SPECIALIZATION OF SOFTMAX ATTENTION HEADS: INSIGHTS FROM THE SINGLE-LOCATION MODEL

Anonymous authors
Paper under double-blind review

ABSTRACT

Multi-head attention enables transformer models to represent multiple attention patterns simultaneously. Empirically, head specialization emerges in distinct stages during training, while many heads remain redundant and learn similar representations. We propose a theoretical model capturing this phenomenon, based on the multi-index and single-location regression frameworks. In the first part, we analyze the training dynamics of multi-head softmax attention under SGD, revealing an initial unspecialized phase followed by a multi-stage specialization phase in which different heads sequentially align with latent signal directions. In the second part, we study the impact of attention activation functions on performance. We introduce the Bayes-softmax attention, which achieves optimal prediction performance in this setting.

1 INTRODUCTION

Multi-head attention is a central architectural ingredient of modern transformer models, enabling multiple attention patterns to coexist within a single layer. Empirical analyses show that attention heads do not all develop simultaneously during training: instead, new specialized heads emerge in distinct stages, often accompanied by sharp changes in loss or behavior suggestive of phase transitions Chen et al. (2024a), with qualitatively new attention behaviors appearing as training progresses Clark et al. (2019); Hoogland et al. (2024); Wang et al. (2025); Chen et al. (2024a); Tigges et al. (2024). At the same time, a substantial fraction of heads in trained models remain redundant and can be removed with little impact on performance Michel et al. (2019); Voita et al. (2019). These observations raise a natural theoretical question: what drives staged head emergence, phase-transition-like behavior, and persistent redundancy in multi-head attention?

Recent theoretical work has begun to elucidate training dynamics and head specialization in simplified attention models. In in-context learning (ICL) of linear regression, linear attention exhibits saddle-to-saddle dynamics in which heads sequentially align with covariance eigenmodes of the data distribution Zhang et al. (2025), while multi-head softmax attention in multi-task linear regression displays short emergence phases during which different heads align with distinct tasks and converge to optimal predictors Chen et al. (2024b). These analyses establish in-context linear regression as a first solvable setting in which head specialization can be studied from first principles. We provide a complementary view on the specialization behavior by considering a model where multi-head attention itself constitutes the sole predictive mechanism and head outputs are uniformly aggregated. By contrasting solvable settings we aspire to identify which aspects of head specialization are shared across models and which depend on architectural or statistical details.

In this work, we study head specialization in a controlled high-dimensional setting by specifying both a probabilistic data model and a minimal attention architecture trained on it. We consider a synthetic sequence-to-token task where the token to be recovered carries a structured signal generated from a multi-index latent model, while all other tokens contain pure noise. We learn this task with a multi-head softmax attention layer trained by stochastic gradient descent (SGD), whose head outputs are uniformly aggregated so that attention itself constitutes the sole predictive mechanism.

Our setting also connects to classical committee-machine and multi-index models: the multi-head architecture plays the role of a committee in which each head acts as a unit with its own parameters, and specialization corresponds to symmetry breaking among these units under SGD Seung et al. (1992); Saad & Solla (1995); Saad et al. (1995); Engel (2001). It is likewise related to recent high-dimensional analyses of SGD in multi-index and sequence single-index models Aubin et al. (2018);

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

Goldt et al. (2019); Ben Arous et al. (2021); Arous et al. (2023); Abbe et al. (2023); Arnaboldi et al. (2025). Our model is also closely related to recent solvable attention and sequence regression frameworks Marion et al. (2025); Troiani et al. (2025); Duranthon et al. (2026); Barnfield et al. (2025); Dohmatob (2025), which analyze high-dimensional training dynamics of simplified attention architectures from first principles.

Within the considered setting, in the high-dimensional limit, the evolution of the attention parameters reduces to a low-dimensional system of order parameters tracking head alignments with the latent signal structure, which captures the full training trajectory. This analysis reveals a two-stage learning dynamics: an initial fast phase where heads develop a common component aligned with the easiest (mean) signal direction, followed by a slower specialization phase in which different heads diverge and align with additional latent directions of the signal, eventually reaching stable specialized configurations. The hierarchy of these specialization events is governed by the latent signal structure, leading to sequential acquisition of increasingly subtle components (see also Saxe et al. (2019) for stage-wise learning in deep linear networks).

We further study how attention activation functions, allowing some heads to be effectively deactivated, influence the model performance. In our setting, redundant or poorly specialized heads can inject persistent variance, making attention normalization a central modeling ingredient. While value or readout transformations can partially mitigate this effect, they act in an input-independent manner and therefore cannot fully suppress noise from redundant heads. By contrast, models that enable head deactivation provide an input-adaptive mechanism for controlling redundancy. Alternative attention activations or extra sink tokens have been studied mainly to prevent over-focusing on irrelevant tokens Kaul et al. (2025); Darcet et al. (2024); Xiao et al. (2024). Here, we analyze such activations as tools for controlling head redundancy and specialization. Finally, Qiu et al. (2025) showed that head-specific gating of attention outputs enhances performance, motivating the deactivation mechanisms without additional parameters.

Our main contributions are as follows.

(i) We introduce a high-dimensional probabilistic framework for training multi-head softmax attention in a sequence-to-token regression task, where attention itself constitutes the sole predictive mechanism, enabling an exact characterization of learning dynamics under SGD.

(ii) We derive a closed system of equations tracking the full evolution of head alignments along the SGD trajectory, and show that training exhibits a fast unspecialized phase followed by a slower hierarchy of specialization events governed by the latent signal structure.

(iii) We analyze the effect of attention normalization on head redundancy, proving that standard softmax is generically suboptimal in this setting. We introduce the Bayes-softmax attention, which, in our setting, reaches the Bayes-risk and prescribes the optimal number of heads and the way to normalize them.

Notations. For a positive integer n we write $[n] = \{1, \dots, n\}$. We use $\delta_{i,j}$ to denote a Kronecker delta that evaluates to 1 if $i = j$ and 0 otherwise. For a positive integer n, $\mathbb{1}_n \in \mathbb{R}^n$ is the vector filled with ones, I_n the $n \times n$ identity matrix, $\mathcal{S}_+^n$ is the set of positive symmetric $n \times n$ matrices. For a matrix A, A_i is its i-th row and $A_{:i}$ is its i-th column. $||\cdot||_2$ is the L_2 norm of a vector, $||\cdot||_F$ is the Frobenius norm of a matrix. We denote a Gaussian law centered at ω with covariance V as $\mathcal{N}(\omega, V)$.

2 Task and data model

We denote a sequence $X \in \mathbb{R}^{L \times D}$ made of L tokens of dimension D. We choose a relevant token X_ϵ by setting a hidden index $\epsilon \in [L]$. This models the fact that, in transformers, the attention is often used to focus and extract one particular token that is needed by the subsequent layers. The relevant token is distinguished from the other tokens by a planted hidden spike $\hat{k} \in \mathbb{R}^D$. We emphasize that the latent ϵ and $\hat{k}$ are different in every sequence/context. Hence, their recovery is akin to a toy in-context learning (ICL) task.

2.1 Probabilistic model of data

Inspired by Marion et al. (2025); Troiani et al. (2025), we introduce a probabilistic data model for this task, that is amenable to analysis in the high-dimensional limit. To define it, we begin by drawing F hidden spikes

$$k_f^* \sim \mathcal{N}\left(0, D^{-1} I_D\right) \quad \text{for } f \in [F]\,. \tag{1}$$

The hidden spikes are common to all the sequences. In the following, we call them directions or features in an interchangeable manner. Each sequence $X \in \mathbb{R}^{L\times D}$ is then sampled as follows. The index ϵ of the relevant token is first taken uniformly at random over $\{1,\dots,L\}$. We take weights $\theta \in \mathbb{R}^F$ for the hidden directions/features according to a distribution P_θ. An effective sequence-dependent direction $\hat{k} \in \mathbb{R}^D$ is then taken as $\hat{k} = \sum_f^F \theta_f k_f^*$. The tokens and the label y are

$$X_\ell \sim \mathcal{N}\left(\delta_{\ell,\epsilon}\hat{k}, I_D\right) \quad \text{for } \ell \in [L]\,, \quad y = X_\epsilon\,. \tag{2}$$

The goal is, given X, to extract the relevant token y. In the following, we will consider different possible choices for P_θ, mainly focusing on:

Flipping spike. For $F = 2$ and signal strengths $\nu_1 > 0, \nu_2 > 0$, $\theta_1 = \sqrt{\nu}_1$ is a constant direction while θ_2 is sampled uniformly from $\{-\sqrt{\nu}_2, +\sqrt{\nu}_2\}$. For $F > 2$ and signal $\nu > 0$, we take $\theta \sim \mathrm{Unif}(\{\sqrt{\nu}e_f\}_{f\in[F]})$ with $\{e_f\}_{f\in[F]}$ the canonical basis of $\mathbb{R}^F$.

Non-isotropic Gaussian. For any F and two signal strengths $\nu_1 \geq \nu_2 > 0$, we take $\theta_f \sim \mathcal{N}(0, \tilde{\nu}_f)$ for all f, with $\tilde{\nu}_1 \geq \tilde{\nu}_2 \geq \ldots \geq \tilde{\nu}_F$ linearly scaled between ν_1 and ν_2. This allows for different signal strengths for the different features. The case $\nu_1 = \nu_2 := \nu$ corresponds to the **isotropic Gaussian**.

MNIST semi-realistic data. We empirically show that our results extend beyond the above data model by considering $\hat{k}$ drawn from the MNIST dataset. Further details are given in Appendix C.

2.2 Learning with attention

We consider the class of estimators $\mathcal{F}_\sigma = \{\hat{y}_{\sigma,k,b,v} : \mathbb{R}^{L\times D} \to \mathbb{R}^D\}_{k\in\mathbb{R}^{H\times D}, b\in\mathbb{R}^H, v\in\mathbb{R}}$ where for H vectors $k_1,\dots,k_H \in \mathbb{R}^D$, H scalar biases $b_1,\dots,b_H$, a scalar v and an activation function $\sigma : (\mathbb{R}^{H\times L}, \mathbb{R}^H, \mathbb{R}, [H]) \to \mathbb{R}^L$, the function $\hat{y}_{\sigma,k,b,v}$ is defined by

$$\hat{y}_{\sigma,k,b,v}(X) = \frac{1}{H}\sum_h^H \sigma(\chi, b, v; h)^\top X\,, \quad \chi_h = Xk_h \in \mathbb{R}^L \quad \text{for } h \in [H]\,. \tag{3}$$

As to the choice of the activation function σ, we focus on the three following cases:

Softmax, the standard choice. For a head h it is defined by $\sigma(\chi, b, v; h)_\ell = e^{\chi_{h\ell}} / \sum_{\ell'=1}^L e^{\chi_{h\ell'}}$. It does not depend on b or v, and the heads only interact by uniformly aggregating their outputs.

Softmax-1, introduced in Kaul et al. (2025), defined by $\sigma(\chi, b, v; h)_\ell = ve^{\chi_{h\ell}} / \left(e^{b_h} + \sum_{\ell'=1}^L e^{\chi_{h\ell'}}\right)$. Notice that the original article takes $e^{b_h} = 1$, which is equivalent to our formulation up to additional head-dependent biases in the attention scores. $\sum_\ell \sigma(\chi, b, v; h)_\ell$ can become smaller than 1, thus "deactivating" some heads. We allow a global rescaling by a factor v to compensate inactive heads.

Bayes-softmax (or B-softmax) which normalizes each head by the output of the H heads. It does not depend on v, but we allow for head-dependent biases. The form is, as motivated later, given by

$$\sigma(\chi, b, v; h)_\ell = \frac{e^{\chi_{h\ell}+b_h}}{\frac{1}{H}\sum_{h'}^H \sum_{\ell'=1}^L e^{\chi_{h'\ell'}+b_{h'}}}\,. \tag{4}$$

Our attention model can be viewed as a simplification of a cross-attention module where a query embedding is independent of the input sequence. Consider a single head attention layer with activation σ, scaling λ (which in practice is usually set to $1/\sqrt{D}$), key and query matrices $K, Q \in \mathbb{R}^{D\times p}$, value matrix is identity $V = I_D$, and let X_{query} be a query embedding independent of the input sequence $\mathrm{Attn}_{\sigma,K,Q,V}(X) = \sigma\left(\lambda(XK)(Q^\top X_{\text{query}})\right)^\top XV = \sigma\left(\lambda\sum_{i=1}^p a_i(XK_{:i})\right)^\top X$, where we denoted $Q^\top X_{\text{query}} = a \in \mathbb{R}^p$ and $K_{:i} \in \mathbb{R}^D$ for $i = 1,\dots,p$ are the columns of the key matrix K. We get attention scores that linearly depend on the input sequence, with the key-vector $k = \lambda\sum_{i=1}^p a_i K_{:i}$. Such attention is used, for example, in the decoder of DEtection TRansformer (Carion et al., 2020), where the input sequence represents encoded image, and the query represents an object to detect. The assumption about the value matrix being identity will be relaxed in section 4 on head deactivation where we will further consider the activation **softmax-v**, a softmax with non-uniform aggregation by learned value weights $\vec{v} \in \mathbb{R}^H$, defined by $\sigma(\chi, b, \vec{v}; h)_\ell = v_h e^{\chi_{h\ell}} / \sum_{\ell'=1}^L e^{\chi_{h\ell'}}$.

The estimator $\hat{y}$ is trained by SGD. The weights k_h are initialized at random $k_h^{(0)} \sim \mathcal{N}(0, \eta^2 D^{-1} I_D)$ with $\eta \in \mathbb{R}^+$ independent of D controlling the initial norm. The biases are initialized to 0 and v to 1. At time t, the weight update is

$$k_h^{(t+1)} = k_h^{(t)} - \gamma\nabla_{k_h^{(t)}}\mathcal{L}^{(t)}, \quad b_h^{(t+1)} = b_h^{(t)} - \gamma\nabla_{b_h^{(t)}}\mathcal{L}^{(t)} \quad \text{for } h \in [H] \tag{5}$$

and $v^{(t+1)} = v^{(t)} - \gamma \nabla_{v^{(t)}} \mathcal{L}^{(t)}$, where $\gamma > 0$ is the learning rate and

$$\mathcal{L}^{(t)} = \frac{1}{N_b} \sum_{\mu=1}^{N_b} \frac{1}{D} ||y^{\mu,(t)} - \hat{y}_{\sigma,k^{(t)},b^{(t)},v^{(t)}}(X^{\mu,(t)})||_2^2 \tag{6}$$

is the empirical loss over a batch of N_b sequences $\{(X^{\mu,(t)}, y^{\mu,(t)})\}_{\mu \in [N_b]}$ drawn iid according to the model eq. (2). We set $N = tN_b$ the total number of samples. Notice that, because of the independence of the batches, on average, the estimator minimizes the population loss

$$\mathcal{E}_\sigma(k, b, v) = \frac{1}{D} \mathbb{E}_{X,y} \left[||y - \hat{y}_{\sigma,k,b,v}(X)||_2^2 \right] . \tag{7}$$

3 Dynamics of the heads

We consider the limit of large embedding dimension $D \to \infty$ and constant sequence length, number of spikes, number of heads, signal strength, and initialization, i.e., $L, F, H, ||\theta||_2, \eta = \Theta(1)$. A first consequence of this limit is that the population loss $\mathcal{E}_\sigma$ eq. (7) can be expressed in terms of a few *order parameters* (or *sufficient statistics*).

Proposition 3.1 (Reparametrized loss). *The loss of the attention $(k, b, v) \mapsto \mathcal{E}_\sigma(k, b, v)$ can be reparametrized as a function $(m, r, b, v) \in (\mathbb{R}^{H \times F}, \mathcal{S}_+^H, \mathbb{R}^H, \mathbb{R}) \mapsto \tilde{\mathcal{E}}_\sigma(m, r, b, v)$ of the following order parameters, for $h, h' \in [H]$, $f, f' \in [F]$:*

$$m_{hf} = (k_h)^\top k_f^*, \quad q_{hh'} = (k_h)^\top k_{h'}, \quad p_{ff'} = (k_f^*)^\top k_{f'}^*, \quad r = q - mp^{-1}m^\top \tag{8}$$

Let $\epsilon \sim \mathrm{Unif}(\{1, \dots, L\})$, $\theta \sim P_\theta$ and conditionally on ϵ and θ, $\chi_{:\ell}^* \sim \mathcal{N}(\delta_{\ell,\epsilon}\theta, I_F)$ and $\xi_h \sim \mathcal{N}(0, I_L)$ for $\ell \in [L], h \in [H]$. For any $s \in \mathbb{R}^{H \times H}$ such that $r = ss^\top$, the reparametrized loss is

$$\tilde{\mathcal{E}}_\sigma(m, r, b, v) = \mathbb{E}_{\epsilon,\theta,\chi^*,\xi} \left[\sum_\ell^L \left(\delta_{\ell,\epsilon} - \frac{1}{H} \sum_h^H \sigma(\chi, b, v; h)_\ell \right)^2 \right], \quad \text{where } \chi = m\chi^* + s\xi \, . \tag{9}$$

A proof is given in App. A.1. Here m_{hf} quantifies the alignment between the head h and the hidden direction f, while q is the overlap between the heads. r is the amplitude of the components that are orthogonal to the spikes; $r = 0$ whenever k is in the span of k^*. In our model, $p \approx I_F$ because the hidden directions are independently sampled from a high-dimensional normal distribution. Last, we parameterize r by s to simplify the formulation of the following results. s is defined up to rotation $s \mapsto sO$ for an orthogonal matrix O that we will fix at initialization.

The motivation to introduce these order parameters is that the learning dynamics of the attention can be expressed in closed-form over m and s. We consider the limit of gradient flow on the population loss, obtained by taking $\gamma N_b^{-1} = o(D^{-1})$ together with $N_b \geq 1$ and $\gamma = \mathcal{O}(1)$, as detailed e.g. in Arnaboldi et al. (2024). In particular, this encompasses the two following cases: i) large batches $N_b = \Theta(D)$ and small learning rates $\gamma = o(1)$; and ii) constant-size batches $N_b = \Theta(1)$ and vanishing learning rates $\gamma = o(D^{-1})$. This scaling allows us to derive the correspondence between the online SGD and the gradient flow, similarly to Saad et al. (1995); Ben Arous et al. (2021).

Proposition 3.2 (Effective dynamics). *Consider the dynamics of SGD defined by eq. (5). Let $\tau = \gamma t$ be the effective time. Fix $s(0) = r(0)^{1/2}$. Then, in the space of the order parameters, the dynamics is given by*

$$\frac{\partial}{\partial \tau} m(\tau) = -\nabla_m \tilde{\mathcal{E}}_\sigma(m(\tau), r(\tau), b(\tau), v(\tau)), \quad \frac{\partial}{\partial \tau} s(\tau) = -\nabla_s \tilde{\mathcal{E}}_\sigma(m(\tau), r(\tau), b(\tau), v(\tau)), \tag{10}$$

and $r(\tau) = s(\tau)s(\tau)^\top$, $\partial_\tau b = -\nabla_b \tilde{\mathcal{E}}_\sigma$ and $\partial_\tau v = -\nabla_v \tilde{\mathcal{E}}_\sigma$.

A proof of this result is given in App. A.2. Similar results, on the equivalence between the dynamics of SGD in the high-dimensional space and the space of order parameters, have been derived for fully-connected two-layer neural networks in e.g. Saad et al. (1995); Ben Arous et al. (2021); Abbe et al. (2023); Arnaboldi et al. (2024) and in Arnaboldi et al. (2025) for sequence single-index models. As a numerical check, we provide simulations in Fig. 1 and in Figs. 9 and 10 in the Appendix. The agreement between the theory and the simulations is very good.

A consequence of Prop. 3.2 is the characterization of two distinct phases in the learning, derived in App. A.3. At random initialization m concentrates around 0. In the first phase, the heads do not specialize and move collectively towards the mean of the signal in a few time steps. The dynamics is controlled by the gradient of the loss in the direction of the mean signal.

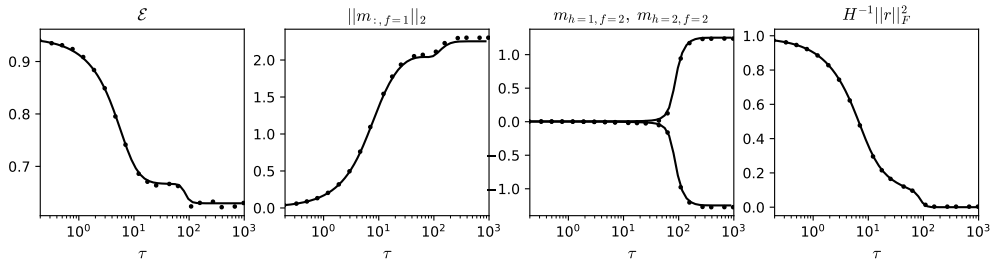


Figure 1: Asymptotic description of the attention trained by SGD. We compare numerical simulations at finite $D = 10^4$ (dots) and the theoretical description stated in proposition 3.2 (continuous lines). Left: loss; center left: alignment along $f = 1$ the constant direction; center right: alignment along $f = 2$ the "flipping sign" direction; right: orthogonal component. We consider sequence length $L = 10$, $H = 2$ heads with σ softmax attention, $F = 2$ features, θ drawn from the flipping spike distribution, with signal strengths $\nu_1 = \nu_2 = 2$. Initialization $\eta = 1$.

