A THEORETICAL ANALYSIS ON THE EMERGENCE OF IMPLICIT CURRICULUM IN TRANSFORMERS

Anonymous authors
Paper under double-blind review

ABSTRACT

Empirical studies of language-model pretraining suggest an implicit curriculum: localization and retrieval capabilities often emerge before capabilities that compose retrieved information. Existing theory, however, typically analyzes these ingredients separately, studying attention-based token selection with simple readouts or nonlinear composition on fixed representations, leaving open whether joint training can itself induce a structured acquisition order. We study this question in a stylized shared-attention multi-task Transformer, in which a softmax attention layer, shared across tasks, feeds task-specific ReLU heads. The model is trained by stochastic gradient descent on a fixed mixture of two tasks whose input sequences contain informative tokens among distractors: a token-selection task, labeled by the sign of a single informative token, and an XOR task, labeled by the XOR of the signs of two informative tokens. Both tasks are present throughout training, and the shared attention-weight parameter and both heads are updated in parallel, without an externally imposed curriculum. Our main result proves that the joint trajectory nevertheless exhibits a definite learning order: token selection reaches small test error while the XOR error is still bounded away from zero, and the XOR error becomes small only at a later iteration. The mechanism is that token-selection learning amplifies attention to a positional signal shared by both tasks, while parity cancellation and cumulative stability control keep XOR updates from disrupting this amplification. Experiments on synthetic data from the model and MNIST-based token sequences support the predicted learning order. As supporting empirical context, preliminary measurements on open intermediate large language model (LLM) checkpoints show a compatible pattern in which perception-like skills emerge earlier than composition-like skills.

1 INTRODUCTION

The success of deep learning is widely attributed to its hierarchical (feature) learning (Bengio et al., 2013; LeCun et al., 2015; Poggio et al., 2017; Allen-Zhu & Li, 2023), in which shallow layers extract task-relevant signals from the input (perception) while deeper layers transform those signals into task-specific outputs (reasoning). This functional division mirrors the structure of the tasks themselves: identifying which inputs carry signal logically precedes working out what those signals imply, and has been studied across vision (Krizhevsky et al., 2012; He et al., 2016), language (Devlin et al., 2019; Floridi & Chiriatti, 2020), and audio (Hannun et al., 2014). Earlier vision research instantiated the hierarchy concretely as pattern recognition feeding logical composition over the recognized parts (Zeiler & Fergus, 2014; Olah et al., 2017; Bau et al., 2017; Kornblith et al., 2019; Allen-Zhu & Li, 2023; Barak et al., 2022; Abbe et al., 2022), supplying both empirical and theoretical evidence that feature learning is essential where shallow or kernel methods fail.

Despite the extensive characterization of the hierarchical learning in earlier research, few theoretical studies have analyzed how perception and reasoning are acquired together from training, especially in transformer-type models: each capability has been addressed in isolation. One line establishes feature learning and benign-overfitting guarantees for attention-based models, characterizing how attention selects relevant tokens (Magen et al., 2024; Sakamoto & Sato, 2025); a complementary line analyzes feature learning in two-layer ReLU networks on logical tasks such as XOR, treating the input representation as given (Frei et al., 2023; Glasgow, 2023; Xu et al., 2024; Meng et al., 2024). A third line studies compositional learning more abstractly (Abbe et al., 2023; Wang et al., 2025;

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.

Panigrahi et al., 2025; Kim & Suzuki, 2025), but obtains its results by freezing one side, decoupling layers in their analysis, or imposing an explicit data curriculum. None of these settings reflect how language models are actually pretrained, where data of widely varying complexity is presented in arbitrary order yet skill acquisition still follows a reproducible sequence (Liu et al., 2026).

We address this gap by analyzing a stylized shared-attention multi-task Transformer, consisting of a shared softmax attention module followed by task-specific ReLU multilayer perceptron (MLP) layers, trained under gradient descent on two related tasks with distinct perception and reasoning demands. Within this architecture, pattern recognition specializes to attention's job of identifying which tokens carry signal (token selection), and logical composition specializes to the MLP's job of computing the XOR of the attended signals. Both specializations are classic in literature (Frei et al., 2023; Glasgow, 2023; Magen et al., 2024; Sakamoto & Sato, 2025). Crucially, our setup mirrors how language models are actually pretrained: no curriculum is imposed on the optimizer, and both pattern recognition (e.g. information extraction and extractive question answering) and logical composition (e.g., arithmetic reasoning, multi-hop question answering) subproblems are present concurrently from the first step of training to the last.

Our main result is that, despite the absence of any externally imposed ordering, training on this task produces a two-phase trajectory in which perception is acquired before reasoning. Along our training dynamic analysis, learning the token-selection task increases the attention mass on position-relevant signal, thereby strengthening the effective learning signal seen by the XOR head. Conversely, symmetry at initialization and parity cancellation make the XOR-induced update to the shared attention small, so XOR learning does not affect the attention localization established by token-selection task.

- We propose a stylized theoretical setup with hierarchical task structure for studying the implicit curriculum behavior on a shared-attention multi-task Transformer (Section 3).
- We rigorously prove (Theorem 4.1) that the model achieves low error on token selection before learning XOR, revealing a provable timescale separation. We further show that XOR-only training cannot achieve the same progress within this regime (Theorem 4.2).
- We empirically validate the predicted ordering on synthetic data sampled from our theoretical setting (Figure 1), and further recover the same ordering on an MNIST-based XOR task in both single- and multi-layer Transformers (Figure 2 and 3), indicating that the two-phase trajectory generalizes beyond our stylized theoretical setting.
- Using intermediate checkpoints from open pretraining runs on real data, we observe that perception-like capabilities saturate substantially earlier than reasoning-like ones during real LLM training, despite the absence of any explicit curriculum in the data (Figure 4).

Taken together, these results suggest a mechanistic account of implicit curriculum in attention-based models. The observed perception-before-reasoning order does not require an externally imposed curriculum; it can arise from staggered learning timescales across the two tasks, so that the easier perception task achieves low test error before the slower reasoning task is learned. Our LLM-checkpoint experiment provides preliminary evidence that the same ordering appears in real pretraining.

2 Related Work

Training Dynamics of Neural Network. The training dynamics of neural networks beyond the lazy / neural tangent kernel (NTK) linearization have been studied in several lines whose techniques and conclusions inform our analysis. The early NTK line (Jacot et al., 2018; Du et al., 2019; Allen-Zhu et al., 2019) shows that wide networks trained by gradient descent can be analyzed as kernel regression, but this regime is now known to be insufficient for tasks requiring genuine feature learning (Yehudai & Shamir, 2019; Ghorbani et al., 2021; Daniely & Malach, 2020). Mean-field and feature-learning analyses (Mei et al., 2018; Damian et al., 2022; Ba et al., 2022; Bietti et al., 2022; Mousavi-Hosseini et al., 2023) characterize how shallow networks acquire task-relevant features beyond NTK, with sharp results on XOR-style targets where kernel methods provably fail (Frei et al., 2023; Glasgow, 2023; Xu et al., 2024; Meng et al., 2024; Barak et al., 2022; Abbe et al., 2022). A complementary line on attention-based models studies the dynamics of token selection: max-margin

and support vector machine (SVM)-style implicit-bias analyses of self-attention (Ataee Tarzanagh et al., 2023; Tarzanagh et al., 2023; Vasudeva et al., 2024; Sheen et al., 2024), and the convergence or benign overfitting of single-head softmax attention (Oymak et al., 2023; Magen et al., 2024; Jiang et al., 2024; Sakamoto & Sato, 2025).

Compositional Tasks. The theoretical study of compositional tasks has developed along several lines whose models and separations motivate our analysis. The classical curriculum-learning perspective (Bengio et al., 2009) has recently been formalized in stylized Boolean settings: curriculum schedules can yield provable training-time separations for parity learning (Abbe et al., 2023), while staircase-structured functions show how gradient-based methods progressively build higher-order features from lower-order ones (Abbe et al., 2021). A related line studies transformers on canonical compositional targets, including parities, k-fold compositions, and k-hop induction, establishing both positive depth requirements and negative expressivity or optimization barriers (Sanford et al., 2024; Peng et al., 2024; Chen et al., 2024; Abbe et al., 2024). A complementary line analyzes intermediate supervision and implicit curricula: scratchpads and chain-of-thought mechanisms can improve compositional reasoning or sample efficiency (Nye et al., 2021; Wen et al., 2024; Feng et al., 2023), progressive distillation can induce curriculum-like behavior in shallow networks (Panigrahi et al., 2025), and next-token or autoregressive training can exhibit failure modes on compositional chains (Bachmann & Nagarajan, 2024).

3 Preliminary

Notations. We write $\|\mathbf{v}\|_2$ for the Euclidean norm of a vector $\mathbf{v}$ and $\|\mathbf{A}\|$ for the spectral norm of a matrix $\mathbf{A}$. We write $\mathbb{S}$ for the softmax map, applied row-wise to matrix inputs. In addition to the typical $O(\cdot)$ notation, for two functions f, g, we use the shorthand $f \lesssim g$ to indicate that $f \le Cg$ for an absolute constant C. We use $\widetilde{O}(\cdot)$, $\widetilde{\Omega}(\cdot)$, and $\widetilde{\Theta}(\cdot)$ for the corresponding asymptotic notations ignoring polylogarithmic factors. We use C to denote a universal constant that is large enough in our setting.

Data model. Let $\mathbf{X} = (\mathbf{x}_1, \dots, \mathbf{x}_T)^\top \in \mathbb{R}^{T\times d}$, where $T \ge 3$. Fix pairwise orthogonal $\mathbf{p}, \boldsymbol{\mu}_1, \boldsymbol{\mu}_2 \in \mathbb{R}^d$ with $\|\mathbf{p}\|_2 > 0$ and $\|\boldsymbol{\mu}_1\|_2 = \|\boldsymbol{\mu}_2\|_2 = 1$. The token-selection distribution $\mathcal{P}_1$ draws independent $c \sim \mathrm{Unif}\{1, 2\}$ and $z \sim \mathrm{Unif}\{-1, 1\}$, and sets

$$\mathbf{x}_1 = \mathbf{p} + z\boldsymbol{\mu}_c + \boldsymbol{\epsilon}_1, \qquad \mathbf{x}_i = \boldsymbol{\epsilon}_i \ (2 \le i \le T), \qquad y = z, \qquad \boldsymbol{\epsilon}_i \overset{\text{i.i.d.}}{\sim} \mathcal{N}(\mathbf{0}, \sigma_\epsilon^2 \mathbf{I}_d).$$

The XOR distribution $\mathcal{P}_2$ draws independent uniform signs z_1, z_2 and sets

$$\begin{aligned} \mathbf{x}_1 &= \mathbf{p} + z_1\boldsymbol{\mu}_1 + \boldsymbol{\epsilon}_1, & \mathbf{x}_2 &= \mathbf{p} + z_2\boldsymbol{\mu}_2 + \boldsymbol{\epsilon}_2, \\ \mathbf{x}_i &= \boldsymbol{\epsilon}_i \quad (3 \le i \le T), & y &= z_1 z_2, \end{aligned}$$

where $\boldsymbol{\epsilon}_1, \boldsymbol{\epsilon}_2 \sim \mathcal{N}(\mathbf{0}, \sigma_\epsilon^2(\mathbf{I}_d - \mathbf{p}\mathbf{p}^\top/\|\mathbf{p}\|_2^2))$ and $\boldsymbol{\epsilon}_i \sim \mathcal{N}(\mathbf{0}, \sigma_\epsilon^2 \mathbf{I}_d)$ for $i \ge 3$. All noises, signs, and class choices within an example are independent; examples are independent.

Network Model. A standard single-head self-attention layer takes the form $f_{\mathrm{SA}}(\mathbf{X}) = \mathbb{S}(\mathbf{X}\mathbf{W}_Q\mathbf{W}_K^\top\mathbf{X}^\top)\mathbf{X}\mathbf{W}_V\mathbf{W}_o$, with trainable matrices $\mathbf{W}_Q, \mathbf{W}_K, \mathbf{W}_V, \mathbf{W}_o$ (Vaswani et al., 2017). Many architectures instead read out from a designated token $\mathbf{x}_0 \in \mathbb{R}^d$, as in the [CLS] construction of BERT (Devlin et al., 2019) and ViT (Dosovitskiy et al., 2021), with related input-side parameterizations used in prompt and prefix tuning (Lester et al., 2021; Li & Liang, 2021). Motivated by this setup, theoretical analyses merge $\mathbf{W}_Q\mathbf{W}_K^\top$ into $\mathbf{W}_{QK}$ and further reduce $(\mathbf{W}_{QK}, \mathbf{x}_0)$ to a single trainable attention vector $\mathbf{q} := \mathbf{W}_{QK}^\top\mathbf{x}_0$ (Magen et al., 2024); Lemma 1 of Ataee Tarzanagh et al. (2023) establishes a one-to-one correspondence between its gradient iterates and those of the full $\mathbf{W}_{QK}$ parametrization.

We adopt this reduced architecture, with shared representation

$$\widetilde{\mathbf{x}} := \mathbf{X}^\top\mathbb{S}(\mathbf{X}\mathbf{q}).$$

The representation $\widetilde{\mathbf{x}}$ is shared by the downstream token-selection and XOR tasks. For $k \in \{1, 2\}$, let $\mathbf{W}_k \in \mathbb{R}^{m\times d}$ have rows $\mathbf{w}_{k,j}^\top$, and define

$$f_k(\mathbf{X}; \mathbf{q}, \mathbf{W}_k) = \mathbf{a}^\top\phi(\mathbf{W}_k\widetilde{\mathbf{x}}).$$

Here $\phi(v) = \max\{v, 0\}$ and $\phi'(0) = 0$. Set $\sigma_w = (md)^{-1/2}$ and initialize $\mathbf{q}(0) = \mathbf{0}$, $(\mathbf{W}_k(0))_{jl} \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, \sigma_w^2)$, and $a_j \overset{\text{i.i.d.}}{\sim} \mathrm{Unif}\{-m^{-1/2}, m^{-1/2}\}$. The two head matrices are independent of one another and of the shared vector $\mathbf{a}$, which remains fixed.

This is a multi-task learning model: the two tasks share the attention weight $\mathbf{q}$ and use private prediction heads $\mathbf{W}_1, \mathbf{W}_2$. Shared representations with task-specific prediction modules also appear in multi-task Transformers (Bhattacharjee et al., 2022). We train both heads and the shared attention weight at every iteration, with all gradients evaluated at the same current state. This differs from the head-then-representation updates and subsequent frozen-representation adaptation analyzed by Collins et al. (2024). Our focus is the order in which token selection and XOR are learned along a single SGD trajectory with a fixed task mixture.

Loss and training. Fix integers $b_1, b_2 \geq 1$, batch size $B = b_1 + b_2$, mixture weight $\lambda = b_2/B$, and learning rate $\eta > 0$. At each iteration τ, draw fresh independent batches $D_{k,\tau} \sim \mathcal{P}_k^{b_k}$, independently of the past. For $\ell(v) = \log(1 + e^{-v})$, define

$$\widehat{\mathcal{L}}_{k,\tau}(\mathbf{q}, \mathbf{W}_k) = \frac{1}{b_k} \sum_{(\mathbf{X},y)\in D_{k,\tau}} \ell(y f_k(\mathbf{X}; \mathbf{q}, \mathbf{W}_k)), \quad \widehat{\mathcal{L}}_\tau = (1-\lambda)\widehat{\mathcal{L}}_{1,\tau} + \lambda\widehat{\mathcal{L}}_{2,\tau}.$$

All gradients below are evaluated at the iteration-τ state:

$$\begin{aligned}
\mathbf{q}(\tau+1) &= \mathbf{q}(\tau) - \eta\big[(1-\lambda)\nabla_{\mathbf{q}}\widehat{\mathcal{L}}_{1,\tau} + \lambda\nabla_{\mathbf{q}}\widehat{\mathcal{L}}_{2,\tau}\big], \\
\mathbf{W}_1(\tau+1) &= \mathbf{W}_1(\tau) - \eta(1-\lambda)\nabla_{\mathbf{W}_1}\widehat{\mathcal{L}}_{1,\tau}, \\
\mathbf{W}_2(\tau+1) &= \mathbf{W}_2(\tau) - \eta\lambda\nabla_{\mathbf{W}_2}\widehat{\mathcal{L}}_{2,\tau}.
\end{aligned}$$

At iteration τ, training has used $B\tau$ examples and τ updates.

The task-specific updates run in parallel throughout training and are coupled through the shared effective input $\widetilde{\mathbf{x}}(\tau) = \mathbf{X}^\top \mathbb{S}(\mathbf{X}\mathbf{q}(\tau))$, which depends on the current attention-weight iterate. No externally imposed curriculum is placed on the data: the model trains on a fixed mixture of the two sources from the first step to the last, mirroring the regime in which language models are pretrained in practice.

Assumptions. Let $\delta \in (0, 10^{-1})$ be a target failure probability. The assumptions below specify the data, batch composition, and width; the learning rate will be selected in the main theorem.

Assumption 3.1 (Parameter conditions)**.** Let $C_\ell, C_{\mathrm{alloc}} > 0$ be sufficiently large universal constants, $c_u > 0$ a sufficiently small universal constant, and $\sigma_\epsilon > 0$. We assume:

(i) **Dimension.** $d \geq \max\Big\{\sigma_\epsilon^{-2},\ T(\log d)^{-9/8},\ c_u^{-1/2}\lambda m^{1/2}(\|\mathbf{p}\|_2^2 \log\log T)^{1/4}\Big\}$.

(ii) **Sequence length.** $T \geq \max\big\{d(\log d)^{17/16},\ \exp(2^{-1}\|\mathbf{p}\|_2^2)\big\}$.

(iii) **Noise variance.** $\sigma_\epsilon^2 \leq 2d^{-1}$.

(iv) **Position signal.** $\|\mathbf{p}\|_2^2 \geq (\log d)^{3/4}$.

(v) **Width.** $m \geq C_\ell \lambda^{-2} d^2 (\|\mathbf{p}\|_2^2 \log\log T)^{-2}\big[(\|\mathbf{p}\|_2^2 \log\log T)^{1/2} + \log(C_{\mathrm{alloc}}\delta^{-1})\big]$.

(vi) **Batch size.** $B \leq d^{1/2}$.

(vii) **Task proportions.** The per-task batch sizes satisfy $b_1 \leq 3b_2$ and $b_2 \leq b_1$.

This is the long-context, retrieval-heavy regime where attention aggregates many tokens: T is at least linear in d and at most sub-exponential in $d^{1/2}$. All constants C, c below are universal. Confidence enters through the displayed logarithmic term in the width condition.

4 Main Results

Our results separate the learning times of the two tasks under simultaneous training. The mixed learner first attains small token-selection error while its XOR error remains bounded away from zero, and later attains small XOR error. Training on XOR alone fails within the same budget.

Theorem 4.1 (Informal version of Theorem A.1, training with mixed tasks succeeds with two phrases)**.** *Under Assumption 3.1, select a learning rate $d^{-6} \le \eta \le d^{-3}$ and run the mixed SGD algorithm in Section 3. For $k \in \{1, 2\}$, define the conditional test error*

$$\mathrm{err}_k(\tau) := \Pr_{(\mathbf{X},y)\sim\mathcal{P}_k}\{\mathrm{sign}\, f_k(\mathbf{X}; \mathbf{q}(\tau), \mathbf{W}_k(\tau)) \neq y\}, \qquad \mathrm{sign}(0) = 1,$$

where the test example is independent of the realized training history. For universal constants $C, c > 0$, with probability at least $1 - \delta$, there are iterations $1 \le \tau_A < \tau_X \le CT/(\eta\|\mathbf{p}\|_2)$ such that

$$\mathrm{err}_1(\tau_A) + \mathrm{err}_2(\tau_X) \le C \exp\left\{-c\|\mathbf{p}\|_2\sqrt{\log\log T}\right\},$$
$$\inf_{0\le\tau\le\tau_A} \mathrm{err}_2(\tau) \ge \frac{1}{4}.$$

The infimum is over integer iterations. Both tasks are trained throughout with fixed batch composition and learning rate.

Theorem 4.2 (Informal version of Theorem A.2, training with only XOR task fails)**.** *Under the same assumptions and learning-rate choice, train the attention weight and XOR head using only XOR examples, with the same initialization, batch size, and learning rate as the mixed learner. Let $\mathrm{err}'_2(\tau)$ be its conditional test error on a fresh XOR example. With probability at least $1 - \delta$, jointly with the upper-bound event, for every integer $0 \le \tau \le \tau_X$,*

$$\mathrm{err}'_2(\tau) \ge \frac{1}{2} - \frac{C}{\sigma_\epsilon\sqrt{T}} - Ce^{-cd} \ge \frac{1}{4}.$$

At each common iteration, the two runs have used the same number of updates and the same total number of examples. The formal lower bound holds on the entire deterministic budget containing τ_X (Theorem A.2).

Theorem 4.1 describes a learning order that emerges within simultaneous optimization: low token-selection error precedes low XOR error, although neither task is withheld from training. Theorem 4.2 shows that the token-selection data are essential for learning XOR within this budget. These statements concern test errors at the indicated checkpoints; they do not require the attention mass on relevant tokens to approach one.

This result gives a concrete mechanism for an implicit curriculum. Empirically, language models often acquire token-level or information-localization capabilities before skills that combine multiple pieces of information (Liu et al., 2026; Allen-Zhu & Li, 2024; 2025). Here such an ordering arises under a fixed data mixture through the coupling of shared attention and task-specific heads.

5 Theoretical Analysis

We outline the proofs of Theorems 4.1 and 4.2 under Assumption 3.1 and the stated learning-rate conditions. We use physical time $t = \eta\tau$; $\asymp$ denotes comparison up to universal constant factors.

5.1 Attention amplification and parity-induced learning

Write $s_i(\mathbf{X}, \mathbf{q}) = [\mathbb{S}(\mathbf{X}\mathbf{q})]_i$ and $\nu_0 = \sigma_\epsilon/\sqrt{T}$. The token-selection population reference of Appendix E has $\mathbf{q}_{\mathrm{sh}} = \gamma\mathbf{p}/\|\mathbf{p}\|_2^2$ and semantic row coordinate $\mathrm{sign}(a_j)b$ along $(\boldsymbol{\mu}_1 + \boldsymbol{\mu}_2)/\sqrt{2}$, after rescaling rows by $\sqrt{m}$. Its head–attention-weight feedback satisfies

$$\dot{b} \asymp \frac{e^\gamma}{T}, \qquad \dot{\gamma} \asymp \|\mathbf{p}\|_2^2 b\frac{e^\gamma}{T}, \qquad b^2 \asymp \frac{\gamma}{\|\mathbf{p}\|_2^2} \quad (\gamma \ge 1). \tag{1}$$

Thus token-selection feature growth amplifies attention, which accelerates subsequent head updates. For the joint population reference $\bar{\mathbf{q}}(t) = \bar{\gamma}(t)\mathbf{p}/\|\mathbf{p}\|_2^2$, define $K(t) = e^{\bar{\gamma}(t)}/(\sigma_\epsilon\sqrt{T})$. On typical XOR inputs, $s_1 = s_2 \asymp \nu_0 K(t)$ and the conditional perpendicular noise scale $\nu = (\sigma_\epsilon^2 \sum_i s_i^2)^{1/2} \asymp \nu_0$. The *semantic learning clock* is

$$A(t) := \frac{\lambda}{2\sqrt{2\pi}} \int_0^t \mathbb{E}_{\text{pos}}\left[\frac{s_1^2}{\nu}\right] \mathrm{d}u, \qquad A(t_K) \asymp \frac{\lambda K}{\|\mathbf{p}\|_2 \sqrt{\log(\sigma_\epsilon\sqrt{T}K)}}. \tag{2}$$

Here the expectation averages fresh XOR position coordinates at $\bar{\mathbf{q}}(u)$, and t_K is the first time $K(t) = K$. The clock estimate holds at the checkpoint levels below, by the time change in (1) and comparison with the joint reference (Proposition D.1).

Parity averaging cancels forces depending on only one semantic sign. Define $\mathbf{J} = \boldsymbol{\mu}_1\boldsymbol{\mu}_2^\top + \boldsymbol{\mu}_2\boldsymbol{\mu}_1^\top$ and $\varepsilon_j = \sqrt{m}\, a_j$. The leading XOR trajectory is $\widetilde{\mathbf{w}}_{2,j}(t) = e^{\varepsilon_j A(t)\mathbf{J}}\mathbf{w}_{2,j}(0)$: the two readout signs create opposite correlations between semantic coordinates (Lemma H.3). Let $\widetilde{\mathbf{W}}_2$ collect these rows and, for $\mathbf{v} \in \mathbb{R}^d$, define $\widetilde{f}_t(\mathbf{v}) = \mathbf{a}^\top \phi(\widetilde{\mathbf{W}}_2(t)\mathbf{v})$. For deterministic $\mathbf{v}$, put $v_i = \boldsymbol{\mu}_i^\top \mathbf{v}$. Averaging over the Gaussian seeds and readout signs, Lemma I.1 gives

$$\mathbb{E}_{\text{init}} \widetilde{f}_t(\mathbf{v}) = \frac{2\sinh(2A(t))v_1v_2}{\sqrt{2\pi d}\big(\|e^{A(t)\mathbf{J}}\mathbf{v}\|_2 + \|e^{-A(t)\mathbf{J}}\mathbf{v}\|_2\big)}. \tag{3}$$

The value at $\mathbf{v} = \mathbf{0}$ is zero. Thus training produces a product feature. At the checkpoints $K_A = \sqrt{K_X}$ and $K_X = c_*\|\mathbf{p}\|_2\sqrt{\log\log T}$, with $c_* > 0$ small enough that $A \le 1$, set $\tau_A = \lceil t_{K_A}/\eta \rceil$ and $\tau_X = \lceil t_{K_X}/\eta \rceil$. The typical relevant-token masses are

$$s_1(\mathbf{X}, \mathbf{q}(\tau_A)) \asymp \nu_0 K_A \quad (\mathcal{P}_1), \qquad (s_1 + s_2)(\mathbf{X}, \mathbf{q}(\tau_X)) \asymp \nu_0 K_X \quad (\mathcal{P}_2). \tag{4}$$

The token-selection margin already has scale $\nu_0 K_A/\|\mathbf{p}\|_2$. For XOR, the learned signal has scale $\nu_0 A K^2/d$, compared with initialization fluctuations $\nu_0/\sqrt{m}$. The width conditions give

$$\frac{\sqrt{m}\, A(t_{K_A})K_A^2}{d} \le C\sqrt{c_u}, \tag{5}$$

$$\frac{mA(t_{K_X})^2 K_X^4}{d^2} \ge cC_\ell\Big[(\|\mathbf{p}\|_2^2 \log\log T)^{1/2} + \log(C_{\text{alloc}}/\delta)\Big]. \tag{6}$$

Anti-concentration of the label-independent initialization output keeps XOR error at least $1/4$ through τ_A (Proposition I.5); late concentration proves success once the tracking errors below are controlled (Proposition J.3). Specifically, with $E_X := mA(\eta\tau_X)^2 K_X^4/d^2$,

$$\mathrm{err}_2(\tau_X) \le Ce^{-cd^{1/4}} + Ce^{-cK_X^2} + e^{-cE_X} + 2e^{-cK_X^4}. \tag{7}$$

5.2 Tracking simultaneous SGD at the classification scale

Softmax invariance to common translations leaves an independent Gaussian component of variance ν_0^2 in the representation, in all directions for $\mathcal{P}_1$ and in $\mathbf{p}^\perp$ for $\mathcal{P}_2$ (Lemma C.5). Independence from the entire centered random map of $\mathbf{q}$ permits smoothing of moving gates. For the token-selection reference, $c(t)$ is the coordinate of its $\sqrt{m}$-rescaled head rows along the unit direction $\mathbf{p}/\|\mathbf{p}\|_2$. Gaussian convolution gives, for universal $c_0 > 0$,

$$c(t) \le \frac{C}{\|\mathbf{p}\|_2}\left(1 + \int_2^{e^{\gamma(t)}/(\sigma_\epsilon\sqrt{T})} e^{-c_0u^2}\,\mathrm{d}u\right) \le \frac{C}{\|\mathbf{p}\|_2} \quad \left(\frac{e^{\gamma(t)}}{\sigma_\epsilon\sqrt{T}} \ge 2\right). \tag{8}$$

This sharper position scale, rather than $b(t) \asymp \sqrt{\gamma}/\|\mathbf{p}\|_2$, saves a $\log\log T$ factor in the curvature integral.

Lemma 5.1 (Informal version of Lemma E.7)**.** *There is a uniform envelope $L(t)$ for the local joint-drift and individual-row self-error coefficients on the comparison neighborhoods, such that*

$$\int_0^{\eta\tau_X} L(t)\,\mathrm{d}t \le \frac{\log d}{128}, \qquad \eta\sum_{\tau=0}^{\tau_X-1} L(\eta\tau) \le \frac{\log d}{128}. \tag{9}$$

Here $L(t)$ bounds the joint-state error coefficient and each individual row's self-error coefficient in the drift comparison estimates.

This limits error amplification to $d^{1/128}$. A stopped comparison then controls the attention weight and both heads throughout simultaneous SGD. Parity further cancels the lower-order position force of the XOR head. Write $\Delta \mathbf{W}_2(\tau) := \mathbf{W}_2(\tau) - \widetilde{\mathbf{W}}_2(\eta\tau)$ and define $\mathbf{\Pi}_s := \boldsymbol{\mu}_1\boldsymbol{\mu}_1^\top + \boldsymbol{\mu}_2\boldsymbol{\mu}_2^\top$ and $\mathbf{\Pi}_N := \mathbf{I}_d - \mathbf{p}\mathbf{p}^\top/\|\mathbf{p}\|_2^2 - \mathbf{\Pi}_s$. Here $\|\cdot\|_F$ denotes the Frobenius norm.

Lemma 5.2 (Informal version of Theorem H.1)**.** *With probability at least $1-\delta$, jointly with the attention-weight comparison, for every integer $0 \le \tau \le \tau_X$,*

$$\|\Delta \mathbf{W}_2(\tau)\mathbf{\Pi}_s\|_F + \|\Delta \mathbf{W}_2(\tau)\mathbf{\Pi}_N\|_F \le \frac{CA(\eta\tau)}{d} + Cd^{-11/8},$$
$$\left\|(\mathbf{W}_2(\tau) - \mathbf{W}_2(0))\frac{\mathbf{p}}{\|\mathbf{p}\|_2}\right\|_2 \le Cd^{-11/8}. \tag{10}$$

Direction-dependent tracking bounds the prediction discrepancy by $\nu_0\{CA(t)(1+K(t))/d + Cd^{-5/4}\}$ in fresh-input root-mean-square norm (see (126)), controlling the logistic residual within the same bootstrap. To retain exponential error bounds, we use Gaussian projections rather than the full noise norm.

Lemma 5.3 (Informal version of Lemma C.2)**.** *For fixed $\mathbf{W}, \widehat{\mathbf{W}} \in \mathbb{R}^{m\times d}$, put $\Delta\mathbf{W} = \mathbf{W} - \widehat{\mathbf{W}}$. If $\mathbf{v} \sim \mathcal{N}(\boldsymbol{b}, \mathbf{\Sigma})$ has mean $\boldsymbol{b} \in \mathbb{R}^d$ and covariance $\mathbf{\Sigma} \succeq 0$, then for $r > 0$,*

$$\Pr\left\{\left|\mathbf{a}^\top[\phi(\mathbf{W}\mathbf{v}) - \phi(\widehat{\mathbf{W}}\mathbf{v})]\right| > \|\Delta\mathbf{W}\boldsymbol{b}\|_2 + r\right\} \le 2\exp\left\{-\frac{r^2}{4\|\Delta\mathbf{W}\mathbf{\Sigma}^{1/2}\|_F^2}\right\}. \tag{11}$$

Here $\mathbf{\Sigma}^{1/2}$ is the positive semidefinite square root. The bound is zero when its denominator vanishes.

Conditioning on training, fresh positions, and signs permits dependent trained rows. At τ_X, the leading head error is smaller than the learned signal by a factor $1/K_X$; the Gaussian remainder costs $2e^{-cK_X^4}$.

5.3 The obstruction for XOR-only training

For the XOR-only attention weight $\mathbf{q}'(\tau)$ and head $\mathbf{W}_2'(\tau)$, take the budget $H_d = \lceil C_H T/(\eta\|\mathbf{p}\|_2)\rceil$ containing τ_X and radius $Q = C_0\sigma_\epsilon^2\eta H_d/\sqrt{d}$, with sufficiently large universal C_H, C_0. Parity suppresses the population head drift near initialization. A stopped martingale argument controls fresh-batch fluctuations, keeps the head small, and consequently limits attention-weight movement. Lemma K.3 gives, with high probability,

$$\sup_{0\le\tau\le H_d}\|\mathbf{q}'(\tau)\|_2 \le Q, \qquad \sup_{0\le\tau\le H_d}\|\mathbf{W}_2'(\tau)\| \le \frac{C}{\sqrt{d}}, \qquad \sigma_\epsilon Q\sqrt{T} = o(1). \tag{12}$$

The suprema are over integer iterations. The remaining obstruction is uniform over this attention-weight ball, independently of the prediction head. On a typical input event, flipping one semantic sign changes the perpendicular pool by $O(1/T)$; translation against its common Gaussian noise costs $O((1/T)/\nu_0) = O(1/(\sigma_\epsilon\sqrt{T}))$ in total variation. Since this noise has no position component, we additionally transport the scalar Gaussian distractor positions to match the position coordinate (Lemma K.4). Pairing opposite-label clusters yields

$$\sup_{\|\mathbf{q}\|_2\le Q}\mathrm{TV}(\mathrm{Law}(\widetilde{\mathbf{x}}(\mathbf{X},\mathbf{q}) \mid y=1), \mathrm{Law}(\widetilde{\mathbf{x}}(\mathbf{X},\mathbf{q}) \mid y=-1)) \le \frac{C}{\sigma_\epsilon\sqrt{T}} + Ce^{-cd}, \tag{13}$$

where $\widetilde{\mathbf{x}}(\mathbf{X},\mathbf{q}) = \mathbf{X}^\top\mathbb{S}(\mathbf{X}\mathbf{q})$, Law denotes distribution and TV total variation distance (Lemma K.5). Balanced labels make the Bayes error one half minus one half of this distance. Conditional on the realized training trajectory, the fresh test example retains its original law. The uniform representation bound therefore applies to every data-dependent attention weight and head, giving

$$\inf_{0\le\tau\le H_d}\mathrm{err}_2'(\tau) \ge \frac{1}{2} - \frac{C}{\sigma_\epsilon\sqrt{T}} - Ce^{-cd}. \tag{14}$$

No union bound over attention weights or iterations is needed. Unlike early mixed-training failure, this obstruction persists even for an optimal classifier of the pooled representation.

6 EXPERIMENTS

The theory predicts a specific learning order: the model first learns to select the informative tokens, and only then learns the compositional rule over the shared representation. We test this prediction at three levels: data sampled from the theoretical model, MNIST-derived visual tokens, and open language-model pretraining checkpoints.

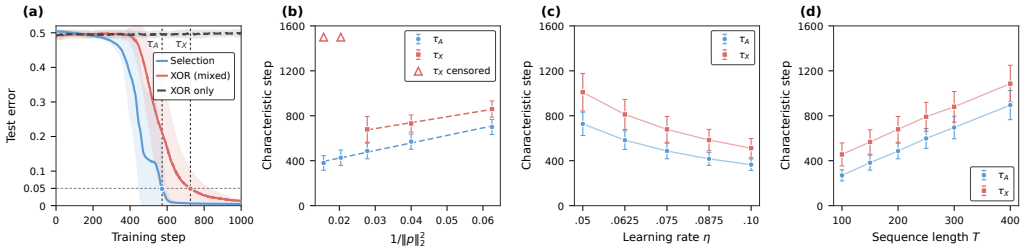

Figure 1: **Synthetic training dynamics and parameter sweeps.** Here τ_A and τ_X are the first steps at which the corresponding test errors fall below 0.05 and remain below 0.05 through final step. (a) Test errors under joint and XOR-only training, with τ_A and τ_X marked by dashed lines. (b)–(d) Characteristic steps under sweeps over $1/\|\mathbf{p}\|_2^2$, learning rate η, and sequence length T, respectively. Triangles indicate runs where τ_X exceeds the step budget.

Synthetic data. We train the shared-attention model online with fresh synthetic batches at every step; full settings are given in Appendix M.1.

Figure 1(a) corroborates the learning-order separation in Theorem 4.1: token selection reaches near-zero test error at τ_A while XOR remains near chance, and XOR improves only later at τ_X. In contrast, XOR-only training stays near chance throughout the budget, consistent with Theorem 4.2. Figures 1(b)–(d) show that the ordering $\tau_A < \tau_X$ persists across sweeps over $\|\mathbf{p}\|_2$, η, and T, indicating that the two-phase learning dynamics are robust to these parameter choices.

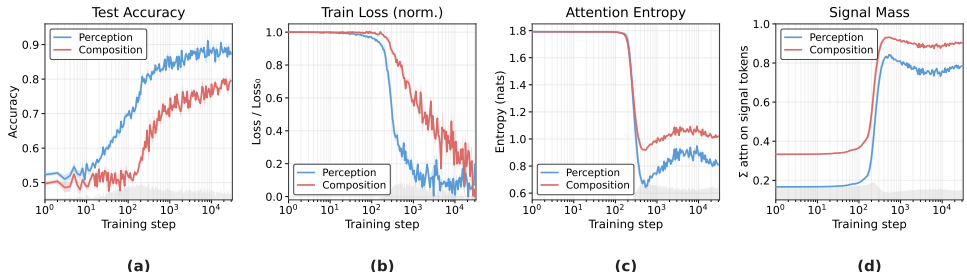

Figure 2: **MNIST single-layer dynamics.** Blue denotes the perception head and red denotes the composition head. The x-axis is logarithmic training step. (a) Test accuracy. (b) Normalized training loss. (c) Entropy of pooling attention weights. (d) Pooling attention mass on signal-token positions. Shaded bands show a 5-point rolling standard deviation.

MNIST data. We next replace the Gaussian signals with MNIST-derived visual tokens. A signal digit u carries parity label $s(u) = +1$ if u is even and -1 otherwise. Perception examples contain one signal token and predict whether its digit is even or odd, while composition examples contain multiple signal tokens and predict the product of their even/odd labels. The single-layer setting (Figure 2) matches our theory model; the multi-signal setting (Figure 3) uses a multi-layer Transformer closer to modern LLM architectures. Full hyperparameters are in Appendix M.2.

Figure 2 shows the same ordering on MNIST inputs: perception accuracy rises early while composition stays near chance much longer. The attention diagnostics confirm the mechanism: entropy drops and signal-token mass rises before composition accuracy takes off. Figure 3 varies the number of signal tokens in the multi-layer setting; perception remains ahead of composition across all settings,

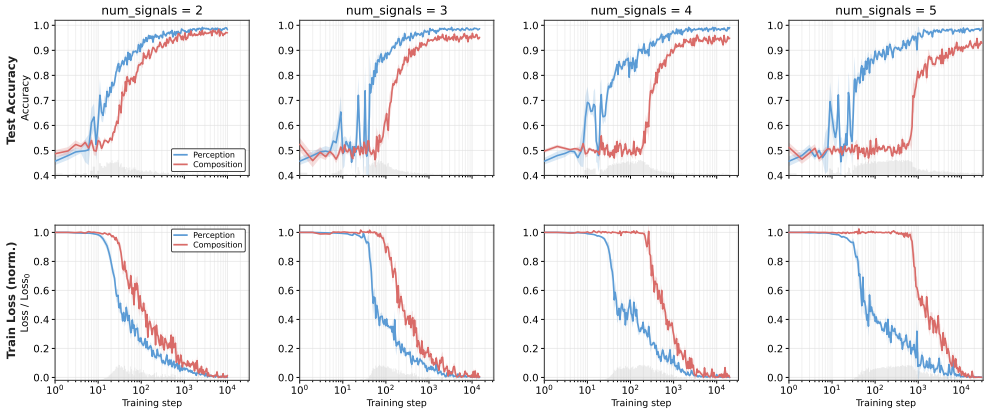

Figure 3: **Multi-layer MNIST sweep over the number of composition signal tokens.** We plot perception and composition test accuracy for 2 to 5 signal tokens.

suggesting that the observed ordering is not merely an artifact of the two-token XOR construction or the shallow model analyzed in theory.

Emergence in real LLM pretraining. We next test whether the same perception-before-reasoning order appears during real LLM pretraining. We evaluate intermediate checkpoints from three open-checkpoint model families: Pythia (Biderman et al., 2023), LLM360 (Liu et al., 2023), and OLMo-2 (OLMo et al., 2024). Amber, the member of the LLM360 family evaluated here, is pretrained on a globally shuffled data mixture without an explicit data curriculum (Liu et al., 2023). Using the benchmark from (Liu et al., 2026), we select four representative subtasks: copy and fact extraction as perception-style tasks, and basic arithmetic and math as reasoning-style tasks. The former mainly require identifying or retrieving relevant information, while the latter require applying an additional rule or computation. Figure 4 shows a consistent temporal separation between perception-style and reasoning-style capabilities. Across all three model families, copy accuracy rises early, and fact extraction improves before arithmetic and math reach high accuracy. The reasoning-style tasks stay low for a longer portion of pretraining and improve only after retrieval-oriented tasks become reliable. Although these checkpoints differ from our controlled theoretical model, the same qualitative ordering suggests that perception-before-reasoning also appears in large-scale language-model pretraining.

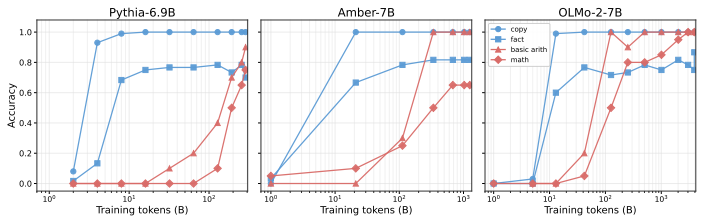

Figure 4: **Capability emergence across real LLM pretraining checkpoints.** Copy and fact extraction are perception-style tasks, while basic arithmetic and math are reasoning-style tasks.

7 Conclusion and Limitations

We showed that an implicit curriculum can arise in joint Transformer training without any explicit data ordering. In a stylized shared-attention multi-task Transformer, gradient descent first learns to attend to relevant tokens, after which the model learns the XOR computation with provable generalization. Experiments on synthetic data, MNIST-based sequences, and pretrained LLM checkpoints exhibited the same qualitative progression, suggesting that the separation between perception-like and reasoning-like learning can emerge naturally from optimization dynamics. Our analysis is deliberately restricted to a controlled setting, where the mechanism can be isolated cleanly. Extending this mechanism to richer architectures, broader task families, and finer-grained capability measures remains an interesting direction.

AI USE STATEMENT

In this work, we used generative AI tools to polish the language of the manuscript. We have reviewed all AI-assisted work and take full responsibility for the final content of this work.

REFERENCES

Emmanuel Abbe, Enric Boix-Adsera, Matthew Brennan, Guy Bresler, and Dheeraj Nagaraj. The staircase property: How hierarchical structure can guide deep learning. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2021.

Emmanuel Abbe, Enric Boix-Adserà, and Theodor Misiakiewicz. The merged-staircase property: a necessary and nearly sufficient condition for SGD learning of sparse functions on two-layer neural networks. In *Conference on Learning Theory*, volume 178, pp. 4782–4887. PMLR, 2022.

Emmanuel Abbe, Elisabetta Cornacchia, and Aryo Lotfi. Provable advantage of curriculum learning on parity targets with mixed inputs. *Advances in Neural Information Processing Systems*, 36: 24291–24321, 2023.

Emmanuel Abbe, Samy Bengio, Aryo Lotfi, Colin Sandon, and Omid Saremi. How far can transformers reason? the globality barrier and inductive scratchpad. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2024.

Tom Alberts and Davar Khoshnevisan. Calculus on gauss space: An introduction to gaussian analysis. *Online lecture notes*, 10, 2018.

Zeyuan Allen-Zhu and Yuanzhi Li. Backward feature correction: How deep learning performs deep (hierarchical) learning. In *Conference on Learning Theory*, volume 195, pp. 4598–4598. PMLR, 2023.

Zeyuan Allen-Zhu and Yuanzhi Li. Physics of language models: part 3.1, knowledge storage and extraction. In *Proceedings of the 41st International Conference on Machine Learning*, pp. 1067–1077, 2024.

Zeyuan Allen-Zhu and Yuanzhi Li. Physics of language models: Part 3.2, knowledge manipulation. In *International Conference on Learning Representations*, 2025.

Zeyuan Allen-Zhu, Yuanzhi Li, and Zhao Song. A convergence theory for deep learning via over-parameterization. In *Proceedings of the 36th International Conference on Machine Learning (ICML)*, volume 97 of *Proceedings of Machine Learning Research*, pp. 242–252, 2019.

Davoud Ataee Tarzanagh, Yingcong Li, Xuechen Zhang, and Samet Oymak. Max-margin token selection in attention mechanism. *Advances in neural information processing systems*, 36:48314–48362, 2023.

Zhangir Azerbayev, Hailey Schoelkopf, Keiran Paster, Marco Dos Santos, Stephen McAleer, Albert Q Jiang, Jia Deng, Stella Biderman, and Sean Welleck. Llemma: An open language model for mathematics. *arXiv preprint arXiv:2310.10631*, 2023.

Jimmy Ba, Murat A. Erdogdu, Taiji Suzuki, Zhichao Wang, Denny Wu, and Greg Yang. High-dimensional asymptotics of feature learning: How one gradient step improves the representation. In *Advances in Neural Information Processing Systems*, volume 35, 2022.

Gregor Bachmann and Vaishnavh Nagarajan. The pitfalls of next-token prediction. In *International Conference on Machine Learning (ICML)*. PMLR, 2024.

Boaz Barak, Benjamin L Edelman, Surbhi Goel, Sham Kakade, Eran Malach, and Cyril Zhang. Hidden progress in deep learning: SGD learns parities near the computational limit. In *Advances in Neural Information Processing Systems*, volume 35, pp. 21750–21764, 2022.

David Bau, Bolei Zhou, Aditya Khosla, Aude Oliva, and Antonio Torralba. Network dissection: Quantifying interpretability of deep visual representations. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 6541–6549, 2017.

Yoshua Bengio, Jérôme Louradour, Ronan Collobert, and Jason Weston. Curriculum learning. In *Proceedings of the 26th Annual International Conference on Machine Learning (ICML)*, pp. 41–48, 2009. doi: 10.1145/1553374.1553380.

Yoshua Bengio, Aaron Courville, and Pascal Vincent. Representation learning: A review and new perspectives. *IEEE transactions on pattern analysis and machine intelligence*, 35(8):1798–1828, 2013.

Deblina Bhattacharjee, Tong Zhang, Sabine Süsstrunk, and Mathieu Salzmann. Muit: An end-to-end multitask learning transformer. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 12021–12031. IEEE, 2022.

Stella Biderman, Hailey Schoelkopf, Quentin Gregory Anthony, Herbie Bradley, Kyle O'Brien, Eric Hallahan, Mohammad Aflah Khan, Shivanshu Purohit, USVSN Sai Prashanth, Edward Raff, et al. Pythia: A suite for analyzing large language models across training and scaling. In *International conference on machine learning*, pp. 2397–2430. PMLR, 2023.

Alberto Bietti, Joan Bruna, Clayton Sanford, and Min Jae Song. Learning single-index models with shallow neural networks. In *Advances in Neural Information Processing Systems*, volume 35, 2022.

Lijie Chen, Binghui Peng, and Hongxun Wu. Theoretical limitations of multi-layer Transformer. *arXiv preprint arXiv:2412.02975*, 2024.

Liam Collins, Hamed Hassani, Mahdi Soltanolkotabi, Aryan Mokhtari, and Sanjay Shakkottai. Provable multi-task representation learning by two-layer relu neural networks. *Proceedings of machine learning research*, 235:9292, 2024.

Alex Damian, Jason D. Lee, and Mahdi Soltanolkotabi. Neural networks can learn representations with gradient descent. In *Proceedings of the 35th Conference on Learning Theory (COLT)*, volume 178 of *Proceedings of Machine Learning Research*, pp. 5413–5452, 2022.

Amit Daniely and Eran Malach. Learning parities with neural networks. *Advances in Neural Information Processing Systems*, 33:20356–20365, 2020.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pp. 4171–4186. Association for Computational Linguistics, 2019.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations*, 2021.

Simon S. Du, Xiyu Zhai, Barnabás Póczos, and Aarti Singh. Gradient descent provably optimizes over-parameterized neural networks. In *International Conference on Learning Representations (ICLR)*, 2019.

Guhao Feng, Bohang Zhang, Yuntian Gu, Haotian Ye, Di He, and Liwei Wang. Towards revealing the mystery behind chain of thought: A theoretical perspective. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2023.

Luciano Floridi and Massimo Chiriatti. Gpt-3: Its nature, scope, limits, and consequences. *Minds and machines*, 30(4):681–694, 2020.

Spencer Frei, Niladri S Chatterji, and Peter L Bartlett. Random feature amplification: Feature learning and generalization in neural networks. *Journal of Machine Learning Research*, 24(303):1–49, 2023.

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.

Behrooz Ghorbani, Song Mei, Theodor Misiakiewicz, and Andrea Montanari. Linearized two-layers neural networks in high dimension. *The Annals of Statistics*, 49(2):1029–1054, 2021.

Margalit Glasgow. Sgd finds then tunes features in two-layer neural networks with near-optimal sample complexity: A case study in the xor problem. *arXiv preprint arXiv:2309.15111*, 2023.

Awni Hannun, Carl Case, Jared Casper, Bryan Catanzaro, Greg Diamos, Erich Elsen, Ryan Prenger, Sanjeev Satheesh, Shubho Sengupta, Adam Coates, and Andrew Y Ng. Deep speech: Scaling up end-to-end speech recognition. *arXiv preprint arXiv:1412.5567*, 2014.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 770–778, 2016.

Arthur Jacot, Franck Gabriel, and Clément Hongler. Neural tangent kernel: Convergence and generalization in neural networks. In *Advances in Neural Information Processing Systems*, volume 31, 2018.

Jiarui Jiang, Wei Huang, Miao Zhang, Taiji Suzuki, and Liqiang Nie. Unveil benign overfitting for transformer in vision: Training dynamics, convergence, and generalization. *Advances in Neural Information Processing Systems*, 37:135464–135625, 2024.

Juno Kim and Taiji Suzuki. Transformers provably solve parity efficiently with chain of thought. In *The Thirteenth International Conference on Learning Representations*, 2025.

Simon Kornblith, Jonathon Shlens, and Quoc V Le. Do better ImageNet models transfer better? In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 2661–2671, 2019.

Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in Neural Information Processing Systems*, volume 25, pp. 1097–1105, 2012.

Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. Deep learning. *nature*, 521(7553):436–444, 2015.

Brian Lester, Rami Al-Rfou, and Noah Constant. The power of scale for parameter-efficient prompt tuning. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pp. 3045–3059. Association for Computational Linguistics, 2021.

Jeffrey Li, Alex Fang, Georgios Smyrnis, Maor Ivgi, Matt Jordan, Samir Gadre, Hritik Bansal, Etash Guha, Sedrick Keh, Kushal Arora, et al. Datacomp-lm: In search of the next generation of training sets for language models. *Advances in Neural Information Processing Systems*, 37:14200–14282, 2024.

