

SINGLE-HEAD ATTENTION IN HIGH DIMENSIONS: A THEORY OF GENERALIZATION, WEIGHTS SPECTRA, AND SCALING LAWS

Anonymous authors

Paper under double-blind review

ABSTRACT

Trained attention layers exhibit striking and reproducible spectral structure of the weights, including low-rank collapse, bulk deformation, and isolated spectral outliers, yet the origin of these phenomena and their implications for generalization remain poorly understood. We study empirical risk minimization in a single-head tied-attention layer trained on synthetic high-dimensional sequence tasks generated from the attention-indexed model. Using tools from random matrix theory, spin-glass theory, and approximate message passing, we obtain an exact high-dimensional characterization of training and test error, interpolation and recovery thresholds, and the spectrum of the key and query matrices. Our theory predicts the full singular-value distribution of the trained query–key map—including low-rank structure and isolated spectral outliers—in qualitative agreement with observations in more realistic transformers. Finally, for targets with power-law spectra, we show that learning proceeds through sequential spectral recovery, leading to the emergence of power-law scaling laws.

1 INTRODUCTION

Attention mechanisms are the backbone of modern ML (Vaswani et al., 2017), yet their theoretical foundations remain elusive. Notably, trained attention layers exhibit non-random spectral patterns (Martin et al., 2021) and “stepwise” emergence during training (Wei et al., 2022). This motivates our central inquiry: *Can one propose and analytically solve a simplified model of attention in the high-dimensional regime that allows to explore the interplay between spectral properties and generalization?*

Our main contributions are

- (i) **Asymptotics and inductive bias.** We derive exact formulas for training/test errors and recovery thresholds. We show that gradient-based algorithms converge to these solutions and demonstrate that factorized training (query/key matrices) provides a superior inductive bias over direct training of their product.
- (ii) **Spectrum–generalization Link.** We provide an analytic characterization of the learned weight spectrum—predicting structured bulks, low-rank collapse, and outliers. This theory quantitatively links the learned spectrum to generalization performance, aligning with empirical transformer observations.
- (iii) **Emergence and scaling Laws.** In regimes with power-law target spectra, we show that spectral modes are recovered sequentially. This isolates how target structure translates into sharp emergence phenomena and observable power-law scaling in attention layers.

2 SETTINGS

Task and architecture. We consider a sequence-to-sequence (seq2seq) and a sequence-to-label (seq2lab) supervised learning task. Our analyses applies to both these cases. In the seq2seq task we consider a dataset $\mathcal{D} = \{\mathbf{x}_{\text{in}}^\mu, \mathbf{x}_{\text{out}}^\mu\}_{\mu=1}^n$ where both the input and output data $\mathbf{x}_{\text{in}}^\mu, \mathbf{x}_{\text{out}}^\mu \in \mathbb{R}^{T \times d}$ are sequences of T tokens, each given through an embedding vector $\mathbf{x}_a^\mu \in \mathbb{R}^d$ for $1 \leq a \leq T$. In the seq2lab task, instead we consider a dataset $\mathcal{D} = \{\mathbf{x}_{\text{in}}^\mu, y^\mu\}_{\mu=1}^n$ where the output data $y^\mu \in \mathbb{R}^{T \times T}$ are matrices of $T \times T$ pair-wise tokens comparisons.

In both cases, we aim at learning the input-output relationship through a parametrized function $\hat{f}(\mathbf{x}_{\text{in}}; W)$ of the form (respectively for the two tasks)

$$\begin{cases} \hat{f}_{\text{sq}}(\mathbf{x}_{\text{in}}; W) = A_W(\mathbf{x}_{\text{in}})\mathbf{x}_{\text{in}} \\ \hat{f}_{\text{lb}}(\mathbf{x}_{\text{in}}; W) = A_W(\mathbf{x}_{\text{in}}) \end{cases}, \quad \text{where} \quad A_W(\mathbf{x}_{\text{in}}) = \sigma_\beta \left(\frac{\mathbf{x}_{\text{in}} W W^T \mathbf{x}_{\text{in}}^T - \mathbb{E}_{\text{tr}}[\mathbf{x} W W^T \mathbf{x}^T]}{d\sqrt{p}} \right) \quad (1)$$

with weights $W \in \mathbb{R}^{d \times p}$. σ_β is the row-wise softmax activation at inverse temperature $\beta > 0$, and $A_W(\mathbf{x})$ is the attention matrix with tied key and query matrices. Thus, in the seq2seq task, \hat{f} is a tied attention layer with identity value matrix. In the seq2lab case, instead, \hat{f} outputs directly the attention matrix. Finally, \mathbb{E}_{tr} is the empirical average over \mathbf{x} in the training set, and the corresponding term plays the role of a batch-centering, ensuring a mean-zero input to the activation.

We learn W by *empirical risk minimization* of the square loss with ℓ_2 regularization (or equivalently, weight decay) that is commonly used in practice in large language models, i.e. $\hat{W} = \arg \min_W \mathcal{L}(W) = \arg \min_W [\mathcal{L}^{\text{data}}(W) + \mathcal{R}(W)]$ where respectively for the two tasks

$$\begin{cases} \mathcal{L}_{\text{sq}}^{\text{data}}(W) := \frac{1}{d} \sum_{\mu=1}^n \|\mathbf{x}_{\text{out}}^\mu - \hat{f}_{\text{sq}}(\mathbf{x}_{\text{in}}^\mu; W)\|_F^2, \\ \mathcal{L}_{\text{lb}}^{\text{data}}(W) := \sum_{\mu=1}^n \|y^\mu - \hat{f}_{\text{lb}}(\mathbf{x}_{\text{in}}^\mu; W)\|_F^2, \end{cases} \quad (2)$$

and $\mathcal{R}(W) = \lambda \|W\|_F^2$. We measure the performance of the learned function \hat{f} through the test errors

$$e_{\text{test}}(\hat{f}) = \frac{1}{d} \mathbb{E}_{\mathbf{x}_{\text{in}}, \mathbf{x}_{\text{out}}} \|\mathbf{x}_{\text{out}} - \hat{f}(\mathbf{x}_{\text{in}})\|_F^2 \quad \text{or} \quad e_{\text{test}}(\hat{f}) = \mathbb{E}_{\mathbf{x}_{\text{in}}, y} \|y - \hat{f}(\mathbf{x}_{\text{in}})\|_F^2, \quad (3)$$

where \mathbb{E} stands for an average over an appropriate test set. We remark that our analytical framework can be extended to a larger class of losses and regularization, see Section 3.

Input data model. To allow for analytical tractability, we assume that the input sequences $\mathbf{x}_{\text{in}}^\mu$ are Gaussian, i.e., that each token is independently given by $\mathbf{x}_{\text{in},a}^\mu \sim \mathcal{N}(0, \mathbb{I}_d)$ for $1 \leq a \leq T$. In this case, it can be shown that the centering term in Eq. (1) concentrates to $\mathbb{I}_T \text{Tr}(W W^T)$ in the regime $n \gg 1$. We remark that the high-dimensional setting we consider exhibits universality way beyond Gaussian data, in the same spirit as in Assumption 2.2 of Xu et al. (2025) and as proven by a number of recent works, e.g. Montanari & Saeed (2022); Dudeja et al. (2023); Dandi et al. (2023); Lu & Yau (2025).

Target function model. Following the successes of the single-index models, we consider a target function that lies within the expressivity class of the architecture in Eq. (1), and restrict the class of (possibly noisy) target functions to (respectively for the seq2seq and seq2lab tasks)

$$\mathbf{x}_{\text{out}}^\mu = \sigma_{\beta_0}(R(\mathbf{x}_{\text{in}})) \mathbf{x}_{\text{in}} \quad \text{and} \quad y^\mu = \sigma_{\beta_0}(R(\mathbf{x}_{\text{in}})), \quad (4)$$

where $R(\mathbf{x}_{\text{in}}) \in \mathbb{R}^{T \times T}$ is a centered pre-activation matrix, and β_0 a softmax temperature possibly different from the learner's one β . We consider this model as we want to focus on the learnability of the token-to-token correlations exhibited by the output sequences. Concretely, we choose:

$$R(\mathbf{x}_{\text{in}})_{ab} = \frac{\mathbf{x}_{\text{in},a}^T S_0 \mathbf{x}_{\text{in},b} - \delta_{ab} \text{Tr}(S_0)}{\sqrt{d}} + \sqrt{\frac{\Delta}{2 - \delta_{ab}}} \xi^\mu, \quad (5)$$

where S_0 is the target function weight matrix of rank p_0 . We denoted $\xi^\mu \in \mathbb{R}^{T \times T}$ to be a symmetric standard Gaussian noise $\xi_{ab} = \xi_{ba} \sim N(0, 1)$ for all $1 \leq a \leq b \leq T$, and the normalization of the variance ensures that the signal-to-noise ratio is uniform across tokens. The intuition behind Eq. (5) is that the learning model is structurally limited to expressing models of attention that are bilinear in the input sequences: for this reason, we model all higher-order dependencies as (Gaussian) noise ξ . For the test set, we consider samples distributed with the same distribution as the training set.

3 ASYMPTOTICS OF TRAIN AND TEST ERROR

Our main technical result is the characterization of the properties of the global minima of the empirical loss of Eq. (2) in the high-dimensional limit where the token embedding dimension $d \rightarrow \infty$ with quadratic number of samples and proportional ranks, i.e., the joint limit $d, n, p \rightarrow +\infty$ with

$$\alpha := n/d^2, \quad \kappa := p/d, \quad \kappa_0 := p_0/d \quad (6)$$

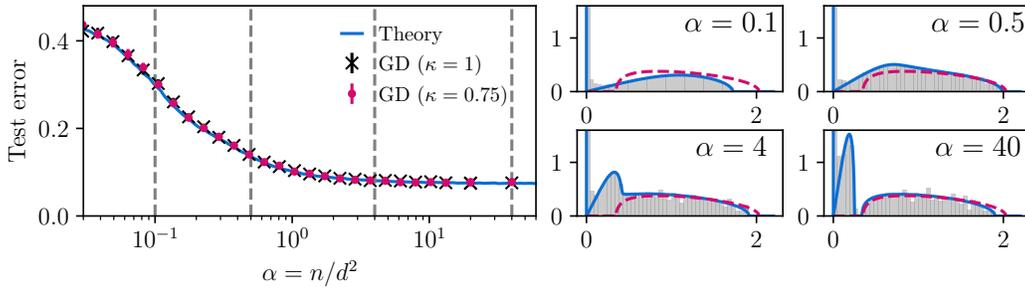


Figure 1: (Left) Test error of the ERM estimator Eq. (2) (Claim B.1, Eq. (14)) compared with Adam simulations at $d = 100$ (64 instances, error bars = standard deviation) as a function of the number of samples $\alpha = n/d^2$, where we use $\kappa = 0.75, 1$ (model width) and parameters $\lambda = 0.01, \Delta = 0.5, T = 2, \beta = \beta_0 = 1$ for the MP target (Section 3, $\kappa_0 = 0.5$). Theory and simulations agree. (Right) Singular value spectrum of the trained weights from theory (blue, Claim 4.1, Eq. (7)) vs. Adam simulations (grey histograms) at $\alpha = 0.05, 0.5, 4, 40$ ($d = 200, 64$ runs, 2000 samples in the test set). The asymptotic spectrum of the target is shown in red dashed. The theory also captures the delta peak at zero. For large α , the spectrum splits into two bulks.

are $\mathcal{O}(1)$ ratios. We further assume that in the same limit the empirical spectral density of S_0 (which may be either random or deterministic) converges to a limiting distribution μ_0 with a finite first and second moment. The length of the sequence $T = \mathcal{O}(1)$ is considered finite in our work in line with other works on sequence multi-index or attention-indexed models (Cui et al., 2024; Troiani et al., 2025; Boncoraglio et al., 2025). Our first result is Claim B.1, which predicts the train and test error of the global minimum of (2).

To illustrate our results, we consider the *Marchenko-Pastur (MP) target* case, where the weights S_0 are such that the limiting distribution satisfies $\mu_0(x) = \sqrt{\kappa_0} \mu_{\text{M.P.}}(\sqrt{\kappa_0}x)$, where $\mu_{\text{M.P.}}$ is the Marchenko-Pastur distribution (Marchenko & Pastur, 1967) with normalized rank $0 < \kappa_0 < 1$, and $Q_0 = 1 + \kappa_0$. For example, this is the case in which the target attention matrix has the form $S_0 = W_0 W_0^T / \sqrt{p_0 d}$, where $W_0 \in \mathbb{R}^{d \times m_0}$ has i.i.d. components extracted from a distribution with zero mean and unit variance, and $\kappa_0 = p_0/d = \text{rank}(S_0)/d$. We believe that the overall phenomenology we showcase here is valid qualitatively for generic S_0 with $\text{rank}(S_0) = \kappa_0 d$ and positive eigenvalues bounded away from the origin. Additionally, we consider the case of matched softmax temperatures $\beta = \beta_0 = 1$ (noticing that a mismatch in temperatures can always be reabsorbed in a rescaling of the regularization). In Figure 1, we plot the test error predicted by our theory as a function of the sample ratio $\alpha = n/d^2$ for $\Delta = 0.5$, which perfectly matched the experiments of gradient descent.

See Appendix B for the formula of the asymptotics and further discussions.

4 SPECTRUM AND GENERALIZATION

Our second result is the characterization of the spectrum of the ERM estimator (2) and its implication for generalization.

Claim 4.1 (Spectra of ERM). *Under the same assumptions and notations of Claim B.1, let \hat{W} be a global minimizer of ERM in the high-dimensional limit. Then, in distribution,*

$$\frac{1}{\sqrt{pd}} \hat{W}^T \hat{W} \stackrel{d}{=} \eta \text{ReLU}(S_0 + \delta Z - \epsilon I_d), \quad (7)$$

where $Z \sim \text{GOE}(d)$, the ReLU is applied on the spectrum, and we defined $\eta = \hat{m}^* / \hat{\Sigma}^*$, $\delta = \sqrt{\hat{q}^*} / \hat{m}^*$, $\epsilon = 2\tilde{\lambda} / \hat{m}^*$. In particular, the empirical spectral density of $\hat{W}^T \hat{W} / \sqrt{pd}$ converges for $d \rightarrow \infty$ to a deterministic limit fully determined by (η, δ, ϵ) and the spectral law of S_0 .

A key consequence of Claim B.1 is that both the learned spectrum and the test error are governed by the scalar parameters (η, δ, ϵ) arising from the minimization of (9). This yields an explicit link between generalization and spectral structure: δ controls the effective noise level due to finite training set size, while ϵ controls the regularization strength, setting small eigenvalues to zero and inducing low effective rank and spectral outliers.

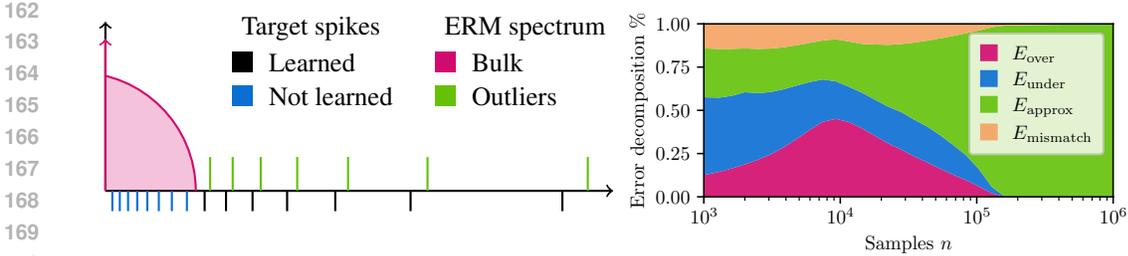


Figure 2: Qualitative representation of the spectral density and error decomposition of Section 4. (Top) Spectrum of the target (below the horizontal axis) and of the ERM (above the axis). The ERM spectrum is composed of outliers (learned target features) and a noise bulk (magenta). (Bottom) Error decomposition for power-law target $\gamma = 0.75$, $d = 200$, $\Delta = 0.5$, $T = 2$, $\lambda = 1/d$ (see Appendix J), expressed as fraction of the total error (shown in Figure 4).

Spectral law for low-rank targets. We start by showcasing the resulting spectral behavior in the case of MP target in the quadratic sample regime (6). Figure 1 compares the spectrum of the learned weights with that of target weights (as well as with numerical simulations with gradient descent). We observe an agreement between theory and experiments, and we clearly see that for this target function learning happens through a sequence of qualitative behaviors (rank-collapse, bleed-out, appearance of outliers) aligning more and more with the target distribution as the number of samples increases. This is in qualitative agreement with spectra observed experimentally in real architectures (Martin et al., 2021; Martin & Mahoney, 2021; Staats et al., 2024).

From spectrum to generalization. We make the connection between spectral properties and generalization explicit by studying how the test error approaches its limiting values as the number of samples increases. To do so, we need to restrict our attention to targets with eigenvalues $\{s_i\}_{i=1}^p$, which we interpret as a set of learnable features that decay faster than $i^{-1/2}$ (including sub-linear rank targets $p_0 = o(d)$).

Assuming further that the learned weights are well aligned with the target (small test error), the learned spectrum admits a simple and interpretable form (see sketch in Figure 2). It consists of a semicircular bulk, originating from noise and finite-sample effects, together with a finite number of outlying eigenvalues that are in one-to-one correspondence with the recovered target eigenvalues $\{s_i\}$, up to a regularization-dependent rescaling and thresholding: spectral outliers can be interpreted as learned features, while the bulk captures spurious directions induced by noise. Up to the leading order, the test error at the empirical minimum can then be written

$$e_{\text{test}}(\hat{W}) = e_0 + e_1(n, d, \lambda) \quad \text{with} \quad e_1 = c_1(E_{\text{over}} + E_{\text{under}} + E_{\text{approx}}) + c_2 E_{\text{mismatch}}, \quad (8)$$

where e_0 is a saturation value independent of n, d, λ , and the decay term e_1 depends explicitly on the learned spectrum. c_1, c_2 depend only on the activations and the label noise. This excess error decomposes into four contributions:

- E_{over} is an *overfitting term*, controlled by the noisy bulk of the learned spectrum and arising from fitting spurious directions induced by finite samples;
- E_{under} is an *underfitting term*, corresponding to target directions that are not recovered due to noise or regularization, consisting of the eigenvalues still trapped in the bulk.
- E_{approx} is an *approximation term*, measuring residual error along recovered spectral outliers, i.e. learned features;
- Finally, E_{mismatch} is a *mismatch term*, which appears when the saturation value e_0 does not correspond to a local minimum of the generalization error.

Explicit expressions for these terms are given in Appendix H. Taken together, this decomposition shows that spectral outliers correspond to genuinely learned target features, and that increasing the number of recovered outliers systematically improves generalization. Conversely, the bulk of small eigenvalues represents pure noise arising from finite samples: the learned spectrum provides a compact and interpretable summary of both the inductive bias induced by the regularization and its consequences for generalization.

As an application of the error decomposition, we can derive exact learning curves and universal exponents for power-law targets. See Appendix C.

216
217
218
219
220
221
222
223
224
225
226
227
228
229
230
231
232
233
234
235
236
237
238
239
240
241
242
243
244
245
246
247
248
249
250
251
252
253
254
255
256
257
258
259
260
261
262
263
264
265
266
267
268
269

REFERENCES

- 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*, pp. 4782–4887. PMLR, 2022.
- 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*, pp. 1199–1227. PMLR, 2023.
- Luca Arnaboldi, Bruno Loureiro, Ludovic Stephan, Florent Krzakala, and Lenka Zdeborová. Asymptotics of SGD in sequence-single index models and single-layer attention networks. *NeurIPS 2025*, 2025. arXiv preprint arXiv:2506.02651.
- Sanjeev Arora and Anirudh Goyal. A theory for emergence of complex skills in language models. *arXiv preprint arXiv:2307.15936*, 2023.
- Sanjeev Arora, Nadav Cohen, Wei Hu, and Yuping Luo. Implicit regularization in deep matrix factorization. *Advances in neural information processing systems*, 32, 2019.
- Alexander Atanasov, Jacob A Zavatone-Veth, and Cengiz Pehlevan. Scaling and renormalization in high-dimensional regression. *arXiv preprint arXiv:2405.00592*, 2024.
- 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.
- Yasaman Bahri, Ethan Dyer, Jared Kaplan, Jaehoon Lee, and Utkarsh Sharma. Explaining neural scaling laws. *Proceedings of the National Academy of Sciences*, 121(27):e2311878121, 2024.
- 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.
- G erard Ben Arous, Murat A Erdogdu, N Mert Vural, and Denny Wu. Learning quadratic neural networks in high dimensions: Sgd dynamics and scaling laws. *arXiv preprint arXiv:2508.03688*, 2025. *NeurIPS 2025*.
- Rapha el Berthier, Andrea Montanari, and Phan-Minh Nguyen. State evolution for approximate message passing with non-separable functions. *Information and Inference: A Journal of the IMA*, 9(1):33–79, March 2020. ISSN 2049-8764, 2049-8772.
- Rapha el Berthier, Andrea Montanari, and Kangjie Zhou. Learning time-scales in two-layers neural networks. *Foundations of Computational Mathematics*, pp. 1–84, 2024.
- Alberto Bietti, Joan Bruna, and Loucas Pillaud-Vivien. On learning Gaussian multi-index models with gradient flow part i: General properties and two-timescale learning. *Communications on Pure and Applied Mathematics*, 2025.
- Erwin Bolthausen. An iterative construction of solutions of the tap equations for the sherrington–kirkpatrick model. *Communications in Mathematical Physics*, 325(1):333–366, 2014.
- Fabrizio Boncoraglio, Emanuele Troiani, Vittorio Erba, and Lenka Zdeborov a. Bayes optimal learning of attention-indexed models. *NeurIPS 2025*, 2025. arXiv:2506.01582.
- Blake Bordelon, Alexander Atanasov, and Cengiz Pehlevan. A dynamical model of neural scaling laws. In *International Conference on Machine Learning*, pp. 4345–4382. PMLR, 2024.
- 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.
- Andrea Caponnetto and Ernesto De Vito. Optimal rates for the regularized least-squares algorithm. *Foundations of Computational Mathematics*, 7(3):331–368, 2007.

