STAGEWISE LEARNING OF BALANCED ORTHOGONAL MIXTURES BY FINITE-WIDTH TWO-HEAD TRANSFORMERS

Anonymous authors
Paper under double-blind review

ABSTRACT

Understanding how attention learns relationships between tokens is central to interpreting Transformers. We study context–signal correspondence learning in balanced orthogonal mixtures. Prior work introduced a three-stage population gradient-flow protocol for this setting. Each input contains three tokens: a context, its matching classification signal, and a distractor. The prior analysis relies on cancellation identities specific to two mixture components ($K = 2$), and its finite-width guarantee uses a sufficient width that grows as the initialization scale decreases. Under the same training setting, we prove high-probability convergence of logistic risk to zero for every fixed number $K \geq 2$ of mixture components in a finite-width two-head softmax Transformer. For fixed K, bias, admissible initial readout-training duration, and confidence, the sufficient width is uniform over sufficiently small positive initialization scales and independent of the target risk. Our analysis combines joint Gram dynamics and relative perturbation estimates to control multidimensional attention growth, and handles asynchronous ReLU crossings along the actual random trajectory. We establish feature separation at the end of attention training. Subsequent readout training yields permanent activation of all ReLU gates and exact logistic-risk and max-margin readout asymptotics. For $K \geq 3$, we further quantify attention-probability gaps, including one-sided cubic lower bounds. Finally, successful learning does not require monotone improvement of the bias-augmented feature margin. At fixed admissible parameters, including a positive initialization scale, the probability of successful learning with an initial margin decrease remains bounded away from zero for all sufficiently large widths. With high probability, the margin nevertheless increases throughout the remaining attention-training interval after a delay that vanishes with width at fixed confidence.

Keywords: Transformers; Softmax attention; Training dynamics; Gradient flow; Feature learning.

1 INTRODUCTION

Attention must use context to distinguish a relevant signal from a distractor. One existing approach analyzes a two-head softmax Transformer in a balanced orthonormal mixture model using prescribed three-stage population gradient flow (Yang et al., 2025, Sections 2–4). In its three-token setting, each input contains a context, its matching classification signal, and a distractor; the matching signal determines the binary label. The number K counts mixture components, or context–signal correspondences, while the number of tokens per input remains three. Balance and orthogonality isolate correspondence learning from frequency imbalance and geometric correlations.

Our guarantees concern this prescribed three-stage population gradient-flow protocol, with Stage 2 attention training restricted to the conflict-conditioned population, where the matching and distracting signals have opposite signs. The complete initialization, trained parameter blocks, objectives, and stage clocks are specified in Sections 2–3.

Prior analysis of this staged Transformer establishes a learning guarantee for $K = 2$, although the model family and a two-head representation extend to arbitrary K. For $K \geq 3$, special two-

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.

component score cancellations no longer hold, so the symmetric reduction does not directly yield a training guarantee. Its finite-width perturbation analysis also uses a sufficient width that grows as the initialization scale shrinks (Yang et al., 2025, Sections 4–6 and Appendix G). The remaining challenge is to prove learning for every fixed K along the actual random factor trajectory, controlling the stage interfaces and gate transitions while keeping the sufficient width uniform over small initialization scales.

We retain this model family and training protocol, and develop controls for these obstacles with sufficient parameter ranges and stage durations specified here. Joint Gram dynamics track all directions in the fastest-growing space, whose dimension is $K(K-1)/2$ per head. A sharp Stage 1 readout estimate and relative perturbation bounds control the logarithmically longer Stage 2 horizon as initialization shrinks. A common crossing orientation in adjacent ReLU gate configurations handles asynchronous transitions and leads to permanent activation. Together, these techniques establish high-probability learning for every fixed $K \geq 2$ in one finite-width two-head network. They also give a sufficient width uniform over small positive initialization scales, under the fixed-parameter choices stated below. Our contributions are as follows.

Contributions. *A complete learning guarantee with uniform sufficient width.* For every fixed $K \geq 2$, a two-head network learns from independent Gaussian key and query factors with high probability (Theorem 4.1). One actual finite-width trajectory reaches separating features, crosses its ReLU gates, and attains permanent activation and vanishing risk without parameter resets. For fixed K, shared positive bias, Stage 1 duration, and confidence, the sufficient width is uniform over sufficiently small positive initialization scales and independent of target risk. A sharp bound on the Stage 1 readout error and relative Gram control support this guarantee over the growing Stage 2 horizon.

Structure of the learned representation and readout. After permanent activation, the actual frozen features determine exact logistic-risk, readout-velocity, and max-margin vector limits, with quantitative eventual signal-coordinate bounds (Theorem 4.2). For $K \geq 3$, Stage 2 endpoint probability comparisons resolve the attention structure through first- and second-order gaps and one-sided cubic bounds (Theorem 4.3).

An initial geometric decrease on successful trajectories. The reference augmented margin grows strictly during Stage 2; on actual trajectories, this margin grows with high probability for the remainder of Stage 2 after a delay vanishing with width at fixed confidence. At fixed admissible parameters and positive initialization scale, an initial-decrease event has probability bounded away from zero and lies inside the learning event under the common width conditions (Theorem 4.4). Successful learning therefore coexists with transient geometric deterioration. The main learning guarantee requires neither the higher-order attention refinement nor whole-stage margin monotonicity.

1.1 Related work

Mixture classification and the staged setting. In joint gradient-descent experiments, Yang et al. (2025, Section 3.1 and Figure 1) observe that readout weights begin aligning with the signals early, while appreciable attention-score changes occur later, coinciding with the decline in conflicting-sample loss. Consistent-sample loss decreases earlier. Attention develops positive associations within matching context–signal pairs and negative associations across groups. These observations motivate their three-stage theoretical protocol: readout alignment, attention learning on conflicting samples, and renewed readout training to reduce the loss further (Yang et al., 2025, Section 3.2 and Algorithm 1). Under this training setting, they establish a finite-width learning guarantee for two mixture components ($K = 2$), with a sufficient width that grows as the initialization scale decreases (Yang et al., 2025, Sections 4–6 and Appendix G). We adopt this experimentally motivated training setting, including conflict-conditioned attention training, with sufficient parameter ranges and stage durations specified here. Routed transformer experts study multicomponent learning through specialization and learned routing (Li et al., 2025).

Phases of attention and feature learning. Chen et al. (2024) analyze warm-up, emergence, and convergence with headwise task allocation in multi-task in-context regression. Yang et al. (2024) establish two phases under simultaneous attention and readout training for word co-occurrence. In a Markovian prediction model, Chen et al. (2026) identify a focus–dilution cycle through stagewise

linearization. These analyses concern phases arising under coupled training. Here the three stages are prescribed by the algorithm. Kim & Suzuki (2024) study nonlinear MLP features followed by linear attention in mean-field and two-timescale limits; our analysis follows finite-width factorized softmax attention through gate transitions.

Parameterization and attention dynamics. Sheen et al. (2024) derive nuclear-norm implicit regularization for separate key and query training under specified data, initialization, and directional-convergence assumptions. Relative query–key and output–value learning speeds affect attention concentration (Vashisht & Ramaswamy, 2026); high-dimensional analyses distinguish direct, tied, and untied attention through symmetry breaking and fast–slow dynamics (Xu et al., 2026). For regression, Goel et al. (2026, Theorem 2) obtain geometric decay of optimization error using an expected-gradient oracle, spectral initialization, regularization, and preconditioning, with a residual finite-context bias term. Here both factors follow ordinary Euclidean gradient flow; direct score optimization and preconditioning define different dynamics.

Margins and the final readout. Ataee Tarzanagh et al. (2023b) connect token selection to attention margins and joint head/attention regularization paths; Ataee Tarzanagh et al. (2023a) distinguish factorized and direct-score parameterizations along vanishing-regularization paths. Max-margin alignment of separable logistic training is classical (Soudry et al., 2018). Here separation and permanent activation are reached by the actual nonlinear trajectory; exact limits and positive signed coordinates then characterize its learned readout. The augmented-margin theorem concerns transient feature geometry during attention learning, including initial decreases on trajectories that subsequently learn.

2 SETTINGS

The orthogonal mixture family, two-head ReLU architecture, initialization scheme, and staged population protocol follow Yang et al. (2025, Section 2 and Algorithm 1). We use balanced three-token support to isolate correspondence learning from frequency and geometric correlations, and specify sufficient parameter ranges and stage clocks for fixed (K, b, T_1).

2.1 DATA DISTRIBUTION AND NOTATION

Fix an integer $K \geq 2$, $d \geq 2K$, and orthonormal vectors $c_1, \ldots, c_K, v_1, \ldots, v_K \in \mathbb{R}^d$. Write

$$U = \operatorname{span}\{c_i, v_i\}_{i=1}^K, \quad c_\Sigma = \sum_i c_i, \quad v_\Sigma = \sum_i v_i, \quad T = I_K - \frac{\mathbf{1}\mathbf{1}^\top}{K}, \quad Q_0 = \frac{\mathbf{1}\mathbf{1}^\top}{K}.$$

In the ordered basis $(c_1, \ldots, c_K, v_1, \ldots, v_K)$ of U, let $J = \operatorname{diag}(I_K, -I_K)$ denote the classification-sign involution. The population is uniform over the $N = 4K(K-1)$ labeled types

$$X = [c_k, y v_k, \epsilon v_j], \qquad k \neq j, \quad y, \epsilon \in \{-1, +1\}, \qquad \text{label } y. \tag{1}$$

Types with $\epsilon = y$ are consistent and those with $\epsilon = -y$ are conflicting; the two subpopulations each have probability $1/2$. Subscripts C, V index the group and classification blocks. Bars denote reference quantities. Norms are Euclidean for vectors, with Frobenius or operator norms specified for matrices. Unless further dependence is stated, constants are selected for fixed (K, b, T_1).

2.2 ARCHITECTURE, OBJECTIVES, AND GATE CONVENTION

Each head $h \in \{+, -\}$ has factors $W_{K,h}, W_{Q,h} \in \mathbb{R}^{m \times d}$ and readout $w_h \in \mathbb{R}^d$. With a fixed shared bias $b > 0$, define

$$P_h = W_{K,h}^\top W_{Q,h}, \qquad Z_h(X) = \sum_{\ell=1}^{3} X \operatorname{softmax}(X^\top P_h x_\ell), \tag{2}$$

$$g_{h,X} = \langle w_h, Z_h(X) \rangle + b, \qquad f(X) = \operatorname{ReLU}(g_{+,X}) - \operatorname{ReLU}(g_{-,X}). \tag{3}$$

There is no additional score scaling. Both factors retain their Euclidean training dynamics; their internal dimension m controls Gram concentration, while the head count stays two. The values are

the input tokens, giving $Z_h \in U$ and $\|Z_h\| \le 3$; summing over all queries gives token-permutation invariance (Remark A.2). The positive bias activates every gate at zero readout initialization.

For $\ell(u) = \log(1 + e^{-u})$, the two objectives are

$$R_{\text{full}} = \frac{1}{N} \sum_{X} \ell(yf(X)), \qquad R_{\text{conf}} = \frac{2}{N} \sum_{X:\epsilon=-y} \ell(yf(X)).$$

Flows have unit learning rates and use the backpropagation convention $D_{\text{ReLU}}(u) = \mathbf{1}\{u > 0\}$, including $D_{\text{ReLU}}(0) = 0$. A *strict trajectory* is locally absolutely continuous and satisfies this specified negative-gradient equation almost everywhere. Uniqueness through gate crossings is proved below.

The gated trainable feature is $F_{\text{gate}} = (\rho_+ Z_+, -\rho_- Z_-)$, with $\rho_h = \mathbf{1}\{g_{h,X} > 0\}$. Appending the affine offset $b(\rho_+ - \rho_-)$ gives F_{aug}, whose geometric margin is defined in (20). Once all gates activate, the offset cancels and the readout acts on $F_{\text{all}} = (Z_+, -Z_-)$. Appendix A summarizes these three feature representations.

3 THREE-STAGE TRAINING ALGORITHM

Algorithm 1 specifies the parameter blocks and objectives; the durations below are selected for our finite-width guarantee.

Initialize $w_+ = w_- = 0$ and all four factors with independent $\mathcal{N}(0, \omega^2/m)$ entries. This places attention near uniform aggregation without prescribing a readout direction. The algorithm passes the actual parameters between stages:

Algorithm 1: Staged population gradient flow
1. Signal alignment. Freeze the factors; train both readouts on R_{full} for time T_1.
2. Attention learning. Freeze the actual readouts; train all four factors on R_{conf} for duration $T_*(\omega)$.
3. Readout convergence. Freeze the actual factors; train both readouts on R_{full} for all subsequent time.

Each stage starts from the preceding stage's actual endpoint, using a single initialization draw without resetting parameters. At zero scores, conflicting classification contributions cancel along the symmetric classification readout; consistent types provide its initial alignment. Conditioning in Stage 2 concentrates attention updates on the unresolved correspondence. Theorem 4.1 establishes separation and permanent gate activation for this prescribed protocol under the sufficient choices below.

Stage clocks. Choose T_1 after the zero-score reference's downward wrong-gate crossing, and define $\mu > 0$ by

$$\begin{aligned} T_1 > T_0 &:= K \left[\frac{b}{2} + \frac{e^{2b} - 1}{4} \right], \\ T_1 - T_0 &= 4K \left[\mu + \frac{e^{2b}(e^{5\mu} - 1)}{5} \right]. \end{aligned} \tag{4}$$

Set the reference endpoint and growth rate to

$$\beta = \frac{b}{2} + 2\mu, \quad \overline{W} = (\mu c_\Sigma + \beta v_\Sigma, \mu c_\Sigma - \beta v_\Sigma), \quad \Lambda_K = \frac{2\mu}{9(K-1)}. \tag{5}$$

For the restricted factors, form each joint Gram matrix

$$C_h = \begin{pmatrix} A_h & P_h^U \\ (P_h^U)^\top & B_h \end{pmatrix} = [W_{K,h}|_U \;\; W_{Q,h}|_U]^\top [W_{K,h}|_U \;\; W_{Q,h}|_U]. \tag{6}$$

The reference trains on R_{conf} with frozen $\overline{W}$, starting from $\overline{C}_h(0) = \omega^2 I_{4K}$, realizable in finite width whenever $m \ge 4K$. Its closed Gram flow defines $T_*(\omega)$ as the first time

$$z(t) := \frac{\operatorname{tr}(T(\overline{P}_+^U(t))_{CV})}{K-1} = c_0 \tag{7}$$

for a sufficiently small fixed $c_0 > 0$. Thus the clock is deterministic and reference-defined, not a sampled-network stopping rule: the actual finite-width factors need not attain contrast exactly c_0 at T_*, but are followed for this same prescribed duration. The learned feature geometry is controlled at the reference contrast; the finite-width argument transfers it to the actual endpoint. Write $R_3(t)$ for the full-population risk in local Stage 3 time t.

4 MAIN RESULTS

We state the learning guarantee, readout limits, attention structure, and transient feature geometry. Learning and positive-margin constants are selected for fixed (K, b, T_1) before ω, confidence, width, the Gaussian draw, and terminal risk. The adverse-event probability uses separate constants selected also for fixed (c_0, ω).

4.1 LEARNING EVERY FIXED MIXTURE

Theorem 4.1 (Finite-width staged learning)**.** *In the setting of Sections 2–3, there exist positive constants $c_0, \omega_K, \rho_K, H_K, \sigma_K, C_w$, depending only on (K, b, T_1), with $\rho_K \le 1$ and $\omega_K \le c_0^{3/2}$, such that the following holds. For any*

$$0 < \omega \le \omega_K, \quad 0 < \delta < 1, \quad m \ge \max\left\{4K, \frac{8}{\rho_K^2}\log\frac{L_K}{\delta}\right\}, \quad L_K = 64K^2 + 16K, \tag{8}$$

write $\eta_m = \sqrt{8\log(L_K/\delta)/m}$. On one initialization event of probability at least $1 - \delta$:

(i) Stage 1 has a unique strict trajectory with finitely many classical pieces. Every wrong-consistent gate crosses downward exactly once within $O(\omega^2\eta_m)$ of T_0; all other gates remain active. Its actual endpoint satisfies

$$\|W^1 - \overline{W}\| \le C_w\omega^2\eta_m.$$

Here a wrong-consistent gate is the positive-head gate when $y = \epsilon = -1$, or the negative-head gate when $y = \epsilon = +1$.

(ii) Stage 2 has a unique classical factor trajectory throughout the prescribed duration, with

$$\frac{1}{\Lambda_K}\log\frac{c_0}{\omega^2} \le T_* \le \frac{1}{\Lambda_K}\log\frac{4c_0}{\omega^2}. \tag{9}$$

Every full-support gate has the mixed sign pattern and absolute slack at least σ_K: both heads are active on conflicts, and only the correct-label head is active on consistent types. At the endpoint the actual gated trainable features have signed margin at least $c_0/(36\sqrt{2K})$ in direction $(c_\Sigma, c_\Sigma)/\sqrt{2K}$.

(iii) Stage 3 has a unique global strict trajectory with finitely many classical pieces. Every wrong-consistent gate crosses upward exactly once in a common finite layer, with no simultaneity requirement. Every other gate remains active. There is no recrossing or Zeno accumulation, and all gates are strictly active for every local Stage 3 time $t \ge H_K$.

(iv) Define the actual frozen labeled features, their minimum-norm hull point, and margin by

$$a_X = y(Z_+(X), -Z_-(X)), \quad v_* = \operatorname*{arg\,min}_{v \in \mathrm{conv}\{a_X\}} \|v\|, \quad \gamma = \|v_*\|. \tag{10}$$

Then $\gamma^2 \ge c_0^2/(162K)$. With

$$\begin{aligned} W_{\mathrm{ref}} &= \sqrt{2K(\mu^2 + \beta^2)}, \qquad R_{\max} = \log 2 + 3\sqrt{2}(W_{\mathrm{ref}} + 1), \\ a_K &= \frac{c_0^2}{162K(1 + NR_{\max})^2}, \qquad R_H = R_3(H_K), \end{aligned} \tag{11}$$

one has $0 < R_H \le R_{\max}$ and, for every $t \ge H_K$,

$$R_3(t) \le \frac{1}{R_H^{-1} + a_K(t - H_K)} \le \frac{1}{R_{\max}^{-1} + a_K(t - H_K)}. \tag{12}$$

In particular, the same fixed-width network satisfies $R_3(t) \to 0$, and its risk is below any $\varepsilon > 0$ at $t = H_K + (a_K\varepsilon)^{-1}$.

For the conflict $X = [c_1, v_1, -v_2]$, zero-score aggregation makes the two classification signals cancel along the symmetric readout. At the leading reference endpoint, both classification queries contribute positively to the difference between heads in the context direction. Their sum creates separation without requiring the two query probabilities to be equal, an identity specific to $K = 2$. The Gram and readout error bounds transfer this reference mechanism to the actual finite-width trajectory. Appendix C.5 gives the calculation and its error control.

4.2 Geometry and asymptotics of the learned readout

Permanent activation connects the nonlinear learning guarantee to logistic readout dynamics on the actual frozen features. The next result identifies their exact risk and vector limits and the signal coordinates that eventually grow.

Theorem 4.2 (Readout asymptotics and signal coordinates)**.** *For every $K \geq 2$, there is a joint choice of the constants and event in Theorem 4.1 for which the following also holds. The unique minimum-norm hard-margin solution is*

$$\theta_{\mathrm{mm}} = \frac{v_*}{\gamma^2} = \operatorname*{arg\,min}_{\theta:\ \langle \theta, a_X \rangle \geq 1\ \forall X} \|\theta\|. \tag{13}$$

For the actual readout pair $W(t) = (w_+(t), w_-(t))$,

$$tR_3(t) \longrightarrow \gamma^{-2}, \qquad t\dot{W}(t) \longrightarrow \theta_{\mathrm{mm}}, \qquad \frac{W(t)}{\log t} \longrightarrow \theta_{\mathrm{mm}}. \tag{14}$$

Let

$$\mathcal{E} = \{(c_i, 0), (0, c_i), (v_i, 0), (0, -v_i) : 1 \leq i \leq K\}, \quad s_0 = \min\left\{\frac{c_0}{9K}, \frac{2c_0^2}{81K^2}\right\}.$$

Then $\langle \theta_{\mathrm{mm}}, e \rangle \geq s_0/36$ for all $e \in \mathcal{E}$. There exists $T_{\mathrm{align},K} > \max\{1, H_K\}$, independent of ω and the draw, such that for every $t \geq T_{\mathrm{align},K}$ and every $e \in \mathcal{E}$,

$$\frac{s_0}{72t} \leq \langle \dot{W}(t), e \rangle, \qquad \frac{s_0}{72} \log t \leq \langle W(t), e \rangle \leq \left(\frac{9\sqrt{2K}}{c_0} + 1\right) \log t. \tag{15}$$

The coordinate bounds apply from $T_{\mathrm{align},K}$ onward and allow unequal actual coordinates.

4.3 Additional attention structure beyond two components

The next result resolves attention comparisons that survive when the two-component cancellation disappears. Appendix D.1 explains their leading terms and the special $K = 2$ cancellation.

Theorem 4.3 (Attention structure for $K \geq 3$)**.** *For $K \geq 3$, there is a joint choice of the constants and initialization event in Theorems 4.1–4.2 for which the following inequalities hold at the actual Stage 2 endpoint. They remain true throughout Stage 3 because the attention factors are frozen. For each conflicting type let*

$$p_{h,a|\ell}(k, y, j) = \left[\operatorname{softmax}(X^\top P_h(T_*) x_\ell)\right]_a.$$

The first index after the head selects the key; the index after the vertical bar selects the query. Simultaneously for every conflicting type,

$$y(p_{+,1|\ell} - p_{-,1|\ell}) \geq \frac{2c_0}{25K}, \qquad \ell = 2, 3, \tag{16}$$

$$p_{+,1|2}(k, +1, j) - p_{+,1|3}(k, +1, j) \geq \frac{7c_0}{225K}. \tag{17}$$

Define $D_\ell = p_{+,2|\ell} - p_{+,3|\ell}$ and $G_1 = y(p_{+,1|1} - p_{-,1|1})$. Then

$$D_2 \leq -\frac{(K-2)c_0^2}{210K}, \qquad D_3 \geq \frac{(K-2)c_0^2}{210K}, \tag{18}$$

$$yD_1 \leq -\frac{(K-2)c_0^3}{352800K}, \qquad G_1 \geq \frac{c_0^3}{529200}. \tag{19}$$

The inequalities in (19) *are one-sided lower bounds on the indicated signed magnitudes. They do not assert $\Theta(c_0^3)$ uniformly over the allowed initialization scales. No assertion of this theorem is required at $K = 2$.*

Quantifiers and width. The witness c_0 is selected for fixed (K, b, T_1) before $\omega, \delta, m, \varepsilon$. The width in (8) contains neither ω nor ε; ρ_K incorporates all fixed-parameter tolerances, without an explicit polynomial dependence on K. Uniform constants give a probability guarantee at each allowed ω, not one event asserted simultaneously for all scales. The common local Stage 3 clocks $H_K, T_{\mathrm{align},K}$ come from compactness. Table 2 summarizes dependencies and selection order. Sufficient total time is $T_1 + T_*(\omega) + H_K + (a_K\varepsilon)^{-1}$, including the logarithmically growing Stage 2 duration.

4.4 Transient feature-margin dynamics

Endpoint separation does not require monotone geometric improvement throughout Stage 2. To study the fixed affine offset as well as the gated features, define

$$\begin{aligned} F_{\mathrm{aug}}(X;t) &= \big(\rho_+ Z_+, -\rho_- Z_-, b(\rho_+ - \rho_-)\big), \\ \gamma_{\mathrm{aug}}(t) &= \max_{u\in\mathbb{R}^{2d+1}:\,\|u\|=1} \min_X y\langle u, F_{\mathrm{aug}}(X;t)\rangle. \end{aligned} \tag{20}$$

This margin measures the separation geometry of the augmented feature set, allowing the last coefficient of the separating vector u to be optimized. The actual output uses the fixed last coefficient 1; the bias is not additionally trained. The quantity tests geometric monotonicity during attention learning, and its decrease alone does not imply higher loss or lower classification accuracy. It differs from the gated and final homogeneous readout margins in Definition A.1.

Theorem 4.4 (Reference growth, initial decreases, and delayed growth)**.** *For each fixed $K \geq 2$, there exist constants satisfying Theorem 4.1 and positive $B_{\mathrm{post}}, \varepsilon_{\mathrm{post}}$ for which the following hold. Write $\bar{F}_{\mathrm{aug}}$ and $\bar{\gamma}_{\mathrm{aug}}$ for the augmented feature and margin along the exact isotropic reference defined in Section 3. Put $h_{\mathrm{geom}} = \sqrt{2/K}/9$ (a geometric constant, unrelated to the factor width m) and use local Stage 2 time. For every allowed (ω, δ, m) satisfying Theorem 4.1, parts (i)–(ii) hold on its stated initialization event; part (iii) has the separate fixed-parameter quantifiers displayed below.*

(i) On the exact isotropic reference, $\bar{\gamma}_{\mathrm{aug}}(0) = 0$ and

$$\bar{\gamma}'_{\mathrm{aug}}(t) \geq \tfrac{1}{2} h_{\mathrm{geom}} \Lambda_K \omega^2 \cosh(\Lambda_K t) > 0 \qquad (0 < t \leq T_*).$$

(ii) On the same initialization event as Theorem 4.1, define the deterministic delay

$$\tau_m = -\frac{1}{2\Lambda_K} \log\left(1 - \frac{B_{\mathrm{post}}\eta_m}{\varepsilon_{\mathrm{post}}}\right), \qquad 0 < \tau_m \leq \frac{B_{\mathrm{post}}\eta_m}{\Lambda_K \varepsilon_{\mathrm{post}}} < T_*. \tag{21}$$

Here ρ_K is chosen so the logarithm's ratio is at most $1/2$. Throughout $[\tau_m, T_*]$,

$$\gamma_{\mathrm{aug}}(t) \geq \tfrac{1}{2} h_{\mathrm{geom}} \omega^2 \sinh(\Lambda_K t) \qquad (\tau_m \leq t \leq T_*), \tag{22}$$

$$\frac{d^+}{dt}\gamma_{\mathrm{aug}}(t) \geq \tfrac{1}{4} h_{\mathrm{geom}} \Lambda_K \omega^2 \cosh(\Lambda_K t) \qquad (\tau_m \leq t < T_*). \tag{23}$$

Thus $\tau_m = O_{K,b,T_1}(\sqrt{\log(L_K/\delta)/m})$, with sufficient width uniform in ω.

(iii) For each fixed allowed (c_0, ω), there exist $p_{\mathrm{adv}} > 0$ and m_{adv}, depending only on fixed (K, b, T_1, c_0, ω) and the reference constants, but not on d, the orthonormal embedding, or a requested confidence level, such that every $m \geq m_{\mathrm{adv}}$ has probability at least p_{adv} of an actual trajectory with $\gamma_{\mathrm{aug}}(0) > 0$ and $\frac{d^+}{dt}\gamma_{\mathrm{aug}}(0) < 0$. Its margin strictly decreases on a positive initial interval, while at the prescribed endpoint

$$\min_X y \left\langle (c_\Sigma, c_\Sigma, 0)/\sqrt{2K}, F_{\mathrm{aug}}(X;T_*)\right\rangle \geq c_0/(36\sqrt{2K}).$$

Here d^+/dt denotes the right derivative, whose existence at positive-margin times follows from Lemma I.1. Let E_m be the initial-decrease event constructed in Appendix I.1, and $\mathcal{G}_{m,\delta}$ the learning event. That appendix proves $E_m \subseteq \mathcal{G}_{m,\delta}$. Under the common width requirement

$$m \geq \max\left\{m_{\mathrm{adv}}, 4K, \frac{8}{\rho_K^2}\log\frac{L_K}{\delta}\right\},$$

every trajectory in E_m initially loses augmented margin, grows throughout $[\tau_m, T_*]$, and has $R_3(t) \to 0$. Its probability is at least p_{adv} for the fixed parameters of part (iii). Thus successful learning can include an initial geometric decrease. Increasing width shortens the interval before guaranteed growth at fixed confidence, but cannot give arbitrary confidence in whole-stage nondecrease. On this constructed event, the initial margin is $\Theta(\omega^2/\sqrt{m})$ at fixed parameters. Its total drop during the positive decreasing interval supplied by Appendix I.1 is bounded above by $O(\omega^2/\sqrt{m})$; a matching lower bound on the drop or its duration is not asserted.

5 Proof outline

We first control the actual Stage 1 interface through all adjacent gate configurations, obtaining $\|W^1 - \overline{W}\| = O(\omega^2\eta_m)$. We then propagate its error across the full growing Gram space and the logarithmic Stage 2 horizon. Write $C = (C_+, C_-)$ with the product Frobenius norm. On the conflict-active branch, put $N_h = \nabla_{P_h^U} R_{\mathrm{conf}}$ and $S_h = \left(\begin{smallmatrix} 0 & N_h \\ N_h^\top & 0 \end{smallmatrix}\right)$. Superscript 0 denotes evaluation at zero scores and frozen $\overline{W}$. Euclidean factor training gives

$$\dot{C}_h = -S_h C_h - C_h S_h, \tag{24}$$

$$N_+^0 = \begin{pmatrix} 0 & -\alpha T \\ \alpha T/2 & 0 \end{pmatrix}, \quad N_-^0 = -N_+^0, \quad \alpha = \frac{\mu}{9(K-1)}. \tag{25}$$

In particular, $\dot{P}_h^U = -A_h N_h - N_h B_h$: the evolving self-Grams enter the score dynamics. The zero-score linearization has top rate $\Lambda_K = 2\alpha$ and a $K(K-1)/2$-dimensional top eigenspace per head; its full semigroup bound controls arbitrary Gram perturbations. Put $q(t) = \omega^2 e^{\Lambda_K t}$, $d_0 = \|C(0) - \overline{C}(0)\|_F$, and $d_w = \|W^1 - \overline{W}\|$, with time zero at Stage 2. Duhamel estimates give the logarithmic first hit and

$$e = \frac{d_0}{\omega^2} + (1 + T_*)d_w, \qquad \|C(t) - \overline{C}(t)\|_F \le B_e\, e\, q(t) \quad (0 \le t \le T_*). \tag{26}$$

Why the width is uniform in the initialization scale. Gaussian concentration gives $d_0/\omega^2 = O(\eta_m)$, while the Stage 1 interface estimate gives $d_w = O(\omega^2\eta_m)$. Combining it with (9), for constants fixed with (K, b, T_1),

$$(1 + T_*)d_w \le C\eta_m\omega^2\big(1 + \log(1/\omega)\big) \le C'\eta_m. \tag{27}$$

Hence $e = O(\eta_m)$ uniformly over the allowed scales, and (26) keeps the Gram error proportional to $q(t)$ over the entire growing time window. An $O(\eta_m)$ interface bound alone would give $O(\eta_m[1 + \log(1/\omega)])$ through the same estimate, which does not yield a width bound uniform in ω. The sharper interface control and relative Gram estimate together preserve the width quantifier.