Raymond Li, Loubna Ben Allal, Yangtian Zi, Niklas Muennighoff, Denis Kocetkov, Chenghao Mou, Marc Marone, Christopher Akiki, Jia Li, Jenny Chim, et al. Starcoder: may the source be with you! *arXiv preprint arXiv:2305.06161*, 2023.

Xiang Lisa Li and Percy Liang. Prefix-tuning: Optimizing continuous prompts for generation. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pp. 4582–4597. Association for Computational Linguistics, 2021.

Emmy Liu, Kaiser Sun, Millicent Li, Isabelle Lee, Lindia Tjuatja, Jen tse Huang, and Graham Neubig. What do language models learn and when? the implicit curriculum hypothesis. In *Third Conference on Language Modeling*, 2026. URL https://openreview.net/forum?id=wjTjjHstuY.

Zhengzhong Liu, Aurick Qiao, Willie Neiswanger, Hongyi Wang, Bowen Tan, Tianhua Tao, Junbo Li, Yuqi Wang, Suqi Sun, Omkar Pangarkar, et al. Llm360: Towards fully transparent open-source llms. *arXiv preprint arXiv:2312.06550*, 2023.

Roey Magen, Shuning Shang, Zhiwei Xu, Spencer Frei, Wei Hu, and Gal Vardi. Benign overfitting in single-head attention. *arXiv preprint arXiv:2410.07746*, 2024.

Song Mei, Andrea Montanari, and Phan-Minh Nguyen. A mean field view of the landscape of two-layer neural networks. *Proceedings of the National Academy of Sciences*, 115(33):E7665–E7671, 2018.

Xuran Meng, Difan Zou, and Yuan Cao. Benign overfitting in two-layer relu convolutional neural networks for xor data. In *Proceedings of the 41st International Conference on Machine Learning*, pp. 35404–35469, 2024.

Alireza Mousavi-Hosseini, Sejun Park, Manuela Girotti, Ioannis Mitliagkas, and Murat A. Erdogdu. Neural networks efficiently learn low-dimensional representations with SGD. In *International Conference on Learning Representations (ICLR)*, 2023.

Maxwell Nye, Anders Johan Andreassen, Guy Gur-Ari, Henryk Michalewski, Jacob Austin, David Bieber, David Dohan, Aitor Lewkowycz, Maarten Bosma, David Luan, Charles Sutton, and Augustus Odena. Show your work: Scratchpads for intermediate computation with language models. *arXiv preprint arXiv:2112.00114*, 2021.

Chris Olah, Alexander Mordvintsev, and Ludwig Schubert. Feature visualization. *Distill*, 2(11):e7, 2017.

Team OLMo, Pete Walsh, Luca Soldaini, Dirk Groeneveld, Kyle Lo, Shane Arora, Akshita Bhagia, Yuling Gu, Shengyi Huang, Matt Jordan, et al. 2 olmo 2 furious. *arXiv preprint arXiv:2501.00656*, 2024.

Samet Oymak, Ankit Singh Rawat, Mahdi Soltanolkotabi, and Christos Thrampoulidis. On the role of attention in prompt-tuning. In *International Conference on Machine Learning*, pp. 26724–26768. PMLR, 2023.

Abhishek Panigrahi, Bingbin Liu, Sadhika Malladi, Andrej Risteski, and Surbhi Goel. Progressive distillation induces an implicit curriculum. In *The Thirteenth International Conference on Learning Representations*, 2025.

Guilherme Penedo, Quentin Malartic, Daniel Hesslow, Ruxandra Cojocaru, Alessandro Cappelli, Hamza Alobeidli, Baptiste Pannier, Ebtesam Almazrouei, and Julien Launay. The refinedweb dataset for falcon llm: outperforming curated corpora with web data, and web data only. *arXiv preprint arXiv:2306.01116*, 2023.

Binghui Peng, Srini Narayanan, and Christos Papadimitriou. On limitations of the transformer architecture. In *Conference on Language Modeling (COLM)*, 2024.

Tomaso Poggio, Hrushikesh Mhaskar, Lorenzo Rosasco, Brando Miranda, and Qianli Liao. Why and when can deep-but not shallow-networks avoid the curse of dimensionality: a review. *International Journal of Automation and Computing*, 14(5):503–519, 2017.

Keitaro Sakamoto and Issei Sato. Benign overfitting in token selection of attention mechanism. In *Forty-second International Conference on Machine Learning*, 2025.

Clayton Sanford, Daniel Hsu, and Matus Telgarsky. Transformers, parallel computation, and logarithmic depth. In *International Conference on Machine Learning (ICML)*, pp. 43276–43327. PMLR, 2024.

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.

Luca Soldaini, Rodney Kinney, Akshita Bhagia, Dustin Schwenk, David Atkinson, Russell Authur, Ben Bogin, Khyathi Chandu, Jennifer Dumas, Yanai Elazar, et al. Dolma: An open corpus of three trillion tokens for language model pretraining research. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 15725–15788, 2024.

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

Bhavya Vasudeva, Puneesh Deora, and Christos Thrampoulidis. Implicit bias and fast convergence rates for self-attention. *Transactions on Machine Learning Research*, 2024.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in Neural Information Processing Systems*, volume 30, pp. 5998–6008, 2017.

Zixuan Wang, Eshaan Nichani, Alberto Bietti, Alex Damian, Daniel Hsu, Jason D Lee, and Denny Wu. Learning compositional functions with transformers from easy-to-hard data. In *The Thirty Eighth Annual Conference on Learning Theory*, pp. 5632–5711. PMLR, 2025.

Maurice Weber, Daniel Y Fu, Quentin Anthony, Yonatan Oren, Shane Adams, Anton Alexandrov, Xiaozhong Lyu, Huu Nguyen, Xiaozhe Yao, Virginia Adams, et al. Redpajama: an open dataset for training large language models. *Advances in neural information processing systems*, 37: 116462–116492, 2024.

Kaiyue Wen, Huaqing Zhang, Hongzhou Lin, and Jingzhao Zhang. From sparse dependence to sparse attention: Unveiling how chain-of-thought enhances transformer sample efficiency. *arXiv preprint arXiv:2410.05459*, 2024.

Zhiwei Xu, Yutong Wang, Spencer Frei, Gal Vardi, and Wei Hu. Benign overfitting and grokking in relu networks for xor cluster data. In *12th International Conference on Learning Representations, ICLR 2024*, 2024.

Gilad Yehudai and Ohad Shamir. On the power and limitations of random features for understanding neural networks. In *Advances in Neural Information Processing Systems*, volume 32, 2019.

Matthew D Zeiler and Rob Fergus. Visualizing and understanding convolutional networks. In *European Conference on Computer Vision*, pp. 818–833. Springer, 2014.

CONTENTS

1 Introduction **1**

2 Related Work **2**

3 Preliminary **3**

4 Main Results **5**

5 Theoretical Analysis **5**

5.1 Attention amplification and parity-induced learning 5

5.2 Tracking simultaneous SGD at the classification scale 6

5.3 The obstruction for XOR-only training 7

6 Experiments **8**

7 Conclusion and Limitations **9**

A Formal main results **17**

B Notation for the weak-curriculum proof **18**

C Perturbation and concentration lemmas **19**

C.1 Gaussian gates and absolute curvature 21

C.2 Local flows and equivariance 24

D Joint trajectory and proof organization **27**

E Reference progress and cumulative curvature **28**

E.1 Invariant shell and exact population equations 28

E.2 Signed gate identities 30

E.3 Reflection and the initialization cone 30

E.4 One stopped shell: barriers and both clocks 31

E.5 Fresh Gaussian moments with all upper tails included 34

E.6 Operator moments, parity moments, and their integrals 35

E.7 Directional curvature on the same stopped interval 36

E.8 Curvature on complete secants 37

F Joint population and SGD continuation **40**

F.1 A single predictable stop 40

F.2 The error recursion 41

F.3 Common noise on the stopped state tube 41

F.4 Finite width: concentrate accumulated forcing 41

F.5 Fresh batches and Euler discretization . . . 42
F.6 Parity, covariance, and an embedded population reference . . . 44
F.7 Inputs on every secant and grid point . . . 45
F.8 The population and SGD continuations . . . 46
F.9 Checkpoint rounding and auxiliary classification . . . 47

G A parity-sensitive off-axis force estimate 48

H Anisotropic tracking of the XOR head 52
H.1 Gaussian empirical moments with constant coefficients . . . 54
H.2 The exact axial field . . . 55
H.3 Fresh-batch and discretization errors . . . 56
H.4 Logistic output feedback and closure . . . 58

I Classification along the characteristic trajectory 60
I.1 An exact common test family and a position event . . . 60
I.2 Mean and uniform second moment of the characteristic . . . 62
I.3 Initialization anti-concentration and the entire early prefix . . . 64

J Exponential transfer to the trained XOR classifier 67
J.1 Fresh-test events and pointwise attention-weight replacement . . . 67

K The XOR-only comparator on the same training budget 71
K.1 Comparator and a general sufficient condition . . . 71
K.2 Parity suppression for one actual row . . . 71
K.3 Aggregate logistic drift . . . 73
K.4 Actual simultaneous trajectory control . . . 73
K.5 Label indistinguishability in the whole attention-weight ball . . . 75
K.6 Completion and specialization to the primary range . . . 77

L Proof of the mixed-training upper bound 78

M More Details on Experiment 79
M.1 Synthetic Data . . . 79
M.2 MNIST Data . . . 80
M.3 Emergence in Real LLM Pretraining . . . 81

A FORMAL MAIN RESULTS

The data distributions, model, and mixed SGD algorithm are those in Section 3. The test errors are defined in Theorems 4.1 and 4.2. The ReLU map ϕ acts coordinatewise on vector inputs. All constants denoted by C, c, including fixed subscripts, are universal; parameter dependence is displayed explicitly, and δ may vary with the other parameters.

For the proof, write $P = \|\mathbf{p}\|_2$ and abbreviate

$$\rho = \sigma_w\sqrt{md} = 1, \qquad G = \log\log T, \qquad K_X = c_* P\sqrt{G},$$
$$K_A = K_X^{1/2}, \qquad L_d = 1 + \log d, \qquad \ell_\delta = \log(C_{\mathrm{alloc}} d/\delta),$$

where $c_* > 0$ is a sufficiently small universal checkpoint constant. The checkpoints are proof quantities and do not alter SGD. Indeed, its dimension bound gives the upper bound on T and the lower bound on σ_ϵ^2, while $T \geq \exp(P^2/2)$ gives the upper bound on P^2. Together with the initialization and learning-rate choice, the remaining conditions become

$$\rho = 1, \quad \tfrac{1}{4} \leq \lambda \leq \tfrac{1}{2}, \quad 2 \leq B \leq \sqrt{d}, \quad d^{-6} \leq \eta \leq d^{-3}, \tag{15}$$

$$\frac{C_\ell d^2\{(P^2G)^{1/2} + \log(C_{\mathrm{alloc}}/\delta)\}}{\lambda^2(P^2G)^2} \leq m \leq \frac{c_u d^2}{\lambda^2(P^2G)^{1/2}}. \tag{16}$$

The width-dependent term in the dimension condition is equivalent to the upper bound in (16); the two batch-size inequalities are equivalent to $1/4 \leq \lambda \leq 1/2$. These proof ranges describe the same region as Assumption 3.1.

Theorem A.1 (Mixed-training upper bound)**.** *For universal $c_*, C_H, C, c > 0$, under Assumption 3.1, the following holds for every $\delta \in (0, 1/10)$. Select $d^{-6} \leq \eta \leq d^{-3}$ and run mixed SGD. With probability at least $1 - \delta/2$ over initialization and training, there are deterministic integers*

$$1 \leq \tau_A < \tau_X \leq H_d := \left\lceil \frac{C_H T}{\eta P} \right\rceil$$

such that

$$\mathrm{err}_1(\tau_A) + \mathrm{err}_2(\tau_X) \leq C\exp\{-c(P^2G)^{1/2}\}, \tag{17}$$

$$\inf_{0 \leq \tau \leq \tau_A} \mathrm{err}_2(\tau) \geq \tfrac{1}{4}. \tag{18}$$

The infimum ranges over integer iterations. The checkpoints are fixed by a population reference; the algorithm uses a fixed mixture and learning rate throughout. The proof is in Appendix L.

Theorem A.2 (XOR-only lower bound)**.** *Under Assumption 3.1 and the learning-rate choice of Theorem A.1, initialize an XOR-only run with the same $(\mathbf{q}(0), \mathbf{W}_2(0), \mathbf{a})$ and use the same B, η. At each step simultaneously update its attention weight and head on the average logistic loss of B fresh XOR examples. Conditional on initialization, this stream is independent of the mixed-training stream. With probability at least $1 - \delta/2$,*

$$\inf_{0 \leq \tau \leq H_d} \mathrm{err}_2'(\tau) \geq \frac{1}{2} - \frac{C}{\sigma_\epsilon\sqrt{T}} - Ce^{-cd} \geq \frac{1}{4}.$$

The two formal theorems hold jointly with probability at least $1 - \delta$. Both runs use $B\tau$ examples and τ updates at every common integer iteration τ.

Proof. Apply Theorem K.6 with failure probability $\delta/2$. A union bound with Theorem A.1 proves the joint statement; independence of their success events is unnecessary. □

Lemma A.3 (Consequences of the parameter conditions)**.** *Under Assumption 3.1 and the learning-rate choice in (15), the following hold after universal choices of the assumption constants:*

$$cL_d^{3/4}\log L_d \leq P^2G \leq CL_d\log L_d, \qquad \ell_\delta \leq CL_d^{3/2}(\log L_d)^{3/2}, \tag{19}$$

$$L_d^{24}\ell_\delta \leq cd^{1/32}, \qquad Cd^{121/64}L_d^{12}\ell_\delta \leq m \leq d^2. \tag{20}$$

The small constant c and large constant C in (20) can be chosen universally in advance by increasing the existing width coefficient C_ℓ. The same choice can enforce any prescribed numerical lower threshold on d. Every logarithm of a fixed polynomial in $d, m, T, B, \eta^{-1}, H_d, \delta^{-1}$ is at most $C\ell_\delta$.

Proof. The batch-size bound in Assumption 3.1, together with $b_1, b_2 \geq 1$, gives $d \geq B^2 \geq 4$. The data ranges then give $G \asymp \log L_d$ and the first bound in (19). Take $C_\ell \geq 1$ and $c_u \leq 1$; the following estimates are uniform in any subsequent increase of C_ℓ. The two width bounds in Assumption 3.1, written as (16), imply

$$\begin{aligned}(P^2G)^{1/2} + \log(C_{\mathrm{alloc}}/\delta) &\leq (c_u/C_\ell)(P^2G)^{3/2},\\ \ell_\delta &= \log d + \log(C_{\mathrm{alloc}}/\delta)\\ &\leq L_d + C(P^2G)^{3/2}\\ &\leq CL_d^{3/2}(\log L_d)^{3/2}.\end{aligned}$$

In particular, the width and task-proportion conditions in Assumption 3.1 give

$$\begin{aligned}m &\geq cd^2/(P^2G)^{3/2},\\ \frac{m}{d^{121/64}L_d^{12}\ell_\delta} &\geq \frac{cd^{7/64}}{(P^2G)^{3/2}L_d^{12}\ell_\delta}\\ &\geq \frac{cd^{7/64}}{L_d^{15}(\log L_d)^3}\\ &\geq C.\end{aligned}$$

The last inequality and $CL_d^{51/2}(\log L_d)^{3/2} \leq cd^{1/32}$ hold above a fixed numerical threshold, since each positive power of d dominates the displayed logarithmic factors. The upper width bound in Assumption 3.1 gives $m \leq d^2$ because $\lambda \geq 1/4$ and P^2G has the displayed lower bound. Finally, Assumption 3.1, the learning-rate choice in (15), and the definition of H_d give $T \leq d^2$, $B \leq d^{1/2}$, $\eta^{-1} \leq d^6$, and $H_d \leq Cd^8$. Taking logarithms proves the last assertion without fixing δ.

It remains to derive the numerical threshold from Assumption 3.1. Fix all constants and thresholds required by the preceding estimates. Given any fixed $D \geq 4$, if $d \leq D$, the width and data bounds imply

$$\begin{aligned}\frac{C_\ell}{c_u} &\leq P^2G\\ &\leq 2\log T\ \log\log T\\ &\leq 2\log\big[D(\log D)^{9/8}\big]\ \log\log\big[D(\log D)^{9/8}\big].\end{aligned}$$

The first inequality follows by retaining the $(P^2G)^{1/2}$ term in (16); the last uses monotonicity for $d \geq 4$. Choose C_ℓ/c_u larger than the final numerical expression. Then $d > D$, as required. This last increase of C_ℓ only strengthens the width lower bound and preserves all estimates already established; it introduces neither a separate dimension assumption nor dependence on δ in a universal constant. □

B Notation for the weak-curriculum proof

We use the data laws, model, initialization, and SGD of Section 3. Throughout, we impose Assumption 3.1 and the learning-rate choice in (15), and use the equivalent proof ranges above. In this appendix, $\sigma = \sigma_\epsilon$, n is an iteration, and $t = n\eta$ is physical time. The suffixes A and X denote tasks 1 and 2. Set $\mathbf{U}_k = \sqrt{m}\,\mathbf{W}_k$, and write $\mathbf{u}_{k,j}^\top$ for its jth row. Thus $\mathbf{u}_{k,j,0} \sim \mathcal{N}(\mathbf{0}, \rho^2\mathbf{I}_d/d)$ and $\epsilon_j = \sqrt{m}\,a_j \in \{-1, 1\}$. Let $\omega = (\epsilon, \mathbf{u}_{A,0}, \mathbf{u}_{X,0})$ denote a seed tuple. Write φ and Φ for the standard normal density and distribution function. Let $\mathbf{e}_p = \mathbf{p}/P$, $\mathbf{e}_s = (\boldsymbol{\mu}_1 + \boldsymbol{\mu}_2)/\sqrt{2}$, $\mathcal{S} = \mathrm{span}\{\boldsymbol{\mu}_1, \boldsymbol{\mu}_2\}$, $\mathcal{N} = \mathrm{span}\{\mathbf{p}, \boldsymbol{\mu}_1, \boldsymbol{\mu}_2\}^\perp$, and let $\mathbf{\Pi}_E$ be the orthogonal projector onto a subspace E. We abbreviate $\mathbf{\Pi}_\perp = \mathbf{I}_d - \mathbf{e}_p\mathbf{e}_p^\top$, $\mathbf{\Pi}_p = \mathbf{e}_p\mathbf{e}_p^\top$ and $\mathbf{\Pi}_s = \mathbf{\Pi}_{\mathcal{S}}$. For any row array $\mathbf{V}$ with rows $\mathbf{v}_j^\top$, write

$$F_{\mathbf{V}}(\mathbf{h}) = \frac{1}{m}\sum_{j=1}^m \epsilon_j[\mathbf{v}_j^\top\mathbf{h}]_+, \qquad \|\mathbf{V}\|_{m,2}^2 = \frac{1}{m}\sum_j \|\mathbf{v}_j\|_2^2, \qquad \mathbf{h}(\mathbf{X}, \mathbf{q}) = \widetilde{\mathbf{x}}.$$

The attention weights are $s_i = [\mathbb{S}(\mathbf{X}\mathbf{q})]_i$; their covariance is $\mathbf{C}(\mathbf{X}, \mathbf{q}) = D_{\mathbf{q}}\mathbf{h} = \sum_i s_i(\mathbf{x}_i - \mathbf{h})(\mathbf{x}_i - \mathbf{h})^\top$. For $y \in \{-1, 1\}$, put $a_y(v) = y/(1 + e^{yv})$. Write $\nu_0 = \sigma/\sqrt{T}$, $\pi_A = 1 - \lambda$, and $\pi_X = \lambda$. Write $\mathcal{D}_A = \mathcal{P}_1$, $\mathcal{D}_X = \mathcal{P}_2$, $\mathrm{err}_A(n) = \mathrm{err}_1(n)$, $\mathrm{err}_X(n) = \mathrm{err}_2(n)$, and $\mathrm{err}'_X(n) = \mathrm{err}'_2(n)$. The iteration checkpoints n_A, n_X are the τ_A, τ_X of Theorem A.1. The symbol $\mathrm{polylog}(d)$ denotes a fixed universal power of $1 + \log d$; it is used only on terms separated by a fixed power of d.

For $\mathbf{b}_k(\mathbf{h}) = m^{-1} \sum_j \epsilon_j \mathbf{u}_{k,j} \mathbf{1}_{\mathbf{u}_{k,j}^\top \mathbf{h} > 0}$, the scaled-coordinate form of the mixed update in Section 3 is

$$\mathbf{q}_{n+1} - \mathbf{q}_n = \eta \sum_k \frac{\pi_k}{b_k} \sum_{(\mathbf{X}, y) \in D_{k,n}} a_y(f_{k,n}(\mathbf{X})) \mathbf{C}(\mathbf{X}, \mathbf{q}_n) \mathbf{b}_{k,n}(\mathbf{h}), \tag{21}$$

$$\mathbf{u}_{k,j,n+1} - \mathbf{u}_{k,j,n} = \frac{\eta \pi_k \epsilon_j}{b_k} \sum_{(\mathbf{X}, y) \in D_{k,n}} a_y(f_{k,n}(\mathbf{X})) \mathbf{h}(\mathbf{X}, \mathbf{q}_n) \mathbf{1}_{\mathbf{u}_{k,j,n}^\top \mathbf{h}(\mathbf{X}, \mathbf{q}_n) > 0}. \tag{22}$$

Here $b_A = b_1$, $b_X = b_2$, and $D_{A,n} = D_{1,n}$, $D_{X,n} = D_{2,n}$; all gradients use the iteration-n state.

C Perturbation and concentration lemmas

Lemma C.1 (Output perturbation). *Let $(\mathbf{u}, \mathbf{v}, \epsilon)$ be any coupled row law with $|\epsilon| \le 1$, independent of a random input $\mathbf{h}$. Define $F_{\mathbf{u}}(\mathbf{h}) = \mathbb{E}[\epsilon[\mathbf{u}^\top \mathbf{h}]_+ \mid \mathbf{h}]$, $F_{\mathbf{v}}$ analogously, $D^2 = \mathbb{E}\|\mathbf{u} - \mathbf{v}\|_2^2$, and $B_{\mathbf{u}}^2 = \|\mathbb{E}\mathbf{u}\mathbf{u}^\top\|$. Then*

$$\|F_{\mathbf{u}}(\mathbf{h}) - F_{\mathbf{v}}(\mathbf{h})\|_{L^2}^2 \le \|\mathbb{E}\mathbf{h}\mathbf{h}^\top\| D^2. \tag{23}$$

Suppose the entire fresh process $\{\mathbf{h}(\mathbf{q} + a\mathbf{v}_0) : 0 \le a \le 1\}$ is independent of the row law, and its paths are absolutely continuous with derivative $\mathbf{C}(\mathbf{q} + a\mathbf{v}_0)\mathbf{v}_0$ almost everywhere, and the bound below is finite. For every deterministic displacement $\mathbf{v}_0$,

$$\|F_{\mathbf{u}}(\mathbf{h}(\mathbf{q} + \mathbf{v}_0)) - F_{\mathbf{u}}(\mathbf{h}(\mathbf{q}))\|_{L^2} \le B_{\mathbf{u}} \|\mathbf{v}_0\|_2 \int_0^1 \left(\mathbb{E}\|\mathbf{C}(\mathbf{q} + a\mathbf{v}_0)\|^2\right)^{1/2} da. \tag{24}$$

Both statements include empirical row laws.

Proof. Independence of the fresh input and Jensen's inequality give

$$\begin{aligned}
\mathbb{E}|F_{\mathbf{u}}(\mathbf{h}) - F_{\mathbf{v}}(\mathbf{h})|^2 &\le \mathbb{E}\big(\mathbb{E}[|(\mathbf{u} - \mathbf{v})^\top \mathbf{h}| \mid \mathbf{h}]\big)^2 \\
&\le \mathbb{E}|(\mathbf{u} - \mathbf{v})^\top \mathbf{h}|^2 \\
&= \mathbb{E}[(\mathbf{u} - \mathbf{v})^\top (\mathbb{E}\mathbf{h}\mathbf{h}^\top)(\mathbf{u} - \mathbf{v})] \\
&\le \|\mathbb{E}\mathbf{h}\mathbf{h}^\top\| D^2.
\end{aligned}$$

Writing $\Delta\mathbf{h} = \mathbf{h}(\mathbf{q} + \mathbf{v}_0) - \mathbf{h}(\mathbf{q})$, Cauchy–Schwarz and Minkowski yield

$$\begin{aligned}
\|F_{\mathbf{u}}(\mathbf{h}(\mathbf{q} + \mathbf{v}_0)) - F_{\mathbf{u}}(\mathbf{h}(\mathbf{q}))\|_{L^2} &\le \big\|\mathbb{E}[|\mathbf{u}^\top \Delta\mathbf{h}| \mid \Delta\mathbf{h}]\big\|_{L^2} \\
&\le B_{\mathbf{u}} \|\Delta\mathbf{h}\|_{L^2} \\
&\le B_{\mathbf{u}} \int_0^1 \|\mathbf{C}(\mathbf{q} + a\mathbf{v}_0)\mathbf{v}_0\|_{L^2} \, da \\
&\le B_{\mathbf{u}} \|\mathbf{v}_0\|_2 \int_0^1 (\mathbb{E}\|\mathbf{C}(\mathbf{q} + a\mathbf{v}_0)\|^2)^{1/2} \, da.
\end{aligned}$$

□

Lemma C.2 (Gaussian row-error transfer). *Let $\mathbf{U}, \mathbf{V} \in \mathbb{R}^{m \times d}$ have row differences $\mathbf{e}_j = \mathbf{u}_j - \mathbf{v}_j$ and the same signs ϵ_j. For any $\mathbf{b} \in \mathbb{R}^d$ and positive semidefinite $\mathbf{\Sigma}$, let $\mathbf{h} = \mathbf{b} + \mathbf{\Sigma}^{1/2}\mathbf{g}$, where $\mathbf{g} \sim \mathcal{N}(\mathbf{0}, \mathbf{I}_d)$. Put $M = m^{-1} \sum_j |\mathbf{e}_j^\top \mathbf{b}|$ and $D_\Sigma^2 = m^{-1} \sum_j \mathbf{e}_j^\top \mathbf{\Sigma} \mathbf{e}_j$. For every $r > 0$,*

$$\mathbb{P}\{|F_{\mathbf{U}}(\mathbf{h}) - F_{\mathbf{V}}(\mathbf{h})| > M + r\} \le 2 \exp\{-r^2/(4D_\Sigma^2)\}. \tag{25}$$

The right-hand side is zero when $D_\Sigma = 0$. No independence among rows is required. In the original coordinates $\mathbf{U} = \sqrt{m}\,\mathbf{W}$, $\mathbf{V} = \sqrt{m}\,\widehat{\mathbf{W}}$, and $a_j = \epsilon_j/\sqrt{m}$, the same bound holds with

$$M \le \|(\mathbf{W} - \widehat{\mathbf{W}})\mathbf{b}\|_2, \qquad D_\Sigma = \|(\mathbf{W} - \widehat{\mathbf{W}})\mathbf{\Sigma}^{1/2}\|_F.$$

Thus Lemma 5.3 holds with $c = 1/4$.

Proof. Let $r_j = \|\mathbf{\Sigma}^{1/2}\mathbf{e}_j\|_2$ and $\bar{r} = m^{-1}\sum_j r_j$. For $\bar{r} > 0$, omit zero rows and set $w_j = r_j/(m\bar{r})$ and $g_j = \mathbf{e}_j^\top \mathbf{\Sigma}^{1/2}\mathbf{g}/r_j$. Each g_j is standard Gaussian. Thus, for $S = m^{-1}\sum_j |\mathbf{e}_j^\top \mathbf{\Sigma}^{1/2}\mathbf{g}|$,

$$\begin{aligned}
\mathbb{E}e^{S^2/(4D_\Sigma^2)} &\le \mathbb{E}e^{S^2/(4\bar{r}^2)} \\
&= \mathbb{E}\exp\left\{\tfrac{1}{4}\left(\sum_j w_j|g_j|\right)^2\right\} \\
&\le \sum_j w_j \mathbb{E}e^{g_j^2/4} \\
&= \sqrt{2} \le 2.
\end{aligned}$$

Since $|F_{\mathbf{U}}(\mathbf{h}) - F_{\mathbf{V}}(\mathbf{h})| \le M + S$, exponential Markov proves the claim. If $\bar{r} = 0$, then $S = 0$ almost surely. □

Lemma C.3 (Hilbert-space Bernstein inequality)**.** *Let $(\mathbf{S}_l)_{l=0}^L$ be a martingale in a real Hilbert space $\mathcal{H}$, with $\mathbf{S}_0 = 0$, increments $\boldsymbol{\xi}_l$, and filtration $(\mathcal{F}_l)$. Suppose, for an $a > 0$, predictable $v_l \ge 0$, and deterministic $v \ge 0$,*

$$\mathbb{E}[\|\boldsymbol{\xi}_l\|_{\mathcal{H}}^r \mid \mathcal{F}_{l-1}] \le \frac{r!}{2}a^{r-2}v_l \quad (r \ge 2), \qquad \sum_{l=1}^L v_l \le v.$$

Then, for $x > 0$,

$$\mathbb{P}\{\max_{l \le L}\|\mathbf{S}_l\|_{\mathcal{H}} \ge x\} \le 2\exp\left\{-\frac{x^2}{2(v + ax)}\right\}.$$

For every $u > 0$, the maximum is at most $\sqrt{2vu} + 2au$ with probability at least $1 - 2e^{-u}$.

Proof. Fix $0 < \theta < a^{-1}$ and write $F(\mathbf{x}) = \cosh(\theta\|\mathbf{x}\|_{\mathcal{H}})$. The inequalities $D^2F(\mathbf{x})[\mathbf{z},\mathbf{z}] \le \theta^2\|\mathbf{z}\|_{\mathcal{H}}^2 F(\mathbf{x})$ and $F(\mathbf{x} + t\mathbf{z}) \le e^{\theta t\|\mathbf{z}\|_{\mathcal{H}}}F(\mathbf{x})$ imply

$$\begin{aligned}
\mathbb{E}[F(\mathbf{S}_l) \mid \mathcal{F}_{l-1}] &\le F(\mathbf{S}_{l-1})\left[1 + \sum_{r\ge 2}\frac{\theta^r}{r!}\mathbb{E}[\|\boldsymbol{\xi}_l\|_{\mathcal{H}}^r \mid \mathcal{F}_{l-1}]\right] \\
&\le F(\mathbf{S}_{l-1})\left[1 + \frac{v_l}{2}\sum_{r\ge 2}\theta^r a^{r-2}\right] \\
&= F(\mathbf{S}_{l-1})\left[1 + \frac{\theta^2 v_l}{2(1-\theta a)}\right] \\
&\le F(\mathbf{S}_{l-1})\exp\left\{\frac{\theta^2 v_l}{2(1-\theta a)}\right\}.
\end{aligned}$$

The associated nonnegative supermartingale and $\cosh t \geq e^t/2$ give

$$\begin{aligned}\mathbb{P}\{\max_l \|\mathbf{S}_l\|_{\mathcal{H}} \geq x\} &\leq 2\exp\left\{-\theta x + \frac{\theta^2 v}{2(1-\theta a)}\right\} \\ &= 2\exp\left\{-\frac{x^2}{v+ax} + \frac{x^2 v}{2(v+ax)^2(1-ax/(v+ax))}\right\} \\ &= 2\exp\left\{-\frac{x^2}{2(v+ax)}\right\} \qquad \left(\theta = \frac{x}{v+ax}\right) \\ &\leq 2e^{-u} \qquad \left(x = \sqrt{2vu} + 2au\right).\end{aligned}$$

If $v = 0$, every increment is zero almost surely. □

Lemma C.4 (Triangular stability of a stopped recursion)**.** *Let $\mathcal{H}$ be a Hilbert space with row projections $P_j : \mathcal{H} \to \mathbb{R}^d$. Suppose $\mathbf{x}_{n\wedge\tau} = \mathbf{F}_n + \eta \sum_{l<n} \mathbf{1}_{l<\tau}\mathbf{d}_l$ for $0 \leq n \leq N$, where $\tau \leq N+1$, $\eta > 0$, and*

$$\|\mathbf{d}_l\|_{\mathcal{H}} \leq L_l \|\mathbf{x}_l\|_{\mathcal{H}}, \tag{26}$$

$$\|P_j \mathbf{d}_l\|_2 \leq a_l \|P_j \mathbf{x}_l\|_2 + c_l \|\mathbf{x}_l\|_{\mathcal{H}} \quad (l < \tau). \tag{27}$$

All coefficients are nonnegative. Define $D_H = \max_n \|\mathbf{F}_n\|_{\mathcal{H}}$, $D_\infty = \max_{n,j} \|P_j \mathbf{F}_n\|_2$, $\Lambda = \eta \sum_{l<N} \mathbf{1}_{l<\tau} L_l$, $\mathcal{A} = \eta \sum_{l<N} \mathbf{1}_{l<\tau} a_l$, and $\mathcal{C} = \eta \sum_{l<N} \mathbf{1}_{l<\tau} c_l$. Then

$$E_H := \max_n \|\mathbf{x}_{n\wedge\tau}\|_{\mathcal{H}} \leq D_H e^{\Lambda}, \qquad E_\infty := \max_{n,j} \|P_j \mathbf{x}_{n\wedge\tau}\|_2 \leq e^{\mathcal{A}}(D_\infty + \mathcal{C} E_H). \tag{28}$$

Proof. Discrete Gronwall gives, for every $n \leq N$,

$$\begin{aligned}\|\mathbf{x}_{n\wedge\tau}\|_{\mathcal{H}} &\leq \|\mathbf{F}_n\|_{\mathcal{H}} + \eta \sum_{l<n} \mathbf{1}_{l<\tau} \|\mathbf{d}_l\|_{\mathcal{H}} \\ &\leq D_H + \eta \sum_{l<n} \mathbf{1}_{l<\tau} L_l \|\mathbf{x}_l\|_{\mathcal{H}} \\ &\leq D_H \prod_{l<n} (1 + \eta \mathbf{1}_{l<\tau} L_l) \\ &\leq D_H e^{\Lambda}.\end{aligned}$$

Applying the same argument after projecting onto a row yields

$$\begin{aligned}\|P_j \mathbf{x}_{n\wedge\tau}\|_2 &\leq D_\infty + \eta \sum_{l<n} \mathbf{1}_{l<\tau} (a_l \|P_j \mathbf{x}_l\|_2 + c_l E_H) \\ &\leq D_\infty + \mathcal{C} E_H + \eta \sum_{l<n} \mathbf{1}_{l<\tau} a_l \|P_j \mathbf{x}_l\|_2 \\ &\leq (D_\infty + \mathcal{C} E_H) \prod_{l<n} (1 + \eta \mathbf{1}_{l<\tau} a_l) \\ &\leq (D_\infty + \mathcal{C} E_H) e^{\mathcal{A}}.\end{aligned}$$

The last active increment is included, so the bounds also hold at a first exit. □

C.1 Gaussian gates and absolute curvature

Lemma C.5 (A common Gaussian translation)**.** *Let $\mathbf{x}_i = \mathbf{v}_i + \boldsymbol{\eta}_i$, where the array $(\mathbf{v}_i)_i$ is independent of $\boldsymbol{\eta}_i \overset{\text{i.i.d.}}{\sim} \mathcal{N}(\mathbf{0}, \sigma^2\boldsymbol{\Pi})$ and $\boldsymbol{\Pi}$ is an orthogonal projector. Set $\mathbf{g}_0 = T^{-1}\sum_i \boldsymbol{\eta}_i$ and $\widetilde{\mathbf{x}}_i = \mathbf{x}_i - \mathbf{g}_0$. Then, simultaneously for every attention weight,*

$$\mathbf{h}(\mathbf{X}, \mathbf{q}) = \mathbf{c}(\mathbf{q}) + \mathbf{g}_0, \qquad \mathbf{c}(\mathbf{q}) = \mathbf{h}(\widetilde{\mathbf{X}}, \mathbf{q}), \qquad \mathbf{g}_0 \sim \mathcal{N}(\mathbf{0}, \nu_0^2\boldsymbol{\Pi}),$$

and $\mathbf{g}_0$ is independent of the process $(\mathbf{c}(\mathbf{q}), \mathbf{C}(\mathbf{X}, \mathbf{q}))_{\mathbf{q}}$.

Proof. For $\widetilde{\boldsymbol{\eta}}_i = \boldsymbol{\eta}_i - \mathbf{g}_0$, Gaussianity and $\mathrm{Cov}(\mathbf{g}_0, \widetilde{\boldsymbol{\eta}}_i) = \mathbf{0}$ give independence of $\mathbf{g}_0$ and $\widetilde{\mathbf{X}}$. Moreover,

$$\begin{aligned} s_i(\mathbf{X}, \mathbf{q}) &= \frac{e^{\widetilde{\mathbf{x}}_i^\top \mathbf{q} + \mathbf{g}_0^\top \mathbf{q}}}{\sum_l e^{\widetilde{\mathbf{x}}_l^\top \mathbf{q} + \mathbf{g}_0^\top \mathbf{q}}} \\ &= s_i(\widetilde{\mathbf{X}}, \mathbf{q}), \\ \mathbf{h}(\mathbf{X}, \mathbf{q}) &= \sum_i s_i(\widetilde{\mathbf{X}}, \mathbf{q})(\widetilde{\mathbf{x}}_i + \mathbf{g}_0) = \mathbf{c}(\mathbf{q}) + \mathbf{g}_0, \\ \mathbf{C}(\mathbf{X}, \mathbf{q}) &= \sum_i s_i(\widetilde{\mathbf{X}}, \mathbf{q})(\widetilde{\mathbf{x}}_i - \mathbf{c})(\widetilde{\mathbf{x}}_i - \mathbf{c})^\top = \mathbf{C}(\widetilde{\mathbf{X}}, \mathbf{q}). \end{aligned}$$

All quantities on the right depend only on $\widetilde{\mathbf{X}}$. □

The following tools use a row law independent of the fresh example; $\langle\cdot\rangle$ denotes its expectation, including an empirical average as a special case. The readout sign is $\epsilon \in \{-1, 1\}$. Use $F(\mathbf{h}) = \langle \epsilon[\mathbf{u}^\top \mathbf{h}]_+ \rangle$ and the row-label notation $\mathbf{u}_j$ when referring to an individual row. For a row variation $\mathbf{e}$, put $e_{\mathbf{U}}^2 = \langle \|\mathbf{e}\|_2^2 \rangle$. Define

$$B_{\mathbf{U}}^2 = \|\langle \mathbf{u}\mathbf{u}^\top \rangle\|, \quad V = \mathbb{E}\|\mathbf{h}\|_2^2, \quad C_i = \mathbb{E}\|\mathbf{C}\|^i \ (i = 1, 2), \quad K_{\text{tok}} = \max_i \|\mathbf{x}_i - \mathbf{h}\|_2.$$

For several heads, the product Hilbert norm is

$$\|(\mathbf{v}, (\mathbf{e}_k)_k)\|_{\mathcal{H}}^2 = \|\mathbf{v}\|_2^2 + \sum_k \langle \|\mathbf{e}_k\|_2^2 \rangle_k.$$

Lemma C.6 (Gaussian gate moments)**.** *Suppose $\mathbf{h} = \mathbf{c} + \mathbf{g}_0$ with $\mathbf{g}_0 \sim \mathcal{N}(\mathbf{0}, \nu_0^2\boldsymbol{\Pi})$ independent of $(\mathbf{c}, \mathbf{C})$, where $\nu_0 > 0$ and $\mathbf{C} = \mathbf{C}^\top$. Assume $0 < a_0 \le r_j := \|\boldsymbol{\Pi}\mathbf{u}_j\|_2$ and $\|\mathbf{u}_j\|_2 \le R_0$. For a universal $C > 0$,*

$$\|\mathbb{E}[\delta_0(\mathbf{u}_j^\top \mathbf{h})\mathbf{h}\mathbf{h}^\top]\| \le H_g := \frac{C}{\nu_0 a_0}\left[\nu_0^2 + (1 + R_0/a_0)^2\|\mathbb{E}\mathbf{c}\mathbf{c}^\top\|\right], \tag{29}$$

$$\|\mathbb{E}[\delta_0(\mathbf{u}_j^\top \mathbf{h})\mathbf{C}\mathbf{u}_j\mathbf{u}_j^\top \mathbf{C}]\| \le Q_j := \frac{C}{\nu_0 r_j}\|\mathbb{E}[\mathbf{C}\mathbf{u}_j\mathbf{u}_j^\top \mathbf{C}]\|, \tag{30}$$

$$\|\mathbb{E}\langle \delta_0(\mathbf{u}^\top \mathbf{h})\mathbf{C}\mathbf{u}\mathbf{u}^\top \mathbf{C}\rangle\| \le Q_{\text{av}} := \frac{C}{\nu_0}\|\mathbb{E}[\mathbf{C}\mathbf{A}_{\mathbf{U}}\mathbf{C}]\|, \tag{31}$$

where $\mathbf{A}_{\mathbf{U}} = \langle \mathbf{u}\mathbf{u}^\top / \|\boldsymbol{\Pi}\mathbf{u}\|_2 \rangle$. Here δ_0 denotes the Dirac delta at zero, interpreted by the conditional-density formula below. The same bounds hold with δ_0 replaced by any nonnegative probability kernel supported on $[-\varepsilon, \varepsilon]$, provided $\varepsilon \le a_0\nu_0$.

Proof. For a fixed row $\mathbf{u}$, write $r = \|\boldsymbol{\Pi}\mathbf{u}\|_2$ and set

$$\mathbf{S}_{\mathbf{u}} = \mathbf{I}_d - \frac{(\boldsymbol{\Pi}\mathbf{u})\mathbf{u}^\top}{r^2}, \qquad \mathbf{P}_{\mathbf{u}} = \boldsymbol{\Pi} - \frac{(\boldsymbol{\Pi}\mathbf{u})(\boldsymbol{\Pi}\mathbf{u})^\top}{r^2}.$$

Conditionally on the residual variables, Gaussian regression gives

$$\begin{aligned} p_{\mathbf{u}}(z \mid \mathbf{c}) &= \frac{1}{\nu_0 r}\varphi\left(\frac{z - \mathbf{u}^\top \mathbf{c}}{\nu_0 r}\right), \\ \mathbb{E}[\mathbf{h} \mid \mathbf{u}^\top \mathbf{h} = z, \mathbf{c}] &= \mathbf{S}_{\mathbf{u}}\mathbf{c} + z\boldsymbol{\Pi}\mathbf{u}/r^2, \\ \mathrm{Cov}(\mathbf{h} \mid \mathbf{u}^\top \mathbf{h} = z, \mathbf{c}) &= \nu_0^2 \mathbf{P}_{\mathbf{u}}. \end{aligned}$$

Since $\|\mathbf{S_u}\| \le 1 + R_0/a_0$, for every vector $\mathbf{v}$,

$$\begin{aligned}\mathbb{E}[\delta_0(\mathbf{u}^\top\mathbf{h})(\mathbf{v}^\top\mathbf{h})^2] &= \mathbb{E}\big[p_{\mathbf{u}}(0 \mid \mathbf{c})\{(\mathbf{v}^\top\mathbf{S_u c})^2 + \nu_0^2\mathbf{v}^\top\mathbf{P_u v}\}\big] \\ &\le \frac{C}{\nu_0 a_0}\mathbf{v}^\top(\mathbf{S_u}\mathbb{E}[\mathbf{cc}^\top]\mathbf{S_u}^\top + \nu_0^2\mathbf{P_u})\mathbf{v} \\ &\le \frac{C}{\nu_0 a_0}\left[(1+R_0/a_0)^2\|\mathbb{E}\mathbf{cc}^\top\| + \nu_0^2\right]\|\mathbf{v}\|_2^2 \\ &= H_g\|\mathbf{v}\|_2^2.\end{aligned}$$

Independence of $\mathbf{C}$ and $\mathbf{g}_0$ similarly gives

$$\begin{aligned}\mathbb{E}\langle\delta_0(\mathbf{u}^\top\mathbf{h})(\mathbf{v}^\top\mathbf{Cu})^2\rangle &= \mathbb{E}\langle p_{\mathbf{u}}(0 \mid \mathbf{c})(\mathbf{v}^\top\mathbf{Cu})^2\rangle \\ &\le \frac{C}{\nu_0}\mathbb{E}\left\langle\frac{(\mathbf{v}^\top\mathbf{Cu})^2}{\|\mathbf{\Pi u}\|_2}\right\rangle \\ &= \frac{C}{\nu_0}\mathbf{v}^\top\mathbb{E}[\mathbf{C}\mathbf{A}_{\mathbf{U}}\mathbf{C}]\mathbf{v} \\ &\le Q_{\mathrm{av}}\|\mathbf{v}\|_2^2.\end{aligned}$$

The unaveraged calculation proves (30). At $|z| \le \varepsilon$, the additional squared conditional mean is at most $2\varepsilon^2\|\mathbf{v}\|_2^2/a_0^2 \le 2\nu_0^2\|\mathbf{v}\|_2^2$. Integrating the same inequalities against the kernel proves the last claim. □

Lemma C.7 (Absolute population curvature)**.** *For the softmax pool, suppose the hypotheses of Lemma C.6 hold throughout a complete state segment, with fixed a_0, R_0, and suppose the following expression has a finite uniform bound on the segment:*

$$L = C[H_g + Q_{\mathrm{av}} + C_1 + B_{\mathbf{U}}\mathbb{E}(K_{\mathrm{tok}}\|\mathbf{C}\|) + V + B_{\mathbf{U}}^2 C_2] \tag{32}$$

That bound is a Lipschitz constant in $\mathcal{H}$ for the negative gradient of $\mathcal{R} = \mathbb{E}\ell(yF(\mathbf{h}))$. For several heads, sum the bounds with their nonnegative source weights. No convexity of $\mathcal{R}$ is required.

Proof. Let ψ_ε be a smooth nonnegative probability kernel supported on $[-\varepsilon, \varepsilon]$, and use $\phi_\varepsilon = [\cdot]_+ * \psi_\varepsilon$ in place of ReLU. Write $g_\varepsilon = \phi_\varepsilon' \in [0,1]$. For a variation $(\mathbf{v}, \mathbf{e})$, Cauchy–Schwarz gives

$$\begin{aligned}|dF_\varepsilon| &= |\langle\epsilon g_\varepsilon(\mathbf{u}^\top\mathbf{h})(\mathbf{e}^\top\mathbf{h} + \mathbf{u}^\top\mathbf{Cv})\rangle| \\ &\le \langle|\mathbf{e}^\top\mathbf{h}|\rangle + \langle|\mathbf{u}^\top\mathbf{Cv}|\rangle \\ &\le e_{\mathbf{U}}\|\mathbf{h}\|_2 + (\mathbf{v}^\top\mathbf{C}\langle\mathbf{uu}^\top\rangle\mathbf{Cv})^{1/2} \\ &\le e_{\mathbf{U}}\|\mathbf{h}\|_2 + B_{\mathbf{U}}\|\mathbf{C}\|\|\mathbf{v}\|_2.\end{aligned}$$

The second variations are

$$\begin{aligned}d^2F_\varepsilon &= \langle\epsilon\psi_\varepsilon(\mathbf{u}^\top\mathbf{h})(\mathbf{e}^\top\mathbf{h} + \mathbf{u}^\top\mathbf{Cv})^2\rangle \\ &\quad + \langle\epsilon g_\varepsilon(\mathbf{u}^\top\mathbf{h})\{2\mathbf{e}^\top\mathbf{Cv} + \mathbf{u}^\top D_{\mathbf{q}}^2\mathbf{h}[\mathbf{v},\mathbf{v}]\}\rangle, \\ D_{\mathbf{q}}^2\mathbf{h}[\mathbf{v},\mathbf{v}] &= \sum_i s_i(\mathbf{x}_i - \mathbf{h})\{\mathbf{v}^\top(\mathbf{x}_i - \mathbf{h})\}^2, \\ \|D_{\mathbf{q}}^2\mathbf{h}[\mathbf{v},\mathbf{v}]\|_2 &\le K_{\mathrm{tok}}\mathbf{v}^\top\mathbf{Cv} \le K_{\mathrm{tok}}\|\mathbf{C}\|\|\mathbf{v}\|_2^2.\end{aligned}$$

Use $|\ell'| \le 1$, $0 \le \ell'' \le 1/4$, and the positive gate measure:

$$\begin{aligned}|D^2\mathcal{R}_\varepsilon[(\mathbf{v},\mathbf{e}),(\mathbf{v},\mathbf{e})]| &\le \tfrac{1}{4}\mathbb{E}(dF_\varepsilon)^2 + \mathbb{E}|d^2F_\varepsilon| \\ &\le \tfrac{V}{2}e_{\mathbf{U}}^2 + \tfrac{1}{2}B_{\mathbf{U}}^2C_2\|\mathbf{v}\|_2^2 + 2H_g e_{\mathbf{U}}^2 + 2Q_{\mathrm{av}}\|\mathbf{v}\|_2^2 \\ &\quad + 2C_1 e_{\mathbf{U}}\|\mathbf{v}\|_2 + B_{\mathbf{U}}\mathbb{E}(K_{\mathrm{tok}}\|\mathbf{C}\|)\|\mathbf{v}\|_2^2 \\ &\le (2H_g + C_1 + V/2)e_{\mathbf{U}}^2 \\ &\quad + (2Q_{\mathrm{av}} + C_1 + B_{\mathbf{U}}\mathbb{E}(K_{\mathrm{tok}}\|\mathbf{C}\|) + B_{\mathbf{U}}^2C_2/2)\|\mathbf{v}\|_2^2.\end{aligned}$$

The absolute quadratic-form bound controls the self-adjoint Hessian operator. For seed laws, first compute on bounded row variations; density and polarization extend the bounded symmetric form to $\mathcal{H}$. Integrating along the segment gives the claimed Lipschitz bound uniformly in $\varepsilon \le a_0\nu_0$. Each gate has zero probability of equality; dominated convergence sends the smoothed gradient to the ReLU population gradient at both endpoints. Passing to the limit in the integrated bound proves the claim. □

Lemma C.8 (Derivative of one row force)**.** *Under the hypotheses of Lemma C.7, let* $g_j = \mathbf{1}_{\mathbf{u}_j^\top \mathbf{h}>0}$ *and* $\mathbf{b}_j = \epsilon_j \mathbb{E}[a_y(F)\mathbf{h} g_j]$*. For a perturbation* $(\mathbf{v}, \mathbf{e})$*, its variation satisfies*

$$\|D\mathbf{b}_j[\mathbf{v},\mathbf{e}]\|_2 \le H_g\|\mathbf{e}_j\|_2 + \tfrac{1}{4}Ve_{\mathbf{U}} + J_j\|\mathbf{v}\|_2, \qquad J_j = C_1 + \sqrt{H_g Q_j} + \tfrac{1}{4}B_{\mathbf{U}}\sqrt{VC_2}. \quad (33)$$

The derivative can equivalently be read as the integrated inequality along the complete segment, with uniform coefficient bounds.

