GRADIENT FLOW POLARIZES SOFTMAX OUTPUTS TOWARDS LOW-ENTROPY SOLUTIONS

Anonymous authors
Paper under double-blind review

ABSTRACT

Sparse attention patterns are prevalent in trained transformers, yet the mechanisms that favor their emergence during training remain poorly understood. In this work, we investigate the role of softmax in shaping these patterns by studying the gradient flow dynamics of an attention head trained to produce a fixed bias vector for all inputs, a task that dense and sparse attention solve equally well. Leveraging symmetries of this task, we show that the dynamics of the head simplify to a *value–softmax* model $\mathbf{V}\boldsymbol{\sigma}(\boldsymbol{a})$, where $\mathbf{V}$ and $\boldsymbol{a}$ are a learnable value matrix and attention vector. We prove that gradient flow on this model drives the attention scores toward low entropy: under logistic loss they converge to a one-hot vector, and under square loss we obtain quantitative polarization bounds. Finally, we use experiments in trained transformers to examine how the choice of attention nonlinearity affects the influence of individual tokens and sparse attention patterns.

1 INTRODUCTION

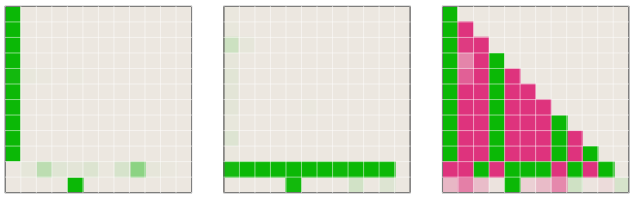
Despite the widespread adoption of large language models (LLMs), our understanding of the internal mechanisms by which they process information remains limited. In particular, the self-attention mechanism and the representations it induces play a central role in model behavior, yet many aspects of their functioning remain in the early stages of understanding. Recent work has begun to shed light on this opacity by identifying recurring learned structures within transformer architectures. Induction heads (Olsson et al., 2022) are a key example of capability-driving circuits, while attention sinks (Xiao et al., 2024) — where disproportionate attention weight concentrates on a small number of tokens — represent a distinct organizational phenomenon that has attracted significant interest.

Building on these efforts, identified components are often examined in isolation to assess how architectural design choices, training algorithms, and hyperparameters, as well as properties of the training data, contribute to their emergence. A theoretical understanding of these factors is a necessary step toward improving the robustness, efficiency, and safety of large language models. In this work, we pursue this objective by studying a broader phenomenon that underlies many of these components: *low-entropy, or sparse, attention patterns*, in which the attention distribution concentrates most of its mass on a small number of tokens.

An example of low-entropy attention is the attention sink phenomenon (Xiao et al., 2024), where a large fraction of the attention mass concentrates on a single token, often the first token in the sequence. One possible explanation is functional: attending strongly to a fixed token can produce a predictable contribution to the attention output, for instance acting as a default bias or offset. But this does not explain why the model realizes this function through a sparse attention pattern. The same contribution could, in principle, be produced by dense attention with appropriately chosen value vectors. Indeed, Figure 1 shows that different attention parameterizations can solve the same problem with very different levels of sparsity. This raises a broader question: *when low-entropy attention is not explicitly required by the task, why does it emerge? In particular, is sparsity a functional necessity, or an implicit preference induced by optimization and parameterization?*

We investigate this question through a simple task in which an attention head is trained to output a constant bias vector for all inputs. Using the symmetries of this task, the head's gradient-flow dynamics reduce exactly to those of a value–softmax model $\mathbf{V}\boldsymbol{\sigma}(\boldsymbol{a})$ for specific initial conditions. We therefore study a simplified value–softmax model, and analyze whether gradient-based training where $\mathbf{V}$ and $\boldsymbol{a}$ are jointly trained exhibits an implicit bias toward sparsification even when many dense

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.



(a) Softmax (b) Sigmoid (c) Linear

Figure 1: Representative attention patterns of the 2nd Transformer layer solving an induction task. Sigmoid and linear attentions are trained without additional normalization (see Table 1). An attention sink clearly emerges for softmax but not other attentions.

solutions exist. Concretely, we consider logistic and square losses, characterize their gradient-flow dynamics, and complement the theory with experiments in trained transformers.

Our contributions are:

- We identify a constant-bias task, on which the dynamics of a single attention head reduce exactly to the value–softmax model. For logistic loss, we show that, among many possible solutions, gradient flow is biased toward sparse solutions: the softmax output converges to a *one-hot* vector.
- We extend the analysis to regression, where polarization is dual to optimization effort: the sparsification rate is controlled by $\int \|\nabla \ell\|_1 \, ds$, implying that slower-converging problems, such as those with ill-conditioned data, yield sharper attention.
- We discuss the pitfalls of this sparsification bias, whereby a disproportionate share of the decision-making burden is assigned to a single token. We also provide empirical evidence showing how attention normalization affects the prevalence of attention sinks in an induction-head task.

1.1 Related Work

Our work connects the emergence of sparse attention patterns with the implicit bias of gradient-based training and the dynamics of jointly learned values and scores.

Sparse attention patterns. Softmax transformers often learn sparse attention patterns, whose study has driven progress in mechanistic interpretability (Olsson et al., 2022; Lindsey et al., 2025). Singh et al. (2024); Zucchet et al. (2025b); Yüksel et al. (2025); Varre et al. (2025) investigate the conditions under which these patterns emerge. Nanda et al. (2023) identify subcircuits for factual recall, and Zucchet et al. (2025a) analyze the evolution of attention on this task. In multi-head attention, Chen et al. (2024a;b) establish a complementary winner-take-all allocation of tasks under gradient flow for in-context multi-task linear regression, where polarization occurs across heads.

Attention sinks and massive activations. Xiao et al. (2024) observed that a large share of attention often falls on the first token, which they termed an attention sink. Subsequent work showed that sinks can also occur at other positions (Yu et al., 2024; Sun et al., 2024). Several explanations have been proposed: Sun et al. (2024) associate sinks with massive activations that provide implicit attention biases, while Xiao et al. (2024); Gu et al. (2025) emphasize the role of softmax normalization. Gu et al. (2025) also examine how sink formation depends on the data distribution, optimization, loss, and architecture. Zhang et al. (2025) describe a mechanism for clustering hidden representations, and Barbero et al. (2025) relate sinks to the prevention of representational collapse. Using a Markovian data model, Guo et al. (2024a) show that a head can attend to sinks on some input domains but not others. Attention sinks have also motivated practical improvements in key–value cache compression (Guo et al., 2024b), long-context generation (Xiao et al., 2024; Yang et al., 2025), quantization (Son et al., 2024), and feature map smoothness (Darcet et al., 2024).

Softmax alternatives and entropy collapse. Alternatives to softmax include sparsemax (Martins & Astudillo, 2016), linear attention (Katharopoulos et al., 2020; Qin et al., 2022; Wortsman et al., 2023), and sigmoid attention (Ramapuram et al., 2025), with computational efficiency and parallelizability motivating many of these designs. Softmax produces substantially lower attention entropy than linear alternatives (Hong & Lee, 2025), but the consequences depend on the setting. Zhai et al. (2023) link entropy collapse to training instability and develop methods to prevent it, whereas Zhang et al. (2024) design an attention operation that preserves useful low-entropy patterns. To explain localization,

Deng et al. (2024) study the effect of logits growing logarithmically with sequence length, and Bao et al. (2024) relate attention localization to the eigenspectrum of query–key matrices.

Implicit bias and polarization dynamics. Closely related to our analysis, Sheen et al. (2024) study implicit regularization under gradient flow in a one-layer softmax attention model and establish convergence toward sparse solutions. Our analysis focuses on how the joint evolution of values and scores reinforces attention ordering and quantifies polarization. The resulting equations also resemble replicator dynamics in evolutionary game theory (Taylor & Jonker, 1978; Smith, 1982; Sorin, 2020) and models of opinion polarization (Proskurnikov et al., 2015). In replicator dynamics, growth depends on fitness relative to the population average. A related comparison appears in the softmax score dynamics; here, the signal being compared is itself shaped by learning the values.

2 Problem Setup

Notation. We write $[n] = \{0, \dots, n-1\}$, X_i and X^j for a matrix's i-th column and j-th row, and $\|\cdot\|$ for the Euclidean or Frobenius norm. We use $\operatorname{diag}(a)$ for a diagonal matrix, $\mathbf{1}$ for an all-ones vector, and $u \odot v$ for the elementwise product. For a vector $a \in \mathbb{R}^m$, softmax is defined coordinatewise by

$$\boldsymbol{\sigma}(a)_i = \frac{\exp(a_i)}{\sum_{j=0}^{m-1} \exp(a_j)}, \qquad a \in \mathbb{R}^m.$$

Inputs and an attention head. Let $x_1, \dots, x_T \in \mathbb{R}^d$ be semantic tokens. At position i, we concatenate x_i with a one-hot positional encoding $e_i \in \mathbb{R}^{T+1}$ to obtain $z_i = (x_i, e_i)$. A separate query token $q = (x_q, e_q)$ attends to this context, with $e_q = e_{T+1}$. Writing $Z \in \mathbb{R}^{T \times (d+T+1)}$ for the matrix with rows $z_i^\top$, the head outputs

$$\widehat{\boldsymbol{\beta}} = W_V Z^\top \boldsymbol{\sigma}(Z W_A q), \tag{1}$$

where $W_V \in \mathbb{R}^{p \times (d+T+1)}$ is the value map and $W_A = QK^\top \in \mathbb{R}^{(d+T+1)\times(d+T+1)}$ is the attention matrix. We train W_V and the merged matrix W_A directly; separate query and key factors induce a different parameter flow (Corollary B.2).

A constant-bias task. We study a task in which the head is trained toward the same target $\boldsymbol{\beta}_* \in \mathbb{R}^p$ for all inputs. The semantic tokens $x_1, \dots, x_T, x_q$ are sampled independently from $N(0, I_d)$, and the population objective is $\mathbb{E}\,\ell(\widehat{\boldsymbol{\beta}}, \boldsymbol{\beta}_*)$. We allow any loss differentiable in its first argument, and write $\nabla\ell$ for that gradient. Our main examples are square loss, $\ell(\beta, \boldsymbol{\beta}_*) = \frac{1}{2}\|\beta - \boldsymbol{\beta}_*\|^2$, and logistic loss, $\ell(\beta, \boldsymbol{\beta}_*) = \log(1 + \exp(-\langle \boldsymbol{\beta}_*, \beta \rangle))$, which encourages a large margin along the fixed direction $\boldsymbol{\beta}_*$. Both dense and sparse attention can produce an input-independent output. This task therefore allows us to ask whether training favors sparse attention even when sparsity is unnecessary for fitting the target.

Symmetry of the task. Let $X \in \mathbb{R}^{T \times d}$ have rows $x_i^\top$. Splitting the parameter matrices into semantic and positional blocks gives

$$\widehat{\boldsymbol{\beta}} = \big(\mathbf{V} + W X^\top\big)\boldsymbol{\sigma}\big(X W_{xx} x_q + W_{px} x_q + X w_p + \boldsymbol{a}\big). \tag{2}$$

Here $\mathbf{V} \in \mathbb{R}^{p \times T}$ supplies the positional values and $\boldsymbol{a} \in \mathbb{R}^T$ supplies the positional scores for the query at e_q. The remaining blocks are $W \in \mathbb{R}^{p \times d}$, $W_{xx} \in \mathbb{R}^{d \times d}$, $W_{px} \in \mathbb{R}^{T \times d}$, and $w_p \in \mathbb{R}^d$. The term $W x_i$ contributes the value of semantic token x_i, while $x_i^\top W_{xx} x_q$ measures its learned semantic correlation with the query. The terms $W_{px} x_q$ and $X w_p$ couple semantics to position.

Because the target is independent of the semantic tokens, the head can solve the task using only its positional parameters. Define

$$\mathcal{N} = \{W = W_{xx} = W_{px} = w_p = 0\}.$$

On this set, the output is $\mathbf{V}\boldsymbol{\sigma}(\boldsymbol{a})$ for every input. Moreover, centering of the semantic tokens and independence of the query make the population gradients of all semantic blocks vanish on $\mathcal{N}$. Thus a head initialized on $\mathcal{N}$ remains there throughout training.

Proposition 2.1 (Exact reduction for a general loss)**.** *Assume the population gradients exist and differentiation commutes with expectation. The set $\mathcal{N}$ is invariant under gradient flow of $\mathbb{E}\,\ell(\widehat{\boldsymbol{\beta}}, \boldsymbol{\beta}_*)$. On $\mathcal{N}$, with $\boldsymbol{s} = \boldsymbol{\sigma}(\boldsymbol{a})$ and $P_{\boldsymbol{s}} = \operatorname{diag}(()\,\boldsymbol{s}) - \boldsymbol{s}\boldsymbol{s}^\top$, the flow is exactly*

$$\dot{\mathbf{V}} = -\nabla\ell(\mathbf{V}\boldsymbol{s}, \boldsymbol{\beta}_*)\boldsymbol{s}^\top, \qquad \dot{\boldsymbol{a}} = -P_{\boldsymbol{s}} \mathbf{V}^\top \nabla\ell(\mathbf{V}\boldsymbol{s}, \boldsymbol{\beta}_*). \tag{3}$$

Appendix B.2 proves the reduction and studies initialization near $\mathcal{N}$. For square loss, semantic blocks initialized at size ε remain of order ε over any fixed finite horizon, while their effect on the positional trajectory is $O(\varepsilon^2)$ (Theorem B.5). The same appendix gives a first-order expansion for logistic loss.

The value–softmax model. The reduction leads to the model $\boldsymbol{\beta} = \mathbf{V}\boldsymbol{\sigma}(\boldsymbol{a})$, with trainable value matrix $\mathbf{V}$ and score vector $\boldsymbol{a}$. For the polarization analysis, we take $T = p$, so $\mathbf{V} \in \mathbb{R}^{p\times p}$ and $\boldsymbol{a} \in \mathbb{R}^p$, and write $\mathcal{L}(\mathbf{V}, \boldsymbol{a}) = \ell(\mathbf{V}\boldsymbol{\sigma}(\boldsymbol{a}), \boldsymbol{\beta}_*)$. As in other re-parameterized models (Li et al., 2022) and the unconstrained features model (Mixon et al., 2020; Zhu et al., 2021), treating these quantities as free parameters makes the training dynamics tractable. Here they describe the attention head exactly on $\mathcal{N}$. Many pairs $(\mathbf{V}, \boldsymbol{a})$ yield the same output: most correspond to dense attention patterns, while some can be sparse.

Gradient flow. We study which representations are selected by continuous-time gradient descent:

$$\dot{\mathbf{V}} = -\nabla_{\mathbf{V}}\mathcal{L}(\mathbf{V}, \boldsymbol{a}), \qquad \dot{\boldsymbol{a}} = -\nabla_{\boldsymbol{a}}\mathcal{L}(\mathbf{V}, \boldsymbol{a}). \tag{4}$$

The chain rule yields (3). Its softmax Jacobian $P_{\boldsymbol{s}}$ compares each value-dependent signal with its attention-weighted mean, coupling the evolution of the scores. Section 3 examines how this interaction shapes the attention weights as the values and scores are trained jointly.

3 Polarizing Effect of Softmax

Jointly training values and attention scores creates a feedback: coordinates receiving more attention learn faster, and the resulting differences in the values influence how attention is allocated. We first examine this feedback under logistic loss, where it leads to a one-hot limit. We then establish a polarization bound for general losses with sign-preserving gradients, apply it to square-loss regression, and discuss the role of the attention nonlinearity.

3.1 Logistic loss

We first take the logistic loss along the fixed direction $\boldsymbol{\beta}_* \in \mathbb{R}^p$ of the constant-bias task in Section 2. As a function of $\mathbf{V}$ and $\boldsymbol{a}$, it reads

$$\mathcal{L}(\mathbf{V}, \boldsymbol{a}) = \log\left(1 + \exp\langle \boldsymbol{\beta}_*, -\mathbf{V}\boldsymbol{\sigma}(\boldsymbol{a})\rangle\right), \tag{5}$$

The gradient flow dynamics are given by

$$\mathrm{d}\mathbf{V} = \gamma(\boldsymbol{\beta})\,\boldsymbol{\beta}_*\boldsymbol{s}^\top \mathrm{d}t, \qquad \mathrm{d}\boldsymbol{a} = \gamma(\boldsymbol{\beta})\,\left[\mathrm{diag}(\boldsymbol{s}) - \boldsymbol{s}\boldsymbol{s}^\top\right]\mathbf{V}^\top\boldsymbol{\beta}_*\mathrm{d}t,$$

where $\gamma(\boldsymbol{\beta}) = (1 + \exp\langle \boldsymbol{\beta}_*, \boldsymbol{\beta}\rangle)^{-1}$. Letting $\boldsymbol{u} = \mathbf{V}^\top\boldsymbol{\beta}_*$, the dynamics reduce to a coupled system over two vectors $\boldsymbol{a}, \boldsymbol{u} \in \mathbb{R}^p$

$$\dot{\boldsymbol{u}} = \gamma(\boldsymbol{\beta})\,\|\boldsymbol{\beta}_*\|^2\boldsymbol{s}, \qquad \dot{\boldsymbol{a}} = \gamma(\boldsymbol{\beta})\,\left[\mathrm{diag}(\boldsymbol{s}) - \boldsymbol{s}\boldsymbol{s}^\top\right]\boldsymbol{u}. \tag{6}$$

How values and scores reinforce one another. The two equations in (6) describe complementary effects. A coordinate receiving more attention has a faster-growing value projection, since $\dot{u}_j = \gamma(\boldsymbol{\beta})\|\boldsymbol{\beta}_*\|^2 s_j$. In turn, each score responds to the corresponding projection relative to its attention-weighted average:

$$\dot{a}_j = \gamma(\boldsymbol{\beta})s_j(u_j - \bar{u}), \qquad \bar{u} = \langle \boldsymbol{u}, \boldsymbol{s}\rangle.$$

Thus a projection above the average raises its score, while one below the average lowers it. This comparison with the mean is reminiscent of replicator dynamics (Sorin, 2020). Here the projections themselves are learned: attention shapes the values, and the values shape attention. The following results show how this feedback amplifies an initial difference between the projections.

Repulsion between the coordinates. We now show that the dynamics induce repulsion between the coordinates of the attention scores $\boldsymbol{s}$ and the value projections $\boldsymbol{u}$ under the following initialization.

Assumption 3.1. The parameters $\mathbf{V}, \boldsymbol{a}$ are initialized such that $\boldsymbol{a}(0) = 0$ and $\boldsymbol{u}_0(0) > \boldsymbol{u}_1(0) > \ldots > \boldsymbol{u}_{p-1}(0)$ for the projection $\boldsymbol{u} = \mathbf{V}^\top\boldsymbol{\beta}_*$.

The scores start uniformly, while differences between the initial value projections break the symmetry. Distinct projections can always be indexed in this order; for a nonzero target direction, they are distinct almost surely under an initialization with a joint density.

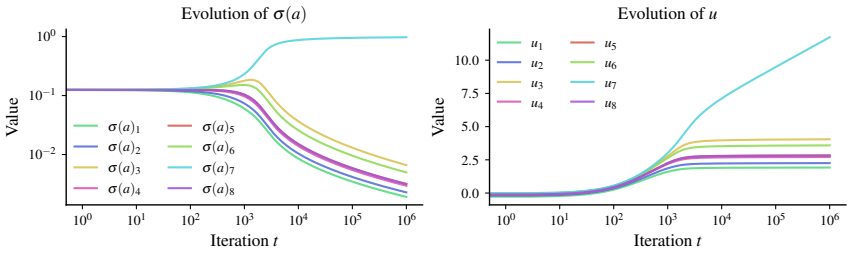


Figure 2: Experimental verification of Theorem 3.3 using logistic loss. The plot shows the evolution of attention scores $\sigma(a)$ and value projections u over time. As predicted, the attention scores converge to a one-hot vector (blue line goes to 1, others to 0) and the value projections preserve their order with maximal coordinate diverging.

Theorem 3.2. *Consider gradient flow on the loss (5) under Assumption 3.1. Then, for all $t > 0$,*

(a) ***Order preservation.*** *The order of the projections $\boldsymbol{u}_i(t)$ and attention scores $\boldsymbol{s}_i(t)$ is preserved*

$$\boldsymbol{u}_0(t) > \boldsymbol{u}_1(t) > \ldots > \boldsymbol{u}_{p-1}(t), \qquad \boldsymbol{s}_0(t) > \boldsymbol{s}_1(t) > \ldots > \boldsymbol{s}_{p-1}(t).$$

(b) ***Repulsion.*** *The difference between projections $\boldsymbol{u}$ grows over time: for any $i < j$ and any $t > 0$,*

$$\frac{d}{dt}\left(\boldsymbol{u}_i(t) - \boldsymbol{u}_j(t)\right) > 0.$$

The two orderings arise together. Initially, the larger projections produce larger score derivatives; once the attention weights are ordered, the larger weights make the corresponding projections grow faster. To formalize this feedback, the proof tracks $\Phi_{ij} = (e^{-a_j} - e^{-a_i})(u_i - u_j)$ for $i < j$. This potential starts at zero and becomes strictly positive, preventing either ordering from reversing. The increasing projection gaps then provide the separation needed to quantify polarization.

Sparsity of the attention scores. We now quantify the polarization and its limit.

Theorem 3.3. *Consider gradient flow for the loss (5) under Assumption 3.1. Let $\delta = \min_{i<j}\left(\boldsymbol{u}_i(0) - \boldsymbol{u}_j(0)\right)$. Then we have*

(a) ***Vanishing loss.*** $\lim_{t\to\infty} L(\boldsymbol{\beta}(t)) = 0$.

(b) ***Decay of non-maximal scores.*** *For any $j \neq 0$ and any $t > 0$,*

$$\frac{\boldsymbol{s}_j(t)}{\boldsymbol{s}_0(t)} \leqslant \left(1 + {}^{\delta}\!/_{p} \int_0^t \gamma(\boldsymbol{\beta}(s))ds\right)^{-1}.$$

(c) ***One-hot limit.*** *The polarization coefficient $\int_0^t \gamma(\boldsymbol{\beta}(s))ds = \Theta(\log t)$, and consequently the attention scores converge to a one-hot vector, i.e.,*

$$\lim_{t\to\infty} \boldsymbol{s}_0(t) = 1, \quad \textit{and} \quad \lim_{t\to\infty} \boldsymbol{s}_j(t) = 0, \quad \textit{for } j \neq 0.$$

The integral $\int_0^t \gamma(\boldsymbol{\beta}(h))\,\mathrm{d}h$ measures how much the feedback accumulates during training. Although the loss and the instantaneous gradient vanish, this integral diverges, so every non-maximal attention weight becomes negligible relative to the leading one (Figure 2). In the representation $\boldsymbol{\beta} = \sum_j s_j \boldsymbol{v}_j$, the mixing weights therefore approach a vertex of the simplex. The sparsity conclusion concerns these weights; the values continue to evolve. Appendix C compares this behavior with linear networks.

3.2 Beyond logistic loss

For logistic loss, the output gradient always points along $\boldsymbol{\beta}_*$, allowing us to summarize the values by a single projection $\boldsymbol{u}$. A general loss need not have a fixed gradient direction. Nevertheless, the same feedback operates row by row whenever the gradient coordinates retain their signs.

Let $g(t) = \nabla_{\boldsymbol{\beta}} \ell(\boldsymbol{\beta}(t), \boldsymbol{\beta}_*)$, with $\boldsymbol{\beta} = \mathbf{V}\boldsymbol{s}$, and write $\boldsymbol{v}^i$ for row i of $\mathbf{V}$. We consider

$$\dot{\mathbf{V}} = -g\boldsymbol{s}^\top, \qquad \dot{\boldsymbol{a}} = -P(\boldsymbol{s})\mathbf{V}^\top g, \qquad P(\boldsymbol{s}) = \operatorname{diag}(\boldsymbol{s}) - \boldsymbol{s}\boldsymbol{s}^\top, \tag{7}$$

These are the gradient-flow equations for the value–softmax model.

The rows follow the scores, and the scores follow the rows. Suppose g_i keeps its initial sign, and orient row i in its direction of motion by setting $\widetilde{v}^{\,i} = \epsilon_i \boldsymbol{v}^i$, where $\epsilon_i = -\operatorname{sign}(g_i(0))$. Then

$$\dot{\widetilde{v}}^{\,i} = |g_i| \boldsymbol{s}.$$

Each signed row moves in the direction of the attention vector: an entry receiving more attention grows faster. Starting from $\mathbf{V}(0) = 0$, an ordering of the scores therefore induces the same ordering in every active signed row. Conversely, the signal driving the scores is

$$w = -\mathbf{V}^\top g = \sum_i |g_i| \widetilde{v}^{\,i}, \qquad \dot{a}_j = s_j\big(w_j - \langle \boldsymbol{s}, w \rangle\big).$$

It is a nonnegative combination of those rows, so their ordering is passed back to the scores. This mutual reinforcement preserves the order and reduces the attention assigned to each coordinate relative to the leading one. The next theorem quantifies the reduction through the cumulative ℓ_1 norm of the output gradient.

Theorem 3.4 (Polarization under sign-preserving gradients)**.** *Consider (7) for a continuously differentiable loss, with $\mathbf{V}(0) = 0$ and $a_0(0) > \cdots > a_{p-1}(0)$. Let $I = \{i : g_i(0) \neq 0\}$ be nonempty. Assume that on $[0, T]$ each g_i, $i \in I$, retains its initial sign (zeros are allowed), and $g_i \equiv 0$ for $i \notin I$. Set $\epsilon_i = -\operatorname{sign}(g_i(0))$ and $q_l = s_l/s_0$. Then:*

(a) ***Order preservation.*** *For $0 < t \leqslant T$, the scores and every active signed row satisfy*

$$a_0(t) > \cdots > a_{p-1}(t), \qquad \epsilon_i \mathbf{V}_{i0}(t) > \cdots > \epsilon_i \mathbf{V}_{i,p-1}(t) \quad (i \in I).$$

(b) ***Polarization.*** *Each q_l, $l \geqslant 1$, is nonincreasing on $[0, T]$. For any $0 < t_0 \leqslant T$, define the positive row separation*

$$d_l(t_0) = \min_{i \in I} \epsilon_i\big(\mathbf{V}_{i0}(t_0) - \mathbf{V}_{il}(t_0)\big) > 0.$$

For $t_0 \leqslant t \leqslant T$,

$$\frac{s_l(t)}{s_0(t)} \leqslant \left[\frac{s_0(t_0)}{s_l(t_0)} + \frac{d_l(t_0)}{p} \int_{t_0}^{t} \big\|\nabla_{\boldsymbol{\beta}} \ell(\boldsymbol{\beta}(h), \boldsymbol{\beta}_*)\big\|_1 \, \mathrm{d}h\right]^{-1}. \tag{8}$$

The positive time t_0 lets the initially equal value entries separate. Thereafter, their separation is bounded below, and the total gradient magnitude controls how much further the attention ratios decrease. If the sign condition holds for all time and the integral diverges, the attention converges to a one-hot vector. A finite integral instead gives a quantitative polarization bound. The proof in Appendix D.2 follows the coupled row and score gaps described above.

Square-loss regression. Let $X \in \mathbb{R}^{n \times p}$ have full column rank and let $X\boldsymbol{\beta}_*$ be the target. The loss is

$$\mathcal{L}(\mathbf{V}, \boldsymbol{a}) = \tfrac{1}{2} \|X\boldsymbol{\beta}_* - X\mathbf{V}\boldsymbol{\sigma}(\boldsymbol{a})\|^2. \tag{9}$$

Let $\lambda_{\min} > 0$ be the smallest eigenvalue of $X^\top X$.

Assumption 3.5. $\mathbf{V}(0) = 0$ and $a_0(0) > a_1(0) > \cdots > a_{p-1}(0)$.

Let τ be the first time an initially nonzero residual coordinate changes sign (in the spectral coordinates defined in Appendix D.3), with $\tau = \infty$ if none does, and write $q_l = s_l/s_0$.