At the reference first hit, put $z = c_0$ and $r = \mathrm{tr}(T(\overline{P}_+^U(T_*))_{VC})/(K-1)$. The endpoint estimate is

$$P^*(z, r) = \begin{pmatrix} 0 & zT \\ rT & 0 \end{pmatrix},$$
$$|r| + \|\overline{P}_+^U(T_*) - P^*\|_{\mathrm{op}} + \|\overline{P}_-^U(T_*) + P^*\|_{\mathrm{op}} \le C_P z^2. \tag{28}$$

The mixed gate pattern gives the full-support gated separator; hull perturbation yields $\gamma^2 \ge c_0^2/(162K)$. In Stage 3, positive crossing orientations in every adjacent gate configuration permit asynchronous returns to an invariant all-active basin and exclude recrossing. Equal biases then cancel. On these actual frozen features, the closest-point inequality bounds risk, and the logistic-weighted feature barycenter converges to v_*. This gives $\theta_{\mathrm{mm}} = v_*/\gamma^2$ and all three limits in (14). Eventual coordinate bounds follow from the minimum-norm point $\overline{v}_*$ of the reference hull $\mathrm{conv}\{y(\overline{Z}_+(X), -\overline{Z}_-(X))\}$ at the Stage 2 endpoint. Its coordinate decomposition and reserves are

$$\overline{v}_* = (A_{\mathrm{hull}} c_\Sigma + B_{\mathrm{hull}} v_\Sigma,\ A_{\mathrm{hull}} c_\Sigma - B_{\mathrm{hull}} v_\Sigma),$$
$$A_{\mathrm{hull}} = \frac{2z}{9K} + O(z^2), \qquad B_{\mathrm{hull}} = \frac{4z^2}{81K^2} + O(z^3). \tag{29}$$

Appendices A–B establish the crossing and Stage 1 interface estimates, Appendix C proves Gram growth and separation, and Appendix E gives permanent activation and the readout tail. Appendix G assembles the common constants; Appendix I treats transient margins. The attention refinement and saved-matrix certificates are in Appendices D and F, respectively.

6 NUMERICAL EVIDENCE

Appendix H reports numerical methods, tables, and complete cohorts. These experiments explore the predicted mechanisms at attention contrasts beyond the proof's displayed sufficient construction; Appendix H.5 explains the settings and purposes of these comparisons. Unless stated otherwise, $K = 3$. Table 1 summarizes the core confirmation results and widths.

The initialization-scale, width, and margin scans each use 20 seed groups, paired within that scan across its scales or widths. The scale scan rescales a common Gaussian draw, and the width scan uses nested row prefixes; configurations within a seed group are therefore dependent. The five readout continuations use specified width-scan endpoints. Complete exploratory cohorts and unsuccessful fine-attention comparisons remain in Appendix H.

Table 1: Core $K = 3$ confirmation results. Endpoint success means a positive separation witness and correct classification of all 24 population types. Width-scan counts follow increasing m.

Diagnostic	Width m and design	Saved result
Initialization scales (Theorem 4.1)	262144; 20 seeds at five scales	$100/100$ endpoint successes.
Fine attention (Theorem 4.3)	$2^8, 2^{10}, \ldots, 2^{18}$; 20 seeds per width	$120/120$ endpoint successes; simultaneous bound counts $1, 1, 2, 9, 18, 20$ out of 20.
Readout tail (Theorem 4.2)	262144; five continuations	About 2.33% vector error at log time $s = 512$.
Augmented margin (Theorem 4.4)	$64, 262144$; 20 paired seeds	First-sample decreases in $12/20$, $9/20$ cases; $40/40$ endpoint successes.

The width-scan endpoints also give $720/720$ correct context-switch reversals. The readout continuations track risk, velocity, and vector asymptotics; $s = 512$ corresponds to $t \approx 2.28 \times 10^{222}$. Finite margin samples do not certify initial derivatives or whole-interval growth. Reference-path and cross-component comparisons provide additional context.

7 DISCUSSION AND CONCLUSION

Under the conditions of Theorem 4.1, the prescribed two-head factor dynamics learn every fixed-K mixture with high probability. The complete guarantee links random initialization, multidimensional attention growth, actual gate crossings, and vanishing risk with a sufficient width uniform over small initialization scales. The actual frozen features also determine the exact risk and readout limits of Theorem 4.2, while for $K \geq 3$ the attention endpoint satisfies the probability-gap bounds of Theorem 4.3. On a set of successful trajectories whose probability stays bounded below at fixed admissible parameters, the augmented feature margin initially decreases before entering the guaranteed Stage 2 growth interval. Successful representation learning therefore coexists with transient deterioration of this feature geometry.

The results concern the prescribed staged population algorithm on balanced, orthogonal, three-token support, with a fixed positive bias and zero initial readouts. Constants and common clocks are selected for fixed (K, b, T_1); the bounds do not specify polynomial dependence on K.

AI USE STATEMENT

AI tools were used to polish the language of the manuscript. The authors take responsibility for all content.

REPRODUCIBILITY STATEMENT

Sections 2–3 specify the population, parameterization, initialization, objectives, and clocks. The appendices contain the complete proofs and numerical methods. The selected supplement supplies the compact small-width endpoint evidence, 30 exploratory and 120 confirmation width-scan configurations, 100 initialization-scale configurations, the 24 reference-path configurations (including six reused source records), 40 confirmation augmented-margin trajectories with the separate exploratory cohort, five readout continuations, and nine cross-component trajectories. Context-switch evaluations and readout continuations reuse identified endpoints. The evidence index distinguishes saved-array recomputation, exported historical summaries, retained-experiment matrix certificates, and integration code. No saved-matrix certificate encloses the continuous ODE trajectory.

REFERENCES

Davoud Ataee Tarzanagh, Yingcong Li, Christos Thrampoulidis, and Samet Oymak. Transformers as Support Vector Machines. *arXiv preprint arXiv:2308.16898*, 2023a. doi: 10.48550/arXiv.2308.16898. URL https://neurips.cc/virtual/2023/81690. NeurIPS 2023 M3L Workshop poster; full version 3, revised February 2024: https://arxiv.org/abs/2308.16898v3.

Davoud Ataee Tarzanagh, Yingcong Li, Xuechen Zhang, and Samet Oymak. Max-Margin Token Selection in Attention Mechanism. In *Advances in Neural Information Processing Systems*, volume 36, pp. 48314–48362, 2023b. doi: 10.52202/075280-2096. URL https://proceedings.neurips.cc/paper_files/paper/2023/hash/970f59b22f4c72aec75174aae63c7459-Abstract-Conference.html.

Siyu Chen, Heejune Sheen, Tianhao Wang, and Zhuoran Yang. Training Dynamics of Multi-Head Softmax Attention for In-Context Learning: Emergence, Convergence, and Optimality. In *Conference on Learning Theory*, volume 247 of *Proceedings of Machine Learning Research*, pp. 4573. PMLR, 2024. URL https://proceedings.mlr.press/v247/siyu24a.html. Extended abstract; full version: https://arxiv.org/abs/2402.19442v2; arXiv version DOI: https://doi.org/10.48550/arXiv.2402.19442.

Zheng-An Chen, Pengxiao Lin, Zhi-Qin John Xu, and Tao Luo. Focus and Dilution: The Multi-stage Learning Process of Attention. In *International Conference on Machine Learning*, 2026. URL https://icml.cc/virtual/2026/poster/64653. Accepted; preprint: https://arxiv.org/abs/2605.01199v1; arXiv version DOI: https://doi.org/10.48550/arXiv.2605.01199.

Rick Durrett. *Probability: Theory and Examples*. Cambridge University Press, fifth edition, 2019. doi: 10.1017/9781108591034. URL https://www.cambridge.org/core/books/probability/DD9A1907F810BB14CCFF022CDFC5677A/listing.

Gautam Goel, Mahdi Soltanolkotabi, and Peter Bartlett. Training Dynamics of Softmax Self-Attention: Fast Global Convergence via Preconditioning. *arXiv preprint arXiv:2603.01514*, 2026. doi: 10.48550/arXiv.2603.01514. URL https://arxiv.org/abs/2603.01514v1.

Juno Kim and Taiji Suzuki. Transformers Learn Nonlinear Features In Context: Nonconvex Mean-field Dynamics on the Attention Landscape. In *International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 24527–24561. PMLR, 2024. URL https://proceedings.mlr.press/v235/kim24af.html. arXiv version DOI: https://doi.org/10.48550/arXiv.2402.01258.

Hongbo Li, Qinhang Wu, Sen Lin, Yingbin Liang, and Ness B. Shroff. Mixture-of-Transformers Learn Faster: A Theoretical Study on Classification Problems. *arXiv preprint arXiv:2510.27004*, 2025. doi: 10.48550/arXiv.2510.27004. URL https://arxiv.org/abs/2510.27004v1.

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. doi: 10.48550/arXiv.2403.08699. URL https://arxiv.org/abs/2403.08699v1.

Daniel Soudry, Elad Hoffer, Mor Shpigel Nacson, Suriya Gunasekar, and Nathan Srebro. The Implicit Bias of Gradient Descent on Separable Data. *Journal of Machine Learning Research*, 19 (70):1–57, 2018. URL https://jmlr.org/papers/v19/18-188.html. arXiv version DOI: https://doi.org/10.48550/arXiv.1710.10345.

Rahul Vashisht and Harish G. Ramaswamy. Faster Query-Key Learning Sharpens Attention in Self-Attention Models. In *International Conference on Machine Learning*, 2026. URL https://icml.cc/virtual/2026/poster/60639. Accepted; preprint: https://arxiv.org/abs/2608.06776v1; arXiv version DOI: https://doi.org/10.48550/arXiv.2608.06776.

Yizhou Xu, Margarita Sagitova, Lenka Zdeborová, and Florent Krzakala. High-Dimensional Learning Dynamics of Attention-Indexed Models. *arXiv preprint arXiv:2609.03858*, 2026. doi: 10.48550/arXiv.2609.03858. URL https://arxiv.org/abs/2609.03858v1.

Hongru Yang, Bhavya Kailkhura, Zhangyang Wang, and Yingbin Liang. Training Dynamics of Transformers to Recognize Word Co-occurrence via Gradient Flow Analysis. In *Advances in Neural Information Processing Systems*, volume 37, pp. 46047–46117, 2024. doi: 10.52202/079017-1465. URL https://proceedings.neurips.cc/paper_files/paper/2024/hash/520416e27d3b0cef3cd70a083e2991c7-Abstract-Conference.html.

Hongru Yang, Zhangyang Wang, Jason D. Lee, and Yingbin Liang. Transformers Provably Learn Two-Mixture of Linear Classification via Gradient Flow. In *International Conference on Learning Representations*, pp. 8889–8946, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/hash/1904149f1dc50ffae420983c090fb727-Abstract-Conference.html.

A READOUT FEATURES, PERMUTATION INVARIANCE, AND DIRECTED CROSSINGS

A.1 READOUT FEATURES AND TOKEN-ORDER INVARIANCE

Definition A.1 (Gated and frozen readout features)**.** Write $Z_h(X;t)$ for the aggregate at the current parameter state and $W(t) = (w_+(t), w_-(t))$. Define the gate indicators and trainable gated features by

$$\rho_{h,X}(t) = \mathbf{1}\{g_{h,X}(t) > 0\}, \qquad F_{\mathrm{gate}}(X;t) = \big(\rho_{+,X}(t)Z_+(X;t), -\rho_{-,X}(t)Z_-(X;t)\big).$$

Their full-support unit-norm feature margin is

$$\gamma_{\mathrm{gate}}(t) = \max_{\substack{\theta\in\mathbb{R}^{2d}\\ \|\theta\|=1}} \min_X y\langle\theta, F_{\mathrm{gate}}(X;t)\rangle,$$

where the minimum runs over every labeled type in (1). The model output is

$$f_t(X) = \langle W(t), F_{\mathrm{gate}}(X;t)\rangle + b\big(\rho_{+,X}(t) - \rho_{-,X}(t)\big).$$

The last term is a fixed-bias affine offset within each gate pattern; it is not a trainable feature coordinate.

At the local Stage 2 endpoint $T_* = T_*(\omega)$ defined in equation (7), define separately

$$F_{\mathrm{all}}(X) = \big(Z_+(X;T_*), -Z_-(X;T_*)\big).$$

These features remain frozen throughout Stage 3. After all gates become active, the biases cancel and the actual output is the homogeneous readout $f_t(X) = \langle W(t), F_{\mathrm{all}}(X)\rangle$.

The endpoint margin in Theorem 4.1(ii) concerns F_{gate} and is witnessed by the unit direction stated there. The quantity γ in Theorem 4.1(iv) and Theorem 4.2 concerns the labeled frozen features $a_X = yF_{\mathrm{all}}(X)$. These are the features governing the homogeneous readout after permanent activation of all gates. The three feature geometries have distinct roles:

- F_{gate}: the fixed affine offset lies outside the feature; this geometry supplies the Stage 2 fixed-direction separator.
- F_{aug}: its last coordinate is optimized in the geometric margin, while its coefficient in the model remains fixed at 1; this geometry governs the transient margin in Theorem 4.4.
- F_{all}: equal biases cancel after full activation; these frozen features govern the readout limits in Theorem 4.2.

Remark A.2 (Token-order invariance). The canonical order in (1) fixes notation. For arbitrary positive integers d, L, m and $b > 0$, consider equally weighted all-query aggregation with no positional encodings, masks, or position-dependent query weights. Write $S(B)_{a,\ell} = e^{B_{a,\ell}}/\sum_{r=1}^{L} e^{B_{r,\ell}}$ for $B \in \mathbb{R}^{L\times L}$ and $Z_P(X) = XS(X^\top PX)\mathbf{1}_L$ for $X \in \mathbb{R}^{d\times L}$, $P \in \mathbb{R}^{d\times d}$. If $\Pi e_j = e_{\sigma(j)}$ is a permutation matrix, then

$$S(\Pi^\top B\Pi) = \Pi^\top S(B)\Pi, \qquad Z_P(X\Pi) = Z_P(X).$$

In any finite indexed labeled population (X_i, y_i), replace fixed masses $p_i \geq 0$, $\sum_i p_i = 1$, by fixed label-preserving permutation masses $p_{i,\Pi} \geq 0$ with $\sum_\Pi p_{i,\Pi} = p_i$. Aggregates, gates, outputs, logistic losses, and strict per-type fields agree at every state, including gate zeros. Averaging preserves full risks and fields, and normalized conditional risks and fields on any fixed positive-mass subset of canonical indices with all permutation copies included. This covers both risks here; the conflicting subset has mass $1/2$.

With identical initialization and prescribed trained and frozen blocks, the locally absolutely continuous strict solution sets coincide. Unique paths and their reference clocks, including first contrast hits, are unchanged. Attention indices retain semantic roles: new entry (a, ℓ) is old entry $(\sigma(a), \sigma(\ell))$. No factor symmetry is assumed.

Proof. The permutation convention gives $(\Pi^\top B\Pi)_{a,\ell} = B_{\sigma(a),\sigma(\ell)}$. Reindexing the denominator gives $S(\Pi^\top B\Pi)_{a,\ell} = e^{B_{\sigma(a),\sigma(\ell)}}/\sum_r e^{B_{\sigma(r),\sigma(\ell)}} = S(B)_{\sigma(a),\sigma(\ell)}$. Substitution of $B = X^\top PX$, using $\Pi\Pi^\top = I_L$ and $\Pi\mathbf{1}_L = \mathbf{1}_L$, proves aggregate invariance.

Positive softmax denominators make this identity smooth. Differentiation through $P_h = W_{K,h}^\top W_{Q,h}$ gives equal factor derivatives; equal aggregates give equal readout derivatives. Thus gate values and all parameter derivatives agree. With $q(u) = (1 + e^u)^{-1}$, the strict field in coordinate θ_r is $y_i q(y_i f)[D_{\mathrm{ReLU}}(g_+)\partial_{\theta_r} g_+ - D_{\mathrm{ReLU}}(g_-)\partial_{\theta_r} g_-]$. Its terms agree even at zeros because $D_{\mathrm{ReLU}}(0) = 0$; an absent parameter has zero gate derivative.

For each common loss or field coordinate H_i, fixed-mass averaging gives $\sum_i \sum_\Pi p_{i,\Pi} H_i = \sum_i p_i H_i$. Restriction to the selected indices and division by their unchanged positive mass proves the conditional assertion, including repeated indexed copies. Freezing sets the same field coordinates to zero. Hence either equation is satisfied almost everywhere by exactly the same paths, proving the solution-set and clock assertions. □

A.2 Directed crossings under arbitrary finite feature perturbations

For two feature families indexed by head h and labeled type i, write $d_Z := \max_{h,i} \|Z_{h,i} - \bar{Z}_{h,i}\|$. For fixed gate bits ρ, put

$$a_i^\rho = y_i(\rho_{+,i} Z_{+,i}, -\rho_{-,i} Z_{-,i}), \qquad c_i^\rho = y_i b(\rho_{+,i} - \rho_{-,i}), \qquad F_Z^\rho(W) = \frac{1}{N} \sum_i \frac{a_i^\rho}{1 + e^{\langle W, a_i^\rho \rangle + c_i^\rho}}.$$

Every such smooth extension satisfies $\|F_Z^\rho\| \le M_0 = 3\sqrt{2}$ and $\|DF_Z^\rho\| \le L_0 = 9/2$. If $\|V\| \le R_0$, then

$$\|F_Z^\rho(W) - F_{\bar{Z}}^\rho(V)\| \le L_0 \|W - V\| + B_0 d_Z, \qquad B_0 = \sqrt{2} + 3R_0/2. \tag{30}$$

Indeed, $\|a_i^\rho\| \le 3\sqrt{2}$, $\|a_i^\rho(Z) - a_i^\rho(\bar{Z})\| \le \sqrt{2} d_Z$, and the logistic multiplier has derivative bounded by $1/4$.

Lemma A.3 (Finite directed event)**.** *Assume $\|Z_{h,i}\|, \|\bar{Z}_{h,i}\| \le 3$. Fix a locally absolutely continuous reference path $\bar{W} : [0, H] \to \mathbb{R}^{2d}$, a nonempty finite fragile set $\mathcal{A}$, an orientation $\eta \in \{-1, 1\}$, and $\tau \in (0, H)$. Let ρ^- and ρ^+ set every nonfragile bit to one and each fragile bit to $(1 - \eta)/2$ and $(1 + \eta)/2$, respectively. Require that the reference obey the same cell equations being compared:*

$$\dot{\bar{W}} = F_{\bar{Z}}^{\rho^-}(\bar{W}) \quad \text{a.e. on } (0, \tau), \qquad \dot{\bar{W}} = F_{\bar{Z}}^{\rho^+}(\bar{W}) \quad \text{a.e. on } (\tau, H).$$

Every nonfragile gate is positive on $[0, H]$. For every fragile gate a, require $\eta g_a(\bar{W}(t), \bar{Z}) < 0$ before τ, $g_a(\bar{W}(\tau), \bar{Z}) = 0$, and $\eta g_a(\bar{W}(t), \bar{Z}) > 0$ after τ. Suppose every assignment of the fragile bits in the full cube $\{0, 1\}^{|\mathcal{A}|}$, with the nonfragile bits fixed, satisfies at the crossing point

$$\eta \langle \bar{Z}_{h,i}, (F_{\bar{Z}}^\rho(\bar{W}(\tau)))_h \rangle > 0, \qquad (h, i) \in \mathcal{A}.$$

There exist $\epsilon_0, C_t, C_W > 0$, depending only on the fixed reference data and path, such that every $\epsilon = d_Z + \|W(0) - \bar{W}(0)\| \le \epsilon_0$ gives a unique locally absolutely continuous strict solution on $[0, H]$. Each fragile gate has exactly one crossing time t_a, with $|t_a - \tau| \le C_t \epsilon$, and $\sup_{0 \le t \le H} \|W(t) - \bar{W}(t)\| \le C_W \epsilon$. Every nonfragile gate remains positive. There are at most $|\mathcal{A}|$ distinct event times; simultaneous crossings are permitted.

Proof. Finiteness and continuity give a state ball of radius r and a feature neighborhood in which all nonfragile gates are positive and every oriented fragile normal velocity in every adjacent cell is at least $\nu > 0$. Choose a fixed outer window $[\tau - \sigma, \tau + \sigma] \subset (0, H)$ with $M_0 \sigma < r/8$. Outside this window the reference has a positive minimum absolute gate slack.

Compare only the incoming smooth extensions until their first event. Taking $R_0 = \|\bar{W}(0)\| + M_0 H$, equation (30) gives

$$\|W^-(t) - \bar{W}^-(t)\| \le C_{\mathrm{pre}} \epsilon, \quad C_{\mathrm{pre}} = e^{L_0 H}(1 + B_0 H), \quad |g_a(W^-, Z) - g_a(\bar{W}^-, \bar{Z})| \le G\epsilon,$$

where $G = 3C_{\mathrm{pre}} + R_0$. Near τ the oriented reference gate has slope at least ν, including along its smooth incoming extension. With $A = 1 + 2G/\nu$, the actual first zero therefore lies between $\tau - A\epsilon$ and $\tau + A\epsilon$. At that time all remaining oriented deficits lie in $[-D\epsilon, 0]$, with $D = G + 3M_0 A$.

No continuation can leave the chosen state ball before $\tau + \sigma$: its distance from the reference crossing state is at most

$$(C_{\mathrm{pre}} + 2M_0 A)\epsilon + M_0 \sigma < r/4$$

after decreasing ϵ. Every competing strict solution consequently satisfies $(\eta g_a)' \geq \nu$ almost everywhere in that window. A gate equal to zero at t_0 must obey $\eta g_a(t) \geq \nu(t - t_0) > 0$ immediately afterwards. This forces its outgoing bit, excludes waiting and sliding, and fixes the outgoing smooth flow by uniqueness. Already crossed gates cannot return to zero. Each event uses at least one previously uncrossed gate, so no infinite sequence of events is possible.

All crossings are over by $t_s = \tau + B_t\epsilon$, where $B_t = A + D/\nu + 1$. Bounded speeds control the discrepancy across the whole event layer by $(C_{\mathrm{pre}}+2M_0(A+B_t))\epsilon$. From t_s onward both paths use the same outgoing extension. One application of (30) gives an $O(\epsilon)$ state bound up to H. Decreasing ϵ until the induced gate error is less than half the reference slack outside the outer window rules out any later exit. The same construction proves uniqueness against every locally absolutely continuous strict competitor. When $\epsilon = 0$, the positive oriented normal velocity forces the unique outgoing branch directly; no zero-width event-window argument is used. □

For Stage 1, equation (32) supplies the hypothesis with orientation -1 and a normal reserve $3q(2b)/(4K)$. For Stage 3 the reserve is obtained below with orientation $+1$. Thus a symmetric reference alone is never used as a uniqueness proof for a nonsymmetric sampled network.

B Stage 1 calculation and strict gate perturbations

B.1 Exact scalar trajectory

At $P_h = 0$, $Z_h = c_k + yv_k + \epsilon v_j$. While all gates are active and $W = (vv_\Sigma, -vv_\Sigma)$, consistent margins equal $4v$, while conflict aggregates have zero classification sum. Population averaging gives $\dot{v} = [K(1 + e^{4v})]^{-1}$. Integrating from $v = 0$ to $v = b/2$ gives

$$T_0 = K \int_0^{b/2} (1 + e^{4s})\, ds = K \left[\frac{b}{2} + \frac{e^{2b} - 1}{4} \right].$$

After the wrong gates close, write $W = (uc_\Sigma + vv_\Sigma, uc_\Sigma - vv_\Sigma)$. The remaining consistent margin is $u + 2v + b$, and the population gradient gives $\dot{v} = 2\dot{u}$ and $\dot{u} = [4K(1 + e^{u+2v+b})]^{-1}$. Thus $v = b/2 + 2u$ and integrating produces (4).

B.2 Why simultaneous reference switches are stable

For $n_{k,j} = c_k - v_k - v_j$,

$$\langle n_{k,j}, n_{a,b} \rangle = 2\mathbf{1}_{k=a} + \mathbf{1}_{k=b} + \mathbf{1}_{j=a} + \mathbf{1}_{j=b}. \tag{31}$$

These entries are nonnegative, and summing over $a \neq b$ gives $5(K - 1)$. At the reference switch, put $q_b = (1 + e^{2b})^{-1}$. For a subset S of the zero wrong gates declared active, the normal velocity is

$$\dot{g}_{k,j} = -\frac{3q_b}{4K} - \frac{q_b}{4K(K-1)} \sum_{(r,s)\in S} \langle n_{k,j}, n_{r,s} \rangle.$$

The subset sum ranges from 0 to $5(K - 1)$, so every such cell satisfies

$$-\frac{2q_b}{K} \leq \dot{g}_{\mathrm{wrong}} \leq -\frac{3q_b}{4K} < 0. \tag{32}$$

The bound is uniform over all adjacent cells, not merely along the incoming symmetric trajectory.

A finite collection of smooth vector fields with a common strictly oriented normal near each active boundary cannot wait or slide on that boundary. Each nearby trajectory crosses the boundary only in the prescribed direction. Between events standard smooth dependence applies; over a finite set of possible events this gives finite-time state and crossing-time errors bounded by the perturbation size. This argument also handles coincident reference events that split under perturbation. Lemma A.3 supplies the local construction and uniqueness argument. Its Stage 1 application and the Gaussian initialization bounds follow below.

We derive the Gaussian initialization event with the constants in (8). Extend the ordered orthonormal basis $c_1, \ldots, c_K, v_1, \ldots, v_K$ of U to an orthonormal basis of $\mathbb{R}^d$. Under this change of coordinates,

each row of any one factor retains density proportional to $\exp(-m\|r\|^2/(2\omega^2))$: the norm is unchanged and the absolute Jacobian determinant is one. This density factors into the coordinate densities. Consequently, the $2K$ restricted columns of each factor are independent centered Gaussian vectors in $\mathbb{R}^m$, each with covariance $(\omega^2/m)I_m$. Different rows and different factor matrices remain independent.

For a vector g with m independent standard Gaussian entries, direct integration gives

$$\mathbb{E}e^{\theta\|g\|^2} = \prod_{r=1}^{m} \frac{1}{\sqrt{2\pi}} \int_{\mathbb{R}} e^{-(1-2\theta)s^2/2}\, ds = (1-2\theta)^{-m/2}, \qquad \theta < \frac{1}{2}.$$

For $0 < u < 1$, exponential Markov with $\theta = u/[2(1+u)]$ gives

$$\Pr\{\|g\|^2 \geq m(1+u)\} \leq \exp\left[-\frac{m}{2}\{u - \log(1+u)\}\right] \leq e^{-mu^2/8}.$$

The last inequality follows from

$$\frac{d}{du}\left(u - \log(1+u) - \frac{u^2}{4}\right) = \frac{u}{1+u} - \frac{u}{2} \geq 0 \qquad (0 \leq u \leq 1),$$

and the expression being differentiated vanishes at zero. For the lower tail, use $\theta = -u/[2(1-u)]$ to obtain

$$\Pr\{\|g\|^2 \leq m(1-u)\} \leq \exp\left[-\frac{m}{2}\{-u - \log(1-u)\}\right] \leq e^{-mu^2/4}.$$

Here

$$\frac{d}{du}\left(-u - \log(1-u) - \frac{u^2}{2}\right) = \frac{u}{1-u} - u \geq 0 \qquad (0 \leq u < 1),$$

and again the expression vanishes at zero. Rescaling to a restricted Gaussian column $v = (\omega/\sqrt{m})g$ therefore yields

$$\Pr\left\{\left|\|v\|^2 - \omega^2\right| > \omega^2 u\right\} \leq 2e^{-mu^2/8}, \qquad 0 < u \leq 1.$$

At $u = 1$, the upper-tail choice $\theta = 1/4$ remains valid, and the strict lower-tail event is empty.

For two independent restricted columns v, w with this covariance, put $v_+ = (v+w)/\sqrt{2}$ and $v_- = (v-w)/\sqrt{2}$. Each has the same Gaussian marginal as v, and

$$\langle v, w\rangle = \frac{\|v_+\|^2 - \|v_-\|^2}{2}.$$

If both squared norms differ from ω^2 by at most $\omega^2 u$, this identity bounds $|\langle v, w\rangle|$ by $\omega^2 u$. Their marginal tail bounds and a union bound give

$$\Pr\{|\langle v, w\rangle| > \omega^2 u\} \leq 4e^{-mu^2/8}, \qquad 0 < u \leq 1.$$

No independence between v_+ and v_- is needed for this bound. It applies to every off-diagonal self-Gram entry. It also applies to every cross-Gram entry because the key and query factors are independent.

There are four symmetric $2K \times 2K$ self-Grams and two unrestricted $2K \times 2K$ cross-Grams. The number of distinct entries to test is

$$4\frac{(2K)(2K+1)}{2} + 2(2K)^2 = 16K^2 + 4K.$$

Each diagonal entry has the squared-column tail bound, and each other tested entry has the inner-product bound. Thus every tested entry has failure probability at most $4e^{-mu^2/8}$.

