MAXIMUM-MARGIN CLASSIFICATION WITH JOINTLY TRAINED SOFTMAX ATTENTION

Anonymous authors
Paper under double-blind review

ABSTRACT

We study the implicit bias of joint training in a single-layer, single-head softmax attention model for binary classification, with trainable query, key, value, output, and classifier factors. For exponential and logistic losses, we show that, once the training loss falls below a specified threshold, gradient flow and sufficiently small fixed-step gradient descent drive the loss to zero and monotonically increase a smoothed normalized margin. We further prove that the effective classifier converges in direction. More precisely, after normalization to unit margin, it converges to the unique Euclidean minimum-norm separator of all limiting signed features generated by the training trajectory. Under additional geometric conditions on the data and initialization, we prove for both losses that attention selects prescribed tokens, identifying the limiting classifier as the hard-margin SVM on those tokens. These results establish a maximum-margin characterization for joint attention and classifier training even though the model is not globally homogeneous. Experiments on MNIST illustrate normalized-margin growth and attention concentration, while a controlled example demonstrates token selection and classifier alignment with the predicted SVM.

1 INTRODUCTION

Understanding which solutions are selected by gradient-based training is a central question in the theory of neural networks. In classification, fitting the training labels does not by itself determine the direction of the learned predictor. A line of work has studied this implicit selection through margin maximization: gradient descent on separable linear logistic regression converges in direction to the Euclidean hard-margin separator (Soudry et al., 2018), while homogeneous neural networks admit normalized-margin and max-margin KKT characterizations (Lyu & Li, 2020; Ji & Telgarsky, 2020).

This question is particularly subtle for softmax attention, a central component of transformers (Vaswani et al., 2017). The query and key determine how tokens contribute to the representation, while the output head determines their value for classification. Existing analyses connect attention training to token-selection margins with a fixed prediction head (Tarzanagh et al., 2023; Sheen et al., 2024); joint training has also been studied under structured data and specialized updates (Vasudeva et al., 2025). When all factors are trained by ordinary gradient flow or fixed-step gradient descent, the representation and classifier evolve together. We study the classifier direction selected by this coupled training and the conditions under which its limiting features can be identified as selected tokens.

We consider a single-layer, single-head classifier

$$p_i = \mathrm{softmax}\left(\frac{X_i K Q^\top q_i}{\sqrt{d}}\right), \qquad h_i = X_i^\top p_i, \qquad f_i = h_i^\top V W a.$$

The effective classifier $v = VWa$ acts on the learned features h_i; Section 3 specifies the factorized architecture. Unlike a globally homogeneous network, rescaling the query and key changes the relative token weights. The output head remains cubic: $f_i(A, tH) = t^3 f_i(A, H)$ for $A = (Q, K)$ and $H = (V, W, a)$. This partial homogeneity determines the head normalization.

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.

1.1 Main Results

We study gradient flow and gradient descent on the model above with exponential or logistic loss. Following Lyu & Li (2020), we assume that the training loss satisfies $n\mathcal{L}(\Theta(t_0)) < \ell(0)$ at some finite time t_0. This condition guarantees a positive margin on every training example. Our results describe the subsequent evolution of the classifier and the learned features. Part I establishes convergence of the classifier direction and characterizes its limit through a maximum-margin problem. Part II gives additional conditions under which attention selects prescribed tokens, thereby identifying the features in that problem.

Our Contribution. The output head is homogeneous of degree three, so its natural margin scale is $r^{3/2}$, where $r = (\|V\|_F^2 + \|W\|_F^2 + \|a\|^2)/3$. We prove that a smooth margin normalized by this scale increases under gradient flow and sufficiently small fixed-step gradient descent. The loss tends to zero, the head factors asymptotically balance (Theorems 4.1–4.2), and the same analysis bounds loss decay and weight growth (Theorem 4.3).

Margin growth alone does not imply convergence of the head direction. Exact softmax identities control the attention scale, and a desingularizing gradient inequality gives finite length of the normalized head trajectory. Consequently, the effective classifier, rescaled to unit margin, converges to the unique minimum-norm separator of all limiting signed aggregate features (Theorem 4.4). This is a convex SVM problem, although its constraints depend on the training trajectory and the features need not have a single limit.

Under additional geometric and initialization conditions, including a finite covariance certificate, attention selects prescribed tokens (Theorems 5.1 and 5.3). The SVM from Part I then becomes the hard-margin SVM on those tokens.

Experiments. On binary MNIST, the smooth normalized margin grows under five-factor training, while different query initialization scales produce distinct levels of attention concentration. A complementary three-token example satisfies the finite covariance certificate and shows increasing selection gaps and classifier rotation toward a predetermined SVM solution (Section 6).

2 Related Work

Maximum-margin dynamics. For linear logistic regression, gradient descent selects the Euclidean hard-margin direction (Soudry et al., 2018). In homogeneous networks, Lyu & Li (2020) prove eventual monotonicity of a smoothed normalized margin and show that limit directions are aligned with KKT points of a parameter-space margin problem. Layer balancing (Du et al., 2018) and alignment in deep linear networks (Ji & Telgarsky, 2019) reveal further structure, while Ji & Telgarsky (2020) establish directional convergence for broad homogeneous models using definable gradient inequalities. Loss tails and optimization geometry further affect the selected direction (Nacson et al., 2019; Gunasekar et al., 2018a). The implicit regularizer also depends on parameterization, as linear convolutional networks show (Gunasekar et al., 2018b). In our model, the readout is homogeneous for fixed attention, but the features move during training.

Nonhomogeneous models and learned representations. Cai et al. (2025) prove directional convergence for near-homogeneous networks under global controls on the Euler error and gradients. Diverging query–key scores need not satisfy those controls for exact softmax attention. Other studies describe how attention and embeddings encode topic structure (Li et al., 2023) and how linearized training can progress from condensation to rank collapse (Chen & Luo, 2025). We analyze the margin dynamics of the classifier along the coupled attention trajectory.

Attention margins and token selection. Tarzanagh et al. (2023) relate gradient-descent token selection to an attention SVM with a fixed head; their joint head–attention result concerns ridge-constrained minimizers under an optimal-token margin condition. Sheen et al. (2024) study separate key and query factors under gradient flow with a fixed head, obtaining a nuclear-norm characterization under structured geometry and an assumed margin-normalized limit. For joint training, Vasudeva et al. (2025) use a direct attention matrix and linear head with normalized updates in a structured two-token regime. Julistiono et al. (2026) extend token selection to mirror-descent geometry and also examine a jointly optimized direct attention matrix and decoder. We instead train all

five factors with ordinary gradient flow or fixed-step gradient descent, first characterizing the classifier on trajectory-generated features and then identifying these features under finite cooperative conditions.

3 PRELIMINARIES

3.1 DATA AND JOINTLY TRAINED SOFTMAX ATTENTION

We consider a binary training set $\{(X_i, y_i)\}_{i=1}^n$, where $y_i \in \{-1, +1\}$ and

$$X_i = (x_{i1}, \dots, x_{is})^\top \in \mathbb{R}^{s\times d}$$

is a sequence of s token vectors in $\mathbb{R}^d$. The query vector for X_i is denoted by $q_i \in \mathbb{R}^d$; it may, for example, be the last token x_{is}. The single-layer, single-head Transformer classifier has separate trainable projections with dimensions

$$Q, K \in \mathbb{R}^{d\times d}, \qquad V \in \mathbb{R}^{d\times d_1}, \qquad W \in \mathbb{R}^{d_1\times d_2}, \qquad a \in \mathbb{R}^{d_2}.$$

Write $A = (Q, K)$ for the attention parameters, $H = (V, W, a)$ for the output-head factors, and $\Theta = (A, H)$ for the full parameter vector. Here V is the value projection, W is the attention output projection, and a is the classifier.

For sample i, the attention scores, probabilities, aggregate feature, and scalar prediction are

$$z_i(A) = \frac{X_i K Q^\top q_i}{\sqrt{d}} \in \mathbb{R}^s, \qquad p_i(A) = \mathrm{softmax}(z_i(A)) \in \mathbb{R}^s, \tag{1}$$

$$h_i(A) = X_i^\top p_i(A) \in \mathbb{R}^d, \qquad f_\Theta(X_i) = h_i(A)^\top v(H), \qquad v(H) = VWa \in \mathbb{R}^d. \tag{2}$$

Thus $p_i(A)$ is a probability vector, and $h_i(A)$ is a convex combination of the input tokens. All five blocks are trained jointly, so the classifier acts on features that move with attention.

The signed margin of sample i and the minimum training margin are

$$m_i(\Theta) = y_i f_\Theta(X_i) = y_i h_i(A)^\top VWa, \qquad q(\Theta) = \min_{i\in[n]} m_i(\Theta).$$

The predictor is cubic in the output-head factors:

$$f_{(A,tH)}(X_i) = t^3 f_{(A,H)}(X_i), \qquad t > 0.$$

This identity determines the normalization used below.

3.2 LOSSES, SCALES, AND NORMALIZED MARGINS

Let $\ell : \mathbb{R} \to \mathbb{R}_+$ be either the exponential loss or the logistic loss,

$$\ell_{\exp}(z) = e^{-z}, \qquad \ell_{\log}(z) = \log(1 + e^{-z}),$$

and define the empirical loss by

$$\mathcal{L}(\Theta) = \frac{1}{n}\sum_{i=1}^n \ell(m_i(\Theta)).$$

The natural radius of the cubic output head is

$$r(H) = \frac{\|V\|_F^2 + \|W\|_F^2 + \|a\|^2}{3}, \qquad \rho(H) = \|VWa\|, \qquad u(H) = \frac{VWa}{\|VWa\|} \quad (VWa \neq 0).$$

The cubic head scale is $r^{3/2}$.

Let $\ell^{-1} : (0, \ell(0)) \to (0, \infty)$ denote the inverse loss restricted to positive margins. When $n\mathcal{L}(\Theta) < \ell(0)$, the smooth minimum-margin surrogate is

$$\alpha(\Theta) = \ell^{-1}\big(n\mathcal{L}(\Theta)\big) = \begin{cases} -\log\big(n\mathcal{L}(\Theta)\big), & \ell = \ell_{\exp}, \\ -\log\big(e^{n\mathcal{L}(\Theta)} - 1\big), & \ell = \ell_{\log}. \end{cases}$$

Since ℓ is decreasing, $\alpha(\Theta) \le q(\Theta)$. The normalized smooth margin is

$$\Psi(\Theta) = \frac{\alpha(\Theta)}{r(H)^{3/2}}.$$

The head has degree three, so $r^{3/2}$ is its natural normalizer; the attention parameters remain unnormalized.

3.3 Optimization dynamics

We study full-batch gradient flow

$$\dot{\Theta}(t) = -\nabla\mathcal{L}(\Theta(t))$$

and gradient descent

$$\Theta_{k+1} = \Theta_k - \eta_k \nabla\mathcal{L}(\Theta_k).$$

All gradients are taken with respect to the standard Euclidean inner product (Frobenius for matrix blocks). In the theoretical results, time is reset at the low-loss checkpoint specified below.

4 Part I: Implicit bias of the jointly trained head

In this section, we establish margin growth and directional convergence for the jointly trained head. In the limit, the effective classifier normalized to unit margin solves a hard-margin problem on the features generated by training (Theorem 4.4). Part II gives conditions that identify those features.

4.1 Model properties and training assumptions

The analysis uses two properties of the model in Section 3: its prediction is smooth in all five blocks, and $f_{(A,tH)}(X_i) = t^3 f_{(A,H)}(X_i)$ for $t > 0$. These follow from exact softmax and the factorization VWa. The loss and checkpoint conditions are as follows.

(A1) Classification loss. The loss is either $\ell_{\exp}$ or $\ell_{\log}$ from Section 3.

(A2) Low-loss checkpoint. At a finite time t_0,

$$n\mathcal{L}(\Theta(t_0)) < 1 \quad (\ell = \ell_{\exp}), \qquad n\mathcal{L}(\Theta(t_0)) < \log 2 \quad (\ell = \ell_{\log}).$$

Both losses have exponential tails. The threshold in (A2) makes the smooth margin α positive and therefore ensures correct classification of every training sample. Following Lyu & Li (2020), we study the dynamics after this checkpoint and reset time so that $t_0 = 0$.

For gradient descent, we use the same checkpoint and additionally require:

(S1) Small fixed step size. The learning rate is constant, $\eta_k \equiv \eta$, with $0 < \eta \le \eta_\star$, where $\eta_\star > 0$ depends on the checkpoint and the finite data set.

The step-size bound controls the discrete remainder relative to the increase of the normalized margin; Appendix B constructs one bound valid for the whole trajectory.

4.2 Margin growth from partial homogeneity

Partial homogeneity gives a normalized-margin law for gradient flow, even though the full model is not homogeneous.

Theorem 4.1 (Gradient flow)**.** *For the model of Section 3, under* (A1)–(A2)*, gradient flow exists for all $t \ge 0$. Its loss and output-head radius satisfy*

$$\mathcal{L}(t) \longrightarrow 0, \qquad r(t) \longrightarrow \infty, \qquad \Psi(t) \uparrow \gamma_\infty > 0.$$

Moreover, the three output factors asymptotically balance and align:

$$\frac{\|V(t)\|_F^2}{r(t)}, \ \frac{\|W(t)\|_F^2}{r(t)}, \ \frac{\|a(t)\|^2}{r(t)} \longrightarrow 1, \qquad \frac{\rho(t)}{r(t)^{3/2}} \longrightarrow 1, \qquad \frac{q(t)}{\rho(t)} \longrightarrow \gamma_\infty.$$

In addition, the attention factors are asymptotically smaller than the output-head factors:

$$\|(Q(t), K(t))\| = o\Big(\sqrt{r(t)}\Big), \qquad \|Q(t)K(t)^\top\|_F = o(r(t)).$$

Euler's identity gives the exact growth law for the head radius,

$$\dot{r}(t) = -\frac{2}{n}\sum_{i=1}^{n} \ell'\big(m_i(t)\big) m_i(t) > 0.$$

The attention contribution to the derivative of Ψ is also nonnegative, so the moving features preserve this monotonicity.

For gradient descent, the second-order remainder is controlled relative to the first-order increase. A sufficiently small fixed step preserves monotonicity of Ψ and yields the same limiting laws.

Theorem 4.2 (Gradient descent)**.** *For the model of Section 3, under* (A1)–(A2) *and* (S1)*, the normalized margin itself is nondecreasing:*

$$\Psi(\Theta_{k+1}) \geq \Psi(\Theta_k).$$

Furthermore,

$$\mathcal{L}(\Theta_k) \longrightarrow 0, \qquad r(\Theta_k) \longrightarrow \infty, \qquad \Psi(\Theta_k) \longrightarrow \gamma_\infty > 0,$$

and all balance, margin, and relative attention-scale conclusions in Theorem 4.1 hold with t replaced by k.

Convergence rates. The same estimates give bounds on loss decay and weight growth. Write $\mathcal{T} = t$ for gradient flow and $\mathcal{T} = \eta k$ for fixed-step gradient descent.

Theorem 4.3 (Convergence rates)**.** *Under* (A1)–(A2) *for the model of Section 3, and additionally* (S1) *for GD, there exist constants $c_1, c_2 > 0$ such that, for all sufficiently large $\mathcal{T}$,*

$$\frac{c_1}{\mathcal{T}(\log \mathcal{T})^{8/3}} \leq \mathcal{L}(\Theta(\mathcal{T})) \leq \frac{c_2}{\mathcal{T}(\log \mathcal{T})^{4/3}}, \qquad \|\Theta(\mathcal{T})\| \sim \sqrt{3}\,\gamma_\infty^{-1/3}(\log \mathcal{T})^{1/3}.$$

4.3 Directional convergence and asymptotic margin bias

The limiting margin problem involves every accumulation point of the signed aggregate features. Set $s_i(t) = y_i h_i(A(t))$ and

$$\mathcal{Z}_\infty = \{z : \exists i,\ t_j \to \infty \text{ such that } s_i(t_j) \to z\}.$$

For gradient descent, t_j denotes iteration indices. Since every h_i lies in the convex hull of the tokens of sample i, $\mathcal{Z}_\infty$ is nonempty and compact.