Corollary 3.6. *Consider gradient flow for (9) under Assumption 3.5, with X of full column rank and $\boldsymbol{\beta}_* \neq 0$.*

(a) ***Optimization.*** *For all $t \geqslant 0$,*

$$\mathcal{L}(\mathbf{V}(t), \boldsymbol{a}(t)) \leqslant \mathcal{L}(\mathbf{V}(0), \boldsymbol{a}(0)) \exp\left(-\frac{2\lambda_{\min}}{p} t\right).$$

(b) ***Eigenvalue dependence.*** *If $\tau = \infty$ and the target $X\boldsymbol{\beta}_*$ has a component of magnitude $m > 0$ along a unit left singular vector of X associated with singular value $\sqrt{\lambda_{\min}}$, then*

$$q_l(\infty) \leqslant \left[\frac{1}{q_l(0)} + \frac{(1 - q_l(0))m^2}{2p^2 \lambda_{\min}}\right]^{-1}.$$

In particular, for fixed initial scores and m bounded away from zero, $q_l(\infty) = O(\lambda_{\min})$.

Polarization–optimization duality. The residual drives both the reduction of the loss and the separation of attention weights: polarization accumulates through $\int \|\boldsymbol{r}(h)\|_1 \, \mathrm{d}h$ while the fit improves. A small eigenvalue slows learning in the corresponding direction, requiring more accumulated residual to fit a target component of a given magnitude. Part (b) quantifies the resulting stronger polarization. The $O(\lambda_{\min})$ conclusion applies when this target component remains bounded away from zero. Unlike logistic loss, the regression residual is integrable: the scores converge to finite limits, yielding polarized but strictly positive attention weights. See Appendix D.3 for the proof and Figure 9 for numerical examples.

For two score coordinates, a separate invariant proves global order without a residual sign condition.

Proposition 3.7 (Two-coordinate regression)**.** *Under Assumption 3.5 with $p = 2$ and X of full column rank,*

$$a_0(t) - a_1(t) \geqslant a_0(0) - a_1(0) > 0 \qquad \textit{for all } t \geqslant 0.$$

Thus the score order persists globally, even beyond the interval of sign-preserving residuals.

The proof uses a conserved relation between the value-column difference and the score gap (Appendix D.4).

3.3 Other attention maps

We now ask which parts of the polarization mechanism depend on the softmax map. We first consider replacing softmax with elementwise nonlinearities such as the sigmoid activation $\sigma_{\mathrm{sig}}(a)_i = 1/1+\exp(-a_i)$ or ReLU. For both regression and logistic losses, we do not observe the same polarization under sigmoid or ReLU (Figure 10 in Appendix H.1). Intuitively, polarization is driven by the mean-centering term in the softmax Jacobian, and elementwise nonlinearities induce no analogous "fitness-minus-average-fitness" interaction across coordinates.

We next consider alternative *normalization* maps, which retain this interaction structure. Let f be monotonically increasing, write $Z_f(\boldsymbol{a}) = \sum_{j=1}^{p} f(\boldsymbol{a}_j)$, and define $\boldsymbol{\sigma}_f : \mathbb{R}^p \to \mathbb{R}^p$ by $\boldsymbol{\sigma}_f(\boldsymbol{a})_i = f(\boldsymbol{a}_i)\, Z_f(\boldsymbol{a})^{-1}$. Writing $\boldsymbol{s}_f := \boldsymbol{\sigma}_f(\boldsymbol{a})$, the gradient-flow dynamics for the logistic loss become

$$\mathrm{d}\boldsymbol{u} = \gamma(\boldsymbol{\beta})\, \|\boldsymbol{\beta}_*\|^2 \boldsymbol{s}_f \, \mathrm{d}t, \qquad \mathrm{d}\boldsymbol{a} = \gamma(\boldsymbol{\beta})\, Z_f(\boldsymbol{a})^{-1}\, f'(\boldsymbol{a}) \odot (\boldsymbol{u} - \langle \boldsymbol{u}, \boldsymbol{s}_f \rangle \, \mathbf{1}) \, \mathrm{d}t.$$

This retains the replicator-type structure, and an analogue of Theorem 3.2 can be established. The polarization strength now depends on f and f'. For instance, with $f(x) = x$, which does not enforce positivity, the polarization is typically insufficient to yield one-hot scores; with $f(x) = x^2$, $\boldsymbol{s}_f$ is driven toward a one-hot limit (Figure 10).

The same score–value coupling appears in both losses, but it leads to different conclusions. In the logistic model, the driving gradient persists long enough for one softmax coordinate to capture all mass. In square-loss regression, the residual decays to zero and the scores approach a finite limit; our quantitative polarization statement applies while the residual coordinates keep their signs. This distinction matters when using the reduced model to interpret trained attention heads.

4 Implications of Polarization for Token Influence and Attention Sparsity

In this section, we discuss how softmax-induced sparsity leads to an imbalance in the influence of individual tokens on the model's predictions, and how it is consistent with attention sinks and the associated massive activations.

4.1 Imbalance in token influence

Classification with self-attention. One major implication of our results is the emergence of tokens that exert a disproportionate influence on the model's output. When the attention weights converge to a one-hot vector, as in Theorem 3.3, the output effectively ignores all context except the single max-logit token. Consequently, the model's predictions become highly sensitive to perturbations of that specific token.

We construct a simple classification dataset as follows. We partition a vocabulary V into K disjoint sets $V^{(1)}, \dots, V^{(K)}$ of equal size, and an input of class k is a sequence of n tokens sampled from $V^{(k)}$. As the sets are disjoint, any one of the input tokens unambiguously predicts the label, so any attention map allows one to solve the classification problem.

Due to the low-entropy bias of softmax, however, we expect the attention weights to concentrate on a few tokens, effectively ignoring the others (Figure 15, Appendix H.3). We test this in an adversarial out-of-distribution setting, where we are allowed to change one token of the input with the goal of flipping the classifier's prediction. Figure 3a reports the success rates. As expected, the predictions of the softmax model are easily reversed; the effect gradually vanishes as more heads and layers are added to the model, since different heads, although each of them is sparse, concentrate their attention on distinct tokens.

4.2 Attention sinks and massive activations

Attention sinks. Attention sinks (Xiao et al., 2024) are attention patterns in which the maximal attention falls on a specific token, often a special token such as BOS or a delimiter or punctuation token. The bias hypothesis recalled in the introduction does not by itself explain why a single token is chosen: a pattern distributed over a set of tokens, for example a mixture spread across BOS and a few delimiter or punctuation tokens, would create a similar bias (Sun et al., 2024; Zhang et al., 2025; Gu et al., 2025). Yet empirically, sinks concentrate on a single token rather than spreading across these functionally equivalent candidates, which is exactly the behavior the polarization theorem predicts.

To see why optimization favors sinks, consider the value–softmax model with several queries, $\boldsymbol{\beta} = \boldsymbol{\sigma}(\mathbf{A})\mathbf{V}$, where $\mathbf{A} \in \mathbb{R}^{T\times T}$ is the attention score matrix (softmax applied row-wise) and $\mathbf{V} \in \mathbb{R}^{T\times d}$ are the value vectors. We optimize $(\mathbf{V}, \mathbf{A})$ under the per-position logistic loss $\mathcal{L}(\mathbf{V}, \mathbf{A}) = \frac{1}{T}\sum_t \ell(\boldsymbol{\beta}[t], \boldsymbol{\beta}_*)$, so that each position t produces the same target $\boldsymbol{\beta}_* \in \mathbb{R}^d$. Starting from $\mathbf{A}(0) = 0$ and a strictly ordered $\boldsymbol{u}(0) = \mathbf{V}(0)\boldsymbol{\beta}_*$, as in Assumption 3.1, all rows receive the same gradient and remain equal. After rescaling time by T, their common flow is the logistic value–softmax flow with $\|\boldsymbol{\beta}_*\|^2$ replaced by $T\|\boldsymbol{\beta}_*\|^2$, so Theorem 3.3 applies: the softmax weights in every row converge to the same one-hot vector, yielding an attention sink. Attention sinks can therefore emerge as a consequence of the optimization dynamics induced by the softmax. Which position becomes the sink, however, is decided here by the initialization, through the largest coordinate of $\boldsymbol{u}(0)$. The setting thus leaves out a second important aspect of attention sinks: in real data, sequences share fixed tokens such as BOS or delimiters, and sinks consistently form on particular ones among them. Explaining which of these shared tokens is preferred is left for future work.

Massive activations. Sun et al. (2024) report that attention sinks are also associated with massive activations in certain feature dimensions, i.e., regimes in which a small fraction of activations take values significantly larger than the rest. A reformulation of the value–softmax viewpoint provides intuition for how this phenomenon can arise from the same polarizing dynamics. For an intermediate residual-state matrix $R \in \mathbb{R}^{T\times d}$ and last-token residual state r, the attention update is

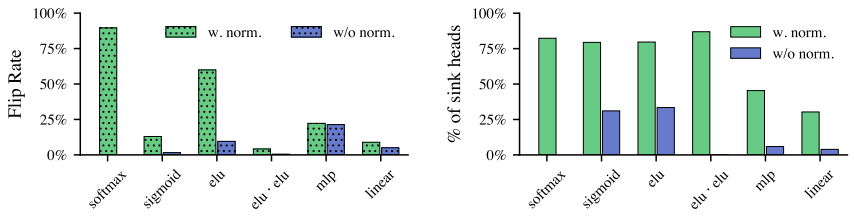
$$R_+ = \boldsymbol{\sigma}\left(RQK^\top R^\top\right) RV, \qquad r_+ = V^\top R^\top \boldsymbol{\sigma}(RKQ^\top r).$$

Here, the product is between the residual-state matrix and the softmax output, as in the value–softmax model, except that the softmax input is no longer free. This motivates the tied formulation $\mathcal{L}(\mathbf{R}, \boldsymbol{a}) = \ell\big(\mathbf{R}\boldsymbol{\sigma}(\mathbf{R}^\top \boldsymbol{a}), \boldsymbol{\beta}_*\big)$, where ℓ is the logistic loss. The following proposition formalizes this behavior for the tied parameterization.

Proposition 4.1. *Consider gradient flow on $\mathcal{L}(\mathbf{R}, \boldsymbol{a}) = \ell(\mathbf{R}\boldsymbol{\sigma}(\mathbf{R}^\top \boldsymbol{a}), \boldsymbol{\beta}_*)$ under the logistic loss, with $\|\boldsymbol{\beta}_*\| = 1$. Set $\boldsymbol{v} \overset{\text{def}}{=} \mathbf{R}^\top \boldsymbol{\beta}_*$ and $c \overset{\text{def}}{=} \langle \boldsymbol{\beta}_*, \boldsymbol{a}\rangle$, and decompose $\mathbf{R} = \boldsymbol{\beta}_* \boldsymbol{v}^\top + \mathbf{R}_\perp$, $\boldsymbol{a} = c\boldsymbol{\beta}_* + \boldsymbol{a}_\perp$ into components parallel and orthogonal to $\boldsymbol{\beta}_*$. If $\mathbf{R}_\perp(0) = 0$, $\boldsymbol{a}_\perp(0) = 0$, $c(0) > 0$, and $\boldsymbol{v}(0)$ has a unique maximizer at coordinate 0, then $\boldsymbol{s}(t) \to e_0$ as $t \to \infty$.*

The proof is given in Appendix F, which also covers rank-one perturbations of the orthogonal components. Figure 8 (Appendix H.1) shows the resulting polarization in both the softmax scores and the residual-state embeddings, suggesting that, under the tied parameterization, attention sinks and massive activations emerge from the same polarizing dynamics.

Induction heads acting as attention sinks. To test whether the results proven for the value–softmax model are reflected in more general transformer setups, we run experiments on more realistic



(a) (b)

Figure 3: Token influence and attention sinks, compared across attention functions. **(a)** Success rates of one-token adversarial flips in a one-layer, one-head classifier. **(b)** Proportion of sink heads in the second layer of transformers trained on the induction task.

transformer models (details in Appendix G). First, we create a simple setup where attention sinks emerge. We train two-layer transformers with a variable number of attention heads on an induction task, a simple variant of the Bigram-Backcopy task of Guo et al. (2024a) and Bietti et al. (2024). Each input sequence begins with tokens sampled uniformly at random and later repeats a token whose earlier successor identifies the next token to predict, which can be solved by a construction known as an induction head (Olsson et al., 2022). An example input is shown below, with the repeated token and its successor marked.

BOS 4 2 6 12 8 11 12* 8

At the positions where tokens are sampled at random, induction heads are not needed and can act as a fixed attention bias, consistent with the emergence of attention sinks described above. On this task, we evaluate the formation of attention sinks and its dependence on the attention functions listed in Table 1. We call an attention head a *sink head* if, at the positions where tokens are sampled uniformly at random, it assigns more than 90% of its total attention weight to the first token, and we report the average proportion of such heads in Figure 3b. In line with Section 3.3, the normalized positive attention variants have a consistently higher proportion of sinks than their unnormalized counterparts or the non-positive variants (mlp, linear).

Pretrained LLMs. We further verify the toy-setup observations on real-world LLMs by comparing softmax and unnormalized-sigmoid 7B-parameter models from Ramapuram et al. (2025), matched on hyperparameters and compute. Running a forward pass on random Pile samples (Gao et al., 2020) and measuring the sparsity of each head as the average attention weight on its max-logit token, Figure 20 (Appendix H.4) shows that softmax heads are substantially sparser, in line with the findings of Gu et al. (2025) on 1B-parameter models.

5 Conclusion

A simple regression example captures the question behind this work: setting every column of $\mathbf{V}$ equal to $\boldsymbol{\beta}_*$ gives $\mathbf{V}\boldsymbol{\sigma}(\boldsymbol{a}) = \boldsymbol{\beta}_*$ for any attention vector. Dense attention can therefore fit the target just as well as sparse attention. Yet training the values and scores jointly can favor polarized solutions. We explain this preference through their mutual reinforcement: attention determines how quickly the values learn, and the resulting value differences reinforce the attention ordering.

Starting from a head trained toward the same target for all inputs, we derive an exact value–softmax model on an invariant set and analyze this feedback under gradient flow. It leads to a one-hot limit for logistic loss and quantitative polarization bounds for regression under a residual sign condition. The experiments examine the same question in trained transformers, showing how the choice of attention nonlinearity affects sparse attention patterns and the influence of individual tokens. Understanding how this preference interacts with input-dependent targets and multiple heads remains a natural next step.

AI use statement

Generative AI tools contributed to some of the technical results of this paper. For the polarization result for general losses (Theorem 3.4), the authors developed the key ideas, namely the ℓ_1-norm control of the gradient and the observations underlying the ordering lemmas, and provided them to an AI assistant, which helped substantially in formulating the theorem and writing its proof. The

perturbation analysis of the value–softmax model near the invariant set (Theorem B.5) was likewise developed by the authors in interaction with an AI assistant. AI tools were also used for writing and editing the manuscript, for producing illustrations, and for implementing experiments. The authors are responsible for all technical results and for the final content of the paper.

REFERENCES

Han Bao, Ryuichiro Hataya, and Ryo Karakida. Self-attention networks localize when qk-eigenspectrum concentrates. In *International Conference on Machine Learning*, pp. 2903–2922. PMLR, 2024.

Federico Barbero, Alvaro Arroyo, Xiangming Gu, Christos Perivolaropoulos, Michael Bronstein, Petar Veličković, and Razvan Pascanu. Why do llms attend to the first token? *Conference on Language Modeling*, 2025.

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

Siyu Chen, Heejune Sheen, Tianhao Wang, and Zhuoran Yang. Training dynamics of multi-head softmax attention for in-context learning: Emergence, convergence, and optimality (extended abstract). In *Annual Conference Computational Learning Theory*, 2024a.

Siyu Chen, Heejune Sheen, Tianhao Wang, and Zhuoran Yang. Training dynamics of multi-head softmax attention for in-context learning: Emergence, convergence, and optimality. *arXiv preprint arXiv:2402.19442*, 2024b.

Timothée Darcet, Maxime Oquab, Julien Mairal, and Piotr Bojanowski. Vision transformers need registers. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=2dnO3LLiJ1.

Yichuan Deng, Zhao Song, Jing Xiong, and Chiwun Yang. How sparse attention approximates exact attention? your attention is naturally n^C-sparse. *arXiv preprint arXiv:2404.02690*, 2024.

Leo Gao, Stella Biderman, Sid Black, Laurence Golding, Travis Hoppe, Charles Foster, Jason Phang, Horace He, Anish Thite, Noa Nabeshima, et al. The pile: An 800gb dataset of diverse text for language modeling. *arXiv preprint arXiv:2101.00027*, 2020.

Xiangming Gu, Tianyu Pang, Chao Du, Qian Liu, Fengzhuo Zhang, Cunxiao Du, Ye Wang, and Min Lin. When attention sink emerges in language models: An empirical view. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=78Nn4QJTEN.

Tianyu Guo, Druv Pai, Yu Bai, Jiantao Jiao, Michael I Jordan, and Song Mei. Active-dormant attention heads: Mechanistically demystifying extreme-token phenomena in llms. *arXiv preprint arXiv:2410.13835*, 2024a.

Zhiyu Guo, Hidetaka Kamigaito, and Taro Watanabe. Attention score is not all you need for token importance indicator in KV cache reduction: Value also matters. In Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen (eds.), *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 21158–21166, Miami, Florida, USA, November 2024b. Association for Computational Linguistics. doi: 10.18653/v1/2024.emnlp-main.1178. URL https://aclanthology.org/2024.emnlp-main.1178/.

Jonghyun Hong and Sungyoon Lee. Variance sensitivity induces attention entropy collapse and instability in transformers. In Christos Christodoulopoulos, Tanmoy Chakraborty, Carolyn Rose, and Violet Peng (eds.), *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 8360–8378, Suzhou, China, November 2025. Association for Computational Linguistics. ISBN 979-8-89176-332-6. doi: 10.18653/v1/2025.emnlp-main.421. URL https://aclanthology.org/2025.emnlp-main.421/.

Ziwei Ji and Matus Telgarsky. Gradient descent aligns the layers of deep linear networks. In *International Conference on Learning Representations*, 2019. URL https://openreview.net/forum?id=HJflg30qKX.

Angelos Katharopoulos, Apoorv Vyas, Nikolaos Pappas, and François Fleuret. Transformers are RNNs: Fast autoregressive transformers with linear attention. In Hal Daumé III and Aarti Singh (eds.), *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pp. 5156–5165. PMLR, 13–18 Jul 2020. URL https://proceedings.mlr.press/v119/katharopoulos20a.html.

Zhiyuan Li, Tianhao Wang, Jason D. Lee, and Sanjeev Arora. Implicit bias of gradient descent on reparametrized models: on equivalence to mirror descent. In *Proceedings of the 36th International Conference on Neural Information Processing Systems*, NIPS '22, Red Hook, NY, USA, 2022. Curran Associates Inc. ISBN 9781713871088.

Jack Lindsey, Wes Gurnee, Emmanuel Ameisen, Brian Chen, Adam Pearce, Nicholas L. Turner, Craig Citro, David Abrahams, Shan Carter, Basil Hosmer, Jonathan Marcus, Michael Sklar, Adly Templeton, Trenton Bricken, Callum McDougall, Hoagy Cunningham, Thomas Henighan, Adam Jermyn, Andy Jones, Andrew Persic, Zhenyi Qi, T. Ben Thompson, Sam Zimmerman, Kelley Rivoire, Thomas Conerly, Chris Olah, and Joshua Batson. On the biology of a large language model. *Transformer Circuits Thread*, 2025. URL https://transformer-circuits.pub/2025/attribution-graphs/biology.html.

Andre Martins and Ramon Astudillo. From softmax to sparsemax: A sparse model of attention and multi-label classification. In *International conference on machine learning*, pp. 1614–1623. PMLR, 2016.

Dustin G. Mixon, Hans Parshall, and Jianzong Pi. Neural collapse with unconstrained features. *arXiv preprint arXiv:2011.11619*, 2020.

Neel Nanda, Senthooran Rajamanoharan, János Kramár, and Rohin Shah. Fact finding: Attempting to reverse-engineer factual recall on the neuron level, December 2023. URL https://www.lesswrong.com/posts/iGuwZTHWb6DFY3sKB/fact-finding-attempting-to-reverse-engineer-factual-recall. LessWrong.

Catherine Olsson, Nelson Elhage, Neel Nanda, Nicholas Joseph, Nova DasSarma, Tom Henighan, Ben Mann, Amanda Askell, Yuntao Bai, Anna Chen, et al. In-context learning and induction heads. *CoRR*, 2022.

Anton V Proskurnikov, Alexey S Matveev, and Ming Cao. Opinion dynamics in social networks with hostile camps: Consensus vs. polarization. *IEEE Transactions on Automatic Control*, 61(6): 1524–1536, 2015.

Zhen Qin, Weixuan Sun, Hui Deng, Dongxu Li, Yunshen Wei, Baohong Lv, Junjie Yan, Lingpeng Kong, and Yiran Zhong. cosformer: Rethinking softmax in attention. In *International Conference on Learning Representations*, 2022.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. Language models are unsupervised multitask learners. 2019.

Jason Ramapuram, Federico Danieli, Eeshan Gunesh Dhekane, Floris Weers, Dan Busbridge, Pierre Ablin, Tatiana Likhomanenko, Jagrit Digani, Zijin Gu, Amitis Shidani, et al. Theory, analysis, and best practices for sigmoid self-attention. In *The Thirteenth International Conference on Learning Representations*, 2025.

Heejune Sheen, Siyu Chen, Tianhao Wang, and Harrison H Zhou. Implicit regularization of gradient flow on one-layer softmax attention. *arXiv preprint arXiv:2403.08699*, 2024.

Aaditya K Singh, Ted Moskovitz, Felix Hill, Stephanie CY Chan, and Andrew M Saxe. What needs to go right for an induction head? a mechanistic study of in-context learning circuits and their formation. In *International Conference on Machine Learning*, pp. 45637–45662. PMLR, 2024.

John Maynard Smith. Evolution and the theory of games. In *Did Darwin get it right? Essays on games, sex and evolution*, pp. 202–215. Springer, 1982.

Seungwoo Son, Wonpyo Park, Woohyun Han, Kyuyeun Kim, and Jaeho Lee. Prefixing attention sinks can mitigate activation outliers for large language model quantization. In Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen (eds.), *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 2242–2252, Miami, Florida, USA, November 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.emnlp-main.134. URL https://aclanthology.org/2024.emnlp-main.134/.

Sylvain Sorin. Replicator dynamics: Old and new. *Journal of Dynamics & Games*, 7(4), 2020.

Mingjie Sun, Xinlei Chen, J Zico Kolter, and Zhuang Liu. Massive activations in large language models. In *First Conference on Language Modeling*, 2024. URL https://openreview.net/forum?id=F7aAhfitX6.

Peter D Taylor and Leo B Jonker. Evolutionary stable strategies and game dynamics. *Mathematical biosciences*, 40(1-2):145–156, 1978.

Aditya Varre, Gizem Yüce, and Nicolas Flammarion. Learning in-context n-grams with transformers: Sub-n-grams are near-stationary points. *arXiv preprint arXiv:2508.12837*, 2025.

Mitchell Wortsman, Jaehoon Lee, Justin Gilmer, and Simon Kornblith. Replacing softmax with relu in vision transformers. *arXiv preprint arXiv:2309.08586*, 2023.

Guangxuan Xiao, Yuandong Tian, Beidi Chen, Song Han, and Mike Lewis. Efficient streaming language models with attention sinks. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=NG7sS51zVF.

Shuai Yang, Yuying Ge, Yang Li, Yukang Chen, Yixiao Ge, Ying Shan, and Ying-Cong Chen. Seed-story: Multimodal long story generation with large language model. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV) Workshops*, pp. 1850–1860, October 2025.

Zhongzhi Yu, Zheng Wang, Yonggan Fu, Huihong Shi, Khalid Shaikh, and Yingyan Celine Lin. Unveiling and harnessing hidden attention sinks: Enhancing large language models without training through attention calibration. In *Forty-first International Conference on Machine Learning*, 2024. URL https://openreview.net/forum?id=DLTjFFiuUJ.

Oğuz Kaan Yüksel, Rodrigo Alvarez Lucendo, and Nicolas Flammarion. Incremental learning of sparse attention patterns in transformers. In *EurIPS 2025 Workshop on Principles of Generative Modeling (PriGM)*, 2025.

Shuangfei Zhai, Tatiana Likhomanenko, Etai Littwin, Dan Busbridge, Jason Ramapuram, Yizhe Zhang, Jiatao Gu, and Josh M. Susskind. Stabilizing transformer training by preventing attention entropy collapse. In *ICML*, 2023. URL https://arxiv.org/abs/2303.06296.

Michael Zhang, Kush Bhatia, Hermann Kumbong, and Christopher Re. The hedgehog & the porcupine: Expressive linear attentions with softmax mimicry. In *The Twelfth International Conference on Learning Representations*, 2024.

Stephen Zhang, Mustafa Khan, and Vardan Papyan. Attention sinks: A 'catch, tag, release' mechanism for embeddings. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025. URL https://openreview.net/forum?id=r8UWp9JeJi.

Zhihui Zhu, Tianyu Ding, Jinxin Zhou, Xiao Li, Chong You, Jeremias Sulam, and Qing Qu. A geometric analysis of neural collapse with unconstrained features. In *Advances in Neural Information Processing Systems*, volume 34, pp. 29820–29834, 2021.

Nicolas Zucchet, Jörg Bornschein, Stephanie CY Chan, Andrew K Lampinen, Razvan Pascanu, and Soham De. How do language models learn facts? dynamics, curricula and hallucinations. *CoRR*, 2025a.

Nicolas Zucchet, Francesco d'Angelo, Andrew K Lampinen, and Stephanie CY Chan. The emergence of sparse attention: impact of data distribution and benefits of repetition. *arXiv preprint arXiv:2505.17863*, 2025b.

Organization of the appendix. Section A fixes notation and collects the identities for the value–softmax flow that are used throughout: the gradients, the shift invariance of the scores, the replicator form of the score dynamics, and the evolution of the output. Section B proves the exact reduction of a single attention head to the value–softmax model (Proposition 2.1) and the stability of this reduction (Theorem B.5). Section C proves the logistic-loss results (Theorems 3.2 and 3.3). Section D treats general losses and square-loss regression: an ordered-fitness criterion, the proof of Theorem 3.4, and the proofs of Corollary 3.6 and Proposition 3.7. Section E extends the no-crossing property to other normalizations, and Section F analyzes the tied model of Proposition 4.1. Sections G and H contain the experimental details and additional experimental results.

A Preliminaries on the Value–Softmax Model

Standing notation. Indices start at 0, so $[p] = \{0, \dots, p-1\}$. For a matrix $\mathbf{V} \in \mathbb{R}^{p\times p}$ we write $\mathbf{V}_j$ for its j-th column, $\boldsymbol{v}^i$ for its i-th row, and $\mathbf{V}_{ij}$ for its entries. We always write

$$\boldsymbol{s} = \boldsymbol{\sigma}(\boldsymbol{a}) \in \Delta^{p-1}, \qquad \boldsymbol{\beta} = \mathbf{V}\boldsymbol{s}, \qquad P_{\boldsymbol{s}} \stackrel{\text{def}}{=} \operatorname{diag}(\boldsymbol{s}) - \boldsymbol{s}\boldsymbol{s}^\top, \qquad Z \stackrel{\text{def}}{=} \sum_{m\in[p]} e^{\boldsymbol{a}_m},$$

so that $\boldsymbol{s}_m = e^{\boldsymbol{a}_m}/Z$; the main text also writes $P(\boldsymbol{s})$ for $P_{\boldsymbol{s}}$. The loss is $\mathcal{L}(\mathbf{V}, \boldsymbol{a}) = \ell(\boldsymbol{\beta}, \boldsymbol{\beta}_*)$ for a loss $\ell : \mathbb{R}^p \times \mathbb{R}^p \to \mathbb{R}$ differentiable in its first argument, with $\nabla\ell$ its gradient in that argument, and

$$\boldsymbol{r} \stackrel{\text{def}}{=} -\nabla\ell(\boldsymbol{\beta}, \boldsymbol{\beta}_*)$$

denotes the negative output gradient (for the square loss, the residual). When the scores are ordered with $\boldsymbol{a}_0$ largest, we write $q_l \stackrel{\text{def}}{=} \boldsymbol{s}_l/\boldsymbol{s}_0 = e^{\boldsymbol{a}_l - \boldsymbol{a}_0}$ for the score ratios. For $f : \mathbb{R}^m \to \mathbb{R}^n$, $\frac{\partial f}{\partial x} \in \mathbb{R}^{n\times m}$ denotes the Jacobian. Unless stated otherwise, dots denote time derivatives along the gradient flow

$$\frac{\mathrm{d}\mathbf{V}}{\mathrm{d}t} = -\nabla_{\mathbf{V}}\mathcal{L}(\mathbf{V}, \boldsymbol{a}), \qquad \frac{\mathrm{d}\boldsymbol{a}}{\mathrm{d}t} = -\nabla_{\boldsymbol{a}}\mathcal{L}(\mathbf{V}, \boldsymbol{a}). \tag{10}$$

All statements of this section hold verbatim for $\mathbf{V} \in \mathbb{R}^{p\times T}$ and $\boldsymbol{a} \in \mathbb{R}^T$, with T in place of p wherever p counts score coordinates; this is the form used in Section B.