Proposition 3.3 (Unspecialized phase)**.** *Let $\mathbb{E}\theta = (\mathbb{E}_{P_\theta}\, \theta_f)_{f\in[F]} \in \mathbb{R}^F$ be the mean of the feature weights and assume it does not vanish $\mathbb{E}\theta \neq 0$. There is a finite time $\tau^{\rm u} = \Theta(1)$ for which there is a $x \in \mathbb{R}^+, x = \Theta(1)$ such that, for all $h \in [H]$, $m_h(\tau^{\rm u}) = x\mathbb{E}\theta + \mathcal{O}(D^{-1/2})$.*

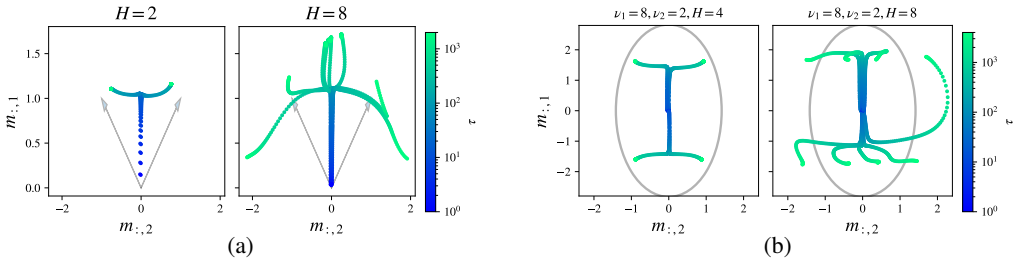
This result comes from the facts that the space of unspecialized m, r and b is invariant by the gradient flow, and that at $m = 0$ the gradient does not vanish and points towards $\mathbb{E}\theta$, as stated in Lemmas A.7 and A.8 in App. A.4. Learning the mean direction $\mathbb{E}\theta$ is fast and in total it requires N larger than $\Theta(D)$ samples to recover this direction. This is depicted in Fig. 1, where we consider the flipping sign distribution, where $\mathbb{E}\,\theta_1 > 0$ and $\mathbb{E}\,\theta_2 = 0$. In the 2nd panel, a time $\tau^{\rm u} \approx 1$ is enough to start learning the direction $f = 1$. In the case when $\mathbb{E}\theta = 0$, m does not move until the next phase. The biases b stay unspecialized: around $\tau^{\rm u}$ there is $\tilde{b} \in \mathbb{R}$ such that $b_h = \tilde{b}$ for all h, with for the B-softmax $\tilde{b} = 0$ and for the softmax-1 $\tilde{b}$ is such that on average the output of the attention scales to 1. Since the space of unspecialized m is invariant under the gradient flow, the specialization of the heads can occur only due to the asymmetry introduced at initialization or by the SGD updates. Therefore, a different rescaling of magnetizations m_{hf} is needed to study the diffusive regime of the dynamics as suggested by Arous et al. (2023). We derive the limiting dynamics in Lemma A.5 and it leads to the following proposition:

Proposition 3.4 (Specialization phase)**.** *Let $P_{\perp\mathbb{E}\theta}$ be the projection onto the space orthogonal to $\mathbb{E}\theta$. Assume that $P_{\perp\mathbb{E}\theta}\,\mathrm{Cov}(\theta)P_{\perp\mathbb{E}\theta}^\top$ is not fully degenerated. Assume that $||\mathbb{E}\theta||_2$ and η are small enough, independently of D. There is a time $\tau^{\rm s} = \Theta(\log D)$ at which, for some $h \in [H]$, $||P_{\perp\mathbb{E}\theta}\, m_h(\tau^{\rm s})||_2 = \Theta(1)$ and the heads start specializing: $||m_h - m_{h'}||_2 = \Theta(1)$ for some $h \neq h'$.*

Learning the directions orthogonal to $\mathbb{E}\theta$ and specializing the heads is slower and in total it requires N larger than $\Theta(D\log D)$ samples to start specializing. As depicted in Fig. 1, 3rd panel, time $\tau^{\rm s} \approx 10$ is needed at $D = 10^4$, during which the alignment with the orthogonal direction and the difference between the heads plateau at 0. $\tau^{\rm s}$ is the time needed to escape the saddle at $P_{\perp\mathbb{E}\theta}m = 0$; its precise value depends on the details of the initialization $P_{\perp\mathbb{E}\theta}m(\tau = 0)$ and varies between runs. In particular, to obtain a match between the theory and the simulation one has to initialize m and r to their empirical values at $\tau = 0$. Prop. 3.4 is derived for small enough $||\mathbb{E}\theta||_2$ and η to ensure that the specialization happens close to the initialization and can be analytically tracked. Such a hypothesis is commonly made, e.g. in Zhang et al. (2025), Pesme & Flammarion (2023) and references therein. Still, Prop. 3.4 holds for larger values, as numerically shown by Figs. 1 to 3 and in Appendix E.1.

These two stages are further illustrated in Fig. 2a for the flipping sign distribution, for different H, where we see that in the first phase of training, all the heads align with the constant direction, and only after some time do they begin to diverge along the "flipping sign" direction. When increasing the number of heads, some heads specialize and diverge more from the average direction. The specialization of the heads versus the two signal strengths and versus H is further numerically analyzed in Appendix E.2.

Phase transition. The (fast) specialization of the heads and the acquisition of new behaviours is often called phase transition, breakthrough or sudden emergence Chen et al. (2024a). Our model allows us to clarify the nature of the dynamical transition around $\tau^{\rm s}$. The transition from unspecialized to specialized heads is truly a dynamical phase transition, in the sense of statistical physics. More



(a) (b)

Figure 2: Evolution of the H heads at $F = 2$, softmax activation, $L = 4$, and $\eta = 1$. Left: for the flipping spike distribution, $\nu_1 = \nu_2 = 2$, and right for the non-isotropic Gaussian distribution, $\nu_1 = 8, \nu_2 = 2$.

precisely, we conjecture that, under the right rescaling when $D \to \infty$, the specialization time τ^{s} concentrates to a deterministic value and the specialization transition around it is sharp. This conjecture relies on a heuristic argument and is supported by numerics detailed in Appendix B and Fig. 6. It moreover extends to semi-realistic data, where D is fixed, by considering the limit of $\eta \to 0$, as we show in App. B Fig. 7, similarly to the sharp transition that appears in Zhang et al. (2025) at small initialization.

Conjecture 3.1 (Dynamical phase transition)**.** *Take $\delta > 0$ small enough, independent of D. Define τ^{s} as the first time τ such that there are some h, h' such that the heads are specialized $||m_h - m_{h'}||_2 > \delta$. Under the rescaling $\tilde{\tau}^{\mathrm{s}} = \tau^{\mathrm{s}}(\log(\sqrt{D}/\eta))^{-1}$, in the limit $D \to \infty$, $\tilde{\tau}^{\mathrm{s}}$ converges to a deterministic value, independent of the realization of the data k^*, X, y, of the initialization $k^{(0)}$ and of D, η, δ.*

Sequential specialization. We give insights on how the specialization of the heads occurs during the specialization phase. The following discussion is based on an analytical result Lemma 3.1 for the initial times of the specialization, and for the later times relies on heuristic arguments on the structure of the landscape and on a numerical integration of Prop. 3.2.

Result 3.1 (Sequential specialization)**.** *During and after τ^{s}, the specialization of the heads occurs in a sequential way, via a saddle-to-saddle dynamics, learning the eigenvectors $(e_f)_{f\in[F]}$ of $\mathrm{Cov}\,\theta$ from the largest eigenvalue to the smallest. Moreover, softmax and softmax-1 learn mixtures of $(e_f)_f$.*

This is a key connection with what Hoogland et al. (2024); Wang et al. (2025) observe in practice, where the attention first learns easy tasks such as bigram statistics and then harder such that n-grams and induction. Such sequential specialization is also described in Zhang et al. (2025) for ICL of linear regression. This sequential learning is shown for the non-isotropic Gaussian distribution, where $\mathbb{E}\theta = 0$ and all the features f have different signal strengths in Figs. 2b and 3, as well as in Appendix Figs. 10 and 13. In the case where two features have the same signal strength, they are learned at the same time, as shown in Fig. 13 top.

Compared to Zhang et al. (2025); Chen et al. (2024b), we observe that each head of the attention does not focus on a single direction $(e_f)_f$. Instead, the softmax and the softmax-1 learn the 2^F possible combinations $\pm e_1 \pm e_2 \ldots \pm e_F$ when H is large enough. This is shown on Fig. 2b at $F = 2$ and $H = 4$ and in the inset of Fig. 3 left at $F = 3$ and $H = 8$, as well as in Appendix Fig. 13. A similar hierarchical learning is described in Saxe et al. (2019) for deep linear fully-connected neural networks; yet notice that, contrary to this work, we do not impose any strong structure on the data P_θ since we only require anisotropy. Rather, the hierarchical learning comes from the structure of the softmax multi-head attention itself. The behaviour of the B-softmax is different: the heads tend to learn single directions $\pm e_f$, as shown in Fig. 3 right; further insights are given by Prop. 4.2 in the next part and an extension to semi-realistic data is shown in Fig. 5.

Result 3.1 comes from the following analysis. The time to escape the saddle and to specialize is controlled by the curvature of the loss in the different directions: the more negatively curbed the faster. We state the following lemma on the Hessian before the specialization time τ^{s}, when the attention is not yet specialized, and after the time τ^{u}, so b and v for the softmax-1 reached their unspecialized minimum. This lemma is a simplification of the Lemma A.11 given in Appendix A.4. We assume that $||\mathbb{E}\theta||_2$ and η are small enough so we can expand the loss around $m \approx 0$ and at $r = 0$.

Lemma 3.1 (Hessian before specialization)**.** *Consider $m \in \mathbb{R}^{H\times F}$ orthogonal to $\mathbb{E}\theta$ i.e. $m\mathbb{E}\theta = 0$. Assume that b is not specialized, and for the softmax-1 assume that b and v reached the fixed-point*

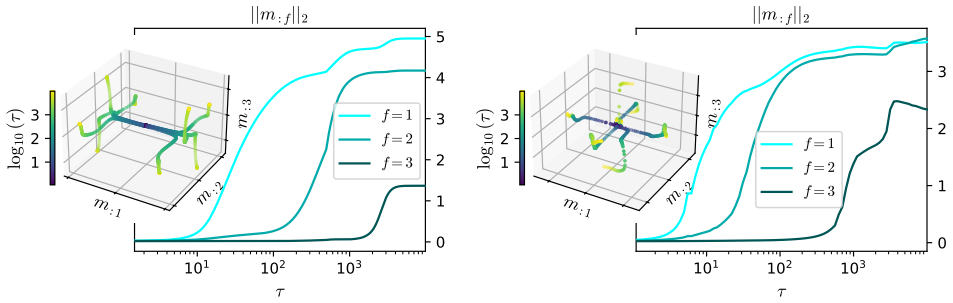


Figure 3: Evolution of the $H = 8$ heads for the non-isotropic Gaussian distribution at $F = 3$, according to Prop. 3.2. Left: σ softmax; right: σ B-softmax. $L = 5$, $\nu_1 = 20$, $\nu_2 = 1$ and $\eta = 1$.

described by Lemma A.10. The loss can be expanded in m as the quadratic form

$$\tilde{\mathcal{E}}_\sigma(m, 0, b, v) = \tilde{\mathcal{E}}_\sigma(0, 0, b, v) \tag{11}$$
$$+ \left(\mathbb{1}_H \mathbb{1}_H^\top \otimes (c_1^{(2)} I_F + (c_2^{(2)} + c_4^{(2)}) \operatorname{Cov} \theta) - (c_3^{(2)} + c_4^{(2)}) I_H \otimes \operatorname{Cov} \theta\right) \cdot (m, m) + \mathcal{O}(||m||_F^4)$$

with the tensorial product $\otimes$ between the spaces $\mathbb{R}^H$ and $\mathbb{R}^F$, and $c_1^{(2)}, c_2^{(2)}, c_3^{(2)} \in \mathbb{R}$ strictly positive and $c_4^{(2)} \in \mathbb{R}$ positive for all $L \geq 3$.

The first descent directions of the magnetization m in the feature space $\mathbb{R}^F$ are thus the eigenvectors of $\operatorname{Cov} \theta$ with largest eigenvalue, as stated by Result 3.1.

We can provide a more precise description of the dynamics. In the head space $\mathbb{R}^H$, according to Lemma 3.1 the first descent directions are all the directions orthogonal to $\mathbb{1}_H$; that is to say the heads $m_{:f}$ tend to split evenly across each feature f. While Lemma 3.1 applies at the beginning of specialization, Lemma A.12 in Appendix A.4 further extends the analysis to later times. The outcome is that, assuming that the features $\bar{f}_1, \ldots, \bar{f}_n$ are already partially learned and that the feature f is not yet learned, $m_{:f}$ tends to grow orthogonally to the magnetizations $m_{:\bar{f}_i}$ of each already-learned feature. This description corresponds to what is shown on Fig. 2b and 3.

Later dynamics, excursions. The later specialization dynamics, and in particular the final split of the heads, depend on the initial condition $k^{(0)}$; it is thus challenging to accurately and entirely describe. As can be seen in Fig. 2b at $H = 8$, because of the stochasticity in initialization, at the beginning of the learning the heads may not split evenly along a direction f, even if $P_\theta(\theta_f)$ is symmetric in this direction. However, later the misplaced heads can rearrange, by performing an *excursion*, so that at the end the split is even and the loss optimized. This shows that the structure of the multi-head softmax helps SGD to navigate the landscape and not to stay stuck in local minima. Moreover, in the next section Fig. 4 we provide a numerical guarantee that the Bayes-softmax reaches the global minimum of the loss whenever H is large enough, which echoes the result of Chen et al. (2024b) on the optimality of gradient flow.

4 Head deactivation via alternative activation

In this section, we consider the trained models and the achieved loss $\mathcal{E}_\sigma^\infty = \lim_{\tau \to \infty} \mathcal{E}_\sigma(k(\tau), b(\tau), v(\tau))$, for different activations σ. We start by deriving the Bayes risk on our probabilistic data model, motivating the Bayes-softmax attention; we compare the expressivity and the performances of the other activations functions and we highlight the impact of the normalization of the heads.

A natural benchmark is the Bayes risk $\mathcal{E}_{\text{Bayes}} = \frac{1}{D}\mathbb{E}_{k^*,X,y}||y - \hat{y}_{\text{Bayes}}(X, k^*)||_2^2$. It corresponds to the loss of the Bayes estimator $\hat{y}_{\text{Bayes}}$, which is the optimal estimator of y in terms of population loss. We characterize it in the following proposition and derive it in App. A.5.

Proposition 4.1 (Bayes estimator)**.** *Given the spikes $\{k_f^*\}_{f \in [F]} \in \mathbb{R}^{F \times D}$ with Gram matrix p and a sequence $X \in \mathbb{R}^{L \times D}$, the Bayes estimator of the label is*

$$\hat{y}_{\text{Bayes}}(X, k^*) = \sum_\ell^L \frac{\int_\theta \exp\left(\hat{k}(\theta)^\top X_\ell\right) e^{-\frac{\theta^\top p \theta}{2}} P_\theta(d\theta)}{\sum_{\ell'}^L \int_\theta \exp\left(\hat{k}(\theta)^\top X_{\ell'}\right) e^{-\frac{\theta^\top p \theta}{2}} P_\theta(d\theta)} X_\ell \,, \quad \text{where } \hat{k}(\theta) = \sum_f^F \theta_f k_f^*. \tag{12}$$

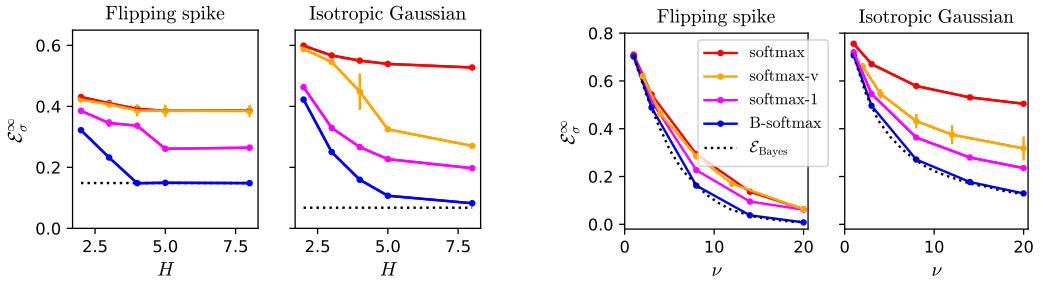


Figure 4: Predicted error $\mathcal{E}_\sigma^\infty$ of the different activation functions. $L = 5$, $\eta = 1$. Left: $F = 4$, $\nu = 10$, varying H; right: $F = 2$, $H = 4$, varying signal strength. We performed 5 independent runs.

In practice, one does not have access to the spikes k^* nor to P_θ, and $\hat{y}_{\text{Bayes}}$ seems to be a purely theoretical estimator. Yet, k^* and P_θ can be learned and $\hat{y}_{\text{Bayes}}$ can be interpreted as the Bayes-softmax attention, which we rewrite as

$$\hat{y}_{\text{B-softmax},k,b}(X) = \sum_\ell^L \frac{\sum_h^H \exp\left(k_h^\top X_\ell + b_h\right)}{\sum_{\ell'}^L \sum_{h'}^H \exp\left(k_{h'}^\top X_{\ell'} + b_{h'}\right)} X_\ell \ . \tag{13}$$

We state the equivalence between the Bayes estimator and the Bayes-softmax attention in the following proposition, which directly follows from a substitution of the given parameters into the B-softmax attention model in high dimension (where for Gaussian spikes $p \to I_F$).

Proposition 4.2 (Optimality of the Bayes-softmax attention)**.** *Consider some distribution P_θ with discrete support $\{\theta^h\}_{h\in[H]}$. Then the Bayes-softmax attention with H heads and parameters $k_h = \hat{k}(\theta^h)$ and $b_h = -||\theta^h||_2^2/2 + \log P_\theta(\theta^h)$ achieves the Bayes risk in the limit $D \to \infty$.*

This proposition gives a prescription on the right number H of attention heads: each point of the support of P_θ should correspond to a different attention head. This can be generalized to continuous distribution P_θ by discretizing the integral in Prop. 4.1 and taking H large enough to approximate it correctly. In Fig. 4 we can see that when the distribution of the spikes is discrete, the loss of the B-softmax model reaches a plateau to the Bayes risk when the number of heads is greater than or equal to the number of spikes. This shows that the B-softmax trained with SGD can exactly estimate the hidden parameters of the Bayes-risk and reach the optimality, whenever H is large enough. For a continuous distribution, the loss does not plateau and continues to decrease, and more heads are required to interpolate P_θ and approach the Bayes-risk. According to Fig. 4 center left, at $F = 4$ for Gaussian P_θ, $H = 5$ is already enough to be close to the optimality; and the B-softmax does not seem cursed by the dimensionality of P_θ. The same behaviour qualitatively holds for the softmax and the softmax-1, though the equivalence with the Bayes risk does not hold.

In our setting, heads that are not aligned with the signal $\hat{k}$ introduce noise that cannot be reduced by other means. This misalignment has to be mitigated by the architecture of the attention itself, by "switching off" the heads with low attention scores. While the standard softmax is not able to do so and therefore cannot reach zero error, the softmax-1 and B-softmax can deactivate heads and outperform the standard variant.

We additionally consider the **softmax-v** attention, designed to emulate the effect of value matrix of the traditional attention architecture on the deactivation of heads. In more realistic settings, relevant and irrelevant tokens may span different subspaces in the space of embeddings and can be separated by value matrices, output projections or MLPs. Yet, the attention still plays a central role, as shown by the works on the softmax-1 and attention sinks Kaul et al. (2025); Darcet et al. (2024); Xiao et al. (2024), where these mechanisms do improve the effectiveness of the standard softmax. Since in our model embeddings of both the relevant and irrelevant tokens span the whole $\mathbb{R}^D$ in an isotropic manner, we only consider the scaling effect introduced by the value matrices. The additional parameters in softmax-v model allow to partially cancel outputs of the not specialized heads, but coefficients v_h cannot adapt to the input sequence. However, as stated by the following proposition, proved in App. A.5, the ability to normalize attention scores adaptively based on the current sequence is crucial for performance in our model.

Proposition 4.3 (Expressivity of softmax, softmax-v and softmax-1)**.** *Assume that there is some disjoint $S \subset \mathbb{R}^F$ and $\bar{S} = \{-\theta, \theta \in S\}$ such that $P_\theta(S)$ and $P_\theta(\bar{S})$ are bounded away from zero by some constant. Then the softmax and softmax-v attentions are not well specified, i.e. $\mathcal{E}_{\text{softmax}}(k, 0, 0)$ and $\mathcal{E}_{\text{softmax-v}}(k, 0, 0)$ are bounded away from zero for all k.*

Consider the multi-head ($H \geq 2$) attention equipped with the softmax-1 activation function. Assume that $\theta = B\tilde{\theta}$, where the distribution of $\tilde{\theta}$ over $\mathbb{R}^F$ is independent of $B \in \mathbb{R}$. Then in the limit of large signal $B \to \infty$ (taken after $D \to \infty$) the softmax-1 attention is well specified, i.e. there is k, b and v such that $\mathcal{E}_{\text{softmax-1}}(k, b, v) \to 0$.

Prop. 4.3 is illustrated by Fig. 4. For the flipping sign direction, the signal $\hat{k}$ is restricted to a quadrant of $\mathbb{R}^F$ and therefore all the heads of the softmax(-v) can be positively aligned with it; the softmax(-v) and the softmax-1 have close performances for all signal strengths. This has to be contrasted with the isotropic Gaussian distribution, where the heads cannot always be aligned with $\hat{k}$: the gap between softmax(-v) and softmax-1 increases with the signal strength ν, and the softmax plateaus at large ν.