- 270 Lucas Clarte, Bruno Loureiro, Florent Krzakala, and Lenka Zdeborova. On double-descent in uncer-
271 tainty quantification in overparametrized models. In *Proceedings of The 26th International Con-*
272 *ference on Artificial Intelligence and Statistics*, volume 206 of *Proceedings of Machine Learning*
273 *Research*, pp. 7089–7125. PMLR, 25–27 Apr 2023.
- 274 Hugo Cui, Bruno Loureiro, Florent Krzakala, and Lenka Zdeborová. Generalization error rates
275 in kernel regression: The crossover from the noiseless to noisy regime. *Advances in Neural*
276 *Information Processing Systems*, 34:10131–10143, 2021.
- 277 Hugo Cui, Freya Behrens, Florent Krzakala, and Lenka Zdeborová. A phase transition between
278 positional and semantic learning in a solvable model of dot-product attention. *Advances in Neural*
279 *Information Processing Systems*, 37:36342–36389, 2024.
- 280 Alex Damian, Eshaan Nichani, Rong Ge, and Jason D Lee. Smoothing the landscape boosts the
281 signal for sgd: Optimal sample complexity for learning single index models. *Advances in Neural*
282 *Information Processing Systems*, 36:752–784, 2023.
- 283 Yatin Dandi, Ludovic Stephan, Florent Krzakala, Bruno Loureiro, and Lenka Zdeborová. Univer-
284 sality laws for Gaussian mixtures in generalized linear models. *Advances in Neural Information*
285 *Processing Systems*, 36:54754–54768, 2023.
- 286 Leonardo Defilippis, Bruno Loureiro, and Theodor Misiakiewicz. Dimension-free deterministic
287 equivalents and scaling laws for random feature regression. *Advances in Neural Information*
288 *Processing Systems*, 37:104630–104693, 2024.
- 289 Leonardo Defilippis, Yizhou Xu, Julius Girardin, Emanuele Troiani, Vittorio Erba, Lenka Zde-
290 borová, Bruno Loureiro, and Florent Krzakala. Scaling laws and spectra of shallow neural net-
291 works in the feature learning regime. *arXiv preprint arXiv:2509.24882*, 2025.
- 292 Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep
293 bidirectional transformers for language understanding. In *Proceedings of the 2019 conference of*
294 *the North American chapter of the association for computational linguistics: human language*
295 *technologies, volume 1 (long and short papers)*, pp. 4171–4186, 2019.
- 296 David L Donoho. Compressed sensing. *IEEE Transactions on information theory*, 52(4):1289–
297 1306, 2006.
- 298 David L Donoho et al. High-dimensional data analysis: The curses and blessings of dimensionality.
299 *AMS math challenges lecture*, 1(2000):32, 2000.
- 300 Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas
301 Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An
302 image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint*
303 *arXiv:2010.11929*, 2020.
- 304 Rishabh Dudeja, Yue M. Lu, and Subhabrata Sen. Universality of approximate message passing
305 with semirandom matrices. *The Annals of Probability*, 51(5):1616–1683, 2023.
- 306 Vittorio Erba, Emanuele Troiani, Lenka Zdeborová, and Florent Krzakala. The nuclear route:
307 Sharp asymptotics of erm in overparameterized quadratic networks. *NeurIPS 2025*, 2025.
308 arXiv:2505.17958.
- 309 Tomer Galanti, Zachary S Siegel, Aparna Gupte, and Tomaso A Poggio. Sgd and weight decay
310 secretly minimize the rank of your neural network. In *NeurIPS 2024 Workshop on Mathematics*
311 *of Modern Machine Learning*, 2024.
- 312 Cédric Gerbelot and Raphaël Berthier. Graph-based approximate message passing iterations. *Infor-*
313 *mation and Inference: A Journal of the IMA*, 12(4):2562–2628, 2023.
- 314 Suriya Gunasekar, Blake E Woodworth, Srinadh Bhojanapalli, Behnam Neyshabur, and Nati Sre-
315 bro. Implicit regularization in matrix factorization. *Advances in neural information processing*
316 *systems*, 30, 2017.

- 324 Suriya Gunasekar, Jason Lee, Daniel Soudry, and Nathan Srebro. Characterizing implicit bias in
325 terms of optimization geometry. In *International Conference on Machine Learning*, pp. 1832–
326 1841. PMLR, 2018.
- 327 Jordan Hoffmann, Sebastian Borgeaud, Arthur Mensch, Elena Buchatskaya, Trevor Cai, Eliza
328 Rutherford, Diego de Las Casas, Lisa Anne Hendricks, Johannes Welbl, Aidan Clark, et al. An
329 empirical analysis of compute-optimal large language model training. *Advances in neural infor-*
330 *mation processing systems*, 35:30016–30030, 2022.
- 331 Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang,
332 Weizhu Chen, et al. Lora: Low-rank adaptation of large language models. *ICLR*, 1(2):3, 2022.
- 333 Jared Kaplan, Sam McCandlish, Tom Henighan, Tom B Brown, Benjamin Chess, Rewon Child,
334 Scott Gray, Alec Radford, Jeffrey Wu, and Dario Amodei. Scaling laws for neural language
335 models. *arXiv preprint arXiv:2001.08361*, 2020.
- 336 Diederik P Kingma and Jimmy Lei Ba. Adam: A method for stochastic gradient descent. In *ICLR:*
337 *international conference on learning representations*, pp. 1–15, 2015.
- 338 Seijin Kobayashi, Yassir Akram, and Johannes Von Oswald. Weight decay induces low-rank atten-
339 tion layers. *Advances in Neural Information Processing Systems*, 37:4481–4510, 2024.
- 340 Frederik Kunstner and Francis Bach. Scaling laws for gradient descent and sign descent for linear
341 bigram models under Zipf’s law. *arXiv preprint arXiv:2505.19227*, 2025.
- 342 Bruno Loureiro, Gabriele Sicuro, Cédric Gerbelot, Alessandro Pocco, Florent Krzakala, and Lenka
343 Zdeborová. Learning Gaussian mixtures with generalized linear models: Precise asymptotics in
344 high-dimensions. *Advances in Neural Information Processing Systems*, 34:10144–10157, 2021.
- 345 Yue M Lu and Horng-Tzer Yau. An equivalence principle for the spectrum of random inner-product
346 kernel matrices with polynomial scalings. *The Annals of Applied Probability*, 35(4):2411–2470,
347 2025.
- 348 Michael Mahoney and Charles Martin. Traditional and heavy tailed self regularization in neural
349 network models. In *International Conference on Machine Learning*, pp. 4284–4293. PMLR,
350 2019.
- 351 Antoine Maillard and Dmitriy Kunisky. Fitting an ellipsoid to random points: predictions using the
352 replica method. *IEEE Transactions on Information Theory*, 70(10):7273–7296, 2024.
- 353 Antoine Maillard, Emanuele Troiani, Simon Martin, Lenka Zdeborová, and Florent Krzakala.
354 Bayes-optimal learning of an extensive-width neural network from quadratically many samples.
355 *Advances in Neural Information Processing Systems*, 37:82085–82132, December 2024.
- 356 Stéphane Mallat. *A wavelet tour of signal processing*. Elsevier, 1999.
- 357 Alexander Maloney, Daniel A Roberts, and James Sully. A solvable model of neural scaling laws.
358 *arXiv preprint arXiv:2210.16859*, 2022.
- 359 VA Marchenko and Leonid A Pastur. Distribution of eigenvalues for some sets of random matrices.
360 *Mat. Sb.(NS)*, 72(114):4, 1967.
- 361 Charles H Martin and Michael W Mahoney. Implicit self-regularization in deep neural networks:
362 Evidence from random matrix theory and implications for learning. *Journal of Machine Learning*
363 *Research*, 22(165):1–73, 2021.
- 364 Charles H Martin, Tongsu Peng, and Michael W Mahoney. Predicting trends in the quality of state-
365 of-the-art neural networks without access to training or testing data. *Nature Communications*, 12
366 (1):4122, 2021.
- 367 Andrea Montanari and Basil N Saeed. Universality of empirical risk minimization. In *Conference*
368 *on Learning Theory*, pp. 4310–4312. PMLR, 2022.
- 369 Behnam Neyshabur, Ryota Tomioka, and Nathan Srebro. Norm-based capacity control in neural
370 networks. In *Conference on learning theory*, pp. 1376–1401. PMLR, 2015.

- 378 Elliot Paquette, Courtney Paquette, Lechao Xiao, and Jeffrey Pennington. 4+3 phases of compute-
379 optimal neural scaling laws. *Advances in Neural Information Processing Systems*, 37:16459–
380 16537, 2024.
- 381 Scott Pesme and Nicolas Flammarion. Saddle-to-saddle dynamics in diagonal linear networks. *Ad-
382 vances in Neural Information Processing Systems*, 36:7475–7505, 2023.
- 383
384 Garvesh Raskutti, Martin J Wainwright, and Bin Yu. Minimax rates of estimation for high-
385 dimensional linear regression over ℓ_q -balls. *IEEE transactions on information theory*, 57(10):
386 6976–6994, 2011.
- 387
388 Allan Raventós, Mansheej Paul, Feng Chen, and Surya Ganguli. Pretraining task diversity and the
389 emergence of non-bayesian in-context learning for regression. *Advances in neural information
390 processing systems*, 36:14228–14246, 2023.
- 391 Yunwei Ren, Eshaan Nichani, Denny Wu, and Jason D Lee. Emergence and scaling laws in sgd
392 learning of shallow neural networks. *arXiv preprint arXiv:2504.19983*, 2025.
- 393
394 Rylan Schaeffer, Brando Miranda, and Sanmi Koyejo. Are emergent abilities of large language
395 models a mirage? *Advances in neural information processing systems*, 36:55565–55581, 2023.
- 396
397 Chongjie Si, Xuankun Yang, Muqing Liu, Yadao Wang, Xiaokang Yang, Wenbo Su, Bo Zheng, and
398 Wei Shen. Weight spectra induced efficient model adaptation. *arXiv preprint arXiv:2505.23099*,
2025.
- 399
400 Daniel Soudry, Elad Hoffer, Mor Shpigel Nacson, Suriya Gunasekar, and Nathan Srebro. The im-
401 plicit bias of gradient descent on separable data. *Journal of Machine Learning Research*, 19(70):
402 1–57, 2018.
- 403
404 Max Staats, Matthias Thamm, and Bernd Rosenow. Small singular values matter: A random matrix
analysis of transformer models. *arXiv preprint arXiv:2410.17770*, 2024.
- 405
406 Mihailo Stojnic. Various thresholds for ℓ_1 -optimization in compressed sensing. *arXiv preprint
arXiv:0907.3666*, 2009.
- 407
408 Christos Thrampoulidis, Samet Oymak, and Babak Hassibi. The Gaussian min-max theorem in the
409 presence of convexity. *arXiv preprint arXiv:1408.4837*, 2014.
- 410
411 Christos Thrampoulidis, Samet Oymak, and Mahdi Soltanolkotabi. Theoretical insights into multi-
412 class classification: A high-dimensional asymptotic view. *Advances in Neural Information Pro-
cessing Systems*, 33:8907–8920, 2020.
- 413
414 Emanuele Troiani, Hugo Cui, Yatin Dandi, Florent Krzakala, and Lenka Zdeborová. Fundamental
415 limits of learning in sequence multi-index models and deep attention networks: high-dimensional
416 asymptotics and sharp thresholds. In *Forty-second International Conference on Machine Learn-
417 ing*, 2025.
- 418
419 Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, et al.
420 Attention is all you need. *Advances in neural information processing systems*, 30(1):261–272,
2017.
- 421
422 Matteo Vilucchio, Yatin Dandi, Cedric Gerbelot, and Florent Krzakala. Asymptotics of non-convex
423 generalized linear models in high-dimensions: A proof of the replica formula. *arXiv preprint
arXiv:2502.20003*, 2025.
- 424
425 Jason Wei, Yi Tay, Rishi Bommasani, Colin Raffel, Barret Zoph, Sebastian Borgeaud, Dani Yo-
426 gatama, Maarten Bosma, Denny Zhou, Donald Metzler, Ed H. Chi, Tatsunori Hashimoto, Oriol
427 Vinyals, Percy Liang, Jeff Dean, and William Fedus. Emergent abilities of large language models.
428 *Trans. Mach. Learn. Res.*, 2022, 2022.
- 429
430 Yizhou Xu, Antoine Maillard, Lenka Zdeborová, and Florent Krzakala. Fundamental limits of
431 matrix sensing: Exact asymptotics, universality, and applications. In *Proceedings of Thirty Eighth
Conference on Learning Theory*, volume 291 of *Proceedings of Machine Learning Research*, pp.
5757–5823. PMLR, 2025.

432 David Yunis, Kumar Kshitij Patel, Samuel Wheeler, Pedro Savarese, Gal Vardi, Karen Livescu,
 433 Michael Maire, and Matthew R Walter. Approaching deep learning through the spectral dynamics
 434 of weights. *arXiv preprint arXiv:2408.11804*, 2024.

435
 436 Yedi Zhang, Aaditya K Singh, Peter E Latham, and Andrew M Saxe. Training dynamics of in-
 437 context learning in linear attention. In *Forty-second International Conference on Machine Learn-*
 438 *ing*, 2025.

440 A RELATED WORK

441
 442
 443 Modern machine learning relies increasingly on attention mechanisms Vaswani et al. (2017), which
 444 form the backbone of state-of-the-art models in natural language processing, vision, and beyond
 445 Devlin et al. (2019); Dosovitskiy et al. (2020); Brown et al. (2020). Despite their empirical success,
 446 many fundamental questions about learning with attention layers remain open.

447 In particular, the weight matrices of trained attention layers display striking and reproducible spec-
 448 tral patterns. Empirical studies report that the singular-value distributions of the query and key
 449 projections are far from random: they exhibit a structured bulk with non-trivial tails Martin et al.
 450 (2021); Staats et al. (2024), consistent with earlier observations of heavy-tailed and compressible
 451 spectra across deep networks Mahoney & Martin (2019). These regularities persist across architec-
 452 tures and training scales, yet their theoretical origin and their implications for generalization remain
 453 largely unexplained. *Why do such spectra emerge in attention weights, and what do they reveal*
 454 *about the inductive biases of the model?*

455 Second, neural scaling laws Kaplan et al. (2020); Brown et al. (2020); Hoffmann et al. (2022) reveal
 456 striking power-law relations between performance and computational or data resources in attention-
 457 based models. Yet, with few exceptions, theoretical understanding of these phenomena remains
 458 largely confined to linearized or lazy regimes, e.g. Caponnetto & De Vito (2007); Cui et al. (2021);
 459 Maloney et al. (2022); Bahri et al. (2024); Paquette et al. (2024); Atanasov et al. (2024); Bordelon
 460 et al. (2024); Kunstner & Bach (2025); Defilippis et al. (2024) that may not capture genuine rep-
 461 resentation learning. At the same time, growing evidence suggests that learning in attention layers
 462 is not smooth: training dynamics often display plateaus and abrupt transitions, with new directions
 463 in representation space emerging sequentially rather than all at once. This stepwise behavior has
 464 been observed both in simplified attention models and in large-scale transformers Wei et al. (2022);
 465 Raventós et al. (2023); Arora & Goyal (2023); Schaeffer et al. (2023); Ren et al. (2025). *Can emer-*
 466 *gence and scaling phenomena of this kind be reproduced and explained within a solvable attention*
 467 *model?*

468 A natural strategy for addressing these questions is to analyze simplified models in high-dimensional
 469 regimes, where the blessing of dimensionality Donoho et al. (2000) enables sharp and tractable
 470 characterizations of learning. This approach has driven major advances in the theory of two-layer
 471 networks via Gaussian single- and multi-index models (see, e.g., Ben Arous et al. (2021); Abbe et al.
 472 (2022); Ba et al. (2022); Arnaboldi et al. (2023); Damian et al. (2023); Bietti et al. (2025); Berthier
 473 et al. (2024)). In these settings, high-dimensional asymptotics have enabled an essentially complete
 474 characterization of empirical risk minimization.

475 In contrast, attention mechanisms remain far less explored within such a high-dimensional frame-
 476 work (with notable exceptions of low-rank attention Cui et al. (2024); Troiani et al. (2025); Arn-
 477 aboldi et al. (2025) and in-context linear regression Zhang et al. (2025)), and, to our knowledge,
 478 there is no high-dimensional asymptotic theory linking spectral properties to generalization.

479 Several empirical studies have further noted that attention projections are approximately low-rank.
 480 Staats et al. (2024) found rapidly decaying spectra in pretrained transformers, while Si et al. (2025)
 481 reported similar profiles in LLaMA models, showing that the low-rank tendency is present already
 482 at pretraining and persists through fine-tuning. Broader surveys confirm stable low-rank structure
 483 during training Yunis et al. (2024), and parameter-efficient fine-tuning methods such as LoRA Hu
 484 et al. (2022) exploit the fact that very low ranks suffice to capture relevant structure. Recent work
 485 further leverages spectral profiles for efficient adaptation Si et al. (2025). While these findings are
 empirical, recent theoretical progress has linked weight decay on the query and key matrices to an
 effective nuclear-norm regularization of their product Kobayashi et al. (2024), thereby formalizing

a low-rank inductive bias in attention, but they do not characterize any explicit relation between this low-rank bias and generalization. The low-rank inductive bias connects directly to earlier analyses of factorized parameterizations in feedforward networks, where weight decay implicitly promotes low-rank solutions through nuclear-norm penalties Gunasekar et al. (2017; 2018); Arora et al. (2019); Galanti et al. (2024).

In contrast to the extensive literature on single- and multi-index models, attention mechanisms remain comparatively underexplored from a high-dimensional theoretical perspective. Existing analyses of attention Cui et al. (2024); Troiani et al. (2025); Arnaboldi et al. (2025) primarily focus on extremely low-rank regimes of order $\mathcal{O}(1)$, which do not capture the rich spectral structure observed empirically in trained attention layers. More recently, attention-indexed models have been introduced Boncoraglio et al. (2025). These models offer a controlled setting in which questions of generalization and representation learning can be addressed. However, existing results in this line focus on Bayes-optimal inference, leaving open the characterization of solutions reached by empirical risk minimization (ERM), which is the regime relevant for trained neural networks.

Our work builds on recent progress in the asymptotic analysis of quadratic neural networks Erba et al. (2025); Xu et al. (2025); Defilippis et al. (2025) to bridge this gap by mapping ERM in attention to an equivalent matrix sensing problem. This reformulation allows us to analyze the learned spectrum using tools from approximate message passing and its connections to convex and non-convex optimization Berthier et al. (2020); Gerbelot & Berthier (2023); Loureiro et al. (2021); Vilucchio et al. (2025).

