A GEOMETRY-DEPENDENT THEORY OF SOFTMAX-ATTENTION IN-CONTEXT LEARNING DYNAMICS

Anonymous authors
Paper under double-blind review

ABSTRACT

Transformers exhibit in-context learning (ICL), the ability to solve new tasks from examples in the prompt without updating their parameters. Existing analyses of how gradient descent trains softmax attention focus largely on linear tasks and orthogonal features, leaving open how nonlinear task functions and non-orthogonal representations shape the training dynamics. In this work, we study one-layer softmax attention trained on in-context regression with nonlinear task functions over non-orthogonal feature prototypes. We first show that the population loss and its gradient decompose exactly, with the task distribution entering only through the covariance of its outputs across prototypes, which we call the task contrast, and the representation entering only through the Gram matrix of the prototypes. Using this decomposition, we prove monotone descent for every full-rank geometry and show that a geometry violating self-similarity can trap training at nonzero loss, whereas small correlations and transitive geometries such as prototypes on a ring converge globally from zero initialization. Under symmetric geometry, we show that training reduces exactly to a scalar recurrence with two stages, a coarse routing stage whose duration grows quadratically with the number of prototypes and an inverse-square refinement stage. We further show that the ratio of task contrast to target accuracy determines which stage is active when the target is reached, which yields two exact thresholds and a crossover at which refinement dominates training time. The learning-rate policy then decides whether stronger task contrast speeds up or slows down convergence. Finally, exact integrations and token-level SGD confirm the predicted boundaries and scaling laws.

1 INTRODUCTION

Transformers (Vaswani et al., 2017) can solve new tasks from a few input–output examples placed in the prompt, without updating their parameters (Brown et al., 2020). This in-context learning (ICL) ability has motivated extensive empirical study (Garg et al., 2022; Xie et al., 2021; Agarwal et al., 2024) and a growing theory of how gradient-based training produces it. For softmax attention, training-dynamics analyses establish convergence for linear regression with orthonormal features (Huang et al., 2024), contextual regression with representations (Yang et al., 2024), multi-head attention on linear tasks (Chen et al., 2024a; He et al., 2025), and chain-of-thought over discrete operators (Li et al., 2025). More recent work proves global convergence of a softmax layer on linear regression (Goel et al., 2026) and shows that attention heads specialize in successive stages (Sagitova et al., 2026; Yüksel et al., 2026). However, these analyses do not characterize how the geometry of nonlinear task functions or of non-orthogonal representations affects the training dynamics. Endpoint analyses characterize the trained attention (Collins et al., 2024) but not how gradient descent reaches it. Both gaps matter in practice. In-context learning often emerges abruptly after a long loss plateau (Reddy, 2024), whether it emerges depends on properties of the training data such as the number of classes (Chan et al., 2022), and the embeddings of trained language models are strongly correlated (Ethayarajh, 2019). Explaining these effects requires a theory of nonlinear task functions, whose outputs vary across shared inputs on a scale we call the task contrast, and of non-orthogonal representations, which couple the gradients of different feature classes. We ask:

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.

Q1: How does the contrast of a nonlinear task prior govern the convergence of softmax-attention ICL training?

Q2: How do non-orthogonal feature representations enter this convergence?

We answer both questions for one-layer softmax attention trained by gradient descent on nonlinear regression tasks over K correlated feature prototypes, a model that captures the clustered and correlated representations of trained networks. Our main contributions are summarized below.

- *The task prior and the representation enter training through separate factors.* We show that the population loss depends on the task distribution only through the covariance of its outputs across prototypes, and that gradient descent transports every class-logit gradient through the prototype Gram matrix on both sides. Using this decomposition, we prove monotone descent for every full-rank Gram matrix and an inverse-accuracy rate once each class attends mostly to its own prototype. We also show that a Gram matrix violating self-similarity can trap gradient descent at nonzero loss and lose a rare class, whereas equal-norm prototypes with small correlations and transitive geometries such as rings converge globally from zero initialization, which points to normalized, well-separated representations as favorable for in-context learning.
- *Under symmetric geometry, training follows two stages with explicit durations.* We show that for equiangular prototypes and every exchangeable task prior, including random single-index and random-network tasks, training from zero reduces exactly to a scalar recurrence. In its first stage, coarse routing moves attention from the uniform initialization to the correct prototype while the loss stays on a plateau whose length grows quadratically with the number of prototypes, a mechanism for the plateaus that precede the emergence of in-context learning in practice. In its second stage, inverse-square refinement reaches a target accuracy in a time inversely proportional to that accuracy. We prove matching bounds for both stages under random prompt counts and show that, at a fixed step size, the task contrast shortens only the first stage.
- *The ratio of task contrast to target accuracy determines which stage is active when the target is reached.* We identify two exact thresholds for this ratio, which separate tasks that are accurate at initialization, during coarse routing, and only during refinement. We further show that refinement dominates training time once the ratio exceeds the upper threshold by an amount of order the number of prototypes, and that the learning-rate policy decides whether stronger task contrast speeds up or slows down convergence. Exact integrations and finite-sample SGD on the token-level model confirm these predictions, also for random correlated representations, imbalanced classes, and noisy inputs.

2 Problem Setup

In-context regression with softmax attention. A task $f : \mathbb{R}^d \to \mathbb{R}$ is drawn from a prior $\mathcal{D}_f$, independently of a prompt with context inputs $x_1, \dots, x_N$ and a query x_{q}. Following Zhang et al. (2024) and Huang et al. (2024), the prompt stacks inputs above labels (with label 0 for the query), the value map reads the label row, and the key and query matrices are merged into a trainable $Q \in \mathbb{R}^{d\times d}$. The query attends only to the N context tokens, so its prediction is

$$\widehat{y}(Q) = \sum_{i=1}^{N} a_i(Q) f(x_i), \qquad a_i(Q) = \frac{\exp(x_i^\top Q x_{\mathrm{q}})}{\sum_{r=1}^{N} \exp(x_r^\top Q x_{\mathrm{q}})}. \tag{1}$$

We train Q by gradient descent on the population loss from $Q^{(0)} = 0$ unless stated otherwise,

$$\mathcal{L}(Q) = \frac{1}{2}\mathbb{E}\big[(\widehat{y}(Q) - f(x_{\mathrm{q}}))^2\big], \tag{2}$$

$$Q^{(t+1)} = Q^{(t)} - \eta \nabla_Q \mathcal{L}(Q^{(t)}), \qquad \eta > 0, \tag{3}$$

where the expectation is over the task, the context inputs, and the query.

Clustered, correlated representations. Trained networks map inputs of the same class to tightly clustered, correlated features (Papyan et al., 2020; Ethayarajh, 2019). We capture both properties

by representing inputs through K feature prototypes $V = (v_1, \dots, v_K)$ with a general Gram matrix $G = V^\top V \succ 0$, so that a token of class z is $x = v_z$ (continuous variation is treated in Appendix F.4). Query classes have probabilities p_k with $p_{\min} = \min_k p_k > 0$, which allows imbalanced, long-tailed data. Context classes are i.i.d. and independent of the query, with counts $n = (n_1, \dots, n_K)$, and a *fixed profile* holds $w = n/N$ fixed. With the class-score matrix $S = V^\top QV$, the total attention that a class-k query pays to class j, and the resulting residual direction, are

$$\pi_{j|k} = \frac{w_j \exp(S_{jk})}{\sum_{r=1}^K w_r \exp(S_{rk})}, \qquad a_k = \pi^{(k)} - e_k. \tag{4}$$

The routing error is $\mathcal{R} = \mathbb{E}_{k\sim p} \|a_k\|_2^2$. For a general G let $\lambda_\perp(G) = \min_{x\perp \mathbf{1}, \|x\|_2=1} x^\top Gx$ and $\omega_G = \lambda_\perp(G)\lambda_{\min}(G)$. A central example is the *equiangular* family

$$v_k^\top v_m = \rho + (1-\rho)\mathbf{1}\{k=m\}, \qquad -\tfrac{1}{K-1} < \rho < 1, \qquad G = \gamma I + \rho \mathbf{1}\mathbf{1}^\top, \tag{5}$$

with contrast scale $\gamma = 1 - \rho = \Delta^2/2$, where $\Delta = \|v_k - v_m\|_2$ for $k \neq m$. It describes trained classifiers, whose centered class means form a simplex with equal pairwise angles (Papyan et al., 2020), once a shared component such as the common direction of language-model embeddings (Mu et al., 2018) is added, which raises ρ.

Task geometry. Let $F_f = (f(v_1), \dots, f(v_K))^\top$ and $P_\perp = I - \mathbf{1}\mathbf{1}^\top/K$. Because attention weights sum to one, a class-k residual is $\widehat{y} - f(v_k) = a_k^\top F_f = a_k^\top P_\perp F_f$, so only centered outputs matter. We call their covariance $\Sigma_f = \mathbb{E}_f[(P_\perp F_f)(P_\perp F_f)^\top]$ the *task contrast*, since it measures how strongly the outputs of a random task vary across prototypes.

Assumption 1 (Task prior)**.** (i) Every task is L-Lipschitz. (ii) There exist $0 < \lambda_f \le \Lambda_f < \infty$ such that $\lambda_f P_\perp \preceq \Sigma_f \preceq \Lambda_f P_\perp$.

Both parts are mild. Part (i) is used only to state results in Lipschitz units. Part (ii) holds unless some fixed contrast $b^\top F_f$ with $b \perp \mathbf{1}$ vanishes for almost every task, that is, unless the prior never distinguishes some pattern of prototypes, and it ensures that every routing error is visible in the loss (Appendix G.1). The next lemma shows that Assumption 1 holds with $\lambda_f = \Lambda_f$ for every prior with nonconstant tasks that treats the prototypes symmetrically, including common nonlinear task families. The task contrast is then isotropic,

$$\Sigma_f = \sigma_f^2 P_\perp. \tag{6}$$

Lemma 1 (Exchangeable task priors)**.** *If the law of F_f is invariant under permutations of its coordinates, then (6) holds with $\sigma_f^2 = \frac{1}{2}\mathbb{E}_f(f(v_1) - f(v_2))^2 \le \frac{1}{2}L^2 \|v_1 - v_2\|_2^2$. For equiangular prototypes, F_f is exchangeable whenever the law of f is invariant under the orthogonal maps that permute the prototypes. Examples include random single-index tasks $f(x) = g(w^\top x + b)$ and random two-layer networks with i.i.d. spherical weights, and the sign family of Proposition 5.*

The proof is given in Appendix C. Such symmetric priors model a diverse task distribution in which no prototype is special, and they are the setting of Section 4. For equiangular prototypes, we write $\sigma_f^2 = c_f^2 L^2 \Delta^2$, where the *contrast fraction* $c_f \le 1/\sqrt{2}$ measures how much of its Lipschitz budget a prior uses. Since the inputs lie on the prototypes, a task affects training only through F_f, so our analysis treats linear and nonlinear priors in a unified way, through the scale of their output contrast.

3 Exact Decomposition and General Convergence

We first ask how the task prior and the representation enter training. The answer is a factorization of the loss and its gradient into two measurable quantities, the task contrast Σ_f and the Gram matrix G. We start by expressing the population loss (2), an average over random tasks, prompts, and queries, in closed form through the attention weights.

Lemma 2 (Exact covariance loss)**.** *For every profile, the population loss (2) is a sum of class losses,*

$$\mathcal{L} = \sum_{k=1}^K \ell_k, \qquad \ell_k = \frac{p_k}{2} a_k^\top \Sigma_f a_k, \tag{7}$$

$$\frac{\lambda_f}{2}\mathcal{R} \le \mathcal{L} \le \frac{\Lambda_f}{2}\mathcal{R}, \qquad \textit{and} \qquad \mathcal{L} = \frac{\sigma_f^2}{2}\mathcal{R} \ \textit{under (6)}. \tag{8}$$

These relations continue to hold after averaging over random context counts.

The proof is given in Appendix C. A task therefore affects training only through how much its outputs vary across prototypes, a quantity that replaces the Lipschitz constant of earlier analyses and can be estimated from data. Since attention depends on Q only through the class scores $S = V^\top QV$, we track the gradient $h_k = \nabla_{S_{\cdot k}}\mathcal{L}$ with respect to the scores of query class k, with $H = (h_1, \dots, h_K)$. Differentiating (7) gives

$$h_k = p_k J_k \Sigma_f a_k, \qquad J_k = \operatorname{diag}(\pi^{(k)}) - \pi^{(k)}\pi^{(k)\top}, \tag{9}$$

where J_k is the Jacobian of the class-k attention. The next proposition shows how the representation enters the update.

Proposition 1 (Exact Gram-coupled update)**.** *For a fixed profile,*

$$\nabla_Q \mathcal{L} = VHV^\top, \tag{10}$$

so that the class scores and the target-to-off-target gaps $d_{k,m} = S_{kk} - S_{mk}$ evolve as

$$S^{(t+1)} = S^{(t)} - \eta\, GH^{(t)}G, \tag{11}$$

$$d_{k,m}^{(t+1)} - d_{k,m}^{(t)} = -\eta\,(e_k - e_m)^\top GH^{(t)}Ge_k. \tag{12}$$

For random counts, $H^{(t)}$ is replaced by its count average.

The proof is given in Appendix C. The representation thus acts on the gradients only through G, once on the query side and once on the key side. For orthonormal features, each class learns independently, as in Huang et al. (2024). With the correlated embeddings found in practice, learning one class also moves the scores of all others, and the next lemma shows that the gradient on the target logit nevertheless keeps a definite sign.

Lemma 3 (Covariance noncancellation)**.** *For every profile, $(h_k)_k = -p_k\pi_{k|k}\, a_k^\top \Sigma_f a_k = -2\pi_{k|k}\ell_k$. Under Assumption 1(ii),*

$$-(h_k)_k \geq p_k \lambda_f\, \pi_{k|k}(1 - \pi_{k|k})^2. \tag{13}$$

The proof is given in Appendix C. The gradient on the target logit therefore always raises the target's own score, in proportion to the current routing error, and this drives attention to concentrate on the correct prototype. Since the objective is also $6\Lambda_f \|G\|_{\mathrm{op}}^2$-smooth in Q (Lemma 4), this sign yields a convergence guarantee for every full-rank geometry under the step size and the routing basin

$$0 < \eta \leq \frac{1}{6\Lambda_f \|G\|_{\mathrm{op}}^2}, \tag{14}$$

$$\mathcal{L}_{t_0} \leq \frac{p_{\min}\lambda_f}{8}. \tag{15}$$

Theorem 1 (General Gram-spectral convergence)**.** *Let $G \succ 0$, fix a positive profile, and assume Assumption 1(ii) and (14). From any initialization, the loss is nonincreasing and*

$$\mathcal{L}_{t+1} \leq \mathcal{L}_t - \frac{\eta}{2}\operatorname{tr}\big[(H^{(t)})^\top GH^{(t)}G\big]. \tag{16}$$

If (15) holds, then $\pi_{k|k}^{(t)} \geq 1/2$ for all k and $t \geq t_0$, and

$$\mathcal{L}_{t+1} \leq \mathcal{L}_t - \frac{\eta\,\omega_G}{2K}\mathcal{L}_t^2. \tag{17}$$

Hence, for $0 < \epsilon < \mathcal{L}_{t_0}$, the loss reaches ϵ within $\frac{2K}{\eta\omega_G}(\frac{1}{\epsilon} - \frac{1}{\mathcal{L}_{t_0}}) + 1$ iterations after t_0. If $H^{(t)}\mathbf{1} = 0$ throughout, ω_G may be replaced by $\lambda_\perp(G)^2$.

The proof is given in Appendix D. The geometry plays two roles. The largest eigenvalue of G limits the stable learning rate, because every gradient is transported through G twice, and the contrast gain ω_G sets the late-phase speed, so well-separated representations train faster. In the basin, the loss also bounds the attention error of every class (Corollary 1 in Appendix D). Whether gradient descent from zero initialization reaches the basin depends on the geometry, as the next example shows.

Proposition 2 (Non-orthogonal geometry can trap gradient descent). *Let $K = 2$, $\Sigma_f = P_\perp$, $w = (\frac{1}{2}, \frac{1}{2})$, $p = (\frac{100}{101}, \frac{1}{101})$, and $G = \begin{pmatrix} 1 & 3 \\ 3 & 10 \end{pmatrix} \succ 0$. For every $\eta > 0$, gradient descent from $Q^{(0)} = 0$ satisfies $\pi^{(t)}_{2|2} \to 0$ and $\mathcal{L}_t \to p_2 > 0$.*

The proof is given in Appendix D.5. Here $\langle v_1, v_1 - v_2 \rangle = -2 < 0$ violates *self-similarity*, i.e., $G_{kk} > G_{km}$ for $m \neq k$, so the first prototype is more aligned with the second than with itself. After transport through G, the gradient of the frequent class then pushes the gap of the rare class in the wrong direction faster than the rare class can correct it, whereas with uniform queries the same G converges numerically. With long-tailed data, a rare class whose representation is dominated by a frequent one can thus be lost from in-context predictions. Equal-norm representations, as encouraged by normalization layers, always satisfy self-similarity, and the next theorem shows that with small correlations they converge globally for arbitrary class frequencies.

Theorem 2 (Global convergence beyond symmetry). *Let the prototypes have equal norms, so that after rescaling η we may write $G = I + E$ with $E_{kk} = 0$, $\max_{j \neq k} |E_{jk}| \leq \bar{\rho}$, and $G \succ 0$. Let the profile be balanced, $\Sigma_f = \sigma_f^2 P_\perp$, and p arbitrary with $p_{\min} = \min_k p_k$. If*

$$\bar{\rho} \leq \frac{p_{\min}}{9K} \qquad \text{and} \qquad 0 < \eta \leq \frac{1}{6\sigma_f^2 \|G\|_{\text{op}}^2}, \tag{18}$$

then gradient descent from $Q^{(0)} = 0$ keeps every diagonal score largest in its column, so $\pi^{(t)}_{k|k} \geq 1/K$ for all k and t, and

$$\mathcal{L}_{t+1} \leq \mathcal{L}_t - c_R \mathcal{L}_t^2, \qquad \mathcal{L}_t \leq \left(\mathcal{L}_0^{-1} + c_R t\right)^{-1}, \qquad c_R = \frac{2\eta\omega_G}{K^2(K-1)}. \tag{19}$$

Every class moreover reaches majority attention by $T_{1/2} \leq \lceil 32(K-2)/(\eta\sigma_f^2 p_{\min}) \rceil + 1$ and keeps it, after which (17) holds with $2K$ replaced by $2(K-1)$ and $\mathcal{L}_t \leq \epsilon$ within $\lceil 2(K-1)/(\eta\omega_G\epsilon) \rceil$ further iterations; for uniform queries the routing bound is $O(K^2/(\eta\sigma_f^2))$, as under symmetry, and $\omega_G \geq (17/18)^2$.

The proof is given in Appendix D. It shows that the region in which every target score is the largest entry of its column is invariant, after which descent drives the loss to zero. The budget scales with $p_{\min}$ because a rare class receives its own gradient at rate $p_{\min}$, while the gradients of frequent classes reach it through G at rate $\bar{\rho}$. For uniform queries it is of order $1/K^2$, the best possible order for an argument that certifies the region from every state (Appendix D.7), and symmetry allows a far larger budget. We call G *transitive* if a permutation group acting transitively on $[K]$ leaves it invariant, i.e., $UGU^\top = G$, as for cyclic concepts that language models represent on circles, such as days of the week (Engels et al., 2025). The equiangular family corresponds to $E = 0$.