Theorem 4.4 (Directional convergence and asymptotic margin bias)**.** *Under* (A1)–(A2) *for the model of Section 3, and additionally* (S1) *for GD, the following statements hold.*

1. *Define the normalized output head by $\bar{H}(t) = H(t)/\sqrt{r(t)}$. There exist a limit $\bar{H}_\infty$ and a unit vector u_∞ such that*

$$\bar{H}(t) \longrightarrow \bar{H}_\infty, \qquad \frac{VWa}{\|VWa\|} \longrightarrow u_\infty, \qquad \frac{VWa}{q} \longrightarrow w_\infty := \frac{u_\infty}{\gamma_\infty}.$$

   In particular,

$$\frac{H}{q^{1/3}} \longrightarrow H_* := \gamma_\infty^{-1/3}\bar{H}_\infty.$$

   For $\mathcal{T} = t$ in flow and $\mathcal{T} = \eta k$ in descent,

$$\frac{q}{\log(2+\mathcal{T})} \to 1, \qquad \frac{VWa}{\log(2+\mathcal{T})} \to w_\infty, \qquad \frac{H}{[\log(2+\mathcal{T})]^{1/3}} \to H_*.$$

2. *The vector w_∞ is the unique solution of*

$$\min_{w \in \mathbb{R}^d} \frac{1}{2}\|w\|^2 \quad \textit{subject to} \quad z^\top w \geq 1 \quad \textit{for every } z \in \mathcal{Z}_\infty. \tag{3}$$

In particular, the limiting unit-margin classifier solves the convex problem (3) even if the representations have multiple accumulation points. Its constraint set depends on the training trajectory.

4.4 WHY THE HEAD DIRECTION CONVERGES

Two ingredients establish directional convergence. For exponential-loss gradient flow, differentiating the smooth margin in log-radius time gives a decomposition involving head motion, attention motion, and entropy. This follows the radial–tangential argument of Lyu & Li (2020, Section 5, Lemma 5.1). Integrating the resulting inequality gives a finite squared-motion budget and hence small-motion subsequences. It does not, by itself, give convergence of the head direction.

Finite path length requires additional control of attention. By the exact softmax pair identities, tokens with a large logit gap have exponentially small probability products. Lemma C.1 then bounds the attention scale strictly below the head scale. After compactifying the trajectory, the desingularizing gradient inequality of Lemma D.1 yields finite length of the normalized head path. Balance and alignment transfer this limit to the effective classifier, even if the attention features have several accumulation points.

To identify the limit, a common small-motion subsequence supplies signed features and loss weights supported on margin-active constraints. Their weighted sum is parallel to the limiting classifier, giving optimality in equation 3; feasibility holds for every limiting feature. Detailed proofs appear in Appendices A–F.

5 PART II: IDENTIFYING THE LIMITING ATTENTION REPRESENTATION

In this section, we give conditions under which the limiting features of Theorem 4.4 are prescribed tokens. The key geometry is a common cone in which every head direction prefers the same token in each sample. Under the conditions below, joint training turns this preference into diverging logit gaps. Vector and matrix inequalities are coordinatewise.

5.1 COOPERATIVE GEOMETRY AND A FINITE CERTIFICATE

The model and losses are those of Section 3, with $s \geq 2$ tokens per sample. For prescribed indices $j_i^\star$, write

$$s_{ij} = y_i x_{ij}, \qquad t_i = y_i q_i, \qquad D_{ij} = s_{ij_i^\star} - s_{ij} \quad (j \neq j_i^\star). \tag{4}$$

Form a single list $R = (g_1, \ldots, g_N)$ of all nonzero signed tokens from all samples and all nonzero columns of V_0, including repeated vectors when they arise. Set $C = \mathrm{cone}\{g_1, \ldots, g_N\}$. We impose the following geometric and checkpoint conditions.

(G1) Positive queries and deficits. $t_i > 0$ for every i, and $D_{ij} > 0$ for every i and $j \neq j_i^\star$.

(G2) Acute cone with uniform reward. Require

$$g_\alpha^\top g_\beta \geq 0 \quad (\alpha, \beta \in [N]), \qquad \epsilon_{\mathrm{rew}} := \min_{i,\, j \neq j_i^\star,\, \alpha} \frac{D_{ij}^\top g_\alpha}{\|g_\alpha\|} > 0. \tag{5}$$

(G3) Finite covariance-action certificate. For every $i \in [n]$, $k \in [d]$, and $\alpha \in [N]$, there exists a nonnegative flow satisfying the allowed-edge, capacity, and demand constraints of Appendix G.1, equation 58. These are ndN separate finite linear feasibility problems constructed from the tokens and g_α.

(G4) Nonnegative, nondegenerate low-loss checkpoint. Require

$$K_0, Q_0, W_0, a_0 \geq 0, \qquad \delta_{\mathrm{row}} := \min_{b \in [d]} \sum_{\ell=1}^{d} (Q_0)_{b\ell}^2 > 0, \qquad n\mathcal{L}(0) < \ell(0). \tag{6}$$

Condition (G1) makes each prescribed gap increase with positive attention motion. Condition (G2) guarantees the common reward bound $D_{ij}^\top u \geq \epsilon_{\mathrm{rew}}$ for every unit head $u \in C$. Acuteness and (G4) preserve this cone. The positive lower bound on the row norms keeps every gap coupled to the attention drive, including when $K_0 = 0$.

The checkpoint in (G4) is precisely the low-loss condition (A2) of Part I; gradient descent also requires (S1). To write the loss gradient in a common form, define the positive derivative mass and

normalized sample weights by

$$\sigma = \frac{1}{n}\sum_i [-\ell'(m_i)], \qquad \pi_i = \frac{-\ell'(m_i)}{n\sigma}, \qquad \sum_i \pi_i = 1.$$

For exponential loss, $\sigma = \mathcal{L}$; for logistic loss, $\sigma = n^{-1}\sum_i (1+e^{m_i})^{-1}$. In either case, the gradient has the common multiplier σ and positive weights π_i.

To see the role of (G3), let $\Sigma_i(\zeta)$ be the covariance of the signed tokens with probabilities proportional to $e^{s_{ij}^\top \zeta}$, as defined in equation 59. The finite checks imply

$$\Sigma_i(\zeta) C \subseteq \mathbb{R}^d_+ \qquad \text{for every } i \in [n] \text{ and } \zeta \in \mathbb{R}^d_+. \tag{7}$$

The attention gradient applies this covariance to the moving head. Thus (G3) keeps the attention motion coordinatewise nonnegative throughout the cone. The witness matches negative token-pair contributions to positive ones whose exponential weights are at least as large. As a result, the head preserves its token preference while attention reinforces it. Appendix G.1 defines the witness and proves equation 7.

5.2 Selection and classifier bias

The following results show that these finite conditions force token selection under both gradient flow and gradient descent.

Theorem 5.1 (Strict selection under joint gradient flow)**.** *For the model above, suppose* (G1)–(G4) *hold for the prescribed indices $j_i^\star$ and full generator list R. Consider unit-rate Euclidean gradient flow. Then the five-factor flow exists for all $t \geq 0$, satisfies the invariance and reward bounds of Lemma G.3, and selects the prescribed tokens: every selected-token logit gap*

$$\Delta_{ij}(t) = \frac{D_{ij}^\top K(t) Q(t)^\top t_i}{\sqrt{d}}, \qquad j \neq j_i^\star, \tag{8}$$

is strictly increasing to $+\infty$. Consequently,

$$p_{ij_i^\star}(t) \uparrow 1, \qquad p_{ij}(t) \longrightarrow 0 \ (j \neq j_i^\star), \qquad h_i(t) \longrightarrow x_{ij_i^\star}. \tag{9}$$

Proof mechanism. Acuteness and the covariance certificate keep the moving head in C and the attention factors nonnegative. The reward and covariance signs make every prescribed gap strictly increase. Divergence follows by considering the clock $\vartheta(t) = \int_0^t \sigma(\tau)\|v(\tau)\|\, d\tau$, which tends to infinity. If a gap remained bounded, the total attention drive would be integrable in this clock. Gronwall's inequality would then bound Q and K, keeping every token probability bounded away from zero. The variance identity gives a positive lower bound on the same drive, a contradiction. Since the normalized loss weights sum to one, no per-sample lower bound is needed. Appendix G gives the full proof.

Corollary 5.2 (Selected-token head SVM)**.** *Under the hypotheses of Theorem 5.1, let $q(t) = \min_i m_i(t)$ as in Section 3. Then*

$$\frac{V(t)W(t)a(t)}{q(t)} \longrightarrow w_\star, \qquad w_\star = \operatorname*{argmin}_{w \in \mathbb{R}^d} \left\{ \frac{1}{2}\|w\|^2 : s_{ij_i^\star}^\top w \geq 1 \text{ for all } i \right\}. \tag{10}$$

The minimizer is unique, and $u(t) \to w_\star/\|w_\star\|$.

Proof. The chosen loss satisfies (A1), and (G4) gives (A2), so Theorem 4.4 applies. By equation 9, its asymptotic signed-feature set is $\mathcal{Z}_\infty = \{s_{ij_i^\star} : i \in [n]\}$. Applying that theorem gives equation 10, uniqueness, and directional convergence. □

Theorem 5.3 (Strict selection under fixed-step gradient descent)**.** *Suppose the data, model, and checkpoint satisfy* (G1)–(G4) *as in Theorem 5.1. Run simultaneous full-batch gradient descent on all five factors with a constant step size satisfying* (S1)*. Then every gap $\Delta_{ij,k}$ in equation 8 is strictly increasing and tends to $+\infty$. The probability and representation limits in equation 9 hold with t replaced by k. Moreover,*

$$\frac{V_k W_k a_k}{q_k} \longrightarrow w_\star, \qquad \frac{V_k W_k a_k}{\log(2+\eta k)} \longrightarrow w_\star,$$

where $w_\star$ is the unique selected-token SVM solution in equation 10.

The gradient-descent proof uses the Part I radius law and the exact simultaneous update of $KQ^\top$; details appear in Appendix G.3.

6 Experiments

In this section, we examine two predictions of the theory: growth of the normalized margin during joint training (Part I), and convergence to a selected-token SVM when the limiting features can be identified (Part II). The first experiment uses MNIST; the second uses a small problem with an explicit SVM solution. All five parameter blocks are trained by full-batch GD on exponential loss.

6.1 Joint training on MNIST

We train on 512 MNIST images of digits 0 and 1 (LeCun et al., 1998). Each image supplies sixteen 7×7 patch tokens and their mean, which also serves as the query; fixed position codes are appended. Training runs for 20000 updates with step size 0.05. We compare $Q_0 = c_Q\bar{Q}_0$ for $c_Q \in \{3, 30\}$, with $K_0 = 0$ and the same initial head within each pair. The two runs start with uniform attention and identical predictions, but their attention updates depend on the query scale. Additional details appear in Appendix H.

Theorem 4.2 predicts that the smooth normalized margin increases once the loss is sufficiently small. Accordingly, Figure 1a–b plots the loss and $\Psi = -\log(n\mathcal{L})/r^{3/2}$, with each margin curve starting after $n\mathcal{L} < 1$. The loss falls and Ψ continues to grow at both query scales. Normalization removes the cubic growth due to rescaling the head, so the increase measures an improvement in margin beyond weight growth alone.

To examine which tokens attention favors, we also measure the maximum attention probability and the relative reward gap. For the normalized signed token reward $b_{ij} = y_i x_{ij}^\top v/\|v\|$, define

$$R_i = \frac{\max_j b_{ij} - \sum_j p_{ij} b_{ij}}{\max_j b_{ij} - \min_j b_{ij}}.$$

A smaller R_i means that the aggregate reward lies closer to the best-token reward, relative to the token reward range. Figure 1c–d shows stronger concentration and smaller reward gaps at the larger query scale. Thus the margin growth in Part I is visible with both diffuse and concentrated attention. Attention maps and margins on the current features appear in Appendix H.

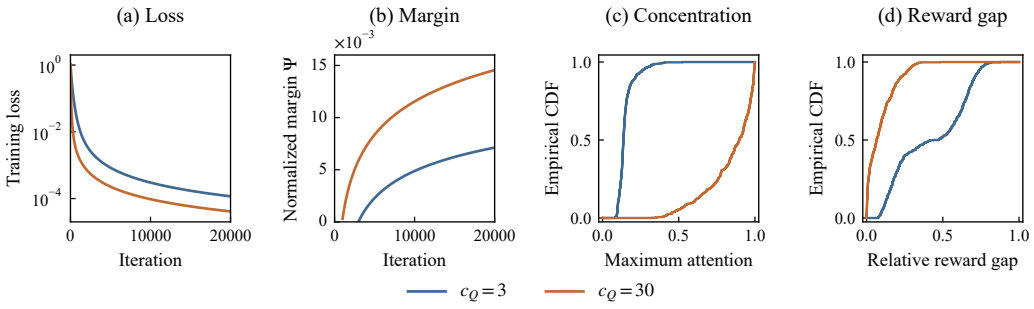


Figure 1: **Margin growth and attention concentration on MNIST.** (a) Training loss. (b) Smooth normalized margin after the low-loss threshold is reached. (c–d) Final empirical CDFs across training images of $\max_j p_{ij}$ and R_i. Larger maximum probabilities indicate stronger concentration; smaller reward gaps indicate closer agreement with the head's token preferences. Colors denote query initialization scales. All panels use one paired initialization; two further pairs are examined in Appendix H.

6.2 Token selection and classifier direction

Theorem 5.3 connects diverging token gaps to a fixed maximum-margin classifier. To examine this connection, we use four binary examples with three two-dimensional tokens each. The data and

initialization satisfy (G1)–(G4), with a nonzero flow in the covariance certificate. Their selected-token SVM is $w_\star = (0.3, 0.4)^\top$. We fix the data and initial query and key, and vary the initial head orientation through $\vartheta \in \{75°, 80°, 85°\}$ at the same initial loss. GD uses step size 0.5 for 20000 updates. The construction and initialization appear in Appendix H.4.

Figure 2 follows the two conclusions of Part II. Panel (a) plots the smallest prescribed-token logit gap over all samples and competitors. A large value means that every prescribed token dominates its competitors in attention probability. Panel (b) measures the angle between v and $w_\star$, while panel (c) measures $\|v/q - w_\star\|/\|w_\star\|$. The latter compares the unit-margin classifier with the actual SVM solution and therefore includes the scale as well as the direction. From all three initial orientations, selection gaps grow and the classifier approaches the same SVM target. These trajectories illustrate how identifying the limiting tokens turns the trajectory-dependent problem in Part I into a fixed margin problem. Step-size comparisons and longer gradient-flow trajectories appear in Appendix H.4.

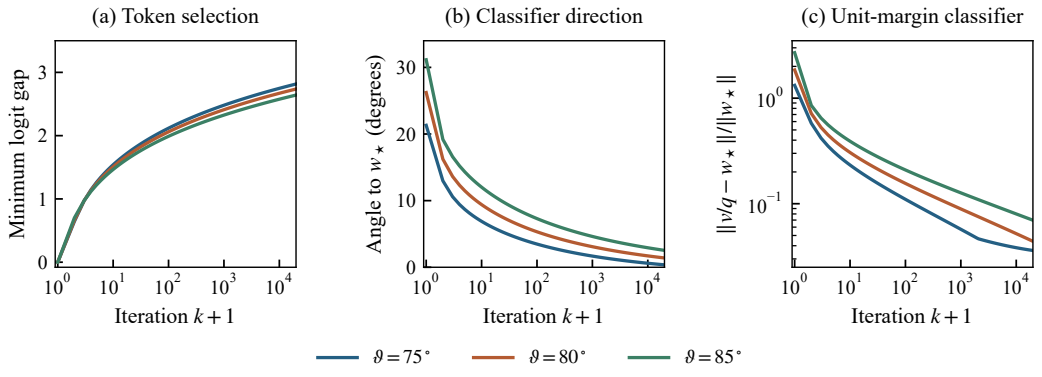


Figure 2: **Token selection and maximum-margin direction under GD.** (a) Minimum prescribed-token logit gap over all samples and competitors. (b) Angle between the effective classifier v and the fixed SVM solution $w_\star$. (c) Relative distance of v/q from $w_\star$, where q is the smallest current training margin. Colors denote the head initialization parameter ϑ; all query, key, and head factors are trained. The logarithmic horizontal axis shows $k + 1$, including initialization at $k = 0$.

7 Conclusion

We have established a maximum-margin characterization for a jointly trained softmax attention classifier under gradient flow and sufficiently small fixed-step gradient descent. Partial homogeneity of the output head gives a monotone normalized margin. Controlling attention motion then yields convergence of the head direction, and the unit-margin classifier converges to the minimum-norm separator of all limiting signed features. Under additional geometric conditions, attention selects prescribed tokens and the limiting problem becomes their hard-margin SVM.

These results leave several natural questions. Can the limiting features be identified from the data and initialization when token preferences change during training? The invariant cone in Part II ensures a common preference throughout the trajectory; allowing this preference to change would broaden the selection result. A further question is whether the loss bounds can be sharpened to matching rates, and whether the sufficient step size can be made explicit. These refinements would give a more quantitative account of the coupled training dynamics.

AI USE STATEMENT

The authors directed the research and used generative AI tools to assist with literature discovery, the exploration and checking of mathematical arguments, and experimental design and analysis. The tools also assisted with code checks and the preparation and revision of selected passages of the manuscript. The authors evaluated the resulting arguments, experiments, and text and take responsibility for the final paper.