Proof. Use the nonnegative mollifier from the proof of Lemma C.7; the following calculation is uniform in its width, and the integrated inequality passes to ReLU by dominated convergence. Cauchy–Schwarz gives $|dF| \le e_{\mathbf{U}}\|\mathbf{h}\|_2 + B_{\mathbf{U}}\|\mathbf{C}\|\|\mathbf{v}\|_2$. Using $|a_y| \le 1$ and $\ell'' \le 1/4$,

$$\begin{aligned} D\mathbf{b}_j[\mathbf{v},\mathbf{e}] &= \epsilon_j \mathbb{E}[-\ell''(yF)dF\,\mathbf{h}g_j + a_y(F)\mathbf{C}\mathbf{v}g_j] \\ &\quad + \epsilon_j \mathbb{E}[a_y(F)\mathbf{h}\delta_0(\mathbf{u}_j^\top\mathbf{h})(\mathbf{e}_j^\top\mathbf{h} + \mathbf{u}_j^\top\mathbf{C}\mathbf{v})]. \end{aligned}$$

The positive gate measure and Cauchy–Schwarz imply

$$\begin{aligned} \|D\mathbf{b}_j[\mathbf{v},\mathbf{e}]\|_2 &\le \tfrac{1}{4}\mathbb{E}[|dF|\|\mathbf{h}\|_2] + C_1\|\mathbf{v}\|_2 + H_g\|\mathbf{e}_j\|_2 + \sqrt{H_gQ_j}\|\mathbf{v}\|_2 \\ &\le \tfrac{1}{4}Ve_{\mathbf{U}} + \tfrac{1}{4}B_{\mathbf{U}}\|\mathbf{v}\|_2\mathbb{E}[\|\mathbf{C}\|\|\mathbf{h}\|_2] + C_1\|\mathbf{v}\|_2 + H_g\|\mathbf{e}_j\|_2 + \sqrt{H_gQ_j}\|\mathbf{v}\|_2 \\ &\le \tfrac{1}{4}Ve_{\mathbf{U}} + \tfrac{1}{4}B_{\mathbf{U}}\sqrt{VC_2}\|\mathbf{v}\|_2 + C_1\|\mathbf{v}\|_2 + H_g\|\mathbf{e}_j\|_2 + \sqrt{H_gQ_j}\|\mathbf{v}\|_2 \\ &= H_g\|\mathbf{e}_j\|_2 + \tfrac{1}{4}Ve_{\mathbf{U}} + J_j\|\mathbf{v}\|_2. \end{aligned}$$

□

C.2 Local flows and equivariance

For compact metric marginal seed spaces Ω_k with probability laws μ_k, let $\epsilon_k : \Omega_k \to \{-1, 1\}$ be continuous. A state consists of $\boldsymbol{\theta} = (\mathbf{q}, (\mathbf{u}_k)_k)$, where $\mathbf{u}_k \in C(\Omega_k; \mathbb{R}^d)$. Define

$$\mathfrak{B} = \mathbb{R}^d \times \prod_k C(\Omega_k; \mathbb{R}^d), \qquad \|\boldsymbol{\theta}\|_{\mathfrak{B}} = \|\mathbf{q}\|_2 + \max_k \sup_{\omega\in\Omega_k} \|\mathbf{u}_k(\omega)\|_2.$$

Each head averages its own marginal law:

$$\begin{aligned} F_k(\mathbf{h}) &= \int \epsilon_k(\omega)[\mathbf{u}_k(\omega)^\top\mathbf{h}]_+ \, d\mu_k(\omega), \\ \mathbf{b}_k(\mathbf{h}) &= \int \epsilon_k(\omega)\mathbf{u}_k(\omega)\mathbf{1}_{\mathbf{u}_k(\omega)^\top\mathbf{h}>0}\, d\mu_k(\omega). \end{aligned}$$

For source weights $\pi_k \ge 0$, $\sum_k \pi_k \le 1$, the population field is

$$\begin{aligned} \boldsymbol{\beta}_q(\boldsymbol{\theta}) &= \sum_k \pi_k \mathbb{E}_{\mathcal{D}_k}[a_y(F_k)\mathbf{C}\mathbf{b}_k], \\ \boldsymbol{\beta}_k(\boldsymbol{\theta})(\omega) &= \pi_k\epsilon_k(\omega)\mathbb{E}_{\mathcal{D}_k}[a_y(F_k)\mathbf{h}\mathbf{1}_{\mathbf{u}_k(\omega)^\top\mathbf{h}>0}]. \end{aligned} \quad (34)$$

The seed law and fresh-data law are separate expectations.

Lemma C.9 (Local existence and uniqueness). *Suppose each source admits the common translation of Lemma C.5, with a fixed projector $\mathbf{\Pi}_k$, $\nu_0 > 0$, and $\mathbb{E}\max_i \|\mathbf{x}_i\|_2^4 < \infty$. The field (34) is locally Lipschitz in $\mathfrak{B}$ on*

$$\mathcal{O} = \left\{ \boldsymbol{\theta} : \min_{k,\omega} \|\mathbf{\Pi}_k \mathbf{u}_k(\omega)\|_2 > 0 \right\}.$$

Every initial state in $\mathcal{O}$ has a unique local solution. At a finite endpoint, a bounded state with a uniform positive projected-row lower bound extends to a longer interval.

Proof. Fix $\boldsymbol{\theta}_0 \in \mathcal{O}$ and choose a ball of radius $r_0 > 0$ on which the row bounds are $a \le \|\mathbf{\Pi}_k \mathbf{u}_k\|_2$ and $\|\mathbf{u}_k\|_2 \le R_0$. Every secant in the ball has the same lower bound. Lemmas C.7 and C.8 give finite constants L_0, S_0 such that

$$\|\boldsymbol{\beta}(\boldsymbol{\theta}) - \boldsymbol{\beta}(\boldsymbol{\vartheta})\|_{\mathfrak{B}} \le L_0 \|\boldsymbol{\theta} - \boldsymbol{\vartheta}\|_{\mathfrak{B}}, \qquad \|\boldsymbol{\beta}(\boldsymbol{\theta})\|_{\mathfrak{B}} \le S_0.$$

The bounds are uniform in ω, and the row formula is continuous in ω; hence the field takes values in $\mathfrak{B}$. Choose h_0 with $h_0 S_0 \le r_0/2$ and $h_0 L_0 \le 1/2$. For continuous paths in this ball, set $(\mathcal{P}\boldsymbol{\vartheta})(t) = \boldsymbol{\theta}_0 + \int_0^t \boldsymbol{\beta}(\boldsymbol{\vartheta}(s))\, ds$. Then $\mathcal{P}$ maps the path ball to itself and

$$\begin{aligned}
\|\mathcal{P}\boldsymbol{\vartheta} - \mathcal{P}\boldsymbol{\psi}\|_{C_t\mathfrak{B}} &\le \sup_{t \le h_0} \int_0^t \|\boldsymbol{\beta}(\boldsymbol{\vartheta}(s)) - \boldsymbol{\beta}(\boldsymbol{\psi}(s))\|_{\mathfrak{B}}\, ds \\
&\le L_0 \sup_{t \le h_0} \int_0^t \|\boldsymbol{\vartheta}(s) - \boldsymbol{\psi}(s)\|_{\mathfrak{B}}\, ds \\
&\le h_0 L_0 \|\boldsymbol{\vartheta} - \boldsymbol{\psi}\|_{C_t\mathfrak{B}} \\
&\le \tfrac{1}{2} \|\boldsymbol{\vartheta} - \boldsymbol{\psi}\|_{C_t\mathfrak{B}}.
\end{aligned}$$

The iterates of $\mathcal{P}$ are Cauchy, since their successive differences form a geometric series. Completeness gives a fixed point, and the same inequality gives uniqueness. If the stated endpoint bounds hold, S_0 remains bounded, so the path is Cauchy at that endpoint. Its limit remains in $\mathcal{O}$ and supplies the next local solution. □

Lemma C.10 (Equivariance of marginal population flows). *Let $\mathbf{H}$ be orthogonal, and let $\chi_k \in \{-1, 1\}$ satisfy $(\mathbf{X}\mathbf{H}^\top, \chi_k y) \overset{\text{law}}{=} (\mathbf{X}, y)$ under $\mathcal{D}_k$. Suppose a measure-preserving homeomorphism τ_k of Ω_k obeys $\epsilon_k(\tau_k\omega) = \chi_k \epsilon_k(\omega)$. Define*

$$\mathcal{T}(\mathbf{q}, (\mathbf{u}_k)_k) = \big(\mathbf{H}\mathbf{q}, [\omega \mapsto \mathbf{H}\mathbf{u}_k(\tau_k^{-1}\omega)]_k\big).$$

Then $\boldsymbol{\beta}(\mathcal{T}\boldsymbol{\theta}) = \mathcal{T}\boldsymbol{\beta}(\boldsymbol{\theta})$. A unique solution with $\mathcal{T}\boldsymbol{\theta}_0 = \boldsymbol{\theta}_0$ remains fixed by $\mathcal{T}$.

Proof. Softmax and change of seed variables give

$$\begin{aligned}
\mathbf{h}(\mathbf{X}\mathbf{H}^\top, \mathbf{H}\mathbf{q}) &= \mathbf{H}\mathbf{h}(\mathbf{X}, \mathbf{q}), & \mathbf{C}(\mathbf{X}\mathbf{H}^\top, \mathbf{H}\mathbf{q}) &= \mathbf{H}\mathbf{C}(\mathbf{X}, \mathbf{q})\mathbf{H}^\top, \\
F_k^{\mathcal{T}}(\mathbf{H}\mathbf{h}) &= \chi_k F_k(\mathbf{h}), & \mathbf{b}_k^{\mathcal{T}}(\mathbf{H}\mathbf{h}) &= \chi_k \mathbf{H}\mathbf{b}_k(\mathbf{h}).
\end{aligned}$$

Since $a_{\chi_k y}(\chi_k f) = \chi_k a_y(f)$, each row and attention-weight drift transforms as in (34). Therefore

$$\begin{aligned}
\mathcal{T}\boldsymbol{\theta}(t) &= \mathcal{T}\boldsymbol{\theta}_0 + \int_0^t \mathcal{T}\boldsymbol{\beta}(\boldsymbol{\theta}(s))\, ds \\
&= \boldsymbol{\theta}_0 + \int_0^t \boldsymbol{\beta}(\mathcal{T}\boldsymbol{\theta}(s))\, ds \\
&= (\mathcal{P}\mathcal{T}\boldsymbol{\theta})(t) \\
&= \boldsymbol{\theta}(t),
\end{aligned}$$

where the last equality is uniqueness on successive local intervals. □

Lemma C.11 (Velocity bounds from token moments)**.** *Suppose the hypotheses of Lemma C.6 hold on a population-flow interval, uniformly over all rows. Let h_* be the number of heads. For $M_0 > 0$, put $\Lambda_* = 1 + a_0^{-1} + R_0 + M_0$. If $M = \max_i \|\mathbf{x}_i\|_2$ satisfies $\|M\|_{L^4} \le M_0 P$, $P \ge 1$, $\nu_0 \le 1$, and $\mathbb{E}\|\mathbf{g}_0\|_2^2 \le M_0 P^2$, then*

$$\|\dot{\boldsymbol{\theta}}\|_{\mathcal{H}} \le C\sqrt{1+h_*}\,\Lambda_*^3 P^2, \qquad \sup_{k,j} \|\dot{\mathbf{u}}_{k,j}\|_2 \le C\Lambda_* P.$$

The population velocity and its sampled characteristics have time Lipschitz constant at most $C(1+h_*)\Lambda_*^{12}P^6/\nu_0$ in $\mathcal{H}$ and in the maximum-row norm.

Proof. Convexity of the pool and its covariance give

$$\begin{aligned} \|\mathbf{h}\|_2 \le M, \qquad \|\mathbf{C}\| \le M^2, \qquad K_{\mathrm{tok}} \le 2M, \\ V + \|\mathbb{E}\mathbf{c}\mathbf{c}^\top\| \le 3\mathbb{E}M^2 + 2\mathbb{E}\|\mathbf{g}_0\|_2^2 \le C\Lambda_*^2 P^2, \\ C_1 + C_2/P^2 + P^{-1}\mathbb{E}(K_{\mathrm{tok}}\|\mathbf{C}\|) \le \mathbb{E}M^2 + P^{-2}\mathbb{E}M^4 + 2P^{-1}\mathbb{E}M^3 \le C\Lambda_*^4 P^2, \\ H_g \le C\Lambda_*^7 P^2/\nu_0, \qquad Q_j + Q_{\mathrm{av}} + L \le C\Lambda_*^7 P^4/\nu_0. \end{aligned}$$

The last line uses $B_{\mathbf{U}} \le R_0$, $\|\mathbf{A}_{\mathbf{U}}\| \le R_0^2/a_0$, and (32). The row and attention-weight fields have norms at most $\mathbb{E}\|\mathbf{h}\|_2 \le \Lambda_* P$ and $R_0\mathbb{E}\|\mathbf{C}\| \le \Lambda_*^3 P^2$, respectively. For sufficiently close times s, t, intervening chords have projected row norm at least $a_0/2$. Thus

$$\begin{aligned} \|\dot{\boldsymbol{\theta}}(t) - \dot{\boldsymbol{\theta}}(s)\|_{\mathcal{H}} &\le (C\Lambda_*^7 P^4/\nu_0)\|\boldsymbol{\theta}(t) - \boldsymbol{\theta}(s)\|_{\mathcal{H}} \\ &\le (C\Lambda_*^7 P^4/\nu_0)\int_s^t \|\dot{\boldsymbol{\theta}}(v)\|_{\mathcal{H}}\,dv \\ &\le (C\Lambda_*^7 P^4/\nu_0)(C\sqrt{1+h_*}\,\Lambda_*^3 P^2)|t-s| \\ &\le C(1+h_*)\Lambda_*^{12}P^6|t-s|/\nu_0. \end{aligned}$$

For each characteristic, (33) gives $J_j \le C\Lambda_*^7 P^3/\nu_0$ and, almost everywhere,

$$\begin{aligned} \|\ddot{\mathbf{u}}_{k,j}\|_2 &\le H_g\|\dot{\mathbf{u}}_{k,j}\|_2 + \tfrac{1}{4}V\|\dot{\mathbf{u}}_k\|_{L^2} + J_j\|\dot{\mathbf{q}}\|_2 \\ &\le (C\Lambda_*^7 P^2/\nu_0)(\Lambda_* P) + C\Lambda_*^2 P^2(\Lambda_* P) + (C\Lambda_*^7 P^3/\nu_0)(\Lambda_*^3 P^2) \\ &\le C\Lambda_*^{10}P^5/\nu_0 \\ &\le C(1+h_*)\Lambda_*^{12}P^6/\nu_0. \end{aligned}$$

Partition the time interval and integrate these local bounds. Uniformity over the seed support gives the sampled-row conclusion. □

Lemma C.12 (Integral Euler remainder)**.** *Let $\boldsymbol{\theta}$ be a continuously differentiable path in a normed space $\mathcal{V}$, with $\|\dot{\boldsymbol{\theta}}(t) - \dot{\boldsymbol{\theta}}(s)\|_{\mathcal{V}} \le K|t-s|$. For $N\eta \le R$, define $\mathbf{r}_n = \boldsymbol{\theta}(n\eta) + \eta\dot{\boldsymbol{\theta}}(n\eta) - \boldsymbol{\theta}((n+1)\eta)$. Then $\sum_{n<N}\|\mathbf{r}_n\|_{\mathcal{V}} \le K\eta R/2$.*

Proof. The fundamental theorem of calculus gives

$$\begin{aligned} \sum_{n<N}\|\mathbf{r}_n\|_{\mathcal{V}} &= \sum_{n<N}\left\|\int_0^\eta [\dot{\boldsymbol{\theta}}(n\eta) - \dot{\boldsymbol{\theta}}(n\eta+s)]\,ds\right\|_{\mathcal{V}} \\ &\le \sum_{n<N}\int_0^\eta \|\dot{\boldsymbol{\theta}}(n\eta+s) - \dot{\boldsymbol{\theta}}(n\eta)\|_{\mathcal{V}}\,ds \\ &\le N\int_0^\eta Ks\,ds \\ &= NK\eta^2/2 \le K\eta R/2. \end{aligned}$$

□

D JOINT TRAJECTORY AND PROOF ORGANIZATION

All probability statements below use the initialization and fresh-data laws in Section 3. Conditioning used to construct a deterministic population reference is a proof device and does not change the algorithm. Use the checkpoints $K_A = \sqrt{K_X}$ and $K_X = c_* P\sqrt{G}$ fixed in the main theorem,

$$\Gamma_K = \log(\sigma\sqrt{T}K), \qquad \delta_J = \delta/8.$$

Throughout this part, L_d and ℓ_δ are as in Lemma A.3. All fixed dimension comparisons below follow from Assumption 3.1 through that lemma, with the learning rate chosen as in (15); no separate dimension condition is imposed. The checkpoint scores are of order G. For the deterministic joint population reference constructed below, write

$$\bar{\mathbf{q}}(t) = \bar{\gamma}(t)\mathbf{e}_p/P, \qquad s(t) = e^{\bar{\gamma}(t)}/T, \qquad K(t) = s(t)/\nu_0.$$

At an axial attention weight, conditioning on positions gives the exact XOR pool

$$\mathbf{h} = b\mathbf{e}_p + a(z_1\boldsymbol{\mu}_1 + z_2\boldsymbol{\mu}_2) + \nu\mathbf{z}_\perp, \quad \mathbf{z}_\perp \sim \mathcal{N}(0, \boldsymbol{\Pi}_\perp), \quad a = s_1 = s_2, \quad \nu^2 = \sigma^2 \sum_i s_i^2 \geq \nu_0^2. \tag{35}$$

The Gaussian $\mathbf{z}_\perp$ is independent of signs and position variables. Define

$$A(t) = \frac{\lambda\varphi(0)}{2\rho} \int_0^t \mathbb{E}_{\mathrm{pos}} \frac{a(u)^2}{\nu(u)}\, du. \tag{36}$$

Proposition D.1 (Joint continuation)**.** *Under Assumption 3.1, the initialization and learning rate of Theorem A.1, there are deterministic reference times $0 < t_A < t_X \leq CT/P$ and $n_k = \lceil t_k/\eta \rceil$ such that the following hold with probability at least $1 - \delta_J$. The reference reaches $\bar{\gamma}(t_k) = \Gamma_{K_k}$ and is defined on a buffered interval containing $[0, n_X\eta]$. Its clocks satisfy*

$$A(t_k) \asymp \frac{\lambda K_k}{P\sqrt{\Gamma_{K_k}}}, \quad k \in \{A, X\}, \qquad A(t_X) \leq C\lambda c_*, \tag{37}$$

$$\int_0^{n_X\eta} s\, dt \leq C\sqrt{G}/P, \qquad \int_0^{n_X\eta} s^2\, dt \leq \frac{C\nu_0 K_X}{P\sqrt{G}}, \tag{38}$$

$$\int_0^{n_X\eta} \bar{\gamma}^2 dt \leq CT/P, \qquad \int_0^{n_X\eta} \mathbb{E}\,\|\mathbf{h}_k(\mathbf{X}, \bar{\mathbf{q}}(t))\|_2^2\, dt \leq C/P. \tag{39}$$

Choose c_* as in Section E.7; in particular $A(t_X) \leq 1$. The reference attention weight has $|\dot{\bar{\gamma}}| \leq \mathrm{polylog}(d)$ and $-o(1) \leq \bar{\gamma} \leq \Gamma_{K_X} + 1$. Its characteristics, evaluated at the actual initial seeds, obey uniformly for $0 \leq n \leq n_X$,

$$\begin{aligned} &\|\mathbf{q}_n - \bar{\mathbf{q}}(n\eta)\|_2 + \left(\frac{1}{m}\sum_{k,j} \|\mathbf{u}_{k,j,n} - \bar{\mathbf{u}}_k(n\eta;\omega_j)\|_2^2\right)^{1/2} \\ &\qquad + \max_{k,j} \|\mathbf{u}_{k,j,n} - \bar{\mathbf{u}}_k(n\eta;\omega_j)\|_2 \leq d^{-15/16}. \end{aligned} \tag{40}$$

All row norms are bounded above; auxiliary total row norms and XOR perpendicular row norms are bounded below by fixed positive constants. Uniformly on these iterations,

$$\|\mathbf{W}_{X,n}\| \leq C/\sqrt{d}, \qquad \max_j \|\boldsymbol{\Pi}_{\mathcal{N}^\perp}\mathbf{u}_{X,j,n}\|_2 \leq C\sqrt{\ell_\delta/d},$$

$$\left\| m^{-1}\sum_j \mathbf{u}_{A,j,n}\mathbf{u}_{A,j,n}^\top \right\| \leq CG/P^2. \tag{41}$$

Finally, $\mathrm{err}_A(n_A) \leq Ce^{-cK_A^2}$.

The proof of Proposition D.1 occupies Sections E–F.9.

E REFERENCE PROGRESS AND CUMULATIVE CURVATURE

Throughout this section, impose Assumption 3.1. The auxiliary shell averages equally over $\epsilon = \pm 1$ and uniform unit vectors $\mathbf{n} \in \mathcal{N}$:

$$\mathbf{u}_{\epsilon,\mathbf{n}} = c\mathbf{e}_p + \epsilon b\mathbf{e}_s + r\mathbf{n}, \qquad \mathbf{q} = \gamma\mathbf{e}_p/P, \qquad (b, c, r, \gamma)(0) = (0, 0, \rho, 0).$$

It follows the auxiliary population gradient flow with source weight π_A. Throughout this proof, $\alpha = \sigma\sqrt{T}/P$. Write $s(\gamma) = e^\gamma/T$, $R_K = \inf\{t : \gamma(t) = \Gamma_K\}$, and let R_* be the first hit of $\Gamma_{K_X} + 1$. The scalar $c(t)$ is a shell coordinate; c_* is the fixed checkpoint constant.

Theorem E.1 (Shell progress and sharp clocks)**.** *Under Assumption 3.1, the shell reaches $R_* \le CT/P$. Throughout $[0, R_*]$,*

$$0 \le c \le b, \qquad c \le C/P, \qquad r = \rho + o(1), \qquad b^2 \asymp \gamma/P^2 \quad (\gamma \ge 1).$$

After $K = e^\gamma/(\sigma\sqrt{T})$ reaches 2, also $c \ge c_0/P$. For each endpoint K in this interval with $\Gamma_K \asymp G$,

$$\int_0^{R_K} s\, dt \le C\sqrt{G}/P, \qquad \int_0^{R_K} \frac{s^2}{\nu_0}\, dt \asymp \frac{K}{P\sqrt{\Gamma_K}}. \tag{42}$$

In particular,

$$\int_0^{R_*} s\, dt \le C\sqrt{G}/P, \qquad \int_0^{R_*} s^2 dt \le \frac{C\nu_0 K_X}{P\sqrt{G}}, \qquad \int_0^{R_*} \gamma^j dt \le C(1+j!)T/P \tag{43}$$

for each fixed integer $j \ge 0$. At R_K, the auxiliary shell has error $Ce^{-c_0K^2}$ and margin $c_0\nu_0 K/P$ outside an event of that order.

E.1 INVARIANT SHELL AND EXACT POPULATION EQUATIONS

Lemma E.2 (Invariance of the auxiliary shell)**.** *The auxiliary population vector field with source weight π_A is tangent to the shell defined above whenever $r > 0$. Its locally unique solution starting on the shell therefore stays on the shell while $r > 0$.*

Proof. Let $\mathbf{e}_- = (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)/\sqrt{2}$ and $\mathbf{H}_\mathcal{S} = \mathbf{I}_d - 2\boldsymbol{\Pi}_\mathcal{S}$. The following transformations preserve the auxiliary data and seed laws:

token map $\mathbf{x} \mapsto \mathbf{Hx}$	$y \mapsto \chi y$	$(\epsilon, \mathbf{n}) \mapsto (\chi\epsilon, \mathbf{Hn})$
$\boldsymbol{\mu}_1 \leftrightarrow \boldsymbol{\mu}_2$	$\chi = 1$	$(\epsilon, \mathbf{n})$
$\mathbf{H}\|_\mathcal{N} \in O(\mathcal{N}), \quad \mathbf{H}\|_{\mathcal{N}^\perp} = \mathbf{I}$	$\chi = 1$	$(\epsilon, \mathbf{Hn})$
$\mathbf{H} = \mathbf{H}_\mathcal{S}$	$\chi = -1$	$(-\epsilon, \mathbf{n})$

Write F for the shell output, $\mathbf{b}_F = \nabla_\mathbf{h} F$, and $\mathbf{G}_\epsilon, \mathbf{G}_q$ for its row and attention-weight drifts. The transformation identities of Lemma C.10 give

$$\begin{aligned}
\mathbf{G}_{\chi\epsilon}(\mathbf{Hu}) &= \pi_A\chi\epsilon\mathbb{E}[a_{\chi y}(F(\mathbf{Hh}))\mathbf{Hh}\mathbf{1}_{(\mathbf{Hu})^\top\mathbf{Hh}>0}] \\
&= \pi_A\chi^2\epsilon\mathbf{H}\mathbb{E}[a_y(F(\mathbf{h}))\mathbf{h}\mathbf{1}_{\mathbf{u}^\top\mathbf{h}>0}] \\
&= \mathbf{H}\mathbf{G}_\epsilon(\mathbf{u}), \\
\mathbf{G}_q &= \pi_A\mathbb{E}[\chi a_y(F)\mathbf{H}\mathbf{C}\mathbf{H}^\top\chi\mathbf{H}\mathbf{b}_F] = \mathbf{H}\mathbf{G}_q.
\end{aligned}$$

Class exchange eliminates the $\mathbf{e}_-$ row component. Rotations fixing $\mathbf{n}$ eliminate its orthogonal nuisance components; rotations mapping $\mathbf{n}$ to another direction make the remaining coefficients independent of $\mathbf{n}$. The last reflection identifies the two signs. Hence

$$\mathbf{G}_\epsilon(\mathbf{u}_{\epsilon,\mathbf{n}}) = \dot{c}\,\mathbf{e}_p + \epsilon\dot{b}\,\mathbf{e}_s + \dot{r}\,\mathbf{n}, \qquad \mathbf{G}_q = \dot{\gamma}\,\mathbf{e}_p/P.$$

These are all components of the full auxiliary drift. Local uniqueness from Lemma C.9 identifies the shell solution with that flow. □

Let $G_i \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0,1)$, $\bar{G} = T^{-1}\sum_i G_i$, and

$$t_i = P\mathbf{1}_{i=1} + \sigma G_i, \quad \bar{t} = P/T + \sigma\bar{G}, \quad \delta_i = t_i - \bar{t}, \quad \boldsymbol{\delta} = (\delta_i)_{i=1}^T.$$

At the axial attention weight, define

$$w_i = \frac{e^{\gamma\delta_i/P}}{\sum_l e^{\gamma\delta_l/P}}, \qquad S = w_1, \qquad \omega_2 = \sum_i w_i^2,$$
$$H = \sum_i w_i t_i, \qquad J = \frac{\sum_i w_i^2 t_i}{\omega_2}, \qquad u = S/\sqrt{2}, \qquad \tau = \sigma\sqrt{\omega_2}.$$

The centered Gaussian positions are independent of their average. Conditional on $\boldsymbol{\delta}$, use independent standard Gaussians G_0, G_s and $\mathbf{G}_{\mathcal{N}} \sim \mathcal{N}(\mathbf{0}, \mathbf{I}_{\mathcal{N}})$ to write

$$H = \mu + \nu_0 G_0, \quad \mu = P/T + \sum_i w_i\delta_i, \quad X = u + \tau G_s, \quad \mathbf{z} = \tau\mathbf{G}_{\mathcal{N}}, \quad R_{\mathcal{N}} = \|\mathbf{z}\|_2. \tag{44}$$

Here $X = y\mathbf{e}_s^\top \mathbf{h}$ in distribution; its conditional mean is $S/\sqrt{2}$ for either class. The shell ignores $\mathbf{e}_-$ and is radial in $\mathcal{N}$. Thus, for fixed b, r, conditionally on $\boldsymbol{\delta}$,

$$\begin{aligned} yF(\mathbf{h}) &\overset{\mathrm{d}}{=} f(cH, X, R_{\mathcal{N}}), \\ f(d_0, x, R) &= \tfrac{1}{2}\mathbb{E}_{\mathbf{n}}\{[d_0 + bx + r\mathbf{n}^\top\mathbf{z}]_+ - [d_0 - bx + r\mathbf{n}^\top\mathbf{z}]_+\}, \quad \|\mathbf{z}\|_2 = R. \end{aligned} \tag{45}$$

Write f_d, f_x, f_b, f_r, f_R for the respective derivatives, evaluated at $(d_0, x, R) = (cH, X, R_{\mathcal{N}})$, and set $q_s = f - d_0 f_d$, $r_\ell(v) = (1 + e^v)^{-1}$. Conditional on $(\boldsymbol{\delta}, H)$, abbreviate

$$B = \mathbb{E}[r_\ell(f) f_b], \quad A = \mathbb{E}[r_\ell(f) f_x], \quad D = \mathbb{E}[r_\ell(f) f_d], \quad Q = \mathbb{E}[r_\ell(f) q_s].$$

Lemma E.3 (Reduced population equations)**.** *On the auxiliary shell with $r > 0$, the exact gradient-flow equations are*

$$\dot{b} = \pi_A \mathbb{E}_{\boldsymbol{\delta},H} B,$$
$$\dot{c} = \pi_A \mathbb{E}_{\boldsymbol{\delta},H}[HD], \tag{46}$$
$$\dot{r} = \frac{\pi_A}{r}\mathbb{E}_{\boldsymbol{\delta},H}[Q - bB], \tag{47}$$
$$\dot{\gamma} = \pi_A P^2 \mathbb{E}_{\boldsymbol{\delta},H}[cH'D + (u' - \kappa u)A + \kappa Q], \tag{48}$$

where primes differentiate γ and

$$H' = \frac{1}{P}\sum_i w_i(t_i - H)^2, \qquad \kappa = \frac{J - H}{P}, \qquad u' - \kappa u = \frac{u(t_1 - J)}{P}. \tag{49}$$

For $\gamma \geq 0$, $H' \geq 0$, $\kappa \geq 0$, and $\mu \geq P/T$. At initialization, $\dot{b} = \pi_A/(4\sqrt{2}T)$ and $\dot{c} = \dot{r} = \dot{\gamma} = 0$.

Proof. The coupling preserves the conditional marginal law at every γ, so the risks agree:

$$\mathcal{R} = \mathbb{E}_{\mathcal{D}_A}\ell(yF(\mathbf{h})) = \mathbb{E}\ell(f(cH, X, R_{\mathcal{N}})).$$

Its row variation is

$$D_{\mathbf{u}}\mathcal{R}[\mathbf{e}] = -\mathbb{E}_{\epsilon,\mathbf{n}}\left\langle \epsilon\mathbb{E}_{\mathcal{D}_A}[a_y(F)\mathbf{h}\mathbf{1}_{\mathbf{u}^\top\mathbf{h}>0}], \mathbf{e}\right\rangle.$$

For a parameter variation, orthogonality gives the induced metric

$$\begin{aligned} \|\delta\mathbf{q}\|_2^2 + \mathbb{E}_{\epsilon,\mathbf{n}}\|\delta\mathbf{u}\|_2^2 &= \|\mathbf{e}_p\delta\gamma/P\|_2^2 + \mathbb{E}\|\mathbf{e}_p\delta c + \epsilon\mathbf{e}_s\delta b + \mathbf{n}\delta r\|_2^2 \\ &= (\delta\gamma)^2/P^2 + (\delta c)^2 + (\delta b)^2 + (\delta r)^2, \\ (\dot{b}, \dot{c}, \dot{r}, \dot{\gamma}) &= -\pi_A(\partial_b\mathcal{R}, \partial_c\mathcal{R}, \partial_r\mathcal{R}, P^2\partial_\gamma\mathcal{R}) \\ &= \pi_A\mathbb{E}[r_\ell(f)(f_b, Hf_d, f_r, P^2\partial_\gamma f)]. \end{aligned}$$

The third line uses the tangency already proved, not a projection of the original algorithm. Parameter homogeneity yields $rf_r = f - d_0 f_d - bf_b = q_s - bf_b$, proving the three row equations.

Differentiate the reduced-risk representation (44) with its Gaussian variables fixed. The weights and conditional scales satisfy

$$w_i' = w_i(t_i - H)/P, \qquad H' = \sum_i w_i(t_i - H)^2/P,$$
$$\omega_2' = 2\sum_i w_i^2(t_i - H)/P, \qquad \tau'/\tau = \omega_2'/(2\omega_2) = (J - H)/P = \kappa,$$
$$u' = u(t_1 - H)/P, \qquad X' = u' + \kappa(X - u),$$
$$R_{\mathcal{N}}' = \kappa R_{\mathcal{N}}.$$

The chain rule consequently gives

$$\begin{aligned}\partial_\gamma f &= cH' f_d + X' f_x + R_{\mathcal{N}}' f_R \\ &= cH' f_d + [u' + \kappa(X - u)] f_x + \kappa R_{\mathcal{N}} f_R \\ &= cH' f_d + (u' - \kappa u) f_x + \kappa(X f_x + R_{\mathcal{N}} f_R) \\ &= cH' f_d + (u' - \kappa u) f_x + \kappa q_s.\end{aligned}$$

The last equality is input homogeneity. Substitution gives (48). For the tilted mean $m_t(a) = \sum_i t_i e^{at_i} / \sum_i e^{at_i}$,

$$J - H = \int_{\gamma/P}^{2\gamma/P} \mathrm{Var}_{e^{at}}(t_i)\, da \geq 0, \qquad \mu(\gamma) = P/T + \int_0^\gamma H'(v)\, dv \geq P/T.$$

All first derivatives above have integrable polynomial Gaussian majorants on bounded parameter sets with r bounded below, so differentiation under expectation is justified. Finally, at initialization $f = 0$, $f_b = X/2$, $f_d = f_r = 0$, and $\mathbb{E}X = 1/(\sqrt{2}T)$, giving the stated velocities. □

E.2 Signed gate identities

For fixed $r > 0$ and $R_{\mathcal{N}} > 0$, let F_R denote the CDF of $r\mathbf{n}^\top \mathbf{z}$. In dimension $\dim \mathcal{N} = d - 3 \geq 3$, its density is proportional to $(1 - v^2/(rR_{\mathcal{N}})^2)_+^{(d-6)/2}$ on its support, and is even and nonincreasing on the positive half-line. For $b > 0$ and $x > 0$, put $v = bx$. Integration of the ReLU derivative gives

$$f = \tfrac{1}{2}\int_{d_0 - v}^{d_0 + v} F_R(z)\, dz, \qquad f_d = \tfrac{1}{2}[F_R(d_0 + v) - F_R(d_0 - v)],$$
$$bf_b = xf_x = \tfrac{v}{2}[F_R(d_0 + v) + F_R(d_0 - v)], \qquad q_s = f - d_0 f_d.$$

The following elementary bounds will be used in the shell barriers:

$$0 \leq g := bf_b - d_0 f_d \leq bx, \qquad 0 \leq q_s \leq bx, \qquad 0 \leq f \leq bx. \tag{50}$$

Indeed, for $0 \leq d_0 \leq v$ both coefficients in $2g = (v - d_0)F_R(d_0 + v) + (v + d_0)F_R(d_0 - v)$ are nonnegative. For $d_0 \geq v$, writing $a = d_0 - v$ and using the decreasing density gives

$$\begin{aligned}g &= vF_R(a) - \tfrac{a}{2}[F_R(a + 2v) - F_R(a)] \\ &\geq v[F_R(a) - aF_R'(a)] \\ &= v\left[\tfrac{1}{2} + \int_0^a (F_R'(s) - F_R'(a))\, ds\right] \\ &\geq v/2, \qquad g \leq vF_R(a) \leq v.\end{aligned}$$

Reflection gives $g(-d_0) = v - g(d_0)$. Also f is concave for $d_0 \geq 0$, so $q_s \geq f(0) = v/2$ there; $q_s \leq f \leq v$ and $q_s(-d_0) = v - q_s(d_0)$ give the other bounds. The same CDF formulas yield $f \geq bx/2$, $f_x \geq b/2$, and $f_b \geq x/2$ when $d_0 \geq 0$; the case $b = 0$ follows by continuity.

E.3 Reflection and the initialization cone

For $x > 0$, if $b\tau^2 \leq u/4$, the reflected weight obeys

$$w_R(x) = r_\ell(f(x))p_u(x) - r_\ell(-f(x))p_{-u}(x) \geq 0,$$
$$\tfrac{7}{8} r_\ell(f(x))(p_u(x) - p_{-u}(x)) \leq w_R(x) \leq p_u(x) - p_{-u}(x).$$

Indeed $p_u/p_{-u} = e^{2ux/\tau^2}$, $r_\ell(-f)/r_\ell(f) = e^f$, and $f \le bx \le (2ux/\tau^2)/8$. Stein's identity and $0 \le f_x \le b$ imply

$$\mathbb{E}[X r_\ell(f)] \ge (u - b\tau^2)\mathbb{E} r_\ell(f), \qquad \mathbb{E} r_\ell(f) \ge \tfrac{1}{2} e^{-b(u+\tau\sqrt{2/\pi})}.$$

Thus $0 \le B \le u$, $0 \le Q \le bu$, $D \ge 0$; for $H \ge 0$, $B \ge 3u\mathbb{E} r_\ell(f)/8$ and $A \ge b\mathbb{E} r_\ell(f)/2$. Gaussian radius averaging gives uniformly in d_0, for bounded positive r,

$$0 \le D \le Cbu/\tau. \tag{51}$$

To see this, average f_d over $R_{\mathcal{N}}$ to obtain the difference of two Gaussian CDFs, bounded by $Cbx/(r\tau)$, and use the upper bound on w_R.

Quantitative gate bounds in the small-gate region. At $c = 0$, $f = bx/2$ for every radius, so

$$D(0) = \mathbb{E}_X r_\ell(bX/2)[\Phi(bX/(r\tau)) - \tfrac{1}{2}] \ge c_0 bu/\tau$$

when $b(1 + u/\tau) = o(1)$ and $b(u + \tau) = o(1)$. For a direct lower bound integrate over $I = [\tau, 2\tau]$ if $u \le \tau$, and $I = [u, u + \tau]$ otherwise. In both cases $\int_I x(p_u - p_{-u})dx \ge c_0 u$, the residual is bounded below, and $\Phi(bx/(r\tau)) - 1/2 \ge c_0 bx/\tau$.

For $H \ge 0$ one also has $\mathbb{E}_{X,R_{\mathcal{N}}}[r_\ell(f)g] \ge c_0 bu$. Here is the required quantitative gate step. If $d_0 \ge v$, monotonicity of the density gives $g \ge v[F_R(d_0 - v) - (d_0 - v)F_R'(d_0 - v)] \ge v/2$. If $0 \le d_0 < v$, then $g \ge v(1/2 - v\,\|F_R'\|_\infty) \ge v/4$ when $v\,\|F_R'\|_\infty \le 1/4$. On a typical nuisance radius $\|F_R'\|_\infty \le C/\tau$; the preceding interval I satisfies this condition. Integrating its reflected weight proves the claimed lower bound. The omitted reflected terms are nonnegative.

Position tails retain the vanishing factor. Use the centered event

$$\mathcal{E} = \left\{ \max_i |\sigma(G_i - \bar{G})| \le \frac{P}{10 \log T}, \quad T^{-1} \sum_i \sigma^2 (G_i - \bar{G})^2 \le 2\sigma^2 \right\}.$$

Gaussian and centered chi-square tails, followed by the data bounds in Assumption 3.1, give

$$\begin{aligned} \mathbb{P}(\mathcal{E}^c) &\le 2T \exp\left\{ -\frac{cP^2}{\sigma^2 (\log T)^2} \right\} + Ce^{-cT} \\ &\le d^C \exp\left\{ -\frac{cd}{(\log d)^{5/4}} \right\}. \end{aligned}$$

On $\mathcal{E}$ and $0 \le \gamma \le \Gamma_{K_X} + 1$, $S \asymp s$, $\omega_2 \asymp 1/T$, $t_1 - J \ge P/2$, and $H' \le C(PS + \sigma^2/P)$; also $0 \le \kappa \le C$ and $(t_1 - J)/P \le C$. By (50) and Gaussian radius averaging,

$$|f| \le b|X|, \qquad |f_x| \le b, \qquad |q_s| \le b|X|, \qquad \mathbb{E}_{R_{\mathcal{N}}}|f_d| \le Cb|X|/\tau, \qquad \tau \ge \nu_0.$$

On the stopped region below, each integrand F for $\dot{c}$, $\dot{\gamma}$, or a quadratic norm derivative has, after radius averaging, a majorant bW, where $W \ge 0$ has $\mathbb{E}W^2 \le d^C$ by polynomial Gaussian moments. Choose universal $C_0, c_0 > 0$ and define $\varepsilon_{\mathrm{pos}} = d^{C_0} \exp\{-c_0 d/(\log d)^{5/4}\}$. Then

$$\begin{aligned} |\mathbb{E}[\mathbf{1}_{\mathcal{E}^c} F]| &\le b\mathbb{E}[\mathbf{1}_{\mathcal{E}^c} W] \\ &\le b(\mathbb{E}W^2)^{1/2} \mathbb{P}(\mathcal{E}^c)^{1/2} \\ &\le bd^C \exp\left\{ -\frac{cd}{(\log d)^{5/4}} \right\} \\ &\le b\varepsilon_{\mathrm{pos}}. \end{aligned}$$

The same argument bounds the $\dot{b}$ error by $\varepsilon_{\mathrm{pos}}$.

E.4 One stopped shell: barriers and both clocks

Proof of Theorem E.1. Stop at the first exit from $0 \le c \le b$, $b^2 \le K_0 G/P^2$, $\rho/2 \le r \le 2\rho$, $0 \le \gamma \le \Gamma_{K_X} + 1$, or $t \le C_{\mathrm{stop}} T/P$. The fixed K_0 and C_{stop} are chosen strictly larger than the

constants derived below. All estimates use this stop. On its centered-position event, uniformly for $\gamma = O(G)$,

$$S \asymp s, \quad \tau \asymp \nu_0, \quad u = S/\sqrt{2}, \quad H \sim \mathcal{N}(\mu, \nu_0^2), \quad \mu \le C\nu_0(PK + 1 + \alpha\gamma),$$

and $\mu \ge cP\nu_0 K$ for $K \ge 1$. The conditional mean upper bound is pointwise on this centered event: the common Gaussian translation does not enter H', and

$$\mu(0) = P/T, \qquad \mu' = H' \le C(PS + \sigma^2/P).$$

Integration using $S \asymp e^\gamma/T$ gives $\mu \le C(Ps + \sigma^2\gamma/P)$, which implies the displayed envelope. Under Assumption 3.1 and the stopped-shell bounds, $\mu \asymp P\nu_0 K$ on every fixed strip $K \ge 1$, since $\alpha\gamma/P = o(1)$. The exceptional centered positions, with polynomial weights, still cost $\varepsilon_{\mathrm{pos}}$. Since K is polynomial logarithmic, these tails are smaller than all lower forces used below. The relative-tail bounds above show that the errors in $\dot{c}$, $\dot{\gamma}$, and the quadratic norm derivatives retain a factor b.

By Assumption 3.1 and the stopped-shell bounds, the reflection condition $b\tau^2 \le u/4$ and the small-output condition $b(u + \tau) = o(1)$ hold throughout $b^2 \le CG/P^2$. The small-gate condition $b(1 + u/\tau) = o(1)$ is used only where $K \le 1$. At $c = 0$, averaging the spherical gate over its Gaussian radius gives

$$\mathbb{E}_R f_d(0, x, R) = \Phi(bx/(r\tau)) - \tfrac{1}{2}.$$

Its reflected integral on $[\tau, 2\tau]$ if $u \le \tau$, and on $[u, u + \tau]$ otherwise, is at least

$$\frac{cbu/\tau}{1 + b(1 + u/\tau)}.$$

The residual on this interval is bounded below. Multiplication by the positive mean $\mu \ge P/T$ gives an inward force that dominates $b\varepsilon_{\mathrm{pos}}$, even when the gate difference has saturated.

At $c = b$ use $d_0 = bH$, $v = bX$, and $g = bf_b - d_0 f_d$. The exact inequality $g \ge v/2$ holds whenever $H \ge X > 0$, without any upper bound on bX/τ. For $K \ge 1$, take $X \in [u, u + \tau]$ and $H \ge \mu/2$. The latter event has probability $1 - e^{-cP^2K^2}$, and $H \ge X$ there under Assumption 3.1 and the stopped-shell bounds. This supplies the same lower force $\mathbb{E}[r_\ell(f)g] \ge cbu$. For $K \le 1$, the preceding small-gate bound applies because $b = o(1)$. The identities

$$\dot{c} = \pi_A \mathbb{E}[HD], \qquad \frac{d}{dt}(c^2 - b^2) = -2\pi_A \mathbb{E}[r_\ell(f)g]$$

therefore point inward at $c = 0$ and $c = b > 0$, including the relative position tails. At initialization, $\dot{b} = \pi_A/(4\sqrt{2}T) > 0$ and $\dot{c} = 0$. Thus $0 \le c \le b$ is an invariant cone. The head-growth and forward attention-weight bounds did not use the discarded global small-gate condition: on $H \ge 0$, $f_b \ge X/2$ and $f_x \ge b/2$ for $X > 0$. Using $1/2 \le \pi_A \le 3/4$ from Assumption 3.1, they give

$$c_0 s \le \dot{b} \le Cs, \qquad \dot{\gamma} \ge c_0 P^2 bs. \tag{52}$$

Integrating $\dot{\gamma} \ge cP^2 b\dot{b}$ gives $b^2 \le C\gamma/P^2$. Euler homogeneity and the reflected signs give

$$\frac{d}{dt}(b^2 + c^2 + r^2) = 2\pi_A \mathbb{E}[r_\ell(f)f], \qquad 0 \le \frac{d}{dt}(b^2 + c^2 + r^2) \le Cbs.$$

Consequently, using Assumption 3.1, $\rho^2 \le b^2 + c^2 + r^2 \le \rho^2 + Cb^2$ and $r = \rho + o(1)$, excluding the head-size and radial exits for large K_0. For $\gamma \le 1$, the stopped estimates and Assumption 3.1 give $b, c \le C/P$ and $S \asymp 1/T$, so

$$cH'/\tau \le C\left[\frac{1}{\sigma\sqrt{T}} + \frac{\sigma\sqrt{T}}{P^2}\right] = o(1).$$

Equation (48) gives $\dot{\gamma} \le CP^2 bs \le CP^2 b\dot{b}$ there. Hence $b \ge c_0/P$ at the first $\gamma = 1$ time t_1. Before it, $b \ge c_0 t/T$ and $\gamma \ge c_0 P^2 t^2/T^2$, so $t_1 \le CT/P$. After t_1,

$$dt \le \frac{C}{Ps}\, d\gamma \le \frac{CT}{P} e^{-\gamma} d\gamma. \tag{53}$$

This reaches R_* within $C_{\mathrm{end}}T/P$. Taking $C_{\mathrm{stop}} > C_{\mathrm{end}}$ excludes the time exit. It also gives $\int_0^{R_*} \gamma^j dt \le C(1 + j!)T/P$ for each fixed j.

We record the terminal upper speed used below. For $K \geq 1$, on the centered position event, $\mu \geq cPS$ and $H' \leq CPS$. On $H \geq \mu/2$, the reflected inequality $g = bf_b - cHf_d \geq 0$ gives $cHD \leq bB \leq bu$, and hence $cH'D \leq Cbu$. On $H < \mu/2$, the bound $D \leq Cbu/\tau$ instead gives

$$\mathbb{E}_H[cH'D\mathbf{1}_{H<\mu/2}] \leq Cbu(cH'/\tau)e^{-cP^2K^2} = o(bu).$$

Under Assumption 3.1, the factor cH'/τ is polynomial logarithmic throughout the stop. The other terms of (48) are at most Cbu, and the centered position complement is $b\varepsilon_{\mathrm{pos}}$. Thus

$$\dot{\gamma} \leq CP^2bs \qquad (K \geq 1). \tag{54}$$

We next prove the new position-coordinate estimate. The exact equation is $\dot{c} = \pi_A\mathbb{E}[HD]$. The reflected upper bound $D \leq Cbu/\tau$ gives

$$\dot{c} \leq Cbs(PK + 1 + \alpha\gamma) + b\varepsilon_{\mathrm{pos}}.$$

Divide by (52) and integrate up to $K \leq 2$. Since $dK/d\gamma = K$,

$$c \leq \frac{C}{P} + \frac{C(\Gamma_2 + \alpha\Gamma_2^2)}{P^2} + o(d^{-10}) \leq C/P. \tag{55}$$

The last step uses Assumption 3.1, which gives $G/P = o(1)$ and

$$\frac{\alpha G^2}{P} = \frac{\sigma\sqrt{T}\,G^2}{P^2} \leq C(\log d)^{-3/16}G^2 = o(1).$$

There is also a lower barrier, obtained on the fixed strip $1 \leq K \leq 2$. Restrict H to $I_H = [\mu/2, 2\mu+\nu_0]$ and x to $I_{\mathrm{X}} = [u, u+\tau]$. There is a fixed $\bar{c} > 0$ such that, whenever $c \leq \bar{c}/P$, all Gaussian arguments $(cH \pm bx)/(r\tau)$ on these intervals lie in one fixed bounded interval; this follows from $\mu \asymp P\nu_0K$, $K \leq 2$, and $b = o(1)$. Consequently Gaussian-radius averaging gives $\mathbb{E}_R f_d(cH, x, R) \geq c_2bx/\tau$ there, with c_2 independent of a later choice of barrier constant. The reflected weight obeys $w_R \geq (7/8)r_\ell(f)(p_u - p_{-u})$; uniformly over the radius and $x \in I_{\mathrm{X}}$, $0 \leq f \leq bx = o(1)$, so $r_\ell(f) \geq 1/3$. Since $\int_{I_{\mathrm{X}}} x(p_u - p_{-u})dx \geq c_2u$, it follows that $D(H) \geq c_3bu/\tau$ for $H \in I_H$. Integrating over I_H and subtracting the exponentially smaller negative-H tail gives $\dot{c} \geq c_4bP\nu_0K^2$. Equation (54) gives $\dot{\gamma} \leq CP^2bs$. Thus, until the barrier is reached,

$$dc/d\gamma \geq c_5K/P.$$

Choose once and for all $0 < c_1 < \min\{\bar{c}, c_5/2\}$. If the barrier c_1/P were not reached anywhere on $1 \leq K \leq 2$, then the gain over that whole strip would be at least $c_5P^{-1}\int_1^2 K\,d\log K = c_5/P$. It therefore must be reached by $K = 2$. At this barrier at later K, the same Gaussian gate lower estimate gives a positive force at least $cbP\nu_0K^2e^{-CK^2}$. Its negative-H contribution is at most a polynomial logarithmic factor times $b\nu_0e^{-cP^2K^2}$, and the centered complement is $b\varepsilon_{\mathrm{pos}}$. Both are smaller under Assumption 3.1 and the stopped bounds above. Hence the barrier cannot be crossed downwards.