The width assumption (8) gives

$$0 < \eta_m = \sqrt{\frac{8\log(L_K/\delta)}{m}} \leq \rho_K \leq 1.$$

Let $\mathcal{G}$ be the event that every entry of each of $A_h(0) - \omega^2 I_{2K}$, $B_h(0) - \omega^2 I_{2K}$, and $P_h^U(0)$ has absolute value at most $\omega^2\eta_m$, for both heads. Taking $u = \eta_m$ in the preceding bounds gives

$$\begin{aligned}\Pr(\mathcal{G}^c) &\leq 4(16K^2 + 4K)e^{-m\eta_m^2/8} \\ &= L_K e^{-\log(L_K/\delta)} = \delta, \qquad L_K = 64K^2 + 16K.\end{aligned}$$

On $\mathcal{G}$, every entry of $C_h(0) - \omega^2 I_{4K}$ has absolute value at most $\omega^2\eta_m$. Each joint Gram has $(4K)^2$ entries, including the repeated entries in its transposed blocks. Consequently,

$$\|C_h(0) - \omega^2 I_{4K}\|_F^2 \le (4K)^2\omega^4\eta_m^2.$$

With $C(0) = (C_+(0), C_-(0))$ and $\overline{C}(0) = (\omega^2 I_{4K}, \omega^2 I_{4K})$, the product Frobenius norm satisfies

$$\|C(0) - \overline{C}(0)\|_F \le C_G\omega^2\eta_m, \qquad C_G = 4\sqrt{2}K.$$

We also bound the frozen aggregate error on this same event. The softmax Jacobian is the symmetric matrix $J(p) = \mathrm{diag}(p) - pp^\top$. Its absolute row sum is

$$p_i(1-p_i) + \sum_{j\neq i} p_i p_j = 2p_i(1-p_i) \le \frac{1}{2}.$$

Choosing a coordinate of maximum absolute value in an eigenvector bounds the absolute value of its eigenvalue by the maximum absolute row sum. Hence $\|J(p)\|_{\mathrm{op}} \le 1/2$. Integrating the derivative along a line segment proves that softmax is globally $1/2$-Lipschitz in Euclidean norm. Since the columns of X are orthonormal and each query has norm one, summing the three query bounds gives

$$\|Z_h(P) - Z_h(\widetilde{P})\| \le \frac{3}{2}\|P - \widetilde{P}\|_{\mathrm{op}}. \tag{33}$$

For the Stage 1 error, the entrywise event gives a sharper bound directly. Write $x(X) = c_k + yv_k + \epsilon v_j$, the aggregate at zero score. Every coordinate of $X^\top P_h(0)x_\ell$ is a restricted cross-Gram entry up to a sign, so on $\mathcal{G}$,

$$\|X^\top P_h(0)x_\ell\|_2 \le \sqrt{3}\,\omega^2\eta_m.$$

The map $X : \mathbb{R}^3 \to U$ is an isometric embedding. Therefore

$$\|X\,\mathrm{softmax}(X^\top P_h(0)x_\ell) - \tfrac{1}{3}X\mathbf{1}\| \le \frac{\sqrt{3}}{2}\omega^2\eta_m.$$

Summing over the three queries yields

$$\alpha_Z := \max_{h,X}\|Z_h(X) - x(X)\| \le \frac{3\sqrt{3}}{2}\omega^2\eta_m.$$

The factors and these aggregates remain frozen throughout Stage 1.

Let W^0 denote the zero-score reference trajectory calculated above, with $W^0(T_1) = \overline{W}$. Applying Lemma A.3 with orientation -1 and the normal reserve in (32) gives constants $\alpha_0, C_{\mathrm{shadow}}, C_{\mathrm{switch}} > 0$ such that, for $\alpha_Z \le \alpha_0$,

$$\sup_{0\le t\le T_1}\|W(t) - W^0(t)\| \le C_{\mathrm{shadow}}\alpha_Z, \qquad |t_a - T_0| \le C_{\mathrm{switch}}\alpha_Z$$

for every wrong-consistent crossing time t_a. Decrease the radius so that all such crossings lie in $(0, T_1)$. The same lemma gives uniqueness, one downward crossing of each wrong-consistent gate, no recrossing, and positivity of every other gate. These constants depend only on (K, b, T_1): the features lie in U, zero initial readouts stay in U, and the isometry from $\mathbb{R}^{2K}$ to U transports the reference trajectory, gate pairings, and readout equations.

In selecting the fixed tolerance ρ_K, require

$$\frac{3\sqrt{3}}{2}\omega_K^2\rho_K \le \alpha_0.$$

This condition is compatible with the constant selection in Appendix G and is imposed before choosing the actual initialization scale. Since $\omega \le \omega_K$ and $\eta_m \le \rho_K$, the deterministic estimates apply on $\mathcal{G}$. Choose $C_w \ge (3\sqrt{3}/2)C_{\mathrm{shadow}}$. Then

$$\|W^1 - \overline{W}\| \le C_w\omega^2\eta_m, \qquad |t_a - T_0| \le \frac{3\sqrt{3}}{2}C_{\mathrm{switch}}\omega^2\eta_m.$$

These are the Stage 1 endpoint and crossing bounds of Theorem 4.1. The initial Gram bounds and all Stage 1 conclusions hold for the sampled factors and endpoint readouts on the single event $\mathcal{G}$.

C FACTORIZED ATTENTION: GROWTH AND RELATIVE PERTURBATIONS

C.1 CLOSED GRAM DYNAMICS AND THE ZERO-SCORE DERIVATIVE

Write the restricted factors as K_h, Q_h, so $P_h = K_h^\top Q_h$. With $N_h = \nabla_{P_h} R_{\mathrm{conf}}$,

$$\dot{K}_h = -Q_h N_h^\top, \quad \dot{Q}_h = -K_h N_h, \quad \dot{A}_h = -N_h P_h^\top - P_h N_h^\top,$$
$$\dot{P}_h = -A_h N_h - N_h B_h, \qquad \dot{B}_h = -N_h^\top P_h - P_h^\top N_h.$$

These are exactly (24); direct optimization of P_h would produce a different equation.

The conditional population and $\overline{W}$ are invariant under simultaneous component permutations, and the Gram field is equivariant under this action. Moreover, $Z_{JPJ}(X) = JZ_P(JX)$ and $\bar{w}_- = J\bar{w}_+$ make the field equivariant under swapping heads and conjugating by J. The isotropic initial Grams are fixed by both actions. Uniqueness therefore preserves component-permutation symmetry and $\bar{C}_- = \mathrm{diag}(J,J)\bar{C}_+ \,\mathrm{diag}(J,J)$, hence $\bar{P}_- = J\bar{P}_+ J$. These symmetries concern the reference trajectory; no symmetry projection is imposed on the sampled factors.

For a conflict let $\Delta = v_k - v_j$ and $u = c_k + y\Delta$. At zero score the softmax Jacobian is $I_3/3 - \mathbf{1}\mathbf{1}^\top/9$. For the positive-head aggregate readout $H_+ = \langle \overline{w}_+, Z_+ \rangle$ one obtains

$$\nabla_{P_+} H_+(0) = \frac{1}{3} d_+ u^\top, \quad d_+ = \frac{2\mu}{3} c_k + \beta(v_k + v_j) - \frac{y\mu}{3}\Delta.$$

The initial conflict margin is zero, so the logistic factor for the positive head is $-y/2$. Averaging over y and ordered pairs (k,j), the terms proportional to β cancel under pair reversal. The ordered-pair identity is an identity of CV coefficient blocks:

$$\big[\mathbb{E}_{k\neq j}\big[c_k(v_k - v_j)^\top\big]\big]_{CV} = \frac{T}{K-1}, \qquad [M]_{CV} := (c_a^\top M v_b)_{a,b=1}^K.$$

Equivalently, the ambient $d \times d$ expectation is

$$\mathbb{E}_{k\neq j}\big[c_k(v_k - v_j)^\top\big] = \frac{1}{K-1}\sum_{a,b=1}^{K} T_{ab} c_a v_b^\top.$$

Substitution yields (25).

C.2 REFERENCE GROWTH AND THE FIRST CONTRAST HIT

Lemma C.1 (Reference first hit on a prescribed comparison horizon)**.** *Fix (K, b, T_1) as in Section 3, with $K \geq 2$. Use the product Frobenius norm on the two symmetric joint Grams, and put*

$$\Lambda = \Lambda_K, \qquad W_b = 1 + \|\overline{W}\|, \qquad k_0 = \min\{\mu + b, 3\mu\}, \qquad M = 2\sqrt{8K}.$$

Let $A_0 \geq 1$ be a constant, fixed at these parameters, such that the analytic conflict-active Gram field satisfies

$$G(C) = \mathcal{L}_0 C + \mathcal{R}(C), \qquad \|\mathcal{R}(C)\| \leq A_0 \|C\|^2 \quad (\|C\| \leq 1).$$

Such a constant exists because $G(0) = 0$ and $DG(0) = \mathcal{L}_0$; for example, enlarge $2(9W_b^2/16 + 18W_b)$ to be at least one if necessary. Set

$$D = \frac{A_0 M^2}{\Lambda}, \qquad c_* = \min\left\{\frac{1}{2}, \frac{1}{4M}, \frac{M}{16D}, \frac{1}{64D}, \frac{k_0}{12MW_b}\right\}.$$

For every $0 < c_0 \leq c_*$ and $0 < \omega \leq c_0^{3/2}$, define the comparison horizon and growing scale without referring to a contrast hit:

$$H_b = \frac{1}{\Lambda}\log\frac{4c_0}{\omega^2}, \qquad q(t) = \omega^2 e^{\Lambda t}.$$

The reference Gram solution exists on $[0, H_b]$, and throughout this interval

$$\|\overline{C}(t)\| \leq Mq(t), \qquad \|\overline{C}(t) - C_{\mathrm{lin}}(t)\| \leq Dq(t)^2, \qquad C_{\mathrm{lin},h}(t) = \omega^2 e^{-2tS_h^0}. \tag{34}$$

All reference full-support gates keep the mixed sign pattern with absolute slack at least $k_0/2$. The contrast has a first hit $T_*(\omega)$ of c_0, and

$$\frac{1}{\Lambda}\log\frac{c_0}{\omega^2} \le T_*(\omega) \le H_b.$$

At that hit, equation (28) *holds with a constant C_P depending only on (K, b, T_1); the choice $C_P = 1 + 80D$ suffices.*

Proof. Equation (25) gives eigenvalues $\pm\alpha$ and $\pm\alpha/2$, each of multiplicity $K-1$, and four zero eigenvalues for each S_h^0. Thus the largest eigenvalue of $\mathcal{L}_0 : C_h \mapsto -S_h^0 C_h - C_h S_h^0$ on symmetric matrices is $2\alpha = \Lambda$. Self-adjointness implies $\|e^{t\mathcal{L}_0}\| = e^{\Lambda t}$. Since the initial Gram pair has norm $\sqrt{8K}\,\omega^2 = M\omega^2/2$, the linear solution has norm at most $Mq(t)/2$.

Work first with the smooth active extension, up to a putative first exit from $\|\overline{C}(t)\| \le Mq(t)$ on $[0, H_b]$. On this interval $Mq(t) \le 4Mc_0 \le 1$, so Duhamel's formula gives

$$\|\overline{C}(t) - C_{\mathrm{lin}}(t)\| \le A_0 M^2 \int_0^t e^{\Lambda(t-s)} q(s)^2\, ds \le Dq(t)^2.$$

Consequently,

$$\|\overline{C}(t)\| \le \{M/2 + 4Dc_0\}q(t) \le 3Mq(t)/4.$$

This strict improvement excludes a norm exit. The resulting bounded solution extends to H_b by the continuation theorem for a smooth finite-dimensional ODE. No hit time was used to define the interval or close this estimate.

At zero score, the two conflict gates equal $\mu + b$, each wrong-consistent gate equals -3μ, and each correct-consistent gate equals $5\mu + 2b$. Equation (33) and the Gram bound change any one gate by at most

$$\frac{3}{2}W_b\|\overline{C}(t)\| \le 6MW_b c_0 \le k_0/2.$$

Thus the smooth extension has exactly the prescribed mixed signs throughout $[0, H_b]$ and is the desired reference flow. Any factor realization has bounded restricted squared norms, given by the self-Gram traces, and consequently continues to the same horizon. Uniqueness of the closed Gram equation makes the reference independent of its factor realization.

The linear standard cross-score coefficients are

$$z_{\mathrm{lin}}(t) = \omega^2 \sinh(\Lambda t) = q(t)/2 - \omega^4/(2q(t)), \qquad r_{\mathrm{lin}}(t) = -\omega^2 \sinh(\Lambda t/2).$$

Extracting the standard CV coefficient is a linear functional of norm $1/\sqrt{K-1} \le 1$ on that score block. Hence $|z(t) - z_{\mathrm{lin}}(t)| \le Dq(t)^2$. For every time with $q(t) \le c_0$,

$$z(t) \le c_0/2 + Dc_0^2 < c_0.$$

At the comparison horizon,

$$z(H_b) \ge 2c_0 - \frac{\omega^4}{8c_0} - 16Dc_0^2 \ge \left(2 - \frac{1}{8} - \frac{1}{4}\right)c_0 > c_0,$$

where $\omega^4 \le c_0^6$, $c_0 \le 1$, and $Dc_0 \le 1/64$ were used. Continuity gives the first hit and both time bounds, without assuming monotonicity of the nonlinear contrast.

For completeness, the norm estimate also explains the nonleading reference terms. The initial identity commutes with S_h^0, so its linear components outside the leading symmetric-Gram eigenspace grow at most as $\omega^2 e^{\Lambda t/2}$. The nonlinear error is bounded by $Dq(t)^2$. This statement concerns the isotropic reference; it does not impose a smaller growth exponent on arbitrary perturbations.

At T_*, $q(T_*) \le 4c_0$ and $|r_{\mathrm{lin}}(T_*)| \le \omega\sqrt{c_0} \le c_0^2$. Let r be the standard VC coefficient of $\overline{P}_+(T_*)$. The linear-score error bounds both $|r - r_{\mathrm{lin}}|$ and $|c_0 - z_{\mathrm{lin}}|$ by $Dq(T_*)^2$. Therefore

$$|r| \le (1 + 16D)c_0^2, \qquad \|\overline{P}_+ - P^*(c_0, r)\|_{\mathrm{op}} \le 32Dc_0^2, \qquad \|\overline{P}_- + P^*(c_0, r)\|_{\mathrm{op}} \le 32Dc_0^2.$$

The last inequality uses $P_{\mathrm{lin},-} = -P_{\mathrm{lin},+}$ and the same error estimate for the other head. Their sum proves equation (28) with the stated C_P. All constants were fixed before c_0 and ω. □

C.3 A UNIFORM BRIDGE FROM ACTUAL PARAMETERS

The actual readout is frozen at W^1. Because $G(0, W) = 0$ for every W, the forcing induced by changing $\overline{W}$ to W^1 is $O(d_w\|C\|)$, rather than a constant error. Applying Duhamel to the full Gram pair yields

$$\|C(t) - \overline{C}(t)\|_F \le B\left[\frac{d_0}{\omega^2} + (1 + T_*)d_w\right] q(t)$$

for sufficiently small bracketed quantity. The nonlinear coefficient is integrable because $\int_0^{T_*} q(t)\,dt \le 4c_0/\Lambda_K$. Moreover,

$$\omega^2(1 + T_*) \le B_{\text{clock}} := 1 + \frac{\log 4 + \exp(-1)}{\Lambda_K}$$

after the fixed smallness choices. Stage 1 therefore supplies an effective error at most $(C_G + B_{\text{clock}}C_w)\eta_m$. At the endpoint, (33) gives

$$d_Z = \max_{h,X}\|Z_h(X) - \overline{Z}_h(X)\| \le 6c_0C_{\text{stage}}\eta_m \tag{35}$$

for a constant C_{stage} independent of ω. Appendix C.4 proves the relative estimate, gate preservation and factor continuation, with explicit smallness conditions. Appendix G gives the order of constant selection.

C.4 RELATIVE ERROR OVER THE ENTIRE ATTENTION INTERVAL

Use the product Frobenius norm on the two joint Grams. Let $\Lambda = \Lambda_K$, $q(t) = \omega^2 e^{\Lambda t}$, and suppose the reference satisfies $\|\bar{C}(t)\| \le M_r q(t)$ up to T_*, with $q(T_*) \le 4z$. The full Gram linearization L is self-adjoint and $\|e^{tL}\| = e^{\Lambda t}$. In the conflict-active extension write

$$G(C, \bar{W}) = LC + \mathcal{R}(C), \qquad \|\mathcal{R}(C) - \mathcal{R}(D)\| \le A(\|C\| + \|D\|)\|C - D\|.$$

Smooth score derivatives and the bilinear Gram equation also give

$$\|G(C, W^1) - G(C, \bar{W})\| \le Bd_w\|C\|, \qquad d_w = \|W^1 - \bar{W}\|. \tag{36}$$

For example, with $W_b = 1 + \|\bar{W}\|$ a valid choice is $A = 2(9W_b^2/16 + 18W_b)$ and $B = 2(2 + 2W_b)$ for $d_w \le 1$. These follow from $\|DZ_P\| \le 3/2$, $\|D^2Z_P\| \le 18$ and the first two logistic derivative bounds. The factor $\|C\|$ in (36) is essential.

Set $e = d_0/\omega^2 + (1 + T_*)d_w$ and $E(t) = \|C(t) - \bar{C}(t)\|/q(t)$. On a tentative interval with $E \le B_e e \le 1$, subtraction of the two full-space equations and Duhamel's formula give

$$E(t) \le \frac{d_0}{\omega^2} + BM_r d_w t + \int_0^t \{A(2M_r + 1)q(u) + Bd_w\}E(u)\,du. \tag{37}$$

Choose $B_e = 4(1 + BM_r)$, and before choosing ω require

$$z \le \frac{\Lambda \log 2}{8A(2M_r + 1)}, \qquad e \le \min\{B_e^{-1}, \log 2/(2B), 1\}.$$

Since $\int_0^{T_*} q \le 4z/\Lambda$ and $T_*d_w \le e$, the Gronwall multiplier is at most two. Equation (37) improves the tentative bound to $E \le B_e e/2$.

The full-support reference gates have slack at least $k_0/2$, where $k_0 = \min\{\mu+b, 3\mu\}$. The difference of corresponding gate values is at most

$$3d_w + \tfrac{3}{2}W_b\|C - \bar{C}\| \le (3 + 6W_bB_ez)e.$$

Choose e still smaller so that the last quantity is at most $k_0/4$. This also verifies the initial signs. There can be no first norm or gate exit. Restricted factor norms are controlled by the traces of their self-Grams; components orthogonal to the signal space are constant. Thus the factor trajectory cannot escape in finite time and extends uniquely to the prescribed reference clock. In particular,

$$d_Z \le 6B_ezе.$$

This proof permits arbitrary actual Gram errors in the top eigenspace; it makes no transverse-decay assumption.

The Stage 1 estimates $d_0 \le C_G\omega^2\eta_m$ and $d_w \le C_w\omega^2\eta_m$ yield

$$e \le (C_G + B_{\mathrm{clock}}C_w)\eta_m, \qquad B_{\mathrm{clock}} = 1 + \frac{\log 4 + \exp(-1)}{\Lambda},$$

because $\omega^2(1+T_*) \le B_{\mathrm{clock}}$. All permitted relative errors can therefore be fixed before ω. The clock itself still diverges as $\omega \downarrow 0$.

To make this selection explicit, use $M_r = M$ from Lemma C.1 and set

$$C_{\mathrm{stage}} = \max\{1, B_e(C_G + B_{\mathrm{clock}}C_w)\}.$$

Here C_G is the concentration constant from Appendix B, and C_w is the fixed Stage 1 endpoint constant. Together with $B_e = 4(1+BM)$, these constants depend only on (K, b, T_1) and can all be fixed before choosing c_0. Thus the later conditions on c_0 and ρ_K introduce no circular dependence.

C.5 The leading separating direction

For the positive conflict $X = [c_1, v_1, -v_2]$, zero-score aggregation gives $Z_+ = Z_- = c_1 + v_1 - v_2$. The symmetric classification readout sees cancelling signed contributions and predicts $f(X) = 0$. At the leading reference endpoint $P_+^{(0)} = \left(\begin{smallmatrix} 0 & c_0T \\ 0 & 0 \end{smallmatrix}\right)$, $P_-^{(0)} = -P_+^{(0)}$, the two classification queries give positive-head group-key logits $(K-1)c_0/K$ and c_0/K. Writing $h(s) = e^s/(e^s+2)$, the group contribution in direction $\theta_C = (c_\Sigma, c_\Sigma)/\sqrt{2K}$ is

$$\frac{h((K-1)c_0/K) - h(-(K-1)c_0/K) + h(c_0/K) - h(-c_0/K)}{\sqrt{2K}} = \frac{4c_0}{9\sqrt{2K}} + O(c_0^3).$$

The reference score remainder is $O(c_0^2)$, and the actual perturbation is controlled by the width event. Theorem 4.1(ii) retains a positive gated margin. After the remaining gates reopen, part (iv) gives separation of the full frozen features; Stage 3 then drives every type's loss to zero. This argument uses the sum of query contributions, not the special $K = 2$ equality of the two query probabilities.

C.6 The full-population gated margin

Lemma C.2 (Actual gated margin on the full population)**.** *Fix (K, b, T_1) with $K \ge 2$, and choose the reference constants as in Lemma C.1, using the already fixed $C_{\mathrm{stage}} \ge 1$ in equation* (35)*. In addition to the reference and relative-error requirements, choose*

$$c_0 \le \min\left\{c_*, \frac{1}{48C_P}\right\}, \qquad \rho_K \le \frac{1}{192C_{\mathrm{stage}}}.$$

On the common initialization event, suppose the actual Stage 2 trajectory has the mixed sign pattern and the endpoint estimate

$$d_Z = \max_{h,X}\|Z_h(X) - \overline{Z}_h(X)\| \le 6c_0C_{\mathrm{stage}}\eta_m, \qquad \eta_m \le \rho_K.$$

Then the unit vector $\theta_C = (c_\Sigma, c_\Sigma)/\sqrt{2K}$ satisfies

$$\min_X y\langle\theta_C, F_{\mathrm{gate}}(X; T_*)\rangle \ge \frac{c_0}{8\sqrt{2K}} \ge \frac{c_0}{36\sqrt{2K}}.$$

These additional choices are independent of ω, the Gaussian draw, and the terminal risk.

Proof. Compare the actual endpoint with the leading scores

$$P_+^{(0)} = \begin{pmatrix} 0 & c_0T \\ 0 & 0 \end{pmatrix}, \qquad P_-^{(0)} = -P_+^{(0)},$$

and let $Z_h^{(0)}(X)$ be their aggregates. Equation (28) and the global aggregate Lipschitz bound imply

$$\Delta_Z := \max_{h,X}\|Z_h(X) - Z_h^{(0)}(X)\| \le \frac{3}{2}C_Pc_0^2 + 6c_0C_{\mathrm{stage}}\eta_m \le c_0/16.$$

For any type $X = [c_k, yv_k, \epsilon v_j]$, all these aggregates lie in $\mathrm{span}\{c_k, v_k, v_j\}$. Thus the group coefficient $G_h(X) = \langle c_\Sigma, Z_h(X)\rangle = \langle c_k, Z_h(X)\rangle$ differs from its leading counterpart by at most Δ_Z, without an extra factor $\sqrt{K}$.

For a conflict, set $a = (K-1)c_0/K$ and $d = c_0/K$. With $h(s) = e^s/(e^s+2)$, the two classification queries and the uniform group query give

$$y\{G_+^{(0)}(X) - G_-^{(0)}(X)\} = h(a) - h(-a) + h(d) - h(-d).$$

For $0 \le s \le 1$,

$$h(s) - h(-s) = \frac{4\sinh s}{5 + 4\cosh s} \ge \frac{4s}{13} \ge \frac{s}{4},$$

using $\sinh s \ge s$ and $\cosh 1 < 2$. Since $a + d = c_0 \le 1$, the leading difference is at least $c_0/4$. Both actual gates are active on conflicts, so

$$y\langle \theta_C, F_{\mathrm{gate}}(X; T_*)\rangle = \frac{y(G_+(X) - G_-(X))}{\sqrt{2K}} \ge \frac{c_0/4 - 2\Delta_Z}{\sqrt{2K}} \ge \frac{c_0}{8\sqrt{2K}}.$$

For a consistent type, only the correct-label head contributes to the actual gated feature. In the leading model its group coefficient is $1/3 + h(s_2) + h(s_3)$ for two logits $s_2, s_3 \in [-c_0, c_0]$. Because $h(s) \ge h(-1) = 1/(1+2e) > 1/7$ on this interval, this coefficient exceeds $1/2$. The signed gated pairing is therefore at least

$$\frac{1/2 - \Delta_Z}{\sqrt{2K}} \ge \frac{7}{16\sqrt{2K}} \ge \frac{c_0}{8\sqrt{2K}}.$$

The sign is positive for either label: for label -1, the minus sign in the negative-head feature cancels the label. These two cases exhaust the full population. Finally, $\|\theta_C\| = 1$ by orthonormality. No factor involving $K - 2$ was used. □

D Quadratic and cubic attention terms

In this appendix assume $K \ge 3$. Permutation invariance decomposes each reference score block into a standard coefficient multiplying T and a trivial coefficient multiplying Q_0. All expansions below concern the exact reference at its first contrast-z hit, with z sufficiently small and $0 < \omega \le z^{3/2}$.

D.1 Attention mechanisms beyond two components

For a conflict $X = [c_k, v_k, -v_j]$, token 1 is context, token 2 the matched signal, and token 3 the distractor. At the Stage 2 endpoint for $K \ge 3$, $D_2 < 0$ and $D_3 > 0$ express opposite positive-head preferences between keys 2 and 3 for the two classification queries; G_1 compares the heads' context-key probabilities for the context query. Separation uses the sum of the context contributions, while this refinement resolves their internal structure. For the leading positive-head score $P^{(0)} = \left(\begin{smallmatrix} 0 & zT \\ 0 & 0 \end{smallmatrix}\right)$ and conflicting type $(k, +1, j)$, the matched-minus-distractor query difference is

$$p^{(0)}_{+,1|2}(k,+1,j) - p^{(0)}_{+,1|3}(k,+1,j) = \frac{2(K-2)}{9K} z + O(z^2).$$

Here the superscript (0) denotes probabilities evaluated at $P^{(0)}$. The first-order coefficient vanishes at $K = 2$. For classification keys, permutation symmetry gives $P_{VV} = aT + b_0 Q_0$ and $(e_k + e_j)^\top P_{VV} e_k = (K-2)a/K + 2b_0/K$; the reference flow has $a = -2z^2/35 + O(z^3)$ and $b_0 = O(\omega^2 z)$. Hence the two signed second-order gaps have opposite signs and magnitude $2(K-2)z^2/(105K) + O(z^3)$, while the one-sided cubic estimates in Theorem 4.3 follow from the quadratic coefficients and a determinant barrier (Appendix D). The $K = 2$ cancellation is structural; finite-width runs need not preserve exact reference equalities.

D.2 Quadratic coefficients

Collect the standard coefficients $x = (P^s_{CC}, P^s_{VV}, A^s_{CV}, B^s_{CV})$. The positive-head score gradient N_+ depends on both scores through the logistic multiplier. Put $D = P^{(0)}/z$ and differentiate the

joint ray $(P_+, P_-) = (\xi D, -\xi D)$:

$$\frac{d}{d\xi} N_{+,CC}(\xi D, -\xi D)\Big|_{\xi=0} = 0, \qquad \frac{d}{d\xi} N_{+,VV}(\xi D, -\xi D)\Big|_{\xi=0} = \frac{\alpha}{6} T.$$

Here is the ordered-pair calculation, including the logistic derivative. For $X = [c_k, yv_k, -yv_j]$, set $\Delta = v_k - v_j$, $S = v_k + v_j$, $a = (K-1)/K$, $d = 1/K$, and $r_{kj} = av_k - dv_j$. Write $m_\xi = y f_{\xi D, -\xi D}(X)$ and $q_\xi = (1 + e^{m_\xi})^{-1}$. With $h(s) = e^s/(e^s+2)$, the context-key logits on the two classification queries are $ya\xi$ and $yd\xi$. Thus

$$m_0 = 0, \qquad m_0' = 2\mu h'(0)(a+d) = \frac{4\mu}{9}, \qquad q_0 = \frac{1}{2}, \quad q_0' = -q_0(1-q_0)m_0' = -\frac{\mu}{9}.$$

Let $G_\xi = \nabla_{P_+} \langle \bar{w}_+, Z_+ \rangle$ on this ray. Differentiating the softmax probabilities gives

$$(G_0)_{VV} = \frac{y\beta}{3} S\Delta^\top - \frac{\mu}{9}\Delta\Delta^\top, \qquad (G_0')_{VV} = -\frac{\beta}{9} S r_{kj}^\top - \frac{y\mu}{27}\Delta r_{kj}^\top,$$
$$(G_\xi)_{CC} = \frac{2\mu}{9} c_k c_k^\top, \qquad (G_0')_{CC} = 0.$$

Here outer products are represented in the corresponding orthonormal context or signal basis. Since $N_+(\xi D, -\xi D) = \mathbb{E}[-yq_\xi G_\xi]$, averaging over the two labels first yields

$$(N_+'(0))_{CC} = 0,$$
$$(N_+'(0))_{VV} = \frac{\mu\beta}{27}\mathbb{E}_{k\neq j}[S\Delta^\top] + \frac{\mu}{54}\mathbb{E}_{k\neq j}[\Delta r_{kj}^\top] = \frac{\mu}{54(K-1)} T = \frac{\alpha}{6} T.$$

Pair reversal makes the first expectation zero; the second has signal coefficient matrix $(a+d)T/(K-1) = T/(K-1)$. The derivatives therefore include both the softmax Jacobian and the logistic weight. We control the remainder on the entire growing interval. In this subsection, $C(t) = (\overline{C}_+(t), \overline{C}_-(t))$ denotes the exact reference Gram pair, $\mathcal{F}$ its analytic field on the active conflict branch, and $L = D\mathcal{F}(0)$. Put $\Lambda = \Lambda_K = 2\alpha$ and $q(t) = \omega^2 e^{\Lambda t}$. Norms of Gram pairs are product Frobenius norms. All constants K_0, C_j and all smallness thresholds used below depend only on (K, μ, b). They are independent of ω, the target contrast z, d, m, and the orthonormal embedding, and are fixed before z and ω are selected.

The full-space linearization in Appendix C satisfies $\|e^{tL}\|_{\mathrm{op}} = e^{\Lambda t}$. Its exact isotropic solution is $(C_{\mathrm{lin}})_h = \omega^2 e^{-2tS_h^0}$. In the positive-head standard sector,

$$(P_{\mathrm{lin}})^s_{CV} = \omega^2 \sinh(\Lambda t), \qquad (P_{\mathrm{lin}})^s_{VC} = -\omega^2 \sinh(\Lambda t/2),$$
$$(A_{\mathrm{lin}})^s_{CC} = (B_{\mathrm{lin}})^s_{VV} = \omega^2 \cosh(\Lambda t), \qquad (A_{\mathrm{lin}})^s_{VV} = (B_{\mathrm{lin}})^s_{CC} = \omega^2 \cosh(\Lambda t/2).$$

All four mixed coordinates defining x vanish in this linear solution. Its trivial self-Grams are $\omega^2 I_2$, and its trivial scores are zero. The negative head follows by conjugation with $\mathrm{diag}(J, J)$. These formulas give $\|C_{\mathrm{lin}}(t)\|_F \le K_0 q(t)$.