REFERENCES

Yuhang Cai, Kangjie Zhou, Jingfeng Wu, Song Mei, Michael Lindsey, and Peter L. Bartlett. Implicit bias of gradient descent for non-homogeneous deep networks. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 6426–6504, 2025.

Zheng-An Chen and Tao Luo. From condensation to rank collapse: A two-stage analysis of transformer training dynamics. In *Advances in Neural Information Processing Systems*, volume 38, 2025.

Didier D'Acunto and Krzysztof Kurdyka. Explicit bounds for the Łojasiewicz exponent in the gradient inequality for polynomials. *Annales Polonici Mathematici*, 87(1):51–61, 2005.

Simon S. Du, Wei Hu, and Jason D. Lee. Algorithmic regularization in learning deep homogeneous models: Layers are automatically balanced. In *Advances in Neural Information Processing Systems*, volume 31, 2018.

Suriya Gunasekar, Jason D. Lee, Daniel Soudry, and Nathan Srebro. Characterizing implicit bias in terms of optimization geometry. In *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pp. 1832–1841, 2018a.

Suriya Gunasekar, Jason D. Lee, Daniel Soudry, and Nathan Srebro. Implicit bias of gradient descent on linear convolutional networks. In *Advances in Neural Information Processing Systems*, volume 31, 2018b.

Ziwei Ji and Matus Telgarsky. Gradient descent aligns the layers of deep linear networks. In *International Conference on Learning Representations*, 2019.

Ziwei Ji and Matus Telgarsky. Directional convergence and alignment in deep learning. In *Advances in Neural Information Processing Systems*, volume 33, 2020.

Addison Kristanto Julistiono, Davoud Ataee Tarzanagh, and Navid Azizan. Optimizing attention with mirror descent: Generalized max-margin token selection. *Journal of Machine Learning Research*, 27(36):1–61, 2026.

Krzysztof Kurdyka. On gradients of functions definable in o-minimal structures. *Annales de l'Institut Fourier*, 48(3):769–783, 1998.

Yann LeCun, Léon Bottou, Yoshua Bengio, and Patrick Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998.

Yuchen Li, Yuanzhi Li, and Andrej Risteski. How do transformers learn topic structure: Towards a mechanistic understanding. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 19689–19729, 2023.

Kaifeng Lyu and Jian Li. Gradient descent maximizes the margin of homogeneous neural networks. In *International Conference on Learning Representations*, 2020.

Mor Shpigel Nacson, Jason D. Lee, Suriya Gunasekar, Pedro Henrique Pamplona Savarese, Nathan Srebro, and Daniel Soudry. Convergence of gradient descent on separable data. In *Proceedings of the Twenty-Second International Conference on Artificial Intelligence and Statistics*, volume 89 of *Proceedings of Machine Learning Research*, pp. 3420–3428, 2019.

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Kopf, Edward Yang, Zach DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. PyTorch: An imperative style, high-performance deep learning library. In *Advances in Neural Information Processing Systems*, volume 32, 2019.

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.

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.

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, 2023.

Bhavya Vasudeva, Puneesh Deora, and Christos Thrampoulidis. Implicit bias and fast convergence rates for self-attention. *Transactions on Machine Learning Research*, 2025.

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

A LOSS ALGEBRA AND FOUNDATIONAL IDENTITIES

In this section, we collect the loss identities and head dynamics needed for Part I, allowing the attention features to move. All vector norms are Euclidean, all matrix norms are Frobenius, and a norm on a tuple of parameter blocks is the product Euclidean norm. Write

$$A = (Q, K), \qquad H = (V, W, a), \qquad r = \frac{\|H\|^2}{3}, \qquad \beta = r^{3/2}, \qquad v = VWa, \qquad \rho = \|v\|.$$

If $C_X = \max_{i,j}\|x_{ij}\|$, then the convex-combination property of softmax gives

$$|m_i| \le C_X \rho, \qquad \rho \le \|V\|_F \|W\|_F \|a\| \le r^{3/2}. \tag{11}$$

A.1 A COMMON REPRESENTATION OF THE TWO LOSSES

We express both losses through the same soft minimum, so the later flow and GD arguments can use one set of weights.

Lemma A.1 (Exact loss normalization)**.** *In the low-loss region of assumption* (A2)*, define*

$$\lambda = \begin{cases} \mathcal{L}, & \ell = \ell_{\exp}, \\ \{n(1+e^{\alpha})\}^{-1}, & \ell = \ell_{\log}, \end{cases}$$

and

$$\omega_i = \begin{cases} \dfrac{e^{-m_i}}{\sum_j e^{-m_j}}, & \ell = \ell_{\exp}, \\ \dfrac{(1+e^{m_i})^{-1}}{(1+e^{\alpha})^{-1}}, & \ell = \ell_{\log}. \end{cases}$$

Let $\kappa = \sum_i \omega_i$, $\pi_i = \omega_i/\kappa$, $C = \sum_i \omega_i m_i$, and $\mathcal{D} = C - \alpha$. Then

$$-\nabla\mathcal{L} = \lambda \nabla\alpha, \qquad \nabla\alpha = \sum_i \omega_i \nabla m_i, \qquad 0 < \lambda \le \mathcal{L} \le 2\lambda, \tag{12}$$

$$0 \le q - \alpha \le \begin{cases} \log n, & \ell_{\exp}, \\ \log(2n), & \ell_{\log}, \end{cases} \qquad \mathcal{L} \le n^{-1} e^{-\alpha}. \tag{13}$$

Moreover, $\kappa = 1$ for the exponential loss and $1 \le \kappa \le 1 + e^{-\alpha} \le 2$ for the logistic loss.

Proof. For the exponential loss the statements follow directly from $\alpha = -\log \sum_i e^{-m_i}$. For the logistic loss,

$$e^{n\mathcal{L}} - 1 = \prod_i (1 + e^{-m_i}) - 1 = \sum_{\varnothing \ne I \subseteq [n]} e^{-m_I}, \qquad m_I = \sum_{i \in I} m_i.$$

Thus, for the exponential loss take $\mathcal{I} = \{\{1\}, \dots, \{n\}\}$, and for the logistic loss take $\mathcal{I}$ to be all nonempty subsets of $[n]$. In both cases,

$$\alpha = -\log \sum_{I \in \mathcal{I}} e^{-m_I}, \quad P_I = \frac{e^{-m_I}}{\sum_J e^{-m_J}}, \quad C = \sum_I P_I m_I, \quad \mathcal{D} = -\sum_I P_I \log P_I. \tag{14}$$

Differentiation gives the displayed gradient formula. In the logistic case, summing P_I over subsets containing i gives exactly the stated ω_i. Equation 14 also proves

$$0 \le \mathcal{D} = C - \alpha \le \log|\mathcal{I}|.$$

For the bound on κ in the logistic case, put $x_i = e^{-m_i}$ and $Z = \prod_i (1 + x_i) - 1 = e^{-\alpha}$. Then

$$\kappa = \frac{1+Z}{Z} \sum_i \frac{x_i}{1+x_i}.$$

The lower bound follows from $1 - \prod_i (1+x_i)^{-1} \le \sum_i x_i/(1+x_i)$; the upper bound follows from $\sum_i x_i/(1+x_i) \le \sum_i x_i \le Z$. The remaining bounds follow from $x/(1+x) \le \log(1+x) \le x \le 2x/(1+x)$ for $0 < x < 1$ and from $e^{-m}/2 \le \log(1+e^{-m}) \le e^{-m}$ when $m > 0$. □

A.2 RADIAL AND TANGENTIAL DYNAMICS

We separate growth of the head radius from motion of its normalized direction and record the identities used throughout Part I.

Set $u_\alpha = \nabla\alpha$ and decompose its head component as

$$(u_\alpha)_H = \frac{C}{r}H + w, \qquad \langle H, w\rangle = 0.$$

Lemma A.2 (Euler identities and normalized dissipation)**.** *Every head block $H_j \in \{V, W, a\}$ satisfies $\langle H_j, (u_\alpha)_{H_j}\rangle = C$. Consequently,*

$$\dot\Theta = \lambda u_\alpha, \qquad \dot r = 2\lambda C > 0, \qquad \dot\alpha = \lambda\|u_\alpha\|^2, \tag{15}$$

$$\dot\Psi = \frac{\lambda}{\beta}\left(\|(u_\alpha)_A\|^2 + \|w\|^2 + \frac{3C\mathcal{D}}{r}\right) \geq 0. \tag{16}$$

For a gradient-descent update $\Theta^+ = \Theta - \eta\nabla\mathcal{L}$,

$$r^+ - r = 2\eta\lambda C + \frac{\eta^2}{3}\|\nabla_H\mathcal{L}\|^2 > 0. \tag{17}$$

Proof. Each composite margin m_I is one-homogeneous in each of V, W, a separately. Euler's identity, averaged with the weights P_I, gives $\langle H_j, (u_\alpha)_{H_j}\rangle = C$. Hence $\langle H, (u_\alpha)_H\rangle = 3C$ and the displayed orthogonal decomposition follows because $\|H\|^2 = 3r$. Differentiating $\Psi = \alpha r^{-3/2}$, using $\nabla_H r = 2H/3$ and $C - \alpha = \mathcal{D}$, proves equation 15–equation 16. Expanding $\|H + \eta\lambda(u_\alpha)_H\|^2/3$ proves equation 17 without discarding its quadratic term. □

For the normalized equations, put

$$\xi = \frac{H}{\sqrt{r}}, \qquad z = r^{-1/2}, \qquad F_i(A, \xi) = y_i h_i(A)^\top \xi_V \xi_W \xi_a.$$

Let

$$\mathscr{F}(A, \xi, z) = -z^3 \log \sum_{I\in\mathcal{I}} e^{-F_I(A,\xi)/z^3}, \qquad F_I = \sum_{i\in I} F_i.$$

Along the trajectory $\mathscr{F} = \Psi$. Define

$$g_A = \nabla_A \mathscr{F}, \qquad d = \nabla_\xi \mathscr{F}, \qquad c = C/\beta, \qquad g_\perp = d - c\xi.$$

Since $c = \Psi + \mathcal{D}/\beta$, the bounds on Ψ and the entropy give fixed positive upper and lower bounds on c along the trajectories established below. Since $\|\xi\|^2 = 3$ and each F_I is cubic in ξ,

$$\langle \xi, d\rangle = 3c, \qquad \langle \xi, g_\perp\rangle = 0. \tag{18}$$

With $\tau = \log r$, direct substitution into Lemma A.2 gives

$$A_\tau = \frac{r}{2c} g_A, \qquad \xi_\tau = \frac{1}{2c} g_\perp, \qquad z_\tau = -\frac{z}{2}, \tag{19}$$

$$\Psi_\tau = \frac{r\|g_A\|^2 + \|g_\perp\|^2}{2c} + \frac{3\mathcal{D}}{2\beta}. \tag{20}$$

The composite representation makes the identity exact for both losses.

The following identities specialize the layer-balancing mechanism of Du et al. (2018) and the deep linear alignment analysis of Ji & Telgarsky (2019) to the two factorized paths of our model.

Lemma A.3 (Balance invariants)**.** *Along gradient flow,*

$$Q^\top Q - K^\top K, \qquad V^\top V - WW^\top, \qquad W^\top W - aa^\top$$

are constant. Under the small-step gradient descent of assumption (S1)*, their cumulative second-order drift is finite. Hence, along either dynamics as $r \to \infty$,*

$$\|V\|_F^2 = r + O(1), \quad \|W\|_F^2 = r + O(1), \quad \|a\|^2 = r + O(1), \quad \frac{\rho}{r^{3/2}} \longrightarrow 1. \tag{21}$$

Proof. The predictor depends on Q, K through $QK^\top$ and on V, W, a through VWa. If J and e denote the loss gradients with respect to these two products, then

$$g_Q = JK, \quad g_K = J^\top Q, \qquad g_V = e(Wa)^\top, \quad g_W = V^\top e\, a^\top, \quad g_a = W^\top V^\top e.$$

The first-order terms in the three Gram differences therefore cancel. For gradient descent only the quadratic terms remain, and their norms are summable under the descent estimate supplied by the sufficiently small step size. Taking traces gives pairwise $O(1)$ differences among the three head-block norms. Any cluster point of $H/\sqrt{r}$ then satisfies

$$\bar{V}^\top \bar{V} = \bar{W}\bar{W}^\top, \qquad \bar{W}^\top \bar{W} = \bar{a}\bar{a}^\top, \qquad \|\bar{V}\|_F = \|\bar{W}\|_F = \|\bar{a}\| = 1.$$

These identities force $\bar{W} = b\bar{a}^\top$ and $\bar{V} = ub^\top$ for unit vectors b, u, so $\|\bar{V}\bar{W}\bar{a}\| = 1$. Compactness proves equation 21. □

Proposition A.4 (Foundational convergence and attention energy)**.** *Under the hypotheses of Theorem 4.1, the flow exists for all time and*

$$\mathcal{L} \to 0, \qquad r \to \infty, \qquad \Psi \uparrow \gamma_\infty \in (0, C_X].$$

Moreover, in the radius clock,

$$\int_{r(0)}^{\infty} \left\| \frac{dA}{dr} \right\|^2 dr < \infty. \tag{22}$$

The corresponding gradient-descent interpolation satisfies the same finite-energy statement. In particular,

$$\|A\| = o(\sqrt{r}), \qquad \|QK^\top\|_F = o(r). \tag{23}$$

Proof. On every finite parameter set the vector field is smooth. If a maximal flow had finite lifetime T_*, loss dissipation would give

$$\|\Theta(t) - \Theta(s)\| \le \sqrt{(t-s)\mathcal{L}(\Theta(0))},$$

so the trajectory would converge to a finite point as $t \uparrow T_*$ and the local solution could be extended. Hence the flow is global. The low-loss region is invariant. Equations 11 and 16 imply $0 < \Psi(0) \le \Psi(t) \le C_X$. If the decreasing loss had a positive limit, then α and r would remain bounded while $\dot{r} = 2\lambda C$ had a positive lower bound, a contradiction. Thus $\mathcal{L} \to 0$, $\alpha \to \infty$, and $r \to \infty$. Equations 13 and 21 give all remaining limits in Theorem 4.1.

Finally, $A_r = g_A/(2c)$ and $\Psi_r \ge \|g_A\|^2/(2c) = 2c\|A_r\|^2$. Since $c \ge \Psi(0)$, integration proves equation 22. Cauchy–Schwarz on $[R, r]$, followed first by $r \to \infty$ and then by $R \to \infty$, proves $\|A\| = o(\sqrt{r})$. The matrix-product bound $\|QK^\top\|_F \le \|A\|^2/2$ gives the second claim. The gradient-descent assertions are proved in Section B below. □

Proof of Theorem 4.1. Global existence, loss convergence, divergence of the output-head radius, and monotonic convergence of the normalized margin are proved in Proposition A.4. The balance identities and the limit $\rho/r^{3/2} \to 1$ are supplied by Lemma A.3. The soft-minimum comparison in Lemma A.1 then gives $q/\rho \to \gamma_\infty$. Finally, equation 23 gives both attention-scale conclusions. These are exactly all assertions of the theorem. □

B FIXED-STEP GRADIENT DESCENT

In this section, we show that a sufficiently small constant step preserves the loss decrease and normalized-margin increase found for gradient flow. The inequalities below express the estimates we prove for ordinary GD; the algorithm does not run a line search. Fix constants $\sigma_L, \sigma_\Psi \in (0, 1)$ and write $g = \nabla\mathcal{L}$ and $D_\Psi = -\langle \nabla\Psi, g\rangle$. We call a candidate $\Theta^+ = \Theta - \eta g$ *safe* when

$$\mathcal{L}(\Theta^+) \le \mathcal{L}(\Theta) - \sigma_L \eta \|g\|^2, \tag{24}$$

$$\Psi(\Theta^+) \ge \Psi(\Theta) + \sigma_\Psi \eta D_\Psi. \tag{25}$$

Lemma B.1 (Uniform relative one-step estimates). *Fix $r_{\min} > 0$, $K_A < \infty$, and $\psi_{\min} > 0$. On the region*

$$r \geq r_{\min}, \qquad \|A\| \leq K_A\sqrt{r}, \qquad \Psi \geq \psi_{\min}, \tag{26}$$

there are constants $C_*, \delta_* > 0$, depending only on this region and the finite data set, such that every positive step satisfying $\chi := \eta\mathcal{L}(1+r)^4 \leq \delta_*$ obeys

$$\mathcal{L}(\Theta - \eta g) \leq \mathcal{L} - \eta(1 - C_*\chi)\|g\|^2, \tag{27}$$

$$\Psi(\Theta - \eta g) - \Psi(\Theta) \geq \eta(1 - C_*\chi)D_\Psi. \tag{28}$$

The constants are uniform even when D_Ψ approaches zero.