For completeness, the upper bound after this strip uses a Gaussian convolution that avoids a bound on the largest semantic sample. Reflection and Gaussian radius averaging give the exact upper envelope

$$D(H) \leq \frac{1}{2}\left[\Phi\left(\frac{cH + bu}{\tau\sqrt{r^2 + b^2}}\right) - \Phi\left(\frac{cH - bu}{\tau\sqrt{r^2 + b^2}}\right)\right]. \tag{56}$$

This follows by integrating f_d against the difference of the two Gaussian semantic densities; the full signed integral is exactly the displayed Gaussian convolution. By Assumption 3.1 and the shell bounds, on $H \geq \mu/2$, $c \geq c_1/P$, $b = o(1)$, and $K \geq 2$, the center cH/τ dominates bu/τ by a factor diverging with d. Thus the Gaussian difference is bounded by $Cbu\tau^{-1}\exp[-c(cH/\tau)^2]$. On the complementary H event the unweighted upper bound is sufficient. Integrating the Gaussian H gives

$$\mathbb{E}[|H|D(H)] \leq CbP\nu_0K^2e^{-cK^2}.$$

For example, completing the square in $\exp[-cc(t)^2H^2/\tau^2]\exp[-(H-\mu)^2/(2\nu_0^2)]$ proves this with uniform constants; the center satisfies $c(t)\mu/\tau \geq cK$ and $c(t) \leq b = o(1)$. Division by (52) now yields

$$dc/d\gamma \leq \frac{CK}{P}e^{-cK^2} + o(d^{-10}).$$

Its integral is $O(1/P)$, because $d\gamma = dK/K$. Together with (55), this proves $c \le C/P$ on the entire interval, without assuming that upper bound in the late convolution.

This sharper position bound provides the missing global reverse comparison. For $K \le 1$, Assumption 3.1 and the position-coordinate estimate give

$$\frac{cH'}{\tau} \le C\left(K + \frac{\alpha}{P}\right) \le C.$$

The attention-weight equation, $D \le Cbu/\tau$, and the remaining signed terms therefore give $\dot\gamma \le CP^2bs$. For $K \ge 1$ this is already the terminal-speed estimate. Hence, on the entire positive-time interval,

$$cP^2b\dot b \le \dot\gamma \le CP^2b\dot b, \qquad b^2 \asymp \gamma/P^2.$$

In particular $dt \asymp d\gamma/(P\sqrt{\gamma}\, s)$ for $\gamma \ge 1$. For $\Gamma \ge 2$,

$$\int_1^\Gamma \frac{e^\gamma}{\sqrt{\gamma}}\, d\gamma \asymp \frac{e^\Gamma}{\sqrt{\Gamma}}.$$

Under Assumption 3.1, the initial interval contributes only $C/(P\sigma\sqrt{T})$ to $\int s^2/\nu_0$, negligible at the asserted endpoints. This proves (42). The first clock follows from $\dot b \asymp s$. Finally, outside the centered-position event's negligible complement, $u/\tau \asymp K$ and $\mu/\nu_0 \gtrsim PK$. The shell output has the sign of X and satisfies $f \ge bX/2$ on $H \ge 0$, $X > 0$. Gaussian tails exclude $H < 0$ and $X < u/2$ with probability Ce^{-cK^2}. Since $b \ge c/P$, the remaining margin is at least $c\nu_0 K/P$. Under Assumption 3.1, the centered-position complement is smaller than this error throughout $K \le eK_{\mathrm{X}}$. No strong selection is used: the relevant mass is $\Theta(\nu_0 K) = o(1)$. □

E.5 Fresh Gaussian moments with all upper tails included

In this subsection the displayed moment estimates are uniform for $1 \le r \le 64$. Their proofs use auxiliary Gaussian and exponential-envelope moments of order at most 1024, obtained directly from the tail bounds below. These are fixed numerical ranges, so the constants are universal. Let $J_0 = 1$ for auxiliary data and $J_0 = 2$ for XOR, and write

$$\mathbf{x}_i = \mathbf{d}_i + \zeta_i \mathbf{e}_p + \boldsymbol{\eta}_i, \qquad \boldsymbol{\eta}_i \overset{\text{iid}}{\sim} \mathcal{N}(0, \sigma^2 \mathbf{\Pi}_{\mathbf{p}^\perp}).$$

The $\mathbf{d}_i$ are the original noiseless tokens. Auxiliary ζ_i are all iid $\mathcal{N}(0,\sigma^2)$; XOR has $\zeta_1 = \zeta_2 = 0$ and independent Gaussian distractor ζ_i. Set $\theta = \gamma/P$, $M = \max_i |\zeta_i|$, and $R_\theta = e^{2\theta M}$. Under Assumption 3.1, Gaussian tails imply

$$\sup_{0 \le \gamma \le \Gamma_{K_{\mathrm{X}}+1}} \mathbb{E} R_\theta^r \le C \quad (1 \le r \le 64). \tag{57}$$

Indeed $\mathbb{P}(M > \sigma\sqrt{2\log(2T)} + \sigma u) \le e^{-u^2/2}$; integrating this tail and using $\theta\sigma\sqrt{\log T} = o(1)$ proves the bound, by Assumption 3.1 and the displayed score range. The distractor denominator is at least $(T - J_0)e^{-\theta M}$, so pointwise

$$w_j \le CsR_\theta \ (j \le J_0), \qquad w_i \le CT^{-1}R_\theta \ (i > J_0), \qquad \|\omega_2\|_{L^r} \le C(T^{-1} + s^2) \le C/T.$$

Let $\mathbf{\Xi}$ have columns $\boldsymbol{\xi}_i = \zeta_i \mathbf{e}_p + \boldsymbol{\eta}_i$. A fixed-unit-vector Gaussian chi-square bound and a $1/2$ sphere net of size at most 5^d give

$$\mathbb{P}\{\|\mathbf{\Xi}\| > C\sigma(\sqrt{T} + \sqrt{d} + \sqrt{t})\} \le e^{-t}.$$

Tail integration and Assumption 3.1 yield $\left\| \left\| T^{-1}\mathbf{\Xi}\mathbf{\Xi}^\top \right\| \right\|_{L^r} \le C\sigma^2$. A separate union bound over the column chi-square tails gives $\|\max_i \|\boldsymbol{\xi}_i\|_2\|_{L^r} \le C$. These statements allow the heterogeneous projected XOR columns. For the attention covariance $\mathbf{C}$,

$$0 \preceq \mathbf{C} \preceq 2\sum_i w_i \mathbf{d}_i\mathbf{d}_i^\top + 2\sum_i w_i \boldsymbol{\xi}_i\boldsymbol{\xi}_i^\top,$$

$$\sum_i w_i \boldsymbol{\xi}_i\boldsymbol{\xi}_i^\top \preceq CR_\theta \left[T^{-1}\mathbf{\Xi}\mathbf{\Xi}^\top + s\sum_{j \le J_0} \boldsymbol{\xi}_j\boldsymbol{\xi}_j^\top \right].$$

Hölder and (57), without independence between weights and noises, give

$$\| \|\mathbf{C}\| \|_{L^r} \le C(\sigma^2 + P^2 s), \qquad \mathbb{E}[K_{\mathrm{tok}} \|\mathbf{C}\|] \le CP(\sigma^2 + P^2 s). \tag{58}$$

Here $K_{\mathrm{tok}} \le 2\max_i \|\mathbf{x}_i\|_2$ has L^r norm at most CP by the Gaussian token tails and Assumption 3.1.

The position tilt is controlled globally. For distractors I, let $D_I(u) = \sum_{i\in I} \zeta_i e^{u\zeta_i} / \sum_{i\in I} e^{u\zeta_i}$. Its derivative is a weighted variance and is bounded by $e^{2uM}|I|^{-1}\sum_{i\in I}\zeta_i^2$. Thus $\|D_I(\theta)\|_{L^2} \le C(\sigma/\sqrt{T} + \sigma^2\theta)$. Writing the full pooled position as relevant signal, relevant noise, and $(1 - \sum_{j\le J_0} w_j)D_I(\theta)$ proves

$$\|H\|_{L^2} \le C(Ps + \nu_0 + \sigma^2\gamma/P). \tag{59}$$

E.6 Operator moments, parity moments, and their integrals

Conditional on positions, semantic signs are centered and the perpendicular noise is Gaussian with covariance $\sigma^2\omega_2\mathbf{\Pi}_{\mathbf{p}^\perp}$. After position averaging, the weight moment bounds give a signal second moment of operator norm at most Cs^2. Consequently (59) gives

$$\left\|\mathbb{E}\mathbf{h}\mathbf{h}^\top\right\| \le C(P^2s^2 + \sigma^2/T + \sigma^4\gamma^2/P^2), \tag{60}$$

$$V_{\mathbf{h}} := \mathbb{E}\left\|\mathbf{h}\right\|_2^2 \le C(P^2s^2 + d\sigma^2/T + \sigma^4\gamma^2/P^2). \tag{61}$$

The signal and position terms have fixed rank; multiplying the first line by d would give an unnecessarily large trace bound. The original common-noise decomposition is $\mathbf{h} = \mathbf{c} + \mathbf{g}_0$ with independent $\mathbf{g}_0 \sim \mathcal{N}(\mathbf{0}, \nu_0^2\mathbf{I}_d)$ for auxiliary data and $\mathbf{g}_0 \sim \mathcal{N}(0, \nu_0^2\mathbf{\Pi}_{\mathbf{p}^\perp})$ for XOR. Hence $\mathbb{E}\mathbf{c}\mathbf{c}^\top \preceq \mathbb{E}\mathbf{h}\mathbf{h}^\top$, and (60) also bounds the mean term in the gate curvature. Distractor XOR position noise is retained.

For XOR the exact conditional pool is

$$\mathbf{h} = H\mathbf{e}_p + a(z_1\boldsymbol{\mu}_1 + z_2\boldsymbol{\mu}_2) + \nu\mathbf{z}_\perp, \quad a = w_1 = w_2, \quad \nu = \sigma\sqrt{\omega}_2 \ge \nu_0, \quad \beta = H/\nu.$$

The weight moment bound and an event of probability tending to one on which the weights are within fixed factors of their noiseless values give

$$g_{\mathrm{X}} := \mathbb{E}_{\mathrm{pos}}\frac{a^2}{\nu} \asymp \frac{s^2}{\nu_0}. \tag{62}$$

For the upper bound simply use $\nu \ge \nu_0$; for the lower bound use $a \ge c_0 s$ and $\omega_2 \le C/T$ on that event. All tails in the weighted upper bound are retained by Cauchy–Schwarz:

$$\begin{aligned}\mathbb{E}\left[\frac{a^2}{\nu}(1+|\beta|)\right] &\le \frac{\mathbb{E}a^2}{\nu_0} + \frac{(\mathbb{E}a^4)^{1/2}(\mathbb{E}H^2)^{1/2}}{\nu_0^2} \\ &\le C\left[1 + \frac{Ps}{\nu_0} + \alpha\gamma\right] g_{\mathrm{X}}.\end{aligned} \tag{63}$$

The growing position tilt $\alpha\gamma$ cannot be omitted near initialization. It is polynomial logarithmic and its curvature contribution is integrated explicitly below.

Higher position moments. The same argument with Hölder exponents enlarged gives, for $1 \le r \le 64$,

$$\|H\|_{L^r} \le C(Ps + \nu_0 + \sigma^2|\gamma|/P), \qquad \|a\|_{L^r} \le Cs.$$

Consequently, writing $k = a/\nu$ and $\beta = H/\nu$, for $1 \le r \le 64$,

$$\mathbb{E}_{\mathrm{pos}}\left[\frac{a^2}{\nu}(1+k+|\beta|)^r\right] \le \mathrm{polylog}(d)\frac{s^2}{\nu_0}, \qquad \mathbb{E}_{\mathrm{pos}}\frac{\nu^2}{\nu_0^2}(1+k^2+\beta^2) \le \mathrm{polylog}(d). \tag{64}$$

Indeed $\nu \ge \nu_0$ bounds every denominator; use the displayed high moments and, under Assumption 3.1, $Ps/\nu_0 + \sigma^2|\gamma|/(P\nu_0) = \mathrm{polylog}(d)$. These bounds also hold for $-1 \le \gamma \le 0$ with $s \asymp 1/T$ by the same exponential moment argument.

Integrating on the already constructed interval. Equations (43), (61), and (58) imply

$$\int_0^{R_*} V_{\mathbf{h}}\,dt \le C\left[P\nu_0 K_{\mathrm{X}}/\sqrt{G} + \frac{d\sigma^2}{P} + \frac{\sigma^4 T}{P^3}\right] \le C/P, \tag{65}$$

$$\int_0^{R_*} \mathbb{E}\omega_2\,dt \le C/P,$$

$$I_C := \int_0^{R_*} \mathbb{E}\left\|\mathbf{C}\right\|^2 dt \le C(\sigma^4 T/P + P^3\nu_0 K_{\mathrm{X}}/\sqrt{G}), \tag{66}$$

$$I_C/\nu_0 \le C(P^2\alpha^3 + P^3 K_{\mathrm{X}}/\sqrt{G}).$$

Here Assumption 3.1 gives $P\nu_0 K_{\mathrm{X}}/\sqrt{G} = c_*\nu_0 P^2 = o(P^{-1})$, $d\sigma^2 = O(1)$, and $\sigma^4 T/P^3 = o(P^{-1})$. The two $1/P$ estimates concern different quantities, ω_2 and $V_{\mathbf{h}}$; the dimension factor in (61) has been retained.

E.7 Directional curvature on the same stopped interval

Lemma E.4 (Directional attention covariance moments)**.** *On the axial interval, for either source and moment orders at most* 64*,*

$$\left\| \|\mathbf{C}\mathbf{e}_p\|_2 \right\|_{L^2} \le C(P^2 s + \sigma^2), \qquad \| \|\mathbf{C}\mathbf{e}_s\|_2 \|_{L^2} \le C(Ps + \sigma^2), \qquad \| \|\mathbf{C}\| \|_{L^2} \le C(P^2 s + \sigma^2).$$

The same bounds hold for each fixed unit semantic direction.

Proof. Expand $\mathbf{C}v = \sum_i w_i \mathbf{x}_i(\mathbf{x}_i^T v) - \mathbf{h}(\mathbf{h}^T v)$ and $\mathbf{x}_i = \mathbf{d}_i + \boldsymbol{\xi}_i$. The deterministic relevant summands have norm at most $CP^2 sR_\theta$ for $v = \mathbf{e}_p$, and $CPsR_\theta$ for a semantic v, because $|\mathbf{d}_i^T \mathbf{e}_s| \le 1$. The mixed deterministic/noise terms have these same upper orders: there are at most two relevant columns, their noise norms have bounded fixed moments, their scalar projections have moments $O(\sigma)$, and $\|R_\theta\|_{L^r} \le C$. The pure noise matrix has fixed operator moments $O(\sigma^2 + s)$ by the weighted Gaussian matrix argument; $s \le Ps$ absorbs its relevant-column term. For the subtracted pooled product, the conditional Gaussian representation gives $\|\mathbf{h}^T \mathbf{e}_s\|_{L^4} \le C(s + \nu_0)$ and $\| \|\mathbf{h}\|_2 \|_{L^4} \le C(Ps + \sqrt{d}\,\nu_0 + \sigma^2\gamma/P)$. Their product is at most $C(Ps + \sigma^2)$ under Assumption 3.1 and the stated range $K = O(P\sqrt{G})$, since $s = \nu_0 \operatorname{polylog}(d)$. The position case is already bounded by the operator-moment estimate. All uses of products apply Hölder with larger fixed moments, so they retain the complete Gaussian tails. □

For the shell, its exact covariance is diagonal in the relevant decomposition:

$$\mathbb{E}\mathbf{u}\mathbf{u}^T = c(t)^2 \mathbf{e}_p \mathbf{e}_p^T + b(t)^2 \mathbf{e}_s \mathbf{e}_s^T + \frac{r(t)^2}{d-3} \mathbf{\Pi}_{\mathcal{N}}.$$

The norm denominator in $\mathbf{A}_{\mathbf{U}}$ is bounded above and below by constants. Hence the directional lemma gives

$$Q_{\mathrm{av}} \le \frac{C}{\nu_0} \left[c(t)^2 (P^4 s^2 + \sigma^4) + b(t)^2 (P^2 s^2 + \sigma^4) + d^{-1} (P^4 s^2 + \sigma^4) \right].$$

Use $c(t)^2 \le C/P^2$, $b(t)^2 \le C\gamma/P^2$, the sharp second clock, $\int \gamma dt \le CT/P$, and Assumption 3.1. At an endpoint $\Gamma \asymp G$ this gives

$$\int Q_{\mathrm{av}} dt \le C \left[\frac{PK}{\sqrt{\Gamma}} + \frac{K\sqrt{\Gamma}}{P} + \alpha^3 \right] + d^{-1} \operatorname{polylog}(d).$$

The static initial XOR law contributes only the displayed d^{-1} remainder. The other five terms in the joint Hessian envelope (32) satisfy

$$\int H_g dt \le C \left[PK/\sqrt{\Gamma} + \alpha + \alpha^3 \right],$$

$$\int C_1 dt \le C(P\sqrt{\Gamma} + P\alpha^2),$$

$$\int B_{\mathbf{U}} \mathbb{E}(K_{\mathrm{tok}} \|\mathbf{C}\|) dt \le C(P\Gamma + P\alpha^2) + d^{-1/2} \operatorname{polylog}(d),$$

$$\int V dt \le C/P, \qquad \int B_{\mathbf{U}}^2 C_2 dt = o(1).$$

For the fourth line use $B_{\mathbf{U}} \le C(b + d^{-1/2})$, $\int bs dt \le Cb(R)^2$, and $\int b dt \le CT/P^2$; the latter follows from $b \le C\sqrt{\gamma}/P$ and $dt \le CTe^{-\gamma} d\gamma/P$. Therefore both the Hilbert curvature budget Λ and the row's own coefficient budget $\mathcal{A}$ obey

$$\Lambda + \mathcal{A} \le C \frac{PK}{\sqrt{\Gamma}} + C \left[P\Gamma + P\alpha^2 + \alpha^3 + \alpha + K\sqrt{\Gamma}/P \right] + o(1). \tag{67}$$

The row-to-attention-weight coefficient $\mathcal{C}$ remains a fixed polynomial logarithm; it is not exponentiated.

E.8 Curvature on complete secants

Lemma E.5 (Curvature under a small state perturbation). *Let $0 < a_0 \le R_0$ and $\nu_0 > 0$. Let $(\mathbf{q}^\circ, \mathbf{U}^\circ)$ be a reference state for one source, independent of a fresh example. Suppose throughout the entire attention-weight neighborhood the same decomposition $\mathbf{h}(\mathbf{X}, \mathbf{q}) = \mathbf{c}(\mathbf{X}, \mathbf{q}) + \mathbf{g}_0$ holds, where $\mathbf{g}_0 \sim \mathcal{N}(\mathbf{0}, \nu_0^2 \mathbf{\Pi})$ is independent of $(\mathbf{c}, D_{\mathbf{q}}\mathbf{h})$ and $\mathbf{\Pi}$ is an orthogonal projector. Assume $a_0 \le \|\mathbf{\Pi}\mathbf{u}_j^\circ\|_2 \le \|\mathbf{u}_j^\circ\|_2 \le R_0$. Set $M = \max_i \|\mathbf{x}_i\|_2$, $\mathbf{C}^\circ = \mathbf{C}(\mathbf{X}, \mathbf{q}^\circ)$, $\mathbf{\Sigma}^\circ = m^{-1} \sum_j \mathbf{u}_j^\circ (\mathbf{u}_j^\circ)^\top$, and $B_0^2 = \|\mathbf{\Sigma}^\circ\|$. For fixed $\kappa \ge 1$, $P \ge 1$, $r_* > 0$, and $D_{\mathrm{cov}} \ge 0$, suppose*

$$\mathbb{E}[e^{4r_* M} \|\mathbf{C}^\circ\|^2] \le \kappa D_{\mathrm{cov}}^2,$$
$$\mathbb{E}[M^2 e^{4r_* M} \|\mathbf{C}^\circ\|^2] \le \kappa P^2 D_{\mathrm{cov}}^2. \tag{68}$$

Let $M_A \ge \|\mathbb{E}\mathbf{c}(\mathbf{X}, \mathbf{q}^\circ)\mathbf{c}(\mathbf{X}, \mathbf{q}^\circ)^\top\|$, $V_0 \ge \mathbb{E}\|\mathbf{h}(\mathbf{X}, \mathbf{q}^\circ)\|_2^2$, and define

$$H_0 = (\nu_0^2 + M_A)/\nu_0, \qquad Q_0 = \|\mathbb{E}[\mathbf{C}^\circ \mathbf{\Sigma}^\circ \mathbf{C}^\circ]\|/\nu_0.$$

Put

$$J_* = (1 + a_0^{-1}) \left(1 + \frac{R_0 + 1}{a_0}\right)^2, \qquad D_* = (1 + \kappa) J_*.$$

Throughout

$$\|\mathbf{q} - \mathbf{q}^\circ\|_2 \le r, \quad \|\mathbf{U} - \mathbf{U}^\circ\|_{m,2} \le r, \quad \max_j \|\mathbf{u}_j - \mathbf{u}_j^\circ\|_2 \le r, \qquad r \le \min\{r_*, 1, a_0/2\},$$

the Hilbert drift coefficient L and the row's own coefficient a_{row} in Lemma C.8 admit the bound

$$L + a_{\mathrm{row}} \le \mathfrak{L}_0 + r\mathfrak{E}_1 + r^2 \mathfrak{E}_2, \tag{69}$$

where

$$\mathfrak{L}_0 = C D_* [H_0 + Q_0 + D_{\mathrm{cov}} + B_0 P D_{\mathrm{cov}} + V_0 + B_0^2 D_{\mathrm{cov}}^2],$$
$$\mathfrak{E}_1 = C D_* P D_{\mathrm{cov}}, \qquad \mathfrak{E}_2 = C D_* D_{\mathrm{cov}}^2 [(1 + P^2 B_0^2)/\nu_0 + 1].$$

The same statement holds for coupled row laws with empirical averages replaced by expectations and maxima by essential suprema.

Proof. Write $\mathbf{v} = \mathbf{q} - \mathbf{q}^\circ$. Softmax likelihood ratios and the variational formula for a variance imply, for $0 \le t \le 1$,

$$e^{-2M\|\mathbf{v}\|_2} \le \frac{s_i(\mathbf{X}, \mathbf{q}^\circ + t\mathbf{v})}{s_i(\mathbf{X}, \mathbf{q}^\circ)} \le e^{2M\|\mathbf{v}\|_2}, \qquad 0 \preceq \mathbf{C}(\mathbf{X}, \mathbf{q}^\circ + t\mathbf{v}) \preceq e^{2Mr} \mathbf{C}^\circ.$$

The third central moment formula, for unit $\mathbf{u}, \mathbf{w}$, gives

$$\begin{aligned} |\mathbf{u}^\top D_{\mathbf{q}} \mathbf{C}[\mathbf{v}] \mathbf{w}| &= \left| \sum_i s_i (\mathbf{u}^\top (\mathbf{x}_i - \mathbf{h}))(\mathbf{w}^\top (\mathbf{x}_i - \mathbf{h}))(\mathbf{v}^\top (\mathbf{x}_i - \mathbf{h})) \right| \\ &\le 2Mr \sum_i s_i |\mathbf{u}^\top (\mathbf{x}_i - \mathbf{h})| \, |\mathbf{w}^\top (\mathbf{x}_i - \mathbf{h})| \\ &\le 2Mr \sqrt{\mathbf{u}^\top \mathbf{C} \mathbf{u}} \sqrt{\mathbf{w}^\top \mathbf{C} \mathbf{w}} \\ &\le 2Mr e^{2Mr} \|\mathbf{C}^\circ\|. \end{aligned}$$

Consequently $\|\Delta \mathbf{C}\| \le 2Mr e^{2Mr} \|\mathbf{C}^\circ\|$ and $\|\Delta \mathbf{h}\|_2 = \|\Delta \mathbf{c}\|_2 \le r e^{2Mr} \|\mathbf{C}^\circ\|$. Moreover, $\mathbf{\Sigma}_{\mathbf{U}} \preceq 2\mathbf{\Sigma}^\circ + 2r^2 \mathbf{I}_d$ and $B_{\mathbf{U}} \le B_0 + r$. Conditional Gaussian regression on the gate, with $\|\mathbf{\Pi}\mathbf{u}_j\|_2 \ge a_0/2$, therefore yields

$$\begin{aligned} H_g &\le \frac{C J_*}{\nu_0} (\nu_0^2 + \|\mathbb{E}\mathbf{c}(\mathbf{X}, \mathbf{q}) \mathbf{c}(\mathbf{X}, \mathbf{q})^\top\|) \\ &\le \frac{C J_*}{\nu_0} (\nu_0^2 + 2M_A + 2\mathbb{E}\|\Delta \mathbf{c}\|_2^2) \\ &\le \frac{C J_*}{\nu_0} (\nu_0^2 + 2M_A + 2\kappa r^2 D_{\mathrm{cov}}^2) \end{aligned}$$

$$\le CD_*(H_0 + r^2 D_{\mathrm{cov}}^2/\nu_0).$$

For the attention-weight gate coefficient, the matrix inequality $(A+B)\Sigma(A+B) \preceq 2A\Sigma A + 2B\Sigma B$ gives

$$\begin{aligned} Q_{\mathrm{av}} &\le \frac{C(1+a_0^{-1})}{\nu_0} \|\mathbb{E}[\mathbf{C}\boldsymbol{\Sigma}_{\mathbf{U}}\mathbf{C}]\| \\ &\le \frac{C(1+a_0^{-1})}{\nu_0} \left(4\|\mathbb{E}[\mathbf{C}^\circ \boldsymbol{\Sigma}^\circ \mathbf{C}^\circ]\| + 4B_0^2 \mathbb{E}\|\Delta\mathbf{C}\|^2 + 2r^2\mathbb{E}\|\mathbf{C}\|^2\right) \\ &\le CD_* Q_0 + \frac{CD_* r^2 (16P^2B_0^2 + 2) D_{\mathrm{cov}}^2}{\nu_0} \\ &\le CD_*[Q_0 + r^2(1+P^2B_0^2) D_{\mathrm{cov}}^2/\nu_0]. \end{aligned}$$

Finally, $V \le 2V_0 + 2\kappa r^2 D_{\mathrm{cov}}^2$, $C_1 \le \sqrt{\kappa} D_{\mathrm{cov}}$, $C_2 \le \kappa D_{\mathrm{cov}}^2$, and $\mathbb{E}[M\|\mathbf{C}\|] \le \sqrt{\kappa} P D_{\mathrm{cov}}$. The second variation and $|\ell'| \le 1$, $\ell'' \le 1/4$ give

$$L \le 2H_g + 2Q_{\mathrm{av}} + C_1 + 2B_{\mathbf{U}}\mathbb{E}[M\|\mathbf{C}\|] + V/2 + B_{\mathbf{U}}^2 C_2/2, \qquad a_{\mathrm{row}} = H_g.$$

Substitution and $(B_0 + r)^2 \le 2B_0^2 + 2r^2$ prove (69), with a fixed coefficient on $\mathfrak{L}_0$. □

Lemma E.6 (Quadrature of a stopped envelope)**.** *Let $F \ge 0$ have bounded variation on $[0, R]$, let $N\eta \le R$, and suppose $c_n \le F(n\eta)$ whenever $n < \tau$. Then*

$$\eta \sum_{n<N} \mathbf{1}_{n<\tau} c_n \le \int_0^R F(t)\,dt + \eta\,\mathrm{TV}_{[0,R]}(F).$$

Proof. Summing over disjoint grid cells gives

$$\begin{aligned} \eta \sum_{n<N} \mathbf{1}_{n<\tau} c_n &\le \eta \sum_{n<N} F(n\eta) \\ &\le \sum_{n<N} \left[\int_{n\eta}^{(n+1)\eta} F(t)\,dt + \eta\,\mathrm{TV}_{[n\eta,(n+1)\eta]}(F) \right] \\ &\le \int_0^{N\eta} F(t)\,dt + \eta\,\mathrm{TV}_{[0,N\eta]}(F) \\ &\le \int_0^R F(t)\,dt + \eta\,\mathrm{TV}_{[0,R]}(F). \end{aligned}$$

□

For the Gaussian token laws under Assumption 3.1, take $r_* = 2d^{-1/4}$ and $D_{\mathrm{cov}} = \sigma^2 + P^2 s$. Gaussian norm tails give $\mathbb{E}e^{8r_*M} \le C$ and $\mathbb{E}[M^4 e^{8r_*M}] \le CP^4$. Together with the axial fourth moment in (58), Cauchy–Schwarz verifies (68) with fixed κ. The shell clocks, $B_0^2 \le C(G/P^2 + d^{-1})$, and Assumption 3.1 imply

$$\begin{aligned} r\int \mathfrak{E}_1 + r^2 \int \mathfrak{E}_2 &\le CrP(\sigma^2 R + P^2 \int s) + Cr^2(1+G)(1+\nu_0^{-1}) \int D_{\mathrm{cov}}^2 \\ &\le CrP^2(\alpha^2 + \sqrt{G}) + Cr^2(1+G)(1+\nu_0^{-1})(\sigma^4 T/P + P^3\nu_0 K_X/\sqrt{G}) \\ &\le C[r(1+\log d)^2 + r^2(1+\log d)^3] \\ &= o(1), \qquad r \in \{d^{-1/4}, 2d^{-2/5}\}. \end{aligned}$$

Choose deterministic envelopes

$$\begin{aligned} M_A &= C(P^2 s^2 + \nu_0^2 + \sigma^4\gamma^2/P^2), \\ V_0 &= C(P^2 s^2 + d\nu_0^2 + \sigma^4\gamma^2/P^2), \\ Q_0 &= \frac{C}{\nu_0}\left[c^2(P^4 s^2 + \sigma^4) + b^2(P^2 s^2 + \sigma^4) + \frac{r_{\mathrm{sh}}^2}{d-3}(P^4 s^2 + \sigma^4)\right], \\ B_0 &= C(c^2 + b^2 + r_{\mathrm{sh}}^2/(d-3))^{1/2}, \end{aligned}$$

where r_{sh} is the shell radius r. They also cover the static initial XOR law and sampled shell covariance. The coarse bounds $|\dot\gamma| \le CP^3$ and $|\dot b| + |\dot c| + |\dot r_{\mathrm{sh}}| \le CP$, together with Assumption 3.1 and the learning-rate choice in (15), give

$$\begin{aligned}\eta \sum_{F\in\{\mathfrak{L}_0,\mathfrak{E}_1,\mathfrak{E}_2\}} \mathrm{TV}(F) &\le C\eta R(1+\nu_0^{-1})(1+P+G)^{12}\\ &\le Cd^{-3}\, d(\log d)^{3/4}\, d(\log d)^{9/16}(\log d)^6\\ &\le Cd^{-1}(1+\log d)^8\\ &= o(1).\end{aligned}$$

Here each envelope is a polynomial in the shell coordinates and s, apart from B_0, whose variation is bounded by that of its Euclidean coordinate vector. Thus no derivative of an actual sample gate is used. Lemma E.6 transfers (67) to the grid, adding only $o(1)$ on every complete stopped secant. In the SGD comparison, the already proved population-to-shell error is $d^{-7/16}$; hence the current $d^{-2/5}$ tube lies in the required $2d^{-2/5}$ shell tube.

Fixing the checkpoint constant. For a comparison radius $r_{\mathrm{tube}} \in \{d^{-1/4}, 2d^{-2/5}\}$, let

$$\mathcal{L}_{r_{\mathrm{tube}}}(t) := \mathfrak{L}_0(t) + r_{\mathrm{tube}}\mathfrak{E}_1(t) + r_{\mathrm{tube}}^2\mathfrak{E}_2(t)$$

be the deterministic envelope in (69), with the shell moment bounds chosen above.

Lemma E.7 (Cumulative curvature)**.** *Under Assumption 3.1 and (15), choose the universal checkpoint constant $c_* > 0$ sufficiently small. By Lemma A.3, both comparison radii satisfy*

$$\int_0^{R_*} \mathcal{L}_{r_{\mathrm{tube}}}(t)\,dt \le \frac{\log d}{128}, \qquad \eta\sum_{n<N}\mathcal{L}_{r_{\mathrm{tube}}}(n\eta) \le \frac{\log d}{128} \quad (N\eta \le R_*). \tag{70}$$

For references satisfying these moment bounds, the same estimates apply to the sums of joint-drift and individual-row self coefficients on any stopped comparison whose complete segments lie in these neighborhoods.

Proof. Assumption 3.1 gives

$$(\log d)^{3/8} \le P \le C(\log d)^{1/2}, \qquad \alpha = \frac{\sigma\sqrt{T}}{P} \le C(\log d)^{3/16}.$$

For $K = eK_{\mathrm{X}}$ and $\Gamma = \Gamma_{K_{\mathrm{X}}} + 1 \asymp G$, the remainder in (67) satisfies

$$\begin{aligned}&P\Gamma + P\alpha^2 + \alpha^3 + \alpha + K\sqrt{\Gamma}/P\\ &\le C\left[PG + \frac{\sigma^2 T}{P} + \frac{(\sigma\sqrt{T})^3}{P^3} + \frac{\sigma\sqrt{T}}{P} + G\right]\\ &\le C\left[(\log d)^{1/2}G + (\log d)^{3/4} + (\log d)^{9/16} + (\log d)^{3/16} + G\right]\\ &= o(\log d).\end{aligned}$$

The leading term is at most $C_0c_*\log d$, including the score buffer and the secant-enlargement constants. Choose $c_* > 0$ so small that $C_0c_* < 1/256$. Assumption 3.1 and Lemma A.3 then give

$$\Lambda + \mathcal{A} < \frac{\log d}{128}, \qquad e^{\Lambda+\mathcal{A}} \le d^{1/128}, \qquad \mathcal{C} \le \mathrm{polylog}(d). \tag{71}$$

Here Λ is the Hilbert curvature integral and $\mathcal{A}$ the individual row's own coefficient integral in Lemma C.4; $\mathcal{C}$ is the cross coefficient and is not exponentiated. The neighborhood corrections and the quadrature error are $o(1)$ by the preceding estimates. Absorbing both in the remaining $(\log d)/256$ proves (70); stopping only removes nonnegative terms. □

This is the formal version of Lemma 5.1. For that application, Proposition D.1 establishes $\eta\tau_X < R_*$ after closing the continuation argument. The proof uses this quantitative bound at the final checkpoint and does not require $\Lambda = o(\log d)$ there. Decrease c_* further, if necessary, so that the feature clock satisfies $A \le 1$ and $\exp(C_{\mathrm{head}}A) \le 2$ for the fixed head-recursion constant C_{head}.

Transfer of the clocks to the joint reference. The population continuation in Section F compares the joint flow with the auxiliary shell and a static initial XOR marginal. Its conclusion is $\sup_{t\le R_*}\|\bar{\mathbf{q}}-\mathbf{q}_{\mathrm{sh}}\|_2 \le d^{-7/16}$. The full population attention weight is axial by its marginal symmetries. Put $\epsilon = Pd^{-7/16} = o(1)$ and let $t_{\mathrm{sh}}(a)$ denote a shell score hitting time. The joint first hit $\bar{R}_K$ obeys

$$t_{\mathrm{sh}}(\Gamma_K-\epsilon) \le \bar{R}_K \le t_{\mathrm{sh}}(\Gamma_K+\epsilon), \qquad e^{-\epsilon}s(\gamma) \le s(\bar{\gamma}) \le e^{\epsilon}s(\gamma).$$

The score buffer contains these times. Upper clocks transfer from the shell; the joint prefix contains the shell strip $[\Gamma_K-1, \Gamma_K-2\epsilon]$, whose occupation and (42) give the same lower clock up to constants. This does not assume monotonicity of $\bar{\gamma}$. The covariance secant estimates above and $D_{\mathbf{q}}\mathbf{h} = \mathbf{C}$ transfer the moment and curvature budgets as well. These are conditional implications from the comparison estimate; the shell inputs used to prove that estimate were established independently on the stopped region.

F Joint population and SGD continuation

F.1 A single predictable stop

In this section N denotes a deterministic grid horizon. Use the empirical version of $\mathcal{H}$ from Section C. Write $F_k = F_{\mathbf{U}_k}$ and $\boldsymbol{\beta}^m$ for the finite-network population SGD drift in this metric.

Set $\chi_0 = C_\chi\sqrt{\ell_\delta/d}$, where C_χ is a sufficiently large universal constant, and define the compact row support

$$\mathcal{K}_0 = \{\mathbf{u}_0 : \|\boldsymbol{\Pi}_{\mathcal{N}^\perp}\mathbf{u}_0\|_2 \le \chi_0,\ |\|\boldsymbol{\Pi}_{\mathcal{N}}\mathbf{u}_0\|_2 - \rho| \le \chi_0\}.$$

For each head, take $\Omega_k = \{-1,1\}\times\mathcal{K}_0$ with the Rademacher–Gaussian marginal law conditioned on $\mathbf{u}_0 \in \mathcal{K}_0$. These laws have full support on Ω_k; essential suprema of continuous characteristic differences therefore equal their suprema. The initial map $(\epsilon,\mathbf{u}_0)\mapsto\mathbf{u}_0$ is continuous and satisfies

$$\min_{k,\omega\in\Omega_k}\|\boldsymbol{\Pi}_k\mathbf{u}_{k,0}(\omega)\|_2 \ge \rho-\chi_0 > \rho/2, \qquad \max_{k,\omega\in\Omega_k}\|\mathbf{u}_{k,0}(\omega)\|_2 \le \rho+2\chi_0.$$

Here $\boldsymbol{\Pi}_A = \mathbf{I}_d$ and $\boldsymbol{\Pi}_X = \boldsymbol{\Pi}_\perp$. Both data laws admit Lemma C.5: use all noise coordinates for A and perpendicular noise coordinates for X. Lemma C.9 therefore supplies a unique local solution with $\bar{\mathbf{q}}(0)=\mathbf{0}$. Write $F_{k,\infty}$ and $\mathbf{b}_{k,\infty}$ for its marginal output and input derivative from (34). Explicitly,

$$\begin{aligned}\dot{\bar{\mathbf{u}}}_k(t;\omega) &= \pi_k\epsilon\mathbb{E}_{\mathcal{D}_k}[a_y(F_{k,\infty})\mathbf{h}(\mathbf{X},\bar{\mathbf{q}})\mathbf{1}_{\bar{\mathbf{u}}_k(t;\omega)^\top\mathbf{h}>0}],\\ \dot{\bar{\mathbf{q}}}(t) &= \sum_k \pi_k\mathbb{E}_{\mathcal{D}_k}[a_y(F_{k,\infty})\mathbf{C}(\mathbf{X},\bar{\mathbf{q}})\mathbf{b}_{k,\infty}].\end{aligned} \tag{72}$$

The two marginal laws and shared attention weight form a closed system; no joint moment of the two head rows occurs. Its law and attention weight are deterministic. The continuation proof below bounds the state and keeps the projected row norms positive, precisely the extension conditions in Lemma C.9. Localize first to any deterministic interval $[0,R]$ on which

$$\inf_{k,t\le R,\omega\in\Omega_k}\|\boldsymbol{\Pi}_k\bar{\mathbf{u}}_k(t;\omega)\|_2 \ge a_0, \qquad \sup_{k,t\le R,\omega\in\Omega_k}\|\bar{\mathbf{u}}_k(t;\omega)\|_2 \le R_0, \tag{73}$$

for fixed bounds holding strictly at initialization. Every grid estimate below uses $N\eta \le R$. Section F.8 closes this localization on $[0,R_*]$; Section F.9 then takes $N=\lfloor R_*/\eta\rfloor$.

Under Assumption 3.1, Lemma A.3 gives $\chi_0 \le 1/8$ and $m \le d^2$. Gaussian coordinate and radial tails, followed by a union over $2m$ rows, make the probability that some row leaves $\mathcal{K}_0$ at most $Cme^{-cd\chi_0^2} \le \delta_J/10$. The one-row cutoff probability is at least $1/2$. On this product event, evaluate the two marginal characteristics at each actual seed pair $\omega_j = (\epsilon_j,\mathbf{u}_{A,j,0},\mathbf{u}_{X,j,0})$; the notation $\bar{\mathbf{u}}_k(t;\omega_j)$ uses only $(\epsilon_j,\mathbf{u}_{k,j,0})$. The sampled pairs are independent across j and share ϵ_j within a pair. This constructs the coupled reference without changing the SGD law. Conditional estimates return to the original law by adding the probability of the product-event complement; unconditioned Gaussian-feature events are intersected separately. Write $\boldsymbol{\theta}_n = (\mathbf{q}_n,(\mathbf{u}_{k,j,n})_{k,j})$ for the joint state, and let $\bar{\boldsymbol{\theta}}_n = (\bar{\mathbf{q}}(n\eta),\bar{\mathbf{u}}_k(n\eta;\omega_j))$ be deterministic population characteristics evaluated at the actual initial seeds. Their law is deterministic; $\bar{\boldsymbol{\theta}}_n$ is a random sample from it. Define

$$\mathbf{x}_n = \boldsymbol{\theta}_n - \bar{\boldsymbol{\theta}}_n, \quad e_n = \|\mathbf{x}_n\|_{\mathcal{H}}, \quad z_n = \max_{k,j}\|\mathbf{x}_{k,j,n}\|_2, \quad r_d = d^{-2/5},$$

$$\tau = \inf\{0 \le n \le N : e_n \ge r_d \text{ or } z_n \ge r_d\}, \qquad \inf \varnothing = N + 1.$$

The filtration $\mathcal{F}_n$ contains initialization and batches before n. Thus $\mathbf{1}_{\{n<\tau\}}$ is measurable before batch n is drawn.

F.2 The error recursion

Put $\mathbf{f}_n = \boldsymbol{\beta}^m(\bar{\boldsymbol{\theta}}_n) - \dot{\bar{\boldsymbol{\theta}}}(n\eta)$, $\mathbf{r}_n = \bar{\boldsymbol{\theta}}_n + \eta\dot{\bar{\boldsymbol{\theta}}}(n\eta) - \bar{\boldsymbol{\theta}}_{n+1}$, and

$$\mathbf{M}_n = \eta \sum_{l<n} \mathbf{1}_{l<\tau}(\widehat{\boldsymbol{\beta}}_l(\boldsymbol{\theta}_l) - \boldsymbol{\beta}^m(\boldsymbol{\theta}_l)).$$

The exact recursion is

$$\mathbf{x}_{n\wedge\tau} = \underbrace{\mathbf{x}_0 + \mathbf{M}_n + \sum_{l<n} \mathbf{1}_{l<\tau}(\mathbf{r}_l + \eta\mathbf{f}_l)}_{\mathbf{F}_n} + \eta \sum_{l<n} \mathbf{1}_{l<\tau}[\boldsymbol{\beta}^m(\boldsymbol{\theta}_l) - \boldsymbol{\beta}^m(\bar{\boldsymbol{\theta}}_l)]. \tag{74}$$

Lemma C.4 applies once the following derivative bounds and forcing budgets are established. Its constants D_H, D_∞ can be bounded by the initial error, maximal martingale partial sums, and accumulated norms of $\mathbf{r}_l + \eta\mathbf{f}_l$. The row-to-attention-weight coefficient enters through $\mathcal{C}E_H$.

F.3 Common noise on the stopped state tube

Apply Lemma C.5 with the two projectors fixed above. For XOR, $\boldsymbol{\eta}_i$ is the perpendicular component of every noise token; the independent position noise of each distractor remains in $\mathbf{v}_i$. Thus the common-noise decomposition is valid off axis for both sources and preserves the original distractor distribution.

Under (73) and before τ, every reference-to-iterate segment satisfies

$$\begin{aligned}\|\boldsymbol{\Pi}_k[(1-t)\bar{\mathbf{u}}_{k,j} + t\mathbf{u}_{k,j}]\|_2 &\ge \|\boldsymbol{\Pi}_k\bar{\mathbf{u}}_{k,j}\|_2 - t\|\boldsymbol{\Pi}_k(\mathbf{u}_{k,j} - \bar{\mathbf{u}}_{k,j})\|_2 \\ &\ge a_0 - \|\mathbf{u}_{k,j} - \bar{\mathbf{u}}_{k,j}\|_2 \\ &\ge a_0 - r_d \\ &\ge a_0/2 \qquad (0 \le t \le 1).\end{aligned}$$

The same segment has row norm at most $R_0 + r_d \le R_0 + 1$. Lemmas C.6–C.8 therefore apply on complete segments. The averaged matrix $\mathbf{A}_\mathbf{U}$ in (31) gives the Hilbert coefficient (32); the unaveraged Q_j gives (33). Lemma E.5 retains these distinct coefficients throughout this tube, so the row-to-attention-weight term enters only through $\mathcal{C}E_H$ in (28).

F.4 Finite width: concentrate accumulated forcing

Fix a deterministic reference time. Reference seed pairs are independent across j; dependence between the two heads through ϵ_j is allowed. Assume their rows have norm at most R_0. For one source write

$$F_m = \frac{1}{m}\sum_j \epsilon_j[\bar{\mathbf{u}}_j^\top \mathbf{h}]_+, \quad F_\infty = \mathbb{E}_\omega \epsilon[\bar{\mathbf{u}}(\omega)^\top \mathbf{h}]_+, \quad \Delta = \|F_m - F_\infty\|_{L^2(D)}.$$

Independence across seed pairs and the variance identity give

$$\mathbb{E}_\omega \Delta^2 \le R_0^2 V/m. \tag{75}$$

The reference and empirical-reference row gates coincide. Thus, for all rows,

$$\|\mathbf{f}_j\|_2 \le \tfrac{1}{4}\sqrt{V}\,\Delta, \qquad \mathbb{E}_\omega \max_j \|\mathbf{f}_j\|_2 \le CR_0V/\sqrt{m}. \tag{76}$$

This uses no independence between a chosen row and Δ.

For the attention weight define $\mathbf{b}_m(\mathbf{h}) = m^{-1}\sum_j \epsilon_j \bar{\mathbf{u}}_j g_j$ and $\mathbf{b}_\infty = \mathbb{E}_\omega \epsilon \bar{\mathbf{u}} g$. Split its forcing into

$$\begin{aligned}\mathbf{f}_q =& \mathbb{E}_D[a_y(F_\infty)\mathbf{C}(\mathbf{b}_m - \mathbf{b}_\infty)] \\ &+ \mathbb{E}_D[(a_y(F_m) - a_y(F_\infty))\mathbf{C}\mathbf{b}_m].\end{aligned}$$

The first term is a centered average of independent vectors

$$\mathbf{k}(\omega) = \mathbb{E}_D[a_y(F_\infty)\mathbf{C}\epsilon\bar{\mathbf{u}}(\omega)g(\omega)], \qquad \|\mathbf{k}(\omega)\|_2 \le R_0 C_1.$$

It has expected norm at most $R_0C_1/\sqrt{m}$. Since $\|\mathbf{b}_m(\mathbf{h})\|_2 \le R_0$, the second term has norm at most $R_0\sqrt{C_2}\Delta/4$. Therefore

$$\mathbb{E}_\omega \|\mathbf{f}_q\|_2 \le \frac{C}{\sqrt{m}}\left(R_0C_1 + R_0^2\sqrt{VC_2}\right). \tag{77}$$

Lemma F.1 (Integrated reference forcing)**.** *For a deterministic reference grid satisfying (73), set*

$$\mathcal{V}_k = \eta\sum_{n<N} V_{k,n}, \qquad \mathcal{J}_k = \eta\sum_{n<N} C_{1,k,n},$$
$$\mathcal{Q}_k = \eta\sum_{n<N} C_{2,k,n}, \quad \mathcal{F}_N = \sum_k (\mathcal{V}_k + \mathcal{J}_k + \sqrt{\mathcal{V}_k\mathcal{Q}_k}).$$

For $0 < \delta_e < 1$, with conditional initialization probability at least $1-\delta_e$ under the product cutoff law,

$$\eta\sum_{n<N}\left(\|\mathbf{f}_n\|_{\mathcal{H}} + \max_{k,j}\|\mathbf{f}_{k,j,n}\|_2\right) \le \frac{C\mathcal{F}_N}{\sqrt{m}}\left(1+\sqrt{\log(1/\delta_e)}\right). \tag{78}$$

The same bound holds after multiplying each summand by a stopping indicator.

Proof. At each time and source put

$$Z_{k,n} = \left\|\frac{1}{m}\sum_j \mathbf{k}_{k,n}(\omega_j) - \mathbb{E}_\omega\mathbf{k}_{k,n}(\omega)\right\|_2.$$

The preceding row bound and attention-weight split give the nonnegative envelope

$$\|\mathbf{f}_n\|_{\mathcal{H}} + \max_{k,j}\|\mathbf{f}_{k,j,n}\|_2 \le C\sum_k\left[Z_{k,n} + (\sqrt{V_{k,n}} + \sqrt{C_{2,k,n}})\Delta_{k,n}\right]. \tag{79}$$

The constant absorbs the fixed row bound R_0. Let Z be η times the sum of the right side over the deterministic grid. Equations (75)–(77) and Cauchy–Schwarz in time imply $\mathbb{E}Z \le C\mathcal{F}_N/\sqrt{m}$. Replacing one entire seed tuple changes $\Delta_{k,n}$ by at most $2R_0\sqrt{V_{k,n}}/m$ and $Z_{k,n}$ by at most $2R_0C_{1,k,n}/m$. Hence it changes Z by at most $C\mathcal{F}_N/m$.

For completeness, reveal the independent tuples successively. The Doob martingale of Z has conditional increment ranges of these lengths; the bounded-range exponential-moment inequality gives

$$\mathbb{E}e^{t(Z-\mathbb{E}Z)} \le \exp\{Ct^2\mathcal{F}_N^2/m\}, \qquad \mathbb{P}\{Z-\mathbb{E}Z > u\} \le \exp\{-cmu^2/\mathcal{F}_N^2\}.$$

Choosing $u = C\mathcal{F}_N\sqrt{\log(1/\delta_e)/m}$ proves the claim. The envelope, rather than a selected row's changing gate, is the bounded-difference functional. Dependence between heads and between grid times is allowed; only the seed tuples are independent. Stopping decreases its nonnegative sum. □

The event is initialization-measurable. Its reference law is exactly the conditioned marginal seed law, so no empirical law or future trajectory success enters the reference definition.

F.5 Fresh batches and Euler discretization

Write $\mathbf{g}_{k,j}(\boldsymbol{\theta};\mathbf{X},y) = \epsilon_j a_y(f_k(\mathbf{X}))\mathbf{h}(\mathbf{X},\mathbf{q})\mathbf{1}_{\mathbf{u}_{k,j}^\top\mathbf{h}(\mathbf{X},\mathbf{q})>0}$ for the single-example scaled-row force. For a source-example drift $\mathbf{g}_k$, fixed batch composition gives exactly

$$\mathbb{E}[\left\|\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^m\right\|_{\mathcal{H}}^2 \mid \mathcal{F}_n] = \frac{1}{B}\sum_k \pi_k\mathbb{E}[\|\mathbf{g}_k - \mathbb{E}[\mathbf{g}_k \mid \mathcal{F}_n]\|_{\mathcal{H}}^2 \mid \mathcal{F}_n]. \tag{80}$$

Here $\mathcal{F}_n$ contains initialization and the batches preceding iteration n, and all example drifts are evaluated at the old state. On the stopped row-norm region, put $K = \max_i\|\mathbf{x}_i\|_2$. Convex pooling and the covariance formula imply

$$\|\mathbf{h}\|_2 \le K, \qquad \|\mathbf{C}\| \le K^2, \qquad \|\mathbf{g}_k\|_{\mathcal{H}}^2 \le C(K^2+K^4), \qquad \|\mathbf{g}_{k,j}\|_2 \le K.$$

For every integer $r \geq 2$, the Gaussian token laws and Assumption 3.1 give

$$\|K\|_{L^r} \leq P + 1 + C\sigma\big(\sqrt{d} + \sqrt{\log(2T)} + \sqrt{r}\big) \leq CP\sqrt{r},$$
$$\|\|\mathbf{g}_k\|_{\mathcal{H}}\|_{L^r} \leq CP^2 r, \qquad \|\|\mathbf{g}_{k,j}\|_2\|_{L^r} \leq CP\sqrt{r}.$$

These estimates follow from Gaussian norm tails, including the union over T tokens; they hold for all r, not just the fixed moment orders used in the axial covariance bounds.

