+1$

end

D.5. 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

$$m^t_{\ell k} = \mathrm{Tr}[\hat{S}^t_\ell S^*_k]/d\,, \qquad q^t_{\ell k} = \mathrm{Tr}[\hat{S}^t_\ell \hat{S}^t_k]/d\,. \tag{100}$$

We now claim, in analogy with [18, 30, 42, 48] 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

$$\mathbb{E}[(\omega^t_{\mu,\,ab})_\ell\,(\omega^t_{\mu,\,ab})_k] = 2q^t_{\ell k}\,, \qquad \mathbb{E}[(\omega^*_{\mu,\,ab})_\ell\,(\omega^t_{\mu,\,ab})_k] = 2m^t_{\ell k}\,, \tag{101}$$

By Nishimori's identities [35] we can assume $m^t = q^t$. The equation (88) is now immediately recovered (modulo the substitution $V \to 2(Q-q^t)$ which will come after)

$$\hat{q}^t_{\ell k} \approx 4\alpha \mathbb{E}_{y,\,\omega^t}\sum_{a\le b}^T \left[g_{\rm out}(y,\omega^t,V^t)^{(\ell)}_{ab}\, g_{\rm out}(y,\omega^t,V^t)^{(k)}_{ab}\right] \tag{102}$$

where

$$y = g\left(\left\{\frac{\omega^*_{ab}}{\sqrt{2-\delta_{ab}}}\right\}^T_{a\le b}\right)\,, \qquad \begin{pmatrix}\omega^t_{ab}\\ \omega^*_{ab}\end{pmatrix} \sim \mathcal{N}\left(0\,, \begin{pmatrix}2q^t & 2m^t\\ 2m^t & 2Q\end{pmatrix}\right) \tag{103}$$

Again as in [18, 30, 42, 48] we will have that in distribution

$$R^t_{ij} = S^*_{ij} + (\hat{q}^t)^{-1}\Xi^t_{ij} \tag{104}$$

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

$$q^t_{\ell k} = \mathbb{E}_{R^t}\,\mathrm{Tr}\left[g_{in}\left(R^t,\hat{q}^t\right)_\ell\, g_{in}\left(R^t,\hat{q}^t\right)_k\right]/d\,, \tag{105}$$

which is exactly the second equation emerging from (56). Notice how the expectation is taken over the random variable R^t_{ij} in (104). The last step is to notice that

$$\hat{C}^t_{\ell k} = \mathrm{Tr}[(\hat{S}^t_\ell - S^*_\ell)(\hat{S}^t_k - S^*_k)]/d^2 = Q - q^t \tag{106}$$

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

Appendix E. 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:

$$S = \frac{1}{\sqrt{rd}} WW^\top \in \mathbb{R}^{d\times d} \quad W \in \mathbb{R}^{d\times r} \quad (W)_{ij} \sim \mathcal{N}(0,1) \tag{107}$$

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.(B).

E.1. Prior channel state equation

Starting from Eq.(56) for $L = 1$ layer, we get::

$$\begin{aligned} I_{\text{in}}(\hat{q}) &= \lim_{d\to\infty} \frac{1}{d^2} \int DY\, \mathcal{Z}_{\text{in}}(Y;\hat{q}) \log \mathcal{Z}_{\text{in}}(Y;\hat{q}) \\ \mathcal{Z}_{\text{in}}(Y;\hat{q}) &= \int dS\, P_0(S) \exp\left(-\frac{d}{2}\left(\hat{Q} + \frac{\hat{q}}{2}\right) \mathrm{Tr}\left(S^T S\right) + \frac{\sqrt{\hat{q}}d}{2} \mathrm{Tr}\left(Y^T S\right)\right). \end{aligned} \tag{108}$$

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

$$\begin{aligned} \int \mathcal{D}Y\, \mathcal{Z}_{\text{in}}(Y;\hat{q}) &= \int \mathcal{D}Y \int dS\, P_0(S) \exp\left(-\frac{\hat{q}d}{4} \mathrm{Tr}(S^\top S) + \frac{\sqrt{\hat{q}}d}{2} \mathrm{Tr}\left(Y^\top S\right) - \frac{1}{4} \mathrm{Tr}\left(Y^\top Y\right)\right) \\ &= \int dS\, P_0(S) \exp\left(\frac{\hat{q}d}{4} \mathrm{Tr}\left(S^\top S\right)\right) \exp\left(-\frac{\hat{q}d}{4} \mathrm{Tr}\left(S^\top S\right)\right) \\ &= \int dS\, P_0(S) = 1 \end{aligned} \tag{109}$$

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

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

Observe

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

Hence

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

Therefore:

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

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

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

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:

$$S = U\,\Lambda\,U^T$$

where $\Lambda = \operatorname{diag}(\lambda_1, \dots, \lambda_d)$. Then

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

Then the exponent

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

becomes

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

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

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

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

$$\Phi_{\mathrm{prior}}(\hat{q}) \;=\; \lim_{d\to\infty} \frac{1}{d^2}\,\mathbb{E}\big[\ln I_0(Y)\big], \tag{110}$$

or more explicitly :

$$\Phi_{\mathrm{prior}}(\hat{q}) = \lim_{d\to\infty} \frac{1}{d^2}\,\mathbb{E}\big[\ln \int P_0(S)\, e^{-\frac{d}{4}\operatorname{Tr}(\sqrt{\hat{q}}\,S - Y)^2}\,\mathrm{d}S\big]. \tag{111}$$