Theorem 3 (Global convergence under transitive geometry). *Let $G = \gamma I + \rho \mathbf{1}\mathbf{1}^\top + E \succ 0$, $\gamma > 0$, be transitive with E symmetric, $E_{kk} = 0$, and $c := \gamma^{-1} \max_k \sum_{m \neq k} |E_{km}| \leq \frac{1}{5}$; let queries be uniform, the profile balanced, $\Sigma_f = \sigma_f^2 P_\perp$, and $0 < \eta \leq K/(6\sigma_f^2 \lambda_{\perp,\max}(G)^2)$, with $\lambda_{\perp,\max}(G)$ the largest eigenvalue of G on $\mathbf{1}^\perp$. Then from $Q^{(0)} = 0$ every gap increases at every step, all classes reach majority attention within $\lceil 16K(K-2)/(\theta_c \eta \sigma_f^2 \gamma^2) \rceil + 1$ iterations, $\theta_c = 1 - 4c - 2c^2$, and keep it, and thereafter $\mathcal{L}_{t+1} \leq \mathcal{L}_t - \eta \lambda_\perp(G)^2 \mathcal{L}_t^2/(2(K-1))$.*

The proof is given in Appendix D.9. Here the common correlation ρ is arbitrary and the asymmetric budget is an absolute constant, because transitivity keeps the trajectory exactly synchronized, i.e., $H^{(t)}\mathbf{1} = 0$, so that every perturbation is measured against a single gradient scale. Our experiments show that the invariant region persists for arbitrary E with $c \leq \frac{1}{2}$ (Section 6), which suggests that a constant budget extends beyond transitive geometries. The next section complements these guarantees with an exact description of the whole trajectory.

4 Global Dynamics under Symmetric Geometry

The guarantees of Section 3 show that training converges, but practitioners also want to know how long it takes and what the loss curve looks like. We answer these questions under symmetric geometry.

It is the geometry toward which trained classifiers converge (Papyan et al., 2020), and it treats all classes alike, which makes it the natural baseline for a diverse task distribution. The update also preserves the symmetry, so the matrix recurrence reduces to a scalar one that we can follow from the first step to any accuracy and bound from both sides. Throughout, we assume equiangular prototypes (5), an isotropic task contrast (6), and uniform query and context classes.

Proposition 3 (Exact symmetric reduction). *Under these conditions, $S^{(t)} = d_t P_\perp + c_t \mathbf{1}\mathbf{1}^\top$ for all t, the shift c_t does not affect attention, and $GH^{(t)}G = \gamma^2 H^{(t)}$. For exactly balanced counts, write $R_t = (K-1)e^{-d_t}$ for the off-target odds, $\alpha_t = 1/(1+R_t)$ for the target attention, and $u_t = 1-\alpha_t$. Then $\mathcal{L}_t = \frac{\sigma_f^2 K}{2(K-1)} u_t^2$ and*

$$d_{t+1} - d_t = \beta\, \alpha_t u_t^2, \qquad R_{t+1} = R_t \exp\Big\{ -\beta \frac{R_t^2}{(1+R_t)^3} \Big\}, \qquad \beta = \eta \sigma_f^2 \gamma^2 \frac{K}{(K-1)^2}. \tag{20}$$

In particular, because both the query probability and the initial target attention equal $1/K$, the first gap increment is

$$d_1 - d_0 = \frac{\eta \sigma_f^2 \gamma^2}{K^2}. \tag{21}$$

The proof is given in Appendix E. Under symmetry, a single gap d_t between the target score and every off-target score describes the whole trajectory, and the recurrence depends on the task prior and the representation only through β. At fixed γ, the common correlation has no effect, since the equiangular dynamics coincide with those of orthogonal features of squared norm γ, so the shared direction of anisotropic embeddings neither helps nor hurts, and only the separation Δ between prototypes matters. The first increment (21) is of order $1/K^2$, because a query belongs to class k with probability $1/K$ and initially pays attention $1/K$ to its own class, which is the origin of the K^2 routing bottleneck.

Proposition 3 assumes exactly balanced prompts. Real prompts have random class counts, and we control their fluctuations through a prompt-length condition, stated as (105) in Appendix E, which holds once the prompt length exceeds K times a logarithmic factor. With a step size

$$0 < \eta \le \frac{c_\eta}{\sigma_f^2 \gamma^2}, \qquad 0 < c_\eta \le \frac{1}{6}, \tag{22}$$

the scalar recurrence then yields a global convergence time.

Theorem 4 (Global convergence under random prompts). *Under (105) with count tolerance $0 < \zeta \le 1/2$, (22), and $0 < \epsilon < \sigma_f^2$, gradient descent from $Q^{(0)} = 0$ reaches $\mathcal{L}_t \le \epsilon$ within*

$$T_\epsilon \le C_\zeta \Big(\frac{K^2}{\eta \sigma_f^2 \gamma^2} + \frac{K}{\eta \gamma^2 \epsilon} \Big) = O_\zeta \Big(\frac{K^2}{\eta \sigma_f^2 \Delta^4} + \frac{K}{\eta \Delta^4 \epsilon} \Big) \tag{23}$$

iterations, where C_ζ depends only on ζ. For an L-Lipschitz prior with contrast fraction c_f, the first term equals $4K^2/(\eta c_f^2 L^2 \Delta^6)$.

The proof is given in Appendix E. The first term in (23) is the time needed to route attention to the correct prototype, and the second is the time needed to refine it, as the next proposition shows.

Proposition 4 (Two dynamical laws). *Let counts be exactly balanced, $K \ge 3$, and let (22) hold. (i)* Coarse routing. *While $R_t \ge 1$, $\beta/16 \le R_t - R_{t+1} \le \beta$. Hence $T_1 := \inf\{t : R_t \le 1\}$ satisfies*

$$\frac{K-2}{\beta} \le T_1 \le \frac{16(K-2)}{\beta} + 1, \qquad \text{so} \qquad T_1 = \Theta\Big(\frac{K^2}{\eta \sigma_f^2 \gamma^2} \Big). \tag{24}$$

(ii) Inverse-square refinement. *While $R_t \le 1$, the potential $Y_t = u_t^{-2}$ satisfies $\beta/4 \le Y_{t+1} - Y_t \le 7\beta$. Under random counts, both bounds hold with ζ-dependent constants (Appendix E).*

The proof is given in Appendix E. While attention is diffuse, the odds decrease by a nearly constant amount per step, so removing the initial $K-1$ units of odds takes of order K/β steps. Since the target attention $1/(1+R_t)$ stays small until R_t is of order one, the loss stays near its initial value for most of this stage, so coarse routing appears as a loss plateau. This offers a mechanism for the

plateaus that precede the abrupt emergence of in-context learning in practice (Reddy, 2024), and it predicts that they lengthen quadratically with the number of classes, a data property known to control whether in-context learning emerges (Chan et al., 2022). Once the target holds majority attention, the potential u_t^{-2} grows linearly, so reaching an attention error u takes of order u^{-2}/β further steps. A larger contrast strengthens the early gradient and shortens routing. In refinement, however, an absolute loss ϵ requires $u_t^2 \asymp \epsilon/\sigma_f^2$, which cancels this gain, so σ_f^2 enters only the first term of (23).

Non-uniform query classes are treated in Appendix E.6. Since both stages have matching upper and lower bounds, the time to reach a given accuracy is determined up to constants once we know which stage is active when the target is reached, which we characterize next.

5 The Geometry–Accuracy Phase Diagram

Section 4 describes the whole loss curve, but in practice training stops once the loss meets a target accuracy ϵ, and the cost of training is the time $T_{\mathrm{acc}} = \inf\{t : \mathcal{L}_t \le \epsilon\}$ needed to reach it. Which of the two stages determines this cost is the question behind choosing a training budget, a learning rate, and the number of classes in the data, and this section answers it. By Lemma 2, reaching ϵ is equivalent to the routing target $\mathcal{R}_t \le 2\epsilon/\sigma_f^2$, which suggests measuring the task contrast in units of the target accuracy through the ratio $q = \sigma_f^2/\epsilon$. Under balanced counts, the losses at initialization and at the crossover $R = 1$ between the two stages, where the target attention equals $1/2$, and the corresponding values of q are

$$\mathcal{L}_0 = \frac{\sigma_f^2(K-1)}{2K}, \qquad \mathcal{L}_{\mathrm{cross}} = \frac{\sigma_f^2 K}{8(K-1)}, \tag{25}$$

$$q_{\mathrm{init}} = \frac{2K}{K-1}, \qquad q_{\mathrm{tail}} = \frac{8(K-1)}{K}. \tag{26}$$

For $q > q_{\mathrm{init}}$ the accuracy target is the state $u_\epsilon = \sqrt{2(K-1)/(Kq)}$, with off-target odds $R_\epsilon = u_\epsilon/(1-u_\epsilon)$ and potential $Y_\epsilon = u_\epsilon^{-2} = Kq/(2(K-1))$.

Comparing the target ϵ with these two reference losses classifies every task.

Theorem 5 (Geometry–accuracy regimes)**.** *Consider equiangular prototypes, isotropic task covariance, uniform queries, exactly balanced counts, $Q^{(0)} = 0$, $K \ge 3$, and (22). (a)* Initial accuracy: $T_{\mathrm{acc}} = 0$ *if and only if* $q \le q_{\mathrm{init}}$. *(b)* Coarse accuracy: *if* $q_{\mathrm{init}} < q \le q_{\mathrm{tail}}$, *then* $T_{\mathrm{acc}} \le T_{\mathrm{I}}$ *and* $T_{\mathrm{acc}} = \Theta\big(1 + (K - 1 - R_\epsilon)/\beta\big)$. *(c)* Tail accuracy: *if* $q > q_{\mathrm{tail}}$, *then* $R_\epsilon < 1$ *and*

$$\frac{Y_\epsilon - 4}{7\beta} - 1 \le T_{\mathrm{acc}} - T_{\mathrm{I}} \le \frac{4(Y_\epsilon - 4)}{\beta} + 1, \tag{27}$$

$$\frac{Y_\epsilon - 4}{\beta} = \frac{K-1}{2\eta\gamma^2\epsilon}\Big(1 - \frac{q_{\mathrm{tail}}}{q}\Big), \tag{28}$$

so the two bounds match up to constants whenever $Y_\epsilon - 4 \ge 14\beta$.

The proof is given in Appendix F. The recurrence (20) is the same smooth map in every regime, so the regimes describe which stage is active when the target is reached, not a bifurcation of the update. Below q_{init}, uniform attention already meets the target. Between the two thresholds, the required target attention is at most one half, so the target is reached no later than majority attention. Above q_{tail}, it is reached only during refinement. The ratio q also determines which stage dominates the training time. When $Y_\epsilon - 4 \ge 14\beta$,

$$T_{\mathrm{I}} = \Theta\Big(\frac{K^2}{\eta\sigma_f^2\gamma^2}\Big), \qquad T_{\mathrm{acc}} - T_{\mathrm{I}} = \Theta\Big[\frac{K}{\eta\gamma^2\epsilon}\Big(1 - \frac{q_{\mathrm{tail}}}{q}\Big)\Big], \tag{29}$$

and their ratio is $\Theta((q - q_{\mathrm{tail}})/K)$. Hence refinement overtakes the K^2 routing bottleneck once $q - q_{\mathrm{tail}} = \Theta(K)$. In practice, this tells where the training budget goes. For moderate accuracy targets the routing plateau dominates, and adding classes lengthens training quadratically. For stringent targets refinement dominates, and training time grows only linearly with the number of classes but inversely with the target accuracy. Whether a sharper task speeds up training depends on the learning-rate policy. Under a common admissible step size, stronger contrast shortens routing as q^{-1},

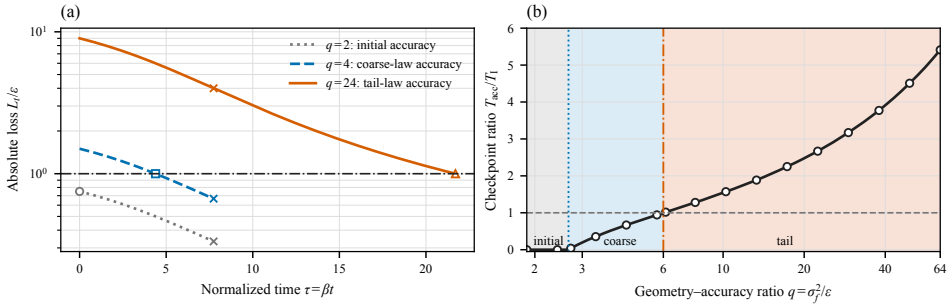


Figure 1: Geometry–accuracy regimes under the full population update ($K = 4$). (a) Loss relative to the target for $q = 2, 4, 24$, with T_{acc} (open markers) and T_{I} (crosses). (b) The ratio $T_{\mathrm{acc}}/T_{\mathrm{I}}$ against q, with the exact boundaries $q_{\mathrm{init}} = 8/3$ and $q_{\mathrm{tail}} = 6$.

while the refinement time is not monotone in q (Appendix G.3). Under the certified curvature-scaled step size $\eta = \Theta((\sigma_f^2\gamma^2)^{-1})$, routing takes $\Theta(K^2)$ iterations for every q, whereas refinement grows as $\Theta(K(q - q_{\mathrm{tail}}))$. Since learning rates are commonly tuned to the largest stable value, this predicts that tasks with sharper outputs can take longer to reach a fixed accuracy, even though they provide a stronger learning signal. In Lipschitz units, both boundaries occur where $c_f L\Delta$ is of order $\sqrt{\epsilon}$ (Corollary 2 in Appendix F).

The classification also holds for multinomial counts (Proposition 11 in Appendix F) and for inputs with label-irrelevant variation, such as noise in embedding directions that the tasks do not use, because training then follows exactly the prototype trajectory (Theorem 6 in Appendix F.4).

6 Experiments

We now test whether these predictions hold, first for the exact population dynamics and then in increasingly realistic training (details and more experiments in Appendix H). We integrate the full $K \times K$ update (11) for equiangular prototypes with $K = 4$ and $\Delta = 0.9$, with tasks drawn from a permuted four-level codebook of values scaled by $L\Delta$, which is an exchangeable prior. With $\epsilon = 0.02$, the cases $q = 2, 4, 24$ fall in the initial, coarse, and tail regimes of Theorem 5, reaching the target before, during, and after the routing plateau (Figure 1a). Across 244 runs with $K \in \{4, 8, 16, 32\}$, the measured regime agrees with the theorem whenever the target is not within one step of a crossing (Figure 1b).

The same integrations confirm the learning-rate prediction. For $K = 32$, raising q eightfold shortens routing eightfold under a common step size, but it grows refinement sixfold under the curvature-scaled step size (Table 3), and independent sweeps recover every predicted exponent with $R^2 > 0.99$ (Table 4).

Finally, we move toward realistic training in three directions (Figure 2). First, finite-sample SGD on the token-level model (1), with a random nonlinear single-index prior, random prompt counts ($N = 64$), and orthogonal continuous nuisance, reproduces the predicted regime in all nine tested configurations (Figure 2a). Its stopping times are within 1.9% of the random-count population dynamics, and its routing time grows as $K^{2.22}$, compared with $K^{2.18}$ for the recurrence (Appendix H.2). With random unit-norm geometry and imbalanced queries, SGD also reproduces all 24 predicted regimes within 5% of its population reference (Appendix H.6). Second, across 1,153 runs from zero initialization with correlated asymmetric geometry ($K \leq 64$), the invariant region held in all 684 runs with correlation mass at most $\gamma/2$, far beyond the certified budgets, and majority attention was reached within 1.83 times the symmetric time. The star geometries of Appendix D.7 keep the region from zero at every tested mass but lose it from the adversarial state above $4(K-1)/K^2$, as predicted, so the initialization determines the budget (Figure 2c, Appendix H.4). Third, with task-relevant input noise, token-level SGD follows the predicted attention trajectory for $r/\Delta \leq 0.2$, and its loss levels off near the within-class label variance, which is the loss of the class-mean predictor (3.67ϵ predicted and 3.61ϵ measured, Figure 2b and Appendix H.7).

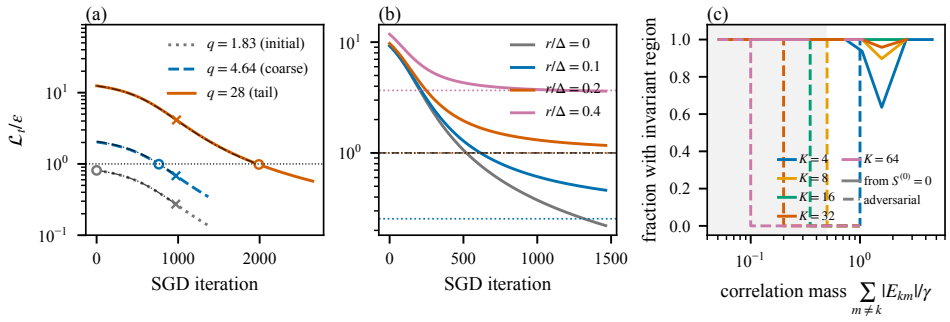


Figure 2: Beyond the assumptions of the theory. (a) Token-level SGD with a nonlinear single-index prior and random counts ($K = 8$) in the three regimes (mean of three seeds), with the population prediction (dash-dotted), $\widehat{T}_{\mathrm{acc}}$ (circles), and $\widehat{T}_{\mathrm{I}}$ (crosses). (b) Task-relevant input noise ($K = 4$, $q = 24$), with the within-class label-variance levels (dotted). (c) Fraction of runs from zero initialization that keep the invariant region against the correlation mass $\max_k \sum_{m \neq k} |E_{km}|/\gamma$ (solid), and the same fraction for the star geometries of Appendix D.7 started from their adversarial state (dashed).

7 Related Work

Theories of ICL span Bayesian views (Xie et al., 2021; Zhang et al., 2025b; Falck et al., 2024), induction heads (Bietti et al., 2023; Nichani et al., 2024; Edelman et al., 2024), expressivity (Akyürek et al., 2023; von Oswald et al., 2023; Bai et al., 2023), and generalization (Collins et al., 2024; Oko et al., 2024). Beyond the training-dynamics analyses discussed in Section 1, related analyses cover linear attention (Zhang et al., 2024; 2025a), next-token prediction (Tian et al., 2023), nonlinear classification (Li et al., 2024), and a large-prompt limit in which softmax attention behaves like linear attention (Boursier & Boyer, 2025). Nonlinear ICL has mostly been studied through what a trained model represents (Shen et al., 2025; Hsu et al., 2026), and not through the training trajectory. Empirically, in-context learning emerges abruptly after loss plateaus (Olsson et al., 2022; Reddy, 2024), whose length and occurrence depend on data properties such as the number of classes and task diversity (Chan et al., 2022; Kim et al., 2024). Our routing stage gives a single-layer mechanism for such plateaus and ties their length to the number of classes, the task contrast, the representation geometry, and the step size. The closest work to ours is Huang et al. (2024), whose balanced orthonormal linear-regression setting corresponds to $G = I$ and $\Sigma_f = P_\perp$ here. We add nonlinear exchangeable priors, general Gram matrices, and two-sided bounds for each stage. Due to space limitations, we provide a more detailed discussion of related work in Appendix B.

8 Conclusion and Limitations

In this work, we gave an exact description of softmax-attention ICL training on nonlinear tasks with non-orthogonal representations, in which the task prior enters through the covariance of its outputs and the representation through its Gram matrix. Under symmetric geometry, training follows a coarse routing stage with a K^2 bottleneck and an inverse-square refinement stage, and the ratio of task contrast to target accuracy determines which stage is active when the target is reached. Beyond symmetry, a Gram matrix that violates self-similarity can trap gradient descent, while small correlations and transitive geometries converge globally. For practice, the analysis identifies two measurable quantities, the task contrast of the data and the Gram matrix of the representations, and predicts from them whether in-context learning emerges, how long its plateau lasts, and whether sharper tasks or larger learning rates speed it up. Our global results cover symmetric, transitive, and near-orthogonal geometries, including mild anisotropy and imbalance (Appendix D.8), and for other geometries they apply once training enters the routing basin. Our experiments indicate convergence far beyond these budgets, and certifying a budget independent of K for arbitrary geometry is an interesting open problem. Our analysis also assumes one attention layer with a merged query–key matrix, prototype-supported inputs with fixed features, noiseless labels, and population gradients, and extending it to learned representations, label noise, and deeper models is a natural next step.