At the same time, the B-softmax is very close to the Bayes risk and achieves better performances, as expected. Independent of its link with the Bayes estimator, the superiority of the B-softmax over the softmax-1 can be explained by its ability to perform normalization adapting not only to the scores of each head separately but also to the "most confident" heads for each input sequence. In App. E.4 we provide additional experiment on the pruning of the trained heads in Fig. 16 and the attention maps for different activation functions in Fig. 15. They show that the B-softmax and the softmax-1 focus more on the single relevant token and are more specialized.

Semi-realistic data. The difference between the activations can also bee observed when training on more realistic data. In Fig. 5 (left) we compare the key-vectors of different heads trained on the MNIST detection task (described in App. C). We see that, as indicated by our discussion in section 3, softmax attention learns representations where digits are mixed together in every head, while B-softmax produces clearly distinct digit patterns in different heads. Moreover, as can be seen in Fig. 5 (right), the final errors of different activation functions behave as expected from our previous analysis, with softmax-1 and B-softmax outperforming standard softmax.

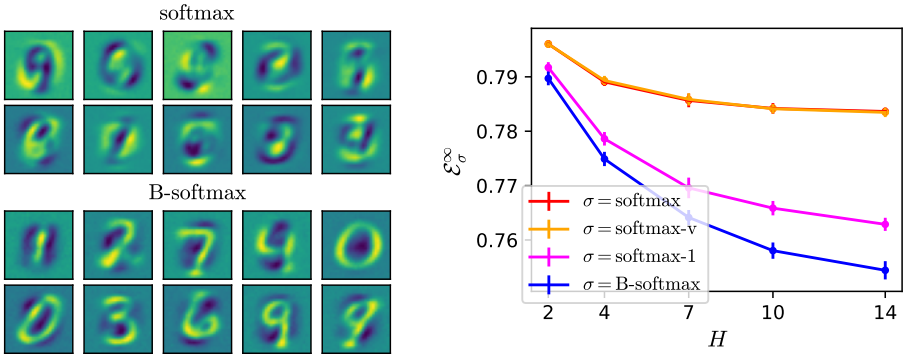


Figure 5: MNIST detection task (described in App. C). Left: learned $(k_h)_{h\in[H]}$ at $H = 10$; right: achieved error $\mathcal{E}_\sigma^\infty$; for the different activation functions after training.

5 Conclusion and Limitations.

We introduced a solvable high-dimensional setup of multi-head attention model and the task where redundant heads induce persistent variance unless explicitly suppressed by the activation. By restricting to this minimal architecture, we obtained an exact description of training dynamics and isolated staged head specialization, head redundancy, and the role of attention normalization under SGD training. Our model is intentionally simplified. It considers a single attention layer, excludes output projections and residual pathways, and focuses on a stylized sequence-to-token regression task. Extending the analysis to deeper architectures and more structured data distributions remains an open direction. As in the theory of the multi-index model Aubin et al. (2018), the population-level training dynamics derived here are expected to translate into sharp sample-complexity transitions under empirical risk minimization, suggesting a route toward understanding data-dependent phase transitions in attention-based models. Finally, comparing the overall phenomenology discovered in our model with recent works on head specialization in in-context learning Zhang et al. (2025); Chen et al. (2024b), we see that stage-wise head emergence appears in both settings. By contrast, the effect of head redundancy is model-dependent: in our setting, redundant heads induce persistent variance unless suppressed by the attention normalization, while in ICL regression, redundant components are asymptotically harmless.

AI USE STATEMENT

We have not used generative AI tools to help develop theoretical models or conceptual frameworks, formulate mathematical claims, provide critical ingredients for proving mathematical claims, assist in the writing of proofs, propose or refine hypotheses, design or provide feedback on research methodology or experiments, implement methods, assist with translation, support qualitative and thematic data analysis, interpret results. Generating synthetic data sets and cleaning and reformating datasets are not applicable to this work.

We used generative AI tools to summarize or analyse existing literature and to search for information, specifically, to find the way of clearly writing the closed form expression (59) by asking an agent to summarize examples in the literature where Kronecker product notation is used for matrix calculus. We have reviewed all AI-assisted work by checking initial literature sources provided by the agent. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

REPRODUCIBILITY STATEMENT

A complete proof of all the claims is included in the appendix with references provided in the main text. For numerical simulations, source code is submitted as supplementary materials. For the experiments setup, details are provided in appendix D.

REFERENCES

Emmanuel Abbe, Enric Boix Adserà, and Theodor Misiakiewicz. Sgd learning on neural networks: leap complexity and saddle-to-saddle dynamics. In *Proceedings of the 36th International Conference on Machine Learning (ICML)*, 2023. arXiv:2302.11055.

Luca Arnaboldi, Yatin Dandi, Florent Krzakala, Bruno Loureiro, Luca Pesce, and Ludovic Stephan. Online learning and information exponents: On the importance of batch size, and time/complexity tradeoffs. *arXiv preprint arXiv:2406.02157*, 2024.

Luca Arnaboldi, Bruno Loureiro, Ludovic Stephan, Florent Krzakala, and Lenka Zdeborová. Asymptotics of SGD in sequence-single index models and single-layer attention networks. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2025. arXiv:2506.02651.

Gerard Ben Arous, Reza Gheissari, and Aukosh Jagannath. High-dimensional limit theorems for sgd: Effective dynamics and critical scaling, 2023. URL https://arxiv.org/abs/2206.04030.

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.

Nicholas Barnfield, Hugo Cui, and Yue M Lu. High-dimensional analysis of single-layer attention for sparse-token classification. arXiv:2509.25153, 2025.

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

Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers. In *Computer Vision – ECCV 2020*, pp. 213–229. Springer International Publishing, 2020. arXiv:2005.12872.

Angelica Chen, Ravid Shwartz-Ziv, Kyunghyun Cho, Matthew L. Leavitt, and Naomi Saphra. Sudden drops in the loss: Syntax acquisition, phase transitions, and simplicity bias in mlms. In *International Conference on Learning Representations (ICLR)*, 2024a. arXiv:2309.07311.

Siyu Chen, Heejune Sheen, Tianhao Wang, and Zhuoran Yang. Training dynamics of multi-head softmax attention for in-context learning: Emergence, convergence, and optimality. In *Conference on Learning Theory (COLT)*, 2024b. arXiv:2402.19442.

Kevin Clark, Urvashi Khandelwal, Omer Levy, and Christopher D Manning. What does bert look at? an analysis of bert's attention. *arXiv preprint arXiv:1906.04341*, 2019.

Timothée Darcet, Maxime Oquab, Julien Mairal, and Piotr Bojanowski. Vision transformers need registers. In *International Conference on Learning Representations (ICLR)*, 2024. arXiv:2309.16588.

Li Deng. The mnist database of handwritten digit images for machine learning research. *IEEE Signal Processing Magazine*, 29(6), 2012.

Elvis Dohmatob. Understanding softmax attention layers: Exact mean-field analysis on a toy problem. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2025.

O. Duranthon, P. Marion, C. Boyer, B. Loureiro, and L. Zdeborová. Statistical advantage of softmax attention: Insights from single-location regression. In *The Fourteenth International Conference on Learning Representations (ICLR)*, 2026. arXiv:2509.21936.

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

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

Jesse Hoogland, George Wang, Matthew Farrugia-Roberts, Liam Carroll, Susan Wei, and Daniel Murfet. Loss landscape degeneracy and stagewise development in transformers. *Transactions on Machine Learning Research (TMLR)*, 2024. arxiv:2402.02364.

Prannay Kaul, Chengcheng Ma, Ismail Elezi, and Jiankang Deng. From attention to activation: Unraveling the enigmas of large language models. In *International Conference on Learning Representations (ICLR)*, volume 2025, pp. 34531–34582, 2025. arXiv:2410.17174.

Pierre Marion, Raphaël Berthier, Gérard Biau, and Claire Boyer. Attention layers provably solve single-location regression. In *The Thirteenth International Conference on Learning Representations (ICLR)*, 2025. arXiv:2410.01537.

Paul Michel, Omer Levy, and Graham Neubig. Are sixteen heads really better than one? In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 32. Curran Associates, Inc., 2019. arXiv:1905.10650.

Scott Pesme and Nicolas Flammarion. Saddle-to-saddle dynamics in diagonal linear networks. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2023. arXiv:2304.00488.

Zihan Qiu, Zekun Wang, Bo Zheng, Zeyu Huang, Kaiyue Wen, Songlin Yang, Rui Men, Le Yu, Fei Huang, Suozhi Huang, Dayiheng Liu, Jingren Zhou, and Junyang Lin. Gated attention for large language models: Non-linearity, sparsity, and attention-sink-free. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems (NeurIPS)*, 2025. arXiv:2505.06708.

David Saad and Sara Solla. Dynamics of on-line gradient descent learning for multilayer neural networks. *Advances in neural information processing systems*, 8, 1995.

David Saad, Sara A Solla, et al. On-line learning in soft committee machines. *Physical review E*, 52 (4):4225–4243, 1995.

Andrew M. Saxe, James L. McClelland, and Surya Ganguli. A mathematical theory of semantic development in deep neural networks. *PNAS*, 116, 2019. arxiv:1810.10531.

Hyunjune Sebastian Seung, Haim Sompolinsky, and Naftali Tishby. Statistical mechanics of learning from examples. *Physical review A*, 45(8):6056, 1992.

Curt Tigges, Michael Hanna, Qinan Yu, and Stella Biderman. Llm circuit analyses are consistent across training and scale. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2024. arXiv:2407.10827.

Emanuele Troiani, Hugo Cui, Yatin Dandi, Florent Krzakala, and Lenka Zdeborová. Fundamental limits of learning in sequence multi-index models and deep attention networks: High-dimensional asymptotics and sharp thresholds. *Forty-second International Conference on Machine Learning (ICML)*, 2025.

Elena Voita, David Talbot, Fedor Moiseev, Rico Sennrich, and Ivan Titov. Analyzing multi-head self-attention: Specialized heads do the heavy lifting, the rest can be pruned. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, 2019. arXiv:1905.09418.

George Wang, Jesse Hoogland, Stan van Wingerden, Zach Furman, and Daniel Murfet. Differentiation and specialization of attention heads via the refined local learning coefficient. In *International Conference on Learning Representations (ICLR)*, 2025. arXiv:2410.02984.

Guangxuan Xiao, Yuandong Tian, Beidi Chen, Song Han, and Mike Lewis. Efficient streaming language models with attention sinks. In *International Conference on Learning Representations (ICLR)*, 2024. arXiv:2309.17453.

Yedi Zhang, Aaditya K. Singh, Peter E. Latham, and Andrew Saxe. Training dynamics of in-context learning in linear attention. In *Proceedings of the 42th International Conference on Machine Learning (ICML)*, 2025. arXiv:2501.16265.

A Proofs of propositions and lemma

In this part we give the different proofs and justifications for our theoretical results of parts 3 and 4. We derive the low-dimensional characterization of the loss in A.1, the low-dimensional characterization of the dynamics in A.2, the two different phases in A.3 and the expressivity and optimality of the activation functions in A.5.

A.1 Low-dimensional characterization of the loss

We start by proving the reparameterization of the loss in terms of the order parameters.

Proof of Proposition 3.1. We first derive the order parameters and the expression for the reparameterized loss. Notice that in the high-dimensional limit when $D \to \infty$, the hidden spikes directions k_f^* are almost orthogonal. Then parameters of the attention model $k_1, \dots, k_H$ can be expressed as follows:

$$k_h = \sum_{f=1}^{F} m_{hf} k_f^* + \sum_{h'=1}^{H} (r^{1/2})_{hh'} k_{h'}^{\perp} \,, \tag{14}$$

where $\{k_{h'}^{\perp}\}_{h' \in [H]}$ are vectors orthogonal to $\{k_f^*\}_{f \in [F]}$ and between each other, and

$$m_{hf} = (k_h)^\top k_f^* \,, \quad h \in [H],\ f \in [F] \qquad\qquad q_{hh'} = (k_h)^\top k_{h'} \,, \quad h, h' \in [H] \tag{15}$$

$$p_{ff'} = (k_f^*)^\top k_{f'}^* \approx \delta_{f,f'} \,, \quad f, f' \in [F] \qquad\qquad r = q - m p^{-1} m^\top \tag{16}$$

Thus, the pre-activations χ_h of the attention head h can be expressed as

$$\chi_h = \sum_{f=1}^{F} m_{hf} X k_f^* + \sum_{h'=1}^{H} (r^{1/2})_{hh'} X k_{h'}^{\perp}, \tag{17}$$

or, if we denote $\chi_f^* = X k_f^*$ and $\xi_{h'} = X k_{h'}^{\perp}$

$$\chi_h = \sum_{f=1}^{F} m_{hf} \chi_f^* + \sum_{h'=1}^{H} (r^{1/2})_{hh'} \xi_{h'}. \tag{18}$$

The risk depends on the parameters only via the projections χ_h, and in the high-dimensional limit the projections χ_f^* and $\xi_{h'}$ are normally distributed, such that all elements of the vectors are independent, $\chi_{f\ell}^* \sim \mathcal{N}(\delta_{\ell,\epsilon^*}\theta_f, 1)$, and $\xi_{h'\ell} \sim \mathcal{N}(0,1)$.

We can now express the loss in terms of the order parameters. We begin by explicitly writing the output of the attention model and expanding the norm squared:

$$\mathcal{E}_\sigma(k,b,v) = \frac{1}{D}\mathbb{E}_{\epsilon,\theta,X}\left[\|X_\epsilon\|_2^2 - \frac{2}{H}\sum_{h=1}^{H}\sigma(\chi,b,v;h)^\top X X_\epsilon \right. \tag{19}$$
$$\left. + \frac{1}{H^2}\sum_{h=1}^{H}\sum_{h'=1}^{H}\sigma(\chi,b,v;h)^\top X X^\top \sigma(\chi,b,v;h')\right].$$

This can be simplified using independence of the tokens X_ℓ and noticing that $\mathbb{E}_X\|X_\ell\|_2^2 \to D$ when $D \to \infty$.

$$\mathcal{E}_\sigma(k,b,v) = \frac{1}{D}\mathbb{E}_{\epsilon,\theta,X}\left[\|X_\epsilon\|_2^2 - \frac{2}{H}\sum_{h=1}^{H}\sigma(\chi,b,v;h)_\epsilon\|X_\epsilon\|_2^2 \right. \tag{20}$$
$$\left. + \frac{1}{H^2}\sum_{h=1}^{H}\sum_{h'=1}^{H}\sum_{\ell=1}^{L}\sigma(\chi,b,v;h)_\ell\sigma(\chi,b,v;h')_\ell\|X_\ell\|_2^2\right]$$
$$= \mathbb{E}_{\epsilon,\theta,X}\sum_{\ell=1}^{L}\left(\delta_{\ell,\epsilon} - \frac{1}{H}\sum_{h=1}^{H}\sigma(\chi,b,v;h)_\ell\right)^2 . \tag{21}$$

Now, we can replace the expectation over X with the expectation over χ^* and ξ. Since $\xi_{:\ell}$ are gaussian vectors with independent elemets, rotating it by an orthogonal matrix O doesn't change the distribution and of $O\xi_{:\ell}$ is also an independent gaussian vector. Therefore, the distribution of $r^{1/2}O\xi_{:\ell}$ is the same as the distribution of $r^{1/2}\xi_{:\ell}$. Thus, the expected error doesn't change if we replace

$$\chi_{:\ell} = \sum_{f=1}^{F}\chi^*_{f\ell}m_{:f} + r^{1/2}\xi_{:\ell}$$

with

$$\chi_{:\ell} = \sum_{f=1}^{F}\chi^*_{f\ell}m_{:f} + r^{1/2}O\xi_{:\ell}$$

for arbitrary orthogonal matrix O. This means, that we can replace $r^{1/2}$ by any matrix s such that $ss^\top = r$.

Using this fact, we get the final expression for $\tilde{\mathcal{E}}$. □

A.2 Low-dimensional characterization of SGD

In this section we prove Proposition 3.2 on how the high-dimensional SGD translates into an effective dynamics for the order parameters.

A.2.1 Informal derivation

In this part we propose an informal derivation, following Saad & Solla (1995); Saad et al. (1995). The learning rate γ and the batch size N_b are such that (cf Arnaboldi et al. (2024)) the weights follow the gradient flow

$$\partial_\tau k = -\nabla_k \mathcal{E}_\sigma(k,b,v)\ . \tag{22}$$

We remind that the pre-activations satisfy $\chi = m\chi^* + s\xi$. We consider the gradient of the error w.r.t. the pre-activations $\chi = k^\top X^\top$, computed as

$$\nabla_k \mathcal{E}_\sigma = (\nabla_\chi \mathcal{E}_\sigma X)^\top\ . \tag{23}$$

We obtain the dynamics over the order parameters as follow. For the magnetization

$$\partial_\tau m = (\partial_\tau k)^\top k^* \tag{24}$$

$$= -\nabla_\chi \mathcal{E}_\sigma X k^* \tag{25}$$
$$= -\nabla_\chi \mathcal{E}_\sigma (\chi^*)^\top \tag{26}$$
$$= -\nabla_m \mathcal{E}_\sigma \tag{27}$$

For the overlap

$$\partial_\tau q = (\partial_\tau k)^\top k + k^\top \partial_\tau k \tag{28}$$
$$= -\nabla_\chi \mathcal{E}_\sigma X k - (\nabla_\chi \mathcal{E}_\sigma X k)^\top \tag{29}$$
$$= -\nabla_\chi \mathcal{E}_\sigma \chi^\top - (\nabla_\chi \mathcal{E}_\sigma \chi^\top)^\top \tag{30}$$

For the orthogonal component

$$\partial_\tau r = \partial_\tau q - m\partial_\tau m^\top - \partial_\tau m m^\top \tag{31}$$
$$= -\nabla_\chi \mathcal{E}_\sigma (\chi - m\chi^*)^\top - (\nabla_\chi \mathcal{E}_\sigma (\chi - m\chi^*)^\top)^\top \tag{32}$$
$$= -\nabla_\chi \mathcal{E}_\sigma (s\xi)^\top - (\nabla_\chi \mathcal{E}_\sigma (s\xi)^\top)^\top \tag{33}$$
$$= -\nabla_s \mathcal{E}_\sigma s^\top - (\nabla_s \mathcal{E}_\sigma s^\top)^\top \tag{34}$$

Since $\partial_\tau r = s\partial_\tau s^\top + \partial_\tau s s^\top$, taking $\partial_\tau s = -\nabla_s \mathcal{E}_\sigma$ satisfies the equation. Notice that we use the fact s is not constrained by any symmetry. Instead, if we had fix $s = r^{1/2}$ for all τ, in general $\partial_\tau s = -\nabla_s \mathcal{E}_\sigma$ would not be possible because $\nabla_s \mathcal{E}_\sigma$ may not be symmetric.

A.2.2 Formal derivation

In this section we formally prove the convergence of the trajectories of the order parameters under SGD to the GF in the limit and derive sample complexity of weak recovery with SGD using Theorem 2.3 from Arous et al. (2023).

To do so, we should prove that the triple: summary statistics (order parameters) $\mathbf{u}^D = (\{m_{hf}\}_{h\in[H],f\in[F]}, \{r_{hh'}\}_{h,h'\in[H]}, \{b_h\}_{h\in[H]}, v)$, loss function (6) and data distribution P_D(as defined in section 2.1) is γ-localizable according to Definition 2.1 in Arous et al. (2023); and we find $\mathbf{h} : \mathbb{R}^C \to \mathbb{R}^C$ and $\mathbf{\Sigma} : \mathbb{R}^C \to \mathbb{R}^{C\times C}$ (where $C = HF + HH + H + 1$) such that the family of summary statistics is asymptotically closable with effective drift $\mathbf{h}$, and effective volatility $\mathbf{\Sigma}$.

Following Arous et al. (2023), we denote

$$H(\{X^\mu, y^\mu\}_{\mu=1}^{N_b}; k, b, v) = \mathcal{L}(\{X^\mu, y^\mu\}_{\mu=1}^{N_b}; k, b, v) - \mathcal{E}(k, b, v).$$

We omit the activation function σ to lighten the notation. We also denote $J_D = \nabla \mathbf{u}^D$, $V(x) = \mathbb{E}[\nabla H(x) \otimes \nabla H(x)]$.