Lemma A.1 (Softmax Jacobian and elementary facts)**.** *For the value–softmax model $\boldsymbol{\beta} = \mathbf{V}\boldsymbol{\sigma}(\boldsymbol{a})$:*

(a) The Jacobian of the softmax map $\boldsymbol{\sigma} : \mathbb{R}^p \to \Delta^{p-1}$ is

$$\frac{\partial\boldsymbol{\sigma}(\boldsymbol{a})}{\partial\boldsymbol{a}} = \operatorname{diag}(\boldsymbol{\sigma}(\boldsymbol{a})) - \boldsymbol{\sigma}(\boldsymbol{a})\boldsymbol{\sigma}(\boldsymbol{a})^\top, \qquad \textit{i.e.} \qquad \frac{\partial\boldsymbol{s}}{\partial\boldsymbol{a}} = P_{\boldsymbol{s}}. \tag{11}$$

(b) The Jacobians of $\boldsymbol{\beta}$ are

$$\frac{\partial\boldsymbol{\beta}}{\partial\boldsymbol{s}} = \mathbf{V}, \qquad \frac{\partial\boldsymbol{\beta}}{\partial\boldsymbol{a}} = \mathbf{V}P_{\boldsymbol{s}}, \qquad \frac{\partial\boldsymbol{\beta}}{\partial\operatorname{vec}(\mathbf{V})} = \boldsymbol{s}^\top \otimes \mathrm{I}_p \in \mathbb{R}^{p\times p^2}, \tag{12}$$

where vec *stacks the columns of* $\mathbf{V}$.

(c) $P_{\boldsymbol{s}}$ is symmetric positive semidefinite, $P_{\boldsymbol{s}}\mathbf{1} = 0$, and $(P_{\boldsymbol{s}}w)_k = \boldsymbol{s}_k\big(w_k - \langle\boldsymbol{s}, w\rangle\big)$ for every $w \in \mathbb{R}^p$.

(d) $\|\boldsymbol{s}\|^2 \geqslant 1/p$ and $\max_k \boldsymbol{s}_k \geqslant 1/p$.

Proof. (a) Differentiating $\boldsymbol{s}_i = e^{\boldsymbol{a}_i}/Z$ gives $\partial\boldsymbol{s}_i/\partial\boldsymbol{a}_k = \boldsymbol{s}_i\delta_{ik} - \boldsymbol{s}_i\boldsymbol{s}_k$. (b) The map $(\mathbf{V}, \boldsymbol{s}) \mapsto \mathbf{V}\boldsymbol{s} = \sum_j \boldsymbol{s}_j\mathbf{V}_j$ is bilinear; the formulas follow from the chain rule and (a). (c) Symmetry is clear from (a), $P_{\boldsymbol{s}}\mathbf{1} = \boldsymbol{s} - \boldsymbol{s}(\boldsymbol{s}^\top\mathbf{1}) = 0$, and $w^\top P_{\boldsymbol{s}}w = \sum_k \boldsymbol{s}_k w_k^2 - \langle\boldsymbol{s}, w\rangle^2 \geqslant 0$ by Jensen's inequality. The coordinate formula is immediate. (d) By Cauchy–Schwarz, $1 = \langle\mathbf{1}, \boldsymbol{s}\rangle^2 \leqslant p\|\boldsymbol{s}\|^2$; and the largest of p numbers summing to 1 is at least $1/p$. □

Lemma A.2 (Gradients for a general loss)**.** *With $\boldsymbol{r} = -\nabla\ell(\boldsymbol{\beta}, \boldsymbol{\beta}_*)$,*

$$\nabla_{\mathbf{V}}\mathcal{L}(\mathbf{V}, \boldsymbol{a}) = -\boldsymbol{r}\boldsymbol{s}^\top, \qquad \nabla_{\boldsymbol{a}}\mathcal{L}(\mathbf{V}, \boldsymbol{a}) = -P_{\boldsymbol{s}}\mathbf{V}^\top\boldsymbol{r}. \tag{13}$$

Hence the gradient flow (10) reads $\dot{\mathbf{V}} = \boldsymbol{r}\boldsymbol{s}^\top$ and $\dot{\boldsymbol{a}} = P_{\boldsymbol{s}}w$, where $w \stackrel{\text{def}}{=} \mathbf{V}^\top\boldsymbol{r}$ is the vector of fitnesses*, and $\dot{\boldsymbol{s}} = P_{\boldsymbol{s}}^2 w$.*

Proof. By the chain rule and Lemma A.1(b), $\frac{\partial \mathcal{L}}{\partial \mathrm{vec}(\mathbf{V})} = -\boldsymbol{r}^\top(\boldsymbol{s}^\top \otimes \mathrm{I}_p) = -(\boldsymbol{s} \otimes \boldsymbol{r})^\top$, which is $-\boldsymbol{r}\boldsymbol{s}^\top$ in matrix form, and $\frac{\partial \mathcal{L}}{\partial \boldsymbol{a}} = -\boldsymbol{r}^\top \mathbf{V} P_{\boldsymbol{s}}$. Transposing and using the symmetry of $P_{\boldsymbol{s}}$ gives (13). Finally $\dot{\boldsymbol{s}} = P_{\boldsymbol{s}}\dot{\boldsymbol{a}}$ by Lemma A.1(a). □

Lemma A.3 (Shift invariance). *(a) $\mathcal{L}(\mathbf{V}, \boldsymbol{a} + c\mathbf{1}) = \mathcal{L}(\mathbf{V}, \boldsymbol{a})$ for every $c \in \mathbb{R}$; consequently $\langle \nabla_{\boldsymbol{a}} \mathcal{L}(\mathbf{V}, \boldsymbol{a}), \mathbf{1} \rangle = 0$.*
(b) Along the gradient flow, $\sum_i \dot{\boldsymbol{a}}_i = 0$, so $\sum_i \boldsymbol{a}_i(t) = \sum_i \boldsymbol{a}_i(0)$ for all $t \geqslant 0$.

Proof. (a) $\boldsymbol{\sigma}(\boldsymbol{a} + c\mathbf{1}) = \boldsymbol{\sigma}(\boldsymbol{a})$; differentiate in c at $c = 0$. (b) follows from (a) and (10). □

The score dynamics of every model in this paper have the replicator form $\dot{\boldsymbol{a}} = P_{\boldsymbol{s}} w$ for some fitness vector w. The following identities are used repeatedly.

Lemma A.4 (Replicator form of the score dynamics). *Let $\boldsymbol{a}$ be continuously differentiable on an interval, with $\dot{\boldsymbol{a}} = P_{\boldsymbol{s}} w$ for a continuous w and $\boldsymbol{s} = \boldsymbol{\sigma}(\boldsymbol{a})$. Then, for all $j, k, l, m \in [p]$:*

(a) $\dot{\boldsymbol{a}}_m = \boldsymbol{s}_m(w_m - \langle \boldsymbol{s}, w \rangle)$ and $\frac{\mathrm{d}}{\mathrm{d}t} e^{-\boldsymbol{a}_m} = -\frac{w_m - \langle \boldsymbol{s}, w \rangle}{Z}$; in particular

$$\frac{\mathrm{d}}{\mathrm{d}t}\big(e^{-\boldsymbol{a}_k} - e^{-\boldsymbol{a}_j}\big) = \frac{w_j - w_k}{Z}.$$

(b) $\boldsymbol{s}_j - \boldsymbol{s}_k = \dfrac{e^{\boldsymbol{a}_j + \boldsymbol{a}_k}}{Z}\big(e^{-\boldsymbol{a}_k} - e^{-\boldsymbol{a}_j}\big)$; in particular $\boldsymbol{s}_j - \boldsymbol{s}_k$ and $\boldsymbol{a}_j - \boldsymbol{a}_k$ have the same sign.

(c) The ratio $q_l = \boldsymbol{s}_l / \boldsymbol{s}_0$ satisfies

$$\dot{q}_l = -q_l \boldsymbol{s}_0 \Big[q_l(w_0 - w_l) + (1 - q_l)\big(w_0 - \langle \boldsymbol{s}, w \rangle\big) \Big]. \tag{14}$$

(d) If, on the interval, $\boldsymbol{s}_0 = \max_k \boldsymbol{s}_k$ and $w_0 = \max_k w_k$, then there

$$\dot{q}_l \leqslant -\frac{q_l^2}{p}(w_0 - w_l) \qquad \text{and} \qquad \frac{\mathrm{d}}{\mathrm{d}t}\frac{1}{q_l} \geqslant \frac{w_0 - w_l}{p}.$$

Proof. (a) The first identity is Lemma A.1(c). Then $\frac{\mathrm{d}}{\mathrm{d}t} e^{-\boldsymbol{a}_m} = -e^{-\boldsymbol{a}_m}\dot{\boldsymbol{a}}_m = -e^{-\boldsymbol{a}_m}\frac{e^{\boldsymbol{a}_m}}{Z}(w_m - \langle \boldsymbol{s}, w \rangle)$; subtract the identities for $m = k$ and $m = j$. (b) $\boldsymbol{s}_j - \boldsymbol{s}_k = (e^{\boldsymbol{a}_j} - e^{\boldsymbol{a}_k})/Z$, and $e^{\boldsymbol{a}_j} - e^{\boldsymbol{a}_k} = e^{\boldsymbol{a}_j + \boldsymbol{a}_k}(e^{-\boldsymbol{a}_k} - e^{-\boldsymbol{a}_j})$. (c) Write $\bar{w} = \langle \boldsymbol{s}, w \rangle$. By (a) and $\boldsymbol{s}_l = q_l \boldsymbol{s}_0$, $\dot{q}_l = q_l(\dot{\boldsymbol{a}}_l - \dot{\boldsymbol{a}}_0) = q_l \boldsymbol{s}_0 \big[q_l(w_l - \bar{w}) - (w_0 - \bar{w}) \big]$, which is (14) after rearranging. (d) Under the hypotheses, $0 < q_l \leqslant 1$, $\boldsymbol{s}_0 \geqslant 1/p$ by Lemma A.1(d), and $w_0 \geqslant \bar{w}$. Both terms in the bracket of (14) are then nonnegative, so $\dot{q}_l \leqslant -q_l^2 \boldsymbol{s}_0 (w_0 - w_l) \leqslant -\frac{q_l^2}{p}(w_0 - w_l)$. Divide by $-q_l^2$. □

Lemma A.5 (Evolution of the output). *Let $A = \mathrm{diag}(A_0, \ldots, A_{p-1})$ with $A_i > 0$, let b be continuous, and suppose $\dot{\mathbf{V}} = A b \boldsymbol{s}^\top$ and $\dot{\boldsymbol{a}} = P_{\boldsymbol{s}} \mathbf{V}^\top b$. Then $\boldsymbol{\beta} = \mathbf{V}\boldsymbol{s}$ satisfies*

$$\dot{\boldsymbol{\beta}} = \mathbf{M}_A(\boldsymbol{s}, \mathbf{V})\, b, \qquad \mathbf{M}_A(\boldsymbol{s}, \mathbf{V}) \overset{\text{def}}{=} \|\boldsymbol{s}\|^2 A + \mathbf{V} P_{\boldsymbol{s}}^2 \mathbf{V}^\top,$$

and

$$\Big\langle b, \dot{\boldsymbol{\beta}} \Big\rangle = \|\boldsymbol{s}\|^2 b^\top A b + \big\| P_{\boldsymbol{s}} \mathbf{V}^\top b \big\|^2 \geqslant \frac{\min_i A_i}{p} \|b\|^2.$$

Proof. $\dot{\boldsymbol{\beta}} = \dot{\mathbf{V}}\boldsymbol{s} + \mathbf{V}\dot{\boldsymbol{s}} = \|\boldsymbol{s}\|^2 A b + \mathbf{V} P_{\boldsymbol{s}} \dot{\boldsymbol{a}} = \|\boldsymbol{s}\|^2 A b + \mathbf{V} P_{\boldsymbol{s}}^2 \mathbf{V}^\top b$, using $\dot{\boldsymbol{s}} = P_{\boldsymbol{s}}\dot{\boldsymbol{a}}$. The quadratic form follows from the symmetry of $P_{\boldsymbol{s}}$, and the lower bound from Lemma A.1(d). □

Corollary A.6 (Loss descent). *Along the gradient flow (10) of $\mathcal{L}(\mathbf{V}, \boldsymbol{a}) = \ell(\mathbf{V}\boldsymbol{\sigma}(\boldsymbol{a}), \boldsymbol{\beta}_*)$,*

$$\frac{\mathrm{d}\ell(\boldsymbol{\beta}, \boldsymbol{\beta}_*)}{\mathrm{d}t} \leqslant -\frac{1}{p} \big\| \nabla \ell(\boldsymbol{\beta}, \boldsymbol{\beta}_*) \big\|^2.$$

Proof. By Lemma A.2, the flow has the form of Lemma A.5 with $A = \mathrm{I}$ and $b = \boldsymbol{r} = -\nabla \ell(\boldsymbol{\beta}, \boldsymbol{\beta}_*)$. Hence $\frac{\mathrm{d}}{\mathrm{d}t}\ell(\boldsymbol{\beta}, \boldsymbol{\beta}_*) = \Big\langle \nabla \ell(\boldsymbol{\beta}, \boldsymbol{\beta}_*), \dot{\boldsymbol{\beta}} \Big\rangle = -\Big\langle \boldsymbol{r}, \dot{\boldsymbol{\beta}} \Big\rangle \leqslant -\frac{1}{p}\|\boldsymbol{r}\|^2$. □

We close with an inequality for co-sorted vectors.

Lemma A.7 (Co-sorted softmax covariance is nonnegative). *Let $\boldsymbol{s} \in \Delta^{p-1}$ and let $\boldsymbol{u}, \boldsymbol{v} \in \mathbb{R}^p$ be sorted in the same order as $\boldsymbol{s}$:*

$$\boldsymbol{s}_0 \geqslant \boldsymbol{s}_1 \geqslant \cdots \geqslant \boldsymbol{s}_{p-1}, \qquad \boldsymbol{u}_0 \geqslant \boldsymbol{u}_1 \geqslant \cdots \geqslant \boldsymbol{u}_{p-1}, \qquad \boldsymbol{v}_0 \geqslant \boldsymbol{v}_1 \geqslant \cdots \geqslant \boldsymbol{v}_{p-1}.$$

Then $\langle P_{\boldsymbol{s}}\boldsymbol{u}, P_{\boldsymbol{s}}\boldsymbol{v}\rangle \geqslant 0$.

Proof. Let $f_k = \boldsymbol{u}_k - \langle \boldsymbol{s}, \boldsymbol{u}\rangle$ and $g_k = \boldsymbol{v}_k - \langle \boldsymbol{s}, \boldsymbol{v}\rangle$. By Lemma A.1(c), $\langle P_{\boldsymbol{s}}\boldsymbol{u}, P_{\boldsymbol{s}}\boldsymbol{v}\rangle = \sum_k \boldsymbol{s}_k^2 f_k g_k$. The sequences f and g are shifts of $\boldsymbol{u}$ and $\boldsymbol{v}$, so they are co-sorted with $\boldsymbol{s}$ and with each other, and $\langle \boldsymbol{s}, f\rangle = \langle \boldsymbol{s}, g\rangle = 0$. The weighted Chebyshev sum inequality with the nonnegative weights $\boldsymbol{s}_k^2$ gives

$$\|\boldsymbol{s}\|^2 \sum_k \boldsymbol{s}_k^2 f_k g_k \;\geqslant\; \Big(\sum_k \boldsymbol{s}_k^2 f_k\Big)\Big(\sum_k \boldsymbol{s}_k^2 g_k\Big). \tag{15}$$

The same inequality applied to the co-sorted sequences $(\boldsymbol{s}_k)$ and (f_k) with the weights $\boldsymbol{s}_k$, which sum to 1, gives $\sum_k \boldsymbol{s}_k \cdot \boldsymbol{s}_k f_k \geqslant \|\boldsymbol{s}\|^2 \langle \boldsymbol{s}, f\rangle = 0$; likewise $\sum_k \boldsymbol{s}_k^2 g_k \geqslant 0$. Hence the right side of (15) is nonnegative, and $\|\boldsymbol{s}\|^2 > 0$ gives the claim. □

B Exact Reduction from Single-Layer Attention

We first record the gradients of a general attention head (Section B.1), and then prove the exact reduction to the value–softmax model together with its stability under small perturbations of the semantic blocks (Section B.2).

B.1 Gradients of a single attention head

In this subsection only, every position is a query, and no assumption is made on the parameters. Let $X \in \mathbb{R}^{T\times d}$ have rows $x_t^\top$, let Q, K be the query and key matrices with merged score matrix $W_A \overset{\text{def}}{=} QK^\top$, let W_V be the value matrix, and write

$$S = XW_AX^\top, \qquad A = \boldsymbol{\sigma}_{\text{row}}(S), \qquad X_+ = A\,XW_V^\top,$$

so that the u-th value vector is $v_u = W_V x_u$ and the t-th output row is $o_t = \sum_u A_{tu} v_u$. Fix a per-position loss ℓ and set

$$\mathcal{L} = \mathbb{E}_X \sum_{t=1}^{T} \ell(o_t, \boldsymbol{\beta}_*), \qquad g_t \overset{\text{def}}{=} \nabla \ell(o_t, \boldsymbol{\beta}_*) \in \mathbb{R}^d, \qquad G \overset{\text{def}}{=} [g_1, \ldots, g_T]^\top \in \mathbb{R}^{T\times d}.$$

Lemma B.1 (Head gradients). *Define the* score sensitivity *matrix $\Gamma \in \mathbb{R}^{T\times T}$ by*

$$\Gamma_{tu} \overset{\text{def}}{=} \frac{\partial \ell(o_t, \boldsymbol{\beta}_*)}{\partial S_{tu}} = A_{tu}\,\langle g_t, v_u - o_t\rangle\,, \qquad \textit{equivalently} \qquad \Gamma_{t\cdot} = \big(\operatorname{diag}(A_{t\cdot}) - A_{t\cdot}A_{t\cdot}^\top\big) XW_V^\top g_t.$$

Then $\Gamma\mathbf{1} = 0$, that is, every row of Γ is mean-centered, and

$$\nabla_{W_A}\mathcal{L} = \mathbb{E}\big[X^\top \Gamma X\big], \qquad \nabla_Q \mathcal{L} = \big(\nabla_{W_A}\mathcal{L}\big)K, \qquad \nabla_K \mathcal{L} = \big(\nabla_{W_A}\mathcal{L}\big)^\top Q,$$

$$\nabla_{W_V}\mathcal{L} = \mathbb{E}\big[G^\top AX\big] = \mathbb{E}\Big[\textstyle\sum_t g_t c_t^\top\Big], \qquad c_t \overset{\text{def}}{=} X^\top A_{t\cdot}^\top = \textstyle\sum_u A_{tu}x_u.$$

Proof. Since $o_t = \sum_w A_{tw} v_w$ and $A_{t\cdot} = \boldsymbol{\sigma}(S_{t\cdot})$, Lemma A.1(a) gives $\partial o_t / \partial S_{tu} = \sum_w v_w (A_{tu}\delta_{uw} - A_{tw}A_{tu}) = A_{tu}(v_u - o_t)$, whence $\Gamma_{tu} = \langle g_t, A_{tu}(v_u - o_t)\rangle$. The Jacobian form follows from $(XW_V^\top g_t)_u = \langle v_u, g_t\rangle$, and mean-centering from $\sum_u \Gamma_{tu} = \langle g_t, \sum_u A_{tu}v_u - o_t\rangle = 0$. Since $S_{tu} = x_t^\top W_A x_u$, we get $\nabla_{W_A} = \sum_{t,u} \Gamma_{tu} x_t x_u^\top = X^\top \Gamma X$, and the gradients in Q and K follow from $W_A = QK^\top$ by the chain rule. For W_V, note that $o_t = W_V c_t$, so $\mathrm{d}\ell(o_t, \boldsymbol{\beta}_*) = \langle g_t, \mathrm{d}W_V\, c_t\rangle = \operatorname{tr}(\mathrm{d}W_V\, c_t g_t^\top)$. □

The next two consequences distinguish the merged parameterization used below from a factored one.

Corollary B.2 (Factored versus merged scores)**.** *Under gradient flow on (Q, K), let $N \overset{\text{def}}{=} \nabla_{W_A}\mathcal{L}$.*

(a) Preconditioned flow. *The merged variable $W_A = QK^\top$ obeys $\dot{W}_A = -\big(QQ^\top N + NKK^\top\big)$, whereas training W_A directly gives $\dot{W}_A = -N$. The two agree, up to a reparameterization of time, only when $QQ^\top$ and $KK^\top$ act as multiples of the identity on the relevant subspaces.*

(b) Conservation. *$\frac{\mathrm{d}}{\mathrm{d}t}\big(Q^\top Q - K^\top K\big) = 0$, so the imbalance $Q^\top Q - K^\top K$ is fixed at initialization.*

Proof. (a) By Lemma B.1, $\dot{Q} = -NK$ and $\dot{K} = -N^\top Q$, so $\dot{W}_A = \dot{Q}K^\top + Q\dot{K}^\top = -NKK^\top - QQ^\top N$. (b) $\frac{\mathrm{d}}{\mathrm{d}t}Q^\top Q = -K^\top N^\top Q - Q^\top NK$ and $\frac{\mathrm{d}}{\mathrm{d}t}K^\top K = -Q^\top NK - K^\top N^\top Q$ coincide. □

Remark B.3. *The matrix Γ has the same "fitness minus average fitness" structure as the value–softmax dynamics (Lemma A.4): row t of Γ is the softmax Jacobian applied to the vector of alignments $\langle g_t, v_u\rangle$. The polarizing mechanism is therefore already visible at the level of a full head; the reduction below adds a regime in which the score and value variables decouple from the input (Proposition 2.1).*

B.2 Reduction to the value–softmax model

We prove the exact reduction for a general loss, then derive the first-order perturbation equation for the logistic loss and the trajectory bounds for the square loss. All three results use the same block decomposition of the head. The bounds hold over different time ranges because the reduced logistic and square-loss flows behave differently at long times.

Setup and blocks. Let $x_1, \dots, x_T$ be centered semantic tokens, and let the query's semantic token x_q be centered and independent of the context. The positional encodings $e_1, \dots, e_T, e_q$ are orthonormal. With $z_i = (x_i, e_i)$ stacked as the rows of Z, $q = (x_q, e_q)$, and X the matrix with rows $x_i^\top$, a head with merged score matrix W_A and value matrix W_V outputs $\widehat{\boldsymbol{\beta}} = W_V Z^\top \boldsymbol{\sigma}(ZW_A q)$. Splitting W_A and W_V into semantic and positional blocks gives (2); the block coordinates are isometric to the corresponding entries of the original parameter matrices. We write $\theta = (\mathbf{V}, \boldsymbol{a})$ for the positional blocks, $\eta = (W, W_{xx}, W_{px}, w_p)$ for the semantic blocks, $\mathcal{N} = \{\eta = 0\}$, and $\mathcal{L}(\theta, \eta) = \mathbb{E}\,\ell(\widehat{\boldsymbol{\beta}}, \boldsymbol{\beta}_*)$ for the population loss. The perturbation results assume independent standard Gaussian semantic tokens; the exact reduction uses only centering and the independence of the query.

Proof of Proposition 2.1. On $\mathcal{N}$ the logits equal $\boldsymbol{a}$, so $\boldsymbol{s} = \boldsymbol{\sigma}(\boldsymbol{a})$ and $\widehat{\boldsymbol{\beta}} = \mathbf{V}\boldsymbol{s}$ are deterministic. Let

$$g = \nabla\ell(\mathbf{V}\boldsymbol{s}, \boldsymbol{\beta}_*), \qquad h = P_{\boldsymbol{s}}\mathbf{V}^\top g.$$

The chain rule gives the population gradients of the semantic blocks at any point of $\mathcal{N}$:

$$\nabla_W\mathcal{L} = \mathbb{E}[g\boldsymbol{s}^\top X], \quad \nabla_{w_p}\mathcal{L} = \mathbb{E}[X^\top h], \quad \nabla_{W_{px}}\mathcal{L} = \mathbb{E}[hx_q^\top], \quad \nabla_{W_{xx}}\mathcal{L} = \mathbb{E}[X^\top hx_q^\top].$$

Because $g, \boldsymbol{s}, h$ are deterministic on $\mathcal{N}$, $\mathbb{E}X = 0$, $\mathbb{E}x_q = 0$, and x_q is independent of X, each of these gradients vanishes. Thus the population vector field is tangent to $\mathcal{N}$. The remaining gradients are $\nabla_{\mathbf{V}}\mathcal{L} = g\boldsymbol{s}^\top$ and $\nabla_{\boldsymbol{a}}\mathcal{L} = h$ (Lemma A.2), which proves (3). The argument uses no property of ℓ beyond differentiability and the interchange of differentiation and expectation. □

Logistic loss: the first-order perturbation. For the logistic objective (5), a small initialization of the semantic blocks produces a first-order perturbation governed by a linear equation along the reduced flow. We state the approximation on finite intervals, where the coefficients of that equation are bounded; the logistic value parameters can grow without bound at long times.