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

$$\Phi_{\mathrm{prior}}(\hat{q}) = \lim_{d\to\infty} \frac{1}{d^2}\,\mathbb{E}_Y \ln I_0(Y) \;=\; -\frac{\hat{q}\,Q}{4} + \frac{1}{2} I_{\mathrm{HCIZ}}\Big(\hat{q}; \mu_0, \mu_0 \boxplus \sigma_{\mathrm{sc},1/\sqrt{\hat{q}}}\Big) + \text{const}, \tag{112}$$

where $Q = 1 + \rho$.Then, one has the relation from [29] :

$$-\frac{1}{2}\Sigma\big(\mu_{\hat{q}}\big) + \frac{1}{4\hat{q}}\mathbb{E}_{\mu_{\hat{q}}}[X^2] - \frac{1}{2} I_{\mathrm{HCIZ}}\Big(\hat{q}; \mu_0, \mu_{\hat{q}}\Big) - \frac{3}{8} + \frac{1}{4}\ln\hat{q} + \frac{1}{4\hat{q}}\mathbb{E}_{\mu_0}[X^2] = 0, \tag{113}$$

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

$$\frac{1}{2} I_{\mathrm{HCIZ}}(\hat{q}; \mu_0, \mu_{\hat{q}}) = -\frac{1}{2}\Sigma(\mu_{\hat{q}}) + \frac{1}{4}\Big[2Q\hat{q} + 1\Big] - \frac{3}{8} - \frac{1}{4}\ln\hat{q}. \tag{114}$$

Plugging back into the free entropy, we obtain

$$\Phi_{\mathrm{prior}}(\hat{q}) = -\frac{\hat{q}\,Q}{4} + \left[-\frac{1}{2}\Sigma(\mu_{\hat{q}}) + \frac{1}{4}(2Q\hat{q}+1) - \frac{3}{8} - \frac{1}{4}\ln\hat{q}\right] + \text{const.} \tag{115}$$

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

$$\frac{\partial\Phi}{\partial\hat{q}} = -\frac{q}{4} + \frac{Q}{4} - \frac{1}{4\hat{q}} - \frac{1}{2}\frac{\partial}{\partial\hat{q}}\Sigma\big(\mu_{\hat{q}}\big) = 0. \tag{116}$$

Using the derivative:

$$\frac{\partial}{\partial\hat{q}}\Sigma\big(\mu_{\hat{q}}\big) = -\frac{2\pi^2}{3\hat{q}^2}\int \mu_{\hat{q}}(x)^3\,dx, \tag{117}$$

this condition becomes

$$-\frac{q}{4} + \frac{Q}{4} - \frac{1}{4\hat{q}} + \frac{\pi^2}{3\hat{q}^2}\int \mu_Y(x)^3\,dx = 0. \tag{118}$$

which is exactly our desired state equation.

To sum up, in the problem

$$Y \;=\; \sqrt{\hat{q}}\,S \;+\; Z, \quad Z \sim \mathrm{GOE}(d), \tag{119}$$

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

$$q = Q - \frac{1}{\hat{q}} + \frac{4\,\pi^2}{3\,\hat{q}^{\,2}}\int \big[\mu_Y(x)\big]^3\,\mathrm{d}x, \tag{120}$$

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

$$\mathcal{R}_{\mu_Y}(z) \;=\; \mathcal{R}_{\mu_S}(z) \;+\; \mathcal{R}_{\sigma_{\mathrm{sc},\,\alpha}}(z). \tag{121}$$

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

$$\mathcal{R}_{\mu_{MP,\rho}}(z) = \frac{\rho}{\sqrt{\rho} - z}. \tag{122}$$

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

$$\mathcal{R}_{\mu_Y}(z) = \frac{\rho}{\sqrt{\rho} - z} \;+\; \alpha^2\,z. \tag{123}$$

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

$$g_{\mu_Y}(z) = \int \frac{\mu_Y(\mathrm{d}x)}{x - z}. \tag{124}$$

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

$$z \;=\; \frac{1}{x} \;+\; \frac{\rho}{\sqrt{\rho} - x} \;+\; \alpha^2\, x. \tag{125}$$

Hence the final polynomial in x is:

$$(\frac{1}{\sqrt{\rho}}\alpha^2)x^3 - (\frac{z}{\sqrt{\rho}} + \alpha^2)x^2 + (z + \frac{1}{\sqrt{\rho}} - \sqrt{\rho})x - 1 = 0 \iff x = g_{\mu_Y}(z). \tag{126}$$

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.

$$\mu_Y(x_0) \;=\; \lim_{\epsilon \to 0^+} \frac{1}{\pi} \operatorname{Im} g_{\mu_Y}(x_0 - i\,\epsilon). \tag{127}$$

E.2. Output channel state equation

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

$$\hat{q} = 4\,\alpha\,\mathbb{E}_{(\eta,y)}\Big[\sum_{a \le b} (g_{out}(y, \omega, V))^2_{ab}\Big] \tag{128}$$

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

$$(g_{out}(y, \omega, V))_{ab} = \partial_{\omega_{ab}} \ln \mathcal{Z}_{out}(y, \omega, V) \tag{129}$$

and:

$$\mathcal{Z}_{out}(y, \omega, V) = \int \prod_{a \le b} dh_{ab} \mathcal{N}(h_{ab}, \omega_{ab}, V)\, \delta\big(y - f(h)\big) \tag{130}$$

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.