Lemma A.1 (Moments of loss deviation partial derivatives)**.** *Let θ be some parameter $\theta \in \{k_{hd}\}_{h\in[H],d\in[D]} \cup \{b_h\}_{h\in[H]} \cup \{v\}$. Denote*

$$Y_\theta = \frac{1}{D}\partial_\theta \|y - \hat{y}_{\sigma,k,b,v}(X)\|^2,$$

where $\hat{y}$ is given by 3, for some fixed values of k, b, v. This random variables have finite absolute moments:

$$\mathbb{E}[|Y_\theta|^q] = O(1), \quad q \in \mathbb{N}.$$

Moreover, even moments of partial derivatives of H are bounded by:

$$\mathbb{E}[(\partial_\theta H)^{2q}] = O(1/N_b^q).$$

Proof. The expressions for partial derivatives of the loss are

$$\partial_{k_{hd'}} \frac{1}{D}\partial_\theta \|y - \hat{y}_{\sigma,k,b,v}(X)\|^2 = \frac{2}{D}\sum_{d=1}^{D}\sum_{\ell=1}^{L} X_{\ell d}\left(\frac{1}{H}\sum_h \sigma_\ell(X; h) - \delta_{\ell\epsilon}\right) \tag{35}$$
$$\sum_{\ell'=1}^{L} X_{\ell' d}\left(\frac{1}{H}\sum_{h'} \partial_{k_{hd'}}\sigma_{\ell'}(X; h')\right) \tag{36}$$

$$= \sum_{l,l',l''}^{L} c_h(X^\mu; l, l', l'') X_{d'l''} \left(\frac{1}{D} \sum_{d}^{D} X_{dl} X_{dl'} \right), \tag{37}$$

$$\partial_{b_h} \frac{1}{D} \partial_\theta \|y - \hat{y}_{\sigma,k,b,v}(X)\|^2 = \sum_{l,l'}^{L} c_h(X; l, l') X_{d'l'} \frac{1}{D} \left(\sum_{d}^{D} X_{dl} \right), \tag{38}$$

where $c_h(X; l, l'), c_{h,d'}(X; l, l', l'')$ are bounded by some constant c_σ due to the activation σ being bounded for fixed values of v and b_h.

Notice, that random variables Y_θ has the form $\frac{1}{D}\sum_d^D A_d$. We can estimate $\mathbb{E}|\frac{1}{D}\sum_d^D A_d|^q \leq \frac{1}{D^q}\sum_{d_1}^D \dots \sum_{d_q}^D \mathbb{E}\left[\prod_{i=1}^q |A_{d_i}|\right]$. In our case $|A_{d_i}| \leq c_\sigma |X_{ld} X_{l'd} X_{l''d'}|$ or $|A_{d_i}| \leq c_\sigma |X_{ld} X_{l'd'}|$, where X_{ld} are normally distributed. In both cases, $\mathbb{E}\left[\prod_{i=1}^q |A_{d_i}|\right] \leq C_\sigma \mathbb{E}\left[\prod_i |X_{l_i d_i} X_{l'_i d_i} X_{l''_i d'_i}|\right]$ (or $\mathbb{E}\left[\prod_i |X_{l_i d_i} X_{l'_i d'_i}|\right]$) $= O(1)$. Summing all this expectations, we get $\mathbb{E}|Y_\theta|^q = O(1)$.

Notice that $\partial_\theta H$ is an average of N_b independent centered random variables $Y_\theta^\mu - \mathbb{E}Y_\theta^\mu$ for $\mu \in [N_b]$. Then

$$\mathbb{E}[(\partial_\theta H)^{2q}] = \frac{1}{N_b^{2q}} \sum_{\mu_1}^{N_b} \dots \sum_{\mu_{2q}}^{N_b} \mathbb{E}\left[\prod_i^{2q} Y^{\mu_i}\right] \tag{39}$$

$$= \frac{1}{N_b^{2q}} \sum_{\mu_1}^{N_b} \dots \sum_{\mu_k}^{N_b} \mathbb{E}\left[\prod_i^{q} (Y^{\mu_i})^2\right] \tag{40}$$

$$= \frac{1}{N_b^{2q}} N_b^q O(1) = O(1/N_b^q), \tag{41}$$

where the second transition is due to Y^{μ_i} and Y^{μ_j} being independent when $\mu_i \neq \mu_j$, and centered. So, if the indices $\mu_1, \dots, \mu_{2q}$ are not "paired", the term is zero in expectation. □

Lemma A.2 (Summary statistics are γ-localizable)**.** *The family of summary statistics* $\mathbf{u}^D$

$$\mathbf{u}^D = (\{m_{hf}\}_{h\in[H], f\in[F]}, \{r_{hh'}\}_{h,h'\in[H]}, \{b_h\}_{h\in[H]}, v), \tag{42}$$

$$\text{where} \quad m_{hf} = \langle k_h, k_f^* \rangle, \quad r_{hh'} = \langle k_h, k_{h'} \rangle - k_h^\top \left(k^* p^{-1} (k^*)^\top\right) k_{h'} \tag{43}$$

is such that the triple ($\mathbf{u}^D$*, $\mathcal{L}$, P_D) is γ-localizable with sequence of compacts $\bar{B}_K(0)$.*

Proof. **Property (1).** Second derivative of the summary statistics in not zero only for r:

$$\nabla_{k_h} r_{hh'} = P k_{h'}, \qquad \nabla_{k_h} r_{hh} = 2P k_h \tag{44}$$

$$\nabla^2_{k_h,k_h} r_{hh'} = 0, \qquad \nabla^2_{k_h,k_h} r_{hh} = 2P \tag{45}$$

$$\nabla^2_{k_h,k_{h'}} r_{hh'} = P \tag{46}$$

where we denote $P = I_D - k^* p^{-1} (k^*)^\top$. We get that the operator norm is bounded by some constant C_K as soon as $\|k_h\|, \|k_f^*\|$ are bounded, which is true on $\bar{B}_K(0)$.

Property (2). First, notice that

$$\|\nabla \Phi\|^2 \leq \sum_{u \in \mathbf{u}_D} |\partial_u \Phi| \|\nabla u\|^2.$$

Here, $\|\nabla u\|^2$ is bounded by constant on the compact. And $\Phi(\mathbf{u}_D)$ is such that

$$\partial_u \Phi = \mathbb{E}_{\epsilon,\theta,\chi_D,\xi_D} \sum_\ell \phi_\ell(\epsilon, \theta, \chi_D, \xi_D; \mathbf{u}_D)(\delta_{u=m_{hf}}(\chi_D)_\ell + \delta_{u=r_{hf}}(\xi_D)_\ell + \delta_{u=b_h,v}),$$

where $|\phi_\ell(\epsilon, \theta, \chi_D, \xi_D; \mathbf{u}_D)| = O(1)$ is bounded and random variables χ_D, ξ_D weakly converge to normal distribution $\chi_D, \xi_D \to \mathcal{N}(0, 1)$, therefore $|\partial_u \Phi(\mathbf{u}_D)| = O(1)$.

Now, to bound $\mathbb{E}[\|\nabla H\|^8]$, by applying Cauchy–Schwartz inequality twice, we get

$$\|\nabla H\|^8 = \left(\sum_\theta (\partial_\theta H)^2\right)^4 \le p_D^3 \sum_{\theta=1}^{p_D} (\partial_\theta L_D)^8 \tag{47}$$

and using A.1, we have

$$\mathbb{E}[(\partial_\theta H)^8] = O\left(\frac{1}{N_b^4}\right), \tag{48}$$

Therefore we get $\mathbb{E}[\gamma^4 \|\nabla L_D\|^8] = O(1)$ and the required bound is satisfied.

Property (3). Notice that $\langle \nabla H, \nabla(b_h/v)\rangle^4 = (\partial_{b_h/v} H)^4 = O(1)$, while

$$\langle \nabla H, \nabla(m_{hf}/r_{hh'})\rangle^4 = \left(\sum_d (k^*_{fd}/(Pk_{h'})_d)\partial_{k_{hd}} H\right)^4 \tag{49}$$

$$\le ((\|k^*\|^2/\|Pk_{h'}\|^2)\sum_d (\partial_{k_{hd}} H)^2)^2 \le C_K D \sum_d (\partial_{k_{hd}} H)^4, \tag{50}$$

and applying lemma A.1, we get $\gamma^2 \mathbb{E}\langle \nabla H, \nabla(b_h/v)\rangle^4 \le \gamma^2 O(D^2 N_b^{-2}) = O(1)$.

Second derivative is only non-zero for the statistics $r_{hh'}$ and we get

$$|\left\langle \nabla^2 r_{hh'}, \nabla H \otimes \nabla H\right\rangle| = |\mathrm{Tr}(\nabla^2 r_{hh'}(\nabla H \otimes \nabla H))| = |(\nabla H)^\top \nabla^2 r_{hh'} \nabla H)| \tag{51}$$

$$= 2(\nabla_{k_h} H)^\top P \nabla_{k_{h'}} H \le 2\|\nabla H\|^2, \tag{52}$$

This gives us that both $\mathbb{E}\left\langle \nabla^2 u, \nabla H \otimes \nabla H\right\rangle = O(DN_b^{-1})$ and $\mathbb{E}\left\langle \nabla^2 u, \nabla H \otimes \nabla H\right\rangle^2 = O(D^2 N_b^{-2})$ (using that $(a+b)^2 \le 2(a^2+b^2)$). Thus, we get $\gamma^3 \mathbb{E}\left\langle \nabla^2 u, \nabla H \otimes \nabla H - V\right\rangle^2 = \gamma O(1) = o(1)$.

□

Lemma A.3 (High-dimensional limit of SGD)**.** *The family of sufficient statistics* (42) *converges (in the sense defined in Theorem 2.3 from Arous et al. (2023)) to the gradient flow* (10) *initialized at* $m = 0, r = \eta^2 I_H, b_h = 0, v = 1$ *for* $h \in [H]$ *when* $D \to \infty$.

Proof. Notice that V consist of the entries of order $O(N_b^{-1})$ and $\partial_i \partial_j u(x)$ is non zero only for statistics $r_{hh'}$ but still bounded on the compact E_K, therefore $\gamma \mathcal{L}_D \mathbf{u} = \gamma O(D/N_b) = o(1)$. Notice that either $\partial_\theta u(x)$ is of order $O(D^{-1/2})$ for $O(D)$ parameters θ or $\partial_\theta u(x)$ is of order $O(1)$ for exactly 1 parameter θ and for all other parameters θ $\partial_\theta u(x) = 0$. Thus, $\sum_{i,j}^D V_{ij} J_{ki} J_{lj} = O(D/N_b)$ and $\gamma J V J^\top = o(1)$.

Applying Theorem 2.3 from Arous et al. (2023), in the limit we have

$$\dot{\mathbf{u}} = \mathbf{h}(\mathbf{u}),$$

where $\mathbf{h}$ is such that:

$$\sup_{x \in \mathbf{u}_D^{-1}(E_K)} \|\mathbf{h}(\mathbf{u}_D(x)) - \mathcal{A}_D \mathbf{u}_D(x)\| \to 0.$$

Notice that

$$\mathcal{A}_D u_l = \sum_i \partial_i \Phi \partial_i u_l = \sum_j \partial_{u_j} \Phi \sum_i \partial_i u_j \partial_i u_l, \tag{53}$$

or in matrix form

$$\mathcal{A}_D \mathbf{u}(x) = (\nabla \mathbf{u})^\top (\nabla \mathbf{u}) \nabla_{\mathbf{u}} \Phi, \tag{54}$$

where $\nabla \mathbf{u} = (\nabla u_1, \ldots, \nabla u_C)$ is a matrix with columns equal to gradients of u w.r.t. the weights.

Notice that $\{\nabla m_{hf}\}_{h\in[H], f\in[F]}$, $\{\nabla r_{hh'}\}_{h,h'\in[H]}$, $\{\nabla b_h\}_{h\in[H]}$, and ∇v are orthogonal and therefore we can consider separately ∇m, ∇r, ∇b and ∇v.

This corresponds to the following system of ODE's:

$$\text{vec}(\dot{m}) = -(\nabla m^\top \nabla m)_{D\to\infty} \nabla_{\text{vec}(m)} \Phi(\mathbf{u}) \tag{55}$$
$$\text{vec}(\dot{r}) = -((\nabla r)^\top \nabla r)_{D\to\infty} \nabla_{\text{vec}(r)} \Phi(\mathbf{u}) \tag{56}$$
$$\dot{b}_h = -\partial_{b_h} \Phi(\mathbf{u}) \tag{57}$$
$$\dot{v} = -\partial_v \Phi(\mathbf{u}). \tag{58}$$

Since $(\nabla m^\top \nabla m)_{D\to\infty} = ((k^* \otimes I_H)^\top (k^* \otimes I_H))_{D\to\infty} = I_{FH}$, the first equation corresponds directly to $\dot{m} = -\nabla_m \Phi$.

At the same time we can write $\nabla r = (I_H \otimes Pk + (I_H \otimes Pk) K_{H,H})$ and $(\nabla r)^\top \nabla r = (I_H \otimes r + r \otimes I_H)(I + K_{H,H})$, where $K_{H,H}$ is the commutation matrix. Combining this results, we get:

$$\text{vec}(\dot{r}) = -(I_H \otimes r + r \otimes I_H)(\nabla_{\text{vec}(r)} \Phi(\mathbf{u}) + K_{H,H} \nabla_{\text{vec}(r)} \Phi(\mathbf{u})). \tag{59}$$

And the final system of ODE's:

$$\dot{m} = -\nabla_m \Phi(\mathbf{u}) \tag{60}$$
$$\dot{r} = -r(\nabla_r \Phi(\mathbf{u}) + (\nabla_r \Phi(\mathbf{u}))^\top) - (\nabla_r \Phi(\mathbf{u}) + (\nabla_r \Phi(\mathbf{u}))^\top) r \tag{61}$$
$$\dot{b} = -\nabla_b \Phi(\mathbf{u}) \tag{62}$$
$$\dot{v} = -\partial_v \Phi(\mathbf{u}). \tag{63}$$

□

Lemma A.4. *Let r be a symmetric matrix such that the evolution of r is given by*

$$\dot{r} = -r(\nabla_r \Phi(r) + (\nabla_r \Phi(r))^\top) - (\nabla_r \Phi(r) + (\nabla_r \Phi(r))^\top) r. \tag{64}$$

Take $s \in \mathbb{R}^{H\times H}$ such that $r = ss^\top$. Then

$$\dot{s} = -\nabla_s \Phi(r), \tag{65}$$

Proof. Let's express the gradient $\nabla_s \Phi(ss^\top)$ using the definition of the differential $d\Phi$:

$$d\Phi = \text{tr}((\nabla_r \Phi)^\top d(r)) = \text{tr}((\nabla_r \Phi)^\top s ds^\top) + \text{tr}((\nabla_r \Phi)^\top ds s^\top) \tag{66}$$
$$= \text{tr}((s^\top (\nabla_r \Phi) + s^\top (\nabla_r \Phi)^\top) ds) = \text{tr}((\nabla_s \Phi)^\top ds). \tag{67}$$

Therefore, we get $\nabla_s \Phi = ((\nabla_r \Phi) + (\nabla_r \Phi)^\top) s$.

Now, we can rewrite the flow as

$$\dot{ss^\top} + \dot{s}s^\top = -s(s^\top \nabla_r \Phi(r) + s^\top (\nabla_r \Phi(r))^\top) - (\nabla_r \Phi(r) s + (\nabla_r \Phi(r))^\top s) s^\top \tag{68}$$
$$= -s(\nabla_s \Phi(r))^\top - \nabla_s \Phi(r) s^\top . \tag{69}$$

From this we obtain that

$$\dot{s} = -\nabla_s \Phi(r) \tag{70}$$

is a valid dynamics that allows us to recover exactly the flow (64) on r. □

A.3 Unspecialized phase, specialization and sample complexity of SGD

In this section we prove Propositions 3.3 and 3.4 on the two phases of the learning. We rely on the technical lemmas part A.4 for the derivatives of the loss.

A.3.1 UNSPECIALIZED PHASE PROP. 3.3

Assuming $\mathbb{E}\theta \neq 0$, according to Lemma A.7 and A.8, the space $\mathcal{M}_{\mathrm{u}}$ of unspecialized m, r and b is invariant by the gradient flow; and moreover the subspace of m proportional to $\mathbb{1}_H(\mathbb{E}\theta)^\top$ is also invariant by the gradient flow. At initialization $m \approx 0$, $r \approx \eta^2 I_H$, $b = 0$ are unspecialized; and the gradient of the loss w.r.t. m does not vanish and points towards $\mathbb{1}_H(\mathbb{E}\theta)^\top$. Consequently only a few time steps τ^{u} are necessary to move in this direction, i.e. a time $\tau^{\mathrm{u}} = \Theta(1)$ is enough so that $m = x\mathbb{1}_H(\mathbb{E}\theta)^\top$ with $x = \Theta(1) \in \mathbb{R}^+$. The total number of required samples is $N = tN_b = \tau^{\mathrm{u}}\gamma^{-1}N_b = \omega(D)$.

As to b and v, for the softmax-1, according to Lemma A.10, the gradient at initialization does not vanish and points towards unspecialized fixed-points that are reached in time $\tau = \Theta(1)$. For the B-softmax the gradient w.r.t. b at initialization is null, b does not move and stays unspecialized.

A.3.2 SPECIALIZATION PHASE PROP. 3.4

To describe the specialization dynamics we rely on Lemma A.11 on the Hessian of the loss at small m and r and unspecialized b. This lemma applies even after initialization, as we justify by the following assumptions.

We assume that $||\mathbb{E}\theta||_2$ is small enough so during the unspecialized phase the growth of m remains bounded and we can consider the loss at $m \approx 0$. We moreover consider a small enough initialization η so $r \propto \eta^2 \approx 0$. By Lemma A.9 on the gradient, $r = 0$ is a fixed-point of the dynamics, and by Lemma A.11 on the Hessian it is a stable or flat fixed-point for m small enough. Consequently $r \approx 0$ holds even after initialization, until m starts growing during the specialization phase. Numerically, we observe that $r = 0$ is a stable fixed-point whose basin of attraction encompasses values of m and r of order one. Last, b stays unspecialized until m grows and specializes, because it corresponds to a stable or flat fixed-point by Lemma A.11.

Applying Lemma A.11, it gives that there exists directions orthogonal to $\mathbb{E}\theta$ in the space of the features, that are descent directions and where the Hessian is not degenerated. More precisely, assuming the heads split evenly and that $m^\top \mathbb{1}_H \approx 0$, the descent directions are the eigenvectors of $\operatorname{Cov}\theta$ projected in the orthogonal space, and the Hessian has strictly negative eigenvalues in all these directions. The heads specialize because they tend to evolve orthogonally to $\mathbb{1}_H$. In the space where $m^\top \mathbb{1}_H \neq 0$ the Hessian has strictly larger eigenvalues and thus it requires more time to move in this direction.

The time necessary to escape the initial saddle is $\tau^{\mathrm{s}} = O(\log D)$ and the total sample complexity is $N = \tau^{\mathrm{s}}\gamma^{-1}N_b = O(D^{1+2\zeta}\log D)$ for arbitrary small $\zeta > 0$, as proven in the following Lemma A.6.

Escaping the invariant subspace. At initialization $m = 0$, and according to lemma A.8, if $\mathbb{E}\theta = 0$, then gradient flow remains stuck at $m = 0$. Therefore, to obtain meaningful analysis of specialization sample complexity, we should, as suggested by Arous et al. (2023), rescale the sufficient statistics to obtain the SDE limiting trajectory of SGD.

Lemma A.5 (SDE limit with rescaled magnetizations)**.** *The family of sufficient statistics with rescaled $m_{hf} \to \tilde{m}_{hf} = D^\zeta m_{hf}$ of the dynamics of SDE with step size $\gamma N_b^{-1} = c_{lr} D^{-1-2\zeta}$ for some constant c_{lr} and arbitrarily small ζ converges to the following system of equations:*

$$d\tilde{m}_{hf} = -\partial_{m_{hf}}\Phi(\mathbf{u})dt + c_\Sigma dB_t \tag{71}$$

$$\dot{r} = -r(\nabla_r\Phi(\mathbf{u}) + (\nabla_r\Phi(\mathbf{u}))^\top) - (\nabla_r\Phi(\mathbf{u}) + (\nabla_r\Phi(\mathbf{u}))^\top)r \tag{72}$$

$$\dot{b} = -\nabla_h\Phi(\mathbf{u}) \tag{73}$$

$$\dot{v} = -\partial_v\Phi(\mathbf{u}), \tag{74}$$

where $c_\Sigma > 0$ is some constant and B_t is a standard Brownian motion in $\mathbb{R}^{HF}$.

Proof. Let $\gamma N_b^{-1} = c_{\mathrm{lr}} D^{-1-2\zeta}$ for some constant c_{lr}, we rescale $m_{hf} \to \tilde{m}_{hf} = D^\zeta m_{hf}$, so that the sufficient statistics are still γ-localizable.

After rescaling $m_{hf} \to \tilde{u} = D^\zeta m_{hf}$, the only property that might break is Property (3), which we avoid by setting learning rate $\gamma N_b = c_{\mathrm{lr}} D^{-1-2\zeta}$ so that it cancels the rescaling coefficient.