Reveal the B examples in iteration n successively, keeping its state fixed. Since $\pi_k/b_k = 1/B$, the centered full-state micro-increment is

$$\boldsymbol{\xi}_{n,b} = \frac{\eta}{B}\mathbf{1}_{\{n<\tau\}}\left(\mathbf{g}_{k(b)}(\boldsymbol{\theta}_n; \mathbf{X}_{n,b}, y_{n,b}) - \mathbb{E}[\mathbf{g}_{k(b)}(\boldsymbol{\theta}_n; \mathbf{X}, y) \mid \mathcal{F}_n]\right).$$

The source $k(b)$ is fixed by the batch composition. Conditional independence of fresh examples and predictability of $\{n < \tau\}$ make these martingale differences for the refined filtration. Conditional centering costs a factor two, and $r! \geq (r/e)^r$ yields

$$\mathbb{E}[\|\boldsymbol{\xi}_{n,b}\|_{\mathcal{H}}^r \mid \text{preceding information}] \leq (C\eta P^2 r/B)^r \leq r!(C_1\eta P^2/B)^r.$$

With $R = N\eta$, Lemma C.3 therefore applies with $a = C_1\eta P^2/B$, $v_l = 2a^2$, and $\sum_l v_l \leq C\eta R P^4/B$. In particular, with probability at least $1 - \delta_b/2$,

$$M_H^* := \max_{n\leq N}\|\mathbf{M}_n\|_{\mathcal{H}} \leq CP^2\left[\sqrt{\frac{\eta R}{B}\log\frac{4}{\delta_b}} + \frac{\eta}{B}\log\frac{4}{\delta_b}\right]. \tag{81}$$

For a single row the same argument uses P in place of P^2; examples of the other source give zero row increments. A union over the $2m$ rows, with total failure at most $\delta_b/2$, gives

$$M_\infty^* := \max_{k,j,n\leq N}\|\mathbf{M}_{k,j,n}\|_2 \leq CP\left[\sqrt{\frac{\eta R}{B}\log\frac{8m}{\delta_b}} + \frac{\eta}{B}\log\frac{8m}{\delta_b}\right]. \tag{82}$$

The bounds hold conditional on every admissible initialization. The maximal inequality already covers all partial sums, and no independence between rows is used. For $\delta_b = \delta_J/10$, Assumption 3.1, Lemma A.3 and $m \leq d^2$ imply

$$M_H^* + M_\infty^* \leq CP^2\left[\sqrt{\eta R\ell_\delta/B} + \eta\ell_\delta/B\right]. \tag{83}$$

No training example is truncated or removed from the update.

Euler error. We verify the moment hypotheses of Lemma C.11. Each token noise has covariance at most $\sigma^2\mathbf{I}_d$, and its noiseless part has norm at most $P + 1$. Gaussian norm tails and a union bound give

$$\mathbb{P}\{M > P + 1 + \sigma\sqrt{d} + \sigma\sqrt{2(\log T + x)}\} \leq e^{-x} \qquad (x \geq 0).$$

Integrating this tail and using Assumption 3.1 yields

$$\begin{aligned}\|M\|_{L^4} &\leq P + 1 + \sigma\sqrt{d} + C\sigma(\sqrt{\log(2T)} + 1)\\ &\leq P + 1 + \sqrt{2} + C\sqrt{\log(2T)/d}\\ &\leq C(P + 1)\\ &\leq CP.\end{aligned}$$

Also, Assumption 3.1 gives $\mathbb{E}\|\mathbf{g}_0\|_2^2 \leq dv_0^2 \leq 2/T$. On the localized interval (73), Lemma C.11 supplies Hilbert and maximum-row velocity Lipschitz constants CP^6/ν_0 for the sampled reference. Apply Lemma C.12 in both norms to obtain

$$\begin{aligned}\sum_{n<N}(\|\mathbf{r}_n\|_{\mathcal{H}} + \max_{k,j}\|\mathbf{r}_{k,j,n}\|_2) &\leq C\eta R P^6/\nu_0\\ &\leq C\eta P^5 T^{3/2}/\sigma\\ &\leq Cd^{-1}L_d^{67/16}.\end{aligned} \tag{84}$$

The last inequality uses Assumption 3.1 and the learning-rate choice in (15); the horizon bound comes from Theorem E.1. The estimate uses population velocities, including an individual bound for each sampled row. It does not require samplewise ReLU gradients to be Lipschitz.

The initialization and stochastic events fit one stop. First intersect the initialization norm, covariance, and forcing events. Conditional on any admissible initialization, (81)–(82) hold for the stopped fresh-batch process. Intersect their success events and use (28) up to the putative exit. Each displayed stochastic module receives $\delta_J/10$, with any remaining allocation unused. There is no union over all visited states.

F.6 Parity, covariance, and an embedded population reference

The axial reduction follows from the marginal equivariance lemma. Besides $\mathbf{H}_{\mathcal{S}}$ from the shell proof, use $\mathbf{H}_{\mathcal{N}} = \mathbf{I}_d - 2\mathbf{\Pi}_{\mathcal{N}}$. The data laws and conditioned seed supports have the symmetries

$\mathbf{H}$	χ_A	χ_X	$\tau_k(\epsilon, \mathbf{u}_0)$
$\mathbf{H}_{\mathcal{S}}$	-1	1	$(\chi_k\epsilon, \mathbf{H}_{\mathcal{S}}\mathbf{u}_0)$
$\mathbf{H}_{\mathcal{N}}$	1	1	$(\epsilon, \mathbf{H}_{\mathcal{N}}\mathbf{u}_0)$

Indeed semantic reflection changes the auxiliary label and preserves $z_1 z_2$; nuisance reflection preserves both labels. Gaussian invariance and the definition of $\mathcal{K}_0$ make each τ_k measure preserving. At initialization, the action in Lemma C.10 satisfies

$$\begin{aligned}(\mathcal{T}\bar{\mathbf{u}}_{k,0})(\epsilon, \mathbf{u}_0) &= \mathbf{H}\bar{\mathbf{u}}_{k,0}(\chi_k\epsilon, \mathbf{H}^\top \mathbf{u}_0)\\ &= \mathbf{H}\mathbf{H}^\top \mathbf{u}_0\\ &= \mathbf{u}_0\\ &= \bar{\mathbf{u}}_{k,0}(\epsilon, \mathbf{u}_0), \qquad \mathbf{H}\bar{\mathbf{q}}(0) = \bar{\mathbf{q}}(0).\end{aligned}$$

Local uniqueness therefore gives $\mathbf{H}_{\mathcal{S}}\bar{\mathbf{q}} = \mathbf{H}_{\mathcal{N}}\bar{\mathbf{q}} = \bar{\mathbf{q}}$. Hence $\mathbf{\Pi}_{\mathcal{S}}\bar{\mathbf{q}} = \mathbf{\Pi}_{\mathcal{N}}\bar{\mathbf{q}} = \mathbf{0}$ and $\bar{\mathbf{q}} = \bar{\gamma}\mathbf{e}_p/P$ up to the continuation endpoint. The two τ_k act on separate marginal spaces; no common sign map on the paired finite array is used. At an axial attention weight, conditional on position noises, the XOR pool is exactly

$$\mathbf{h} = b\mathbf{e}_p + a(z_1\boldsymbol{\mu}_1 + z_2\boldsymbol{\mu}_2) + \nu\mathbf{z}_\perp, \quad y = z_1 z_2.$$

For $u_i = \boldsymbol{\mu}_i^\top \mathbf{u}$ $(i = 1, 2)$, $u_p = \mathbf{e}_p^\top \mathbf{u}$, and $r = \left\|\mathbf{\Pi}_{\mathbf{p}^\perp}\mathbf{u}\right\|_2 \geq a_0$, $\chi_s = \left\|\mathbf{\Pi}_{\mathcal{S}}\mathbf{u}\right\|_2$, $\beta = b/\nu$, $v_i = au_i/(\nu r)$, and $\alpha_{\mathbf{u}} = bu_p/(\nu r)$, Gaussian integration by parts gives the exact half-label force

$$\begin{aligned}\mathbf{f}_{1/2}(\mathbf{u}) =& \frac{b}{2}\mathbf{e}_p \mathbb{E}_z[y\Phi(\alpha_{\mathbf{u}} + v_1 z_1 + v_2 z_2)]\\ &+ \frac{a}{2}\sum_{i=1}^{2}\boldsymbol{\mu}_i \mathbb{E}_z[y z_i \Phi(\alpha_{\mathbf{u}} + v_1 z_1 + v_2 z_2)]\\ &+ \frac{\nu}{2r}\mathbf{\Pi}_{\mathbf{p}^\perp}\mathbf{u}\, \mathbb{E}_z[y\varphi(\alpha_{\mathbf{u}} + v_1 z_1 + v_2 z_2)].\end{aligned}$$

One or two sign differences, using bounded derivatives of Φ, φ, give

$$\left\|\mathbf{f}_{1/2}(\mathbf{u})\right\|_2 \leq C\frac{a^2}{\nu}[\chi_s + (1 + |\beta|)\chi_s^2]. \tag{85}$$

This is a global finite-difference inequality, not a Taylor approximation. Let $\Gamma_* = \Gamma_{K_X} + 1$ and $\beta_* = 1 + PK_X + \alpha\Gamma_*$. The weighted position bounds imply $\mathbb{E}[(a^2/\nu)(1 + |\beta|)] \leq C\beta_* s^2/\nu_0$. Before the population stop, $\chi_s \leq \chi_0 + d^{-1/4}$. Under Assumption 3.1, Lemma A.3 gives

$$\beta_*(\chi_0 + d^{-1/4}) \leq c,$$

so the averaged force is bounded by $Cg(t)\chi_s$, where $g(t) = s(t)^2/\nu_0$ and $\int g \leq Cc_*$.

Let D_{X} be population RMS row displacement from initialization, and let B_{X}^2 be its row covariance operator norm. Deterministically, $B_{\mathrm{X}} \leq B_{\mathrm{X},0} + D_{\mathrm{X}}$. The logistic identity and Cauchy–Schwarz give a per-row residual bound $CB_{\mathrm{X}}V_{\mathrm{X}}$. Consequently

$$D_{\mathrm{X}}(t) \leq C\int_0^t (g + V_{\mathrm{X}})(d^{-1/2} + D_{\mathrm{X}})\, du, \qquad D_{\mathrm{X}} \leq Cc_* d^{-1/2}, \quad B_{\mathrm{X}} = O(d^{-1/2}), \tag{86}$$

since $\int(g + V_{\mathrm{X}}) \leq Cc_* + C/P$. The corresponding individual-row inequality preserves maximum center size $O(\chi_0)$ and lower perpendicular norms. For sampled characteristics, the residual still contains the deterministic B_{X}; the same argument with initial empirical semantic RMS $O(d^{-1/2})$ proves sampled covariance $O(1/d)$ uniformly in time, without a time net.

Embedding the auxiliary shell in the joint state space. Let $\Psi = (\mathbf{q}_{\mathrm{sh}}, \mathbf{u}_{A,\mathrm{sh}}, \mathbf{u}_{\mathrm{X},0})$, retaining the symmetric initial XOR law as a static comparison law. Exactly, for every $\mathbf{h}$,

$$F_{\mathrm{X},0}(\mathbf{h}) = 0, \qquad \mathbb{E}_\omega[\epsilon \mathbf{u}_{\mathrm{X},0} \mathbf{1}_{\mathbf{u}_{\mathrm{X},0}^\top \mathbf{h} > 0}] = 0.$$

Hence its discrepancy from the *full joint* population field is solely

$$\dot{\Psi} = \boldsymbol{\beta}^{\mathrm{joint}}(\Psi) - (0, 0, \lambda \epsilon \mathbf{f}_{1/2}(\mathbf{u}_{\mathrm{X},0}, \mathbf{q}_{\mathrm{sh}})). \tag{87}$$

The forcing has RMS $Cg/\sqrt{d}$ and maximum-row size $Cg\chi_0$. Shared signs cause no correction: the marginal equations and the lifted metric add the two head contributions. No trained parameter is frozen in the actual algorithm.

F.7 Inputs on every secant and grid point

Under Assumption 3.1, Lemma A.3 gives

$$m \geq C d^{121/64} L_d^{12} \ell_\delta, \qquad m \leq d^2, \qquad L_d^{24} \ell_\delta \leq c d^{1/32}.$$

The shell horizon bound in Theorem E.1 gives $R \leq CT/P$, and $\nu_0^{-1} = \sqrt{T}/\sigma$ by definition. We simplify the following bounds using Assumption 3.1. For $s_* := \sup_{0 \leq t \leq R_*} s(t)$, the shell results in Section E give

$$\begin{aligned} &s_* \leq C\nu_0 K_{\mathrm{X}}, && \int s \leq C\sqrt{G}/P, && \int s^2 \leq C\nu_0 K_{\mathrm{X}}/(P\sqrt{G}), \\ &\int V_k \leq C/P, && \int \gamma^2 \leq CT/P, && C_{2,k} \leq C(\sigma^4 + P^4 s^2), \end{aligned} \tag{88}$$

$$\left\| \mathbb{E} \mathbf{c}_k \mathbf{c}_k^\top \right\| \leq C \left(P^2 s^2 + \sigma^2 \mathbb{E} \textstyle\sum_i s_i^2 + \sigma^4 \left\| \mathbf{q} \right\|_2^2 \right).$$

Also $C_{1,k} \leq C(\sigma^2 + P^2 s)$ and $\mathbb{E}[K_{\mathrm{tok}} \left\| \mathbf{C}_k \right\|] \leq CP(\sigma^2 + P^2 s)$. The shell covariance retains its separate position and semantic coefficients $c^2 \leq C/P^2$ and $b^2 \leq C\gamma/P^2$. The weighted position moments in (64) include the Gaussian tails and the distractor position tilt. The directional secant calculation proves (71) before either stop is excluded.

The deterministic envelopes of Lemmas E.5 and E.6 control the integral-to-grid transfer with error $O(d^{-1}(1 + \log d)^8) = o(1)$. The curvature coefficient remains below the strict margin in (71) before either stop is excluded.

For both continuous integrals and the deterministic grid, the preceding clock estimates and Assumption 3.1 give universal envelopes

$$\begin{aligned} &\mathcal{V} = C/P, && \mathcal{J} = C(\sigma^2 T/P + P\sqrt{G}), \\ &\mathcal{Q} = C(\sigma^4 T/P + P^3 \nu_0 K_X/\sqrt{G}), && \mathcal{F} = \mathcal{V} + \mathcal{J} + \sqrt{\mathcal{V}\mathcal{Q}} \leq C L_d^{3/4}. \end{aligned} \tag{89}$$

They dominate $\mathcal{V}_k, \mathcal{J}_k, \mathcal{Q}_k$ and $\mathcal{F}_N$ up to universal factors. For the grid use the shell's monotone envelopes $1, s, s^2, 1 + \gamma^2$; their left sums are bounded by the corresponding integrals. The already controlled population-to-shell score error only multiplies s by a constant. Equivalently, Lemma E.6 bounds the quadrature remainder.

The row cross coefficient also needs an explicit confidence-free bound. On the full stopped secant, the fixed row bounds and projected-row floor give $Q_j \leq CC_2/\nu_0$. Equation (33) and Cauchy–Schwarz in time, or on the grid, imply

$$\mathcal{C} \leq C \left[\mathcal{V} + \mathcal{J} + \sqrt{L_d \mathcal{Q}/\nu_0} + \sqrt{\mathcal{V}\mathcal{Q}} \right] \leq C L_d^{3/2}. \tag{90}$$

Here Assumption 3.1 gives $\mathcal{Q}/\nu_0 \leq C[(\sigma\sqrt{T})^3/P + P^4] \leq CL_d^2$; the established curvature bound gives $\int H_g \leq CL_d$. The same bound holds for population comparisons. In particular it uses $Q_j \leq CC_2/\nu_0$, rather than replacing the covariance by the maximum initial center squared. The constants a_0, R_0, κ, D_* in the secant estimate are universal on these fixed tubes. Thus neither (90) nor the amplification $d^{1/128}$ contains an unrecorded factor depending on ℓ_δ.

F.8 THE POPULATION AND SGD CONTINUATIONS

For population comparisons replace empirical row averages in $\mathcal{H}$ by expectations over the seed coupling, and maximum row error by essential supremum. The same secant proof and continuous Gronwall argument apply. The initial auxiliary row is coupled to $\rho\mathbf{n}$, where $\mathbf{n} = \mathbf{\Pi}_{\mathcal{N}}\mathbf{u}_0 / \|\mathbf{\Pi}_{\mathcal{N}}\mathbf{u}_0\|_2$. Its essential-supremum error is at most $2\chi_0$, whereas

$$\mathbb{E}\|\mathbf{u}_0 - \rho\mathbf{n}\|_2^2 \le C/d.$$

Indeed the unconditioned Gaussian semantic and radial second moments have this order, and division by a cutoff probability at least $1/2$ preserves it. The XOR rows in the two compared states start identically.

Population continuation. Apply the continuous counterpart of Lemma C.4 to (87), stopped when the population-to-shell RMS or maximum-row difference reaches $d^{-1/4}$. Its initial-plus-forcing budgets are $D_H \le C/\sqrt{d}$ and $D_\infty \le C\chi_0$. Equations (71) and (90) give

$$\begin{aligned} E_{\mathrm{pop}} &\le Cd^{1/128}\left[\chi_0 + \frac{1+\mathcal{C}}{\sqrt{d}}\right] \\ &\le Cd^{-63/128}(\sqrt{\ell_\delta} + L_d^{3/2}) \le cd^{-7/16}. \end{aligned} \tag{91}$$

For the last inequality use $\sqrt{\ell_\delta} \le \sqrt{c}\, d^{1/64} L_d^{-12}$ from Assumption 3.1 and Lemma A.3; all remaining factors are fixed powers of L_d. This strict improvement excludes the stop and preserves the positive projected-row floors needed by Lemma C.9. The reference therefore continues through R_*. No prior bound on the evolved XOR attention-weight force was assumed. Equation (86) sharpens XOR covariance afterward; its individual-row version gives maximum center $O(\chi_0)$.

Sampled auxiliary covariance. Write the coupled shell rows as $\mathbf{v}_j = c\mathbf{e}_p + \epsilon_j b\mathbf{e}_s + r\mathbf{n}_j$. The single initial event $\left\| m^{-1}\sum_j \mathbf{n}_j\mathbf{n}_j^\top \right\| \le C/d$ implies

$$\mathbf{\Sigma}_V \preceq 3c^2\mathbf{e}_p\mathbf{e}_p^\top + 3b^2\mathbf{e}_s\mathbf{e}_s^\top + 3r^2m^{-1}\sum_j \mathbf{n}_j\mathbf{n}_j^\top, \qquad \|\mathbf{\Sigma}_V\| \le CG/P^2$$

at every time. The uniform coarse comparison gives $\mathbf{\Sigma}_{\bar{\mathbf{u}}_A} \preceq 2\mathbf{\Sigma}_V + 2d^{-7/8}\mathbf{I}_d$ and the same bound for sampled auxiliary characteristics. The parity argument above gives sampled XOR covariance $O(1/d)$. Both statements hold throughout the reference interval with no time net.

Actual SGD continuation. Use Lemma F.1 with $\delta_e = \delta_J/10$, the two stopped batch bounds, and the deterministic Euler estimate. The initial sampled reference equals the actual initial state. The total forcing budget is at most

$$D_H + D_\infty \le C\left[\mathcal{F}\sqrt{\ell_\delta/m} + P^2\{\sqrt{\eta R\ell_\delta/B} + \eta\ell_\delta/B\} + \eta RP^6/\nu_0\right] =: \mathcal{D}_J.$$

Consequently Lemma C.4 gives

$$E_H + E_\infty \le Cd^{1/128}(1+\mathcal{C})\mathcal{D}_J \le cd^{-15/16}. \tag{92}$$

Using Assumption 3.1, Lemma A.3, and the learning-rate choice in (15), its finite-width, square-root batch, linear batch, and Euler terms are respectively at most

$$\begin{aligned} &CC_m^{-1/2}d^{-15/16}L_d^{-15/4}, \\ &Cd^{-127/128}L_d^{23/8}\sqrt{\ell_\delta} \le C\sqrt{c_{\mathrm{tr}}}d^{-125/128}L_d^{-73/8}, \\ &Cd^{-3+1/128}L_d^{5/2}\ell_\delta, \qquad Cd^{-127/128}L_d^{91/16}. \end{aligned}$$

Here C_m, c_{tr} are the constants in $m \ge C_m d^{121/64}L_d^{12}\ell_\delta$ and $L_d^{24}\ell_\delta \le c_{\mathrm{tr}}d^{1/32}$ supplied by Lemma A.3. Its universal constant choices make the sum smaller than the required fixed multiple of $d^{-15/16}$. The strict bound excludes the $d^{-2/5}$ stop. After it closes, $B_{\mathrm{X,actual}} \le B_{\mathrm{X,sampled}} + E_H = O(d^{-1/2})$. The argument does not impose this sharper covariance on the temporary tube.

The finer structure. The full state controls feedback in the averaged metric. Individual rows have a larger cross sensitivity, but its time integral enters only as forcing in (28). Population sign symmetry makes the embedded reference defect small, and fresh data makes the finite-width and batch fluctuations amenable to cumulative estimates. These are three different cancellations; none requires independence of trained rows.

F.9 Checkpoint rounding and auxiliary classification

We now complete Proposition D.1 from the preceding estimates. All confidence dependence is carried by ℓ_δ and Lemma A.3; the constants in Assumption 3.1 are chosen once to enforce the fixed numerical comparisons.

Initialization events. Gaussian coordinate and chi-square tails show that one row fails the cutoff with probability at most $C\exp(-cd\chi_0^2)$; a union bound over $2m$ rows makes the product failure at most $\delta_J/10$. Conditional nuisance directions remain uniform on the sphere. For a unit vector $\mathbf{v}\in\mathcal{N}$, the random variable $d(\mathbf{v}^\top\mathbf{n})^2$ has moments bounded by $C^r r!$: integrate the spherical-coordinate density $\Gamma((d-3)/2)(1-u^2)^{(d-6)/2}/[\sqrt{\pi}\,\Gamma((d-4)/2)]$. Its exponential moment is therefore bounded at a fixed positive argument. Independence, Chernoff's inequality, and a $1/4$ net of the unit sphere give

$$\mathbb{P}\left\{\left\|m^{-1}\sum_j \mathbf{n}_j\mathbf{n}_j^\top\right\| > C/d\right\} \le \exp(Cd-cm).$$

The identical argument applies to the conditioned Gaussian rows: their conditional one-dimensional exponential moments are bounded by the unconditional moments divided by the one-row cutoff probability, which is at least $1/2$ under Assumption 3.1, by the cutoff estimate above and Lemma A.3. Their covariance is at most $C\mathbf{I}_d/d$. Initial semantic RMS is $O(d^{-1/2})$ by the same moment argument. Under Assumption 3.1, Lemma A.3 gives $m\ge C(d+\ell_\delta)$, so the sum of these failures is at most $\delta_J/10$. Thus all initial covariance and RMS events invoked above hold jointly with the allocated confidence. Conditioning preserves independence across seed pairs because the cutoff is a product event and does not involve the signs.

A common deterministic horizon. Use the shell endpoint R_* from Section E, and let t_k be the joint reference's first hitting time of Γ_{K_k}. The population comparison and the buffered transfer in Section E give

$$\sup_{t\le R_*}|\bar\gamma(t)-\gamma_{\mathrm{sh}}(t)| \le Pd^{-7/16}=o(1), \qquad R_*\le CT/P.$$

The row bounds and $\mathbb{E}\,\|\mathbf{C}\|\le CP^2$ give $\left\|\dot{\tilde{\mathbf{q}}}\right\|_2\le CP^2$ and $|\dot{\tilde\gamma}|\le CP^3$; the shell has the same crude speed bound. At $t_{\mathrm X}$ the shell score is $\Gamma_{K_{\mathrm X}}+o(1)$, so the remaining score buffer takes at least c/P^3 physical time. The learning-rate choice in (15) gives $\eta\le d^{-3}$, while Assumption 3.1 and Lemma A.3 give $d^{-3}\le cP^{-3}$. Hence $(n_{\mathrm X}+1)\eta<R_*$. Run the finite stopped comparison through $\lfloor R_*/\eta\rfloor$; no reference continuation beyond the proved shell interval is required.

Continuity and $\Gamma_{K_{\mathrm X}}>\Gamma_{K_A}>0$ imply $0<t_A<t_{\mathrm X}$. Moreover,

$$t_{\mathrm X}-t_A\ge\frac{\Gamma_{K_{\mathrm X}}-\Gamma_{K_A}}{CP^3}=\frac{\log K_{\mathrm X}}{2CP^3}\gg\eta,$$

where the final comparison uses Assumption 3.1 and the learning-rate choice in (15). Thus $1\le n_A<n_{\mathrm X}$. Rounding changes scores and the feature clock by at most $\eta\,\mathrm{polylog}(d)$; this is negligible in every displayed bound. The shell hitting-time sandwich transfers the prefix clocks and the lower terminal-strip clock to the same joint reference. The finite and population continuations therefore prove (37)–(D.1) under Assumption 3.1 and the learning-rate choice in (15). The constants in (92) may be chosen so that the sum of all three errors in (40) is at most $d^{-15/16}$. Since A is nondecreasing, rounding also preserves the lower clock: $A(n_k\eta)\ge A(t_k)$; its upper increment is at most $C\eta\lambda\nu_0K_X^2$.

Auxiliary success. Put $u_A=n_A\eta$. At this deterministic time the shell score is $\Gamma_{K_A}+o(1)$. Its already proved margin is

$$\mathbb{P}_{\mathcal{D}_A}\{yF_{A,\mathrm{sh}}(\mathbf{h}(\mathbf{X},\mathbf{q}_{\mathrm{sh}}(u_A)))<m_A\}\le Ce^{-cK_A^2}, \qquad m_A=c\nu_0K_A/P. \tag{93}$$

At the same time, the all-tail moments and secant bounds give

$$\sqrt{\|\mathbb{E}\mathbf{h}\mathbf{h}^\top\|} \le C\nu_0(PK_A + 1 + \alpha\Gamma_{K_A}),$$
$$(\mathbb{E}\,\|\mathbf{C}\|^2)^{1/2} \le C\nu_0(P^2K_A + P\alpha), \qquad B_A \le C\sqrt{G}/P.$$

Set

$$D_A = d + P^2K_A^2 + \alpha^2(\Gamma_{K_A} + 1)^2,$$
$$\Theta_A = C\left[P^2(1+\sqrt{G}) + \frac{P}{K_A}\{1 + \alpha(\Gamma_{K_A} + 1 + \sqrt{G})\}\right].$$

Lemma C.1, applied to population versus shell and actual versus sampled reference, bounds their combined L^2 output error, divided by m_A, by $\Theta_A(E_{\mathrm{pop}} + E_H + E_\infty)$. Assumption 3.1 gives $D_A \le Cd$ and $\Theta_A \le CL_d^{3/2}$; the displayed axial moments give $V_A(u_A) \le C\nu_0^2 D_A$.

It remains to control the empirical-reference output at the fixed, deterministic time u_A. Under the product cutoff law its summands are independent across seed tuples and have $L^2(\mathcal{D}_A)$ norm at most $R_0\sqrt{V_A(u_A)}$. Changing one tuple changes

$$\Delta_A = \|F_{A,m}(\mathbf{h}(\mathbf{X}, \bar{\mathbf{q}}(u_A))) - F_{A,\infty}(\mathbf{h}(\mathbf{X}, \bar{\mathbf{q}}(u_A)))\|_{L^2(\mathcal{D}_A)}$$

by at most $2R_0\sqrt{V_A(u_A)}/m$. Equation (75) gives $\mathbb{E}\Delta_A \le R_0\sqrt{V_A(u_A)/m}$. The bounded-difference argument in Lemma F.1, now for this single norm, therefore gives, with conditional failure at most $\delta_J/10$,

$$\Delta_A \le C\sqrt{V_A(u_A)\ell_\delta/m}, \qquad \Delta_A/m_A \le \frac{CP}{K_A}\sqrt{D_A\ell_\delta/m}.$$

No seed is conditioned on subsequent training success. Combining the three output comparisons and using Chebyshev only for the fresh-test output difference in (93) yields

$$\mathrm{err}_A(n_A) \le Ce^{-cK_A^2} + C\left[\Theta_A(E_{\mathrm{pop}} + E_H + E_\infty) + \frac{P}{K_A}\sqrt{D_A\ell_\delta/m}\right]^2$$
$$\le Ce^{-cK_A^2} + CL_d^3 d^{-7/8} + C\frac{P^2 d\ell_\delta}{mK_A^2}.$$

For the new checkpoint, Assumption 3.1 and Lemma A.3 give

$$\frac{P^2}{K_A^2} = \frac{P}{c_*\sqrt{G}} \le CL_d^{1/2}, \qquad \frac{P^2 d\ell_\delta}{mK_A^2} \le Cd^{-57/64}L_d^{-23/2}.$$

Since $K_A^2 = c_*P\sqrt{G} \le C\sqrt{L_d \log L_d} = o(\log d)$, both polynomial remainders are absorbed into $Ce^{-c'K_A^2}$ by Lemma A.3. This proves the auxiliary conclusion without an inverse power of δ_J.

The product cutoff, initial covariance/RMS, integrated forcing, stopped batch, and fixed-time empirical-output modules use at most $\delta_J/10$ each (the batch module includes both norms). Their total is at most δ_J, including the return from the conditioned product law to the original initialization law. The deterministic shell margin is a fresh-test error contribution and does not spend training confidence. This completes Proposition D.1.

G A PARITY-SENSITIVE OFF-AXIS FORCE ESTIMATE

Write $\mathbf{\Pi} = \mathbf{\Pi}_\perp$, and $E_c = \mathrm{span}\{\mathbf{e}_p, \boldsymbol{\mu}_1, \boldsymbol{\mu}_2\}$. Throughout this section the expectation is over one fresh XOR example with its original Gaussian noise law. Put

$$\nu_0 = \frac{\sigma}{\sqrt{T}}, \qquad s_\gamma = \frac{e^\gamma}{T}, \qquad g_\gamma = \frac{s_\gamma^2}{\nu_0}, \qquad \mathbf{f}_{1/2}(\mathbf{u}, \mathbf{q}) = \frac{1}{2}\mathbb{E}[y\mathbf{h}(\mathbf{X}, \mathbf{q})\mathbf{1}_{\{\mathbf{u}^\top \mathbf{h}(\mathbf{X}, \mathbf{q}) > 0\}}].$$

The quantity s_γ here is a deterministic attention scale; it is not the random attention weight of a token.

Lemma G.1 (Off-axis half-label force)**.** *Impose Assumption 3.1. Fix a sufficiently small universal $c_0 > 0$ and a sufficiently large universal $C_0 > 0$ once and for all. Suppose*

$$\mathbf{q}_* = \gamma \mathbf{e}_p / P, \qquad -1 \le \gamma \le \Gamma, \qquad \Gamma \le C_0 \log\log d,$$
$$\|\mathbf{q} - \mathbf{q}_*\|_2 \le \xi \le d^{-15/16}, \qquad c_0 \le \|\mathbf{\Pi u}\|_2 \le \|\mathbf{u}\|_2 \le C_0, \qquad \|\mathbf{\Pi}_{E_c}\mathbf{u}\|_2 \le \chi \le C_0\sqrt{\ell_\delta / d}.$$

Assume also $s_\gamma/\nu_0 \le C_0 K_{\mathrm{X}}$, where $K_{\mathrm{X}} = c_* P\sqrt{G}$. There is a universal constant $C > 0$ such that

$$\|\mathbf{f}_{1/2}(\mathbf{u}, \mathbf{q}) - \mathbf{f}_{1/2}(\mathbf{u}, \mathbf{q}_*)\|_2 \le C L_d^{12} \xi \left\{ g_\gamma(\chi + \sigma) + s_\gamma^2 \right\}. \tag{94}$$

Under Assumption 3.1, Lemma A.3 and the displayed bound on χ ensure $\chi + \sigma \le 1$. The estimate also holds after any orthogonal projection. It is uniform over deterministic admissible $(\mathbf{u}, \mathbf{q}, \mathbf{q}_*)$ in the sense that the same constant bounds each fresh-example expectation. Consequently it applies pointwise to any past-measurable admissible training state; no independence of the trained rows is required.

Proof. We prove moment estimates for the original untruncated Gaussian law. This will also justify every differentiation below.

Removing exactly the common perpendicular Gaussian. Represent the perpendicular noises as independent $\boldsymbol{\eta}_i \sim \mathcal{N}(0, \sigma^2 \mathbf{\Pi})$, $1 \le i \le T$. Independently let $\zeta_1 = \zeta_2 = 0$ and $\zeta_i \sim \mathcal{N}(0, \sigma^2)$ for $i \ge 3$. Define

$$\mathbf{g}_0 = T^{-1} \sum_i \boldsymbol{\eta}_i, \qquad \widetilde{\boldsymbol{\eta}}_i = \boldsymbol{\eta}_i - \mathbf{g}_0,$$
$$\mathbf{d}_1(z) = \mathbf{p} + z_1 \boldsymbol{\mu}_1, \qquad \mathbf{d}_2(z) = \mathbf{p} + z_2 \boldsymbol{\mu}_2, \qquad \mathbf{d}_i(z) = 0 \quad (i \ge 3),$$
$$\widetilde{\mathbf{x}}_i(z) = \mathbf{d}_i(z) + \zeta_i \mathbf{e}_p + \widetilde{\boldsymbol{\eta}}_i, \qquad \mathbf{x}_i(z) = \widetilde{\mathbf{x}}_i(z) + \mathbf{g}_0.$$

Here $\mathbf{g}_0 \sim \mathcal{N}(0, \nu_0^2 \mathbf{\Pi})$ is independent of all centered residuals, position noises, and signs. The shared logit $\mathbf{q}^\top \mathbf{g}_0$ cancels from softmax. Thus, exactly for every attention weight and also for interpolated $z \in [-1, 1]^2$,

$$\mathbf{h}(\mathbf{X}, \mathbf{q}) = \mathbf{c}(z, \mathbf{q}) + \mathbf{g}_0, \qquad \mathbf{c}(z, \mathbf{q}) = \sum_i s_i(z, \mathbf{q}) \widetilde{\mathbf{x}}_i(z), \qquad \mathbf{C}(z, \mathbf{q}) = D_{\mathbf{q}} \mathbf{c} = \sum_i s_i \mathbf{r}_i r_i^\top, \tag{95}$$

where $\mathbf{r}_i = \widetilde{\mathbf{x}}_i - \mathbf{c} = \mathbf{x}_i - \mathbf{h}$. All quantities on the right are independent of $\mathbf{g}_0$. Only a perpendicular common Gaussian has been removed; the relevant-token position noises are still exactly zero.

Set $r = \|\mathbf{\Pi u}\|_2$, $t = \mathbf{u}^\top \mathbf{c}/(\nu_0 r)$, and $\mathbf{k} = \nu_0 \mathbf{\Pi u}/r$. Gaussian regression gives

$$\mathbf{r}_{\mathbf{u}}(\mathbf{c}) := \mathbb{E}_{\mathbf{g}_0}[(\mathbf{c} + \mathbf{g}_0) \mathbf{1}_{\{\mathbf{u}^\top(\mathbf{c} + \mathbf{g}_0) > 0\}}] = \mathbf{c}\Phi(t) + \mathbf{k}\varphi(t), \tag{96}$$

where Φ, φ are the standard normal distribution function and density. For a vector $\mathbf{b}$, write $t_{\mathbf{b}} = \mathbf{u}^\top \mathbf{b}/(\nu_0 r)$. Direct differentiation yields

$$D\mathbf{r}_{\mathbf{u}}[\mathbf{b}] = \mathbf{b}\Phi(t) + (\mathbf{c}\varphi(t) + \mathbf{k}\varphi'(t)) t_{\mathbf{b}}, \tag{97}$$
$$D^2 \mathbf{r}_{\mathbf{u}}[\mathbf{b}, \mathbf{d}] = (\mathbf{b} t_{\mathbf{d}} + \mathbf{d} t_{\mathbf{b}})\varphi(t) + (\mathbf{c}\varphi'(t) + \mathbf{k}\varphi''(t)) t_{\mathbf{b}} t_{\mathbf{d}}, \tag{98}$$
$$D^3 \mathbf{r}_{\mathbf{u}}[\mathbf{b}, \mathbf{d}, \mathbf{f}] = (\mathbf{b} t_{\mathbf{d}} t_{\mathbf{f}} + \mathbf{d} t_{\mathbf{b}} t_{\mathbf{f}} + \mathbf{f} t_{\mathbf{b}} t_{\mathbf{d}})\varphi'(t) + (\mathbf{c}\varphi''(t) + \mathbf{k}\varphi'''(t)) t_{\mathbf{b}} t_{\mathbf{d}} t_{\mathbf{f}}. \tag{99}$$

These identities are valid globally in $\mathbf{c}$; no small-ball assumption on the standardized preactivation is used.

A global softmax and covariance envelope. The following estimates are uniform over $z \in [-1, 1]^2$, all attention weights $\mathbf{q}' = \gamma' \mathbf{e}_p / P + \mathbf{w}$ with γ' between 0 and γ and $\|\mathbf{w}\|_2 \le \xi$, and each fixed unit direction $\mathbf{v}$. Let

$$M_p = \max_i |\zeta_i|, \qquad M_n = \max_i \|\widetilde{\boldsymbol{\eta}}_i + \zeta_i \mathbf{e}_p\|_2,$$
$$Z = \exp(2\Gamma_0 M_p / P + 2\xi(P + 1 + M_n)), \qquad \Gamma_0 = \max\{1, \Gamma\}.$$

Comparing each logit with its noiseless axial logit and bounding the denominator by its $T-2$ distractor terms gives

$$s_i(z,\mathbf{q}') \le C s_\gamma Z \quad (i=1,2), \qquad s_i(z,\mathbf{q}') \le CZ/T \quad (i \ge 3). \tag{100}$$

For $\gamma < 0$ the factor C absorbs $\sup_{\gamma' \in [\gamma,0]} e^{\gamma'-\gamma} \le e$.

For completeness, the displayed moment bounds below hold for $1 \le b \le 64$. Auxiliary Gaussian moments of order at most 1024 are used directly from the following tails. The elementary Gaussian tail and the chi-square moment generating function imply, for $u \ge 0$,

$$\mathbb{P}\{M_p > C\sigma(\sqrt{\log T} + \sqrt{u})\} \le 2e^{-u},$$
$$\mathbb{P}\{M_n > C\sigma(\sqrt{d} + \sqrt{\log T} + \sqrt{u})\} \le 4e^{-u}.$$

The second inequality follows by applying $\mathbb{P}\{\|\mathcal{N}(0,\sigma^2\mathbf{I}_d)\|_2 > \sigma(\sqrt{d}+\sqrt{2u})\} \le e^{-u}$ to each original noise and to $\mathbf{g}_0$, followed by a union bound. Integration of these tails, using $a\sqrt{u} - u \le a^2/2 - u/2$, proves

$$\mathbb{E}Z^b \le C \qquad (1 \le b \le 64). \tag{101}$$

Indeed, Assumption 3.1, Lemma A.3, and the displayed bounds on Γ_0, ξ give $\Gamma_0\sigma\sqrt{\log T}/P \le c$ and $\xi(P + \sigma\sqrt{d} + \sigma\sqrt{\log T}) \le c$ for a fixed sufficiently small universal $c > 0$.

Here and below a supremum over the displayed attention-weight/sign envelope may be inserted inside the moments. The covariance is positive semidefinite and is bounded by its uncentered second moment. Splitting the two relevant columns from the distractor columns in that second moment and using (100), we obtain

$$\|\mathbf{C}(z,\mathbf{q}')\| \le CZ\left[s_\gamma \sum_{i=1}^{2}(P^2 + 1 + \|\widetilde{\boldsymbol{\eta}}_i\|_2^2) + \frac{1}{T}\|\widetilde{\boldsymbol{\Xi}}_D\|^2\right], \tag{102}$$

where $\widetilde{\boldsymbol{\Xi}}_D$ has columns $\widetilde{\boldsymbol{\eta}}_i + \zeta_i \mathbf{e}_p$, $i \ge 3$. The centered perpendicular matrix is the original Gaussian matrix multiplied on the right by a projection. Adding the one position row and restricting to distractor columns cannot increase its operator norm by more than their sum. Consequently

$$\left\|\left\|\widetilde{\boldsymbol{\Xi}}_D\right\|\right\|_{L^b} \le C\sigma(\sqrt{T} + \sqrt{d}).$$

One elementary derivation is to take a $1/4$-net of the unit sphere in $\mathbb{R}^d$, of size at most 9^d, apply the chi-square upper-tail bound to $\|\boldsymbol{\Xi}^\top \mathbf{u}\|_2$ at its net points, and integrate the resulting tail. This argument also covers the projected columns here. Since Assumption 3.1 gives $T \ge d$, Hölder's inequality in (102) gives

$$\left\|\sup_{z,\mathbf{q}'} \|\mathbf{C}(z,\mathbf{q}')\|\right\|_{L^b} \le C(\sigma^2 + P^2 s_\gamma). \tag{103}$$

This is the off-axis extension of the covariance estimate used in Section E.5.

The residual pooled vector has no $\sqrt{d}$ loss. The exact cancellation of the centered perpendicular noises gives

$$\mathbf{c}(z,0) = \frac{2\mathbf{p} + z_1\boldsymbol{\mu}_1 + z_2\boldsymbol{\mu}_2}{T} + \frac{1}{T}\sum_{i=3}^{T}\zeta_i \mathbf{e}_p. \tag{104}$$

In particular $\|\sup_z \|\mathbf{c}(z,0)\|_2\|_{L^b} \le C(P/T + \nu_0)$. Integration along the straight segment from 0 to $\mathbf{q}'$ and (103) now yield

$$\left\|\sup_{z,\mathbf{q}'} \|\mathbf{c}(z,\mathbf{q}')\|_2\right\|_{L^b} \le C\left[P/T + \nu_0 + (\Gamma_0/P + \xi)(\sigma^2 + P^2 s_\gamma)\right] \le \nu_0 L_b, \tag{105}$$

where one may take $L_b = CL_d^3$ for every $1 \le b \le 64$. By Assumption 3.1 and the displayed bounds on $s_\gamma/\nu_0, \Gamma_0, \xi, P/(\sigma\sqrt{T}) \le CL_d^{1/2}$, $\sigma\sqrt{T} + P^2 s_\gamma/\nu_0 \le CL_d^2$, and $\Gamma_0/P + \xi \le CL_d$. Notice that this estimate concerns $\mathbf{c}$, not $\mathbf{h}$: the removed $\mathbf{g}_0$ has norm of order $\nu_0\sqrt{d}$.

There is also a useful dimension-sensitive scalar bound. For every fixed admissible row $\mathbf{u}$, the Gaussian maximal inequality gives

$$\big\|\max_i |\mathbf{u}^\top \widetilde{\mathbf{x}}_i(z)|\big\|_{L^b} \le C\{P\chi + \chi + \sigma\|\mathbf{u}\|_2\sqrt{\log T}\} \le (\chi+\sigma)L_b. \tag{106}$$

The same estimate holds with $\sup_z$ inside the norm. Because $\mathbf{c}$ is a convex combination of the $\widetilde{\mathbf{x}}_i$, it follows that $\|\sup_{z,\mathbf{q}'} |\mathbf{u}^\top \mathbf{r}_i|\|_{L^b} \le (\chi+\sigma)L_b$. This use of a scalar fresh-Gaussian projection is valid for every fixed current row. Replacing it by $\|\mathbf{u}\|_2\|\mathbf{r}_i\|_2$ would lose the essential factor $d^{-1/2}$.

Exact derivatives and their moments. Subscripts $1, 2$ denote differentiation in the interpolated signs, and subscript $\mathbf{v}$ differentiation in a fixed unit attention-weight direction. Put $\beta_i = \boldsymbol{\mu}_i^\top \mathbf{q}'$, so $|\beta_i| \le \xi$, and $\mathbf{r}_* = 2\mathbf{c} - \widetilde{\mathbf{x}}_1 - \widetilde{\mathbf{x}}_2$. Differentiating softmax gives

$$\mathbf{c}_i = s_i(\boldsymbol{\mu}_i + \beta_i \mathbf{r}_i), \qquad \mathbf{c}_\mathbf{v} = \mathbf{C}\mathbf{v}, \tag{107}$$

$$\mathbf{c}_{i\mathbf{v}} = s_i\{(\mathbf{v}^\top \mathbf{r}_i)\boldsymbol{\mu}_i + (\boldsymbol{\mu}_i^\top \mathbf{v})\mathbf{r}_i + \beta_i[(\mathbf{v}^\top \mathbf{r}_i)\mathbf{r}_i - \mathbf{C}\mathbf{v}]\}, \tag{108}$$

$$\mathbf{c}_{12} = s_1 s_2 \mathbf{b}, \qquad \mathbf{b} = -\beta_2\boldsymbol{\mu}_1 - \beta_1\boldsymbol{\mu}_2 + \beta_1\beta_2\mathbf{r}_*. \tag{109}$$

Differentiating the last identity once more gives

$$\begin{aligned}\mathbf{c}_{12\mathbf{v}} = s_1 s_2\{&[\mathbf{v}^\top(\mathbf{r}_1+\mathbf{r}_2)]\mathbf{b} - (\boldsymbol{\mu}_2^\top\mathbf{v})\boldsymbol{\mu}_1 - (\boldsymbol{\mu}_1^\top\mathbf{v})\boldsymbol{\mu}_2 \\ &+ [(\boldsymbol{\mu}_1^\top\mathbf{v})\beta_2 + \beta_1(\boldsymbol{\mu}_2^\top\mathbf{v})]\mathbf{r}_* + 2\beta_1\beta_2\mathbf{C}\mathbf{v}\}.\end{aligned} \tag{110}$$

For example $D_{z_i} s_j = s_j(\mathbf{1}_{\{i=j\}} - s_i)\beta_i$ and $D_\mathbf{v} s_i = s_i \mathbf{v}^\top \mathbf{r}_i$ give the first three identities directly; $D_\mathbf{v}\mathbf{r}_* = 2\mathbf{C}\mathbf{v}$ gives the last one.

Let $\kappa = \chi + \sigma$. Combining (100)–(106) with these exact derivatives gives the following table. Each entry bounds the L^{16} norm, including the supremum over the sign/attention-weight envelope. In this table take $L = CL_d^3$, with a universal constant large enough for the finitely many fixed moments used in Hölder's inequality.

quantity $\mathbf{d}$	$\big\|\|\mathbf{d}\|_2\big\|_{L^{16}}$	$\|\mathbf{u}^\top\mathbf{d}\|_{L^{16}}$
$\mathbf{c}$	$\nu_0 L$	$\nu_0 L$
$\mathbf{c}_\mathbf{v}$	$\nu_0 L$	$\nu_0 L$
$\mathbf{c}_i$	$s_\gamma L$	$s_\gamma \kappa L$
$\mathbf{c}_{i\mathbf{v}}$	$s_\gamma L$	$s_\gamma \kappa L$
$\mathbf{c}_{12}$	$s_\gamma^2 \xi L$	$s_\gamma^2 \xi\kappa L$
$\mathbf{c}_{12\mathbf{v}}$	$s_\gamma^2 L$	$s_\gamma^2 \kappa L$

The powers in the scalar columns can be checked before enlarging L. The preceding estimates give, for each fixed moment order used here,

$$\|\mathbf{r}_i\|_{L^b} \le CL_d^{1/2}, \qquad \|\mathbf{u}^\top\mathbf{r}_i\|_{L^b} \le C\kappa L_d^{1/2}, \qquad \|\mathbf{C}\|_{L^b} \le C\nu_0 L_d^2,$$

with the same sign/attention-weight suprema. Hence the exact derivative identities and Hölder's inequality imply

$$\begin{aligned}\|\mathbf{u}^\top\mathbf{c}_{i\mathbf{v}}/s_i\|_{L^{16}} &\le C\{\chi L_d^{1/2} + \kappa L_d^{1/2} + \xi(\kappa L_d + \nu_0 L_d^2)\} \\ &\le C\kappa L_d^2, \\ \|\mathbf{u}^\top\mathbf{c}_{12\mathbf{v}}/(s_1 s_2)\|_{L^{16}} &\le C\{\xi\kappa L_d + 2\chi + \xi\kappa L_d^{1/2} + \xi^2\nu_0 L_d^2\} \\ &\le C\kappa L_d^2.\end{aligned}$$

Here $\xi \le 1$ and $\nu_0 \le \kappa$ follow from the displayed bound on ξ and Assumption 3.1. Combining with the weight moments proves the displayed scalar entries; the same derivative identities and vector bounds prove the other column.

Taking the mixed sign difference before norms. Let $\mathbf{f}_{\mathrm{full}}(\mathbf{u},\mathbf{q}) = 2\mathbf{f}_{1/2}(\mathbf{u},\mathbf{q})$. Averaging the four signs and applying the fundamental theorem of calculus twice gives

$$D_{\mathbf{v}}\mathbf{f}_{\mathrm{full}}(\mathbf{u},\mathbf{q}') = \frac{1}{4}\mathbb{E}_{\mathrm{res}}\int_{[-1,1]^2}\{D^3\mathbf{r}_{\mathbf{u}}[\mathbf{c}_{\mathbf{v}},\mathbf{c}_1,\mathbf{c}_2] + D^2\mathbf{r}_{\mathbf{u}}[\mathbf{c}_{1\mathbf{v}},\mathbf{c}_2] + D^2\mathbf{r}_{\mathbf{u}}[\mathbf{c}_1,\mathbf{c}_{2\mathbf{v}}] + D^2\mathbf{r}_{\mathbf{u}}[\mathbf{c}_{\mathbf{v}},\mathbf{c}_{12}] + D\mathbf{r}_{\mathbf{u}}[\mathbf{c}_{12\mathbf{v}}]\}\,dz_1dz_2. \tag{111}$$

Here $\mathbb{E}_{\mathrm{res}}$ averages the centered residuals and position noises. The identity uses the exact factor $1/4$ from the four sign patterns. The Gaussian density and its first three derivatives are bounded, and $\|\mathbf{k}\|_2 = \nu_0$, $r \ge c_0$. Substitution of the table into (97)–(99), followed by Hölder, bounds the expected norms of the five displayed groups, respectively, by

$$\begin{aligned} &Cg_\gamma(\kappa+\kappa^2)L^4, \\ &Cg_\gamma(\kappa+\kappa^2)L^3, \qquad Cg_\gamma(\kappa+\kappa^2)L^3, \\ &Cs_\gamma^2\xi(1+\kappa)L^3, \qquad Cs_\gamma^2(1+\kappa)L^2. \end{aligned}$$

For instance, the term $\mathbf{c}_1 t_{\mathbf{c}_{\mathbf{v}}} t_{\mathbf{c}_2}$ has size $s_\gamma \cdot 1 \cdot (s_\gamma\kappa/\nu_0)$, whereas $\mathbf{c}_{\mathbf{v}} t_{\mathbf{c}_1} t_{\mathbf{c}_2}$ has size $\nu_0(s_\gamma\kappa/\nu_0)^2$. These two terms explain the κ and κ^2 factors. Since $\kappa,\xi \le 1$,

$$\|D_{\mathbf{v}}\mathbf{f}_{\mathrm{full}}(\mathbf{u},\mathbf{q}')\|_2 \le CL^4\{g_\gamma\kappa + s_\gamma^2\}.$$

All expressions possess integrable majorants on the full sign/attention-weight envelope by the moment table. Dominated convergence therefore justifies the attention-weight differentiations, residual expectation, and sign integrals. Finally integrate along $\mathbf{q}_* + u(\mathbf{q}-\mathbf{q}_*)$, $0 \le u \le 1$, and divide by two. Orthogonal projections have operator norm at most one, completing the proof of (94). □

H Anisotropic tracking of the XOR head

Only the XOR head is used in this section. Write $\mathbf{u}_{j,n} = \mathbf{u}_{\mathrm{X},j,n}$, and define the fixed operator $\mathbf{J} = \boldsymbol{\mu}_1\boldsymbol{\mu}_2^\top + \boldsymbol{\mu}_2\boldsymbol{\mu}_1^\top$.