Proposition B.4 (Logistic perturbations to first order)**.** *Use the full head (2) with independent standard Gaussian semantic tokens and $\ell(u, \boldsymbol{\beta}_*) = \log(1 + e^{-\langle \boldsymbol{\beta}_*, u\rangle})$. Let $\bar{\theta}(t) = (\bar{\mathbf{V}}(t), \bar{\boldsymbol{a}}(t))$ be the reduced flow on $\mathcal{N}$. For a fixed ζ_0 with $\|\zeta_0\| \leqslant 1$, initialize $\theta_\varepsilon(0) = \bar{\theta}(0)$ and $\eta_\varepsilon(0) = \varepsilon\zeta_0$. Let ζ solve*

$$\dot{\zeta}(t) = -H^{\log}_{\eta\eta}(t)\zeta(t), \qquad \zeta(0) = \zeta_0, \tag{16}$$

where $H^{\log}_{\eta\eta}(t)$ is the Hessian of the population logistic loss with respect to η at $(\bar{\theta}(t), 0)$, given explicitly in (17) below. Then for every finite $H > 0$ there are $C_H < \infty$ and $\varepsilon_H > 0$ such that $0 < \varepsilon \leqslant \varepsilon_H$ implies

$$\sup_{0 \leqslant t \leqslant H} \big(\|\eta_\varepsilon(t) - \varepsilon\zeta(t)\| + \|\theta_\varepsilon(t) - \bar{\theta}(t)\|\big) \leqslant C_H\varepsilon^2.$$

Proof. At a point $(\mathbf{V}, \boldsymbol{a}, 0)$ of $\mathcal{N}$, let $\boldsymbol{s} = \boldsymbol{\sigma}(\boldsymbol{a})$. For a perturbation $\zeta = (\Delta W, \Delta W_{xx}, \Delta W_{px}, \Delta w_p)$ of the semantic blocks, set

$$L_\zeta = X\Delta W_{xx}x_q + \Delta W_{px}x_q + X\Delta w_p, \qquad J\zeta = D_\eta\widehat{\boldsymbol{\beta}}[\zeta] = \Delta W X^\top \boldsymbol{s} + \mathbf{V}P_{\boldsymbol{s}}L_\zeta.$$

For two directions $\zeta = (\Delta W, \dots)$ and $\zeta' = (\Delta W', \dots)$, differentiating (2) gives

$$D^2_\eta\widehat{\boldsymbol{\beta}}[\zeta, \zeta'] = \Delta W X^\top P_{\boldsymbol{s}}L_{\zeta'} + \Delta W' X^\top P_{\boldsymbol{s}}L_\zeta + \mathbf{V}D^2\boldsymbol{\sigma}(\boldsymbol{a})[L_\zeta, L_{\zeta'}].$$

Along the reduced logistic trajectory let $\bar{\boldsymbol{\beta}} = \bar{\mathbf{V}}\bar{\boldsymbol{s}}$ and $\gamma = (1+e^{\langle \boldsymbol{\beta}_*, \bar{\boldsymbol{\beta}}\rangle})^{-1}$. Since $\nabla\ell(\bar{\boldsymbol{\beta}}, \boldsymbol{\beta}_*) = -\gamma\boldsymbol{\beta}_*$ and $\nabla^2\ell(\bar{\boldsymbol{\beta}}, \boldsymbol{\beta}_*) = \gamma(1-\gamma)\boldsymbol{\beta}_*\boldsymbol{\beta}_*^\top$, the Hessian with respect to η is the bilinear form

$$\left\langle \zeta, H^{\log}_{\eta\eta}(t)\zeta' \right\rangle = \gamma(1-\gamma)\,\mathbb{E}\big[\langle \boldsymbol{\beta}_*, J\zeta\rangle\,\langle \boldsymbol{\beta}_*, J\zeta'\rangle\big] - \gamma\left\langle \boldsymbol{\beta}_*, \mathbb{E}[D^2_\eta\widehat{\boldsymbol{\beta}}[\zeta, \zeta']]\right\rangle. \tag{17}$$

Write (F_θ, F_η) for the full negative-gradient vector field. The centering calculation in the proof of Proposition 2.1 gives $F_\eta(\theta, 0) = 0$ and $D_\eta F_\theta(\theta, 0) = 0$. Thus a first-order change in η does not alter the positional trajectory at first order, while its own first variation obeys (16). On any fixed time interval, the reduced trajectory and the needed derivatives of the Gaussian population field are bounded. Taylor's theorem for the parameter-dependent ODE then gives the $O_H(\varepsilon^2)$ remainder uniformly on that interval. □

B.3 Stability of the reduction for regression

For an initialization near $\mathcal{N}$, we compare the full head with the reduced trajectory started from the same positional parameters. Take independent standard Gaussian semantic tokens in (2) and the square loss $\ell(u, \boldsymbol{\beta}_*) = \frac{1}{2}\|u - \boldsymbol{\beta}_*\|^2$ applied directly to the head output. When $T = p$, this is the $X = I$ case of the regression model in Corollary 3.6; the theorem below does not cover perturbations of a head under a general data matrix. We write $\theta = (\mathbf{V}, \boldsymbol{a})$ for the positional parameters, $\eta = (W, W_{xx}, W_{px}, w_p)$ for the semantic ones, and $\bar{\theta}(t)$ for the reduced solution with $\bar{\theta}(0) = \theta(0)$.

Theorem B.5 (Regression perturbations near $\mathcal{N}$). *For each finite horizon $H > 0$, there are positive constants $K_H, L_H, D_H, \varepsilon_H$, depending on the reduced trajectory and the model dimensions, such that $\|\eta(0)\| \leqslant \varepsilon \leqslant \varepsilon_H$ implies, for $0 \leqslant t \leqslant H$,*

$$\|\eta(t)\| \leqslant \varepsilon e^{K_H t}, \qquad \|\theta(t) - \bar{\theta}(t)\| \leqslant D_H\varepsilon^2 \int_0^t e^{L_H(t-u)+2K_H u}\,\mathrm{d}u. \tag{18}$$

Moreover, let ζ denote a first-order change in the semantic blocks η along $\bar{\theta}(t)$, with the positional blocks fixed at first order. This linearized perturbation is bounded for all time: for some finite C depending on the reduced trajectory,

$$\sup_{t\geqslant 0}\|\zeta(t)\| \leqslant \exp\big(CT\|\bar{\mathbf{V}}(0)\boldsymbol{\sigma}(\bar{\boldsymbol{a}}(0)) - \boldsymbol{\beta}_*\|\big)\|\zeta(0)\|. \tag{19}$$

The first estimate controls the actual full trajectory over any prescribed finite interval: a perturbation of size ε in the semantic blocks stays of order ε, while its effect on the positional trajectory is of order ε^2. The second concerns only the first-order perturbation equation along the reduced regression flow and holds for all time because the reduced residual decays exponentially; it does not imply an all-time bound for the nonlinear full head, nor convergence back to $\mathcal{N}$. The finite-horizon first-order equation for the logistic loss is given in Proposition B.4 above.

Proof of Theorem B.5. Write the full negative-gradient vector field as (F_θ, F_η) and the reduced one as $F_{\mathrm{vs}}(\theta) = F_\theta(\theta, 0)$. The centering calculation in the proof of Proposition 2.1 gives $F_\eta(\theta, 0) = 0$ for every θ. It also makes the derivative of the tangential field in every semantic-block direction vanish:

$$D_\eta F_\theta(\theta, 0) = 0.$$

Indeed, every first-order change of the output is a sum of terms linear in X, linear in x_q, or bilinear in the independent centered pair (X, x_q), and the first-order changes of the positional gradients have the same property. For Gaussian semantic tokens the population vector field is smooth on bounded parameter sets: softmax derivatives are bounded and Gaussian moments control the derivatives of the

integrand. On a neighborhood of the reduced trajectory over $0 \leqslant t \leqslant H$, Taylor's theorem and local Lipschitz continuity give constants K_H, L_H, D_H with

$$\|F_\eta(\theta,\eta)\| \leqslant K_H\|\eta\|, \quad \|F_\theta(\theta,\eta) - F_{\mathrm{vs}}(\theta)\| \leqslant D_H\|\eta\|^2, \quad \|F_{\mathrm{vs}}(\theta) - F_{\mathrm{vs}}(\bar\theta)\| \leqslant L_H\|\theta-\bar\theta\|.$$

Thus $\|\eta(t)\| \leqslant \varepsilon e^{K_H t}$. For $e = \theta - \bar\theta$, variation of constants and $e(0) = 0$ yield

$$\|e(t)\| \leqslant D_H\varepsilon^2 \int_0^t e^{L_H(t-u)+2K_H u}\,\mathrm{d}u.$$

For ε_H small enough, these bounds keep the full trajectory inside the neighborhood for $0 \leqslant t \leqslant H$, so the estimates apply throughout the interval and prove (18).

For the all-time linearized bound, let $\bar{\boldsymbol{s}} = \boldsymbol{\sigma}(\bar{\boldsymbol{a}})$ and $\bar{\boldsymbol{r}}(t) = \boldsymbol{\beta}_* - \bar{\mathbf{V}}(t)\bar{\boldsymbol{s}}(t) = -\nabla\ell(\bar{\mathbf{V}}\bar{\boldsymbol{s}}, \boldsymbol{\beta}_*)$. The reduced flow is $\dot{\bar{\mathbf{V}}} = \bar{\boldsymbol{r}}\bar{\boldsymbol{s}}^\top$, $\dot{\bar{\boldsymbol{a}}} = P_{\bar{\boldsymbol{s}}}\bar{\mathbf{V}}^\top\bar{\boldsymbol{r}}$, and Lemma A.5 with $A = \mathrm{I}$ and T positions gives

$$\frac{\mathrm{d}}{\mathrm{d}t}\frac{1}{2}\|\bar{\boldsymbol{r}}\|^2 = -\|\bar{\boldsymbol{s}}\|^2\|\bar{\boldsymbol{r}}\|^2 - \|P_{\bar{\boldsymbol{s}}}\bar{\mathbf{V}}^\top\bar{\boldsymbol{r}}\|^2 \leqslant -\frac{1}{T}\|\bar{\boldsymbol{r}}\|^2, \qquad \text{so} \qquad \|\bar{\boldsymbol{r}}(t)\| \leqslant \|\bar{\boldsymbol{r}}(0)\|e^{-t/T}.$$

Both $\dot{\bar{\mathbf{V}}}$ and $\dot{\bar{\boldsymbol{a}}}$ are therefore integrable, and the reduced trajectory is bounded. At $(\bar\theta(t), 0)$, let $J(t) = D_\eta\widehat{\boldsymbol{\beta}}$ and let $D^2_\eta\widehat{\boldsymbol{\beta}}$ be the second derivative of the head output with respect to η. Since $\widehat{\boldsymbol{\beta}} = \bar{\mathbf{V}}\bar{\boldsymbol{s}}$ is deterministic on $\mathcal{N}$, the Hessian of the population square loss is

$$H_{\eta\eta}(t) = \mathbb{E}[J(t)^\top J(t)] - \langle \bar{\boldsymbol{r}}(t), \mathbb{E}[D^2_\eta\widehat{\boldsymbol{\beta}}(t)]\rangle.$$

The first term is positive semidefinite. The second contracts the curvature of the head output with the reduced residual; along the bounded orbit, Gaussian moments and bounded softmax derivatives bound its operator norm by $C\|\bar{\boldsymbol{r}}(t)\|$ for a finite C. A first-order perturbation in η obeys $\dot\zeta = -H_{\eta\eta}(t)\zeta$, hence

$$\frac{1}{2}\frac{\mathrm{d}}{\mathrm{d}t}\|\zeta\|^2 = -\mathbb{E}\|J(t)\zeta\|^2 + \langle \bar{\boldsymbol{r}}(t), \mathbb{E}[D^2_\eta\widehat{\boldsymbol{\beta}}(t)[\zeta,\zeta]]\rangle \leqslant C\|\bar{\boldsymbol{r}}(t)\|\|\zeta\|^2.$$

Since $\int_0^\infty \|\bar{\boldsymbol{r}}(t)\|\,\mathrm{d}t \leqslant T\|\bar{\boldsymbol{r}}(0)\|$, Grönwall's inequality proves (19) (after enlarging C if needed). □

Near-uniform initialization. For the square-loss reduced flow, a near-uniform start can be followed explicitly to first order.

Corollary B.6 (Near-uniform attention)**.** *Consider the square-loss reduced flow of Proposition 2.1 started at $\mathbf{V}(0) = 0$ and $\boldsymbol{a}(0) = \varepsilon\,\boldsymbol{a}_0$ with $\langle \mathbf{1}, \boldsymbol{a}_0\rangle = 0$. Let $\tilde a_u \stackrel{\mathrm{def}}{=} T\boldsymbol{s}_u - 1$ and $\tau(t) \stackrel{\mathrm{def}}{=} 1 - e^{-t/T}$. Then, for all $t \geqslant 0$,*

$$\tilde a(t) = \tilde a(0)\cosh\big(\|\boldsymbol{\beta}_*\|\,\tau(t)\big) + O(\varepsilon^2).$$

In particular, to first order the ordering of $\boldsymbol{a}$ is preserved, and deviations from uniform attention are amplified by exactly $\cosh(\|\boldsymbol{\beta}_*\|)$ as $t \to \infty$. The amplification depends on the target norm but not on the number of positions T, which only sets the time scale.

Proof sketch. Write $\mathbf{V}_u = \boldsymbol{\beta} + d_u$, where $\boldsymbol{\beta} = \mathbf{V}\boldsymbol{s}$ and $\mathbf{V}_u$ is the u-th column; then $d(0) = 0$ and $\sum_u \boldsymbol{s}_u d_u = 0$. Because $P_{\boldsymbol{s}}\mathbf{1} = 0$, $P_{\boldsymbol{s}}\mathbf{V}^\top\boldsymbol{r}$ depends on $\mathbf{V}$ only through the d_u, and $\sum_u \dot{\boldsymbol{s}}_u = 0$. Expanding to first order in ε gives

$$\dot{\boldsymbol{\beta}} = -\tfrac{1}{T}(\boldsymbol{\beta} - \boldsymbol{\beta}_*), \qquad \dot d_u = -\tfrac{\tilde a_u}{T}(\boldsymbol{\beta} - \boldsymbol{\beta}_*), \qquad \dot{\tilde a}_u = -\tfrac{1}{T}\langle d_u, \boldsymbol{\beta} - \boldsymbol{\beta}_*\rangle,$$

so $\boldsymbol{\beta} - \boldsymbol{\beta}_* = -\boldsymbol{\beta}_* e^{-t/T}$. With $e_u = \langle \boldsymbol{\beta}_*, d_u\rangle$ and $\mathrm{d}\tau = \frac{1}{T}e^{-t/T}\mathrm{d}t$, this becomes $\mathrm{d}\tilde a/\mathrm{d}\tau = e$ and $\mathrm{d}e/\mathrm{d}\tau = \|\boldsymbol{\beta}_*\|^2\tilde a$, hence $\tilde a'' = \|\boldsymbol{\beta}_*\|^2\tilde a$ with $\tilde a'(0) = e(0) = 0$. Since $\tau \in [0,1)$, Grönwall's inequality bounds the neglected $O(\varepsilon^2)$ terms uniformly in t. □

Remark B.7 (Scope)**.** *The exact reduction concerns a separate attended context, an objective that does not depend on the semantic tokens, population gradients, and a merged score matrix. Centered semantic tokens and an independent query suffice for invariance; the perturbation estimates use Gaussian semantic tokens. The nonlinear bound in Theorem B.5 and the logistic expansion in Proposition B.4 are stated on fixed horizons and do not establish attraction to $\mathcal{N}$. Without a causal mask, which position becomes the sink is set by the initial scores $\boldsymbol{a}(0)$.*

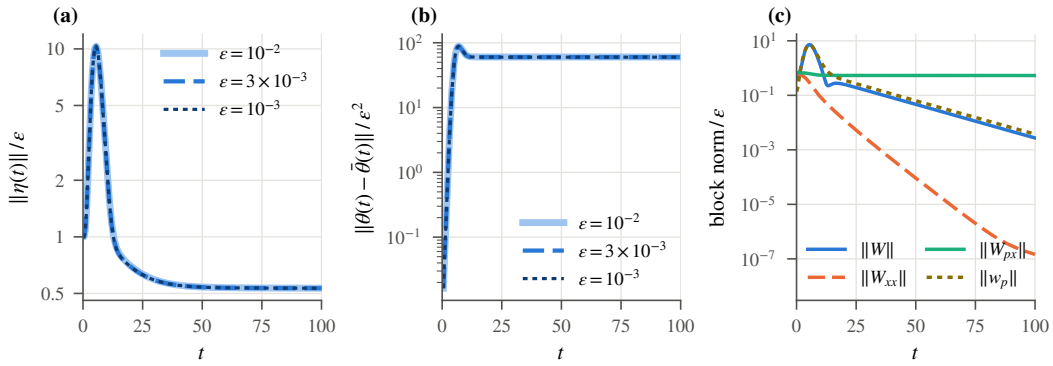


Figure 4: Perturbation off the set $\mathcal{N} = \{\eta = 0\}$ of heads with zero semantic blocks for the attention head of Theorem B.5 ($T = d = 8$, $p = 4$, $\|\boldsymbol{\beta}_*\| = 2$), started at $\|\eta(0)\| = \varepsilon$ and compared with the value–softmax flow $\bar\theta$ from the same $(\mathbf{V}(0), \boldsymbol{a}(0))$. **(a)** Semantic-block perturbation $\|\eta(t)\|/\varepsilon$: the curves for the three values of ε coincide, peak at about 10 near $t \approx 5$ and settle at about 0.53. **(b)** Deviation of the positional parameters, $\|\theta(t) - \bar\theta(t)\|/\varepsilon^2$ (from the first logged step, as it is 0 at $t = 0$): again the curves coincide, confirming the $O(\varepsilon^2)$ scaling, and level off at about 60. **(c)** Blocks of η divided by ε for $\varepsilon = 10^{-3}$: W and w_p decay at rate about 0.056 and W_{xx} at about 0.11, while W_{px} stays at about 0.53. Log vertical axes.

Numerical check. We simulate the population gradient flow of the head in Theorem B.5 with $T = d = 8$, $p = 4$ and $\|\boldsymbol{\beta}_*\| = 2$, starting from a random $\theta(0) = (\mathbf{V}(0), \boldsymbol{a}(0))$ and $\eta(0) = \varepsilon\,\eta_0$ for a fixed unit direction η_0 and $\varepsilon \in \{10^{-2}, 3\times10^{-3}, 10^{-3}\}$, alongside the value–softmax flow $\bar\theta(t)$ from the same $\theta(0)$ (forward Euler, step 0.02, up to $t = 100$). The expectation is a fixed Monte Carlo sample of 2048 draws of (X, x_q), symmetrized over the signs of X and of x_q, so that $\mathcal{N}$ is exactly invariant for the sampled loss as well. Figure 4 shows that $\|\eta(t)\|/\varepsilon$ and $\|\theta(t) - \bar\theta(t)\|/\varepsilon^2$ collapse onto single curves, as predicted by Theorem B.5. The perturbation is first amplified about tenfold and then stays bounded without decaying to zero: W, W_{xx} and w_p decay exponentially, but W_{px} levels off at about $0.5\,\varepsilon$, consistent with a neutral normal direction at the limiting reduced fit. This is a single configuration; the constants, though not the qualitative shape, depend on the Monte Carlo sample size.

C Logistic Loss

Throughout this section, ℓ is the logistic loss along the direction $\boldsymbol{\beta}_*$,

$$\ell(\boldsymbol{\beta}, \boldsymbol{\beta}_*) = \log\big(1 + \exp(-\langle \boldsymbol{\beta}_*, \boldsymbol{\beta}\rangle)\big), \tag{20}$$

so that $\boldsymbol{r} = -\nabla\ell(\boldsymbol{\beta}, \boldsymbol{\beta}_*) = \gamma\,\boldsymbol{\beta}_*$ with

$$\gamma = \gamma(\boldsymbol{\beta}) \overset{\text{def}}{=} \frac{1}{1 + \exp(\langle \boldsymbol{\beta}_*, \boldsymbol{\beta}\rangle)} \in (0, 1).$$

The output gradient always points along $\boldsymbol{\beta}_*$; only its scale γ depends on $\boldsymbol{\beta}$. By Lemma A.2, the flow is $\dot{\mathbf{V}} = \gamma\boldsymbol{\beta}_*\boldsymbol{s}^\top$ and $\dot{\boldsymbol{a}} = \gamma P_{\boldsymbol{s}}\mathbf{V}^\top\boldsymbol{\beta}_*$, and the projection $\boldsymbol{u} = \mathbf{V}^\top\boldsymbol{\beta}_*$ obeys the two-vector system (6):

$$\dot{\boldsymbol{u}} = \gamma\|\boldsymbol{\beta}_*\|^2\boldsymbol{s}, \qquad \dot{\boldsymbol{a}} = \gamma P_{\boldsymbol{s}}\boldsymbol{u}.$$

Thus the scores follow the replicator dynamics of Lemma A.4 with fitness $w = \gamma\boldsymbol{u}$, and $\langle \boldsymbol{\beta}_*, \boldsymbol{\beta}\rangle = \langle \boldsymbol{s}, \boldsymbol{u}\rangle$.

Normalization of the target. In the proofs of Theorem 3.3 and Lemma C.1 we take $\|\boldsymbol{\beta}_*\| = 1$. For a general $\boldsymbol{\beta}_* \neq 0$ the factor $\|\boldsymbol{\beta}_*\|^2$ multiplies $\dot{\boldsymbol{u}}$ only: the ratio bound in Theorem 3.3(b) does not involve it, and in parts (a) and (c) it changes constants only.

C.1 Order preservation

Theorem 3.2. *Consider gradient flow on the loss (5) under Assumption 3.1. Then, for all $t > 0$,*

(a) ***Order preservation.*** *The order of the projections $\boldsymbol{u}_i(t)$ and attention scores $\boldsymbol{s}_i(t)$ is preserved*

$$\boldsymbol{u}_0(t) > \boldsymbol{u}_1(t) > \ldots > \boldsymbol{u}_{p-1}(t), \qquad \boldsymbol{s}_0(t) > \boldsymbol{s}_1(t) > \ldots > \boldsymbol{s}_{p-1}(t).$$

(b) ***Repulsion.*** *The difference between projections $\boldsymbol{u}$ grows over time: for any $i < j$ and any $t > 0$,*

$$\frac{d}{dt}\left(\boldsymbol{u}_i(t) - \boldsymbol{u}_j(t)\right) > 0.$$

Proof. Fix $i < j$. By Lemma A.4(a,b) with $w = \gamma\boldsymbol{u}$,

$$\frac{\mathrm{d}}{\mathrm{d}t}\left(e^{-\boldsymbol{a}_j} - e^{-\boldsymbol{a}_i}\right) = \frac{\gamma}{Z}(\boldsymbol{u}_i - \boldsymbol{u}_j), \qquad \frac{\mathrm{d}}{\mathrm{d}t}(\boldsymbol{u}_i - \boldsymbol{u}_j) = \gamma\left\|\boldsymbol{\beta}_*\right\|^2(\boldsymbol{s}_i - \boldsymbol{s}_j) = \frac{\gamma}{Z}\left\|\boldsymbol{\beta}_*\right\|^2\left(e^{\boldsymbol{a}_i} - e^{\boldsymbol{a}_j}\right).$$

The two differences reinforce each other. To exploit this, consider the Lyapunov function

$$\Phi_{ij}(t) = \left(\boldsymbol{u}_i(t) - \boldsymbol{u}_j(t)\right)\left(e^{-\boldsymbol{a}_j(t)} - e^{-\boldsymbol{a}_i(t)}\right).$$

Using $(e^{\boldsymbol{a}_i} - e^{\boldsymbol{a}_j})(e^{-\boldsymbol{a}_j} - e^{-\boldsymbol{a}_i}) = (e^{\boldsymbol{a}_i} - e^{\boldsymbol{a}_j})^2 e^{-\boldsymbol{a}_i - \boldsymbol{a}_j}$,

$$\frac{\mathrm{d}\Phi_{ij}}{\mathrm{d}t} = \frac{\gamma}{Z}(\boldsymbol{u}_i - \boldsymbol{u}_j)^2 + \frac{\gamma}{Z}\left\|\boldsymbol{\beta}_*\right\|^2 \frac{\left(e^{\boldsymbol{a}_i} - e^{\boldsymbol{a}_j}\right)^2}{e^{\boldsymbol{a}_i}e^{\boldsymbol{a}_j}} \geqslant 0.$$

At $t = 0$ we have $\boldsymbol{a}(0) = 0$, so $\Phi_{ij}(0) = 0$, while $\dot{\Phi}_{ij}(0) = \frac{\gamma}{Z}(\boldsymbol{u}_i(0) - \boldsymbol{u}_j(0))^2 > 0$. Since Φ_{ij} is nondecreasing, $\Phi_{ij}(t) > 0$ for every $t > 0$. Hence neither $\boldsymbol{u}_i - \boldsymbol{u}_j$ nor $\boldsymbol{a}_i - \boldsymbol{a}_j$ vanishes for $t > 0$. By continuity and $\boldsymbol{u}_i(0) > \boldsymbol{u}_j(0)$, we get $\boldsymbol{u}_i(t) > \boldsymbol{u}_j(t)$ for all t, and then $\Phi_{ij} > 0$ forces $e^{-\boldsymbol{a}_j} > e^{-\boldsymbol{a}_i}$, i.e. $\boldsymbol{a}_i(t) > \boldsymbol{a}_j(t)$ and, by Lemma A.4(b), $\boldsymbol{s}_i(t) > \boldsymbol{s}_j(t)$ for $t > 0$. This proves (a).

For (b), by (a),

$$\frac{\mathrm{d}}{\mathrm{d}t}(\boldsymbol{u}_i - \boldsymbol{u}_j) = \gamma\left\|\boldsymbol{\beta}_*\right\|^2(\boldsymbol{s}_i - \boldsymbol{s}_j) > 0 \qquad \text{for } t > 0,$$

so $\boldsymbol{u}_i(t_+) - \boldsymbol{u}_j(t_+) > \boldsymbol{u}_i(t) - \boldsymbol{u}_j(t)$ for all $t_+ > t > 0$. A similar argument holds for $\boldsymbol{a}_i, \boldsymbol{s}_i$. □

C.2 Polarization and the one-hot limit

Theorem 3.3. *Consider gradient flow for the loss (5) under Assumption 3.1. Let $\delta = \min_{i<j}(\boldsymbol{u}_i(0) - \boldsymbol{u}_j(0))$. Then we have*

(a) ***Vanishing loss.*** $\lim_{t\to\infty} L(\boldsymbol{\beta}(t)) = 0$.

(b) ***Decay of non-maximal scores.*** *For any $j \neq 0$ and any $t > 0$,*

$$\frac{\boldsymbol{s}_j(t)}{\boldsymbol{s}_0(t)} \leqslant \left(1 + {}^{\delta}/_{p}\int_0^t \gamma(\boldsymbol{\beta}(s))ds\right)^{-1}.$$

(c) ***One-hot limit.*** *The polarization coefficient $\int_0^t \gamma(\boldsymbol{\beta}(s))ds = \Theta(\log t)$, and consequently the attention scores converge to a one-hot vector, i.e.,*

$$\lim_{t\to\infty} \boldsymbol{s}_0(t) = 1, \quad \textit{and} \quad \lim_{t\to\infty} \boldsymbol{s}_j(t) = 0, \quad \textit{for } j \neq 0.$$

Proof. Assume $\left\|\boldsymbol{\beta}_*\right\| = 1$. By Theorem 3.2, for all $t \geqslant 0$: $\boldsymbol{s}_0$ is the largest score, $\boldsymbol{u}$ is ordered, and the gaps $\boldsymbol{u}_0 - \boldsymbol{u}_l$ are nondecreasing, so $\boldsymbol{u}_0(t) - \boldsymbol{u}_l(t) \geqslant \delta$ for every $l \geqslant 1$.

Step 1: the ratio bound (b). The fitness $w = \gamma\boldsymbol{u}$ is ordered, so Lemma A.4(d) gives

$$\dot{q}_l \leqslant -\frac{\gamma}{p}\, q_l^2\, (\boldsymbol{u}_0 - \boldsymbol{u}_l) \leqslant -\frac{\gamma\delta}{p}\, q_l^2.$$

Integrating from $q_l(0) = 1$ (since $\boldsymbol{a}(0) = 0$),

$$q_l(t) \leqslant \frac{1}{1 + \frac{\delta}{p}\int_0^t \gamma(h)\,\mathrm{d}h},$$

which is (b). The bound is uniform in l; it remains to show that $\int_0^t \gamma$ diverges.

Step 2: divergence of $\int \gamma$. Since $\boldsymbol{u}_j \leqslant \boldsymbol{u}_0$ for all j, we have $\langle \boldsymbol{\beta}_*, \boldsymbol{\beta} \rangle = \langle \boldsymbol{s}, \boldsymbol{u} \rangle \leqslant \boldsymbol{u}_0$, hence $\gamma \geqslant (1 + e^{\boldsymbol{u}_0})^{-1}$. As $\boldsymbol{s}_0 \geqslant 1/p$,

$$\dot{\boldsymbol{u}}_0 = \gamma \boldsymbol{s}_0 \geqslant \frac{\gamma}{p} \geqslant \frac{1}{p\,(1 + e^{\boldsymbol{u}_0})}.$$

Separating variables, $\int_{\boldsymbol{u}_0(0)}^{\boldsymbol{u}_0(t)} (1 + e^z)\,\mathrm{d}z \geqslant t/p$, that is, $\boldsymbol{u}_0(t) + e^{\boldsymbol{u}_0(t)} \geqslant t/p + \boldsymbol{u}_0(0) + e^{\boldsymbol{u}_0(0)}$. Using $z \leqslant e^z$ on the left gives, for a constant c_0 depending only on $\boldsymbol{u}_0(0)$,

$$\boldsymbol{u}_0(t) \geqslant \log\Big(\frac{t}{2p} - c_0\Big) \qquad \text{whenever } \frac{t}{2p} > c_0. \tag{21}$$

On the other hand $\dot{\boldsymbol{u}}_0 = \gamma \boldsymbol{s}_0 \leqslant \gamma$, so

$$\int_0^t \gamma(h)\,\mathrm{d}h \;\geqslant\; \boldsymbol{u}_0(t) - \boldsymbol{u}_0(0) \;\geqslant\; \log\Big(\frac{t}{2p} - c_0\Big) - \boldsymbol{u}_0(0). \tag{22}$$

Consequently

$$\lim_{t \to \infty} \int_0^t \gamma(h)\,\mathrm{d}h = \infty, \tag{23}$$

and Step 1 yields, for large t, the explicit rate

$$q_l(t) \leqslant \Big(1 + \frac{\delta}{p}\Big(\log\Big(\frac{t}{2p} - c_0\Big) - \boldsymbol{u}_0(0)\Big)\Big)^{-1} \longrightarrow 0.$$

Since $\boldsymbol{s}_0 = (1 + \sum_{l \geqslant 1} q_l)^{-1}$, this gives $\boldsymbol{s}_0 \to 1$ and $\boldsymbol{s}_l \to 0$ for $l \geqslant 1$, which is (c).