E.3. 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

$$\mathcal{Z}_{out}(y,\omega,V) = \int \Big[\prod_{a\le b} \mathcal{N}\big(h_{ab};\omega_{ab},V\big)\Big] \prod_{a\le b} \delta\big[y_{ab} - \frac{h_{ab}}{\sqrt{2-\delta_{ab}}}\big]\, \mathrm{d}h. \tag{131}$$

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

$$\mathcal{Z}_{out}(y,\omega,V) = \prod_{a\le b}\Big[\frac{1}{\sqrt{2\pi\, V}}\exp\Big(-\tfrac{(\sqrt{2-\delta_{ab}}y_{ab}-\omega_{ab})^2}{2\,V}\Big)\Big]. \tag{132}$$

Hence

$$\ln \mathcal{Z}_{\text{out}}(y,\omega,V) = \sum_{a\le b}\Big[-\tfrac{1}{2}\ln(2\pi\, V) \;-\; \frac{(\sqrt{2-\delta_{ab}}y_{ab}-\omega_{ab})^2}{2\,V}\Big]. \tag{133}$$

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

$$\omega_{ab} = \sqrt{2q}\;\eta_{ab}, \quad V_{ab} = 2(Q-q), \quad Q \;=\; 1+\rho. \tag{134}$$

Then

$$(g_{out}(y,\omega,V))_{ab} = \frac{\partial}{\partial\omega_{ab}}\ln\mathcal{Z}_{\text{out}}(y,\omega,V) = -\frac{\partial}{\partial\omega_{ab}}\Big[\frac{(\sqrt{2-\delta_{ab}}y_{ab}-\omega_{ab}(\eta))^2}{2\,V}\Big] = +\frac{(\sqrt{2-\delta_{ab}}y_{ab}-\omega_{ab})}{V}. \tag{135}$$

Thus

$$\sum_{a\le b}(g_{\text{out}}(y,\omega,V))^2_{ab} = \sum_{a\le b}\Big(\big[\,\sqrt{2-\delta_{ab}}y_{ab}-\omega_{ab}(\eta)\,\big]\frac{1}{2(Q-q)}\Big)^2. \tag{136}$$

We can compute the expectation:

$$\begin{aligned}&\mathbb{E}\Big[\sum_{a\le b}(g_{\text{out}}(y,\omega,V))^2_{ab}\Big]\\ &= \int\Big(\prod_{a\le b}\mathrm{d}\eta_{ab}\,\frac{e^{-\eta_{ab}^2/2}}{\sqrt{2\pi}}\Big)\int\Big(\prod_{a\le b}\mathrm{d}y_{ab}\Big)\,\mathcal{Z}_{\text{out}}(y,\omega,V)\;\sum_{a\le b}(g_{\text{out}}(y,\omega,V))^2_{ab}.\end{aligned} \tag{137}$$

We can simply use:

$$\int \mathrm{d}y_{ab}\,\frac{1}{\sqrt{2\pi\,V}}\exp\Big(-\frac{(\sqrt{2-\delta_{ab}}y_{ab}-\omega_{ab})^2}{2\,V}\Big)\,\big[\sqrt{2-\delta_{ab}}y_{ab}-\omega_{ab}\big]^2 = V. \tag{138}$$

Therefore,

$$\int\Big(\prod_{a\le b}\mathrm{d}y_{ab}\Big)\,\mathcal{Z}_{\text{out}}(y,\omega,V)\sum_{a\le b}(g_{\text{out}}(y,\omega,V))^2_{ab} \;=\; \sum_{a\le b}\Big[\big(\frac{1}{2(Q-q)}\big)^2\,V\Big]. \tag{139}$$

Thus each term becomes

$$\big(\frac{1}{2(Q-q)}\big)^2\,2(Q-q) \;=\; \frac{1}{2(Q-q)}. \tag{140}$$

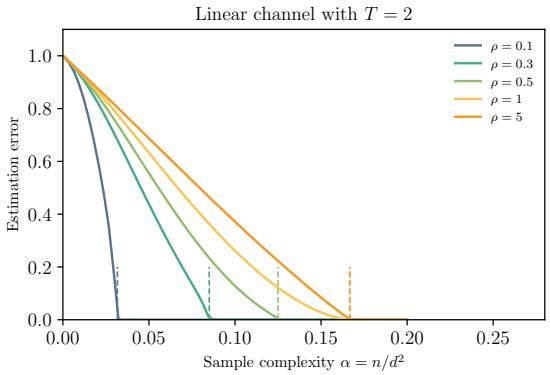

Figure 3: Illustration of the Bayes-optimal error for the linear output channel baseline in Eq.(131), for $T = 2$ tokens and several values of the width ratio $\rho = r/d$. The model reaches zero BO error at finite α. The recovery threshold matches perfectly the one find by the simple counting argument in (144), plotted in short vertical lines.

Hence the entire sum is

$$\sum_{a\leq b} \frac{1}{2(Q-q)} = \frac{T(T+1)}{4}\frac{1}{Q-q}. \tag{141}$$

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:

$$\hat{q} = 4\,\alpha\,\mathbb{E}_{(\eta,y)}\Big[\sum_{a\leq b}(g_{out}(y,\omega,V))^2_{ab}\Big] = 4\alpha\sum_{a\leq b}\Big[\mathbb{E}_{(\eta,y)}g_{out}(y,\omega,V))^2_{ab}\Big] = 4\alpha\frac{T(T+1)}{4(Q-q)} \tag{142}$$

which finally simplifies into the output channel state equation:

$$\hat{q} = \frac{T(T+1)\,\alpha}{Q-q} \tag{143}$$

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 (8) remains unchanged, while we use Eq.(142) for simulating the linear output channel results. We also show in vertical dashed lines the recovery threshold found by the simple counting problem:

$$\frac{T(T+1)}{2}\alpha_{count} = \rho - \frac{\rho^2}{2} \tag{144}$$

The linear output channel matches perfectly the counting recovery threshold, unlike in the softmax case shown in Eq.(8).

