A Phase Transition between Positional and Semantic Learning in a Solvable Model of Dot-Product Attention

Hugo Cui HUGO.CUI@EPFL.CH
Statistical Physics Of Computation laboratory, EPFL, Switzerland

Freya Behrens FREYA.BEHRENS@EPFL.CH
Statistical Physics Of Computation laboratory, EPFL, Switzerland

Florent Krzakala FLORENT.KRZAKALA@EPFL.CH
Information Learning & Physics laboratory, EPFL, Switzerland

Lenka Zdeborová LENKA.ZDEBOROVA@EPFL.CH
Statistical Physics Of Computation laboratory, EPFL, Switzerland

Abstract

A theoretical understanding of how algorithmic abilities emerge in the learning of language models remains elusive. In this work, we provide a tight theoretical analysis of the emergence of semantic attention in a solvable model of dot-product attention and consider a non-linear self-attention layer with trainable tied and low-rank query and key matrices. In the asymptotic limit of high-dimensional data and a comparably large number of training samples we provide a tight closed-form characterization of the global minimum of the non-convex empirical loss landscape. We show that this minimum corresponds to either a positional attention mechanism (with tokens attending to each other based on their respective positions) or a semantic attention mechanism (with tokens attending to each other based on their meaning), and evidence an emergent phase transition from the former to the latter with increasing sample complexity.

1. Introduction

Self-attention layers [60] have been instrumental in advancing the abilities of language models, as they provide an efficient method of extracting information from sentences – both the information encoded in the ordering (i.e. *positions*) of the words, and that encoded in the meaning (i.e. *semantics*) of the words. In theory, attention layers can learn to leverage both types of information, by having tokens attend to each other based on their respective positions (a mechanism called *positional attention* in [27], through some form of positional encodings [24, 28, 46, 50, 52]) and/or respective meanings (henceforth referred to as *semantic attention*). A growing body of work on mechanistic interpretability aims to empirically understand which precise algorithmic mechanisms a neural network learns and has shown that attention layers are able to implement a wide range of different algorithms using both positional and semantic attributes of the inputs [1, 11, 38, 43, 44, 51, 65]. Simultaneously, empirical studies have provided evidence for the emergence of specific algorithmic mechanisms (abilities) in the learning of language models that lead to qualitative improvements of the model capabilities [51, 61, 62]. Despite these efforts, it remains an open question how to theoretically characterise the conditions under which such an ability emerges in the model. Even the nature of this algorithmic emergence is unclear, i.e. whether it constitutes a fast but smooth change in performance or it is due to a sharp boundary between fundamentally different regimes of learning [48]. While there is

a plethora of work on the theory of attention investigating various aspects such as their expressivity [16, 18, 23], inductive bias [47, 56, 57], training dynamics [9, 27, 29, 59], and in-context learning [3, 8, 21, 30, 64], these studies do not allow to capture sharp changes in the behaviour of attention mechanisms such as phase transitions or do not capture an emergent phenomenon [3, 18, 27, 45].

In our work, we take inspiration from physics, where a similar theoretical questions about the nature of phase transitions and emergent phenomena were posed a century ago for models of interacting particles, such as the famous Ising model [25, 39] and more recently in the theory of feed-forward fully connected neural networks, e.g. [2, 4, 5, 22, 33, 49, 53]. Here, for the first time, we bring this type of study to the analysis of neural networks with attention layers. We introduce and analyse a tractable model that permits a sharp high-dimensional characterisation for attention layers. In particular, we describe a model with a single self-attention layer with tied, low-rank query and key matrices, with Gaussian input data and realizable labels:

(1) We show that this model exhibits a phase transition in terms of sample complexity between a semantic and a positional mechanism.
(2) We analyse this model in the asymptotic limit where the embedding dimension d of the tokens and the number n of training samples grow proportionally to infinity and provide a *tight closed-form characterization* of the test error and training loss achieved at the minima of the non-convex empirical loss.
(3) Using this high-dimensional characterization, we locate the positional-semantic phase transition, thus providing the first theoretical result about the *emergence of sharp phase transitions* in a model of dot-product attention.

2. Tied low-rank attention model

Input data model We consider a model of embedded sentences with uncorrelated (1-gram) words. More precisely, a sentence $\boldsymbol{x} \in \mathbb{R}^{L\times d}$, where L is the sentence length and d represents the embedding dimension, consists of L tokens $\{\boldsymbol{x}_\ell\}_{1\le\ell\le L}$ independently drawn from a Gaussian distribution $\boldsymbol{x}_\ell \sim \mathcal{N}(0, \boldsymbol{\Sigma}_\ell)$ with covariance $\boldsymbol{\Sigma}_\ell \in \mathbb{R}^{d\times d}$. In the following, we denote the probability distribution of $\boldsymbol{x}$ as p_x. Note that while this sentence model does not involve in itself statistical correlations between tokens, the task (target function) will entail interactions between different tokens.

Target function The target function (teacher) is assumed to be of the form

$$y(x) = \mathtt{T}\left[\frac{1}{\sqrt{d}}\boldsymbol{x}\boldsymbol{Q}_\star\right]\boldsymbol{x}, \tag{1}$$

for a function $\mathtt{T} : \mathbb{R}^{L\times r_t} \to \mathbb{R}^{L\times L}$. The term $\mathtt{T}\left[1/\sqrt{d}\boldsymbol{x}\boldsymbol{Q}_\star\right] \in \mathbb{R}^{L\times L}$ in (1) should be interpreted as the target attention matrix, which mixes the tokens of the input $\boldsymbol{x}$. This attention matrix is parametrized by the target weights $\boldsymbol{Q}_\star \in \mathbb{R}^{d\times r_t}$.

Tied attention We consider the learning of the target (1) by a single attention layer

$$f_{\boldsymbol{Q}}(x) = \mathtt{S}\left[\frac{1}{\sqrt{d}}(\boldsymbol{x}+\boldsymbol{p})\boldsymbol{Q}\right](\boldsymbol{x}+\boldsymbol{p}). \tag{2}$$

In (2), $\boldsymbol{p} \in \mathbb{R}^{L\times d}$ is a *fixed* matrix, corresponding to positional encodings, and $\boldsymbol{Q} \in \mathbb{R}^{d\times r_s}$ is a trainable weight matrix. We denote subsequently $\boldsymbol{p}_\ell \in \mathbb{R}^d$ the $\ell-$th row of $\boldsymbol{p}$. Like the target (1), the parametric function (2) takes the form of a data-dependent attention matrix $\mathtt{S}\left[1/\sqrt{d}(\boldsymbol{x}+\boldsymbol{p})\boldsymbol{Q}\right] \in \mathbb{R}^{L\times L}$ mixing the tokens of the input $\boldsymbol{x}$.

Empirical risk minimization We study the learning of the attention layer (2), when a training set $\mathcal{D} = \{\boldsymbol{x}^\mu, y(\boldsymbol{x}^\mu)\}_{\mu=1}^n$ with n independently sampled sentences $\{\boldsymbol{x}^\mu\}_{\mu=1}^n$, and the associated labels $\{y(\boldsymbol{x}^\mu)\}_{\mu=1}^n$, is available. The target (1) can be learnt by carrying out an empirical risk minimization, with the generalization error measured at test time by the mean squared error (MSE)

$$\hat{\boldsymbol{Q}} = \underset{\boldsymbol{Q}\in\mathbb{R}^{d\times r}}{\operatorname{argmin}} \left[\sum_{\mu=1}^{n} \frac{1}{2d} \left\|y(\boldsymbol{x}^\mu) - f_{\boldsymbol{Q}}(\boldsymbol{x}^\mu)\right\|^2 + \frac{\lambda}{2}\|\boldsymbol{Q}\|^2\right] \;;\; \epsilon_g \equiv \frac{1}{dL}\mathbb{E}_{\boldsymbol{x}\sim p_x}\left\|y(\boldsymbol{x}) - f_{\hat{\boldsymbol{Q}}}(\boldsymbol{x})\right\|^2. \tag{3}$$

3. Closed-form characterization of the training loss

We analyze the learning problem (3) in the limit where the embedding dimension d and the number of training samples n jointly tend to infinity, while their ratio $\alpha = n/d$, the sample complexity, stays of order $\Theta_d(1)$. We further assume the sentence length L, the ranks r_s, r_t of the weights $\boldsymbol{Q}, \boldsymbol{Q}_\star$, and the norm of the positional embeddings $\|\boldsymbol{p}\|$ to be $\Theta_d(1)$. This limit has been considered in a stream of previous works (e.g. [14, 32, 36]) and allows to derive closed-form characterization of the ERM problem (3). It also exhibits a particularly rich learning phenomenology which we further explore in Section 4.

The main technical result of the present work is a closed-formed characterization of the test MSE and training loss (3) achieved in the high-dimensional limit for the model (2) trained via the empirical risk minimization of (3), and we state it stated in Appendix A. The derivation is exploiting a mapping of the model (2) to a (variant of) a Generalized Linear Model (GLM) [34, 37]. The summary statistics characterized by the self-consistent state evolution equations (8) [26] asymptotically describe the fixed points of a Generalized Approximate Message Passing (GAMP) algorithm [42] (A). The fixed points of GAMP in turn correspond to critical (zero-gradient) points of the non-convex empirical loss landscape (3). Therefore, while the technical Result 1 is stated as a characterization of the global minimum of (3), which is the main concern of the present work, solutions of (8) also describe local minima and saddles. Note that Appendix B provides an alternative derivation of the result for different losses using the replica method from statistical physics [40]. The methodology we use is similar to many recent work that study asymptotics of a large number of high-dimensional problems, e.g. [7, 15, 17, 19, 31].

4. Positional-to-semantic phase transition

Rank one dot-product attention In the following, we turn to a special case of tied low-rank attention (2) – namely a dot-product attention layer, with the student

$$\mathtt{S}\left[\frac{1}{\sqrt{d}}(\boldsymbol{x}+\boldsymbol{p})\boldsymbol{Q}\right] = \mathrm{softmax}\left(\frac{1}{d}(\boldsymbol{x}+\boldsymbol{p})\boldsymbol{Q}\boldsymbol{Q}^\top(\boldsymbol{x}+\boldsymbol{p})^\top\right), \tag{4}$$

and a specific case of target attention matrix (1) of the form

$$\mathtt{T}\left[\frac{1}{\sqrt{d}}\boldsymbol{x}\boldsymbol{Q}_\star\right] = (1-\omega)\mathrm{softmax}\left(\frac{1}{d}\boldsymbol{x}\boldsymbol{Q}_\star\boldsymbol{Q}_\star^\top\boldsymbol{x}^\top\right) + \omega A\cdot \tag{5}$$

with $A \in \mathbb{R}^{L\times L}$ a fixed matrix. In (5), the parameter $\omega \in [0,1]$ tunes the relative strength of the dot-product term and the fixed matrix term, and interpolates between a fully positional and a fully semantic task: With $\omega = 0$ we have a purely semantic target as the $i,j-$th element of the score matrix $\mathrm{softmax}(1/d\boldsymbol{x}\boldsymbol{Q}_\star\boldsymbol{Q}_\star^\top\boldsymbol{x}^\top)$ only depends on the tokens $\boldsymbol{x}_i, \boldsymbol{x}_j$ and not explicitely on their respective placements i, j inside the sentence. For $\omega = 1$, the attention matrix A associated thereto is purely

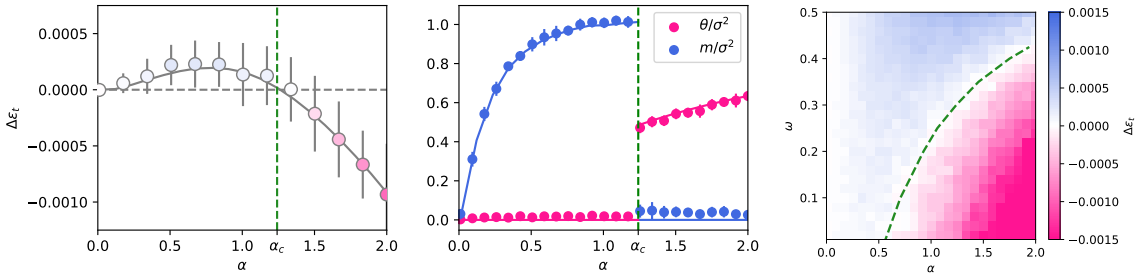

Figure 1: **Mixed positional/semantic teacher for $\omega = 0.3$.** Setting is $r_s = r_t = 1, L = 2, A = ((0.6, 0.4), (0.4, 0.6)), \boldsymbol{\Sigma}_1 = \boldsymbol{\Sigma}_2 = 0.25\mathbb{I}_d$, $\boldsymbol{p}_1 = \mathbf{1}_d = -\boldsymbol{p}_2$ and $\boldsymbol{Q}_\star \sim \mathcal{N}(0, \mathbb{I}_d)$. We compare the solutions of (8) in Result 1 (solid lines) with the models (2) trained using gradient descent initialized resp. at $\boldsymbol{Q}_\star$ and at $\boldsymbol{p}_1$ (markers). The green dashed line is the theoretical prediction for the threshold $\alpha_c(\omega)$ above which the semantic solution of (8) in Result 1 has lower loss than the positional solution. **(left)** Difference in training loss $\Delta\epsilon_t$ for $\omega = 0.3$. **(center)** overlap θ between the learnt weights $\hat{\boldsymbol{Q}}$ and the target weights $\boldsymbol{Q}_\star$ Overlap m between the learnt weights $\hat{\boldsymbol{Q}}$ and the positional embedding p_1, where only the solution of (8) corresponding to the lowest found training loss is represented. **(right)** Empirical difference in training loss for a range of ω, α.

positional, in the sense that A_{ij} is a function of i, j but not of $\boldsymbol{x}_i, \boldsymbol{x}_j$. To complete the learning task, a *positional* mechanism then needs to be learnt. The parameter ω thus allows to tune the amount of semantic/positional content in the target (5), and thus the extent to which the task requires the model to implement semantic attention (small ωs) or rather positional attention (large ωs). In the following, for definiteness, we further assume $r_s = r_t = 1$ and set $\boldsymbol{Q}_\star$ to be a fixed random Gaussian vector drawn from $\mathcal{N}(0, \mathbb{I}_d)$, and choose the positional encodings $\boldsymbol{p}_1 = -\boldsymbol{p}_2 = \mathbf{1}_d$. Finally, for simplicity, we consider sentences with two tokens $L = 2$ and isotropic token covariances $\boldsymbol{\Sigma}_1 = \boldsymbol{\Sigma}_2 = \sigma^2 \mathbb{1}_d$.

Semantic and positional mechanisms The summary statistics θ_ℓ, m_ℓ describing the global minimizer of the empirical loss minimization (3) of the dot-product attention (4) on the target (5) are captured alongside the corresponding test error (3) and training loss (3), by Result 1. The solution of the system of equations (8) is not unique, and different stable fixed points describe different corresponding critical points of the non-convex empirical loss landscape (3). In practice, we find two solutions of (8), corresponding to two minima associated with different mechanisms implemented by the dot-product attention (4) when approximating the target (5):

–*Positional solution* One solution of (8) correspond to vanishing overlap $\theta = 0$ between the trained weights $\hat{\boldsymbol{Q}}$ and the semantic target weights $\boldsymbol{Q}_\star$, and non-zero $m > 0$ between the trained weights $\hat{\boldsymbol{Q}}$ and the positional embedding $\boldsymbol{p}_1 = -\boldsymbol{p}_2$. Consequently, the argument of the dot-product attention $\hat{\boldsymbol{Q}}(\boldsymbol{x} + \boldsymbol{p})$ has a sizeable token-independent –thus positional– contribution $\hat{\boldsymbol{Q}}\boldsymbol{p}$, alongside a token-dependent semantic part $\hat{\boldsymbol{Q}}\boldsymbol{x}$. Because of the positional terms, the resulting learnt attention matrix $\mathrm{softmax}(1/d(\boldsymbol{x} + \boldsymbol{p})\hat{\boldsymbol{Q}}\hat{\boldsymbol{Q}}^\top(\boldsymbol{x} + \boldsymbol{p})^\top)$ implements a partly positional mechanism.

