Specialization of softmax attention heads: insights from the high-dimensional single-location model.

Margarita Sagitova, Odilon Duranthon, Lenka Zdeborová
Statistical Physics of Computation Laboratory, EPFL

The problem

Goal: to understand head specialization in multi-head softmax attention.
Contributions: – A simple model for specialization;
– Characterization of the learning dynamics under population gradient flow;
– Mitigation of the variance induced by redundant heads, via alternative activation functions.

Probabilistic model of data

First, we sample **hidden spikes** (common for all sequences):

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

Then, for each sequence, we sample relevant token index $\epsilon \sim \text{Uniform}\{1, \dots, L\}$ and an effective **sequence-dependent direction**:

$$\hat{k} = \sum_f^F \theta_f k_f^* \quad \text{where } \theta \in \mathbb{R}^F, \quad \theta \sim P_\theta$$

The sequence X and the label y are defined as:

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

Attention model

A simplified model of sequence-to-token attention:

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

Learned parameters $k_h \in \mathbb{R}^D, b \in \mathbb{R}^H, v \in \mathbb{R}$.

Softmax. Standard activation in modern LLMs:

$$\sigma(\chi, b, v; h)_\ell = e^{\chi_{h\ell}} / \sum_{\ell'=1}^L e^{\chi_{h\ell'}}$$

Softmax-1. Leads to head deactivation when attention scores are low (can be viewed as adding special zero token to the sequence):

$$\sigma(\chi, b, v; h)_\ell = v e^{\chi_{h\ell}} / \left(e^{b_h} + \sum_{\ell'=1}^L e^{\chi_{h\ell'}}\right)$$

Bayes-softmax. Allows heads interaction, dynamically rescaling importance of heads:

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

Asymptotic characterization

We consider the limit $D \to \infty$ and $L, F, H = \Theta(1)$.
The population loss can be expressed in terms of the **order parameters**:

$$m_{hf} = (k_h)^\top k_f^*, \qquad q_{hh'} = (k_h)^\top k_{h'},$$
$$r = (q - mm^\top)^{1/2},$$

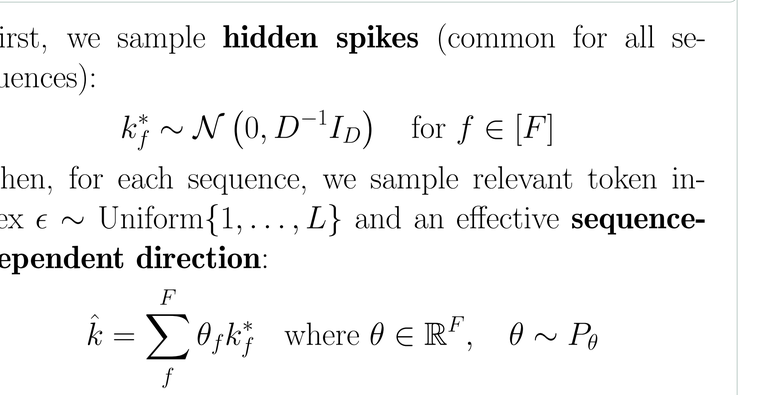
for $h, h' \in [H]$, $f, f' \in [F]$.

$$\mathcal{E}_\sigma(k, b, v) = \frac{1}{D}\mathbb{E}_{X,y}\left[||y - \hat{y}_{\sigma,k,b,v}(X)||_2^2\right]$$
$$\Downarrow$$
$$\tilde{\mathcal{E}}_\sigma(m, r, b, v)$$
$$= \mathbb{E}_{\epsilon,\theta,\chi}\left[\sum_\ell^L \left(\delta_{\ell,\epsilon} - \frac{1}{H}\sum_h^H \sigma(\chi, b, v; h)_\ell\right)^2\right]$$

Equivalence between high-D and low-D gradient flows on $\mathcal{E}_\sigma(k)$ or $\tilde{\mathcal{E}}_\sigma(m, r)$.

Unpecialized/specialized phases

In a first phase, the heads do not specialize and **move collectively towards the mean of the signal** in a few time steps $\tau = \Theta_D(1)$. The dynamics is controled by the gradient of the loss in the direction of the mean signal.



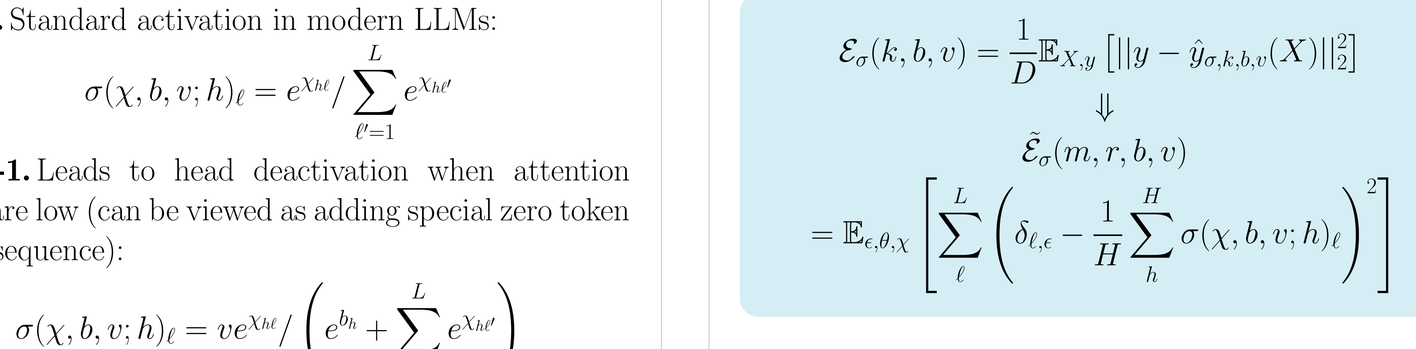


Fig. 2: Dots: numerical simulations at finite $D = 10^4$. Lines: theory. $H = F = 2$, σ softmax.