E.4. Softmax output channel for generic number of tokens T

We compute the quantity:

$$\mathcal{Z}_{out}(y,\omega,V) = \int \prod_{a\leq b} dh_{ab}\,\frac{1}{\sqrt{2\pi V_{ab}}}e^{-\frac{(h_{ab}-\omega_{ab})^2}{2V_{ab}}}\prod_{a\leq b}\delta(y_{ab} - \text{Softmax}\{\frac{\beta}{\sqrt{2-\delta_{ab}}}h_{ab}\}). \tag{145}$$

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:

$$\mathcal{Z}_{out}(y,\omega,V) = \int \prod_{a\leq b} dh_{ab}\delta(y-\sigma(\frac{h_{ab}}{\tau_{ab}}) \prod_{a\leq b} \mathcal{N}(h_{ab},\omega_{ab},V_{ab}) \tag{146}$$

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:

$$\begin{aligned}\mathcal{Z}_{out}(y,\omega,V) &= \int \prod_{a\leq b} dz_{ab}\,\delta\big(y-\sigma(z)\big) \prod_{a\leq b} \mathcal{N}\big(z_{ab},\tfrac{\omega_{ab}}{\tau_{ab}},\tfrac{V_{ab}}{\tau^2_{ab}}\big) = \\ &= \int \prod_{a\leq b<T} dt_{ab}\,\mathcal{N}\big(t_{ab},\tfrac{\omega_{ab}}{\tau_{ab}}-s_a,\tfrac{V_{ab}}{\tau^2_{ab}}\big) \prod_{a=1}^{T} ds_a\,\mathcal{N}\big(s_a,\tfrac{\omega_{aT}}{\tau_{aT}},\tfrac{V_{aT}}{\tau^2_{aT}}\big)\end{aligned} \tag{147}$$

where in the last equality we introduced the inverse mapping of the row-wise softmax function, defined in Eq.(14). In particular, we introduce:

$$e^{\beta t_{ab}} = \frac{e^{\beta z_{ab}}}{e^{\beta z_{aT}}} = \frac{e^{\beta z_{ab}}}{\sum_{b=1}^{T} e^{\beta z_{ab}}}\Big(\frac{e^{\beta z_{aT}}}{\sum_{b=1}^{T} e^{\beta z_{ab}}}\Big)^{-1} = \frac{y_{ab}}{y_{aT}} \quad \forall a\leq b<T \tag{148}$$

which leads to:

$$t_{ab} = \frac{1}{\beta}\log(\frac{y_{ab}}{y_{aT}}) = \phi_{ab}(y) \quad \forall a\leq b<T \tag{149}$$

while for $b=T$:

$$\frac{y_{Ta}}{y_{TT}} = \frac{e^{\beta z_{Ta}}}{e^{\beta z_{TT}}} = \frac{e^{\beta z_{aT}}}{e^{\beta z_{TT}}} = e^{\beta(s_a-s_{TT})} \to s_a = s_{TT}+\phi_{Ta}(y) \quad \forall a<T \tag{150}$$

having introduced the change of variables:

$$z_{ab}\to t_{ab} = z_{ab}-z_{aT} \to z_{ab} = t_{ab}+s_a = \phi_{ab}+\phi_{Ta}+s_{TT} \quad \forall a\leq b<T \tag{151}$$

and

$$z_{aT}\to s_a = z_{aT} \to z_{aT} = s_a = s_{TT}+\phi_{Ta} \quad \forall a<T \tag{152}$$

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.(147) to that of one simple scalar integral in the variable $x = s_T$, namely:

$$\begin{aligned}\mathcal{Z}_{out}(y,\omega,V) &= \int dx\mathcal{N}(x,\tilde{\omega}_{TT},\tilde{V}_{TT}) \prod_{a=1}^{T-1} \mathcal{N}(x+\phi_{Ta}(y),\tilde{\omega}_{aT},\tilde{V}_{aT}) \\ &\quad\times \prod_{a\leq b<T} \mathcal{N}(\phi_{ab}(y)+\phi_{Ta}(y)+x,\tilde{\omega}_{ab},\tilde{V}_{ab}) \\ &= \int dx\exp\Big\{-\frac{1}{2}\Big[\sum_{a=1}^{T}\frac{(x+\phi_{Ta}-\tilde{\omega}_{aT})^2)}{\tilde{V}_{aT}} + \sum_{a\leq b<T}\frac{(\phi_{ab}+\phi_{Ta}+x-\tilde{\omega}_{ab})^2}{\tilde{V}_{ab}}\Big]\Big\}\end{aligned} \tag{153}$$

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

$$-\frac{1}{2}\Big[x^2\Big(\sum_{a\leq b}\tilde{V}_{ab}^{-1}\Big)+2x\Big(\sum_{a=1}^{T}\frac{\phi_{Ta}-\tilde{\omega}_{aT}}{\tilde{V}_{aT}}+\sum_{a\leq b<T}\frac{\phi_{ab}+\phi_{Ta}-\tilde{\omega}_{aT}}{\tilde{V}_{ab}}\Big) + \Big(\sum_{a=1}^{T}\frac{(\phi_{Ta}-\tilde{\omega}_{aT})^2}{\tilde{V}_{aT}}+\sum_{a\leq b<T}\frac{(\phi_{ab}+\phi_{Ta}-\tilde{\omega}_{ab})^2}{\tilde{V}_{ab}}\Big)\Big] \tag{154}$$

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

$$\log\mathcal{Z}=\frac{1}{2\tilde{V}}\Big[\sum_{a=1}^{T}\frac{\phi_{Ta}-\tilde{\omega}_{aT}}{\tilde{V}_{aT}}+\sum_{a\leq b<T}\frac{\phi_{ab}+\phi_{Ta}-\tilde{\omega}_{aT}}{\tilde{V}_{ab}}\Big]^2 - \frac{1}{2}\Big[\sum_{a=1}^{T}\frac{(\phi_{Ta}-\tilde{\omega}_{aT})^2}{\tilde{V}_{aT}}+\sum_{a\leq b<T}\frac{(\phi_{ab}-\phi_{Ta}-\tilde{\omega}_{ab})^2}{\tilde{V}_{ab}}\Big]+cost \tag{155}$$