–*Semantic solution* Another solution of the system of equations (8) is associated with a vanishing overlap $m = 0$ between the learnt weights $\hat{\boldsymbol{Q}}$ and the positional embeddings, and a finite overlap $\theta > 0$ with the target weights $\boldsymbol{Q}_\star$. Therefore the resulting learnt attention matrix $\mathrm{softmax}(1/d(\boldsymbol{x} + \boldsymbol{p})\hat{\boldsymbol{Q}}\hat{\boldsymbol{Q}}^\top(\boldsymbol{x} + \boldsymbol{p})^\top) \approx \mathrm{softmax}(1/d\boldsymbol{x}\hat{\boldsymbol{Q}}\hat{\boldsymbol{Q}}^\top\boldsymbol{x}^\top)$ is largely semantic.

Positional-to-semantic phase transition While the system of self-consistent equations (8) may admit other solutions, we did not find solutions with lower training loss than the two aforedescribed fixed points. Which of these solution corresponds to the global minimum – and thus the solution of the optimization (3)– depends on the sample complexity α and the positional/semantic parameter ω (5). For a fixed parameter ω in (5), an analysis of equations (8) (Appendix B), reveals that for a sizeable range of ω there exists a threshold α_c for the sample complexity so that for $\alpha < \alpha_c$, the global minimum of (3) corresponds to a positional mechanism, and is described by the positional solution of (8) of Result 1 with $\theta = 0, m > 0$. For $\alpha > \alpha_c$, the global minimum of (3) corresponds to a semantic mechanism, and is described by the semantic solution of (8) of Result 1 with $\theta > 0, m = 0$.

The dot-product attention thus displays *a phase transition in sample complexity from a positional mechanism to a semantic mechanism*, implementing the simpler positional mechanism when having access to small amounts of data, and only learning the semantic content of the target (5) when presented sufficient data. The critical sample complexity α_c generically grows with the positionality ω of the target function (5), as the semantic content – i.e. the first term of (5)– is less apparent for larger ω, and thus requires larger amounts of data to be identified and approximated by the dot-product attention (4). In Fig. 1 (left) for $\omega = 0.3$ the difference in training loss $\Delta\epsilon_t$ between the positional and semantic solutions of (8) is represented, alongside the difference in training loss at convergence experimentally reached by gradient descent. For small (resp. large) sample complexity $\alpha < \alpha_c$ (resp. $\alpha > \alpha_c$), the training loss of the positional (resp. semantic) minimum is lower, and thus corresponds to the global minimum.

Experimentally, the positional minimum can be reached for $\alpha < \alpha_c$ via gradient descent by initializing the weights $\boldsymbol{Q}$ of the attention (4) close to the positional embedding $\boldsymbol{p}_1$. By the same means, the semantic minimum can be reached from an initialization at the teacher weights $\boldsymbol{Q}_\star$ (5). Note that the semantic initialization is informed in nature, in that it necessitates the knowledge of the target parameters $\boldsymbol{Q}_\star$. We conduct numerical experiments from a random initialization of $\boldsymbol{Q}$ in Appendix C.3, and show that the dynamics may reach either of the local minima, or get stuck in a different one.

In Fig. 1 (center), we compare our analytical characterizations for different metrics at the global mimimum – the summary statistics θ, m (middle), and the test MSE (right)–, with the corresponding experimental estimates, obtained by optimizing (3) with the `Pytorch` implementation of gradient descent, from a positional (resp. semantic) initialization for $\alpha < \alpha_c$ (resp. $\alpha > \alpha_c$), displaying overall good agreement. In Appendix C.1 we further verify that in the scaling limit of our analysis, namely $n, d \to \infty$ for $\alpha = O(1)$, the agreement improves with growing n, d. Overall, we obsevere the emergence of semantic learning as a function of the task and the sample complexity in Fig. 1.

Conclusion

We explored the interplay between positional and semantic attention, through the prism of tied low-rank self-attention in high dimensions. In a theoretically controlled setting, we characterized the global optimum of the empirical loss, when learning a target attention layer. This global optimum was found to correspond to either a positional or a semantic mechanism, with a phase transition between the two mechanisms occurring as the sample complexity increases. We believe the present asymptotic analysis of the inner workings of attention mechanisms opens up exciting research directions. Considering untied query and key matrices, appending a readout network after the attention layer, or addressing more practical training procedures such as masked language modelling, are some possible extensions which will hopefully pave the way towards a satisfactory theoretical comprehension of attention mechanisms.

References

[1] Sanjeev Arora and Anirudh Goyal. A theory for emergence of complex skills in language models. *arXiv preprint arXiv:2307.15936*, 2023.

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

[3] Yu Bai, Fan Chen, Huan Wang, Caiming Xiong, and Song Mei. Transformers as statisticians: Provable in-context learning with in-context algorithm selection. *arXiv preprint arXiv:2306.04637*, 2023.

[4] Jean Barbier, Florent Krzakala, Nicolas Macris, Léo Miolane, and Lenka Zdeborová. Optimal errors and phase transitions in high-dimensional generalized linear models. *Proceedings of the National Academy of Sciences*, 116(12):5451–5460, 2019.

[5] Adriano Barra, Giuseppe Genovese, Peter Sollich, and Daniele Tantari. Phase transitions in restricted boltzmann machines with generic priors. *Physical Review E*, 96(4):042156, 2017.

[6] Mohsen Bayati and Andrea Montanari. The dynamics of message passing on dense graphs, with applications to compressed sensing. *IEEE Transactions on Information Theory*, 57(2): 764–785, 2011.

[7] Mohsen Bayati and Andrea Montanari. The lasso risk for gaussian matrices. *IEEE Transactions on Information Theory*, 58(4):1997–2017, 2011.

[8] Alberto Bietti, Vivien Cabannes, Diane Bouchacourt, Herve Jegou, and Leon Bottou. Birth of a transformer: A memory viewpoint. *Advances in Neural Information Processing Systems*, 36, 2024.

[9] Enric Boix-Adsera, Etai Littwin, Emmanuel Abbe, Samy Bengio, and Joshua Susskind. Transformers learn through gradual rank increase. *arXiv preprint arXiv:2306.07042*, 2023.

[10] 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.

[11] Angelica Chen, Ravid Schwartz-Ziv, Kyunghyun Cho, Matthew L Leavitt, and Naomi Saphra. Sudden drops in the loss: Syntax acquisition, phase transitions, and simplicity bias in mlms. *arXiv preprint arXiv:2309.07311*, 2023.

[12] Lucas Clarté, Bruno Loureiro, Florent Krzakala, and Lenka Zdeborová. Theoretical characterization of uncertainty in high-dimensional linear classification. *Machine Learning: Science and Technology*, 4(2):025029, 2023.

[13] Elisabetta Cornacchia, Francesca Mignacco, Rodrigo Veiga, Cédric Gerbelot, Bruno Loureiro, and Lenka Zdeborová. Learning curves for the multi-class teacher–student perceptron. *Machine Learning: Science and Technology*, 4(1):015019, 2023.

[14] Hugo Cui and Lenka Zdeborová. High-dimensional asymptotics of denoising autoencoders. *arXiv preprint arXiv:2305.11041*, 2023.

[15] David Donoho and Andrea Montanari. High dimensional robust m-estimation: Asymptotic variance via approximate message passing. *Probability Theory and Related Fields*, 166:935–969, 2016.

[16] Benjamin L Edelman, Surbhi Goel, Sham Kakade, and Cyril Zhang. Inductive biases and variable creation in self-attention mechanisms. In *International Conference on Machine Learning*, pages 5793–5831. PMLR, 2022.

[17] Melikasadat Emami, Mojtaba Sahraee-Ardakan, Parthe Pandit, Sundeep Rangan, and Alyson Fletcher. Generalization error of generalized linear models in high dimensions. In *International Conference on Machine Learning*, pages 2892–2901. PMLR, 2020.

[18] Hengyu Fu, Tianyu Guo, Yu Bai, and Song Mei. What can a single attention layer learn? a study through the random features lens. *arXiv preprint arXiv:2307.11353*, 2023.

[19] Cedric Gerbelot, Alia Abbara, and Florent Krzakala. Asymptotic errors for teacher-student convex generalized linear models (or: How to prove kabashima's replica formula). *IEEE Transactions on Information Theory*, 69(3):1824–1852, 2022.

[20] Yehoram Gordon. On milman's inequality and random subspaces which escape through a mesh in rn. In *Geometric Aspects of Functional Analysis: Israel Seminar (GAFA) 1986–87*, pages 84–106. Springer, 1988.

[21] Tianyu Guo, Wei Hu, Song Mei, Huan Wang, Caiming Xiong, Silvio Savarese, and Yu Bai. How do transformers learn in-context beyond simple functions? a case study on learning with representations. *arXiv preprint arXiv:2310.10616*, 2023.

[22] Géza Györgyi. First-order transition to perfect generalization in a neural network with binary synapses. *Physical Review A*, 41(12):7097, 1990.

[23] Michael Hahn. Theoretical limitations of self-attention in neural sequence models. *Transactions of the Association for Computational Linguistics*, 8:156–171, 2020.

[24] Adi Haviv, Ori Ram, Ofir Press, Peter Izsak, and Omer Levy. Transformer language models without positional encodings still learn positional information. In Yoav Goldberg, Zornitsa Kozareva, and Yue Zhang, editors, *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 1382–1390, Abu Dhabi, United Arab Emirates, December 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.findings-emnlp.99. URL https://aclanthology.org/2022.findings-emnlp.99.

[25] Ernst Ising. Contribution to the theory of ferromagnetism. *Z. Phys*, 31(1):253–258, 1925.

[26] Adel Javanmard and Andrea Montanari. State evolution for general approximate message passing algorithms, with applications to spatial coupling. *Information and Inference: A Journal of the IMA*, 2(2):115–144, 2013.

[27] Samy Jelassi, Michael Sander, and Yuanzhi Li. Vision transformers provably learn spatial structure. *Advances in Neural Information Processing Systems*, 35:37822–37836, 2022.

[28] Amirhossein Kazemnejad, Inkit Padhi, Karthikeyan Natesan, Payel Das, and Siva Reddy. The impact of positional encoding on length generalization in transformers. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023. URL https://openreview.net/forum?id=Drrl2gcjzl.

[29] Hongkang Li, Meng Wang, Sijia Liu, and Pin-Yu Chen. A theoretical understanding of shallow vision transformers: Learning, generalization, and sample complexity. *arXiv preprint arXiv:2302.06015*, 2023.

[30] Yingcong Li, Muhammed Emrullah Ildiz, Dimitris Papailiopoulos, and Samet Oymak. Transformers as algorithms: Generalization and stability in in-context learning. In *International Conference on Machine Learning*, pages 19565–19594. PMLR, 2023.

[31] Bruno Loureiro, Gabriele Sicuro, Cedric Gerbelot, Alessandro Pacco, Florent Krzakala, and Lenka Zdeborová. Learning gaussian mixtures with generalized linear models: Precise asymptotics in high-dimensions. In M. Ranzato, A. Beygelzimer, Y. Dauphin, P.S. Liang, and J. Wortman Vaughan, editors, *Advances in Neural Information Processing Systems*, volume 34, pages 10144–10157. Curran Associates, Inc., 2021. URL https://proceedings.neurips.cc/paper_files/paper/2021/file/543e83748234f7cbab21aa0ade66565f-Paper.pdf.

[32] Bruno Loureiro, Gabriele Sicuro, Cédric Gerbelot, Alessandro Pacco, Florent Krzakala, and Lenka Zdeborová. Learning gaussian mixtures with generalized linear models: Precise asymptotics in high-dimensions. *Advances in Neural Information Processing Systems*, 34:10144–10157, 2021.

[33] Antoine Maillard, Bruno Loureiro, Florent Krzakala, and Lenka Zdeborová. Phase retrieval in high dimensions: Statistical and computational phase transitions. *Advances in Neural Information Processing Systems*, 33:11071–11082, 2020.

[34] Peter McCullagh. *Generalized linear models*. Routledge, 2019.

[35] Marc Mézard, Giorgio Parisi, and Miguel Angel Virasoro. *Spin glass theory and beyond: An Introduction to the Replica Method and Its Applications*, volume 9. World Scientific Publishing Company, 1987.

[36] Francesca Mignacco, Florent Krzakala, Yue Lu, Pierfrancesco Urbani, and Lenka Zdeborova. The role of regularization in classification of high-dimensional noisy gaussian mixture. In *International conference on machine learning*, pages 6874–6883. PMLR, 2020.

[37] John Ashworth Nelder and Robert WM Wedderburn. Generalized linear models. *Journal of the Royal Statistical Society Series A: Statistics in Society*, 135(3):370–384, 1972.

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

Circuits Thread, 2022. https://transformer-circuits.pub/2022/in-context-learning-and-induction-heads/index.html.

[39] Lars Onsager. Crystal statistics. i. a two-dimensional model with an order-disorder transition. *Physical Review*, 65(3-4):117, 1944.

[40] Giorgio Parisi. Toward a mean field theory for spin glasses. *Physics Letters A*, 73(3):203–205, 1979.

[41] Giorgio Parisi. Order parameter for spin-glasses. *Physical Review Letters*, 50(24):1946, 1983.

[42] Sundeep Rangan, Philip Schniter, Erwin Riegler, Alyson K Fletcher, and Volkan Cevher. Fixed points of generalized approximate message passing with arbitrary matrices. *IEEE Transactions on Information Theory*, 62(12):7464–7474, 2016.

[43] Allan Raventós, Mansheej Paul, Feng Chen, and Surya Ganguli. Pretraining task diversity and the emergence of non-bayesian in-context learning for regression. In A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine, editors, *Advances in Neural Information Processing Systems*, volume 36, pages 14228–14246. Curran Associates, Inc., 2023. URL https://proceedings.neurips.cc/paper_files/paper/2023/file/2e10b2c2e1aa4f8083c37dfe269873f8-Paper-Conference.pdf.

[44] Gautam Reddy. The mechanistic basis of data dependence and abrupt learning in an in-context classification task. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=aN4Jf6Cx69.

[45] Riccardo Rende, Federica Gerace, Alessandro Laio, and Sebastian Goldt. Optimal inference of a generalised Potts model by single-layer transformers with factored attention. *arXiv preprint arXiv:2304.07235*, 2023.

[46] Anian Ruoss, Grégoire Delétang, Tim Genewein, Jordi Grau-Moya, Róbert Csordás, Mehdi Bennani, Shane Legg, and Joel Veness. Randomized positional encodings boost length generalization of transformers. In Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki, editors, *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 1889–1903, Toronto, Canada, July 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.acl-short.161. URL https://aclanthology.org/2023.acl-short.161.

[47] Arda Sahiner, Tolga Ergen, Batu Ozturkler, John Pauly, Morteza Mardani, and Mert Pilanci. Unraveling attention via convex duality: Analysis and interpretations of vision transformers. In *International Conference on Machine Learning*, pages 19050–19088. PMLR, 2022.

[48] Rylan Schaeffer, Brando Miranda, and Sanmi Koyejo. Are emergent abilities of large language models a mirage? *Advances in Neural Information Processing Systems*, 36, 2024.

[49] Henry Schwarze. Learning a rule in a multilayer neural network. *Journal of Physics A: Mathematical and General*, 26(21):5781, 1993.

[50] Peter Shaw, Jakob Uszkoreit, and Ashish Vaswani. Self-Attention with Relative Position Representations. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 464–468. Association for Computational Linguistics, 2018. doi: 10.18653/v1/N18-2074. URL https://aclanthology.org/N18-2074.

[51] Aaditya K Singh, Stephanie C.Y. Chan, Ted Moskovitz, Erin Grant, Andrew M Saxe, and Felix Hill. The transient nature of emergent in-context learning in transformers. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023. URL https://openreview.net/forum?id=Of0GBzow8P.