AI USE STATEMENT

Generative AI tools were used in preparing this manuscript, and we disclose their role in accordance with the ICLR 2027 AI Policy for Authors. Large language models (OpenAI Codex and Anthropic Claude) assisted with (i) drafting and editing parts of the manuscript; (ii) auditing the proofs and numerically re-verifying the stated identities and inequalities; and (iii) implementing and debugging experiment code. We have reviewed all AI-assisted work. We take responsibility for the final content of this work, including text, claims, and artifacts produced with the aid of generative AI.

ETHICS STATEMENT

This work is theoretical and uses only synthetic data generated from the stated model. It involves no human subjects, personal data, or deployed systems, and we foresee no direct ethical concerns.

REPRODUCIBILITY STATEMENT

The model assumptions are stated in Section 2, theorem-specific hypotheses in each statement, and complete proofs of every result appear in Appendices C–F. Appendix H specifies every experimental configuration. The supplementary material contains deterministic scripts that regenerate all figures and tables on a CPU, together with raw result files, fixed seeds, and machine-readable integrity checks.

REFERENCES

Rishabh Agarwal, Avi Singh, Lei Zhang, Bernd Bohnet, Luis Rosias, Stephanie Chan, Biao Zhang, Ankesh Anand, Zaheer Abbas, Azade Nova, et al. Many-shot in-context learning. *Advances in Neural Information Processing Systems*, 37:76930–76966, 2024.

Ekin Akyürek, Dale Schuurmans, Jacob Andreas, Tengyu Ma, and Denny Zhou. What learning algorithm is in-context learning? investigations with linear models. In *International Conference on Learning Representations*, 2023.

Yu Bai, Fan Chen, Huan Wang, Caiming Xiong, and Song Mei. Transformers as statisticians: Provable in-context learning with in-context algorithm selection. *Advances in Neural Information Processing Systems*, 36:57125–57211, 2023.

Alberto Bietti, Vivien Cabannes, Diane Bouchacourt, Herve Jegou, and Leon Bottou. Birth of a transformer: A memory viewpoint. *Advances in Neural Information Processing Systems*, 36: 1560–1588, 2023.

Blake Bordelon, Mary I. Letey, and Cengiz Pehlevan. Theory of scaling laws for in-context regression: Depth, width, context and time. *arXiv preprint arXiv:2510.01098*, 2025.

Etienne Boursier and Claire Boyer. Softmax as linear attention in the large-prompt regime: a measure-based perspective. *arXiv preprint arXiv:2512.11784*, 2025.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *Advances in Neural Information Processing Systems*, 33:1877–1901, 2020.

Stephanie C. Y. Chan, Adam Santoro, Andrew K. Lampinen, Jane X. Wang, Aaditya Singh, Pierre H. Richemond, James McClelland, and Felix Hill. Data distributional properties drive emergent in-context learning in transformers. In *Advances in Neural Information Processing Systems*, 2022.

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

Siyu Chen, Heejune Sheen, Tianhao Wang, and Zhuoran Yang. Unveiling induction heads: Provable training dynamics and feature learning in transformers. In *Advances in Neural Information Processing Systems*, 2024b.

Xiang Cheng, Yuxin Chen, and Suvrit Sra. Transformers implement functional gradient descent to learn non-linear functions in context. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 8002–8037. PMLR, 2024.

Liam Collins, Advait Parulekar, Aryan Mokhtari, Sujay Sanghavi, and Sanjay Shakkottai. In-context learning with transformers: Softmax attention adapts to function lipschitzness. In *Advances in Neural Information Processing Systems*, volume 37, 2024. doi: 10.52202/079017-2942.

Damai Dai, Yutao Sun, Li Dong, Yaru Hao, Shuming Ma, Zhifang Sui, and Furu Wei. Why can gpt learn in-context? language models implicitly perform gradient descent as meta-optimizers. *arXiv preprint arXiv:2212.10559*, 2022.

Sara Dragutinović, Andrew M Saxe, and Aaditya K Singh. Softmax $\geq$ linear: Transformers may learn to classify in-context by kernel gradient descent. *arXiv preprint arXiv:2510.10425*, 2025.

Ezra Edelman, Nikolaos Tsilivis, Benjamin Edelman, Eran Malach, and Surbhi Goel. The evolution of statistical induction heads: In-context learning markov chains. *Advances in Neural Information Processing Systems*, 37:64273–64311, 2024.

Joshua Engels, Eric J. Michaud, Isaac Liao, Wes Gurnee, and Max Tegmark. Not all language model features are one-dimensionally linear. In *International Conference on Learning Representations*, 2025.

Kawin Ethayarajh. How contextual are contextualized word representations? comparing the geometry of BERT, ELMo, and GPT-2 embeddings. In *Conference on Empirical Methods in Natural Language Processing and International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, 2019.

Fabian Falck, Ziyu Wang, and Chris Holmes. Is in-context learning in large language models bayesian? a martingale perspective. *arXiv preprint arXiv:2406.00793*, 2024.

Shivam Garg, Dimitris Tsipras, Percy S Liang, and Gregory Valiant. What can transformers learn in-context? a case study of simple function classes. *Advances in Neural Information Processing Systems*, 35:30583–30598, 2022.

Cole Gibson, Wenping Cui, and Gautam Reddy. Distinct mechanisms underlying in-context learning in transformers. *arXiv preprint arXiv:2604.12151*, 2026.

Gautam Goel, Mahdi Soltanolkotabi, and Peter Bartlett. Training dynamics of softmax self-attention: Fast global convergence via preconditioning. *arXiv preprint arXiv:2603.01514*, 2026.

Jianliang He, Xintian Pan, Siyu Chen, and Zhuoran Yang. In-context linear regression demystified: Training dynamics and mechanistic interpretability of multi-head softmax attention. In *International Conference on Machine Learning*. PMLR, 2025.

Alexander Hsu, Zhaiming Shen, Wenjing Liao, and Rongjie Lai. Understanding in-context learning for nonlinear regression with transformers: Attention as featurizer. *arXiv preprint arXiv:2605.05176*, 2026.

Yu Huang, Yuan Cheng, and Yingbin Liang. In-context convergence of transformers. In *International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 19660–19722. PMLR, 2024.

Samy Jelassi, Michael Sander, and Yuanzhi Li. Vision transformers provably learn spatial structure. *Advances in Neural Information Processing Systems*, 35:37822–37836, 2022.

Hong Jun Jeon, Jason D Lee, Qi Lei, and Benjamin Van Roy. An information-theoretic analysis of in-context learning. *arXiv preprint arXiv:2401.15530*, 2024.

Yanhao Jin, Krishnakumar Balasubramanian, and Lifeng Lai. In-context learning for mixture of linear regressions: Existence, generalization and training dynamics. *arXiv preprint arXiv:2410.14183*, 2024.

Jaeyeon Kim, Sehyun Kwon, Joo Young Choi, Jongho Park, Jaewoong Cho, Jason D. Lee, and Ernest K. Ryu. Task diversity shortens the ICL plateau. *arXiv preprint arXiv:2410.05448*, 2024.

Juno Kim and Taiji Suzuki. Transformers learn nonlinear features in context: Nonconvex mean-field dynamics on the attention landscape. In *International Conference on Machine Learning*, 2024.

Hongkang Li, Meng Wang, Songtao Lu, Xiaodong Cui, and Pin-Yu Chen. How do nonlinear transformers learn and generalize in in-context learning? In *International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 28734–28783. PMLR, 2024.

Hongkang Li, Songtao Lu, Pin-Yu Chen, Xiaodong Cui, and Meng Wang. Training nonlinear transformers for chain-of-thought inference: A theoretical generalization analysis. In *International Conference on Learning Representations*, 2025.

Yuchen Li, Yuanzhi Li, and Andrej Risteski. How do transformers learn topic structure: Towards a mechanistic understanding. In *International Conference on Machine Learning*, pp. 19689–19729. PMLR, 2023.

Ziqian Lin and Kangwook Lee. Dual operating modes of in-context learning. In *Forty-first International Conference on Machine Learning*, 2024.

Yue M Lu, Mary I Letey, Jacob A Zavatone-Veth, Anindita Maiti, and Cengiz Pehlevan. Asymptotic theory of in-context learning by linear attention. *arXiv preprint arXiv:2405.11751*, 2024.

Yuankai Luo, Hongkang Li, Lei Shi, and Xiao-Ming Wu. Enhancing graph transformers with hierarchical distance structural encoding. *Advances in Neural Information Processing Systems*, 37: 57150–57182, 2024.

Arvind Mahankali, Tatsunori B Hashimoto, and Tengyu Ma. One step of gradient descent is provably the optimal in-context learner with one layer of linear self-attention. *arXiv preprint arXiv:2307.03576*, 2023.

Nischal Mainali and Lucas Teixeira. Exact learning dynamics of in-context learning in linear transformers and its application to non-linear transformers. *arXiv preprint arXiv:2504.12916*, 2025.

Ashok Vardhan Makkuva, Marco Bondaschi, Adway Girish, Alliot Nagle, Hyeji Kim, Michael Gastpar, and Chanakya Ekbote. Local to global: Learning dynamics and effect of initialization for transformers. *Advances in Neural Information Processing Systems*, 37:86243–86308, 2024.

Jiaqi Mu, Suma Bhat, and Pramod Viswanath. All-but-the-top: Simple and effective postprocessing for word representations. In *International Conference on Learning Representations*, 2018.

Quan Nguyen and Thanh Nguyen-Tang. How transformers learn in-context recall tasks? optimality, training dynamics and generalization. *arXiv preprint arXiv:2505.15009*, 2025.

Eshaan Nichani, Alex Damian, and Jason D Lee. How transformers learn causal structure with gradient descent. In *International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 38018–38070. PMLR, 2024.

Kazusato Oko, Yujin Song, Taiji Suzuki, and Denny Wu. Pretrained transformer efficiently learns low-dimensional target functions in-context. *Advances in Neural Information Processing Systems*, 37:77316–77365, 2024.

Catherine Olsson, Nelson Elhage, Neel Nanda, Nicholas Joseph, Nova DasSarma, Tom Henighan, Ben Mann, Amanda Askell, Yuntao Bai, Anna Chen, et al. In-context learning and induction heads. *arXiv preprint arXiv:2209.11895*, 2022.

Vardan Papyan, X. Y. Han, and David L. Donoho. Prevalence of neural collapse during the terminal phase of deep learning training. *Proceedings of the National Academy of Sciences*, 117(40): 24652–24663, 2020.

Zhen Qin, Jinxin Zhou, Jiachen Jiang, and Zhihui Zhu. On the convergence of gradient descent on learning transformers with residual connections. *arXiv preprint arXiv:2506.05249*, 2025.

Nived Rajaraman, Marco Bondaschi, Ashok Vardhan Makkuva, Kannan Ramchandran, and Michael Gastpar. Transformers on markov data: Constant depth suffices. *Advances in Neural Information Processing Systems*, 37:137521–137556, 2024.

Gautam Reddy. The mechanistic basis of data dependence and abrupt learning in an in-context classification task. In *International Conference on Learning Representations*, 2024.

Ruifeng Ren and Yong Liu. Towards understanding how transformers learn in-context through a representation learning lens. *Advances in Neural Information Processing Systems*, 37:892–933, 2024.

M. Sagitova, O. Duranthon, and L. Zdeborová. Specialization of softmax attention heads: insights from the high-dimensional single-location model. *arXiv preprint arXiv:2603.03993*, 2026.

Wei Shen, Ruida Zhou, Jing Yang, and Cong Shen. On the training convergence of transformers for in-context classification. *arXiv preprint arXiv:2410.11778*, 2024.

Zhaiming Shen, Alexander Hsu, Rongjie Lai, and Wenjing Liao. Understanding in-context learning on structured manifolds: Bridging attention to kernel methods. *arXiv preprint arXiv:2506.10959*, 2025.

Haoyuan Sun, Ali Jadbabaie, and Navid Azizan. On the role of transformer feed-forward layers in nonlinear in-context learning. *arXiv preprint arXiv:2501.18187*, 2025.

Yuandong Tian, Yiping Wang, Beidi Chen, and Simon S. Du. Scan and snap: Understanding training dynamics and token composition in 1-layer transformer. In *Advances in Neural Information Processing Systems*, 2023.

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

Maria-Luiza Vladarean, Xuhui Zhang, and Suvrit Sra. On learning linear dynamical systems in context with attention layers. In *International Conference on Learning Representations*, 2026.

Johannes von Oswald, Eyvind Niklasson, Ettore Randazzo, João Sacramento, Alexander Mordvintsev, Andrey Zhmoginov, and Max Vladymyrov. Transformers learn in-context by gradient descent. In *International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 35151–35174. PMLR, 2023.

Xinyi Wang, Wanrong Zhu, Michael Saxon, Mark Steyvers, and William Yang Wang. Large language models are latent variable models: Explaining and finding good demonstrations for in-context learning. In *Advances in Neural Information Processing Systems*, volume 36, 2023.

Zixuan Wang, Stanley Wei, Daniel Hsu, and Jason D. Lee. Transformers provably learn sparse token selection while fully-connected nets cannot. In *International Conference on Machine Learning*, 2024.

Sang Michael Xie, Aditi Raghunathan, Percy Liang, and Tengyu Ma. An explanation of in-context learning as implicit bayesian inference. *arXiv preprint arXiv:2111.02080*, 2021.

Tong Yang, Yu Huang, Yingbin Liang, and Yuejie Chi. In-context learning with representations: Contextual generalization of trained transformers. *Advances in Neural Information Processing Systems*, 37, 2024. doi: 10.52202/079017-2727.

Oğuz Kaan Yüksel, Rodrigo Alvarez Lucendo, and Nicolas Flammarion. Incremental learning of sparse attention patterns in transformers. In *International Conference on Machine Learning*, 2026.

Ruiqi Zhang, Spencer Frei, and Peter L Bartlett. Trained transformers learn linear models in-context. *Journal of Machine Learning Research*, 25(49):1–55, 2024.

Yedi Zhang, Aaditya K Singh, Peter E Latham, and Andrew Saxe. Training dynamics of in-context learning in linear attention. In *International Conference on Machine Learning*. PMLR, 2025a.

Yufeng Zhang, Fengzhuo Zhang, Zhuoran Yang, and Zhaoran Wang. What and how does in-context learning learn? bayesian model averaging, parameterization, and generalization. In *International Conference on Artificial Intelligence and Statistics*, volume 258 of *Proceedings of Machine Learning Research*, pp. 1684–1692. PMLR, 2025b.

A Overview of the Appendix

Appendix B compares our setting and results with prior analyses. Appendix C proves the exact decomposition of Section 3 and the task-prior results of Section 2, including a Lipschitz sign family. Appendix D proves the general Gram-spectral theorem, the trap, and the global results beyond symmetry (Theorems 2 and 3) with their extensions. Appendix E proves the symmetric reduction and the global convergence theorem of Section 4. Appendix F proves the phase diagram, its random-count version, and nuisance invariance. Appendix G collects additional discussion, including why pointwise nondegeneracy cannot replace Assumption 1(ii). Appendix H gives experimental details and additional results.

B Extended Related Work

Theories of in-context learning. One line of work interprets ICL as implicit Bayesian inference, formalizing prompting as approximate model averaging over latent hypotheses (Xie et al., 2021; Zhang et al., 2025b; Wang et al., 2023; Falck et al., 2024). A related literature studies transformers as sequence models through Markov chains and induction heads, explaining how attention recovers latent transition structures and token-level recurrence (Nichani et al., 2024; Bietti et al., 2023; Edelman et al., 2024; Rajaraman et al., 2024; Ren & Liu, 2024; Li et al., 2023; Makkuva et al., 2024). The function-learning perspective initiated by Garg et al. (2022) demonstrated that transformers can learn simple function classes from prompts. Subsequent work analyzed expressive power (Akyürek et al., 2023; von Oswald et al., 2023; Cheng et al., 2024; Luo et al., 2024; Sun et al., 2025; Dai et al., 2022; Bai et al., 2023) and generalization (Jelassi et al., 2022; Li et al., 2024; Collins et al., 2024; Oko et al., 2024; Jeon et al., 2024), showing that trained transformers can implement estimators such as ridge regression and gradient descent. Several of these works consider nonlinear tasks (Cheng et al., 2024; Sun et al., 2025; Collins et al., 2024; Oko et al., 2024), but they do not analyze gradient-based attention dynamics; Li et al. (2024) do, for in-context classification with a nonlinear transformer, and we return to them below.

Training dynamics of ICL. Recent work characterizes gradient-descent training of transformers for ICL (Huang et al., 2024; Zhang et al., 2024; Yang et al., 2024; Li et al., 2025; Edelman et al., 2024; Chen et al., 2024a; Zhang et al., 2025a), complementing analyses of converged or Bayes-optimal predictors (Mahankali et al., 2023; Lu et al., 2024; Lin & Lee, 2024). Huang et al. (2024) give the first rigorous analysis of softmax-attention dynamics for linear regression and show that task-aligned tokens come to dominate attention. Their orthonormal linear model avoids both the Gram coupling and the task-contrast cancellation that arise for nonlinear tasks under non-orthogonal features. Subsequent analyses consider contextual regression with representations (Yang et al., 2024) and chain-of-thought over discrete operators (Li et al., 2025), where tasks admit low-dimensional linear representations or finite dictionaries. Other analyses cover multi-head softmax or linear attention (Chen et al., 2024a; Zhang et al., 2025a), classification (Li et al., 2024; Shen et al., 2024; Dragutinović et al., 2025), mixtures of regressions (Jin et al., 2024), recall (Nguyen & Nguyen-Tang, 2025), anisotropic linear regression (He et al., 2025), exact linear-transformer dynamics (Mainali & Teixeira, 2025), and Markov induction circuits (Gibson et al., 2026). Beyond in-context regression, softmax-attention training dynamics have been characterized for next-token prediction, where attention scans toward discriminative tokens (Tian et al., 2023), for sparse token selection (Wang et al., 2024), and for induction heads with feature learning (Chen et al., 2024b); Kim & Suzuki (2024) analyze nonlinear feature learning under linear attention in a mean-field limit.

Table 1 places our setting among prior training-dynamics analyses of in-context learning.

Softmax training dynamics in the past year. Several analyses have since pushed the softmax case further. Goel et al. (2026) prove convergence to globally optimal self-attention parameters at a geometric rate for a softmax layer trained on linear regression, by relating the finite-sample problem to nonconvex matrix factorization and combining preconditioning, regularization, and a data-dependent spectral initialization; their guarantee is global in the loss but tied to that initialization, whereas our trajectories start at $Q^{(0)} = 0$ and the geometry enters through G alone. Sagitova et al. (2026) and Yüksel et al. (2026) establish stage-wise dynamics of a different kind: attention heads start undifferentiated and specialize one after another, in a high-dimensional single-location model and on

Table 1: Training-dynamics analyses of ICL. The last column indicates whether the analysis characterizes how task geometry determines the stage at which a target accuracy is reached.

Reference	Attention	Task class	Non-orthogonal	Accuracy regimes
Zhang et al. (2024)	linear	linear regression	N.A.	×
Li et al. (2024)	softmax	binary classification	×	×
Nichani et al. (2024)	softmax	causal/Markov sequences	N.A.	×
Huang et al. (2024)	softmax	linear regression	×	×
Yang et al. (2024)	softmax	representation regression	✓	×
Chen et al. (2024a)	softmax	multi-task linear	N.A.	×
Li et al. (2025)	softmax	discrete operators	×	×
Goel et al. (2026)	softmax	linear regression	N.A.	×
Sagitova et al. (2026)	softmax	single-location regression	N.A.	×
Yüksel et al. (2026)	softmax	high-order Markov chains	N.A.	×
Ours	softmax	nonlinear, covariance-diverse	✓	✓

high-order Markov chains, the latter with stage-wise convergence proved for a reduced differential system. Their stages index which head learns which pattern, while our two laws index how far the routing of a single head has progressed toward a target accuracy. Boursier & Boyer (2025) show that softmax attention concentrates on a linear operator in the large-prompt limit, with bounds on outputs and gradients, so that linear-attention analyses transfer to softmax for long prompts; we instead keep the softmax nonlinearity at fixed prompt length, which is what creates the K^2 routing bottleneck. Bordelon et al. (2025) derive scaling laws for deep linear attention on in-context regression in depth, width, context, and training time, and Qin et al. (2025) prove linear convergence of gradient descent for a one-layer transformer with residual connections, identifying the conditioning of the attention output matrix as the controlling quantity.