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:

$$g_{out}(y,\omega,V)_{ab}=\partial_{\omega_{ab}}\log\mathcal{Z}_{\text{out}}(y,\omega,V) \tag{156}$$

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.(5) for an arbitrary number of tokens, substituting back the original V and ω:

$$V(g_{out})_{ij}=-\frac{\tau_{ij}}{T^2}\Big[\sum_{a\leq b}^{T}\tau_{ab}^2\phi_{Ta}-\sum_{a\leq b}^{T}\tau_{ab}\omega_{ab}+\sum_{a\leq b}^{T-1}\tau_{ab}^2\phi_{ab}\Big]+\tau_{ij}\phi_{Ti}-\omega_{ij}+\delta(j<T)\phi_{ij}\tau_{ij} \tag{157}$$

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

$$\phi_{ab}=h_{ab}-h_{aT}\quad h\sim\mathcal{N}(\tilde{\omega},\tilde{V})\rightarrow\tau_{ab}h_{ab}=\sqrt{2q}\,\eta_{ab}+\sqrt{V}\,\xi_{ab}\quad a\leq b\leq T \tag{158}$$

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

$$\phi_{Ta}=h_{Ta}-h_{TT}=h_{aT}-h_{TT} \tag{159}$$

We thus substitute these relationships inside Eq.(157) and finally compute $\mathbb{E}_{\eta,\xi}\sum_{a\leq b}(g_{out})_{ab}^2$. After a long but simple algebraic calculation, it is possible to show that the denoiser function reduces to simply:

$$
\begin{aligned}
V(g_{out})_{ij} =& = \tau_{ij}\sqrt{V}\xi_{TT} - \frac{\tau_{ij}}{T^2}\sum_{a\leq b}^{T^2}\tau_{ab}\sqrt{V}\xi_{ab} + \frac{\tau_{ij}}{\tau_{iT}}\sqrt{V}\xi_{iT} \\
&\quad - \frac{\tau_{ij}}{\tau_{TT}}\sqrt{V}\xi_{TT} + \delta(j<T)\sqrt{V}\xi_{ij} - \delta(j<T)\frac{\tau_{ij}}{\tau_{iT}}\sqrt{V}\xi_{iT} \\
&= -\frac{\tau_{ij}}{T^2}\sum_{a\leq b}^{T^2}\tau_{ab}\sqrt{V}\xi_{ab} + \sqrt{V}\xi_{iT}\delta(j=T) + \delta(j<T)\sqrt{V}\xi_{ij} \\
&= -\frac{\tau_{ij}}{T^2}\sum_{a\leq b}^{T^2}\tau_{ab}\sqrt{V}\xi_{ab} + \sqrt{V}\xi_{ij}
\end{aligned}
\tag{160}
$$

which finally gives:

$$
\begin{aligned}
\mathbb{E}_{\eta,\xi}V\sum_{i\leq j}^{T}(g_{out})_{ij}^2 =& = \frac{T(T+1)}{2} - \frac{2}{T^2}\sum_{i\leq j}^{T}\sum_{a\leq b}^{T}\tau_{ij}\tau_{ab}\mathbb{E}\xi_{ij}\xi_{ab} + \frac{1}{T^4}\sum_{i\leq j}^{T}\sum_{a\leq b}^{T}\sum_{c\leq d}^{T}\tau_{ij}^2\tau_{ab}\tau_{cd}\mathbb{E}\xi_{ab}\xi_{cd} = \\
&= \frac{T(T+1)}{2} - \frac{2}{T^2}\sum_{i\leq j}^{T}\tau_{ij}^2 + \frac{1}{T^4}\sum_{i\leq j}^{T}\sum_{a\leq b}^{T}\tau_{ij}^2\tau_{ab}^2 = \\
&= \frac{T^2+T-2}{2}
\end{aligned}
\tag{161}
$$

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:

$$
\hat{q} = 4\alpha\mathbb{E}_{\eta,\xi}\sum_{i\leq j}^{T}(g_{out})_{ij}^2 = \frac{4\alpha(T^2+T-2)}{2V} = \frac{\alpha(T^2+T-2)}{Q-q} \tag{162}
$$

which is the result presented in the main text in Eq.(8). 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 (145) 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.

E.5. Hardmax output channel for 2 tokens

We now discuss the hardmax output channel case, in the special case of $T = 2$ tokens. Following Eq.(6) in the main text, we need to compute the quantity:

$$
\mathcal{Z}_{out}(y,\omega,V) = \int\prod_{a\leq b}dh_{ab}\,\frac{1}{\sqrt{2\pi V_{ab}}}e^{-\frac{(h_{ab}-\omega_{ab})^2}{2V_{ab}}}\prod_{a\leq b}\delta(y_{ab} - \sigma_{\text{hard}}(\{\frac{1}{\sqrt{2-\delta_{ab}}}h_{ab}\}_a)_b. \tag{163}
$$

with :

$$
\sigma_{\text{hard}}(z_1\dots z_T)_i = \delta(i = \arg\max_j x_j) \tag{164}
$$

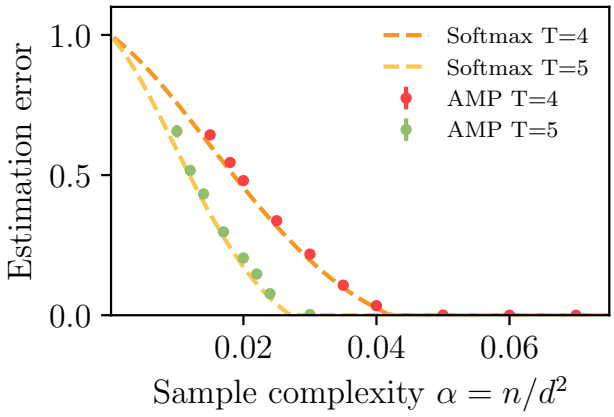

Figure 4: Comparison between the fixed points solutions of the state equations for a softmax output channel in Eq. (8) and Eq. (162) for $T = 4, 5$ tokens. We compare the theoretical solution with their corresponding AMP algorithm run over 16 different realizations and with $d = 120$. The error bars in the AMP dots are computed with respect to the mean value.