To specify the Taylor remainder, define on a fixed small Gram ball

$$\mathcal{Q}_2(H) = \tfrac{1}{2} D^2\mathcal{F}(0)[H,H], \qquad \mathcal{R}_3(H) = \mathcal{F}(H) - LH - \mathcal{Q}_2(H).$$

Bounded third derivatives on that ball, and the bilinear representation of $\mathcal{Q}_2$, give

$$\|\mathcal{R}_3(H)\|_F \le C_1 \|H\|_F^3,$$
$$\|\mathcal{Q}_2(H) - \mathcal{Q}_2(\widetilde{H})\|_F \le C_1(\|H\|_F + \|\widetilde{H}\|_F)\|H - \widetilde{H}\|_F,$$
$$\|\mathcal{F}(H) - LH\|_F \le C_1 \|H\|_F^2.$$

Choose the ball small enough that (33) preserves the full-support mixed gate signs. On a tentative interval with $\|C(t)\|_F \le 2K_0 q(t)$ and $q(t) \le q_0$, Duhamel's formula gives

$$\|C(t) - C_{\mathrm{lin}}(t)\|_F \le 4C_1 K_0^2 \int_0^t e^{\Lambda(t-u)} q(u)^2 \, du$$
$$\le C_2 q(t)^2, \qquad C_2 = \frac{4C_1 K_0^2}{\Lambda}.$$

Take $q_0 > 0$ so small that $2K_0 q_0$ lies inside the chosen ball and $C_2 q_0 \le K_0/2$. This improves the tentative norm bound to $3K_0 q(t)/2$. The gate signs cannot change inside the ball. The traces of the self-Grams bound the restricted factor norms, and the orthogonal factor components are constant. Thus neither a norm exit, a gate exit, nor a finite-time failure of factor continuation can occur before $q = q_0$. We have consequently established, uniformly throughout that envelope,

$$\|C(t)\|_F \le 2K_0 q(t), \qquad \|C(t) - C_{\mathrm{lin}}(t)\|_F \le C_2 q(t)^2.$$

Define a joint Gram direction $\mathcal{D} = (\mathcal{D}_+, \mathcal{D}_-)$ separately from the score direction D above. In the standard sector, ordered as key C, key V, query C, query V, set

$$(\mathcal{D}_+)^s = \begin{pmatrix} 1 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 1 \end{pmatrix}.$$

Each scalar entry here multiplies T in signal coordinates. The trivial sector is zero, and $\mathcal{D}_- = \mathrm{diag}(J, J)\mathcal{D}_+ \,\mathrm{diag}(J, J)$. The score pair of $\xi\mathcal{D}$ is $(\xi D, -\xi D)$. The exact linear formulas retain the reverse, trivial, and decreasing hyperbolic terms and give

$$\begin{aligned} \|C_{\mathrm{lin}}(t) - (q(t)/2)\mathcal{D}\|_F &\le C_3 \left[\omega^2 e^{\Lambda t/2} + \omega^2 + \frac{\omega^4}{q(t)}\right] \\ &\le C_4 \omega \sqrt{q(t)}, \end{aligned}$$

where the last inequality uses $q(t) \ge \omega^2$. Hence

$$\|C(t) - (q(t)/2)\mathcal{D}\|_F \le C_5 \left[\omega\sqrt{q(t)} + q(t)^2\right].$$

Reduce the fixed envelope if necessary so that $(q/2)\mathcal{D}$ also lies in the Taylor ball.

Let $\mathcal{X}$ be the bounded linear extraction of the four positive-head standard coordinates:

$$\mathcal{X}(C) = \big((P_+)^s_{CC}, (P_+)^s_{VV}, (A_+)^s_{CV}, (B_+)^s_{CV}\big), \qquad x(t) = \mathcal{X}(C(t)).$$

The exact block equations in Appendix C give $\mathcal{X}(LC) = \alpha M x$. At $\xi\mathcal{D}$, the only nonzero positive-head standard Gram coordinates are $A^s_{CC} = B^s_{VV} = P^s_{CV} = \xi$. The two loss-gradient derivatives computed above give quadratic mixed-field components $(0, -\alpha\xi^2/6, -\alpha\xi^2/6, 0)$. Thus $\mathcal{X}(\mathcal{Q}_2(\xi\mathcal{D})) = \alpha f \xi^2$, and substitution into the Gram equation gives the forced system

$$\dot{x} = \alpha M x + \alpha f (q/2)^2 + \mathcal{E}(t), \quad M = \begin{pmatrix} 0 & 0 & -1/2 & 1 \\ 0 & 0 & 1 & -1/2 \\ -1/2 & 1 & 0 & 0 \\ 1 & -1/2 & 0 & 0 \end{pmatrix}, \quad f = \begin{pmatrix} 0 \\ -1/6 \\ -1/6 \\ 0 \end{pmatrix}. \tag{38}$$

Here the actual remainder is

$$\mathcal{E}(t) = \mathcal{X}\left(\mathcal{Q}_2(C(t)) - \mathcal{Q}_2((q(t)/2)\mathcal{D}) + \mathcal{R}_3(C(t))\right).$$

The Taylor bounds and the preceding Gram estimates imply

$$\begin{aligned} \|\mathcal{E}(t)\|_2 &\le C_6 \left\{ q(t) \left[\omega\sqrt{q(t)} + q(t)^2\right] + q(t)^3 \right\} \\ &\le C_7 \left[\omega q(t)^{3/2} + q(t)^3\right]. \end{aligned}$$

In particular, a reverse component of size $\omega\sqrt{q}$ contributes to this remainder only after multiplication by another factor of q.

The eigenvalues of M are $\pm 1/2, \pm 3/2$, below the exponent 4 of the quadratic forcing in α time. Since M is symmetric,

$$\|e^{\alpha t M}\|_{\mathrm{op}} \le e^{3\Lambda t/4} \qquad (t \ge 0).$$

Solving

$$(4I - M)v = f \quad \Longrightarrow \quad v = \left(\frac{1}{105}, -\frac{2}{35}, -\frac{2}{35}, \frac{1}{105}\right)$$

identifies the forced coefficients. The function $x_{\mathrm{app}}(t) = v(q(t)/2)^2$ solves (38) with $\mathcal{E} = 0$, because $\frac{d}{dt}(q/2)^2 = 4\alpha(q/2)^2$. Its initial value is $v\omega^4/4$, whereas $x(0) = 0$. Variation of constants therefore gives

$$\|x(t) - x_{\mathrm{app}}(t)\|_2 \le \frac{\|v\|_2}{4}\omega^4 e^{3\Lambda t/4} + C_7 \int_0^t e^{3\Lambda(t-u)/4} \left[\omega q(u)^{3/2} + q(u)^3\right] du.$$

For every $p > 3/4$, inserting $q(u) = q(t)e^{-\Lambda(t-u)}$ yields the elementary bound

$$\int_0^t e^{3\Lambda(t-u)/4} q(u)^p \, du = q(t)^p \int_0^t e^{-(p-3/4)\Lambda(t-u)} \, du \le \frac{q(t)^p}{(p-3/4)\Lambda}.$$

Apply it with $p = 3/2$ and $p = 3$. The initial particular-solution error obeys

$$\omega^4 e^{3\Lambda t/4} = \omega^{5/2} q(t)^{3/4} \le \omega q(t)^{3/2},$$

again because $q(t) \ge \omega^2$. We obtain the uniform integrated estimate

$$\|x(t) - v(q(t)/2)^2\|_2 \le C_8 \left[\omega q(t)^{3/2} + q(t)^3\right].$$

The trivial sector requires a separate evolution estimate. Let $P_{\mathrm{triv}}, A_{\mathrm{triv}}, B_{\mathrm{triv}}$ be its positive-head 2×2 coefficient matrices, and put

$$C_{\mathrm{triv}} = \begin{pmatrix} A_{\mathrm{triv}} & P_{\mathrm{triv}} \\ P_{\mathrm{triv}}^\top & B_{\mathrm{triv}} \end{pmatrix}, \qquad S_{\mathrm{triv}} = \begin{pmatrix} 0 & N_{\mathrm{triv}} \\ N_{\mathrm{triv}}^\top & 0 \end{pmatrix}.$$

Permutation invariance and $TQ_0 = Q_0 T = 0$ reduce the exact Gram equation to

$$\dot{C}_{\mathrm{triv}} = -S_{\mathrm{triv}} C_{\mathrm{triv}} - C_{\mathrm{triv}} S_{\mathrm{triv}}, \qquad C_{\mathrm{triv}}(0) = \omega^2 I_4, \qquad P_{\mathrm{triv}}(0) = 0.$$

The zero-score gradient in (25) has no trivial component. Bounded first derivatives of the score gradient and $\|C(t)\|_F \le 2K_0 q(t)$ consequently give $\|S_{\mathrm{triv}}(t)\|_{\mathrm{op}} \le C_9 q(t)$. The integral norm inequality for this matrix equation gives

$$\|C_{\mathrm{triv}}(t)\|_F \le 2\omega^2 \exp\left(2C_9 \int_0^t q(u)\, du\right) \le C_{10}\omega^2,$$

since $\int_0^t q(u)\, du = (q(t) - \omega^2)/\Lambda \le q_0/\Lambda$. Integrating the matrix equation once more yields

$$\|C_{\mathrm{triv}}(t) - \omega^2 I_4\|_F \le 2C_9 C_{10}\omega^2 \int_0^t q(u)\, du \le C_{11}\omega^2 q(t).$$

Its initial cross block is zero, so $\|P_{\mathrm{triv}}(t)\|_F \le C_{11}\omega^2 q(t)$. Head conjugacy gives the corresponding bound for the negative head.

Finally, let T_z be the actual first contrast hit of this exact reference trajectory; $T_z = T_*$ when $z = c_0$. Choose fixed positive thresholds $z_{\mathrm{jet}} < 1$ and ω_{jet} within the first-hit ranges of Appendix C, with $4z_{\mathrm{jet}} \le q_0$. For $0 < z \le z_{\mathrm{jet}}$ and $0 < \omega \le \min\{\omega_{\mathrm{jet}}, z^{3/2}\}$, the preceding first-hit bounds give $z \le q(T_z) \le 4z$. The contrast extraction of the full Gram error gives

$$\left| z(t) - \frac{q(t)}{2} \right| \le \frac{\omega^4}{2q(t)} + C_{12} q(t)^2,$$

because $\omega^2 \sinh(\Lambda t) = q(t)/2 - \omega^4/(2q(t))$. At the hit, writing $q_z = q(T_z)$, this implies

$$\left|(q_z/2)^2 - z^2\right| \le (q_z/2 + z)\left[\frac{\omega^4}{2q_z} + C_{12} q_z^2\right] \le C_{13}(\omega^4 + z^3).$$

Combining this conversion with the integrated mixed estimate and the trivial-sector bound gives

$$\|x(T_z) - vz^2\|_2 \le C_{14}\left[\omega z^{3/2} + z^3 + \omega^4\right] \le C_{15} z^3,$$
$$\|P_{\mathrm{triv}}(T_z)\|_F \le 4C_{11}\omega^2 z.$$

Here $\omega z^{3/2} \le z^3$ and $\omega^4 \le z^6 \le z^3$. Thus the integrated remainder is $O(z^3)$ at the actual hit, and the trivial score coefficients are $O(\omega^2 z)$. Consequently

$$P^s_{CC} = \frac{z^2}{105} + O(z^3), \qquad P^s_{VV} = -\frac{2z^2}{35} + O(z^3). \tag{39}$$

The coefficients in (39) enter the determinant argument in Appendix D.4.

For the two classification queries, the leading score has equal logits for keys 2 and 3, so their probability difference starts at second order. Substituting the P^s_{VV} coefficient in the two softmax denominators and using the uniform Taylor remainder above gives, at the exact reference hit,

$$\overline{D}_2 = -\frac{2(K-2)}{105K} z^2 + O(z^3), \qquad \overline{D}_3 = \frac{2(K-2)}{105K} z^2 + O(z^3). \tag{40}$$

The signs in (40) hold for both labels and are preserved by head conjugacy. Indeed, the classification-key logits are $(P_{kk}, -P_{jk})$ for query 2 and $(-P_{kj}, P_{jj})$ for query 3: the label factors cancel, and head conjugacy preserves the VV block. The positive softmax denominators may change, but the numerator signs do not. Combining the score-tube error with the $O(z^3)$ Taylor remainder gives the finite-width one-sided bounds in Theorem 4.3.

D.3 A DETERMINANT BARRIER FOR THE REVERSE SCORE

Let C_s be the 4×4 standard Gram matrix and $\mathcal{J}_4 = \mathrm{diag}(I_2, -I_2)$. Let S_s denote the standard-sector restriction of S_+. Along the reference flow, $C_s(t) = \omega^2 \Phi_s(t)\Phi_s(t)^\top$, where $\dot{\Phi}_s = -S_s(t)\Phi_s$ and $\Phi_s(0) = I_4$. The continuous coefficient $S_s(t)$ has zero trace, so Liouville's formula gives $\det \Phi_s(t) = 1$ at every finite time. Thus $C_s(t) \succ 0$ throughout the reference horizon of Lemma C.1. The off-diagonal form of S implies the preserved identity

$$C_s \mathcal{J}_4 C_s = \omega^4 \mathcal{J}_4. \tag{41}$$

Write its 2×2 score block as $P_s = \left(\begin{smallmatrix} a & z \\ r & b \end{smallmatrix}\right)$. At a zero of $\det P_s$, positive definiteness and (41) yield

$$\frac{d}{dt} \det P_s = -2\omega^2 \langle \mathrm{cof}(P_s), N_s \rangle_F. \tag{42}$$

Indeed, let A_s, B_s be the self-Gram blocks of C_s. Equation (41) and positive definiteness imply

$$A_s = (\omega^4 I + P_s P_s^\top)^{1/2}, \qquad B_s = (\omega^4 I + P_s^\top P_s)^{1/2}.$$

At a rank-one boundary, $P_s^\top \mathrm{cof}(P_s) = 0$ and $\mathrm{cof}(P_s) P_s^\top = 0$. Thus $A_s \,\mathrm{cof}(P_s) = \omega^2 \,\mathrm{cof}(P_s)$ and $\mathrm{cof}(P_s) B_s = \omega^2 \,\mathrm{cof}(P_s)$. Pairing $\dot{P}_s = -A_s N_s - N_s B_s$ with the cofactor gives (42). If $P_s = 0$, both sides vanish directly. The rescaled linear solution first gives $\det P_s > 0$ on a fixed positive-time launch interval. At any proposed later first zero, $r = ab/z = O(q^3)$, while $N^s_{VC} = \alpha/2 + O(q)$. The cofactor pairing is strictly negative, so (42) is positive, ruling out such a first zero.

At the hit, (39) implies, after reducing the fixed contrast threshold,

$$a \ge \frac{z^2}{210}, \qquad b \le -\frac{z^2}{35}, \qquad r < \frac{ab}{z} \le -\frac{z^3}{7350}.$$

Writing r_0 for the trivial reverse coefficient, one has $r_0 = O(\omega^2 z) = O(z^4)$. The group-query classification-key logit difference therefore has label-oriented sign determined by $[(K-2)r + 2r_0]/K$. Bounded exponential denominators turn this into a one-sided cubic lower bound for the probability-gap magnitude. Head conjugacy also gives a positive cubic lower bound for the group-key between-head gap. Appendix D.4 supplies the separate initial-sign argument and explains the one-sided conclusion.

D.4 Initial sign and the one-sided reverse-score bound

The zero-boundary identity (42) alone cannot establish the initial sign of $\det P_s$. Here is the separate launch argument. Let $\epsilon = \omega^2$, $t_c = \log 2/\Lambda$, and

$$P_0(t) = \begin{pmatrix} 0 & \sinh(2\alpha t) \\ -\sinh(\alpha t) & 0 \end{pmatrix}.$$

The reference Duhamel estimate and its differentiated equation give a constant C_{launch} independent of ω such that, on $[0, t_c]$,

$$\|(P_s/\epsilon)' - P_0'\|_F \le 4C_{\mathrm{launch}}\epsilon, \qquad \|P_s/(\epsilon t) - P_0(t)/t\|_F \le 4C_{\mathrm{launch}}\epsilon.$$

Both matrices vanish initially, so the second inequality follows by integration; the quotient extends continuously at zero. On this fixed interval $\det(P_0(t)/t) \ge 2\alpha^2$ and $\|P_0(t)/t\|_F \le 3\alpha$. For 2×2 matrices,

$$|\det(A+E) - \det A| \le \|A\|_F\|E\|_F + \tfrac{1}{2}\|E\|_F^2.$$

Choosing $\epsilon \le \alpha/(16C_{\mathrm{launch}})$ makes this error at most $25\alpha^2/32$, proving $\det P_s(t) > 0$ on $(0, t_c]$.

For $t \ge t_c$ up to the first hit, the linear solution and remainder bound in Lemma C.1 give

$$\frac{z_t}{q(t)} \ge \frac{1 - e^{-2\Lambda t}}{2} - Dq(t) \ge \frac{3}{8} - \frac{1}{16} \ge \frac{1}{4}.$$

Here $q(t) \le 4z$ and $Dz \le 1/64$ are already part of that lemma's contrast budget. The diagonal score coefficients a_s, b_s are $O(q(t)^2)$, and $N^s_{VC} = \alpha/2 + O(q(t))$. At a proposed zero determinant, $r = a_s b_s / z_t = O(q(t)^3)$. Hence

$$\langle \mathrm{cof}\, P_s, N_s\rangle_F = b_s N^s_{CC} - rN^s_{CV} - z_t N^s_{VC} + a_s N^s_{VV} \le -\alpha q(t)/8 + O(q(t)^2) < 0.$$

The balanced-Gram identity makes (42) strictly positive there. A differentiable function previously positive has nonpositive left derivative at its first zero, a contradiction. This proves positivity through the first contrast hit without assuming it along the incoming path.

At the hit, the quadratic score estimates then imply $r < -z^3/7350$. This is a *one-sided cubic lower bound on* $-r$. The linear solution already contains $-\omega^2 \sinh(\Lambda t/2)$, whose endpoint scale is $\omega\sqrt{z}$; the allowed condition $\omega \le z^{3/2}$ only bounds this by z^2. The present argument therefore gives no uniform upper bound $-r = O(z^3)$, and no $\Theta(z^3)$ assertion follows. The attention inequalities in Theorem 4.3 use only the proved sign and lower bound.

D.5 Reference probability bounds for the group query

Write the positive-head reverse block as $P_{VC} = rT + r_0 Q_0$ and put

$$u = (P_{CC})_{kk}, \qquad d = \frac{(K-1)r + r_0}{K}, \qquad e = \frac{-r + r_0}{K}.$$

For the group query c_k, the positive-head logits are $(u, yd, -ye)$; head conjugacy gives $(u, -yd, ye)$ for the negative head. With $C_{\mathrm{rev}} = 4C_{11}$, the previous estimates give $r < -z^3/7350$ and $|r_0| \le C_{\mathrm{rev}}\omega^2 z \le C_{\mathrm{rev}} z^4$. Reduce the fixed contrast threshold so that all these logits have absolute value at most $\log 2$ and

$$2C_{\mathrm{rev}} z \le \frac{K-2}{14700}.$$

Then

$$d + e = \frac{(K-2)r + 2r_0}{K} \le -\frac{(K-2)z^3}{14700K}, \qquad e - d = -r > 0.$$

Let $L_y = e^u + e^{yd} + e^{-ye}$. Since $L_y \le 6$ and the derivative of the exponential is at least $1/2$ between any two of these logits, the mean-value theorem gives

$$-y\overline{D}_1 = -y\frac{e^{yd} - e^{-ye}}{L_y} \ge \frac{|d+e|}{12} \ge \frac{(K-2)z^3}{176400K}.$$

For the group-key comparison between heads, direct subtraction gives

$$\overline{G}_1 = \frac{2e^u(\sinh e - \sinh d)}{L_y L_{-y}} \geq \frac{e-d}{36} = \frac{-r}{36} \geq \frac{z^3}{264600}.$$

Here $2e^u \geq 1$ and $\sinh e - \sinh d \geq e - d$ because $\cosh s \geq 1$. In particular, the second bound does not require a sign assumption on the trivial coefficient r_0.

For the classification-query comparison in (17), take the leading score $P_+^{(0)} = z \left(\begin{smallmatrix} 0 & T \\ 0 & 0 \end{smallmatrix}\right)$, set $a = (K-1)z/K$ and $d = z/K$. For a positive label and a conflicting type, the two nonzero logits in the context-key row are a for query 2 and d for query 3, while the other two token logits are zero. Hence

$$p^{(0)}_{+,1|2} = h(a), \qquad p^{(0)}_{+,1|3} = h(d), \qquad h(s) = \frac{e^s}{e^s + 2}.$$

These identities hold for the leading score; the reference and finite-width perturbations are controlled separately below. For $0 < z \leq 1$, $h'(s) = 2e^s/(e^s+2)^2 \geq 1/5$ on $[0,1]$, and therefore

$$p^{(0)}_{+,1|2} - p^{(0)}_{+,1|3} \geq \frac{a-d}{5} = \frac{(K-2)z}{5K}.$$

For the negative label, the label-oriented leading-score gap is $h(-d) - h(-a) > 0$. Both gaps have first-order term $2(a-d)/9$, but their finite values generally differ. Equation (17) uses only the positive-label comparison above. Each softmax probability changes by at most half the score perturbation, so that comparison transfers with the separate Taylor and finite-width error budgets below. This calculation uses the classification queries $\ell = 2, 3$ and is separate from the group-query calculation above.

D.6 Transfer to the actual attention

For the reference, the leading first-order and quadratic expansions, together with the two bounds just proved, allow the fixed contrast threshold to be chosen so that every comparison has at least twice the corresponding margin in Theorem 4.3. Define

$$\Delta_K(z) = \min\left\{\frac{2z}{25K}, \frac{7z}{225K}, \frac{(K-2)z^2}{210K}, \frac{(K-2)z^3}{352800K}, \frac{z^3}{529200}\right\} > 0.$$

The full Gram perturbation estimate gives

$$\max_h \|P_h(T_*) - \overline{P}_h(T_*)\|_{\mathrm{op}} \leq C_{\mathrm{sc}} z \eta_m, \qquad C_{\mathrm{sc}} = 4B_e(C_G + B_{\mathrm{clock}} C_w),$$

with a constant independent of the subsequent choice of ω. Each sample matrix X has orthonormal columns and each query has unit norm. The softmax Jacobian has operator norm at most $1/2$, so each probability changes by at most half this score error. Every comparison above is a difference of two probabilities; its total change is therefore at most $C_{\mathrm{sc}} z \eta_m$, uniformly over all types and both labels. After fixing c_0 below the common contrast thresholds and then choosing $\omega_K \leq \min\{c_0^{3/2}, \omega_{\mathrm{jet}}\}$ within the launch and continuation thresholds, choose

$$\rho_K \leq \min\left\{\rho_K^{\mathrm{base}}, \frac{\Delta_K(c_0)}{C_{\mathrm{sc}} c_0}\right\},$$

where $\rho_K^{\mathrm{base}} > 0$ already enforces the all-K stage and readout requirements. Then $\eta_m \leq \rho_K$ preserves every stated margin. The resulting width condition is independent of ω and uses the same initialization event; no separate probability event or later increase of the width is needed.

E Feature geometry and the permanent readout tail

E.1 Reference hull computation

Average positive-head aggregates over ordered component pairs and define

$$Z_F = \overline{Z}(+,-) - \overline{Z}(-,+), \qquad Z_G = \overline{Z}(+,+) - \overline{Z}(-,-).$$

Here the two signs indicate (y, ϵ); the bar denotes the pair average in this paragraph. The unique minimum-norm point is fixed by the simultaneous component permutations and the head involution. Averaging a convex representation over this group preserves the point and makes its weights constant on each orbit, without requiring uniqueness of those weights. Convexity then reduces the point to

$$v_F + xd, \quad 0 \le x \le \frac{1}{2}, \qquad v_F = (Z_F/2, JZ_F/2), \quad d = (Z_G - Z_F, J(Z_G - Z_F)).$$

To display the orbit sum explicitly, first take the leading score $P_+ = zD$. Set $a = (K-1)/K$, $d = 1/K$, $h(s) = e^s/(e^s+2)$, and

$$C_{y,\epsilon}(z) = \frac{1}{3} + h(yaz) + h(-\epsilon dz), \qquad L_{y,\epsilon}(z) = \frac{3 - C_{y,\epsilon}(z)}{2}.$$

The context query contributes $1/3$ to each token; the other queries give equal probabilities to the two signal keys. Hence the ordered-pair average is

$$\frac{1}{K(K-1)} \sum_{k \ne j} \big[C_{y,\epsilon} c_k + L_{y,\epsilon}(y v_k + \epsilon v_j)\big] = \frac{C_{y,\epsilon}}{K} c_\Sigma + \frac{(y+\epsilon) L_{y,\epsilon}}{K} v_\Sigma.$$

Writing a superscript ray for these leading-score coefficients, subtraction of the four sign orbits gives

$$\begin{aligned}
A_F^{\mathrm{ray}} &= \frac{h(az) - h(-az) + h(dz) - h(-dz)}{2K} = \frac{2z}{9K} + O(z^3), \\
B_F^{\mathrm{ray}} &= 0, \\
A_D^{\mathrm{ray}} &= \frac{2}{K}\big[h(-dz) - h(dz)\big] = -\frac{8z}{9K^2} + O(z^3), \\
B_D^{\mathrm{ray}} &= \frac{1}{K}\big[6 - C_{+,+} - C_{-,-}\big] = \frac{4}{K} + O(z^2).
\end{aligned}$$

The actual reference score differs from zD by $O(z^2)$, so bounded feature derivatives transfer these leading coefficients. The sharper classification cancellation is detailed next. Write $Z_F/2 = A_F c_\Sigma + B_F v_\Sigma$ and $Z_G - Z_F = A_D c_\Sigma + B_D v_\Sigma$. The required expansions are

$$\begin{aligned}
A_F &= \frac{2z}{9K} + O(z^2), & B_F &= O(z^3), \\
A_D &= -\frac{8z}{9K^2} + O(z^2), & B_D &= \frac{4}{K} + O(z^2).
\end{aligned}$$

The $O(z^3)$ bound for B_F is essential. Its full derivative at zero vanishes: for an arbitrary score perturbation E, the relevant ordered-pair term contains $(v_k + v_j)^\top E(v_k - v_j)$, which changes sign under $k \leftrightarrow j$. The pure-top ray has the same cancellation, controlling the remaining terms in the score tube.

It follows that

$$\langle v_F, d\rangle = -\frac{32z^2}{81K^2} + O(z^3), \qquad \|d\|^2 = \frac{32}{K} + O(z^2), \qquad x_* = -\frac{\langle v_F, d\rangle}{\|d\|^2} = \frac{z^2}{81K} + O(z^3).$$

This lies strictly inside $[0, 1/2]$ and gives (29). In particular, the positive classification coordinate is genuinely quadratic and cannot be inferred from a first-order separator alone.

E.2 Stability without symmetry of the actual features

Suppose each head aggregate changes by at most d_Z. The Hausdorff distance between the two labeled feature hulls is at most $\sqrt{2} d_Z$. Projection inequalities for their respective minimum-norm points give

$$\|v_* - \overline{v}_*\|^2 \le (\gamma + \overline{\gamma})\sqrt{2} d_Z \le 12 d_Z, \tag{43}$$

because every labeled feature has norm at most $3\sqrt{2}$. Thus a fixed sufficiently small feature perturbation preserves positivity of all prescribed signed signal coordinates of the minimum-norm hull point. Since $\gamma^2 \le 18$, the coordinate bounds for θ_{mm} follow from those for v_*. The small tolerance required here is compatible with (35) after the width is selected once.

For completeness, (13) follows directly from projection geometry. The minimizing property of v_* gives $\langle v_*, a_X\rangle \ge \gamma^2$, so v_*/γ^2 is feasible. Every feasible θ obeys $\langle \theta, v_*\rangle \ge 1$ and hence $\|\theta\| \ge 1/\gamma$. The proposed vector attains this value and strict convexity gives uniqueness.

E.3 UNIFORM GEOMETRY ON THE CLOSED SCORE TUBE

The readout argument uses only the score tube, simultaneous component permutation symmetry, head conjugacy, and the common launch $\overline{W}$. It does not require a member of the closed tube to be realized by a Stage 2 trajectory. Indeed, write

$$D = \begin{pmatrix} 0 & T \\ 0 & 0 \end{pmatrix}, \qquad P_+ = zD + E_z, \qquad \|E_z\|_{\mathrm{op}} \le Cz^2, \qquad P_- = JP_+J.$$

The separately displayed reverse coefficient in (28) is included in E_z. All derivatives of the finite-population aggregation map are uniformly bounded on a fixed neighborhood of zero.

For arbitrary P, define $B_F(P) = K^{-1}v_\Sigma^\top\{\overline{Z}_P(+,-) - \overline{Z}_P(-,+)\}/2$, where the bar denotes the ordered-pair average. On the symmetric tube this projection is the common classification coefficient. For every matrix direction E,

$$DB_F(0)[E] = \frac{1}{3K}\mathbb{E}_{k\neq j}(v_k + v_j)^\top E(v_k - v_j) = 0.$$

The last equality follows by interchanging k and j. Moreover, $B_F(zD) = 0$ for every z: the two classification keys have equal logits under each query on this ray. The mean-value formula and $DB_F(0) = 0$ therefore give, uniformly on the entire closed tube,

$$|B_F(zD + E_z)| \le C'\big(z\|E_z\|_{\mathrm{op}} + \|E_z\|_{\mathrm{op}}^2\big) = O(z^3).$$

The other hull and mixed-branch coefficient expansions follow from the same uniformly bounded derivatives. To make the mixed-branch signs explicit, set

$$\begin{gathered} Z_C = \overline{Z}(+,+), \qquad Z_W = \overline{Z}(-,-), \\ A = \|Z_C\|^2, \quad B = \langle Z_C, Z_F\rangle, \quad E = \|Z_F\|^2, \\ a = \langle Z_W, Z_C\rangle, \qquad b_W = \langle Z_W, Z_F\rangle. \end{gathered}$$

In particular $E \ge 0$ exactly. The finite crossing, every-adjacent-cell normal test, and no-recrossing barrier consequently hold for every member of this closed family, after one uniform smallness choice of z for fixed (K, b, T_1).