Due to the rescaling, for the block of parameters $\{m_{hf}\}_{h\in[H],f\in[F]}$ we get $\gamma JVJ^\top = O(1)$

Using Taylor expansion of V:

$$V_{k_{hd},k_{h'd'}} = \mathbb{E}[\partial_{k_{hd}}H(\mathbf{u})\partial_{k_{h'd'}}H(\mathbf{u})] = \mathbb{E}[\partial_{k_{hd}}H(D^{-\zeta}\tilde{m},r,b,v)\partial_{k_{h'd'}}H(D^{-\zeta}\tilde{m},r,b,v)] \quad (75)$$

$$= \mathbb{E}\left[\left(\partial_{k_{hd}}H(0) + D^{-\zeta}\left\langle\tilde{m},\nabla_{\tilde{m}}\partial_{k_{hd}}H(m_1)\right\rangle\right)\left(\partial_{k_{h'd'}}H(0) + D^{-\zeta}\left\langle\tilde{m},\nabla_{\tilde{m}}\partial_{k_{h'd'}}H(m_2)\right\rangle\right)\right] \quad (76)$$

$$= \mathbb{E}[\partial_{k_{hd}}H(0)\partial_{k_{h'd'}}H(0)] \quad (77)$$

$$+ D^{-2\zeta}\mathbb{E}\partial_{k_{h'd'}}H(0)\left\langle\tilde{m},\nabla_m\partial_{k_{hd}}H(m_1)\right\rangle + D^{-2\zeta}\mathbb{E}\partial_{k_{hd}}H(0)\left\langle\tilde{m},\nabla_m\partial_{k_{h'd'}}H(m_2)\right\rangle \quad (78)$$

$$+ D^{-4\zeta}\mathbb{E}\left\langle\tilde{m},\nabla_m\partial_{k_{hd}}H(m_1)\right\rangle\left\langle\tilde{m},\nabla_m\partial_{k_{h'd'}}H(m_2)\right\rangle \quad (79)$$

It's easy to see that $\mathbb{E}\partial_{k_{hd}}H(0)\left\langle\tilde{m},\nabla_m\partial_{k_{h'd'}}H(m_2)\right\rangle$ and $\mathbb{E}\left\langle\tilde{m},\nabla_m\partial_{k_{hd}}H(m_1)\right\rangle\left\langle\tilde{m},\nabla_m\partial_{k_{h'd'}}H(m_2)\right\rangle$ are of order $O(N_b^{-1})$ on the compact E_K. Indeed, $\|\tilde{m}\|$ is bounded by constant on E_K and $\nabla_m\partial_{k_{hd}}H(m_1)$ is a random variable of the similar form as $\partial_{k_{hd}}H$, consisting of the sum of products of bounded functions of random variables and normal random variables X_{ld}.

We have $V_{k_{hd},k_{h'd'}} = \mathbb{E}[\partial_{k_{hd}}H(0)\partial_{k_{h'd'}}H(0)] + O(D^{-2\zeta}N_b^{-1})$

Now, explicitly writing $\gamma JVJ^\top$, we get:

$$\gamma(JVJ^\top)_{m_{hf},m_{h'f'}} = \gamma D^{2\zeta}\sum_{k_{hd},k_{h'd'}} V_{k_{hd},k_{h'd'}}k^*_{fd}k^*_{f'd'} \quad (80)$$

$$= \gamma D^{2\zeta}\sum_{k_{hd},k_{h'd'}} \mathbb{E}[\partial_{k_{hd}}H(0)\partial_{k_{h'd'}}H(0)]k^*_{fd}k^*_{f'd'} + \gamma O(D^1N_b^{-1}) \quad (81)$$

$$= \gamma D^{2\zeta}\sum_{k_{hd},k_{h'd'}} \mathbb{E}[\partial_{k_{hd}}H(0)\partial_{k_{h'd'}}H(0)]k^*_{fd}k^*_{f'd'} + o(1). \quad (82)$$

Therefore, we get that $\mathbf{\Sigma}$ is constant matrix, independent of $\mathbf{u}$.

Placing it all together, we have a system

$$d\tilde{m}_{hf} = -\partial_{m_{hf}}\Phi(\mathbf{u})dt + c_\Sigma dB_t \quad (83)$$

$$\dot{r} = -r(\nabla_r\Phi(\mathbf{u}) + (\nabla_r\Phi(\mathbf{u}))^\top) - (\nabla_r\Phi(\mathbf{u}) + (\nabla_r\Phi(\mathbf{u}))^\top)r \quad (84)$$

$$\dot{b}_h = -\partial_{b_h}\Phi(\mathbf{u}) \quad (85)$$

$$\dot{v} = -\partial_v\Phi(\mathbf{u}) \quad (86)$$

□

Lemma A.6 (Sample complexity of weak recovery with SGD)**.** *Starting from initialization $m_{hf} = 0$ when initial gradient $\nabla_m\Phi(m=0) = 0$, the sample complexity of reaching $m_{hf} = \Theta(1)$ independent of D with high probability $1-o(1)$ is $O(D^{1+2\zeta}\log D)$ for arbitrary small $\zeta > 0$.*

Proof. By lemma A.5, for $\tilde{m}_{hf}$ around zero, we can write

$$d\tilde{m} \simeq -\nabla_m^2\Phi(0)\tilde{m}dt + c_\Sigma dB_t,$$

and projecting on the eigenvectors of $\nabla_m^2\Phi(0)$ with strictly negative eigenvalues (given by Lemma A.11), we get for some $c_v > 0$

$$d\left\langle v,\tilde{m}\right\rangle = c_v\left\langle v,\tilde{m}\right\rangle dt + c_{\Sigma,v}dBt.$$

This is a mean-repellent process, such that $\left\langle v,\tilde{m}\right\rangle(t) = c_{\Sigma,v}\int_0^t e^{c_v(t-s)}dB_s$, i.e. $\mathbb{E}|\left\langle v,\tilde{m}\right\rangle(t)| = \Theta(e^{c_vt})$ and $\mathbb{E}\left\langle v,\tilde{m}\right\rangle(t)^2 = \Theta(e^{2c_vt})$.

Now, the probability bound for a Gaussian variable $\left\langle v,\tilde{m}\right\rangle(t)$ to stay in the interval $(-\delta,\delta)$:

$$\mathbb{P}(|\left\langle v,\tilde{m}\right\rangle(t)| < \delta_D) \le \sqrt{\frac{2}{\pi}}\frac{\delta_D}{\sigma(\left\langle v,\tilde{m}\right\rangle(t))} = \Theta(\delta_De^{-c_vt}).$$

Using weak convergence of sufficient statistics $m_{hf}^{(D)}(t)$ to the limit with continuous cumulative distribution function, we get uniform convergence of CDFs, and setting $\delta_D = \delta D^{2\zeta}$ and $t = \frac{2\zeta+a}{c_v}\log D$ we get for D large enough

$$\mathbb{P}(|\langle v, \tilde{m}_D\rangle(t)| < \delta_D) = \mathbb{P}(|\langle v, m_D\rangle(t)| < \delta) \leq \Theta(D^{2\zeta}e^{-c_v(2\zeta+a)/c_v \log D}) = o(1)$$

Thus, we get that it takes $O(\log D)$ to escape from the fixed point with high probability. And the total time complexity is $O(D^{1+2\zeta}\log D)$ for arbitrary small $\zeta > 0$. □

A.4 Derivatives of the loss

We derive the technical results about the gradient, the hessian and the 4th order derivative of the loss $\tilde{\mathcal{E}}$ in the space of the order parameters. For the derivative of the activation function we use the notation

$$\partial_{h'\ell'}\sigma(\chi, b, v; h)_\ell = \frac{\partial}{\partial \chi_{h'\ell'}}\sigma(\chi, b, v; h)_\ell \,. \tag{87}$$

Lemma A.7 (Invariance of the unspecialized manifold by gradient descent.)**.** *Consider the unspecialized manifold $\mathcal{M}_{\rm u}$, where the heads are not specialized, defined by $m = \mathbb{1}_H\tilde{m}^\top$ with $\tilde{m} \in \mathbb{R}^F$, $r = \tilde{r}_1 I_H + \tilde{r}_2 \mathbb{1}_H\mathbb{1}_H^\top$ with $\tilde{r}_1 > 0, \tilde{r}_1 + H\tilde{r}_2 > 0$ and $b = \tilde{b}\mathbb{1}_H$ with $\tilde{b} \in \mathbb{R}$. $\mathcal{M}_{\rm u}$ is invariant by the gradient descent eq.* (10)*.*

Proof. The loss is invariant by permutation of the heads and therefore on $\mathcal{M}_{\rm u}$ $\nabla_{m_h}\tilde{\mathcal{E}}_\sigma$, $\partial_{r_{hh}}\tilde{\mathcal{E}}_\sigma$, $\partial_{r_{h\neq h'}}\tilde{\mathcal{E}}_\sigma$ and $\partial_{b_h}\tilde{\mathcal{E}}_\sigma$ do not depend on h. □

Lemma A.8 (Gradient of the loss at initialization and in the unspecialized phase)**.** *Let $\mathbb{E}\theta = (\mathbb{E}_{P_\theta}\theta_f)_{f\in[F]} \in \mathbb{R}^F$ be the mean of the signal. Take m, r and b on the unspecialized manifold $\mathcal{M}_{\rm u}$. Take all the heads aligned with $\mathbb{E}\theta$, i.e. take $x \in \mathbb{R}$ and $m_h = x\mathbb{E}\theta$ for all h. There is $c^{(1)}(x, r, b, v) \in \mathbb{R}$ such that for all h*

$$\nabla_{m_h}\tilde{\mathcal{E}}_\sigma(m, r, b, v) = -c^{(1)}(x, r, b, v)\mathbb{E}\theta \,. \tag{88}$$

Moreover at initialization $c^{(1)}(0, r, 0, 1) > 0$.

Proof. We compute the gradient of the loss in the space of the order parameters. We remind that the reparameterized loss is

$$\tilde{\mathcal{E}}_\sigma(m, r, b, v) = \mathbb{E}_{\epsilon,\theta,\chi^*,\xi}\left[\sum_\ell^L\left(\delta_{\ell,\epsilon} - \frac{1}{H}\sum_{h'}^H \sigma(\chi, b, v; h')_\ell\right)^2\right] \tag{89}$$

$$\chi_{h\ell} = \sum_f^F m_{hf}\chi^*_{f\ell} + \sum_{h'}^H r_{hh'}\xi_{h'} \,, \quad h \in [H], \ell \in [L] \,. \tag{90}$$

with $\epsilon \sim \mathrm{Unif}(\{1, \dots, L\})$, $\theta \sim P_\theta$ and conditionally on ϵ and θ, $\chi^*_{:,\ell} \sim \mathcal{N}(\delta_{\ell,\epsilon}\theta, I_F)$ and $\xi_{:,\ell} \sim \mathcal{N}(0, I_H)$ for $\ell \in [L]$. The gradient is

$$\nabla_{m_h}\tilde{\mathcal{E}}_\sigma(m, r, b, v) = 2\mathbb{E}_{\epsilon,\theta,\chi^*,\xi}\sum_\ell^L\left(\frac{1}{H}\sum_{h'}^H \sigma(\chi, b, v; h')_\ell - \delta_{\ell,\epsilon}\right)\sum_{h',\ell'}^{H,L}\partial_{h\ell'}\sigma(\chi, b, v; h')_\ell\chi^*_{:,\ell'}. \tag{91}$$

We show the gradient is collinear to $\mathbb{E}\theta$ if all heads align with $\mathbb{E}\theta$, i.e. $m_h = x\mathbb{E}\theta$ for all $h \in [H]$. Let $w \in \mathbb{R}^F$ be orthogonal to $\mathbb{E}\theta$.

$$w^\top\nabla_{m_h}\tilde{\mathcal{E}}_\sigma(m, r, b, v) = 2\mathbb{E}_{\epsilon,\theta,\chi,\xi}\sum_\ell^L\left(\frac{1}{H}\sum_{h'}^H \sigma(\chi, b, v; h')_\ell - \delta_{\ell,\epsilon}\right)\sum_{h',\ell'}^{H,L}\partial_{h\ell'}\sigma(\chi, b, v; h')_\ell w^\top\chi^*_{:,\ell'} \tag{92}$$

Then by orthogonality $w^\top \chi^*_{:\ell'}$ and $\chi_{h\ell} = x(\mathbb{E}\theta)^\top \chi^*_{:\ell} + \dots$ are independent Gaussian random variables for all ℓ, ℓ'. Thus, we can factorize the expectation. Since $\mathbb{E}\, w^\top \chi^*_{:\ell'} = w^\top \mathbb{E}\theta = 0$, we have $w^\top \nabla_{m_h} \tilde{\mathcal{E}}_\sigma(m, r, b, v) = 0$. Moreover we consider the unspecialized manifold, and so there is a same $c^{(1)}(x, r, b, v) \in \mathbb{R}$ for all h such that $\nabla_{m_h} \tilde{\mathcal{E}}_\sigma(m, r, b, v) = -c^{(1)}(x, r, b, v)\mathbb{E}\theta$.

At initialization $b = 0$, $v = 1$ and $m = 0$, and we have $\chi_{h'}$ independent of $\{\chi^*_h\}_{h\in[H]}$ for all h'; thus

$$\nabla_{m_h} \tilde{\mathcal{E}}_\sigma(0, r, 0, 1) = 2\mathbb{E}_{\epsilon,\xi} \sum_\ell^L \left(\frac{1}{H} \sum_{h'}^H \sigma(\chi, 0, 1; h')_\ell - \delta_{\ell,\epsilon} \right) \sum_{h',\ell'}^{H,L} \partial_{h\ell'} \sigma(\chi, 0, 1; h')_\ell \mathbb{E}_{\theta,\chi} \chi^*_{:,\ell'} \tag{93}$$

$$= 2\mathbb{E}_{\epsilon,\xi} \sum_\ell^L \left(\frac{1}{H} \sum_{h'}^H \sigma(\chi, 0, 1; h')_\ell - \delta_{\ell,\epsilon} \right) \sum_{h'}^H \partial_{h\epsilon} \sigma(\chi, 0, 1; h')_\ell \mathbb{E}\theta \tag{94}$$

$$= 2\mathbb{E}_{\epsilon,\xi} \left(\frac{1}{H} \sum_{h'}^H \sigma(\chi, 0, 1; h')_\epsilon - 1 \right) \sum_{h'}^H \partial_{h\epsilon} \sigma(\chi, 0, 1; h')_\epsilon \mathbb{E}\theta \tag{95}$$

$$+ 2\mathbb{E}_{\epsilon,\xi} \sum_{\ell\neq\epsilon}^L \frac{1}{H} \sum_{h'}^H \sigma(\chi, 0, 1; h')_\ell \sum_{h'}^H \partial_{h\epsilon} \sigma(\chi, 0, 1; h')_\ell \mathbb{E}\theta$$

The two pre-factors in front of $\mathbb{E}\theta$ are negative because

$$\frac{1}{H} \sum_{h'}^H \sigma(\chi, 0, 1; h')_\epsilon - 1 < 0 \qquad \sum_{h'}^H \partial_{h\epsilon} \sigma(\chi, 0, 1; h')_\epsilon > 0 \tag{96}$$

$$\frac{1}{H} \sum_{h'}^H \sigma(\chi, 0, 1; h')_\ell > 0 \qquad \sum_{h'}^H \partial_{h\epsilon} \sigma(\chi, 0, 1; h')_\ell < 0 \tag{97}$$

for all $\epsilon \neq \ell$, h, χ and for the three different activation functions σ. Consequently $c^{(1)}(0, r, 0, 1) > 0$. □

Lemma A.9 (Gradient of the loss with respect to s at small s)**.** *Consider $r = s = 0$, then for all m*

$$\nabla_s \tilde{\mathcal{E}}_\sigma(m, 0, b, v) = 0. \tag{98}$$

Proof. The loss is a symmetric function of s: it is invariant by the change of variables $(s, \xi) \mapsto (-s, -\xi)$. Therefore its gradient is null at $s = 0$. □

Lemma A.10 (Gradient of the loss with respect to b and v at small m and s for the softmax-1)**.** *Consider $m = 0$, $r = s = 0$ and σ to be the softmax-1. Take unspecialized heads i.e. $b = \tilde{b}\mathbb{1}_H$ for $\tilde{b} \in \mathbb{R}$. Then*

$$\nabla_{\tilde{b}} \tilde{\mathcal{E}}_\sigma(0, 0, b, v) = -2 \left(\frac{Lv}{L + e^{\tilde{b}}} - 1 \right) \frac{v}{H} \frac{e^{\tilde{b}}}{(L + e^{\tilde{b}})^2} \tag{99}$$

$$\nabla_v \tilde{\mathcal{E}}_\sigma(0, 0, b, v) = 2 \left(\frac{Lv}{L + e^{\tilde{b}}} - 1 \right) \frac{1}{L + e^{\tilde{b}}} \tag{100}$$

The fixed-points of this system satisfy $Lv = L + e^{\tilde{b}}$ and they are attractive. Last at initialization $b = 0$ and $v = 1$ and

$$\nabla_{\tilde{b}} \tilde{\mathcal{E}}_\sigma(0, 0, 0, 1) > 0 \tag{101}$$

$$\nabla_v \tilde{\mathcal{E}}_\sigma(0, 0, 0, 1) < 0 \tag{102}$$

Proof. The proof is a straightforward computation of the derivatives. □

Lemma A.11 (Hessian of the loss at small m and s)**.** *Consider $m \in \mathbb{R}^{H\times F}$ in the space orthogonal to $\mathbb{E}\theta$ i.e. $m_h^\top \mathbb{E}\theta = 0$ for all h. Take $s \in \mathbb{R}^{H\times H}$ and $r = ss^\top$. Assume that b is not specialized, i.e.*

$b_h = \tilde{b}$ *for all* h*. For the softmax-1 assume that* b *and* v *reached the fixed-point described by Lemma A.10 i.e.* $Lv = L + e^{\bar{b}}$*. The loss around* $m \approx 0$*,* $s \approx 0$*,* b *and* v *can be expanded as*

$$\begin{aligned}\tilde{\mathcal{E}}_\sigma(m, r, b + \bar{b}, v) &= \tilde{\mathcal{E}}_\sigma(0, 0, b, v) + c_1^{(2)} \mathbb{1}_H^\top s s^\top \mathbb{1}_H + c_5^{(2)} (\bar{b}^\top \mathbb{1}_H)^2 \\ &+ \left(\mathbb{1}_H \mathbb{1}_H^\top \otimes (c_1^{(2)} I_F + (c_2^{(2)} + c_4^{(2)}) \operatorname{Cov} \theta) - (c_3^{(2)} + c_4^{(2)}) I_H \otimes \operatorname{Cov} \theta \right) \cdot (m, m) \\ &+ \mathcal{O}((||s||_F^2 + ||b||_2^2 + ||m||_F^2)^2)\end{aligned} \tag{103}$$

with $\otimes$ *the tensorial product between the spaces* $\mathbb{R}^H$ *and* $\mathbb{R}^F$*, and* $c_1^{(2)}, c_2^{(2)}, c_3^{(2)} \in \mathbb{R}$ *strictly positive and* $c_4^{(2)} \in \mathbb{R}, c_5^{(2)} \in \mathbb{R}$ *positive for all* $L \geq 3$*.*

Proof. $\nabla^2_{s,m} \tilde{\mathcal{E}}_\sigma(0, 0, b, v) = 0$ because the gradient w.r.t. s brings a ξ while the gradient w.r.t. m brings a χ^*. ξ is centered and independent of χ^* thus the expectation is null. The same reasoning holds for $\nabla^2_{s,b} \tilde{\mathcal{E}}_\sigma(0, 0, b, v) = 0$. We also have $\nabla^2_{m,b} \tilde{\mathcal{E}}_\sigma(0, 0, b, v) \cdot (m, b) = 0$ because we consider $m \mathbb{E}\theta = 0$.

For the Hessian w.r.t. s we expand the loss around $\chi = s\xi \approx 0$:

$$\tilde{\mathcal{E}}_\sigma(0, r, b, v) = \mathbb{E}_{\epsilon,\xi} \left[\sum_\ell^L \left(\delta_{\ell,\epsilon} - \frac{1}{H} \sum_h^H \sigma(s\xi, b, v; h)_\ell \right)^2 \right] \tag{104}$$

$$= \mathbb{E}_{\epsilon,\xi} \left[\sum_\ell^L \left(\delta_{\ell,\epsilon} - \frac{1}{H} \sum_h^H \left(\sigma(0, b, v; h)_\ell + \sum_{h',\ell'} \partial_{h'\ell'} \sigma(0, b, v; h)_\ell s_{h'}^\top \xi_{:\ell'} \right. \right. \right. \tag{105}$$

$$\left. \left. \left. + \frac{1}{2} \sum_{h',\ell',h'',\ell''} \partial^2_{h'\ell',h''\ell''} \sigma(0, b, v; h)_\ell s_{h'}^\top \xi_{:\ell'} s_{h''}^\top \xi_{:\ell''} \right) \right)^2 \right]$$

$$= \mathbb{E}_{\epsilon,\xi} \left[\sum_\ell^L \left(\delta_{\ell,\epsilon} - H^{-1} \sum_h \sigma(0, b, v; h)_\ell \right)^2 - \sum_\ell^L \frac{1}{H} \left(\delta_{\ell,\epsilon} - H^{-1} \sum_h \sigma(0, b, v; h)_\ell \right) \right. \tag{106}$$

$$\left. \times \sum_{h,h',\ell',h'',\ell''} \partial^2_{h'\ell',h''\ell''} \sigma(0, b, v; h)_\ell s_{h'}^\top \xi_{:\ell'} s_{h''}^\top \xi_{:\ell''} + \sum_\ell^L \frac{1}{H^2} \left(\sum_{h,h',\ell'} \partial_{h'\ell'} \sigma(0, b, v; h)_\ell s_{h'}^\top \xi_{:\ell'} \right)^2 \right]$$

$$= \tilde{\mathcal{E}}_\sigma(0, 0, b, v) + \sum_\ell^L \frac{1}{H} \left(H^{-1} \sum_h \sigma(0, b, v; h)_\ell - \delta_{\ell,1} \right) \sum_{h,h',h'',\ell'} \partial^2_{h'\ell',h''\ell'} \sigma(0, b, v; h)_\ell s_{h'}^\top s_{h'}$$

$$+ \sum_\ell^L \frac{1}{H^2} \sum_{h_1,h_1',h_2,h_2',\ell'} \partial_{h_1'\ell'} \sigma(0, b, v; h_1)_\ell \partial_{h_2'\ell'} \sigma(0, b, v; h_2)_\ell s_{h_1'}^\top s_{h_2'} \tag{107}$$

$$= \tilde{\mathcal{E}}_\sigma(0, 0, b, v) + \frac{1}{H^2} \sum_{h',h''} \sum_{\ell,\ell'} \left(\sum_h \partial_{h'\ell'} \sigma(0, b, v; h)_\ell \right) \left(\sum_h \partial_{h''\ell'} \sigma(0, b, v; h)_\ell \right) s_{h'}^\top s_{h''} \tag{108}$$

where we took the expectation over $\xi_\ell \sim \mathcal{N}(0, I_H)$ and discarded the term of order one in s with null expectation. We simplified $\sum_\ell^L \left(H^{-1} \sum_h \sigma(0, b, v; h)_\ell - \delta_{\ell,1} \right) \times (\text{function independent of } \ell) = 0$, using that b is not specialized and using the fixed-point condition for b and v for the softmax-1. Consequently,

$$\tilde{\mathcal{E}}_\sigma(0, r, b, v) = \tilde{\mathcal{E}}_\sigma(0, 0, b, v) + c_1^{(2)} \sum_{h',h''} (s s^\top)_{h'h''} \tag{109}$$

with the constant

$$c_1^{(2)} = \frac{1}{H^2} \sum_{\ell,\ell'} \left(\sum_h \partial_{1\ell'} \sigma(0, b, v; h)_\ell \right)^2 > 0 \,. \tag{110}$$

We chose the particular index 1 for the derivative because of the permutation invariance with respect to the heads.