Nonlinear in-context regression beyond dynamics. Nonlinear ICL has mostly been analyzed through what a trained transformer represents rather than how it is trained. Shen et al. (2025) bridge attention to kernel methods for Hölder functions on manifolds, with generalization bounds governed by the intrinsic dimension, and Hsu et al. (2026) construct transformers whose attention realizes polynomial or spline features and bound the resulting regression error. Vladarean et al. (2026) give optimal weight constructions for linear attention learning linear dynamical systems in context. These results describe a solution rather than the trajectory reaching it, which is the gap our analysis addresses for nonlinear priors and non-orthogonal prototypes.

Relation to Huang et al. (2024). With orthonormal prototypes and linear tasks $f(x) = w^\top x$ with $w \sim \mathcal{N}(0, I_d)$, one has $G = I$, $F_f \sim \mathcal{N}(0, I_K)$, and $\Sigma_f = P_\perp$, so the balanced version of this setting is the special case $\gamma = \sigma_f^2 = 1$ of Section 4. In that case our recurrence gives $T_\epsilon = O(K^2/\eta + K/(\eta\epsilon))$ with two-sided bounds for each stage. We do not claim a formal recovery of their theorem, whose conditions and parameterization details differ, and we do not treat their imbalanced-feature regime. Our additions are the exact covariance–Gram update, which isolates how nonlinear task priors and non-orthogonal features enter; the general Gram-spectral theorem; exchangeable nonlinear priors; and the geometry–accuracy phase diagram.

Relation to Collins et al. (2024). Collins et al. show that softmax attention trained for ICL regression adapts its effective window to the Lipschitz constant of the task and the noise level, characterizing a population-risk minimizer. Our results are complementary. We analyze the gradient-descent trajectory rather than its endpoint, under noiseless labels and prototype-organized inputs, and our boundaries describe when a finite accuracy target is reached along that trajectory. An endpoint characterization does not determine the K^2 routing bottleneck, the refinement law, or the dependence of convergence time on the learning-rate policy.

C Proofs for the Exact Decomposition and Task Priors

C.1 Proof of Lemma 2

Proof. Because $\mathbf{1}^\top \pi^{(k)} = \mathbf{1}^\top e_k = 1$, one has $a_k \perp \mathbf{1}$ and hence

$$a_k^\top F_f = a_k^\top P_\perp F_f. \tag{30}$$

Since the task is independent of the prompt, taking the task expectation gives

$$\frac{p_k}{2}\mathbb{E}_f[(a_k^\top F_f)^2] = \frac{p_k}{2} a_k^\top \mathbb{E}_f[(P_\perp F_f)(P_\perp F_f)^\top] a_k = \frac{p_k}{2} a_k^\top \Sigma_f a_k = \ell_k. \tag{31}$$

Summing over k proves (7). Assumption 1(ii) yields

$$\frac{p_k \lambda_f}{2} \|a_k\|_2^2 \le \ell_k \le \frac{p_k \Lambda_f}{2} \|a_k\|_2^2 . \tag{32}$$

Summing (32) proves (8). Under (6), both inequalities are equalities with coefficient $\sigma_f^2/2$. Count averaging preserves every relation by linearity. □

C.2 Proof of Proposition 1

Proof. For fixed w, write $s_k = Se_k = V^\top Q v_k$ for the logits of query class k. The class attention is $\pi^{(k)} = \mathrm{softmax}(s_k + \log w)$, and its Jacobian with respect to s_k is J_k. Differentiating (7) gives

$$\nabla_{s_k} \ell_k = p_k J_k \Sigma_f a_k = h_k, \tag{33}$$

because J_k is symmetric. Since $ds_k = V^\top (dQ) v_k$,

$$d\mathcal{L} = \sum_{k=1}^{K} h_k^\top V^\top (dQ) v_k = \mathrm{tr}\big[(VHV^\top)^\top dQ\big] , \tag{34}$$

which proves (10). Multiplying the update (3) on the left by $V^\top$ and on the right by V gives

$$S^{(t+1)} = S^{(t)} - \eta V^\top V H^{(t)} V^\top V, \tag{35}$$

which is (11). Left multiplication by $(e_k - e_m)^\top$ and right multiplication by e_k prove (12). For random counts, the loss is a finite average over count vectors, so differentiation and count expectation commute, and linearity replaces H by $\mathbb{E}_n[H(n)]$. A class with $n_j = 0$ has weight $w_j = 0$ and simply drops out of the softmax. □

C.3 Proof of Lemma 3

Proof. The kth row of the softmax Jacobian satisfies

$$e_k^\top J_k = \pi_{k|k} e_k^\top - \pi_{k|k} \pi^{(k)\top} = \pi_{k|k} (e_k - \pi^{(k)})^\top = -\pi_{k|k} a_k^\top . \tag{36}$$

Substitution into (9) gives

$$(h_k)_k = -p_k \pi_{k|k} a_k^\top \Sigma_f a_k = -2\pi_{k|k} \ell_k. \tag{37}$$

Because $a_k \perp \mathbf{1}$, Assumption 1(ii) implies

$$a_k^\top \Sigma_f a_k \ge \lambda_f \|a_k\|_2^2 \ge \lambda_f (1 - \pi_{k|k})^2, \tag{38}$$

which proves (13). The fixed-profile quantity in (37) is nonpositive for every count vector, so its count average has the same sign. □

C.4 Proof of Lemma 1

Proof. Let $\bar{F} = P_\perp F_f$ and $M = \mathbb{E}_f[\bar{F}\bar{F}^\top] = \Sigma_f$, which is finite under Assumption 1(ii); no moment assumption on the uncentered vector F_f is needed. Permutation matrices commute with $P_\perp$, so exchangeability of F_f gives $PMP^\top = \mathbb{E}_f[(P\bar{F})(P\bar{F})^\top] = M$ for every permutation matrix P. Hence all diagonal entries of M are equal and, because permutations act transitively on ordered pairs

of distinct indices, all off-diagonal entries are equal. Thus $M = aI + b\mathbf{1}\mathbf{1}^\top$ with $a = M_{11} - M_{12}$. Since $M = P_\perp M P_\perp$ and $P_\perp \mathbf{1} = 0$,

$$\Sigma_f = P_\perp M P_\perp = aP_\perp, \qquad a = \mathbb{E}_f[\bar{F}_1^2] - \mathbb{E}_f[\bar{F}_1\bar{F}_2] = \tfrac{1}{2}\mathbb{E}_f(\bar{F}_1 - \bar{F}_2)^2 = \tfrac{1}{2}\mathbb{E}_f(F_{f,1} - F_{f,2})^2, \tag{39}$$

where the second identity uses $\mathbb{E}_f[\bar{F}_1^2] = \mathbb{E}_f[\bar{F}_2^2]$ and the third uses $\bar{F}_1 - \bar{F}_2 = F_{f,1} - F_{f,2}$. The upper bound follows from $|f(v_1) - f(v_2)| \le L \|v_1 - v_2\|_2$.

For equiangular prototypes and a permutation ϖ of $[K]$, define a linear map on $\mathcal{U} = \operatorname{span}(V)$ by $T_\varpi v_k = v_{\varpi(k)}$. It is well defined because $G \succ 0$, and it preserves inner products because $G_{\varpi(j)\varpi(k)} = G_{jk}$. Extending T_ϖ by the identity on $\mathcal{U}^\perp$ gives an orthogonal map O_ϖ of $\mathbb{R}^d$ with $F_{f\circ O_\varpi} = (f(v_{\varpi(1)}), \dots, f(v_{\varpi(K)}))^\top$. If $f \circ O_\varpi$ and f have the same law, this permutation of F_f has the law of F_f; hence F_f is exchangeable.

For $f(x) = g(w^\top x + b)$ with w uniform on a sphere and independent of b, we have $f \circ O(x) = g((O^\top w)^\top x + b)$, and $O^\top w$ has the law of w. The same argument applies to $\sum_j c_j g(w_j^\top x + b_j)$ with i.i.d. spherical w_j independent of (c_j, b_j). These tasks are Lipschitz with constants $\operatorname{Lip}(g)\|w\|_2$ and $\sum_j |c_j| \operatorname{Lip}(g) \|w_j\|_2$, respectively. The sign family below is exchangeable by construction. □

C.5 A Lipschitz Sign Family

The following family attains a constant contrast fraction and certifies that Assumption 1 holds with $\sigma_f^2 = \Theta(L^2\Delta^2)$. Draw s uniformly from

$$\mathcal{S}_K = \{-1, 1\}^K \setminus \{\mathbf{1}, -\mathbf{1}\}, \tag{40}$$

fix $c_0 \in (0, 1]$, and set

$$2\tau = c_0 L\Delta, \qquad f_s(v_j) = \tau s_j, \tag{41}$$

$$f_s(x) = \min_{j\in[K]} \{\tau s_j + L\|x - v_j\|_2\}. \tag{42}$$

Proposition 5 (Lipschitz sign family)**.** *For equiangular prototypes, each f_s is globally L-Lipschitz, every prototype has another prototype whose value differs by $c_0 L\Delta$, and*

$$\Sigma_f = \sigma_f^2 P_\perp, \qquad \sigma_f^2 = \tau^2 \kappa_K, \qquad \kappa_K = \frac{2^{K-1}}{2^{K-1} - 1}. \tag{43}$$

Its contrast fraction is therefore $c_f = c_0\sqrt{\kappa_K}/2$.

Proof. For every pair of distinct prototypes (the case $j = m$ is trivial),

$$|f_s(v_j) - f_s(v_m)| \le 2\tau = c_0 L\Delta \le L\|v_j - v_m\|_2. \tag{44}$$

Thus the prescribed values form an L-Lipschitz map on the finite prototype set. The function in (42) is a minimum of L-Lipschitz cones and is therefore L-Lipschitz. At a prototype v_i, the finite Lipschitz inequality gives

$$\tau s_j + L\|v_i - v_j\|_2 \ge \tau s_i \tag{45}$$

for every j, while $j = i$ gives equality. Hence the extension agrees with (41).

Every $s \in \mathcal{S}_K$ contains both signs. For every i there is therefore a j with $s_j = -s_i$, and

$$|f_s(v_i) - f_s(v_j)| = 2\tau = c_0 L\Delta. \tag{46}$$

It remains to compute the task second moment. Uniform signs over all 2^K patterns satisfy $\sum_s ss^\top = 2^K I$. Removing $s = \mathbf{1}$ and $s = -\mathbf{1}$ gives

$$\mathbb{E}[ss^\top \mid s \ne \pm\mathbf{1}] = \frac{2^K I - 2\mathbf{1}\mathbf{1}^\top}{2^K - 2}. \tag{47}$$

Multiplication by $P_\perp$ on both sides yields

$$\mathbb{E}[(P_\perp s)(P_\perp s)^\top] = \frac{2^{K-1}}{2^{K-1} - 1} P_\perp. \tag{48}$$

Since $F_f = \tau s$, (43) follows, and $c_f^2 = \sigma_f^2/(L^2\Delta^2) = c_0^2\kappa_K/4$. □

D Proofs for General Gram Matrices

D.1 Smoothness and Descent

Lemma 4 (Smoothness and descent)**.** *Under Assumption 1(ii), the fixed-profile objective is* $6\Lambda_f \|G\|_{\mathrm{op}}^2$*-smooth in* Q*. Consequently, every step size satisfying* (14) *yields* (16)*.*

Proof. For $\pi(z) = \mathrm{softmax}(z)$ and $J(z) = \mathrm{diag}(\pi(z)) - \pi(z)\pi(z)^\top$, the following Euclidean bounds hold:

$$\|\pi(z) - \pi(z')\|_2 \le \|z - z'\|_2 , \tag{49}$$

$$\|J(z)\|_{\mathrm{op}} \le 1, \tag{50}$$

$$\|J(z) - J(z')\|_{\mathrm{op}} \le 3\,\|z - z'\|_2 . \tag{51}$$

The first two follow from the covariance form of J and the mean-value theorem. For the third, $\|\mathrm{diag}(\pi - \pi')\|_{\mathrm{op}} \le \|\pi - \pi'\|_2$ and $\left\|\pi\pi^\top - \pi'\pi'^\top\right\|_{\mathrm{op}} \le (\|\pi\|_2 + \|\pi'\|_2)\,\|\pi - \pi'\|_2 \le 2\,\|\pi - \pi'\|_2$. Also, $\|a_k\|_2 \le \sqrt{2}$ and $\|\Sigma_f\|_{\mathrm{op}} \le \Lambda_f$. For the logit gradient $h_k(z) = p_k J(z)\Sigma_f(\pi(z) - e_k)$,

$$\begin{aligned}\|h_k(z) - h_k(z')\|_2 &\le p_k\Lambda_f\,\|J(z) - J(z')\|_{\mathrm{op}}\,\|a_k(z)\|_2 + p_k\Lambda_f\,\|J(z')\|_{\mathrm{op}}\,\|\pi(z) - \pi(z')\|_2 \\ &\le 6p_k\Lambda_f\,\|z - z'\|_2 .\end{aligned} \tag{52}$$

The full score loss is therefore $6\Lambda_f$-smooth in Frobenius norm.

The linear map $\mathcal{T} : Q \mapsto V^\top QV$ satisfies

$$\|\mathcal{T}(Q)\|_F \le \|V\|_{\mathrm{op}}^2\,\|Q\|_F = \|G\|_{\mathrm{op}}\,\|Q\|_F . \tag{53}$$

Composing a smooth function with a linear map multiplies its smoothness constant by the squared operator norm. Hence the Q-objective is $6\Lambda_f\|G\|_{\mathrm{op}}^2$-smooth. Under (14), the descent lemma gives

$$\mathcal{L}_{t+1} \le \mathcal{L}_t - \frac{\eta}{2}\,\|\nabla_Q \mathcal{L}_t\|_F^2 . \tag{54}$$

Finally, Proposition 1 implies

$$\|\nabla_Q \mathcal{L}_t\|_F^2 = \left\|VH^{(t)}V^\top\right\|_F^2 = \mathrm{tr}\left[(H^{(t)})^\top GH^{(t)}G\right], \tag{55}$$

which proves (16). □

D.2 Proof of Theorem 1

Proof. Monotonicity and (16) follow from Lemma 4. Every column of H lies in $\mathbf{1}^\perp$ because $\mathbf{1}^\top J_k = 0$. Define $B := H^\top GH \succeq 0$. Then

$$\mathrm{tr}(H^\top GHG) = \mathrm{tr}(BG) \ge \lambda_{\min}(G)\,\mathrm{tr}(B), \qquad \mathrm{tr}(B) = \sum_{k=1}^{K} h_k^\top Gh_k \ge \lambda_\perp(G)\,\|H\|_F^2 . \tag{56}$$

Together these inequalities give the factor ω_G. If $H\mathbf{1} = 0$ as well, then $B\mathbf{1} = 0$, so the range of B lies in $\mathbf{1}^\perp$; applying the contrast-subspace Rayleigh bound to $\mathrm{tr}(BG)$ gives the sharper factor $\lambda_\perp(G)^2$. This argument does not require G to preserve $\mathbf{1}^\perp$.

Suppose (15) holds. From (32),

$$\mathcal{L}_{t_0} \ge \ell_k^{(t_0)} \ge \frac{p_{\min}\lambda_f}{2}(1 - \pi_{k|k}^{(t_0)})^2. \tag{57}$$

Thus $\pi_{k|k}^{(t_0)} \ge 1/2$ for every k. Loss descent preserves (15), so the same conclusion holds for all $t \ge t_0$. Lemma 3 now gives

$$|(h_k^{(t)})_k| = 2\pi_{k|k}^{(t)}\ell_k^{(t)} \ge \ell_k^{(t)}. \tag{58}$$

Consequently,

$$\left\|H^{(t)}\right\|_F^2 \geq \sum_{k=1}^{K}(h_k^{(t)})_k^2 \geq \sum_{k=1}^{K}(\ell_k^{(t)})^2 \geq \frac{1}{K}\Big(\sum_{k=1}^{K}\ell_k^{(t)}\Big)^2 = \frac{\mathcal{L}_t^2}{K}. \tag{59}$$

Combining (54), (55), (56), and (59) proves (17).

Let $c = \eta\omega_G/(2K)$. If $\mathcal{L}_{t+1} = 0$, the target is reached. Otherwise, monotonicity and (17) imply

$$\frac{1}{\mathcal{L}_{t+1}} - \frac{1}{\mathcal{L}_t} = \frac{\mathcal{L}_t - \mathcal{L}_{t+1}}{\mathcal{L}_t\mathcal{L}_{t+1}} \geq c\frac{\mathcal{L}_t}{\mathcal{L}_{t+1}} \geq c. \tag{60}$$

Summing (60) from t_0 until the first time the loss is at most ϵ proves the hitting-time bound of Theorem 1. □

D.3 Attention Concentration

The loss bounds the attention error of every class.

Corollary 1 (Attention concentration)**.** *Under Assumption 1(ii), for every fixed profile, iterate, and class,*

$$1 - \pi_{k|k}^{(t)} \leq \sqrt{2\mathcal{L}_t/(p_k\lambda_f)}. \tag{61}$$

Proof. For each k, (32) and $\ell_k \leq \mathcal{L}_t$ give

$$\mathcal{L}_t \geq \frac{p_k\lambda_f}{2}\left\|a_k^{(t)}\right\|_2^2 \geq \frac{p_k\lambda_f}{2}(1 - \pi_{k|k}^{(t)})^2. \tag{62}$$

Rearranging proves (61). □

D.4 The Cone Condition

Let $\lambda_{\text{cone}} = \min_k \min_{b\in\Delta_K,\, b_k=0}(b - e_k)^\top\Sigma_f(b - e_k)$, where Δ_K is the probability simplex. If $\pi_{k|k} < 1$, write $u_k = 1 - \pi_{k|k}$ and $b = (\pi^{(k)} - \pi_{k|k}e_k)/u_k$. Then $b \in \Delta_K$, $b_k = 0$, and $a_k = u_k(b - e_k)$, so

$$\ell_k \geq \frac{p_k\lambda_{\text{cone}}}{2}(1 - \pi_{k|k})^2. \tag{63}$$

In the proofs of Lemma 3, Theorem 1, and Corollary 1, the lower covariance bound enters only through this inequality: it gives the basin property (57) and (62). The smoothness bound uses only $\|\Sigma_f\|_{\text{op}} \leq \Lambda_f$, and (58)–(59) do not use it. The results therefore hold with λ_f replaced by λ_{cone}. The cone condition can hold when Σ_f is singular. If $F_f = Az$ with $\mathbb{E}zz^\top = I_m$ and rows $A_1, \ldots, A_K \in \mathbb{R}^m$, then $(b - e_k)^\top\Sigma_f(b - e_k) = \left\|\sum_j b_jA_j - A_k\right\|_2^2$, so $\lambda_{\text{cone}} > 0$ exactly when no row lies in the convex hull of the others. For example, the vertices of a regular K-gon give $\operatorname{rank}\Sigma_f = 2$ and $\lambda_{\text{cone}} > 0$.