Proof. Softmax and all of its derivatives are bounded functions of its logits. The logits are quadratic in A, while VWa is cubic in H. Repeated use of the product rule on the cone equation 26 gives, on a fixed relative neighbourhood of the current point,

$$\|\nabla m_i\| \leq C(1+r)^2, \quad \|\nabla^2 m_i\| \leq C(1+r)^{5/2}, \quad \|\nabla^2\Psi\| \leq C(1+r)^{5/2}. \tag{29}$$

Both losses satisfy $|\ell'| \leq \ell$ and $|\ell''| \leq \ell$. Consequently, along the update segment,

$$\|\nabla^2\mathcal{L}\| \leq C\mathcal{L}(1+r)^4, \tag{30}$$

provided χ is below a fixed constant. Indeed, the update length is at most $C\eta\mathcal{L}(1+r)^2$; integration of $\nabla \log \mathcal{L}$ shows that the loss at every point of the segment is within an $e^{C\chi}$ factor of its initial value. Taylor's theorem and equation 30 prove equation 27.

For the margin estimate, put $u = \nabla\alpha$, decompose $u_H = (C/r)H + w$ with $w \perp H$, and let

$$t = \eta\lambda, \qquad s_0 = 1 + tC/r, \qquad A^+ = A + tu_A, \qquad \widetilde{H} = H + (t/s_0)w.$$

The actual head update is $H^+ = s_0\widetilde{H}$. The function increment splits exactly into a nonradial and a radial part:

$$\Psi(A^+, H^+) - \Psi(A, H) = \Psi(A^+, \widetilde{H}) - \Psi(A, H) + \Psi(A^+, s_0\widetilde{H}) - \Psi(A^+, \widetilde{H}).$$

At the initial point,

$$\nabla_A\Psi = u_A/\beta, \qquad \nabla_H\Psi = w/\beta + \mathcal{D}H/(r\beta).$$

Orthogonality therefore gives the linear term $t(\|u_A\|^2 + \|w\|^2/s_0)/\beta$. By equation 29, the absolute remainder is at most

$$C(1+r)^{5/2}t^2(\|u_A\|^2 + \|w\|^2).$$

Since $s_0 - 1 = O(\chi)$ and $t\beta(1+r)^{5/2} \leq C\chi$, this proves

$$\Psi(A^+, \widetilde{H}) - \Psi(A, H) \geq (1 - C\chi)\frac{t}{\beta}(\|u_A\|^2 + \|w\|^2). \tag{31}$$

For the radial part, every composite margin is multiplied by z^3 at $(A^+, z\widetilde{H})$, and direct differentiation gives

$$\frac{d}{dz}\Psi(A^+, z\widetilde{H}) = \frac{3\mathcal{D}_z}{z^4\widetilde{r}^{3/2}}.$$

The composite margins change by at most $C\chi$ on the auxiliary segments. To compare the entropies, write M_I for these margins and $\bar{M} = \sum_I P_I M_I$. Direct differentiation gives

$$\frac{\partial\mathcal{D}}{\partial M_I} = P_I(\bar{M} - M_I).$$

Subtracting $\min_I M_I$ from all margins leaves this identity and the entropy unchanged. After this shift, $M_I \geq 0$ and $\mathcal{D} = \bar{M} + \log\sum_I e^{-M_I} \geq \bar{M}$. Hence $\|\nabla_M\mathcal{D}\|_1 \leq 2\bar{M} \leq 2\mathcal{D}$. Integrating gives $\mathcal{D}_z \geq e^{-C\chi}\mathcal{D}$. If the composite set has only one element, its entropy is identically zero and the radial estimate holds directly. Since $s_0 - 1 = tC/r$ and $\widetilde{r}/r = 1 + O(\chi^2)$, integration in z yields

$$\Psi(A^+, s_0\widetilde{H}) - \Psi(A^+, \widetilde{H}) \geq (1 - C\chi)\frac{3tC\mathcal{D}}{r\beta}. \tag{32}$$

Combining equation 31 and equation 32 with Lemma A.2 proves equation 28. □

Proposition B.2 (Safe fixed steps and discrete foundational convergence)**.** *There exists $\eta_* > 0$, depending only on the low-loss checkpoint and the finite data set, such that every fixed step $0 < \eta \le \eta_*$ is safe at every iteration. Along the resulting ordinary gradient-descent sequence,*

$$\mathcal{L}_k \downarrow 0, \qquad r_k \uparrow \infty, \qquad \Psi_k \uparrow \gamma_\infty > 0, \tag{33}$$

$$\sum_k \eta \|g_k\|^2 < \infty, \qquad \sum_k \frac{\|A_{k+1} - A_k\|^2}{r_{k+1} - r_k} < \infty, \tag{34}$$

and every conclusion of Theorem 4.2 holds.

Proof. Assume inductively that a finite prefix is safe. Equations 17 and 25 give $r_k \ge r_0$ and $\Psi_k \ge \Psi_0$. Moreover,

$$\frac{\|A_{k+1} - A_k\|^2}{r_{k+1} - r_k} \le \frac{\Psi_{k+1} - \Psi_k}{2\sigma_\Psi \Psi_0}.$$

Thus every safe finite prefix satisfies

$$\sum_{j<k} \frac{\|A_{j+1} - A_j\|^2}{r_{j+1} - r_j} \le E_0 := \frac{C_X - \Psi_0}{2\sigma_\Psi \Psi_0}.$$

Cauchy–Schwarz then gives the prefix-uniform bound

$$\|A_k\| \le \|A_0\| + \sqrt{E_0(r_k - r_0)} \le K_A \sqrt{r_k}, \quad K_A := \frac{\|A_0\|}{\sqrt{r_0}} + \sqrt{E_0}.$$

Hence every endpoint of a safe prefix lies in one fixed cone of the form equation 26. On that cone,

$$\mathcal{L}_k (1 + r_k)^4 \le \frac{1}{n} e^{-\Psi_0 r_k^{3/2}} (1 + r_k)^4 \le M_* < \infty.$$

Choose η_* so that $\eta_* M_* \le \delta_*$ and $C_* \eta_* M_* \le \min(1 - \sigma_L, 1 - \sigma_\Psi)$. Lemma B.1 makes the next step safe. Induction proves safety for the entire ordinary fixed-step trajectory, without assuming any of its asymptotic properties.

Summing equation 24 proves the first energy estimate in equation 34; the displayed prefix estimate proves the second. If $\mathcal{L}_k$ had a positive limit, then α_k and, because $\Psi_k \ge \Psi_0$, also r_k would be bounded. But equation 17 would then give $r_{k+1} - r_k \ge c\eta > 0$, a contradiction. This proves equation 33.

For each Gram difference in Lemma A.3, the linear terms cancel and the one-step drift is quadratic. Therefore

$$\sum_k \|\Delta B_k\| \le \eta^2 \sum_k \|g_k\|^2 < \infty.$$

The Gram differences converge to finite matrices, giving the balance limits and $\rho_k / r_k^{3/2} \to 1$. The bounded gap between q_k and α_k gives $q_k / \rho_k \to \gamma_\infty$. Finally, equation 34 and Cauchy–Schwarz in the radius clock give $A_k = o(\sqrt{r_k})$ and $\|Q_k K_k^\top\|_F = o(r_k)$. □

Lemma B.3 (Exact normalized GD update)**.** *Let $\varepsilon_k = \eta \lambda_k \sqrt{r_k}$ and $\Delta\tau_k = \log(r_{k+1}/r_k)$. Then*

$$A_{k+1} = A_k + \varepsilon_k r_k g_{A,k}, \qquad H_{k+1} = \sqrt{r_k}(\xi_k + \varepsilon_k d_k), \tag{35}$$

$$\frac{r_{k+1}}{r_k} = 1 + 2\varepsilon_k c_k + \frac{\varepsilon_k^2}{3}\|d_k\|^2, \qquad \|\xi_{k+1} - \xi_k\| \le \varepsilon_k \|g_{\perp,k}\|. \tag{36}$$

Furthermore, $\varepsilon_k \to 0$, $\Delta\tau_k \asymp \varepsilon_k$, and $\sum_k \varepsilon_k = \infty$.

Proof. Substitution of $\nabla_H \alpha = rd$ and $\nabla_A \alpha = \beta g_A$ into the update proves equation 35; expanding the head norm proves the first identity in equation 36. Since $d_k = c_k \xi_k + g_{\perp,k}$ with orthogonal summands, exact normalization gives

$$\xi_{k+1} = \frac{\xi_k + a_k g_{\perp,k}}{\sqrt{1 + a_k^2 \|g_{\perp,k}\|^2 / 3}}, \qquad a_k = \frac{\varepsilon_k}{1 + \varepsilon_k c_k},$$

which implies the stated distance bound. Finally, $\lambda_k \le n^{-1} e^{-\alpha_k}$ and $\alpha_k \ge \Psi_0 r_k^{3/2}$ imply $\varepsilon_k \to 0$. The positive upper and lower bounds on c_k, together with bounded d_k, make $\Delta\tau_k \asymp \varepsilon_k$. Since $r_k \to \infty$, $\sum_k \Delta\tau_k = \infty$, and hence $\sum_k \varepsilon_k = \infty$. □

Proof of Theorem 4.2. Proposition B.2 constructs a single positive threshold η_* for which every iterate is safe. Hence equation 25 proves monotonicity of Ψ_k, while the same proposition proves loss convergence, radius divergence, balance, the two margin limits, and the relative attention-scale estimates. This proves every assertion of the theorem for an ordinary fixed step. □

C AUTOMATIC CONTROL OF THE ATTENTION SCALE

In this section, we strengthen $A = o(\sqrt{r})$ to a strict power bound, as needed for the compactification in Appendix D. The finite set of quadratic score differences in the exact softmax gradient provides this bound.

Lemma C.1 (Automatic sublinear attention growth)**.** *For either loss and both optimization schemes of Part I, there are $C < \infty$ and $\theta \in [1/2, 1)$ such that, for all sufficiently large r,*

$$1 + \|A(r)\|^2 \le C r^{\theta}(1 + \log r). \tag{37}$$

For gradient descent, $A(r)$ is the piecewise-affine interpolation in the output-head radius. Consequently, $\|QK^\top\|_F = O(r^{\theta}(1 + \log r)) = o(r)$.

We first prove the estimate for arbitrary finite quadratic channels, then apply it to the score differences. The argument does not assume an attention growth rate.

Lemma C.2 (Controlled quadratic channels)**.** *Let $\ell_1, \dots, \ell_N$ be fixed homogeneous quadratic forms on $\mathbb{R}^m$, and fix $r_0 \ge 1$. Suppose that a locally absolutely continuous curve $A : [r_0, \infty) \to \mathbb{R}^m$ satisfies*

$$\int_{r_0}^{\infty} \|A'(r)\|^2 dr < \infty$$

and, almost everywhere,

$$A' = \sum_{j \in S(r)} a_j(r) \nabla \ell_j(A) + R(r), \tag{38}$$

where $S(r) \subseteq [N]$ may vary, $|a_j(r)| \le C_0$,

$$|\ell_j(A(r))| \le C_0(1 + \log r) \quad (j \in S(r)), \qquad \|R(r)\| \le \mu(r)\|A(r)\|,$$

and $0 \le \mu \in L^1([r_0, \infty)) \cap L^\infty([r_0, \infty))$. Then for some $\theta \in [1/2, 1)$,

$$1 + \|A(r)\|^2 = O\big(r^{\theta}(1 + \log r)\big)\,. \tag{39}$$

The conclusion is uniform over arbitrary switching among the finitely many subsets $S(r)$; no derivative or convergence of the coefficients is required.

Proof. Adjoin a constant coordinate $X = (A, 1)$, put $s = \|X\|^2$ and $v = X/\sqrt{s}$, and extend the quadratic forms to ignore the last coordinate. For each fixed subset S, define

$$f_S(v, a) = \sum_{j \in S} a_j \ell_j(v).$$

Uniformly for $\|v\| = 1$ and bounded a, there are $C < \infty$ and $\theta \in [1/2, 1)$ such that

$$|f_S(v, a)|^{\theta} \le C\left(\|\nabla_v^{\mathrm{sph}} f_S(v, a)\| + \|(\ell_j(v))_{j \in S}\|\right). \tag{40}$$

We use the polynomial gradient inequality of D'Acunto & Kurdyka (2005, Main Theorem): if a polynomial P of degree $d \ge 2$ in M variables satisfies $P(x_0) = 0$ and $\nabla P(x_0) = 0$, then, near x_0,

$$|P(x)|^{\vartheta} \le C\|\nabla P(x)\|, \qquad \vartheta = 1 - \frac{1}{d(3d-3)^{M-1}} < 1.$$

The critical point need not be isolated. Apply this result to

$$P_S(v, a, t) = f_S(v, a) + t(\|v\|^2 - 1)$$

and the compact graph

$$K_S = \{(v, a, -f_S(v, a)) : \|v\| = 1,\ |a_j| \le C_0\}.$$

Every critical point of P_S has value zero: its a-derivatives force $\ell_j(v) = 0$ and its t-derivative forces $\|v\| = 1$. On K_S, quadratic homogeneity gives

$$P_S = f_S, \qquad \nabla P_S = (\nabla_v^{\mathrm{sph}} f_S, (\ell_j(v))_{j\in S}, 0).$$

A finite collection of local gradient inequalities covers the critical points in K_S. On the compact complement of those neighborhoods, $\|\nabla P_S\|$ has a positive minimum and $|P_S|$ is bounded, so enlarging the constant extends the estimate to all of K_S. Taking a common exponent in $[1/2, 1)$ and a common constant over the finitely many subsets proves equation 40. The coefficients are variables in this inequality; their derivatives along the curve do not enter the argument. In particular, the polynomial P_S has degree three in $m + |S| + 2$ variables. The external theorem therefore provides an exponent depending only on these dimensions, before the finite maximum over S is taken.

Let $V_{\mathrm{low}} = \sum_{j\in S} a_j \nabla \ell_j(A)$ and $E = (R, 0)$. Quadratic homogeneity gives

$$f_S(v, a) = \frac{s'_{\mathrm{low}}}{4s}, \quad \nabla_v^{\mathrm{sph}} f_S = \frac{P_{v^\perp}(V_{\mathrm{low}}, 0)}{\sqrt{s}}, \quad \|(\ell_j(v))_{j\in S}\| \le \frac{C(1 + \log r)}{s}.$$

Writing $p = 1/\theta \in (1, 2]$, $h = \|A'\|$, and $L(r) = C(1 + \log r)$, inequality equation 40, together with $\|E\| \le \mu\sqrt{s}$, first yields

$$s' \le C\left[s^{1-p/2}h^p + L^p s^{1-p} + (\mu + \mu^p)s\right].$$

Since μ is bounded, $\mu^p \le \|\mu\|_\infty^{p-1}\mu$. Absorbing this factor into the constant gives

$$s' \le C\left[s^{1-p/2}h^p + L^p s^{1-p} + \mu s\right]. \tag{41}$$

For $Y = s^{p/2}$, this becomes

$$Y' \le C\left[h^p + \frac{L^p}{Y} + \mu Y\right].$$

Put $M(r) = C\int_{r_0}^r \mu(u)\,du$ and $Z = e^{-M}Y$. Since M is bounded, $Z' \le C_1 h^p + C_1 L^p/Z$ after enlarging C_1. A positive supersolution is

$$W(r) = Z(r_0) + C_1\int_{r_0}^r h(u)^p\,du + \left(1 + 2C_1\int_{r_0}^r L(u)^p\,du\right)^{1/2}.$$

Indeed, $W' \ge C_1 h^p + C_1 L^p/W$ and $W(r_0) > Z(r_0)$. Scalar comparison and boundedness of e^M give

$$Y(r) \le C\left(1 + \int_{r_0}^r h^p ds + \sqrt{\int_{r_0}^r L(s)^p ds}\right).$$

Hölder's inequality and the finite L^2 energy imply

$$\int_{r_0}^r h^p ds = O(r^{1-p/2}), \qquad \sqrt{\int_{r_0}^r L(s)^p ds} = O(\sqrt{r}(1 + \log r)^{p/2}).$$

The second term dominates. Raising the result to the power $2/p$ proves equation 39. □

Lemma C.3 (Exact softmax channel decomposition)**.** *For each i and $j < k$, define the logit gap and normalized reward gap*

$$\ell_{ijk}(A) = \frac{q_i^\top Q K^\top (x_{ij} - x_{ik})}{\sqrt{d}}, \qquad \Delta b_{ijk} = y_i(\xi_V \xi_W \xi_a)^\top (x_{ij} - x_{ik}).$$

Then the normalized attention gradient has the exact representation

$$g_A = \sum_i \omega_i \sum_{j<k} p_{ij} p_{ik} \Delta b_{ijk} \nabla_A \ell_{ijk}(A). \tag{42}$$

Proof. For a probability vector $p = \mathrm{softmax}(z)$ and values b_j, the covariance identity

$$\nabla \sum_j p_j b_j = \sum_{j<k} p_j p_k (b_j - b_k) \nabla (z_j - z_k)$$

follows by expanding $\mathrm{diag}(p) - pp^\top$. Apply it to each normalized signed margin F_i and then sum with the exact weights ω_i of Lemma A.1. □

Proof of Lemma C.1. We first treat gradient flow. Proposition A.4 gives

$$A_r = \frac{g_A}{2c}, \qquad 0 < c_- \le c \le c_+ < \infty, \qquad \int^\infty \|A_r\|^2 dr < \infty.$$

At radius r, retain a token pair when $p_{ij}p_{ik} \ge r^{-2}$. For every retained pair,

$$|\ell_{ijk}(A)| = \left|\log \frac{p_{ij}}{p_{ik}}\right| \le -\log(p_{ij}p_{ik}) \le 2\log r. \tag{43}$$

The coefficients of the retained quadratic gradients in $A_r = g_A/(2c)$ are bounded: ω_i and $1/c$ are bounded, $p_{ij}p_{ik} \le 1$, and Δb_{ijk} is bounded because $\|\xi\|^2 = 3$ and the data set is finite. For discarded pairs, $p_{ij}p_{ik} < r^{-2}$, while $\|\nabla \ell_{ijk}(A)\| \le C\|A\|$. Their total contribution is therefore a remainder satisfying

$$\|R(r)\| \le C r^{-2} \|A(r)\|.$$

All hypotheses of Lemma C.2 now hold, which proves equation 37. The product estimate follows from $\|QK^\top\|_F \le \|A\|^2/2$.