The interface supplied by Proposition D.1. We use L_d and ℓ_δ from the parameter ledger in Lemma A.3. Under Assumption 3.1 and the learning-rate choice in (15), it gives

$$m \ge Cd^{121/64}L_d^{12}\ell_\delta, \qquad L_d^{24}\ell_\delta \le cd^{1/32}, \qquad m \le d^2. \tag{112}$$

All fixed dimension comparisons in this section follow from Assumption 3.1 through Lemma A.3. Fix the failure allocation $\delta_H = \delta/8$. Set $N_{\mathrm{X}} = n_{\mathrm{X}}$ and $R = N_{\mathrm{X}}\eta \le CT/P$. The reference of Proposition D.1 is defined on $[0, R+\eta]$ by the buffer argument in Section F.9. Write it as $\bar{\mathbf{q}}(t) = \gamma(t)\mathbf{e}_p/P$. At this attention weight, conditional on the original position noises, the exact fresh XOR pool is

$$\mathbf{h}_\gamma = b\mathbf{e}_p + a(z_1\boldsymbol{\mu}_1 + z_2\boldsymbol{\mu}_2) + \nu\mathbf{z}, \qquad \mathbf{z} \sim \mathcal{N}(0, \boldsymbol{\Pi}_{\mathbf{p}^\perp}), \qquad y = z_1z_2, \tag{113}$$

where $\mathbf{z}, z_1, z_2$ are independent, $a = s_1 = s_2$, and

$$\nu^2 = \sigma^2\left(2a^2 + \sum_{i\ge3} s_i^2\right), \qquad b = 2Pa + \sum_{i\ge3} s_i\zeta_i, \qquad \zeta_i = \mathbf{e}_p^\top\boldsymbol{\epsilon}_i.$$

Define

$$s_\gamma = e^\gamma/T, \qquad K_\gamma = s_\gamma/\nu_0, \qquad \nu_0 = \sigma/\sqrt{T}, \tag{114}$$

$$\bar{g}(\gamma) = \frac{\varphi(0)}{2}\mathbb{E}_{\mathrm{pos}}\frac{a^2}{\nu}, \qquad A(t) = \frac{\lambda}{\rho}\int_0^t \bar{g}(\gamma(v))\,dv, \qquad A_n = A(n\eta). \tag{115}$$

Let $\mathcal{E}_J$ denote the event of Proposition D.1, with $\mathbb{P}(\mathcal{E}_J^c) \le \delta_J$. On this event, for every $n \le N_{\mathrm{X}}$,

$$\|\mathbf{q}_n - \bar{\mathbf{q}}(n\eta)\|_2 \le d^{-15/16}, \qquad c \le \|\boldsymbol{\Pi}_{\mathbf{p}^\perp}\mathbf{u}_{j,n}\|_2 \le \|\mathbf{u}_{j,n}\|_2 \le C, \tag{116}$$

$$\max_j \|\boldsymbol{\Pi}_{\mathcal{N}^\perp}\mathbf{u}_{j,n}\|_2 \le \chi := C\sqrt{\ell_\delta/d}, \qquad \|\mathbf{W}_{\mathrm{X},n}\| \le Cd^{-1/2}. \tag{117}$$

The reference bounds used below are

$$-1 \le \gamma \le CG, \quad |\dot\gamma| \le CP^3, \quad \max_{t\le R+\eta} K_{\gamma(t)} \le CK_{\mathrm{X}}, \quad A(R+\eta) \le a_* < 1. \tag{118}$$

The fixed constant a_* is chosen small enough for the Gronwall amplification below to be at most two. The derivative bound follows directly from bounded reference rows: $\|\dot{\bar{\mathbf{q}}}\|_2 \le C\mathbb{E}\|D_{\mathbf{q}}\mathbf{h}\| \le CP^2$.

We record explicit versions of the position moments used here. For each fixed $1 \le r \le 64$, the moment estimates in Section E.5 and (62)–(64) give

$$\bar g(\gamma) \asymp s_\gamma^2/\nu_0, \qquad \mathbb{E}_{\mathrm{pos}}\left[\frac{a^2}{\nu}\left(1+\frac{a}{\nu}+\frac{|b|}{\nu}\right)^r\right] \le C\bar g(\gamma)L_d^{2r}, \tag{119}$$

$$\mathbb{E}_{\mathrm{pos}}\nu^2 \le C\nu_0^2, \qquad \mathbb{E}_{\mathrm{pos}}a^2 \le Cs_\gamma^2, \qquad \mathbb{E}_{\mathrm{pos}}b^2 \le C\nu_0^2L_d^4. \tag{120}$$

To see the powers, use $\nu \ge \nu_0$ and

$$\|a\|_{L^r} \le Cs_\gamma, \qquad \|b\|_{L^r} \le C(Ps_\gamma + \nu_0 + \sigma^2|\gamma|/P).$$

By Assumption 3.1 and (118), the normalized last bound is at most $C(1+PK_{\mathrm{X}}+\sigma\sqrt{T}\,G/P) \le CL_d^2$. Hölder's inequality, with the finitely many fixed moment orders enlarged as needed, proves (119). In the axial expansion only its case $r = 2$ will be used.

Uniformly on the attention-weight segments in (116),

$$\left(\mathbb{E}\sup_{\mathbf{q}'\in[\mathbf{q}_n,\bar{\mathbf{q}}(n\eta)]}\|D_{\mathbf{q}}\mathbf{h}(\mathbf{X},\mathbf{q}')\|^2\right)^{1/2} \le C\nu_0L_d^2,$$
$$\|\mathbb{E}\mathbf{h}(\mathbf{X},\mathbf{q}_n)\mathbf{h}(\mathbf{X},\mathbf{q}_n)^\top\|^{1/2} \le C\nu_0L_d^2. \tag{121}$$

Indeed Assumption 3.1, (118), and (103) give $(\sigma^2+P^2s_\gamma)/\nu_0 = \sigma\sqrt{T}+P^2K_\gamma \le CL_d^2$. The second estimate holds at the axis by (120); integrating the first estimate along the attention-weight segment extends it to $\mathbf{q}_n$. These are unconditional fresh-input moments. In particular, no pointwise assertion $\nu \asymp \nu_0$ is made for every position realization, and no probability estimate is conditioned on $\mathcal{E}_J$.

Use the same actual Gaussian seeds to define

$$\widetilde{\mathbf{u}}_{j,n} = e^{\epsilon_j A_n \mathbf{J}}\mathbf{u}_{j,0}, \qquad \mathbf{e}_{j,n} = \mathbf{u}_{j,n} - \widetilde{\mathbf{u}}_{j,n}, \qquad D_v(n) = \|\mathbf{E}_n\mathbf{\Pi}_v\|_{m,2} \quad (v = s, N, p). \tag{122}$$

Here $\widetilde{\mathbf{U}}_n$ and $\mathbf{E}_n$ have rows $\widetilde{\mathbf{u}}_{j,n}^\top$ and $\mathbf{e}_{j,n}^\top$, respectively, and $\mathbf{\Pi}_N = \mathbf{\Pi}_{\mathcal{N}}$. The initialization discrepancy is exactly zero. Since $\mathbf{J}\mathbf{e}_p = \mathbf{0}$, the reference keeps the initial position coordinates. Consequently, the position error has the exact original-coordinate expression

$$\|(\mathbf{W}_2(n) - \mathbf{W}_2(0))\mathbf{e}_p\|_2^2 = \frac{1}{m}\sum_j |\mathbf{e}_{j,n}^\top\mathbf{e}_p|^2 = D_p(n)^2. \tag{123}$$

Theorem H.1 (Anisotropic head tracking). *Under Assumption 3.1, the learning-rate choice in (15), and the resulting bounds (112), there is an event $\mathcal{E}_H$ with $\mathbb{P}(\mathcal{E}_H^c) \le \delta_H$ such that on $\mathcal{E}_J \cap \mathcal{E}_H$, simultaneously for $0 \le n \le N_{\mathrm{X}}$,*

$$D_s(n) + D_N(n) \le CA_n/d + Cd^{-11/8}, \tag{124}$$

$$D_p(n) \le Cd^{-11/8}, \tag{125}$$

$$\frac{\left\|f_{\mathrm{X},n}(\mathbf{X}) - F_{\widetilde{\mathbf{U}}_n}(\mathbf{h}(\mathbf{X},\bar{\mathbf{q}}(\eta n)))\right\|_{L^2(\mathcal{D}_{\mathrm{X}})}}{\nu_0} \le \frac{CA_n(1+K_{\gamma(\eta n)})}{d} + Cd^{-5/4}. \tag{126}$$

The constant multiplying A_n/d is independent of d. The event $\mathcal{E}_H$ consists of Gaussian-seed events, the characteristic output event of Lemma I.2, and a predictable stopped batch-noise event. In particular no batch martingale is conditioned on $\mathcal{E}_J$. Together with Proposition D.1 and (123), the position estimate gives Lemma 5.2 on the corresponding joint event; the failure allocations are combined in Appendix L.

H.1 Gaussian empirical moments with constant coefficients

Lemma H.2 (Weighted radius and product moments). *For the actual iid Gaussian seeds $\mathbf{u}_{j,0} \sim \mathcal{N}(0, \rho^2 \mathbf{I}_d/d)$, an event of probability at least $1 - \delta_H/4$ has the following properties simultaneously for every $A \in [0,1]$. Put $\mathbf{v}_j = e^{\epsilon_j A \mathbf{J}} \mathbf{u}_{j,0}$ and $r_j = \|\mathbf{\Pi}_{\mathbf{p}^\perp} \mathbf{v}_j\|_2$. Then*

$$c \le r_j \le \|\mathbf{v}_j\|_2 \le C, \qquad \max_j \|\mathbf{\Pi}_{\mathcal{N}^\perp} \mathbf{v}_j\|_2 \le C\sqrt{\ell_\delta/d}, \tag{127}$$

$$\big\| \|\mathbf{\Pi}_s \mathbf{v}_j\|_2 \, |r_j - \rho| \big\|_{m,2} \le C/d, \qquad \|v_{j,1} v_{j,2}\|_{m,2} \le C/d,$$

$$\big\| \|\mathbf{\Pi}_{\mathcal{N}^\perp} \mathbf{v}_j\|_2^3 \big\|_{m,2} \le C d^{-3/2}. \tag{128}$$

For scalar row-indexed quantities, $\|\cdot\|_{m,2}$ denotes the empirical root mean square. In particular, the coefficients multiplying d^{-1} in (128) do not depend on confidence.

Proof. Write $\mathbf{u}_0 = \rho \mathbf{z}/\sqrt{d}$ with $\mathbf{z} \sim \mathcal{N}(0, \mathbf{I}_d)$ and $r_0 = \|\mathbf{\Pi}_{\mathbf{p}^\perp} \mathbf{u}_0\|_2$. For every integer $k \ge 1$, Gaussian integration and the chi-square exponential moment give

$$\mathbb{E}\|\mathbf{\Pi}_{\mathcal{N}^\perp} \mathbf{u}_0\|_2^{2k} \le (Ck)^k d^{-k}, \qquad \mathbb{E}|r_0^2 - \rho^2|^{2k} \le (Ck)^{2k} d^{-k}. \tag{129}$$

Indeed, for $Y \sim \chi^2_{d-1}$,

$$\|Y - (d-1)\|_{L^{2k}} \le C(\sqrt{dk} + k), \qquad \|r_0^2 - \rho^2\|_{L^{2k}} \le Ck/\sqrt{d}.$$

Since $|r_0 - \rho| \le |r_0^2 - \rho^2|/\rho$, Hölder applied to

$$Y_1 = d^2 \|\mathbf{\Pi}_s \mathbf{u}_0\|_2^2 (r_0 - \rho)^2, \qquad Y_2 = d^2 \|\mathbf{\Pi}_s \mathbf{u}_0\|_2^4,$$
$$Y_3 = d^3 \|\mathbf{\Pi}_{\mathcal{N}^\perp} \mathbf{u}_0\|_2^6.$$

gives, for every integer $k \ge 1$,

$$\mathbb{E} Y_i^k \le (Ck)^{3k}, \qquad \mathbb{E} Y_i^2 \le C \quad (i = 1, 2, 3). \tag{130}$$

Taking k proportional to t in Markov's inequality proves $\mathbb{P}(Y_i > Ct^3) \le e^{-t}$ for $t \ge 1$.

Truncate the independent copies $Y_{i,j}$ at $C\ell_\delta^3$. Since Assumption 3.1 gives $m \le d^2$ through (112), the constant in this cutoff can be fixed so that any truncation has probability at most $\delta_H/32$. Their truncated means and second moments are bounded by universal constants. Scalar Bernstein therefore bounds each empirical average by

$$C + C\sqrt{\ell_\delta/m} + C\ell_\delta^4/m \le C$$

outside probability $\delta_H/32$. The last inequality follows from $m \ge C\ell_\delta^4$, a consequence of Assumption 3.1 through (112). Truncation only decreases the means; no tail bias is needed for this upper bound.

Gaussian coordinate tails give the center-coordinate maximum. Chi-square tails give a constant lower bound on the nuisance norm and a constant upper bound on the total norm. Their joint failure is at most Cme^{-cd} plus an arbitrarily small fixed allocation of δ_H from the coordinate maximum. Assumption 3.1, through Lemma A.3, makes $Cme^{-cd} \le \delta_H/32$. Allocate the finitely many events above within $\delta_H/4$.

The map $e^{\epsilon A \mathbf{J}}$ fixes $\mathcal{N}$ and $\mathbf{e}_p$, and has norm at most e on $\mathcal{S}$. Uniformly for $A \le 1$,

$$\|\mathbf{\Pi}_s \mathbf{v}_j\|_2 \le e\|\mathbf{\Pi}_s \mathbf{u}_{j,0}\|_2, \qquad |r_j^2 - r_{j,0}^2| \le CA\|\mathbf{\Pi}_s \mathbf{u}_{j,0}\|_2^2.$$

On the nuisance lower-norm event,

$$\|\mathbf{\Pi}_s \mathbf{v}_j\|_2 |r_j - \rho| \le C\|\mathbf{\Pi}_s \mathbf{u}_{j,0}\|_2 |r_{j,0} - \rho| + CA\|\mathbf{\Pi}_s \mathbf{u}_{j,0}\|_2^3.$$

The three empirical averages already controlled prove (128); also $|v_{j,1} v_{j,2}| \le \|\mathbf{\Pi}_s \mathbf{v}_j\|_2^2/2$. No net in A or replacement of an empirical product moment by a maximum-row bound is used. □

H.2 THE EXACT AXIAL FIELD

Lemma H.3 (Global axial expansion and cubic position force)**.** *For the conditional law* (113)*, define*

$$\mathbf{f}_{1/2}(\mathbf{u}) = \tfrac{1}{2}\mathbb{E}_{z,\mathbf{z}}[y\mathbf{h}_\gamma \mathbf{1}_{\{\mathbf{u}^\top \mathbf{h}_\gamma > 0\}}], \quad r = \left\|\mathbf{\Pi}_{\mathbf{p}^\perp}\mathbf{u}\right\|_2, \quad g = \frac{\varphi(0)a^2}{2\nu}.$$

If $c \le r \le \|\mathbf{u}\|_2 \le C$ *and* $\|\mathbf{\Pi}_{\mathcal{N}^\perp}\mathbf{u}\|_2 \le \chi \le 1$*, then*

$$\mathbf{f}_{1/2}(\mathbf{u}) = g\left(\frac{\mathbf{J}\mathbf{u}}{r} - \frac{u_1 u_2 \mathbf{\Pi}_{\mathbf{p}^\perp}\mathbf{u}}{r^3}\right) + \mathbf{r}_{\mathrm{ax}}(\mathbf{u}), \tag{131}$$

$$\|\mathbf{r}_{\mathrm{ax}}(\mathbf{u})\|_2 \le Cg\left(1 + a/\nu + |b|/\nu\right)^2 \chi^3, \tag{132}$$

$$|\mathbf{e}_p^\top \mathbf{f}_{1/2}(\mathbf{u})| \le Cg(b/\nu)^2 |u_p u_1 u_2|/r^3. \tag{133}$$

The bounds are global in a/ν *and* b/ν*.*

Proof. Put $\alpha = bu_p/(\nu r)$ and $v_i = au_i/(\nu r)$. Gaussian regression gives the exact identity

$$\mathbf{f}_{1/2}(\mathbf{u}) = \tfrac{1}{2}\mathbb{E}_z z_1 z_2 \big[(b\mathbf{e}_p + a(z_1\boldsymbol{\mu}_1 + z_2\boldsymbol{\mu}_2))\Phi(\alpha + v_1 z_1 + v_2 z_2) + \nu \mathbf{\Pi}_\perp \mathbf{u}\, \varphi(\alpha + v_1 z_1 + v_2 z_2)/r\big]. \tag{134}$$

The identities for one and two Rademacher differences are

$$\mathbb{E}_{z_i} z_i F(c + v_i z_i) = \frac{v_i}{2}\int_{-1}^{1} F'(c + v_i u)\, du,$$

$$\mathbb{E}_z z_1 z_2 F(\alpha + v_1 z_1 + v_2 z_2) = \frac{v_1 v_2}{4}\int_{[-1,1]^2} F''(\alpha + v_1 u + v_2 v)\, du\, dv.$$

Since φ and φ'' are even, their bounded second derivatives give, for every real x,

$$|\varphi(x) - \varphi(0)| \le Cx^2, \qquad |\varphi''(x) + \varphi(0)| \le Cx^2.$$

The two semantic coefficients of (134) thus equal $g\mathbf{J}\mathbf{u}/r$ with error at most

$$Cg\,\|\mathbf{\Pi}_s \mathbf{u}\|_2\,(\alpha^2 + v_1^2 + v_2^2).$$

The Gaussian-vector coefficient equals the displayed radial term, with error at most $Cg|u_1 u_2|(\alpha^2 + v_1^2 + v_2^2)$. These errors are bounded by the right side of (132).

For the position term, write $\Psi(\alpha) = \mathbb{E}_z z_1 z_2 \Phi(\alpha + v_1 z_1 + v_2 z_2)$. Pairing z with $-z$ and using $\Phi(x) + \Phi(-x) = 1$ gives $\Psi(0) = 0$. The double-difference identity applied to Φ' gives $|\Psi'(\alpha)| \le C|v_1 v_2|$, because Φ''' is bounded. Hence

$$|\mathbf{e}_p^\top \mathbf{f}_{1/2}(\mathbf{u})| = |b\Psi(\alpha)|/2 \le C|b\alpha v_1 v_2| = Cg(b/\nu)^2 |u_p u_1 u_2|/r^3.$$

This also bounds the position contribution to (132). The integral identities justify the estimates without any small-standardized-center restriction. □

Lemma H.4 (Axial drift discrepancy in empirical RMS)**.** *On the seed event of Lemma H.2, let* $\mathbf{v}_j = e^{\epsilon_j A\mathbf{J}}\mathbf{u}_{j,0}$*,* $A \le 1$*, and let* $\mathbf{u}_j$ *obey the row bounds in* (116)–(117)*. Put* $D = \|\mathbf{U} - \mathbf{V}\|_{m,2}$ *and* $D_s = \|(\mathbf{U} - \mathbf{V})\mathbf{\Pi}_s\|_{m,2}$*. Then*

$$\left\|\mathbb{E}_{\mathrm{pos}}\mathbf{f}_{1/2}(\mathbf{u}_j) - \frac{\bar{g}}{\rho}\mathbf{J}\mathbf{v}_j\right\|_{m,2} \le C\bar{g}\left(D + \frac{1}{d} + L_d^4 \chi^3\right), \tag{135}$$

$$\left\|\mathbf{e}_p^\top \mathbb{E}_{\mathrm{pos}}\mathbf{f}_{1/2}(\mathbf{u}_j)\right\|_{m,2} \le C\bar{g}L_d^4\chi^3. \tag{136}$$

Proof. Write $r_j = \left\|\mathbf{\Pi}_{\mathbf{p}^\perp}\mathbf{u}_j\right\|_2$ and $\widetilde{r}_j = \left\|\mathbf{\Pi}_{\mathbf{p}^\perp}\mathbf{v}_j\right\|_2$. The leading semantic term satisfies

$$\left\|\frac{\mathbf{J}\mathbf{u}_j}{r_j} - \frac{\mathbf{J}\mathbf{v}_j}{\rho}\right\|_{m,2} \le CD_s + C\chi D + C\left\|\|\mathbf{\Pi}_s\mathbf{v}_j\|_2\,|\widetilde{r}_j - \rho|\right\|_{m,2}$$
$$\le CD + C/d.$$

The first line follows by inserting $\mathbf{J}\mathbf{v}_j/r_j$ and $\mathbf{J}\mathbf{v}_j/\widetilde{r}_j$, and by $|r_j - \widetilde{r}_j| \le \|\mathbf{u}_j - \mathbf{v}_j\|_2$. For the radial term, the bounded row norms and the identity

$$u_{j,1}u_{j,2} - v_{j,1}v_{j,2} = (u_{j,1} - v_{j,1})u_{j,2} + v_{j,1}(u_{j,2} - v_{j,2})$$

give

$$\|u_{j,1}u_{j,2}\|_{m,2} \le C/d + C\chi D_s.$$

The two constants C/d are exactly the RMS moments from Lemma H.2. Average (132) over positions using (119); this costs only $C\bar{g}L_d^4\chi^3$. The position estimate follows directly from (133). The leading semantic and radial vectors both have zero position coordinate. □

H.3 Fresh-batch and discretization errors

Let $\mathcal{F}_n$ contain the initialization and all batches before iteration n. The scaled-row recursion (22) specializes to

$$\mathbf{u}_{j,n+1} = \mathbf{u}_{j,n} + \frac{\eta\lambda\epsilon_j}{b_{\mathrm{X}}} \sum_{(\mathbf{X},y)\in D_{X,n}} a_y(f_{\mathrm{X},n}(\mathbf{X}))\mathbf{h}(\mathbf{X},\mathbf{q}_n)\mathbf{1}_{\{\mathbf{u}_{j,n}^\top \mathbf{h}(\mathbf{X},\mathbf{q}_n)>0\}}. \tag{137}$$

The common learning rate in the original $\mathbf{W}$ coordinates produces precisely (137); no additional factor m or m^{-1} occurs.

Lemma H.5 (Uniform stopped row noise and characteristic Euler error)**.** *There is an event of probability at least $1 - \delta_H/4$ on which the stopped martingale in (137), in the row Hilbert norm $\|\cdot\|_{m,2}$, satisfies*

$$\max_{n\le N_{\mathrm{X}}} \|\mathbf{M}_n\|_{m,2} \le C\left(\sqrt{\frac{\eta\lambda\ell_\delta}{BP}} + \frac{\eta P\ell_\delta}{B}\right) \le d^{-23/16}. \tag{138}$$

On the seed event of Lemma H.2, moreover,

$$\sum_{n<N_{\mathrm{X}}} \left\|\widetilde{\mathbf{U}}_{n+1} - \widetilde{\mathbf{U}}_n - \eta\frac{\lambda}{\rho}\bar{g}(\gamma(n\eta))\big(\epsilon_j(\mathbf{J}\tilde{\mathbf{u}}_{j,n})^\top\big)_{j=1}^m\right\|_{m,2} \le d^{-2}, \tag{139}$$

$$\sup_{n\le N_{\mathrm{X}}} \left|\eta\frac{\lambda}{\rho}\sum_{l<n}\bar{g}(\gamma(l\eta)) - A_n\right| \le d^{-2}. \tag{140}$$

The same Euler estimate holds after orthogonal projection.

Proof. Stop before the first iteration at which a past-measurable bound in (116)–(117) fails, and at N_{X}. Write κ for this stopping time. The fresh-input estimates are uniform over each deterministic state in this tube, so they apply conditionally on $\mathcal{F}_n$ when $n < \kappa$. No estimate is conditioned on $\mathcal{E}_J$.

At the axial attention weight, (113) and (120) give

$$\begin{aligned}\mathbb{E}\|\mathbf{h}_\gamma\|_2^2 &= (d-1)\mathbb{E}_{\mathrm{pos}}\nu^2 + 2\mathbb{E}_{\mathrm{pos}}a^2 + \mathbb{E}_{\mathrm{pos}}b^2\\ &\le C\nu_0^2(d + L_d^4) \le Cd\nu_0^2.\end{aligned} \tag{141}$$

By (121),

$$\|\mathbf{h}(\mathbf{X},\mathbf{q}_n) - \mathbf{h}(\mathbf{X},\bar{\mathbf{q}}(n\eta))\|_{L^2} \le Cd^{-15/16}\nu_0 L_d^2.$$

Consequently

$$\mathbb{E}[\|\mathbf{h}(\mathbf{X},\mathbf{q}_n)\|_2^2 \mid \mathcal{F}_n] \le Cd\nu_0^2, \qquad n < \kappa. \tag{142}$$

Within each iteration reveal its $b_{\mathrm{X}} = \lambda B$ fresh XOR examples one at a time, keeping the state fixed. For one example let $\mathbf{Y}$ be the row array with jth row

$$\epsilon_j a_y(f_{\mathrm{X},n}(\mathbf{X}))\mathbf{h}(\mathbf{X},\mathbf{q}_n)^\top \mathbf{1}_{\{\mathbf{u}_{j,n}^\top \mathbf{h}(\mathbf{X},\mathbf{q}_n)>0\}}.$$

Then $\|\mathbf{Y}\|_{m,2} \le M_{\mathrm{tok}} := \max_i \|\mathbf{x}_i\|_2$ and its conditional second moment is at most $Cd\nu_0^2$. The centered microincrement has coefficient η/B, because $\eta\lambda/b_{\mathrm{X}} = \eta/B$. Writing $\mathbf{M}_n$ for their stopped sum at iteration boundaries, the conditional variance budget is bounded deterministically by

$$V \le CN_{\mathrm{X}}\lambda B\frac{\eta^2}{B^2}d\nu_0^2 = \frac{C\eta\lambda Rd\nu_0^2}{B} \le \frac{C\eta\lambda}{BP}. \tag{143}$$

The last step uses $d\sigma^2 \le 2$ from Assumption 3.1 and $R \le CT/P$ from Proposition D.1. Independence is used only between fresh examples, not between rows.

Set

$$p_0 = \frac{\delta_H}{C(N_{\mathrm{X}}B+1)^2}, \qquad M_\delta = C\left(P + 1 + \sigma\sqrt{d} + \sigma\sqrt{\log(T/p_0)}\right).$$

Gaussian token tails give $\mathbb{P}(M_{\mathrm{tok}} > M_\delta) \le p_0$. Assumption 3.1, the learning-rate range in (15), and $R \le CT/P$ from Proposition D.1 imply $\log(T/p_0) \le C\ell_\delta$, so $M_\delta \le CP$ by (112). For the proof only, truncate $\mathbf{Y}$ to $\mathbf{Y}\mathbf{1}_{\{M_{\mathrm{tok}} \le M_\delta\}}$ and recenter conditionally on the past. Its microincrements have norm at most $2\eta M_\delta/B$ and the same upper variance budget V. Bounded jumps and variance imply the moment hypothesis of Lemma C.3. With failure at most $\delta_H/32$, all partial sums of the truncated centered martingale are at most

$$C\left(\sqrt{V\log(C/\delta_H)} + \frac{\eta M_\delta}{B}\log(C/\delta_H)\right).$$

The probability that any revealed example was truncated is at most $N_{\mathrm{X}}Bp_0 \le \delta_H/32$. On the event none was truncated, the original centered martingale differs only by the predictable centering bias. Since the Gaussian token law and Assumption 3.1 give $\mathbb{E}M_{\mathrm{tok}}^2 \le CP^2$, every prefix of this bias has norm at most

$$C\lambda RP\sqrt{p_0} \le \frac{C\eta P\sqrt{\delta_H}}{B} \le \frac{C\eta P}{B}.$$

Thus the training algorithm is unchanged, and the original stopped martingale obeys the first inequality of (138). Allocate the displayed failures within $\delta_H/4$. For its second inequality, use $\eta \le d^{-3}$ from (15) and $\ell_\delta \le cd^{1/32}L_d^{-24}$ from Assumption 3.1 via Lemma A.3 to obtain

$$\sqrt{\frac{\eta\lambda\ell_\delta}{BP}} \le Cd^{-95/64}L_d^{-12}, \qquad \frac{\eta P\ell_\delta}{B} \le Cd^{-95/32}PL_d^{-24}.$$

Both are smaller than a fixed fraction of $d^{-23/16}$ by Lemma A.3.

For the deterministic Euler estimate put $M_\zeta = \max_{i\ge 3}|\zeta_i|$. The exact axial logits give

$$|\partial_\gamma \log a| \le 2(1 + M_\zeta/P), \qquad |\partial_\gamma \log \nu| \le 2(1 + M_\zeta/P).$$

The same fixed-order Gaussian maximum moments as in (119), together with Assumption 3.1, imply

$$|\bar{g}'(\gamma)| \le C\mathbb{E}_{\mathrm{pos}}\left[\frac{a^2}{\nu}(1 + M_\zeta/P)\right] \le C\nu_0 K_{\mathrm{X}}^2 \le C\nu_0 L_d^2.$$

Together with $|\dot{\gamma}| \le CP^3$, this gives $|A'| + |A''| \le C\nu_0 L_d^4$. Twice differentiating the characteristic yields

$$\widetilde{\mathbf{u}}_j'' = [\epsilon_j A''\mathbf{J} + (A')^2\mathbf{J}^2]\widetilde{\mathbf{u}}_j.$$

The seed event gives $\|\widetilde{\mathbf{U}}\|_{m,2} \le C$, and hence $\sup_{t\le R+\eta}\|\widetilde{\mathbf{U}}''(t)\|_{m,2} \le C\nu_0 L_d^4$. Taylor's integral formula bounds the cumulative Euler error by

$$C\eta R\nu_0 L_d^4 \le C\eta L_d^5 \le d^{-2},$$

Here $R \le CT/P$ from Proposition D.1 and Assumption 3.1 give $R\nu_0 \le C\sigma\sqrt{T}/P \le CL_d$; the final bound also uses $\eta \le d^{-3}$ from (15). The same estimate bounds the clock quadrature error using A'', which proves (139)–(140). Orthogonal projection cannot increase the error. Almost-everywhere derivatives and their integral representations suffice here. □

H.4 Logistic output feedback and closure

Lemma H.6 (Prediction and logistic residual bounds)**.** *For arbitrary deterministic row arrays* $\mathbf{U}, \mathbf{V}$*, set* $\mathbf{E} = \mathbf{U} - \mathbf{V}$*, write* $\mathbf{e}_j^\top$ *for its* j*th row, and define* $D_v = \|\mathbf{E}\mathbf{\Pi}_v\|_{m,2}$ *for* $v = s, N, p$*. The conditional law* (113) *satisfies*

$$\frac{\|F_{\mathbf{U}}(\mathbf{h}_\gamma) - F_{\mathbf{V}}(\mathbf{h}_\gamma)\|^2_{L^2(z,\mathbf{z})}}{\nu^2} \le D_N^2 + (1 + a^2/\nu^2)D_s^2 + (b^2/\nu^2)D_p^2. \tag{144}$$

After position averaging,

$$\frac{\|F_{\mathbf{U}}(\mathbf{h}_\gamma) - F_{\mathbf{V}}(\mathbf{h}_\gamma)\|_{L^2(\mathcal{D}_{\mathrm{X}})}}{\nu_0} \le CD_N + C(1 + K_\gamma)D_s + CL_d^2 D_p. \tag{145}$$

For the actual trained rows on $\mathcal{E}_J$*,*

$$\frac{\|F_{\mathbf{U}_n}(\mathbf{h}(\mathbf{X}, \mathbf{q}_n)) - F_{\mathbf{U}_n}(\mathbf{h}(\mathbf{X}, \bar{\mathbf{q}}(\eta n)))\|_{L^2}}{\nu_0} \le Cd^{-23/16}L_d^2. \tag{146}$$

Finally, for any fixed row gate $g_j(\mathbf{X}) \in \{0, 1\}$*,*

$$\left\| \frac{1}{2}\mathbb{E}[\mathbf{h}(\mathbf{X}, \mathbf{q}_n) g_j(\mathbf{X}) \tanh(f_{\mathrm{X},n}(\mathbf{X})/2)] \right\|_2 \le \frac{1}{4} \left\|\mathbb{E}\mathbf{h}\mathbf{h}^\top\right\|^{1/2} \|f_{\mathrm{X},n}\|_{L^2}. \tag{147}$$

Proof. ReLU Lipschitzness and Jensen's inequality imply $|F_{\mathbf{U}}(\mathbf{h}) - F_{\mathbf{V}}(\mathbf{h})|^2 \le m^{-1}\sum_j |(\mathbf{u}_j - \mathbf{v}_j)^\top \mathbf{h}|^2$. For each fixed row difference, the expectation of its squared inner product under (113) is exactly

$$\nu^2 \|\mathbf{\Pi}_{\mathcal{N}}\mathbf{e}_j\|_2^2 + (\nu^2 + a^2)\|\mathbf{\Pi}_s\mathbf{e}_j\|_2^2 + b^2|\mathbf{e}_p^\top \mathbf{e}_j|^2.$$

This proves (144); use (120) and the triangle inequality for (145).

For attention-weight replacement, Cauchy–Schwarz over rows gives pointwise $|F_{\mathbf{U}_n}(\mathbf{h}) - F_{\mathbf{U}_n}(\mathbf{h}')| \le \|\mathbf{W}_{\mathrm{X},n}\|\,\|\mathbf{h} - \mathbf{h}'\|_2$. Integrate $D_{\mathbf{q}}\mathbf{h}$ on the attention-weight segment and use (116), (117), and (121). The product of the two dimension factors is $d^{-15/16}d^{-1/2} = d^{-23/16}$.

For a unit vector $\mathbf{v}$, scalar Cauchy–Schwarz gives

$$|\mathbb{E}[(\mathbf{v}^\top \mathbf{h}) g_j \tanh(f/2)]| \le \sqrt{\mathbb{E}(\mathbf{v}^\top \mathbf{h})^2}\,\|\tanh(f/2)\|_{L^2} \le \tfrac{1}{2}\left\|\mathbb{E}\mathbf{h}\mathbf{h}^\top\right\|^{1/2}\|f\|_{L^2}.$$

Take the supremum over $\mathbf{v}$ and include the coefficient $1/2$. The covariance operator is used here; replacing it by its trace would lose a factor $\sqrt{d}$. □

Proof of Theorem H.1. Intersect the seed event of Lemma H.2, the stopped-noise event of Lemma H.5, and the characteristic-output event of Lemma I.2, using failure allocation $\delta_H/4$ for each. Their intersection is $\mathcal{E}_H$ and has failure at most δ_H. Only afterwards intersect with $\mathcal{E}_J$. The stopped martingale then equals the actual martingale through the required horizon.

Define finite prefix maxima

$$\mathcal{M}_n = \max_{l \le n} \frac{\|F_{\mathbf{U}_l}(\mathbf{h}(\mathbf{X}, \bar{\mathbf{q}}(l\eta)))\|_{L^2}}{\nu_0}, \qquad \mathcal{D}_n = \max_{l \le n} \|\mathbf{E}_l\|_{m,2}.$$

The identity $a_y(f) = y/2 - \frac{1}{2}\tanh(f/2)$ splits the conditional mean update into the half-label field and the logistic residual. First apply Lemma G.1 to replace the actual attention weight by the axial reference. By Assumption 3.1 and (117), $\chi + \sigma + \nu_0 \le C\sqrt{\ell_\delta/d}$; the clock and its Riemann estimate therefore give, for every row and prefix,

$$\begin{aligned}
&\eta\lambda \sum_{l<n} \|\mathbf{f}_{1/2}(\mathbf{u}_{j,l}, \mathbf{q}_l) - \mathbf{f}_{1/2}(\mathbf{u}_{j,l}, \bar{\mathbf{q}}(l\eta))\|_2 \\
&\quad \le Cd^{-15/16}L_d^{12}(\chi + \sigma + \nu_0)\eta\lambda \sum_{l<n} s^2_{\gamma(l\eta)}/\nu_0 \\
&\quad \le Cd^{-23/16}L_d^{12}\sqrt{\ell_\delta}.
\end{aligned} \tag{148}$$

The term s_γ^2 in that lemma is smaller than s_γ^2/ν_0 by a factor ν_0. Comparability in (119) and (140) bound the clock sum by a universal constant.

Using $d\sigma^2 \le 2$ and the task-proportion bounds in Assumption 3.1, together with $R \le CT/P$ from Proposition D.1, set

$$q_H = Cd^{-23/16}L_d^2, \qquad \varepsilon_H = C\lambda R\nu_0^2 L_d^2 \le \frac{CL_d^2}{dP}.$$

Equations (146), (121), and (147) give the cumulative full-row logistic residual bound

$$C\eta\lambda \sum_{l<n} \nu_0^2 L_d^2(\mathcal{M}_n + q_H) \le \varepsilon_H \mathcal{M}_n + d^{-2}. \tag{149}$$

Indeed $\varepsilon_H q_H \le Cd^{-39/16}L_d^4/P \le d^{-2}$ by Lemma A.3. The same bound holds after position projection.

Subtract the characteristic Euler recursion from (137) before taking norms. Put $\alpha_l = \eta\lambda\bar{g}(\gamma(l\eta))/\rho$ and define

$$b_H = C\left[d^{-23/16}L_d^{12}\sqrt{\ell_\delta} + d^{-3/2}L_d^4\ell_\delta^{3/2} + d^{-2}\right].$$

Assumption 3.1, through Lemma A.3, implies the useful bound

$$b_H \le Cd^{-91/64}. \tag{150}$$

For its first term use $L_d^{12}\sqrt{\ell_\delta} \le cd^{1/64}$. For the second use $L_d^4\ell_\delta^{3/2} \le cd^{3/64}L_d^{-32}$. These estimates retain the confidence logarithm explicitly.

Lemmas H.4 and H.5, together with (148)–(149), give

$$\|\mathbf{E}_n\|_{m,2} \le C\sum_{l<n}\alpha_l\|\mathbf{E}_l\|_{m,2} + \frac{C}{d}\sum_{l<n}\alpha_l + b_H + \varepsilon_H\mathcal{M}_n, \tag{151}$$

$$D_p(n) \le b_H + \varepsilon_H\mathcal{M}_n. \tag{152}$$

The axial remainder has size $CL_d^4\chi^3 \le Cd^{-3/2}L_d^4\ell_\delta^{3/2}$ after its bounded clock integral. The martingale is at most $d^{-23/16}$ and the Euler error at most d^{-2}, both included in b_H. The leading C/d coefficient is the product RMS coefficient in Lemma H.2; it carries no factor L_d or ℓ_δ.

The clock quadrature estimate gives $\sum_{l<n}\alpha_l \le A_n + d^{-2}$. Choose a_* once so that

$$\exp\left(C\sum_{l<N_{\mathrm{X}}}\alpha_l\right) \le \exp(Ca_* + Cd^{-2}) \le 2$$

by the choice of a_* and the dimension consequences of Assumption 3.1 in Lemma A.3. For each fixed prefix, $\mathcal{M}_n$ is a finite constant upper bound for its residual budgets. Discrete Gronwall therefore yields

$$\|\mathbf{E}_n\|_{m,2} \le CA_n/d + Cb_H + C\varepsilon_H\mathcal{M}_n, \tag{153}$$

$$D_p(n) \le Cb_H + C\varepsilon_H\mathcal{M}_n. \tag{154}$$

We close the output budget using the full-law estimate (174). Take its characteristic-clock endpoint to be a_* and its semantic endpoint to be CK_{X}. Assumption 3.1, via Lemma A.3, bounds the position-center envelope by

$$R_{\mathrm{pos}} = C(PK_{\mathrm{X}} + \sqrt{\log(2T)} + G\sigma\sqrt{T}/P) \le CL_d^2.$$

Thus $R_{\mathrm{pos}}^2 + K_{\mathrm{X}}^2 \le d$, its grid logarithm is at most $C\ell_\delta$, and its original-position tail satisfies $P\nu_0^{-1}T^{-11/2} \le Cd^{-4}$. Lemma I.2 supplies, on its allocated initialization event,

$$\max_{l\le N_{\mathrm{X}}} \frac{\|F_{\widetilde{\mathbf{U}}_l}(\mathbf{h}(\mathbf{X},\bar{\mathbf{q}}(l\eta)))\|_{L^2}}{\nu_0} \le Q_H,$$

$$Q_H = C\left[\frac{a_*(1+K_{\mathrm{X}}^2)}{d} + \sqrt{\frac{\ell_\delta}{m}} + d^{-3} + P\nu_0^{-1}T^{-11/2}\right]$$

$$\le C\left[\frac{L_d^2}{d} + d^{-121/128}L_d^{-6}\right]. \tag{155}$$

The last line uses the lower width in Assumption 3.1, through (112). In particular the empirical fluctuation is kept as $d^{-121/128}L_d^{-6}$; it is not replaced by d^{-1} times a fixed logarithmic power.

Apply the anisotropic estimate (145) to (153)–(154), and take the maximum over the prefix. This gives

$$\mathcal{M}_n \le Q_H + \frac{Ca_*(1+K_{\mathrm{X}})}{d} + CL_d^2 b_H + CL_d^2 \varepsilon_H \mathcal{M}_n.$$

Here Assumption 3.1 and Lemma A.3 give $CL_d^2\varepsilon_H \le CL_d^4/(dP) \le 1/2$. Every $\mathcal{M}_n$ is finite, so direct algebra, without an additional stop, gives

$$\mathcal{M}_n \le C\left[\frac{L_d^2}{d} + d^{-121/128}L_d^{-6}\right]. \tag{156}$$

Substituting this estimate in the residual yields

$$\varepsilon_H \mathcal{M}_n \le C\left[\frac{L_d^4}{d^2 P} + \frac{d^{-249/128}L_d^{-4}}{P}\right] \le Cd^{-3/2}.$$

Together with (150), equations (153)–(154) prove (124)–(125), using $D_s + D_N \le \sqrt{2}\|\mathbf{E}_n\|_{m,2}$ and $d^{-91/64} \le d^{-11/8}$.

Finally, keeping the local clock in the anisotropic prediction bound,

$$\frac{\|F_{\mathbf{U}_n}(\mathbf{h}_{\gamma_n}) - F_{\widetilde{\mathbf{U}}_n}(\mathbf{h}_{\gamma_n})\|_{L^2}}{\nu_0} \le \frac{CA_n(1+K_{\gamma_n})}{d} + CL_d^2(b_H + \varepsilon_H \mathcal{M}_n).$$

The last term, together with the attention-weight replacement q_H, is at most $Cd^{-5/4}$ by the displayed dimension powers. This proves (126), uniformly at every iteration including the entire early prefix. □

I Classification along the characteristic trajectory

Write $\mathbf{v}_j = \mathbf{u}_{\mathrm{X},j,0} \sim \mathcal{N}(\mathbf{0}, \rho^2 \mathbf{I}_d/d)$ and let ϵ_j be the independent Rademacher readout signs, independent of these Gaussian rows. The current initialization has $\rho = 1$; retaining ρ makes the output scale explicit. Define

$$\widetilde{F}_A(\mathbf{h}) = \frac{1}{m}\sum_{j=1}^{m} \epsilon_j [(e^{\epsilon_j A \mathbf{J}}\mathbf{v}_j)^\top \mathbf{h}]_+, \qquad F_0 = \widetilde{F}_0, \qquad I_0 = \rho/\sqrt{m}. \tag{157}$$

All initialization events below are proved for this original joint law. They may subsequently be intersected with the trajectory events; their probabilities are not conditioned on successful training or on the other head. Every constant C, c in this section is universal.

I.1 An exact common test family and a position event

Set $N_D = T-2$, $\tau_p = \sigma/P$, and let $v_i \sim \mathcal{N}(0, \tau_p^2)$, $3 \le i \le T$, be the distractor position coordinates divided by P. At $\mathbf{q}_\gamma = \gamma \mathbf{e}_p/P$, put

$$\begin{aligned} D_\gamma &= 2e^{2\gamma} + \sum_{i=3}^{T} e^{2\gamma v_i}, & L_\gamma &= 2e^{\gamma} + \sum_{i=3}^{T} e^{\gamma v_i}, \\ \nu_\gamma &= \frac{\sigma\sqrt{D_\gamma}}{L_\gamma}, & k_\gamma &= \frac{e^\gamma}{\sigma\sqrt{D_\gamma}}, \\ u_\gamma &= \frac{P(2e^\gamma + \sum_{i=3}^{T} v_i e^{\gamma v_i})}{\sigma\sqrt{D_\gamma}}. \end{aligned} \tag{158}$$

Conditional on $\mathbf{v}_{\mathrm{pos}} = (v_i)_{i\ge 3}$, the normalized axial XOR pool has the exact law

$$\mathbf{v}_\gamma = \mathbf{g}_\perp + u_\gamma \mathbf{e}_p + k_\gamma(z_1\boldsymbol{\mu}_1 + z_2\boldsymbol{\mu}_2), \qquad y = z_1 z_2, \qquad \mathbf{g}_\perp \sim \mathcal{N}(\mathbf{0}, \boldsymbol{\Pi}_{\mathbf{p}^\perp}). \tag{159}$$

The Gaussian, semantic signs, and position variables are independent. Using the same triple at every γ preserves each fresh-population marginal. Cauchy–Schwarz applied to the T attention weights gives

$$\nu_\gamma \geq \nu_0 = \sigma/\sqrt{T} \quad \text{for every position realization and every } \gamma. \tag{160}$$

Fix $M = 22$ and set $b_T = \sqrt{2(M+3)\log(2T)}$. The fresh position event is

$$\mathcal{P} = \left\{ \max_i |v_i| \leq \tau_p b_T, \quad \sum_i v_i^2 \leq 4N_D\tau_p^2, \quad |\sum_i v_i| \leq \tau_p\sqrt{N_D}\, b_T \right\}. \tag{161}$$

For T above a universal constant, Gaussian tails and $\mathbb{E}e^{\|\mathbf{z}\|_2^2/4} = 2^{N_D/2}$ for $\mathbf{z} \sim \mathcal{N}(\mathbf{0}, \mathbf{I}_{N_D})$ give $\mathbb{P}(\mathcal{P}^c) \leq CT^{-22}$. For the following finite bounds it suffices that

$$\Gamma \geq 1, \qquad \Gamma\tau_p b_T \leq c, \qquad e^{2\Gamma}/T \leq c. \tag{162}$$

Define

$$K_* = \max\{1, Ce^\Gamma/(\sigma\sqrt{T})\}, \qquad \alpha = \sigma\sqrt{T}/P, \qquad R = C(PK_* + b_T + \Gamma\alpha) \geq 1. \tag{163}$$

On $\mathcal{P}$, uniformly for $-1 \leq \gamma \leq \Gamma$,

$$\nu_0 \leq \nu_\gamma \leq C\nu_0, \qquad c\frac{e^\gamma}{\sigma\sqrt{T}} \leq k_\gamma \leq C\frac{e^\gamma}{\sigma\sqrt{T}} \leq K_*, \qquad |u_\gamma| \leq R. \tag{164}$$

Indeed the distractor exponentials are between fixed positive constants, so $D_\gamma, L_\gamma \asymp T$, and

$$\left|\sum_i v_i e^{\gamma v_i}\right| \leq \left|\sum_i v_i\right| + C|\gamma|\sum_i v_i^2.$$

Differentiating the exact formulas gives

$$\begin{aligned} \left|D_\gamma'/(2D_\gamma)\right| &\leq C(\sigma^2K_*^2 + \tau_p b_T) \leq C, \\ |k_\gamma'| &\leq CK_*, \qquad\qquad |u_\gamma'| \leq C(PK_* + \alpha + R). \end{aligned} \tag{165}$$

For the last bound, the derivative of $2e^\gamma + \sum_i v_i e^{\gamma v_i}$ is bounded by $C(e^\Gamma + T\tau_p^2)$; differentiating its denominator contributes $C|u_\gamma|$. The segment from $\mathbf{g}_\perp$ to $\mathbf{v}_0$, followed by the path over $[-1, \Gamma]$, consequently has total length at most

$$R_{\text{len}} = C(1+\Gamma)(PK_* + \alpha + R + K_*). \tag{166}$$

Assumption 3.1 implies (162) for $\Gamma \leq \Gamma_{K_X} + 2$ by Lemma A.3. The event $\mathcal{P}$ concerns a fresh example; its failure is not an initialization or training confidence allocation.