B TECHNICAL RESULTS CONCERNING THE ASYMPTOTICS

Claim B.1 (Train and test error of ERM). *Define $\tilde{\lambda} = \sqrt{\kappa}\lambda$ and $\mu_\delta = \mu_0 \boxplus \mu_{\text{sc},\delta}$, where \boxplus is the free convolution and $\mu_{\text{sc},\delta}(x) = \sqrt{4\delta^2 - x^2}/(2\pi\delta^2)$ the semicircle distribution of radius 2δ for $\delta > 0$, and F_δ its c.d.f.. Define also $\tilde{\sigma}_\beta(A) = \sigma_\beta(\{\sqrt{1 + \delta_{ab}} A_{ab}\}_{ab})$ for $A \in \mathbb{R}^{T \times T}$, and where σ_β is applied row-wise. Call $(\Sigma^*, m^*, q^*, \hat{\Sigma}^*, \hat{m}^*, \hat{q}^*)$ the global minimizer of*

$$\Phi(\Sigma, m, q, \hat{\Sigma}, \hat{m}, \hat{q}) = \frac{\hat{q}\Sigma + 2\hat{m}m - \hat{\Sigma}q}{4} + \frac{n}{d^2} \mathcal{M}(\Sigma, m, q) - \frac{\hat{m}^2}{4\hat{\Sigma}} J\left(\frac{\sqrt{\hat{q}}}{\hat{m}}, \frac{2\tilde{\lambda}}{\hat{m}}\right), \quad (9)$$

where

$$\mathcal{M}(\Sigma, m, q) = \mathbb{E}_{z_0, z} \inf_{h \in \mathcal{S}(T)} \left\{ \frac{1}{2\Sigma} \sum_{\substack{a,b=1 \\ a \leq b}}^T (h_{ab} - z_{ab})^2 + \sum_{a,b=1}^T [\tilde{\sigma}_{\beta_0}(z_0)_{ab} - \tilde{\sigma}_\beta(h)_{ab}]^2 \right\}, \quad (10)$$

$$J(\delta, \epsilon) = \int_\epsilon^{+\infty} \mu_\delta(dx) (x - \epsilon)^2,$$

$\mathcal{S}(T)$ is the set of $T \times T$ symmetric matrices, and where $\mathbb{E}_{z_0, z}$ denotes the average over $\{z_{ab}, z_{0,ab}\}_{1 \leq a \leq b \leq T}$, which are Gaussian random variables

$$\begin{bmatrix} z_{0,ab} \\ z_{ab} \end{bmatrix} \sim \mathcal{N}\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}; \begin{bmatrix} Q_0 + \Delta/2 & m \\ m & q \end{bmatrix}\right), \quad (11)$$

independently for each pair a, b (up to symmetry $z_{ab} = z_{ba}$ and $z_{0,ab} = z_{0,ba}$). Assume that at the global minimum the replicon condition Vilucchio et al. (2025) is satisfied:

$$2\alpha \mathbb{E}_{z_0, z} \sum_{\substack{a,b,c,d=1 \\ a \leq b, c \leq d}}^T \left(\frac{\partial_{z_{ab}} p_{cd} - \delta_{ac} \delta_{bd}}{\Sigma} \right)^2 \int \mu_{\sqrt{\hat{q}}/\hat{m}}(dx) \mu_{\sqrt{\hat{q}}/\hat{m}}(dy) \frac{(\xi(x) - \xi(y))^2}{\hat{\Sigma}^2(x-y)^2} < 1, \quad (12)$$

where $\xi(x) = \text{ReLU}(x - 2\tilde{\lambda}/\hat{m})$ and

$$p(z_0, z, \Sigma) = \operatorname{arginf}_{h \in \mathcal{S}(T)} \left\{ \frac{1}{2\Sigma} \sum_{\substack{a,b=1 \\ a \leq b}}^T (h_{ab} - z_{ab})^2 + \sum_{a,b=1}^T [\tilde{\sigma}_{\beta_0}(z_0)_{ab} - \tilde{\sigma}_\beta(h)_{ab}]^2 \right\}. \quad (13)$$

Then, for all values of $\alpha, \lambda > 0$, $\Delta \geq 0$ and $\kappa \geq 1 - F_{\sqrt{\hat{q}/\hat{m}}}(2\tilde{\lambda}/\hat{m})$ any global minimum \hat{W} of Eq. (2) satisfies

$$\lim_{d \rightarrow \infty} \mathbb{E} e_{\text{test}}(\hat{W}) = \mathbb{E}_{z_0, z} \sum_{a, b=1}^T [\tilde{\sigma}_{\beta_0}(z_0)_{ab} - \tilde{\sigma}_{\beta}(z)_{ab}]^2, \quad (14)$$

$$\lim_{d \rightarrow \infty} d^{-2} \mathbb{E} \mathcal{L}(\hat{W}) = \Phi(\Sigma, m, q, \hat{\Sigma}, \hat{m}, \hat{q}),$$

where all order parameters $(\Sigma, m, q, \hat{\Sigma}, \hat{m}, \hat{q})$ are evaluated at the global minimum defined above.

Claim B.1 characterizes the global minimum of (2) through the solution of a six-dimensional optimization problem (9), whose numerical solution we discuss in Appendix J. To provide a bit more intuition, we remark the parameters m^* , q^* and Σ^* measure, respectively, the target-learner overlap $\text{Tr}(S_0^\top S)/d$, the learner norm $\text{Tr}(S^\top S)/d$ and the trace of the inverse Hessian of the loss Clarte et al. (2023) as customary in approximate-message-passing analysis, where we defined $S = \hat{W}\hat{W}^\top/\sqrt{pd}$. We also stress that Claim B.1 can be adapted to more general losses $\sum_{\mu=1}^n \ell(\mathbf{x}_\mu S_0 \mathbf{x}_\mu^\top; \mathbf{x}_\mu S \mathbf{x}_\mu^\top)$ (i.e., depending bi-linearly on the data) and regularizations inducing a spectral penalty on the effective matrix $S = W^\top W/\sqrt{pd}$. This includes models with mismatched nonlinearities between data and learner, classification tasks, and more. We provide the equations for the more general case in Appendix E.2.1.

We formulate the result as a claim rather than a theorem, as its complete proof would require a substantial technical development that closely follows standard arguments from high-dimensional inference. We present a proof sketch highlighting the key steps in Appendix E. All components of the argument rely on established techniques from random matrix theory and approximate message passing. Additionally, we stress that we evaluated condition (12) numerically for all cases shown in the plots, and we found that it was always satisfied.

Sharp learning curves at quadratic sample complexity. In Figure 1 (and Figure 6 in the Appendix J) we plot the test error predicted by our theory as a function of the sample ratio $\alpha = n/d^2$ for $\Delta = 0.5$. We observe a monotone decreasing behavior at large values of regularization λ (Figure 1), while at lower values of λ a characteristic interpolation peak (with oddly asymmetric shape) appears around the maximal number of samples that can be perfectly fit by the model architecture, the so-called interpolation threshold (Figure 6). Claim B.1 allows to derive an analytic prediction for the interpolation threshold, i.e. the value α_{interp} before which the ERM estimator for $\lambda \rightarrow 0^+$ achieves zero training loss. In the noiseless case $\Delta = 0$, we can also predict analytically the perfect recovery threshold (Corollary B.2), i.e. the value α_{perf} after which the ERM estimator achieves perfect generalization. B.1.

Corollary B.2 (Interpolation and perfect recovery thresholds). *Consider the setting of Section 2 with $\lambda \rightarrow 0^+$, $\Delta \geq 0$ and $T \geq 2$. Then, there exists a value of sample ratio α_{interp} such that the training loss at its global minimum is zero for $\alpha < \alpha_{\text{interp}}$ (perfect fit of the training set), and strictly positive for $\alpha > \alpha_{\text{interp}}$, and α_{interp} satisfies*

$$\alpha_{\text{interp}} = \frac{\partial_1 J(\bar{\delta}, 0)}{2\bar{\delta}(T^2 + T - 2)} \quad (15)$$

where $\bar{\delta}$ is the solution of

$$Q_0 + \frac{\Delta}{2} = J(\bar{\delta}, 0) - \frac{\bar{\delta}}{2} \partial_1 J(\bar{\delta}, 0) \quad (16)$$

and J is defined in Eq. (10). Moreover, if $\Delta = 0$, there exists a value of sample ratio α_{perf} such that the test error at the global minimum is zero for $\alpha > \alpha_{\text{perf}}$ (perfect generalization), and strictly positive for $\alpha < \alpha_{\text{perf}}$, determined as follows. Call \bar{c} the solution of the equation

$$M_{\text{s.c.}}^{(1)}(\bar{c}) - \bar{c} M_{\text{s.c.}}^{(0)}(\bar{c}) + \frac{\bar{c}}{1 - \kappa_0} = 0, \quad (17)$$

where

$$M_{\text{s.c.}}^{(k)}(x) = \int_{-2}^x dx \mu_{\text{sc}}(x) x^k \quad (18)$$

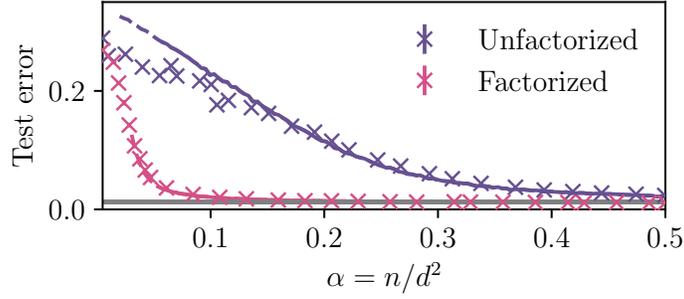


Figure 3: Test error as a function of α for standard factorized query-key training Eq. (2) and direct (non-factorized) training of the attention matrix Eq. (21), both at optimal regularization (selected by cross-validation). The factorized parameterization consistently achieves significantly lower test error across all α . Solid lines show theoretical predictions (Claim B.1 and Appendix G), while dots correspond to numerical experiments with Adam at $d = 100$, averaged over 8 runs with 2000 test samples. Parameters are $\kappa_0 = 0.05$, $\Delta = 0.05$, $T = 2$, and $\beta = \beta_0 = 1$.

for $0 < \kappa_0 < 1$. Then,

$$\alpha_{\text{perf}} = \frac{1 - (1 - \kappa_0)^2 \left(M_{\text{s.c.}}^{(2)}(c) - cM_{\text{s.c.}}^{(1)}(c) \right)}{T^2 + T - 2}. \quad (19)$$

Corollary B.2 follows from a mapping of the theory of softmax attention to that of linear attention that holds for $\alpha \leq \alpha_{\text{interp}}$ and $\lambda \rightarrow 0^+$, allowing to adapt results from Erba et al. (2025), as detailed in Appendix F.

Behavior of gradient descent. A natural question is whether gradient-based methods can reach global minima of the non-convex loss in Eq. (2), or whether they remain stuck in spurious local minima. In Fig. 1 and 6 we compare our theory with numerical results of runs of Adam Kingma & Ba (2015) (details in Appendix J) run on the loss of Eq. (2) with $d = 100$ and averaging over 32 different instances, and $p = d$. We observe an excellent match for both the test error and the training loss for both $\kappa = 1$ and $\kappa = 3/4$.

Inductive bias: weight decay implies nuclear norm on attention. The ℓ_2 regularization over the weights W naturally translates to a nuclear norm regularization in the equivalent generalized matrix problem, see Kobayashi et al. (2024) and Appendix D, naturally favoring model weights configurations with an effective lower width (i.e., implementable with fewer hidden units): a weight decay in W thus implies a low-rank learned attention matrix. This partly explains why learning with low-rank weights (usually done for computational reasons) does not negatively affect generalization despite reducing expressivity. We quantify this bias in Claim B.1, where we find that all attention learners with rank satisfying $p/d \geq 1 - F_{\sqrt{q}/\hat{m}}(2\hat{\lambda}/\hat{m})$ achieve the same error. This implies that mildly rank-constrained to massively over-parametrized architectures incur no penalty in terms of train and test error.

Comparison with non-factorized parameterization. A natural baseline to consider is learning a non-factorized attention S through the empirical risk minimization of the model (e.g. in the seq2seq task)

$$\hat{f}_{\text{L2}}(\mathbf{x}_{\text{in}}; S) = \sigma_{\beta} \left(\frac{\mathbf{x}_{\text{in}} S \mathbf{x}_{\text{in}}^T - \mathbb{E}_{\text{tr}} \text{Tr}(\mathbf{x}_{\text{in}} S \mathbf{x}_{\text{in}}^T)}{\sqrt{d}} \right) \mathbf{x}_{\text{in}}, \quad (20)$$

where we impose that S is symmetric (but neither PSD nor factorized), and we train on the loss

$$\mathcal{L}_{\text{L2}}(S) := \frac{1}{d} \sum_{\mu=1}^n \|\mathbf{x}_{\text{out}}^{\mu} - \hat{f}_{\text{L2}}(\mathbf{x}_{\text{in}}^{\mu}; S)\|_2^F + \lambda \|S\|_F^2. \quad (21)$$

The Frobenius-regularized estimator underperforms the nuclear-norm-regularized one whenever the target is sufficiently low-rank. In the noisy case, we compare the test error of the factorized ERM

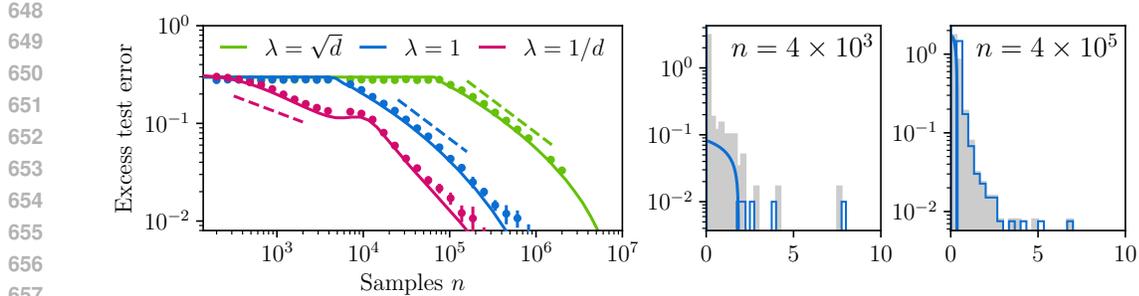


Figure 4: (Left) Excess test error (Eq. (3) minus its value at $n \rightarrow +\infty$) of the ERM estimator Eq. (2) (Claim B.1, Eq. (14)) compared with Adam simulations at $d = 200$ (8 instances, error bars = standard deviation) as a function of sample number n , where we use power-law target (Section C, decay exponent $\gamma = 0.75$), and $\Delta = 0.5$, $T = 2$, $\beta = \beta_0 = 1$. We plot three values of $\lambda = 1/d, 1, \sqrt{d}$. One can see clearly several scaling regimes in different regions, with different decay exponents. We highlight those of them (in dashed lines) in which the error decays as $n^{-f(\gamma)}$ with non-trivial dependency on γ , namely $f(\gamma) = 1 - 1/(2\gamma)$ for $\lambda = 1/d$, and $f(\gamma) = 2 - 1/\gamma$ for $\lambda = 1, \sqrt{d}$. (Right) Eigenvalue spectrum of the ERM estimator from theory (blue, Claim 4.1) vs. Adam simulations (grey histograms) at $n = 4 \cdot 10^3, 4 \cdot 10^5$ ($d = 400$, single runs). For large n the spectrum develops a heavy tail.

estimator of Eq. (2) with the non-factorized Frobenius estimator of Eq. (21) at optimal regularization (cross-validated, Fig. 3 left). The factorized model consistently outperforms, fully exploiting the induced low-rank bias. In the extreme low-rank regime $p_0 \ll d$, it achieves vanishing error at sample scale $\mathcal{O}(dp_0)$, whereas the non-factorized model requires $\mathcal{O}(d^2)$, see Fig. 8 left in Appendix J.

C POWER-LAW TARGETS AND SCALING LAWS

We now focus on targets whose attention map exhibits power-law structure, as discussed for multi-index models in Ren et al. (2025); Ben Arous et al. (2025); Defilippis et al. (2025). Concretely, we assume that the target matrix S_0 has a power-law spectrum, with eigenvalues $\{\sqrt{d}i^{-\gamma}\}_{i=1}^d$ with $\gamma > 1/2$, corresponding to a sequence of signal directions with progressively decreasing strength. Such spectra are natural in structured or multi-scale tasks and provide a minimal setting in which emergence and scaling phenomena can be studied analytically. All details of the derivations are reported in Appendix I.

Sequential spectral recovery and heavy-tailed spectra. Under the power-law assumption, the learned spectral law (7) features a progressive recovery of signal modes. The effective threshold ϵ acts as a moving cutoff along the target spectrum: as the sample complexity α increases (or the regularization decreases), ϵ shifts toward smaller eigenvalues, allowing increasingly weaker spectral directions of S_0 to be incorporated into the learned representation. Learning, therefore, proceeds through a sequence of regimes where a growing fraction of the dominant spectral modes contributes to generalisation.

This mechanism extends the spectral phenomenology discussed earlier. For targets with Marchenko–Pastur–type spectra, the same process leads to the emergence of a bulk of spectral outliers as learning progresses. Instead, when the target attention map has a power-law spectrum the progressive recovery of spectral modes produces isolated outliers, reproducing one of the most robust empirical observations in trained attention layers (Figure 4 and Martin et al. (2021); Staats et al. (2024)), and leads in some regimes to a learned spectrum with power-law tails. The heavy-tailed behavior arises not from isolated spikes, but from the accumulation of an increasing number of recovered modes across scales.

Specifying our results to power-law targets also leads to an interpretable mechanism for emergence Wei et al. (2022); Schaeffer et al. (2023); Ren et al. (2025). The underfitting and approximation terms of (8) are controlled by the cutoff scale ϵ , while the overfitting contribution is governed by the bulk noise level δ . As n, d, λ varies, different contributions dominate the excess risk, leading to

sharp crossovers between learning regimes. These crossovers correspond to the successive inclusion of new spectral modes into the learned attention map, providing a first-principles explanation of emergence phenomena in feature-learning models. From this perspective, the diversity of spectral behaviors observed in modern attention architectures would reflect differences in task structure, rather than qualitative changes in optimization dynamics or architectural design.

Power-law learning curves and universal exponents. The same emergence mechanism directly yields power-law scaling of the test error with the number of samples. In particular, the excess risk obeys $e_{\text{test}} - e_0 \asymp n^{-f(\gamma)}$ where the exponent $f(\gamma) > 0$ depends on the tail exponent γ of the target spectrum, and on the value of n, d, λ . Remarkably, these rates coincide with those previously obtained for LASSO and matrix compressed sensing by Defilippis et al. (2025), supporting their universality, and read:

$$e_{\text{test}} = e_0 + \Theta\left(\left(\frac{n}{d}\right)^{-1+\frac{1}{2\gamma}} + \left(\frac{n}{d^2}\right)^{\frac{2}{5}}\right) \quad (22)$$

for $d \ll n \ll d^2$, $\lambda \ll \sqrt{\frac{n}{d^2}}$ and

$$e_{\text{test}} - e'_0 = \begin{cases} \Theta\left(\frac{d^2}{n}\right), & n \gg d^2, \lambda \ll \sqrt{\frac{n}{d^2}}, \\ \Theta\left(\left(\frac{\lambda d^{3/2}}{n}\right)^{2-\frac{1}{\gamma}}\right), & \max\left(\sqrt{\frac{n}{d^2}}, \frac{n}{d^{\gamma+3/2}}\right) \ll \lambda \ll \frac{n}{d}, \\ \Theta\left(\left(\frac{\lambda d^2}{n}\right)^2\right), & \sqrt{\frac{n}{d^2}} \ll \lambda \ll \frac{n}{d^{\gamma+3/2}}. \end{cases} \quad (23)$$

where e_0 and $e'_0 \leq e_0$ refer to the two saturation errors. The derivation is detailed in Appendix I, and the results are illustrated in Fig. 4 as well as in a phase diagram in Fig. 5 in Appendix I. The key mechanism underlying these universal scaling laws is thus the combination of two ingredients: *quasi-sparsity* of the target representation, encoded by a heavy-tailed spectrum that induces an ordered hierarchy of feature strengths Mallat (1999); Donoho (2006), and the implicit *rank-sparsity* bias induced by attention. Under optimal regularization, these yield an excess generalization error scaling as $n^{-1+1/(2\gamma)}$, matching the classical minimax rate achieved by the LASSO Raskutti et al. (2011).

D INDUCTIVE NUCLEAR NORM BIAS FOR TIED AND UNTIED WEIGHTS

We briefly recall here the mathematical reasons behind the appearance of the *inductive nuclear norm*, following arguments that go back to classical variational characterizations of atomic norms, and were emphasized in the deep learning context by Neyshabur et al. (2015); Gunasekar et al. (2017); Soudry et al. (2018); Pesme & Flammarion (2023). These arguments are quite general: whenever a predictor is parameterized in a factorized form (for instance, as the product of two matrices, or as a diagonal interaction of two vectors), an ℓ_2 -type penalty on the factors induces, via an AM–GM–style variational identity, an ℓ_1 or nuclear norm penalty on the effective predictor. More recent works Maillard & Kunisky (2024); Xu et al. (2025) have also used these ideas.