Unsigned normal and margin reserves. For every individual reference positive-head aggregate,

$$\overline{Z}_+(X) = c_k + yv_k + \epsilon v_j + O(z), \qquad \overline{v}_{*,+} = A_{\mathrm{hull}}c_\Sigma + B_{\mathrm{hull}}v_\Sigma,$$

and hence

$$\langle \overline{Z}_+(X), \overline{v}_{*,+}\rangle = \frac{2z}{9K} + O(z^2).$$

Head conjugacy gives the same formula for the negative head, with no extra label sign in the gate normal. Choose z uniformly over the closed family so that $\overline{\kappa} \ge z/(9K) > 0$. Under a feature perturbation d_Z, the change of each such product is at most $3\sqrt{2}d_Z + 3\sqrt{12d_Z}$ by (43). Thus a fixed smaller feature neighborhood preserves a positive actual κ, as required for strict entry into the actual invariant basin.

Choose the same z so that the reference group and prescribed signed classification coordinates are at least $z/(9K)$ and $2z^2/(81K^2)$, respectively. With s_0 as in Theorem 4.2, the choice $d_Z \le s_0^2/48$ gives

$$\|v_* - \overline{v}_*\| \le \frac{s_0}{2} \le \frac{z}{18K} \le \frac{\overline{\gamma}}{2}, \qquad \overline{\gamma} \ge \frac{\sqrt{2K}\,z}{9K}.$$

Consequently

$$\gamma \ge \frac{\overline{\gamma}}{2} \ge \frac{z}{\sqrt{162K}}, \qquad \langle v_*, e\rangle \ge \frac{s_0}{2}, \qquad \langle \theta_{\mathrm{mm}}, e\rangle \ge \frac{s_0}{36} \quad (e \in \mathcal{E}),$$

where the last bound uses $\gamma^2 \le 18$. Thus the common launch radius preserves the explicit margin reserve and all the stated signed coordinate reserves. The allowed actual features can be enclosed in a closed bounded neighborhood of the compact reference feature family; realizability of every point in that neighborhood is not needed.

E.4 Reference return to the all-active region

With the orbit averages just defined, the positive readout on the mixed reference branch is

$$w_+(t) = w_+(0) + X(t)Z_C + Y(t)Z_F, \qquad \dot{X} = \frac{1}{4}q(m_C), \quad \dot{Y} = \frac{1}{4}q(m_F), \quad q(u) = (1+e^u)^{-1}.$$

The two margins and the wrong gate are affine in (X, Y):

$$m_C = m_C^0 + AX + BY, \qquad m_F = m_F^0 + BX + EY,$$
$$g_W = g_W^0 + aX + b_W Y.$$

The coefficient estimates are

$$A = 5/K + O(z), \qquad a = -3/K + O(z), \qquad E = O(z^2),$$
$$B = 4z/(9K) + O(z^2), \qquad b_W = 4z/(9K) + O(z^2).$$

So B and b_W have the same leading term but need not be exactly equal. Initially $m_C^0 = 5\mu + 2b + O(z)$, $m_F^0 = O(z)$, and $g_W^0 = -3\mu + O(z)$.

If g_W stayed negative forever, it would bound Y above by an affine function of X. Bounded X would then bound both margins and contradict $\dot{X} > 0$ uniformly on a bounded state. Unbounded X forces $m_C - m_F \to \infty$, hence $dY/dX = q(m_F)/q(m_C) \to \infty$, contradicting that affine upper bound. A finite crossing follows.

The quantitative argument also gives, at the first crossing τ,

$$\frac{q(m_C(\tau))}{q(m_F(\tau))} \le C_p z^p = o(z), \qquad 1 < p < 8/5.$$

This makes all adjacent wrong-gate normal velocities positive. On the all-active reference branch the ratio of the consistent and conflicting weights stays below $z/36$ by a barrier argument, keeping the gate derivatives positive. Appendix E.7 expands the crossing-ratio estimate, every-adjacent-cell test and no-recrossing barrier.

E.5 A Lyapunov basin for nonsymmetric features

Let the actual frozen head features have norm at most $M = 3$, and assume both separation and positive gate-direction products

$$\kappa := \min_{h,X} \langle \widetilde{Z}_{h,X}, v_* \rangle > 0, \quad \widetilde{Z}_{+,X} = (Z_+(X), 0), \quad \widetilde{Z}_{-,X} = (0, Z_-(X)).$$

This is a condition on unsigned gate normals; the label sign is already in a_X and is not inserted again here. The reference geometry supplies this positivity, and sufficiently small perturbations preserve it.

In the smooth all-active flow, put

$$q_i = (1 + e^{\langle W, a_i \rangle})^{-1}, \quad Q = \sum_i q_i, \quad p_i = q_i/Q, \quad v = \sum_i p_i a_i, \quad \frac{ds}{dt} = Q/N.$$

Primes in the next display denote derivatives in s. For $u_i = \langle a_i, v \rangle$, $\mu_0 = \|v\|^2$, and $\mu_2 = \sum_i p_i^2 u_i$,

$$Q' = -Q(\mu_0 - Q\mu_2), \tag{44}$$
$$p_i' = p_i\{\mu_0 - Q\mu_2 - (1 - Qp_i)u_i\}.$$

With $\phi = \|v\|^2/2$, direct differentiation gives

$$\phi' = -\operatorname{Var}_p(u) + Q\left(\sum_i p_i^2 u_i^2 - \mu_0\mu_2\right) \le 8M^4 Q, \qquad R' = -Q\|v\|^2/N \le -Q\gamma^2/N.$$

Therefore

$$\mathcal{L} = \phi - \gamma^2/2 + \frac{8NM^4}{\gamma^2} R \quad \text{is nonincreasing.} \tag{45}$$

The projection inequality implies $\|v - v_*\|^2 \le 2(\phi - \gamma^2/2)$. A sufficiently small sublevel of (45) consequently makes every gate's s-velocity at least $\kappa/2$. The sublevel is therefore a permanent all-active basin. Lemma E.1 states the actual-feature threshold and proves permanent continuation, together with an explicit conditional decay bound.

E.6 EXACT ASYMPTOTICS AND UNIFORMITY

The rescaled time satisfies $s \to \infty$: otherwise W would remain bounded, all q_i would be bounded away from zero, and s could not stay finite. Since $R' \le -Q\gamma^2/N$, one has $\int Q\, ds < \infty$.

Choose hull weights p^* representing v_*. Differentiating the relative entropy with (44) gives

$$\frac{d}{ds}\,\mathrm{KL}(p^*\|p) \le -(\phi - \gamma^2/2) + 4M^2Q.$$

Hence the nonnegative hull gap is integrable. Its derivative is bounded, so it tends to zero; therefore $v \to v_*$. As the risk vanishes, $Q/(NR) \to 1$. Returning to ordinary time,

$$\frac{d}{dt}\frac{1}{R} = \left(\frac{Q}{NR}\right)^2 \|v\|^2 \longrightarrow \gamma^2, \qquad \dot W = (Q/N)v.$$

These identities yield $tR \to \gamma^{-2}$, $t\dot W \to v_*/\gamma^2$, and, by integration, $W/\log t \to v_*/\gamma^2$.

Uniformity requires an additional argument. At fixed $z = c_0$, the closed reference score family in (28) is compact. Each member has a finite crossing sequence and a strict entry into the permanent basin. Finite-time dependence and directed switching provide a neighborhood with the same qualitative entry, and a finite subcover supplies a common radius and entry time. On a compact family of smooth starts in the permanent basin, $\mathcal{L}(t)$ is continuous in the parameters, nonincreasing in t, and tends pointwise to zero. A finite-cover argument upgrades this convergence to uniform convergence. Applying it to the preceding asymptotic identities and the uniform positive signed coordinates of the minimum-norm hull point gives one $T_{\mathrm{align},K}$. Appendix E.9 gives the finite-cover construction, uses a full bounded ball of post-entry states, and derives common tail-error envelopes.

E.7 THE UPWARD CROSSING AND THE ABSENCE OF RECROSSING

Use the coefficients and orbit averages in Appendix E. The positive reference mixed branch has $\dot X = q(m_C)/4$ and $\dot Y = q(m_F)/4$, with $B, b_W > 0$, $a < 0$, $A - B > 0$ and $B - E > 0$ for sufficiently small z. For a nonwrong positive-head aggregate n_X, the coefficient bounds are

$$\langle n_X, Z_C\rangle = \begin{cases} 5/K + O(z), & y = \epsilon = +1, \\ 1/K + O(z), & \epsilon = -y, \end{cases} \qquad \langle n_X, Z_F\rangle = 4z/(9K) + O(z^2).$$

Indeed $n_X = c_k + yv_k + \epsilon v_j + O(z)$, $Z_C = (c_\Sigma + 2v_\Sigma)/K + O(z)$, and $Z_F = 4zc_\Sigma/(9K) + O(z^2)$. All displayed coefficients are positive for sufficiently small z, uniformly over the reference score tube. Hence the derivative of each nonwrong positive-head gate is

$$\dot g_X = \tfrac{1}{4}\{q(m_C)\langle n_X, Z_C\rangle + q(m_F)\langle n_X, Z_F\rangle\} > 0.$$

The negative-head statement follows by head conjugacy. Every nonwrong gate starts positive, so the first possible boundary is the wrong-consistent gate.

The affine-upper-bound contradiction in Appendix E proves that this gate reaches zero at finite τ. To establish the direction at that event, put $d = -g_W^0 > 0$, $S = b_W Y$, $k = B/b_W$ and $F(X) = e^{-kS(X)}$. Since $m_F \ge m_F^0$ and $m_F^0 = O(z)$, direct substitution into the multiplier ratio gives, with a uniformly bounded positive factor if necessary,

$$-F'(X) \le CB(1 + e^{m_C^0 + AX}), \qquad 1 - F(x) \le CBx + \frac{CBe^{m_C^0}}{A}(e^{Ax} - 1).$$

Indeed $Fe^{BY} = 1$, while the denominator $1 + e^{m_F}$ is at least one; the displayed inequality even holds with $C = 1$. At the crossing $S = d - ax \ge d$. Since d and k have fixed positive lower bounds, $1 - F(x) \ge \eta_0 > 0$. It follows that $x \ge A^{-1}\log(1/B) - C_0$.

At this boundary, write x for the scalar trajectory coordinate (the symbol X continues to denote a population type). Then

$$m_C - m_F = m_C^0 - m_F^0 + \frac{(B-E)d}{b_W} + \left(A - B - \frac{(B-E)a}{b_W}\right)x.$$

The coefficient of x tends to $8/K$, whereas $A \to 5/K$. Also $m_F \geq m_F^0 = O(z)$. Hence for every fixed $p \in (1, 8/5)$,

$$\rho_\tau = \frac{q(m_C(\tau))}{q(m_F(\tau))} \leq C_p z^p = o(z). \tag{46}$$

Write $n = K(K-1)$ and let n_r be the individual positive-head wrong-consistent aggregates. At the simultaneous reference zero, choosing any subset S of these zero gates to be active leaves the loss values, and hence the logistic multipliers q_s, q_c, unchanged. Its positive-head velocity is

$$V_S = \tfrac{1}{4}(q_s Z_C + q_c Z_F) - \frac{q_s}{4n} \sum_{r \in S} n_r.$$

With $c = \|Z_W\|^2$, permutation symmetry gives, for every wrong gate p,

$$\langle n_p, V_{\text{all}} \rangle = \frac{q_c}{4}\{b_W - (c-a)\rho_\tau\} = q_c\{z/(9K) + o(z)\} > 0.$$

At zero scores, all pairwise $\langle n_p, n_r \rangle$ are nonnegative. At the allowed reference endpoint each is at least $-Cz$. Removing any subset of active wrong gates can therefore decrease the normal velocity by at most $Cq_s z/4 = O(q_c z^{p+1})$. This is smaller than its positive $q_c z$ leading term. All adjacent patterns point upward in both heads. Their finite minimum is the transverse reserve required by Lemma A.3.

On the outgoing reference branch $w'_+ = (q_s Z_G + q_c Z_F)/4$. At a putative first boundary $\rho = q_s/q_c = z/36$, the coefficient expansions give

$$\dot{m}_s \geq q_c z/(12K), \qquad |\dot{m}_c| \leq C q_c z^2.$$

Here $q_s \leq z/36$, so $(\log \rho)' = -(1-q_s)\dot{m}_s + (1-q_c)\dot{m}_c < 0$ for small z. Equation (46) starts strictly inside this barrier. It follows that $\rho \leq z/36$ forever. For each individual gate normal n_X, the same expansions give

$$\langle n_X, Z_G \rangle \geq -9/K, \qquad \langle n_X, Z_F \rangle \geq z/(3K).$$

Consequently $\dot{g}_X \geq q_c z/(48K) > 0$. The reference therefore has no recrossing and is globally all-active after τ. These statements hold for every fixed $K \geq 2$ throughout the sufficiently small quadratic score tube; no $K \geq 3$ attention bound is used.

E.8 An actual-feature invariant region and a quantitative tail lemma

Let arbitrary frozen labeled features $a_i = y_i(Z_{+,i}, -Z_{-,i})$ satisfy $\|Z_{h,i}\| \leq M$, and let $\gamma = \|v_*\| > 0$. Consider the smooth logistic flow from any finite state W_0, without an initial gate-activity requirement. Define q_i, Q, p_i, v, s as in Appendix E and put

$$g = \tfrac{1}{2}(\|v\|^2 - \gamma^2), \quad A_* = \frac{8NM^4}{\gamma^2}, \quad \mathcal{L} = g + A_* R, \qquad \lambda = \frac{\gamma^2}{1 + NR_0}, \quad R_0 = R(0).$$

Lemma E.1 (Smooth tail and permanent ReLU continuation)**.** *For any convex representation $v_* = \sum_i p_i^* a_i$, let $D_0 = \mathrm{KL}(p^* \| p(0))$ with zero weights omitted, and set*

$$C_* = D_0 + \frac{(4NM^2 + A_*)R_0}{\lambda}.$$

Then $\mathcal{L}$ is nonnegative and nonincreasing, and

$$\mathcal{L}(s) \leq \frac{C_*}{s}, \qquad \mathcal{L}(t) \leq \frac{2M^2(1+NR_0)C_*}{\log(1 + 2M^2 R_0 t)} \quad (t > 0). \tag{47}$$

In particular, $tR \to \gamma^{-2}$, $t\dot{W} \to v_*/\gamma^2$ and $W/\log t \to v_*/\gamma^2$. If also $\kappa = \min_{h,i}\langle Z_{h,i}, v_{*,h} \rangle > 0$, all gates at a given finite time are positive, and

$$\mathcal{L} \leq \epsilon_* := \min\left\{ \frac{\kappa^2}{8M^2}, \frac{8M^4 \log 2}{\gamma^2} \right\}, \tag{48}$$

then the strict ReLU flow coincides with this smooth continuation forever, and every gate is strictly increasing thereafter.

Proof. Projection on the feature hull gives $\|v - v_*\|^2 \le \|v\|^2 - \gamma^2 = 2g$. The normalized-flow identities in (44) imply

$$g' \le 8M^4 Q, \qquad R' = -Q\|v\|^2/N, \qquad \mathcal{L}' \le 0,$$

where these primes denote s derivatives. The relation $q_i = 1 - e^{-\ell_i}$ for $\ell_i = \log(1 + e^{-\langle W, a_i\rangle})$ yields

$$\frac{R}{1 + NR_0} \le Q/N \le R.$$

It follows that $R(s) \le R_0 e^{-\lambda s}$ and $\int_0^\infty Q\, ds \le NR_0/\lambda$. Normalized time is unbounded: a bounded s would bound W through $dW/ds = v$, forcing every q_i away from zero and contradicting bounded s.

Writing $u_i = \langle a_i, v\rangle$ and $\mu_2 = \sum_i p_i^2 u_i$, direct differentiation gives

$$D' = \langle v_*, v\rangle - \|v\|^2 + Q\left(\mu_2 - \sum_i p_i^* p_i u_i\right) \le -g + 4M^2 Q.$$

This holds without positive or unique optimal convex weights. Since $D \ge 0$, integration gives $\int_0^\infty g\, ds \le D_0 + 4NM^2 R_0/\lambda$. Adding $A_* \int R$ proves $\int_0^\infty \mathcal{L}\, ds \le C_*$. Monotonicity implies $s\mathcal{L}(s) \le C_*$. In physical time,

$$\frac{d}{dt}\frac{1}{R} = \left(\frac{Q}{NR}\right)^2 \|v\|^2 \le 2M^2,$$

so $R(t) \ge (R_0^{-1} + 2M^2 t)^{-1}$. Integrating $ds/dt \ge R/(1 + NR_0)$ proves the second bound in (47). In particular $v \to v_*$ and $R \to 0$. Since $e^{-NR} \le Q/(NR) \le 1$, integration of $(1/R)' \to \gamma^2$ gives $tR \to \gamma^{-2}$. The exact identity $\dot{W} = (Q/N)v$ then yields $t\dot{W} \to v_*/\gamma^2$. Integrating this velocity limit gives $W/\log t \to v_*/\gamma^2$.

Finally, (48) gives $R \le \log 2/N$, hence each $q_i \le 1/2$, and $\|v - v_*\| \le \kappa/(2M)$. Thus

$$\langle Z_{h,i}, v_h\rangle \ge \kappa/2, \qquad \dot{g}_{h,i} = (Q/N)\langle Z_{h,i}, v_h\rangle > 0$$

at every finite time. The sublevel is invariant under the smooth flow. Starting with positive gates there is no first gate exit, so the smooth continuation is the unique strict ReLU continuation. The values v_*, γ, κ in this argument all belong to the actual frozen features. □

E.9 Uniformity and finite-time interpretation

For the fixed positive contrast z, the reference score triples in (28), with permutation symmetry, head conjugacy and exact contrast, form a closed bounded subset of a finite-dimensional space. Their aggregate image $\mathcal{B}_z$ is compact. For each $B \in \mathcal{B}_z$, Appendix E.7 and Lemma E.1 give a reference finite time H_B with positive gates and strict tail entry. Hull projection is continuous under feature perturbation by (43); thus γ and κ are also continuous. Lemma A.3 transfers the finite reference path to all sufficiently close actual features and launches. Shrinking its neighborhood gives strict entry into each actual family's own sublevel at H_B.

Let δ_B be such a total feature-plus-launch radius. Take a finite subcover of $\mathcal{B}_z$ by its relative balls of radii $\delta_B/4$, with centers B_j. With s_0 as in Theorem 4.2, choose

$$r_{\mathrm{read}} = \min\{1, s_0^2/48, \delta_{B_1}/4, \dots, \delta_{B_J}/4\}, \qquad H = \max_j H_{B_j}.$$

An actual launch within r_{read} of any reference lies strictly within some center's δ_{B_j} neighborhood by the triangle inequality. Openness and forward invariance therefore give entry into the permanent tail by the common time H for the entire family.

At H every actual readout lies in the ball $\|W(H)\| \le B_H := \|\bar{W}\| + 1 + 3\sqrt{2}H$. The allowed actual features are contained in a compact family, with $\gamma \ge g_0 := z/\sqrt{162K} > 0$, as shown in Appendix E.3. Apply the smooth result to the product of this entire feature family and the full ball of radius B_H.

For an explicit uniform tail bound, consider any smooth starts with $\|W_0\| \le B_H$ and $\gamma \ge g_0$. Set

$$B_f = \sqrt{2}M, \qquad r_- = \log(1 + e^{-B_f B_H}), \qquad r_+ = \log 2 + B_f B_H,$$
$$A_0 = 8NM^4/g_0^2, \qquad \lambda_0 = g_0^2/(1 + Nr_+), \qquad D_0^+ = \log\{N(1 + e^{B_f B_H})\},$$
$$C_0 = D_0^+ + (4NM^2 + A_0)r_+/\lambda_0, \qquad a = g_0^2/(1 + Nr_+)^2.$$

The entropy bound $D_0 \le \log(1/\min_i p_i(0)) \le D_0^+$ avoids selecting optimal weights continuously. Lemma E.1 and the reciprocal-risk inequality give common decreasing envelopes

$$\bar{L}(u) = \min\left\{M^2 + A_0 r_+, \frac{2M^2(1 + Nr_+)C_0}{\log(1 + 2M^2 r_- u)}\right\}, \qquad \bar{R}(u) = (r_+^{-1} + au)^{-1}.$$

For $u \ge 1$ define

$$E(u) = \frac{1}{ur_-} + \frac{2M^2}{\sqrt{u}} + 2\bar{L}(\sqrt{u}) + 4NM^2\bar{R}(\sqrt{u}).$$

To check this envelope, $\alpha = (1/R)'$ satisfies $|\alpha - \gamma^2| \le 2\bar{L} + 4NM^2\bar{R}$ and independently $|\alpha - \gamma^2| \le 2M^2$. Integrating on $[0, \sqrt{u}]$ and $[\sqrt{u}, u]$ gives $|1/(uR) - \gamma^2| \le E(u)$. Once $E(u) \le g_0^2/2$,

$$|uR - \gamma^{-2}| \le 2E(u)/g_0^4, \tag{49}$$

$$\|u\dot{W} - \theta_{\mathrm{mm}}\| \le \frac{2\sqrt{2\bar{L}(u)}}{g_0^2} + B_f\left(\frac{2E(u)}{g_0^4} + \frac{2N\bar{R}(u)}{g_0^2}\right) =: \bar{h}(u). \tag{50}$$

Both bounds tend to zero uniformly and allow degenerate or changing support sets. For $u \ge U \ge 2$ beyond that threshold, integrating the velocity error gives

$$\left\|\frac{W(u)}{\log u} - \theta_{\mathrm{mm}}\right\| \le \frac{I_U}{\log u} + \bar{h}(U)\frac{\log(u/U)}{\log u}, \quad I_U \le B_H + \frac{B_f}{a}\log(1 + ar_+U) + \frac{\log U}{g_0}. \tag{51}$$

First choose U to make $\bar{h}(U)$ small, and then u to make $I_U/\log u$ small. This proves a common finite clock for the logarithmic limit as well.

Here $u = t - H$ is elapsed smooth-tail time. For finite-time uses, the clock shift must be retained. Define the two envelopes explicitly by $E_R(u) = 2E(u)/g_0^4$ and $E_V(u) = \bar{h}(u)$. Then

$$|tR_3(t) - \gamma^{-2}| \le (1 + H/u)E_R(u) + H/(u\gamma^2),$$
$$\|t\dot{W}(t) - \theta_{\mathrm{mm}}\| \le (1 + H/u)E_V(u) + H\|\theta_{\mathrm{mm}}\|/u.$$

The logarithmic estimate also includes the factor $\log u/\log t$ and the additive correction $\|\theta_{\mathrm{mm}}\|\log(1 + H/u)/\log t$. Thus the same limits hold in local Stage 3 time. Pairing them with each prescribed signed coordinate, whose limiting value is at least $s_0/36$, gives a common $T_{\mathrm{align},K}$ and the lower bounds $s_0/(72t)$ and $s_0 \log t/72$.

Equation (51) bounds a finite-time deviation from an asymptotic limit. It is not an ODE discretization error bound. It also does not assert a convergent constant residual in $W(u) - \theta_{\mathrm{mm}}\log u$ for every actual feature family. The displayed constants are explicit conditional on H, B_H, g_0; the earlier reference-family entry radius and clock remain existential.

F EXACT FINITE-MATRIX CERTIFICATES

For a saved feature matrix A interpreted entrywise as exact binary64 rational values, any exact feasible vector $\hat{\theta}$ and nonnegative dual vector λ give

$$D_{\mathrm{dual}} = \mathbf{1}^\top\lambda - \tfrac{1}{2}\|A^\top\lambda\|^2 \le p_* := \min_{A\theta \ge \mathbf{1}} \tfrac{1}{2}\|\theta\|^2 \le U = \tfrac{1}{2}\|\hat{\theta}\|^2.$$

This follows by completing the square and using $\lambda^\top(A\theta - \mathbf{1}) \ge 0$. If $D_{\mathrm{dual}} > 0$, then $(2U)^{-1} \le \gamma^2 \le (2D_{\mathrm{dual}})^{-1}$. These checks can succeed even when the optimizer did not report success.

These certificates concern the saved matrices. To transfer them to exact ODE endpoints one additionally needs a proved feature error $\max_i\|a_i - \tilde{a}_i\| \le e$. A separator feasible for $\tilde{A}$ then has true margins at least $1 - e\|\hat{\theta}\|$, and the two hull distances differ by at most e. These finite-matrix checks are distinct from the analytic proofs of the all-K and infinite-time theorems.

G CONSTANT SELECTION AND PROOF ASSEMBLY

The compatibility of constants matters because an existential contrast cannot be arbitrarily reduced after invoking a theorem valid only at that selected witness. The selection proceeds in this order.

Table 2: Selection order and dependencies. The base parameters (K, b, T_1) are fixed throughout. The last row uses a separately fixed positive initialization scale.

Quantity	Selection and dependence	Nature of the guarantee
c_0	First: below the common reference, stage, and margin thresholds	Fixed positive witness; selected before $\omega, \delta, m, \varepsilon$.
ω_K	Next: below $c_0^{3/2}$ and the common launch and expansion thresholds	Every $0 < \omega \le \omega_K$ is allowed with the same sufficient-width constants.
$H_K, T_{\mathrm{align},K}$	At fixed c_0, from the compact readout family, before the actual ω and draw	Common local Stage 3 entry and coordinate times; finite-cover choices, not explicit numerical budgets.
ρ_K	After the common feature and readout tolerances are fixed	One tolerance for all stages and joint conclusions, independent of $\omega, \delta, m, \varepsilon$; no explicit polynomial dependence on K.
m	After $0 < \delta < 1$: equation (8)	Explicit sufficient width in terms of the selected ρ_K and L_K, independent of ω and target risk.
$p_{\mathrm{adv}}, m_{\mathrm{adv}}$	Separately, after fixing allowed (c_0, ω)	Positive probability and an existential width threshold from the Gaussian limit and scale conditions; independent of requested confidence, with no uniform claim as $\omega \downarrow 0$.

The common-width eventwise conclusion combines the two width thresholds as stated after Theorem 4.4. The table summarizes existing choices; it does not assert a single random event valid for all ω.

1. Fix (K, b, T_1) and determine $(\mu, \beta, \Lambda_K, \overline{W})$. Obtain positive thresholds for the reference attention bootstrap, the actual all-K bridge, and the all-K readout family.
2. Before choosing c_0, also obtain the reference-margin and actual-margin thresholds of Appendix I. Choose $c_0 \le 1/4$ below these and the preceding thresholds, including $\min\{c_*, 1/(48C_P)\}$ from Lemmas C.1 and C.2. For $K \ge 3$, include the attention thresholds of orders one, two, and three at this step. No factor involving $K - 2$ is imposed at $K = 2$.
3. Choose $\omega_K \le \min\{c_0^{3/2}, 1\}$ below the reference, reference-margin, and, where applicable, higher-order expansion thresholds. At fixed c_0, obtain a common readout launch radius, H_K, and $T_{\mathrm{align},K}$ from the compact family.
4. Choose $\rho_K > 0$ so that Stage 1 concentration, the scaled error in (26), and the readout launch distance
$$d_Z + \|W^1 - \overline{W}\| \le (6c_0 C_{\mathrm{stage}} + C_w \omega_K^2)\eta_m$$
meet all fixed tolerances, and impose $\rho_K \le 1/(192C_{\mathrm{stage}})$ for the gated margin. Also impose $\rho_K \le \rho_{\mathrm{post}}$ from Appendix I, which ensures the deterministic delay and relative directional estimates. For $K \ge 3$, impose the extra order-c_0 and order-c_0^2 tolerances needed for the attention gaps. Every tolerance is independent of the subsequently chosen ω.
5. Choose (ω, δ, m) as in (8). The single Gaussian event supplies the launch estimates. All subsequent statements follow deterministically for the network obtained from that initialization.

This proves Theorems 4.1–4.3 and the positive assertions of Theorem 4.4 upon combining the component results, without an additional random event or a later width increase. The adverse probability

in Theorem 4.4(iii) is separately obtained at each fixed allowed (c_0, ω); its event has the same prescribed training protocol and a positive separated endpoint.

For the risk constant, let $a = \langle w_+, Z_+(X)\rangle$ and $c = \langle w_-, Z_-(X)\rangle$. The common bias cancels in the Lipschitz comparison, without requiring either gate to be active:

$$|f(X)| = |\operatorname{ReLU}(a+b) - \operatorname{ReLU}(c+b)| \le |a-c| \le 3(\|w_+\| + \|w_-\|) \le 3\sqrt{2}\|W\|.$$

Thus $R_3(0) \le R_{\max}$. The risk is nonincreasing along the finitely piecewise classical strict flow. For $t \ge H_K$, if $\ell_i = \log(1 + e^{-\langle W, a_i\rangle})$, then

$$q_i = 1 - e^{-\ell_i} \ge \frac{\ell_i}{1+\ell_i} \ge \frac{\ell_i}{1+NR_H}.$$

Together with the hull projection inequality, this proves (12) with its stated R_H denominator and the deterministic constant a_K.

H Numerical experiments and verification

We examine finite-width learning, fine attention structure, readout asymptotics, and transient augmented margins. Unless stated otherwise, we use $K = 3$, $b = 0.5$, $\mu = 0.13$, $c_0 = 0.02$, and $\omega_0 = c_0^{3/2}/4$, with zero initial readouts and independent Gaussian key and query factors. The standard readout-alignment time is $T_1 \approx 9.57159$, and the reference-defined attention-training duration at ω_0 is $T_* \approx 781.64101$. The parameter blocks and objectives follow Section 3. The experiments explore the predicted mechanisms beyond the displayed sufficient contrast construction; Appendix H.5 explains this relation for each design. The following subsections give the numerical observations; complete cohorts, methods, and saved-matrix checks follow in this appendix.