D.5 Proof of Proposition 2

Proof. Write $d_1 = S_{11} - S_{21}$ and $d_2 = S_{22} - S_{12}$, so that under the balanced profile $\pi_{1|1} = \alpha_1 = \text{sigmoid}(d_1)$ and $\pi_{2|2} = \alpha_2 = \text{sigmoid}(d_2)$. With $\Sigma_f = P_\perp$, a direct computation gives $h_1 = r_1(-1, 1)^\top$ and $h_2 = r_2(1, -1)^\top$, where $r_i = 2p_i\alpha_i(1 - \alpha_i)^2$. Hence (11) yields

$$d_1^{(t+1)} - d_1^{(t)} = 5\eta\,(r_1 - 3r_2), \qquad d_2^{(t+1)} - d_2^{(t)} = 5\eta\,(-3r_1 + 10r_2). \tag{64}$$

We show that the cone $\mathcal{C} = \{d_1 \geq 0,\ d_2 \leq -\frac{5}{2}d_1\}$ is invariant. On $\mathcal{C}$, $\alpha_2(1 - \alpha_2)^2 = e^{d_2}/(1 + e^{d_2})^3 \leq e^{d_2} \leq e^{-5d_1/2}$ and $\alpha_1(1 - \alpha_1)^2 = e^{d_1}/(1 + e^{d_1})^3 \geq e^{-2d_1}/8$. Hence $r_2/r_1 \leq 8e^{-d_1/2}p_2/p_1 \leq 0.08$. Consequently $d_1^{(t+1)} - d_1^{(t)} \geq 5\eta r_1(1 - 0.24) > 0$ and

$$\Big(d_2 + \tfrac{5}{2}d_1\Big)^{(t+1)} - \Big(d_2 + \tfrac{5}{2}d_1\Big)^{(t)} = 5\eta\Big(-\tfrac{1}{2}r_1 + \tfrac{5}{2}r_2\Big) \leq 5\eta r_1(-0.5 + 0.2) < 0. \tag{65}$$

Since $S^{(0)} = 0$ lies in $\mathcal{C}$, so does every iterate. The sequence $d_1^{(t)}$ increases; if it had a finite limit, r_1 would stay bounded below and its increments could not vanish. Hence $d_1^{(t)} \to \infty$ and $d_2^{(t)} \leq -\frac{5}{2} d_1^{(t)} \to -\infty$, so $\alpha_1 \to 1$ and $\alpha_2 \to 0$. By Lemma 2, $\mathcal{L}_t = p_1(1-\alpha_1)^2 + p_2(1-\alpha_2)^2 \to p_2$. □

The query imbalance is essential to this example. With uniform queries $p = (\frac{1}{2}, \frac{1}{2})$ and the same G, the loss falls below $3 \times 10^{-3} < p_{\min}\lambda_f/8$ within 3×10^5 iterations of the step (14), so Theorem 1 applies from that time and the loss converges to zero.

D.6 Proof of Theorem 2

Throughout, $\Omega = \{S : S_{kk} \geq S_{jk} \text{ for all } j \neq k\}$ and $u_k = 1 - \pi_{k|k}$; we write $\pi = \pi^{(k)}$ when the query class is clear. We first record the structure of H on Ω.

Lemma 5 (Structure on Ω). *Let $w_j = 1/K$, let $\Sigma_f = \sigma_f^2 P_\perp$, and let $S \in \Omega$. Write $\alpha_m = \pi_m^{(m)}$, $u_m = 1 - \alpha_m$, and $A_m = -(h_m)_m$. Then for every m, (a) $\alpha_m = \max_r \pi_r^{(m)} \geq 1/K$ and $\|\pi^{(m)}\|_2^2 \leq \alpha_m$; (b) $(h_m)_r \geq \sigma_f^2 p_m (\pi_r^{(m)})^2 \geq 0$ for $r \neq m$, so that, with $b_{rm} = (h_m)_r$,*

$$h_m = \sum_{r \neq m} b_{rm}(e_r - e_m), \qquad b_{rm} \geq 0, \qquad \sum_{r \neq m} b_{rm} = A_m = \tfrac{1}{2}\|h_m\|_1 ; \tag{66}$$

(c) $\|a_m\|_2^2 = 1 - 2\alpha_m + \|\pi^{(m)}\|_2^2$, and $u_m^2 \leq \|a_m\|_2^2 \leq \min\{2u_m^2, u_m\}$; (d) $A_m = \sigma_f^2 p_m \alpha_m \|a_m\|_2^2 \leq \sigma_f^2 p_m \alpha_m u_m \leq \sigma_f^2 p_m/4$; (e) if $d_{k,j} = 0$ for some $j \neq k$, then $\pi_j^{(k)} = \alpha_k$ and

$$A_k + b_{jk} = \sigma_f^2 p_k \alpha_k, \qquad A_k - b_{jk} \leq \sigma_f^2 p_k \alpha_k (1 - 2\alpha_k). \tag{67}$$

Proof. With $w_j = 1/K$ we have $\pi_r^{(m)} = e^{S_{rm}} / \sum_q e^{S_{qm}}$, so on Ω the mth coordinate is maximal, $\alpha_m \geq 1/K$, and $\|\pi\|_2^2 \leq (\max_q \pi_q) \sum_q \pi_q = \alpha_m$, which is (a). Isotropy and $a_m \perp \mathbf{1}$ give $h_m = \sigma_f^2 p_m J_m a_m$, hence $(h_m)_r = \sigma_f^2 p_m \pi_r (\pi_r + \alpha_m - \|\pi\|_2^2)$ for $r \neq m$, and (a) gives the sign in (b); the column sums to zero because $\mathbf{1}^\top J_m = 0$, so the off-target entries carry all the mass and $\|h_m\|_1 = 2A_m$, which is (66). Expanding $\|\pi - e_m\|_2^2$ gives the identity in (c), which may also be written $\|a_m\|_2^2 = u_m^2 + \sum_{r \neq m} \pi_r^2$; hence $\|a_m\|_2^2 \geq u_m^2$, while $\sum_{r \neq m} \pi_r^2 \leq (\sum_{r \neq m} \pi_r)^2 = u_m^2$ gives $\|a_m\|_2^2 \leq 2u_m^2$, and (a) gives $\|a_m\|_2^2 \leq 1 - 2\alpha_m + \alpha_m = u_m$. Lemma 3 gives $A_m = \sigma_f^2 p_m \alpha_m \|a_m\|_2^2$, so (c) and $\alpha_m u_m \leq 1/4$ give (d). Finally, if $d_{k,j} = 0$ then $\pi_j^{(k)} = \pi_k^{(k)} = \alpha_k$, so $b_{jk} = \sigma_f^2 p_k \alpha_k (2\alpha_k - \|\pi\|_2^2)$ and $A_k = \sigma_f^2 p_k \alpha_k (1 - 2\alpha_k + \|\pi\|_2^2)$; adding gives the identity in (e), and subtracting, with $\|\pi\|_2^2 \leq \alpha_k$, gives its bound. □

Lemma 6 (Boundary projection). *Let $G_{kk} = 1$ for every k, as after the rescaling in Theorem 2, and let $F_{k,j}(S) = -(g_k - g_j)^\top H(S) g_k$ be the gap velocity, so that $d_{k,j}^{(t+1)} = d_{k,j}^{(t)} + \eta F_{k,j}(S^{(t)})$ by (12). Then $\|Dh_k\|_{\mathrm{op}} \leq 3\Lambda_f p_k$, where Dh_k is the derivative of h_k with respect to the kth score column. Consequently, under (14), if $F_{k,j} \geq 0$ at every state of Ω with $d_{k,j} = 0$, for every pair $j \neq k$, then Ω is forward invariant.*

Proof. Write $a = \pi^{(k)} - e_k$ and $b = J_k \Sigma_f a$. Differentiating $h_k = p_k J_k \Sigma_f a$ gives

$$Dh_k = p_k \big[J_k \Sigma_f J_k + \mathrm{diag}(b) - b\pi^{(k)\top} - \pi^{(k)} b^\top \big]. \tag{68}$$

For every x, $x^\top J_k x = \mathrm{Var}_{r \sim \pi^{(k)}}(x_r) \leq \frac{1}{4}(\max_r x_r - \min_r x_r)^2 \leq \frac{1}{2}\|x\|_2^2$, so $\|J_k\|_{\mathrm{op}} \leq 1/2$; with $\|a\|_2 \leq \sqrt{2}$ this gives $\|b\|_2 \leq \Lambda_f/\sqrt{2}$. The four terms of (68) therefore have operator norms at most $\Lambda_f/4$, $\Lambda_f/\sqrt{2}$, $\Lambda_f/\sqrt{2}$, and $\Lambda_f/\sqrt{2}$, so $\|Dh_k\|_{\mathrm{op}} \leq p_k \Lambda_f (\frac{1}{4} + \frac{3}{\sqrt{2}}) < 3\Lambda_f p_k$.

Fix $S \in \Omega$ and $j \neq k$, and let $\bar{S}$ agree with S except that $\bar{S}_{jk} = S_{kk}$; then $\bar{S} \in \Omega$ and $d_{k,j}(\bar{S}) = 0$. Only the kth column changes, and only by $d_{k,j}(S)$ in one coordinate, so only h_k changes,

by at most $3\Lambda_f p_k d_{k,j}(S)$ in Euclidean norm. Since $G_{kk} = 1$, the induced change in $F_{k,j}$ is $-(g_k - g_j)^\top [h_k(S) - h_k(\bar{S})]$, whence

$$|F_{k,j}(S) - F_{k,j}(\bar{S})| \le 3\sqrt{2}\,\Lambda_f p_k \,\|G\|_{\mathrm{op}}\, d_{k,j}(S). \tag{69}$$

Using (14) and $\|G\|_{\mathrm{op}} \ge G_{kk} = 1$,

$$d_{k,j}(S^+) = d_{k,j}(S) + \eta F_{k,j}(S) \ge \eta F_{k,j}(\bar{S}) + \Big(1 - \frac{p_k}{\sqrt{2}\,\|G\|_{\mathrm{op}}}\Big) d_{k,j}(S) \ge 0. \tag{70}$$

Induction over t gives the claim. □

Proof of Theorem 2. Write $g_k = Ge_k = e_k + \epsilon_k$, where $(\epsilon_k)_k = 0$ and $\|\epsilon_k\|_\infty \le \bar{\rho}$. Because G is symmetric, one step changes $d_{k,j} = S_{kk} - S_{jk}$ by $\Delta d_{k,j} = -\eta(g_k - g_j)^\top H g_k$, that is,

$$\begin{aligned}\Delta d_{k,j} = {} & \underbrace{-\eta\big[(h_k)_k - (h_k)_j\big]}_{A} \underbrace{-\eta(\epsilon_k - \epsilon_j)^\top h_k}_{C} \\ & \underbrace{-\eta \sum_{m \ne k} (\epsilon_k)_m \big[(h_m)_k - (h_m)_j\big]}_{D_1} \underbrace{-\eta \sum_{m \ne k} (\epsilon_k)_m (\epsilon_k - \epsilon_j)^\top h_m}_{D_2}.\end{aligned} \tag{71}$$

By Lemma 5(b), $A = \eta[A_k + b_{jk}]$, while $\|\epsilon_k - \epsilon_j\|_\infty \le 2\bar{\rho}$ and (66) give

$$|C| \le 2\eta\bar{\rho}\,\|h_k\|_1 = 4\eta\bar{\rho}A_k. \tag{72}$$

For $m \notin \{k, j\}$ both k and j are off-target for query m, so $|(h_m)_k| + |(h_m)_j| = b_{km} + b_{jm} \le A_m$; for $m = j$ the same sum is at most $b_{kj} + A_j \le 2A_j$. With Lemma 5(d),

$$|D_1| \le \eta\bar{\rho}\Big[\sum_{m \ne k} A_m + A_j\Big] \le \frac{\eta\bar{\rho}\sigma_f^2}{4}\big[(1-p_k) + p_j\big], \qquad |D_2| \le 4\eta\bar{\rho}^2 \sum_{m \ne k} A_m \le \eta\bar{\rho}^2\sigma_f^2(1-p_k). \tag{73}$$

Part (i). By Lemma 6 it suffices to bound $F_{k,j}$ below on the face $d_{k,j} = 0$, where Lemma 5(e) gives $A_k + b_{jk} = \sigma_f^2 p_k \alpha_k$ and hence $A_k \le \sigma_f^2 p_k \alpha_k$. Dividing (72)–(73) by $\eta\sigma_f^2 p_k \alpha_k$ and using $\alpha_k \ge 1/K$ and $p_k \ge p_{\min}$,

$$\frac{F_{k,j}}{\sigma_f^2 p_k \alpha_k} \;\ge\; 1 - 4\bar{\rho} - \frac{\bar{\rho}K}{p_{\min}}\Big(\frac{1}{2} + \bar{\rho}\Big). \tag{74}$$

Since $\bar{\rho} \le p_{\min}/(9K) \le 1/(9K^2) \le 1/36$, the right side is at least $1 - \frac{1}{9} - \frac{1}{9}(\frac{1}{2} + \frac{1}{36}) = 1 - \frac{55}{324} > \frac{4}{5}$. Every face velocity is therefore positive, so Ω is forward invariant by Lemma 6. As $S^{(0)} = 0 \in \Omega$, every iterate lies in Ω, and Lemma 5(a) gives $\pi^{(t)}_{k|k} \ge 1/K$.

Part (ii). Each column of H sums to zero, so Cauchy–Schwarz gives $\sum_{r \ne k} (h_k)_r^2 \ge (h_k)_k^2/(K-1)$ and hence $\|h_k\|_2^2 \ge \frac{K}{K-1}(h_k)_k^2$. Lemma 3 and $\pi_{k|k} \ge 1/K$ give $|(h_k)_k| = 2\pi_{k|k}\ell_k \ge 2\ell_k/K$, so

$$\|H\|_F^2 \ge \frac{4}{K(K-1)} \sum_{k=1}^{K} \ell_k^2 \ge \frac{4\mathcal{L}^2}{K^2(K-1)}. \tag{75}$$

Combining (54), (55), and (56) gives $\mathcal{L}_{t+1} \le \mathcal{L}_t - c_R\mathcal{L}_t^2$; dividing by $\mathcal{L}_t\mathcal{L}_{t+1}$ and summing yields the closed form.

Part (iii). No gap moves far in one step: $\|h_m\|_2 \le p_m\sigma_f^2\,\|J_m\|_{\mathrm{op}}\,\|a_m\|_2 \le p_m\sigma_f^2/\sqrt{2}$, so $\|H\|_F \le \sigma_f^2/\sqrt{2}$ and

$$|\Delta d_{k,j}| \le \eta\,\|g_k - g_j\|_2\,\|H\|_{\mathrm{op}}\,\|g_k\|_2 \le \sqrt{2}\,\eta\,\|G\|_{\mathrm{op}}^2\,\|H\|_F \le \eta\sigma_f^2\,\|G\|_{\mathrm{op}}^2 \le \frac{1}{6}. \tag{76}$$

Suppose $\alpha_k \le 2/3$. Then $u_k \ge 1/3$, so Lemma 5(c),(d) give $A_k \ge \sigma_f^2 p_k \alpha_k u_k^2 \ge \sigma_f^2 p_k/(9K)$, while $p_j \le 1-(K-1)p_{\min}$ gives $(1-p_k)+p_j \le 2-Kp_{\min}$. Dividing (72)–(73) by ηA_k, the total relative error is at most

$$4\bar\rho + \frac{9K\bar\rho}{4p_{\min}}\big[2-Kp_{\min}+4\bar\rho(1-p_{\min})\big] \le \frac{1}{2}, \tag{77}$$

because the left side increases in $\bar\rho$ and, at $\bar\rho = p_{\min}/(9K)$ with $x = Kp_{\min} \le 1$, is at most $\frac{1}{2} + x[\frac{5}{9K^2} - \frac{1}{4}] \le \frac{1}{2}$. Hence $\Delta d_{k,j} \ge \eta A_k/2$ for every $j \ne k$ whenever $\alpha_k \le 2/3$.

Let $R_k = u_k/\alpha_k = \sum_{j\ne k} e^{-d_{k,j}}$. While $R_k \ge 1$ we have $\alpha_k \le 1/2$ and $u_k \ge 1/2$, and every gap of column k grows by at least $x_k := \eta A_k/2 \ge \eta\sigma_f^2 p_k \alpha_k u_k^2/2$, so $R_k^+ \le R_k e^{-x_k}$. Since $x_k \le \eta\sigma_f^2 p_k/8 \le 1$, the bound $1-e^{-x} \ge x/2$ gives

$$R_k - R_k^+ \ge \frac{R_k x_k}{2} \ge \frac{\eta\sigma_f^2 p_k u_k^3}{4} \ge \frac{\eta\sigma_f^2 p_k}{32}. \tag{78}$$

At $S^{(0)} = 0$ we have $R_k = K-1$, so $R_k \le 1$ within $\lceil 32(K-2)/(\eta\sigma_f^2 p_{\min})\rceil$ steps, and one further step gives $\alpha_k \ge 1/2$. Majority persists: if $\alpha_k \in [1/2, 2/3]$ then every gap of column k increases, so R_k decreases; and if $\alpha_k > 2/3$ then $R_k < 1/2$, so (76) gives $R_k^+ \le e^{1/6} R_k < e^{1/6}/2 < 1$. Once $\alpha_k \ge 1/2$ for every k, Lemma 3 gives $|(h_k)_k| \ge \ell_k$, and the display in part (ii) gives $\|H\|_F^2 \ge \frac{K}{K-1}\sum_k \ell_k^2 \ge \mathcal{L}^2/(K-1)$, which is (17) with $2K$ replaced by $2(K-1)$; summing reciprocals as in Theorem 1 gives the stated tail time.

Finally, $\bar\rho \le 1/(9K^2)$ and Gershgorin's theorem give $\lambda_{\min}(G) \ge 1-(K-1)\bar\rho \ge 17/18$ and $\|G\|_{\mathrm{op}} \le 19/18$, so $\omega_G \ge \lambda_{\min}(G)^2 \ge (17/18)^2$. □

D.7 The Sharp Correlation Threshold

Theorem 2 assumes $\bar\rho \le p_{\min}/(9K)$ because part (iii) needs the margin (77). Forward invariance of Ω by itself holds under the weaker budget $\bar\rho \le 2p_{\min}/K$, and no budget of this form can exceed $4/K^2$.

Proposition 6 (Invariance threshold, sharp up to a factor of two)**.** *Assume the hypotheses of Theorem 2 with (18) replaced by $\bar\rho \le 2p_{\min}/K$ and the same step size. Then Ω is forward invariant, $\pi_{k|k}^{(t)} \ge 1/K$ for all k and t, and (19) holds with $\omega_G \ge 1/4$.*

Proof. By Lemma 6 it suffices to prove $F_{k,j} \ge 0$ on the face $d_{k,j} = 0$. Put $r = G_{kj}$ and $v = g_k - g_j$, so $v_k = 1-r$, $v_j = r-1$, and $|v_m| \le 2\bar\rho$ otherwise. Expanding $F_{k,j} = -\sum_m G_{mk} v^\top h_m$ and using $G_{kk} = 1$ gives three groups.