[52] Koustuv Sinha, Amirhossein Kazemnejad, Siva Reddy, Joelle Pineau, Dieuwke Hupkes, and Adina Williams. The curious case of absolute position embeddings. In Yoav Goldberg, Zornitsa Kozareva, and Yue Zhang, editors, *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 4449–4472, Abu Dhabi, United Arab Emirates, December 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.findings-emnlp.326. URL https://aclanthology.org/2022.findings-emnlp.326.

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

[54] Mihailo Stojnic. Meshes that trap random subspaces. *arXiv preprint arXiv:1304.0003*, 2013.

[55] Mihailo Stojnic. Upper-bounding l1-optimization weak thresholds. *arXiv preprint arXiv:1303.7289*, 2013.

[56] Davoud Ataee Tarzanagh, Yingcong Li, Christos Thrampoulidis, and Samet Oymak. Transformers as support vector machines. *arXiv preprint arXiv:2308.16898*, 2023.

[57] Davoud Ataee Tarzanagh, Yingcong Li, Xuechen Zhang, and Samet Oymak. Max-margin token selection in attention mechanism. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023.

[58] Christos Thrampoulidis, Samet Oymak, and Babak Hassibi. The gaussian min-max theorem in the presence of convexity. *arXiv preprint arXiv:1408.4837*, 2014.

[59] Yuandong Tian, Yiping Wang, Beidi Chen, and Simon Du. Scan and snap: Understanding training dynamics and token composition in 1-layer transformer. *arXiv preprint arXiv:2305.16380*, 2023.

[60] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. Attention is all you need. In I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017. URL https://proceedings.neurips.cc/paper_files/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf.

[61] Taylor Webb, Keith J Holyoak, and Hongjing Lu. Emergent analogical reasoning in large language models. *Nature Human Behaviour*, 7(9):1526–1541, 2023.

[62] Jason Wei, Yi Tay, Rishi Bommasani, Colin Raffel, Barret Zoph, Sebastian Borgeaud, Dani Yogatama, Maarten Bosma, Denny Zhou, Donald Metzler, et al. Emergent abilities of large language models. *arXiv preprint arXiv:2206.07682*, 2022.

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

[64] Ruiqi Zhang, Spencer Frei, and Peter L Bartlett. Trained transformers learn linear models in-context. *arXiv preprint arXiv:2306.09927*, 2023.

[65] Ziqian Zhong, Ziming Liu, Max Tegmark, and Jacob Andreas. The clock and the pizza: Two stories in mechanistic explanation of neural networks, 2023.

Appendix A. Derivation of Main Result

In this Appendix, we provide a detailed derivation our works main technical result, as stated in subsection A.1. In subsection A.3 we introduce a Generalized Approximate Message Passing algorithm (GAMP) [42]. Subsection A.4 then establishes that equations (8) of Result 1 track the dynamics of summary statistics describing the GAMP algorithm. In particular, the equations (8) describe the fixed points of GAMP. Finally, subsection A.5 shows that fixed points of GAMP correspond to critical (zero-gradient) points of the empirical loss landscape (3), thus establishing that equations 8 of Result 1 describe fixed points of GD.

A.1. Main Result

Assumption 1 *The covariances $\{\boldsymbol{\Sigma}_\ell\}_{\ell=1}^L$ admit a common set of eigenvectors $\{\boldsymbol{e}_i\}_{i=1}^d$. We further note $\{\lambda_i^\ell\}_{i=1}^d$ the eigenvalues of $\boldsymbol{\Sigma}_\ell$. The eigenvalues $\{\lambda_i^\ell\}_{i=1}^d$ and the projection of the positional embedding $\{\boldsymbol{p}_\ell\}_{\ell=1}^L$ and the teacher columns $\{\boldsymbol{Q}_j^\star\}_{j=1}^{r_t}$ on the eigenvectors $\{\boldsymbol{e}_i^\top \boldsymbol{p}_\ell\}_{i,\ell}$, $\{\boldsymbol{e}_i^\top \boldsymbol{Q}_j^\star\}_{i,j}$ are assumed to admit a well-defined joint distribution ν as $d \to \infty$ – namely, for $\gamma = (\gamma_1, ..., \gamma_L) \in \mathbb{R}^L$,$\pi = (\pi_1, ..., \pi_{r_t}) \in \mathbb{R}^{r_t}$ and $\tau = (\tau_1, ..., \tau_L) \in \mathbb{R}^L$:*

$$\frac{1}{d}\sum_{i=1}^{d}\prod_{\ell=1}^{L}\delta\left(\lambda_i^\ell - \gamma_\ell\right)\delta\left(\sqrt{d}\boldsymbol{e}_i^\top \boldsymbol{p}_\ell - \tau_\ell\right)\prod_{j=1}^{r_t}\delta\left(\boldsymbol{e}_i^\top \boldsymbol{Q}_j^\star - \pi_j\right) \xrightarrow{d\to\infty} \nu\left(\gamma, \tau, \pi\right). \tag{6}$$

Result 1 *Under Assumption 1, in the limit $n, d \to \infty$, $\|\boldsymbol{p}\|, n/d, L, r_s, r_t = \Theta_d(1)$, the summary statistics*

$$\begin{aligned} \rho_\ell &\equiv \frac{\boldsymbol{Q}_\star^\top \boldsymbol{\Sigma}_\ell \boldsymbol{Q}_\star}{d} \in \mathbb{R}^{r_t \times r_t}, & q_\ell &\equiv \frac{\hat{\boldsymbol{Q}}^\top \boldsymbol{\Sigma}_\ell \hat{\boldsymbol{Q}}}{d} \in \mathbb{R}^{r_s \times r_s}, \\ m_\ell &\equiv \frac{\hat{\boldsymbol{Q}}^\top \boldsymbol{p}_\ell}{d} \in \mathbb{R}^{r_s}, & \theta_\ell &\equiv \frac{\hat{\boldsymbol{Q}}^\top \boldsymbol{\Sigma}_\ell \boldsymbol{Q}_\star}{d} \in \mathbb{R}^{r_s \times r_t} \end{aligned} \tag{7}$$

concentrate in probability, and are solutions of the set of finite-dimensional self-consistent equations

$$\begin{cases} q_\ell = \int d\nu(\gamma,\tau,\pi)\gamma_\ell \left(\lambda \mathbb{I}_r + \sum_{\kappa=1}^L \gamma_\kappa \hat{V}_\kappa\right)^{-1}\left(\sum_{\kappa=1}^L \gamma_\kappa \hat{q}_\kappa + \left(\sum_{\kappa=1}^L \hat{m}_\kappa \tau_\kappa + \gamma_\kappa \hat{\theta}_\kappa \cdot \pi\right)^{\otimes 2}\right)\left(\lambda \mathbb{I}_r + \sum_{\kappa=1}^L \gamma_\kappa \hat{V}_\kappa\right)^{-1} \\ V_\ell = \int d\nu(\gamma,\tau,\pi)\gamma_\ell \left(\lambda \mathbb{I}_r + \sum_{\kappa=1}^L \gamma_\kappa \hat{V}_\kappa\right)^{-1} \\ m_\ell = \int d\nu(\gamma,\tau,\pi)\tau_\ell \left(\lambda \mathbb{I}_r + \sum_{\kappa=1}^L \gamma_\kappa \hat{V}_\kappa\right)^{-1}\left(\sum_{\kappa=1}^L \hat{m}_\kappa \tau_\kappa + \gamma_\kappa \hat{\theta}_\kappa \cdot \pi\right) \\ \theta_\ell = \int d\nu(\gamma,\tau,\pi)\gamma_\ell \left(\lambda \mathbb{I}_r + \sum_{\kappa=1}^L \gamma_\kappa \hat{V}_\kappa\right)^{-1}\left(\sum_{\kappa=1}^L \hat{m}_\kappa \tau_\kappa + \gamma_\kappa \hat{\theta}_\kappa \cdot \pi\right)\pi^\top. \end{cases} \tag{8}$$

$$\begin{cases} \hat{q}_\ell = \alpha \mathbb{E}_{\Xi,U} V_\ell^{-1}\left(\text{prox}(\Xi, U)_\ell - q_\ell^{\frac{1}{2}}\xi_\ell - m_\ell\right)^{\otimes 2} V_\ell^{-1} \\ \hat{V}_\ell = \hat{\theta}_\ell \theta_\ell^\top q_\ell^{-1} - \alpha \mathbb{E}_{\Xi,U} V_\ell^{-1}\left(\text{prox}(\Xi, U)_\ell - q_\ell^{\frac{1}{2}}\xi_\ell - m_\ell\right)\xi_\ell^\top q_\ell^{-\frac{1}{2}} \\ \hat{m}_\ell = \alpha \mathbb{E}_{\Xi,U} V_\ell^{-1}\left(\text{prox}(\Xi, U)_\ell - q_\ell^{\frac{1}{2}}\xi_\ell - m_\ell\right) \\ \hat{\theta}_\ell = \alpha \mathbb{E}_{\Xi,U} V_\ell^{-1}\left(\text{prox}(\Xi, U)_\ell - q_\ell^{\frac{1}{2}}\xi_\ell - m_\ell\right)\left(u_\ell - \xi_\ell^\top q_\ell^{-1/2}\theta_\ell\right)^\top \left(\rho_\ell - \theta_\ell^\top q_\ell^{-1}\theta_\ell\right)^{-1} \end{cases} \tag{9}$$

In (8), $U = \{u_\ell\}_{\ell=1}^L$ and $\Xi = \{\xi_\ell\}_{\ell=1}^L$, with $u_\ell \sim \mathcal{N}(\xi_\ell^\top q_\ell^{-1/2}\theta_\ell, \rho_\ell - \theta_\ell^\top q_\ell^{-1}\theta_\ell)$ and $\xi_\ell \sim \mathcal{N}(0, \mathbb{I}_{r_s})$, and $\cdot^{\otimes 2}$ denotes the outer product of a vector with itself. Finally, the resolvents $\{\mathrm{prox}(\Xi, U)_\ell\}_{\ell=1}^L$ are defined as the minimizers of the Moreau envelope

$$\mathcal{M}(\Xi, U) = \inf_{z_1,\dots,z_L} \left\{ \sum_{\ell=1}^L \mathrm{Tr}\left[V_\ell^{-1} \left(x_\ell - q_\ell^{1/2}\xi_\ell - m_\ell \right)^{\otimes 2} \right] + \mathrm{Tr}\left[\mathtt{S}(Z)\rho_\Sigma \mathtt{S}(Z)^\top \right] - 2\mathrm{Tr}\left[\mathtt{T}(U)\rho_\Sigma \mathtt{S}(Z)^\top \right] \right\}.$$

We noted $Z \in \mathbb{R}^{L\times r_s}$ (resp. $U \in \mathbb{R}^{L\times r_t}$) the matrix whose rows are z_ℓ (resp. u_ℓ) and:

$$\rho_\Sigma \equiv \mathrm{diag}\left[\left(\int d\nu(\gamma,\tau,\pi)\gamma_\ell \right)_{\ell=1}^L \right] \in \mathbb{R}^{L\times L}. \tag{10}$$

In the same limit, the test error (3) converges in probability to

$$\epsilon_g = \mathbb{E}_h \mathrm{Tr}\left[\mathtt{S}[h]\rho_\Sigma \mathtt{S}[h]^\top \right] + \mathbb{E}_{h^\star} \mathrm{Tr}\left[\mathtt{T}[h^\star]\rho_\Sigma \mathtt{T}[h^\star]^\top \right] - 2\mathbb{E}_{h,h^\star} \mathrm{Tr}\left[\mathtt{S}[h]\rho_\Sigma \mathtt{T}[h^\star]^\top \right]. \tag{11}$$

where the average bears on $h \in \mathbb{R}^{L\times r_s}, h^\star \in \mathbb{R}^{L\times r_t}$ with independent rows with statistics

$$(h_\ell, h_\ell^\star) \sim \mathcal{N}\left[\begin{pmatrix} m_\ell \\ 0 \end{pmatrix}, \left(\begin{array}{c|c} q_\ell & \theta_\ell \\ \hline \theta_\ell^\top & \rho_\ell \end{array} \right) \right] \tag{12}$$

Finally, the training loss ϵ_t converges in probability to

$$\epsilon_t = \alpha \mathbb{E}_{Y,\Xi}\mathcal{M} - \frac{1}{2}\sum_{\ell=1}^L \mathrm{Tr}[\hat{q}_\ell V_\ell] + \frac{\lambda}{2}\int d\nu(\gamma,\tau)\mathrm{Tr}\left[\left(\lambda + \sum_{\ell=1}^L \gamma_\ell \hat{V}_\ell \right)^{-1} \left(\sum_{\ell=1}^L \gamma_\ell \hat{q}_\ell + \left(\sum_{\ell=1}^L \tau_\ell \hat{m}_\ell + \hat{\theta}_\ell \cdot \pi \right)^{\otimes 2} \right) \right]. \tag{13}$$

A.2. Notations

For simplicity, we place ourselves in the setting $r_s = 1$ explored in Section 4 of the main text, but allow the length L of the sentences to be arbitrary, and allow a generic learning model $\mathtt{S}$ (2), i.e. not necessarily the dot-product attention model analyzed in Section 4. The case $r_s \geq 2$ follows identical derivation steps, modulo the replacement of all variables by tensor objects. We provide another alternative derivation of Result 1 in full generality in Appendix B, using the replica method from statistical physics. Let us note $\{X_\ell\}_{1\leq\ell\leq L}$ a series of L $n \times d$ matrices, with X_ℓ corresponding to the $\ell-$th rows (tokens) of each input sentence x^μ stacked vertically, and normalized by $\sqrt{d}$. We denote $\tilde{X}_\ell \equiv X_\ell + P_\ell$, where $P \in \mathbb{R}^{n\times d}$ is the matrix with all rows equal to the $\ell-$th positional encoding p_ℓ. Let us further define $\rho \in \mathbb{R}^{n\times L\times L}$ the tensor corresponding to the sequence of n matrices $\frac{1}{d}x^\mu(x^\mu)^\top \in \mathbb{R}^{L\times L}$. Finally, let us denote $T \in \mathbb{R}^{n\times L\times L}$ the tensor so that the $\mu-$th row of T satisfies $y(x^\mu) = T^\mu x^\mu$, see equation (1). In other words, T corresponds to the concatenation of the target attention matrices.

Before detailing the derivation, we first highlight a simplifying observation. Note that a loss item can be expanded as

$$\begin{aligned} \frac{1}{d}\left\| y(x^\mu) - \mathtt{S}\left[\frac{1}{\sqrt{d}}(x^\mu + p)Q \right] x^\mu \right\|^2 = &\|y(x^\mu)\|^2 + \mathrm{Tr}\,\mathtt{S}\left[\frac{1}{\sqrt{d}}(x^\mu + p)Q \right] \rho_\Sigma \mathtt{S}\left[\frac{1}{\sqrt{d}}(x^\mu + p)Q \right]^\top \\ &- 2\,\mathrm{Tr}\,\mathtt{T}\left[\frac{1}{\sqrt{d}}x_\ell Q_\star \right] \rho_\Sigma \mathtt{S}\left[\frac{1}{\sqrt{d}}(x_\ell^\mu + p_\ell)Q \right]^\top, \end{aligned} \tag{14}$$

where we used that with high probability in the considered asymptotic limit, for all $1 \leq \mu \leq n$,

$$xx^\top = (x+p)(x+p)^\top = x(x+p)^\top = \rho_\Sigma. \tag{15}$$

Since the first term of (14) does not depend on the weights Q, it can be without loss of generality substracted from the loss. Without loss of generality, one can thus consider the equivalent empirical risk minimization problem

$$\hat{\boldsymbol{Q}} = \underset{\boldsymbol{Q}\in\mathbb{R}^{d\times r}}{\operatorname{argmin}} \sum_{\mu=1}^{n} \frac{1}{2d} \left[\operatorname{Tr} \mathtt{S}\left[\tfrac{1}{\sqrt{d}}(x^\mu+p)Q\right]\rho_\Sigma \mathtt{S}\left[\tfrac{1}{\sqrt{d}}(x^\mu+p)Q\right]^\top - 2\operatorname{Tr}\mathtt{T}\left[\tfrac{1}{\sqrt{d}}x_\ell Q_\star\right]\rho_\Sigma \mathtt{S}\left[\tfrac{1}{\sqrt{d}}(x_\ell^\mu+p_\ell)Q\right]^\top \right] + \frac{\lambda}{2}\|\boldsymbol{Q}\|^2. \tag{16}$$

The risks (16) and (3) are equivalent, and we shall use the former in the following.