H.1 Finite-width learning and initialization-scale uniformity

To isolate the initialization scale, we fix $m = 262144$ and $c_0 = 0.02$ and use 20 paired initializations at each of $\omega/\omega_0 \in \{1, 1/4, 1/16, 1/64, 1/256\}$. Each scale uses its own prescribed reference duration. All 100 configurations have positive endpoint separation witnesses and correctly classify all 24 population types. The median operator-norm score error relative to the reference stays between approximately 9.479×10^{-5} and 9.490×10^{-5} across the five scales, while the attention-training duration increases from about 781.64 to 1549.43 (Table 4). This finite scan illustrates the initialization-uniform width behavior of Theorem 4.1; it does not numerically establish its continuum of scales or target-risk quantifier.

Additional trajectories with $K = 2, 4, 8$, three initializations per K, reach states with all gates active and continue to lower risk through the saved time $t_3 = 10^8$. Their complete gate and risk records are reported in the experimental appendix. These cross-component runs provide finite instances of the learning behavior for different fixed values of K.

H.2 Context-dependent classification and fine attention structure

We next distinguish endpoint learning from recovery of the finer inequalities in Theorem 4.3. Twenty confirmation seed groups are paired across widths $m \in \{256, 1024, 4096, 16384, 65536, 262144\}$, giving 120 configurations. All configurations separate and correctly classify the population. The numbers satisfying all six grouped numerical attention bounds are respectively $1, 1, 2, 9, 18, 20$ out of 20; the corresponding counts for all six signs are $1, 1, 3, 14, 20, 20$. The six groups summarize all seven displayed inequalities. Table 6 shows that endpoint learning occurs at widths for which fine attention structure is not yet consistently resolved.

For the isotropic reference, we compare $c_0 \in \{0.02, 0.01, 0.005\}$ along the paths $\omega = c_0^{3/2}/4$ and $\omega = c_0^{5/2}/4$. Tables 7 and 8 give the complete signed minima across components and attention types. In these three-point scans, $-D_2$ shows an approximately quadratic reference trend, whereas G_1 exhibits different apparent scaling along the two initialization paths. The one-sided cubic bound on G_1 does not assert a uniform $\Theta(c_0^3)$ law over those paths.

To test context dependence directly, we keep both signed signal tokens and their positions fixed and switch the context. Across the 120 confirmation endpoints, all 720 tested pairs reverse their prediction correctly. The context-key term accounts for approximately 99.8819%–100.1184% of the signed output change. Values above 100% reflect cancellation by the remaining signed contributions, not probabilities above one. The decomposition retains the actual gates and bias.

H.3 Risk and max-margin readout asymptotics

We continue the actual readout trajectories from five prespecified confirmation endpoints at $m = 262144$, with each trajectory's learned attention frozen. The maximum-margin baseline is computed from its own actual frozen features. With t denoting local readout-training time, the numerical diagnostics are

$$\widehat{E}_R(t) = \left| tR_3(t)\widehat{\gamma}^2 - 1 \right|,$$
$$\widehat{E}_V(t) = \frac{\|t\dot{W}(t) - \widehat{\theta}_{\mathrm{mm}}\|}{\|\widehat{\theta}_{\mathrm{mm}}\|},$$
$$\widehat{E}_W(t) = \frac{\|W(t)/\log t - \widehat{\theta}_{\mathrm{mm}}\|}{\|\widehat{\theta}_{\mathrm{mm}}\|},$$

where hats denote numerical maximum-margin quantities. The corresponding exact limits are stated in Theorem 4.2.

After integration to $t = 10^8$, we continue in the equivalent variable $s = \log t$. The normalized risk approaches one, and the readout-velocity error relative to the numerical optimizer reaches approximately 10^{-6} (Table 9). Across the five trajectories, the readout-vector error decreases from approximately 64.8% at $t = 10^8$ to 2.33% at $s = 512$. Thus the three diagnostics exhibit substantially different finite-time behavior.

The final point corresponds to $t \approx 2.28 \times 10^{222}$ and is used to diagnose asymptotics rather than practical training speed. Optimization certificates and numerical consistency checks are reported separately: the velocity error relative to the numerical optimizer is not a certified error bound relative to the exact maximum-margin vector, and saved-matrix certificates do not enclose the continuous ODE trajectory.

H.4 Initial augmented-margin decreases on successful trajectories

We measure the optimized augmented margin defined in (20), including optimization over its last separating coordinate; the network bias itself remains fixed. Twenty confirmation seed groups are paired at $m = 64$ and $m = 262144$. At the first saved positive attention-training times, respectively $t = 0.001$ and $t = 0.000015625$, the margin is below its initial value for $12/20$ and $9/20$ trajectories (Table 10). One additional large-width trajectory rises first and subsequently falls below its initial value within the sampled window.

All 40 confirmation trajectories are continued without resets to the prescribed attention endpoint, where they have positive separation and correctly classify all 24 population types. Complete confirmation records and high-precision checks are supplied in the supplement. Two illustrative trajectories are additionally continued to low readout risk.

These observations illustrate early augmented-margin decreases coexisting with subsequent endpoint learning, as in the phenomenon addressed by Theorem 4.4(iii). A decrease at a finite saved time does not by itself certify a negative initial right derivative. The two widths use different sampling windows, so their frequencies do not estimate a width-uniform adverse-event probability or the deterministic delay in part (ii). The reference derivative bound in part (i) and the full post-delay growth statement in part (ii) are established analytically and are not directly tested by this experiment.

H.5 Common setup and diagnostic definitions

The main numerical experiments use the exact balanced population of $N = 4K(K-1)$ input types, $d = 2K$, $b = 0.5$, $\mu = 0.13$, $c_0 = 0.02$, and $\omega_0 = c_0^{3/2}/4$ unless a scan specifies otherwise.

The four key/query factors have independent Gaussian entries with variance ω^2/m; readouts start at zero. Parameters pass between stages without resets. Algorithm 1 uses the full population in the readout stages and the conflict-conditioned population in attention training. The derivative of ReLU at zero is zero. The standard $K = 3$ clocks are $T_1 = 9.571586568$ and $T_* = 781.6410063$; attention stops at this reference-defined duration, not at a random network's own first hit.

Table 3: Distinct geometric diagnostics. The separator's last coordinate is optimized for the augmented margin; the network bias is never trained.

Observable	Feature set and optimization	Use
$M(t)$	Gated trainable features in the fixed direction $(c_\Sigma, c_\Sigma)/\sqrt{2K}$	Positive endpoint separation witness
γ	Maximum unit-norm margin of actual frozen $(Z_+, -Z_-)$	Final homogeneous readout limits
$\gamma_{\mathrm{aug}}(t)$	Maximum unit-norm margin of $(\rho_+ Z_+, -\rho_- Z_-, b(\rho_+ - \rho_-))$	Transient attention-stage geometry

In particular,

$$M(t) = \min_X y \left\langle (c_\Sigma, c_\Sigma)/\sqrt{2K}, F_{\mathrm{gate}}(X;t) \right\rangle .$$

The small-width endpoint check integrates actual factors using float64 CUDA RK4 with step 0.5 in Stage 2 and event-located DOP853 in the readout stages. The high-width experiments use the closed joint-Gram equations induced by the same Euclidean factor flow, initialized from actual Gaussian factors. This closure retains both self-Grams; it neither optimizes the score matrix as a free parameter nor imposes reference symmetry. Its state dimension is independent of width at fixed K.

For the paired width scan, Stage 1 uses DOP853 with relative/absolute tolerances $10^{-11}/10^{-14}$, maximum step 0.5, and gate-event grouping tolerance 10^{-12}. Stage 2 uses $10^{-11}/(\omega^2 10^{-12})$ and maximum step 5. Each module retains its own solver settings and event records. Reference, scale, margin, readout, and cross-component experiments use the configurations documented below and in the accompanying index. Post-processing, context switches, matrix certificates, and continuations do not add independent training initializations.

Relation to the sufficient theoretical regime. The population, architecture, initialization, and stage protocol follow the theoretical setting, while the numerical contrasts exceed the proof's displayed sufficient construction. Appendix C selects $c_* \le 1/(64D)$, where $D = A_0 M^2/\Lambda$, $A_0 \ge 1$, and $M^2 = 32K$. Consequently its displayed construction requires

$$c_0 \le c_* \le \frac{\Lambda}{2048K} = \frac{\mu}{9216K(K-1)} \approx 2.351 \times 10^{-6} \qquad (K = 3,\ \mu = 0.13),$$

whereas the main numerical setting uses $c_0 = 0.02$. At $\mu = 0.13$, this ceiling is at most 7.054×10^{-6} over $K = 2, 3, 4, 8$, below even the smallest reference-scan contrast 0.005. Thus none of the reported training cohorts or reference paths instantiates this full sufficient construction; shrinking ω or increasing width alone does not change that contrast comparison. The displayed ceiling is only a coarse upper bound on the admissible contrast, not a complete sufficient condition or an executable choice of c_0; all remaining constraints in Appendix G must also hold. The full construction is sufficient, not necessary for learning. The experiments examine how its mechanisms persist at larger contrasts. The purpose of each design is specified below; conditional readout diagnostics and finite-matrix certificates address their own stated hypotheses. Finite observations do not certify infinite-time limits, probability quantifiers, or continuous ODE error bounds.

H.6 Finite-width and cross-component learning

Small-width endpoint check. At $K = 3$ and widths $64, 128, 256$, the earlier cohort has positive endpoint separation and correct classification of all 24 types in all 60 configurations (20 per width; seeds 20260910 and 2026091401–2026091419); its factor-flow settings and complete endpoint records are supplied for this compact check. At $c_0 = 0.02$, this tests whether endpoint separation also occurs with substantially smaller factor widths, beyond the displayed sufficient construction.

Fixed-width initialization scan. Twenty seed groups, 2026092401–2026092420, are paired across five initialization scales at $m = 262144$. For each seed, PCG64 draws the two-head key array and then the query array, each of shape $(2, m, 6)$. A single standard Gaussian Gram matrix is rescaled by ω^2 at each scale. Actual Stage 1 is performed from zero readouts, followed by Stage 2 with that run's readouts frozen. Each scale has its own prescribed reference clock. Keeping $c_0 = 0.02$ fixed while reducing ω probes whether separation and reference tracking persist over longer attention phases at one width; the contrast remains outside the displayed sufficient construction.

Table 4: Fixed $c_0 = 0.02$ and width 262144. Score error is $\max_h \|P_h^U(T_*) - \bar{P}_h^U(T_*)\|_{\mathrm{op}}$, without division by c_0. The last column preserves the finer sign outcomes.

ω/ω_0	T_*	Separated and correct	Median score error	Six signs
1	781.6410	20/20	9.489878×10^{-5}	20/20
1/4	973.5894	20/20	9.481357×10^{-5}	20/20
1/16	1165.5378	20/20	9.479533×10^{-5}	20/20
1/64	1357.4862	20/20	9.479222×10^{-5}	20/20
1/256	1549.4347	20/20	9.479216×10^{-5}	18/20

At the smallest scale, seeds 2026092406 and 2026092417 have negative G_1, approximately -1.10×10^{-8} and -8.25×10^{-9}, respectively. These outcomes remain in the complete cohort; stable endpoint learning does not imply that all fine attention signs are stable. The complete scale records include the readout-interface error and the Gaussian Gram deviation. Four tighter-tolerance endpoint comparisons agree to floating-point precision, without providing a rigorous ODE enclosure.

Cross-component trajectories. The width-262144 trajectories use seeds 2026092601–2026092603, with separate PCG64 streams derived from each (K, seed) pair. The common $c_0 = 0.02$ exceeds the displayed ceiling at every tested K; this comparison explores learning and gate activation as the number of correspondences grows while the number of heads stays two. All nine reach the full-activation region and remain there through the saved horizon $t_3 = 10^8$. Table 5 reports the observed times and risks; the complete per-seed gate and risk records accompany the supplement. The entry time is the infimum of the right-hand all-active region: a gate exactly at a crossing is zero. Finite observations cannot certify future permanent activation.

Table 5: Cross-component trajectories, width 262144, three per K. All times are local Stage 3 times.

K	First all-active entry time	Median $R_3(10^8)$
2	6555.85–6581.06	0.0005082221
4	13171.03–13195.25	0.001017490
8	26393.72–26418.87	0.002048006

For these nine runs, Stage 2 uses DOP853 on the normalized Gram variable $U(t) = C(t)/(\omega^2 e^{\Lambda_K t})$, with relative and absolute tolerances 10^{-11} and 10^{-13}, and maximum step $0.07/\Lambda_K$ in local Stage 2 time. The actual Stage 1 readouts stay fixed, and integration ends at the saved reference clock. The original driver records these settings directly; the supplement's cross-component design and per-run records identify the source snapshot and its hash.

The corresponding product of entry time and each trajectory's frozen feature margin squared lies in $[0.129726, 0.130728]$. This describes these nine instances, without establishing a general complexity law in K.

H.7 Fine attention structure and mechanism

Paired width scan. The five exploratory seeds 2026092201–2026092205 and the twenty confirmation seeds 2026092301–2026092320 are kept separate. Widths are

$256, 1024, 4096, 16384, 65536, 262144$. For each seed, key and query arrays are drawn at the largest width; width m takes the first m rows and scales them by $\omega/\sqrt{m}$. Thus widths are paired within a seed. There are 30 exploratory and 120 confirmation configurations, not 150 independent seeds. All 150 endpoints separate and classify correctly. At the common $c_0 = 0.02$, the scan explores how width affects classification and the finer attention gaps beyond the displayed construction. Passing the endpoint inequalities is a diagnostic of the learned representation.

Let H be the minimum signed group-key probability difference between heads for classification queries and J the minimum positive-label group-key difference between the two classification queries. Together with $-D_2, D_3, -yD_1, G_1$, these define six groups, with minima taken over every applicable conflict type. The H group covers two displayed inequalities, so the six groups cover all seven inequalities of Theorem 4.3. Each numerical bound is evaluated at the prescribed c_0.

Table 6: Complete paired-width counts. Exploratory and confirmation denominators are separate; no configuration is excluded.

m	Exploratory signs	Exploratory bounds	Confirmation signs	Confirmation bounds
256	0/5	0/5	1/20	1/20
1024	0/5	0/5	1/20	1/20
4096	1/5	0/5	3/20	2/20
16384	3/5	1/5	14/20	9/20
65536	5/5	5/5	20/20	18/20
262144	5/5	5/5	20/20	20/20

For exploratory seed 2026092204 at $m = 256, 65536$, tighter tolerances change endpoint attention probabilities by at most 1.23×10^{-15}. The saved-data verifier independently regenerates the Gaussian initial Grams and recomputes the endpoint diagnostics.

Reference paths and reused source records. The deterministic isotropic reference uses the reference readout and $C_h(0) = \omega^2 I$; it is not a Gaussian sample. The new comparison uses $K = 2, 3, 4, 8$, $c_0 = 0.02, 0.01, 0.005$, and the two paths $\omega = c_0^{3/2}/4$ and $\omega = c_0^{5/2}/4$. Even the smallest contrast exceeds the displayed ceiling for each tested K. These deterministic paths isolate cancellation and gap-scaling mechanisms; they do not sample the random finite-width learning event. Six $K = 3$ endpoints are explicitly reused from the earlier reference archive; the other 18 are new reference configurations. Tables 7 and 8 list all six minima for all 24 cases.

The approximately second- or third-order trends are three-point empirical comparisons, not uniform asymptotic equivalences. At $K = 2$, the reference J, D_2, D_3 cancel, but this does not imply that every fine comparison vanishes. Theorem 4.3 only asserts its inequalities for $K \geq 3$.

Context switches and signed output decomposition. Each of the 120 confirmation endpoints supplies six context-switch pairs. These post-processing comparisons inherit the width scan's larger contrast and examine how the learned representation uses context to select a signal. Both signed signal tokens remain in their physical positions; only the context changes, which reverses the relevant label. The output is decomposed using each head's actual gate, all queries, each key's attention probability, and its readout value; the signed bias term is retained. The reconstructed output agrees within 4.44×10^{-16} in the independent check. All 720 pairs reverse correctly. The context-key share of the signed output change ranges from 99.8819% to 100.1184%. In the finer context-query comparison, 43 of 1440 conflict terms are negative. These terms are retained: correct classification and a dominant summed context contribution do not require every finer term to be positive.

Table 7: All 24 deterministic reference endpoints, with $\omega = c_0^p/4$: the first three signed minima. H and J denote the head-gap and cross-query group-key comparisons. Six $K = 3$ records reuse saved endpoints, identified in the evidence index. Zeros are recorded numerical values; the structural $K = 2$ cancellations concern J, D_2, D_3. The remaining comparisons appear in Table 8.

K	p	c_0	H	J	$-D_2$
2	1.5	0.02	4.44×10^{-3}	0	0
2	1.5	0.01	2.22×10^{-3}	0	0
2	1.5	0.005	1.11×10^{-3}	0	0
2	2.5	0.02	4.44×10^{-3}	0	0
2	2.5	0.01	2.22×10^{-3}	0	0
2	2.5	0.005	1.11×10^{-3}	0	0
3	1.5	0.02	2.96×10^{-3}	1.49×10^{-3}	2.55×10^{-6}
3	1.5	0.01	1.48×10^{-3}	7.42×10^{-4}	6.36×10^{-7}
3	1.5	0.005	7.41×10^{-4}	3.71×10^{-4}	1.59×10^{-7}
3	2.5	0.02	2.96×10^{-3}	1.49×10^{-3}	2.54×10^{-6}
3	2.5	0.01	1.48×10^{-3}	7.42×10^{-4}	6.34×10^{-7}
3	2.5	0.005	7.41×10^{-4}	3.71×10^{-4}	1.59×10^{-7}
4	1.5	0.02	2.22×10^{-3}	2.23×10^{-3}	3.83×10^{-6}
4	1.5	0.01	1.11×10^{-3}	1.11×10^{-3}	9.56×10^{-7}
4	1.5	0.005	5.56×10^{-4}	5.56×10^{-4}	2.38×10^{-7}
4	2.5	0.02	2.22×10^{-3}	2.23×10^{-3}	3.81×10^{-6}
4	2.5	0.01	1.11×10^{-3}	1.11×10^{-3}	9.53×10^{-7}
4	2.5	0.005	5.56×10^{-4}	5.56×10^{-4}	2.38×10^{-7}
8	1.5	0.02	1.11×10^{-3}	3.34×10^{-3}	5.79×10^{-6}
8	1.5	0.01	5.56×10^{-4}	1.67×10^{-3}	1.44×10^{-6}
8	1.5	0.005	2.78×10^{-4}	8.34×10^{-4}	3.58×10^{-7}
8	2.5	0.02	1.11×10^{-3}	3.34×10^{-3}	5.76×10^{-6}
8	2.5	0.01	5.56×10^{-4}	1.67×10^{-3}	1.43×10^{-6}
8	2.5	0.005	2.78×10^{-4}	8.34×10^{-4}	3.58×10^{-7}

Table 8: The remaining three signed minima for the same 24 deterministic reference endpoints as Table 7. All configurations, including recorded zeros, are retained.

K	p	c_0	D_3	$-yD_1$	G_1
2	1.5	0.02	0	2.30×10^{-11}	1.57×10^{-5}
2	1.5	0.01	0	7.20×10^{-13}	3.93×10^{-6}
2	1.5	0.005	0	2.25×10^{-14}	9.82×10^{-7}
2	2.5	0.02	0	1.67×10^{-16}	3.15×10^{-7}
2	2.5	0.01	0	0	3.94×10^{-8}
2	2.5	0.005	0	0	4.93×10^{-9}
3	1.5	0.02	2.56×10^{-6}	7.86×10^{-6}	1.57×10^{-5}
3	1.5	0.01	6.37×10^{-7}	1.96×10^{-6}	3.93×10^{-6}
3	1.5	0.005	1.59×10^{-7}	4.91×10^{-7}	9.82×10^{-7}
3	2.5	0.02	2.54×10^{-6}	1.58×10^{-7}	3.15×10^{-7}
3	2.5	0.01	6.35×10^{-7}	1.97×10^{-8}	3.94×10^{-8}
3	2.5	0.005	1.59×10^{-7}	2.46×10^{-9}	4.93×10^{-9}
4	1.5	0.02	3.85×10^{-6}	1.18×10^{-5}	1.57×10^{-5}
4	1.5	0.01	9.57×10^{-7}	2.95×10^{-6}	3.93×10^{-6}
4	1.5	0.005	2.39×10^{-7}	7.37×10^{-7}	9.82×10^{-7}
4	2.5	0.02	3.83×10^{-6}	2.36×10^{-7}	3.15×10^{-7}
4	2.5	0.01	9.54×10^{-7}	2.96×10^{-8}	3.94×10^{-8}
4	2.5	0.005	2.38×10^{-7}	3.69×10^{-9}	4.93×10^{-9}
8	1.5	0.02	5.82×10^{-6}	1.77×10^{-5}	1.57×10^{-5}
8	1.5	0.01	1.44×10^{-6}	4.42×10^{-6}	3.93×10^{-6}
8	1.5	0.005	3.59×10^{-7}	1.10×10^{-6}	9.82×10^{-7}
8	2.5	0.02	5.78×10^{-6}	3.55×10^{-7}	3.15×10^{-7}
8	2.5	0.01	1.44×10^{-6}	4.43×10^{-8}	3.94×10^{-8}
8	2.5	0.005	3.58×10^{-7}	5.54×10^{-9}	4.93×10^{-9}

H.8 READOUT ASYMPTOTICS

The five continuations reuse width-262144 confirmation seeds 2026092301–2026092305. Each begins from its own actual Stage 2 endpoint and freezes its learned attention. The source attention runs use $c_0 = 0.02$ outside the displayed sufficient construction. Fixing their actual learned features isolates the conditional readout predictions and the different convergence rates of the three diagnostics. The homogeneous signed feature matrix is $A_X = y(Z_+(X), -Z_-(X))$, with a separate numerical maximum-margin solution $\widehat{\theta}_{\mathrm{mm}}$ for each trajectory. Ten saved joins between the source endpoint, ordinary Stage 3 integration, and logarithmic continuation match exactly.

Ordinary readout training is continued to $t_3 = 10^8$; subsequent integration uses $s = \log t_3$, with $dW/ds = t_3\dot{W}$. Stable exponential expressions evaluate risk and velocity without first underflowing the loss. The log-time solver is event-resolved Radau with an analytic Jacobian, relative/absolute tolerances $10^{-9}/10^{-11}$, and maximum log-time step 8. The full three-diagnostic records appear in Table 9.

Table 9: Ranges across all five readout continuations, using each trajectory's own numerical maximum-margin solution. Log time is $s = \log t_3$; $s = 512$ corresponds to approximately 2.2844×10^{222} units of physical time. A common exponent multiplies both endpoints of each displayed interval. These are numerical errors relative to the computed optimizer.

s	t_3	Risk error $\widehat{E}_R$
18.4207	1.000×10^{8}	$[7.537, 7.594]\times10^{-3}$
32	7.896×10^{13}	$[1.22, 4.92]\times10^{-7}$
64	6.235×10^{27}	$[1.567, 5.262]\times10^{-7}$
128	3.888×10^{55}	$[1.567, 5.262]\times10^{-7}$
256	1.511×10^{111}	$[1.567, 5.261]\times10^{-7}$
512	2.284×10^{222}	$[1.567, 5.26]\times10^{-7}$

s	Velocity error $\widehat{E}_V$	Vector error $\widehat{E}_W$
18.4207	$[6.815, 6.998]\times10^{-3}$	$[6.477, 6.482]\times10^{-1}$
32	$[1.915, 5.588]\times10^{-5}$	$[3.726, 3.729]\times10^{-1}$
64	$[0.9383, 1.566]\times10^{-6}$	$[1.863, 1.864]\times10^{-1}$
128	$[0.9431, 1.568]\times10^{-6}$	$[9.315, 9.322]\times10^{-2}$
256	$[0.9422, 1.568]\times10^{-6}$	$[4.657, 4.661]\times10^{-2}$
512	$[0.9404, 1.566]\times10^{-6}$	$[2.329, 2.33]\times10^{-2}$

At $s = 512$, the velocity error relative to the numerical optimizer ranges from 9.40×10^{-7} to 1.57×10^{-6}; the normalized risk at saved $s \geq 32$ lies between 0.99999947 and 0.99999988. The much slower vector-amplitude convergence accounts for the remaining roughly 2.33% vector error. The saved gates are all active, and all signed readout coordinates are positive in these reported tail states. Positivity alone does not certify every eventual coordinate constant in Theorem 4.2.

The maximum-margin certificates concern these five actual frozen matrices, with relative optimizer uncertainty bounded by approximately 4.59×10^{-6}. Thus the approximately 10^{-6} velocity diagnostic against $\widehat{\theta}_{\mathrm{mm}}$ is not itself a bound against the exact optimizer. A pointwise check of the ordinary/log-time vector-field identity on 45 saved states has relative discrepancy below 2.39×10^{-13}. The first-seed tighter-tolerance log-time continuation differs in W by at most 6.79×10^{-13} relatively. A separately integrated overlap comparison is documented in the source run, but its second solution array was not retained; that particular discrepancy cannot be independently recomputed without integration.

H.9 TRANSIENT AUGMENTED MARGINS

The exploratory cohort has 12 width-64 seeds 2026092401–2026092412, with first-sample decreases in 7/12 cases. The confirmation cohort uses 20 separate seed labels 2026092501–2026092520 paired at widths 64 and 262144. These counts are not pooled. The lower-width early

grid is $0, 0.001, 0.01, 0.1, 0.3, 1, 3, 10$; the higher-width grid is that grid divided by 64. The full-support augmented margin optimizes its last separating coordinate, while the model bias remains fixed. At $c_0 = 0.02$, these paired widths explore whether early sampled geometric deterioration coexists with endpoint learning beyond the displayed construction; the separate exploratory cohort addresses the same mechanism.

Table 10: Complete confirmation counts for the optimized augmented margin. The first positive times differ between the two widths.

m	First positive time	First-point decrease	Any sampled decrease	Endpoint learning
64	0.001	12/20	12/20	20/20
262144	0.000015625	9/20	10/20	20/20

Each continuation starts from the final saved Gram state and Stage 2 time of its corresponding early trajectory and carries forward the actual Stage 1 readout without parameter resets. All 40 saved continuation readouts agree elementwise with their source records. The endpoint optimized margin ranges are $[0.00277446802, 0.00422391798]$ at width 64 and $[0.00362830661, 0.00364541560]$ at width 262144. The complete 320 early observations, 220 later saved states, per-seed decisions, and source identifiers are supplied in machine-readable tables. The early optimized matrices and their certificates remain separately available, including the 96 exploratory observations.

The two illustrative continuations use the lowest seed label at each width with a first-point decrease and a successful endpoint: 2026092502 at width 64 and 2026092503 at width 262144. This selection does not represent every initialization. At high width, seed 2026092504 initially rises and only later falls below its initial margin; it is included in the 10/20 window count, not the 9/20 first-point count.

The two selected cases also have readout continuations to $t_3 = 10^8$, with risks 0.00090258749 and 0.00076475976, respectively; both classify all 24 types correctly and have 48 active gates at the saved endpoint. These are two illustrative continuations, not a 40-run Stage 3 success rate. Full first-point decisions were rechecked with 80-digit feature evaluation; independent 90-digit recomputation preserved all 40 decisions. Two tighter integration checks give maximal saved Gram and readout differences of 1.17×10^{-21} and 5.56×10^{-16}. These finite comparisons do not certify a negative initial right derivative, a width-uniform adverse-event probability, or the post-delay growth theorem.

H.10 Numerical reliability and saved-matrix certificates

The following checks concern only the retained augmented-margin, readout, and cross-component experiments. Appendix F supplies the primal–dual construction used for the actual frozen readout matrices. These checks directly verify finite-matrix inequalities under their algebraic hypotheses. Their validity for a saved matrix does not require its generating trajectory to lie in the sufficient training-parameter construction.

For augmented-margin states, a nonnegative rational simplex vector λ gives $p = A^\top \lambda$, $U = \|p\|^2$ and $q = \min_i a_i^\top p$. If $q > 0$, the finite-matrix margin obeys $q^2/U \le \gamma_{\mathrm{aug}}^2 \le U$. A saved-state decrease is certified by comparing the later upper bound with the initial lower bound. All 636 binary64 augmented-matrix certificates pass; the maximum relative squared-margin interval width is 2.52×10^{-56}. Certificates for Decimal-evaluated matrices are kept distinct from those for binary64 matrices. Neither kind is an interval enclosure of real softmax evaluation or of the exact ODE.

The original matrices, solver settings, rational bounds, and coverage limits are indexed separately for augmented-margin and homogeneous readout certificates. A certificate for one feature matrix does not transfer to another feature definition or state. Experiment-specific numerical comparisons are reported with the corresponding experiments above.

Table 11: Certificates for distinct saved finite matrices. The final column is a rounded descriptive maximum of relative primal–dual gaps (homogeneous features) or squared-margin interval widths (augmented features). Exact rational bounds and definitions remain in the supplied records; rounded values here are not new exact certificates.

Certified object	Count	Maximum relative gap
Augmented features (all saved states)	636	2.52×10^{-56}
Homogeneous readout features (five continuations)	5	2.10×10^{-11}
Homogeneous readout features (cross-component runs)	9	5.16×10^{-11}

I TRANSIENT FEATURE-MARGIN DYNAMICS

This appendix studies the full-support augmented features in equation (20). Throughout, time is local Stage 2 time, $\Lambda = \Lambda_K$, and $h_{\mathrm{geom}} = \sqrt{2/K}/9$ is a fixed geometric constant (not a width-dependent quantity). All factors and frozen readouts are the actual ones passed from Stage 1 unless a reference is explicitly specified.

Notation. Throughout this appendix write

$$z_X(t) := yF_{\mathrm{aug}}(X;t), \qquad \bar{z}_X(t) := y\bar{F}_{\mathrm{aug}}(X;t),$$

and let $\mathcal{H}(t) = \mathrm{conv}\{z_X(t) : X\}$ denote the labeled-feature hull. A bar always denotes the isotropic Stage 2 reference trajectory. The augmented coordinate is retained in this appendix; thus these vectors live in $\mathbb{R}^{2d+1}$ and are different from the homogeneous Stage 3 features.