For gradient descent, interpolate A_k affinely as a function of r on $[r_k, r_{k+1}]$. The exact update

$$r_{k+1} - r_k = 2c_k t_k + \frac{t_k^2}{3r_k}\|d_k\|^2, \qquad t_k = \eta\lambda_k\beta_k,$$

and the discrete attention-energy inequality give

$$\int^\infty \|A'_{\mathrm{int}}(r)\|^2 dr = \sum_k \frac{\|A_{k+1} - A_k\|^2}{r_{k+1} - r_k} < \infty.$$

On $[r_k, r_{k+1}]$, retain the pairs with $p_{ij,k}p_{ih,k} \ge r_k^{-2}$. Their coefficients have the extra factor $t_k/(r_{k+1} - r_k) \le 1/(2c_k)$ and are therefore bounded. Since $\|\Delta A_k\| \le C t_k \|A_k\|$ and $t_k \to 0$, $\|A_{\mathrm{int}}(r)\| \asymp \|A_k\|$ on a sufficiently late tail. Quadraticity allows gradients at A_k to be translated to the interpolated point with remainder

$$\|R_{\mathrm{int}}(r)\| \le C(r^{-2} + t_k)\|A_{\mathrm{int}}(r)\|.$$

The normalized-margin lower bound gives

$$t_k \le C\eta r_k^{3/2} \exp(-\Psi_0 r_k^{3/2}) \longrightarrow 0.$$

We also need integrability in the radius clock rather than only pointwise smallness. The safe loss inequality and the polynomial gradient bounds give

$$(r_{k+1} - r_k)^2 \le C\eta r_k \big(\mathcal{L}_k - \mathcal{L}_{k+1}\big).$$

The margin bound gives $r_k \le C[1 + \log(1/\mathcal{L}_k)]^{2/3}$. The function $F(x) = [1 + \log(1/x)]^{2/3}$ is decreasing on $(0, \mathcal{L}_0]$, so

$$\sum_k r_k(\mathcal{L}_k - \mathcal{L}_{k+1}) \le C \sum_k \int_{\mathcal{L}_{k+1}}^{\mathcal{L}_k} F(x)\,dx \le C \int_0^{\mathcal{L}_0} F(x)\,dx < \infty.$$

Thus $\sum_k (r_{k+1} - r_k)^2 < \infty$. The exact radius identity and the positive lower bound for c_k then imply

$$\sum_k t_k(r_{k+1} - r_k) \le C \sum_k (r_{k+1} - r_k)^2 < \infty.$$

Thus the remainder coefficient is both bounded and integrable after interpolation. Finally, every logit gap is quadratic in A, and on the (k)-th interval its change is bounded by

$$C\|A_k\|\|A_{k+1} - A_k\| + C\|A_{k+1} - A_k\|^2 \le C\|A_k\|^2 t_k(1 + t_k) = o(1).$$

Consequently equation 43 remains valid up to a fixed additive constant. Lemma C.2 applies to the interpolation and proves the discrete claim. □

D FINITE LENGTH AND CONVERGENCE OF THE NORMALIZED HEAD

In this section, we use the strict attention growth bound to prove that the normalized head has finite path length. The weighted attention term enters the same argument, so we can obtain both continuous and discrete directional convergence without assuming that attention itself converges.

Lemma D.1 (Joint finite length)**.** *Under the conditions of Part I,*

$$\int^{\infty} \left(\left\| \frac{d\bar{H}}{d\tau} \right\| + \frac{1}{\sqrt{r}} \left\| \frac{dA}{d\tau} \right\| \right) d\tau < \infty, \tag{44}$$

with the analogous summable increments for gradient descent.

The next compactification makes the gradient inequality applicable on a bounded open domain. Its derivative bound is where $\theta < 1$ is used.

Lemma D.2 (A compactification adapted to the trajectory)**.** *There is a definable C^1 diffeomorphism $\chi : \mathbb{R}^{\dim A} \to \{Y : \|Y\| < 1\}$ and a constant $c_\chi > 0$ such that, along the trajectory for all sufficiently large r,*

$$r\, s_{\min}(D\chi(A(r)))^2 \geq c_\chi. \tag{45}$$

Proof. By Lemma C.1, choose a rational $a \in (\theta/2, 1/2)$ and C such that $\|A(r)\| \leq Cr^a$ eventually. Put $\nu = 1/(2a) > 1$ and

$$h(R) = 1 - (1+R)^{1-\nu}, \qquad \chi(A) = \begin{cases} h(\|A\|)A/\|A\|, & A \neq 0, \\ 0, & A = 0. \end{cases}$$

The formula extends differentiably through the origin, with derivative $(\nu - 1)I$, and is a definable radial diffeomorphism onto the open unit ball. Its radial singular value is $h'(R) = (\nu-1)(1+R)^{-\nu}$. Concavity of h and $h(0) = 0$ imply $h(R)/R \geq h'(R)$, so the tangential singular values are no smaller. Consequently,

$$r\, s_{\min}(D\chi(A(r)))^2 \geq (\nu-1)^2 r(1+Cr^a)^{-2\nu} \geq c_\chi > 0,$$

where $2a\nu = 1$. □

Let $Y = \chi(A)$ and

$$\widetilde{\mathscr{F}}(Y, \xi, z) = \mathscr{F}(\chi^{-1}(Y), \xi, z).$$

The function is definable in the o-minimal structure containing the real exponential. No extension to $\|Y\| = 1$ or $z = 0$ is required.

Lemma D.3 (Desingularized tail estimate)**.** *There are $a_0 > 0$ and a continuous, increasing, concave function $\varphi : [0, a_0) \to \mathbb{R}_+$, C^1 on $(0, a_0)$ with $\varphi(0) = 0$, such that at every sufficiently late trajectory point with $\Psi < \gamma_\infty$,*

$$\varphi'(\gamma_\infty - \Psi) \left(\|\nabla_Y \widetilde{\mathscr{F}}\|^2 + \|g_\perp\|^2 + |\partial_z \mathscr{F}|^2 \right)^{1/2} \geq 1. \tag{46}$$

Proof. We use Kurdyka (1998, Theorem 1) in the following form. If U is a bounded open definable set and $f : U \to (0, \infty)$ is C^1 and definable, then there exist $b, \delta_0 > 0$ and a positive, strictly increasing, definable C^1 function ψ such that

$$\psi'(f(x))\|\nabla f(x)\| \geq b \quad \text{for } x \in U, \quad 0 < f(x) < \delta_0.$$

This statement concerns f on U; continuity on $\overline{U}$ is not a hypothesis.

Put $f(Y, x, z) = \gamma_\infty - \widetilde{\mathscr{F}}(Y, \sqrt{3}x/\|x\|, z)$ and choose $z_0, \delta > 0$. Apply the theorem on

$$U = \{(Y, x, z) : \|Y\| < 1, 1 < \|x\| < 2,\ 0 < z < z_0,\ 0 < f(Y, x, z) < \delta\}.$$

The domain is bounded, open, and definable in $\mathbb{R}_{\exp}$. The function is positive and C^1 there: χ^{-1} is C^1 on the open ball, and the soft minimum is smooth for $z > 0$. Every sufficiently late trajectory point with $\Psi < \gamma_\infty$ has a lift to U at any fixed $\|x\| \in (1, 2)$. At such a lift, the chain rule gives

$$\nabla_Y f = -\nabla_Y \widetilde{\mathscr{F}}, \qquad \nabla_x f = -\frac{\sqrt{3}}{\|x\|} g_\perp, \qquad \partial_z f = -\partial_z \mathscr{F}.$$

Thus $\|\nabla f\|$ is bounded above and below by fixed positive multiples of the norm in equation 46. Restricting the theorem to these points and rescaling ψ gives the claimed lower bound. This step uses definability of the function and domain, not of the training trajectory.

To obtain concavity, shrink the interval until ψ' is monotone, as permitted by o-minimal monotonicity. If ψ' is nonincreasing, ψ is already concave. Otherwise ψ' is bounded above near zero; replacing $\psi(t)$ by Ct with C at least that bound preserves the inequality and gives a concave function. Positivity and monotonicity give a finite limit at zero. Subtracting this limit and extending by zero yields φ. □

Proof of Lemma D.1. The chain rule and equation 45 imply

$$\|\nabla_Y \widetilde{\mathscr{F}}\|^2 \le Cr\|g_A\|^2.$$

Moreover, $\partial_z \mathscr{F} = -3z^2\mathcal{D}$, and $\mathcal{D}$ is uniformly bounded. Since $z \to 0$, $z^4\mathcal{D}^2 \le Cz^3\mathcal{D}$ along the tail. Hence, with

$$\mathcal{E} = \|g_\perp\|^2 + r\|g_A\|^2 + 3cz^3\mathcal{D},$$

the norm in equation 46 is at most $C\sqrt{\mathcal{E}}$. Equation 20 is exactly $\Psi_\tau = \mathcal{E}/(2c)$. Therefore

$$-\frac{d}{d\tau}\varphi(\gamma_\infty - \Psi) = \varphi'(\gamma_\infty - \Psi)\Psi_\tau \ge c_0\sqrt{\mathcal{E}}.$$

Integration gives

$$\int^\infty \left(\|g_\perp\| + \sqrt{r}\|g_A\|\right) d\tau < \infty.$$

Using equation 19 gives equation 44. If Ψ attains γ_∞ at finite time, monotonicity and equation 20 make $\mathcal{E} = 0$ on the remaining tail, so the same conclusion follows directly.

For gradient descent, define

$$\mathcal{E}_k = \|g_{\perp,k}\|^2 + r_k\|g_{A,k}\|^2 + 3c_k z_k^3 \mathcal{D}_k.$$

The exact identity

$$-\langle \nabla\Psi_k, \nabla\mathcal{L}_k\rangle = \lambda_k\sqrt{r_k}\,\mathcal{E}_k$$

and the safe margin inequality equation 25 give

$$\Psi_{k+1} - \Psi_k \ge \sigma_\Psi \varepsilon_k \mathcal{E}_k, \qquad \sum_k \varepsilon_k \mathcal{E}_k < \infty. \tag{47}$$

Apply Lemma D.3 at the iterates. The compactifier bound and $z_k^4\mathcal{D}_k^2 \le Cz_k^3\mathcal{D}_k$ imply

$$1 \le C\varphi'(\gamma_\infty - \Psi_k)\sqrt{\mathcal{E}_k}$$

for all sufficiently large k. Since φ is concave and $\Psi_{k+1} \ge \Psi_k$,

$$\begin{aligned}\varphi(\gamma_\infty - \Psi_k) - \varphi(\gamma_\infty - \Psi_{k+1}) &\ge \varphi'(\gamma_\infty - \Psi_k)(\Psi_{k+1} - \Psi_k)\\ &\ge c\varepsilon_k\sqrt{\mathcal{E}_k}.\end{aligned}$$

Here concavity gives the first inequality by applying the tangent-line bound at $\gamma_\infty - \Psi_k$ to the smaller argument $\gamma_\infty - \Psi_{k+1}$. Telescoping therefore proves

$$\sum_{k\ge K} \varepsilon_k\sqrt{\mathcal{E}_k} \le C\varphi(\gamma_\infty - \Psi_K) < \infty. \tag{48}$$

For completeness, the normalization in Lemma B.3 has $a_k = \varepsilon_k/(1 + \varepsilon_k c_k) \le \varepsilon_k$. Orthogonality of ξ_k and $g_{\perp,k}$ gives

$$\|\xi_{k+1} - \xi_k\|^2 = 6\left(1 - \frac{1}{\sqrt{1 + a_k^2\|g_{\perp,k}\|^2/3}}\right) \le a_k^2\|g_{\perp,k}\|^2.$$

Together with the exact attention update, this yields

$$\|\xi_{k+1} - \xi_k\| \le \varepsilon_k\|g_{\perp,k}\|, \qquad \frac{\|A_{k+1} - A_k\|}{\sqrt{r_k}} = \varepsilon_k\sqrt{r_k}\|g_{A,k}\|.$$

Both terms are bounded by $\varepsilon_k\sqrt{\mathcal{E}_k}$. Summation of equation 48 yields

$$\sum_k \left(\|\xi_{k+1} - \xi_k\| + \frac{\|A_{k+1} - A_k\|}{\sqrt{r_k}} \right) < \infty,$$

which is the discrete form of the claim. If Ψ_k reaches its limit at a finite index, equation 47 makes $\mathcal{E}_k = 0$ thereafter and the conclusion is immediate. □

Proposition D.4 (Directional convergence)**.** *The limits of $H/\sqrt{r}$, $VWa/\|VWa\|$, VWa/q, and $H/q^{1/3}$ in part 1 of Theorem 4.4 hold. Appendix F supplies the time normalizations.*

Proof. Lemma D.1 makes $\xi = H/\sqrt{r}$ a Cauchy curve, so $\xi \to \bar{H}_\infty$. Lemma A.3 forces every limiting balanced head to have the rank-one form

$$\bar{H}_\infty = (u_\infty b_\infty^\top,\, b_\infty c_\infty^\top,\, c_\infty)$$

for unit vectors $u_\infty, b_\infty, c_\infty$. Thus $VWa/\|VWa\| \to u_\infty$. Finally, $q/\rho \to \gamma_\infty$ by equation 13 and equation 21, so

$$\frac{VWa}{q} \to \frac{u_\infty}{\gamma_\infty}, \qquad \frac{H}{q^{1/3}} \to \gamma_\infty^{-1/3}\bar{H}_\infty.$$

□

E Maximum margin on asymptotic features

In this section, we identify the limit of the effective classifier. Attention may have several accumulation points, so we take one sequence on which both the signed features and the loss weights converge. The resulting certificate proves optimality for the SVM constrained by every asymptotic feature, not only those on that sequence.

Put $s_i = y_i h_i(A)$ and $b = \sum_i \pi_i s_i$. Recall that $\pi_i = \omega_i/\kappa$ is a probability distribution for both losses.

Lemma E.1 (A common limiting certificate)**.** *There is a common sequence tending to infinity along which*

$$s_i \to s_i^*, \qquad \pi_i \to \pi_i^*, \qquad i \in [n],$$

and

$$\sum_i \pi_i^* s_i^* = \gamma_\infty u_\infty, \qquad \sum_i \pi_i^* = 1, \qquad \pi_i^* > 0 \implies (s_i^*)^\top u_\infty = \gamma_\infty. \tag{49}$$

Proof. The continuous clock $\tau = \log r$ is unbounded, and in GD $\sum_k \varepsilon_k = \infty$ by Lemma B.3. The finite-length estimate therefore supplies a common sequence on which $g_\perp \to 0$ and $\sqrt{r}\, g_A \to 0$. Compactness of the token convex hulls and of the probability simplex then gives a further common subsequence on which all s_i and π_i converge.