Case 1 (Tied weight): Let $W \in \mathbb{R}^{d \times p}$ and $M := W^\top W \in \mathbb{R}^{p \times p}$. Assume the loss depends on W only via M , i.e. there exists $\Phi : \mathbb{S}_+^p \rightarrow \mathbb{R}$ such that

$$\mathcal{L}(W) = \Phi(W^\top W) + \lambda \|W\|_F^2.$$

Then the following problems are equivalent in optimal value, and their minimizers correspond via $M = W^\top W$:

$$\min_{W \in \mathbb{R}^{d \times p}} \Phi(W^\top W) + \lambda \|W\|_F^2 \equiv \min_{M \succeq 0, \text{rank}(M) \leq d} \Phi(M) + \lambda \|M\|_\star. \quad (\star)$$

Proof. Since $M = W^\top W \succeq 0$ with $\text{rank}(M) \leq d$, and if $s_i(W)$ are the singular values of W , then the eigenvalues of M are $s_i(W)^2$. Hence

$$\|W\|_F^2 = \sum_i s_i(W)^2 = \sum_i \lambda_i(M) = \|M\|_\star.$$

Because Φ depends only on M , we can replace W by M and obtain (\star) . Moreover, any feasible $M \succeq 0$ with $\text{rank}(M) \leq d$ admits a factorization $M = W^\top W$; choosing such a W makes the two objectives equal. \square

Case 2 (Untied weights): Let $U \in \mathbb{R}^{p \times r}$, $V \in \mathbb{R}^{n \times r}$, and $M := UV^\top \in \mathbb{R}^{p \times n}$. Assume the loss depends on (U, V) only through the product M , i.e.

$$\mathcal{L}(U, V) = \Psi(UV^\top) + \frac{\lambda}{2}(\|U\|_F^2 + \|V\|_F^2),$$

for some (possibly nonconvex) function $\Psi : \mathbb{R}^{p \times n} \rightarrow \mathbb{R}$. We show the exact equivalence

$$\min_{U, V} \Psi(UV^\top) + \frac{\lambda}{2}(\|U\|_F^2 + \|V\|_F^2) \equiv \min_{M \in \mathbb{R}^{p \times n}} \Psi(M) + \lambda \|M\|_\star. \quad (\dagger)$$

Matrix AM–GM analogue (variational identity for $\|\cdot\|_\star$). The key inequality is the matrix analogue of AM–GM: for any factorization $M = UV^\top$,

$$\|M\|_\star \leq \frac{1}{2}(\|U\|_F^2 + \|V\|_F^2), \quad (\text{A})$$

with equality attained by a specific SVD-based factorization.

Proof of (A). Use the dual characterization of the nuclear norm:

$$\|M\|_\star = \max_{\|A\|_2 \leq 1} \langle M, A \rangle = \max_{\|A\|_2 \leq 1} \langle UV^\top, A \rangle = \max_{\|A\|_2 \leq 1} \langle U, AV \rangle,$$

where $\langle X, Y \rangle := \text{Tr}(X^\top Y)$ and $\|\cdot\|_2$ is the spectral norm. For any A with $\|A\|_2 \leq 1$, apply Cauchy–Schwarz and the scalar AM–GM (or $ab \leq \frac{1}{2}(a^2 + b^2)$):

$$\langle U, AV \rangle \leq \|U\|_F \|AV\|_F \leq \|U\|_F \|A\|_2 \|V\|_F \leq \|U\|_F \|V\|_F \leq \frac{1}{2}(\|U\|_F^2 + \|V\|_F^2).$$

Taking the maximum over $\|A\|_2 \leq 1$ yields (A).

Tightness of (A). Let the compact SVD of M be $M = P\Sigma Q^\top$ with $\Sigma = \text{diag}(\sigma_1, \dots, \sigma_\rho)$, $\sigma_i \geq 0$. Choose the *balanced SVD factorization*

$$U := P\Sigma^{1/2}, \quad V := Q\Sigma^{1/2}.$$

Then $M = UV^\top$ and

$$\|U\|_F^2 + \|V\|_F^2 = \text{Tr}(\Sigma) + \text{Tr}(\Sigma) = 2 \sum_{i=1}^{\rho} \sigma_i = 2\|M\|_\star.$$

Hence, equality holds in (A):

$$\|M\|_\star = \min_{M=UV^\top} \frac{1}{2}(\|U\|_F^2 + \|V\|_F^2), \quad (\text{B})$$

and the minimum is attained by the balanced SVD factorization.

Equivalence of objectives. Because Ψ depends on (U, V) only through $M = UV^\top$, we may fix M and minimize the regularizer over all factorizations $M = UV^\top$. By (B),

$$\inf_{U, V: UV^\top = M} \frac{\lambda}{2}(\|U\|_F^2 + \|V\|_F^2) = \lambda \|M\|_\star.$$

Therefore

$$\inf_{U, V} \Psi(UV^\top) + \frac{\lambda}{2}(\|U\|_F^2 + \|V\|_F^2) = \inf_M \Psi(M) + \lambda \|M\|_\star,$$

which is exactly (\dagger) .

Notice that no convexity of Ψ is required for the equality of optimal values; convexity only affects algorithmic guarantees.

E DERIVATION OF CLAIM B.1 AND PROOF SKETCH

Note that Claim 1 requires evaluating the replicon condition (12), which guarantees convergence of the AMP algorithm Bolthausen (2014). Although a general proof is difficult and direct evaluation can be numerically demanding, in our experiments AMP did converge (the replicon condition being a convergence criterion for AMP), and gradient-based algorithms reached the same fixed point. This provides strong empirical evidence that the replicon condition is satisfied in our setting.

In this paper, we provide an informal justification for Claim B.1. Claim B.1 can be proven by going through the following steps:

- Reduction to the PSD generalized matrix estimation problem

$$\tilde{\mathcal{L}}(S) := \frac{1}{d} \sum_{\mu=1}^n \|\tilde{\sigma}_\beta(\{\text{Tr}[SX_{ab}(\mathbf{x}^\mu)]\}_{ab})\mathbf{x}^\mu - \tilde{\sigma}_{\beta_0}(\{\text{Tr}[S_0X_{ab}(\mathbf{x}^\mu)]\}_{ab})\mathbf{x}^\mu\|_F^2 + \sqrt{pd} \lambda \text{Tr}(S) \quad (24)$$

where $X_{ab}(\mathbf{x}) = (\mathbf{x}_a\mathbf{x}_b^T + \mathbf{x}_b\mathbf{x}_a^T - 2\delta_{ab}\mathbb{I}_d)/\sqrt{2d(1+\delta_{ab})}$, $S = W^TW/\sqrt{pd} \succeq 0$ and $\tilde{\sigma}(A) = \sigma(\{\sqrt{1+\delta_{ab}}A_{ab}\}_{ab})$. This step is discussed in Appendix D.

- Asymptotic equivalence of sequence-to-sequence and sequence-to-attention tasks. One needs to show (as done for e.g. in Cui et al. (2024)) that Eq. (24) is asymptotically the same as

$$\tilde{\mathcal{L}}(S) := \sum_{\mu=1}^n \|\tilde{\sigma}_\beta(\{\text{Tr}[SX_{ab}(\mathbf{x}^\mu)]\}_{ab}) - \tilde{\sigma}_{\beta_0}(\{\text{Tr}[S_0X_{ab}(\mathbf{x}^\mu)]\}_{ab})\|_F^2 + d\tilde{\lambda} \text{Tr}(S), \quad (25)$$

where we called $\tilde{\lambda} = \sqrt{p/d}\lambda = \sqrt{\kappa}\lambda$. To do this, one should use a law-of-large-numbers kind of argument giving

$$\frac{1}{d} \|F(\mathbf{x})\mathbf{x}\|_F^2 = \frac{1}{d} F(\mathbf{x})\mathbf{x}\mathbf{x}^T F(\mathbf{x})^T \approx F(\mathbf{x})F(\mathbf{x})^T \quad (26)$$

where we used that $\mathbb{E}_{\mathbf{x}}\mathbf{x}\mathbf{x}^T/d = \mathbb{I}_T$.

- Gaussian universality, i.e. replacement of each matrix $X_{ab}(\mathbf{x}^\mu)$ by a random independent Wigner matrix G_{ab}^μ . This step is *a priori* non-trivial, and its proof would require generalizing the arguments of Maillard et al. (2024); Xu et al. (2025); Erba et al. (2025) to the case of multiple tokens (the case $T = 1$ is instead included in Erba et al. (2025)). We do not foresee any technical roadblock here, as this step requires only promoting scalar outputs to finite-dimensional vectorial ones in all such proofs.
- Analysis of the final empirical risk minimization problem

$$\tilde{\mathcal{L}}(S) := \frac{1}{d} \sum_{\mu=1}^n \|\sigma_\beta(\{\text{Tr}[SG_{ab}^\mu]\}_{ab}) - \sigma_{\beta_0}(\{\text{Tr}[S_0G_{ab}^\mu]\}_{ab})\|_F^2 + d\tilde{\lambda} \text{Tr}(S). \quad (27)$$

The asymptotic analysis of this loss can be done by adapting the analysis in Erba et al. (2025) (corresponding to the single token case $T = 1$ with linear activation) to multiple tokens, in the same spirit as the generalization of single-index models to multiple tokens performed in Cui et al. (2024). In practice this reduces to:

- Writing down an appropriate Approximate Message Passing (AMP) algorithm (which we present explicitly in the following) whose fixed point are local minima of the loss in Eq. (27).
- Using AMP theory to derive a set of low-dimensional state evolution equations that track AMP step-by-step, and use them to characterize the properties of its fixed points.
- Among the fixed points of state evolution, select the one with the lowest value of the training loss (describing then the properties of global minima).

The only non-trivial point here is that this program is *a priori* that for non-convex losses such as Eq. (27) AMP may not converge signaling the onset of so-called replica symmetry breaking. Nonetheless, in Vilucchio et al. (2025) the authors rigorously show that the described procedure is correct whenever the replicon condition in Eq. (12) is satisfied.

864 Again one would need to generalize their proof (an AMP upper bound coupled with a
865 lower-bound in the line of Stojnic (2009); Thrampoulidis et al. (2014; 2020)) to multiple
866 tokens, but that should not pose any roadblock.

867 We devote the rest of the section to write down the suitable AMP algorithm used in the sketch of the
868 proof, in the more generic setting of multi-token generalized matrix sensing.

871 E.1 GENERIC SETTING

872 Consider the data model

$$873 y^\mu = g_0(\{\text{Tr}(S_0 Z_a^\mu)\}_{a=1}^{L_{\text{in}}}), \quad (28)$$

874 where Z_a^μ are $\text{GOE}(d)$ matrices, $S_0 \in \mathbb{R}^{d \times d}$ and $g_0 : \mathbb{R}^{L_{\text{in}}} \rightarrow \mathbb{R}^{L_{\text{out}}}$. g_0 can be a stochastic
875 function, in which case we assume that its stochasticity is independent and identical for each sample
876 μ . Consider the empirical risk minimization problem over $d \times d$ symmetric matrices

$$877 \hat{S} = \arg \min_{S \in \mathcal{C}} L(S), \quad L(S) = \sum_{\mu} \ell\left(y^\mu; \{\text{Tr}(S Z_a^\mu)\}_{a=1}^{L_{\text{in}}}\right) + R(S), \quad (29)$$

878 for $\ell : \mathbb{R}^{L_{\text{out}}} \times \mathbb{R}^{L_{\text{in}}} \rightarrow \mathbb{R}$, $R : \mathbb{R}^{d \times d} \rightarrow \mathbb{R}$ a rotationally invariant regularization (i.e. $R(S) =$
881 $R(OSO^T)$ for all d -dimensional rotation matrices O) and \mathcal{C} is a rotationally-invariant subset of the
882 set of $d \times d$ symmetric matrices.

883 **Special case I: main text model.** In the main text, we consider $L_{\text{in}} = T(T+1)/2$ (interpreted
884 as $T \times T$ symmetric matrices), $L_{\text{out}} = T^2$, g_0 is the row-wise softmax with Gaussian noise in the
885 input

$$886 g_0(A) = \sigma\left(\left\{\frac{A_{ab} + \sqrt{\Delta}\xi_{ab}}{\sqrt{2 - \delta_{ab}}}\right\}_{ab}\right) \quad (30)$$

887 where $\xi^\mu \in \mathbb{R}^{T \times T}$ is a symmetric standard Gaussian noise $\xi_{ab} = \xi_{ba} \sim N(0, 1)$ for all $1 \leq a \leq$
888 $b \leq T$,

$$889 \ell(y, \hat{y}) = \sum_{a,b=1}^T (y_{ab} - \hat{y}_{ab})^2 \quad \text{and} \quad R(S) = d\tilde{\lambda} \text{Tr}(S). \quad (31)$$

890 Finally, \mathcal{C} is the set of PSD symmetric $d \times d$ matrices.

891 **Special case II: main text model with linear attention.** In this case, we consider $L_{\text{in}} = T(T+1)/2$
892 and $L_{\text{out}} = T(T+1)/2$ (interpreted as $T \times T$ symmetric matrices), g_0 is

$$893 g_0(A)_{ab} = \frac{A_{ab} + \sqrt{\Delta}\xi_{ab}}{\sqrt{2 - \delta_{ab}}}, \quad (32)$$

894 where $\xi^\mu \in \mathbb{R}^{T \times T}$ is a symmetric standard Gaussian noise $\xi_{ab} = \xi_{ba} \sim N(0, 1)$ for all $1 \leq a \leq$
895 $b \leq T$, and

$$896 \ell(y, \hat{y}) = \sum_{a,b=1}^T (y_{ab} - \hat{y}_{ab})^2 \quad \text{and} \quad R(S) = d\tilde{\lambda} \text{Tr}(S). \quad (33)$$

897 Finally, \mathcal{C} is the set of PSD symmetric $d \times d$ matrices. The case of single token $T = 1$ reduces to
898 Erba et al. (2025).

909 E.2 MAPPING TO A VECTOR-WEIGHTS MODEL WITH COUPLED REGULARIZATION

910 We consider the mapping from $\text{vec} : \text{Sym}_d \rightarrow \mathbb{R}^{d(d+1)/2}$ (which conveniently maps the Frobenius
911 scalar product in Sym_d given by $\langle A; B \rangle = \text{Tr}(AB)$ to the standard Euclidean scalar product in
912 $\mathbb{R}^{d(d+1)/2}$) given by

$$913 \text{vec}(A)_{(ab)} = \langle b^{(ab)}; A \rangle = \sqrt{2 - \delta_{ab}} A_{ab}, \quad (34)$$

914 under the choice of orthonormal basis

$$915 b_{ij}^{(aa)} = \delta_{ia}\delta_{ja}, \quad b_{ij}^{(ab)} = \frac{\delta_{ia}\delta_{jb} + \delta_{ib}\delta_{ja}}{\sqrt{2}}. \quad (35)$$

Here (ab) stands for the ordered pair of $1 \leq a \leq b \leq d$, and we denote A_{ij} as the i, j entry of a matrix A , while as $A_{(ab)}$ the component of matrix A onto the basis element $b^{(ab)}$. Let us denote $d(d+1)/2 = D$ (we use $D \approx d^2/2$ as we are interested in the leading order in d).

Our input data is given, for each sample μ , by L_{in} $d \times d$ symmetric matrices $\{Z_a\}_{a=1}^{L_{\text{in}}}$ that in the Gaussian equivalent model we treat as independent GOEs. Then the sensing vectors satisfy

$$A_{a,(ij)} = \sqrt{\frac{d}{2}} \text{vec}(Z_a)_{(ij)} \sim N(0, \mathbb{I}_D) \quad (36)$$

for $1 \leq a \leq L_{\text{in}}$, and (ij) means the dimension indices $1 \leq i \leq j \leq d$. Moreover, if we define

$$w_{(ij)} = \sqrt{\frac{d}{2}} \text{vec}(S)_{(ij)}, \quad (37)$$

we have that

$$\begin{aligned} \text{Tr}(SZ_a) &= \sum_{(ij)} Z_{a,(ij)} S_{(ij)} = \sqrt{\frac{2}{d}} \sum_{(ij)} A_{a,(ij)} S_{(ij)} = \frac{\sqrt{2}}{\sqrt{D}} \sum_{(ij)} A_{a,(ij)} w_{(ij)}, \\ \frac{1}{d} \text{Tr}(S^2) &= \frac{1}{d} \sum_{(ij)} S_{(ij)}^2 = \frac{2}{d^2} \sum_{(ij)} w_{(ij)}^2 = \frac{1}{D} \sum_{(ij)} w_{(ij)}^2. \end{aligned} \quad (38)$$

In this formulation, under Gaussian equivalence, we see that our model is just a sequence single-index model of the form Cui et al. (2024) with coupled prior term.

Finally, the loss in the vectorial model reads

$$\tilde{\ell}(y = g_0(z_0), z) = \ell(y = g_0(\sqrt{2}z_0), \sqrt{2}z) \quad (39)$$

due to the factor $\sqrt{2}$ in Eq. (38).

E.2.1 APPROXIMATE MESSAGE PASSING

Under the mapping of Section E.2 we can directly combine the single token derivation of AMP given in Erba et al. (2025) (which holds for generic single-token losses and generic rotational-invariant regularization), with multi-token AMP given in Cui et al. (2024) to obtain the following AMP algorithm with fixed points given by local minima of the loss in Eq. (29). We stress that the multi-token extension can be mapped to the standard treatment of AMP, amounting to a non-separable non-linearity along the samples dimension. Hence the mapping from ERM to AMP of Erba et al. (2025) can be adapted to the multi-token case.

AMP in vector notation. In vector notation, the labels are generated as

$$y = g_0(\{\sqrt{2}D^{-1/2} \sum_{(ij)} A_{\mu,a,(ij)}^0 w_{(ij)}\}_{a=1}^{L_{\text{in}}}) \quad (40)$$

and the AMP reads (here $\mu = 1, \dots, n$, $a = 1, \dots, L_{\text{in}}$ and $1 \leq i \leq j \leq d$)

$$\begin{aligned} w_{(ij)}^t &= \sqrt{\frac{d}{2}} \text{vec} \left(\phi \left(\sqrt{\frac{2}{d}} \text{mat}(\Gamma^{t-1}), \Lambda^{t-1} \right) \right)_{(ij)} \\ V^t &= \frac{1}{D} (\text{div} \phi) \left(\sqrt{\frac{2}{d}} \text{mat}(\Gamma^{t-1}), \Lambda^{t-1} \right) \\ \omega_{\mu,a}^t &= \frac{1}{\sqrt{D}} \sum_{(ij)} A_{\mu,a,(ij)} w_{(ij)}^t - \theta(t \geq 1) V^t f_{\mu,a}^{t-1} \\ f_{\mu,a}^t &= \frac{\text{prox}(y_\mu, \omega_\mu^t, V^t)_a - \omega_{\mu,a}^t}{V} \\ \Lambda^t &= -\frac{1}{D} \sum_{a=1}^{L_{\text{in}}} \sum_{\mu=1}^n \partial_{\omega_{\mu,a}} f_{\mu,a}^t \\ \Gamma_{(ij)}^t &= \frac{1}{\sqrt{D}} \sum_{a=1}^{L_{\text{in}}} \sum_{\mu=1}^n A_{\mu,a,(ij)} f_{\mu,a}^t + \Lambda^t w_{(ij)}^t \end{aligned} \quad (41)$$

972 where

$$973 \text{prox}(y, \omega, V) = \underset{h \in \mathbb{R}^{L_{\text{in}}}}{\text{arginf}} \left\{ \frac{1}{2V} \sum_{a=1}^{L_{\text{in}}} (h_a - \omega_a)^2 + \ell(y, \sqrt{2}h) \right\} \quad (42)$$

$$974 \phi(M = ODO^T, \Lambda) = O \underset{T \in \mathcal{C}}{\text{argmin}} \left[\frac{1}{d^2} R(T) + \frac{\Lambda}{4} \sum_{i=1}^d T_i^2 - \frac{1}{2} \sum_{i=1}^d T_i D_i \right] O^T$$

975 where M is a d -dimensional symmetric matrix with eigen-decomposition $M = ODO^T$.