Lemma I.1 (Closest-point margin and one-sided sensitivity)**.** *Let $z_1, \dots, z_n$ be C^1 maps from an open interval I to a Euclidean space and put*

$$\gamma(t) = \max_{\|u\|=1} \min_i \langle u, z_i(t)\rangle, \quad q(t,\alpha) = \sum_i \alpha_i z_i(t), \quad \Delta_n = \{\alpha \ge 0 : \sum_i \alpha_i = 1\}.$$

At a time t_0 with $\gamma(t_0) > 0$, let $A(t_0) = \arg\min_{\alpha\in\Delta_n} \|q(t_0,\alpha)\|$ and let q_0 be the closest point. Then q_0 is independent of the minimizing representation, $\gamma(t_0) = \|q_0\|$, and the unique maximizing unit vector is $u_0 = q_0/\|q_0\|$. Moreover

$$\gamma'_+(t_0) = \min_{\alpha\in A(t_0)} \sum_i \alpha_i \langle u_0, z_i'(t_0)\rangle, \qquad \gamma'_-(t_0) = \max_{\alpha\in A(t_0)} \sum_i \alpha_i \langle u_0, z_i'(t_0)\rangle, \tag{M.1}$$

where the second expression is the left derivative in the convention $\gamma'_-(t_0) = \lim_{h\downarrow0}(\gamma(t_0) - \gamma(t_0 - h))/h$. In particular, if every minimizing representation has the same weighted velocity, then γ is differentiable at t_0. Every $\alpha \in A(t_0)$ is supported on the exposed face $\langle u_0, z_i(t_0)\rangle = \gamma(t_0)$: a strictly larger projection forces $\alpha_i = 0$.

Proof. Let $C(t) = \mathrm{conv}\{z_i(t)\}$. Positivity of $\gamma(t_0)$ separates 0 from $C(t_0)$. The squared norm has a unique minimizer q_0 on this compact convex set. The projection inequality $\langle q_0, z - q_0\rangle \ge 0$ for $z \in C(t_0)$ gives $\min_i \langle q_0/\|q_0\|, z_i\rangle \ge \|q_0\|$. Conversely, for any unit u and any convex representation of q_0, $\min_i \langle u, z_i\rangle \le \langle u, q_0\rangle \le \|q_0\|$. Thus the two quantities are equal and equality in Cauchy–Schwarz gives uniqueness of the maximizing direction. The support assertion follows by averaging the projection inequalities with any $\alpha \in A(t_0)$.

The minimum remains positive in a neighborhood of t_0. For a fixed $\alpha \in A(t_0)$, differentiating $\|q(t,\alpha)\|$ gives the weighted velocity in (M.1), hence the right derivative is at most its minimum. For the reverse inequality choose a minimizing α_h at $t_0 + h$ and pass to a convergent subsequence in the compact simplex. Its limit belongs to $A(t_0)$, and uniform differentiability of the norm near the nonzero point q_0 gives the opposite bound. Replacing h by $-h$ gives the left formula. This also proves the final assertion. □

Lemma I.2 (Relative closest-hull stability)**.** *Let $C, \bar{C}$ be compact convex sets in a Euclidean space with Hausdorff distance at most δ, and let $p, \bar{p}$ be their closest points to the origin. If $\bar{g} = \|\bar{p}\| > 0$, then*

$$|\|p\| - \bar{g}| \le \delta, \qquad \|p - \bar{p}\|^2 \le (\|p\| + \bar{g})\delta. \tag{M.2}$$

Consequently, whenever $r := \delta/\bar{g} \le 1/2$, $p \neq 0$ *and*

$$\left\| \frac{p}{\|p\|} - \frac{\bar{p}}{\bar{g}} \right\| \le \sqrt{10r}. \tag{M.3}$$

The same assertions hold for finite hulls when each generator is perturbed by at most δ.

Proof. Choose points of either set within δ of the other. Minimality of the two projections gives the first inequality in (M.2). The projection inequalities $\langle p, x - p\rangle \ge 0$ and $\langle \bar{p}, \bar{x} - \bar{p}\rangle \ge 0$, applied to cross-near points, give $\|p - \bar{p}\|^2 \le (\|p\| + \bar{g})\delta$. If $r \le 1/2$, then $\|p\| \le (1 + r)\bar{g}$ and normalizing the two vectors yields $\|p/\|p\| - \bar{p}/\bar{g}\| \le 2\|p - \bar{p}\|/\bar{g} \le 2\sqrt{(2 + r)r} \le \sqrt{10r}$. Taking the same convex coefficients proves the finite-hull Hausdorff bound. □

The reference orbit and its launch. We next record the calculation used in the positive part of Theorem 4.4. Put

$$a(t) = \frac{1}{|\mathcal{C}|} \sum_{X \in \mathcal{C}} \bar{z}_X(t), \qquad b_c(t) = \frac{1}{|\mathcal{S}|} \sum_{X \in \mathcal{S}} \bar{z}_X(t),$$

where $\mathcal{C}$ and $\mathcal{S}$ are respectively the conflict and consistent type sets. Let

$$e_C = \frac{(c_\Sigma, c_\Sigma, 0)}{\sqrt{2K}}, \quad e_V = \frac{(v_\Sigma, -v_\Sigma, 0)}{\sqrt{2K}}, \quad e_b = (0, 0, 1), \quad a_0 = \frac{1}{\sqrt{2K}}.$$

These vectors are orthonormal. At zero scores, the mixed gates give

$$a(0) = 0, \qquad b_c(0) = b_0 = a_0 e_C + 2a_0 e_V + b e_b. \tag{M.4}$$

Fix once and for all a positive contrast threshold c_{post} below the reference and endpoint thresholds selected in the preceding appendices, and write $Q_{\mathrm{post}} = 4c_{\mathrm{post}}$. The theorem chooses $0 < c_0 \le c_{\mathrm{post}}$, so $q(T_*) \le Q_{\mathrm{post}}$ is uniform in the subsequently chosen c_0 and ω. For the positive-head linear score write

$$H = \begin{pmatrix} 0 & zT \\ rT & 0 \end{pmatrix}, \qquad P_+(H) = H, \quad P_-(H) = -H.$$

The correct token signs are essential: direct differentiation of the three-token softmax at zero gives

$$Da(0)[H, -H] = h_{\mathrm{geom}}(2z - r)e_C, \qquad h_{\mathrm{geom}} = \frac{\sqrt{2/K}}{9}. \tag{52}$$

Indeed, for a conflict put $\Delta = v_k - v_j$ and $u = c_k + y\Delta$. The group coordinate has derivative

$$D\langle c_\Sigma, Z_H\rangle(0) = \frac{1}{3} c_k^\top H u - \frac{1}{9} u^\top H u = \frac{y(2z - r)}{9},$$

while the classification coordinate is independent of y after averaging over ordered pairs and hence cancels after multiplication by the label. The two heads and the normalization of the conflict average supply the factor $1/K$ and the displayed e_C coefficient.

Set

$$q(t) = \omega^2 e^{\Lambda_K t}, \quad \ell(t) = \omega^2 \sinh(\Lambda_K t), \quad s(t) = \omega^2 \sinh(\Lambda_K t/2), \quad \nu(t) = \ell'(t).$$

The linear Gram orbit has $z = \ell$, $r = -s$. In the following display $C(t)$ and $C_{\mathrm{lin}}(t)$ denote the reference Gram pair and its linearization. By Lemma C.1 and the smooth active Gram equation, there is a constant C depending only on (K, b, T_1) such that, up to T_*,

$$\|C(t)\| \le Cq(t), \qquad \|C(t) - C_{\mathrm{lin}}(t)\| \le Cq(t)^2, \tag{53}$$

$$\|C'(t) - C'_{\mathrm{lin}}(t)\| \le Cq(t)^2, \qquad \|C'(t)\| \le Cq(t). \tag{M.5}$$

The velocity error follows by substituting $C' - C'_{\mathrm{lin}} = L(C - C_{\mathrm{lin}}) + R(C)$ in the reference field, where $\|R(C)\| \le C\|C\|^2$. Since the score pair starts at zero, integrating $\|P'\| \le Cq$ also gives $\|P(t)\| \le C\ell(t)$. At time zero the exact score gradient is evaluated at zero scores and the self-Grams are $\omega^2 I$, so $P'(0) = P'_{\mathrm{lin}}(0)$ exactly. For the fixed-gate orbit map,

$$a' - Da(0)[P'_{\mathrm{lin}}] = Da(0)[P' - P'_{\mathrm{lin}}] + (Da(P) - Da(0))[P'] = O(q^2).$$

Both $a - Da(0)[P_{\mathrm{lin}}]$ and $b_c - b_0$ vanish at zero. Integrating $\int_0^t q(u)^2 du = q(t)\ell(t)/\Lambda_K$, and using $q/\nu \le 2/\Lambda_K$ and the bounded first derivative of b_c, gives

$$\|a(t) - h_{\mathrm{geom}}(2\ell + s)e_C\| \le Cq\ell, \qquad \|a'(t) - h_{\mathrm{geom}}(2\nu + s')e_C\| \le Cq\nu, \tag{M.6}$$

$$\|b_c(t) - b_0\| \le C\ell, \qquad \|b_c'(t)\| \le C\nu. \tag{M.7}$$

These are relative estimates at the launch because q/ℓ is not bounded there; the error is $O(q\ell)$ rather than $O(q^2)$.

Choose the additional reference contrast threshold $c_{\mathrm{mono}} > 0$ so small that (M.6)–(M.7), together with $q(T_*) \le 4c_0$, imply for $t > 0$

$$\langle a, b_c - a\rangle \ge h_{\mathrm{geom}} a_0 \ell, \qquad \langle a, a'\rangle \ge 2h_{\mathrm{geom}}^2 \ell\nu, \qquad h_{\mathrm{geom}}\ell \le \|a\| \le 3h_{\mathrm{geom}}\ell. \tag{M.8}$$

This choice is made before c_0 and ω, and all mixed reference gates remain fixed by the preceding gate-slack estimate.

Lemma I.3 (Reference full-support growth)**.** *Under the preceding threshold choice, $\bar\gamma_{\mathrm{aug}}(0) = 0$ and, for $0 < t \le T_*$,*

$$\bar\gamma_{\mathrm{aug}}(t) = \|a(t)\|, \qquad \bar\gamma_{\mathrm{aug}}'(t) \ge \frac{h_{\mathrm{geom}}}{2}\nu(t), \quad \bar\gamma_{\mathrm{aug}}'(0+) = \frac{5h_{\mathrm{geom}}}{2}\Lambda_K\omega^2. \tag{M.9}$$

Every consistent type has strictly larger signed projection than the conflicting types, and every closest-point simplex is supported on conflicts.

Proof. Let $Jc_i = c_i$ and $Jv_i = -v_i$. Simultaneous index permutations and the orthogonal action $(u_+, u_-, u_b) \mapsto (Ju_-, Ju_+, u_b)$ preserve the reference labeled-feature hull. They act transitively on conflicting types and on consistent types separately, including at $K = 2$. On $U \oplus U \oplus \mathbb{R}$, their group average is the orthogonal projection onto

$$E = \{(Ac_\Sigma + Bv_\Sigma, Ac_\Sigma - Bv_\Sigma, c) : A, B, c \in \mathbb{R}\}.$$

It maps the full hull onto the segment $[a, b_c]$, which itself lies in the hull. The unique minimum-norm point of the full hull is fixed by the group and hence is the minimum point of this segment. By (M.8), the derivative at a of $\|a + x(b_c - a)\|^2$ is positive, so that point is a. The closest-point identity in Lemma I.1 gives $\bar\gamma = \|a\|$ and the unique direction $\bar u = a/\|a\|$. The first inequality in (M.8) is exactly the strict projection gap for consistent types. The second and the upper norm bound give

$$\bar\gamma' = \frac{\langle a, a'\rangle}{\|a\|} \ge \frac{2h_{\mathrm{geom}}}{3}\nu \ge \frac{h_{\mathrm{geom}}}{2}\nu.$$

At zero, $a(0) = 0$ and the bias direction has zero conflict projection and positive consistent projection, so the margin is zero. Equation (52) with $z'(0) = \Lambda_K\omega^2$ and $r'(0) = -\Lambda_K\omega^2/2$ gives the launch value in (M.9). The support claim follows from the last assertion of Lemma I.1. The group action is independent of time, so it also permutes the reference feature velocities. Since $\bar u$ is group-fixed, every conflict has $\langle \bar u, \bar z_X'\rangle = \langle \bar u, a'\rangle = \bar\gamma_{\mathrm{aug}}'$. This pointwise velocity identity is used in the perturbation argument below. □

Lemma I.4 (Deterministic delayed-growth bridge)**.** *Let $\bar C$ be the reference trajectory above and let an actual Stage 2 trajectory start from arbitrary restricted factors and frozen readouts. Set*

$$e = \frac{\|C(0) - \bar C(0)\|_F}{\omega^2} + (1 + T_*)\|W - \bar W\|.$$

For the constants used in the preceding relative-perturbation appendix we write

$$M_r := M, \qquad A_{\mathrm{eff}} := A, \qquad B_{\mathrm{force}} := B, \qquad B_{\mathrm{eff}} := B_e,$$

where M, A, B, B_e are the explicitly selected reference-envelope, nonlinear-Lipschitz, readout-forcing, and relative-error constants there. The renamed constants retain their parameter dependence from Appendix C. There are constants $e_{\mathrm{eff}}, c_{\mathrm{eff}} > 0$, depending only on (K, b, T_1), such that if $c_0 \le c_{\mathrm{eff}}$ and $e \le e_{\mathrm{eff}}$, then the actual flow exists on $[0, T_*]$, has the reference mixed gates with uniform slack, and for every $\tau \in (0, T_*)$ satisfying

$$e \le \varepsilon_{\mathrm{post}}(1 - e^{-2\Lambda_K\tau}), \tag{M.10}$$

one has on $[\tau, T_*]$

$$\gamma_{\mathrm{aug}}(t) \geq \frac{h_{\mathrm{geom}}}{2}\omega^2 \sinh(\Lambda_K t), \tag{M.11}$$

$$\frac{d^+}{dt}\gamma_{\mathrm{aug}}(t) \geq \frac{h_{\mathrm{geom}}}{4}\Lambda_K \omega^2 \cosh(\Lambda_K t), \qquad t < T_*. \tag{M.12}$$

In addition, after decreasing c_{eff} *and* e_{eff} *if needed, the endpoint obeys the deterministic full-support separator*

$$\min_X y \left\langle \frac{(c_\Sigma, c_\Sigma, 0)}{\sqrt{2K}}, F_{\mathrm{aug}}(X; T_*) \right\rangle \geq \frac{c_0}{8\sqrt{2K}}. \tag{M.13a}$$

For every $t \in [\tau, T_*]$*, each consistent type has signed projection at least* $1/(6\sqrt{2K})$ *above the margin, so every closest-point simplex is supported on conflicts. One valid choice is*

$$\begin{aligned} \varepsilon_{\mathrm{post}} &= \min\left\{ e_{\mathrm{eff}}, \frac{h_{\mathrm{geom}}}{4C_Z}, \frac{h_{\mathrm{geom}}\theta^2}{20C_Z}, \frac{a_0}{24C_Z Q_{\mathrm{post}}}, \frac{h_{\mathrm{geom}}\Lambda_K}{16C_{Z,\mathrm{dot}}} \right\}, \\ \theta &= \min\left\{ 1, \frac{a_0}{12M_F}, \frac{h_{\mathrm{geom}}\Lambda_K}{16C_{\mathrm{refvel}}} \right\}. \end{aligned} \tag{M.13b}$$

where $Q_{\mathrm{post}} = 4c_{\mathrm{post}}$*,* $M_F = \sqrt{18 + b^2}$ *and the finite positive constants* $C_Z, C_{Z,\mathrm{dot}}, C_{\mathrm{refvel}}$ *are the softmax value and velocity coefficients obtained below. In the notation of the preceding appendices,* $M_r = M$*,* $A = A_{\mathrm{eff}}$*,* $B = B_{\mathrm{force}}$ *and* $B_e = B_{\mathrm{eff}}$*; decreasing* e_{eff} *if necessary enforces* $B_e e_{\mathrm{eff}} \leq 1/2$ *and* $e_{\mathrm{eff}} \leq 1/(192B_e)$*.*

Proof. On the fixed mixed-gate branch the closed Gram equation has the form

$$\dot{C} = LC + R(C, W), \quad \|e^{tL}\| \leq e^{\Lambda_K t}, \quad \|R(C, \bar{W}) - R(D, \bar{W})\| \leq A_{\mathrm{eff}}(\|C\| + \|D\|)\|C - D\|,$$

and $\|R(C, W) - R(C, \bar{W})\| \leq B_{\mathrm{force}}\|W - \bar{W}\|\|C\|$. The last estimate uses $R(0, W) = 0$, which is why readout error is multiplied by the Gram norm. Duhamel and $q(t) = \omega^2 e^{\Lambda_K t}$ give, on a standard first-exit interval,

$$\|C(t) - \bar{C}(t)\| \leq B_e e q(t), \quad \|C'(t) - \bar{C}'(t)\| \leq C_{\dot{e}} e q(t), \tag{M.14}$$

where a valid choice is

$$C_{\dot{e}} = \left[\Lambda_K + A_{\mathrm{eff}}(2M_r + 1)Q_{\mathrm{post}}\right]B_{\mathrm{eff}} + B_{\mathrm{force}}(M_r + 1).$$

provided $e \leq e_{\mathrm{eff}}$; the integral is closed because $\int_0^{T_*} q \leq 4c_0/\Lambda_K$. The same argument gives the reference gate slack and classical continuation. This is precisely the relative estimate, with the symbols mapped as stated in the lemma; no transverse decay or symmetry of the actual draw is used.

We verify the endpoint assertion without a Gaussian hypothesis. Decrease c_{eff} so that $c_{\mathrm{eff}} \leq \min\{1, 1/(48C_P)\}$ and use $e_{\mathrm{eff}} \leq 1/(192B_e)$. At T_* compare with

$$P_+^{(0)} = \begin{pmatrix} 0 & c_0 T \\ 0 & 0 \end{pmatrix}, \qquad P_-^{(0)} = -P_+^{(0)}.$$

The reference tube, (M.14), and the aggregate Lipschitz bound give

$$\Delta_Z := \max_{h,X} \|Z_{P_h(T_*)}(X) - Z_{P_h^{(0)}}(X)\| \leq \tfrac{3}{2}C_P c_0^2 + 6B_e e c_0 \leq c_0/16.$$

Every aggregate lies in its token span, so its group sum equals its c_k coefficient; its error is at most Δ_Z, without a factor $\sqrt{K}$. For a conflict let $a = (K-1)c_0/K$, $d = c_0/K$ and $h(s) = e^s/(e^s + 2)$. The leading signed difference of the two group coefficients is

$$h(a) - h(-a) + h(d) - h(-d) \geq (a + d)/4 = c_0/4,$$

because $h(s) - h(-s) = 4\sinh s/(5 + 4\cosh s) \geq s/4$ for $0 \leq s \leq 1$. Both conflict gates are active, so the actual signed projection in (M.13a) is at least $(c_0/4 - 2\Delta_Z)/\sqrt{2K} \geq c_0/(8\sqrt{2K})$. For a consistent type only its correct-label head contributes. Its leading group coefficient is $1/3 + h(s_2) + h(s_3) > 1/2$, where $s_2, s_3 \in [-c_0, c_0]$ and $h(-1) > 1/7$. Its actual signed projection is at least $(1/2 - \Delta_Z)/\sqrt{2K} \geq c_0/(8\sqrt{2K})$. The label cancels the negative-head feature sign for

negative consistent types. The separator has unit norm and zero bias coordinate, proving (M.13a) for the full augmented hull on every admissible deterministic path.

The global softmax derivative bounds

$$\|DZ_P[H]\| \le \tfrac{3}{2}\|H\|, \qquad \|D^2 Z_P[H,H']\| \le 18\|H\|\|H'\|$$

turn (M.14) into, uniformly in X,

$$\|z_X - \bar{z}_X\| \le C_Z e q(t), \quad \|z'_X - \bar{z}'_X\| \le C_{Z,\mathrm{dot}} e q(t), \quad \|\bar{z}'_X\| \le C_{\mathrm{refvel}} q(t). \tag{M.15}$$

Explicit choices, made before c_0 and ω, are

$$C_{\dot r} = \Lambda_K M_r + A_{\mathrm{eff}} M_r^2 Q_{\mathrm{post}}, \qquad C_Z = 3B_e/2,$$
$$C_{\mathrm{refvel}} = 3C_{\dot r}/2, \qquad C_{Z,\mathrm{dot}} = 3C_{\dot e}/2 + 18 B_e C_{\dot r} Q_{\mathrm{post}}.$$

Indeed $\|\bar{C}'\| \le C_{\dot r} q$ follows from the reference field. For the velocity difference split each head derivative into $DZ_{P_h}[P'_h - \bar{P}'_h] + (DZ_{P_h} - DZ_{\bar{P}_h})[\bar{P}'_h]$. The second term is at most $18\|P_h - \bar{P}_h\|\|\bar{P}'_h\|$; the product-norm Cauchy–Schwarz inequality and $q \le Q_{\mathrm{post}}$ give the displayed coefficient. Equal fixed gate bits remove inactive heads and leave the bias-coordinate error and velocity zero. Let $p, \bar{p}$ be the closest points of the actual and reference hulls. By (M.8), $\|\bar{p}\| \ge h_{\mathrm{geom}} \ell(t)$, while the generator error in (M.15) is $\delta_H = C_Z e q(t)$. For $t \ge \tau$, (M.10) and $q/\ell \le 2/(1 - e^{-2\Lambda_K \tau})$ imply

$$r = \frac{\delta_H}{\|\bar{p}\|} \le \frac{2C_Z e}{h_{\mathrm{geom}}(1 - e^{-2\Lambda_K \tau})} \le \frac{2C_Z \varepsilon_{\mathrm{post}}}{h_{\mathrm{geom}}} \le \min\{1/2, \theta^2/10\}. \tag{M.16}$$

Lemma I.2 gives the margin lower bound in (M.11) and $\|p/\|p\| - \bar{p}/\|\bar{p}\|\| \le \theta$.

The reference consistent projection gap is at least $a_0/3$. Hence, using the feature bound M_F and (M.13b), each consistent type satisfies

$$\langle p/\|p\|, z_X\rangle - \|p\| \ge a_0/3 - M_F\theta - 2C_Z e Q_{\mathrm{post}} \ge a_0/6. \tag{M.17}$$

Thus all optimal simplex representations use only conflicts. For every conflict, (M.9), (M.15), and the direction bound yield

$$\langle p/\|p\|, z'_X\rangle \ge \frac{h_{\mathrm{geom}}}{2}\nu(t) - [\theta C_{\mathrm{refvel}} + C_{Z,\mathrm{dot}} e] q(t) \ge \frac{h_{\mathrm{geom}}}{4}\nu(t), \tag{M.18}$$

by (M.13b) and $q/\nu \le 2/\Lambda_K$. Lemma I.1 now gives (M.12) for the right derivative, and the integration lemma below, with $c_{\mathrm{vel}} = h_{\mathrm{geom}}/4$, gives the finite-interval increment. Strict gate slack and the smooth fixed-gate equation permit local C^1 extensions at the finite endpoints, so the sensitivity lemma applies at every required one-sided time. No continuity of optimal simplex weights or of margin derivatives is assumed. □

Lemma I.5 (Positive weighted velocity integrates)**.** *On a compact interval where the margin is positive, if every closest-point simplex has weighted velocity at least $c(t) \ge 0$, then the margin is nondecreasing. If $c(t) = c_{\mathrm{vel}}\Lambda_K \omega^2 \cosh(\Lambda_K t)$ for a constant $c_{\mathrm{vel}} > 0$, then*

$$\gamma(t_2) - \gamma(t_1) \ge c_{\mathrm{vel}}\omega^2[\sinh(\Lambda_K t_2) - \sinh(\Lambda_K t_1)].$$

Proof. By Lemma I.1, the right derivative is the minimum of the weighted velocities. On a compact interval the C^1 generators are Lipschitz, and taking the finite minimum and the maximum over unit vectors preserves the bound $|\gamma(t) - \gamma(s)| \le \max_i \|z_i(t) - z_i(s)\|$. Thus γ is absolutely continuous. Its ordinary derivative agrees almost everywhere with the right derivative and is at least $c(t)$. Integration proves monotonicity and, for the stated continuous $c(t)$, the displayed hyperbolic-sine increment. □

Completion of the delayed-growth assertion. On the common Stage 1 Gaussian event used by Theorem 4.1, the existing concentration estimate gives

$$\frac{\|C(0) - \bar{C}(0)\|}{\omega^2} \le C_G \eta_m, \qquad \|W^1 - \bar{W}\| \le C_w \omega^2 \eta_m, \qquad \omega^2(1 + T_*) \le B_{\mathrm{clock}}.$$

Set $B_{\mathrm{post}} = C_G + B_{\mathrm{clock}} C_w$. Then $e \le B_{\mathrm{post}} \eta_m$. Choose

$$\rho_{\mathrm{post}} \le \min\{1, 2\alpha_0/(3\sqrt{3}), \varepsilon_{\mathrm{post}}/(2B_{\mathrm{post}})\}.$$

With $\eta_m = \sqrt{8\log(L_K/\delta)/m}$, the theorem's width condition ensures $B_{\mathrm{post}}\eta_m/\varepsilon_{\mathrm{post}} \le 1/2$, so the delay in equation (21) is well defined. It satisfies

$$\varepsilon_{\mathrm{post}}(1 - e^{-2\Lambda_K \tau_m}) = B_{\mathrm{post}}\eta_m, \qquad \tau_m \le \frac{B_{\mathrm{post}}\eta_m}{\Lambda_K \varepsilon_{\mathrm{post}}}.$$

The joint choices $c_0 \le 1/4$ and $\omega \le c_0^{3/2}$, together with the reference first-hit lower bound, give

$$T_* \ge \Lambda_K^{-1}\log(c_0/\omega^2) \ge \Lambda_K^{-1}\log(1/c_0^2) \ge \Lambda_K^{-1}\log 16.$$

Thus $0 < \tau_m \le \log 2/(2\Lambda_K) < T_*$, and even the upper bound $B_{\mathrm{post}}\eta_m/(\Lambda_K\varepsilon_{\mathrm{post}}) \le 1/(2\Lambda_K)$ lies below T_*. The additional α_0 condition is the Stage 1 aggregate tolerance: since $\omega \le 1$, it ensures $(3\sqrt{3}/2)\omega^2\eta_m \le \alpha_0$ on the same event. Applying Lemma I.4 with $\tau = \tau_m$ proves Theorem 4.4(ii), with sufficient width uniform in ω, on the same event $\mathcal{G}$.

I.1 An adverse launch and the width obstruction

This subsection proves the negative clause of Theorem 4.4 under Algorithm 1, with Stage 2 initialized at the actual Stage 1 endpoint. All statements below concern the augmented gated feature

$$F_{\mathrm{aug}}(X;t) = \big(\rho_+(X,t)Z_{P_+(t)}(X), -\rho_-(X,t)Z_{P_-(t)}(X), b(\rho_+(X,t) - \rho_-(X,t))\big), \tag{54}$$

where $\rho_h(X,t) = \mathbf{1}\{\langle w_h^1, Z_{P_h(t)}(X)\rangle + b > 0\}$. The margin is the full-support margin

$$\gamma_{\mathrm{aug}}(t) = \max_{\|u\|_2=1} \min_X y_X \langle u, F_{\mathrm{aug}}(X;t)\rangle, \qquad N = 4K(K-1), \tag{55}$$

and the minimum includes both consistent and conflicting types.

Proposition I.6 (Local decrease of the full-support margin)**.** *Fix $K \ge 2$, $b > 0$, and a Stage 1 duration $T_1 > T_0$ as in Theorem 4.1. Fix the allowed c_0 and $\omega > 0$ before choosing an initialization. For every width $m \ge 4K$, there is $\varepsilon_{\mathrm{neg}} > 0$, depending only on these fixed parameters and the constants in the reference and bridge lemmas, with the following property. For every $0 < \varepsilon \le \varepsilon_{\mathrm{neg}}$, there exist real factors with restricted self-Grams $A_h = B_h = \omega^2 I_{2K}$ such that*

$$P_+|_U = \begin{pmatrix} 0 & -\varepsilon\omega^2 T \\ 0 & 0 \end{pmatrix}, \qquad P_-|_U = JP_+|_U J, \qquad T = I_K - \frac{1}{K}\mathbf{1}\mathbf{1}^\top, \tag{56}$$

and the actual Stage 1 endpoint and unchanged factors admit the unique Stage 2 path on $[0, T_*]$ with the mixed gate pattern and the prescribed positive gate slack. At its launch,

$$\gamma_{\mathrm{aug}}(0) = 2h_{\mathrm{geom}}\kappa_{\mathrm{proj}}\,\varepsilon\omega^2 + O(\varepsilon^2\omega^4) > 0, \tag{57}$$

$$\gamma'_{\mathrm{aug}}(0+) = -\frac{5}{2}h_{\mathrm{geom}}\Lambda_K\kappa_{\mathrm{proj}}\,\omega^2 + O(\varepsilon\omega^4) < 0, \tag{58}$$

where

$$h_{\mathrm{geom}} = \frac{\sqrt{2/K}}{9}, \quad e_C = \frac{(c_\Sigma, c_\Sigma, 0)}{\sqrt{2K}}, \quad e_V = \frac{(v_\Sigma, -v_\Sigma, 0)}{\sqrt{2K}}, \quad e_b = (0,0,1), \tag{59}$$

$$b_0 = \Big(\frac{c_\Sigma}{2K} + \frac{v_\Sigma}{K}, \frac{c_\Sigma}{2K} - \frac{v_\Sigma}{K}, b\Big) = a_0 e_C + 2a_0 e_V + b e_b, \quad a_0 = (2K)^{-1/2}, \tag{60}$$

and

$$\Pi_b = I - \frac{b_0 b_0^\top}{\|b_0\|_2^2}, \qquad \kappa_{\mathrm{proj}} = \|\Pi_b e_C\|_2 = \sqrt{\frac{4 + 2Kb^2}{5 + 2Kb^2}} > 0. \tag{61}$$

Here Π_b is the feature-space orthogonal projection perpendicular to b_0, distinct from the signal-index projector Q_0 in the main text. In particular, deleting the last (bias) coordinate from b_0 leaves $a_0 e_C + 2a_0 e_V$, whose e_V component is nonzero and orthogonal to e_C; the projection of the context direction therefore remains nonzero. Here $b > 0$ remains fixed in the model and supplies the initial gate activation and crossing control used in the staged analysis. The $O(\cdot)$ constants are uniform in c_0, ω, d, m after the fixed parameter choices. The positive launch margin is attained by a full-support simplex certificate: the total weight on the consistent orbit is positive and is $O(\varepsilon\omega^2)$. There is a $\tau_{\mathrm{neg}} \in (0, T_*]$ on which γ_{aug} is positive and strictly decreasing.