For the Hessian w.r.t. m we perform a similar derivation, expanding the loss around $\chi = m\chi^* \approx 0$. We take in account the fact that $\chi^*_{f\ell} \sim \mathcal{N}(\delta_{\ell,\epsilon}\theta_f, 1)$.

$$\tilde{\mathcal{E}}_\sigma(m,0,b,v) = \mathbb{E}_{\epsilon,\theta,\chi^*}\left[\sum_\ell^L \left(\delta_{\ell,\epsilon} - \frac{1}{H}\sum_h^H \sigma(m\chi^*, b, v; h)_\ell\right)^2\right] \tag{111}$$

$$= \mathbb{E}_{\epsilon,\theta,\chi^*}\left[\sum_\ell^L \left(\delta_{\ell,\epsilon} - \frac{1}{H}\sum_h^H \left(\sigma(0,b,v;h)_\ell + \sum_{h',\ell'} \partial_{h'\ell'}\sigma(0,b,v;h)_\ell m_{h'}^\top \chi^*_{:\ell'}\right.\right.\right. \tag{112}$$

$$\left.\left.\left. + \frac{1}{2}\sum_{h',\ell',h'',\ell''} \partial^2_{h'\ell',h''\ell''}\sigma(0,b,v;h)_\ell m_{h'}^\top \chi^*_{:\ell'} m_{h''}^\top \chi^*_{:\ell''}\right)\right)^2\right]$$

$$= \mathbb{E}_{\epsilon,\theta,\chi^*}\left[\sum_\ell^L \left(\delta_{\ell,\epsilon} - H^{-1}\sum_h \sigma(0,b,v;h)_\ell\right)^2 + \sum_\ell^L \frac{1}{H^2}\left(\sum_{h,h',\ell'} \partial_{h'\ell'}\sigma(0,b,v;h)_\ell m_{h'}^\top \chi^*_{:\ell'}\right)^2\right.$$

$$- 2\sum_\ell^L \frac{1}{H}\left(\delta_{\ell,\epsilon} - H^{-1}\sum_h \sigma(0,b,v;h)_\ell\right)\sum_{h,h',\ell'} \partial_{h'\ell'}\sigma(0,b,v;h)_\ell m_{h'}^\top \chi^*_{:\ell'} \tag{113}$$

$$\left. - \sum_\ell^L \frac{1}{H}\left(\delta_{\ell,\epsilon} - H^{-1}\sum_h \sigma(0,b,v;h)_\ell\right)\sum_{h,h',\ell',h'',\ell''} \partial^2_{h'\ell',h''\ell''}\sigma(0,b,v;h)_\ell m_{h'}^\top \chi^*_{:\ell'} m_{h''}^\top \chi^*_{:\ell''}\right]$$

$$= \tilde{\mathcal{E}}_\sigma(0,0,b,v) + 2\sum_\ell^L \frac{1}{H}\left(H^{-1}\sum_h \sigma(0,b,v;h)_\ell - \delta_{\ell,1}\right)\sum_{h,h'} \partial_{h'1}\sigma(0,b,v;h)_\ell m_{h':}^\top \mathbb{E}\theta \tag{114}$$

$$+ \sum_\ell^L \frac{1}{H}\left(H^{-1}\sum_h \sigma(0,b,v;h)_\ell - \delta_{\ell,1}\right)\sum_{h,h',\ell',h''} \partial^2_{h'\ell',h''\ell'}\sigma(0,b,v;h)_\ell \sum_{f,f'}^F m_{h'f} m_{h''f'}(\delta_{f,f'} + \delta_{\ell',1}\mathbb{E}\theta_f\theta_{f'})$$

$$+ \sum_\ell^L \frac{1}{H^2}\sum_{h_1,h_1',h_2,h_2',\ell'} \partial_{h_1'\ell'}\sigma(0,b,v;h_1)_\ell \partial_{h_2'\ell'}\sigma(0,b,v;h_2)_\ell \sum_{f,f'}^F m_{h_1'f} m_{h_2'f'}(\delta_{f,f'} + \delta_{\ell',1}\mathbb{E}\theta_f\theta_{f'})$$

$$= \tilde{\mathcal{E}}_\sigma(0,0,b,v) + \sum_\ell^L \frac{1}{H}\left(H^{-1}\sum_h \sigma(0,b,v;h)_\ell - \delta_{\ell,1}\right)\sum_{h,h'} \partial^2_{h'1,h''1}\sigma(0,b,v;h)_\ell \sum_{f,f'}^F m_{h'f} m_{h''f'}\mathbb{E}\theta_f\theta_{f'} \tag{115}$$

$$+ \frac{1}{H^2}\sum_{h,h'}\sum_{\ell,\ell'}\left(\sum_h \partial_{h'\ell'}\sigma(0,b,v;h)_\ell\right)\left(\sum_h \partial_{h''\ell'}\sigma(0,b,v;h)_\ell\right)\sum_{f,f'}^F m_{h'f} m_{h''f'}(\delta_{f,f'} + \delta_{\ell',1}\mathbb{E}\theta_f\theta_{f'})$$

We consider the space orthogonal to $\mathbb{E}\theta$, i.e. $m_{h:}^\top \mathbb{E}\theta = 0$ for all h. We simplified $\sum_\ell^L \left(H^{-1}\sum_h \sigma(0,b,v;h)_\ell - \delta_{\ell,1}\right) \times (\text{function independent of } \ell) = 0$. We introduce the covariance of θ. Consequently,

$$\tilde{\mathcal{E}}_\sigma(m,0,b,v) = \tilde{\mathcal{E}}_\sigma(0,0,b,v) + c_1^{(2)}\sum_{h',h''} m_{h'}^\top m_{h''} + (c_2^{(2)} + c_4^{(2)})\sum_{h',h''} m_{h'}^\top \operatorname{Cov}(\theta) m_{h''}$$
$$- (c_3^{(2)} + c_4^{(2)})\sum_{h'} m_{h'}^\top \operatorname{Cov}(\theta) m_{h'} \tag{116}$$

$$c_1^{(2)} = \frac{1}{H^2}\sum_{\ell,\ell'}\left(\sum_h \partial_{1\ell'}\sigma(0,b,v;h)_\ell\right)^2 > 0 \tag{117}$$

$$c_2^{(2)} = \frac{1}{H^2}\sum_{\ell}\left(\sum_h \partial_{11}\sigma(0,b,v;h)_\ell\right)^2 > 0 \tag{118}$$

$$c_3^{(2)} = -\frac{1}{H}\sum_\ell^L \left(H^{-1}\sum_h \sigma(0,b,v;h)_\ell - \delta_{\ell,1}\right)\sum_h \partial^2_{11,11}\sigma(0,b,v;h)_\ell \tag{119}$$

$$c_4^{(2)} = \frac{1}{H}\sum_{\ell}^{L}\left(H^{-1}\sum_h \sigma(0,b,v;h)_\ell - \delta_{\ell,1}\right)\sum_h \partial^2_{11,21}\sigma(0,b,v;h)_\ell \tag{120}$$

so the loss is

$$\begin{aligned}\tilde{\mathcal{E}}_\sigma(m,0,b,v) &= \tilde{\mathcal{E}}_\sigma(0,0,b,v) \\ &+ \left(\mathbb{1}_H\mathbb{1}_H^\top \otimes (c_1^{(2)} I_F + (c_2^{(2)} + c_4^{(2)})\operatorname{Cov}\theta) - (c_3^{(2)} + c_4^{(2)}) I_H \otimes \operatorname{Cov}\theta\right)\cdot(m,m)\,.\end{aligned} \tag{121}$$

Between $c_3^{(2)}$ and $c_4^{(2)}$ we distinguished the cases $h' = h''$ and $h' \neq h''$. We compute these two constants for the different activation functions. We have

σ	$c_3^{(2)}$	$c_4^{(2)}$
softmax	$H^{-1}(L-1)L^{-2}(1-2L^{-1})$	0
softmax-1	$vH^{-1}(L-1)L^{-1}(L+e^{\tilde{b}})^{-1}(1-2(L+e^{\tilde{b}})^{-1})$	0
B-softmax	$H^{-1}(L-1)L^{-2}(1-2(HL)^{-1})$	$(L-1)L^{-1}2(HL)^{-2}$

Consequently for all $L \geq 3$ one has $c_3^{(2)} > 0$ and $c_4^{(2)} \geq 0$.

For the Hessian w.r.t. b we compute that for the softmax-1

$$\partial^2_{b_h,b'_h}\tilde{\mathcal{E}}_\sigma(0,0,b,v) = 2\frac{v^2 L}{H^2}\frac{e^{2\tilde{b}}}{(L+e^{\tilde{b}})^4} \tag{122}$$

and for the B-softmax

$$\partial^2_{b_h,b'_h}\tilde{\mathcal{E}}_\sigma(0,0,b,v) = 0\,. \tag{123}$$

□

Lemma A.12 (Hessian of the loss at small but finite m)**.** *We take σ softmax. We assume that $\mathbb{E}\theta = 0$ and that the θ_f are independent. Let n be an integer, $\bar{f}_1, \ldots, \bar{f}_n \in [F]^n$ all different, $\bar{m} \in \mathbb{R}^{H\times F}$ and assume that $\bar{m}_{:f} = 0$ for all $f \notin \{\bar{f}_1, \ldots, \bar{f}_n\}$. Pick $f \notin \{\bar{f}_1, \ldots, \bar{f}_n\}$; then the Hessian of the loss is*

$$\begin{aligned}\nabla^2_{m_{:f},m_{:f}}\tilde{\mathcal{E}}_\sigma(\bar{m},0,0,0) &= \nabla^2_{m_{:f},m_{:f}}\tilde{\mathcal{E}}_\sigma(0,0,0,0) + \sum_i^n c_{1,i}^{(4)}||\bar{m}_{:\bar{f}_i}||_2^2 I_H \\ &+ \sum_i^n c_{2,i}^{(4)}\operatorname{Diag}(\bar{m}_{:\bar{f}_i}^{\odot 2}) + \sum_i^n c_{3,i}^{(4)}\bar{m}_{:\bar{f}_i}\bar{m}_{:\bar{f}_i}^\top + M + \mathcal{O}(||\bar{m}||_F^4)\end{aligned} \tag{124}$$

where, for a vector v, $v^{\odot 2}$ is the element-wise square and $\operatorname{Diag}(v)$ the diagonal matrix which diagonal is v. M is a quadratic form that cancels when $m_{:f}^\top \mathbb{1}_H = 0$ and $\bar{m}^\top \mathbb{1}_H = 0$, and where $c_{1,i}^{(4)} > 0, c_{2,i}^{(4)} \in \mathbb{R}, c_{3,i}^{(4)} > 0$ for all $L \geq 3$ does not depend on $\bar{m}$.

Proof. The proof is a Taylor expansion of the loss to the 2nd order in m_f around $\bar{m}$ and to the 2nd order in $\bar{m}$ around 0. We take $m \in \mathbb{R}^{H\times F}$ with $m_{\bar{f}_i} = 0$ for all $\bar{f}_i$; it encompasses the case of a matrix where only the f-th column is not null equal to m_f. Since we consider σ softmax, to lighten the notation we write $\sigma(\chi_h)$ for $\sigma(\chi, b, v; h)$.

$$\tilde{\mathcal{E}}_\sigma(m+\bar{m},0,0,0) = \mathbb{E}_{\epsilon,\theta,\chi^*}\left[\sum_\ell^L\left(\delta_{\ell,\epsilon} - \frac{1}{H}\sum_h^H \sigma(m_h^\top\chi^*)_\ell\right)^2\right] \tag{125}$$

$$\begin{aligned}&= \mathbb{E}_{\epsilon,\theta,\chi^*}\left[\sum_\ell^L\left(\delta_{\ell,\epsilon} - \frac{1}{H}\sum_h^H\left(\sigma(\bar{m}_h^\top\chi^*)_\ell + \sum_{\ell'}\partial_{\ell'}\sigma(\bar{m}_h^\top\chi^*)_\ell m_h^\top\chi^*_{:\ell'}\right.\right.\right. \\ &\left.\left.\left. + \frac{1}{2}\sum_{\ell',\ell''}\partial^2_{\ell'\ell''}\sigma(\bar{m}_h^\top\chi^*)_\ell m_h^\top\chi^*_{:\ell'}m_h^\top\chi^*_{:\ell''}\right)\right)^2\right]\end{aligned} \tag{126}$$

$$= \mathbb{E}_{\epsilon,\theta,\chi^*}\left[\sum_\ell^L\left(\delta_{\ell,\epsilon}-\frac{1}{H}\sum_h^H\sigma(\bar{m}_h^\top\chi^*)_\ell\right)^2+\sum_\ell^L\frac{1}{H^2}\left(\sum_{h,\ell'}\partial_{\ell'}\sigma(\bar{m}_h^\top\chi^*)_\ell m_h^\top\chi^*_{:\ell'}\right)^2\right. \tag{127}$$

$$-2\sum_\ell^L\left(\delta_{\ell,\epsilon}-\frac{1}{H}\sum_h^H\sigma(\bar{m}_h^\top\chi^*)_\ell\right)\frac{1}{H}\sum_{h,\ell'}\partial_{\ell'}\sigma(\bar{m}_h^\top\chi^*)_\ell m_h^\top\chi^*_{:\ell'}$$

$$\left.-\sum_\ell^L\left(\delta_{\ell,\epsilon}-\frac{1}{H}\sum_h^H\sigma(\bar{m}_h^\top\chi^*)_\ell\right)\frac{1}{H}\sum_{h,\ell',\ell''}\partial^2_{\ell'\ell''}\sigma(\bar{m}_h^\top\chi^*)_\ell m_h^\top\chi^*_{:\ell'}m_h^\top\chi^*_{:\ell''}\right]$$

$$= \tilde{\mathcal{E}}_\sigma(\bar{m},0,0,0)+-2\mathbb{E}_{\epsilon,\theta,\chi^*}\sum_\ell^L\left(\delta_{\ell,\epsilon}-\frac{1}{H}\sum_h^H\sigma(\bar{m}_h^\top\chi^*)_\ell\right)\frac{1}{H}\sum_h\partial_1\sigma(\bar{m}_h^\top\chi^*)_\ell\underbrace{m_h^\top\mathbb{E}\theta}_{=0} \tag{128}$$

$$\underbrace{-\mathbb{E}_{\epsilon,\theta,\chi^*}\sum_\ell^L\left(\delta_{\ell,\epsilon}-\frac{1}{H}\sum_h^H\sigma(\bar{m}_h^\top\chi^*)_\ell\right)\frac{1}{H}\sum_{h,\ell'}\partial^2_{\ell'\ell'}\sigma(\bar{m}_h^\top\chi^*)_\ell\sum_{f,f'}^F m_{hf}m_{hf'}(\delta_{f,f'}+\delta_{\ell',1}\mathbb{E}\theta_f\theta_{f'})}_{(1)}$$

$$\underbrace{+\mathbb{E}_{\epsilon,\theta,\chi^*}\sum_\ell^L\frac{1}{H^2}\sum_{h,h',\ell'}\partial_{\ell'}\sigma(\bar{m}_h^\top\chi^*)_\ell\partial_{\ell'}\sigma(\bar{m}_{h'}^\top\chi^*)_\ell\sum_{f,f'}^F m_{hf}m_{h'f'}(\delta_{f,f'}+\delta_{\ell',1}\mathbb{E}\theta_f\theta_{f'})}_{(2)}$$

where we used the independence of the θ_f to factorize the expectation. We expand with respect to $\bar{m}$, discarding the 1st order terms because $\mathbb{E}\theta=0$.

$$(1) = c_2^{(2)}\sum_h m_h^\top\operatorname{Cov}(\theta)m_h \tag{129}$$

$$+\mathbb{E}_{\epsilon,\theta,\chi^*}\sum_\ell^L\frac{1}{2H}\sum_{h',\ell'',\ell'''}\partial^2_{\ell''\ell'''}\sigma(0)_\ell\bar{m}_{h'}^\top\chi^*_{:\ell''}\bar{m}_{h'}^\top\chi^*_{:\ell'''}\frac{1}{H}\sum_{h,\ell'}\partial^2_{\ell'\ell'}\sigma(0)_\ell m_h^\top(I_F+\delta_{\ell',1}\operatorname{Cov}(\theta))m_h$$

$$+\mathbb{E}_{\epsilon,\theta,\chi^*}\sum_\ell^L\frac{1}{H}\underbrace{\sum_{h',\ell'''}\partial_{\ell'''}\sigma(0)_\ell\bar{m}_{h'}^\top}_{=0}\chi^*_{:\ell'''}\frac{1}{H}\sum_{h,\ell',\ell''}\partial^3_{\ell'\ell'\ell''}\sigma(0)_\ell\bar{m}_h^\top\chi^*_{:\ell''}m_h^\top(I_F+\delta_{\ell',1}\operatorname{Cov}(\theta))m_h$$

$$+\mathbb{E}_{\epsilon,\theta,\chi^*}\sum_\ell^L(\sigma(0)_\ell-\delta_{\ell,1})\frac{1}{2H}\sum_{h,\ell',\ell'',\ell'''}\partial^4_{\ell'\ell'\ell''\ell'''}\sigma(0)_\ell\bar{m}_h^\top\chi^*_{:\ell''}\bar{m}_h^\top\chi^*_{:\ell'''}m_h^\top(I_F+\delta_{\ell',1}\operatorname{Cov}(\theta))m_h$$

$$= c_2^{(2)}\sum_h m_h^\top\operatorname{Cov}(\theta)m_h \tag{130}$$

$$+\sum_\ell^L\frac{1}{2H^2}\sum_{h',\ell''}\partial^2_{\ell''\ell''}\sigma(0)_\ell\bar{m}_{h'}^\top(I_F+\delta_{\ell'',1}\operatorname{Cov}(\theta))\bar{m}_{h'}\sum_{h,\ell'}\partial^2_{\ell'\ell'}\sigma(0)_\ell m_h^\top(I_F+\delta_{\ell',1}\operatorname{Cov}(\theta))m_h$$

$$+\sum_\ell^L(\sigma(0)_\ell-\delta_{\ell,1})\sum_{\ell',\ell''}\frac{1}{2H}\sum_h\partial^4_{\ell'\ell'\ell''\ell''}\sigma(0)_\ell\bar{m}_h^\top(I_F+\delta_{\ell'',1}\operatorname{Cov}(\theta))\bar{m}_h m_h^\top(I_F+\delta_{\ell',1}\operatorname{Cov}(\theta))m_h$$

$$= c_2^{(2)}\sum_h m_h^\top\operatorname{Cov}(\theta)m_h+\sum_\ell^L\frac{1}{2H^2}\sum_{h'}\partial^2_{1,1}\sigma(0)_\ell\bar{m}_{h'}^\top\operatorname{Cov}(\theta)\bar{m}_{h'}\sum_h\partial^2_{1,1}\sigma(0)_\ell m_h^\top\operatorname{Cov}(\theta)m_h$$

$$+\sum_\ell^L(\sigma(0)_\ell-\delta_{\ell,1})\sum_{\ell',\ell''}\frac{1}{2H}\sum_h\partial^4_{\ell'\ell'\ell''\ell''}\sigma(0)_\ell\bar{m}_h^\top(I_F+\delta_{\ell'',1}\operatorname{Cov}\theta)\bar{m}_h m_h^\top(I_F+\delta_{\ell',1}\operatorname{Cov}\theta)m_h \tag{131}$$

where we discarded a term because of $\bar{m}_{:f}^\top\mathbb{1}_H=0$ and used that $\sum_{\ell'}\partial^2_{\ell'\ell'}\sigma(0)_\ell=0$ for all ℓ. Consequently there are constants $\tilde{c}^{(4)}_{1,i}>0, \tilde{c}^{(4)}_{2,i}, \tilde{c}^{(4)}_{3,i}\in\mathbb{R}$ for all $L\geq 3$ independent of $\bar{m}$ such that

$$(1) = c_2^{(2)}(I_H\otimes\operatorname{Cov}\theta)\cdot(m,m)+\sum_i^n\tilde{c}^{(4)}_{1,i}|\bar{m}_{:\bar{f}_i}||_2^2\,(I_H\otimes\operatorname{Cov}\theta)\cdot(m,m)$$

$$+\sum_{i}^{n}\mathrm{Diag}(\bar{m}_{:\bar{f}_i}^{\odot 2})\otimes\left(\tilde{c}_{2,i}^{(4)}I_F+\tilde{c}_{3,i}^{(4)}\,\mathrm{Cov}\,\theta\right)\cdot(m,m)\,. \tag{132}$$