976 **AMP in matrix notation.** If we map back all quantities to their original matrix shape, we obtain
977 the following equivalent AMP algorithm. The labels are generated as

$$978 y^\mu = g_0(\{\text{Tr}(S_0 Z_a^\mu)\}_{a=1}^{L_{\text{in}}}) \quad (43)$$

979 and the AMP reads (here $\mu = 1, \dots, n$, $a = 1, \dots, L_{\text{in}}$ and $1 \leq i, j \leq d$)

$$980 S^t = \phi(\Theta^{t-1}, \Lambda^{t-1})$$

$$981 V^t = \frac{2}{d^2} (\text{div} \phi)(\Theta^{t-1}, \Lambda^{t-1})$$

$$982 \Omega_{\mu,a}^t = \text{Tr}(S^t Z_a^\mu) - \theta(t \geq 1) V^t F_{\mu,a}^{t-1}$$

$$983 F_{\mu,a}^t = \frac{\text{prox}_{\text{mat}}(y^\mu, \Omega_{\mu,a}^t, V^t)_a - \Omega_{\mu,a}^t}{V} \quad (44)$$

$$984 \Lambda^t = -\frac{2}{d^2} \sum_{a=1}^{L_{\text{in}}} \sum_{\mu=1}^n \partial_{\Omega_{\mu,a}} F_{\mu,a}^t$$

$$985 \Theta_{ij}^t = \frac{1}{d} \sum_{a=1}^{L_{\text{in}}} \sum_{\mu=1}^n Z_{a,ij}^\mu F_{\mu,a}^t + \Lambda^t S_{ij}^t$$

986 where ϕ is the same as in the vector case, while

$$987 \text{prox}(y, \omega, V)_{\text{mat}} = \underset{h \in \mathbb{R}^{L_{\text{in}}}}{\text{arginf}} \left\{ \frac{1}{4V} \sum_{a=1}^{L_{\text{in}}} (h_a - \omega_a)^2 + \ell(y, h) \right\}. \quad (45)$$

988 **State evolution.** The iterations of both algorithms can be tracked by the following state evolution
989 equations.

$$990 \begin{cases} \hat{q}^t &= \frac{2n}{d^2 (\hat{\Sigma}^t)^2} \mathbb{E}_{z_0, z} \sum_{a=1}^{L_{\text{in}}} (p_a - z_a)^2 \\ \hat{\Sigma}^t &= \frac{2n}{d^2 \hat{\Sigma}^t} \left[L_{\text{in}} - \mathbb{E}_{z_0, z} \sum_{a=1}^{L_{\text{in}}} \frac{Q_0 z_a p_a - m(z_0)_a p_a}{Q_0 q - m^2} \right] \\ \hat{m}^t &= \frac{2n}{d^2 \hat{\Sigma}^t} \mathbb{E}_{z_0, z} \sum_{a=1}^{L_{\text{in}}} \frac{q(z_0)_a p_a - m z_a p_a}{Q_0 q - m^2} \\ m^{t+1} &= -\partial_{\hat{m}} \Psi(\hat{\Sigma}^t, \hat{q}^t, \hat{m}^t) \\ q^{t+1} &= 2\partial_{\hat{\Sigma}} \Psi(\hat{\Sigma}^t, \hat{q}^t, \hat{m}^t) \\ \Sigma^{t+1} &= -2\partial_{\hat{q}} \Psi(\hat{\Sigma}^t, \hat{q}^t, \hat{m}^t) \end{cases} \quad (46)$$

991 where

$$992 p_a = \underset{h \in \mathbb{R}^{L_{\text{in}}}}{\text{arginf}} \left\{ \frac{1}{2\Sigma^t} \sum_{a=1}^{L_{\text{in}}} (h_a - \omega_a)^2 + \ell(g_0(\sqrt{2}\{z_0\}), \sqrt{2}h) \right\}, \quad (47)$$

$$993 \Psi(\hat{\Sigma}, \hat{q}, \hat{m}) = \frac{2}{d} \mathbb{E}_D \min_{T \in \mathcal{C}} \left[\frac{1}{d^2} R(T) + \frac{\hat{\Sigma}}{4} \sum_{i=1}^d T_i^2 - \frac{1}{2} \sum_{i=1}^d T_i D_i \right],$$

994 where the average \mathbb{E}_D is over the spectrum D of the matrix $\hat{m}S_0 + \sqrt{\hat{q}}Z$ with $Z \sim \text{GOE}(d)$,
995 while $\mathbb{E}_{z_0, z}$ is over two Gaussian vectors $z_0, z \in \mathbb{R}^{L_{\text{in}}}$ such that, independently for all components
996 $a = 1, \dots, L_{\text{in}}$

$$997 \begin{bmatrix} (z_0)_a \\ z_a \end{bmatrix} \sim \mathcal{N} \left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}; \begin{bmatrix} Q_0 & m^t \\ m^t & q^t \end{bmatrix} \right), \quad (48)$$

998 and over any stochasticity of the activation g_0 .

Observables. Call T^* is the optimizer in Ψ of Eq. (47). Then, by state evolution, at the global minimum \hat{W} of Eq. (29), we have that the spectral density of $\hat{W}\hat{W}^T/\sqrt{pd}$ converges to the one of T^* . Moreover, pre-activations converge to Gaussians of the form

$$\begin{bmatrix} \text{Tr}(S_0 Z_a) \\ \text{Tr}(S Z_a) \end{bmatrix} \sim \mathcal{N} \left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}; \begin{bmatrix} 2Q_0 & 2m \\ 2m & 2q \end{bmatrix} \right) \quad (49)$$

when evaluated on data Z that is not included in the training set. Otherwise, the target's pre-activations are still Gaussian with zero mean and variance $2Q_0$, but the learned model's pre-activations are given by

$$\text{Tr}(S Z_a) \sim \sqrt{2} p_a. \quad (50)$$

With that in mind, we have that the training error (without regularization term) at a fixed point of AMP is given by

$$e_{\text{train}} = \mathbb{E}_{z_0, z} \sum_{a=1}^{L_{\text{in}}} \ell \left(g_0(\sqrt{2}z_0); \sqrt{2}p \right), \quad (51)$$

The regularization part of the loss can be computed as $d^{-2}R(T^*)$, and any test loss (intended as a comparison of the output of the target and learned function) can be instead computed as

$$e_{\text{test}} = \mathbb{E}_{z_0, z} \sum_{a=1}^{L_{\text{in}}} \ell_{\text{test}} \left(g_0(\sqrt{2}z_0); \sqrt{2}z \right). \quad (52)$$

All averages are as defined for the state evolution equations.

Variational formulation. It can be checked Cui et al. (2024) that Eq. (9) is stationary at the fixed point of Eq. (46), and that it matches the training loss.

Replicon condition. The replicon condition is the linear stability condition of AMP under perturbations of a fixed points. A derivation is given in Erba et al. (2025). In our case, we have

$$\frac{2\alpha}{\hat{\Sigma}^2 \hat{\Sigma}^2} \mathbb{E}_{z_0, z} \sum_{\substack{a, b, c, d=1 \\ a \leq b, c \leq d}}^T (\partial_{z_{ab}} p_{cd} - \delta_{ac} \delta_{bd})^2 \frac{2}{d^2} \mathbb{E}_D \sum_{\substack{1 \leq i \leq j \leq d \\ 1 \leq k \leq l \leq d}} \partial_{M_{ij}} \phi(M = ODO^T, \hat{\Sigma}) < 1 \quad (53)$$

where the averages are the same as for state evolution.

E.3 SIMPLIFICATIONS IN THE MAIN TEXT SETTING

In the main text, we consider $L_{\text{in}} = T(T+1)/2$ (interpreted as $T \times T$ symmetric matrices), $L_{\text{out}} = T^2$, g_0 is the row-wise softmax with Gaussian noise in the input

$$g_0(A) = \sigma_{\beta_0} (\{[A_{ab} + \sqrt{\Delta} \xi_{ab}] / \sqrt{2 - \delta_{ab}}\}_{ab}), \quad (54)$$

where $\xi \in \mathbb{R}^{T \times T}$ is a symmetric standard Gaussian noise $\xi_{ab} = \xi_{ba} \sim N(0, 1)$ for all $1 \leq a \leq b \leq T$,

$$\ell(y, z) = \sum_{a, b=1}^T (y_{ab} - \sigma_{\beta} (z_{ab} / \sqrt{2 - \delta_{ab}}))^2 \quad \text{and} \quad R(S) = d\tilde{\lambda} \text{Tr}(S). \quad (55)$$

In this case,

$$\begin{aligned} \phi(M = ODO^T, \Lambda) &= O \text{diag} \left(\frac{\text{ReLU}(D_{ii} - 2\tilde{\lambda})}{\Lambda} \Big|_{i=1, \dots, d} \right) O^T \\ \Psi(\hat{\Sigma}, \hat{q}, \hat{m}) &= -\frac{\hat{m}^2}{2\hat{\Sigma}} J(\sqrt{\hat{q}}/\hat{m}, 2\tilde{\lambda}/\hat{m}) \\ J(a, b) &= \int_b^{+\infty} dx \mu_a(x) (x - b)^2. \end{aligned} \quad (56)$$

where $\mu_a = \mu_{\text{sc}} \boxplus \mu_0$. Notice that the noise can be treated simply by altering the second moment of z_0 in the state evolution equations from Q_0 to $Q_0 \rightarrow Q_0 + \Delta/2$ (where the factor 2 comes from gathering the $\sqrt{2}$ factor present in the state evolution equations).

Observables. The training and test errors are given by

$$\begin{aligned}
e_{\text{train}} &= \mathbb{E}_{z_0, z} \sum_{a=1}^{L_{\text{in}}} \|\tilde{\sigma}_{\beta_0}(\sqrt{2}z_0) - \tilde{\sigma}_{\beta}(\sqrt{2}p)\|_F^2, \\
e_{\text{test}} &= \mathbb{E}_{z_0, z} \sum_{a=1}^{L_{\text{in}}} \|\tilde{\sigma}_{\beta_0}(\sqrt{2}z_0) - \tilde{\sigma}_{\beta}(\sqrt{2}z)\|_F^2.
\end{aligned} \tag{57}$$

The spectral density of $\hat{W}\hat{W}^T/\sqrt{pd}$ converges to that of T^* , which is a shifted, rescaled and cropped version of the spectral density of $S_0 + \sqrt{\hat{q}}/\hat{m}Z$ for $Z \sim \text{GOE}(d)$, giving the expression in the main text.

State evolution. This gives the following state equations (at the fixed point)

$$\begin{aligned}
q &= \frac{\hat{m}^2}{\hat{\Sigma}^2} J\left(\frac{\sqrt{\hat{q}}}{\hat{m}}, \frac{2\tilde{\lambda}}{\hat{m}}\right) \\
m &= \frac{\hat{m}J\left(\frac{\sqrt{\hat{q}}}{\hat{m}}, \frac{2\tilde{\lambda}}{\hat{m}}\right) - \tilde{\lambda}\partial_2 J\left(\frac{\sqrt{\hat{q}}}{\hat{m}}, \frac{2\tilde{\lambda}}{\hat{m}}\right) - \frac{\sqrt{\hat{q}}}{2}\partial_1 J\left(\frac{\sqrt{\hat{q}}}{\hat{m}}, \frac{2\tilde{\lambda}}{\hat{m}}\right)}{\hat{\Sigma}} \\
\Sigma\hat{\Sigma} &= \frac{\hat{m}}{2\sqrt{\hat{q}}}\partial_1 J\left(\frac{\sqrt{\hat{q}}}{\hat{m}}, \frac{2\tilde{\lambda}}{\hat{m}}\right) \\
\Sigma\hat{\Sigma} &= 2\alpha L - 2\alpha\mathbb{E}_{z_0, z} \sum_{a=1}^{L_{\text{in}}} \frac{Q_0 z_a p_a - m z_{0,a} p_a}{Q_0 q - m^2} \\
\Sigma\hat{m} &= 2\alpha\mathbb{E}_{z_0, z} \sum_{a=1}^{L_{\text{in}}} \frac{q z_{0,a} p_a - m z_a p_a}{Q_0 q - m^2} \\
\hat{q}\Sigma^2 &= 2\alpha\mathbb{E}_{z_0, z} \sum_{a=1}^{L_{\text{in}}} (p_a - z_a)^2
\end{aligned} \tag{58}$$

where $\alpha = n/d^2$ and

$$p_a = \operatorname{arginf}_{h \in \mathbb{R}^{L_{\text{in}}}} \left\{ \frac{1}{2\Sigma} \sum_{a=1}^{L_{\text{in}}} (h_a - \omega_a)^2 + \sum_{a=1}^{L_{\text{out}}} (\tilde{\sigma}_{\beta_0}(\sqrt{2}z_0)_a - \tilde{\sigma}_{\beta}(\sqrt{2}h)_a)^2 \right\}. \tag{59}$$

For the replicon, the only thing that changes is that the regularization dependent part becomes the same as presented in Erba et al. (2025), giving the expression of Claim B.1.

Small regularization limit $\lambda \rightarrow 0^+$. To study the small regularization limit before interpolation (where a vanishing regularization would lead to degenerate global minima), one can perform the

change of variables $\hat{m} \rightarrow \hat{m}\tilde{\lambda}$, $\hat{q} \rightarrow \hat{q}\tilde{\lambda}^2$, $\Sigma \rightarrow \Sigma/\tilde{\lambda}$ and $\hat{\Sigma} \rightarrow \hat{\Sigma}\tilde{\lambda}$ to obtain

$$\begin{aligned}
q &= \frac{\hat{m}^2}{\hat{\Sigma}^2} J\left(\frac{\sqrt{\hat{q}}}{\hat{m}}, \frac{2}{\hat{m}}\right) \\
m &= \frac{\hat{m} J\left(\frac{\sqrt{\hat{q}}}{\hat{m}}, \frac{2}{\hat{m}}\right) - \tilde{\lambda} \partial_2 J\left(\frac{\sqrt{\hat{q}}}{\hat{m}}, \frac{2}{\hat{m}}\right) - \frac{\sqrt{\hat{q}}}{2} \partial_1 J\left(\frac{\sqrt{\hat{q}}}{\hat{m}}, \frac{2}{\hat{m}}\right)}{\hat{\Sigma}} \\
\Sigma \hat{\Sigma} &= \frac{\hat{m}}{2\sqrt{\hat{q}}} \partial_1 J\left(\frac{\sqrt{\hat{q}}}{\hat{m}}, \frac{2}{\hat{m}}\right) \\
\Sigma \hat{\Sigma} &= 2\alpha L - 2\alpha \mathbb{E}_{z_0, z} \sum_{a=1}^{L_{\text{in}}} \frac{Q_0 z_a p_a - m z_{0,a} p_a}{Q_0 q - m^2} \\
\Sigma \hat{m} &= 2\alpha \mathbb{E}_{z_0, z} \sum_{a=1}^{L_{\text{in}}} \frac{q z_{0,a} p_a - m z_a p_a}{Q_0 q - m^2} \\
\hat{q} \Sigma^2 &= 2\alpha \mathbb{E}_{z_0, z} \sum_{a=1}^{L_{\text{in}}} (p_a - z_a)^2
\end{aligned} \tag{60}$$

with

$$p_a = \operatorname{arginf}_{h \in \mathbb{R}^{L_{\text{in}}}} \left\{ \frac{\tilde{\lambda}}{2\Sigma} \sum_{a=1}^{L_{\text{in}}} (h_a - \omega_a)^2 + \sum_{a=1}^{L_{\text{out}}} (\tilde{\sigma}_{\beta_0}(\sqrt{2}z_0)_a - \tilde{\sigma}_{\beta}(\sqrt{2}h)_a)^2 \right\} \tag{61}$$

We then see that for $\tilde{\lambda} \rightarrow 0^+$ the last expression reduces to

$$p_a = \operatorname{arginf}_{h \in \mathbb{R}^{L_{\text{in}}}} \left\{ \sum_{a=1}^{L_{\text{out}}} (\tilde{\sigma}_{\beta_0}(\sqrt{2}z_0)_a - \tilde{\sigma}_{\beta}(\sqrt{2}h)_a)^2 \right\} \tag{62}$$

with the prescription that if such arginf is degenerate, one should pick the arginf closest to z in L2 distance. This set of equations is valid as long as $\Sigma > 0$ and not diverging. At interpolation we expect $\Sigma \rightarrow 0$, as there the curvature of the loss (proportional to Σ^{-1}) diverges.

After interpolation, there is no need to change variable. One can put directly $\lambda \rightarrow 0^+$ in the original Eq. (46). This set of equations is valid as long as $\Sigma > 0$ and not diverging. At interpolation we expect $\Sigma \rightarrow +\infty$, as there the curvature of the loss (proportional to Σ^{-1}) goes to zero.

F REDUCTION TO LINEAR ATTENTION

Let us consider the setting of the main text in the $\lambda \rightarrow 0^+$ limit, before interpolation (i.e. when there exists multiple sets of weights that perfectly fit the training dataset). Then, we need to solve Eq. (68). We now show that here, for the case of softmax-softmax studied in the main text, the equations reduce to a rescaled version of the state equations for the case of linear single token attention Erba et al. (2025). We consider w.l.o.g. the case $\beta = \beta_0$. Let us consider the proximal

$$p_a = \operatorname{arginf}_{h \in \mathbb{R}^{L_{\text{in}}}} \left\{ \sum_{a=1}^{L_{\text{out}}} (\tilde{\sigma}_{\beta_0}(z_0)_a - \tilde{\sigma}_{\beta}(h)_a)^2 \right\}, \tag{63}$$

where we recall that $\tilde{\sigma}_{\beta}(A)_{ab} = \sigma_{\beta}(\sqrt{1 + \delta_{ab}} A_{ab})$ for any matrix $A \in \mathbb{R}^{T \times T}$. For the softmax activation, it is easy to see that the minimum is achieved at

$$p_{ab} = z_{0,ab} + \frac{\bar{a}}{\sqrt{1 + \delta_{ab}}} \tag{64}$$

for any scalar \bar{a} . Thus, we have a continuum of global minima, and we need to pick the one closest in L2 distance to z , giving

$$p_{ab} = z_{0,ab} - \frac{2}{T^2} \sum_{c \leq d} \frac{z_{0,cd} - z_{cd}}{\sqrt{(1 + \delta_{cd})(1 + \delta_{ab})}} \tag{65}$$

which provides striking simplifications in the equations, that reduce to

$$\begin{cases} q &= \frac{\hat{m}^2}{\hat{\Sigma}^2} J\left(\frac{\sqrt{\hat{q}}}{\hat{m}}, \frac{2}{\hat{m}}\right) \\ m &= \frac{\hat{m} J\left(\frac{\sqrt{\hat{q}}}{\hat{m}}, \frac{2}{\hat{m}}\right) - \hat{\lambda} \partial_2 J\left(\frac{\sqrt{\hat{q}}}{\hat{m}}, \frac{2}{\hat{m}}\right) - \frac{\sqrt{\hat{q}}}{2} \partial_1 J\left(\frac{\sqrt{\hat{q}}}{\hat{m}}, \frac{2}{\hat{m}}\right)}{\hat{\Sigma}} \\ \Sigma \hat{\Sigma} &= \frac{\hat{m}}{2\sqrt{\hat{q}}} \partial_1 J\left(\frac{\sqrt{\hat{q}}}{\hat{m}}, \frac{2}{\hat{m}}\right) \\ \Sigma \hat{\Sigma} &= 2\alpha(L_{\text{in}} - 1) \\ \hat{m}\Sigma &= 2\alpha(L_{\text{in}} - 1) \\ \hat{q}\Sigma^2 &= 2\alpha(L_{\text{in}} - 1)(Q_0 - 2m + q) \end{cases} \quad (66)$$

It is easy to see that instead, in the linear case, $p = z_0$, leading to the set of equations

$$\begin{cases} q &= \frac{\hat{m}^2}{\hat{\Sigma}^2} J\left(\frac{\sqrt{\hat{q}}}{\hat{m}}, \frac{2}{\hat{m}}\right) \\ m &= \frac{\hat{m} J\left(\frac{\sqrt{\hat{q}}}{\hat{m}}, \frac{2}{\hat{m}}\right) - \hat{\lambda} \partial_2 J\left(\frac{\sqrt{\hat{q}}}{\hat{m}}, \frac{2}{\hat{m}}\right) - \frac{\sqrt{\hat{q}}}{2} \partial_1 J\left(\frac{\sqrt{\hat{q}}}{\hat{m}}, \frac{2}{\hat{m}}\right)}{\hat{\Sigma}} \\ \Sigma \hat{\Sigma} &= \frac{\hat{m}}{2\sqrt{\hat{q}}} \partial_1 J\left(\frac{\sqrt{\hat{q}}}{\hat{m}}, \frac{2}{\hat{m}}\right) \\ \Sigma \hat{\Sigma} &= 2\alpha L_{\text{in}} \\ \hat{m}\Sigma &= 2\alpha L_{\text{in}} \\ \hat{q}\Sigma^2 &= 2\alpha L_{\text{in}}(Q_0 - 2m + q). \end{cases} \quad (67)$$

We thus see that, calling $\alpha_{\text{soft}} = (L_{\text{in}} - 1)\alpha$ and $\alpha_{\text{linear}} = L_{\text{in}}\alpha$, one get the same equations as the single-token case Erba et al. (2025).

This implies immediately Corollary B.2, giving in particular the expression for the interpolation threshold up to which the mapping discussed in this Appendix holds. It also implies immediately that the low-rank limit $\kappa_0 \rightarrow 0$ reduces to Result 2 of Erba et al. (2025) (modulo a rescaling).