In a second phase, the heads **specialize in other directions**, in $\tau = \Theta_D(\log D)$ time steps. The escape dynamics from the unspecialized saddle is controled by the Hessian of the loss.

Head deactivation via alternative activations

When the weight distribution is not restricted to a half-space $\mathbb{R}^F$, the model with **softmax activation cannot achieve zero loss** for any value of k, while there is a model with **softmax-1 activation that reaches zero loss in the limit of large signal**.

Given the spikes $\{k_f^*\}_{f\in[F]} \in \mathbb{R}^{F\times D}$ and a sequence $X \in \mathbb{R}^{L\times D}$, the Bayes estimator of the label is

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

$$P(\epsilon = \ell | X, k^*) = \frac{\int_\theta \exp\left(\hat{k}(\theta)^T X_\ell\right) e^{-\frac{||\theta||_2^2}{2}} P_\theta(d\theta)}{\sum_{\ell'}^L \int_\theta \exp\left(\hat{k}(\theta)^T X_{\ell'}\right) e^{-\frac{||\theta||_2^2}{2}} P_\theta(d\theta)}$$

If the weight distribution has discrete support, **Bayes-softmax achieves the Bayes risk**.

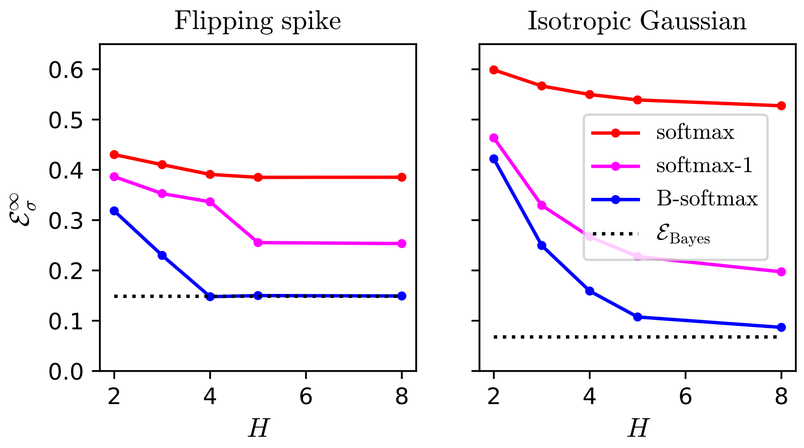


Fig.3. Error after training. $F = 4$.

Sequential specialization

The heads learn the features f sequentially, from the ones with the largest signal Var θ_f to the ones with smaller signal.

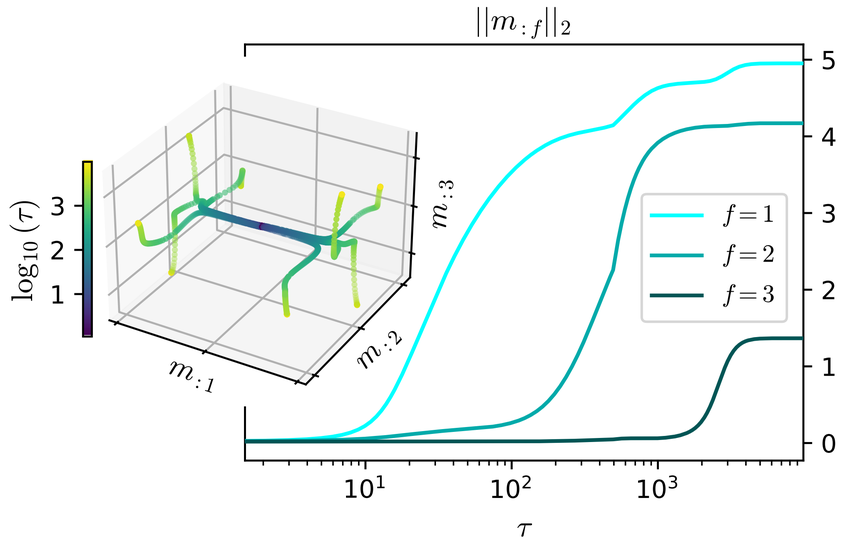


Fig. 4: Non-isotropic Gaussian distribution at $F = 3$. $H = 8$, σ softmax.

Available at arxiv:2603.03993

References

[1] Siyu Chen et al. "Training Dynamics of Multi-Head Softmax Attention for In-Context Learning: Emergence, Convergence, and Optimality". In: *Conference on Learning Theory (COLT)*. arXiv:2402.19442. 2024.
[2] O. Duranthon et al. "Statistical Advantage of Softmax Attention: Insights from Single-Location Regression". In: *The Fourteenth International Conference on Learning Representations (ICLR)*. arXiv:2509.21936. 2026.
[3] Pierre Marion et al. "Attention layers provably solve single-location regression". In: *The Thirteenth International Conference on Learning Representations (ICLR)*. arXiv:2410.01537. 2025.
[4] Yedi Zhang et al. "Training Dynamics of In-Context Learning in Linear Attention". In: *Proceedings of the 42th International Conference on Machine Learning (ICML)*. arXiv:2501.16265. 2025.