To identify the limit of b, let $j = \xi_W \xi_a$. Differentiation of the normalized margins in the V block gives

$$(g_\perp)_V = \kappa b j^\top - c\xi_V.$$

Projecting after right multiplication by j onto $u^\perp$ yields

$$(I - uu^\top)(g_\perp)_V j = \kappa \|j\|^2 (I - uu^\top) b.$$

Balance bounds $\|j\|$ away from zero. Hence $g_\perp \to 0$ implies that every limit of b is parallel to u_∞. Moreover,

$$u^\top b = \frac{C}{\kappa\rho} = \frac{\alpha + \mathcal{D}}{\kappa\rho} \longrightarrow \gamma_\infty,$$

since $\kappa \to 1$, $\mathcal{D}$ is bounded, and $\alpha/\rho \to \gamma_\infty$. Hence the limit of b is $\gamma_\infty u_\infty$.

Every feature limit satisfies $u_\infty^\top s_i^* \geq \gamma_\infty$ because q/ρ tends to γ_∞. If $\pi_i^* > 0$ and the inequality were strict, then $m_i - q$ would grow linearly in ρ. Choose any minimum-margin index j at each point of the sequence. For exponential loss, $\pi_i/\pi_j = e^{-(m_i - q)}$; for logistic loss,

$$\frac{\pi_i}{\pi_j} = \frac{1 + e^q}{1 + e^{m_i}} \leq 2e^{-(m_i - q)}.$$

Since $\pi_j \leq 1$, either bound forces $\pi_i \to 0$. This proves the last implication in equation 49. □

Proof of part 2 of Theorem 4.4. Every $z \in \mathcal{Z}_\infty$ obeys $u_\infty^\top z \geq \gamma_\infty$, so $w_\infty = u_\infty/\gamma_\infty$ is feasible for equation 3. For any other feasible w, Lemma E.1 gives

$$\gamma_\infty u_\infty^\top w = \sum_i \pi_i^* (s_i^*)^\top w \geq 1.$$

Thus $\|w\| \geq 1/\gamma_\infty = \|w_\infty\|$. Strict convexity makes w_∞ the unique optimizer. The multipliers π_i^*/γ_∞^2 give the corresponding finite-support KKT certificate. □

Completion of the proof of Theorem 4.4. Part 1 follows from Proposition D.4, the time-radius laws in the proof of Theorem 4.3, and $q/r^{3/2} \to \gamma_\infty$. Part 2 follows from Lemma E.1 and strict convexity of the asymptotic-feature SVM. The conclusions cover gradient flow and fixed-step gradient descent without requiring attention convergence. □

F TIME SCALES AND CONVERGENCE RATES

In this section, we convert the head-radius clock used in the proofs into physical training time. The same calculation gives the loss, margin, and parameter growth rates in Theorems 4.3 and 4.4.

Proof of Theorem 4.3. For both losses,

$$\lambda^{-1} = ne^{\alpha}(1 + o(1)), \qquad \mathcal{L} = n^{-1}e^{-\alpha}(1 + o(1)), \qquad c \to \gamma_\infty.$$

Under gradient flow, $\dot{r} = 2\lambda c r^{3/2}$, whence

$$\frac{dt}{dr} = \frac{ne^{\alpha(r)}}{2c(r)r^{3/2}}(1 + o(1)).$$

Since $\alpha/r^{3/2} = \Psi \to \gamma_\infty$, integration of upper and lower exponential bounds gives

$$\log t \sim \gamma_\infty r^{3/2}, \qquad r \sim \gamma_\infty^{-2/3}(\log t)^{2/3}. \tag{50}$$

Together with equation 21 and equation 23, this proves

$$\|\Theta(t)\| \sim \sqrt{3}\,\gamma_\infty^{-1/3}(\log t)^{1/3}.$$

For the loss bounds, let

$$\mathcal{P} = \frac{\|\nabla\alpha\|^2}{r^2} = 3c^2 + \|g_\perp\|^2 + r\|g_A\|^2.$$

Finite data and $\|A\| = o(\sqrt{r})$ give $3c^2 \leq \mathcal{P} \leq Cr^2$. Direct differentiation gives

$$\frac{d}{dt}\frac{1}{\mathcal{L}} = r^2\mathcal{P}(1 + o(1)).$$

Equation 50 also proves the required slow variation. For every fixed $b \in (0, 1)$, $r(s)/r(t) \to 1$ uniformly for $bt \leq s \leq t$: both logarithms differ from $\log t$ by at most the constant $|\log b|$. Monotonicity of r therefore gives, for every fixed $p > 0$,

$$(1 - b)t\, r(bt)^p \leq \int_{bt}^{t} r(s)^p ds \leq t\, r(t)^p.$$

Divide by $tr(t)^p$, let $t \to \infty$, and then let $b \downarrow 0$ to obtain $\int_0^t r(s)^p ds \sim tr(t)^p$. Integrating the preceding two-sided bounds now yields

$$\frac{c_1}{tr(t)^4} \leq \mathcal{L}(t) \leq \frac{c_2}{tr(t)^2}.$$

Substitution of equation 50 proves the two exponents in Theorem 4.3.

For fixed-step gradient descent, put $T_k = 1 + \eta k$ and $\vartheta_k = \lambda_k/\mathcal{L}_k$. The derivative bounds in Lemma B.1, Taylor's theorem, and $\eta\mathcal{L}_k(1 + r_k)^4 \to 0$ give the uniform one-step expansions

$$\mathcal{L}_{k+1} = \mathcal{L}_k - \eta\|\nabla\mathcal{L}_k\|^2(1 + o(1)), \tag{51}$$

$$\frac{1}{\mathcal{L}_{k+1}} - \frac{1}{\mathcal{L}_k} = \eta\vartheta_k^2 r_k^2 \mathcal{P}_k(1 + o(1)), \tag{52}$$

$$r_{k+1} - r_k = 2\eta\lambda_k c_k r_k^{3/2}(1 + o(1)). \tag{53}$$

Here $\vartheta_k \to 1$. To justify the second line, the first line implies $\mathcal{L}_{k+1}/\mathcal{L}_k \to 1$, and $\|\nabla \mathcal{L}_k\|^2 = \lambda_k^2 r_k^2 \mathcal{P}_k$. For the last line, use the exact radius identity equation 17; the ratio of its quadratic term to its positive linear term is $O(\varepsilon_k) = o(1)$ by Lemma B.3.

Since $\lambda_k^{-1} = ne^{\alpha_k}(1 + o(1))$, equation 53 is equivalent to

$$\eta = \frac{ne^{\alpha_k}}{2c_k r_k^{3/2}}(1 + o(1))(r_{k+1} - r_k). \tag{54}$$

Fix $\epsilon > 0$. Eventually $(\gamma_\infty - \epsilon) r_k^{3/2} \le \alpha_k \le (\gamma_\infty + \epsilon) r_k^{3/2}$. Summing the upper bound from equation 54 up to k gives

$$\limsup_{k\to\infty} \frac{\log T_k}{r_k^{3/2}} \le \gamma_\infty + \epsilon.$$

For the reverse bound, retain only indices whose initial radius lies in $[(1-b)r_k, r_k)$ for fixed $b \in (0,1)$. Since $(r_{j+1} - r_j)/r_j \to 0$, the sum of their radius increments is at least $br_k/2$ eventually, whereas their exponential factor is at least $\exp((\gamma_\infty - \epsilon)(1-b)^{3/2} r_k^{3/2})$. Thus

$$\liminf_{k\to\infty} \frac{\log T_k}{r_k^{3/2}} \ge (\gamma_\infty - \epsilon)(1-b)^{3/2}.$$

Letting $\epsilon, b \downarrow 0$ proves

$$\log T_k \sim \gamma_\infty r_k^{3/2}, \qquad r_k \sim \gamma_\infty^{-2/3} (\log T_k)^{2/3}. \tag{55}$$

The discrete slow-variation estimate used below is

$$\sum_{j<k} \eta r_j^p \sim T_k r_k^p, \qquad p > 0. \tag{56}$$

The upper bound follows from monotonicity. On the terminal interval $bT_k \le T_j < T_k$, equation 55 gives $r_j/r_k \to 1$ uniformly; this interval has total step length $(1-b)T_k + O(\eta)$. This proves the matching lower bound after $b \downarrow 0$.

Finally, $3c_k^2 \le \mathcal{P}_k \le C r_k^2$ and equation 52–equation 56 give

$$cT_k r_k^2 \le \frac{1}{\mathcal{L}_k} \le CT_k r_k^4$$

for all sufficiently large k. Inverting these inequalities and using equation 55 proves the two loss bounds in Theorem 4.3. The balance limit and $A_k = o(\sqrt{r_k})$ give

$$\|\Theta_k\| \sim \sqrt{3r_k} \sim \sqrt{3}\, \gamma_\infty^{-1/3} (\log T_k)^{1/3},$$

which completes the discrete proof. □

G Proof of strict token selection

In this section, we prove strict selection for the five-factor flow and simultaneous fixed-step GD. A finite covariance certificate and a shared attention drive are the two main ingredients. Throughout, R and C are the generator list and cone of Section 5; $r = (\|V\|_F^2 + \|W\|_F^2 + \|a\|^2)/3$.

G.1 Finite covariance-action certificate

We turn the finite witness condition (G3) into a covariance sign valid for every nonnegative attention score vector.

Definition G.1 (Covariance-action witness)**.** Use the full generator list R from Section 5, including every nonzero signed token across all samples and every nonzero column of V_0. For every $i \in [n]$, $k \in [d]$, $\alpha \in [N]$, and unordered pair $e = \{j, h\}$ with $1 \le j < h \le s$, define

$$b_{ie} = s_{ij} + s_{ih}, \qquad c_{ie}^{k\alpha} = (s_{ij,k} - s_{ih,k}) g_\alpha^\top (s_{ij} - s_{ih}). \tag{57}$$

For each fixed triple (i, k, α), let $\mathcal{P}$ and $\mathcal{N}$ be the pairs with strictly positive and strictly negative coefficients, respectively; zero coefficients are omitted. A witness for this triple consists of numbers $F_{ef} \geq 0$ for $e \in \mathcal{P}$, $f \in \mathcal{N}$, subject to $F_{ef} = 0$ unless $b_{ie} \geq b_{if}$ coordinatewise, and

$$\sum_{f \in \mathcal{N}} F_{ef} \leq c_{ie}^{k\alpha} \quad (e \in \mathcal{P}), \qquad \sum_{e \in \mathcal{P}} F_{ef} \geq -c_{if}^{k\alpha} \quad (f \in \mathcal{N}). \tag{58}$$

Check (G3) requires a witness for each of the ndN triples, with separate capacities and flows for each triple. An empty negative set needs no flow; an empty positive set with a nonempty negative set is infeasible. All coefficients and allowed edges are determined by the finite tokens and V_0. The witnesses are chosen once, independently of scoring vectors, time, loss weights, or any prospective trajectory bound.

We next show why these finite witnesses control every nonnegative scoring vector, rather than just the scores at initialization. For a scoring vector $\zeta \in \mathbb{R}^d$, define

$$\begin{aligned} Z_i(\zeta) &= \sum_{j=1}^{s} e^{s_{ij}^\top \zeta}, \qquad p_{ij}(\zeta) = \frac{e^{s_{ij}^\top \zeta}}{Z_i(\zeta)}, \\ \bar{s}_i(\zeta) &= \sum_j p_{ij}(\zeta) s_{ij}, \quad \Sigma_i(\zeta) = \sum_j p_{ij}(\zeta) s_{ij} s_{ij}^\top - \bar{s}_i(\zeta) \bar{s}_i(\zeta)^\top. \end{aligned} \tag{59}$$

Expanding the last product and collecting unordered pairs gives

$$\Sigma_i(\zeta) = \sum_{j<h} p_{ij}(\zeta) p_{ih}(\zeta) (s_{ij} - s_{ih})(s_{ij} - s_{ih})^\top. \tag{60}$$

Thus the coefficients in equation 57 obey the exact identity

$$\big(\Sigma_i(\zeta) g_\alpha\big)_k = \frac{\sum_e c_{ie}^{k\alpha} e^{b_{ie}^\top \zeta}}{Z_i(\zeta)^2}. \tag{61}$$

There is no factor $1/2$ because each pair is counted once.

Fix a triple (i, k, α) and a feasible flow in equation 58. For each negative node f, its incoming flow is positive. Multiply all its incoming flows by $(-c_{if}^{k\alpha})/(\sum_e F_{ef}) \leq 1$. This makes its demand an equality, without increasing any capacity usage or changing allowed edges. With the resulting flow, the numerator is exactly

$$\begin{aligned} \sum_e c_{ie}^{k\alpha} e^{b_{ie}^\top \zeta} = \sum_{e \in \mathcal{P}} \left(c_{ie}^{k\alpha} - \sum_{f \in \mathcal{N}} F_{ef} \right) e^{b_{ie}^\top \zeta} \\ + \sum_{e \in \mathcal{P},\, f \in \mathcal{N}} F_{ef} \left(e^{b_{ie}^\top \zeta} - e^{b_{if}^\top \zeta} \right) \geq 0 \qquad (\zeta \geq 0). \end{aligned} \tag{62}$$

The first sum is nonnegative by capacity feasibility. Each term of the second sum is nonnegative because an allowed edge has $b_{ie} - b_{if} \geq 0$, so its inner product with $\zeta \geq 0$ is nonnegative. Division by $Z_i^2 > 0$ and linearity in the generator give

$$\Sigma_i(\zeta) w \geq 0 \qquad (\zeta \geq 0,\ w \in C). \tag{63}$$

There are ndN problems, each with at most $\binom{s}{2}^2$ candidate edges. The construction uses no sampling over ζ; rational inputs and flows admit exact feasibility checks.

Remark G.2 (Scope of the certificate). The flow feasibility test is sufficient for $\Sigma_i(\zeta) C \subseteq \mathbb{R}^d_+$ on the nonnegative orthant. It does not require the covariance matrix itself to be entrywise nonnegative.

We also need a reward bound for every direction in the moving head cone, not just for its initial generators. The low-loss condition implies $v_0 \neq 0$: otherwise every initial margin is zero, giving $n\mathcal{L}(0) = n$ for exponential loss or $n \log 2$ for logistic loss, contrary to (G4). In particular $N \geq 1$. For any $w \in C$, choose a representation $w = \sum_\alpha a_\alpha \widehat{g}_\alpha$ with $a_\alpha \geq 0$ and $\widehat{g}_\alpha = g_\alpha / \|g_\alpha\|$. Pairwise nonnegative generator inner products imply

$$\|w\|^2 \geq \sum_\alpha a_\alpha^2, \qquad D_{ij}^\top w \geq \epsilon_{\mathrm{rew}} \sum_\alpha a_\alpha \geq \epsilon_{\mathrm{rew}} \|w\|. \tag{64}$$

The first inequality follows by expanding the squared norm; the last follows from the triangle inequality. The same expansion gives $w^\top \widetilde{w} \geq 0$ for all $w, \widetilde{w} \in C$. The cone is closed: for a convergent sequence in C, the first inequality bounds its nonnegative representation coefficients; a convergent subsequence of those coefficients represents the limit. These facts do not impose coordinatewise signs on w.

G.2 The exact five-block flow and strict selection

We use the covariance sign to keep the five factors in an invariant cone, then show that the resulting attention drive selects the prescribed tokens.

Lemma G.3 (Invariant cone and preserved rewards)**.** *Under the hypotheses of Theorem 5.1, on the maximal forward interval the factors K, Q, W, a are entrywise nonnegative and nondecreasing, every column of V remains in C, and $v = VWa \in C \setminus \{0\}$. Moreover, $\|v(t)\| \geq \|v(0)\| > 0$ and $D_{ij}^\top u \geq \epsilon_{\mathrm{rew}}$ for $u = v/\|v\|$.*

Proof of Lemma G.3. We start with the unnormalized factor equations. Set $M = KQ^\top$ and $\zeta_i = Mt_i/\sqrt{d}$. The signs cancel in the logits because $x_{ij} = y_i s_{ij}$ and $q_i = y_i t_i$. Thus the probabilities in the original model equal those in equation 59 at ζ_i, and $y_i h_i = \bar{s}_i$. Define

$$m_i = v^\top \bar{s}_i, \qquad \sigma = \frac{1}{n}\sum_i [-\ell'(m_i)], \qquad \pi_i = \frac{-\ell'(m_i)}{n\sigma}, \qquad b_{\mathrm{av}} = \sum_i \pi_i \bar{s}_i, \qquad v = VWa. \tag{65}$$

For either loss, $\sigma > 0$, the π_i are positive and sum to one, and $-\nabla\mathcal{L} = \sigma \sum_i \pi_i \nabla m_i$. For exponential loss $\sigma = \mathcal{L}$; for logistic loss, $\sigma = n^{-1}\sum_i (1 + e^{m_i})^{-1}$. In the notation of Lemma A.1, $\sigma = \lambda\kappa$ and $\pi_i = \omega_i/\kappa$ for both losses. Differentiating the softmax mean gives $\partial \bar{s}_i / \partial \zeta_i = \Sigma_i$. The chain rule, including both factors of M, now gives

$$\dot{V} = \sigma b_{\mathrm{av}} (Wa)^\top, \qquad \dot{W} = \sigma V^\top b_{\mathrm{av}} a^\top, \qquad \dot{a} = \sigma W^\top V^\top b_{\mathrm{av}}, \tag{66}$$

$$\dot{K} = \mathcal{B}Q, \qquad \dot{Q} = \mathcal{B}^\top K, \qquad \mathcal{B} = \frac{\sigma}{\sqrt{d}} \sum_i \pi_i \Sigma_i v t_i^\top. \tag{67}$$

Here $\mathcal{B}$ is a matrix, distinct from the attention tuple $A = (Q, K)$.