Step 3: vanishing loss (a). We have

$$\langle \boldsymbol{\beta}_*, \boldsymbol{\beta}(t) \rangle = \sum_j \boldsymbol{s}_j(t) \boldsymbol{u}_j(t) \leqslant \boldsymbol{u}_0(t), \qquad \sum_j \boldsymbol{s}_j(t) \boldsymbol{u}_j(t) \geqslant \frac{1}{p} \boldsymbol{u}_0(t),$$

and $\boldsymbol{u}_0(t) \to \infty$ by (21). Therefore $\langle \boldsymbol{\beta}_*, \boldsymbol{\beta}(t) \rangle \to \infty$, and by (20), $\ell(\boldsymbol{\beta}(t), \boldsymbol{\beta}_*) = \log(1 + e^{-\langle \boldsymbol{\beta}_*, \boldsymbol{\beta}(t) \rangle}) \to 0$. □

The next lemma refines Theorem 3.3 for the non-maximal coordinates.

Lemma C.1 (Rates for the non-maximal coordinates)**.** *Under the assumptions of Theorem 3.3:*

(a) The projections $\mathbf{u}_i(t)$, $i \neq 0$, stay bounded: there is $C > 0$ with $\mathbf{u}_i(t) \leqslant C$ for all $t \geqslant 0$. Since $\boldsymbol{u}_0 \to \infty$, the matrix $\mathbf{V}$ is asymptotically close to a rank-one matrix.

(b) The attention scores satisfy $\boldsymbol{s}_i = O(1/\log^2 t)$ for $i \neq 0$.

Proof. Assume $\|\boldsymbol{\beta}_*\| = 1$. From (22) we have $\gamma = O(1/t)$. By Lemma A.4, $\frac{\mathrm{d}}{\mathrm{d}t}(\boldsymbol{u}_0 - \boldsymbol{u}_i) = \gamma(\boldsymbol{s}_0 - \boldsymbol{s}_i) = \gamma \boldsymbol{s}_0 (1 - q_i)$. As $q_i \to 0$, there is t_0 after which $1 - q_i \geqslant 1/2$, so for $t \geqslant t_0$,

$$\frac{\mathrm{d}}{\mathrm{d}t}(\boldsymbol{u}_0 - \boldsymbol{u}_i) \geqslant \frac{\gamma}{2p}, \qquad \boldsymbol{u}_0(t) - \boldsymbol{u}_i(t) \geqslant \frac{1}{2p} \int_{t_0}^t \gamma(h)\,\mathrm{d}h + c_1 = \Omega(\log t).$$

Using this in (14) with $w = \gamma \boldsymbol{u}$ and the asymptotic bounds on γ and $\boldsymbol{u}_0 - \boldsymbol{u}_i$,

$$\dot{q}_i = -\gamma q_i \boldsymbol{s}_0 \Big[q_i(\boldsymbol{u}_0 - \boldsymbol{u}_i) + (1 - q_i)\big(\boldsymbol{u}_0 - \langle \boldsymbol{u}, \boldsymbol{s} \rangle\big) \Big] \geqslant -c\,\frac{\gamma}{p}\, q_i^2 \log t \geqslant -c\, q_i^2\, \frac{\log t}{t}.$$

Integrating,

$$-\Big(\frac{1}{q_i(t)} - 1\Big) \geqslant -c\,\frac{(\log t)^2}{2}, \qquad q_i(t) \leqslant \frac{1}{1 + c(\log t)^2/2} = O\Big(\frac{1}{\log^2 t}\Big),$$

which gives (b), as $\boldsymbol{s}_i \leqslant q_i$. For (a), $\dot{\boldsymbol{u}}_i = \gamma \boldsymbol{s}_i \leqslant \gamma q_i \leqslant c/(t \log^2 t)$, so

$$\boldsymbol{u}_i(t) - \boldsymbol{u}_i(t_0) \leqslant c \int_{t_0}^t \frac{\mathrm{d}h}{h \log^2 h} = c\Big(\frac{1}{\log t_0} - \frac{1}{\log t}\Big) \leqslant c'.$$ □

Remark C.2 (Comparison with linear networks). *When $\boldsymbol{\sigma}$ is replaced by the identity map, the model reduces to a linear network, and gradient flow is known to converge to a rank-one factorization (see, e.g., Ji & Telgarsky, 2019). Under the softmax parameterization, $\mathbf{V}(t)$ also becomes approximately rank one (Lemma C.1). The crucial difference is that the attention scores $\boldsymbol{s}(t)$ converge to a one-hot vector. This does not happen in linear networks, and it forces the dominant right singular direction of $\mathbf{V}$ to align with a coordinate axis rather than a generic direction.*

D General Losses and Square-Loss Regression

For a general differentiable loss, Lemma A.2 gives the flow $\dot{\mathbf{V}} = \boldsymbol{r}\boldsymbol{s}^\top$, $\dot{\boldsymbol{a}} = P_{\boldsymbol{s}} w$ with fitness $w = \mathbf{V}^\top \boldsymbol{r} = -\mathbf{V}^\top \nabla \ell(\boldsymbol{\beta}, \boldsymbol{\beta}_*)$, and Lemma A.4 shows that $\dot{\boldsymbol{a}}_k = \boldsymbol{s}_k(w_k - \langle \boldsymbol{s}, w\rangle)$ is a replicator equation. Unlike the logistic case, the direction of $\boldsymbol{r}$ changes during training. We give two sufficient conditions for polarization, an ordering condition on the fitness (Section D.1) and a sign condition on the output gradient (Section D.2), and apply them to the square loss (Sections D.3 and D.4).

D.1 An ordered-fitness criterion

Theorem D.1 (Polarization under ordered fitness). *Assume $\boldsymbol{a}_0(0) > \boldsymbol{a}_1(0) > \cdots > \boldsymbol{a}_{p-1}(0)$, and suppose that on $[0, T]$ the fitness is ordered in the same way, $w_0(t) \geqslant w_1(t) \geqslant \ldots \geqslant w_{p-1}(t)$. Then, for all $t \in [0, T]$:*

(a) Order preservation. *$\boldsymbol{a}_0(t) > \boldsymbol{a}_1(t) > \cdots > \boldsymbol{a}_{p-1}(t)$, and $t \mapsto e^{-\boldsymbol{a}_l(t)} - e^{-\boldsymbol{a}_k(t)}$ is nondecreasing for $k < l$.*

(b) Polarization. *For every $l \geqslant 1$,*

$$\frac{\boldsymbol{s}_l(t)}{\boldsymbol{s}_0(t)} \leqslant \left(\frac{\boldsymbol{s}_0(0)}{\boldsymbol{s}_l(0)} + \frac{1}{p} \int_0^t \big(w_0(h) - w_l(h)\big)\, \mathrm{d}h \right)^{-1}.$$

Proof. (a) By Lemma A.4(a), $\frac{\mathrm{d}}{\mathrm{d}t}\big(e^{-\boldsymbol{a}_l} - e^{-\boldsymbol{a}_k}\big) = (w_k - w_l)/Z \geqslant 0$ for $k < l$, so these gaps stay above their positive initial values. (b) By (a), $\boldsymbol{s}_0$ is the largest score, and w_0 is the largest fitness by assumption. Lemma A.4(d) gives $\frac{\mathrm{d}}{\mathrm{d}t}(1/q_l) \geqslant (w_0 - w_l)/p$; integrate, using $1/q_l(0) = \boldsymbol{s}_0(0)/\boldsymbol{s}_l(0)$. □

Corollary D.2 (Logistic loss; recovers Theorem 3.3(b)). *For the logistic loss (5), $w = \gamma(\boldsymbol{\beta})\, \boldsymbol{u}$ with $\boldsymbol{u} = \mathbf{V}^\top \boldsymbol{\beta}_*$. Under Assumption 3.1, Theorem 3.2(b) gives $\boldsymbol{u}_0 - \boldsymbol{u}_l \geqslant \delta \stackrel{\text{def}}{=} \min_{i<j}(\boldsymbol{u}_i(0) - \boldsymbol{u}_j(0))$ for all t, so Theorem D.1 applies with $T = \infty$ and yields $\boldsymbol{s}_l/\boldsymbol{s}_0 \leqslant \big(1 + \frac{\delta}{p}\int_0^t \gamma\big)^{-1}$.*

D.2 Sign-preserving gradients

Proof of Theorem 3.4. We prove the result for the slightly more general flow $\dot{\mathbf{V}} = -Ag\boldsymbol{s}^\top$, $\dot{\boldsymbol{a}} = -P_{\boldsymbol{s}}\mathbf{V}^\top g$, where $A = \operatorname{diag}((\,) A_0, \ldots, A_{p-1})$ has positive entries. Taking $A = I_p$ gives the theorem; the diagonal version will be used only for the regression change of coordinates below. For the proof, write $b = -g = -\nabla_{\boldsymbol{\beta}} \ell(\boldsymbol{\beta}, \boldsymbol{\beta}_*)$ and $B_i(t) = \int_0^t |b_i(h)|\, \mathrm{d}h$. Thus $\epsilon_i = \operatorname{sign}(b_i(0))$ and $\dot{\mathbf{V}} = Ab\boldsymbol{s}^\top$. Let $w = \mathbf{V}^\top b$, so that $\dot{\boldsymbol{a}} = P_{\boldsymbol{s}} w$ and Lemma A.4 applies. On $[0, T]$ we have $\epsilon_i b_i = |b_i|$ for $i \in I$ and $b_i = 0$ for $i \notin I$. Fix two score coordinates $j < k$, and for each $i \in I$ define

$$D_{jk}^{(i)} = \epsilon_i(\mathbf{V}_{ij} - \mathbf{V}_{ik}), \qquad \Delta_{jk} = e^{-\boldsymbol{a}_k} - e^{-\boldsymbol{a}_j}.$$

Initially $\Delta_{jk}(0) > 0$, whereas every $D_{jk}^{(i)}(0) = 0$ because $\mathbf{V}(0) = 0$. Since $\dot{\mathbf{V}}_{ij} = A_i b_i \boldsymbol{s}_j$ and $w_j - w_k = \sum_{i \in I} b_i(\mathbf{V}_{ij} - \mathbf{V}_{ik})$, Lemma A.4(a,b) gives the exact dynamics on $[0, T]$:

$$\dot{D}_{jk}^{(i)} = A_i |b_i| \frac{e^{\boldsymbol{a}_j + \boldsymbol{a}_k}}{Z} \Delta_{jk}, \qquad \dot{\Delta}_{jk} = \frac{w_j - w_k}{Z} = \frac{1}{Z} \sum_{i \in I} |b_i| D_{jk}^{(i)}. \tag{24}$$

This is a linear cooperative system: its off-diagonal coefficients are nonnegative, so its nonnegative orthant is forward invariant. Hence $D_{jk}^{(i)} \geqslant 0$ and $\Delta_{jk} \geqslant \Delta_{jk}(0) > 0$, so $\boldsymbol{a}_j > \boldsymbol{a}_k$ and $\boldsymbol{s}_j > \boldsymbol{s}_k$

throughout the interval. Whenever $B_i(t) > 0$, integrating the first equation shows $D_{jk}^{(i)}(t) > 0$. Moreover

$$w_j - w_k = \sum_{i \in I} |b_i| D_{jk}^{(i)} \geqslant 0,$$

so the fitness w is ordered. Lemma A.4(c,d) then gives

$$\dot{q}_l = -q_l \boldsymbol{s}_0 \Big[q_l (w_0 - w_l) + (1 - q_l)\big(w_0 - \langle \boldsymbol{s}, w \rangle\big) \Big] \leqslant -\frac{q_l^2}{p}(w_0 - w_l). \tag{25}$$

In particular q_l is nonincreasing, and

$$\frac{\mathrm{d}}{\mathrm{d}t} \frac{1}{q_l} \geqslant \frac{1}{p} \sum_{i \in I} |b_i| D_{0l}^{(i)}. \tag{26}$$

For $t \geqslant t_0$ the row gaps are nondecreasing, so $D_{0l}^{(i)}(t) \geqslant d_l(t_0) > 0$, and $\sum_{i \in I} |b_i| D_{0l}^{(i)} \geqslant d_l(t_0) \|b\|_1$. Integrating (26) proves (8).

For the bound from time zero, $\boldsymbol{s}_0 - \boldsymbol{s}_l = \boldsymbol{s}_0(1 - q_l) \geqslant (1 - q_l(0))/p$, since q_l is nonincreasing and $\boldsymbol{s}_0 \geqslant 1/p$. Therefore

$$D_{0l}^{(i)}(t) = A_i \int_0^t |b_i(h)| \big(\boldsymbol{s}_0(h) - \boldsymbol{s}_l(h)\big)\, \mathrm{d}h \geqslant \frac{A_i(1 - q_l(0))}{p} B_i(t).$$

Substituting this in (26) and using $\int_0^t |b_i(h)| B_i(h)\, \mathrm{d}h = B_i(t)^2/2$ gives the additional estimate

$$\frac{1}{q_l(t)} \geqslant \frac{1}{q_l(0)} + \frac{1 - q_l(0)}{2p^2} \sum_{i \in I} A_i B_i(t)^2. \tag{27}$$

We use this estimate below to express regression polarization in terms of value-row norms. □

Remark D.3 (KL divergence)**.** *For an objective such as $\ell(\boldsymbol{\beta}, p_*) = -\langle p_*, \log \boldsymbol{\beta} \rangle$ on its positive domain, Theorem 3.4 applies to any trajectory whose active gradient coordinates retain their signs. Figure 7 illustrates one such numerical behavior; the sign condition must be checked along the trajectory before invoking the theorem.*

D.3 Square-loss regression

The whitened flow. Gradient flow on (9) reads

$$\dot{\mathbf{V}} = X^\top X(\boldsymbol{\beta}_* - \mathbf{V}\boldsymbol{s})\boldsymbol{s}^\top, \qquad \dot{\boldsymbol{a}} = P_{\boldsymbol{s}} \mathbf{V}^\top X^\top X(\boldsymbol{\beta}_* - \mathbf{V}\boldsymbol{s}).$$

With $X^\top X = U \Lambda U^\top$, multiplying the first equation by $\Lambda^{1/2} U^\top$ and substituting $\mathbf{V} \mapsto \Lambda^{1/2} U^\top \mathbf{V}$, $\boldsymbol{\beta}_* \mapsto \Lambda^{1/2} U^\top \boldsymbol{\beta}_*$ gives the whitened flow

$$\dot{\mathbf{V}} = \Lambda \boldsymbol{r} \boldsymbol{s}^\top, \qquad \dot{\boldsymbol{a}} = P_{\boldsymbol{s}} \mathbf{V}^\top \boldsymbol{r}, \qquad \boldsymbol{r} = \boldsymbol{\beta}_* - \mathbf{V}\boldsymbol{s}. \tag{28}$$

Here $\Lambda = \operatorname{diag}(\lambda_0, \ldots, \lambda_{p-1})$ and we reuse $\mathbf{V}, \boldsymbol{\beta}_*$ for the transformed variables. For $I = \{i : (\boldsymbol{\beta}_*)_i \neq 0\}$, the sign-preserving horizon is precisely

$$\tau = \sup\{T \geqslant 0 : \operatorname{sign}((\boldsymbol{\beta}_*)_i) r_i(t) \geqslant 0 \text{ for all } i \in I,\ 0 \leqslant t \leqslant T\}.$$

This is the diagonal extension proved in Appendix D.2, with $A = \Lambda$ and $g = -\boldsymbol{r}$ (equivalently, $b = \boldsymbol{r}$ in the proof notation). Row by row,

$$\dot{\boldsymbol{v}}^i = \lambda_i \boldsymbol{r}_i \boldsymbol{s}, \qquad \dot{\boldsymbol{r}}_i = -\lambda_i \boldsymbol{r}_i \|\boldsymbol{s}\|^2 - \sum_j \boldsymbol{r}_j \left\langle P_{\boldsymbol{s}} \boldsymbol{v}^i, P_{\boldsymbol{s}} \boldsymbol{v}^j \right\rangle,$$

where $\boldsymbol{r}_i = (\boldsymbol{\beta}_*)_i - \langle \boldsymbol{v}^i, \boldsymbol{s} \rangle$. There is no closed-form solution; the proofs below track signed differences of entries within each row.

Lemma D.4 (Convergence to zero loss)**.** *Let X have full column rank, so $\lambda_{\min} > 0$, and consider the whitened flow (28). Let $L(t) = \mathcal{L}(\mathbf{V}(t), \boldsymbol{a}(t)) = \frac{1}{2}\|\boldsymbol{r}(t)\|^2$ with $\boldsymbol{r} = \boldsymbol{\beta}_* - \mathbf{V}\boldsymbol{s}$. Then*

$$\dot{L}(t) \leqslant -\frac{2\lambda_{\min}}{p} L(t), \qquad \text{so} \qquad L(t) \leqslant L(0) \exp\Big(-\frac{2\lambda_{\min}}{p} t\Big) \xrightarrow[t \to \infty]{} 0.$$

In particular $\boldsymbol{r}(t) \to 0$, $\mathbf{V}(t) \to \mathbf{V}^\star$, and $\boldsymbol{a}(t) \to \boldsymbol{a}^\star$ for finite limits satisfying $\mathbf{V}^\star \boldsymbol{\sigma}(\boldsymbol{a}^\star) = \boldsymbol{\beta}_*$.

Proof. Since $\dot{\boldsymbol{r}} = -\dot{\boldsymbol{\beta}}$, Lemma A.5 with $A = \Lambda$ and $b = \boldsymbol{r}$ gives

$$\dot{L} = -\left\langle \boldsymbol{r}, \dot{\boldsymbol{\beta}} \right\rangle = -\|\boldsymbol{s}\|^2 \boldsymbol{r}^\top \Lambda \boldsymbol{r} - \left\|P_{\boldsymbol{s}} \mathbf{V}^\top \boldsymbol{r}\right\|^2 \leqslant -\frac{\lambda_{\min}}{p}\|\boldsymbol{r}\|^2 = -\frac{2\lambda_{\min}}{p} L.$$

Grönwall's lemma yields the exponential decay. Then $\|\boldsymbol{r}(t)\|$ is integrable, so $\dot{\mathbf{V}} = \Lambda \boldsymbol{r}\boldsymbol{s}^\top$ is integrable and $\mathbf{V}(t) \to \mathbf{V}^\star$. Since $\mathbf{V}(t)$ stays bounded, $\dot{\boldsymbol{a}} = P_{\boldsymbol{s}} \mathbf{V}^\top \boldsymbol{r}$ is also integrable. Hence $\boldsymbol{a}(t) \to \boldsymbol{a}^\star$ and $\boldsymbol{s}(t) \to \boldsymbol{\sigma}(\boldsymbol{a}^\star)$, and $\boldsymbol{r}(t) \to 0$ gives the stated identity. □

Corollary 3.6. *Consider gradient flow for (9) under Assumption 3.5, with X of full column rank and $\boldsymbol{\beta}_* \neq 0$.*

(a) ***Optimization.*** *For all $t \geqslant 0$,*

$$\mathcal{L}\left(\mathbf{V}(t), \boldsymbol{a}(t)\right) \leqslant \mathcal{L}\left(\mathbf{V}(0), \boldsymbol{a}(0)\right) \exp\left(-\frac{2\lambda_{\min}}{p} t\right).$$

(b) ***Eigenvalue dependence.*** *If $\tau = \infty$ and the target $X\boldsymbol{\beta}_*$ has a component of magnitude $m > 0$ along a unit left singular vector of X associated with singular value $\sqrt{\lambda_{\min}}$, then*

$$q_l(\infty) \leqslant \left[\frac{1}{q_l(0)} + \frac{(1 - q_l(0))m^2}{2p^2 \lambda_{\min}}\right]^{-1}.$$

In particular, for fixed initial scores and m bounded away from zero, $q_l(\infty) = O(\lambda_{\min})$.

Proof of Corollary 3.6. Part (a) is Lemma D.4. The loss bound also gives global existence: $\boldsymbol{r}$ stays bounded, hence $\dot{\mathbf{V}} = \Lambda \boldsymbol{r}\boldsymbol{s}^\top$ is bounded, and then $\dot{\boldsymbol{a}} = P_{\boldsymbol{s}} \mathbf{V}^\top \boldsymbol{r}$ is bounded on every finite interval.

For part (b), let $I = \{i : (\boldsymbol{\beta}_*)_i \neq 0\}$. If $i \notin I$, then $\boldsymbol{v}^i(0) = 0$ and $\boldsymbol{r}_i = -\left\langle \boldsymbol{v}^i, \boldsymbol{s} \right\rangle$, and uniqueness of solutions implies $\boldsymbol{v}^i(t) = 0$ and $\boldsymbol{r}_i(t) = 0$ for all t. For $i \in I$, $\boldsymbol{r}_i(0) = (\boldsymbol{\beta}_*)_i \neq 0$, so $\tau > 0$. On $[0, \tau]$, apply the diagonal extension in the proof of Theorem 3.4 with $A = \Lambda$ and $g = -\boldsymbol{r}$ (equivalently, $b = \boldsymbol{r}$ in the proof notation). Writing $B_i(t) = \int_0^t |\boldsymbol{r}_i(h)| \, \mathrm{d}h$, every active $B_i(t_0)$ is positive for $0 < t_0 \leqslant \tau$, so the order and ℓ_1 bounds from that proof apply.

On this interval $\boldsymbol{r}_i$ keeps its initial sign and $\boldsymbol{v}^i(0) = 0$, so every entry of $\boldsymbol{v}^i(t) = \lambda_i \int_0^t \boldsymbol{r}_i(h)\boldsymbol{s}(h)^\top \, \mathrm{d}h$ has that sign. As $\sum_m \boldsymbol{s}_m(h) = 1$,

$$\left\|\boldsymbol{v}^i(t)\right\|_1 = \lambda_i \int_0^t |\boldsymbol{r}_i(h)| \, \mathrm{d}h = \lambda_i B_i(t). \tag{29}$$

Substituting into (27) gives

$$\frac{1}{q_l(t)} \geqslant \frac{1}{q_l(0)} + \frac{1 - q_l(0)}{2p^2} \sum_{i \in I} \frac{\|\boldsymbol{v}^i(t)\|_1^2}{\lambda_i}. \tag{30}$$

When τ is finite, the bound extends to $t = \tau$ by continuity. To obtain the eigenvalue bound, a unit left singular vector u_* associated with $\sqrt{\lambda_{\min}}$ satisfies $|u_*^\top X \boldsymbol{\beta}_*| = |(\Lambda^{1/2} U^\top \boldsymbol{\beta}_*)_{i_*}|$ for the corresponding eigenvector basis; thus the target component in the corollary is exactly the transformed target component used here. Take $\tau = \infty$ and an index i_* with $\lambda_{i_*} = \lambda_{\min}$ and $|(\boldsymbol{\beta}_*)_{i_*}| = m$. Convergence of the fit implies $\|\boldsymbol{v}^{i_*}(\infty)\|_1 \geqslant |\langle \boldsymbol{v}^{i_*}(\infty), \boldsymbol{s}(\infty) \rangle| = m$. Keeping this term in (30) proves the claimed bound. □

Remark D.5 (Eigenvalue dependence)**.** *We justify the eigenvalue claim following Corollary 3.6. Suppose $\tau = \infty$. Then Lemma D.4 gives $\boldsymbol{r}(t) \to 0$ exponentially, so (29) has a finite limit. Also $\left\langle \boldsymbol{v}^i(t), \boldsymbol{s}(t) \right\rangle \to (\boldsymbol{\beta}_*)_i$ and $|\left\langle \boldsymbol{v}^i(t), \boldsymbol{s}(t) \right\rangle| \leqslant \left\|\boldsymbol{v}^i(t)\right\|_1$, whence $\left\|\boldsymbol{v}^{i_*}(\infty)\right\|_1 \geqslant |(\boldsymbol{\beta}_*)_{i_*}|$. The bound (30) implies $q_l(\infty) = O(\lambda_{\min})$ along a family only when this transformed target component and the initial score gap have uniform positive lower bounds. This condition matters: if the original target is fixed, its component after the change of variables can instead be $O(\sqrt{\lambda_{\min}})$.*

The residual sign condition of Corollary 3.6 can be replaced by a weaker condition on the correlations of the residual over time.

Corollary D.6 (Regression under a nonnegative residual kernel)**.** *Consider the whitened flow (28), $\dot{\mathbf{V}} = \Lambda \boldsymbol{r}\boldsymbol{s}^\top$ with $\boldsymbol{r} = \boldsymbol{\beta}_* - \mathbf{V}\boldsymbol{s}$, under Assumption 3.5 ($\mathbf{V}(0) = 0$, $\boldsymbol{a}(0)$ strictly ordered). Define the residual kernel and its horizon*

$$K(t,h) \stackrel{\text{def}}{=} \boldsymbol{r}(h)^\top \Lambda\, \boldsymbol{r}(t), \qquad T_K \stackrel{\text{def}}{=} \sup\{T \geqslant 0 : K(t,h) \geqslant 0 \ \ \forall\, 0 \leqslant h \leqslant t \leqslant T\}.$$

Then $T_K > 0$ whenever $\boldsymbol{\beta}_* \neq 0$, and T_K is at least the first time some coordinate of $\boldsymbol{r}$ changes sign. Moreover:

(a) Vanishing loss. *For all $t \geqslant 0$, $\mathcal{L}(t) \leqslant \mathcal{L}(0)\exp(-2\lambda_{\min} t/p)$.*
(b) Order preservation. *For $t < T_K$, $\boldsymbol{a}_0(t) > \cdots > \boldsymbol{a}_{p-1}(t)$.*
(c) Polarization. *For $t < T_K$ and $l \geqslant 1$,*

$$\frac{\boldsymbol{s}_l(t)}{\boldsymbol{s}_0(t)} \leqslant \left(\frac{\boldsymbol{s}_0(0)}{\boldsymbol{s}_l(0)} + \frac{1}{p}\int_0^t \int_0^{t'} K(t',h)\big(\boldsymbol{s}_0(h) - \boldsymbol{s}_l(h)\big)\,\mathrm{d}h\,\mathrm{d}t' \right)^{-1}.$$

Proof. Since $\mathbf{V}(0) = 0$, we have $\mathbf{V}(t) = \int_0^t \Lambda \boldsymbol{r}(h)\boldsymbol{s}(h)^\top \mathrm{d}h$, and therefore

$$w(t) = \mathbf{V}(t)^\top \boldsymbol{r}(t) = \int_0^t K(t,h)\, \boldsymbol{s}(h)\,\mathrm{d}h.$$

(a) is Lemma D.4. (b, c) Let $t_1 = \sup\{t < T_K : \boldsymbol{a}$ is strictly ordered on $[0,t]\}$. On $[0,t_1)$, $\boldsymbol{s}(h)$ is ordered, so $w(t)$, a nonnegative combination of ordered vectors, is ordered, and $w_0 - w_l = \int_0^t K(t,h)(\boldsymbol{s}_0(h) - \boldsymbol{s}_l(h))\,\mathrm{d}h \geqslant 0$. By Theorem D.1(a), the gaps $e^{-\boldsymbol{a}_l} - e^{-\boldsymbol{a}_k}$ do not decrease on $[0,t_1)$, so $\boldsymbol{a}$ is still strictly ordered at t_1, and by continuity slightly beyond. Hence $t_1 = T_K$, and (c) follows from Theorem D.1(b). (Horizon.) $K(0,0) = \boldsymbol{\beta}_*^\top \Lambda \boldsymbol{\beta}_* > 0$ and K is continuous, so $T_K > 0$. If no coordinate of $\boldsymbol{r}$ has changed sign by time t, every term of $K(t,h) = \sum_i \lambda_i \boldsymbol{r}_i(h)\boldsymbol{r}_i(t)$ is nonnegative. □

Remark D.7 (Kernel form for a general loss)**.** *The kernel form extends to any loss with $\mathbf{V}(0) = 0$: $w(t) = \int_0^t \langle \nabla \ell(\boldsymbol{\beta}(h), \boldsymbol{\beta}_*), \nabla \ell(\boldsymbol{\beta}(t), \boldsymbol{\beta}_*)\rangle\ \boldsymbol{s}(h)\,\mathrm{d}h$. Thus polarization and order preservation hold for as long as gradients at different times stay positively correlated. The invariant of Proposition 3.7 establishes global score order for the square loss when $p = 2$, including trajectories beyond T_K; for larger p, the kernel criterion remains a sufficient condition.*

D.4 Two score coordinates

Proof of Proposition 3.7. Let $d = \boldsymbol{a}_0 - \boldsymbol{a}_1$, $g = \mathbf{V}_0 - \mathbf{V}_1$ (difference of the two columns), and $Q = e^{-\boldsymbol{a}_1} - e^{-\boldsymbol{a}_0}$. By Lemma A.3(b), $\dot{\boldsymbol{a}}_0 + \dot{\boldsymbol{a}}_1 = 0$, so $c = e^{\boldsymbol{a}_0 + \boldsymbol{a}_1}$ is constant. With $w = \mathbf{V}^\top \boldsymbol{r}$, Lemma A.4(a,b) gives

$$\dot{g} = \Lambda \boldsymbol{r}(\boldsymbol{s}_0 - \boldsymbol{s}_1) = \frac{cQ}{Z}\Lambda \boldsymbol{r}, \qquad \dot{Q} = \frac{w_0 - w_1}{Z} = \frac{\boldsymbol{r}^\top g}{Z}.$$

Consequently,

$$\frac{\mathrm{d}}{\mathrm{d}t}\big(g^\top \Lambda^{-1} g\big) = 2\frac{cQ}{Z} g^\top \boldsymbol{r} = c\frac{\mathrm{d}}{\mathrm{d}t}Q^2.$$

Since $g(0) = 0$ and $cQ^2 = 2(\cosh d - 1)$, integration yields

$$g(t)^\top \Lambda^{-1} g(t) = c\big(Q(t)^2 - Q(0)^2\big) = 2\big(\cosh d(t) - \cosh d(0)\big).$$

The left side is nonnegative, so $\cosh d(t) \geqslant \cosh d(0)$. As $d(0) > 0$ and d is continuous, d cannot reach the branch $d \leqslant -d(0)$ without passing through values with $\cosh d < \cosh d(0)$. Thus $d(t) \geqslant d(0) > 0$ for all t. The argument places no condition on the signs of the $\boldsymbol{r}_i(t)$. □

E Other Normalizations

We consider normalizations other than softmax and show that the no-crossing property persists. Let f be monotonically increasing and differentiable, and define $\boldsymbol{\sigma}_f : \mathbb{R}^p \to \mathbb{R}^p$ by

$$\boldsymbol{\sigma}_f(\boldsymbol{a})_i = \frac{f(\boldsymbol{a}_i)}{Z_f(\boldsymbol{a})}, \qquad Z_f(\boldsymbol{a}) = \sum_{j\in[p]} f(\boldsymbol{a}_j),$$

with $\boldsymbol{s}_f = \boldsymbol{\sigma}_f(\boldsymbol{a})$. For the logistic loss, with $\boldsymbol{u} = \mathbf{V}^\top \boldsymbol{\beta}_*$ as in Section C, the gradient flow is

$$\mathrm{d}\boldsymbol{u} = \gamma(\boldsymbol{\beta})\big\|\boldsymbol{\beta}_*\big\|^2 \boldsymbol{s}_f \,\mathrm{d}t, \qquad \mathrm{d}\boldsymbol{a} = \gamma(\boldsymbol{\beta})\frac{f'(\boldsymbol{a})}{Z_f(\boldsymbol{a})} \odot \big(\boldsymbol{u} - \langle \boldsymbol{u}, \boldsymbol{s}_f\rangle \,\mathbf{1}\big)\mathrm{d}t, \tag{31}$$

where $f'(\boldsymbol{a})$ is applied coordinatewise.