I.2 Mean and uniform second moment of the characteristic

Lemma I.1 (Exact characteristic mean)**.** *For $0 \le A \le 1$ and $\mathbf{h} \neq \mathbf{0}$, write $h_i = \boldsymbol{\mu}_i^\top \mathbf{h}$. Then*

$$\begin{aligned}\overline{F}_A(\mathbf{h}) := \mathbb{E}_{\epsilon,\mathbf{V}} \widetilde{F}_A(\mathbf{h}) &= \frac{\rho}{2\sqrt{2\pi d}}(\|e^{A\mathbf{J}}\mathbf{h}\|_2 - \|e^{-A\mathbf{J}}\mathbf{h}\|_2) \\ &= \frac{2\rho \sinh(2A) h_1 h_2}{\sqrt{2\pi d}(\|e^{A\mathbf{J}}\mathbf{h}\|_2 + \|e^{-A\mathbf{J}}\mathbf{h}\|_2)}. \end{aligned} \tag{167}$$

Both sides are zero at $\mathbf{h} = \mathbf{0}$. For $A > 0$ the mean has the sign of $h_1 h_2$, and

$$\frac{c\rho A |h_1 h_2|}{\sqrt{d}\|\mathbf{h}\|_2} \le |\overline{F}_A(\mathbf{h})| \le \frac{C\rho A |h_1 h_2|}{\sqrt{d}\|\mathbf{h}\|_2}. \tag{168}$$

For $d \ge 6$, $A_* \le 1$, $K_* \ge 1$, and $\mathbf{h}(u,k,z) = \mathbf{g}_\perp + u\mathbf{e}_p + k(z_1\boldsymbol{\mu}_1 + z_2\boldsymbol{\mu}_2)$,

$$\left\| \sup_{\substack{0 \le A \le A_*,\ u \in \mathbb{R} \\ 0 \le k \le K_*,\ z \in \{-1,1\}^2}} |\overline{F}_A(\mathbf{h}(u,k,z))| \right\|_{L^2(\mathbf{g}_\perp)} \le \frac{C\rho A_*(1+K_*^2)}{d}. \tag{169}$$

Proof. The positive-part mean of a centered Gaussian of standard deviation s is $s/\sqrt{2\pi}$. Average the two readout signs and use

$$\|e^{\pm A\mathbf{J}}\mathbf{h}\|_2^2 = \|\mathbf{h}\|_2^2 + (\cosh(2A) - 1)(h_1^2 + h_2^2) \pm 2\sinh(2A) h_1 h_2.$$

This proves (167); the eigenvalues of $e^{\pm A\mathbf{J}}$ lie in $[e^{-1}, e]$, proving (168). For the stronger uniform bound, its numerator is at most $CA_*(\|\mathbf{\Pi}_{\mathcal{S}}\mathbf{g}_\perp\|_2 + K_*)^2$, while its denominator contains $\|\mathbf{\Pi}_{\mathcal{N}}\mathbf{g}_\perp\|_2$. The two projections are independent and direct integration of the chi-square density gives

$$\mathbb{E}\|\mathbf{\Pi}_{\mathcal{N}}\mathbf{g}_\perp\|_2^{-2} = \frac{1}{d-5}, \qquad \mathbb{E}(\|\mathbf{\Pi}_{\mathcal{S}}\mathbf{g}_\perp\|_2 + K_*)^4 \le C(1+K_*^4).$$

Their product proves (169). □

We use an elementary all-order concentration bound. If independent, centered scalar variables satisfy $\mathbb{E}e^{X_j^2/s^2} \le 2$, then

$$\mathbb{P}\left\{ \left| m^{-1} \sum_j X_j \right| > t \right\} \le 2e^{-cmt^2/s^2}. \tag{170}$$

To verify it, symmetrize with independent copies X_j' by Jensen. The exponential-square assumption bounds the even moments of $X_j - X_j'$ by $2q!(Cs^2)^q$. Expanding its hyperbolic cosine and using $(2q)! \ge (q!)^2$ gives $\mathbb{E}e^{t(X_j - X_j')} \le e^{Cs^2t^2}$, after enlarging C. Multiply the moment generating functions and optimize. Consequently a union bound and tail integration, for every real $r \ge 1$, give

$$\left\| \max_{1 \le a \le N_0} \left| m^{-1} \sum_j X_{j,a} \right| \right\|_{L^r} \le Cs\sqrt{\frac{r + \log(2N_0)}{m}}. \tag{171}$$

Independence is required across j, not across the grid index a. For example this follows by splitting the tail integral at $Cs\sqrt{\log(2N_0)/m}$ and integrating the remaining Gaussian tail. These statements hold at all orders; they do not use the fixed-order moment convention for the token estimates.

For $A_* \le 1$ and $R, K_* \ge 1$, choose an endpoint-containing rectangular grid for (A, u, k, z) and define

$$\Delta = \frac{d^{-3}}{1 + R + K_*}, \qquad N_0 \le 4(2 + A_*/\Delta)(2 + 2R/\Delta)(2 + K_*/\Delta), \qquad \ell_0 = r + \log(2N_0). \tag{172}$$

The factor 4 includes all semantic sign pairs.

Lemma I.2 (Uniform characteristic output budget). *Let $d \geq 6$, $A_* \leq 1$, $R, K_* \geq 1$, and $r \geq 2$. Outside an initialization event of probability at most e^{-r},*

$$\left\| \sup_{\substack{0\leq A\leq A_* \\ |u|\leq R,\ 0\leq k\leq K_*,\ z}} |\widetilde{F}_A(\mathbf{h}(u,k,z))| \right\|_{L^2(\mathbf{g}_\perp)} \leq C\rho \left\{ \frac{A_*(1+K_*^2)}{d} + \left(1+\frac{R+K_*}{\sqrt{d}}\right)\sqrt{\frac{\ell_0}{m}} + \Delta(1+K_*)\left(1+\sqrt{\frac{r}{md}}\right)\right\}. \tag{173}$$

Suppose also (162), (163), $P \geq 1$, $d\sigma^2 \leq 2$, and $\sigma^2 \log(2T) \leq P^2$. On an event of probability at least $1 - e^{-r} - Ce^{-cmd}$, the original axial input law satisfies

$$\sup_{\substack{0\leq A\leq A_* \\ -1\leq\gamma\leq\Gamma}} \frac{\|\widetilde{F}_A(\mathbf{h}(\mathbf{X},\mathbf{q}_\gamma))\|_{L^2(\mathbf{X})}}{\nu_0} \leq Q,$$

$$Q := C\rho\left\{ \frac{A_*(1+K_*^2)}{d} + \left(1+\frac{R+K_*}{\sqrt{d}}\right)\sqrt{\frac{\ell_0}{m}} + \Delta(1+K_*)\left(1+\sqrt{\frac{r}{md}}\right) + P\nu_0^{-1}T^{-11/2}\right\}. \tag{174}$$

In particular the failure is at most ε if $r = \log(4/\varepsilon)$ and $Ce^{-cmd} \leq \varepsilon/2$.

Proof. For fixed $\mathbf{g}_\perp$, a centered row output at any grid point has exponential-square parameter $s(\mathbf{g}_\perp) = C\rho(\|\mathbf{g}_\perp\|_2 + R + K_*)/\sqrt{d}$. Conditional on either readout sign its underlying linear functional is Gaussian with that variance bound; taking a positive part and centering only changes the universal constant. Thus (171) and Minkowski with exponent $r/2 \geq 1$ yield

$$\begin{aligned} &\left\| \left\| \max_{\text{grid}} |\widetilde{F}_A(\mathbf{h}) - \overline{F}_A(\mathbf{h})| \right\|_{L^2(\mathbf{g}_\perp)} \right\|_{L^r(\epsilon,\mathbf{V})} \\ &\leq \left(\mathbb{E}_{\mathbf{g}_\perp} \left\| \max_{\text{grid}} |\widetilde{F}_A(\mathbf{h}) - \overline{F}_A(\mathbf{h})| \right\|^2_{L^r(\epsilon,\mathbf{V})} \right)^{1/2} \\ &\leq C\rho\left(1+\frac{R+K_*}{\sqrt{d}}\right)\sqrt{\ell_0/m}. \end{aligned} \tag{175}$$

The supremum is inside the Gaussian integral. For interpolation, changing u, k changes a row output by at most $C\|\mathbf{v}_j\|_2(|\Delta u| + |\Delta k|)$; changing A changes it by at most

$$C\|\mathbf{\Pi}_{\mathcal{S}}\mathbf{v}_j\|_2(\|\mathbf{\Pi}_{\mathcal{S}}\mathbf{g}_\perp\|_2 + K_*)|\Delta A|.$$

The means obey the same inequalities after taking expectation. Therefore the integrated $L^2(\mathbf{g}_\perp)$ interpolation error is at most $C\Delta(1+K_*)(\bar{V}+\rho)$, pointwise in initialization, where

$$\bar{V} = m^{-1}\sum_j \|\mathbf{v}_j\|_2 \leq \left(m^{-1}\sum_j \|\mathbf{v}_j\|_2^2\right)^{1/2} = \frac{\rho}{\sqrt{md}}\chi_{md}, \qquad \|\bar{V}\|_{L^r} \leq C\rho(1+\sqrt{r/(md)}).$$

The last estimate follows by integrating the chi-square exponential-moment tail $\mathbb{P}(\chi_n > \sqrt{n}+\sqrt{2t}) \leq e^{-t}$; it holds for every $r \geq 1$. Add this estimate to (175) and apply the r-th moment inequality at e times its bound. Lemma I.1 proves (173).

On $\mathcal{P}$, homogeneity and (164) transfer this same event to the random position-dependent parameters u_γ, k_γ. This is valid because the supremum was inside $L^2(\mathbf{g}_\perp)$. For the complement, the Gaussian seed norm event $m^{-1}\sum_j \|\mathbf{v}_j\|_2^2 \leq C\rho^2$ fails with probability at most Ce^{-cmd}. On this event $|\widetilde{F}_A(\mathbf{h})|^2 \leq C\rho^2\|\mathbf{h}\|_2^2$ for every $A \leq 1$. Conditional on positions, the original pool obeys

$$\mathbb{E}_{\mathbf{g}_\perp,z}\|\mathbf{h}(\mathbf{X},\mathbf{q}_\gamma)\|_2^2 \leq C\{(P + P\max_i |v_i|)^2 + 1 + d\sigma^2\} =: \mathcal{H}_{\text{pos}}, \qquad \mathbb{E}\mathcal{H}_{\text{pos}}^2 \leq CP^4.$$

Indeed the weights are nonnegative and sum to one, while $\mathbb{E}\max_i |v_i|^4 \le C\tau_p^4 \log^2(2T)$. Uniformly in A, γ, Cauchy–Schwarz consequently gives

$$\frac{\|\mathbf{1}_{\mathcal{P}^c}\widetilde{F}_A(\mathbf{h}(\mathbf{X},\mathbf{q}_\gamma))\|_{L^2}}{\nu_0} \le C\rho P\nu_0^{-1}\mathbb{P}(\mathcal{P}^c)^{1/4} \le C\rho P\nu_0^{-1}T^{-11/2}.$$

This is an integrated fresh-input tail, not a seed failure probability. Adding the two parts proves (174). □

I.3 Initialization anti-concentration and the entire early prefix

Lemma I.3 (A label-independent initialization baseline)**.** *There are universal $a_0, c, C > 0$ such that, for $m \ge d \ge C$, $B_0 = F_0(\mathbf{g}_\perp)$ satisfies*

$$\mathbb{P}_{\epsilon,\mathbf{V}}\{\mathbb{P}_{\mathbf{g}_\perp}(|B_0| \le a_0 I_0) > 1/16\} \le Ce^{-cd}. \tag{176}$$

This assertion uses iid random readout signs. Moreover, for every $r \ge 1$, an initialization event of failure at most e^{-r} satisfies

$$\mathbb{E}_{\mathbf{g}_\perp,\mathbf{v}_{\mathrm{pos}},z}\left[\mathbf{1}_{\mathcal{P}}\sup_{-1\le\gamma\le\Gamma}|F_0(\mathbf{v}_\gamma) - B_0|\right] \le CI_0R_{\mathrm{len}}\sqrt{r/d}. \tag{177}$$

For every fixed initialization B_0 is independent of the balanced fresh label $y = z_1z_2$.

Proof. Identify $\mathbf{p}^\perp$ with $\mathbb{R}^{d-1}$ and set $\mathbf{z}_j = \epsilon_j\mathbf{\Pi}_{\mathbf{p}^\perp}\mathbf{v}_j$, $\mathbf{s} = \sum_j \mathbf{z}_j$. These Gaussian rows are independent of the entire sign array. Let $\mathbf{Z}$ have rows $\mathbf{z}_j^\top$. Exactly,

$$F_0(\mathbf{g}) = \frac{\mathbf{s}^\top\mathbf{g}}{2m} + \mathcal{R}(\mathbf{g}), \qquad \mathcal{R}(\mathbf{g}) = \frac{1}{2m}\sum_j \epsilon_j|\mathbf{z}_j^\top\mathbf{g}|.$$

Except with probability Ce^{-cd},

$$\|\mathbf{s}\|_2 \ge \rho\sqrt{m}/2, \qquad \|\mathbf{Z}\|^2 \le C\rho^2 m/d.$$

The first bound is a chi-square lower tail in dimension $d-1$. For the second, apply the chi-square exponential moment to $\|\mathbf{Z}\mathbf{x}\|_2^2$ on a $1/4$-net of the unit sphere, of size at most 9^{d-1}. Taking the upper threshold sufficiently large gives failure $e^{Cd-c'm} \le Ce^{-cd}$ when $m \ge d$.

Condition on such $\mathbf{Z}$ and put $\mathbf{e} = \mathbf{s}/\|\mathbf{s}\|_2$, $s = \|\mathbf{s}\|_2/(2m) \ge I_0/4$, and $a_j = \mathbf{z}_j^\top\mathbf{e}$. The matrix

$$(\mathbf{M})_{ij} = a_ia_j\mathbb{E}_{\mathbf{g}}[\mathrm{sign}(\mathbf{z}_i^\top\mathbf{g})\,\mathrm{sign}(\mathbf{z}_j^\top\mathbf{g})]$$

is positive semidefinite, and

$$4m^2\|\partial_{\mathbf{e}}\mathcal{R}\|_{L^2(\mathbf{g})}^2 = \boldsymbol{\epsilon}^\top\mathbf{M}\boldsymbol{\epsilon}, \qquad \mathrm{tr}\,\mathbf{M} = \|\mathbf{Z}\mathbf{e}\|_2^2 \le C\rho^2 m/d.$$

For an auxiliary $\boldsymbol{\xi} \sim \mathcal{N}(\mathbf{0}, \mathbf{I}_m)$, Gaussian linearization and $\cosh x \le e^{x^2/2}$ give

$$\begin{aligned}\mathbb{E}_\epsilon e^{t\boldsymbol{\epsilon}^\top\mathbf{M}\boldsymbol{\epsilon}} &= \mathbb{E}_{\boldsymbol{\xi}}\prod_j \cosh(\sqrt{2t}\,[\mathbf{M}^{1/2}\boldsymbol{\xi}]_j)\\ &\le \mathbb{E}_{\boldsymbol{\xi}}e^{t\boldsymbol{\xi}^\top\mathbf{M}\boldsymbol{\xi}} = \det(\mathbf{I}_m - 2t\mathbf{M})^{-1/2} \le e^{1/2}, \qquad t = (4\,\mathrm{tr}\,\mathbf{M})^{-1}.\end{aligned}$$

Indeed $\|\mathbf{M}\| \le \mathrm{tr}\,\mathbf{M}$, and $-\log(1-x) \le 2x$ for $0 \le x \le 1/2$. If the trace is zero the derivative vanishes identically. Otherwise, exponential Markov gives, for any fixed small universal $\varepsilon_0 > 0$,

$$\mathbb{P}_\epsilon\{\|\partial_{\mathbf{e}}\mathcal{R}\|_{L^2}^2 > \varepsilon_0 I_0^2 \mid \mathbf{Z}\} \le e^{1/2 - c\varepsilon_0 d}.$$

For a fixed initialization on these events, let $h = a_0I_0$, $\psi_h(x) = \max\{-h, \min\{x, h\}\}$ and $p = \mathbb{P}_{\mathbf{g}}(|F_0(\mathbf{g})| \le h)$. Gaussian integration by parts yields

$$\begin{aligned}sp &\le \mathbb{E}|\mathbf{e}^\top\mathbf{g}|\,h + \sqrt{p}\,\|\partial_{\mathbf{e}}\mathcal{R}\|_{L^2} \le h + \sqrt{p}\,\|\partial_{\mathbf{e}}\mathcal{R}\|_{L^2},\\ p &\le 2h/s + \|\partial_{\mathbf{e}}\mathcal{R}\|_{L^2}^2/s^2 \le 8a_0 + 16\varepsilon_0.\end{aligned}$$

The Lipschitz chain rule justifies this calculation. The nonzero clipping levels have zero Gaussian mass: on each of the finitely many gate cones F_0 is linear and homogeneous. Choose a_0, ε_0 so that the last bound is at most $1/16$; their fixed values give (176).

For the path bound, condition on a fresh absolutely continuous path $\mathbf{x}(t)$ independent of initialization. Its row derivative is $Y_j(t) = \epsilon_j \mathbf{1}_{\{\mathbf{v}_j^\top \mathbf{x}(t)>0\}} \mathbf{v}_j^\top \mathbf{x}'(t)$. The independent signs give

$$\begin{aligned}\mathbb{E}_{\epsilon,\mathbf{v}} e^{aY_j(t)} &= \mathbb{E}_{\mathbf{v}} \cosh\bigl(a\mathbf{1}_{\{\mathbf{v}^\top \mathbf{x}(t)>0\}} \mathbf{v}^\top \mathbf{x}'(t)\bigr) \\ &\le \mathbb{E}_{\mathbf{v}} \cosh(a\mathbf{v}^\top \mathbf{x}'(t)) = \exp\left(\frac{a^2\rho^2 \|\mathbf{x}'(t)\|_2^2}{2d}\right).\end{aligned}$$

Thus, for every $r \ge 1$,

$$\|DF_0(\mathbf{x}(t))[\mathbf{x}'(t)]\|_{L^r(\epsilon,\mathbf{V})} \le CI_0\sqrt{r/d}\,\|\mathbf{x}'(t)\|_2. \tag{178}$$

Integrate along the segment and full path in (166). Minkowski over path length and then fresh inputs bounds the L^r norm of the left side of (177) by its displayed right side. The r-th moment inequality proves the claimed failure e^{-r} after absorbing a factor e. The chain rule holds almost everywhere, including paths meeting zero since their velocity is zero almost everywhere on a zero level set. The random signs were not conditioned away. □

Lemma I.4 (Uniform early characteristic drift)**.** *Let $d \ge 6$, $A_* \le 1$, $R, K_* \ge 1$, and $D_A(\mathbf{h}) = \widetilde{F}_A(\mathbf{h}) - F_0(\mathbf{h})$. For every fixed sign array and every $r \ge 1$, with Gaussian-row failure at most e^{-r},*

$$\begin{aligned}&\mathbb{E}_{\mathbf{g}_\perp} \sup_{\substack{0\le A\le A_* \\ |u|\le R,\ 0\le k\le K_*,\ z}} |D_A(\mathbf{h}(u,k,z))| \\ &\quad\le \frac{C\rho A_*(1+K_*^2)}{d} + \frac{C\rho A_*(1+K_*)\sqrt{\ell_0}}{\sqrt{md}} + C\rho\Delta(1+K_*)\left(1+\sqrt{\frac{r}{md}}\right),\end{aligned} \tag{179}$$

where (172) defines Δ, N_0, ℓ_0. The coefficient of the first term has no confidence or grid factor.

Proof. Center conditionally on the entire sign array, writing $\mu_\epsilon(A,\mathbf{h}) = \mathbb{E}_{\mathbf{V}}[D_A(\mathbf{h}) \mid \boldsymbol{\epsilon}]$. For one sign its Gaussian-row mean is

$$\frac{\epsilon\rho}{\sqrt{2\pi d}}(\|e^{\epsilon A\mathbf{J}}\mathbf{h}\|_2 - \|\mathbf{h}\|_2).$$

The absolute value of its A derivative is bounded by $C\rho\|\mathbf{\Pi}_{\mathcal{S}}\mathbf{h}\|_2^2/(\sqrt{d}\|\mathbf{\Pi}_{\mathcal{N}}\mathbf{g}_\perp\|_2)$. Integrate in A, take the supremum, and use independence of the semantic and nuisance Gaussian projections to obtain

$$\mathbb{E}_{\mathbf{g}_\perp} \sup_{A,u,k,z} |\mu_\epsilon(A,\mathbf{h})| \le \frac{C\rho A_*(1+K_*^2)}{d}.$$

Only $\mathbb{E}\|\mathbf{\Pi}_{\mathcal{N}}\mathbf{g}_\perp\|_2^{-1} \le C/\sqrt{d}$ is used here; no high inverse-chi moment is required.

For fixed $\mathbf{g}_\perp$, a row increment has absolute value at most $CA_*\|\mathbf{\Pi}_{\mathcal{S}}\mathbf{v}_j\|_2(\|\mathbf{\Pi}_{\mathcal{S}}\mathbf{g}_\perp\|_2 + K_*)$. Its centered exponential-square parameter is therefore at most $C\rho A_*(\|\mathbf{\Pi}_{\mathcal{S}}\mathbf{g}_\perp\|_2 + K_*)/\sqrt{d}$. Equation (171) and Minkowski yield

$$\left\| \mathbb{E}_{\mathbf{g}_\perp} \max_{\text{grid}} |D_A(\mathbf{h}) - \mu_\epsilon(A,\mathbf{h})| \right\|_{L^r(\mathbf{V}|\epsilon)} \le \frac{C\rho A_*(1+K_*)\sqrt{\ell_0}}{\sqrt{md}}.$$

The interpolation argument in Lemma I.2, now in $L^1(\mathbf{g}_\perp)$, bounds its remaining $L^r(\mathbf{V} \mid \epsilon)$ norm by $C\rho\Delta(1+K_*)(1+\sqrt{r/(md)})$. Combine these two moment bounds before applying the r-th moment inequality, and keep the deterministic conditional mean outside it. This proves (179), with constants uniform in the sign array. □

Proposition I.5 (A finite early-prefix certificate)**.** *Let $0 < \varepsilon < 1/4$, $r = \log(4/\varepsilon)$, $m \ge d \ge C$, and $Ce^{-cd} \le \varepsilon/2$. Assume (162) and $CT^{-22} \le 1/16$. Fix deterministic $A_* \le 1, \Gamma, R, K_*$ as above. Consider any realized trajectory whose early prefix satisfies*

$$-1 \le \gamma_n \le \Gamma, \qquad 0 \le A_n \le A_*, \qquad 0 \le n \le n_A, \tag{180}$$

and

$$\frac{\|f_{\mathrm{X},n}(\mathbf{X}) - \widetilde{F}_{A_n}(\mathbf{h}(\mathbf{X}, \mathbf{q}_{\gamma_n}))\|_{L^2(\mathbf{X})}}{\nu_0} \le H_*, \qquad 0 \le n \le n_A, \tag{181}$$

for a deterministic H_*. If

$$\frac{H_*}{I_0} + CR_{\mathrm{len}}\sqrt{r/d} + \frac{CA_*(1+K_*^2)\sqrt{m}}{d} + CA_*(1+K_*)\sqrt{\ell_0/d} + C\sqrt{m}\,\Delta(1+K_*)\left(1+\sqrt{\frac{r}{md}}\right) \le a_0/8, \tag{182}$$

there is an initialization event of probability at least $1-\varepsilon$ on whose intersection with these trajectory conclusions,

$$\inf_{0 \le n \le n_A} \mathrm{err}_{\mathrm{X}}(n) \ge 1/4. \tag{183}$$

No independence of trained parameters is assumed.

Proof. Intersect the baseline, initial-path, and drift events. Their total failure is at most $Ce^{-cd} + 2e^{-r} \le \varepsilon$. For each fixed n, conditional Gaussian disintegration realizes the original axial pool by (159); any remaining fresh token coordinates retain their conditional law. Set $V_n = f_{\mathrm{X},n}(\mathbf{X})/\nu_{\gamma_n}$. Its sign equals the actual prediction sign, and (160) implies

$$\mathbb{E}|V_n - \widetilde{F}_{A_n}(\mathbf{v}_{\gamma_n})| \le \|f_{\mathrm{X},n} - \widetilde{F}_{A_n}(\mathbf{h}(\mathbf{X}, \mathbf{q}_{\gamma_n}))\|_{L^2}/\nu_0 \le H_*.$$

The uniform path and drift bounds, together with (182), give $\mathbb{E}[\mathbf{1}_{\mathcal{P}}|V_n - B_0|] \le a_0 I_0/8$. Consequently, on the same event for every prefix index,

$$\begin{aligned}\mathbb{P}\{\mathrm{sign}(V_n) \ne \mathrm{sign}(B_0)\} &\le \mathbb{P}(\mathcal{P}^c) + \mathbb{P}(|B_0| \le a_0 I_0) + \frac{\mathbb{E}[\mathbf{1}_{\mathcal{P}}|V_n - B_0|]}{a_0 I_0} \\ &\le 1/16 + 1/16 + 1/8 = 1/4.\end{aligned}$$

For each fixed initialization B_0 is independent of the balanced label and has error exactly $1/2$, including $\mathrm{sign}(0) = 1$. This proves (183). The characteristic events were uniform over the deterministic parameter boxes, so adaptive selection of A_n, γ_n creates no time union bound. The disintegration need only preserve each marginal, not a common coupling of all unused token noises. □

Application at the early checkpoint. Recall $K_A = \sqrt{K_{\mathrm{X}}}$ and apply Proposition I.5 with $\varepsilon = \delta/8$, $\Gamma = \Gamma_{K_A} + 1$, $K_* = CK_A$, and

$$A_* = \frac{C\lambda K_A}{P\sqrt{G}}, \qquad H_* \le \frac{CA_*(1+K_*)}{d} + Cd^{-5/4},$$

as supplied by (37) and (126), respectively. Since $K_A^6 = K_{\mathrm{X}}^3 = c_*^3(P^2G)^{3/2}$, the upper width in Assumption 3.1, written as (16), gives

$$m \le \frac{c_{\mathrm{e}} P^2 G d^2}{\lambda^2 K_A^6}, \qquad c_{\mathrm{e}} = c_u c_*^3. \tag{184}$$

Choose the universal upper-width coefficient c_u after c_* so that c_{e} is sufficiently small. This is the width condition in Assumption 3.1, with no additional confidence-dependent restriction. Indeed,

$$\begin{aligned}\frac{CA_*(1+K_*^2)\sqrt{m}}{d} &\le C\sqrt{c_{\mathrm{e}}}, \\ H_*/I_0 &\le C\sqrt{c_{\mathrm{e}}}/K_A + Cd^{-1/4}.\end{aligned} \tag{185}$$

Under Assumption 3.1, $R_{\mathrm{len}} \le CL_d^3$, $A_*(1+K_*) \le CL_d^3$, and $\log(2N_0) \le CL_d$. Under Assumption 3.1, Lemma A.3 gives $r + \log(2N_0) \le C\ell_\delta$ and

$$CL_d^3\sqrt{\ell_\delta/d} \le Cd^{-31/64}L_d^{-9}, \qquad C\sqrt{m}\,\Delta(1+K_*)\left(1+\sqrt{r/(md)}\right) \le Cd^{-2}.$$

The same consequences of Assumption 3.1 give $m \ge d$, $Ce^{-cd} \le \varepsilon/2$ and all finite position conditions. Choose c_u so the two constant contributions in (185) use at most half the budget in (182); then enlarge C_ℓ so the displayed vanishing bounds use the rest. Thus Proposition I.5 gives the entire early-prefix conclusion with additional initialization failure probability at most $\delta/8$.

J Exponential transfer to the trained XOR classifier

The classification transfer uses the row errors themselves and the Gaussian fresh-test law. The resulting tail bounds hold for arbitrary dependence among the trained rows. Independence is used only for the initialization seeds defining the characteristic at a deterministic reference clock.

Apply Lemma C.2 with $\mathbf{b} = b\mathbf{e}_p + a(z_1\boldsymbol{\mu}_1 + z_2\boldsymbol{\mu}_2)$ and $\boldsymbol{\Sigma} = \nu^2\boldsymbol{\Pi}_\perp$. For row errors $\mathbf{e}_j$, define $D_\perp^2 = m^{-1}\sum_j \|\boldsymbol{\Pi}_\perp\mathbf{e}_j\|_2^2$. The deterministic term in that lemma satisfies

$$\begin{aligned} M &= m^{-1}\sum_j |\mathbf{e}_j^\top \mathbf{b}| \\ &\le |b|m^{-1}\sum_j |\mathbf{e}_j^\top \mathbf{e}_p| + |a|m^{-1}\sum_j |\mathbf{e}_j^\top(z_1\boldsymbol{\mu}_1 + z_2\boldsymbol{\mu}_2)| \\ &\le |b|D_p + \sqrt{2}|a|D_s, \\ D_\Sigma^2 &= \nu^2 D_\perp^2. \end{aligned}$$

Conditioning on training permits arbitrary dependence among the row errors.

J.1 Fresh-test events and pointwise attention-weight replacement

The events below are defined on an independent fresh example and leave the training distribution unchanged. In the application below, the data ranges follow from Assumption 3.1; the reference trajectory supplies the score bounds.

Lemma J.1 (Strong fresh-test events)**.** *Suppose $\sigma^2 \asymp d^{-1}$, $T = d\,\mathrm{polylog}(d)$, $(\log d)^{3/4} \le P^2 \le 2\log T$, and $|\gamma| \le C\log\log T$. Write $s = e^\gamma/T$, $K = s/\nu_0$, and assume $K \le (\log d)^C$. There is a position-only event $\mathcal{P}_+$ with*

$$\mathbb{P}(\mathcal{P}_+^c) \le Ce^{-cd^{1/4}}$$

on which, uniformly over these axial scores,

$$\nu \asymp \nu_0, \qquad k = a/\nu \asymp K, \qquad |b|/\nu_0 \le C\{PK + \alpha|\gamma| + d^{1/8}\} \le Cd^{1/8}, \qquad \alpha = \sigma\sqrt{T}/P.$$

There is a second, full-token event $\mathcal{T}_+$ with $\mathbb{P}(\mathcal{T}_+^c) \le Ce^{-cd}$ such that, uniformly over every attention weight $\mathbf{q}'$ satisfying $\|\mathbf{q}' - \gamma\mathbf{e}_p/P\|_2 \le \xi = d^{-15/16}$,

$$\|D_{\mathbf{q}}\mathbf{h}(\mathbf{X}, \mathbf{q}')\| \le C(\sigma^2 + P^2 s). \tag{186}$$

Both events are independent of initialization and training.

Proof. Let $\zeta_i = \mathbf{e}_p^\top\boldsymbol{\epsilon}_i$ be the independent distractor position noises. Fix a sufficiently large absolute C_0 and set

$$\mathcal{P}_+ = \left\{ \max_{i\ge 3}|\zeta_i| \le C_0, \quad N_D^{-1}\sum_{i\ge 3}\zeta_i^2 \le C_0\sigma^2, \quad \left|N_D^{-1}\sum_{i\ge 3}\zeta_i\right| \le \nu_0 d^{1/8} \right\}.$$

A scalar Gaussian union bound gives Ce^{-cd} for failure of the first condition, the chi-square exponential moment gives Ce^{-cT} for the second, and the sample mean is exactly $\mathcal{N}(0, \sigma^2/N_D)$, giving $2e^{-cd^{1/4}}$ for the third. Their sum proves the claimed failure bound.

For $\theta = \gamma/P$,

$$|\theta| \max_i |\zeta_i| \le C\frac{\log\log T}{P} \le C(\log\log T)(\log d)^{-3/8} = o(1).$$

Thus the sums $\sum_{i\ge 3} e^{r\theta\zeta_i}$, $r = 1, 2$, are comparable to T. Also $e^{2\gamma}/T = \sigma^2 K^2 = o(1)$, so the exact pool formulas give $\nu \asymp \nu_0$ and $a/\nu \asymp K$. For

$$D(\theta) = \frac{\sum_{i\ge 3} \zeta_i e^{\theta\zeta_i}}{\sum_{i\ge 3} e^{\theta\zeta_i}},$$

its derivative is the weighted variance. On the entire segment between 0 and θ, its absolute value is at most $CN_D^{-1}\sum_i \zeta_i^2 \le C\sigma^2$. Since $D(0)$ is the sample mean,

$$|D(\theta)| \le \nu_0 d^{1/8} + C\sigma^2|\gamma|/P.$$

The exact pooled position is $b = 2Pa + (1-2a)D(\theta)$. Divide by ν_0 and use $a \le C\nu_0 K$ to obtain the center bound.

For the second event, let $\boldsymbol{\xi}_i$ denote the original Gaussian noise in token i, with its appropriate projected covariance for the two relevant tokens, and let $\boldsymbol{\Xi}_D$ have the distractor noises as columns. Take

$$\mathcal{T}_+ = \left\{ \max_{1\le i\le T} \|\boldsymbol{\xi}_i\|_2 \le C_1, \quad T^{-1}\|\boldsymbol{\Xi}_D\|^2 \le C_1\sigma^2 \right\}$$

for a sufficiently large absolute C_1. The chi-square exponential moment and a union over T columns give Ce^{-cd} for failure of the first condition. For the second, apply the same chi-square upper tail to $\|\boldsymbol{\Xi}_D^\top \mathbf{v}\|_2^2$ at a $1/4$-net of the unit sphere in $\mathbb{R}^d$, with at most 9^d points. Choosing C_1 large gives $Ce^{-cT+Cd} \le Ce^{-c'T}$ because $T/d \to \infty$.

On $\mathcal{T}_+$, for a distractor,

$$|\mathbf{q}'^\top \boldsymbol{\xi}_i| \le |\gamma|C_1/P + \xi C_1 = o(1),$$

and the deviation of a relevant token's logit from γ is at most $\xi(P+1+C_1) = o(1)$; its Gaussian position coordinate is exactly zero. Consequently, uniformly throughout the whole attention-weight ball,

$$s_i(\mathbf{X}, \mathbf{q}') \le Cs \quad (i = 1, 2), \qquad s_i(\mathbf{X}, \mathbf{q}') \le C/T \quad (i \ge 3).$$

The attention covariance is bounded by the uncentered second moment. The two relevant token norms are at most CP, and the distractor second moment is bounded by $CT^{-1}\boldsymbol{\Xi}_D\boldsymbol{\Xi}_D^\top$. This gives (186) directly. The events depend only on the fresh example, so they are independent of the trained state. □

Lemma J.2 (Population-moment concentration)**.** *Let $\mathbf{M}_+, \mathbf{M}_- \in \mathbb{R}^{d\times d}$ satisfy $\|\mathbf{M}_\pm\| \le e$. Let $\mathbf{u}_{j,0} \sim \mathcal{N}(\mathbf{0}, \rho^2\mathbf{I}_d/d)$ and ϵ_j be mutually independent Gaussian rows and Rademacher signs. For an independent random vector $\mathbf{h}$, set*

$$f(\mathbf{h}) = \frac{1}{m}\sum_j \epsilon_j[\mathbf{u}_{j,0}^\top \mathbf{M}_{\epsilon_j}\mathbf{h}]_+, \qquad \bar{f}(\mathbf{h}) = \mathbb{E}_{\text{init}} f(\mathbf{h}).$$

Let $\mathcal{Q}$ be an event in the input probability space. Suppose that, for $r \ge 2$ and $v, \Lambda_r > 0$,

$$\mathbf{1}_{\mathcal{Q}}\|\mathbf{h}\|_2 \le v\sqrt{d}, \qquad \|\mathbf{1}_{\mathcal{Q}}\mathbf{w}^\top \mathbf{M}_\pm \mathbf{h}\|_{L^r} \le v\Lambda_r\|\mathbf{w}\|_2 \quad (\mathbf{w} \in \mathbb{R}^d).$$

For $\zeta \in (0,1)$, write $\ell = \log(4/\zeta)$. With initialization probability at least $1-\zeta$,

$$\|\mathbf{1}_{\mathcal{Q}}(f - \bar{f})\|_{L^r} \le \frac{C\rho v}{\sqrt{m}}\left(\sqrt{r} + \sqrt{\ell} + \Lambda_r\sqrt{\ell/d}\right). \tag{187}$$

The norm is over the original input law, with the indicator retained.

Proof. Condition on all signs and write $c_\epsilon(\mathbf{h}) = \mathbb{E}_{\mathrm{G}} f(\mathbf{h})$ and $Z_r = \|\mathbf{1}_{\mathcal{Q}}(f - c_\epsilon)\|_{L^r}$, where $\mathbb{E}_{\mathrm{G}}$ averages the Gaussian rows. Centered positive-part Gaussian variables are sub-Gaussian, so Tonelli

and Jensen give

$$\begin{aligned}\mathbb{E}_{\mathrm{G}} Z_r &\le (\mathbb{E}_{\mathrm{G}} Z_r^r)^{1/r} \\ &= (\mathbb{E}_{\mathbf{h}} \mathbf{1}_{\mathcal{Q}} \mathbb{E}_{\mathrm{G}} |f(\mathbf{h}) - c_\epsilon(\mathbf{h})|^r)^{1/r} \\ &\le C\rho\sqrt{r/(md)} (\mathbb{E}_{\mathbf{h}} \mathbf{1}_{\mathcal{Q}} \|\mathbf{h}\|_2^r)^{1/r} \\ &\le C\rho v\sqrt{r/m}.\end{aligned}$$

Write $\mathbf{u}_{j,0} = \rho \mathbf{g}_j/\sqrt{d}$ and let $\mathbf{G}$ collect the standard Gaussian rows. For another row array $\mathbf{G}'$, ReLU Lipschitzness and the directional moment assumption imply

$$\begin{aligned}|Z_r(\mathbf{G}) - Z_r(\mathbf{G}')| &\le \|\mathbf{1}_{\mathcal{Q}}(f_{\mathbf{G}} - f_{\mathbf{G}'})\|_{L^r} \\ &\le \frac{\rho}{m\sqrt{d}} \sum_j \|\mathbf{1}_{\mathcal{Q}} (\mathbf{g}_j - \mathbf{g}_j')^\top \mathbf{M}_{\epsilon_j} \mathbf{h}\|_{L^r} \\ &\le \frac{\rho v \Lambda_r}{m\sqrt{d}} \sum_j \|\mathbf{g}_j - \mathbf{g}_j'\|_2 \\ &\le \frac{\rho v \Lambda_r}{\sqrt{md}} \|\mathbf{G} - \mathbf{G}'\|_F.\end{aligned}$$

Gaussian Lipschitz concentration (Alberts & Khoshnevisan, 2018, Chapter 5, Theorem 1.1) therefore gives, conditionally on every sign vector, with failure at most $2e^{-\ell}$,

$$Z_r \le \frac{C\rho v}{\sqrt{m}} \left(\sqrt{r} + \Lambda_r \sqrt{\ell/d}\right).$$

The remaining sign bias is exact. With $\bar{\epsilon} = m^{-1} \sum_j \epsilon_j$,

$$\begin{aligned}c_\epsilon(\mathbf{h}) - \bar{f}(\mathbf{h}) &= \frac{\rho \bar{\epsilon}}{2\sqrt{2\pi d}} (\|\mathbf{M}_+ \mathbf{h}\|_2 + \|\mathbf{M}_- \mathbf{h}\|_2), \\ \|\mathbf{1}_{\mathcal{Q}}(c_\epsilon - \bar{f})\|_{L^r} &\le C\rho v |\bar{\epsilon}| \\ &\le C\rho v \sqrt{\ell/m}, \\ \|\mathbf{1}_{\mathcal{Q}}(f - \bar{f})\|_{L^r} &\le Z_r + C\rho v\sqrt{\ell/m}.\end{aligned}$$

The second inequality fails with probability at most $2e^{-\ell}$ by Hoeffding's inequality. A union bound proves the claim. □

Proposition J.3 (Late XOR error)**.** *Under Assumption 3.1, at the deterministic late reference checkpoint n_{X}, write $A = A_{n_{\mathrm{X}}}$ and $K = K_{\mathrm{X}} = c_* P\sqrt{G}$. Suppose Proposition D.1 and Theorem H.1 supply $c\lambda c_* \le A \le C\lambda c_* \le a_* < 1$,*

$$D_s + D_N \le CA/d + Cd^{-11/8}, \qquad D_p \le Cd^{-11/8},$$

and

$$\|\mathbf{q}_{n_{\mathrm{X}}} - \bar{\mathbf{q}}_{n_{\mathrm{X}}}\|_2 \le d^{-15/16}, \qquad \|\mathbf{W}_{\mathrm{X},n_{\mathrm{X}}}\| \le Cd^{-1/2}.$$

Let $\zeta \in (0,1)$ and assume

$$E := \frac{mA^2K^4}{d^2}, \qquad C\log(4/\zeta) \le E \le cd.$$

There is an initialization event of probability at least $1-\zeta$ such that, on its intersection with these trajectory conclusions, the actual classifier satisfies

$$\mathrm{err}_{\mathrm{X}}(n_{\mathrm{X}}) \le Ce^{-cd^{1/4}} + Ce^{-cK^2} + e^{-cE} + 2e^{-cK^4}. \tag{188}$$

All constants in the exponents are independent of d, m. Fresh-test probabilities are conditional on the realized training history and do not cost another training or initialization failure event.

Proof. Use $\mathcal{P}_+$ from Lemma J.1. Conditional on its position variables, write the exact axial pool as

$$\mathbf{h} = \nu \mathbf{v}, \qquad \mathbf{v} = \mathbf{z} + u\mathbf{e}_p + k(z_1 \boldsymbol{\mu}_1 + z_2 \boldsymbol{\mu}_2), \qquad \mathbf{z} \sim \mathcal{N}(0, \boldsymbol{\Pi}_{\mathbf{p}^\perp}),$$

where $\nu \asymp \nu_0$, $k \asymp K$, and $u^2 + K^2 \le Cd^{1/4} = o(d)$. Define the good fresh-test event

$$\mathcal{Q} = \mathcal{P}_+ \cap \{z_i \boldsymbol{\mu}_i^\top \mathbf{v} \ge k/2 \ (i = 1, 2), \ \|\mathbf{v}\|_2 \le C\sqrt{d}\}.$$

The two scalar Gaussian tails and the Gaussian norm tail imply

$$\mathbb{P}(\mathcal{Q}^c) \le Ce^{-cd^{1/4}} + Ce^{-cK^2} + Ce^{-cd}.$$

The exact characteristic-mean formula in Lemma I.1 gives

$$y\overline{F}_A(\mathbf{h}) \ge \mu := c_1 \rho \nu_0 A K^2/d \quad \text{on } \mathcal{Q}, \tag{189}$$

for an absolute sufficiently small $c_1 > 0$.

For a deterministic vector $\mathbf{w}$ and $r \ge 2$, the conditional Gaussian representation on $\mathcal{P}_+$ yields

$$\begin{aligned}\|\mathbf{1}_{\mathcal{Q}} \mathbf{w}^\top e^{\pm A\mathbf{J}} \mathbf{h}\|_{L^r} &\le \|\mathbf{1}_{\mathcal{P}_+} \mathbf{w}^\top e^{\pm A\mathbf{J}} [b\mathbf{e}_p + a(z_1\boldsymbol{\mu}_1 + z_2\boldsymbol{\mu}_2) + \nu\mathbf{z}]\|_{L^r} \\ &\le C\nu_0(d^{1/8} + K)\|\mathbf{w}\|_2 + C\nu_0 \|\mathbf{w}^\top e^{\pm A\mathbf{J}} \mathbf{z}\|_{L^r} \\ &\le C\nu_0(d^{1/8} + K + \sqrt{r})\|\mathbf{w}\|_2.\end{aligned}$$

Thus Lemma J.2, with $\mathbf{M}_\pm = e^{\pm A\mathbf{J}}$, $v = C\nu_0$, and $\Lambda_r = C(d^{1/8} + K + \sqrt{r})$, gives, with probability $1 - \zeta$,

$$\begin{aligned}W_r := \|\mathbf{1}_{\mathcal{Q}}(\widetilde{F}_A - \overline{F}_A)\|_{L^r} &\le \frac{C\rho\nu_0}{\sqrt{m}} \left[\sqrt{r} + \sqrt{\ell} + \sqrt{(d^{1/4} + K^2 + r)\ell/d}\right] \\ &\le C\rho\nu_0 \sqrt{(r+\ell)/m}, \qquad \ell = \log(4/\zeta), \quad 2 \le r \le d.\end{aligned}$$

Here $K^2 \le Cd$ follows from Assumption 3.1 through Lemma A.3. Choose $r = \kappa E$ with a sufficiently small universal $\kappa > 0$. Increasing the universal constant in $E \ge C\ell$ ensures $r \ge 2$ and $2W_r/\mu \le 1/2$; $E \le cd$ ensures $r \le d$. Consequently

$$\begin{aligned}\mathbb{P}_{\mathbf{X}}\{\mathcal{Q}, \ y\widetilde{F}_A(\mathbf{h}) \le \mu/2\} &\le \mathbb{P}_{\mathbf{X}}\{\mathcal{Q}, |\widetilde{F}_A(\mathbf{h}) - \overline{F}_A(\mathbf{h})| \ge \mu/2\} \\ &\le (2W_r/\mu)^r \\ &\le 2^{-r} \le e^{-cE}.\end{aligned} \tag{190}$$

The reference $A, \bar{\mathbf{q}}_{n_X}$ and $\mathcal{Q}$ are deterministic before initialization. This event is intersected with the trajectory events only after its probability is established. The exact common bias from the readout signs is included in Lemma J.2.

Next fix any realized training trajectory on $\mathcal{E}_J \cap \mathcal{E}_H$ from Proposition D.1 and Theorem H.1. Set $\mathbf{u}_j = \mathbf{u}_{\mathrm{X},j,n_X}$ and $\mathbf{v}_j = e^{\epsilon_j A\mathbf{J}} \mathbf{u}_{j,0}$, so both arrays are now fixed. The perpendicular RMS error satisfies $D_\perp \le CA/d$ by $A \ge c\lambda c_*$ and Assumption 3.1, using Lemma A.3. On $\mathcal{P}_+$, the deterministic center contribution obeys

$$|b|D_p + \sqrt{2}aD_s \le C\nu_0\{d^{-5/4} + AK/d + Kd^{-11/8}\} = o(\mu).$$

Here $a = \nu k$, $|b| \le C\nu_0 d^{1/8}$, and A^{-1}, K are polynomial logarithmic; the leading ratio is $C/K = o(1)$. Assumption 3.1, through Lemma A.3, makes this contribution at most $\mu/16$. Lemma C.2, with $t = \mu/16$, therefore gives

$$\mathbb{P}_{\mathbf{X}}\{\mathcal{P}_+, \ |F_{\mathbf{U}}(\mathbf{h}(\mathbf{X}, \bar{\mathbf{q}}_{n_X})) - F_{\mathbf{V}}(\mathbf{h}(\mathbf{X}, \bar{\mathbf{q}}_{n_X}))| > \mu/8\} \le 2e^{-cK^4}.$$

This bound conditions only on positions when testing the Gaussian noise. No conditioning on the semantic-margin event or the full-token event is required, and correlation between $\mathbf{U}$ and $\mathbf{V}$ is harmless.

Finally, ReLU Lipschitzness and Cauchy–Schwarz over the rows imply

$$|F_{\mathbf{U}}(\mathbf{h}) - F_{\mathbf{U}}(\mathbf{h}')| \le \|\mathbf{W}_X\| \|\mathbf{h} - \mathbf{h}'\|_2.$$

On $\mathcal{T}_+$, integrate (186) along the segment from $\bar{\mathbf{q}}_{n_X}$ to $\mathbf{q}_{n_X}$:

$$\begin{aligned}|F_{\mathbf{U}}(\mathbf{h}(\mathbf{X}, \mathbf{q}_{n_X})) - F_{\mathbf{U}}(\mathbf{h}(\mathbf{X}, \bar{\mathbf{q}}_{n_X}))| &\le Cd^{-1/2} d^{-15/16}(\sigma^2 + P^2 s) \\ &\le C\nu_0 d^{-23/16}\{\sigma\sqrt{T} + P^2 K\} = o(\mu).\end{aligned}$$

The braces are polynomial logarithmic under Assumption 3.1, while the reference clock controls A^{-1}; the remaining factor is $d^{-7/16}$ relative to the margin. Lemma A.3, under Assumption 3.1, therefore makes the discrepancy at most $\mu/8$ throughout $\mathcal{T}_+$. Its exceptional probability is Ce^{-cd}.

Outside the separately bounded events, the actual signed score is at least $\mu/2 - \mu/8 - \mu/8 = \mu/4 > 0$. A union bound with (190) and the two transfer bounds proves (188), absorbing Ce^{-cd} into $Ce^{-cd^{1/4}}$. □

K THE XOR-ONLY COMPARATOR ON THE SAME TRAINING BUDGET

K.1 COMPARATOR AND A GENERAL SUFFICIENT CONDITION

The comparator sees only the original law $\mathcal{D}_{\mathrm{X}}$. It has one attention weight $\mathbf{q}'_n$ and one head $\mathbf{W}'_n$, the same width, fixed readout signs $a_j = \epsilon_j/\sqrt{m}$, common learning rate η, and batch size B as the joint learner. Use the coupling $\mathbf{q}'_0 = \mathbf{q}_0 = 0$ and $\mathbf{W}'_0 = \mathbf{W}_{\mathrm{X},0}$; the marginal entries of $\mathbf{W}'_0$ are independent $\mathcal{N}(0, \rho^2/(md))$. At every iteration draw B independent fresh XOR examples, independently of the past. Write

$$f'_n(\mathbf{X}) = \sum_{j=1}^{m} a_j[(\mathbf{w}'_{j,n})^\top \mathbf{h}(\mathbf{X}, \mathbf{q}'_n)]_+, \qquad r(u) = (1+e^u)^{-1}.$$

Both parameter blocks are updated at their old state:

$$\mathbf{W}'_{n+1} - \mathbf{W}'_n = \frac{\eta}{B}\sum_{i=1}^{B} y_{ni} r(y_{ni} f'_n(\mathbf{X}_{ni})) \nabla_{\mathbf{W}'} f'_n(\mathbf{X}_{ni}),$$

$$\mathbf{q}'_{n+1} - \mathbf{q}'_n = \frac{\eta}{B}\sum_{i=1}^{B} y_{ni} r(y_{ni} f'_n(\mathbf{X}_{ni})) \mathbf{C}(\mathbf{X}_{ni}, \mathbf{q}'_n) \nabla_{\mathbf{h}} f'_n(\mathbf{X}_{ni}). \tag{191}$$

The ReLU derivative at zero is zero. The fresh population error is $\mathrm{err}'_{\mathrm{X}}(n) = \mathbb{P}_{(\mathbf{X},y)\sim\mathcal{D}_{\mathrm{X}}}(\mathrm{sign}(f'_n(\mathbf{X})) \neq y)$, conditional on the training state, with $\mathrm{sign}(0) = 1$. Thus N iterations use exactly BN fresh examples for either learner. The lower bound below holds for every fixed realization of the readout signs.

In the trajectory proof below, t and k index integer iterations. For this section only, suppress primes in the proof, write $\mathbf{e} = \mathbf{e}_p$,. The geometry and the Gaussian noise law remain those of the main setting. For an arbitrary deterministic horizon $N \geq 1$, confidence $\delta_L \in (0, 1/10)$, and physical time $L = \eta N$, set

$$g = \frac{1}{\sigma T^{3/2}}, \qquad Q = K_0 \frac{\rho\sigma^2 L}{\sqrt{d}}, \qquad R_{\mathrm{X}} = P + 1 + \sigma\sqrt{d + \log T}, \qquad \ell_* = \log\frac{32m}{\delta_L}. \tag{192}$$

Proposition K.1 (General comparator certificate)**.** *There are universal constants $K_0, C, c, c_{\mathrm{tail}} > 0$ such that the following is sufficient for the comparator (191):*

$$m \geq d, \quad T \geq Cd, \quad d \geq C\log\frac{m}{\delta_L}, \quad P + 1 \leq c\sigma\sqrt{T}, \quad \sigma Q\sqrt{T} \leq c, \tag{193}$$

$$\frac{gL}{\rho} + \frac{\sigma^2 dL}{T} \leq c, \qquad \left(\frac{LR_{\mathrm{X}}\sqrt{d}}{\rho} + \frac{R_{\mathrm{X}}^2}{\sigma^2}\right) e^{-c_{\mathrm{tail}} d} \leq c\delta_L, \tag{194}$$

$$\eta \leq c\min\left\{\frac{\rho^2 T}{\sigma^2 d^2 L\ell_*}, \frac{\rho\sqrt{T}}{\sigma d\ell_*}\right\}. \tag{195}$$

Here $Q, L, R_{\mathrm{X}}, \ell_*$ are explicit functions of the primitive parameters and the prescribed horizon, not assumptions on the subsequent trajectory. With probability at least $1 - \delta_L$, simultaneously for $0 \leq n \leq N$,

$$\|\mathbf{q}'_n\|_2 \leq Q, \qquad \|\mathbf{W}'_n\| \leq \frac{C\rho}{\sqrt{d}}, \qquad \mathrm{err}'_{\mathrm{X}}(n) \geq \frac{1}{2} - \frac{C}{\sigma\sqrt{T}} - Ce^{-cd}. \tag{196}$$

Its proof occupies the next three subsections. Constants are chosen in the order specified in the stopping argument; increasing the universal C and decreasing the universal c in the sufficient conditions supplies all required strict inequalities.