G NON-FACTORIZED ATTENTION

The case of non-factorized attention falls in the formalism of Appendix E.1, with state evolution equations given by

$$\begin{aligned} q &= \frac{\hat{m}^2}{(\hat{\Sigma} + 4\tau)^2} \left(Q_0 + \frac{\hat{q}}{\hat{m}^2} \right) \\ m &= \frac{Q_0 \hat{m}}{\hat{\Sigma} + 4\tau} \\ \Sigma &= \frac{1}{\hat{\Sigma} + 4\tau} \\ \Sigma \hat{\Sigma} &= 2\alpha L - 2\alpha \mathbb{E}_{z_0, z} \sum_{a=1}^{L_{\text{in}}} \frac{Q_0 z_a p_a - m z_{0,a} p_a}{Q_0 q - m^2} \\ \Sigma \hat{m} &= 2\alpha \mathbb{E}_{z_0, z} \sum_{a=1}^{L_{\text{in}}} \frac{q z_{0,a} p_a - m z_a p_a}{Q_0 q - m^2} \\ \hat{q}\Sigma^2 &= 2\alpha \mathbb{E}_{z_0, z} \sum_{a=1}^{L_{\text{in}}} (p_a - z_a)^2 \end{aligned} \quad (68)$$

with τ being the Frobenius regularization. The equations can be derived by noticing that

$$\begin{aligned}
\Psi(\hat{\Sigma}, \hat{q}, \hat{m}) &= \frac{2}{d} \mathbb{E}_D \min_{T \in \mathcal{C}} \left[\frac{1}{d^2} R(T) + \frac{\hat{\Sigma}}{4} \sum_{i=1}^d T_i^2 - \frac{1}{2} \sum_{i=1}^d T_i D_i \right] \\
&= \frac{1}{d} \mathbb{E}_D \min_{T \in \mathcal{C}} \left[\frac{\hat{\Sigma} + 4\tau}{2} \sum_{i=1}^d T_i^2 - \sum_{i=1}^d T_i D_i \right] \\
&= -\frac{1}{2(\hat{\Sigma} + 4\tau)} \frac{1}{d} \mathbb{E}_D \sum_{i=1}^d T_i^2 \\
&= -\frac{Q_0 \hat{m}^2 + \hat{q}}{2(\hat{\Sigma} + 4\tau)}
\end{aligned} \tag{69}$$

while the loss-related equations are the same as in the case of the main text.

In the limit $\tau \rightarrow 0^+$, the last three equations reduce as in the case of Appendix F, giving a set of equations that can be solved and gives (in the noiseless case)

$$\|\hat{S} - S_0\|_F^2 = Q_0 - 2m + q = Q_0(1 - 2\alpha) \tag{70}$$

giving the strong recovery at $\alpha = 1/2$ and proving the claim that if $\kappa_0 \rightarrow 0^+$ learning for the non-factorized model happens only at scale $\alpha = \mathcal{O}(1)$, i.e. $n = \mathcal{O}(d^2)$.

H DERIVATION OF THE ERROR DECOMPOSITION OF SECTION 4

To begin with, the generalization error is given by

$$\mathcal{I}(m, q) = \mathbb{E}_{z_0, z, \xi} \|\sigma_1(z_0 + \sqrt{\Delta/2}\zeta) - \sigma_2(z)\|^2, \tag{71}$$

where $(z_0, z) \sim \mathcal{N}\left(0, \begin{pmatrix} Q_0 & m \\ m & q \end{pmatrix}\right)$, $\zeta \sim \mathcal{N}(0, 1)$ and for simplicity we denote $\sigma_1 := \tilde{\sigma}_{\beta_0}$ to be the teacher activation and $\sigma_2 := \tilde{\sigma}_{\beta}$ to be the student activation.

We further assume that Claim B.1 is valid for any n, d large enough, as specified by the following conjecture.

Conjecture H.1. *Let $\lambda > 0$, $\Delta \geq 0$ and consider $n, d \gg 1$ sufficiently large. Then with high probability the generalization error at the global minimum is given by*

$$e_{\text{test}}(\hat{W}) = \mathcal{I}(m_0, q_0) + (\mathcal{I}(m^*, q^*) - \mathcal{I}(m_0, q_0))(1 + o_{n,d}(1)), \tag{72}$$

where m_0, q_0 independent of n, d are the zeroth orders of m^*, q^* (i.e. $m^* = m_0(1 + o_{n,d}(1))$, $q^* = q_0(1 + o_{n,d}(1))$) and m^*, q^* are obtained as the global minimizer of

$$\Phi(\Sigma, m, q, \hat{\Sigma}, \hat{m}, \hat{q}) = \frac{\hat{q}\Sigma + 2\hat{m}m - \hat{\Sigma}q}{4} + \frac{n}{d^2} \mathcal{M}(\Sigma, m, q) - \frac{\hat{m}^2}{4\hat{\Sigma}} J\left(\frac{\sqrt{\hat{q}}}{\hat{m}}, \frac{2\tilde{\lambda}}{\hat{m}}\right). \tag{73}$$

In another word, we conjecture that the generalization error is given by

$$e_{\text{test}}(\hat{W}) \approx \mathcal{I}(m^*, q^*), \tag{74}$$

up to the first order and m^*, q^* are obtained by minimizing (9) at large but finite n, d .

H.1 SHARP MINIMUM: $\Sigma \ll 1$

When $\Sigma \ll 1$ we approximately have

$$\mathcal{M}(\Sigma, m, q) \approx \mathbb{E}_{z_0, z} (\|\sigma_1(z_0) - \sigma_2(z)\|^2 - 4\Sigma \|\mathcal{J}_{\sigma_2}(z)(z - z_0)\|^2), \tag{75}$$

where z, z_0 are regarded as $\frac{1}{2}T(T+1)$ -dimensional vector and thus $\mathcal{J}_{\sigma_2} \in \mathbb{R}^{\frac{1}{2}T(T+1) \times \frac{1}{2}T(T+1)}$ denotes the derivative of σ_1 .

Under the assumption that the cosine similarity between the teacher and the student is close to 1, i.e.,

$$q - \frac{m^2}{Q_0} \ll Q_0, \quad (76)$$

we can write $z_0 = \sqrt{\frac{\Delta}{2}}\zeta_1 + \sqrt{Q_0}\zeta_0$ and $z = \frac{m}{\sqrt{Q_0}}\zeta_0 + \sqrt{q - \frac{m^2}{Q_0}}\zeta$ with $\zeta, \zeta_0, \zeta_1 \sim \mathcal{N}(0, 1)$ independent of each other and then expand (75) as

$$\begin{aligned} & \mathbb{E}_{z_0, z} \|\sigma_1(z_0) - \sigma_2(z)\|^2 \\ & \approx \mathbb{E}_{\zeta, \zeta_0, \zeta_1} \left\| \sigma_1 \left(\sqrt{\frac{\Delta}{2}}\zeta_1 + \sqrt{Q_0}\zeta_0 \right) - \sigma_2 \left(\frac{m}{\sqrt{Q_0}}\zeta_0 \right) + \mathcal{J}_{\sigma_2} \left(\frac{m}{\sqrt{Q_0}}\zeta_0 \right) \sqrt{q - \frac{m^2}{Q_0}}\zeta \right\|^2 \\ & = \mathbb{E}_{\zeta_0, \zeta_1} \left\| \sigma_1 \left(\sqrt{\frac{\Delta}{2}}\zeta_1 + \sqrt{Q_0}\zeta_0 \right) - \sigma_2 \left(\frac{m}{\sqrt{Q_0}}\zeta_0 \right) \right\|^2 + \left(q - \frac{m^2}{Q_0} \right) \mathbb{E}_{\zeta_0} \left\| \mathcal{J}_{\sigma_2} \left(\frac{m}{\sqrt{Q_0}}\zeta_0 \right) \right\|_F^2, \end{aligned} \quad (77)$$

and

$$\begin{aligned} & \mathbb{E}_{z_0, z} \|\mathcal{J}_{\sigma_2}(z)(z - z_0)\|^2 \\ & = \mathbb{E}_{\zeta, \zeta_0, \zeta_1} \left\| \mathcal{J}_{\sigma_2} \left(\sqrt{\frac{\Delta}{2}}\zeta_1 + \sqrt{Q_0}\zeta_0 \right) \left((\sqrt{Q_0} - \frac{m}{\sqrt{Q_0}})\zeta_0 + \sqrt{\frac{\Delta}{2}}\zeta_1 - \sqrt{q - \frac{m^2}{Q_0}}\zeta \right) \right\|^2 \\ & \approx \mathbb{E}_{\zeta_0} \|\mathcal{J}_{\sigma_2}(\sqrt{Q_0 + \Delta/2}\zeta_0)\|_F^2 \left(q - \frac{m^2}{Q_0} \right) + \mathbb{E}_{\zeta_0, \zeta_1} \left\| \mathcal{J}_{\sigma_2} \left(\sqrt{\frac{\Delta}{2}}\zeta_1 + \sqrt{Q_0}\zeta_0 \right) \left((\sqrt{Q_0} - \frac{m}{\sqrt{Q_0}})\zeta_0 + \sqrt{\frac{\Delta}{2}}\zeta_1 \right) \right\|_F^2, \end{aligned} \quad (78)$$

which gives

$$\mathcal{M}(\Sigma, m, q) \approx \kappa_0(m) + \kappa_1(m)\left(q - \frac{m^2}{Q_0}\right) - 4\Sigma\kappa_2(m) \quad (79)$$

up to the first order of Σ and $q - \frac{m^2}{Q_0}$, where

$$\kappa_0(m) := \mathbb{E}_{\zeta_0, \zeta_1} \left\| \sigma_1 \left(\sqrt{Q_0}\zeta_0 + \sqrt{\frac{\Delta}{2}}\zeta_1 \right) - \sigma_2 \left(\frac{m}{\sqrt{Q_0}}\zeta_1 \right) \right\|^2 = \mathcal{I}(m, m^2/Q_0), \quad (80)$$

$$\kappa_1(m) := \mathbb{E}_{\zeta} \left\| \mathcal{J}_{\sigma_2} \left(\frac{m}{\sqrt{Q_0}}\zeta \right) \right\|_F^2 \quad (81)$$

and

$$\kappa_2(m) := \mathbb{E}_{\zeta, \zeta_0} \left\| \mathcal{J}_{\sigma_2} \left(\sqrt{Q_0}\zeta_0 + \sqrt{\frac{\Delta}{2}}\zeta \right) \left((\sqrt{Q_0} - \frac{m}{\sqrt{Q_0}})\zeta_0 + \sqrt{\frac{\Delta}{2}}\zeta \right) \right\|_F^2. \quad (82)$$

Then the saddle point equations for (9) are

$$\begin{cases} \hat{\Sigma} = 4\alpha\kappa_1 \\ \hat{m} = 2\alpha\left(\kappa_1 \frac{2m}{Q_0} + \kappa_0'(m)\right) \\ \hat{q} = 4\alpha\kappa_2 \\ m = \partial_{\hat{m}} \frac{\hat{m}^2}{2\hat{\Sigma}} J\left(\frac{\sqrt{\hat{q}}}{\hat{m}}, \frac{2\hat{\lambda}}{\hat{m}}\right) \\ q = \left(\frac{\hat{m}}{\hat{\Sigma}}\right) J\left(\frac{\sqrt{\hat{q}}}{\hat{m}}, \frac{2\hat{\lambda}}{\hat{m}}\right) \\ \Sigma = \frac{\hat{m}^2}{\hat{\Sigma}} \partial_{\hat{q}} J\left(\frac{\sqrt{\hat{q}}}{\hat{m}}, \frac{2\hat{\lambda}}{\hat{m}}\right). \end{cases} \quad (83)$$

At the leading order we have $J \approx Q_0$ because the estimator is given by $\hat{W}^T \hat{W} / \sqrt{pd} = \text{ReLU}\left(\frac{\hat{m}}{\hat{\Sigma}}S_0 + \frac{\sqrt{\hat{q}}}{\hat{\Sigma}}Z - \frac{2\hat{\lambda}}{\hat{\Sigma}}\right)$ in distribution and J describes the second moment of $\text{ReLU}\left(S_0 + \frac{\sqrt{\hat{q}}}{\hat{m}}Z - \frac{2\hat{\lambda}}{\hat{m}}\right)$. Then $m^2 \approx qQ_0$ is equivalent to $J \approx Q_0$. One can also check at if we use $J \approx Q_0$, we have $\text{Tr}[\hat{W}^T \hat{W} S_0 / \sqrt{pd}] = \frac{\hat{m}}{\hat{\Sigma}}Q_0 = m$ in the leading order.

Then taking the first two equations of (83) into the last three equations of (83) gives the first order solutions

$$\begin{cases} m_0 &= \frac{2\kappa_1 m_0 / Q_0 + \kappa'_0(m_0)}{2\kappa_1} Q_0, \\ q_0 &= \left(\frac{2\kappa_1 m_0 / Q_0 + \kappa'_0(m_0)}{2\kappa_1} \right)^2 Q_0, \\ \Sigma &= 0 \end{cases} \quad (84)$$

consistent with the ansatz $q_0 = \frac{m_0^2}{Q_0}$ and $\Sigma \ll 1$. Then the leading order solution is given by $q_0 = \frac{m_0^2}{Q_0}$ with m_0 satisfying

$$\kappa'_0(m_0) = 0, \quad (85)$$

or equivalently

$$m_0 = \arg \inf_m \mathcal{I}(m, m^2 / Q_0). \quad (86)$$

This gives the zeroth order generalization error

$$e_0 = \inf_m \mathcal{I}(m, m^2 / Q_0). \quad (87)$$

Now we consider the first order correction. The excess risk is given by

$$\begin{aligned} e_1 &:= \frac{\partial \mathcal{I}(m, q)}{\partial m} \Big|_{m_0, q_0} m_1 + \frac{\partial \mathcal{I}(m, q)}{\partial q} \Big|_{m_0, q_0} q_1 \\ &= \frac{\chi}{\eta} (q_1 - 2m_1), \end{aligned} \quad (88)$$

where m_1, q_1 are the first order corrections, $\eta := \frac{\hat{m}}{\hat{\Sigma}} = \frac{m_0}{Q_0}$ and

$$\chi := \mathbb{E} \left[\text{Tr} \left(\mathcal{J}_{\sigma_1} \left(z_0 + \sqrt{\frac{\Delta}{2}} \xi \right) \mathcal{J}_{\sigma_2} (\eta z_0)^T \right) \right]. \quad (89)$$

The first order corrections of fourth and fifth equations of (83) are

$$\begin{aligned} m_1 &= \frac{\hat{m}}{\hat{\Sigma}} (J(\delta, \epsilon) - Q_0) - \frac{1}{2\hat{\Sigma}} (\sqrt{\hat{q}} \partial_1 J(\delta, \epsilon) + \tilde{\lambda} \partial_2 J(\delta, \epsilon)) \\ &= \eta (J(\delta, \epsilon) - Q_0) - \frac{\eta}{2} (\delta \partial_1 J(\delta, \epsilon) + \epsilon \partial_2 J(\delta, \epsilon)) \end{aligned} \quad (90)$$

and

$$q_1 = \left(\frac{\hat{m}}{\hat{\Sigma}} \right)^2 (J(\delta, \epsilon) - Q_0) = \eta^2 (J(\delta, \epsilon) - Q_0), \quad (91)$$

and thus

$$e_1 = \eta \chi ((\delta \partial_1 + \epsilon \partial_2 - 1) J(\delta, \epsilon) + Q_0), \quad (92)$$

where we define

$$\delta := \frac{\sqrt{\hat{q}}}{\hat{m}} = \frac{\sqrt{\kappa_2(m_0)}}{\sqrt{\alpha \eta \kappa_1(m_0)}} \quad (93)$$

and

$$\epsilon := \frac{2\tilde{\lambda}}{\hat{m}} = \frac{\tilde{\lambda}}{2\alpha \kappa_1(m_0) \eta}. \quad (94)$$

Therefore, the excess risk can be written as

$$e_1 := \eta \chi ((\delta \partial_1 + \epsilon \partial_2 - 1) J(\delta, \epsilon) + Q_0). \quad (95)$$

Finally, assume that the spectrum is given by the BBP approximation, i.e., for the under-regularization case

$$J(\delta, \epsilon) = \frac{1}{d} \sum_{i=1}^{K(\delta)} \left(s_i + \frac{\delta^2}{s_i} - \epsilon \right)^2 + \delta^2 \int_{\epsilon/\delta}^2 \mu_{\text{sc}}(x) (x - \epsilon/\delta)^2 \quad (96)$$

with a cutoff $K(\delta) \ll d$ satisfying $s_{K(\delta)} = \delta$, or for the over-regularization case

$$J(\delta, \epsilon) = \frac{1}{d} \sum_{i=1}^{K(\delta, \epsilon)} \left(s_i + \frac{\delta^2}{s_i} - \epsilon \right)^2 \quad (97)$$

with a cutoff $K(\delta, \epsilon) \ll d$ satisfying $s_{K(\delta, \epsilon)} + \frac{\delta^2}{s_{K(\delta, \epsilon)}} - \epsilon = 0$. Then we can borrow Result 3 of Defilippis et al. (2025) to write

$$e_1 = \eta\chi \left(\underbrace{\delta^2 \int_{\epsilon/\delta}^2 \mu_{sc}(dx) \left(x - \frac{\epsilon}{\delta} \right)^2}_{\text{overfitting (learned noise)}} + \underbrace{\frac{1}{d} \delta K'(\delta) (2\delta - \epsilon)^2 + \frac{1}{d} \sum_{i=K(\delta)+1}^d s_i^2}_{\text{underfitting (not learned features)}} + \underbrace{\frac{1}{d} \sum_{i=1}^{K(\delta)} \left[\left(\frac{\delta^2}{s_i} - \epsilon \right)^2 + \frac{\delta^2}{s_i} \left(s_i + \frac{\delta^2}{s_i} - \epsilon \right) \right]}_{\text{approximation error for learned features}} \right) \quad (98)$$

for the under-regularization case and

$$e_1 = \eta\chi \left(\underbrace{\frac{1}{d} \sum_{i=K(\delta, \epsilon)+1}^d s_i^2}_{\text{underfitting (not learned features)}} + \underbrace{\frac{1}{d} \sum_{i=1}^{K(\delta, \epsilon)} \left[\left(\frac{\delta^2}{s_i} - \epsilon \right)^2 + \frac{\delta^2}{s_i} \left(s_i + \frac{\delta^2}{s_i} - \epsilon \right) \right]}_{\text{approximation error for learned features}} \right) \quad (99)$$

for the over-regularization case. This finishes the derivation of the error decomposition for the case $\Sigma \ll 1$.

When is it the global minimum? One thing to notice is that the plateau (87) is not necessarily equal to the oracle generalization error

$$\min_{m, q: m^2 \leq q Q_0} \mathcal{I}(m, q). \quad (100)$$

Thus it only describes one saddle point rather than the global minimum.

Here we only point out that a sufficient condition for (87) to be the oracle generalization error is when the activations are conservative monotone operators, i.e., if

$$\sigma_1 = \nabla \Phi_1, \quad \sigma_2 = \nabla \Phi_2 \quad (101)$$

for functions Φ_1, Φ_2 which are both convex or both concave. This can be easily proven by

$$\begin{aligned} \frac{\partial \mathcal{I}(m, q)}{\partial m} &= -\mathbb{E}_{z_0, z} \frac{\partial \sigma_1(z_0 + \sqrt{\Delta/2}\xi)^T \sigma_2(z)}{\partial m} \\ &= -\sum_{k=1}^d \mathbb{E} \left[\frac{\partial^2}{\partial y_k \partial z_k} (\sigma_1(y)^T \sigma_2(z)) \right] \\ &= -\mathbb{E} \left[\text{Tr} \left(\mathcal{J}_{\sigma_1}(y) (\mathcal{J}_{\sigma_1}(z))^T \right) \right] \leq 0, \end{aligned} \quad (102)$$

where we denote $y := z_0 + \sqrt{\Delta/2}\xi$ and use the Stein's lemma. The last inequality is because if σ_1, σ_2 are conservative monotone, $\mathcal{J}_{\sigma_1}, \mathcal{J}_{\sigma_1}$ are positive/negative semi-definite. Then the oracle generalization error is achieved when $m = \sqrt{q Q_0}$, which proves that e_0 is equal to the oracle generalization error.

The condition can be relaxed to row-wise or column-wise conservative monotone, because we can write

$$\mathcal{I}(m, q) = \sum_{\ell=1}^d \mathbb{E}_{z_0, z, \zeta} (\sigma_{1\ell}(z_0 + \sqrt{\Delta/2}\zeta) - \sigma_{2\ell}(z))^2. \quad (103)$$

Then the softmax activation satisfies this condition because it is row-wise conservative monotone: $\text{softmax}(x) = \nabla \log \sum_i e^{x_i}$.