Finally, for arguments $T \in \mathbb{R}^{L\times L}, \rho \in \mathbb{R}^{L\times L}, \omega \in \mathbb{R}^L, V \in \mathbb{R}^{L\times L}$ we introduce the resolvent

$$\operatorname{prox}(T,\rho,\omega,V) \equiv \underset{x=\{x_\ell\in\mathbb{R}\}_{\ell=1}^L}{\operatorname{arginf}} \left\{ \sum_{\ell,\kappa=1}^{L} (x_\ell-\omega_\ell)(V^{-1})_{\ell\kappa}(x_\kappa-\omega_\kappa) - 2\operatorname{Tr}\Big[\mathtt{S}[x]\rho T^\top\Big] + \operatorname{Tr}\Big[\mathtt{S}[x]\rho \mathtt{S}[x]^\top\Big] \right\} \tag{17}$$

Note that the latter part of the bracketed term corresponds to the simplified loss (14) derived in the beginning of Appendix B, which is the one we shall without loss of generality consider in the present appendix. For ease of presentation, we place ourselves under Assumption 1, where all the input covariances $\{\Sigma_\ell\}_\ell$ are codiagonalizable. In the following, without loss of generality, we thus assume them diagonal, by placing ourselves in the common basis $\{e_i\}_{1\leq i\leq d}$ of Assumption 1.

A.3. AMP algorithm

We are now in a position to state the AMP algorithm:

Algorithm 1 GAMP

Inputs : $\{\tilde{X}_\ell \in \mathbb{R}^{n\times d}\}_{\ell=1}^L, T \in \mathbb{R}^{n\times L\times L}, \rho \in \mathbb{R}^{n\times L\times L}$
Initialize $\hat{Q}^0 = \sim \mathcal{N}(0, \mathbb{I}_d), \hat{c}^0 = \mathbb{I}_d, \{f_\ell^0 = 0_n\}_{\ell=1}^L$
for $t \leq t_{\max}$ **do**

$$\forall 1 \leq \ell, \kappa \leq L, \;\; V_{\ell\kappa}^t = (\tilde{X}_\ell \odot \tilde{X}_\kappa)\hat{c}^t$$

$$\forall 1 \leq \ell \leq L, \;\; \omega_\ell^t = \tilde{X}_\ell \hat{Q}^t - \sum_{\kappa=1}^L V_{\ell\kappa}^t f_\kappa^{t-1}$$

$$\forall 1 \leq \ell \leq L, \;\; f_\ell^t = \sum_\kappa (V^{-1})_{\ell\kappa}(\mathrm{prox}(T, \rho, \omega^t, V^t)_\kappa - \omega_\kappa^t)$$

$$\forall 1 \leq \ell, \kappa \leq L, \;\; g_{\ell\kappa}^t = \partial_{\omega_\ell} f_\kappa^t$$

$$A^t = -\sum_{\ell,\kappa=1}^L (\tilde{X}_\ell \odot \tilde{X}_\kappa)^\top g_{\ell\kappa}^t$$

$$b^t = \sum_{\ell=1}^L \tilde{X}_\ell^\top f_\ell^t + A^t \odot \hat{Q}^t$$

$$\hat{Q}^{t+1} = (\lambda \mathbb{I}_d + A^t)^{-1} b^t$$

$$\hat{c}^{t+1} = (\lambda \mathbb{I}_d + A^t)^{-1}$$

end for

return Estimator $\hat{Q}$

The GAMP algorithm can be derived in standard fashion from the Belief Propagation (BP) algorithm, see e.g. [6, 42] or [63] for an overview. Compared to the standard GAMP iterations for Generalized linear models, one needs to account for the fact that there exist different sources of data X_ℓ (corresponding to the $\ell-$ th tokens of each input sentence), and for the fact that the output of the equivalent GLM are $\mathbb{R}^{L\times L}$- valued attention matrices. In the following subsection, we show that the fixed points of GAMP 1 correspond to critical points of the empirical loss (3), i.e. fixed points of Gradient Descent (GD), allowing to connect Result 1 to our numerical experiments using GD.

A.4. State evolution

In this section we show that the dynamics of the GAMP Algorithm 1 are tracked by the summary statistics of Result 1. In particular, the equations (8) describe the statistics of the GAMP fixed points. To see this, it is convenient to take as a starting point the relaxed Belief Propagation (rBP) equations, which are a step upstream in the derivation of the GAMP iterations, and which are asymptotically equivalent– see e.g. [63] for a review or e.g. [12], Appendix A, for a detailed walkthrough. The rBP equations read

As conventional, we note $\cdot_\mu$ the version of a variable $\cdot_{\mu\to i}$ where the summation also encompasses the index i, and $\cdot_i$ the version of a variable $\cdot_{i\to\mu}$ where the summation also encompasses the index μ. Note that in all cases above the two variables differ by at most $\Theta_d(1/\sqrt{d})$.

Concentration of $(V_{\mu\to i}^t)_{\ell\kappa}$ We first study the statistics of $V_{\mu\to i}^t, A_{i\to\mu}^t$, remembering that the data $\tilde{x}_{\ell i}^\mu \equiv {}^{(x_\ell^\mu)_i}/\sqrt{d} + {}^{(p_\ell)_i}/\sqrt{d}$ in the notation of the main text, with $(x_\ell^\mu)_i = \Theta_d(1), \;\; (p_\ell)_i = \Theta_d(1/\sqrt{d})$.

Algorithm 2 rBP

Inputs : $\{\tilde{X}_\ell \in \mathbb{R}^{n\times d}\}_{\ell=1}^L, T \in \mathbb{R}^{n\times L\times L}, \rho \in \mathbb{R}^{n\times L\times L}$

Initialize $\forall 1\leq \mu\leq n,\ \ 1\leq i\leq d,\ \ \hat{Q}^0_{i\to\mu}=0, \hat{c}^0_{i\to\mu}=1, \{f^0_{\ell\mu\to i}=0\}_{\ell=1}^L$

for $t\leq t_{\max}$ **do**

$\forall 1\leq \ell,\kappa\leq L, 1\leq\mu\leq n, 1\leq i\leq d,\ \ (V^t_{\mu\to i})_{\ell\kappa} = \sum_{j\neq i}(\tilde{x}^\mu_{\ell j})(\tilde{x}^\mu_{\kappa j})\hat{c}^t_{j\to\mu}$

$\forall 1\leq \ell, 1\leq\mu\leq n, 1\leq i\leq d,\ \ \omega^t_{\ell,\mu\to i} = \sum_{j\neq i}\tilde{x}^\mu_{\ell,i}\hat{Q}_{j\to\mu}$

$\forall 1\leq \ell, 1\leq\mu\leq n, 1\leq i\leq d,\ \ f^t_{\ell,\mu\to i} = \sum_\kappa (V^{-1}_{\mu\to i})_{\ell\kappa}(\mathrm{prox}(T_\mu,\rho_\mu,\omega^t_{\mu\to i},V^t_{\mu\to i})_\kappa - \omega^t_{\kappa,\mu\to i})$

$\forall 1\leq \ell,\kappa\leq L, 1\leq\mu\leq n, 1\leq i\leq d,\ \ g^t_{\ell\kappa,\mu\to i} = \partial_{\omega_\ell} f^t_{\kappa\mu\to i}$

$\forall 1\leq \mu\leq n, 1\leq i\leq d, A^t_{i\to\mu} = -\sum_{\ell,\kappa=1}^L\sum_{\nu\neq\mu}(\tilde{x}^\nu_{\ell i})(\tilde{x}^\nu_{\kappa i})g^t_{\ell\kappa,\nu\to i}$

$\forall 1\leq \mu\leq n, 1\leq i\leq d,\ \ b^t_{i\to\mu} = \sum_{\ell=1}^L\sum_{\nu\neq\mu} x^\nu_{\ell i} f^t_{\ell,\nu\to i}$

$\forall 1\leq \mu\leq n, 1\leq i\leq d,\ \ \hat{Q}^{t+1}_{i\to\mu} = (\lambda\mathbb{I}_d + A^t_{i\to\mu})^{-1} b^t_{i\to\mu}$

$\forall 1\leq \mu\leq n, 1\leq i\leq d,\ \ \hat{c}^{t+1}_{i\to\mu} = (\lambda\mathbb{I}_d + A^t_{i\to\mu})^{-1}$

end for

return Estimator $\hat{Q}$

Replacing in the rBP updates:

$$
\begin{aligned}
(V^t_{\mu\to i})_{\ell\kappa} &= \sum_{j\neq i}(\tilde{x}^\mu_{\ell j})(\tilde{x}^\mu_{\kappa j})\hat{c}^t_{j\to\mu} \\
&= \underbrace{\frac{1}{d}\sum_{j\neq i}(x^\mu_\ell)_j(x^\mu_\kappa)_j\hat{c}^t_{j\to\mu}}_{\delta_{\ell\kappa}\Theta_d(1)+(1-\delta_{\ell\kappa})\Theta_d(1/\sqrt{d})} + \underbrace{\frac{1}{d}\sum_{j\neq i}(x^\mu_\ell)_j(p_\kappa)_j\hat{c}^t_{j\to\mu} + (\ell\leftrightarrow\kappa)}_{\Theta_d(1/d)} + \underbrace{\frac{1}{d}\sum_{j\neq i}(p_\ell)_j(p_\kappa)_j\hat{c}^t_{j\to\mu}}_{\Theta_d(1/d)} \\
&= \delta_{\ell\kappa}\frac{1}{d}\sum_j(\Sigma_\ell)_{jj}\hat{c}^t_j \equiv V^t_\ell
\end{aligned}
\tag{18}
$$

Distribution of $\omega^t_{\ell,\mu\to i}$ Let us first introduce the teacher local field

$$
h_{\mu,\ell} = \sum_i (x^\mu_\ell)_i Q^\star_i. \tag{19}
$$

Like e.g. [12], Appendix A, we first ascertain the joint distribution of $h_{\mu,\ell}, \omega^t_{\ell,\mu\to i}$ with respect to the data. These variables have mean

$$
\mathbb{E}[\omega^t_{\ell,\mu\to i}] = \frac{p_\ell^\top \hat{Q}^t}{\sqrt{d}} \equiv m^t_\ell \tag{20}
$$

and respective variance

$$\mathbb{V}[\omega^t_{\ell,\mu\to i}\omega^t_{\kappa,\nu\to j}] = \delta_{\mu\nu}\delta_{\ell\kappa}\frac{1}{d}\sum_{i,j}\hat{Q}^t_i(\Sigma_\ell)_{ij}\hat{Q}^t_j \equiv \delta_{\mu\nu}\delta_{\ell\kappa}q^t_\ell \tag{21}$$

$$\mathbb{E}[h_{\mu\ell}h_{\nu\kappa}] = \delta_{\mu\nu}\delta_{\ell\kappa}\frac{1}{d}\sum_{i,j}Q^\star_i(\Sigma_\ell)_{ij}Q^\star_j \equiv \delta_{\mu\nu}\delta_{\ell\kappa}\rho_\ell \tag{22}$$

$$\mathbb{E}[h_{\mu\ell}\omega^t_{\kappa,\nu\to j}] = \delta_{\mu\nu}\delta_{\ell\kappa}\frac{1}{d}\sum_{i,j}Q^\star_i(\Sigma_\ell)_{ij}\hat{Q}^t_j \equiv \delta_{\mu\nu}\delta_{\ell\kappa}\theta^t_\ell \tag{23}$$

Distribution of $b^t_{i\to\mu}$ Let us ascertain the distribution of $b^t_{i\to\mu}$.

$$\begin{aligned} b^t_{i\to\mu} &= \sum_\ell\sum_{\nu\neq\mu}(\tilde{x}^\nu_\ell)_i\underbrace{(V^t_\ell)^{-1}\left(\mathrm{prox}(T_\nu,\rho_\nu,\{\omega^t_{\kappa,\nu\to i}\}_\kappa,V^t_{\nu\to i})_\ell - \omega^t_{\ell,\nu\to i}\right)}_{\equiv\tilde{\mathrm{prox}}(T_\nu,\rho_\nu,\{\omega^t_{\kappa,\nu\to i}\}_\kappa,V^t_{\nu\to i})_\ell} \\ &= \sum_\ell\sum_{\nu\neq\mu}{}^1\!/\!\sqrt{d}((x^\nu_\ell)_i+(p_\ell)_i)\Bigg[\tilde{\mathrm{prox}}(\mathtt{T}[\{h_{\nu\to i,\kappa}\}_\kappa],\rho_\nu,\{\omega^t_{\kappa,\nu\to i}\}_\kappa,V^t_{\nu\to i})_\ell \\ &\quad + {}^1\!/\!\sqrt{d}\sum_\gamma(x^\nu_\gamma)_iQ^\star_i\partial_{h_\gamma}\tilde{\mathrm{prox}}(\mathtt{T}[\{h_{\nu\to i,\kappa}\}_\kappa],\rho_\nu,\{\omega^t_{\kappa,\nu\to i}\}_\kappa,V^t_{\nu\to i})_\ell\Bigg], \end{aligned} \tag{24}$$

leading asymptotically to

$$\begin{aligned} \mathbb{E}[b^t_{i\to\mu}] &= \sum_\ell(\sqrt{d}p_\ell)_i\underbrace{\alpha\mathbb{E}_{H=\{h_\kappa\}\Xi=\{\xi_\kappa\}}\tilde{\mathrm{prox}}(\mathtt{T}[H],\rho_\Sigma,\{m^t_\kappa+\sqrt{q^t_\kappa}\xi_\kappa\}_\kappa,\{V^t_\kappa\}_\kappa)_\ell}_{\equiv\hat{m}^t_\ell} \\ &\quad + Q^\star_i\sum_\ell(\Sigma_\ell)_{ii}\underbrace{\alpha\mathbb{E}_{H,\Xi}\partial_{h_\ell}\tilde{\mathrm{prox}}(\mathtt{T}[H],\rho_\Sigma,\{m^t_\kappa+\sqrt{q^t_\kappa}\xi_\kappa\}_\kappa,\{V^t_\kappa\}_\kappa)_\ell}_{\equiv\hat{\theta}^t_\ell} \end{aligned} \tag{25}$$

where the expectations bear over $\xi_\ell\sim\mathcal{N}(0,1)$ and $h_\ell\sim\mathcal{N}(\xi_\ell\theta^t_\ell/\sqrt{q^t_\ell},\rho_\ell-{}^{(\theta^t_\ell)^2}\!/\!_{q^t_\ell})$. The variance is given by

$$\mathbb{V}[b^t_i,b^t_j] = \delta_{ij}\sum_\ell(\Sigma_\ell)_{ii}\underbrace{\alpha\mathbb{E}_{H,\Xi}\tilde{\mathrm{prox}}(\mathtt{T}[H],\rho_\Sigma,\{m^t_\kappa+\sqrt{q^t_\kappa}\xi_\kappa\}_\kappa,\{V^t_\kappa\}_\kappa)^2_\ell}_{\equiv\hat{q}^t_\ell} \tag{26}$$

Concetration of $A^t_{i\to\mu}$ Similarly to the derivation for $V^t_{\mu\to i}$, $A^t_{i\to\mu}$ concentrates to

$$A^t_{i\to\mu} = \sum_\ell\underbrace{-\alpha\frac{1}{V^t_\ell}\left(\mathbb{E}_{H,\Xi}\partial_{\omega_\ell}\mathrm{prox}(\mathtt{T}[H],\rho_\Sigma,\{m^t_\kappa+\sqrt{q^t_\kappa}\xi_\kappa\}_\kappa,\{V^t_\kappa\}_\kappa)_\ell-1\right)}_{\equiv\hat{V}^t_\ell}(\Sigma_\ell)_{ii} \tag{27}$$