Lemma E.1 (No crossing for general normalizations)**.** *Consider the gradient flow* (31) *for the logistic loss $\ell(\boldsymbol{\beta}, \boldsymbol{\beta}_*)$ under Assumption 3.1. Then, for all $t \geqslant 0$,*

$$\boldsymbol{u}_0(t) > \boldsymbol{u}_1(t) > \cdots > \boldsymbol{u}_{p-1}(t), \qquad \boldsymbol{a}_0(t) \geqslant \boldsymbol{a}_1(t) \geqslant \cdots \geqslant \boldsymbol{a}_{p-1}(t).$$

Proof. We write the proof for $\big\|\boldsymbol{\beta}_*\big\| = 1$; a general norm only multiplies the second term of $\mathrm{d}\Phi_{ij}$ below by $\big\|\boldsymbol{\beta}_*\big\|^2$. Write $Z_f = Z_f(\boldsymbol{a})$, and let $G(x) = \int_0^x \frac{\mathrm{d}v}{f'(v)}$. Coordinatewise, (31) reads

$$\mathrm{d}\boldsymbol{u}_i = \frac{\gamma}{Z_f} f(\boldsymbol{a}_i)\,\mathrm{d}t, \qquad \mathrm{d}G(\boldsymbol{a}_i) = \frac{\mathrm{d}\boldsymbol{a}_i}{f'(\boldsymbol{a}_i)} = \frac{\gamma}{Z_f}\big(\boldsymbol{u}_i - \langle \boldsymbol{u}, \boldsymbol{s}_f\rangle\big)\mathrm{d}t.$$

For any i, j, taking differences,

$$\mathrm{d}\big(G(\boldsymbol{a}_i) - G(\boldsymbol{a}_j)\big) = \frac{\gamma}{Z_f}(\boldsymbol{u}_i - \boldsymbol{u}_j)\,\mathrm{d}t, \qquad \mathrm{d}(\boldsymbol{u}_i - \boldsymbol{u}_j) = \frac{\gamma}{Z_f}\big(f(\boldsymbol{a}_i) - f(\boldsymbol{a}_j)\big)\mathrm{d}t.$$

The potential $\Phi_{ij} = \big(G(\boldsymbol{a}_i) - G(\boldsymbol{a}_j)\big)(\boldsymbol{u}_i - \boldsymbol{u}_j)$ therefore satisfies

$$\mathrm{d}\Phi_{ij} = \frac{\gamma}{Z_f}\Big[(\boldsymbol{u}_i - \boldsymbol{u}_j)^2 + \big(G(\boldsymbol{a}_i) - G(\boldsymbol{a}_j)\big)\big(f(\boldsymbol{a}_i) - f(\boldsymbol{a}_j)\big)\Big]\mathrm{d}t.$$

Since f is increasing, $f' > 0$, so G is increasing as well. We assume that it does not hit 0 in finite time, so the second term in the bracket is nonnegative. Hence $\mathrm{d}\Phi_{ij} \geqslant 0$, the potential Φ_{ij} does not decrease, and, as in the proof of Theorem 3.2, the no-crossing property holds for all $t \geqslant 0$. □

F Polarization in the Tied Value–Softmax Model

This section proves Proposition 4.1 (Proposition F.3) and extends it to rank-one orthogonal perturbations (Proposition F.4). Throughout, ℓ is the logistic loss (20) with $\big\|\boldsymbol{\beta}_*\big\| = 1$, and

$$\boldsymbol{\beta} = \mathbf{R}\boldsymbol{\sigma}(\mathbf{R}^\top \boldsymbol{a}), \qquad \boldsymbol{s} = \boldsymbol{\sigma}(\mathbf{R}^\top \boldsymbol{a}), \qquad \boldsymbol{v} = \mathbf{R}^\top \boldsymbol{\beta}_*, \qquad \gamma = \big(1 + e^{\langle \boldsymbol{\beta}_*, \boldsymbol{\beta}\rangle}\big)^{-1}.$$

We decompose the parameters into components parallel and orthogonal to $\boldsymbol{\beta}_*$:

$$\mathbf{R} = \boldsymbol{\beta}_* \boldsymbol{v}^\top + \mathbf{R}_\perp, \qquad \boldsymbol{a} = c\,\boldsymbol{\beta}_* + \boldsymbol{a}_\perp, \qquad c = \langle \boldsymbol{\beta}_*, \boldsymbol{a}\rangle,$$

with $\mathbf{R}_\perp = (\mathrm{I} - \boldsymbol{\beta}_*\boldsymbol{\beta}_*^\top)\mathbf{R}$ and $\boldsymbol{a}_\perp = (\mathrm{I} - \boldsymbol{\beta}_*\boldsymbol{\beta}_*^\top)\boldsymbol{a}$.

Lemma F.1 (Tied dynamics)**.** *Gradient flow on $\ell(\mathbf{R}\boldsymbol{\sigma}(\mathbf{R}^\top \boldsymbol{a}), \boldsymbol{\beta}_*)$ reads*

$$\dot{\mathbf{R}} = \gamma\big[\boldsymbol{\beta}_* \boldsymbol{s}^\top + \boldsymbol{a}(P_{\boldsymbol{s}}\boldsymbol{v})^\top\big], \qquad \dot{\boldsymbol{a}} = \gamma\,\mathbf{R}P_{\boldsymbol{s}}\boldsymbol{v}.$$

In the decomposition above,

$$\dot{\boldsymbol{v}} = \gamma\big[\boldsymbol{s} + cP_{\boldsymbol{s}}\boldsymbol{v}\big], \qquad \dot{c} = \gamma\,\boldsymbol{v}^\top P_{\boldsymbol{s}}\boldsymbol{v},$$
$$\dot{\mathbf{R}}_\perp = \gamma\,\boldsymbol{a}_\perp(P_{\boldsymbol{s}}\boldsymbol{v})^\top, \qquad \dot{\boldsymbol{a}}_\perp = \gamma\,\mathbf{R}_\perp P_{\boldsymbol{s}}\boldsymbol{v},$$

the matrix $\mathbf{R}_\perp\mathbf{R}_\perp^\top - \boldsymbol{a}_\perp\boldsymbol{a}_\perp^\top$ is conserved, and $S_v \overset{\text{def}}{=} \mathbf{1}^\top \boldsymbol{v}$ satisfies $\dot{S}_v = \gamma$.

Proof. A variation $(\mathrm{d}\mathbf{R}, \mathrm{d}\boldsymbol{a})$ changes the output by $\mathrm{d}\boldsymbol{\beta} = \mathrm{d}\mathbf{R}\,\boldsymbol{s} + \mathbf{R}P_{\boldsymbol{s}}(\mathrm{d}\mathbf{R}^\top \boldsymbol{a} + \mathbf{R}^\top \mathrm{d}\boldsymbol{a})$. Pairing with $-\nabla\ell(\boldsymbol{\beta}, \boldsymbol{\beta}_*) = \gamma\boldsymbol{\beta}_*$ and using $\mathbf{R}^\top\boldsymbol{\beta}_* = \boldsymbol{v}$,

$$-\mathrm{d}\ell = \gamma\big[\langle \boldsymbol{\beta}_*\boldsymbol{s}^\top + \boldsymbol{a}(P_{\boldsymbol{s}}\boldsymbol{v})^\top, \mathrm{d}\mathbf{R}\rangle + \langle \mathbf{R}P_{\boldsymbol{s}}\boldsymbol{v}, \mathrm{d}\boldsymbol{a}\rangle\big],$$

which gives the flow. Multiplying by $\boldsymbol{\beta}_*^\top$ gives the equations for $\boldsymbol{v}$ and c, and applying $\mathrm{I} - \boldsymbol{\beta}_*\boldsymbol{\beta}_*^\top$ gives those for $\mathbf{R}_\perp$ and $\boldsymbol{a}_\perp$. The derivative of $\mathbf{R}_\perp\mathbf{R}_\perp^\top - \boldsymbol{a}_\perp\boldsymbol{a}_\perp^\top$ is $\gamma\big[\boldsymbol{a}_\perp(P_{\boldsymbol{s}}\boldsymbol{v})^\top\mathbf{R}_\perp^\top + \mathbf{R}_\perp P_{\boldsymbol{s}}\boldsymbol{v}\boldsymbol{a}_\perp^\top - \mathbf{R}_\perp P_{\boldsymbol{s}}\boldsymbol{v}\boldsymbol{a}_\perp^\top - \boldsymbol{a}_\perp(P_{\boldsymbol{s}}\boldsymbol{v})^\top\mathbf{R}_\perp^\top\big] = 0$. Finally, $\mathbf{1}^\top\boldsymbol{s} = 1$ and $\mathbf{1}^\top P_{\boldsymbol{s}} = 0$ give $\dot{S}_v = \gamma$. □

Lemma F.2 (Divergence of S_v and $\int\gamma\,\mathrm{d}t$)**.** *If the logits and the components of $\boldsymbol{v}$ are ordered the same way throughout training, then:*

(a) $S_v(t) \to \infty$ *as* $t \to \infty$*;*

(b) $\displaystyle\int_0^\infty \gamma(t)\,\mathrm{d}t = \infty.$

Proof. Since $\dot{S}_v = \gamma$ and $\gamma = (1 + e^m)^{-1}$ with $m = \langle \boldsymbol{s}, \boldsymbol{v}\rangle$, the weighted-mean inequality gives $m > S_v/p$. Hence $\dot{S}_v < (1 + e^{S_v/p})^{-1}$. Integrating both sides,

$$\int_{S_v(0)}^{S_v(T)} \big(1 + e^{S/p}\big)\,\mathrm{d}S > T \quad\Longrightarrow\quad p\,e^{S_v(T)/p} \gtrsim T,$$

so $S_v(T) \gtrsim p\ln(T/p) \to \infty$. Part (b) follows since $\int_0^\infty \gamma\,\mathrm{d}t = \int_0^\infty \dot{S}_v\,\mathrm{d}t = S_v(\infty) - S_v(0) = \infty$. □

F.1 Parallel initialization

Proposition F.3 (Polarization from a parallel initialization)**.** *Suppose $\mathbf{R}_\perp(0) = \boldsymbol{a}_\perp(0) = 0$ and $c(0) > 0$. Assume the initial logits $\boldsymbol{z}(0) = c(0)\boldsymbol{v}(0)$ are sorted with a unique maximizer at coordinate 0. Then $\boldsymbol{s}(t) \to e_0$ as $t \to \infty$.*

Up to the labeling of the non-maximal coordinates, Proposition F.3 is Proposition 4.1.

Proof. Invariant. By Lemma F.1, the condition $\mathbf{R}_\perp = \boldsymbol{a}_\perp = 0$ is preserved for all t. Thus $\boldsymbol{z} \overset{\text{def}}{=} \mathbf{R}^\top\boldsymbol{a} = c\boldsymbol{v}$ throughout. With $m = \langle \boldsymbol{s}, \boldsymbol{z}\rangle$ and $\nu = \boldsymbol{z}^\top P_{\boldsymbol{s}}\boldsymbol{z}$, the evolution is

$$\dot{\boldsymbol{z}} = \gamma\Big[c\,\boldsymbol{s} + cP_{\boldsymbol{s}}\boldsymbol{z} + \frac{\nu}{c^3}\boldsymbol{z}\Big], \qquad \dot{c} = \frac{\gamma\nu}{c^2}.$$

Since $\nu \geqslant 0$, c is nondecreasing; write $c_0 = c(0) > 0$.

The gap equation. For $j \neq 0$, define the gap $\xi_j = \boldsymbol{z}_j - \boldsymbol{z}_0 \leqslant 0$, set $\bar{\xi} = \sum_{k\neq 0}\boldsymbol{s}_k\xi_k$, and let $\mathcal{A} = 1 + \sum_{k\neq 0} e^{\xi_k}$, so that $\boldsymbol{s}_0 = \mathcal{A}^{-1}$ and $\boldsymbol{s}_j = e^{\xi_j}/\mathcal{A}$. Using $\dot{\boldsymbol{z}}_i = \gamma\big[c\boldsymbol{s}_i + c\boldsymbol{s}_i(\boldsymbol{z}_i - m) + \frac{\nu}{c^3}\boldsymbol{z}_i\big]$ and $m = \boldsymbol{z}_0 + \bar{\xi}$,

$$\dot{\xi}_j = \gamma\Big[c\big(\boldsymbol{s}_j(1 + \xi_j - \bar{\xi}) - \boldsymbol{s}_0(1 - \bar{\xi})\big) + \frac{\nu}{c^3}\xi_j\Big]. \tag{32}$$

The gaps are nonincreasing. Since $\xi_j \leqslant 0$, we have $\boldsymbol{s}_j \leqslant \boldsymbol{s}_0$ and $\bar{\xi} \leqslant 0$. Therefore

$$\boldsymbol{s}_j(1 + \xi_j - \bar{\xi}) \leqslant \boldsymbol{s}_j \leqslant \boldsymbol{s}_0 \leqslant \boldsymbol{s}_0(1 - \bar{\xi}),$$

so the bracketed c-term in (32) is nonpositive. The term $\frac{\nu}{c^3}\xi_j$ is nonpositive as well. Hence $\dot{\xi}_j \leqslant 0$, with strict inequality whenever $\xi_j < 0$: gaps that are initially negative remain negative.

The gaps diverge. Using $1 - \bar{\xi} \geqslant 1$, $e^{\xi_j}\xi_j \leqslant 0$, and $\mathcal{A} \geqslant 1$,

$$\boldsymbol{s}_j(1 + \xi_j - \bar{\xi}) - \boldsymbol{s}_0(1 - \bar{\xi}) = \frac{(1 - \bar{\xi})(e^{\xi_j} - 1) + e^{\xi_j}\xi_j}{\mathcal{A}} \leqslant \frac{e^{\xi_j} - 1}{\mathcal{A}} \leqslant e^{\xi_j} - 1.$$

Combined with $\frac{\nu}{c^3}\xi_j \leqslant 0$ and $c \geqslant c_0$, (32) gives

$$\dot{\xi}_j \leqslant \gamma c_0\big(e^{\xi_j} - 1\big).$$

Let h solve $\dot{h} = \gamma c_0(e^h - 1)$ with $h(0) = \xi_j(0) < 0$. By the comparison principle $\xi_j(t) \leqslant h(t)$, and separation of variables yields

$$h(t) = -\log\Big(1 + \big(e^{-\xi_j(0)} - 1\big)\exp\Big(c_0 \int_0^t \gamma(u)\,\mathrm{d}u\Big)\Big).$$

Since $e^{-\xi_j(0)} > 1$ and $\int_0^\infty \gamma\,\mathrm{d}t = \infty$ (Lemma F.2), $h(t) \to -\infty$, hence $\xi_j(t) \to -\infty$.

Therefore, for $j \neq 0$, $\boldsymbol{s}_j = e^{\xi_j}/\mathcal{A} \leqslant e^{\xi_j} \to 0$, so $\boldsymbol{s}_0 = 1 - \sum_{j\neq 0} \boldsymbol{s}_j \to 1$, that is, $\boldsymbol{s}(t) \to e_0$. □

F.2 Rank-one orthogonal component

We now allow nonzero orthogonal components but assume that the conserved quantity $\mathbf{R}_\perp \mathbf{R}_\perp^\top - \boldsymbol{a}_\perp \boldsymbol{a}_\perp^\top$ vanishes, as it does approximately for small-scale initializations. Then $\mathbf{R}_\perp$ is a rank-one matrix aligned with $\boldsymbol{a}_\perp$. We show that polarization persists and that the orthogonal part becomes negligible compared with the parallel part $\boldsymbol{\beta}_* \boldsymbol{v}^\top$ as $\|\boldsymbol{v}\| \to \infty$. We rescale time by

$$\tau(t) = \int_0^t \gamma(u)\,\mathrm{d}u,$$

and write $'$ for differentiation in τ. By Lemma F.2, $\tau(t) \to \infty$.

Proposition F.4 (Polarization with a rank-one perturbation)**.** *Assume*

$$\mathbf{R}_\perp \mathbf{R}_\perp^\top - \boldsymbol{a}_\perp \boldsymbol{a}_\perp^\top = 0, \qquad \boldsymbol{a}_\perp(0) \neq 0,$$

and write $\mathbf{R}_\perp = \boldsymbol{a}_\perp \boldsymbol{\omega}^\top$ *with* $\|\boldsymbol{\omega}\| = 1$ *and* $\rho = \|\boldsymbol{a}_\perp\|^2$. *Suppose* $c(0) > 0$, $\|\boldsymbol{v}(0)\| + \rho(0) \leqslant \varepsilon$, *and the initial logits* $\boldsymbol{z}(0) = c(0)\boldsymbol{v}(0) + \rho(0)\boldsymbol{\omega}(0)$ *have a unique maximizer at coordinate* 0, *that is,*

$$\boldsymbol{z}_0(0) > \boldsymbol{z}_1(0) > \cdots > \boldsymbol{z}_j(0) > \cdots, \qquad \boldsymbol{v}_0(0) > \boldsymbol{v}_1(0) > \cdots > \boldsymbol{v}_j(0) > \cdots.$$

Assume furthermore that the initial differences in the parallel components dominate the orthogonal perturbation: for every $i, j \neq 0$,

$$c(0)\big(\boldsymbol{v}_i(0) - \boldsymbol{v}_j(0)\big) \geqslant 4\rho(0)\,|\boldsymbol{\omega}_i(0) - \boldsymbol{\omega}_j(0)|.$$

Then, for ε *sufficiently small,* $\boldsymbol{s}(\tau) \to e_0$. *Moreover,* $\rho(\tau)$ *and* $\boldsymbol{\omega}(\tau)$ *converge,* $\|\boldsymbol{v}(\tau)\| \to \infty$, *and*

$$\frac{\|\mathbf{R}_\perp(\tau)\|_F}{\|\boldsymbol{\beta}_* \boldsymbol{v}(\tau)^\top\|_F} = \frac{\sqrt{\rho(\tau)}}{\|\boldsymbol{v}(\tau)\|} \longrightarrow 0.$$

Proof. Invariant. The invariant $\mathbf{R}_\perp \mathbf{R}_\perp^\top = \boldsymbol{a}_\perp \boldsymbol{a}_\perp^\top$ implies $\operatorname{rank}(\mathbf{R}_\perp) \leqslant 1$, so $\mathbf{R}_\perp = \boldsymbol{a}_\perp \boldsymbol{\omega}^\top$ with $\|\boldsymbol{\omega}\| = 1$, and the pre-softmax scores are

$$\boldsymbol{z} = \mathbf{R}^\top \boldsymbol{a} = c\boldsymbol{v} + \mathbf{R}_\perp^\top \boldsymbol{a}_\perp = c\boldsymbol{v} + \rho\boldsymbol{\omega}.$$

By Lemma F.1, in τ-time the full system reads

$$\begin{aligned} \boldsymbol{v}' &= \boldsymbol{s} + cP_{\boldsymbol{s}}\boldsymbol{v}, \qquad & c' &= \nu_v, \\ \rho' &= 2\rho\, C_{\omega v}, \qquad & \boldsymbol{\omega}' &= (\mathrm{I} - \boldsymbol{\omega}\boldsymbol{\omega}^\top) P_{\boldsymbol{s}}\boldsymbol{v}, \end{aligned} \tag{33}$$

with $\boldsymbol{s} = \boldsymbol{\sigma}(\boldsymbol{z})$, $\nu_v = \boldsymbol{v}^\top P_{\boldsymbol{s}}\boldsymbol{v} \geqslant 0$, and $C_{\omega v} = \boldsymbol{\omega}^\top P_{\boldsymbol{s}}\boldsymbol{v}$.

Idea of the proof. The main point is that $\boldsymbol{v}$ and ρ grow on different time scales: $\|\boldsymbol{v}'\| = O(1)$ while $|\rho'| = O(\rho)$. Hence $\boldsymbol{v}$ grows linearly and ρ logarithmically. Before ρ grows, polarization makes the $C_{\omega v}$ term negligible, so the parallel components dominate their orthogonal counterparts.

The differences of the logits. Differentiating $\boldsymbol{z}_i = c\boldsymbol{v}_i + \rho\boldsymbol{\omega}_i$ along (33),

$$\boldsymbol{z}_i' = c\boldsymbol{s}_i + (c^2 + \rho)(P_{\boldsymbol{s}}\boldsymbol{v})_i + \nu_v \boldsymbol{v}_i + \rho C_{\omega v}\boldsymbol{\omega}_i. \tag{34}$$

The differences of the parallel components are $\boldsymbol{v}_i' - \boldsymbol{v}_0' = (\boldsymbol{s}_i - \boldsymbol{s}_0) + c\big[(P_{\boldsymbol{s}}\boldsymbol{v})_i - (P_{\boldsymbol{s}}\boldsymbol{v})_0\big]$. For $j \neq 0$, set $\xi_j = \boldsymbol{z}_j - \boldsymbol{z}_0$, $\pi_j = \boldsymbol{v}_j - \boldsymbol{v}_0$, $\delta_j = \boldsymbol{\omega}_j - \boldsymbol{\omega}_0$, and $\bar{\pi} = \sum_{k\neq 0} \boldsymbol{s}_k \pi_k$. Using the identity $(P_{\boldsymbol{s}}\boldsymbol{v})_j - (P_{\boldsymbol{s}}\boldsymbol{v})_0 = \boldsymbol{s}_j(\pi_j - \bar{\pi}) + \boldsymbol{s}_0\bar{\pi}$, (34) gives

$$\xi_j' = (\boldsymbol{s}_j - \boldsymbol{s}_0)\big[c - (c^2 + \rho)\bar{\pi}\big] + \big[(c^2 + \rho)\boldsymbol{s}_j + \nu_v\big]\pi_j + \rho C_{\omega v}\delta_j. \tag{35}$$

The ordered region. Define the set

$$\mathcal{O} = \big\{\tau \geqslant 0 : \xi_j(\tau') < 0 \text{ and } \pi_j(\tau') < 0 \text{ for all } j \neq 0,\ \tau' \in [0,\tau]\big\}.$$

On $\mathcal{O}$, $\boldsymbol{s}_j < \boldsymbol{s}_0$ and $\bar{\pi} < 0$, so $c - (c^2+\rho)\bar{\pi} > c > 0$. Hence the first term in (35) is strictly negative, and the second is nonpositive since $\pi_j < 0$. Only the perturbation $\rho C_{\omega v}\delta_j$ is unsigned, with

$$|\rho C_{\omega v}\delta_j| \leqslant \rho\big\|P_{\boldsymbol{s}}\boldsymbol{v}\big\|\,|\delta_j| \leqslant \tfrac{1}{2}\rho\big\|\tilde{\boldsymbol{v}}\big\|\,|\delta_j|, \tag{36}$$

where $\tilde{\boldsymbol{v}} = \boldsymbol{v} - \mu\mathbf{1}$ with $\mu = \frac{1}{p}\mathbf{1}^\top\boldsymbol{v}$, using $P_{\boldsymbol{s}}\boldsymbol{v} = P_{\boldsymbol{s}}\tilde{\boldsymbol{v}}$ and $\big\|P_{\boldsymbol{s}}\big\|_{\mathrm{op}} \leqslant \frac{1}{2}$.

Linear growth of the mean. Since $\mathbf{1}^\top\boldsymbol{s} = 1$ and $\mathbf{1}^\top P_{\boldsymbol{s}} = 0$,

$$\mu' = \tfrac{1}{p}, \qquad \mu(\tau) = \mu(0) + \tfrac{\tau}{p}, \tag{37}$$

so $\big\|\boldsymbol{v}(\tau)\big\| \to \infty$.