We prove cone invariance before normalizing the head. On the closed convex set where $K, Q, W, a \geq 0$ and all columns of V belong to C, we have $\zeta_i \geq 0$ and $v, b_{\mathrm{av}} \in C$. Acuteness gives $V^\top b_{\mathrm{av}} \geq 0$, while equation 63 gives $\mathcal{B} \geq 0$. Equations 66–67 therefore give nonnegative derivatives for K, Q, W, a, and each column derivative of V lies in C.

We establish invariance in finite nonnegative coordinates. Let R be the matrix with columns g_α, and fix nonnegative representations $s_{ij} = \mathsf{R} d_{ij}$, including $d_{ij} = 0$ for a zero token. Write $V_0 = \mathsf{R} B_0$ with $B_0 \geq 0$. Introduce $V = \mathsf{R}B$ and evolve

$$\dot{B} = \sigma \left(\sum_{i,j} \pi_i p_{ij} d_{ij} \right) (Wa)^\top \tag{68}$$

with the other four equations unchanged. On the nonnegative orthant of (B, K, Q, W, a) every derivative is nonnegative by the preceding sign checks. A locally Lipschitz vector field with this boundary property preserves that orthant: on any bounded interval, the squared distance to the orthant has upper derivative bounded by a constant times that distance, by projection onto the orthant and local Lipschitz continuity; Gronwall and zero initial distance give zero distance throughout. The lifted field is smooth, and its image solves the original equations. Uniqueness gives invariance on the original maximal interval. The lift cannot end earlier there: on a finite interval where the original parameters are bounded, the right side of equation 68 is bounded as well.

For $t_2 \geq t_1$, integration and closedness of C imply that each column of $V(t_2) - V(t_1)$ lies in C. Also $b_H(t) := W(t)a(t)$ is nonnegative and entrywise nondecreasing, by the product rule. Consequently

$$v(t_2) - v(t_1) = \big(V(t_2) - V(t_1)\big) b_H(t_2) + V(t_1)\big(b_H(t_2) - b_H(t_1)\big) \in C. \tag{69}$$

Pairing this increment with $v(t_1) \in C$ proves $\|v(t_2)\|^2 \geq \|v(t_1)\|^2$. Hence $\rho = \|v\| \geq \rho_0 > 0$. Only now define $u = v/\rho$; equation 64 yields $D_{ij}^\top u \geq \epsilon_{\mathrm{rew}}$ for all times in the maximal interval.

□

Proof of Theorem 5.1. Lemma G.3 supplies invariance and a nonzero head on the maximal interval. We first show that this interval is infinite.

The original vector field is smooth on the entire finite-dimensional parameter space. Its gradient-flow energy identity is $\dot{\mathcal{L}} = -\|\dot{\Theta}\|^2$. Thus, for $0 \le t_1 < t_2$ in the maximal interval, integral Cauchy–Schwarz gives

$$\int_{t_1}^{t_2} \|\dot{\Theta}\|\, dt \le \sqrt{(t_2 - t_1) \int_{t_1}^{t_2} \|\dot{\Theta}\|^2 dt} \le \sqrt{(t_2 - t_1)\mathcal{L}(0)}. \tag{70}$$

If the maximal endpoint were finite, this bound would make Θ Cauchy at that endpoint. Its finite limit would admit a local smooth solution extending the trajectory, a contradiction. All finite-time probabilities are therefore strictly positive.

To measure cumulative attention motion, define

$$\vartheta(t) = \int_0^t \sigma(\tau)\rho(\tau)\, d\tau, \qquad C_X = \max_{i,j} \|s_{ij}\| > 0, \quad T_X = \max_i \|t_i\|. \tag{71}$$

The derivative of this clock is positive. Pairing each equation in equation 66 with its factor shows that all three squared head norms have derivative $2\sigma v^\top b_{\mathrm{av}}$. It follows that

$$\dot{r} = 2\sigma v^\top b_{\mathrm{av}}, \qquad 0 \le \frac{dr}{d\vartheta} = 2u^\top b_{\mathrm{av}} \le 2C_X. \tag{72}$$

The lower bound uses acuteness; the upper bound uses Cauchy–Schwarz and the convex-combination bound $\|b_{\mathrm{av}}\| \le C_X$. If $\vartheta(\infty) = \Lambda < \infty$, then $r \le r(0) + 2C_X\Lambda =: R_H$. The product norm inequality and arithmetic–geometric mean inequality give $\rho \le \|V\|_F\|W\|_F\|a\| \le r^{3/2}$. Hence $m_i \le M_H := C_X R_H^{3/2}$. In either case,

$$\sigma \ge \begin{cases} e^{-M_H}, & \ell = \ell_{\exp}, \\ (1 + e^{M_H})^{-1}, & \ell = \ell_{\log}. \end{cases}$$

Thus $\dot{\vartheta} = \sigma\rho$ has a fixed positive lower bound on an infinite physical-time interval, contradicting finite Λ. Therefore $\vartheta \to \infty$.

We now bound each token gap by a common scalar drive. Use a subscript ϑ for differentiation in this clock. Equation 67 becomes

$$G = \frac{1}{\sqrt{d}} \sum_i \pi_i \Sigma_i u t_i^\top \ge 0, \qquad K_\vartheta = GQ, \quad Q_\vartheta = G^\top K, \quad M_\vartheta = KK^\top G + GQQ^\top. \tag{73}$$

Fix one nonselected index j_i for each sample. Put $w_i = \Sigma_i u \ge 0$. Because $|u_k| \le 1$, the scalar pair identity equation 60 implies

$$\begin{aligned} \|\Sigma_i u\|_1 &= \sum_k (w_i)_k \ge \sum_k |u_k|(w_i)_k \ge u^\top \Sigma_i u \\ &= \sum_{j<h} p_{ij}p_{ih}\big(u^\top(s_{ij} - s_{ih})\big)^2 \ge p_{ij_i^\star} p_{ij_i} \epsilon_{\mathrm{rew}}^2 > 0. \end{aligned} \tag{74}$$

The last inequality retains the square associated with the selected token and j_i. Thus the sum of all entries of G is

$$S = \sum_{k,b} G_{kb} = \frac{1}{\sqrt{d}} \sum_i \pi_i \|t_i\|_1 \|\Sigma_i u\|_1 > 0. \tag{75}$$

Cone invariance gives $Q(t) \ge Q_0 \ge 0$ entrywise. Squaring each entry and summing along row b gives

$$(QQ^\top)_{bb} = \sum_{\ell=1}^d Q_{b\ell}(t)^2 \ge \sum_{\ell=1}^d (Q_0)_{b\ell}^2 \ge \delta_{\mathrm{row}} > 0. \tag{76}$$

Also $Q \geq 0$ makes every entry of $QQ^\top$ nonnegative. Since $t_i > 0$, dropping its nonnegative off-diagonal contributions yields

$$(QQ^\top t_i)_b \geq (QQ^\top)_{bb}(t_i)_b \geq \delta_{\mathrm{row}}(t_i)_b. \tag{77}$$

The last inequality uses equation 76. Pairing equation 73 with $D_{ij}, t_i > 0$ and dropping the nonnegative $KK^\top G$ term gives

$$\Delta_{ij,\vartheta} \geq \frac{\delta_{\mathrm{row}}}{\sqrt{d}} D_{ij}^\top G t_i \geq \kappa_{ij} S > 0, \qquad \kappa_{ij} = \frac{\delta_{\mathrm{row}}}{\sqrt{d}} \min_k D_{ij,k} \min_b (t_i)_b > 0. \tag{78}$$

The second bound compares each nonnegative entry of G with the two positive coordinate minima. Since $\dot{\vartheta} > 0$, all gaps are strictly increasing in physical time. The identity

$$p_{ij_i^\star} = \left(1 + \sum_{j \neq j_i^\star} e^{-\Delta_{ij}}\right)^{-1} \tag{79}$$

then makes every selected-token probability strictly increasing as well.

It remains to show that strict increase continues without a finite upper limit. Suppose one gap is bounded above by a finite number H_Δ. Integrating equation 78 gives

$$I_S := \int_0^\infty S \, d\vartheta \leq \frac{H_\Delta - \Delta_{ij}(0)}{\kappa_{ij}} < \infty. \tag{80}$$

Because $G \geq 0$, $\|G\|_{\mathrm{op}} \leq \|G\|_F \leq S$. For $E = \|K\|_F^2 + \|Q\|_F^2$, Frobenius Cauchy–Schwarz and $2ab \leq a^2 + b^2$ give

$$E_\vartheta = 4\langle K, GQ\rangle_F \leq 4S\|K\|_F\|Q\|_F \leq 2SE, \qquad E \leq E(0)e^{2I_S} =: E_*. \tag{81}$$

The last inequality is Gronwall's inequality. Therefore $\|M\|_{\mathrm{op}} \leq \|K\|_F\|Q\|_F \leq E_*/2$. Every logit has absolute value at most $B_* = C_X T_X E_*/(2\sqrt{d})$, so the numerator lower bound e^{-B_*} and denominator upper bound se^{B_*} yield

$$p_{ij} \geq \epsilon_p := e^{-2B_*}/s > 0. \tag{82}$$

Substitute this bound into equation 74 and equation 75. Using only $\sum_i \pi_i = 1$ gives

$$S \geq \frac{\epsilon_p^2 \epsilon_{\mathrm{rew}}^2}{\sqrt{d}} \sum_i \pi_i \|t_i\|_1 \geq \frac{\epsilon_p^2 \epsilon_{\mathrm{rew}}^2}{\sqrt{d}} \min_i \|t_i\|_1 > 0. \tag{83}$$

An infinite ϑ interval with this fixed positive lower bound contradicts equation 80. Thus every gap tends to infinity. Equation 79 proves selected-token concentration, and finite token sums give $h_i \to x_{ij_i^\star}$. □

Remark G.4 (Constants in the contradiction). The cone, reward gap, LP witnesses, and row-norm floor are fixed by the data and checkpoint. Bounds introduced after assuming a finite clock or bounded gap may depend on that assumed bound. They are used only to reach a contradiction; no individual loss weight π_i is assumed bounded below.

G.3 Fixed-step gradient descent

We adapt the cone and attention-drive argument to simultaneous GD updates under the stated step-size bound.

Proof of Theorem 5.3. Reset the iteration index to the checkpoint. Write $\rho_k = \|v_k\|$, $u_k = v_k/\rho_k$, $h_k = \eta\sigma_k\rho_k$, and

$$G_k = \frac{1}{\sqrt{d}} \sum_i \pi_{i,k} \Sigma_{i,k} u_k t_i^\top, \qquad S_k = \sum_{a,b} (G_k)_{ab}.$$

The updates are simultaneous and use the old iterate on every right-hand side. The cone induction below ensures that $\rho_k > 0$, making u_k well defined.

Assume at an iterate that $K, Q, W, a \geq 0$ and every column of V belongs to C. The signed softmax mean b_{av} and $v = VWa$ belong to C. Because $K, Q \geq 0$ and $t_i > 0$, every scoring vector $KQ^\top t_i/\sqrt{d}$ is nonnegative. Equation 63 therefore gives $\Sigma_i v \geq 0$. The attention gradient matrix $\mathcal{B}$ in equation 67 is entrywise nonnegative. The exact simultaneous GD updates are

$$V^+ = V + \eta\sigma b_{\mathrm{av}}(Wa)^\top, \qquad W^+ = W + \eta\sigma V^\top b_{\mathrm{av}} a^\top, \qquad a^+ = a + \eta\sigma W^\top V^\top b_{\mathrm{av}},$$
$$K^+ = K + \eta\mathcal{B}Q, \qquad Q^+ = Q + \eta\mathcal{B}^\top K.$$

Acuteness gives $V^\top b_{\mathrm{av}} \geq 0$; hence W^+, a^+, K^+, Q^+ are nonnegative and every column of $V^+ - V$ belongs to C. Moreover, $b_H = Wa$ is nonnegative and $b_H^+ - b_H \geq 0$ entrywise. Consequently,

$$v^+ - v = (V^+ - V)b_H^+ + V(b_H^+ - b_H) \in C.$$

Pairing with $v \in C$ proves $\|v^+\| \geq \|v\|$. The low-loss checkpoint has $v_0 \neq 0$, so $\rho_k \geq \rho_0 > 0$ by induction. The same induction gives $Q_k \geq Q_0$ and thus preserves equation 76; the reward bound $D_{ij}^\top u_k \geq \epsilon_{\mathrm{rew}}$ also holds at every iterate.

Next we show that the discrete attention clock has infinite total length. This is the only step of the selection argument that needs the Part I radius law. Both losses satisfy (A1), and (G4) implies (A2) for the chosen loss. Together with (S1), these give the hypotheses of Theorem 4.2. Thus $r_k \to \infty$, $\rho_k/r_k^{3/2} \to 1$, and $q_k/r_k^{3/2} \to \gamma_\infty > 0$. For either loss, $-\ell'(m) \leq e^{-m}$, so $\sigma_k \leq e^{-q_k}$ and $\sigma_k\sqrt{r_k} \to 0$. The exact radius identity equation 17 and the head updates above give

$$r_{k+1} - r_k = 2h_k u_k^\top b_{\mathrm{av},k} + \frac{\eta^2}{3}\|\nabla_H \mathcal{L}_k\|^2, \qquad \|\nabla_H \mathcal{L}_k\|^2 \leq C\sigma_k^2 r_k^2.$$

The gradient bound follows from $\|b_{\mathrm{av},k}\| \leq C_X$ and $\|V\|_F\|W\|_F$, $\|V\|_F\|a\|$, $\|W\|_F\|a\| \leq 3r/2$. Since $u_k^\top b_{\mathrm{av},k} \leq C_X$ and $\rho_k \sim r_k^{3/2}$, the quadratic term divided by h_k is $O(\eta\sigma_k\sqrt{r_k}) = o(1)$. Therefore, eventually $r_{k+1} - r_k \leq (2C_X + 1)h_k$. The divergence of r_k implies

$$\sum_{k=0}^{\infty} h_k = \infty. \tag{84}$$

We apply equation 74 at each iterate. It gives $S_k > 0$ and, whenever every token probability has a fixed positive lower bound ϵ_p, gives

$$S_k \geq \frac{\epsilon_p^2 \epsilon_{\mathrm{rew}}^2}{\sqrt{d}} \min_i \|t_i\|_1 > 0. \tag{85}$$

Since $\mathcal{B}_k = \sigma_k\rho_k G_k$, the attention updates become $K_{k+1} = K_k + h_k G_k Q_k$ and $Q_{k+1} = Q_k + h_k G_k^\top K_k$. For $M_k = K_k Q_k^\top$,

$$M_{k+1} - M_k = h_k(G_k Q_k Q_k^\top + K_k K_k^\top G_k) + h_k^2 G_k Q_k K_k^\top G_k. \tag{86}$$

All entries of the final term are nonnegative. The coordinate row-norm bound equation 77, positive D_{ij}, t_i, and $G_k \geq 0$ therefore give exactly the lower bound from equation 78:

$$\Delta_{ij,k+1} - \Delta_{ij,k} \geq h_k \kappa_{ij} S_k > 0, \qquad j \neq j_i^\star. \tag{87}$$

The odds identity equation 79 shows that each selected probability is strictly increasing.

Strictly increasing gaps could still have finite limits. Suppose one did. Summing equation 87 would give $\sum_k h_k S_k < \infty$. Because $G_k \geq 0$, $\|G_k\|_{\mathrm{op}} \leq S_k$. Stack K_k and Q_k as a block vector. Their update matrix has off-diagonal blocks G_k and $G_k^\top$ and operator norm at most S_k. Thus, for $E_k = \|K_k\|_F^2 + \|Q_k\|_F^2$,

$$E_{k+1} \leq (1 + h_k S_k)^2 E_k, \qquad E_k \leq E_0 \exp\left(2\sum_{\ell<k} h_\ell S_\ell\right).$$

Hence K_k, Q_k and M_k are bounded. As in equation 82, all logits are bounded and every token probability has a uniform positive lower bound. Equation 85 then yields $S_k \geq s_0 > 0$ for every k, contradicting equation 84 and $\sum_k h_k S_k < \infty$. Every gap diverges. Equation equation 79 gives selected-token concentration, and the finite token sums give $h_i(A_k) \to x_{ij_i^\star}$.