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

$$y_{11} = \Theta(h_{11} - h_{12}) \tag{165}$$

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

$$\Theta(u) = \begin{cases} 1, & u > 0, \\ 0, & u < 0. \end{cases} \tag{166}$$

For the computation of the quantity in Eq.(163), it is convenient to make a change of variables by introducing the differences:

$$u = h_{11} - \frac{h_{12}}{\sqrt{2}}, \quad v = h_{22} - \frac{h_{12}}{\sqrt{2}}. \tag{167}$$

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

$$\begin{aligned} I_{\text{out}}(\eta, y) = \int dh_{12}\, \mathcal{N}(h_{12}; \omega_{12}, V_{12}) \left\{ \int_{u \in \mathcal{R}(y_{11})} du\, \mathcal{N}(u + \frac{h_{12}}{\sqrt{2}}; \omega_{11}, V_{11}) \right\} \\ \times \left\{ \int_{v \in \mathcal{R}(y_{22})} dv\, \mathcal{N}(v + \frac{h_{12}}{\sqrt{2}}; \omega_{22}, V_{22}) \right\}, \end{aligned} \tag{168}$$

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

$$\mathcal{N}(u + h_{12}; \omega_{11}, V_{11}) = \mathcal{N}(u;\, \omega_{11} - \frac{h_{12}}{\sqrt{2}}, V_{11}), \tag{169}$$

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

$$\mathcal{Z}_{out}(y, \omega, V) = \int dh_{12}\, \mathcal{N}(h_{12}; \omega_{12}, V_{12})\; F_{11}(\frac{h_{12}}{\sqrt{2}}; \omega)\; F_{22}(\frac{h_{12}}{\sqrt{2}}; \omega)\,, \tag{170}$$

with

$$F_{11}(\frac{h_{12}}{\sqrt{2}}; \omega) = \int_{u \in \mathcal{R}(y_{11})} du\, \mathcal{N}(u;\, \omega_{11} - \frac{h_{12}}{\sqrt{2}}, V_{11}) = \Phi\Big(s_{11} \frac{\omega_{11} - \frac{h_{12}}{\sqrt{2}}}{\sqrt{V_{11}}}\Big), \tag{171}$$

$$F_{22}(\frac{h_{12}}{\sqrt{2}}; \omega) = \int_{v \in \mathcal{R}(y_{22})} dv\, \mathcal{N}(v;\, \omega_{22} - \frac{h_{12}}{\sqrt{2}}, V_{22}) = \Phi\Big(s_{22} \frac{\omega_{22} - \frac{h_{12}}{\sqrt{2}}}{\sqrt{V_{22}}}\Big), \tag{172}$$

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:

$$\mathcal{Z}_{\rm out}(y, \omega, V) = \int_{-\infty}^{+\infty} dh_{12}\, \mathcal{N}(h_{12}; \omega_{12}, V_{12})\; \Phi\Big(s_{11} \frac{\omega_{11} - \frac{h_{12}}{\sqrt{2}}}{\sqrt{V_{11}}}\Big)\, \Phi\Big(s_{22} \frac{\omega_{22} - \frac{h_{12}}{\sqrt{2}}}{\sqrt{V_{22}}}\Big) \tag{173}$$

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,

$$\mathcal{Z}_{\rm out}(y, \omega, V) = \mathbb{E}_Z\Big[\Phi\big(u_1 - \lambda_1 Z\big)\, \Phi\big(u_2 - \lambda_2 Z\big)\Big], \tag{174}$$

where

$$u_1 = s_{11}\, \frac{\sqrt{2}\omega_{11} - \omega_{12}}{\sqrt{2V_{11}}}, \quad u_2 = s_{22}\, \frac{\sqrt{2}\omega_{22} - \omega_{12}}{\sqrt{2V_{22}}}, \tag{175}$$

$$\lambda_1 = s_{11} \sqrt{\frac{V_{12}}{2V_{11}}}, \qquad \lambda_2 = s_{22} \sqrt{\frac{V_{12}}{2V_{22}}}. \tag{176}$$

A classical identity for jointly Gaussian variables gives

$$\mathbb{E}_Z\big[\Phi(a + bZ)\, \Phi(c + dZ)\big] = \Phi_2\Big(\tfrac{a}{\sqrt{1+b^2}},\, \tfrac{c}{\sqrt{1+d^2}};\, \tfrac{bd}{\sqrt{(1+b^2)(1+d^2)}}\Big). \tag{177}$$

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

$$\mathcal{Z}_{out}(y, \omega, V) = \Phi_2\Big(\kappa_1,\, \kappa_2;\, c\Big) \tag{178}$$

with the compact parameters

$$\kappa_1 = s_{11}\, \frac{\sqrt{2}\omega_{11} - \omega_{12}}{\sqrt{2V_{11} + V_{12}}}, \qquad \kappa_2 = s_{22}\, \frac{\sqrt{2}\omega_{22} - \omega_{12}}{\sqrt{2V_{22} + V_{12}}}, \tag{179}$$

$$c = s_{11}s_{22}\,\frac{V_{12}}{\sqrt{(2V_{11}+V_{12})(2V_{22}+V_{12})}} = s_{11}s_{22}\,\frac{1}{3} \quad (V_{11}=V_{22}=V_{12}). \tag{180}$$

We can hence compute denoising function:

$$(g_{out}(y,\omega,V))_{ab} = \frac{\partial}{\partial \omega_{ab}} \ln \mathcal{Z}_{out}(y,\omega,V). \tag{181}$$