Turning to the second term,

$$(2)=c_1^{(2)}\sum_{h,h'}m_h^\top m_{h'}+c_3^{(2)}\sum_{h,h'}m_h^\top\,\mathrm{Cov}(\theta)m_{h'} \tag{133}$$

$$+\sum_{\ell}^{L}\frac{1}{H^2}\sum_{h,h',\ell',\ell''}\partial^2_{\ell'\ell''}\sigma(0)_\ell\partial^2_{\ell'\ell''}\sigma(0)_\ell\bar{m}_h^\top(I_F+\delta_{\ell'',1}\,\mathrm{Cov}(\theta))\bar{m}_{h'}m_h^\top(I_F+\delta_{\ell',1}\,\mathrm{Cov}(\theta))m_{h'}$$

where we kept only terms with even number of h and h' because of $m_{:f}^\top\mathbb{1}_H=0$ and took the expectation. Consequently there are constants $\tilde{c}_{4,i}^{(4)}>0,\tilde{c}_{5,i}^{(4)}>0$ for all $L\geq 2$ independent of $\bar{m}$ such that

$$(2)=\mathbb{1}_H\mathbb{1}_H^\top\otimes(c_1^{(2)}I_F+c_3^{(2)}\,\mathrm{Cov}\,\theta)\cdot(m,m)+\sum_{i}^{n}\bar{m}_{:\bar{f}_i}\bar{m}_{:\bar{f}_i}^\top\otimes\left(\tilde{c}_{4,i}^{(4)}I_F+\tilde{c}_{5,i}^{(4)}\,\mathrm{Cov}\,\theta\right)\cdot(m,m) \tag{134}$$

Assembling the two parts together we obtain

$$\tilde{\mathcal{E}}_\sigma(m+\bar{m},0,0,0)=\tilde{\mathcal{E}}_\sigma(\bar{m},0,0,0)+\tilde{\mathcal{E}}_\sigma(m,0,0,0)-\tilde{\mathcal{E}}_\sigma(0,0,0,0) \tag{135}$$

$$+\sum_{i}^{n}\tilde{c}_{1,i}^{(4)}||\bar{m}_{:\bar{f}_i}||_2^2\,(I_H\otimes\mathrm{Cov}\,\theta)\cdot(m,m)+\sum_{i}^{n}\mathrm{Diag}(\bar{m}_{:\bar{f}_i}^{\odot 2})\otimes\left(\tilde{c}_{2,i}^{(4)}I_F+\tilde{c}_{3,i}^{(4)}\,\mathrm{Cov}\,\theta\right)\cdot(m,m)$$

$$+\sum_{i}^{n}\bar{m}_{:\bar{f}_i}\bar{m}_{:\bar{f}_i}^\top\otimes\left(\tilde{c}_{4,i}^{(4)}I_F+\tilde{c}_{5,i}^{(4)}\,\mathrm{Cov}\,\theta\right)\cdot(m,m)+\mathcal{O}(||\bar{m}||_F^4,||m||_F^4)\,.$$

For a particular feature f and a particular magnetization $m_{:f}$ we extract the corresponding term to obtain

$$\nabla^2_{m_f,m_f}\tilde{\mathcal{E}}_\sigma(\bar{m},0,0,0)=\nabla^2_{m_f,m_f}\tilde{\mathcal{E}}_\sigma(0,0,0,0)+\sum_{i}^{n}c_{1,i}^{(4)}||\bar{m}_{:\bar{f}_i}||_2^2I_H \tag{136}$$

$$+\sum_{i}^{n}c_{2,i}^{(4)}\mathrm{Diag}(\bar{m}_{:\bar{f}_i}^{\odot 2})+\sum_{i}^{n}c_{3,i}^{(4)}\bar{m}_{:\bar{f}_i}\bar{m}_{:\bar{f}_i}^\top+M+\mathcal{O}(||\bar{m}||_F^4)$$

with $c_{1,i}^{(4)}=\mathrm{Var}(\theta_f)\tilde{c}_{1,i}$, $c_{2,i}^{(4)}=\tilde{c}_{2,i}^{(4)}+\tilde{c}_{3,i}^{(4)}\,\mathrm{Var}(\theta_f)$ and $c_{3,i}^{(4)}=\tilde{c}_{4,i}^{(4)}+\tilde{c}_{5,i}^{(4)}\,\mathrm{Var}(\theta_f)$. □

A.5 Expressiveness and performances of different activation functions

In this part we derive the results about the Bayes risk and the expressivity of the softmax and softmax-1 activation functions.

Proof of Proposition 4.1. The Bayes-optimal estimator is the posterior mean under our probabilistic model. It is given by the conditional probability $P(\epsilon=\ell|X,\{k_f^*\}_{f\in[F]})$:

$$\hat{y}_{\mathrm{Bayes}}(X,k^*)=\sum_{\ell}^{L}P(\epsilon=\ell|X,k^*)X_\ell\,. \tag{137}$$

We can compute it using the Bayes formula:

$$P(\epsilon=\ell|X,\{k_f^*\}_{f\in[F]})=\frac{P(\epsilon=\ell,X|\{k_f^*\}_{f\in[F]})}{P(X|\{k_f^*\}_{f\in[F]})}\,. \tag{138}$$

The numerator is computed as

$$P(\epsilon=\ell,X|\{k_f^*\}_{f\in[F]})=\int_\theta P(\epsilon=\ell,\theta,X|\{k_f^*\}_{f\in[F]})\,d\theta \tag{139}$$

$$= \int_\theta P(X|\epsilon = \ell, \theta, \{k_f^*\}_{f\in[F]}) P(\epsilon = \ell) P(\theta)\, d\theta\,, \tag{140}$$

and denoting the Gram matrix of $\{k_f^*\}_{f\in[F]}$ as p, we get

$$P(X|\epsilon = \ell, \theta, \{k_f^*\}_{f\in[F]}) = \frac{1}{Z} e^{-\frac{\|X_\ell - \sum_f \theta_f k_f^*\|_2^2}{2}} \prod_{\ell'\neq\ell}^{L} e^{-\frac{\|X_{\ell'}\|_2^2}{2}} \tag{141}$$

$$= \exp\left(-\frac{\|\sum_f \theta_f k_f^*\|_2^2 - 2X_\ell^\top(\sum_f \theta_f k_f^*)}{2}\right) \frac{1}{Z} \prod_{\ell'}^{L} e^{-\frac{\|X_{\ell'}\|_2^2}{2}} \tag{142}$$

$$= \exp -\frac{\theta^\top p \theta}{2} \exp X_\ell^\top(\hat{k}(\theta)) \cdot \frac{1}{Z} \prod_{\ell'}^{L} e^{-\frac{\|X_{\ell'}\|_2^2}{2}}\,. \tag{143}$$

And denominator is a sum over ℓ' of the conditional $P(\epsilon = \ell, X|\{k_f^*\}_{f\in[F]})$. Combining it all together, we get

$$P(\epsilon = \ell|X, \{k_f^*\}_{f\in[F]}) = \frac{\int_\theta \exp -\frac{\theta^\top p\theta}{2} \exp X_\ell^\top(\hat{k}(\theta)) P(\theta) d\theta}{\sum_{\ell'}^{L} \int_\theta \exp -\frac{\theta^\top p\theta}{2} \exp X_{\ell'}^\top(\hat{k}(\theta)) P(\theta) d\theta}. \tag{144}$$

□

Proof of Proposition 4.3. We show that the softmax (and softmax-v) attention is not well specified for our data model, while the softmax-1 is well specified.

The intuition is that for softmax one head cannot be good at the same time on a spike $\hat{k}$ and on the opposite $-\hat{k}$, and has to return noise in some cases. Assume that the softmax(-v) attention is well-specified, i.e. the reparameterized loss is $\tilde{\mathcal{E}}_\sigma(m, r, 0, 0) \approx 0$. Take disjoint $S \subset \mathbb{R}^F$ and $\bar{S} = \{-\theta, \theta \in S\}$ such that $P_\theta(S) > 0$ and $P_\theta(\bar{S}) > 0$. Recall that by our characterization

$$\tilde{\mathcal{E}}_\sigma(m, r, 0, 0) = \mathbb{E}_{\epsilon,\theta,\chi,\xi}\left[\sum_{\ell}^{L}\left(\delta_{\ell,\epsilon} - \sum_{h}^{H} v_h \sigma(\chi_h)_\ell\right)^2\right] \approx 0 \tag{145}$$

$$\chi_{h\ell} = \sum_{f}^{F} m_{hf}\chi_{f\ell}^* + \sum_{h'}^{H}(r^{1/2})_{hh'}\xi_{h'}\,, \tag{146}$$

where $v_h = \frac{1}{H}$ for softmax or some arbitrary constants for softmax-v attention. Since we assumed that the model is well-specified, we get $\sum_h^H v_h \sigma(\chi_h)_\ell \approx \delta_{\ell,\epsilon}$ almost surely, which means that it should hold for any $\{\chi_{f\ell}^*\}_{f\in[F],\ell\in[L]}$, because softmax is a continuous function. Notice that summing over $\ell \in [L]$ we get $\sum_h v_h = 1$.

Now, if there exists $\theta \in S$, such that the norm $\|\theta\|_2 = B_1$ for some constant B_1, the equation can not hold simultaneously for all realisations of the random variables $\chi_{f\ell}^*, \xi_{h\ell}$, therefore we would only consider the case of $\|\theta\|_2 = B$ where $B \to \infty$.

In this case, let's fix some $\theta \in S$, then the variables become split in three groups $H^+ = \{h : \sum_f m_{hf}\theta_f > 0\}$, $H^0 = \{h : \sum_f m_{hf}\theta_f = 0\}$ and $H^- = \{h : \sum_f m_{hf}\theta_f < 0\}$, and consequently applying softmax leads to

$$\sum_{h\in H^+} v_h \sigma(\chi_h)_\ell + \sum_{h\in H^0} v_h \sigma(\chi_h)_\ell + \sum_{h\in H^-} v_h \sigma(\chi_h)_\ell \tag{147}$$

$$= \sum_{h\in H^0} v_h \sigma(\chi_h)_\ell + \sum_{h\in H^-} v_h \sigma(\chi_h)_\ell = 0. \tag{148}$$

Let $h^- \in H^-, h^0 \in H^0$. Notice that $\sigma(\chi_{h^-})_\ell = \exp \chi_{h^-\ell} / \sum_{\ell'\neq\epsilon} \exp \chi_{h^-\ell'}$, while $\sigma(\chi_{h^0})_\ell = \exp \chi_{h^0\ell} / \left(\exp \chi_{h^0\epsilon} + \sum_{\ell'\neq\epsilon} \exp \chi_{h^0\ell'}\right)$. Since $\chi_{h^0\epsilon}$ is independent from $\chi_{h^0\ell}$ for $\ell \neq \epsilon$, $\sigma(\chi_{h^-})_\ell$

and $\sigma(\chi_{h^0})_\ell$ are linearly independent positive random variables and thus both sets of coefficients v_h for $h \in H^-$ and $h \in H^0$ should be 0. Applying the same reasoning to the case of flipped spike θ, we get that $v_h = 0$ for $h \in H^+$ as well. This contradicts with the fact that $\sum_h v_h = 1$.

Instead, the softmax-1 is well specified, because it allows a head not to return noise in case it is not well aligned with any token. Let's take $H = 2$ opposite heads $m_1 \in \mathbb{R}^F$ and $m_2 = -m_1$ and choose m_1 such that the hyperplane $m_1^\top \tilde{\theta} = 0$ has a null probability.

Notice that this can be always achieved for a discrete distribution or a continuous distribution as long as it has F-dimensional support. For a discreet distribution, a space cannot be covered by a finite union of hyperplanes and therefore we can pick m_1 from the space excluding the orthogonal hyperplanes of $\tilde{\theta}$ in the distribution domain. For a continuous distribution in the space $\mathbb{R}^F$ any hyperplane has zero probability and we can pick arbitrary m_1.

Then one has that $m_1^\top \tilde{\theta} = c_{\tilde{\theta}} > 0$ for some $c_{\tilde{\theta}}$ independent of B and consequently $m_1^\top \theta = B c_{\tilde{\theta}}$ has a large signal with a probability going to 1.

Take $b_1 = b_2 = B^{3/2}$ the biases of the softmax-1, $v = H$, and scale m_1 and m_2 as B. By symmetry assume that $m_1^\top \theta > 0$. Then one has the scalings

$$\chi_{1\epsilon} = m_1^\top \theta + \mathcal{O}(B) = \Theta(B^2) \qquad \chi_{1,\ell\neq\epsilon} = m_1^\top \chi_\ell^* + \mathcal{O}(1) = \Theta(B) \tag{149}$$

$$\chi_{2\epsilon} = m_2^\top \theta + \mathcal{O}(B) \ll -1 \qquad \chi_{2,\ell\neq\epsilon} = m_2^\top \chi_\ell^* + \mathcal{O}(1) = \Theta(B) \tag{150}$$

Consequently the attention scores are

$$\sigma(\chi_1)_\epsilon = H \frac{e^{\Theta(B^2)}}{e^{\Theta(B^{3/2})} + e^{\Theta(B^2)} + e^{\Theta(B)}} \to 1 \quad \sigma(\chi_1)_{\ell\neq\epsilon} = H \frac{e^{\Theta(B)}}{e^{\Theta(B^{3/2})} + e^{\Theta(B)}} \to 0 \quad \sigma(\chi_2)_\ell \to 0 \tag{151}$$

$$\frac{1}{H}\sum_h^H \sigma(\chi_h)_\epsilon \to 1 \qquad \frac{1}{H}\sum_h^H \sigma(\chi_h)_{\ell\neq\epsilon} \to 0 \tag{152}$$

and $\tilde{\mathcal{E}}_\sigma(m, r, b, v) = 0$.

□

B DYNAMICAL PHASE TRANSITION

We provide a heuristic argument and numerical evidence for the presence of the dynamical phase transition stated in Conjecture 3.1. The specialization dynamics is controlled by the escape from the unspecialized saddle point of the loss. Having $m(\tau^{\mathrm{u}})$ the alignment at the unspecialized saddle point, expanding m around it, according to Prop. 3.2 the dynamics is

$$\frac{\partial}{\partial \tau} m(\tau) = -\nabla_m \tilde{\mathcal{E}}_\sigma(m(\tau)) \tag{153}$$

$$= -\nabla^2_{m,m} \tilde{\mathcal{E}}_\sigma(m(\tau^{\mathrm{u}})) \cdot (m(\tau) - m(\tau^{\mathrm{u}})) \, . \tag{154}$$

We integrate it to obtain the exponential escape

$$m(\tau) = e^{-(\tau-\tau^{\mathrm{u}})\nabla^2_{m,m}\tilde{\mathcal{E}}_\sigma(m(\tau^{\mathrm{u}}))} \cdot m(\tau^{\mathrm{u}}) \, . \tag{155}$$

We now consider the eigendirection $s \in \mathbb{R}^{H\times F}$ of minus the Hessian $-\nabla^2_{m,m}\tilde{\mathcal{E}}_\sigma(m(\tau^{\mathrm{u}}))$ with most positive eigenvalue c. As we show in Lemma A.11, s corresponds to a direction of specialization of the heads. We call m_s the projection of m onto s. The dynamics is then, up to terms that take longer to grow

$$m_s(\tau) = e^{c(\tau-\tau^{\mathrm{u}})} m_s(\tau^{\mathrm{u}}) \, . \tag{156}$$

For $\delta > 0$ small enough (independently of D, η) we define the specialization time τ^{s} to be such that $||m_s(\tau^{\mathrm{s}})||_F = \delta$, i.e. the amplitude in the specialization direction is large enough. We have

$$\tau^{\mathrm{s}} = \tau^{\mathrm{u}} + \frac{1}{c}\left(\log\delta - \log||m_s(\tau^{\mathrm{u}})||_F\right) \, . \tag{157}$$

The initial specialized component m_s is a random variable that depends on the random initialization of $k^\star$. We consider the limits $D \to \infty$, and later $\eta \to 0$. Its amplitude scales like $m_s = D^{-1/2}\eta\tilde{m}_s \to 0$ with $\tilde{m}_s = \Theta_{D,\eta}(1)$. We rescale the time as $\tilde{\tau} = \tau \log(\sqrt{D}/\eta)^{-1}$. According to Prop. 3.3 the unspecialized time is constant $\tau^{\mathrm{u}} = \Theta_{D,\eta}(1)$. We obtain

$$\tilde{\tau}^{\mathrm{s}} = \frac{1}{c} + \frac{1}{c\log(\sqrt{D}/\eta)}\left(c\tau^{\mathrm{u}} + \log\delta - \log||\tilde{m}_s(\tau^{\mathrm{u}})||_F\right) \, . \tag{158}$$

This expression shows that $\tilde{\tau}^{\mathrm{s}}$ concentrates to a deterministic quantity $1/c$ and that the specialization transition is sharp at $\tilde{\tau}^{\mathrm{s}}$. Indeed, in the limit $D \to \infty$ we have that $\tilde{\tau}^{\mathrm{s}} \to 1/c$. c is a random quantity that depends on the realization of the data and of k^*; its variance is $1/\sqrt{D}$ and thus it concentrates. Moreover, $\tilde{\tau}^{\mathrm{s}}$ does not depend on δ to the leading order and the transition is thus sharp.

We can consider a more common scenario where the dataset is given and D is fixed. In this case, one can still take the limit of small initialization $\eta \to 0$. The same analysis leading to Eq. (158) holds. The transition is still sharp at $\tilde{\tau}^{\mathrm{s}} = 1/c$ but c now admits random fluctuations from one realization of the data to another. We illustrate this in Fig. 6 for the flipping spike at $F = H = 2$. We consider several independent realizations of the data and the SGD. We show that for various D, for η going to 0, the empirical means of $\tilde{\tau}^{\mathrm{s}}$ collapse to the same value as $\log(\sqrt{D}/\eta) \to 0$. The empirical variance of $\tilde{\tau}^{\mathrm{s}}$ decreases as $\log(\sqrt{D}/\eta) \to 0$ but goes to 0 only when $D \to \infty$.

We show that the insights of our analysis extend to semi-realistic data, considering sequences based on the MNIST detection task described in Appendix C. $D = \tilde{D} = 784$ is fixed and we take the limit of small initializations $\eta \to 0$. We define each feature/direction f to be the average digit f, for $f = 0, \dots, 9$. In Fig. 7 we see that the rescaled specialization times for each feature concentrate to a deterministic value, independent of the run and the initial condition, different for each f.

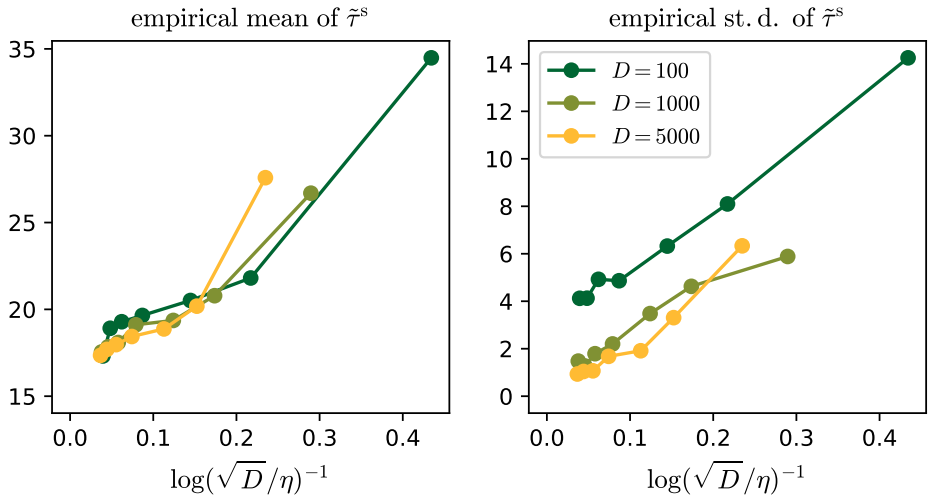


Figure 6: Phase transition: concentration of the specialization time $\tilde{\tau}^s$. Numerical simulations of SGD. We consider sequence length $L = 5$, $H = 2$ heads with σ softmax attention, $F = 2$ features, θ drawn from the flipping spike distribution, with signal strengths $\nu_1 = \nu_2 = 2$. The threshold to determine the specialization is $\delta = 0.6$. The means and variances are empirically computed over 100 independent runs for each point.