Finally, Theorem 4.4 applies to this small-step GD trajectory and identifies the limiting classifier as $w_\star$ in equation 10. The time-radius estimate equation 55, together with $q_k/r_k^{3/2} \to \gamma_\infty$ and $V_k W_k a_k/q_k \to w_\star$, gives $q_k/\log(2 + \eta k) \to 1$ and the stated log-time limit. □

H ADDITIONAL EXPERIMENTAL DETAILS

In this section, we describe the MNIST measurements and give an explicit three-token instance for the strict-selection result.

H.1 DATA, MODEL, AND PAIRED INITIALIZATION

We retain unmodified MNIST digits 0 and 1 and select 256 training and 250 test images per class without replacement, using NumPy seed 2028. Labels are -1 and $+1$, respectively. Pixel intensities are divided by 255; sixteen nonoverlapping 7×7 blocks and their mean form 17 tokens. Appending fixed one-hot position codes gives $d = 66$. The mean token is the query. Thus this is the five-factor model of Section 3, with head widths 32 and 16, and no residual, normalization, feed-forward sublayer, or bias.

For factor seeds 23, 37, and 51, let $\bar{Q}_0$ have independent $N(0, 1/d)$ entries. We set $Q_0 = c_Q \bar{Q}_0$, $K_0 = 0$, and compare $c_Q = 3$ and 30. The head factors have independent centered Gaussian entries, with standard deviation 1.5 divided by the square root of their input width. Within each seed the two regimes have identical initial head factors and predictions. All five blocks are updated simultaneously by full-batch GD on exponential loss, with fixed step 0.05 for 20000 updates. We use PyTorch (Paszke et al., 2019) with float32 arithmetic on CUDA and record every 100 updates. A smaller-step run with $\eta = 0.025$ and 40000 updates uses the same accumulated step size and gives similar attention concentration and reward gaps for $c_Q = 30$.

H.2 SAMPLE-LEVEL MEASUREMENTS AND DISPLAY

We measure $\Psi = -\log(n\mathcal{L})/r^{3/2}$ once $n\mathcal{L} < 1$. Figure 1 uses factor seed 23 in both regimes, the same initialization used for the fixed-image maps. Its displayed margin curves start at the first saved low-loss checkpoints: update 3000 for $c_Q = 3$ and 1000 for $c_Q = 30$. The normalized margin increases at every recorded low-loss checkpoint in all six MNIST runs. Define the normalized signed token reward and its relative reward gap (regret) by

$$b_{ij} = y_i x_{ij}^\top \frac{v}{\|v\|}, \qquad R_i = \frac{\max_j b_{ij} - \sum_j p_{ij} b_{ij}}{\max_j b_{ij} - \min_j b_{ij}}.$$

The denominator is positive for every measured sample. This scale-free diagnostic is zero when attention is supported on maximal-reward tokens; it does not require their maximizer to be unique. We compute the empirical CDFs of R_i and $\max_j p_{ij}$ over all 512 training images separately for each seed. The trajectories in Figure 1 use seed 23 without averaging or smoothing; its CDFs are exact empirical step functions for that run. Figure 4 compares the SVM trajectories across the three paired initializations.

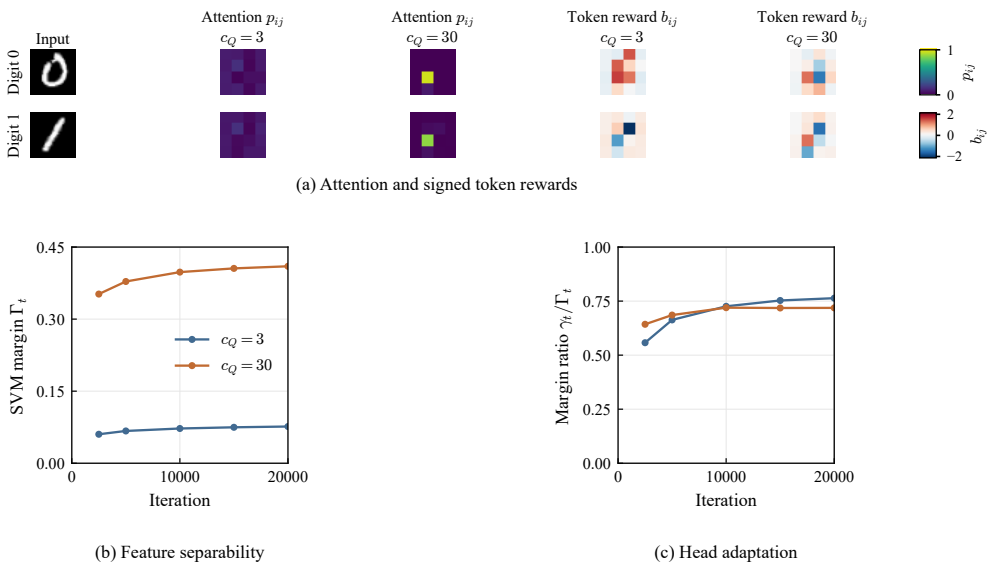
The maps in Figure 3 use original training indices 192 and 205, the smallest original training index of each class in the sampled subset. The maps use factor seed 23 and compare $c_Q = 3$ with $c_Q = 30$ at update 20000. Each grid shows the sixteen image coordinates without renormalizing after omitting the query token. Attention uses a common $[0, 1]$ scale and rewards a common symmetric scale over both images and initialization regimes. The scalar measurements use all 17 tokens.

H.3 EVOLVING-FEATURE SVM DIAGNOSTIC

We compare the trained classifier with what the current representation can achieve. At updates 2500, 5000, 10000, 15000, and 20000, we solve

$$w_t^{\mathrm{inst}} = \arg\min_w \tfrac{1}{2}\|w\|^2 \quad \text{subject to} \quad y_i h_i(t)^\top w \geq 1 \quad (i \in [n]).$$

The plotted achievable margin is $\Gamma_t = 1/\|w_t^{\mathrm{inst}}\|$; the attained fraction is $\min_i y_i h_i(t)^\top v(t)/(\|v(t)\|\Gamma_t)$. The curves start at update 2500, after all current heads separate the training set. Here the constraints are evaluated at the current checkpoint. Feasibility, stationarity, and relative primal–dual gap are checked numerically; the last two residuals must be below 10^{-4}.



(a) Attention and signed token rewards

(b) Feature separability

(c) Head adaptation

Figure 3: **Attention maps and current-feature margins on MNIST.** (a) Final attention and signed token rewards for two fixed images under the two query scales. Attention maps and reward maps each share a color scale; the mean query token is omitted without renormalizing. (b) Maximum unit-norm margin on the current aggregate features. (c) Fraction of that margin attained by the learned head. The curves use the same paired initialization as Figure 1.

Figure 4 compares the two quantities across all three paired initializations. The larger query scale produces features with a larger achievable margin in every pair. The attained fraction depends on initialization: improved feature separability does not imply a uniformly larger margin ratio.

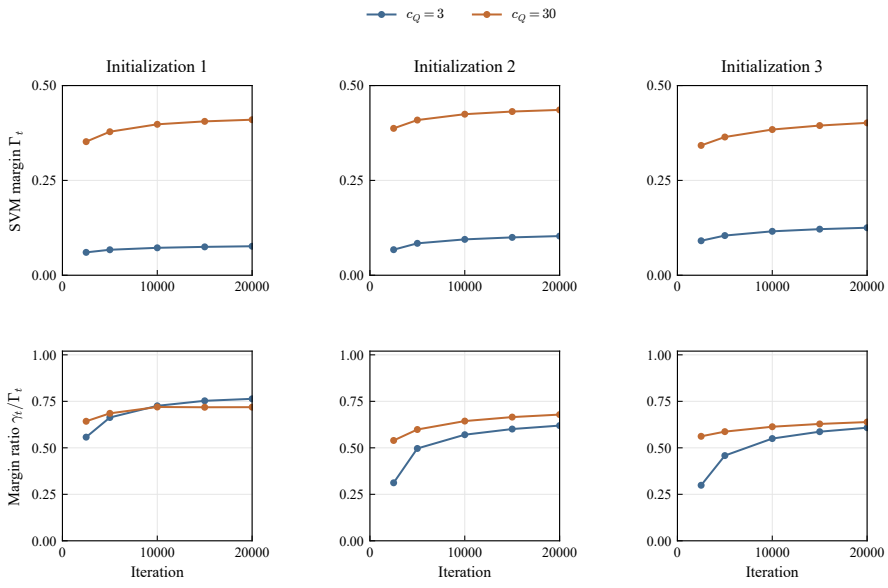


Figure 4: **Representation and head margins across initializations.** Columns correspond to factor seeds 23, 37, and 51, respectively; each column compares the two query scales with the same initial head. Top: maximum unit-norm margin on the current aggregate features. Bottom: fraction of that margin attained by the trained head. Colors denote query scales and markers denote SVM evaluations. The first column is the paired run shown in Figure 1.

H.4 A THREE-TOKEN INSTANCE WITH A NONZERO FLOW CERTIFICATE

We specify a four-example problem for which the preferred tokens and the selected-token SVM are known before training. Set $d = 2$, take labels $(+1, -1, +1, -1)$, and let the signed selected tokens and deficits be

$$S = \begin{pmatrix} 2.8 & 1.0 \\ 1.2 & 2.2 \\ 2.4 & 0.7 \\ 0.8 & 1.9 \end{pmatrix}, \qquad D^{(1)} = \begin{pmatrix} 0.8 & 0.2 \\ 0.2 & 0.8 \\ 0.6 & 0.4 \\ 0.4 & 0.6 \end{pmatrix}, \qquad D^{(2)} = \begin{pmatrix} 0.2 & 0.8 \\ 0.8 & 0.2 \\ 0.4 & 0.6 \\ 0.6 & 0.4 \end{pmatrix}.$$

For example i, the query is $q_i = y_i S_i$ and its three tokens are $X_i = y_i(S_i, S_i - D_i^{(1)}, S_i - D_i^{(2)})$; the first token is prescribed. The signed query $t_i = y_i q_i = S_i$ and both deficits are strictly positive, as required by (G1). We set

$$Q_0 = 3 \begin{pmatrix} 2 & 0.12 \\ 0.12 & 2 \end{pmatrix}, \qquad K_0 = 0.$$

To vary the initial head direction, let $e = (1, 1)^\top/\sqrt{2}$ and $u_\vartheta = (\cos\vartheta, \sin\vartheta)^\top$, and initialize

$$V_0 = s_\vartheta(u_\vartheta e^\top + 0.02I), \qquad W_0 = s_\vartheta(ee^\top + 0.02I), \qquad a_0 = s_\vartheta e, \qquad \vartheta \in \{75^\circ, 80^\circ, 85^\circ\}.$$

For each orientation, the unique positive scale s_ϑ is chosen so that $n\mathcal{L}(0) = 0.95$. The initial effective classifier is proportional to $u_\vartheta + 0.02e$, so its direction varies while the initial total loss is fixed. Both head matrices have full rank. All signed tokens and columns of V_0 are positive; their cone is acute and has positive reward against every deficit, giving (G2). The nonnegative factors, positive query row norms, and initial loss give (G4). All runs start with uniform attention.

Here the covariance certificate has a simple explicit form. Denote $s_0 = S_i$, $s_1 = S_i - D_i^{(1)}$, $s_2 = S_i - D_i^{(2)}$ and, for a generator g and coordinate k, write $c_{ab} = (s_{a,k} - s_{b,k})g^\top(s_a - s_b)$ as in equation 57. The coefficients c_{01} and c_{02} are positive, while c_{12} may be negative. Their token-sum vectors satisfy $b_{01} - b_{12} = D_i^{(2)} > 0$ and $b_{02} - b_{12} = D_i^{(1)} > 0$, so either positive pair may fund the negative one. For $i = 1, 2$, the two coordinate sums $c_{01} + c_{02} + c_{12}$ equal $1.04g_1 - 0.04g_2$ and $-0.04g_1 + 1.04g_2$; for $i = 3, 4$ they equal $0.56g_1 + 0.44g_2$ and $0.44g_1 + 0.56g_2$. Every generator here has $1/26 < g_1/g_2 < 26$, so these sums are positive. Whenever $c_{12} < 0$, the flows

$$F_{01,12} = \frac{-c_{12}c_{01}}{c_{01} + c_{02}}, \qquad F_{02,12} = \frac{-c_{12}c_{02}}{c_{01} + c_{02}}$$

meet the capacity and demand constraints in equation 58; otherwise zero flow suffices. The certificate is genuinely stronger than termwise nonnegativity: for $i = 1$, $k = 1$, and $g = (1.2, 2.2)$, the three coefficients are $(1.12, 0.40, -0.36)$.

The selected-token hard-margin problem has the explicit solution $w_\star = (0.3, 0.4)^\top$. Indeed,

$$Sw_\star = (1.24, 1.24, 1, 1)^\top, \qquad w_\star = \tfrac{1}{16}S_3 + \tfrac{3}{16}S_4,$$

which gives feasibility, complementary slackness, and stationarity with positive multipliers on the two support vectors.

We train all five factors using full-batch GD in double precision, with step 0.5 for 20000 updates. Figure 2 tracks the minimum of the eight prescribed-token gaps, the angle between v and $w_\star$, and $\|v/q - w_\star\|/\|w_\star\|$. Halving the step to 0.25 and doubling the number of updates gives similar trajectories at equal accumulated step size. At every update in these runs, the loss and smooth-margin inequalities equation 24 and equation 25 hold numerically with $\sigma_L = \sigma_\Psi = 1/2$; all prescribed-token gaps increase. These are checks on the simulated interval; the sufficient threshold $\eta_\star$ in (S1) is not evaluated.

For the same initializations, we integrate unit-rate gradient flow to time 10^8 using DOP853 in double precision on $\tau = \log(1 + t)$, with relative and absolute tolerances 2×10^{-8} and 2×10^{-10}. For the middle orientation, tightening both tolerances by a factor of 100 changes the final gap, angle, and relative SVM distance by less than 10^{-8}. Figure 5 gives these longer trajectories and the GD step-size comparison. The angular error can be nonmonotone when the classifier passes through the SVM direction.

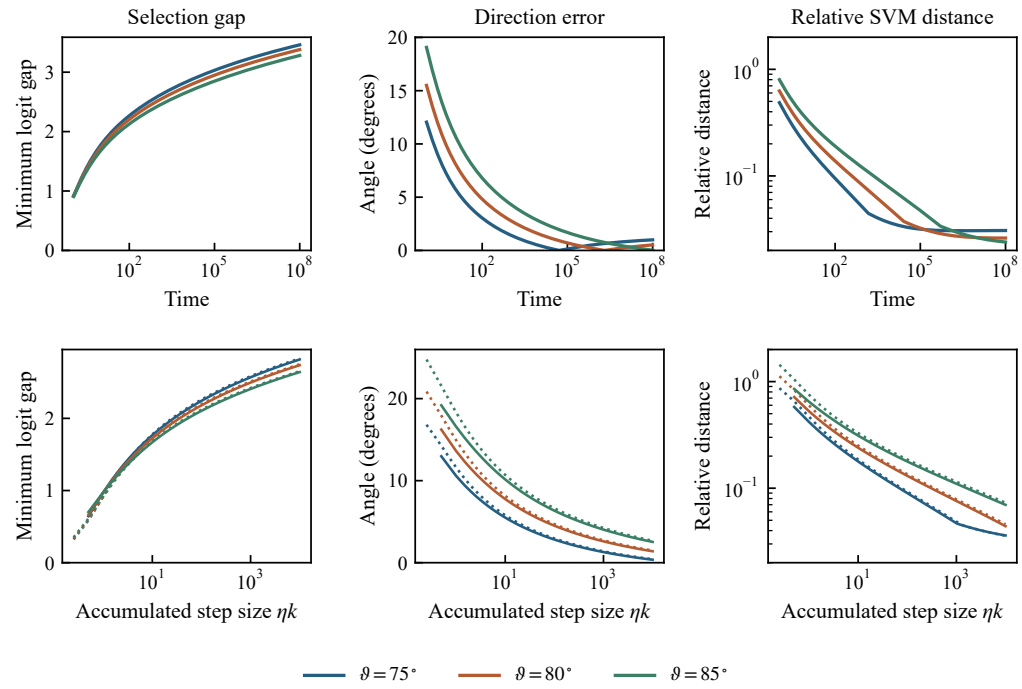


Figure 5: **Longer trajectories and step-size comparison.** Top: gradient flow from the three head orientations of Figure 2. Bottom: GD with step 0.5 (solid) and 0.25 (dotted), plotted against ηk. Columns show the minimum selection gap, angle to $w_\star$, and relative unit-margin distance. Colors have the same meaning in both rows.