Recovering Result 1 Wrapping up, we now massage these equations to recover equations (8) from Result 1 of the main text. Starting from (18):

$$\begin{aligned} V^t_\ell &= \frac{1}{d}\sum_j(\Sigma_\ell)_{jj}\frac{1}{\lambda+\sum_\kappa\hat{V}^{t-1}_\kappa(\Sigma_\kappa)} \\ &= \int d\nu(\gamma,\tau)\gamma_\ell\left(\lambda+\sum_\kappa\hat{V}^{t-1}_\kappa\gamma_\kappa\right)^{-1}. \end{aligned} \tag{28}$$

Next, for q_ℓ^t (21):

$$q_\ell^t = \frac{1}{d}\sum_i (\Sigma_\ell)_{ii}\left(\left(\sum_\kappa (\sqrt{d}(p_\kappa)_i \hat{m}_\kappa^{t-1} + Q_i^\star(\Sigma_\kappa)_{ii}\hat{\theta}_\kappa\right)^2 + (\Sigma_\kappa)_{ii}\hat{q}_\kappa^{t-1}\right)\left(\lambda + \sum_\kappa \hat{V}_\kappa^{t-1}(\Sigma_\kappa)\right)^{-2}$$
$$= \int d\nu(\gamma,\tau,\pi)\gamma_\ell\left(\left(\sum_\kappa \hat{m}_\kappa^{t-1}\tau_\kappa + \hat{\theta}_\kappa\gamma_\kappa\pi_\kappa\right)^2 + \gamma_\kappa\hat{q}_\kappa^{t-1}\right)\left(\lambda + \sum_\kappa \hat{V}_\kappa^{t-1}\gamma_\kappa\right)^{-2} \tag{29}$$

For θ_ℓ^t(21):

$$\theta_\ell^t = \frac{1}{d}\sum_1 (\Sigma_\ell)_{ii} Q_i^\star\left(\sum_\kappa (\sqrt{d}(p_\kappa)_i \hat{m}_\kappa^{t-1} + Q_i^\star(\Sigma_\kappa)_{ii}\hat{\theta}_\kappa\right)\left(\lambda + \sum_\kappa \hat{V}_\kappa^{t-1}(\Sigma_\kappa)\right)^{-1} + o_d(1)$$
$$= \int d\nu(\gamma,\tau,\pi)\gamma_\ell\pi_\ell\left(\sum_\kappa \hat{m}_\kappa^{t-1}\tau_\kappa + \hat{\theta}_\kappa\gamma_\kappa\pi_\kappa\right)\left(\lambda + \sum_\kappa \hat{V}_\kappa^{t-1}\gamma_\kappa\right)^{-1}. \tag{30}$$

Finally for m_ℓ^t (20):

$$m_\ell^t = \frac{1}{d}\sum_i (\sqrt{d}p_\ell)_i\left(\sum_\kappa (\sqrt{d}(p_\kappa)_i \hat{m}_\kappa^{t-1} + Q_i^\star(\Sigma_\kappa)_{ii}\hat{\theta}_\kappa\right)\left(\lambda + \sum_\kappa \hat{V}_\kappa^{t-1}(\Sigma_\kappa)\right)^{-1} + o_d(1)$$
$$= \int d\nu(\gamma,\tau,\pi)\tau_\ell\left(\sum_\kappa \hat{m}_\kappa^{t-1}\tau_\kappa + \hat{\theta}_\kappa\gamma_\kappa\pi_\kappa\right)\left(\lambda + \sum_\kappa \hat{V}_\kappa^{t-1}\gamma_\kappa\right)^{-1}. \tag{31}$$

For $\hat{m}_\ell^t$ (25):

$$\hat{m}_\ell^t = \alpha\mathbb{E}_{H,\Xi}\frac{1}{V_\ell^t}\left[\mathrm{prox}_\ell - \sqrt{q_\ell^t}\xi_\ell - m_\ell^t\right], \tag{32}$$

while for $\hat{\theta}_\ell^t$ (25):

$$\hat{\theta}_\ell^t = \alpha\mathbb{E}_{H,\Xi}\frac{1}{V_\ell^t}\partial_{h_\ell}\left[\mathrm{prox}_\ell - \sqrt{q_\ell^t}\xi_\ell - m_\ell^t\right]$$
$$= \alpha\mathbb{E}_{H,\Xi}\frac{1}{V_\ell^t}\frac{h_\ell - \theta_\ell^t/\sqrt{q_\ell^t}\xi_\ell}{\rho_\ell - (\theta_\ell^t)^2/q_\ell^t}\left[\mathrm{prox}_\ell - \sqrt{q_\ell^t}\xi_\ell - m_\ell^t\right]. \tag{33}$$

Now turning to $\hat{q}_\ell^t$:

$$\hat{q}_\ell^t = \alpha\mathbb{E}_{H,\Xi}\left[\left(\frac{1}{V_\ell^t}\mathrm{prox}_\ell - \sqrt{q_\ell^t}\xi_\ell - m_\ell^t\right)^2\right]. \tag{34}$$

Finally, for $\hat{V}_\ell^t$ (27):

$$\begin{aligned}
\hat{V}_\ell^t &= -\alpha \mathbb{E}_{H,\Xi} \frac{1}{V_\ell^t} \left[\partial_{\omega_\ell} \mathrm{prox}_\ell - 1\right] \\
&= -\alpha \mathbb{E}_{H,\Xi} \frac{1}{V_\ell^t} \left[\frac{1}{\sqrt{q_\ell^t}} \partial_\xi (\mathrm{prox}_\ell - \sqrt{q_\ell^t}\xi_\ell - m_\ell)\right] \\
&= \alpha \mathbb{E}_{H,\Xi} \frac{1}{\sqrt{q_\ell^t} V_\ell^t} \left[\frac{\theta_\ell^t}{\sqrt{q_\ell^t} V_\ell^t} \left(\frac{h_\ell - \sqrt{q_\ell^t}\xi_\ell}{\rho_\ell - (\theta_\ell^t)^2/q_\ell^t} - \xi\right) (\mathrm{prox}_\ell - \sqrt{q_\ell^t}\xi_\ell - m_\ell)\right] \\
&= \frac{\theta_\ell^t \hat{\theta}_\ell^t}{q_\ell^t} - \alpha \mathbb{E}_{H,\Xi} \frac{1}{\sqrt{q_\ell^t} V_\ell^t} (\mathrm{prox}_\ell - \sqrt{q_\ell^t}\xi_\ell - m_\ell)\xi_\ell
\end{aligned} \tag{35}$$

Summary : State evolution equations The state evolution equations asymptotically describing the dynamics of the GAMP algorithm 1 thus read

$$\begin{cases}
V_\ell^t = \int d\nu(\gamma,\tau)\gamma_\ell \left(\lambda + \sum_\kappa \hat{V}_\kappa^{t-1}\gamma_\kappa\right)^{-1} \\
q_\ell^t = \int d\nu(\gamma,\tau,\pi)\gamma_\ell \left(\left(\sum_\kappa \hat{m}_\kappa^{t-1}\tau_\kappa + \hat{\theta}_\kappa\gamma_\kappa\pi_\kappa\right)^2 + \gamma_\kappa \hat{q}_\kappa^{t-1}\right)\left(\lambda + \sum_\kappa \hat{V}_\kappa^{t-1}\gamma_\kappa\right)^{-2} \\
\theta_\ell^t = \int d\nu(\gamma,\tau,\pi)\gamma_\ell\pi_\ell \left(\sum_\kappa \hat{m}_\kappa^{t-1}\tau_\kappa + \hat{\theta}_\kappa\gamma_\kappa\pi_\kappa\right)\left(\lambda + \sum_\kappa \hat{V}_\kappa^{t-1}\gamma_\kappa\right)^{-1} \\
m_\ell^t = \int d\nu(\gamma,\tau,\pi)\tau_\ell \left(\sum_\kappa \hat{m}_\kappa^{t-1}\tau_\kappa + \hat{\theta}_\kappa\gamma_\kappa\pi_\kappa\right)\left(\lambda + \sum_\kappa \hat{V}_\kappa^{t-1}\gamma_\kappa\right)^{-1}.
\end{cases} \tag{36}$$

$$\begin{cases}
\hat{V}_\ell^t = \frac{\theta_\ell^t\hat{\theta}_\ell^t}{q_\ell^t} - \alpha \mathbb{E}_{H,\Xi} \frac{1}{\sqrt{q_\ell^t}V_\ell^t}(\mathrm{prox}_\ell - \sqrt{q_\ell^t}\xi_\ell - m_\ell)\xi_\ell \\
\hat{q}_\ell^t = \alpha \mathbb{E}_{H,\Xi}\left[\left(\frac{1}{V_\ell^t}\mathrm{prox}_\ell - \sqrt{q_\ell^t}\xi_\ell - m_\ell^t\right)^2\right] \\
\hat{\theta}_\ell^t = \alpha \mathbb{E}_{H,\Xi} \frac{1}{V_\ell^t} \frac{h_\ell - \theta_\ell^t/\sqrt{q_\ell^t}\xi_\ell}{\rho_\ell - (\theta_\ell^t)^2/q_\ell^t}\left[\mathrm{prox}_\ell - \sqrt{q_\ell^t}\xi_\ell - m_\ell^t\right] \\
\hat{m}_\ell^t = \alpha \mathbb{E}_{H,\Xi} \frac{1}{V_\ell^t}\left[\mathrm{prox}_\ell - \sqrt{q_\ell^t}\xi_\ell - m_\ell^t\right]
\end{cases} \tag{37}$$

which exactly recovers equations (8) of Result 1 of the main text, for the case $r_s = 1$ considered in the present Appendix. Again, we mention that the case $r_s \geq 2$ should follow straightforwardly with the exact same derivation steps, using tensor variables (see e.g. [13]). This subsection has thus established that the equations (8) (with time indices) describe the summary statistics capturing the dynamics of GAMP iterations 1. In particular, (8) describe the fixed points of GAMP. The next subsection further shows that the (stable) fixed points of GAMP correspond to critical (zero-gradient) points of the empirical landscape (3), i.e. fixed points of gradient descent. Finally, we provide in Appendix B an alternative derivation of the state evolution equations (8)(36), using the replica method from statistical physics [40, 41].

A.5. Fixed points of GAMP are fixed points of GD

In this subsection, we show that fixed points of GAMP 1, as asymptotically described by (8) in Result 1, correspond to critical (zero gradient) points of the empirical landscape (3). Again, we present the

result for $r_s = 1$ for clarity, the generalization to $r_s \geq 2$ being straightforward (see e.g. [13]). In the previous notations, let us denote the (simplified, see (14)) empirical loss as

$$L(\{\tilde{X}_\ell Q\}_\ell) + g(Q) \tag{38}$$

where we introduced the shorthands

$$L(\{h_\ell\} \in \mathbb{R}^n) \equiv \sum_{\mu=1}^{n} \left(-2 \operatorname{Tr}\left[\mathtt{S}[\{h_\ell^\mu\}_\ell] \rho_\mu T_\mu^\top \right] + \operatorname{Tr}\left[\mathtt{S}[\{h_\ell^\mu\}_\ell] \rho_\mu \mathtt{S}[\{h_\ell^\mu\}_\ell]^\top \right] \right) \tag{39}$$

$$g(Q) \equiv \frac{\lambda}{2} \|Q\|^2, \tag{40}$$

i.e. respectively the simplified empirical loss (3) and the regularization, as functions with matrix arguments. The empirical minimization problem (3) can thus be written compactly as

$$\hat{Q} = \underset{Q \in \mathbb{R}^d}{\operatorname{argmin}} \left\{ L(\{\tilde{X}_\ell Q\}_\ell) + r(Q) \right\} \tag{41}$$

with the critical (zero-gradient) condition being given by

$$\sum_{\ell=1}^{L} \tilde{X}_\ell^\top \partial_\ell L(\{\tilde{X}_\ell Q\}_\ell) + \partial g(Q) \overset{!}{=} 0. \tag{42}$$

Let us choose a diagonal definite $A \in \mathbb{R}^{d \times d}$, and a sequence $\{V_\mu\}_{1 \leq \mu \leq n}$ of symmetric definite $L \times L$ matrices. Group them into a block diagonal matrix $\check{V} \in \mathbb{R}^{Ln \times Ln}$, so that the $\mu-$ th block of $\check{V}$ corresponds to V_μ. It shall prove useful to further introduce the matrices $\check{\tilde{X}} \in \mathbb{R}^{d \times nL}$ (resp. $\check{\partial} L(\check{\tilde{X}}) \in \mathbb{R}^{nL}$), defined as the concatenation of the matrices $\tilde{X}_1, ..., \tilde{X}_L$ (resp. $\partial_1 L, ..., \partial_L L$), viewed as n blocks of length L. Then without loss of generality the zero-gradient condition can be rewritten as

$$\check{\tilde{X}}^\top V^{-1} \left(V \check{\partial} L(\check{\tilde{X}}) - \check{\tilde{X}} Q \right) + A(A^{-1} \partial g(Q) + Q) \overset{!}{=} \check{\tilde{X}}^\top V^{-1} \check{\tilde{X}} Q + AQ. \tag{43}$$

Similarly to [13], let us introduce

$$\check{\omega} \equiv V \check{\partial} L(\check{\tilde{X}}) - \check{\tilde{X}} Q. \tag{44}$$

This can be written in terms of a resolvent as

$$\check{\tilde{X}} Q = \operatorname{prox}(\check{\omega}) \tag{45}$$

where

$$\operatorname{prox}(\check{\omega}) \in \mathbb{R}^{nL} = \underset{\check{x} \in \mathbb{R}^{nL}}{\operatorname{argmin}} \left\{ \frac{1}{2} \|\check{x} - \check{\omega}\|_V^2 + L(\check{x}) \right\} \tag{46}$$

which corresponds to (17). Similarly, we denote

$$b \equiv A^{-1} \partial g(Q) + Q \tag{47}$$

So that

$$Q = \mathrm{prox}_g(b) = \underset{x\in\mathbb{R}^d}{\mathrm{argmin}} \left\{ \frac{1}{2}\|x-b\|^2_{A^{-1}} + g(x) \right\} \tag{48}$$

In the particular case of an ℓ_2 regularization $g(\cdot) = \lambda/2\|\cdot\|^2$, note that

$$\mathrm{prox}_g(b) = (\lambda \mathbb{I}_d + A)^{-1} Ab. \tag{49}$$

The zero-gradient condition can now be rewritten as

$$\begin{cases} \check{\tilde{X}}^\top V^{-1} \left(\mathrm{prox}(\check{\omega}) - \check{\omega}\right) = A(b - \mathrm{prox}_g(b)) \\ \check{\tilde{X}} \mathrm{prox}_g(b) = \mathrm{prox}(\check{\omega}) \end{cases} \tag{50}$$

One is now in a position to expand the concatenated variables $\check{\cdot}$ into a sequence of L $n-$dimensional parameters. For $u = \mathrm{prox}(\check{\omega}), \check{\omega}$ let us denote $u_{\mu\ell}$ $(1 \le \mu \le n, 1 \le \ell \le L)$ the $\ell-$th component of the $\mu-$th block. Introduce

$$f_{\mu\ell} \equiv \sum_\kappa (V_\mu^{-1})_{\ell\kappa} (\mathrm{prox}(\check{\omega})_{\mu\kappa} - \check{\omega}_{\mu\kappa}). \tag{51}$$

Denote $f_\ell \equiv (f_{\mu\ell})_{1\le\mu\le n} \in \mathbb{R}^n, \omega_\ell \equiv (\omega_{\mu\ell})_{1\le\mu\le n} \in \mathbb{R}^n$. The system of equations (50) can then be rewritten as (further redefining $b \leftarrow Ab$):

$$\begin{cases} \sum_\ell \tilde{X}_\ell^\top f_\ell = b - A(\lambda \mathbb{I}_d + A)^{-1} b \\ \tilde{X}_\ell (\lambda \mathbb{I}_d + A)^{-1} b = \sum_\kappa V_{\ell\kappa} f_\kappa - \omega_\ell \end{cases} \tag{52}$$

We used the assumption that $g(\cdot)$ is an ℓ_2 regularization. Finally, introducing $\hat{Q} = \mathrm{prox}_g(A^{-1}b) = (\lambda \mathbb{I}_d + A)^{-1} b$, on reaches

$$\begin{cases} \sum_\ell \tilde{X}_\ell^\top f_\ell = b - A\hat{Q} \\ \tilde{X}_\ell \hat{Q} = \sum_\kappa V_{\ell\kappa} f_\kappa - \omega_\ell \end{cases} \tag{53}$$

which corresponds to the fixed-point equations of GAMP (Algorithm 1). This finishes to show the correspondence between the fixed points of GAMP and the critical points of the empirical landscape (3). To summarize, we have shown that equations (8) describe the zero-gradient points of the empirical loss landscape (3), i.e. fixed points of GD.