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

$$\frac{\partial}{\partial \omega_{11}}\Phi_2(\kappa_1,\kappa_2;\rho_{12}) = \frac{\partial \kappa_1}{\partial \omega_{11}}\,\phi_2(\kappa_1,\kappa_2;\rho_{12}), \quad \frac{\partial \kappa_1}{\partial \omega_{11}} = \frac{\sqrt{2}s_{11}}{\sqrt{2V_{11}+V_{12}}}. \tag{182}$$

The four independent derivatives are therefore

$$g_{out}(y,\omega,V)_{11} = \frac{\sqrt{2}s_{11}}{\sqrt{2V_{11}+V_{12}}}\,\frac{\phi_2(\kappa_1,\kappa_2;\rho_{12})}{\Phi_2(\kappa_1,\kappa_2;\rho_{12})} \tag{183}$$

$$g_{out}(y,\omega,V)_{22} = \frac{\sqrt{2}s_{22}}{\sqrt{2V_{22}+V_{12}}}\,\frac{\phi_2(\kappa_2,\kappa_1;\rho_{12})}{\Phi_2(\kappa_1,\kappa_2;\rho_{12})} \tag{184}$$

$$g_{out}(y,\omega,V)_{12} = -\left(\frac{s_{11}}{\sqrt{2V_{11}+V_{12}}}+\frac{s_{22}}{\sqrt{2V_{22}+V_{12}}}\right)\frac{\phi_2(\kappa_1,\kappa_2;\rho_{12})}{\Phi_2(\kappa_1,\kappa_2;\rho_{12})} \tag{185}$$

This expression can be compactly rewritten as:

$$g_{out}(y,\omega,V)_{ab} = \frac{1}{\sqrt{6(Q-q)}}\frac{\phi(k_1,k_2,c)}{\Phi(k_1,k_2,c)}\begin{pmatrix}\sqrt{2}s_1 & -(s_1+s_2)\\ -(s_1+s_2) & \sqrt{2}s_2\end{pmatrix}_{ab}, \tag{186}$$

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 B). Moreover, $s_a = 2(y_{aa}-1)$, $k_a = s_a(\sqrt{2}\omega_{aa}-\omega_{12})/\sqrt{6(Q-q)}$, $c = s_1s_2/3$ and $\omega_{ab} = \sqrt{2q}\,\eta_{ab}$.

E.6. Generalization error and sequence-to-sequence version of the model

In this section we draw some consideration on the generalization error of the model, in the setting of a self-attention layer as in (19) and its sequence-to-sequence version as in (20).

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:

$$E_{est} = \frac{1}{d}\|S^* - \hat{S}\|_F^2 = Q - q \tag{187}$$

Regarding the generalization error, we may instead want to compute and plot a different quantity, namely:

$$\mathcal{E}_{gen}(\hat{y}) = \mathbb{E}_{\mathcal{D},S^*}\mathbb{E}_{y_{new},\boldsymbol{x}_{new}}||\hat{y}(\boldsymbol{x}_{new},\mathcal{D}) - y_{new}||_F^2\,, \tag{188}$$

with:

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

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 $h_{ab} = \frac{x_a S x_b^\top - \delta_{ab} \operatorname{Tr} S}{\sqrt{d}}$ and get the expression:

$$\mathcal{E}_{\rm gen} = \sum_{a,b} \mathbb{E}_{x_{ab}} \int dh_{ab} d\hat{h}_{ab} \|\sigma(\frac{h_{ab}}{\sqrt{2-\delta_{ab}}}) - \sigma(\frac{\hat{h}_{ab}}{\sqrt{2-\delta_{ab}}})\|^2 \delta(h_{ab} - \frac{x_a S x_b^\top - \delta_{ab} \operatorname{Tr} S}{\sqrt{d}}) \delta(\hat{h}_{ab} - \frac{x_a \hat{S} x_b^\top - \delta_{ab} \operatorname{Tr} \hat{S}}{\sqrt{d}}) \tag{189}$$

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

$$\mathbb{E}_{x_{ab}} \delta(h_{ab} - \frac{x_a S x_b^\top - \delta_{ab} \operatorname{Tr} S}{\sqrt{d}}) \delta(\hat{h}_{ab} - \frac{x_a \hat{S} x_b^\top - \delta_{ab} \operatorname{Tr} \hat{S}}{\sqrt{d}}) = \mathcal{N}\Big(\begin{pmatrix} h_{ab} \\ \hat{h}_{ab} \end{pmatrix}, \begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} q & q \\ q & Q^* \end{pmatrix} 2\Big) = P(h_{ab}, \hat{h}_{ab}) \tag{190}$$

Then, the overall generalization error is given by

$$\mathcal{E}_{\rm gen} = \sum_{a,b=1}^{T} \mathbb{E}_{(h_{ab},\hat{h}_{ab}) \sim P(h_{ab},\hat{h}_{ab})} \Big[\sigma(\{\frac{h_{ab}}{\sqrt{2-\delta_{ab}}}\}_{ab}) - \sigma(\{\frac{\hat{h}_{ab}}{\sqrt{2-\delta_{ab}}}\}_{ab})\Big]^2 \tag{191}$$

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 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 $x^\mu_{ai} \sim \mathcal{N}(0,1)$, we use the fact that

$$\sum_{i=1}^{d} x_{t'i} x_{t''i} \approx \delta_{tt'} \tag{192}$$

with high probability when d is large. Hence:

$$||(A - \hat{A})x||_F^2 = \sum_{t,i} [\sum_{t'} (A_{tt'} - \hat{A}_{tt'}) x_{t'i}]^2 = \sum_{t,i} \sum_{t',t''} (A_{tt'} - \hat{A}_{tt})(A_{tt''} - \hat{A}_{tt''}) x_{t',i} x_{t'',i} \tag{193}$$

but using the concentration property of x we finally get:

$$||(A - \hat{A})x||_F^2 = \sum_{t,i} \sum_{t',t''} (A_{tt'} - \hat{A}_{tt'})(A_{tt''} - \hat{A}_{tt''}) x_{t',i} x_{t'',i} = \sum_{t,t'} (A_{tt''} - \hat{A}_{tt'})^2 = ||A - \hat{A}||_F^2 \tag{194}$$

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.