The proposition gives a deterministic family of initial factors. Below, strict decrease is extended to a joint-Gram ball whose probability is positive at every sufficiently large width under Gaussian initialization.

Realization and the actual Stage 1 interface. Let $s = \varepsilon\omega^2$. Write the first score block in (56) as $P_+ = sM_+$, where $M_+ = \left(\begin{smallmatrix} 0 & -T \\ 0 & 0 \end{smallmatrix}\right)$, and let $M_- = JM_+J$. Since $\|M_h\|_{\mathrm{op}} = 1$ and $0 < \varepsilon < 1$, $I - \varepsilon^2 M_h^\top M_h$ is positive definite. On U, choose the $4K$-row factor realization

$$K_h = \omega \begin{bmatrix} I_{2K} \\ 0 \end{bmatrix}, \qquad Q_h = \omega \begin{bmatrix} \varepsilon M_h \\ (I_{2K} - \varepsilon^2 M_h^\top M_h)^{1/2} \end{bmatrix}. \tag{62}$$

Append zero rows when $m > 4K$, and extend both factors by zero on $U^\perp$. Then $K_h^\top K_h = Q_h^\top Q_h = \omega^2 I_U$ and $K_h^\top Q_h = P_h$. Thus the score perturbation is realizable by the factorized model with the displayed self- and cross-Gram blocks.

At zero scores the aggregate is $Z_0(X) = x_1 + x_2 + x_3$ (the three uniform softmax weights sum to one), with the corresponding notation of the main text. The global softmax bound gives

$$\max_{h,X} \|Z_{P_h}(X) - Z_0(X)\|_2 \le \frac{3}{2}\|P_h\|_{\mathrm{op}} = \frac{3}{2}s. \tag{63}$$

The Stage 1 directed-crossing lemma therefore applies for sufficiently small s. It gives the unique strict readout trajectory from the actual zero-readout launch, including its finite crossing history, and

$$\|W^1 - \overline{W}\|_2 \le C_{\mathrm{sh}} s, \qquad C_{\mathrm{sh}} = \tfrac{3}{2} C_{\mathrm{shadow}}, \qquad \overline{W} = (\mu c_\Sigma + \beta v_\Sigma, \mu c_\Sigma - \beta v_\Sigma). \tag{64}$$

The factors in (62) are unchanged after Stage 1. The comparison quantity

$$e_{\mathrm{joint}} = \frac{\|C(0) - \overline{C}(0)\|_F}{\omega^2} + (1 + T_*)\|W^1 - \overline{W}\|_2 \tag{65}$$

is $O(\varepsilon)$: the first term is $2\sqrt{K-1}\,\varepsilon$ and the second is bounded by the clock constant times $C_{\mathrm{sh}}\varepsilon$. Decrease $\varepsilon_{\mathrm{neg}}$ so that e_{joint} is within the hypothesis of Lemma I.4. That lemma gives the unique actual factorized Stage 2 continuation from W^1 on $[0, T_*]$, all full-support mixed gates, and the same fixed positive slack as in the reference.

For this explicit family the frozen scores commute with simultaneous index permutations and obey $P_- = JP_+J$. The full-population Stage 1 equation is equivariant under those permutations and under $(w_+, w_-) \mapsto (Jw_-, Jw_+)$, accompanied by $(y, \epsilon) \mapsto (-y, -\epsilon)$. Its unique strict trajectory starts at zero, so the actual endpoint readouts are permutation-invariant and $w_-^1 = Jw_+^1$. The conditional Stage 2 population has the same symmetries. With these frozen readouts the smooth joint-Gram equation $\dot{C}_h = -S_h C_h - C_h S_h$ is equivariant and its initial Grams are fixed by the transformations. Uniqueness therefore preserves the score symmetries throughout $[0, T_*]$. This uses no rowwise symmetry of the chosen factor realization. The stability interface itself applies also without these symmetries.

Full-hull geometry at the launch. For a type X, put $z_X = y_X F_{\mathrm{aug}}(X; 0)$. Simultaneously permuting the K group indices and applying $(u_+, u_-, u_b) \mapsto (Ju_-, Ju_+, u_b)$ leaves the labeled feature set invariant. This finite group has two orbits: all conflicting types and all consistent types. Its fixed subspace within $U \oplus U \oplus \mathbb{R}$ is

$$E = \{(Ac_\Sigma + Bv_\Sigma, Ac_\Sigma - Bv_\Sigma, c) : A, B, c \in \mathbb{R}\}. \tag{66}$$

Let v_{conf} and v_{cons} be the uniform means of z_X over the two orbits. Averaging over the group is the orthogonal projection onto E, and maps the full convex hull onto the segment $[v_{\mathrm{conf}}, v_{\mathrm{cons}}]$. The squared norm has a unique minimizer on the compact full hull; group invariance puts it in E, and the segment is itself contained in the hull. Hence the full-hull closest point is exactly the closest point on this segment, which incorporates both orbits and all $4K(K-1)$ support types.

For the following zero-score calculation, regard these orbit means as functions of the score pair $\mathbf{P} = (P_+, P_-)$ with the mixed gates fixed. At $\mathbf{P} = 0$,

$$v_{\mathrm{conf}}\big|_{\mathbf{P}=0} = 0, \qquad v_{\mathrm{cons}}\big|_{\mathbf{P}=0} = b_0 = \left(\frac{c_\Sigma}{2K} + \frac{v_\Sigma}{K}, \frac{c_\Sigma}{2K} - \frac{v_\Sigma}{K}, b\right). \tag{67}$$

The first-order orbit calculation used below differentiates with respect to the score pair, not time. For $H = \left(\begin{smallmatrix} 0 & z^T \\ r^T & 0 \end{smallmatrix}\right)$,

$$D_{\mathbf{P}} v_{\mathrm{conf}}|_{\mathbf{P}=0}\,[H, -H] = h_{\mathrm{geom}}(2z - r)e_C. \tag{68}$$

For $\Delta = v_k - v_j$ and $u = c_k + y\Delta$, the zero-score three-token derivative is

$$DZ_0(X)[H] = \frac{1}{3}XX^\top Hu - \frac{1}{9}u(u^\top Hu).$$

Using $c_k^\top Hu = yz$ in the ordered-pair coordinates and $u^\top Hu = y(z+r)$ (equivalently applying the block convention in (68)) gives the coefficient $(2z-r)/9$ after the group average. The classification component is independent of y and vanishes after multiplying by the label and averaging.

At the actual time-zero launch, $z = -s$ and $r = 0$ in (56). Taylor expansion of the finite analytic softmax map therefore gives

$$v_{\mathrm{conf}}(0) = -2h_{\mathrm{geom}}se_C + O(s^2), \qquad v_{\mathrm{cons}}(0) = b_0 + O(s). \tag{69}$$

Closest point and the complete KKT certificate. Put $d = v_{\mathrm{cons}}(0) - v_{\mathrm{conf}}(0)$. The coefficient of the closest point on the affine line is

$$\theta = -\frac{\langle v_{\mathrm{conf}}(0), d\rangle}{\|d\|_2^2} = \frac{2h_{\mathrm{geom}}a_0}{\|b_0\|_2^2}s + O(s^2). \tag{70}$$

Since $h_{\mathrm{geom}}, a_0 > 0$, $\theta \in (0,1)$ for sufficiently small s. The segment closest point is consequently

$$p = v_{\mathrm{conf}}(0) + \theta d = -2h_{\mathrm{geom}}s\Pi_b e_C + O(s^2), \qquad \|p\|_2 = 2h_{\mathrm{geom}}\kappa_{\mathrm{proj}}s + O(s^2). \tag{71}$$

The closest-point characterization (Lemma I.1) gives $\gamma_{\mathrm{aug}}(0) = \|p\|_2$, with maximizing unit vector $u = p/\|p\|_2 = -\Pi_b e_C/\kappa_{\mathrm{proj}} + O(s)$. Moreover $\langle p, d\rangle = 0$ exactly. The full-support simplex certificate assigns total weight θ uniformly to the consistent orbit and total weight $1-\theta$ uniformly to the conflicting orbit. Every type has signed projection $\|p\|_2^2$ against p, because the group orbit mean is its projection and both orbit means have the same pairing with p. The consistent orbit contributes its bias coordinate with positive mass $\theta = O(s)$, and the conflicting orbit has mass $1-\theta$.

Launch velocity and strict decrease. Let $N_h = \nabla_{P_h}\mathcal{R}_{\mathrm{conf}}$. The factorized Gram equation is

$$\dot{P}_h = -N_hB_h - A_hN_h. \tag{72}$$

At the isotropic zero-score reference with readouts $\overline{W}$,

$$N_+(0,\overline{W}) = \begin{pmatrix} 0 & -\alpha_K T \\ \alpha_K T/2 & 0 \end{pmatrix}, \quad N_-(0,\overline{W}) = -N_+(0,\overline{W}), \quad \alpha_K = \frac{\Lambda_K}{2}. \tag{73}$$

The actual Stage 1 endpoint differs by $O(s)$ and all gates are strict, so smoothness of the conditional score gradient and (72) imply

$$\dot{P}_+(0) = \begin{pmatrix} 0 & \Lambda_K\omega^2 T \\ -\Lambda_K\omega^2 T/2 & 0 \end{pmatrix} + O(\omega^2 s), \qquad \dot{P}_-(0) = -\dot{P}_+(0) + O(\omega^2 s). \tag{74}$$

Applying (68) to this velocity gives

$$\dot{v}_{\mathrm{conf}}(0) = \frac{5}{2}h_{\mathrm{geom}}\Lambda_K\omega^2 e_C + O(\omega^2 s), \qquad \|\dot{v}_{\mathrm{cons}}(0)\|_2 = O(\omega^2). \tag{75}$$

The remainders account for both the score Taylor expansion and the actual Stage 1 endpoint error.

On a sufficiently short right interval the segment coefficient remains interior and the closest point remains nonzero. Differentiating its norm, and using $\langle p, d\rangle = 0$, yields

$$\gamma'_{\mathrm{aug}}(0+) = \langle u, (1-\theta)\dot{v}_{\mathrm{conf}}(0) + \theta\dot{v}_{\mathrm{cons}}(0)\rangle, \qquad u = \frac{p}{\|p\|_2}. \tag{76}$$

Since $u = -\Pi_b e_C/\kappa_{\mathrm{proj}} + O(s)$ and $\langle \Pi_b e_C, e_C\rangle = \kappa_{\mathrm{proj}}^2$, substitution gives

$$\gamma'_{\mathrm{aug}}(0+) = -\frac{5}{2}h_{\mathrm{geom}}\Lambda_K\kappa_{\mathrm{proj}}\omega^2 + O(\omega^2 s),$$

which is (58) because $s = \varepsilon\omega^2$. Shrinking $\varepsilon_{\mathrm{neg}}$ makes this strictly negative. Continuity of the smooth fixed-gate trajectory and the finite-hull closest-point map then gives a $\tau_{\mathrm{neg}} > 0$ such that the margin is positive and strictly decreases throughout $[0, \tau_{\mathrm{neg}}]$.

An adverse ball for nonsymmetric initializations. We extend the strict-decrease construction to a ball in the full joint-Gram space, which will determine its probability under iid Gaussian factors. Define, in the ordered (C, V) basis,

$$D_{\mathrm{P}} = \begin{pmatrix} 0 & T \\ -T/2 & 0 \end{pmatrix}, \qquad H_{\mathrm{adv},+} = \begin{pmatrix} 0 & -D_{\mathrm{P}} \\ -D_{\mathrm{P}}^{\top} & 0 \end{pmatrix}, \qquad H_{\mathrm{adv},-} = -H_{\mathrm{adv},+}, \tag{77}$$

and $H_{\mathrm{adv}} = (H_{\mathrm{adv},+}, H_{\mathrm{adv},-})$. Define

$$\Omega = \omega^2, \quad d_F = \frac{5h_{\mathrm{geom}}}{2} e_C, \quad d_{\perp} = \Pi_b d_F,$$
$$\chi_{\mathrm{adv}} = \|d_{\perp}\| = \frac{5h_{\mathrm{geom}}}{2} \kappa_{\mathrm{proj}} > 0, \quad a_{\mathrm{vel}} = \Lambda_K \Omega \chi_{\mathrm{adv}}.$$

Lemma I.7 (A full joint-Gram adverse ball)**.** *Fix the allowed (K, b, T_1, c_0, ω). There are $0 < r_{\mathrm{adv}} \le 1$ and $\lambda_{\mathrm{adv}} > 0$, independent of d, m, the embedding, and the factor realization, such that every actual initial joint-Gram pair obeying*

$$\left\| \lambda^{-1} \left[\frac{C(0)}{\omega^2} - (I_{4K}, I_{4K}) \right] - H_{\mathrm{adv}} \right\|_F \le r_{\mathrm{adv}}, \qquad 0 < \lambda \le \lambda_{\mathrm{adv}}, \tag{78}$$

has a unique actual Stage 1 trajectory through T_1 and the unique classical Stage 2 continuation on $[0, T_*]$ supplied by Lemma I.4, including its mixed gates and endpoint separator. At the Stage 2 launch,

$$\gamma_{\mathrm{aug}}(0) \ge \frac{\chi_{\mathrm{adv}}}{4} \Omega \lambda > 0, \qquad \gamma'_{\mathrm{aug}}(0+) \le -\frac{a_{\mathrm{vel}}}{2} < 0.$$

For some sample-dependent $\tau \in (0, T_*]$, the margin stays positive on $[0, \tau]$ and, for every $0 \le t_1 \le t_2 \le \tau$,

$$\gamma_{\mathrm{aug}}(t_2) - \gamma_{\mathrm{aug}}(t_1) \le -\frac{a_{\mathrm{vel}}}{4}(t_2 - t_1).$$

No equality of actual self-Grams, head conjugacy, or permutation symmetry is assumed.

Proof. All constants in this proof are selected for the fixed positive Ω. Initially take any candidate radius $r \le 1$ and put $s = \Omega\lambda$ and $H_{\mathrm{norm}} = \|H_{\mathrm{adv}}\|_F = \sqrt{5(K-1)}$. Equation (78) gives

$$\|C(0) - \Omega(I, I)\|_F \le s(H_{\mathrm{norm}} + 1), \qquad \max_{h,X} \|Z_{P_h(0)}(X) - Z_0(X)\| \le \tfrac{3}{2} s(H_{\mathrm{norm}} + 1).$$

For sufficiently small λ, the deterministic Stage 1 stability lemma therefore gives its unique strict zero-readout trajectory and

$$d_w := \|W^1 - \bar{W}\| \le C_{w,\mathrm{adv}} s, \qquad C_{w,\mathrm{adv}} = \tfrac{3}{2} C_{\mathrm{shadow}}(H_{\mathrm{norm}} + 1).$$

The factors are unchanged at the interface, and hence

$$e \le \lambda \big[H_{\mathrm{norm}} + 1 + (1 + T_*) C_{w,\mathrm{adv}} \Omega \big].$$

Choose an upper scale so that this is at most e_{eff} and the Stage 1 aggregate tolerance holds. Lemma I.4 applies to every such initial state, uniformly for $r \le 1$, and supplies the full Stage 2 interval, gate slack, and deterministic endpoint separator.

For each sufficiently small λ, introduce the exact comparison Gram pair

$$C_{\lambda}^{a} = \Omega(I, I) + s H_{\mathrm{adv}}.$$

Its self-Grams are ΩI and its scores are $(-sD_{\mathrm{P}}, +sD_{\mathrm{P}})$. Since $\|D_{\mathrm{P}}\|_{\mathrm{op}} = 1$ and $\lambda < 1$, the realization in (62), with $M_+ = -D_{\mathrm{P}}$ and $\varepsilon = \lambda$, gives real factors. The preceding interface applies to their actual Stage 1 history as well. This comparison center has a reverse block $sT/2$; it is distinct from the CV-only family in Proposition I.6.

Let q_λ be the closest point of its full launch feature hull and $u_\lambda = q_\lambda / \|q_\lambda\|$. The same symmetric full-hull calculation now uses $z = -s$, $r = s/2$ in (68). Thus its conflict and consistent means satisfy $v_{\mathrm{conf}}^{a} = -sd_F + O(s^2)$ and $v_{\mathrm{cons}}^{a} = b_0 + O(s)$. Since $\langle d_F, b_0 \rangle > 0$, the segment coefficient is interior and equals $s\langle d_F, b_0 \rangle / \|b_0\|^2 + O(s^2)$. Consequently

$$q_\lambda = -sd_{\perp} + O(s^2), \qquad \frac{\chi_{\mathrm{adv}}}{2} s \le \|q_\lambda\| \le 2\chi_{\mathrm{adv}} s, \qquad u_\lambda \longrightarrow u_{\mathrm{adv}} := -d_{\perp}/\chi_{\mathrm{adv}}.$$

These bounds hold after a fixed reduction of the upper scale.

Let p be the closest point of an arbitrary actual launch hull in the prospective ball. Its score-pair difference from C^a_λ is at most rs. Both hulls have the same fixed mixed gates, including identical bias coordinates, so their generator differences and Hausdorff distance are at most Lrs, where $L = 3/2$. Lemma I.2 gives

$$\big|\|p\| - \|q_\lambda\|\big| \le Lrs, \qquad \|p - q_\lambda\|^2 \le (\|p\| + \|q_\lambda\|)Lrs. \tag{79}$$

Require $Lr \le \chi_{\mathrm{adv}}/4$. Then

$$\frac{\chi_{\mathrm{adv}}}{4}s \le \|p\| \le 3\chi_{\mathrm{adv}}s, \qquad \|p - q_\lambda\| \le \sqrt{5\chi_{\mathrm{adv}}Lr}\,s, \qquad \left\|\frac{p}{\|p\|} - u_\lambda\right\| \le 4\sqrt{\frac{5Lr}{\chi_{\mathrm{adv}}}}.$$

The last bound follows by normalization using $2\|p - q_\lambda\|/\|q_\lambda\|$. In particular the actual margin is positive and equals $\|p\|$, and its maximizing unit vector is $u = p/\|p\|$. Retaining the small closest-point norms in (79) makes this directional bound independent of s.

Let V^0_X denote the labeled feature velocity at zero scores, self-Grams ΩI, and readouts $\bar{W}$, using the conditional factor equation. Put $V_{\max} = \max_X \|V^0_X\|$ and $M_V = V_{\max} + 1$. The time-independent index and head-label symmetries permute these velocities on conflicts and fix u_{adv}. By (73) and (68), their conflict mean is $\Lambda_K \Omega d_F$. Therefore every conflict satisfies

$$\langle u_{\mathrm{adv}}, V^0_X\rangle = -\Lambda_K \Omega \chi_{\mathrm{adv}} = -a_{\mathrm{vel}}.$$

Write $V_X = (y_X F_{\mathrm{aug}}(X;t))'|_{t=0}$ for actual launch velocities. Uniformly over the prospective ball, scores, self-Grams, and actual Stage 1 readouts differ from this limiting state by $O(s)$. On the strict mixed cell, $N_h(P, W^1)$ and DZ_{P_h} are smooth and $P'_h = -N_h B_h - A_h N_h$. Their derivatives are bounded on a fixed neighborhood, so

$$\max_X \|V_X - V^0_X\| \le Cs \longrightarrow 0.$$

This estimate includes the actual Stage 1 readout error.

Choose $\zeta = a_{\mathrm{vel}}/[8(V_{\max} + 1)]$. First fix $r > 0$ sufficiently small that $Lr \le \chi_{\mathrm{adv}}/4$ and $4\sqrt{5Lr/\chi_{\mathrm{adv}}} \le \zeta/2$; call this radius r_{adv}. Next reduce the upper scale so that $\|u_\lambda - u_{\mathrm{adv}}\| \le \zeta/2$ and $\max_X \|V_X - V^0_X\| \le \min\{1, a_{\mathrm{vel}}/8\}$. It follows that $\|V_X\| \le M_V$ and, for every conflict,

$$\langle u, V_X\rangle \le -a_{\mathrm{vel}} + \zeta V_{\max} + a_{\mathrm{vel}}/8 \le -3a_{\mathrm{vel}}/4.$$

For any simplex representation $p = \sum_X \alpha_X y_X F_{\mathrm{aug}}(X;0)$, the last feature coordinate is zero on conflicts and exactly b on consistent types. Thus its total consistent weight is

$$a_{\mathrm{cons}} := \sum_{X \text{ consistent}} \alpha_X = p_b/b \le \|p\|/b \le 3\chi_{\mathrm{adv}}s/b.$$

Reduce the upper scale so that this is at most $a_{\mathrm{vel}}/[4(M_V + a_{\mathrm{vel}})]$. Every closest-point representation then has weighted velocity

$$\sum_X \alpha_X \langle u, V_X\rangle \le -\tfrac{3}{4}a_{\mathrm{vel}}(1 - a_{\mathrm{cons}}) + M_V a_{\mathrm{cons}} \le -a_{\mathrm{vel}}/2.$$

The margin sensitivity lemma gives the right derivative bound. Its open-time hypothesis holds by a short backward extension of the smooth Stage 2 equation from the strict launch state, used only to evaluate this local derivative. The actual Stage 1 history remains unchanged.

For each such actual state, the feature values and velocities are continuous for small positive time. The hull estimate makes $p(t)$ and its unit direction continuous while the margin remains positive. At each nearby time, every closest-point representation still obeys $0 \le a_{\mathrm{cons}}(t) \le \min\{1, \|p(t)\|/b\}$. Its weighted velocity is bounded above by

$$\begin{aligned} &A(t)(1 - a_{\mathrm{cons}}(t)) + B(t)a_{\mathrm{cons}}(t),\\ &A(t) = \max_{X \text{ conflict}} \langle u(t), z'_X(t)\rangle,\\ &B(t) = \max_{X \text{ consistent}} \langle u(t), z'_X(t)\rangle. \end{aligned}$$

The maximum of this affine expression over the displayed interval of possible a_{cons} is continuous and at launch is at most $-a_{\mathrm{vel}}/2$. Shrinking to some $[0,\tau] \subset [0,T_*]$ keeps it at most $-a_{\mathrm{vel}}/4$ and keeps the margin positive. The finite-feature margin is locally Lipschitz, since $|\gamma(t)-\gamma(t')| \le \max_X \|z_X(t)-z_X(t')\|$. It is therefore absolutely continuous. Its ordinary derivative agrees almost everywhere with the right derivative, which the sensitivity lemma bounds above by $-a_{\mathrm{vel}}/4$. Integration proves the asserted bound for every pair of times. This step requires no continuous choice of individual simplex weights.

The finitely many scale restrictions just imposed define a positive λ_{adv} after r_{adv} is fixed. All calculations take place in the fixed signal coordinates and are uniform in d, m and the realization. The resulting τ need not be uniform over samples. □

Gaussian joint-Gram fluctuations. We now match the adverse ball to the actual iid factor initialization. Put $q = 4K$. For each head h and row r, let $g_{h,r} \sim N(0, I_q)$ be independent and define

$$S_{h,m} = \frac{1}{m}\sum_{r=1}^{m} g_{h,r} g_{h,r}^{\top}, \qquad G_m = \left(\sqrt{m}(S_{+,m} - I_q), \sqrt{m}(S_{-,m} - I_q)\right).$$

We derive the required fixed-dimensional Gaussian limit from the scalar central limit theorem and the Cramér–Wold criterion (Durrett, 2019, Theorems 3.4.1 and 3.10.6). Let Y_r collect, for both heads, the coordinates

$$(g_{h,r,i}^2 - 1)/\sqrt{2} \quad (i = 1,\dots,q), \qquad g_{h,r,i} g_{h,r,j} \quad (1 \le i < j \le q).$$

There are $q(q+1)$ coordinates. Independence, symmetry, and $\mathbb{E}g^2 = 1$, $\mathbb{E}g^4 = 3$ give $\mathbb{E}Y_r = 0$ and $\mathrm{Cov}(Y_r) = I$. For $X_{m,r} = Y_r/\sqrt{m}$, the covariance sum is I and

$$\sum_{r=1}^{m} \mathbb{E}\left[\|X_{m,r}\|_2^2 \mathbf{1}\{\|X_{m,r}\|_2 \ge \delta\}\right] = \mathbb{E}\left[\|Y_1\|_2^2 \mathbf{1}\{\|Y_1\|_2 \ge \delta\sqrt{m}\}\right] \longrightarrow 0$$

for every $\delta > 0$. Thus the rowwise independence and Lindeberg condition hold in fixed dimension. For every fixed vector $v \in \mathbb{R}^{q(q+1)}$, the scalar random variables $\langle v, Y_r\rangle$ are iid, centered, and have variance $\|v\|^2$. For $v \ne 0$, the scalar central limit theorem gives $m^{-1/2}\sum_r \langle v, Y_r\rangle \Rightarrow N(0, \|v\|^2)$. For $v = 0$ the assertion is immediate. The Cramér–Wold criterion therefore yields joint convergence of $m^{-1/2}\sum_r Y_r$ to $N(0, I)$. The linear map that multiplies diagonal coordinates by $\sqrt{2}$ and reflects upper triangles into the lower triangles gives

$$G_m \Rightarrow G, \qquad G_{h,ii} \sim N(0,2), \quad G_{h,ij} \sim N(0,1)\ (i<j), \tag{80}$$

with all displayed upper-triangular coordinates independent across heads.

For each head, concatenate the restrictions of its key and query row to the $2K$-dimensional orthonormal signal span. The resulting $4K$-vector has exactly the law $(\omega/\sqrt{m})g_{h,r}$: orthonormal restriction preserves the Gaussian covariance and the original key and query factors are independent. Consequently, for the actual restricted joint Grams,

$$\sqrt{m}\left[\frac{C(0)}{\omega^2} - (I_{4K}, I_{4K})\right] \overset{\mathrm{law}}{=} G_m \tag{81}$$

for every finite m, every $d \ge 2K$, and every orthonormal embedding of U. Thus G_m is the exact finite-width joint-Gram fluctuation, and (80) describes its limiting law.

Let

$$E_m = \left\{ \left\| \sqrt{m}\left[\frac{C(0)}{\omega^2} - (I, I)\right] - H_{\mathrm{adv}} \right\|_F < r_{\mathrm{adv}} \right\}. \tag{82}$$

By (80), the limiting pair G has a strictly positive continuous density on the product of symmetric matrix spaces, so

$$\Pr(E_m) \longrightarrow p_0 := \Pr(\|G - H_{\mathrm{adv}}\|_F < r_{\mathrm{adv}}) > 0. \tag{83}$$

The ball boundary has zero Gaussian measure, so the continuity-set form of weak convergence (Durrett, 2019, Theorem 3.10.1(v)) gives convergence of these probabilities. Choose $p_{\mathrm{adv}} = p_0/2$ and an integer $m_{\mathrm{adv}} \ge 4K$ so large that, for all $m \ge m_{\mathrm{adv}}$, $m^{-1/2} \le \lambda_{\mathrm{adv}}$ and $\Pr(E_m) \ge p_{\mathrm{adv}}$. On E_m, apply Lemma I.7 with $\lambda = m^{-1/2}$. Hence for every $m \ge m_{\mathrm{adv}}$ there is a fixed positive probability p_{adv} of a positive but initially decreasing full-support margin, while the same trajectory has the positive Stage 2 endpoint separator supplied by Lemma I.4. The threshold m_{adv} is obtained existentially from this weak convergence and the scale condition $m^{-1/2} \le \lambda_{\mathrm{adv}}$.

Duration of the initial layer. On E_m, the quantitative decreasing interval of Lemma I.7 obeys

$$0 < \gamma_{\mathrm{aug}}(\tau) \le 3\chi_{\mathrm{adv}}\omega^2 m^{-1/2} - \tfrac{1}{4}a_{\mathrm{vel}}\tau, \qquad \tau < \frac{12}{\Lambda_K\sqrt{m}}.$$

This bounds the interval with the stated uniform negative speed. To control all later decreases, put $C_{\mathrm{adv}} = \|H_{\mathrm{adv}}\|_F + 1 + (1+T_*)C_{w,\mathrm{adv}}\omega^2$. The proof of the same lemma gives $e \le C_{\mathrm{adv}}/\sqrt{m}$. For sufficiently large m, the deterministic bridge therefore guarantees strict growth on $[\widehat{\tau}_m, T_*]$, where

$$\widehat{\tau}_m = -\frac{1}{2\Lambda_K}\log\!\left(1 - \frac{C_{\mathrm{adv}}}{\varepsilon_{\mathrm{post}}\sqrt{m}}\right) = O_{K,b,T_1,c_0,\omega}(m^{-1/2}) < T_*.$$

The positive probability bound therefore coexists with an initial layer whose duration vanishes as the width grows.

Probability bound and eventwise conclusions. Let S_m denote the event that the actual staged path exists through T_* and that γ_{aug} is nondecreasing on the whole Stage 2 interval. The events E_m are disjoint from S_m, so for every $m \ge m_{\mathrm{adv}}$,

$$\Pr(S_m) \le 1 - p_{\mathrm{adv}}.$$

Therefore no function $M(\delta)$ can make $\Pr(S_m) \ge 1-\delta$ for every $m \ge M(\delta)$ and every $0 < \delta < p_{\mathrm{adv}}$, with the fixed (K, b, T_1, c_0, ω) above. The constants p_{adv} and m_{adv} are selected for these fixed parameters; this conclusion is not uniform in ω, and it gives no width-independent lower bound on the duration of the decreasing interval. The adverse ball also lies in the learning event $\mathcal{G} = \mathcal{G}_{m,\delta}$ of Appendix B. Indeed, (77) gives $\|H_{\mathrm{adv}}\|_{\max} = 1 - 1/K < 1$, and $r_{\mathrm{adv}} \le 1$. Every entry of both joint Grams on E_m consequently satisfies

$$\max_h \|C_h(0) - \omega^2 I_{4K}\|_{\max} < \frac{2\omega^2}{\sqrt{m}} < \omega^2\eta_m, \qquad \eta_m\sqrt{m} = \sqrt{8\log(L_K/\delta)} > 2.$$

These are all the self- and cross-Gram conditions defining $\mathcal{G}$. Hence $E_m \subseteq \mathcal{G}$ for every $0 < \delta < 1$. Under the common width requirement

$$m \ge \max\left\{m_{\mathrm{adv}}, 4K, \frac{8}{\rho_K^2}\log\frac{L_K}{\delta}\right\},$$

every trajectory in E_m has the permanent activation and risk convergence of Theorem 4.1, as well as the initial margin decrease.