A.6. Towards a rigorous proof of result 1

While the connection between the GAMP fixed point and the extrema of the loss is sound, and has been at the roots of many rigorous results for convex losses, see e.g. [7, 15, 17, 19, 31], there exist technical difficulties in adapting these rigorous arguments to the present setting, and a fully rigorous proof would warrant sizable work. While we leave this challenging task for future work, we wish to discuss how it can be potentially achieved. The first task would require the proof of point-wise

convergence of GAMP, as indeed, the identification of the GAMP estimates with the one of the extrema of the loss function requires to be at the fixed point of the iteration. This difficulty, discussed in detail in, e.g. [10, 19, 31], can be in principle addressed by computing the convergence criterion from the state evolution equations (see [10] the discussion in Lemma 7 in [19]), a criterion sometimes called the "replicon" in the context of replica theory [35].

Provided the replicon criterion is satisfied, all converging fixed point described by our theory thus correspond rigorously to fixed point of the loss. The last task would be to prove that the minimum of the loss is indeed the fixed point we found with minimum energy. A potential strategy to prove this would be to use the Gordon-Minimax approach of [20]. While it is used in many situations for convex problems (e.g. [54, 55, 58]), only one side would be required for our (non-convex) problem thanks to the GAMP matching bound. We hope that our results would provide inspiration for further research in this direction.

Appendix B. Derivation of the Main Result with the replica method

In the Appendix we provide an alternative derivation of Result 1, which sharply characterizes the global minimum of the empirical loss (3), using the heuristic replica method from statistical physics [40, 41] in its replica-symmetric formulation. First observe that for any test function $\phi(\hat{Q})$ of the minimizer $\hat{Q}$ of (3),

$$\phi(\hat{Q}) = \lim_{\beta\to\infty} \mathbb{E}_{\mathcal{D}} \frac{1}{Z} \int dQ \phi(Q) e^{-\beta \mathcal{R}[Q]}, \tag{54}$$

where we denoted $R[Q]$ the empirical loss (3), and

$$Z \equiv \int dQ e^{-\beta \mathcal{R}[Q]} \tag{55}$$

the normalization factor, also known as the *partition function* in statistical physics. We remind that $\mathcal{D}$ refers to the training set. In order to access key summary statistics and learning metrics associated to $\hat{Q}$, it is therefore reasonable to seek to compute the generating function associated to the measure (54), namely $\mathbb{E} \ln Z$. Such computations can be addressed using the *replica* method from statistical physics [40, 41], building on the identity

$$\ln Z = \lim_{s\to 0} \frac{Z^s - 1}{s}. \tag{56}$$

The backbone of the derivation thus lies in the computation of $\mathbb{E}Z^s$.Below, we detail the derivation for a generic convex regularizer $g : \mathbb{R}^d \to \mathbb{R}_+$ and later specialize to the case of ℓ_2 regularization. The replicated partition function thus reads

$$\mathbb{E}Z^s = \int \prod_{a=1}^{s} dQ_a e^{-\beta \sum_{a=1}^{s} g(Q_a)} \prod_{\mu=1}^{n} \mathbb{E}_x e^{-\beta \sum_{a=1}^{s} \left(\mathrm{Tr}\, \mathtt{S}\left[\frac{1}{\sqrt{d}}(x+p)Q_a\right] \rho_\Sigma \mathtt{S}\left[\frac{1}{\sqrt{d}}(x+p)Q_a\right]^\top - 2\,\mathrm{Tr}\, \mathtt{T}\left[\frac{1}{\sqrt{d}} x_\ell Q_\star\right] \rho_\Sigma \mathtt{S}\left[\frac{1}{\sqrt{d}}(x+p)Q_a\right]^\top \right)}. \tag{57}$$

Introduce the local fields

$$h^a \equiv \frac{xQ_a}{\sqrt{d}} \in \mathbb{R}^{L\times r}, \qquad h^\star \equiv \frac{xQ_\star}{\sqrt{d}} \in \mathbb{R}^{L\times t} \tag{58}$$

and the overlaps

$$m_a \equiv \frac{pQ_a}{\sqrt{d}} \in \mathbb{R}^{L\times r}, \tag{59}$$

with rows m_a^ℓ. These fields have statistics

$$\mathbb{E}_x[h_\ell^a(h_\kappa^b)^\top] = \delta_{\ell\kappa}\frac{Q_a^\top \Sigma_\ell Q_b}{d} \equiv q_{ab}^\ell \tag{60}$$

$$\mathbb{E}_x[h_\ell^\star(h_\kappa^\star)^\top] = \delta_{\ell\kappa}\frac{Q_\star^\top \Sigma_\ell Q_\star}{d} \equiv \rho_\ell \tag{61}$$

$$\mathbb{E}_x[h_\ell^a(h_\kappa^\star)^\top] = \delta_{\ell\kappa}\frac{Q_a^\top \Sigma_\ell Q_\star}{d} \equiv \theta_a^\ell. \tag{62}$$

Thus

$$\begin{aligned}\mathbb{E}Z^s = \int dm d\hat{m} d\theta d\hat{\theta} dq d\hat{q} \underbrace{e^{-d\sum_a\sum_\ell\left[\hat{m}_a^{\ell\top} m_a^\ell + \mathrm{Tr}\left(\theta_a^\ell \hat{\theta}_a^{\ell\top}\right)\right] - d\sum_\ell \sum_{1\le a\le b\le s} \mathrm{Tr}\left(q_{ab}^\ell \hat{q}_{ab}^{\ell\top}\right)}}_{e^{sd\Psi_t}} \\ \underbrace{\int \prod_{a=1}^s dQ_a e^{-\beta\sum_{a=1}^s g(Q_a) + \sum_a\sum_\ell(\sqrt{d}\hat{m}_a^{\ell\top} Q_a^\top p_\ell + \mathrm{Tr}\left[\theta_a^\ell Q_\star^\top \Sigma_\ell Q_a\right]) + \sum_{1\le a\le b\le s}\sum_\ell \mathrm{Tr}\left[q_{ab}^\ell Q_b^\top \Sigma_\ell Q_a\right]}}_{e^{sd\Psi_Q}} \\ \underbrace{\left[\mathbb{E}_{h^\star,\{h_a\}_{a=1}^s} e^{-\beta\sum_{a=1}^s \left(\mathrm{Tr}\,\mathtt{S}[h^a+m^a]\rho_\Sigma \mathtt{S}[h^a+m^a]^\top - 2\,\mathrm{Tr}\,\mathtt{T}[h^\star]\rho_\Sigma \mathtt{S}[h^a+m^a]^\top\right)}\right]^{\alpha d}}_{e^{s\alpha d\Psi_y}},\end{aligned} \tag{63}$$

where we decomposed the replicated free entropy into the trace, entropic and energetic potentials Ψ_t, Ψ_Q, Ψ_y. Note that all exponents are scaling with $d \to \infty$. Therefore the integral in (63) can be computed using a Laplace saddle-point approximation.

B.1. Replica-Symmetric ansatz

We have thus rephrased the analysis of the measure (54) as a optimization problem over the order parameters $\{q_{ab}^\ell, \theta_a^\ell, m_a\}$, and the associated conjugate variables. However, these still represent $2L(s^2+1)+2s$ variables, and $s \to 0$. In order to make progress, we assume that the maximizer is of *replica-symmetric* (RS) form [40, 41]

$$q_{ab}^\ell = (r_\ell - q_\ell)\delta_{ab} + q_\ell \tag{64}$$

$$m_a^\ell = m_\ell \tag{65}$$

$$\theta_a^\ell = \theta_\ell \tag{66}$$

$$\hat{q}_{ab}^\ell = -\left(\hat{r}_\ell/2 + \hat{q}_\ell\right) + \hat{q}_\ell \tag{67}$$

$$\hat{m}_a^\ell = \hat{m}_\ell \tag{68}$$

$$\hat{\theta}_a^\ell = \hat{\theta}_\ell \tag{69}$$

The RS ansatz holds in a number of machine learning settings, notably for convex problems and Bayes-optimal settings, see e.g. [63] for a review. In the present setting, since the empirical loss (3) is non-convex, we emphasize that the RS ansatz constitutes a heuristic technical assumption of our analysis.

B.2. Entropic potential

We now turn to the entropic potential Ψ_w. It is convenient to introduce the variance order parameter

$$\hat{V}_\ell \equiv \hat{r}_\ell + \hat{q}_\ell. \tag{70}$$

The entropic potential can then be expressed as

$$\begin{aligned}
&e^{\beta s d \Psi_Q} \\
&= \int \prod_{a=1}^{s} dQ_a e^{-\beta \sum_a g(Q^a) + \sum_{\ell=1}^{L}\sum_{a=1}^{s}\left(\sqrt{d}\hat{m}_\ell^\top Q_a^\top p_\ell + \mathrm{Tr}\left[\hat{Q}_\star^\top \Sigma_\ell Q_a\right]\right) - \frac{1}{2}\sum_{\ell=1}^{L}\sum_{a=1}^{s}\mathrm{Tr}\left[\hat{V}_\ell Q_a \boldsymbol{\Sigma}_\ell Q_a^\top\right] + \frac{1}{2}\sum_{\ell=1}^{L}\sum_{a,b}\mathrm{Tr}\left[\hat{q}_\ell Q_a \boldsymbol{\Sigma}_\ell Q_b^\top\right]} \\
&= \int \prod_{\ell=1}^{L} D\boldsymbol{\Xi}_\ell \\
&\qquad \left[\int dQ e^{-\beta g(Q) - \frac{1}{2}\mathrm{Tr}\left[\sum_{\ell=1}^{L}\hat{V}_\ell Q \boldsymbol{\Sigma}_\ell Q^\top\right] + \left(\sum_{\ell=1}^{L}\left(\sqrt{d}\hat{m}_\ell p_\ell^\top + \hat{\theta}_\ell Q_\star^\top \Sigma_\ell\right) + \sum_{\ell=1}^{L}\boldsymbol{\Xi}_\ell \odot (\hat{q}_k \otimes \boldsymbol{\Sigma}_k)^{\frac{1}{2}}\right)\odot Q}\right]^s \\
&= \mathbb{E}_\Xi \left[\int dQ e^{-\beta g(Q) - \frac{1}{2}Q\odot\left[\sum_{\ell=1}^{L}\hat{V}_\ell \otimes \boldsymbol{\Sigma}_\ell\right]\odot Q + \left(\sum_{\ell=1}^{L}\left(\sqrt{d}\hat{m}_\ell p_\ell^\top + \hat{\theta}_\ell Q_\star^\top \Sigma_\ell\right) + \sum_{\ell=1}^{L}\boldsymbol{\Xi}_\ell \odot (\hat{q}_\ell \otimes \boldsymbol{\Sigma}_\ell)^{\frac{1}{2}}\right)\odot Q}\right]^s .
\end{aligned} \tag{71}$$

Therefore

$$\beta\Psi_w = \frac{1}{d}\int \mathbb{E}_\Xi \ln \left[\int dQ e^{-\beta g(Q) - \frac{1}{2}Q\odot\left[\sum_{\ell=1}^{L}\hat{V}_\ell \otimes \boldsymbol{\Sigma}_\ell\right]\odot Q + \left(\sum_{\ell=1}^{L}\left(\sqrt{d}\hat{m}_\ell p_\ell^\top + \hat{\theta}_\ell Q_\star^\top \Sigma_\ell\right) + \sum_{\ell=1}^{L}\boldsymbol{\Xi}_\ell \odot (\hat{q}_\ell \otimes \boldsymbol{\Sigma}_\ell)^{\frac{1}{2}}\right)\odot Q}\right]. \tag{72}$$

For a matrix $\boldsymbol{\Xi} \in \mathbb{R}^{r\times d}$ and tensors $\boldsymbol{A}, \boldsymbol{B} \in \mathbb{R}^{r\times d} \otimes \mathbb{R}^{r\times d}$, we denoted $(\boldsymbol{\Xi}\odot\boldsymbol{A})_{kl} = \sum_{ij}\boldsymbol{\Xi}^{ij}\boldsymbol{A}_{ij,kl}$ and $(\boldsymbol{A}\odot\boldsymbol{B})_{ij,kl} = \sum_{rs}\boldsymbol{A}_{ij,rs}\boldsymbol{B}_{rs,kl}$.

B.3. Energetic potential

The computation of the energetic potential Ψ_y is rather standard and follows the same lines as in e.g. [2], yielding

$$\begin{aligned}\beta\Psi_y = &\int_{\mathbb{R}^{L\times t}} dY\, DZ \int_{\mathbb{R}^{L\times r}} D\Xi \prod_{\ell=1}^{L} \delta\left[y_\ell - (\rho_\ell - \theta_\ell^\top q_\ell^{-1}\theta_\ell)^{\frac{1}{2}} z_\ell - \theta_\ell^\top q_\ell^{\frac{1}{2}}\xi_\ell\right] \\ &\times \ln\left[\int_{\mathbb{R}^{L\times r}} dX \prod_{\ell=1}^{L} \frac{e^{-\frac{1}{2}\left(x_\ell - q_\ell^{\frac{1}{2}}\xi_\ell\right)^\top V_\ell^{-1}\left(x_\ell - q_\ell^{\frac{1}{2}}\xi_\ell\right)}}{\det(2\pi V_\ell)} e^{-\beta \operatorname{Tr} \mathtt{S}[x+m]\rho_\sigma \mathtt{S}[x+m]^\top - 2\operatorname{Tr}\mathtt{T}[y]\rho_\Sigma \mathtt{S}[x+m]^\top}\right] \\ = &\underbrace{\int_{\mathbb{R}^{L\times t}} dY \int_{\mathbb{R}^{L\times r}} D\Xi \prod_{\ell=1}^{L} \frac{e^{-\frac{1}{2}\left(y_\ell - \theta_\ell^\top q_\ell^{\frac{1}{2}}\xi_\ell\right)^\top (\rho_\ell - \theta_\ell^\top q_\ell^{-1}\theta_\ell)^{-1}\left(y_\ell - \theta_\ell^\top q_\ell^{\frac{1}{2}}\xi_\ell\right)}}{\det\left[2\pi(\rho_\ell - \theta_\ell^\top q_\ell^{-1}\theta_\ell)\right]}}_{\equiv \mathbb{E}_{Y,\Xi}} \\ &\times \ln\left[\int_{\mathbb{R}^{L\times r}} dX \prod_{\ell=1}^{L} \frac{e^{-\frac{1}{2}\left(x_\ell - q_\ell^{\frac{1}{2}}\xi_\ell\right)^\top V_\ell^{-1}\left(x_\ell - q_\ell^{\frac{1}{2}}\xi_\ell\right)}}{\det(2\pi V_\ell)} e^{-\beta \operatorname{Tr} \mathtt{S}[x+m]\rho_\sigma \mathtt{S}[x+m]^\top - 2\operatorname{Tr}\mathtt{T}[y]\rho_\Sigma \mathtt{S}[x+m]^\top}\right]\end{aligned} \tag{73}$$