For the target column, (66) and $|v_m| \le 2\bar\rho$ for $m \notin \{k,j\}$ give $-v^\top h_k \ge (1-r)(A_k + b_{jk}) - 2\bar\rho(A_k - b_{jk})$, so (67) and $\alpha_k \ge 1/K$ yield

$$-v^\top h_k \ge \sigma_f^2 p_k \alpha_k\big[1-3\bar\rho+4\bar\rho\alpha_k\big] \ge \frac{\sigma_f^2 p_k}{K}\Big[1-\Big(3-\frac{4}{K}\Big)\bar\rho\Big]. \tag{79}$$

For column j, every coordinate of v lies between v_j and v_k, so $v^\top h_j = \sum_{l\ne j} b_{lj}(v_l - v_j) \in [0, 2(1-r)A_j]$ and hence $-G_{jk} v^\top h_j \ge -2\bar\rho(1-\bar\rho)A_j$. For $m \notin \{k,j\}$, writing $x = G_{mk}$ and $y = G_{mj}$ so that $v_m = x - y$, each coefficient obeys $-x(v_l - v_m) = x^2 - xy - xv_l \ge -\bar\rho(1+\bar\rho)$: for $l \in \{k,j\}$ because $x^2 - |x|(1+2\bar\rho)$ decreases on $[0,\bar\rho]$, and otherwise because $x^2 - 3\bar\rho|x| \ge -2\bar\rho^2$. Summing against $\sum_{l\ne m} b_{lm} = A_m$ and applying Lemma 5(d),

$$\frac{F_{k,j}}{\sigma_f^2} \ge \frac{p_k}{K}\Big[1-\Big(3-\frac{4}{K}\Big)\bar\rho\Big] - \frac{\bar\rho}{4}\big[(1+\bar\rho)(1-p_k) + (1-3\bar\rho)p_{\max}\big]. \tag{80}$$

For $K \ge 3$ we have $\bar\rho \le 2/K^2 \le 1/3$, and substituting $p_{\max} \le 1 - p_{\min}$ makes the right side of (80) at least $f(\bar\rho) = \frac{p_{\min}}{K} - [\frac{p_{\min}}{K}(3-\frac{4}{K}) + \frac{1-p_{\min}}{2}]\bar\rho + \frac{1-p_{\min}}{2}\bar\rho^2$, which decreases on $[0, 2p_{\min}/K]$ and satisfies

$$f\Big(\frac{2p_{\min}}{K}\Big) = \frac{p_{\min}^2}{K}\Big(1 - \frac{4}{K} + \frac{8}{K^2} - \frac{2p_{\min}}{K}\Big) > 0, \tag{81}$$

since $p_{\min} \le 1/K$ and $1 - 4/K + 6/K^2 > 0$. For $K = 2$ the face velocity is $F_{k,j} = 2(1-r)(A_k - rA_j)$ with $A_k = \sigma_f^2 p_k/4$ and $A_j \le \sigma_f^2 p_j/4$, so $r \le p_{\min}/p_{\max}$ suffices, and $|r| \le p_{\min}$ implies this. Gershgorin now gives $\lambda_{\min}(G) \ge 1 - 2(K-1)p_{\min}/K \ge 1/2$ and $\|G\|_{\mathrm{op}} \le 3/2$, so $\omega_G \ge 1/4$; part (ii) of Theorem 2 used only $\pi_{k|k} \ge 1/K$ and therefore still applies. □

Proposition 7 (The budget cannot exceed $4/K^2$)**.** *Let $p = \mathbf{1}/K$, fix k, and let $G = I + \bar{\rho}\sum_{m\neq k}(e_k e_m^\top + e_m e_k^\top)$, which is positive definite exactly when $\bar{\rho}\sqrt{K-1} < 1$. At states of Ω with $\pi^{(k)} = \mathbf{1}/K$ and $\pi^{(m)} \to (e_k + e_m)/2$ for $m \neq k$,*

$$\frac{F_{k,j}}{\sigma_f^2} = \frac{[K - 2\bar{\rho}(K-1)](4 - \bar{\rho}K^2)}{4K^3}, \qquad j \neq k, \tag{82}$$

which is negative for $4/K^2 < \bar{\rho} < 1/\sqrt{K-1}$. Hence Ω is not forward invariant for such G, and the budget of Proposition 6 is optimal up to a factor of two for uniform queries.

Proof. Direct evaluation of $F_{k,j} = -(g_k - g_j)^\top H g_k$ at the stated attention vectors gives (82); replacing each zero attention entry by a small mixture with $\mathbf{1}/K$ realizes these values by finite scores in Ω and preserves the strict sign. □

D.8 Anisotropic Covariance and Imbalanced Profiles

The isotropy and balance assumptions of Theorem 2 can both be relaxed. Neither extension can rely on the off-target signs of Lemma 5(b): for $K = 3$, $\pi = (\frac{1}{2}, \frac{1}{2} - \varepsilon, \varepsilon)$, $v = (-\frac{1}{2} + \frac{1}{\sqrt{12}}, \frac{1}{2} + \frac{1}{\sqrt{12}}, -\frac{2}{\sqrt{12}})$ and $\Sigma_f = P_\perp + \delta v v^\top$ give $(h_1)_3/(p_1\pi_3^{(1)}) \to -\delta\sqrt{3}/4 < 0$ for a class-1 query with attention $\pi^{(1)} = \pi$ as $\varepsilon \to 0$, for every $\delta > 0$. The boundary-projection lemma nevertheless applies, because on a face the softmax Jacobian acts diagonally.

Proposition 8 (Anisotropic task covariance)**.** *Under the remaining hypotheses of Theorem 2 ($G = I + E \succ 0$ with $E_{kk} = 0$ and $\max_{j\neq k}|E_{jk}| \le \bar{\rho}$, arbitrary p), let $w = \mathbf{1}/K$, let $\lambda P_\perp \preceq \Sigma_f \preceq 2\lambda P_\perp$, let $\bar{\rho} \le p_{\min}/(6K)$, and let $0 < \eta \le 1/(6\Lambda_f \|G\|_{\mathrm{op}}^2)$. Then Ω is forward invariant from $Q^{(0)} = 0$, $\pi_{k|k}^{(t)} \ge 1/K$, and (19) holds.*

Proof. Write $B = \Sigma_f - \lambda P_\perp$, so $0 \preceq B \preceq \lambda P_\perp$. On the face $d_{k,j} = 0$ put $v = e_k - e_j$; then $\pi_k = \pi_j = \alpha$ gives $J_k v = \alpha v$ and $v^\top a_k = -1$, while Lemma 5(c) gives $\|a_k\|_2^2 \le 1 - \alpha \le 1$. Hence $-v^\top h_k = -p_k\alpha\, v^\top \Sigma_f a_k = p_k\alpha(\lambda - v^\top B a_k)$. For $0 \preceq C \preceq I$ we have $x^\top C y \le \frac{1}{2}(x^\top y + \|x\|\,\|y\|)$, applied with $C = B/\lambda$, $\|v\| = \sqrt{2}$:

$$-v^\top h_k \ge p_k\alpha\Big[\lambda - \frac{\lambda}{2}(\sqrt{2}-1)\Big] = \frac{3-\sqrt{2}}{2}\lambda p_k \alpha. \tag{83}$$

Since $\|J_m x\|_1 = \mathbb{E}_{\pi^{(m)}}|x_r - \mathbb{E}x| \le \|x\|_2$, splitting h_m into its $\lambda P_\perp$ and B parts gives $\|h_m\|_1 \le 4\lambda p_m \alpha_m u_m^2 + \sqrt{2}\lambda p_m u_m \le M\lambda p_m$ with $M = \frac{16}{27} + \sqrt{2}$. The transported errors are at most $2\bar{\rho}\,\|h_k\|_1 + \bar{\rho}(1+2\bar{\rho})\sum_{m\neq k}\|h_m\|_1$, which after division by $\lambda p_k \alpha \ge \lambda p_k/K$ and $\bar{\rho} \le p_{\min}/(6K) \le 1/24$ is at most $\frac{M}{6} + \frac{13M}{72} = \frac{25M}{72} < 0.697 < \frac{3-\sqrt{2}}{2}$. Every face velocity is positive, so Lemma 6 gives invariance; Lemma 3 is unchanged, so part (ii) of Theorem 2 applies verbatim. □

Proposition 9 (Imbalanced context profile)**.** *Under the remaining hypotheses of Theorem 2 ($G = I + E \succ 0$ with $E_{kk} = 0$ and $\max_{j\neq k}|E_{jk}| \le \bar{\rho}$, arbitrary p, $\Sigma_f = \sigma_f^2 P_\perp$, and the step (18)), let $0 \le \zeta \le 1/4$ and $w_j \in [(1-\zeta)/K, (1+\zeta)/K]$ for every j, and let $\bar{\rho} \le (1-\zeta)p_{\min}/(8K)$. Then Ω is forward invariant from $Q^{(0)} = 0$, the attention floor is $\pi_{k|k}^{(t)} \ge (1-\zeta)/K$, and (19) holds with c_R multiplied by $(1-\zeta)^2$.*

Proof. On Ω, $\pi_k^{(k)} \ge w_k \ge (1-\zeta)/K$, and on the face $d_{k,j} = 0$ the ratio $r = \pi_j^{(k)}/\pi_k^{(k)} = w_j/w_k$ lies in $[\frac{3}{5}, \frac{5}{3}]$. With $\alpha_k = \pi_k^{(k)}$, a direct expansion gives $-(e_k - e_j)^\top J_k a_k = \alpha_k B_0$, so the exact

inward drift $-(e_k - e_j)^\top h_k$ equals $\sigma_f^2 p_k \alpha_k B_0$ with

$$B_0 = 1 + (r^2 + r - 2)\alpha_k + (1 - r)\left\|\pi^{(k)}\right\|_2^2. \tag{84}$$

If $r \geq 1$ then $\|\pi\|_2^2 \leq 1$ gives $B_0 \geq 2 - r \geq 1/3$; if $r \leq 1$ then dropping the last term and using $\alpha_k \leq 1/(1+r)$ gives $B_0 \geq (r^2 + 2r - 1)/(1+r) \geq 7/20 > 1/3$. Off-target gradients may now be negative, but their total absolute mass obeys $\|h_m\|_1 \leq \sigma_f^2 p_m[4\alpha_m u_m^2 + 2(\kappa - 1)\alpha_m u_m] \leq M\sigma_f^2 p_m$ with $\kappa = \frac{1+\zeta}{1-\zeta} \leq \frac{5}{3}$ and $M = \frac{16}{27} + \frac{\kappa - 1}{2} \leq \frac{25}{27}$. Dividing the transported errors by $\sigma_f^2 p_k \alpha_k \geq \sigma_f^2 p_k (1 - \zeta)/K$ bounds them by $\frac{M}{8} + \frac{17M}{128} = \frac{33M}{128} < \frac{1}{3}$, so every face velocity is positive and Lemma 6 applies. The floor $(1-\zeta)/K$ replaces $1/K$ in part (ii), which multiplies c_R by $(1-\zeta)^2$. □

The $1/K$ floor is genuinely unavailable here: a class with $w_k < 1/K$ already has $\pi_{k|k} < 1/K$ at initialization.

D.9 Global Convergence under Transitive Geometry

Throughout this subsection queries are uniform, the profile is balanced, $\Sigma_f = \sigma_f^2 P_\perp$, and $G = \gamma I + \rho \mathbf{1}\mathbf{1}^\top + E \succ 0$ with $\gamma > 0$, E symmetric, and $E_{kk} = 0$; Ω, α_k, u_k, A_k, and b_{rk} are as in Lemma 5, and $c = \gamma^{-1} \max_k \sum_{m \neq k} |E_{km}|$. A permutation matrix U acts on scores by $S \mapsto USU^\top$, which relabels classes simultaneously for queries and contexts. We write $\lambda_{\perp,\max}(G)$ for the largest eigenvalue of G on $\mathbf{1}^\perp$; since $\|E\|_{\mathrm{op}} \leq c\gamma$, every eigenvalue of G on $\mathbf{1}^\perp$ lies in $[\gamma(1-c), \gamma(1+c)]$.

Lemma 7 (Equivariance and exact synchronization)**.** *For every permutation matrix U, $H(USU^\top) = UH(S)U^\top$. If $UGU^\top = G$ for every U in a group Γ acting transitively on $[K]$, then $US_tU^\top = S_t$ for all t and $U \in \Gamma$. Consequently all A_k coincide, $H_t\mathbf{1} = 0$, $G\mathbf{1} \propto \mathbf{1}$, and*

$$GH_tG = (\gamma I + E)H_t(\gamma I + E), \tag{85}$$

where H_t and GH_tG are doubly centered.

Proof. With uniform p and w, relabeling classes permutes the columns of Π and their entries together with the targets e_k, so $\mathcal{L}(USU^\top) = \mathcal{L}(S)$. Since $H = \nabla_S \mathcal{L}$ and $S \mapsto USU^\top$ is an isometry, the chain rule gives $H(USU^\top) = UH(S)U^\top$. Now $S_0 = 0$ is invariant, and if $US_tU^\top = S_t$ then $US_{t+1}U^\top = S_t - \eta(UGU^\top)(UH_tU^\top)(UGU^\top) = S_{t+1}$. Transitivity maps any diagonal entry of H_t to any other, so all A_k are equal; the vectors $H_t\mathbf{1}$ and $G\mathbf{1}$ are Γ-invariant, hence constant, and $\mathbf{1}^\top H_t\mathbf{1} = 0$ forces $H_t\mathbf{1} = 0$. Therefore $\mathbf{1}\mathbf{1}^\top H_t = 0 = H_t\mathbf{1}\mathbf{1}^\top$, which gives (85), and $G\mathbf{1} \propto \mathbf{1}$ together with $\mathbf{1}^\top H_t = 0 = H_t\mathbf{1}$ makes GH_tG doubly centered. □

Lemma 8 (Descent on the contrast subspace)**.** *Suppose $H_t\mathbf{1} = 0$ and $G\mathbf{1} \propto \mathbf{1}$ for all t, and let $0 < \eta \leq K/(6\sigma_f^2 \lambda_{\perp,\max}(G)^2)$. Then*

$$\mathcal{L}_{t+1} \leq \mathcal{L}_t - \frac{\eta}{2}\operatorname{tr}\left[H_t^\top GH_tG\right] \leq \mathcal{L}_t - \frac{\eta\lambda_\perp(G)^2}{2}\|H_t\|_F^2. \tag{86}$$

Proof. Let $A = P_\perp G P_\perp$ and $X = P_\perp S$. The loss and H depend on S only through X, and $X_{t+1} = X_t - \eta AH_tA$ because $P_\perp H_t = H_t$, $H_t\mathbf{1} = 0$, and G preserves $\mathbf{1}^\perp$. On doubly centered matrices the map $Y \mapsto AYA$ is self-adjoint and positive definite with eigenvalues $\lambda_i(A)\lambda_j(A)$, so $\|AYA\|_F^2 \leq \lambda_{\perp,\max}(G)^2 \langle Y, AYA\rangle$ and $\langle Y, AYA\rangle \geq \lambda_\perp(G)^2 \|Y\|_F^2$. By Lemma 6 each h_k is $3\sigma_f^2 p_k = 3\sigma_f^2/K$-Lipschitz in its own score column, and h_k depends on no other column, so $\nabla\mathcal{L} = H$ is $\ell = 3\sigma_f^2/K$-Lipschitz in Frobenius norm. The descent lemma gives

$$\begin{aligned}\mathcal{L}(X - \eta AHA) &\leq \mathcal{L}(X) - \eta\langle H, AHA\rangle + \frac{\ell\eta^2}{2}\|AHA\|_F^2 \\ &\leq \mathcal{L}(X) - \eta\Big(1 - \frac{\ell\eta\lambda_{\perp,\max}(G)^2}{2}\Big)\langle H, AHA\rangle,\end{aligned} \tag{87}$$

and $\ell\eta\lambda_{\perp,\max}(G)^2 \leq \frac{1}{2}$. Finally $\langle H, AHA\rangle = \operatorname{tr}(H^\top GHG)$ for doubly centered H. □

Proof of Theorem 3. By Lemma 7, all A_k equal a common value A_t, and (85) holds. Suppose $S_t \in \Omega$. Expanding (85) in the gap velocity $F_{k,j} = -(e_k - e_j)^\top G H_t G e_k$,

$$F_{k,j} = \gamma^2(A_t + b_{jk}) - \gamma(e_k - e_j)^\top E h_k - \gamma(e_k - e_j)^\top H_t E e_k - (e_k - e_j)^\top E H_t E e_k. \quad (88)$$

By Lemma 5, every entry of H_t has absolute value at most A_t, while the absolute row and column sums of E are at most $c\gamma$. The three error terms are therefore bounded by $2c\gamma^2 A_t$, $2c\gamma^2 A_t$, and $2c^2\gamma^2 A_t$, and since $b_{jk} \geq 0$,

$$F_{k,j} \geq \gamma^2 A_t\,(1 - 4c - 2c^2) = \theta_c \gamma^2 A_t > 0, \quad (89)$$

because $A_t = \sigma_f^2 \alpha_k \|a_k\|_2^2 / K > 0$ at every finite state. Starting from $S_0 = 0 \in \Omega$, induction shows that every gap increases strictly at every step; hence $S_t \in \Omega$ for all t, and majority attention persists once reached.

Routing. Let $R_k = \sum_{j \neq k} e^{-d_{k,j}} = u_k/\alpha_k$. While $R_k \geq 1$ we have $u_k \geq \frac{1}{2}$, and every gap of column k grows by at least $x_t := \eta\theta_c\gamma^2 A_t$, so $R_k^+ \leq R_k e^{-x_t}$. Lemma 5(d) gives $A_t \leq \sigma_f^2/(4K)$, and $\lambda_{\perp,\max}(G) \geq \gamma(1-c)$ gives $x_t \leq \theta_c/(24(1-c)^2) < 1$; with $1 - e^{-x} \geq x/2$ on $[0,1]$ and $A_t \geq \sigma_f^2 \alpha_k u_k^2/K$,

$$R_k - R_k^+ \geq \frac{R_k x_t}{2} \geq \frac{\eta\theta_c\gamma^2\sigma_f^2 u_k^3}{2K} \geq \frac{\eta\theta_c\gamma^2\sigma_f^2}{16K}. \quad (90)$$

Since $R_k = K - 1$ at $t = 0$, $R_k \leq 1$ within $\lceil 16K(K-2)/(\eta\theta_c\gamma^2\sigma_f^2)\rceil$ steps, and one further step gives $\alpha_k \geq \frac{1}{2}$.

Tail. Lemma 8 applies by Lemma 7. Once $\alpha_k \geq \frac{1}{2}$ for every k, part (iii) of the proof of Theorem 2 gives $\|H_t\|_F^2 \geq \mathcal{L}_t^2/(K-1)$, and (86) yields the stated recurrence. □

The equiangular family is the case $E = 0$, for which Proposition 3 gives the exact reduction; weighted rings and tori—prototypes indexed by a cyclic or product group with distance-dependent correlations—are non-equiangular transitive examples. Numerically the certified budget is conservative: on circulant geometries with $K \leq 32$, $\rho \leq 0.95$, and $c \leq \frac{1}{2}$, every gap increased at every step (300 runs), and within the certified budget $c \leq \frac{1}{5}$ the routing time stayed below 0.12 of its bound. Transitivity is essential to this proof, since it forces $H_t\mathbf{1} = 0$ exactly; for arbitrary E the invariant region, monotone loss, and the symmetric routing law are observed empirically whenever $c \leq \frac{1}{2}$ (Appendix H.4), but a proof must control the row sums $H_t\mathbf{1}$ along the trajectory, and we leave the corresponding absolute budget as a conjecture.

D.10 Global Convergence on an Open Neighborhood of the Symmetric Family

Proposition 10 (Open neighborhood)**.** *Fix the balanced profile, uniform queries, $\Sigma_f = \sigma_f^2 P_\perp$, an equiangular Gram matrix G_0, and $0 < \delta < \lambda_{\min}(G_0)$. Let $0 < \eta \leq 1/(6\sigma_f^2(\|G_0\|_{\mathrm{op}} + \delta)^2)$, and let T_0 be the first time at which the symmetric trajectory for G_0 satisfies $\mathcal{L} \leq \sigma_f^2/(16K)$. Set $M = 6\sigma_f^2(\|G_0\|_{\mathrm{op}} + \delta)^2$ and, for symmetric E with $\|E\|_{\mathrm{op}} \leq \delta$, $D(E) = \sqrt{2}\sigma_f^2\,\|E\|_{\mathrm{op}}\,(2\,\|G_0\|_{\mathrm{op}} + \|E\|_{\mathrm{op}})$. If $G = G_0 + E$ satisfies*

$$\frac{D(E)}{M}\left[(1+\eta M)^{T_0} - 1\right] \leq \frac{1}{16\sqrt{2}\,K}, \quad (91)$$

then gradient descent with G from $Q^{(0)} = 0$ reaches the basin (15) by time T_0, and Theorem 1 applies afterwards.