E.7. Gradient descent and the details on the numerical experiments

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

$$\mathcal{L}(W) = \sum_{\mu=1}^{n} \left(y_\mu - \sigma_\beta \left(\frac{\boldsymbol{x}_a W W^\top \boldsymbol{x}_b^\top - \delta_{ab} \operatorname{Tr} W W^\top}{\sqrt{r}\, d} \right) \right)^2 , \tag{195}$$

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. (8). In the latter case, finally, we compute the expectation in Eq. (128) with Monte-Carlo over $n_{samples} = 20000$ 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.

Appendix F. Small/Large width limit of the prior channel

We recall that the state equations are of the form

$$\begin{aligned} Q - q &= \frac{1}{\hat{q}} - \frac{4\pi^2}{3\hat{q}^2}\int dx\, \mu_{1/\hat{q}}(x)^3 \\ \hat{q} &= 2\alpha F(Q-q, q) \end{aligned} \tag{196}$$

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

F.1. Small width limit

We follow [30, 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.

$$\nu(y) = \rho^{-1/2}\mu_{1/\hat{q}}(\rho^{-1/2}y)\,. \tag{197}$$

Notice that this is precisely the ν defined in [30, Eq. 56]. Then we have

$$\begin{aligned} Q - q &= \frac{1}{\hat{q}} - \frac{4\pi^2}{3\hat{q}^2}\int dx\, \mu_{1/\hat{q}}(x)^3 \\ &= \frac{t}{\rho} - \frac{4\pi^2 t^2}{3\rho^2}\rho\int dy\, [\rho^{-1/2}\mu_{1/\hat{q}}(\rho^{-1/2}y)]^3 \\ &= \frac{t}{\rho} - \frac{4\pi^2 t^2}{3\rho}\int dy\, \nu(y)^3 \\ &= \frac{t}{\rho}\left[1 - \frac{4\pi^2 t}{3}\int dy\, \nu(y)^3\right] \\ &\approx \begin{cases} t(2-t) & \text{if} \quad t \le 1 \\ 1 & \text{if} \quad t > 1 \end{cases} \end{aligned} \tag{198}$$

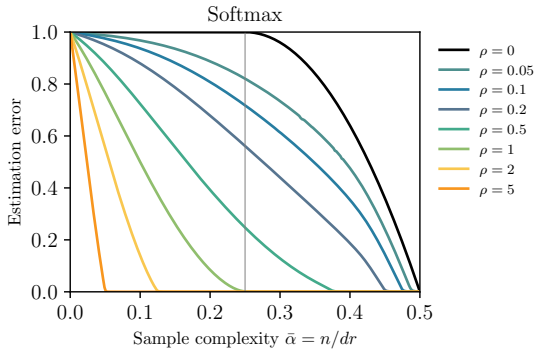

Figure 5: Low width limit of the self-attention model for $L = 1$ layer and $T = 2$ tokens in Eq.(5). We rescale the sample ratio as $\bar{\alpha} = n/dr$ and we plot several values of the width ratio $\rho = r/d$. We correctly predict the weak recovery threshold in Eq.(200).

where we used [30, Eq. 57 and following] to take the limit of small κ at leading order. Thus, the equations can be recast to

$$Q - q = \begin{cases} t(2-t) & \text{if} \quad t \leq 1 \\ 1 & \text{if} \quad t > 1 \end{cases} \qquad t = \frac{1}{2\bar{\alpha} F(Q-q)} \,. \tag{199}$$

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

$$\bar{\alpha} < \frac{1}{2F(1)} \tag{200}$$

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

In Fig.5 we plot the low width behavior of the self-attention model for $L = 1$ layer and $T = 2$ tokens in Eq.(5), for which we recover the simple output channel state equation in Eq.(162), thus giving :

$$F(Q-q, q) = \frac{T^2 + T - 2}{2(Q-q)} \tag{201}$$

F.2. Large width limit for softmax

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

$$\begin{aligned} Q - q &= \frac{1}{\hat{q}} - \frac{4\pi^2}{3\hat{q}^2} \int dx\, \mu_{1/\hat{q}}(x)^3 \\ &= \frac{1}{\hat{q}} \left[1 - \frac{4\pi^2}{3\hat{q}\rho} \int dy\, [\sqrt{\rho}\mu_{1/\hat{q}}(\sqrt{\rho}y)]^3 \right] \\ &= \frac{1}{\hat{q}} \left[1 - \frac{4\pi^2}{3\hat{q}\rho} \int dy\, \mu_{1/\rho\hat{q}}(y)^3 \right] \\ &\approx \frac{1}{\hat{q}} \left[1 - \frac{1}{1+\hat{q}} \right] \\ &\approx \frac{1}{1+\hat{q}}\,, \end{aligned} \tag{202}$$

where we used [30, Section E.2]. Thus, we get the equation

$$\hat{q} = \frac{\alpha(T^2 - T + 2)}{Q - q} = \alpha(T^2 - T + 2)(1 + \hat{q}) \tag{203}$$

so we get

$$\hat{q} = \frac{\alpha(T^2 - T + 2)}{1 - \alpha(T^2 - T + 2)} \tag{204}$$

which gives the large ρ result:

$$\mathrm{M}MSE = \frac{1}{1+\hat{q}} = 1 - \alpha(T^2 - T + 2)\,. \tag{205}$$