The gaps stay negative on an initial interval. On $\mathcal{O}$, the centered and orthogonal variables satisfy

$$\big\|\tilde{\boldsymbol{v}}'\big\| \leqslant c\big\|\tilde{\boldsymbol{v}}\big\| + \tfrac{1}{2}\rho, \quad 0 \leqslant c' \leqslant \tfrac{1}{2}\big\|\tilde{\boldsymbol{v}}\big\|^2, \quad |(\log\rho)'| \leqslant \big\|\tilde{\boldsymbol{v}}\big\|, \quad \big\|\boldsymbol{\omega}'\big\| \leqslant \tfrac{1}{2}\big\|\tilde{\boldsymbol{v}}\big\|. \tag{38}$$

Using $\big\|\tilde{\boldsymbol{v}}(0)\big\| + \rho(0) \leqslant \varepsilon$ and Grönwall's inequality on (38), there exist $T_0 = T_0(c(0)) > 0$ and a constant K independent of ε such that, for $\tau \in [0, T_0]$,

$$\big\|\tilde{\boldsymbol{v}}(\tau)\big\| \leqslant K\varepsilon, \quad \rho(\tau) = \rho(0)(1+O(\varepsilon)), \quad c(\tau) = c(0) + O(\varepsilon^2), \quad |\delta_j(\tau) - \delta_j(0)| \leqslant K\varepsilon\tau.$$

By the initial dominance assumption,

$$-c(0)\pi_j(0) \geqslant 4\rho(0)|\delta_j(0)|, \qquad j \neq 0. \tag{39}$$

Combined with (36) and the bounds above, the perturbation $|\rho C_{\omega v}\delta_j| = O(\varepsilon\rho(0))$ is dominated on $[0, T_0]$ by the strictly negative drift $(\boldsymbol{s}_j - \boldsymbol{s}_0)[c - (c^2+\rho)\bar{\pi}] \leqslant -\frac{1}{2}c_0|\boldsymbol{s}_j - \boldsymbol{s}_0| < 0$, so ξ_j remains strictly negative. Likewise π_j remains strictly negative, since $|\pi_j'| \leqslant \big\|\tilde{\boldsymbol{v}}'\big\|$ (the drift of μ is common to all coordinates). Hence $[0, T_0] \subset \mathcal{O}$ for ε small. A similar analysis can also be done for the difference of any two non-maximal coordinates, showing that the orderings of the $\boldsymbol{v}_j$ and of the scores are preserved on $[0, T_0]$.

Polarization. On $\mathcal{O}$, dropping the nonpositive second term in (35) and applying the algebraic bound used in the proof of Proposition F.3,

$$(\boldsymbol{s}_j - \boldsymbol{s}_0)\big[c - (c^2+\rho)\bar{\pi}\big] \leqslant c\,(\boldsymbol{s}_j - \boldsymbol{s}_0) \leqslant c_0\big(e^{\xi_j} - 1\big),$$

where the last step uses $\boldsymbol{s}_j - \boldsymbol{s}_0 = (e^{\xi_j} - 1)/\mathcal{A} \leqslant e^{\xi_j} - 1$ and $c \geqslant c_0$. Therefore

$$\xi_j' \leqslant c_0\big(e^{\xi_j} - 1\big) + \rho\big\|P_{\boldsymbol{s}}\boldsymbol{v}\big\|\,|\delta_j|. \tag{40}$$

Set $R(\tau) = \rho(\tau)\big\|P_{\boldsymbol{s}}\boldsymbol{v}(\tau)\big\|\,|\delta_j(\tau)|$. By (38) and (37), on $\mathcal{O}$ we have $\rho = O(1)$, $|\delta_j| = O(1)$, and $\big\|P_{\boldsymbol{s}}\boldsymbol{v}\big\| \leqslant \frac{1}{2}\big\|\tilde{\boldsymbol{v}}\big\|$, where $\big\|\tilde{\boldsymbol{v}}'\big\| \leqslant c\big\|\tilde{\boldsymbol{v}}\big\| + \frac{1}{2}\rho$. The bootstrap closes globally because the resulting polarization drives $\big\|P_{\boldsymbol{s}}\boldsymbol{v}\big\|$ to zero exponentially: once $\xi_j(\tau) \leqslant -1$ for all $j \neq 0$, we have $\big\|P_{\boldsymbol{s}}\big\|_{\mathrm{op}} \leqslant C\kappa$ with $\kappa = 1 - \boldsymbol{s}_0$, and a comparison of (40) with $\dot{h} = c_0(e^h - 1)$ yields, for some $\lambda > 0$,

$$\xi_j(\tau) \leqslant -\lambda\tau + O(1), \qquad \boldsymbol{s}_j(\tau) \leqslant Ce^{-\lambda\tau}, \qquad j \neq 0. \tag{41}$$

Combined with $\big\|\boldsymbol{v}(\tau)\big\| = O(1+\tau)$,

$$\big\|P_{\boldsymbol{s}}\boldsymbol{v}(\tau)\big\| \leqslant C\kappa(\tau)\big\|\boldsymbol{v}(\tau)\big\| \leqslant C(1+\tau)e^{-\lambda\tau} \in L^1([0,\infty)), \tag{42}$$

so $\int_0^\infty R(\tau)\,\mathrm{d}\tau < \infty$. The perturbed comparison (40) therefore gives $\xi_j(\tau) \to -\infty$, and $\mathcal{O} = [0,\infty)$.

Convergence of the orthogonal variables. By (42) and the equations for ρ and $\boldsymbol{\omega}$ in (33),

$$\int_0^\infty |(\log\rho)'|\,\mathrm{d}\tau \leqslant 2\int_0^\infty \big\|P_{\boldsymbol{s}}\boldsymbol{v}\big\|\,\mathrm{d}\tau < \infty, \qquad \int_0^\infty \big\|\boldsymbol{\omega}'\big\|\,\mathrm{d}\tau \leqslant \int_0^\infty \big\|P_{\boldsymbol{s}}\boldsymbol{v}\big\|\,\mathrm{d}\tau < \infty.$$

Hence $\rho(\tau) \to \rho_\infty \in (0,\infty)$ and $\boldsymbol{\omega}(\tau) \to \boldsymbol{\omega}_\infty$ with $\big\|\boldsymbol{\omega}_\infty\big\| = 1$.

Dominance of the parallel component. Since $\sqrt{\rho(\tau)} = O(1)$ and $\big\|\boldsymbol{v}(\tau)\big\| \to \infty$ by (37),

$$\frac{\big\|\mathbf{R}_\perp(\tau)\big\|_F}{\big\|\boldsymbol{\beta}_*\boldsymbol{v}(\tau)^\top\big\|_F} = \frac{\sqrt{\rho(\tau)}}{\big\|\boldsymbol{v}(\tau)\big\|} \to 0,$$

and $\boldsymbol{z} = c\boldsymbol{v} + \rho\boldsymbol{\omega} = c\boldsymbol{v} + o(\big\|\boldsymbol{v}\big\|)$, so the asymptotic scores and dynamics are governed by $\boldsymbol{v}$ alone. □

G Additional Experimental Details

Both in the experiment with attention sinks and in the classification experiment we test multiple attention versions following Gu et al. (2025). All attention variants that we use are listed in Table 1.

Name	$\mathrm{sim}(\phi(\mathbf{q}_i), \phi(\mathbf{k}_j))$	Normalization
softmax	$\exp\left(\frac{\mathbf{q}_i \mathbf{k}_j^\top}{\sqrt{d_h}}\right)$	$\sum_{j'=1}^{m} \exp\left(\frac{\mathbf{q}_i \mathbf{k}_{j'}^\top}{\sqrt{d_h}}\right)$
sigmoid	$\sigma\left(\frac{\mathbf{q}_i \mathbf{k}_j^\top}{\sqrt{d_h}}\right)$	$\sum_{j'=1}^{m} \sigma\left(\frac{\mathbf{q}_i \mathbf{k}_{j'}^\top}{\sqrt{d_h}}\right)$
elu	$\mathrm{elu}\left(\frac{\mathbf{q}_i \mathbf{k}_j^\top}{\sqrt{d_h}}\right) + 1$	$\sum_{j'=1}^{m} \left(\mathrm{elu}\left(\frac{\mathbf{q}_i \mathbf{k}_{j'}^\top}{\sqrt{d_h}}\right) + 1\right)$
elu · elu	$\frac{\left(\mathrm{elu}(\mathbf{q}_i) + 1\right)\left(\mathrm{elu}(\mathbf{k}_j) + 1\right)^\top}{\sqrt{d_h}}$	$\sum_{j'=1}^{m} \frac{\left(\mathrm{elu}(\mathbf{q}_i) + 1\right)\left(\mathrm{elu}(\mathbf{k}_{j'}) + 1\right)^\top}{\sqrt{d_h}}$
mlp	$\frac{\mathrm{mlp}(\mathbf{q}_i)\,\mathrm{mlp}(\mathbf{k}_j)^\top}{\sqrt{d_h}}$	$\max\left(\left\lvert \sum_{j'=1}^{m} \frac{\mathrm{mlp}(\mathbf{q}_i)\,\mathrm{mlp}(\mathbf{k}_{j'})^\top}{\sqrt{d_h}} \right\rvert, 1\right)$
linear	$\frac{\mathbf{q}_i \mathbf{k}_j^\top}{\sqrt{d_h}}$	$\max\left(\left\lvert \sum_{j'=1}^{m} \frac{\mathbf{q}_i \mathbf{k}_{j'}^\top}{\sqrt{d_h}} \right\rvert, 1\right)$

Table 1: Attention functions we test in Section 4. When computing the normalization, m = sequence length for non-causal models and $m = i$ for causal models. For all functions but softmax, we test their variants both with and without normalization.

G.1 Induction heads acting as attention sinks

Task format. The structure of each object in the dataset is illustrated in Figure 5. Each object consists of a beginning-of-sequence token, k randomly sampled tokens from the vocabulary $t_1, \ldots, t_k$, a token t'_i uniquely corresponding to a randomly chosen t_i, followed by t_{i+1}. We experimented both with $t'_i = t_i$ and t'_i being a different symbol, observing that the latter typically leads to less attention sinks forming.

The loss is calculated as $L_{\text{total}} = \lambda \cdot (1/k) L_{\text{random}} + L_{\text{induction}}$. We tested $\lambda = 0$ and $\lambda = 0.2$. In the former case, the model is only penalized on incorrect predictions of the last token and not incentivized to keep a constant uniform prediction on random tokens. This leads to all heads duplicating each other and almost no attention sinks for any attention type.

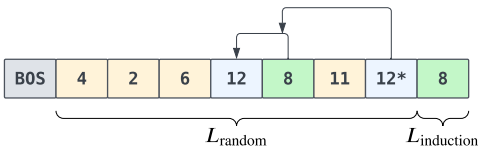


Figure 5: Task format in the induction heads experiment. The arrows illustrate the induction head mechanism.

Hyperparameters.

In this experiment, the model is a standard 2-layer H-head causal Transformer with pre-normalization, similar to GPT-2 Radford et al. (2019). The full hyperparameter and configuration details are available in Table 2.

Metrics.

We discard all runs that did not achieve induction accuracy above 95%. A sink score for each head is measured as an average proportion of attention weight assigned to the BOS token on the positions where no informed prediction is possible. We consider a head *a sink head* if its sink score is higher than 0.9. The computation can be expressed via the following pseudocode:

```
A = attention_matrix  # [layer, head, sample, query_seq_len, key_seq_len]

# index 0 is removed because it's BOS.
# indices -1 and -2 are removed because prediction is not random.
bos_score[L, H] = A[L, H, :, 1 : -2, 0]
total_weight[L, H] = A[L, H, :, 1 : -2, :].sum(axis=-1)

# clipping affects attention types without positivity. doesn't change is_sink.
sink_score[L, H] = (bos_score[L, H] / total_weight[L, H]).mean().clip(0, 1)
is_sink[L, H] = (sink_score[L, H] > 0.9)
```

We only plot the results from the second layer, as all heads in the first layer typically attend to the previous token.

Hyperparameter	**Value**	**Hyperparameter**	**Value**
Layers	2	Epochs	10
Hidden dim	128	Train size	100,000
Feedforward dim	512	Eval size	1,000
Vocab size	51*	Sequence length	13
Causal	`true`	Batch size	256
Use LayerNorm	`true`	Use MLP	`true`
Heads	$\{\mathbf{1}, \mathbf{2}, \mathbf{4}, \mathbf{8}\}$	$t'_i = t_i$	$\{$`true`, **`false`**$\}$
Learning rate	$\{\mathbf{10^{-2}}, \mathbf{10^{-3}}, \mathbf{10^{-4}}\}$	λ	$\{0.0, \mathbf{0.2}\}$
Attention type	all in Table 1	Positional type	$\{$`concat`, **`additive`**, `no`$\}$
Seed	$\{\mathbf{0}, \mathbf{1}, \mathbf{2}, \mathbf{3}, \mathbf{4}\}$		

Table 2: Configuration and hyperparameters for the induction experiment. The bold values indicate the setup shown in Figures 1 and 3b. Results for other setups are additionally presented in Appendix H.2. Vocabulary size is increased by 50 if $t'_i \neq t_i$.

G.2 Difference in attention sparsity between pretrained softmax and sigmoid LLMs

We use the weights and the codebase[1] for the pretrained 7B softmax and sigmoid models released by Ramapuram et al. (2025). We run both models on 1000 samples from a subset[2] of the Pile dataset (Gao et al., 2020) with the maximum sequence length of 256. We measure the sparsity score of each head as the average proportion of total attention weight assigned to the max-logit token. In pseudocode,

```
A = attention_matrix  # [layer, head, sample, query_seq_len, key_seq_len]
max_attn_weight = A.max(axis=-1)
total_attn_weight = A.sum(axis=-1)
sparsity_score[L, H] = (max_attn_weight / total_attn_weight).mean()
```

In Figure 20, we plot the distribution of sparsity scores for all heads in softmax and sigmoid models.

[1] https://github.com/apple/ml-sigmoid-attention

[2] https://huggingface.co/datasets/NeelNanda/pile-10k

G.3 Emerging imbalance in token influence on classifier's predictions

Task format. The task is sequence classification into K labels. The vocabulary is split into K disjoint groups $V^{(1)}, \ldots, V^{(K)}$. Each object with a label k is a random sequence of tokens from the k-th group: $x_i = v_{j_{i1}}^{(k_i)}, \ldots, v_{j_{in}}^{(k_i)}, y_i = k$. `CLS` token is appended to the input. Knowledge of any of the input tokens is sufficient to achieve perfect in-distribution accuracy. The task is illustrated in Figure 6 (left).

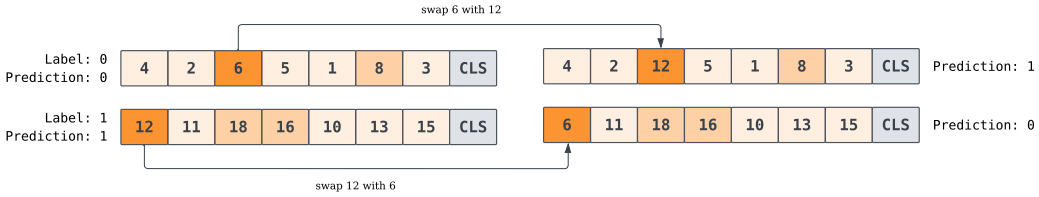


Figure 6: Task format in the classification experiment and the token swap procedure. Cell brightness illustrates the attention weight assigned to each token.

Hyperparameter	Value	Hyperparameter	Value
Layers	$\{\mathbf{1}, 2\}$	Epochs	10
Hidden dim	128	Train size	10,000
Feedforward dim	512	Eval size	1,000
Vocab size	$\{\mathbf{41}, \mathbf{81}, \mathbf{161}\}$	Sequence length	11
Labels	4	Causal	`false`
Batch size	256	Positional type	`no`
Heads	$\{\mathbf{1}, 2, 4\}$	Use MLP	$\{\mathbf{true}, \texttt{false}\}$
Use LayerNorm	$\{\mathbf{true}, \texttt{false}\}$	Learning rate	$\{\mathbf{10^{-2}}, \mathbf{10^{-3}}, \mathbf{10^{-4}}\}$
Attention type	all in Table 1	Seed	$\{\mathbf{0}, \mathbf{1}, \mathbf{2}, \mathbf{3}, \mathbf{4}\}$

Table 3: Configuration and hyperparameters for the classification experiment. The bold values indicate parameters used for the setup in Figures 3a and 15. Results for other setups are additionally provided in Appendix H.3.

Hyperparameters.

We use the same architecture as in the induction head experiment. However, we do not use causal masking or positional encodings as the input is effectively a set. We experiment with different numbers of layers and heads, as well as active and inactive MLPs and LayerNorm layers. Full details are provided in Table 3.

Metrics.

We calculate the out-of-distribution flip rate, presented in Figure 3a, as follows. We repeatedly pick two random sequences with different labels from the evaluation set and run a forward pass on them, extracting the attention weights with the `CLS` token as a query. Then we swap the max-logit tokens for the head (L, H) and check if the model changed the prediction. In pseudocode,

```
X1, X2, Y1, Y2 = pick_samples_with_different_labels()

A1 = attention_matrix(X1)  # [layer, head, query_seq_len, key_seq_len]
A2 = attention_matrix(X2)

# obtaining the max-logit tokens excluding CLS itself
ranking_1[L, H] = A1[L, H,-1,:-1].argsort(reverse=True)
ranking_2[L, H] = A2[L, H,-1,:-1].argsort(reverse=True)

# max-logit token from label 2 + others from label 1
adversarial_1[L, H] = [ranking_2[L, H][0]] + ranking_1[L, H][1:]
```

```
    # max-logit token from label 1 + others from label 2
    adversarial_2[L, H] = [ranking_1[L, H][0]] + ranking_2[L, H][1:]

    prediction_adv_1[L, H] = model(adversarial_1[L, H])
    prediction_adv_2[L, H] = model(adversarial_2[L, H])

    # flipped if swapping one token changes the prediction
    flipped_1[L, H] = (prediction_adv_1[L, H] == Y2)
    flipped_2[L, H] = (prediction_adv_2[L, H] == Y1)
    flip_rate[L, H] = (flip_1 + flip_2) / 2
```

Figure 6 illustrates this procedure. We report *the maximum flip rate* over all heads and layers.

G.4 Compute Resources

The experiments with the value-softmax model were conducted on CPU, the experiments with toy transformers used one A100 node with different experiments running in parallel on separate GPUs. All experiments combined take less than 1 day of walltime.

H Additional Experimental Results

H.1 Value-Softmax Model

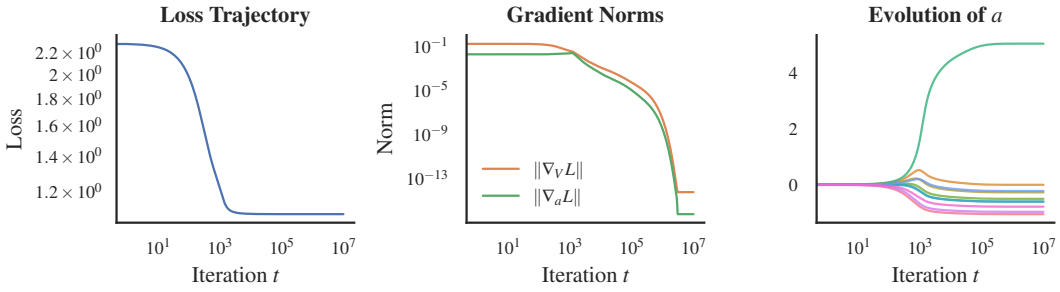


Figure 7: Training dynamics of Value-Softmax model with KL-divergence as loss function.

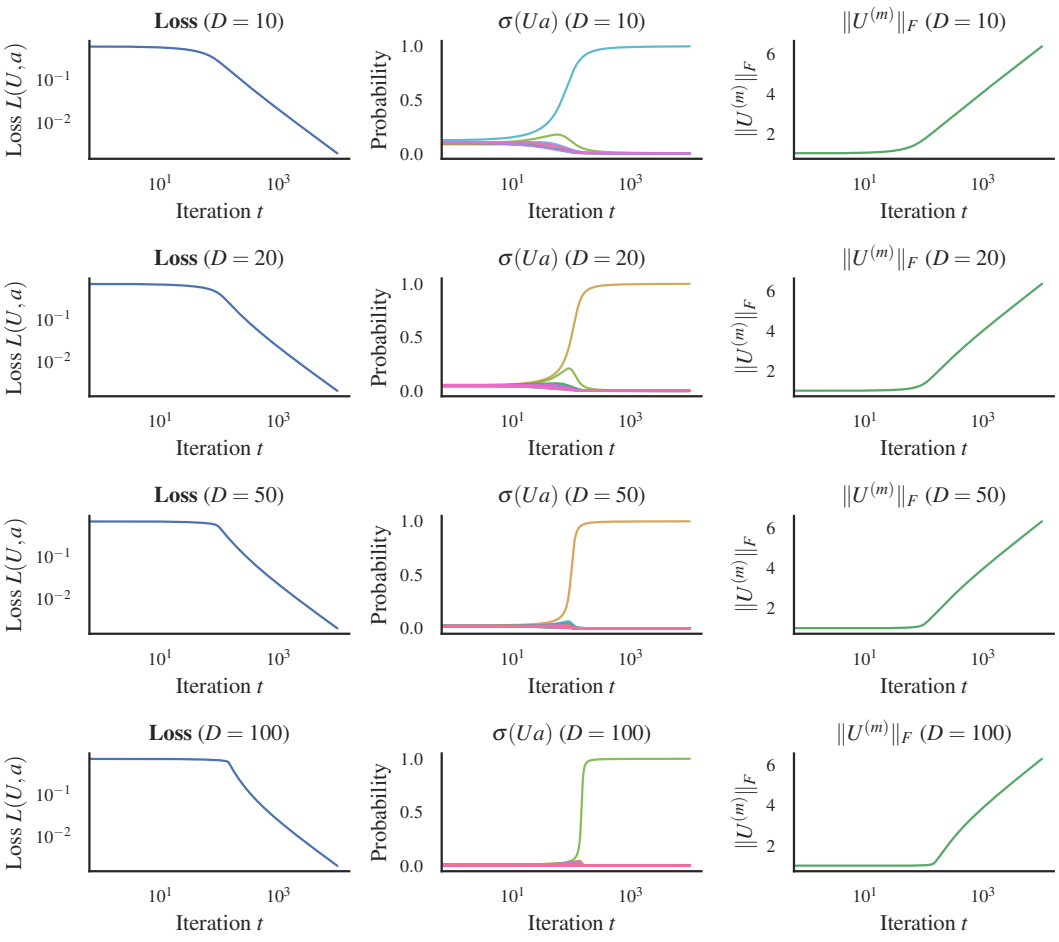


Figure 8: Training dynamics of tied Value-Softmax model ($U\sigma(Ua)$) with logistic loss for different dimensionality D. We denote the column of U corresponding to the largest element of a as $U^{(m)}$. In all cases, a sparse solution is found, and the norm of $U^{(m)}$ diverges.

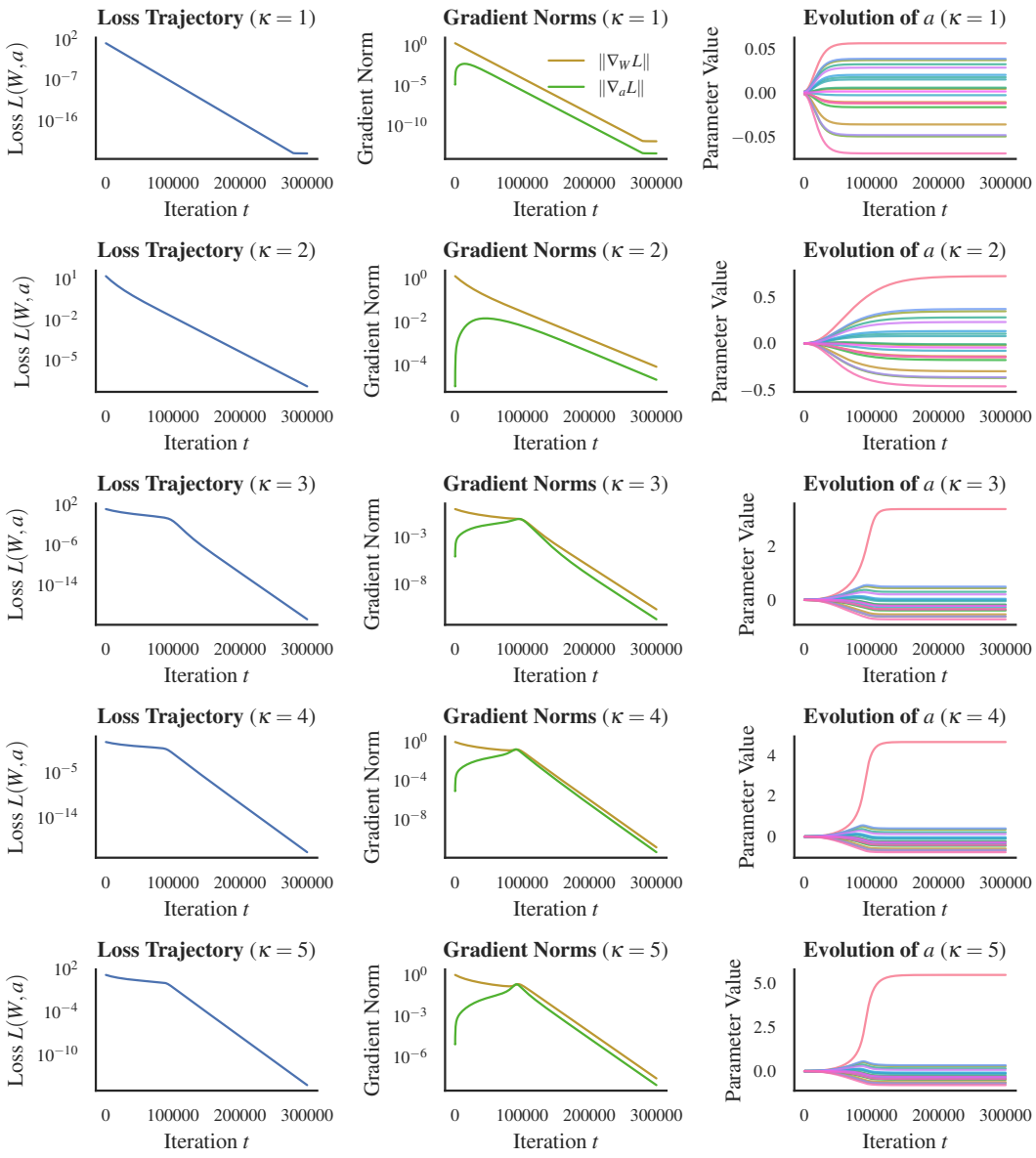
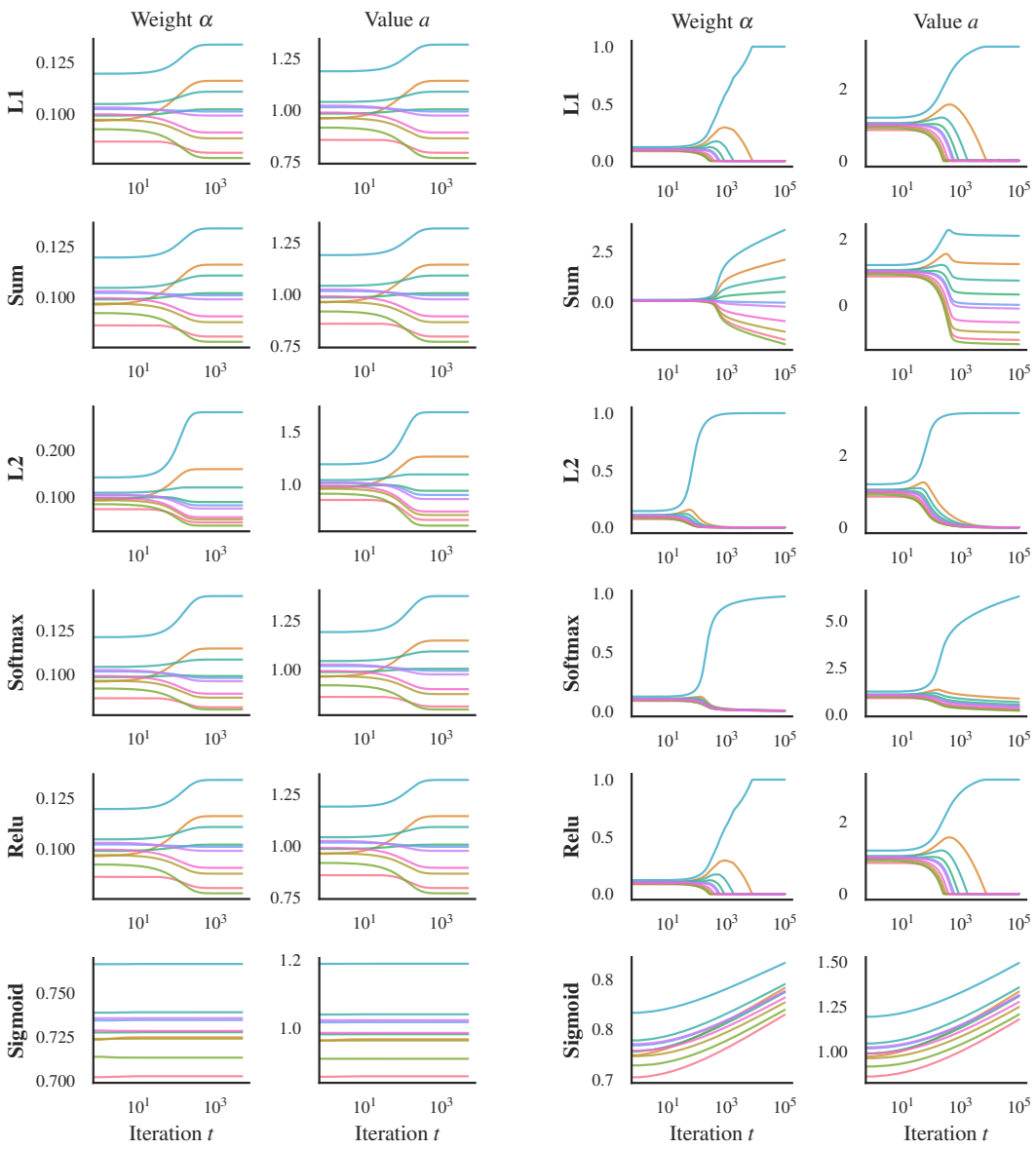


Figure 9: Training dynamics in regression problems with condition number κ from 1 to 5. Higher condition number leads to slower optimization and stronger sparsity.