Proof. Since $\delta < \lambda_{\min}(G_0)$, $G \succ 0$, and the step satisfies (14) for both G and G_0; in particular $\eta\sigma_f^2\gamma^2 \leq 1/6$, so the symmetric trajectory for G_0 converges and T_0 is finite. Let S_t and S_t^0 be the score iterates for G and G_0, and $\Delta_t = S_t - S_t^0$. The score gradient $H(S)$ is $6\sigma_f^2$-Lipschitz in Frobenius norm (Lemma 4) and satisfies $\|H(S)\|_F \leq \sqrt{2}\sigma_f^2$, because $\|h_k\|_2 \leq p_k\,\|J_k\|_{\mathrm{op}}\,\sigma_f^2\,\|a_k\|_2 \leq \sqrt{2}p_k\sigma_f^2$. Using $GHG - G_0HG_0 = EHG_0 + G_0HE + EHE$,

$$\begin{aligned}\|\Delta_{t+1}\|_F &\leq \|\Delta_t\|_F + \eta\left\|G(H(S_t) - H(S_t^0))G\right\|_F + \eta\left\|GH(S_t^0)G - G_0H(S_t^0)G_0\right\|_F \\ &\leq (1+\eta M)\,\|\Delta_t\|_F + \eta D(E). \end{aligned} \quad (92)$$

With $\Delta_0 = 0$, induction gives $\|\Delta_t\|_F \le \frac{D(E)}{M}[(1+\eta M)^t - 1]$. The score loss is $\sqrt{2}\sigma_f^2$-Lipschitz because its gradient is H. Therefore (91) gives $\mathcal{L}(S_{T_0}) \le \mathcal{L}(S^0_{T_0}) + \sigma_f^2/(16K) \le \sigma_f^2/(8K) = p_{\min}\lambda_f/8$, which is (15). Theorem 1 applies from $t_0 = T_0$. □

Because $D(E)$ is proportional to $\|E\|_{\mathrm{op}}$ for small perturbations, (91) contains a nonempty open neighborhood. Its radius can be exponentially small in T_0, so this result shows that global convergence is not an artifact of exact symmetry, not that it holds for moderately asymmetric geometry; the latter is examined numerically in Appendix H.

E Proofs for Symmetric Geometry

E.1 Permutation-Symmetric Score Trajectory

This subsection and the next prove Proposition 3. Under uniform queries, isotropic task covariance, and an exchangeable count law, the population objective, viewed as a function of S, satisfies $\mathcal{L}(PSP^\top) = \mathcal{L}(S)$ for every permutation matrix P, and $G = \gamma I + \rho \mathbf{1}\mathbf{1}^\top$ commutes with every P. Hence the score gradient is equivariant, $H(PSP^\top) = PH(S)P^\top$, and the autonomous update $S \mapsto S - \eta G H(S) G$ maps the commutant $\{S : PSP^\top = S \text{ for all } P\} = \mathrm{span}\{I, \mathbf{1}\mathbf{1}^\top\}$ into itself. Starting from $S^{(0)} = 0$, every iterate therefore satisfies

$$S^{(t)} = d_t P_\perp + c_t \mathbf{1}\mathbf{1}^\top, \tag{93}$$

where the column shift $c_t \mathbf{1}\mathbf{1}^\top$ is softmax-neutral. By equivariance, $H^{(t)}$ also lies in the commutant. Its columns sum to zero because $\mathbf{1}^\top J_k = 0$, which forces $H^{(t)} = h_t P_\perp$ for a scalar h_t. In particular,

$$\mathbf{1}^\top H^{(t)} = 0, \qquad H^{(t)}\mathbf{1} = 0, \tag{94}$$

and, using $G = \gamma I + \rho \mathbf{1}\mathbf{1}^\top$,

$$G H^{(t)} G = \gamma^2 H^{(t)} = \frac{\Delta^4}{4} H^{(t)}. \tag{95}$$

This reduction is a consequence of the exact Gram-coupled update, not an orthogonal approximation. It also shows that, at fixed γ, the common correlation ρ is dynamically neutral, as discussed after Proposition 3.

E.2 Exactly Balanced Recurrence

Suppose that $n_j = N/K$ for every j. For query class k, symmetry gives

$$\pi^{(t)}_{k|k} = \alpha_t, \qquad \pi^{(t)}_{m|k} = \frac{1-\alpha_t}{K-1} \quad (m \ne k). \tag{96}$$

With $u_t = 1 - \alpha_t$,

$$\left\| a_k^{(t)} \right\|_2^2 = \frac{K}{K-1} u_t^2. \tag{97}$$

Lemma 3, $p_k = 1/K$, and isotropy yield

$$(h_k^{(t)})_k = -\frac{\sigma_f^2}{K-1} \alpha_t u_t^2. \tag{98}$$

The off-target coordinates of h_k are equal and sum to $-(h_k)_k$. Combining this fact with (95) gives the exact gap recurrence

$$d_{t+1} - d_t = \eta \sigma_f^2 \gamma^2 \frac{K}{(K-1)^2} \alpha_t u_t^2. \tag{99}$$

At initialization, $\alpha_0 = 1/K$ and $u_0 = (K-1)/K$, so

$$d_1 - d_0 = \frac{\eta \sigma_f^2 \gamma^2}{K^2}, \tag{100}$$

which proves (21). Define $R_t = (K-1)e^{-d_t}$. Then

$$\alpha_t = \frac{1}{1+R_t}, \qquad u_t = \frac{R_t}{1+R_t}. \tag{101}$$

Let

$$\beta := \eta\sigma_f^2\gamma^2\frac{K}{(K-1)^2}, \qquad \phi(R) := \frac{R^2}{(1+R)^3}. \tag{102}$$

Equation (99) becomes

$$R_{t+1} = R_t \exp[-\beta\phi(R_t)], \tag{103}$$

and the exact balanced population loss is

$$\mathcal{L}_t = \frac{\sigma_f^2 K}{2(K-1)}u_t^2. \tag{104}$$

This completes the proof of Proposition 3.

E.3 Comparison for Multinomial Context Counts

For $0 < \zeta \le 1/2$, let $\mathcal{E}_\zeta$ be the event that $|n_j/N - 1/K| \le \zeta/K$ for all j, which by the Chernoff and union bounds fails with probability at most $\nu := 2K\exp(-\zeta^2 N/(3K))$. Theorem 4 assumes the prompt-length condition

$$\nu \le c_\nu \min\left\{\frac{1}{K}, \frac{\epsilon}{\sigma_f^2}\right\}, \qquad c_\nu = \frac{1}{100}, \tag{105}$$

which holds when $N \ge 3K\zeta^{-2}\log[2K/(c_\nu \min\{K^{-1}, \epsilon/\sigma_f^2\})]$. A condition of this type is necessary, because a query class is missing from the prompt with probability $(1-1/K)^N$, so the loss never falls below $\frac{\sigma_f^2}{2}(1-\frac{1}{K})^N$.

For a count vector n with $n_k > 0$, define

$$r_{k,n} := \frac{N-n_k}{n_k}, \qquad c_{k,n} := \frac{r_{k,n}}{K-1}. \tag{106}$$

On $\mathcal{E}_\zeta$ with $\zeta \le 1/2$, one has $c_- \le c_{k,n} \le c_+$ with $c_\pm$ as in Proposition 11, and in particular

$$\frac{1}{3} \le c_{k,n} \le 3. \tag{107}$$

Along (93), the count-conditioned target odds, attention, and residual are

$$R_{k,n} = c_{k,n}R_t, \qquad \alpha_{k,n} = \frac{1}{1+R_{k,n}}, \qquad u_{k,n} = \frac{R_{k,n}}{1+R_{k,n}}. \tag{108}$$

For any allocation of off-target attention mass,

$$u_{k,n}^2 \le \|a_{k,n}\|_2^2 \le 2u_{k,n}^2, \tag{109}$$

so that

$$\phi(R_{k,n}) \le \alpha_{k,n}\|a_{k,n}\|_2^2 \le 2\phi(R_{k,n}). \tag{110}$$

Moreover, for $c \in [1/3, 3]$,

$$\frac{1}{9}\phi(R) \le \phi(cR) \le 9\phi(R). \tag{111}$$

By exchangeability, the count-averaged $H^{(t)}$ keeps the form $h_t P_\perp$. Repeating the target-coordinate calculation gives

$$d_{t+1} - d_t = A\,\mathbb{E}_n\left[\alpha_{k,n}\|a_{k,n}\|_2^2\right], \qquad A := \frac{\eta\sigma_f^2\gamma^2}{K-1}. \tag{112}$$

Since $\alpha\|a\|_2^2 \le 2\alpha(1-\alpha)^2 \le 8/27$, (107)–(111) and $\mathbb{P}(\mathcal{E}_\zeta^c) \le \nu$ imply

$$\frac{A(1-\nu)}{9}\phi(R_t) \le d_{t+1} - d_t \le A\left[18\phi(R_t) + \frac{8\nu}{27}\right]. \tag{113}$$

Assume

$$\nu \le c_\nu \min\left\{\frac{1}{K}, \frac{\epsilon}{\sigma_f^2}\right\}, \qquad c_\nu \le \frac{1}{100}. \tag{114}$$

Then the exceptional-count term is dominated by $\phi(R_t)$ at every step before the stopping rule $R_t \le R_* := \sqrt{\epsilon/(36\sigma_f^2)}$ used in the proof of Theorem 4. In coarse routing, $1 \le R_t \le K-1$ and (116) give $\phi(R_t) \ge 1/(8K)$, whereas $8\nu/27 \le 8c_\nu/(27K)$. In the refinement region $R_* < R_t \le 1$, one has $\phi(R_t) \ge R_t^2/8 > \epsilon/(288\sigma_f^2)$, whereas $8\nu/27 \le 8c_\nu\epsilon/(27\sigma_f^2)$. In both cases $8\nu/27 \le \phi(R_t)$, and therefore

$$\frac{(1-\nu)A}{9}\phi(R_t) \le d_{t+1} - d_t \le 19A\,\phi(R_t) \tag{115}$$

until the stopping rule is met.

E.4 Coarse Routing and Inverse-Square Tail

Initially, $R_0 = K-1$. For $R \ge 1$,

$$\frac{1}{8R} \le \phi(R) \le \frac{1}{R}. \tag{116}$$

Under (22), the one-step exponent is uniformly bounded. Equations (115) and (116), together with $x/2 \le 1 - e^{-x} \le x$ for $0 \le x \le 1$, imply

$$R_t - R_{t+1} = \Theta_\zeta(A) \quad \text{while } R_t \ge 1. \tag{117}$$

Therefore

$$T_{\mathrm{I}} = \Theta_\zeta\left(\frac{K}{A}\right) = \Theta_\zeta\left(\frac{K^2}{\eta\sigma_f^2\gamma^2}\right) \tag{118}$$

for multinomial counts and $K \ge 3$. In the exactly balanced case, the same statement holds with absolute constants (Appendix F).

For the exact balanced recurrence in the tail $R_t \le 1$, define

$$Y_t := u_t^{-2} = \frac{(1+R_t)^2}{R_t^2}, \qquad x_t := \beta\phi(R_t). \tag{119}$$

Equation (103) yields

$$Y_{t+1} - Y_t = R_t^{-2}(e^{2x_t} - 1) + 2R_t^{-1}(e^{x_t} - 1). \tag{120}$$

The upper overshoot bound needed below also holds on the step that enters the tail. Indeed, $\phi(R) \le 4/27$ for every $R > 0$, and $e^z - 1 \le ze^z$ gives

$$Y_{t+1} - Y_t \le 2\beta\left[\frac{e^{8/27}}{(1+R_t)^3} + \frac{e^{4/27}R_t}{(1+R_t)^3}\right] < 4\beta. \tag{121}$$

The step-size condition gives $\beta \le 1$, and $R_t \le 1$ gives $0 \le x_t \le 1/8$. Using $e^z - 1 \ge z$ for the lower bound and $e^z - 1 \le ze^z$ for the upper bound,

$$\frac{\beta}{4} \le Y_{t+1} - Y_t \le 7\beta. \tag{122}$$

For $y_2 > y_1 \ge 4$, let $\vartheta(y) = \inf\{t : Y_t \ge y\}$. The one-step overshoot is at most 7β, so

$$\frac{y_2 - y_1}{7\beta} - 1 \le \vartheta(y_2) - \vartheta(y_1) \le \frac{4(y_2 - y_1)}{\beta} + 1. \tag{123}$$

The constants are not optimized; a direct computation gives $\beta/2 \le Y_{t+1} - Y_t \le 2e^{1/4}\beta$ for $R_t \le 1$.

Under random counts, the drift $x_t = d_{t+1} - d_t$ lies in the range (115) before the stopping rule. Repeating the two estimates with this range shows that, while $R_t \le 1$, Y_t increases by at least $A/20$ and at most $90A$ per step. The same two-sided order therefore holds with β replaced by constant multiples of A.

E.5 Proof of Theorem 4

Proof. Along the symmetric trajectory let $T_{\mathrm{I}} = \inf\{t : R_t \le 1\}$. For $K \ge 3$, the coarse interval contributes $T_{\mathrm{I}} = O_\zeta(K^2/(\eta\sigma_f^2\gamma^2))$ by (118). When $K = 2$, $R_0 = 1$ and $T_{\mathrm{I}} = 0$.

Let $R_* = \sqrt{\epsilon/(36\sigma_f^2)} < 1$ and $\mathcal{L}_{\mathrm{bal}}(R) = \frac{\sigma_f^2 K}{2(K-1)} \frac{R^2}{(1+R)^2} \le \sigma_f^2 R^2$. By Lemma 9, $\mathcal{L}_t \le c_+^2 \frac{K-1}{K-1-\zeta} \mathcal{L}_{\mathrm{bal}}(R_t) + \nu\sigma_f^2$. Since $\zeta \le 1/2$ gives $c_+ \le 3$ and $(K-1)/(K-1-\zeta) \le 2$, and since $\nu\sigma_f^2 \le c_\nu \epsilon$, we obtain $\mathcal{L}_t \le 18\sigma_f^2 R_t^2 + c_\nu\epsilon \le \epsilon$ whenever $R_t \le R_*$.

After T_{I}, $Y_t = (1 + R_t^{-1})^2$ increases by at least $A/20$ per step until $R_t \le R_*$. The latter holds once $Y_t \ge (1 + R_*^{-1})^2$, which is implied by $Y_t \ge 4R_*^{-2} = 144\sigma_f^2/\epsilon$. Starting from $Y_{T_{\mathrm{I}}} \ge 4$, at most

$$\frac{20 \cdot 144\, \sigma_f^2}{A\, \epsilon} + 1 = O\Big(\frac{K}{\eta\gamma^2\epsilon}\Big) \tag{124}$$

further iterations are needed, because $A = \eta\sigma_f^2\gamma^2/(K-1)$. The factor σ_f^2 cancels because the required routing residual is itself of order ϵ/σ_f^2. Adding the two contributions proves (23). The alternative forms follow from $\gamma = \Delta^2/2$ and $\sigma_f^2 = c_f^2 L^2 \Delta^2$. □

E.6 Non-Uniform Query Classes with Orthogonal Features

Remark 1 (Non-uniform query classes). For orthogonal features $G = \gamma I$ with balanced contexts and isotropic covariance, the columns of S decouple and each remains symmetric, so class k follows (20) with its own clock $\beta_k = Kp_k\beta$. Imbalanced query frequencies therefore rescale per-class times by $1/(Kp_k)$. For $\rho \ne 0$, non-uniform p couples the columns, and Proposition 2 shows that strong imbalance combined with adverse geometry can prevent convergence.

For $G = \gamma I$, (11) reads $S^{(t+1)} = S^{(t)} - \eta\gamma^2 H^{(t)}$, and h_k depends only on the kth column of S, so the columns evolve independently. With a balanced profile and $\Sigma_f = \sigma_f^2 P_\perp$, a column of the form $d\,(e_k - \mathbf{1}/K) + c\,\mathbf{1}$ gives equal off-target attention; then a_k, $\Sigma_f a_k = \sigma_f^2 a_k$, and h_k have equal off-target coordinates, so this form is preserved from $S^{(0)} = 0$. Repeating (98)–(99) with p_k in place of $1/K$ gives

$$d_{k,t+1} - d_{k,t} = Kp_k\, \beta\, \alpha_{k,t} u_{k,t}^2. \tag{125}$$

Class k therefore follows (20) with clock $\beta_k = Kp_k\beta$, and $\mathcal{L}_t = \sum_k p_k \frac{\sigma_f^2 K}{2(K-1)} u_{k,t}^2$. For equiangular G with $\rho \ne 0$, $GHG = \gamma^2 H + \gamma\rho\, H\mathbf{1}\mathbf{1}^\top$ whenever $\mathbf{1}^\top H = 0$, and $H\mathbf{1} \ne 0$ for non-uniform p, so the columns couple and the scalar reduction fails.

F Proofs for the Phase Diagram

F.1 Proof of Theorem 5 and Proposition 4

Proof. Under isotropic covariance, Lemma 2 gives

$$\mathcal{L}_t = \frac{\sigma_f^2}{2}\mathcal{R}_t = \frac{\sigma_f^2 K}{2(K-1)} u_t^2. \tag{126}$$

At initialization, $u_0 = (K-1)/K$, which gives the first identity in (25). At the reference state $R = 1$, one has $u = 1/2$ and $Y = 4$, which gives the second. Dividing these two losses by ϵ yields q_{init} and q_{tail} in (26). In particular, $q \le q_{\mathrm{init}}$ is equivalent to $\mathcal{L}_0 \le \epsilon$, which proves part (a).

For $q > q_{\mathrm{init}}$, solving $\mathcal{L}_t = \epsilon$ in (126) gives exactly the target state u_ϵ, with odds R_ϵ and potential Y_ϵ, defined before Theorem 5. Since the loss is increasing in R_t, $T_{\mathrm{acc}} = \inf\{t : R_t \le R_\epsilon\}$. Moreover, $R_\epsilon \ge 1$ if and only if $q \le q_{\mathrm{tail}}$, so $T_{\mathrm{acc}} \le T_{\mathrm{I}}$ in part (b) and $R_\epsilon < 1$ in part (c).

It remains to bound the hitting times. For $R_t \ge 1$, write $x_t = \beta\phi(R_t)$. The step-size condition implies $0 \le x_t \le 1$, while

$$\frac{1}{8} \le R_t\phi(R_t) = \frac{R_t^3}{(1+R_t)^3} \le 1. \tag{127}$$

Since $x/2 \le 1 - e^{-x} \le x$ on this interval, the balanced recurrence gives

$$\frac{\beta}{16} \le R_t - R_{t+1} \le \beta \qquad (R_t \ge 1). \tag{128}$$

Because $q > q_{\text{init}}$ also implies $T_{\text{acc}} \ge 1$, summing until the first crossing of $R_\epsilon \in [1, K-1)$ yields

$$\max\left\{1, \frac{K-1-R_\epsilon}{\beta}\right\} \le T_{\text{acc}} \le \frac{16(K-1-R_\epsilon)}{\beta} + 1, \tag{129}$$

which gives the order in part (b). The same argument with $R_\epsilon = 1$ proves (24), and (122) proves part (ii) of Proposition 4. At $q = q_{\text{tail}}$, $\epsilon = \mathcal{L}_{\text{cross}}$ and the exact loss identity gives $\{\mathcal{L}_t \le \epsilon\} = \{R_t \le 1\}$, so the two stopping times coincide even when the discrete update overshoots the reference state.