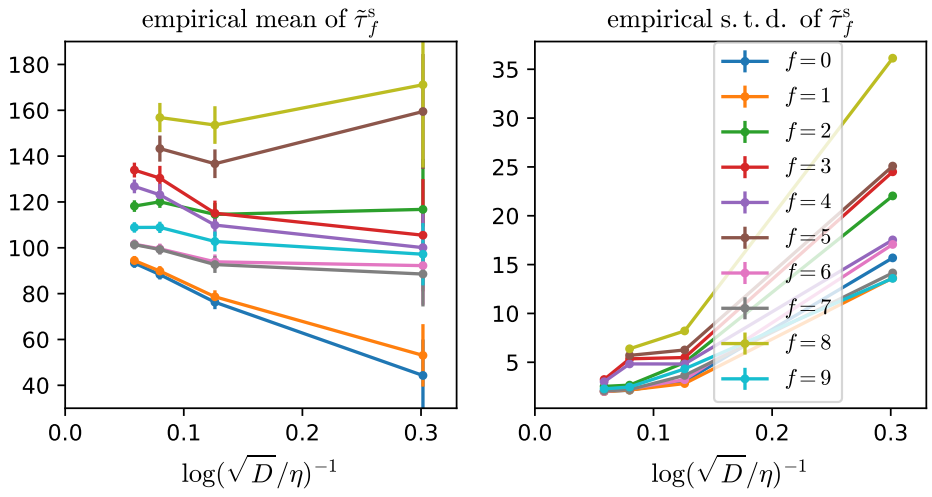


Figure 7: Phase transition: concentration of the specialization time $\tilde{\tau}_f \tau^s$. Numerical simulations of SGD on the MNIST detection task App. C. $H = 10$ heads with σ softmax attention. The threshold to determine the specialization time τ_f^s in each direction f is $||m_{:,f} - H^{-1} \sum_h^H m_{h,f} \mathbb{1}_H||_2 > 1$. The means and variances are empirically computed over 100 independent runs for each point.

C MNIST DIGITS DETECTION

To extend the scope of our theoretical predictions to the case of more complex data, we conducted experiments where the model is trained to detect a handwritten digit from a sequence of patches where the other patches are pure noise. For each sequence we take the relevant token (patch) to be uniformly sampled from the MNIST dataset Deng (2012). We denote the dataset $\mathcal{S} = \{\tilde{X}_\mu \in \mathbb{R}^{\tilde{D}}\}_{\mu\in[\tilde{N}]}$, $\tilde{N} = 10^4$, $\tilde{D} = 784$. The MNIST dataset is normalized in the following way: the values of the pixels are first rescaled as $\tilde{X}_\mu \mapsto \tilde{X}_\mu/255 \in [0, 1]$. The mean direction is then removed according to $\tilde{X}_\mu \mapsto \tilde{X}_\mu - \tilde{N}^{-1}\sum_{\mu'}^{\tilde{N}} \tilde{X}_{\mu'}$. Each sequence is constructed as

$$X_\ell \sim \mathcal{N}(\nu\delta_{\ell,\epsilon}\hat{k}, I_D)\ , \qquad \hat{k} \sim \mathrm{unif}(\mathcal{S})\ . \tag{159}$$

The number of patches is $L = 5$; the signal strength is $\nu = 0.3$. The attention model is the same as in the main eq. (3).

D NUMERICAL SIMULATION DETAILS

For the reparametrized population loss $\tilde{\mathcal{E}}$ estimation we discretize the flow with a step size $\delta = 0.02$. We use Monte Carlo integration with 10^5 samples to compute the expectations. We keep the same samples for each step of the gradient descent, which doesn't affect the behavior of the model as can be seen in Fig. 8.

For the initialization, when not comparing to SGD, we add initial noise $\mathcal{N}(0, 10^{-4})$ to m and r, which allows to break the initial symmetry in the parameters. When comparing to SGD we initialize m and r to their empirical values. For SGD and the loss $\mathcal{E}$, we take $N_b = D$ and learning rate $\gamma = 0.02$.

We provide the code to run our predictions in the supplementary material. Running one gradient descent takes a few minutes to half an hour and a few GBs on a local GPU. Overall we ran a thousand of descents.

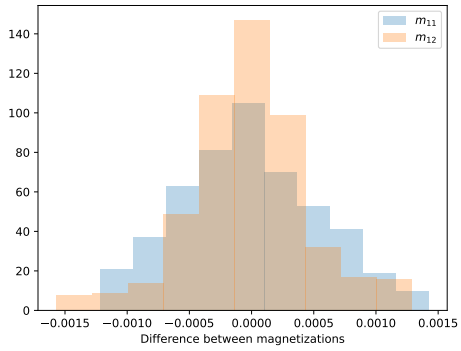


Figure 8: Histogram of the differences between magnetizations of the first head in the model trained with changed MC samples and the same MC samples at each step of the training. The data are sampled from the flipping spike distribution, $H = 2, F = 2, \nu_1 = \nu_2 = 2, L = 8$.

E SUPPLEMENTARY FIGURES

E.1 CHARACTERIZATION OF THE TRAINING DYNAMICS

We compare our theoretical prediction, Prop. 3.2, to the numerical simulation of SGD at finite large D. We consider the different activation functions, on the flipping spike or non-isotropic Gaussian.

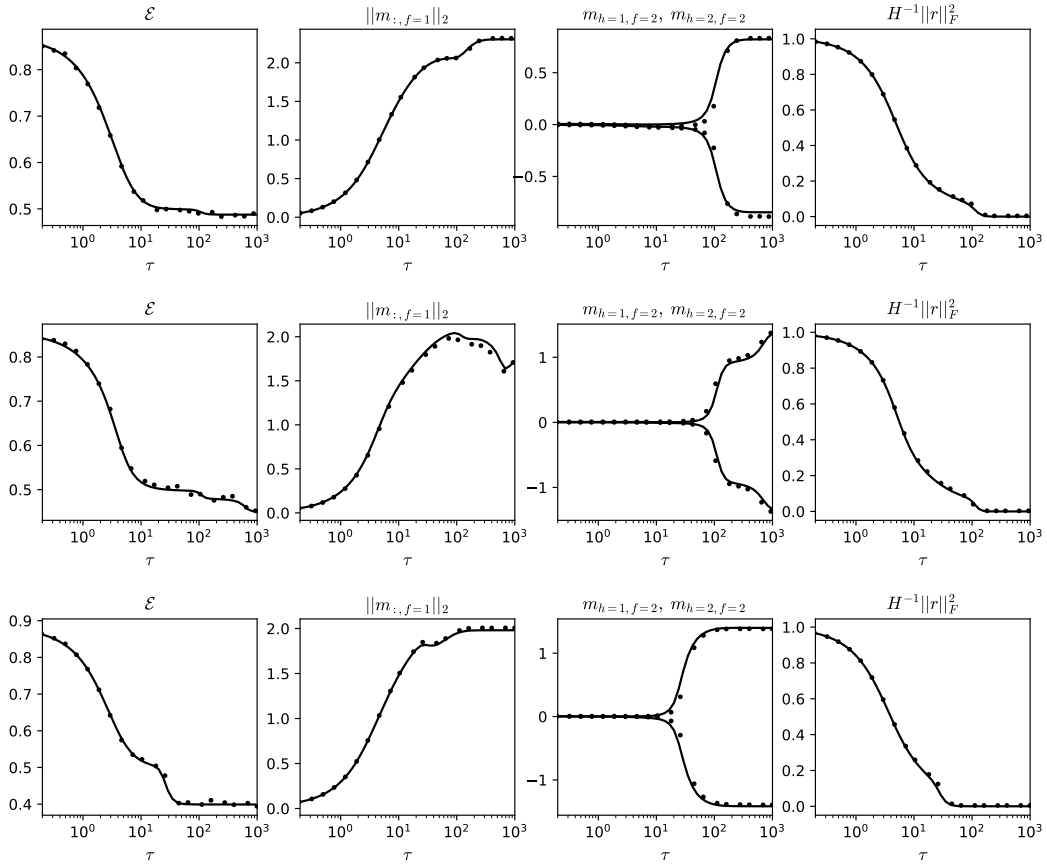


Figure 9: Asymptotic description of the attention trained by SGD. We compare numerical simulations at finite $D = 10^4$ (dots) and the theoretical description stated in proposition 3.2 (continuous lines). We consider sequence length $L = 5$, $H = 2$ heads, $F = 2$ features and θ distributed according to the flipping spike distribution, with signal strengths $\nu_1 = \nu_2 = 2$. Initialization $\eta = 1$. Top: softmax; middle: softmax-1; bottom: B-softmax.

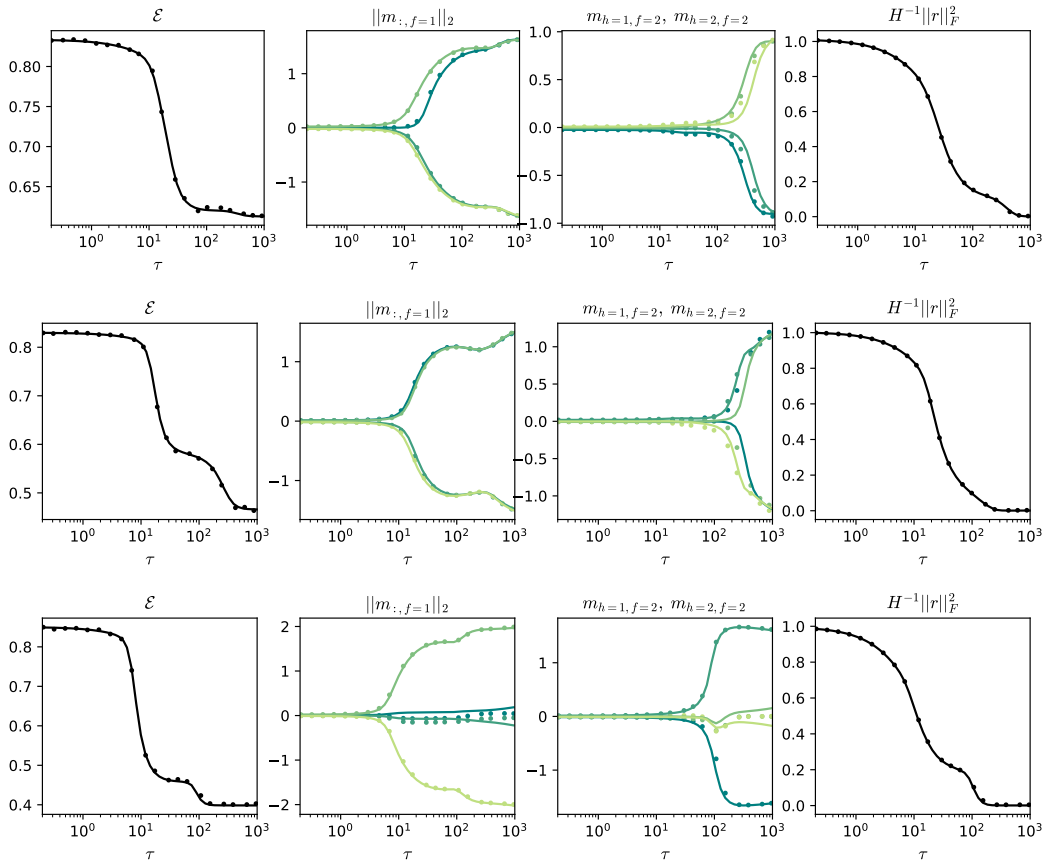


Figure 10: Asymptotic description of the attention trained by SGD. We compare numerical simulations at finite $D = 10^4$ (dots) and the theoretical description stated in proposition 3.2 (continuous lines). We consider sequence length $L = 5$, $H = 4$ heads, $F = 2$ features and θ distributed according to the non-isotropic Gaussian distribution, with signal strengths $\nu_1 = 8, \nu_2 = 2$. Initialization $\eta = 1$. Top: softmax; middle: softmax-1; bottom: B-softmax.

E.2 SPECIALIZATION OF THE HEADS

We provide an additional figure illustrating the specialization of the heads for the flipping spike distribution, depending on the signal strength ν_1 of the average direction and the signal strength ν_2 of the flipping-sign direction. The specialization of the two heads is measured as their cosine similarity and is reported in Fig. 11 for the softmax. It shows that the specialization grows monotonically with ν_1 and decreases monotonically with ν_2, as expected.

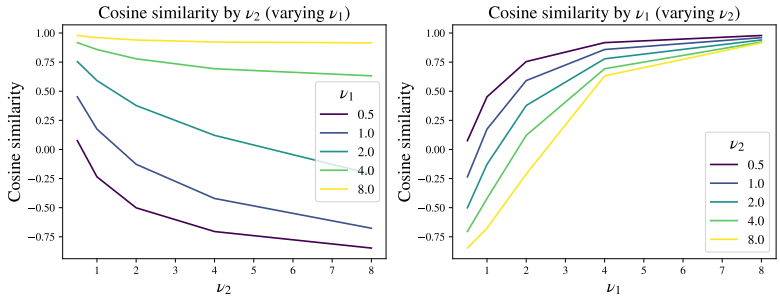


Figure 11: Cosine similarity between heads for softmax attention with $H = 2$, and flipping spike distribution with $F = 2$, $L = 4$ depending on the signal strengths for the constant and flipping-sign directions.

Figure 12 shows that when increasing the number of heads H, some heads specialize and diverge more from the average direction, more than at lower H.

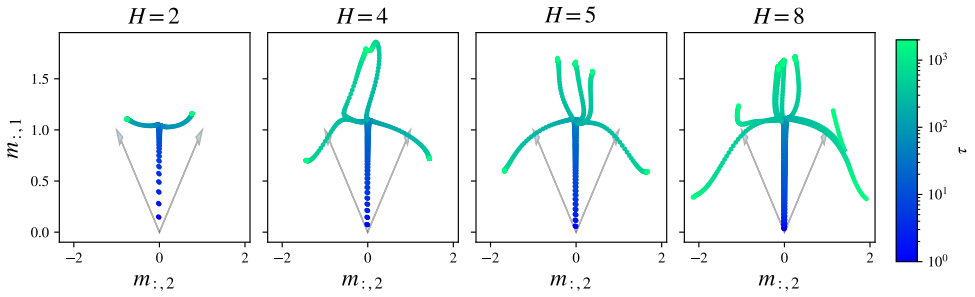


Figure 12: Evolution of the heads in the model of attention with varying number of heads $H \in \{2, 4, 5, 8\}$, and flipping spike distribution with $F = 2$, $L = 4$, with signal strengths $\nu_1 = \nu_2 = 2$. $\eta = 1$.

E.3 SEQUENTIAL SPECIALIZATION

We provide Figs. 13 and 14 that complement the description of the specialization dynamics of section 3, showing the specialization of the heads for an anisotropic Gaussian distribution for different H or the different activation functions. The three different runs in Fig. 14 show that the behaviour of the heads is similar across initializations.

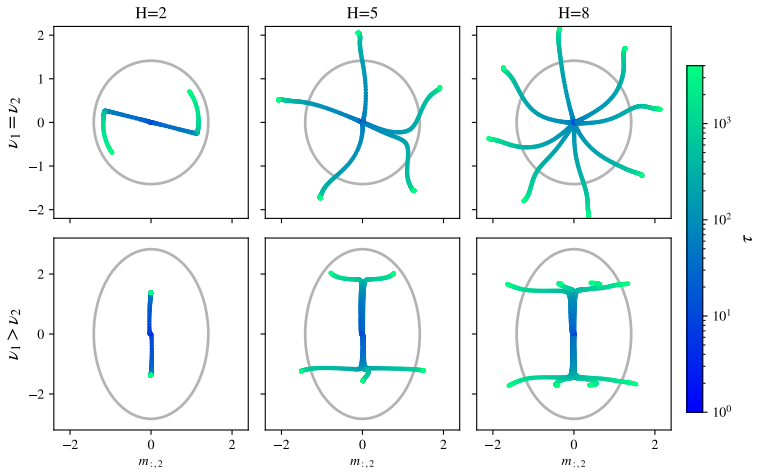


Figure 13: Evolution of the heads attention with varying number of heads $H \in \{2, 5, 8\}$, and non-isotropic Gaussian distribution with $F = 2$, $L = 4$, with signal strengths $\nu_1 = \nu_2 = 2$ (top) and $\nu_1 = 8, \nu_2 = 2$ (bottom). $\eta = 1$.

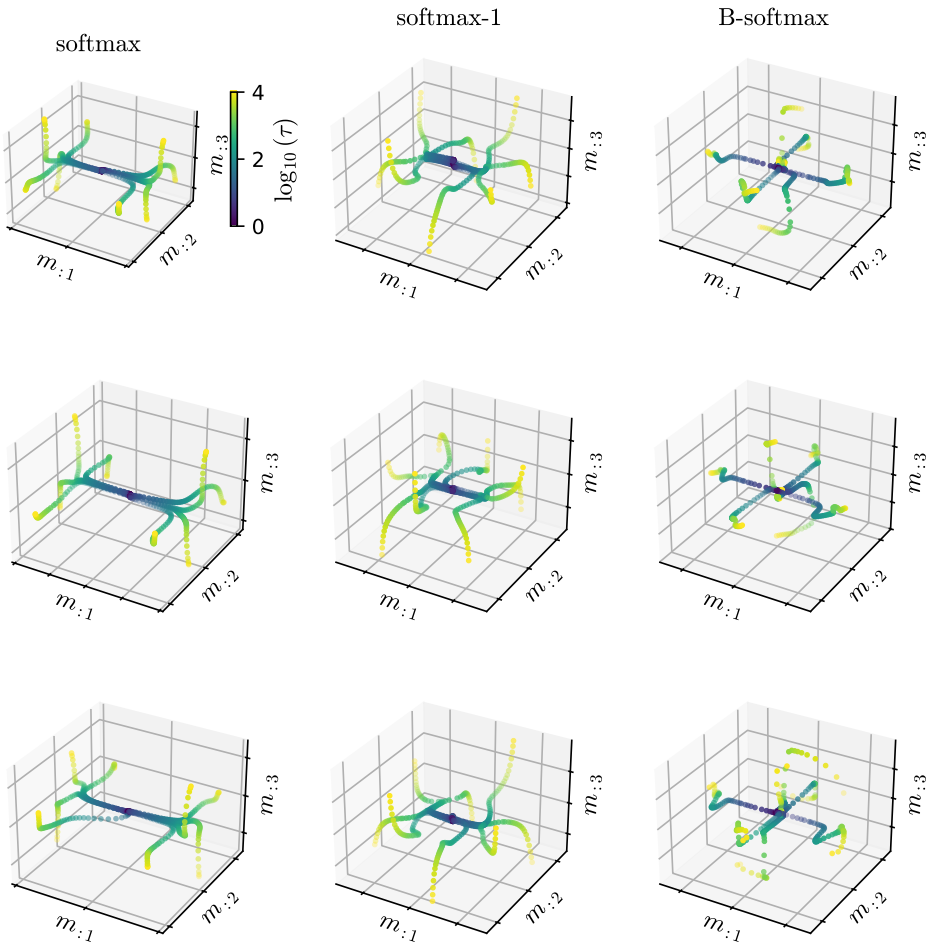


Figure 14: Evolution of the $H = 8$ heads for the non-isotropic Gaussian distribution at $F = 3$. The 2 lines are 2 different runs with different initial conditions.

E.4 Comparison of the different activations after training

In this section we provide further elements of comparison of the different activations after training, complementing the discussion of part 4.

The attention maps of the different attentions are shown in Fig. 15 for a few sequences drawn from the isotropic Gaussian distribution. The attention map of the softmax is noisy: the heads that do not focus on the relevant token have to focus on irrelevant ones, while the softmax-1 reduces attention scores of the "irrelevant" heads, and this effect is even more prominent for the B-softmax.

The softmax-1 and B-softmax rely on the multiple heads in a more optimal and specialized way. To see this, we perform a head pruning experiment and compare the importance of the heads. The pruning of the trained heads is done in a greedy manner and we uniformly rescale the output to keep constant the amplitude of the attention scores. The results of the experiment are reported on Fig. 16. Similarly to previous empirical studies Michel et al. (2019); Voita et al. (2019), we observe that a substantial number of heads can be removed without seriously affecting performance. In our model, in the case of a flipping spike distribution over F spikes or an isotropic Gaussian in dimension F, the number of heads $\tilde{H}$ that can be pruned without significant loss of performance is close to $H - F$; i.e. we can keep one head per each hidden feature approximately. If one prunes more heads $\tilde{H} > H - F$ and removes the heads that are actually necessary for the inference, the softmax-1 and the B-softmax behave differently from the standard softmax. Their performances degrade more severely than softmax and with a larger variance over the repeated runs. This suggests that attention with softmax-1 or B-softmax activation strongly relies on all the necessary heads together, and that these are strongly specialized.

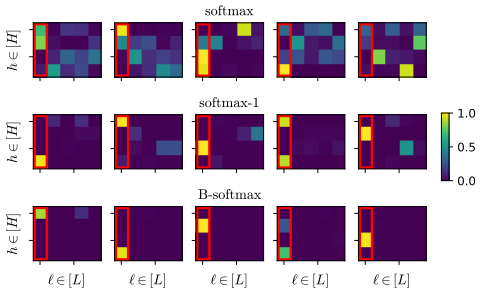


Figure 15: Attention maps of the different activation functions after training. $H = 4$ heads, $L = 5$ and the relevant token is highlighted in the red rectangle. We show the attention maps for five independent sequences. $F = 3$, signal isotropic Gaussian of strength $\nu = 9$.

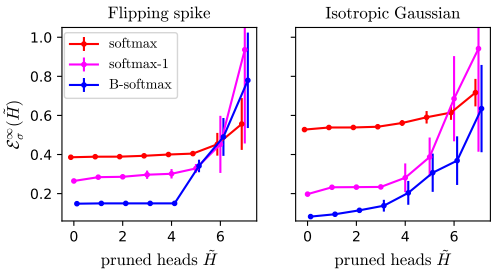


Figure 16: Pruning head experiment. Error $\mathcal{E}_\sigma^\infty(\tilde{H})$ of the different activation functions after training, after pruning $\tilde{H}$ heads over $H = 8$ total heads. $L = 5$, $F = 4$, $\nu = 10$. We performed five training runs with different initial conditions.