1458 **The potential global minimum** In general we can write the SE at $\Sigma \ll 1$ as

$$1459 \begin{cases} \hat{\Sigma} = 4\alpha \frac{\partial \mathcal{I}(m, q)}{\partial m} \\ \hat{m} = -2\alpha \frac{\partial \mathcal{I}(m, q)}{\partial q} \\ \hat{q} = 4\alpha \kappa_2(m, q) \\ m = \partial_{\hat{m}} \frac{\hat{m}^2}{2\hat{\Sigma}} J\left(\frac{\sqrt{\hat{q}}}{\hat{m}}, \frac{2\lambda}{\hat{m}}\right) \\ q = \left(\frac{\hat{m}}{\hat{\Sigma}}\right)^2 J\left(\frac{\sqrt{\hat{q}}}{\hat{m}}, \frac{2\lambda}{\hat{m}}\right) \\ \Sigma = \frac{\hat{m}^2}{\hat{\Sigma}} \partial_{\hat{q}} J\left(\frac{\sqrt{\hat{q}}}{\hat{m}}, \frac{2\lambda}{\hat{m}}\right), \end{cases} \quad (104)$$

1468 where

$$1469 \kappa_2(m, q) := \mathbb{E}_{z, z_0} \|\mathcal{J}_{\sigma_2}(z)(z - z_0)\|^2. \quad (105)$$

1471 For the global minimum, the plateau is given by

$$1472 e_0 = \inf_{m, q: m^2 \leq q Q_0} \mathcal{I}(m, q). \quad (106)$$

1474 Then if the ansatz $m^2 = q Q_0$ is not valid at the global minimum, we must have

$$1476 \frac{\partial \mathcal{I}(m, q)}{\partial m} \Big|_{m_0, q_0} = \frac{\partial \mathcal{I}(m, q)}{\partial q} \Big|_{m_0, q_0} = 0, \quad (107)$$

1478 which gives the first order solution m_0, q_0 . Then we have $\hat{q}_0 = 4\alpha \kappa_2(m_0, q_0)$ and $\hat{m}_0, \hat{\Sigma}_0$ are solved self-consistently from

$$1481 m_0 = \partial_{\hat{m}_0} \frac{\hat{m}_0^2}{2\hat{\Sigma}_0} J\left(\frac{\sqrt{\hat{q}_0}}{\hat{m}_0}, \frac{2\lambda}{\hat{m}_0}\right), \quad q_0 = \left(\frac{\hat{m}_0}{\hat{\Sigma}_0}\right)^2 J\left(\frac{\sqrt{\hat{q}_0}}{\hat{m}_0}, \frac{2\lambda}{\hat{m}_0}\right). \quad (108)$$

1484 The excess risk is then given by

$$1485 e_1 := \frac{1}{2} \frac{\partial^2 \mathcal{I}(m, q)}{\partial m^2} \Big|_{m_0, q_0} m_1^2 + \frac{1}{2} \frac{\partial^2 \mathcal{I}(m, q)}{\partial q^2} \Big|_{m_0, q_0} q_1^2 + \frac{\partial^2 \mathcal{I}(m, q)}{\partial m \partial q} \Big|_{m_0, q_0} m_1 q_1, \quad (109)$$

1488 where m_1, q_1 are the first order corrections. It is left as the future work to explicitly calculate the excess risk in this scenario.

1491 H.2 FLAT MINIMUM: $\Sigma \gg 1$

1493 When $\Sigma \gg 1$, we approximately have

$$1494 \mathcal{M}(\Sigma, m, q) = \frac{1}{2\Sigma} \mathcal{K}(m, q), \quad (110)$$

1497 where

$$1498 \mathcal{K}(m, q) := \mathbb{E}_{z_0, z, \zeta} \inf_{h \in \arg \inf \|\sigma_1(z_0 + \sqrt{\Delta/2}\zeta) - \sigma_2(\cdot)\|^2} \|h - z\|^2. \quad (111)$$

1500 For example, when $\sigma_2(x) = \sigma_1(\beta^{-1}x)$ and σ_1 is a one-to-one mapping, we have

$$1502 \mathcal{M}(\Sigma, m, q) = \mathbb{E}_{z_0, z} \frac{1}{2\Sigma} \sum_{a \leq b}^T (\beta z_{0,ab} - z_{ab})^2 = \frac{T(T+1)}{4\Sigma} (\beta^2 Q_0 + \beta^2 \frac{\Delta}{2} + q - 2\beta m). \quad (112)$$

1505 When $\sigma_2(x) = \sigma_1(\beta^{-1}x)$ and $\sigma_1(z) = \text{softmax}(\{\sqrt{1 + \delta_{ab} z_{ab}}\}_{ab})$, there is a redundant degree of freedom, so we have

$$1508 \mathcal{M}(\Sigma, m, q) = \mathbb{E}_{z_0, z} \inf_{\xi} \frac{1}{2\Sigma} \sum_{a \leq b}^T \left(\beta z_{0,ab} + \frac{\xi}{\sqrt{1 + \delta_{ab}}} - z_{ab} \right)^2 \quad (113)$$

$$1511 = \frac{1}{2\Sigma} \left[\frac{1}{2} T(T+1) - 1 \right] (\beta^2 Q_0 + \beta^2 \frac{\Delta}{2} + q - 2\beta m).$$

By using (75) the SE reduces to

$$\begin{cases} \hat{\Sigma} = \frac{2\alpha}{\Sigma} \frac{\partial \mathcal{K}(m, q)}{\partial q} \\ \hat{m} = -\frac{\alpha}{\Sigma} \frac{\partial \mathcal{K}(m, q)}{\partial m} \\ \hat{q} = 2\alpha \frac{\mathcal{K}(m, q)}{\Sigma^2} \\ m = \partial_{\hat{m}} \frac{\hat{m}^2}{2\Sigma} J\left(\frac{\sqrt{\hat{q}}}{\hat{m}}, \frac{2\lambda}{\hat{m}}\right) \\ q = \left(\frac{\hat{m}}{\Sigma}\right)^2 J\left(\frac{\sqrt{\hat{q}}}{\hat{m}}, \frac{2\lambda}{\hat{m}}\right) \\ \Sigma = \frac{\hat{m}^2}{\Sigma} \partial_{\hat{q}} J\left(\frac{\sqrt{\hat{q}}}{\hat{m}}, \frac{2\lambda}{\hat{m}}\right). \end{cases} \quad (114)$$

Under the ansatz $J \approx Q_0$, we have

$$m_0 = \eta Q_0, \quad q_0 = \eta^2 Q_0 \quad (115)$$

at the leading order by the fourth and fifth equations of (114), satisfying $m_0^2 = q_0 Q_0$, where we denote $\eta := \frac{\hat{m}}{\Sigma} = \frac{m_0}{Q_0}$. They are solved self-consistently with

$$\eta = -\frac{1}{2} \left[\frac{\partial \mathcal{K}(m, q)}{\partial q} \Big|_{m_0, q_0} \right]^{-1} \frac{\partial \mathcal{K}(m, q)}{\partial m} \Big|_{m_0, q_0}, \quad (116)$$

which is obtained by combining the first and the second equations of (114). This gives the plateau $e_0 = \mathcal{I}(m_0, q_0)$. Note that in this case, the plateau $\mathcal{I}(m_0, q_0)$ is non-optimal but does not depend on n, d, λ .

Then we can similarly define the excess risk as

$$\begin{aligned} e_1 &:= \frac{\partial \mathcal{I}(m, q)}{\partial m} \Big|_{m_0, q_0} m_1 + \frac{\partial \mathcal{I}(m, q)}{\partial q} \Big|_{m_0, q_0} q_1 \\ &= -2\chi m_1 + \chi' q_1 \end{aligned} \quad (117)$$

where m_1, q_1 are the first order correction. χ is defined the same way as (89) and

$$\chi' := \mathbb{E} \left[\|\mathcal{J}_{\sigma_2}(\eta z_0)\|_F^2 + \sum_{i=1}^d (\sigma_{2i}(\eta z_0) - \sigma_{1i}(z_0)) \Delta \sigma_i(\eta z_0) \right]. \quad (118)$$

The first order corrections are the same as the previous case

$$m_1 = \eta(J(\delta, \epsilon) - Q_0) - \frac{\eta}{2}(\delta \partial_1 J(\delta, \epsilon) + \epsilon \partial_2 J(\delta, \epsilon)) \quad (119)$$

and

$$q_1 = \eta^2(J(\delta, \epsilon) - Q_0), \quad (120)$$

which gives

$$\begin{aligned} e_1 &= (\eta \chi' - \chi) \eta (J(\delta, \epsilon) - Q_0) + \eta \chi ((\delta \partial_1 + \epsilon \partial_2 - 1) J(\delta, \epsilon) + Q_0) \\ &= \eta \tilde{\chi} (J(\delta, \epsilon) - Q_0) + \eta \chi ((\delta \partial_1 + \epsilon \partial_2 - 1) J(\delta, \epsilon) + Q_0), \end{aligned} \quad (121)$$

where

$$\tilde{\chi} := \frac{1}{2} \sqrt{1 + 2\eta^2} \frac{\partial \mathcal{I}(m, q)}{\partial \ell} \Big|_{m_0, q_0} \quad (122)$$

with ℓ the direction along $m^2 = qQ_0$. There is an additional term compared with the original error decomposition, with coefficient proportional to the derivative along $m^2 = qQ_0$ because

$$\begin{aligned} \frac{\partial \mathcal{I}(m, q)}{\partial \ell} &= \frac{1}{\sqrt{Q_0^2 + 4m_0^2}} \left(Q_0 \frac{\partial \mathcal{I}(m, q)}{\partial m} \Big|_{m_0, q_0} + 2m_0 \frac{\partial \mathcal{I}(m, q)}{\partial q} \Big|_{m_0, q_0} \right) \\ &= \frac{2Q_0}{\sqrt{Q_0^2 + 4m_0^2}} (\eta \chi' - \chi). \end{aligned} \quad (123)$$

This derivative becomes zero if the plateau is equal to (87), and thus evaluates how suboptimal the plateau is.

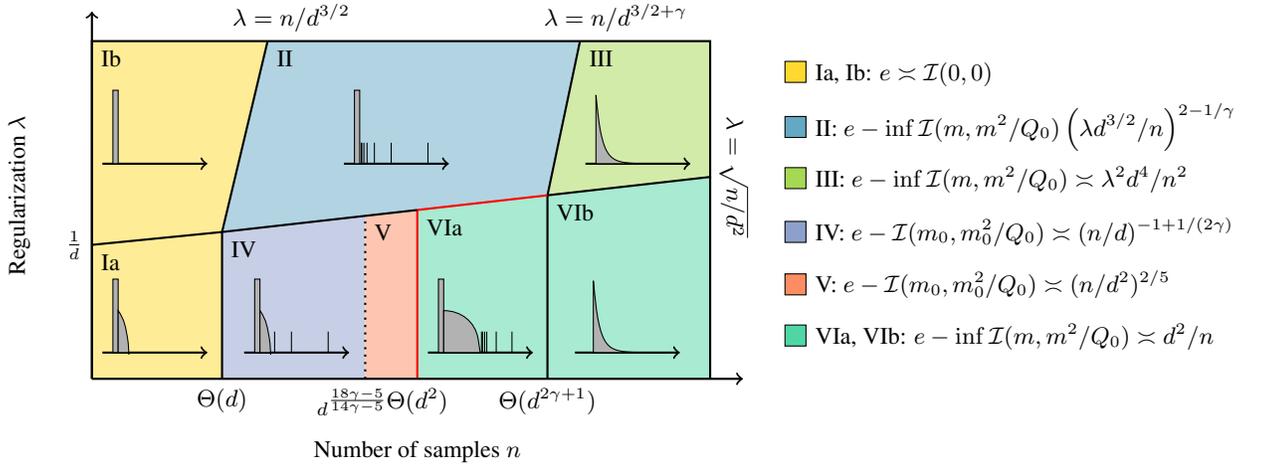


Figure 5: Excess risk rates of as a function of n and $\lambda(n, d)$, with a sketch of the corresponding spectral properties of the learned weights.

Finally, to obtain δ and ϵ , we can use the second and third equations of (114) to obtain

$$\alpha \delta^2 = \alpha \frac{\hat{q}}{\hat{m}^2} = \left[\frac{\partial \mathcal{K}(m, q)}{\partial m} \Big|_{m_0, q_0} \right]^{-2} \mathcal{K}(m_0, q_0). \quad (124)$$

Then using the first and the last equations of (114), we can obtain

$$2\alpha \frac{\partial \mathcal{K}(m, q)}{\partial q} \Big|_{m_0, q_0} = \Sigma \hat{\Sigma} = \frac{1}{\delta} \partial_1 J(\delta, \epsilon). \quad (125)$$

δ, ϵ are solved from (124) and (125).

Similarly, under the BBP approximation, we can write the error decomposition as

$$e_1 = \eta \chi \left(\underbrace{\delta^2 \int_{\epsilon/\delta}^2 \mu_{sc}(dx) \left(x - \frac{\epsilon}{\delta}\right)^2}_{\text{overfitting (learned noise)}} + \underbrace{\frac{1}{d} \delta K'(\delta) (2\delta - \epsilon)^2}_{\text{underfitting (not learned features)}} + \underbrace{\frac{1}{d} \sum_{i=K(\delta)+1}^d s_i^2 + \frac{1}{d} \sum_{i=1}^{K(\delta)} \left[\left(\frac{\delta^2}{s_i} - \epsilon\right) + \frac{\delta^2}{s_i} \left(s_i + \frac{\delta^2}{s_i} - \epsilon\right) \right]}_{\text{approximation error for learned features}} \right) + \eta \tilde{\chi} \left(\underbrace{\frac{1}{d} \sum_{i=1}^{K(\delta)} \left(s_i + \frac{\delta^2}{s_i} - \epsilon\right)^2 + \delta^2 \int_{\epsilon/\delta}^2 \mu_{sc}(x) (x - \epsilon/\delta)^2 - \frac{1}{d} \sum_{i=1}^d s_i^2}_{\text{mismatch of the second moment}} \right) \quad (126)$$

for the under-regularization case and

$$e_1 = \eta \chi \left(\underbrace{\frac{1}{d} \sum_{i=K(\delta, \epsilon)+1}^d s_i^2}_{\text{underfitting (not learned features)}} + \underbrace{\frac{1}{d} \sum_{i=1}^{K(\delta, \epsilon)} \left[\left(\frac{\delta^2}{s_i} - \epsilon\right) + \frac{\delta^2}{s_i} \left(s_i + \frac{\delta^2}{s_i} - \epsilon\right) \right]}_{\text{approximation error for learned features}} \right) + \eta \tilde{\chi} \left(\underbrace{\frac{1}{d} \sum_{i=1}^{K(\delta, \epsilon)} \left(s_i + \frac{\delta^2}{s_i} - \epsilon\right)^2 - \frac{1}{d} \sum_{i=1}^d s_i^2}_{\text{mismatch of the second moment}} \right) \quad (127)$$

for the over-regularization case. This finishes the derivation of the error decomposition for the case $\Sigma \ll 1$.

I ERROR SCALINGS FOR POWER-LAW TEACHERS

In this section we assume that the teacher has a power-law spectrum, i.e., S_0 has eigenvalues $\{\sqrt{di}^{-\gamma}\}_{i=1}^p$ with $p \ll d$. The overall phase diagram is shown in Figure 5, which is the same as that in Defilippis et al. (2025). The detailed derivation is given in the following.

1620 I.1 SHARP MINIMUM: $\Sigma \ll 1$

1621
1622 **Phase Ib: Trivial phase** $\lambda \gg \max\left(\sqrt{\frac{n}{d^2}}, \frac{n}{d^{3/2}}\right)$ According to Defilippis et al. (2025), in this
1623 phase we have $J(\delta, \epsilon) \approx 0$. In this case (83) is still true, but gives the leading order solution

$$1624 \quad m, q, \Sigma = 0, \hat{\Sigma} = 4\alpha\kappa_1(0), \hat{m} = 2\alpha\kappa_0'(0), \hat{q} = 4\alpha\kappa_2(0). \quad (128)$$

1625
1626 This is consistent because $\frac{\lambda}{\hat{m}} \gg \sqrt{d}$. Thus we obtain the trivial generalization error $\mathcal{I}(0, 0)$.

1627
1628 **Phase II: Over-regularization phase** $\max\left(\sqrt{\frac{n}{d^2}}, \frac{n}{d^{\gamma+\frac{3}{2}}}\right) \ll \lambda \ll \frac{d^{3/2}}{n}$ According to Defilippis
1629 et al. (2025), in this phase we have

$$1631 \quad J(\delta, \epsilon) \approx Q_0 + \left(\frac{\gamma+1}{\gamma-1} - \frac{1}{2\gamma-1}\right) \left(\frac{\epsilon}{\sqrt{d}}\right)^{\frac{2\gamma-1}{\gamma}} - \frac{2\epsilon}{\sqrt{d}} \mathbf{1}_{\gamma>1}. \quad (129)$$

1633 Taking it into (95), we have

$$1634 \quad e_1 = \frac{2\eta\chi\gamma}{2\gamma-1} \left(\lambda \frac{d^{3/2}}{2\kappa_1 n}\right)^{\frac{2\gamma-1}{\gamma}}. \quad (130)$$

1635
1636 The condition $\Sigma = \frac{\epsilon}{\delta} \partial_1 J(\delta, \epsilon) \ll 1$ is satisfied.

1637
1638 **Phase III: Intermediate over-regularization phase** $\sqrt{\frac{n}{d^2}} \ll \lambda \ll \frac{n}{d^{\gamma+3/2}}$ According to Defilippis
1639 et al. (2025), in this phase we have

$$1641 \quad J(\delta, \epsilon) \approx Q_0 - \epsilon p^{\min(\gamma, 1)-1} + \lambda^2 \epsilon^2. \quad (131)$$

1642 Taking it into (95), we have

$$1643 \quad e_1 = \frac{\eta\chi\gamma\lambda^2 d^4}{4\kappa_1^2 n^2}. \quad (132)$$

1644
1645 The condition $\Sigma = \frac{\epsilon}{\delta} \partial_1 J(\delta, \epsilon) \ll 1$ is satisfied.

1646
1647 **Phase VI: Large-sample phase** $\lambda \ll \sqrt{\frac{n}{d^2}}$ and $n \gg d^{2\gamma+1}$ According to Defilippis et al.
1648 (2025), in this phase we have

$$1650 \quad J(\delta, \epsilon) = Q_0 + \frac{1}{2} \delta^2 (1 + 4p/d) \approx Q_0 + \frac{\delta^2}{2}, \quad (133)$$

1651 or

$$1652 \quad J(\delta, \epsilon) = Q_0 + \frac{1}{2} \delta^2 + C(\gamma) \left(\frac{\delta}{\sqrt{d}}\right)^{2-\frac{1}{\gamma}}. \quad (134)$$

1653 Taking it into (95), we have

$$1654 \quad e_1 = \frac{\Delta d^2}{4\kappa_1 n} (1 + 4c) \quad \text{or} \quad \frac{\Delta d^2}{4\kappa_1 n}. \quad (135)$$

1655
1656 The condition $\Sigma = \frac{\epsilon}{\delta} \partial_1 J(\delta, \epsilon) = \Theta(1/\alpha) \ll 1$ is satisfied.

1657 I.2 FLAT MINIMUM: $\Sigma \gg 1$

1658
1659 **Phase Ia: Trivial phase** $n \ll d$ and $\lambda \ll \sqrt{\frac{n}{d^2}}$ According to Defilippis et al. (2025), in this
1660 phase we have

$$1661 \quad J(\delta, \epsilon) \approx \int_{\epsilon}^{\delta} \mu_{\text{sc}}(x/\delta) / \delta (x - \epsilon)^2 dx \approx \delta^2 \frac{16t^{7/2}}{105\pi}. \quad (136)$$

1662 (124) and (125) are still true, and give $\delta = \sqrt{\frac{\Delta_1}{4\alpha}}$ and $\delta^2 \frac{16t^{5/2}}{15\pi} = \Delta_2$, with constants defined as

$$1663 \quad \Delta_1 := 8 \left[\frac{\partial \mathcal{K}(m, q)}{\partial m} \Big|_{m_0, q_0} \right]^{-2} \mathcal{K}(m_0, q_0), \quad \Delta_2 := \Delta_1 \frac{\partial \mathcal{K}(m, q)}{\partial q} \Big|_{m_0, q_0}. \quad (137)$$

1664 Thus we have $J \approx 0$, which gives

$$1665 \quad m_0, q_0 = 0 \quad (138)$$

1666 at the leading order. Then we obtain the trivial generalization error $\mathcal{I}(0, 0)$.