K.2 PARITY SUPPRESSION FOR ONE ACTUAL ROW

Let $Q_{\mathcal{S}} = \|\mathbf{\Pi}_{\mathcal{S}}\mathbf{q}\|_2$. For a nonzero $\mathbf{w}$, set

$$R = \|\mathbf{w}\|_2, \quad \chi_{\mathcal{S}} = \|\mathbf{\Pi}_{\mathcal{S}}\mathbf{w}\|_2/R, \quad \chi_p = |\mathbf{e}^\top\mathbf{w}|/R, \quad \kappa_{\mathbf{w}} = \chi_{\mathcal{S}} + Q_{\mathcal{S}}(\sigma + P\chi_p).$$

Lemma K.2. *Under the geometric conditions in (193), if $\|\mathbf{q}\|_2 \le Q$ and $\chi_p \le 1/2$, then*

$$\left\|\mathbb{E}[y\mathbf{h}(\mathbf{X},\mathbf{q})\mathbf{1}_{\{\mathbf{w}^\top \mathbf{h}(\mathbf{X},\mathbf{q})>0\}}]\right\|_2 \le Cg\kappa_{\mathbf{w}} + CR_{\mathrm{X}}e^{-cT}. \tag{197}$$

Proof. Decompose the perpendicular noises into their sample mean and centered residuals:

$$\boldsymbol{\epsilon}_i^\perp = \mathbf{g}_0 + \boldsymbol{\zeta}_i, \qquad \sum_i \boldsymbol{\zeta}_i = 0, \qquad \mathbf{g}_0 \sim \mathcal{N}(\mathbf{0}, \sigma^2\boldsymbol{\Pi}_\perp/T).$$

Gaussian orthogonality makes $\mathbf{g}_0$ independent of the residuals and all position noises. The common $\mathbf{g}_0$ cancels from all attention-logit differences, so $\mathbf{h} = \mathbf{c}(z,\mathbf{q}) + \mathbf{g}_0$ exactly.

Let $\mathcal{E}$ bound the centered perpendicular noise matrix and the distractor position-noise vector by $C_1\sigma\sqrt{T}$ in operator and Euclidean norm, respectively. Gaussian matrix and vector tails give $\mathbb{P}(\mathcal{E}^c) \le Ce^{-cT}$ because $T \ge Cd$. For an iid standard Gaussian matrix $\mathbf{Z} \in \mathbb{R}^{T\times d}$, the bound is $\mathbb{P}(\|\mathbf{Z}\| > \sqrt{T} + \sqrt{d} + u) \le 2e^{-u^2/2}$; centering is an orthogonal projection. The event is independent of $\mathbf{g}_0, z$. On this event, all residual logits are bounded by a universal constant uniformly for $z \in [-1,1]^2$ and $\|\mathbf{q}\|_2 \le Q$. Consequently

$$s_i \le C/T, \qquad \|\mathbf{C}(\mathbf{X}, u\mathbf{q})\| \le C\sigma^2 \quad (0 \le u \le 1). \tag{198}$$

For the covariance bound, subtract $\mathbf{g}_0$ from all tokens. The resulting token matrix has squared operator norm at most $C[\sigma^2 T + (P+1)^2]$; multiply by $\max_i s_i$. Integration of $D_{\mathbf{q}}\mathbf{h} = \mathbf{C}$ then gives

$$\|\mathbf{c}(z,\mathbf{q}) - \mathbf{c}(z,0)\|_2 \le C\sigma^2 Q, \quad \mathbf{c}(z,0) = \frac{2\mathbf{p} + z_1\boldsymbol{\mu}_1 + z_2\boldsymbol{\mu}_2}{T} + b_0\mathbf{e}, \tag{199}$$

where b_0 is Gaussian with variance $\sigma^2(T-2)/T^2$.

Condition on the residuals and position noises. Put $\tau = \sigma\|\boldsymbol{\Pi}_\perp\mathbf{w}\|_2/\sqrt{T}$ and $\mathbf{k} = \sigma^2\boldsymbol{\Pi}_\perp\mathbf{w}/T$. Integrating the common Gaussian exactly gives

$$\mathbf{j}_{\mathbf{w}}(\mathbf{c}) = \mathbb{E}_{\mathbf{g}_0}[(\mathbf{c}+\mathbf{g}_0)\mathbf{1}_{\{\mathbf{w}^\top(\mathbf{c}+\mathbf{g}_0)>0\}}] = \mathbf{c}\Phi(\mathbf{w}^\top\mathbf{c}/\tau) + (\mathbf{k}/\tau)\varphi(\mathbf{w}^\top\mathbf{c}/\tau), \tag{200}$$

Writing $\beta_c = \mathbf{q}^\top\boldsymbol{\mu}_c$ and $\widetilde{\mathbf{x}}_c = \mathbf{x}_c - \mathbf{g}_0$, differentiation in the continuous sign variables yields

$$\begin{aligned}\mathbf{c}_c &= s_c[\boldsymbol{\mu}_c + \beta_c(\widetilde{\mathbf{x}}_c - \mathbf{c})],\\ \mathbf{c}_{12} &= -s_1s_2[\beta_2\boldsymbol{\mu}_1 + \beta_1\boldsymbol{\mu}_2 + \beta_1\beta_2(\widetilde{\mathbf{x}}_1 + \widetilde{\mathbf{x}}_2 - 2\mathbf{c})].\end{aligned} \tag{201}$$

For every fixed $r \le 8$, the following moment estimates hold:

$$\begin{aligned}\|\mathbf{1}_{\mathcal{E}}\|\mathbf{c}\|_2\|_{L^r} &\le C\sigma/\sqrt{T}, & \|\mathbf{1}_{\mathcal{E}}\|\mathbf{c}_c\|_2\|_{L^r} &\le C/T,\\ \left\|\mathbf{1}_{\mathcal{E}}|\mathbf{w}^\top\mathbf{c}_c|\right\|_{L^r} &\le CR\kappa_{\mathbf{w}}/T, & \|\mathbf{1}_{\mathcal{E}}\|\mathbf{c}_{12}\|_2\|_{L^r} &\le CQ_{\mathcal{S}}/T^2.\end{aligned} \tag{202}$$

Indeed, (199) and $\sigma Q\sqrt{T} \le c$ imply the first bound. The vector moment $\|\mathbf{1}_{\mathcal{E}}\|\widetilde{\mathbf{x}}_c - \mathbf{c}\|_2\|_{L^r} \le C(P+1+\sigma\sqrt{d})$ and $Q(P+1+\sigma\sqrt{d}) \le Cc$ imply the other vector bounds from (201). For the scalar projection,

$$\left\|\mathbf{1}_{\mathcal{E}}|\mathbf{w}^\top(\widetilde{\mathbf{x}}_c - \mathbf{c})|\right\|_{L^r} \le CR(\sigma + P\chi_p + \chi_{\mathcal{S}} + \sigma^2 Q).$$

After multiplication by $|\beta_c| \le Q_{\mathcal{S}}$, the $\chi_{\mathcal{S}}$ term is absorbed into $\chi_{\mathcal{S}}$ using $Q \le c/(P+1)$; the $\sigma^2 Q$ term is absorbed into σ using $\sigma Q \le c$. This proves the scalar bound without an extraneous $\sqrt{d}$ factor.

For $x = \mathbf{w}^\top\mathbf{c}/\tau$, direct differentiation of (200) gives, for vector directions $\mathbf{u}, \mathbf{v} \in \mathbb{R}^d$,

$$\begin{aligned}D\mathbf{j}_{\mathbf{w}}(\mathbf{c})[\mathbf{u}] &= \mathbf{u}\Phi(x) + \mathbf{c}\varphi(x)\frac{\mathbf{w}^\top\mathbf{u}}{\tau} + \mathbf{k}\varphi'(x)\frac{\mathbf{w}^\top\mathbf{u}}{\tau^2},\\ D^2\mathbf{j}_{\mathbf{w}}(\mathbf{c})[\mathbf{u},\mathbf{v}] &= \varphi(x)\frac{\mathbf{u}(\mathbf{w}^\top\mathbf{v}) + \mathbf{v}(\mathbf{w}^\top\mathbf{u})}{\tau} + \mathbf{c}\varphi'(x)\frac{(\mathbf{w}^\top\mathbf{u})(\mathbf{w}^\top\mathbf{v})}{\tau^2} + \mathbf{k}\varphi''(x)\frac{(\mathbf{w}^\top\mathbf{u})(\mathbf{w}^\top\mathbf{v})}{\tau^3}.\end{aligned} \tag{203}$$

The parity average is a mixed finite difference:

$$\mathbb{E}_z[y\mathbf{j}_{\mathbf{w}}(\mathbf{c}(z,\mathbf{q}))] = \frac{1}{4}\int_{-1}^{1}\int_{-1}^{1}\{D^2\mathbf{j}_{\mathbf{w}}(\mathbf{c})[\mathbf{c}_1,\mathbf{c}_2] + D\mathbf{j}_{\mathbf{w}}(\mathbf{c})[\mathbf{c}_{12}]\}\,dz_1dz_2.$$

Use $\tau \geq \sqrt{3}\sigma R/(2\sqrt{T})$, $\|\mathbf{k}\|_2 \leq \sigma^2 R/T$, bounded Gaussian density derivatives, and Hölder's inequality in (202). The three Hessian contributions are respectively at most $Cg\kappa_{\mathbf{w}}$, $Cg\kappa_{\mathbf{w}}^2$, and $Cg\kappa_{\mathbf{w}}^2$. The first-derivative contribution is at most $CQ_{\mathcal{S}}/T^2 \leq Cg\sigma Q_{\mathcal{S}}$. Since $\kappa_{\mathbf{w}} \leq C$, their sum is $Cg\kappa_{\mathbf{w}}$. On the bad residual event, bound the original gated vector before differentiating:

$$\mathbb{E}[\max_i \|\mathbf{x}_i\|_2 \mathbf{1}_{\mathcal{E}^c}] \leq (\mathbb{E}\max_i \|\mathbf{x}_i\|_2^2)^{1/2}\mathbb{P}(\mathcal{E}^c)^{1/2} \leq CR_{\mathrm{X}}e^{-cT}.$$

This completes the proof. □

K.3 Aggregate logistic drift

Since $\|\mathbf{a}\|_2 = 1$, for every input $\mathbf{h}$,

$$|f_{\mathbf{W}}(\mathbf{h})| \leq \|\mathbf{W}\|\,\|\mathbf{h}\|_2, \quad \|\nabla_{\mathbf{h}} f_{\mathbf{W}}(\mathbf{h})\|_2 \leq \|\mathbf{W}\|, \quad \|\nabla_{\mathbf{W}} f_{\mathbf{W}}(\mathbf{h})\|_F \leq \|\mathbf{h}\|_2. \tag{204}$$

Let $\mathbf{v} = (v_j)_{j=1}^m$, where $v_j = \mathbf{1}_{\{\mathbf{w}_j^\top \mathbf{h} > 0\}}$. The identity $yr(yf) = y/2 - \tanh(f/2)/2$ gives the population head force

$$\mathbf{G}_{\mathbf{W}} = \tfrac{1}{2}\mathbb{E}[y(\mathbf{a}\odot\mathbf{v})\mathbf{h}^\top] + \mathbf{R}_{\mathbf{W}}, \qquad \|\mathbf{R}_{\mathbf{W}}\|_F \leq \tfrac{1}{4}\|\mathbf{W}\|\,\mathbb{E}\,\|\mathbf{h}\|_2^2.$$

Uniformly on $\|\mathbf{q}\|_2 \leq Q$, $\mathbb{E}\,\|\mathbf{h}\|_2^2 \leq C\sigma^2 d/T$. This follows by integrating $\mathbf{g}_0, b_0$ in (199) on $\mathcal{E}$, then bounding its complement by Gaussian fourth moments and Ce^{-cT}. If every row satisfies $\|\mathbf{w}_j\|_2 \geq \rho/(2\sqrt{m})$ and $|\mathbf{e}^\top \mathbf{w}_j| \leq \|\mathbf{w}_j\|_2/2$, rowwise squaring of Lemma K.2 and Minkowski's inequality imply

$$\|\mathbf{G}_{\mathbf{W}}\|_F \leq Cg\left[\frac{\|\mathbf{W}\mathbf{\Pi}_{\mathcal{S}}\|_F + PQ_{\mathcal{S}}\|\mathbf{W}\mathbf{e}\|_2}{\rho} + \sigma Q_{\mathcal{S}}\right] + C\|\mathbf{W}\|\frac{\sigma^2 d}{T} + CR_{\mathrm{X}}e^{-cT}. \tag{205}$$

For a single row, the coarser parity bound is $(Cg + CR_{\mathrm{X}}e^{-cT})/\sqrt{m}$, and its logistic remainder is at most $C\|\mathbf{W}\|\sigma^2 d/(T\sqrt{m})$.

K.4 Actual simultaneous trajectory control

Lemma K.3. *Under (193)–(195), with probability at least $1-\delta_L$, all $0 \leq t \leq N$ satisfy*

$$\|\mathbf{W}_t - \mathbf{W}_0\|_F \leq C\rho/\sqrt{d}, \quad \|\mathbf{W}_t\| \leq C\rho/\sqrt{d}, \quad \|\mathbf{q}_t\|_2 \leq Q,$$
$$\min_j \|\mathbf{w}_{j,t}\|_2 \geq \rho/(2\sqrt{m}), \qquad |\mathbf{e}^\top \mathbf{w}_{j,t}| \leq \|\mathbf{w}_{j,t}\|_2/2. \tag{206}$$

Proof. Gaussian matrix, vector, and scalar concentration give an initialization event with probability at least $1-\delta_L/8$ on which

$$\|\mathbf{W}_0\| \leq C_2\rho/\sqrt{d}, \qquad \|\mathbf{W}_0\mathbf{\Pi}_{\mathcal{S}}\|_F + \|\mathbf{W}_0\mathbf{e}\|_2 \leq C_2\rho/\sqrt{d},$$
$$3\rho/(4\sqrt{m}) \leq \|\mathbf{w}_{j,0}\|_2 \leq 2\rho/\sqrt{m}, \qquad |\mathbf{e}^\top \mathbf{w}_{j,0}| \leq \rho/(16\sqrt{m}) \quad \text{for every } j. \tag{207}$$

The logarithmic dimension condition absorbs the confidence and row counts. Fix a large C_3 and let $D_* = C_3\rho/\sqrt{d}$. Stop on the first exit from

$$\|\mathbf{W} - \mathbf{W}_0\|_F < D_*, \quad \|\mathbf{q}\|_2 < Q, \quad \|\mathbf{w}_j\|_2 > \rho/(2\sqrt{m}), \quad |\mathbf{e}^\top \mathbf{w}_j| < \|\mathbf{w}_j\|_2/2.$$

We use only these current-state conditions to prove strict slack at any candidate exit time.

The residual event and bounds $\|\mathbf{g}_0\|_2 \leq C\sigma\sqrt{d/T}$, $|b_0| \leq C\sigma\sqrt{d/T}$ imply uniformly over $\|\mathbf{q}\|_2 \leq Q$ that

$$\|\mathbf{h}\|_2 \leq h_* := C_4\sigma\sqrt{d/T}, \qquad \|\mathbf{C}(\mathbf{X},\mathbf{q})\| \leq C_4\sigma^2. \tag{208}$$

Denote this event by $\mathcal{G}$. It depends only on the example, has $\Pr(\mathcal{G}^c) \leq Ce^{-cd}$, and its bounds hold for every $\|\mathbf{q}\|_2 \leq Q$ simultaneously. Put $K(\mathbf{X}) = \max_i \|\mathbf{x}_i\|_2$. Gaussian moments imply, for $r = 1, 2$,

$$\mathbb{E}[K^r \mathbf{1}_{\mathcal{G}^c}] \leq (\mathbb{E}K^{2r})^{1/2}\Pr(\mathcal{G}^c)^{1/2}$$
$$\leq CR_{\mathrm{X}}^r e^{-c_{\mathrm{tail}}d},$$
$$Z_r := \frac{\eta}{B}\sum_{n=0}^{N-1}\sum_{i=1}^{B} K(\mathbf{X}_{ni})^r \mathbf{1}_{\mathcal{G}^c}(\mathbf{X}_{ni}),$$
$$\mathbb{E}Z_r \leq CLR_{\mathrm{X}}^r e^{-c_{\mathrm{tail}}d}.$$

For a sufficiently small universal $\varepsilon_0 > 0$, Markov's inequality and (194) give

$$\Pr\left\{Z_1 > \frac{\varepsilon_0\rho}{\sqrt{d}} \text{ or } Z_2 > \varepsilon_0\sigma^2 L\right\} \le \frac{C}{\varepsilon_0}\left(\frac{LR_{\mathrm{X}}\sqrt{d}}{\rho} + \frac{R_{\mathrm{X}}^2}{\sigma^2}\right)e^{-c_{\mathrm{tail}}d} \le \frac{\delta_L}{8}. \tag{209}$$

Nonnegativity makes these bounds valid for every stopped prefix.

Let τ be the first exit above and write $\mathbf{H}_{ni} = y_{ni}r(y_{ni}f_n(\mathbf{X}_{ni}))\nabla_{\mathbf{W}}f_n(\mathbf{X}_{ni})$, evaluated before the update. By (204), $\|\mathbf{H}_{ni}\|_F \le K(\mathbf{X}_{ni})$ and $\|(\mathbf{H}_{ni})_{j,:}\|_2 \le K(\mathbf{X}_{ni})/\sqrt{m}$. Let $\mathcal{F}_n$ be generated by initialization and the XOR-only batches before iteration n. Under this filtration, define

$$\boldsymbol{\xi}_n = \eta\mathbf{1}_{n<\tau}\left[\frac{1}{B}\sum_{i=1}^{B}\mathbf{H}_{ni}\mathbf{1}_{\mathcal{G}}(\mathbf{X}_{ni}) - \mathbb{E}[\mathbf{H}_{n1}\mathbf{1}_{\mathcal{G}}(\mathbf{X}_{n1}) \mid \mathcal{F}_n]\right], \qquad \mathbf{M}_k = \sum_{n<k}\boldsymbol{\xi}_n.$$

Its increments have norm at most $2\eta h_*$ and conditional second moment at most $C\eta^2 h_*^2$; rowwise the bounds gain factors $m^{-1/2}$ and m^{-1}. The exact stopped update is

$$\mathbf{W}_{k\wedge\tau} - \mathbf{W}_0 = \eta\sum_{n<k}\mathbf{1}_{n<\tau}\mathbb{E}[\mathbf{H}_{n1} \mid \mathcal{F}_n] + \mathbf{M}_k + \mathbf{R}_k^{\mathrm{bad}}, \tag{210}$$

where the last term is the empirical bad-example contribution minus its conditional mean. Uniformly over $k \le N$, (209) gives

$$\begin{aligned}\|\mathbf{R}_k^{\mathrm{bad}}\|_F &\le Z_1 + CLR_{\mathrm{X}}e^{-c_{\mathrm{tail}}d}\\ &\le 2\varepsilon_0\rho/\sqrt{d},\\ \max_j\|(\mathbf{R}_k^{\mathrm{bad}})_{j,:}\|_2 &\le \frac{Z_1 + CLR_{\mathrm{X}}e^{-c_{\mathrm{tail}}d}}{\sqrt{m}}\\ &\le 2\varepsilon_0\rho/\sqrt{md}.\end{aligned}$$

All examples remain in the actual update. The following martingale bound is proved without conditioning on (209).

For bounded centered increments $\boldsymbol{\xi}_t$ in a Hilbert space $\mathcal{H}$, the conditions $\|\boldsymbol{\xi}_t\|_{\mathcal{H}} \le b$ and $\sum_t \mathbb{E}[\|\boldsymbol{\xi}_t\|_{\mathcal{H}}^2 \mid \mathcal{F}_t] \le V$ imply the moment hypothesis of Lemma C.3: for every $r \ge 2$,

$$\mathbb{E}[\|\boldsymbol{\xi}_t\|_{\mathcal{H}}^r \mid \mathcal{F}_t] \le b^{r-2}\mathbb{E}[\|\boldsymbol{\xi}_t\|_{\mathcal{H}}^2 \mid \mathcal{F}_t] \le \frac{r!}{2}b^{r-2}\mathbb{E}[\|\boldsymbol{\xi}_t\|_{\mathcal{H}}^2 \mid \mathcal{F}_t].$$

Consequently

$$\mathbb{P}\left(\max_{k\le N}\left\|\sum_{t<k}\boldsymbol{\xi}_t\right\|_{\mathcal{H}} > \sqrt{2Vx} + 2bx\right) \le 2e^{-x}. \tag{211}$$

Apply this inequality in Frobenius space and to every row, with a union bound over $m+1$ processes. Except with probability $\delta_L/4$, all prefix martingales satisfy

$$\|\mathbf{M}_k\|_F \le Ch_*(\sqrt{\eta L\ell_*} + \eta\ell_*), \qquad \|(\mathbf{M}_k)_j\|_2 \le \frac{Ch_*}{\sqrt{m}}(\sqrt{\eta L\ell_*} + \eta\ell_*). \tag{212}$$

Equation (195) makes these arbitrarily small fixed multiples of $\rho/\sqrt{d}$ and $\rho/\sqrt{md}$. Equation (210) and the bad-example remainder bounds account for every actual update.

Inside the stopped region, $\|\mathbf{W}\| \le (C_2 + C_3)\rho/\sqrt{d}$ and $\|\mathbf{W}\boldsymbol{\Pi}_{\mathcal{S}}\|_F + \|\mathbf{W}\mathbf{e}\|_2 \le C\rho/\sqrt{d}$. Moreover $PQ \le [(P+1)/(\sigma\sqrt{T})](\sigma Q\sqrt{T}) \le c^2$. Thus (205) bounds each population head force by

$$C\left[\frac{g}{\sqrt{d}} + g\sigma Q + \frac{\rho\sigma^2\sqrt{d}}{T}\right] + CR_{\mathrm{X}}e^{-cT}.$$

After multiplication by L, the first and third terms are small multiples of $\rho/\sqrt{d}$. For the second,

$$\frac{g\sigma QL}{\rho/\sqrt{d}} = \frac{gL}{\rho}(\sigma Q\sqrt{T})\sqrt{d/T} \le Cc^2.$$

Together with (212) and the bad-example remainder, this gives $\|\mathbf{W}_k - \mathbf{W}_0\|_F < D_*/2$ at every candidate exit.

For one row, the coarser force bound following (205) gives total population displacement at most

$$\frac{C}{\sqrt{m}}\left[gL + \frac{\rho\sigma^2\sqrt{d}L}{T} + LR_{\mathrm{X}}e^{-cd}\right].$$

Choose the small constants so this, its martingale, and its bad-example remainder sum to less than $\rho/(16\sqrt{m})$. The initialization margins in (207) preserve the row norm and position-angle conditions with strict slack.

For every example, $\|\mathbf{C}(\mathbf{X},\mathbf{q})\| \le K(\mathbf{X})^2$. Separate the good and bad contributions in each actual attention-weight update. Before the stop, $\|\mathbf{W}_n\| \le (C_2+C_3)\rho/\sqrt{d}$, so

$$\begin{aligned}\|\mathbf{q}_{k\wedge\tau}\|_2 &\le \frac{(C_2+C_3)\rho}{\sqrt{d}}(C_4\sigma^2 L + Z_2)\\ &\le \frac{(C_2+C_3)(C_4+\varepsilon_0)\rho\sigma^2 L}{\sqrt{d}}\\ &< \frac{K_0\rho\sigma^2 L}{2\sqrt{d}}\\ &= Q/2,\end{aligned}$$

where $K_0 > 2(C_2+C_3)(C_4+\varepsilon_0)$ is fixed universally. The constant choices are ordered as C_2, C_3, C_4, then K_0, then the small c in (193)–(195). All four exit conditions improve, so no first exit occurs through N. The failure budgets sum to at most δ_L, completing the lemma. □

K.5 Label indistinguishability in the whole attention-weight ball

Lemma K.4 (Scalar Gaussian transport)**.** *Let $Z \sim \mathcal{N}(0,s^2)$, where $s > 0$, and let $u \in C^1(\mathbb{R})$ satisfy $\|u\|_\infty \le A$ and $\|u'\|_\infty \le \kappa < 1/2$, with $A, \kappa \ge 0$. For $T_1(x) = x + u(x)$,*

$$\mathrm{TV}(\mathcal{L}(T_1(Z)), \mathcal{L}(Z)) \le C(A/s + \kappa). \tag{213}$$

Proof. Each map $x \mapsto x + tu(x)$, $0 \le t \le 1$, is a global diffeomorphism with derivative at least $1-\kappa$. Writing φ_s for the density of Z,

$$\begin{aligned}\mathrm{TV}(\mathcal{L}(T_1(Z)), \mathcal{L}(Z)) &= \tfrac{1}{2}\int |\varphi_s(x) - \varphi_s(T_1(x))T_1'(x)|\,dx\\ &\le \tfrac{1}{2}\int |\varphi_s(x) - \varphi_s(T_1(x))|\,dx + \tfrac{\kappa}{2}\int \varphi_s(T_1(x))\,dx\\ &\le \tfrac{A}{2}\int_0^1\int |\varphi_s'(x+tu(x))|\,dx\,dt + \frac{\kappa}{2(1-\kappa)}\\ &\le \frac{A\sqrt{2/\pi}/s + \kappa}{2(1-\kappa)} \le C(A/s+\kappa).\end{aligned}$$

□

Lemma K.5. *If $T \ge Cd$, $P+1 \le c\sigma\sqrt{T}$, and $\sigma Q\sqrt{T} \le c$, then for every deterministic $\|\mathbf{q}\|_2 \le Q$ and every measurable classifier F of the pooled vector,*

$$\mathbb{P}(F(\mathbf{h}(\mathbf{X},\mathbf{q})) \ne y) \ge \frac{1}{2} - \frac{C}{\sigma\sqrt{T}} - Ce^{-cd}. \tag{214}$$

Proof. Decompose perpendicular noises into $\mathbf{g}_0 + \boldsymbol{\zeta}_i$ as before; separately decompose distractor position noises as $b + \Delta_i$, where $\sum_{i\ge 3}\Delta_i = 0$ and $b \sim \mathcal{N}(0, \sigma^2/N_D)$. The variables $\mathbf{g}_0, b$ and the residual collection are independent. Relevant position noises remain exactly zero.

Condition on a residual event of probability at least $1 - Ce^{-cd}$ where

$$\left\|[\boldsymbol{\zeta}_i^\top]_{i\le T}\right\| + \|[\Delta_i]_{i\ge 3}\|_2 \le C_1\sigma\sqrt{T}, \qquad \|\boldsymbol{\zeta}_1\|_2 + \|\boldsymbol{\zeta}_2\|_2 \le C_1\sigma\sqrt{d}.$$

It is independent of $\mathbf{g}_0, b, z$. Set $a = \mathbf{e}^\top\mathbf{q}$, $\beta_c = \mathbf{q}^\top\boldsymbol{\mu}_c$, and

$$\begin{aligned} u_c(z) &= e^{aP+\mathbf{q}^\top\boldsymbol{\zeta}_c+z_c\beta_c}, & A_z &= u_1(z) + u_2(z),\\ B_D &= \sum_{i\ge 3} e^{\mathbf{q}^\top\boldsymbol{\zeta}_i+a\Delta_i}, & \omega_i &= e^{\mathbf{q}^\top\boldsymbol{\zeta}_i+a\Delta_i}/B_D,\\ d_D &= \sum_{i\ge 3}\omega_i\Delta_i, & \mathbf{v}_D &= \sum_{i\ge 3}\omega_i\boldsymbol{\zeta}_i,\\ \mathbf{u}_z &= A_z^{-1}\sum_{c=1}^{2} u_c(z)(\boldsymbol{\zeta}_c + z_c\boldsymbol{\mu}_c). \end{aligned}$$

Retaining the scalar Gaussian inside the softmax, the exact pool is

$$\begin{aligned} S_z(b) &= A_z/(A_z + B_D e^{ab}),\\ \phi_z(b) &= b + d_D + (P - b - d_D)S_z(b),\\ \boldsymbol{\psi}_z(b) &= \mathbf{v}_D + S_z(b)(\mathbf{u}_z - \mathbf{v}_D),\\ \mathbf{h}(\mathbf{X},\mathbf{q}) &= \phi_z(b)\mathbf{e} + \mathbf{g}_0 + \boldsymbol{\psi}_z(b). \end{aligned} \tag{215}$$

Let $B_0 = \sigma\sqrt{d/T}$ and $D_{\mathrm{pool}} = P + B_0 + \sigma^2 Q$. For $|b| \le 3B_0$, every token weight is at most C/T. The normalized distractor weights are at most C/N_D along the segment from zero to $\mathbf{q}$, so their covariance has operator norm at most $C\sigma^2$. Integrating its mean derivative gives

$$|d_D| \le C\sigma^2 Q, \quad \|\mathbf{v}_D\|_2 \le C(\sigma\sqrt{d}/T + \sigma^2 Q), \quad \|\mathbf{u}_z\|_2 \le 1 + C\sigma\sqrt{d}. \tag{216}$$

At $\mathbf{q} = \mathbf{0}$ the two distractor means are respectively zero and $-(\boldsymbol{\zeta}_1 + \boldsymbol{\zeta}_2)/N_D$, which explains the middle bound.

If z' differs from z in one sign, $\log A_z$ changes by at most $2Q$. Differentiating its two-point weighted mean gives

$$|S_z(b) - S_{z'}(b)| \le CQ/T, \qquad \|\mathbf{u}_z - \mathbf{u}_{z'}\|_2 \le C.$$

The hypotheses of this lemma imply $QD_{\mathrm{pool}} \le Cc$ and $Q(1 + \sigma\sqrt{d}) \le Cc$. Since $S_z' = -aS_z(1 - S_z)$, direct differentiation yields

$$\phi_z'(b) = (1 - S_z(b))[1 - aS_z(b)(P - b - d_D)].$$

It follows, throughout $|b| \le 3B_0$, that

$$\begin{aligned} 1/2 \le \phi_z' \le 2, \qquad |\phi_z''| \le CQ/T,\\ |\phi_z - \phi_{z'}| \le CQD_{\mathrm{pool}}/T, \qquad |\phi_z' - \phi_{z'}'| \le CQ/T. \end{aligned} \tag{217}$$

For $|b| \le 2B_0$, define $\mathcal{T}_0(b)$ uniquely by $\phi_{z'}(\mathcal{T}_0(b)) = \phi_z(b)$. Its value lies in $[-3B_0, 3B_0]$: the maps differ by less than $B_0/4$ and their derivatives are at least $1/2$. The inverse function theorem gives

$$|\mathcal{T}_0(b) - b| \le a_0 := CQD_{\mathrm{pool}}/T, \qquad |\mathcal{T}_0'(b) - 1| \le CQ/T.$$

Let χ be a C^1 cutoff equal to one on $[-B_0, B_0]$, zero off $(-2B_0, 2B_0)$, with $|\chi'| \le C/B_0$. Then $\mathcal{T}(b) = b + \chi(b)(\mathcal{T}_0(b) - b)$ is a global increasing diffeomorphism satisfying

$$\|\mathcal{T} - \mathrm{id}\|_\infty \le a_0, \quad \|\mathcal{T}' - 1\|_\infty \le \kappa_{\mathrm{tr}} := C(Q/T + a_0/B_0) < 1/2.$$

Define $\mathbf{d}(b) = \chi(b)[\boldsymbol{\psi}_z(b) - \boldsymbol{\psi}_{z'}(\mathcal{T}(b))]$, extended by zero outside the cutoff support. The preceding estimates give $\|\boldsymbol{\psi}_z - \boldsymbol{\psi}_{z'}\|_2 \le C/T$ and $\left\|\boldsymbol{\psi}_{z'}'\right\|_2 \le CQ(1 + \sigma\sqrt{d})/T \le C/T$, whence $\sup_b \|\mathbf{d}(b)\|_2 \le C/T$.

Apply Lemma K.4 to the scalar transport $\mathcal{T}$.

Let ν be the law of $(\mathbf{g}_0, b)$ and transport it by $\mathcal{T}_{\text{joint}}(\mathbf{g}_0, b) = (\mathbf{g}_0 + \mathbf{d}(b), \mathcal{T}(b))$. Conditional Gaussian translation in $\mathbf{e}^{\perp}$ and (213) give

$$\mathrm{TV}((\mathcal{T}_{\text{joint}})_{\#}\nu, \nu) \le C\left[\frac{1/T}{\sigma/\sqrt{T}} + \frac{a_0}{\sigma/\sqrt{N_D}} + \kappa_{\text{tr}}\right] \le \frac{C}{\sigma\sqrt{T}}. \tag{218}$$

The translation bound follows from the Gaussian KL formula and Pinsker; in orthonormal coordinates on $\mathbf{e}^{\perp}$, the covariance is the nondegenerate matrix $(\sigma^2/T)\mathbf{I}_{d-1}$. For the last inequality set $v = 1/(\sigma\sqrt{T})$ and note

$$a_0/(\sigma/\sqrt{N_D}) \le CvQD_{\text{pool}}, \quad Q/T = v\sigma Q/\sqrt{T}, \quad a_0/B_0 = CvQD_{\text{pool}}/\sqrt{d}.$$

All remaining factors are bounded by the hypotheses.

On $|b| \le B_0$, the transported z' pool equals the original z pool exactly by (215). Outside this interval the probability is at most $2e^{-cd}$. The triangle inequality, data processing, and (218), averaged over the residual event, show

$$\mathrm{TV}(\mathcal{L}(\mathbf{h} \mid z), \mathcal{L}(\mathbf{h} \mid z')) \le C/(\sigma\sqrt{T}) + Ce^{-cd}.$$

Pair the positive-label clusters $\{++, --\}$ with the negative-label clusters $\{+-, -+\}$ by one-sign flips, and use convexity of total variation. The equal-prior Bayes error is $(1 - \mathrm{TV})/2$; all classifiers have at least this error. This proves the lemma. □

K.6 Completion and specialization to the primary range

Proof of Proposition K.1. Lemma K.3 supplies the attention-weight norm and head operator-norm bounds with probability at least $1 - \delta_L$. Lemma K.5 is a deterministic statement for every point of that ball and every measurable classification head. Conditional on the entire realized training trajectory, apply it to each $\mathbf{q}'_n$ and its trained head. The test example is fresh, so conditioning does not change its law. No union bound over training times, heads, or the continuum of attention weights is needed. This proves (196) on the trajectory event. □

Theorem K.6 (XOR-only lower bound in the primary setting)**.** *Under Assumption 3.1 and (15), set $\delta_L = \delta/2$ and use the universal C_H in Theorem A.1. Let*

$$N = H_d = \left\lceil \frac{C_H T}{\eta P} \right\rceil, \qquad R = \eta H_d.$$

With probability at least $1 - \delta_L$, the comparator (191) satisfies

$$\inf_{0 \le n \le H_d} \mathrm{err}'_{\mathrm{X}}(n) \ge \frac{1}{2} - \frac{C}{\sigma\sqrt{T}} - Ce^{-cd} \ge \frac{1}{2} - C\sqrt{\frac{d}{T}} - Ce^{-cd} \ge \frac{1}{4}. \tag{219}$$

Both the attention weight and the head are trained simultaneously throughout this prefix, with all B examples in each batch drawn from $\mathcal{D}_X$.

Proof. We verify Proposition K.1 with $L = R$. The horizon definition, the learning-rate choice in (15), and Assumption 3.1 through Lemma A.3 give

$$R \le C_H T/P + \eta \le CT/P, \qquad d \le m \le d^2, \qquad \log(m/\delta_L) \le C\ell_\delta, \tag{220}$$
$$\ell_* \le C\ell_\delta \le CL_d^{3/2}(\log L_d)^{3/2} \le cd.$$

By Assumption 3.1, write $L_0 = \log d$ and $T = dL_0^{\theta}$, where $17/16 \le \theta \le 9/8$. The geometric conditions follow from

$$\frac{P+1}{\sigma\sqrt{T}} \le CL_0^{(1-\theta)/2} \le CL_0^{-1/32} \le c,$$
$$\sigma Q\sqrt{T} = K_0 \frac{\sigma^3 R\sqrt{T}}{\sqrt{d}} \le C\frac{T^{3/2}}{d^2 P} \le Cd^{-1/2}L_0^{21/16} \le c. \tag{221}$$

Together with $T/d \ge L_0^{17/16}$, these inequalities verify (193). Similarly,

$$gR \le \frac{C}{P\sigma\sqrt{T}} \le CL_0^{-3/8-\theta/2} \le CL_0^{-29/32},$$
$$\frac{\sigma^2 dR}{T} \le C/P \le CL_0^{-3/8}. \tag{222}$$

Their sum is at most the constant in (194) by Assumption 3.1 and Lemma A.3. Its data bounds give $R_X \le CP$, and the initialization has $\rho = 1$, so

$$\begin{aligned}\left(RR_X\sqrt{d} + R_X^2/\sigma^2\right) e^{-c_{\mathrm{tail}}d} &\le C(T\sqrt{d} + dP^2)e^{-c_{\mathrm{tail}}d} \\ &\le Cd^2 e^{-c_{\mathrm{tail}}d} \\ &\le c\exp\{-CL_d^{3/2}(\log L_d)^{3/2}\} \\ &\le c\delta_L.\end{aligned}$$

The third line follows from the dimension consequences of Assumption 3.1 in Lemma A.3 for any prescribed universal C, and the fourth uses (19). Finally, Assumption 3.1 and $R \le CT/P$ give

$$\frac{\rho^2 T}{\sigma^2 d^2 R\ell_*} \ge c\frac{P}{d\ell_\delta},$$
$$\frac{\rho\sqrt{T}}{\sigma d\ell_*} \ge c\frac{\sqrt{T/d}}{\ell_\delta}. \tag{223}$$

The ratios of $\eta \le d^{-3}$ from (15) to these two lower bounds are at most $C\ell_\delta/(d^2P)$ and $Cd^{-3}\ell_\delta\sqrt{d/T}$, respectively. Both are below the universal step-size constant by (19); hence (195) holds. Proposition K.1 proves the first inequality in (219). The remaining inequalities follow from $\sigma^2 \ge 1/d$ and $T/d \ge (\log d)^{17/16}$ in Assumption 3.1, with the dimension comparison supplied by Lemma A.3. □

L Proof of the mixed-training upper bound

Proof of Theorem A.1. Choose the universal checkpoint constant c_* so that the two joint continuations have amplification at most $d^{1/128}$ and the head clock is at most $a_* < 1$. Next choose c_u in Assumption 3.1 sufficiently small for the early comparison below and C_ℓ sufficiently large for the late concentration, Lemma A.3, and all deterministic margin comparisons. These choices are independent of δ.

Apply Proposition D.1 and Theorem H.1, each with its failure allocation $\delta/8$. Their deterministic buffered reference and clock bounds give

$$A_A \asymp \frac{\lambda K_A}{P\sqrt{G}}, \qquad A_X \asymp \frac{\lambda K_X}{P\sqrt{G}} \asymp \lambda c_*. \tag{224}$$

The rounding and quadrature errors are included in those statements. On their common event,

$$D_s(n) + D_N(n) \le CA_n/d + Cd^{-11/8}, \qquad D_p(n) \le Cd^{-11/8}, \qquad 0 \le n \le n_X. \tag{225}$$

For the early prefix, use I_0 from (157). Since $K_A^2 = K_X$, the upper width in Assumption 3.1, together with (224), gives

$$\frac{C\rho A_A(1+K_A^2)/d}{I_0} \le \frac{C\lambda K_A^3\sqrt{m}}{P\sqrt{G}\,d} \le C\sqrt{c_u}, \tag{226}$$
$$\frac{CA_A(1+K_A)/d}{I_0} \le C\sqrt{c_u}/K_A, \tag{227}$$
$$d^{-5/4}/I_0 \le d^{-1/4}.$$

The small-ball threshold in Proposition I.5 is universal. Choose c_u to put the first line below its fixed mean-drift budget. Assumption 3.1, through Lemma A.3, controls the initial-path, centered

characteristic, mesh, and additive tracking budgets in that proposition. It follows, with additional failure at most $\delta/8$, that

$$\inf_{0\le n\le n_A} \mathrm{err}_X(n) \ge \tfrac{1}{4}. \tag{228}$$

The event is uniform over the deterministic early family, so no union bound over SGD iterations is needed.

The auxiliary margin calculation in Section F.9, together with the lower width in Assumption 3.1 via Lemma A.3, gives

$$\begin{aligned}\mathrm{err}_A(n_A) &\le Ce^{-cK_A^2} + CL_d^3 d^{-7/8} + C\frac{P^2 d\ell_\delta}{mK_A^2}\\ &\le Ce^{-cK_A^2} + CL_d^3 d^{-7/8} + Cd^{-57/64}L_d^{-23/2}\\ &\le Ce^{-cK_A^2}.\end{aligned} \tag{229}$$

Indeed, the data bounds in Assumption 3.1 give $P^2/K_A^2 \le CL_d^{1/2}$ and $K_A^2 = c_*(P^2G)^{1/2} \le C\sqrt{L_d \log L_d}$, so Lemma A.3 absorbs both polynomial remainders.

At the late checkpoint set $E_X = mA_X^2K_X^4/d^2$. By (224) and the width interval in Assumption 3.1, written as (16),

$$\begin{aligned}&E_X \asymp \lambda^2 m(P^2G)^2/d^2,\\ &E_X \ge cC_\ell\{(P^2G)^{1/2} + \log(C_{\mathrm{alloc}}/\delta)\},\\ &E_X \le Cc_u(P^2G)^{3/2} \le cd.\end{aligned} \tag{230}$$

The final dimension comparison follows from Lemma A.3 under Assumption 3.1. All comparison constants may depend on the already fixed universal c_*. Increasing C_ℓ ensures the lower threshold of Proposition J.3 with $\zeta = \delta/8$. Its initialization event has additional failure at most $\delta/8$; on intersection with the trajectory events,

$$\begin{aligned}\mathrm{err}_X(n_X) &\le Ce^{-cd^{1/4}} + Ce^{-cK_X^2} + e^{-cE_X} + 2e^{-cK_X^4}\\ &\le Ce^{-c(P^2G)^{1/2}}.\end{aligned} \tag{231}$$

The last inequality uses the data and width consequences of Assumption 3.1 in Lemma A.3 and (230). Together with

$$K_A^2 = K_X = c_*(P^2G)^{1/2}, \tag{232}$$

this proves (17). Fresh-input exceptional sets contribute to the displayed test errors, not to the initialization or training failure probability.

Finally, the data ranges in Assumption 3.1 give $\Gamma_{K_X} - \Gamma_{K_A} = \frac{1}{2}\log K_X \asymp G$. The reference speed and buffer bounds in Proposition D.1 therefore give distinct rounded integers $1 \le n_A < n_X$. Choosing C_H larger than its universal hitting-time constant gives $n_X \le H_d$. The four module failures sum to $4(\delta/8) = \delta/2$. Combining (228) and (231) finishes the proof. □

M More Details on Experiment

M.1 Synthetic Data

Hyperparameters and data. We use ambient dimension $d = 200$, noise scale $\sigma_\epsilon = 0.8$, and deterministic orthogonal signal directions $\boldsymbol{\mu}_1 = 3e_1$, $\boldsymbol{\mu}_2 = 3e_2$, and $\mathbf{p} = |\mathbf{p}|_2 e_3$. The main dynamics experiment uses $T = 200$ and $|\mathbf{p}|_2 = 6.0$. At every training step, we independently generate 256 fresh samples from each of the two token-selection subpopulations and 512 fresh XOR samples. Independent test sets contain 2000 samples per token-selection subpopulation and 2000 XOR samples. Token-selection error is averaged equally over the two subpopulations. Noise is isotropic Gaussian except on the two relevant XOR tokens, where its component along $\mathbf{p}$ is projected out.

Architecture and initialization. The model shares a softmax attention query $\mathbf{q}$ across two separate width-$m = 100$ ReLU heads:

$$\widetilde{\mathbf{x}} = \mathbf{X}^\top \mathbb{S}(\mathbf{X}\mathbf{q}), \qquad f_k(\mathbf{X}) = \mathbf{a}^\top \operatorname{ReLU}(\mathbf{W}_k \widetilde{\mathbf{x}}), \quad k \in \{1, 2\}.$$

The output vector $\mathbf{a}$ is shared by both heads and fixed throughout training, with entries independently sampled from $\{-0.5, +0.5\}$. We initialize $\mathbf{q}(0) = 0$ and draw independent Gaussian matrices $\mathbf{G}_1, \mathbf{G}_2$ with entries distributed as $\mathcal{N}(0, \sigma_w^2)$, where $\sigma_w = 0.05$. The two heads use $\mathbf{W}_1(0) = \mathbf{G}_1$ and $\mathbf{W}_2(0) = \mathbf{G}_2$.

$$\tfrac{1}{2}\mathcal{L}_{\text{selection}} + \tfrac{1}{2}\mathcal{L}_{\text{XOR}},$$

where both losses are averaged over the current batches, with $\mathcal{L}_{\text{selection}}$ weighting the two token-selection subpopulations equally. At each step, $\mathbf{q}$, $\mathbf{W}_1$, and $\mathbf{W}_2$ are updated simultaneously using the same learning rate $\eta = 0.075$. The XOR-only baseline also draws 1024 fresh XOR samples at each step and updates $\mathbf{q}$ and $\mathbf{W}_2$ using $\mathcal{L}_{\text{XOR}}$ alone, with the same initialization, XOR data distribution, test set, learning rate, and training horizon.

Parameter sweeps. We vary one parameter at a time while fixing the remaining parameters at the main-experiment values. The position-signal sweep uses $|\mathbf{p}|_2 \in \{4.0, 5.0, 6.0, 7.0, 8.0\}$. The displayed learning-rate sweep uses $\eta \in \{0.05, 0.0625, 0.075, 0.0875, 0.1\}$, with the same rate applied to all trainable parameters. The sequence-length sweep uses $T \in \{100, 150, 200, 250, 300, 400\}$ while keeping $d = 200$. Each condition is evaluated over 10 random seeds.

$$\tau_k = \min\left\{t \in \{0, \ldots, 1500\} : \operatorname{err}_k(s) \le 0.05 \text{ for all } s \in \{t, \ldots, 1500\}\right\}.$$

If no such step is observed, the corresponding time is right-censored at the training horizon. The evaluation contains 2k data points for each subset of token selection (in total 4k) and 2 k data points of XOR for each seed. Error curves show seed averages with one-standard-deviation shaded bands; characteristic-time error bars show the sample standard deviation across seeds. Conditions with censored XOR times are marked by open triangles, and no complete-case mean is reported for those conditions. Dashed curves in the position-signal panel are least-squares fits against $1/|\mathbf{p}|_2^2$; connecting lines in the learning-rate and sequence-length panels serve only as visual guides.

M.2 MNIST Data

Token construction. Each MNIST image is flattened into a 784-dimensional token, ℓ_2-normalized, and multiplied by signal scale 1.0. Distractor tokens are Gaussian noise vectors with standard deviation 0.10 and are not normalized. We merge the original MNIST train and test pools, split images into 80%/20% train/test, and reserve a separate 20% holdout over digit tuples for out-of-distribution composition evaluation.

Sequence layouts. For the one-hidden-layer dynamics in Figure 2, the sequence length is 6: perception batches contain one signal token and five distractors, composition batches contain two signal tokens and four distractors. For the multi-signal experiment in Figure 3, the sequence length is 8 and the composition task uses $2, 3, 4$, or 5 signal tokens.

one-hidden-layer architecture and optimization. The one-hidden-layer model (Figure 2) computes $\mathbb{S}(\mathbf{X}\mathbf{q})$ with zero-initialized $\mathbf{q}$ and pools $\sum_t \mathbb{S}(\mathbf{X}\mathbf{q})_t \mathbf{x}_t$. The composition head is a width-$m = 256$ ReLU network with output signs fixed at $\pm 1/\sqrt{m}$. We train with SGD, learning rate 1.0, batch size 512, no momentum, no weight decay, for 30000 steps.

Multi-layer architecture and optimization. The multi-signal experiment uses a two-layer Transformer that projects each 784-dimensional token to dimension 128, applies two pre-normalization Transformer blocks with 4 attention heads and feed-forward multiplier 4, and applies a final zero-initialized pooling attention weight before the same perception and composition heads. Trained with Adam, learning rate $3 \cdot 10^{-4}$, batch size 512, no weight decay, with training horizons $10000, 15000, 20000, 30000$ for $2, 3, 4, 5$ signal tokens respectively.

Joint loss. At each step, we draw one perception batch and one composition batch, perform two forward passes through the shared model, and backpropagate $\ell_{\text{percept}} + \ell_{\text{comp}}$, where ℓ_{percept} and ℓ_{comp} are both binary logistic loss.

Evaluation details. The normalized train loss is $(\ell_\tau)/(\ell_0 + 10^{-12})$, with the step-0 value replaced by the first finite loss before normalization. Attention entropy is $-\sum_t w_t \log(w_t + 10^{-8})$ for the pooling weights, averaged over a batch. Signal-token attention mass is the sum of pooling weights at signal-token positions. Shaded bands show a 5-point rolling standard deviation along a single trajectory.

M.3 Emergence in Real LLM Pretraining

Models. All three are dense decoder-only Transformers at the $\sim$7B scale: Pythia-6.9B (Biderman et al., 2023) is GPT-NeoX-style (32 layers, hidden 4096, 32 heads, RoPE, parallel attention/MLP, 50,304-vocab BPE, context 2048), Amber-7B (Liu et al., 2023) follows LLaMA-7B (RMSNorm, SwiGLU, 32,000-vocab tokenizer, context 2048), and OLMo-2-7B (OLMo et al., 2024) adds post-norm RMSNorm, QK-normalization, and z-loss regularization on the logits with a 100,278-vocab tokenizer and context 4096.

Training data and presentation order. None of the three recipes impose a difficulty- or skill-based curriculum on the training stream: Pythia trains on a single global shuffle of The Pile (Gao et al., 2020) ($\sim$300B tokens, $\sim$207B deduplicated, 2049-token packed sequences); Amber consumes $\sim$1.26T tokens from RefinedWeb (Penedo et al., 2023) + RedPajama-v1 (Weber et al., 2024) + StarCoder (Li et al., 2023), pre-shuffled and partitioned into exactly 360 checkpoint chunks (Liu et al., 2023); and OLMo-2 uses a two-stage schedule of $\sim$3.9T tokens on OLMo-Mix-1124 (DCLM (Li et al., 2024), Dolma 1.7 (Soldaini et al., 2024), StarCoder (Li et al., 2023), Proof-Pile-II (Azerbayev et al., 2023)) under document-level shuffle followed by a $\sim$50B-token anneal on Dolmino-Mix-1124 (OLMo et al., 2024) repeated under three different random orders and averaged into a model soup (OLMo et al., 2024); in all cases documents are interleaved by random shuffle rather than sorted by length, domain, or task relevance.

Evaluation task. Following the ElementalTask suite of Liu et al. (2026), we evaluate four representative subtasks that separate localization-style capabilities from computation-style capabilities. *Copy* asks the model to reproduce a given string exactly, testing whether the model can preserve and emit surface-form information from the prompt. *Fact extraction* gives a short passage and asks for an explicitly stated entity, number, or location, testing whether the model can identify the relevant span in context. *Basic arithmetic* presents a one-step natural-language arithmetic question, such as addition or subtraction, and requires the numeric answer. *Math* presents a bare symbolic arithmetic expression and requires evaluating it. We group copy and fact extraction as perception-style tasks because their main burden is selecting or retrieving the relevant input information, while basic arithmetic and math are reasoning-style tasks because they require applying an additional computational rule after the relevant symbols have been identified.