B.4. Zero-temperature limit

We now take the limit $\beta \to \infty$. Rescaling

$$\beta\hat{V}_\ell \leftarrow \hat{V}_\ell, \qquad \frac{1}{\beta}V_\ell \leftarrow V_\ell, \qquad \beta\hat{m}_\ell \leftarrow \hat{m}_\ell, \qquad \beta\hat{\theta}_\ell \leftarrow \hat{\theta}_\ell, \qquad \beta^2\hat{q}_\ell \leftarrow \hat{q}_\ell \tag{74}$$

The entropic potential then reduces to

$$\begin{aligned}\Psi_w = &\frac{1}{2d}\mathbb{E}_\Xi \operatorname{Tr}\left[\left(\sum_{\ell=1}^{L} \hat{V}_\ell \otimes \Sigma_\ell\right) \odot \left(\sum_{\ell=1}^{L}\left(\sqrt{d}\hat{m}_\ell p_\ell^\top + \hat{\theta}_\ell Q_\star^\top \Sigma_\ell\right) + \sum_{\ell=1}^{L} \boldsymbol{\Xi}_\ell \odot (\hat{q}_\ell \otimes \boldsymbol{\Sigma}_\ell)^{\frac{1}{2}}\right)^{\otimes 2}\right] \\ &- \frac{1}{d}\mathbb{E}_\Xi \mathcal{M}_g(\Xi)\end{aligned} \tag{75}$$

where we defined the entropic Moreau enveloppe

$$M_g(\Xi) \equiv \inf_Q \left\{\frac{1}{2}\left\|\left(\sum_{\ell=1}^{L}\hat{V}_\ell \otimes \Sigma_\ell\right)^{1/2}\left(Q - \left(\sum_{\ell=1}^{L}\hat{V}_\ell \otimes \Sigma_\ell\right)^{-1}\left(\sum_{\ell=1}^{L}\left(\sqrt{d}\hat{m}_\ell p_\ell^\top + \hat{\theta}_\ell Q_\star^\top \Sigma_\ell\right) + \sum_{\ell=1}^{L}\boldsymbol{\Xi}_\ell \odot (\hat{q}_\ell \otimes \boldsymbol{\Sigma}_\ell)^{\frac{1}{2}}\right)\right)\right\|^2 + g(Q)\right\}. \tag{76}$$

The energetic potential can be similarly recast into a more compact form

$$\Psi_y = -\mathbb{E}_{Y,\Xi}\mathcal{M}(Y,\Xi) \tag{77}$$

where the Moreau envelope is defined as

$$\mathcal{M}(Y,\Xi) = \inf_X \frac{1}{2}\left\{ \sum_{\ell=1}^{L} \mathrm{Tr}\left[V_\ell^{-1}\left(x_\ell - q_\ell^{1/2}\xi_\ell - m_\ell\right)^{\otimes 2}\right] + \mathrm{Tr}\left[\mathtt{S}(X)\rho_\Sigma \mathtt{S}(X)^\top\right] - 2\,\mathrm{Tr}\left[\mathtt{T}(Y)\rho_\Sigma \mathtt{S}(X)^\top\right]\right\}. \tag{78}$$

B.5. Replica free entropy

One finally reaches an expression for the replica free entropy as

$$\begin{aligned}\Phi =& \frac{1}{2}\sum_{\ell=1}^{L}\left(\mathrm{Tr}\,\hat{V}_\ell q_\ell - \mathrm{Tr}\,\hat{q}_\ell V_\ell\right) - \sum_{\ell=1}^{L}\hat{m}_\ell^\top m_\ell - \sum_{\ell=1}^{L}\mathrm{Tr}\,\hat{\theta}_\ell^\top\theta_\ell - \frac{1}{d}\mathbb{E}_\Xi \mathcal{M}_g(\Xi) \\ &+ \frac{1}{2d}\mathbb{E}_\Xi\,\mathrm{Tr}\left[\left(\sum_{\ell=1}^{L}\hat{V}_\ell\otimes\Sigma_\ell\right)\odot\left(\left(\sqrt{d}\hat{m}_\ell p_\ell^\top + \hat{\theta}_\ell Q_\star^\top\Sigma_\ell\right) + \sum_{\ell=1}^{L}\mathbf{\Xi}_\ell\odot(\hat{q}_\ell\otimes\mathbf{\Sigma}_\ell)^{\frac{1}{2}}\right)^{\otimes 2}\right] - \alpha\mathbb{E}_{y,\xi}\mathcal{M}(y,\xi)\end{aligned} \tag{79}$$

B.6. Saddle-point equations : general regularizer

The extremization of the free entropy (79) yields, similarly to [14], the following system of self-consistent equations on the summary statistics:

$$\begin{cases} V_\ell = \frac{1}{d}\mathbb{E}_{\mathbf{\Xi}}\left[\left(\mathrm{prox}_g\odot(\hat{q}_\ell\otimes\mathbf{\Sigma}_\ell)^{-\frac{1}{2}}\odot(\mathbb{I}_r\otimes\mathbf{\Sigma}_\ell)\right)\mathbf{\Xi}_\ell^\top\right] \\ q_\ell = \frac{1}{d}\mathbb{E}_{\mathbf{\Xi}}\left[\mathrm{prox}_g\mathbf{\Sigma}_\ell\mathrm{prox}_g^\top\right] \\ m_\ell = \frac{1}{\sqrt{d}}\mathbb{E}_\Xi\left[\mathrm{prox}_g p_\ell\right] \\ \theta_\ell = \frac{1}{\sqrt{d}}\mathbb{E}_\Xi\left[\mathrm{prox}_g\Sigma_\ell Q_\star\right] \end{cases} \tag{80}$$

$$\begin{cases} \hat{q}_\ell = \alpha\mathbb{E}_{\Xi,Y}V_\ell^{-1}\left(\mathrm{prox}_\ell - q_\ell^{\frac{1}{2}}\xi_\ell - m_\ell\right)^{\otimes 2}V_\ell^{-1} \\ \hat{V}_\ell = q_\ell^{-1}\hat{\theta}_\ell\theta_\ell^\top - \alpha q_\ell^{-\frac{1}{2}}\mathbb{E}_{\Xi,Y}V_\ell^{-1}\left(\mathrm{prox}_\ell - q_\ell^{\frac{1}{2}}\xi_\ell - m_\ell\right)\xi_\ell^\top \\ \hat{m}_\ell = \alpha\mathbb{E}_{\xi,\eta}V_\ell^{-1}\left(\mathrm{prox}_\ell - q_\ell^{\frac{1}{2}}\xi_\ell - m_\ell\right) \\ \hat{\theta}_\ell = \alpha\mathbb{E}_{\xi,\eta}V_\ell^{-1}\left(\mathrm{prox}_\ell - q_\ell^{\frac{1}{2}}\xi_\ell - m_\ell\right)\left(y_\ell - \xi_\ell^\top q_\ell^{-1/2}\theta_\ell\right)^\top\left(\rho_\ell - \theta_\ell^\top q_\ell^{-1}\theta_\ell\right)^{-1} \end{cases}, \tag{81}$$

where the proximals prox_g and prox_ℓ respectively refer to the arginf in Q (resp. x_ℓ) of the envelopes $\mathcal{M}_g$ (76) (resp. $\mathcal{M}$ 78).

B.7. Saddle-point equations : ℓ_2

We now specialize the saddle-point equations (80) to the case of an ℓ_2 regularizer $g(\cdot) = 1/2\|\cdot\|$ the entropic potential admits the simple form

$$\begin{aligned}\Psi_Q =&\frac{1}{2d}\operatorname{Tr}\left[\left(\lambda\mathbb{I}_r\odot\mathbb{I}_d+\sum_{\ell=1}^{L}\hat{V}_\ell\otimes\mathbf{\Sigma}_\ell\right)^{-1}\odot\left(\sum_{\ell=1}^{L}\hat{q}_\ell\otimes\mathbf{\Sigma}_\ell+\left(\sum_{\ell=1}^{L}(\sqrt{d}\hat{m}_\ell p_\ell^\top+\hat{\theta}_\ell Q_\star^\top\Sigma_\ell\right)^{\otimes 2}\right)\right]\\ =&\frac{1}{2d}\sum_{i=1}^{d}\operatorname{Tr}\left[\left(\lambda+\sum_{\ell=1}^{L}\lambda_i^\ell\hat{V}_\ell\right)^{-1}\left(\sum_{\ell=1}^{L}\lambda_i^\ell\hat{q}_k+\left(\sum_{\ell=1}^{L}(\sqrt{d}\hat{m}_\ell p_\ell^\top e_i+\hat{\theta}_\ell Q_\star^\top\Sigma_\ell e_i\right)\left(\sum_{\ell=1}^{L}(\sqrt{d}\hat{m}_\ell p_\ell^\top e_i+\hat{\theta}_\ell Q_\star^\top\Sigma_\ell e_i\right)^\top\right)\right]\\ \overset{d\to\infty}{=}&\frac{1}{2}\int d\nu(\gamma,\tau,\pi)\operatorname{Tr}\left[\left(\lambda+\sum_{\ell=1}^{L}\gamma_\ell\hat{V}_\ell\right)^{-1}\left(\sum_{\ell=1}^{L}\gamma_\ell\hat{q}_\ell+\left(\sum_{\ell=1}^{L}\tau_\ell\hat{m}_\ell+\gamma_\ell\hat{\theta}_\ell\cdot\pi\right)^{\otimes 2}\right)\right].\end{aligned} \tag{82}$$

The replica free energy thus reads

$$\begin{aligned}\Phi =&\frac{1}{2}\sum_{\ell=1}^{L}\left(\operatorname{Tr}\hat{V}_\ell q_\ell-\operatorname{Tr}\hat{q}_\ell V_\ell\right)-\sum_{\ell=1}^{L}\hat{m}_\ell^\top m_\ell-\sum_{\ell=1}^{L}\operatorname{Tr}\hat{\theta}_\ell^\top\theta_\ell-\alpha\mathbb{E}_{y,\xi}\mathcal{M}(y,\xi)\\ &+\frac{1}{2}\int d\nu(\gamma,\tau,\pi)\operatorname{Tr}\left[\left(\lambda+\sum_{\ell=1}^{L}\gamma_\ell\hat{V}_\ell\right)^{-1}\left(\sum_{\ell=1}^{L}\gamma_\ell\hat{q}_\ell+\left(\sum_{\ell=1}^{L}\tau_\ell\hat{m}_\ell+\gamma_\ell\hat{\theta}_\ell\cdot\pi\right)^{\otimes 2}\right)\right],\end{aligned} \tag{83}$$

leading to the saddle point equations

$$\begin{cases}\hat{q}_\ell=\alpha\mathbb{E}_{\Xi,Y}V_\ell^{-1}\left(\operatorname{prox}_\ell-q_\ell^{\frac{1}{2}}\xi_\ell-m_\ell\right)^{\otimes 2}V_\ell^{-1}\\ \hat{V}_\ell=\hat{\theta}_\ell\theta_\ell^\top q_\ell^{-1}-\alpha\mathbb{E}_{\Xi,Y}V_\ell^{-1}\left(\operatorname{prox}_\ell-q_\ell^{\frac{1}{2}}\xi_\ell-m_\ell\right)\xi_\ell^\top q_\ell^{-\frac{1}{2}}\\ \hat{m}_\ell=\alpha\mathbb{E}_{\xi,\eta}V_\ell^{-1}\left(\operatorname{prox}_\ell-q_\ell^{\frac{1}{2}}\xi_\ell-m_\ell\right)\\ \hat{\theta}_\ell=\alpha\mathbb{E}_{\xi,\eta}V_\ell^{-1}\left(\operatorname{prox}_\ell-q_\ell^{\frac{1}{2}}\xi_\ell-m_\ell\right)\left(y_\ell-\xi_\ell^\top q_\ell^{-1/2}\theta_\ell\right)^\top\left(\rho_\ell-\theta_\ell^\top q_\ell^{-1}\theta_\ell\right)^{-1}\end{cases}$$
$$\begin{cases}q_\ell=\int d\nu(\gamma,\tau,\pi)\gamma_\ell\left(\lambda\mathbb{I}_r+\sum_{\kappa=1}^{L}\gamma_\kappa\hat{V}_\kappa\right)^{-1}\left(\sum_{\kappa=1}^{L}\gamma_\kappa\hat{q}_\kappa+\left(\sum_{\kappa=1}^{L}\hat{m}_\kappa\tau_\kappa+\gamma_\kappa\hat{\theta}_\kappa\cdot\pi\right)^{\otimes 2}\right)\left(\lambda\mathbb{I}_r+\sum_{\kappa=1}^{L}\gamma_\kappa\hat{V}_\kappa\right)^{-1}\\ V_\ell=\int d\nu(\gamma,\tau,\pi)\gamma_\ell\left(\lambda\mathbb{I}_r+\sum_{\kappa=1}^{L}\gamma_\kappa\hat{V}_\kappa\right)^{-1}\\ m_\ell=\int d\nu(\gamma,\tau,\pi)\tau_\ell\left(\lambda\mathbb{I}_r+\sum_{\kappa=1}^{L}\gamma_\kappa\hat{V}_\kappa\right)^{-1}\left(\sum_{\kappa=1}^{L}\hat{m}_\kappa\tau_\kappa+\gamma_\kappa\hat{\theta}_\kappa\cdot\pi\right)\\ \theta_\ell=\int d\nu(\gamma,\tau,\pi)\gamma_\ell\left(\lambda\mathbb{I}_r+\sum_{\kappa=1}^{L}\gamma_\kappa\hat{V}_\kappa\right)^{-1}\left(\sum_{\kappa=1}^{L}\hat{m}_\kappa\tau_\kappa+\gamma_\kappa\hat{\theta}_\kappa\cdot\pi\right)\pi^\top.\end{cases} \tag{84}$$

which finishes to recover (8). Let us finally mention that the update equations (8) for the summary statistics (7) do *not* describe the dynamics of gradient descent, but rather that of an Approximate Message Passing algorithm [6], which we elicit in Appendix A for completeness. q.e.d.