Phase IV and V: Benign and harmful overfitting phase $\lambda \ll \sqrt{\frac{n}{d^2}}$ and $d \ll n \ll d^2$ According to Defilippis et al. (2025), in this phase we have $J(\delta, \epsilon) \approx J_1(\delta, \epsilon) + J_2(\delta, \epsilon)$, where

$$J_2(\delta, \epsilon) := \int_{\epsilon}^{\delta} \mu_{\text{sc}}(x/\delta)/\delta(x-\epsilon)^2 dx \approx \delta^2 \frac{16t^{7/2}}{105\pi} \quad (139)$$

with $t := 2 - \frac{\epsilon}{\delta} \ll 1$ and

$$J_1(\delta, \epsilon) \approx Q_0 + \left(\frac{\delta}{\sqrt{d}}\right)^{2-\frac{1}{\gamma}} \left(-\frac{1}{2\gamma-1} + (\epsilon/\delta)^2 + 2 - 2\frac{\epsilon}{\delta} \frac{1}{1-\gamma} - \frac{2}{1+\gamma} \frac{\epsilon}{\delta} + \frac{1}{1+2\gamma}\right) - \mathbf{1}_{\gamma>1} \zeta(\gamma) \frac{2\epsilon}{\sqrt{d}}. \quad (140)$$

Then we have

$$\delta \partial_1 J(\delta, \epsilon) \approx \delta^2 \frac{16t^{5/2}}{15\pi}. \quad (141)$$

Taking it into (124) and (125), we obtain $\delta = \sqrt{\frac{\Delta_1}{4\alpha}}$ and $\delta^2 \frac{16t^{5/2}}{15\pi} = \Delta_2$. Finally (121) gives the excess risk

$$e_1 = \frac{24\eta\chi\gamma^3}{4\gamma^3 + 4\gamma^2 - \gamma - 1} \left(\frac{d\Delta_1}{4n}\right)^{1-\frac{1}{2\gamma}} + \frac{\eta^2\chi'\Delta_2}{7} \left(\frac{15\pi\Delta_2}{4\Delta_1}\right)^{2/5} \left(\frac{n}{d^2}\right)^{2/5}. \quad (142)$$

The condition

$$\Sigma = \frac{\epsilon}{2\delta} \partial_1 J(\delta, \epsilon) \approx \frac{\Delta_2}{2\lambda\delta} \gg 1 \quad (143)$$

is satisfied because we have $\lambda \ll \sqrt{\alpha}$ in this phase.

Interpolation peak According to Defilippis et al. (2025), at the interpolation peak the spectrum is mainly determined by the noise, which gives

$$J(\delta, \epsilon) \approx Q_0 + \frac{1}{2}\delta^2 - \frac{8}{3\pi}\epsilon\delta \quad \text{or} \quad \frac{1}{2}\delta^2 - \frac{8}{3\pi}\epsilon\delta \quad (144)$$

at the leading order. From the fourth and fifth equations of (114) we have

$$m = \frac{\eta}{2}(2 - \delta\partial_1 - \epsilon\partial_2)J(\delta, \epsilon) = Q_0 \quad \text{or} \quad 0 \quad (145)$$

and

$$q = \eta^2 J(\delta, \epsilon), \quad (146)$$

where $\eta := \frac{\hat{m}}{\Sigma}$ is determined by the first and the second equations of (114), i.e.,

$$\eta = -\frac{1}{2} \left[\frac{\partial \mathcal{K}(m, q)}{\partial q} \right]^{-1} \frac{\partial \mathcal{K}(m, q)}{\partial m}. \quad (147)$$

$\delta := \frac{\sqrt{q}}{\hat{m}}$ is given by the the second and the third equations (114), i.e.,

$$\alpha\delta^2 = \left[\frac{\partial \mathcal{K}(m, q)}{\partial m} \right]^{-2} \mathcal{K}(m, q). \quad (148)$$

The special about the interpolation peak is that we need to expand $\mathcal{M}(\Sigma, m, q)$ to the second order, i.e.,

$$\mathcal{M}(\Sigma, m, q) = \frac{1}{2\Sigma} \mathcal{K}(m, q) + \frac{1}{2\Sigma^2} K_1(m, q), \quad (149)$$

where

$$K_1(m, q) := -\mathbb{E}_{z, z_0} \frac{1}{2} (h_0(z, z_0) - z)^T H(z, z_0)^\dagger (h_0(z, z_0) - z) \quad (150)$$

with

$$h_0(z, z_0) := \arg \inf_{h \in \arg \inf_{\|\sigma_1(z_0) - \sigma_2(\cdot)\|^2} \|h - z\|^2} \|h - z\|^2 \quad (151)$$

1728 and

$$1729 H(z, z_0) = \nabla_h^2 \|\sigma_1(z_0) - \sigma_2(h)\|^2|_{h=h_0(z, z_0)}. \quad (152)$$

1730 Then the first equation of (114) gives

$$1731 \Sigma \hat{\Sigma} = 2\alpha \frac{\partial \mathcal{K}(m, q)}{\partial q} + \frac{2\alpha}{\Sigma} \frac{\partial K_1(m, q)}{\partial q} \quad (153)$$

$$1732 \approx 2\alpha \frac{\partial \mathcal{K}(m, q)}{\partial q} + \frac{8\alpha}{\eta\epsilon\delta} \frac{\partial K_1(m, q)}{\partial q},$$

1733 where we use $\Sigma \approx \frac{\eta\epsilon\delta}{4}$ from the last equation of (114). Then combine it with the the last equation
1734 of (114), we obtain

$$1735 4\alpha\delta^2 \frac{\partial \mathcal{K}(m, q)}{\partial q} + \frac{16\alpha\delta}{\eta\epsilon} \frac{\partial K_1(m, q)}{\partial q} = \delta\partial_1 J(\delta, \epsilon) \approx \delta^2 - \frac{8}{3\pi}\epsilon\delta, \quad (154)$$

1736 which implies

$$1737 \begin{cases} \alpha = \frac{1}{4} \left[\frac{\partial \mathcal{K}(m, q)}{\partial q} \right]^{-1} \\ \frac{16\alpha}{\eta\epsilon} \frac{\partial K_1(m, q)}{\partial q} = -\frac{8}{3\pi}\epsilon. \end{cases} \quad (155)$$

1738 Finally the generalization error is given by $\mathcal{I}(m, q)$, where m, q are solved self-consistently from
1739 (145), (146), (147), (148), (155).

1740 For the softmax activations, M is linear in q , and thus the first equation of (155) gives a simple
1741 formula for the interpolation threshold

$$1742 \alpha = \frac{1}{4} \left[\frac{1}{2} T(T+1) - 1 \right]^{-1} \quad (156)$$

1743 by using (113). Similarly, for invertible activations we have

$$1744 \alpha = \frac{1}{2T(T+1)} \quad (157)$$

1745 by using (112), which recovers the interpolation threshold $\frac{1}{4}$ in Erba et al. (2025) by choosing $T = 1$.

1746 J DETAILS OF THE IMPLEMENTATION AND ADDITIONAL EXPERIMENTS

1747 The analytical predictions in all plots are obtained by iterating the equations in Eq. (46) until conver-
1748 gence. The minimization in Eq. (13) is performed with the minimize package of Scipy, initializing
1749 in a Gaussian of variance 10^{-4} centered on z_0 . The expectations in Eq. (46) are computed using
1750 Monte-Carlo integration with at least 10^4 samples. Even though a single iteration of Eq. (46) takes
1751 typically less than one hour on a standard laptop, for convenience we used 60 nodes with 2 Intel
1752 Xeon 8360Y CPUs. For producing this paper we used approximately 60000 CPU hours including
1753 the initial exploration.

1754 We offer some additional numerical explorations of the model in Eq. (1) that are complementary for
1755 the single-head tied-attention model in the main text.

1756 In Fig. 6 we demonstrate a perfect match between GD and theory for various values of λ , which also
1757 suggests a large but finite interpolation peak at small regularization.

1758 In Fig. 7 we evaluate the replicon condition (12). Let's define the replicon as

$$1759 \text{Replicon} = 1 - 2\alpha \mathbb{E}_{z_0, z} \sum_{\substack{a, b, c, d=1 \\ a \leq b, c \leq d}}^T \left(\frac{\partial_{z_{ab}} p_{cd} - \delta_{ac} \delta_{bd}}{\Sigma} \right)^2 \int \mu_{\sqrt{q}/\hat{m}}(dx) \mu_{\sqrt{q}/\hat{m}}(dy) \frac{(\xi(x) - \xi(y))^2}{\hat{\Sigma}^2(x-y)^2}. \quad (158)$$

1760 In order for our results to hold we need the replicon to be positive.

1761 In Fig. 8 left we display the thresholds for perfect recovery of the target in the limit of vanishing
1762 noise $\Delta \rightarrow 0$ and Marchenko-Pastur target. In Fig. 8 right we compare the prediction of the test
1763 error (14) with the error decomposition (126), which is an expansion around zero test error and thus
1764 qualitative in nature when applied to regions where the test error is far from zero. For the specific
1765 plot we used the following prescriptions:

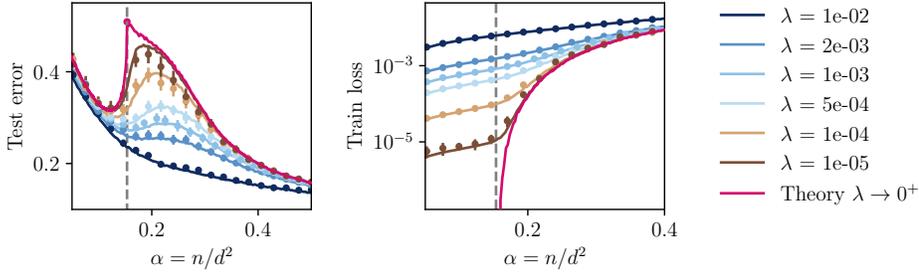


Figure 6: Test error (left) and training loss (right) as functions of the number of samples for decreasing regularization λ ($\Delta = 0.5$, $T = 2$, $\beta = \beta_0 = 1$, $\kappa_0 = 0.5$, $\kappa = 1$). Solid lines: theoretical predictions; dots: Adam simulations at $d = 100$, averaged over 16 runs with 2000 samples in the test set. The gray dashed line marks the analytical interpolation threshold (Corollary B.2), i.e., the largest α for which the ERM estimator fits the training data as $\lambda \rightarrow 0^+$. As expected, the training loss vanishes before interpolation (vertical log scale), while the test error exhibits a non-symmetric interpolation peak, distinct from the usual cusp-like shape.

1. We approximate $\delta K'(\delta) \approx 0$, as it's of a smaller order in d than the other terms for power-law targets;
2. We multiply the overfitting term of (126) by the factor $1 - K(\delta)/d$, to take into account the fact that the bulk will lose mass if a large number of spikes exit from it;
3. We take δ , ϵ , η from the numerical solution of (46);
4. For the mismatch part of the error, we assume that $m = q = 1$ for $n < d^2/4$ instead of using the solution of (46), consistently with the theory (see Eq. (122)).

We remark that even though this procedure is qualitative, the agreement with the pure state evolution error is quite nice.

In Fig. 9 we compare the sequence-to-sequence and sequence-to-label formulations, evaluating test error via Eq. (3) (left) and training loss via Eq. (2) (right), for both noiseless ($\Delta = 0$) and noisy ($\Delta = 0.5$) settings at $T = 2$. The two variants are indistinguishable within error bars across the whole range of $\alpha = n/d^2$, confirming the asymptotic equivalence used in our proof sketch. The learning curves, indeed, exhibit the same phenomenology for the two variants of the model, seq2seq and seq2lab.

We next move to analyze the impact of the inverse temperature used inside the softmax activations in both the model and the target of Eq. (1).

In Fig. 10 we compare the cases in which the inverse temperature matched temperatures $\beta = \beta_0 \in \{0.5, 1, 2\}$ at fixed $T = 2$. Lower temperatures (larger smoothing of the row-wise softmax) systematically yield lower test error for a given α , both without noise (left) and with label noise $\Delta = 0.5$ (right).

To conclude, we vary the number of tokens T to investigate the impact of such parameter in the model's behavior. Finally, in Fig. 11 we vary the number of tokens $T \in \{2, 3, 5\}$ and plot the test error against the rescaled sample ratio $\bar{\alpha} = n/d^2(T(T+1)/2 - 1) = 2n/(T^2 + T - 2)d^2$, motivated by the effective number of scalar constraints in the attention matrix (cf. Appendix F). After this normalization, the learning curves for different T largely align in both the noiseless and noisy cases (left/right), exhibiting the same qualitative dependence on the number of samples and confirming that T primarily rescales the usable information rather than altering the underlying learning dynamics.

1836
1837
1838
1839
1840
1841
1842
1843
1844
1845
1846
1847
1848
1849
1850
1851
1852
1853
1854
1855
1856
1857
1858
1859
1860
1861
1862
1863
1864
1865
1866
1867
1868
1869
1870
1871
1872
1873
1874
1875
1876
1877
1878
1879
1880
1881
1882
1883
1884
1885
1886
1887
1888
1889

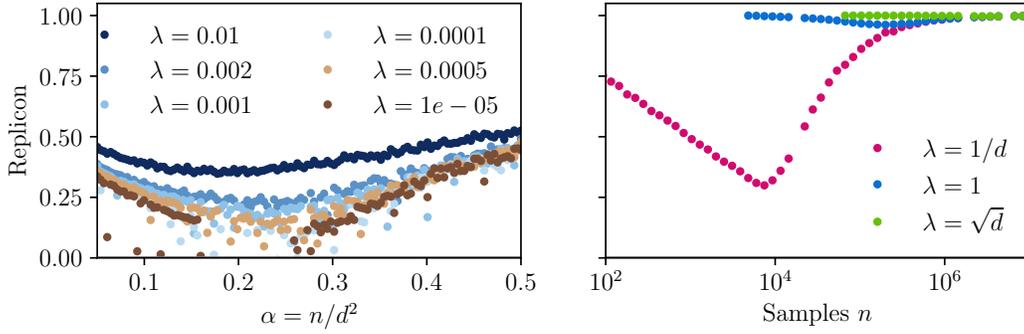


Figure 7: Replicon (158) condition for the same parameters as Figure 6 (left, including the data showed in Figure 1) and ad Figure 4. We see that in general the replicon is well bounded away from zero, meaning that the replicon condition (12) for Claim B.1 is satisfied. We remark that as λ decreases, the replicon approaches lower and lower values around the interpolation threshold, where our numerical solver performs poorly. We conjecture that the condition is satisfied for all values of α .

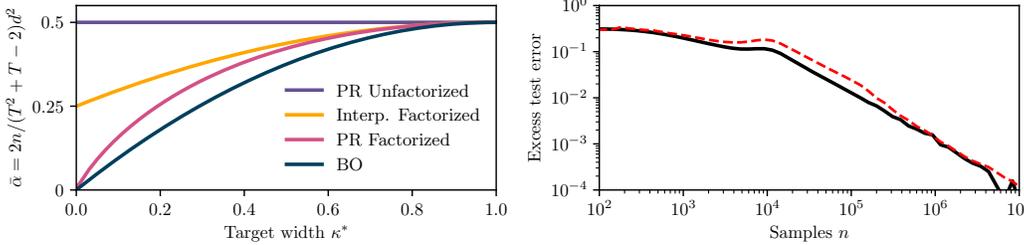


Figure 8: (Left) Analytical thresholds for vanishing regularization in the noiseless case $\Delta = 0$. We plot the Bayes-optimal perfect recovery from Boncoraglio et al. (2025) (lower-bound to the perfect recovery of any estimator), the perfect recovery of the attention model Eq. (2) from Corollary B.2, the interpolation threshold from Corollary B.2, and the perfect recovery of the non-factorized attention model Eq. (21). (Right) Comparison of the excess test error from Claim B.1 (in black) in tis non-asymptotic version as described in Section 4 and its equivalent computed using the error decomposition in Eq. (8). Parameters are $\gamma = 0.75$, $d = 200$, $\Delta = 0.5$, $T = 2$, $\lambda = 1/d$. The specific formula used is Eq. (126).

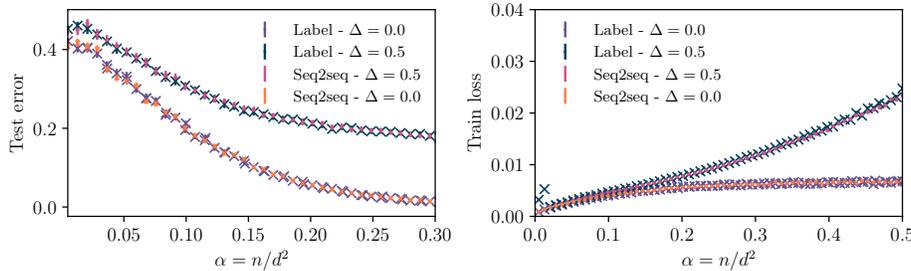


Figure 9: Comparison between the seq2seq model and the seq2lab model, both in Eq. (1). Both panels display the numerical simulations obtain from Adam with $\lambda = 0.01$, $\beta = \beta_0 = 1.0$, $T = 2$. We compare in both panels the noiseless $\Delta = 0.0$ and the noisy version of the model with $\Delta = 0.5$. We show in the (Left panel) the test error of the model computed from the seq2seq and the seq2lab formulas in Eq. (3). In the (Right panel) we show the train loss computed from the formulas in Eq. (2) with Adam. In all curves we average over 32 different realizations, with $d = 100$ and 2000 samples in the test set. Both variants achieve same test error and train loss, both in the noiseless and in the noisy version of the model.

1890
 1891
 1892
 1893
 1894
 1895
 1896
 1897
 1898
 1899
 1900
 1901
 1902
 1903
 1904
 1905
 1906
 1907
 1908
 1909
 1910
 1911
 1912
 1913
 1914
 1915
 1916
 1917
 1918
 1919
 1920
 1921
 1922
 1923
 1924
 1925
 1926
 1927
 1928
 1929
 1930
 1931
 1932
 1933
 1934
 1935
 1936
 1937
 1938
 1939
 1940
 1941
 1942
 1943

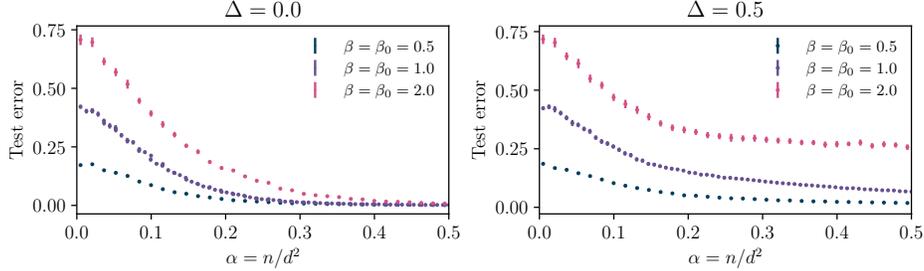


Figure 10: We show the test error for the model in Eq. (1) with respect to $\alpha = n/d^2$. We compute the test error through Adam for different values of the inverse temperatures $\beta = \beta_0$ of the softmax activation in the single-head tied-attention model. In all cases we consider the matched-temperature scenario between the target and the model. Both panels display the numerical simulations obtain from Adam with $\lambda = 0.01$, $\kappa_0 = 0.5$, $\kappa = 1$, $T = 2$. We compare in both panels the noiseless $\Delta = 0.0$ (Left panel) and the noisy version of the model with $\Delta = 0.5$ (Right panel) for three values of the inverse temperatures $\beta = \beta_0 = 0.5, 1, 2.0$. In all curves we average over 32 different realizations, with $d = 100$ and 2000 samples in the test set.

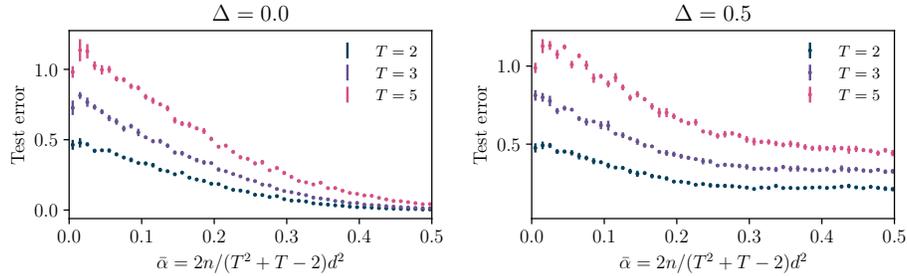


Figure 11: We show the test error computed through Adam for the model in Eq. (1) for a different number of tokens $T \geq 2$. In the x-axis we show the rescaled sample complexity $\bar{\alpha} = n/d^2 \left(\frac{T(T+1)}{2} - 1 \right) = 2n/(T^2 + T - 2)d^2$. Both panels display the numerical simulations obtain from Adam with $\lambda = 0.001$, $\kappa_0 = 0.5$, $\kappa = 1$, $\beta = \beta_0 = 1.0$ and different number of tokens $T = 2, 3, 5$. We compare in both panels the noiseless $\Delta = 0.0$ (Left panel) and the noisy version of the model with $\Delta = 0.5$. In all curves we average over 32 different realizations, with $d = 100$ and 2000 samples in the test set.