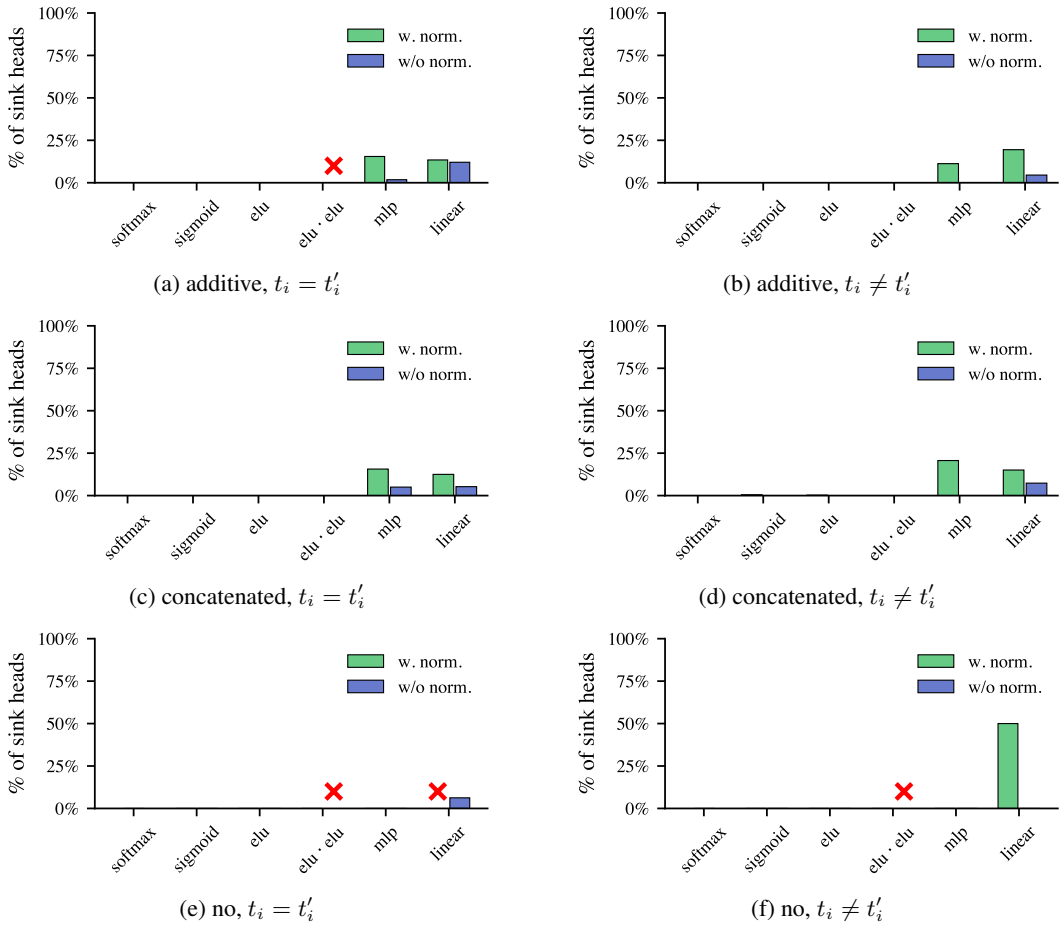


(a) Square loss.

(b) Logistic loss.

Figure 10: Training dynamics of value-softmax model for square and logistic losses and various activation functions. Note that square loss does not give rise to sparse solutions. Additionally, sum and sigmoid do not converge to sparse solutions for logistic loss as well.

H.2 INDUCTION HEADS ACTING AS ATTENTION SINKS



(a) additive, $t_i = t'_i$

(b) additive, $t_i \neq t'_i$

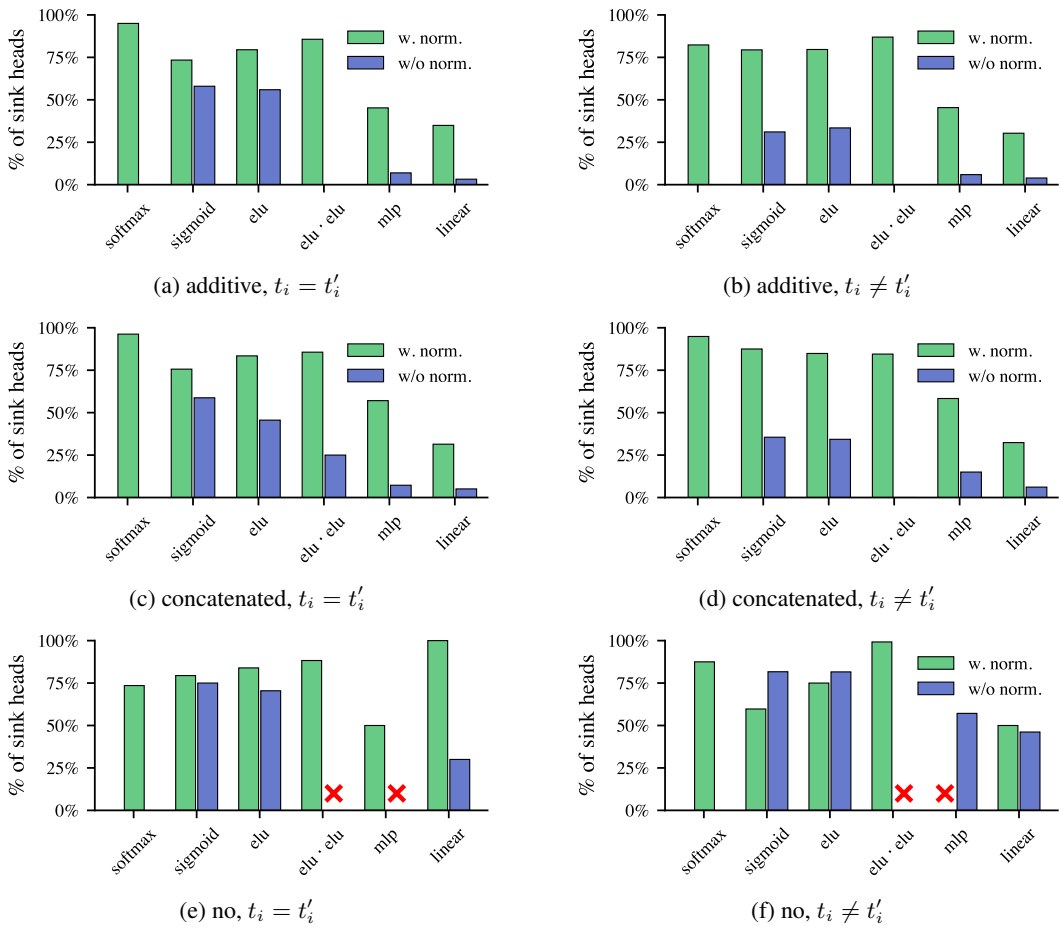
(c) concatenated, $t_i = t'_i$

(d) concatenated, $t_i \neq t'_i$

(e) no, $t_i = t'_i$

(f) no, $t_i \neq t'_i$

Figure 11: Sink proportions in induction-trained Transformer with $\lambda = 0$. Each row represents a type of positional encodings. Setups that result in training failures are marked by red crosses. We observe that under this setup, most heads duplicate each other and do not form sinks.



(a) additive, $t_i = t'_i$

(b) additive, $t_i \neq t'_i$

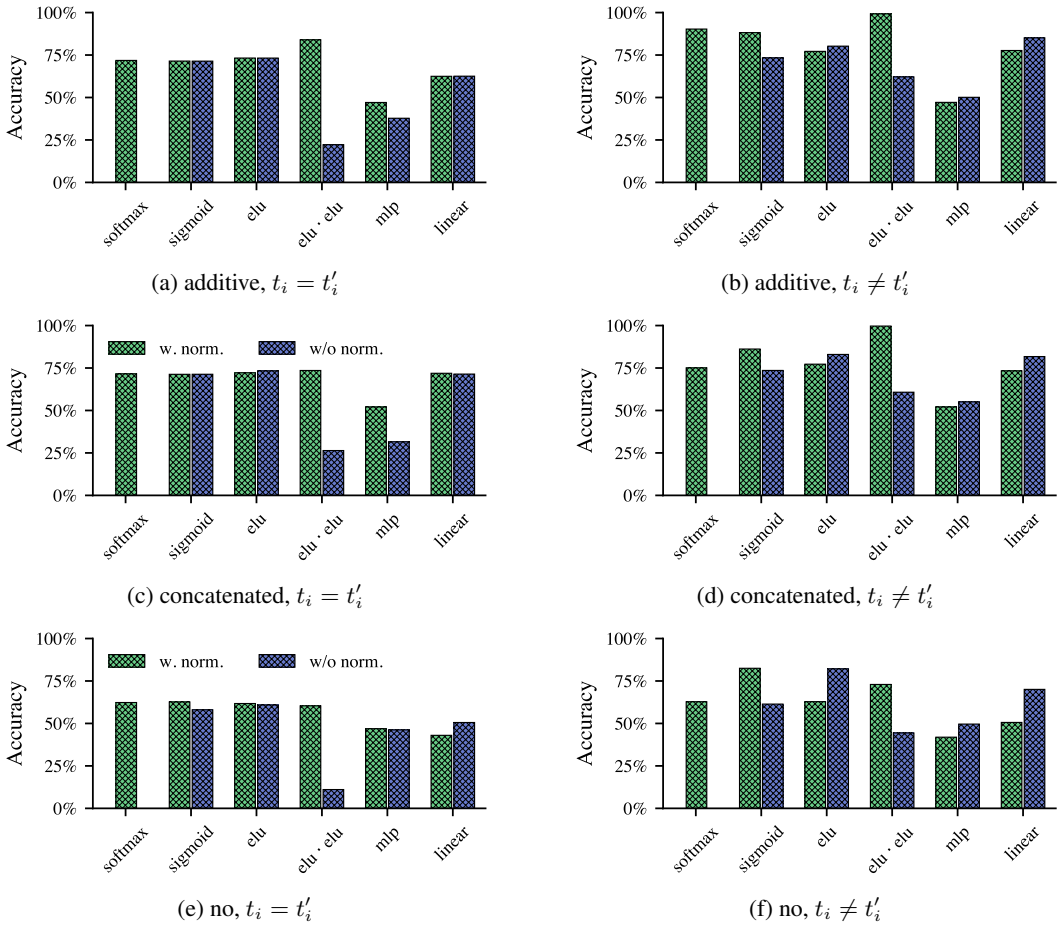
(c) concatenated, $t_i = t'_i$

(d) concatenated, $t_i \neq t'_i$

(e) no, $t_i = t'_i$

(f) no, $t_i \neq t'_i$

Figure 12: Sink proportions in induction-trained Transformer with $\lambda = 0.2$. Each row represents a type of positional encodings. Setups that result in training failures are marked by red crosses. Interestingly, we observe that setups with no explicit positional encodings lead to an increased formation of attention sinks. Note that in this case the model has access to positional information through causal masking, so the task is still solvable.



(a) additive, $t_i = t'_i$

(b) additive, $t_i \neq t'_i$

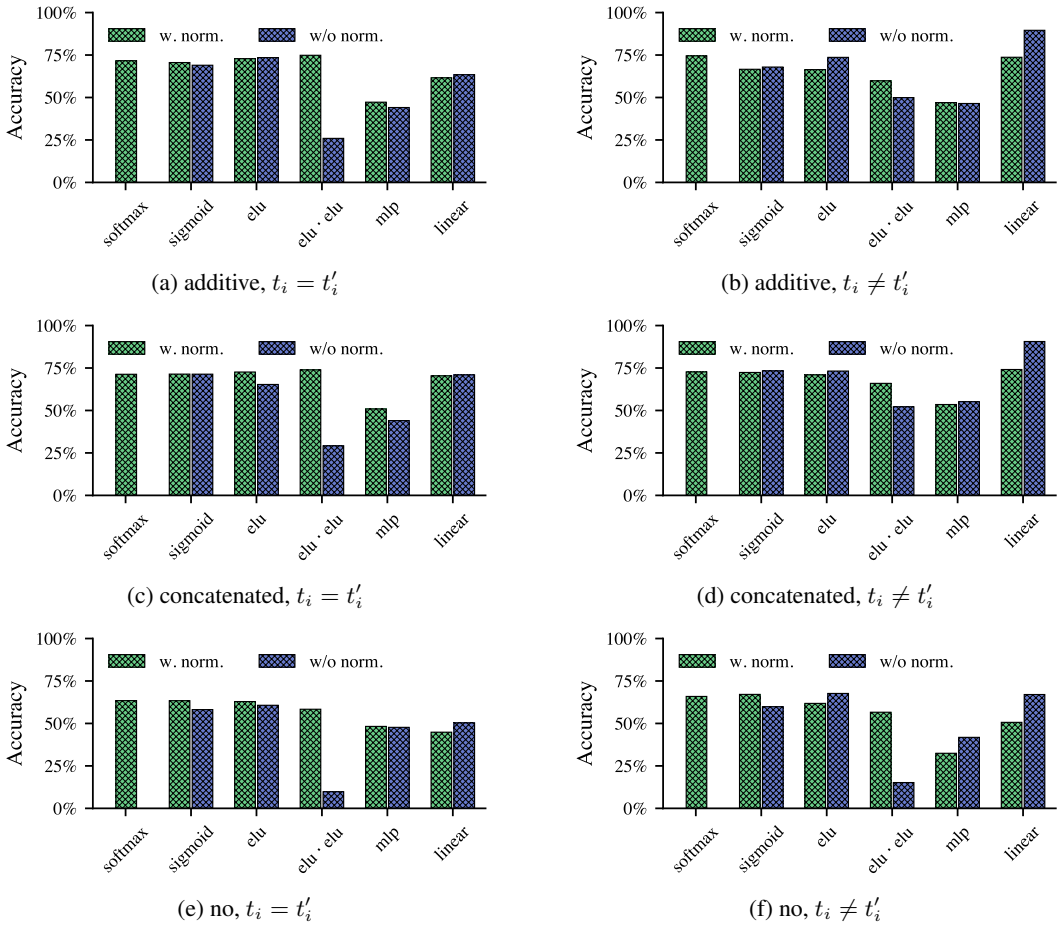
(c) concatenated, $t_i = t'_i$

(d) concatenated, $t_i \neq t'_i$

(e) no, $t_i = t'_i$

(f) no, $t_i \neq t'_i$

Figure 13: Average induction accuracy for models trained with $\lambda = 0$. Each row represents a type of positional encodings.



(a) additive, $t_i = t'_i$

(b) additive, $t_i \neq t'_i$

(c) concatenated, $t_i = t'_i$

(d) concatenated, $t_i \neq t'_i$

(e) no, $t_i = t'_i$

(f) no, $t_i \neq t'_i$

Figure 14: Average induction accuracy for models trained with $\lambda = 0.2$. Each row represents a type of positional encodings.

H.3 EMERGING IMBALANCE IN TOKEN INFLUENCE ON CLASSIFIER'S PREDICTIONS

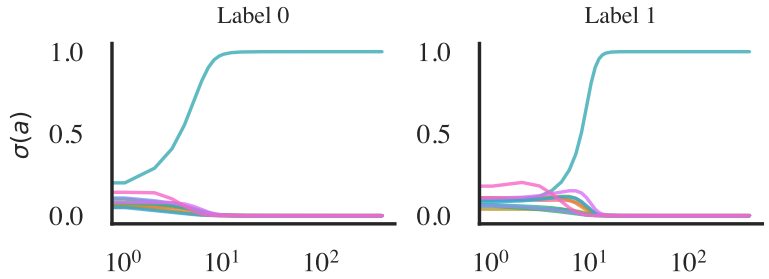
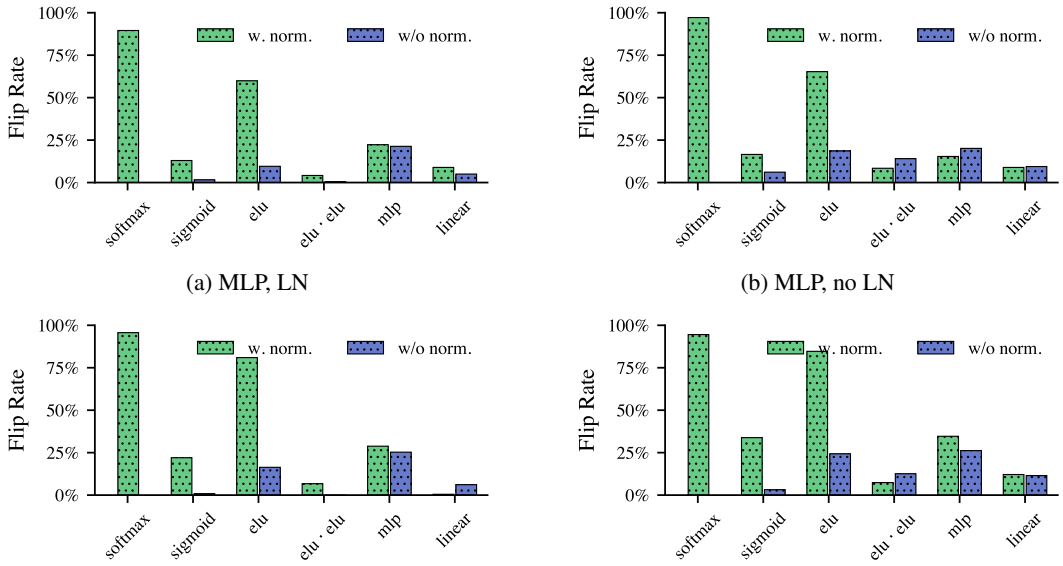


Figure 15: Evolution of attention scores for two differently labeled samples in the classification experiment.



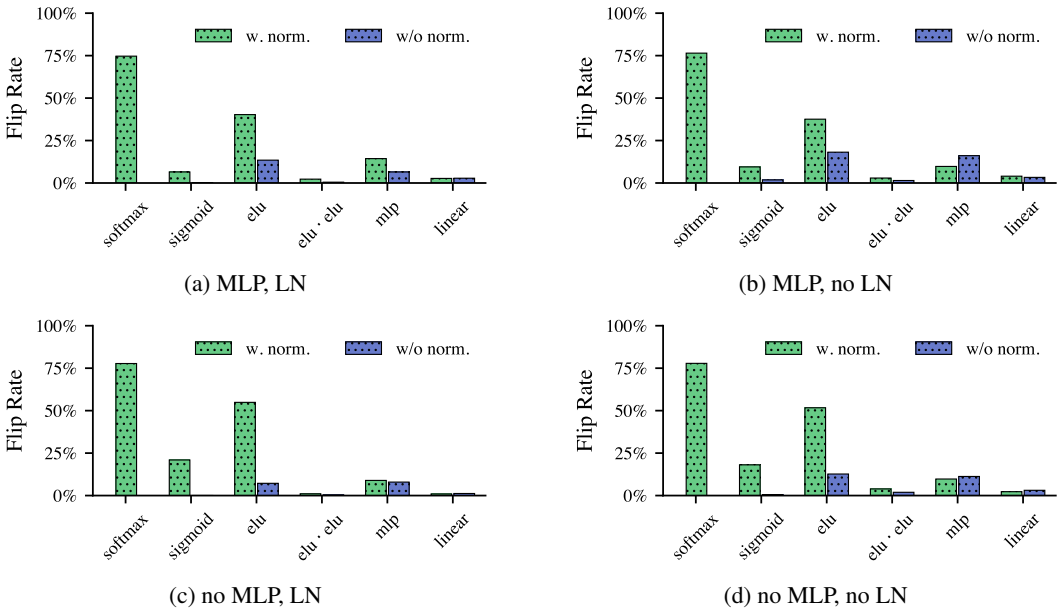
(a) MLP, LN

(b) MLP, no LN

(c) no MLP, LN

(d) no MLP, no LN

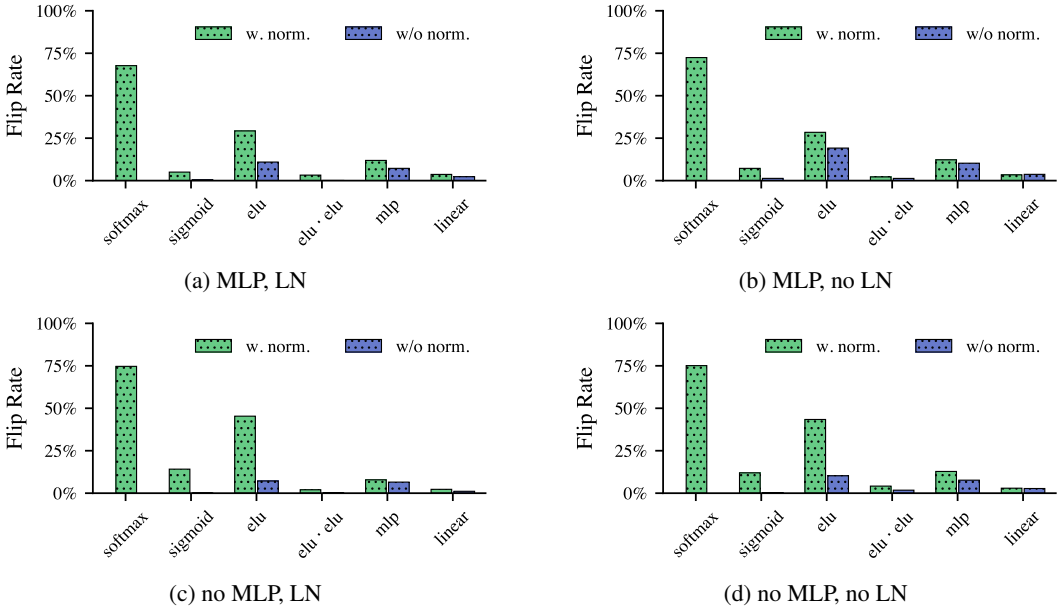
Figure 16: Flip-rate plots for vocabulary size $|\mathcal{V}| = 41$. Rows: MLP vs. no-MLP. Columns: LayerNorm vs. no-LayerNorm.



(a) MLP, LN (b) MLP, no LN

(c) no MLP, LN (d) no MLP, no LN

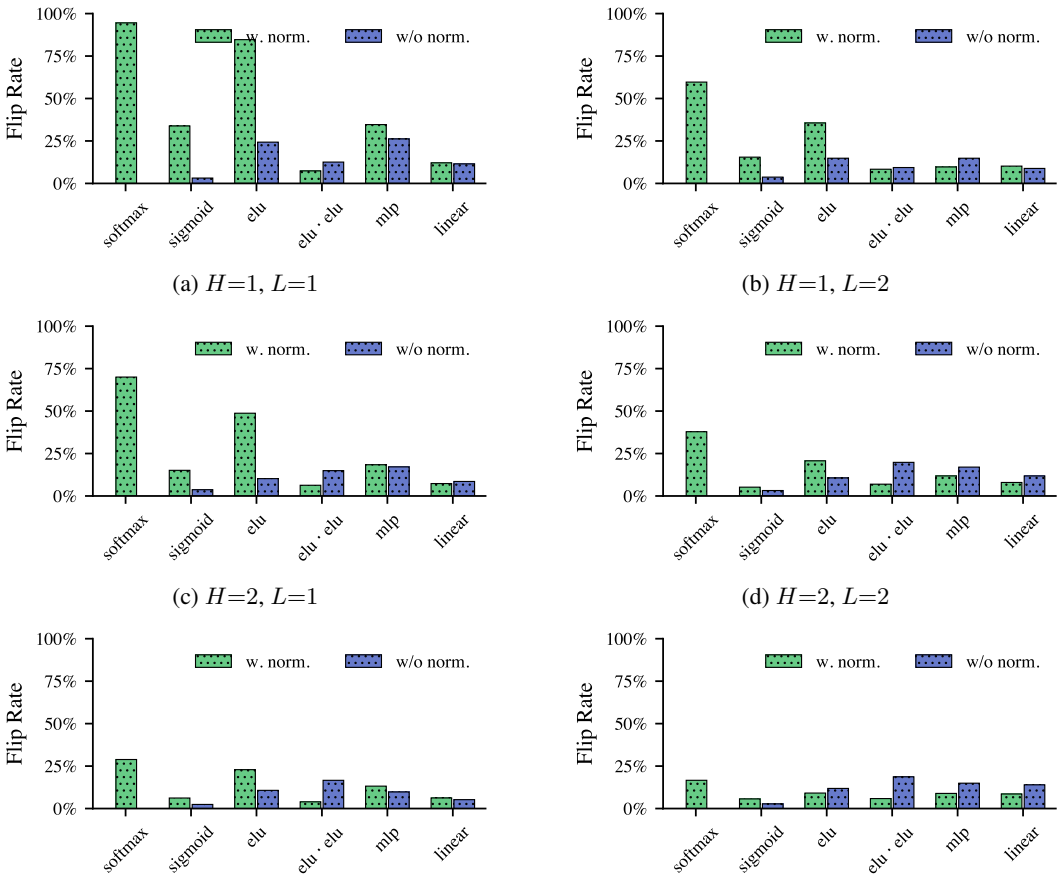
Figure 17: Flip-rate plots for vocabulary size $|\mathcal{V}| = 81$. Rows: MLP vs. no-MLP. Columns: LayerNorm vs. no-LayerNorm.



(a) MLP, LN (b) MLP, no LN

(c) no MLP, LN (d) no MLP, no LN

Figure 18: Flip-rate plots for vocabulary size $|\mathcal{V}| = 161$. Rows: MLP vs. no-MLP. Columns: LayerNorm vs. no-LayerNorm.



(a) H=1, L=1 (b) H=1, L=2

(c) H=2, L=1 (d) H=2, L=2

(e) H=4, L=1 (f) H=4, L=2

Figure 19: Flip-rate plots for $|\mathcal{V}| = 41$ with no MLP and no LayerNorm. Rows: number of heads $H \in \{1, 2, 4\}$. Columns: number of layers $L \in \{1, 2\}$. Increasing the number of heads or layers leads to the diminishing flip rate.

H.4 ATTENTION SINKS IN PRETRAINED LLMS

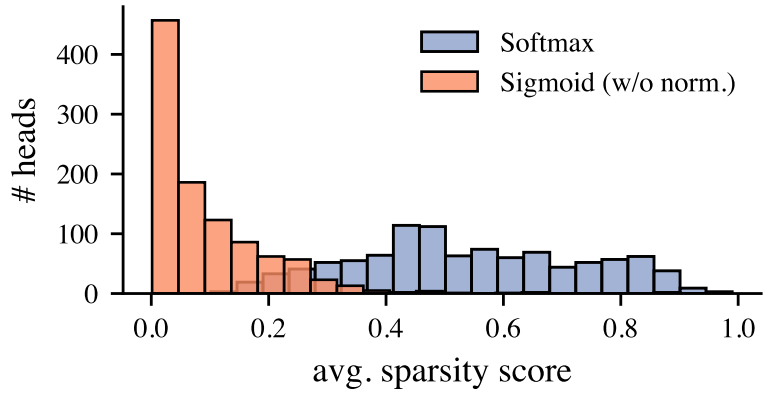


Figure 20: Distribution of attention-head sparsity scores for pretrained softmax/sigmoid LLMs; sparse attention concentrates on few tokens per class.