B.8. test MSE

The generalization performance is measured by the test error

$$\epsilon_g \equiv \mathbb{E}_{\mathcal{D}}\mathbb{E}_x \left\| \mathtt{T}\left[\frac{1}{\sqrt{d}} x Q_\star\right] x - \mathtt{S}\left[\frac{1}{\sqrt{d}}(x+p)\hat{Q}\right](x+p) \right\|^2 . \tag{85}$$

Expliciting this expression in terms of the correlated Gaussian variables $xQ_\star, xQ$ allows to straightforwardly show that ϵ_g admits the sharp asymptotic characterization in terms of the summary statistics characterized by (84):

$$\epsilon_g = \mathbb{E}_X \operatorname{Tr}\left[\mathtt{S}[X]\rho_\Sigma \mathtt{S}[X]^\top\right] + \mathbb{E}_Y \operatorname{Tr}\left[\mathtt{T}[Y]\rho_\Sigma \mathtt{T}[Y]^\top\right] - 2\mathbb{E}_{X,Y} \operatorname{Tr}\left[\mathtt{S}[X]\rho_\Sigma \mathtt{T}[Y]^\top\right], \tag{86}$$

where the average bears on $X \in \mathbb{R}^{L\times r}, Y \in \mathbb{R}^{L\times t}$ with independent rows with statistics

$$(x_\ell, y_\ell) \sim \mathcal{N}\left[\begin{pmatrix} m \\ 0 \end{pmatrix}, \left(\begin{array}{c|c} q_\ell & \theta_\ell \\ \hline \theta_\ell^\top & \rho_\ell \end{array}\right)\right] \tag{87}$$

B.9. Training loss

We finally turn to the training loss. It is reasonable to expect, from statistical physics, that the training loss should be equal to the free energy $-\Phi$ at zero temperature. We provide below an alternative derivation, for simplicity in the case of ℓ_2 regularization $g = {}^1\!/_2\|\cdot\|^2$. First note that the training loss ϵ_t can be expressed as

$$\epsilon_t = -\lim_{\beta\to\infty} \partial_\beta \underbrace{\frac{1}{d}\ln Z(\beta)}_{\Phi(\beta)} \tag{88}$$

Where $\Phi(\beta)$ is the free entropy at finite temperature. The trace potential Ψ_t bears no explicit dependence on β. On the other hand,

$$\begin{aligned}\beta\Psi_Q = &-\frac{1}{2}\ln\det\left[\beta\lambda \mathbb{I}_r \otimes \mathbb{I}_d + \sum_{\ell=1}^L \hat{V}_\ell \otimes \Sigma_\ell\right] \\ &+ \frac{1}{2d}\operatorname{Tr}\left[\left(\beta\lambda \mathbb{I}_r \odot \mathbb{I}_d + \sum_{\ell=1}^L \hat{V}_\ell \otimes \mathbf{\Sigma}_\ell\right)^{-1} \odot \left(\sum_{\ell=1}^L \hat{q}_\ell \otimes \mathbf{\Sigma}_\ell + \left(\sum_{\ell=1}^L (\sqrt{d}\hat{m}_\ell p_\ell^\top + \hat{\theta}_\ell Q_\star^\top \Sigma_\ell\right)^{\otimes 2}\right)\right]\end{aligned} \tag{89}$$

Thus

$$\begin{aligned}\partial_\beta(\beta\Psi_Q) = &-\frac{\lambda}{2}\operatorname{Tr}\left[\beta\lambda \mathbb{I}_r \otimes \mathbb{I}_d + \sum_{\ell=1}^L \hat{V}_\ell \otimes \Sigma_\ell\right]^{-1} \\ &- \frac{\lambda}{2d}\operatorname{Tr}\left[\left(\beta\lambda \mathbb{I}_r \odot \mathbb{I}_d + \sum_{\ell=1}^L \hat{V}_\ell \otimes \mathbf{\Sigma}_\ell\right)^{-2} \odot \left(\sum_{\ell=1}^L \hat{q}_\ell \otimes \mathbf{\Sigma}_\ell + \left(\sum_{\ell=1}^L (\sqrt{d}\hat{m}_\ell p_\ell^\top + \hat{\theta}_\ell Q_\star^\top \Sigma_\ell\right)^{\otimes 2}\right)\right]\end{aligned} \tag{90}$$

Finally, going through the same rescaling steps to take the $\beta \to \infty$ limit,

$$\lim_{\beta\to\infty} \partial_\beta(\beta\Psi_Q) = -\frac{\lambda}{2d}\,\mathrm{Tr}\left[\left(\lambda\mathbb{I}_r \odot \mathbb{I}_d + \sum_{\ell=1}^{L} \hat{V}_\ell \otimes \mathbf{\Sigma}_\ell\right)^{-2} \odot \left(\sum_{\ell=1}^{L} \hat{q}_\ell \otimes \mathbf{\Sigma}_\ell + \left(\sum_{\ell=1}^{L} (\sqrt{d}\hat{m}_\ell p_\ell^\top + \hat{\theta}_\ell Q_\star^\top \Sigma_\ell\right)^{\otimes 2}\right)\right] \tag{91}$$

By the same token, it is straightforward to see that

$$\lim_{\beta\to\infty} \partial_\beta(\beta\Psi_y) = -\mathbb{E}_{Y,\Xi}\left[\mathcal{M}(Y,\Xi) - \frac{1}{2}\sum_{\ell=1}^{L} \mathrm{Tr}\left[\underbrace{V_\ell^{-1}\left(x_\ell - q_\ell^{1/2}\xi_\ell - m_\ell\right)^{\otimes 2}}_{\hat{q}_\ell V_\ell}\right]\right] \tag{92}$$

We used the self-consistent equations (8) to identify the term in underbrace. Putting everything together,

$$\begin{aligned} -\epsilon_t = \lim_{\beta\to\infty} \partial_\beta \Psi(\beta) = &-\frac{\lambda}{2}\int d\nu(\gamma,\tau)\,\mathrm{Tr}\left[\left(\lambda + \sum_{\ell=1}^{L}\gamma_\ell \hat{V}_\ell\right)^{-1}\left(\sum_{\ell=1}^{L}\gamma_\ell\hat{q}_\ell + \left(\sum_{\ell=1}^{L}\tau_\ell\hat{m}_\ell + \hat{\theta}_\ell\cdot\pi\right)^{\otimes 2}\right)\right] \\ &- \alpha\mathbb{E}_{Y,\Xi}\left[\mathcal{M}(Y,\Xi)\right] + \frac{1}{2}\sum_{\ell=1}^{L}\mathrm{Tr}[\hat{q}_\ell V_\ell]. \end{aligned} \tag{93}$$

This constitutes a sharp asymptotic characterization of the training loss ϵ_t as a function of the summary statistics characterized in Result 1.

For completeness, we finally explicit the connection between ϵ_t and the negative free entropy (i.e. the *free energy* in statistical physics). We go back to massage the expression for the free entropy

$$\begin{aligned}
\Phi &= \frac{1}{2}\sum_{\ell=1}^{L}\Big(\operatorname{Tr}\hat{V}_\ell q_\ell - \operatorname{Tr}\hat{q}_\ell V_\ell\Big) - \sum_{\ell=1}^{L}\hat{m}_\ell^\top m_\ell - \sum_{\ell=1}^{L}\operatorname{Tr}\hat{\theta}_\ell^\top\theta_\ell - \alpha\mathbb{E}_{Y,\Xi}\mathcal{M}(Y,\Xi) \\
&\quad + \frac{1}{2}\int d\nu(\gamma,\tau)\operatorname{Tr}\left[\left(\lambda+\sum_{\ell=1}^{L}\gamma_\ell\hat{V}_\ell\right)^{-1}\left(\sum_{\ell=1}^{L}\gamma_\ell\hat{q}_\ell+\left(\sum_{\ell=1}^{L}\tau_\ell\hat{m}_\ell+\gamma_\ell\hat{\theta}_\ell\cdot\pi\right)^{\otimes 2}\right)\right] \\
&= \frac{1}{2}\sum_{\ell=1}^{L}\operatorname{Tr}\Big[\hat{q}_\ell V_\ell+\hat{V}_\ell q_\ell\Big] - \frac{1}{2}\int d\nu(\gamma,\tau)\operatorname{Tr}\left[\left(\lambda+\sum_{\ell=1}^{L}\gamma_\ell\hat{V}_\ell\right)^{-1}\left(\sum_{\ell=1}^{L}\gamma_\ell\hat{q}_\ell+\left(\sum_{\ell=1}^{L}\tau_\ell\hat{m}_\ell+\gamma_\ell\hat{\theta}_\ell\cdot\pi\right)^{\otimes 2}\right)\right] \\
&\quad - \alpha\mathbb{E}_{Y,\Xi}\mathcal{M}(Y,\Xi) \\
&= \frac{1}{2}\sum_{\ell=1}^{L}\operatorname{Tr}[\hat{q}_\ell V_\ell] + \frac{1}{2}\int d\nu(\gamma,\tau)\operatorname{Tr}\left[\left(\lambda+\sum_{\ell=1}^{L}\gamma_\ell\hat{V}_\ell\right)^{-2}\left(\sum_{\ell=1}^{L}\hat{V}_\ell\gamma_\ell-\lambda-\sum_{\ell=1}^{L}\gamma_\ell\hat{V}_\ell\right)\left(\sum_{\ell=1}^{L}\gamma_\ell\hat{q}_\ell+\left(\sum_{\ell=1}^{L}\tau_\ell\hat{m}_\ell+\gamma_\ell\hat{\theta}_\ell\cdot\pi\right)^{\otimes 2}\right)\right] \\
&\quad - \alpha\mathbb{E}_{Y,\Xi}\mathcal{M}(Y,\Xi) \\
&= -\frac{\lambda}{2}\int d\nu(\gamma,\tau)\operatorname{Tr}\left[\left(\lambda+\sum_{\ell=1}^{L}\gamma_\ell\hat{V}_\ell\right)^{-1}\left(\sum_{\ell=1}^{L}\gamma_\ell\hat{q}_\ell+\left(\sum_{\ell=1}^{L}\tau_\ell\hat{m}_\ell+\hat{\theta}_\ell\cdot\pi\right)^{\otimes 2}\right)\right] \\
&\quad - \alpha\mathbb{E}_{Y,\Xi}\left[\mathcal{M}(Y,\Xi)\right] + \frac{1}{2}\sum_{\ell=1}^{L}\operatorname{Tr}[\hat{q}_\ell V_\ell] \\
&= -\epsilon_t
\end{aligned} \tag{94}$$

In other words, the training loss is equal to the zero-temperature free energy.

Appendix C. Supplementary Experiments

C.1. Empirical scaling of $\alpha = d/n$

In the following we verify that our experiments are consistent with the scaling behaviour predicted from the theory. We jointly increase d and n for a fixed value of α. In Fig. 2 we indeed observe the expected behaviour for an exemplary value of $\alpha = 2$. The same holds for the summary statistics θ and m, which concentrate as d and n jointly grow, shown in Fig. 1 (center) in the main text.

C.2. Alternative hyperparameters

We provide supplementary results for different parameter settings. Fig. 3 on the left shows more slices from the phase diagram that appears in the main Fig. 1. For the experimental section of the main text, we chose a specific A for definiteness. In the following, we present the same results for a different A with a stronger off-diagonal and a higher rank,

$$A = \begin{pmatrix} 0.3 & 0.7 \\ 0.8 & 0.2 \end{pmatrix}. \tag{95}$$

In Fig. 4 we present the analogous simulations to Fig. 1. While the global phenomena match the previous example, the details of the transitions location differ.

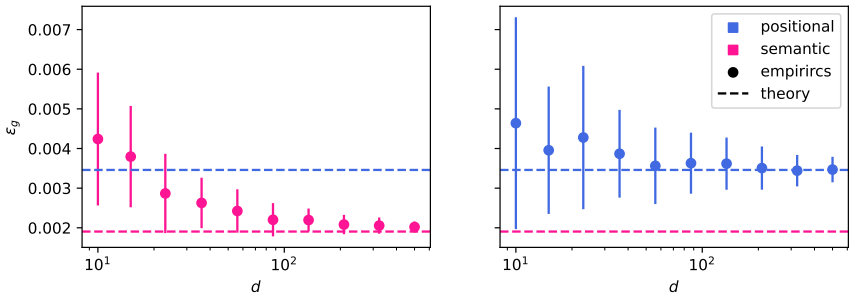

Figure 2: Scaling d and n jointly for $\alpha = 1.5$ approaches the theoretical prediction of the generalization error of the positional and semantic local minima. Experimental settings as in Fig. 1, with 70 runs per datapoint.

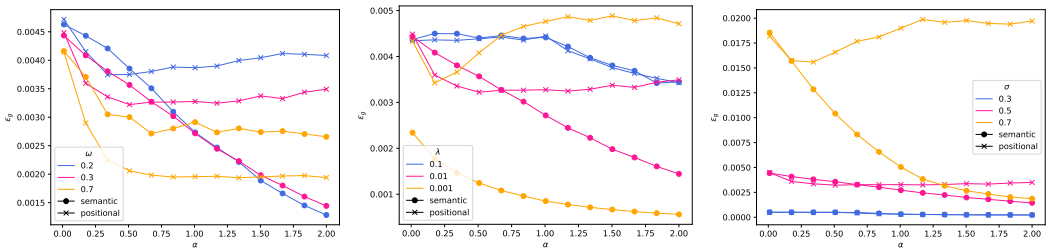

Figure 3: *Alternative Parameters.* Mixed positional/semantic teacher for $\omega = 0.3$. Settings is $r_s = r_t = 1, L = 2, A = ((0.6, 0.4), (0.4, 0.6)), \boldsymbol{\Sigma}_1 = \boldsymbol{\Sigma}_2 = 0.25\mathbb{I}_d$, $\boldsymbol{p}_1 = \mathbf{1}_d = -\boldsymbol{p}_2$ and $\boldsymbol{Q}_\star \sim \mathcal{N}(0, \mathbb{I}_d)$. While keeping all other settings the same, we vary from left to right: The target positionality ω, the student regularizer λ and the standard deviation σ (which is $0.5 = \sqrt{\boldsymbol{\Sigma}_1}$) . Experiment settings as in Fig. 1.

C.3. Uninformed initialization and training via Adam

In or experiments, to obtain the empirical results, we initialize the GD optimizer in an informed fashion, i.e. initializing $\boldsymbol{Q}_\star$ of the student with $r = 1$ as either $\boldsymbol{p}_1$ (positional) or $\boldsymbol{Q}_\star$ (semantics). GD then converges in the two local optima described by our theory.

Since our theory only ascertains that these solutions predicted are indeed fixed points of GD for large sizes, this does not have direct implications for other types of optimization algorithms. In Fig. 5 we show that indeed running the Adam optimizer from an *uninformed* initialization may lead one to either of the local minima for $d = 100$. For larger d we observe the semantic minimum is reached less often than the positional minimum, and a considerable number of times the algorithm simply does not find either of them.

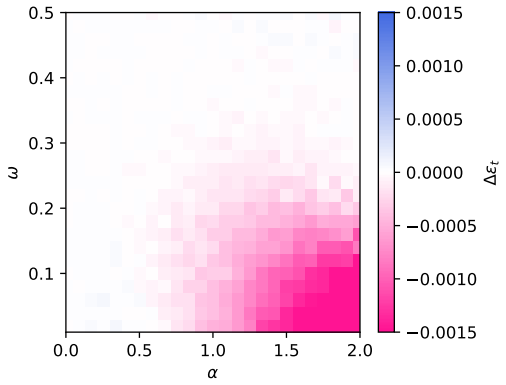

Figure 4: *Alternative Positional Matrix.* $r_s = r_t = 1, L = 2, \mathbf{\Sigma}_1 = \mathbf{\Sigma}_2 = 0.25\mathbb{I}_d$, $\boldsymbol{p}_1 = -\boldsymbol{p}_2$ and $\boldsymbol{p}_1, \boldsymbol{Q}_\star \sim \mathcal{N}(0, \mathbb{I}_d)$ independently. Here, we use a definite matrix A from (95) , which differs form the one used in the main text. Experiments were conducted as in Fig. 1.

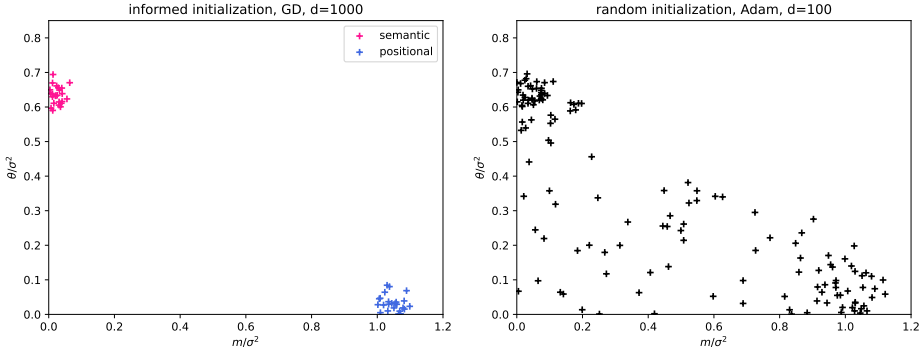

Figure 5: *Comparing GD and Adam.* Settings as in Fig. 1 for the sample complexity $\alpha = 2$. The student parameter $\mathbf{Q}$ is obtained via either **(left)** positional and semantic informed initialization and **(right)** GD training from a random initialization are compared. Each point represents a single run. For the informed GD, we used the same optimization parameters as in Fig. 1 (24 runs per initialization). For Adam we trained on the same data, but for $2,500$ epochs with learning rate $\eta = 0.01$ (showing 140 runs).