If $q > q_{\text{tail}}$, then $Y_\epsilon > 4$. Applying (123) with $y_1 = 4$ and $y_2 = Y_\epsilon$ gives

$$\frac{Y_\epsilon - 4}{7\beta} - 1 \le T_{\text{acc}} - T_{\text{I}} \le \frac{4(Y_\epsilon - 4)}{\beta} + 1. \tag{130}$$

Finally,

$$\frac{Y_\epsilon - 4}{\beta} = \frac{Kq - 8(K-1)}{2(K-1)} \cdot \frac{(K-1)^2}{\eta\sigma_f^2\gamma^2 K} = \frac{K-1}{2\eta\gamma^2\epsilon}\left(1 - \frac{q_{\text{tail}}}{q}\right), \tag{131}$$

which proves (28). When $Y_\epsilon - 4 \ge 14\beta$, the lower bound in (130) is at least $(Y_\epsilon - 4)/(14\beta)$, which establishes the matching order. For $K = 2$, $q_{\text{init}} = q_{\text{tail}} = 4$ and $R_0 = 1$, so the coarse-accuracy regime is empty. □

F.2 Lipschitz Form of the Boundaries

Corollary 2 (Lipschitz form of the boundaries)**.** *For equiangular prototypes and an L-Lipschitz exchangeable prior with contrast fraction c_f, the boundaries of Theorem 5 occur at*

$$L_j = \frac{\sqrt{q_j\,\epsilon}}{c_f\Delta}, \qquad j \in \{\text{init}, \text{tail}\}, \tag{132}$$

so that

$$\frac{c_f L_j \Delta}{\sqrt{\epsilon}} = \sqrt{q_j} = \Theta(1) \qquad \textit{and} \qquad \frac{L_{\text{tail}}}{L_{\text{init}}} = \frac{2(K-1)}{K}. \tag{133}$$

Since $c_f \le 1/\sqrt{2}$, every such prior needs $L\Delta > \sqrt{2q_{\text{tail}}\epsilon}$ to place its accuracy target in the refinement region.

Proof. By definition of the contrast fraction, $\sigma_f^2 = c_f^2 L^2\Delta^2$, so $q = c_f^2 L^2\Delta^2/\epsilon$ is increasing in L. Setting $q = q_j$ gives (132). The first identity in (133) is a rearrangement, $q_j \in [2, 8)$ for $K \ge 2$, and $L_{\text{tail}}^2/L_{\text{init}}^2 = q_{\text{tail}}/q_{\text{init}} = 4(K-1)^2/K^2$. Finally, Lemma 1 gives $\sigma_f^2 \le L^2\Delta^2/2$, so $q > q_{\text{tail}}$ forces $L^2\Delta^2 > 2q_{\text{tail}}\epsilon$. For the sign family, $c_f = c_0\sqrt{\kappa_K}/2$ by (43). □

F.3 Random Prompt Counts

Let $\mathcal{L}_{\text{bal}}(R) = \frac{\sigma_f^2 K}{2(K-1)}\frac{R^2}{(1+R)^2}$ denote the balanced loss at odds R.

Proposition 11 (Phase boundaries under random counts)**.** *In the setting of Theorem 4, the initial loss is exactly $\mathcal{L}_0 = \frac{\sigma_f^2(K-1)}{2K}(1 + \frac{1}{N})$, so $T_{\text{acc}} = 0$ if and only if $q \le q_{\text{init}}N/(N+1)$. Let $\nu_\zeta = \Pr(\mathcal{E}_\zeta^{\text{c}})$, $c_\pm = \frac{K-1\pm\zeta}{(1\mp\zeta)(K-1)}$, $\kappa_- = (1-\nu_\zeta)c_-^2$, and $\kappa_+ = c_+^2\frac{K-1}{K-1-\zeta}$, so that $\kappa_\pm \to 1$ as $\zeta, \nu_\zeta \to 0$. If $q\,(\kappa_+/q_{\text{tail}} + \nu_\zeta) \le 1$, then $T_{\text{acc}} \le T_{\text{I}}$. If $q > q_{\text{tail}}/\kappa_-$, then $\mathcal{L}_t > \epsilon$ whenever $R_t > 1$, so the accuracy target lies in the refinement region.*

Lemma 9 (Loss sandwich under random counts)**.** *In the setting of Theorem 4, let $R_t = (K-1)e^{-d_t}$ along the symmetric trajectory, $\zeta \le 1/2$, and $\nu_\zeta = \Pr(\mathcal{E}_\zeta^{\text{c}})$. Then*

$$(1-\nu_\zeta)c_-^2\,\mathcal{L}_{\text{bal}}(R_t) \le \mathcal{L}_t \le c_+^2\frac{K-1}{K-1-\zeta}\mathcal{L}_{\text{bal}}(R_t) + \nu_\zeta\sigma_f^2. \tag{134}$$

Proof. Fix $n \in \mathcal{E}_\zeta$ and a query class k. By (108), $u_{k,n} = c_{k,n}R_t/(1 + c_{k,n}R_t)$. On $\mathcal{E}_\zeta$, $n_k/N \in [(1-\zeta)/K, (1+\zeta)/K]$, so $c_{k,n} \in [c_-, c_+]$ with $c_- \le 1 \le c_+$. Since $u(R) = R/(1+R)$ is increasing, $u(cR) \le c\,u(R)$ for $c \ge 1$ and $u(cR) \ge c\,u(R)$ for $c \le 1$; hence $c_- u_t \le u_{k,n} \le c_+ u_t$. The off-target mass $u_{k,n}$ is split in proportion to $\omega_m = n_m/(N - n_k)$, whose squares sum to between $1/(K-1)$ and $\max_{m \ne k} \omega_m \le (1+\zeta)/(K-1-\zeta)$. Therefore

$$\frac{K}{K-1} u_{k,n}^2 \le \|a_{k,n}\|_2^2 \le \frac{K}{K-1-\zeta} u_{k,n}^2. \tag{135}$$

Averaging $\frac{\sigma_f^2}{2}\|a_{k,n}\|_2^2$ over uniform k gives a profile loss between $c_-^2 \mathcal{L}_{\mathrm{bal}}(R_t)$ and $c_+^2 \frac{K-1}{K-1-\zeta}\mathcal{L}_{\mathrm{bal}}(R_t)$ on $\mathcal{E}_\zeta$. On $\mathcal{E}_\zeta^{\mathrm{c}}$, the profile loss lies in $[0, \sigma_f^2]$ because $\|a\|_2^2 \le 2$. Averaging over counts proves (134). □

Proof of Proposition 11. At $Q^{(0)} = 0$, $\pi^{(k)} = w$ for every k, so $\|a_k\|_2^2 = 1 - 2w_k + \|w\|_2^2$ and the profile loss is $\frac{\sigma_f^2}{2}(1 - \frac{2}{K} + \|w\|_2^2)$. For multinomial counts, $\mathbb{E}n_j^2 = \frac{N}{K}(1 - \frac{1}{K}) + \frac{N^2}{K^2}$, hence $\mathbb{E}\|w\|_2^2 = \frac{1}{K} + \frac{K-1}{KN}$ and $\mathcal{L}_0 = \frac{\sigma_f^2(K-1)}{2K}(1 + \frac{1}{N})$. This proves the first claim.

If $q(\kappa_+/q_{\mathrm{tail}} + \nu_\zeta) \le 1$, then at $t = T_{\mathrm{I}}$ we have $R_t \le 1$, so $\mathcal{L}_{\mathrm{bal}}(R_t) \le \mathcal{L}_{\mathrm{bal}}(1) = \sigma_f^2/q_{\mathrm{tail}}$, and Lemma 9 gives $\mathcal{L}_{T_{\mathrm{I}}} \le \sigma_f^2(\kappa_+/q_{\mathrm{tail}} + \nu_\zeta) \le \epsilon$. Hence $T_{\mathrm{acc}} \le T_{\mathrm{I}}$. If instead $q > q_{\mathrm{tail}}/\kappa_-$ and $R_t > 1$, then $\mathcal{L}_{\mathrm{bal}}(R_t) > \sigma_f^2/q_{\mathrm{tail}}$, and the lower bound in Lemma 9 gives $\mathcal{L}_t > \kappa_- \sigma_f^2/q_{\mathrm{tail}} > \epsilon$. □

F.4 Task-Irrelevant Continuous Variation

The phase diagram does not require the inputs to lie exactly on the prototypes, as long as the variation carries no task information. Let $\mathcal{U} = \mathrm{span}(V)$ and $x_i = v_{z_i} + \xi_i$ with $\xi_i \in \mathcal{U}^\perp$. Assume that the token pairs (z_i, ξ_i), $i \in [N] \cup \{\mathrm{q}\}$, are mutually independent and independent of the task, with $\mathbb{E}[\xi_i \mid z_i] = 0$ and finite second moments, and that labels are constant along nuisance fibers, $f(v_k + \xi) = f(v_k)$. The independence across tokens is needed, since a context nuisance correlated with the query class injects a nonzero nuisance gradient.

Theorem 6 (Nuisance invariance)**.** *Under these conditions and $Q^{(0)} = 0$, for every t,*

$$Q^{(t)} = P_{\mathcal{U}} Q^{(t)} P_{\mathcal{U}}, \tag{136}$$

$$x_i^\top Q^{(t)} x_{\mathrm{q}} = v_{z_i}^\top Q^{(t)} v_{z_{\mathrm{q}}}. \tag{137}$$

Hence the population trajectory coincides with the prototype trajectory for any nuisance scale.

The result is exact for label-irrelevant variation orthogonal to the prototype span, and task-relevant variation is examined empirically in Appendix H.7.

Proof. We proceed by induction. The claim (136) holds at $t = 0$. Suppose it holds at time t. Since every nuisance vector lies in $\mathcal{U}^\perp$,

$$x_i^\top Q^{(t)} x_{\mathrm{q}} = (v_{z_i} + \xi_i)^\top P_{\mathcal{U}} Q^{(t)} P_{\mathcal{U}} (v_{z_{\mathrm{q}}} + \xi_{\mathrm{q}}) = v_{z_i}^\top Q^{(t)} v_{z_{\mathrm{q}}}. \tag{138}$$

Hence attention weights, predictions, and residuals depend only on the class indices and the task. The single-prompt gradient is $\sum_i \chi_i\, x_i x_{\mathrm{q}}^\top$, where $\chi_i = (\widehat{y} - f(x_{\mathrm{q}}))a_i(f(x_i) - \widehat{y})$ depends only on the class indices and the task. Because the pairs (z_i, ξ_i) are independent across tokens and of the task, $\mathbb{E}[\xi_i \mid z_1, \dots, z_N, z_{\mathrm{q}}, f] = \mathbb{E}[\xi_i \mid z_i] = 0$, and ξ_i is conditionally independent of ξ_{q}. With finite second moments, this gives

$$\mathbb{E}\big[x_i x_{\mathrm{q}}^\top \mid z_1, \dots, z_N, z_{\mathrm{q}}, f\big] = v_{z_i} v_{z_{\mathrm{q}}}^\top. \tag{139}$$

Averaging the single-prompt gradient over nuisances therefore produces exactly the prototype gradient $VHV^\top$, which maps $\mathcal{U}^\perp$ to zero and has range in $\mathcal{U}$. The gradient update preserves (136), and induction proves the claim and (137).

For completeness, fiber-constant tasks can be globally Lipschitz. Let $\mathcal{M}$ be the union of the class supports. For different classes,

$$\|(v_k + \xi) - (v_m + \xi')\|_2^2 = \|v_k - v_m\|_2^2 + \|\xi - \xi'\|_2^2 \ge \|v_k - v_m\|_2^2, \tag{140}$$

where orthogonality removes the cross term. Labels are equal within a fiber. If the prototype values satisfy $|f(v_k) - f(v_m)| \le L \|v_k - v_m\|_2$, as they do for every L-Lipschitz task, the labels define an L-Lipschitz map on $\mathcal{M}$, and the McShane extension

$$\widetilde{f}(x) = \inf_{a \in \mathcal{M}} \{f(a) + L \|x - a\|_2\} \tag{141}$$

is globally L-Lipschitz and agrees with the fiber-constant labels on the whole support. The task covariance on prototype values is unchanged. □

G Additional Discussion

G.1 Why Pointwise Nondegeneracy Does Not Prevent Cancellation

A natural task assumption requires every task to be L-Lipschitz and every prototype to have another prototype whose value differs at the Lipschitz scale. Such pointwise conditions control magnitudes but not signs. Take $K = 3$ equiangular prototypes with pairwise distance Δ and the task values

$$f(v_1) = 0, \qquad f(v_2) = \frac{L\Delta}{2}, \qquad f(v_3) = -\frac{L\Delta}{2}. \tag{142}$$

These values are L-Lipschitz on the prototypes and extend to a globally L-Lipschitz function by the McShane construction. Every prototype has another prototype whose value differs by at least $L\Delta/2$. Under uniform attention, a class-1 query predicts $(0 + L\Delta/2 - L\Delta/2)/3 = 0 = f(v_1)$, so its residual, its class loss, and by Lemma 3 its target-logit gradient all vanish. If the prior places its mass on this task, then $\Sigma_f = F_f F_f^\top$ has rank one and $a_1^\top \Sigma_f a_1 = 0$ at initialization, so no positive lower bound on the task-aligned gradient can follow from pointwise conditions. Assumption 1(ii) excludes this example because it requires $\lambda_f > 0$ in every contrast direction.

G.2 Functional Meaning of the Two Laws

The learned object is a retrieval rule shared across tasks. Because the task is resampled for every prompt, a prototype has no fixed label, and the model cannot memorize a prototype-to-label map; it must use the labels in the prompt. Coarse routing learns *where* to retrieve, raising the target attention from $1/K$ to one half. Refinement suppresses the contamination contributed by labels of other classes. Nonlinear task geometry enters through Σ_f: it sets both the scale of the gradient that drives routing and the attention precision that a given absolute accuracy requires. This is why a stronger contrast can shorten routing and yet leave refinement unchanged or longer, depending on the step-size policy.

G.3 Learning-Rate Policies

With a common admissible step η, the coarse time $\Theta(K^2/(\eta\sigma_f^2\gamma^2))$ decreases as q^{-1} at fixed ϵ; the admissibility condition $\eta\sigma_f^2\gamma^2 \le c_\eta$ must then hold for the largest contrast considered. The separated refinement time has order $(K/(\eta\gamma^2\epsilon))(1 - q_{\mathrm{tail}}/q)$, but this two-sided order does not determine its shape in q. The small-step limit does. Writing $s = 1/u = 1 + 1/R$, the recurrence (20) approaches the flow $\dot{s} = \beta(s-1)^2/s^3$, so that

$$\beta\,(T_{\mathrm{acc}} - T_{\mathrm{I}}) \longrightarrow \mathcal{I}(\sqrt{Y_\epsilon}) - \mathcal{I}(2), \qquad \mathcal{I}(s) = \frac{s^2}{2} + 2s + 3\log(s-1) - \frac{1}{s-1}. \tag{143}$$

At a common step, $\beta \propto q \propto Y_\epsilon$, so the refinement time is proportional to $[\mathcal{I}(s_\epsilon) - \mathcal{I}(2)]/s_\epsilon^2$ with $s_\epsilon = \sqrt{Y_\epsilon}$. This ratio vanishes at $q = q_{\mathrm{tail}}$ ($s_\epsilon = 2$), peaks at $s_\epsilon \approx 4.111$ (that is, $q \approx 4.23\, q_{\mathrm{tail}}$), and decreases to the plateau $1/2$, which it approaches from above. The finite-step recurrence agrees with (143): at $\beta = 10^{-3}$, $K = 32$, and $q = 8q_{\mathrm{tail}}$ it gives 26.713 against the limit 26.714, and a common-step sweep peaks at $q/q_{\mathrm{tail}} = 4.25$ on a grid of spacing 0.25 near the maximum. This explains why the common-step refinement times in Table 3, whose smallest contrast sits at the peak, decrease with q. With the certified curvature-scaled step $\eta = c/(\sigma_f^2\gamma^2)$, $\beta = cK/(K-1)^2$ does not depend on q, the coarse time is $\Theta(K^2)$, and the refinement time is $\Theta(K(q - q_{\mathrm{tail}}))$, which grows linearly in q far beyond the boundary. The step-size policy therefore determines whether stronger task contrast shortens or lengthens refinement.

G.4 Stage Boundaries Are Intrinsic

The crossover $R = 1$, equivalently target attention $1/2$, is where the recurrence changes from an approximately constant decrement of R_t to an approximately constant increment of $Y_t = u_t^{-2}$. It is fixed by the dynamics, not selected from a loss curve, and both T_{I} and T_{acc} are defined by pre-specified thresholds on observable quantities. The three regimes of Theorem 5 are therefore reproducible from attention and loss trajectories alone.

H Experimental Details and Additional Results

H.1 Exact Population Dynamics

Task prior. We use the centered codebook $(-1/2, -1/6, 1/6, 1/2)$ and draw each task by permuting these values uniformly across prototypes. When K is a multiple of four, the codebook is repeated $K/4$ times and scaled by $L\Delta$. The prior is exchangeable, and a direct calculation gives

$$\Sigma_f = c_K (L\Delta)^2 P_\perp, \qquad c_K = \frac{5K}{36(K-1)}. \tag{144}$$

The prototype values satisfy the finite-set L-Lipschitz condition and therefore admit a McShane extension. Unless stated otherwise, prototypes are equiangular, class proportions are balanced, and every step satisfies $\eta\sigma_f^2\gamma^2 \le 1/6$.

Representative trajectories. Figure 1 uses $K = 4$, $\Delta = 0.9$, $\epsilon = 0.02$, and the common step $\eta = 0.2$; only the task contrast changes. The values $q = 2$, 4, and 24 lie below $q_{\mathrm{init}} = 8/3$, between q_{init} and $q_{\mathrm{tail}} = 6$, and above q_{tail}. Table 2 reports the stopping times measured from the full $K \times K$ update. For visualization, each trajectory is continued to T_{I} even when accuracy is attained earlier. Across these runs, the maximum matrix-versus-scalar score discrepancy is below 5.7×10^{-15}, the maximum loss discrepancy is below 2.1×10^{-16}, and every loss trajectory is nonincreasing.

Table 2: Representative geometry–accuracy regimes ($K = 4$, $\Delta = 0.9$, $\epsilon = 0.02$, $\eta = 0.2$).

Regime	q	σ_f^2	T_{acc}	T_{I}
Initially accurate	2	0.04	0	13,273
Coarse accuracy	4	0.08	3,762	6,636
Tail accuracy	24	0.48	3,101	1,106

Accuracy versus attention concentration. The accuracy checkpoint should not be confused with convergence of the attention matrix. For $q = 24$, the loss threshold requires only $\alpha_\epsilon = 1 - \sqrt{2(K-1)/(Kq)} = 0.75$. Figure 3 shows the class attention at the separate checkpoint $T_{\mathrm{conc}}(0.01) = \inf\{t : 1 - \alpha_t \le 0.01\}$, reached at iteration 745,384, where the target attention is 0.990 and each off-target entry is approximately 0.0033.

Boundary sweeps and learning-rate policies. We sweep q around both boundaries for $K \in \{4, 8, 16, 32\}$. Across 244 deterministic runs, the measured regime agrees with Theorem 5 whenever the target is separated from a one-step crossing. To separate the regime boundary from the complexity crossover, we use $K \in \{16, 32, 64\}$, $\Delta = 0.9$, and $\epsilon = 0.01$, and sweep q up to $8K$. The common-step protocol uses $\eta_{\mathrm{fix}} = 0.1/(512\epsilon\gamma^2) \simeq 0.1191$, and the curvature-scaled protocol uses $\eta = 0.1/(\sigma_f^2\gamma^2)$. Table 3 gives representative results for $K = 32$, and Figure 4 displays the sweeps.

Finite-time scaling. We conduct independent K, ϵ, η, Δ, and L sweeps and report descriptive slopes on logarithmic coordinates over the stated grids (Table 4). The modest deviation of the routing exponent in K reflects finite-K constants in the exact recurrence; over $K \in \{4, 8, 12, 16, 24, 32\}$, the ratio T_{I}/K^2 has a coefficient of variation of 8.9% and a maximum deviation of 17.8% from its mean. Varying ϵ at fixed q gives a measured exponent of -0.9999 for T_{acc} in both the coarse case $q = 4$ and the tail case $q = 24$. Varying Δ and ϵ jointly and comparing trajectories in the dimensionless time βt, the relative range of βT_{acc} is below 0.052%.

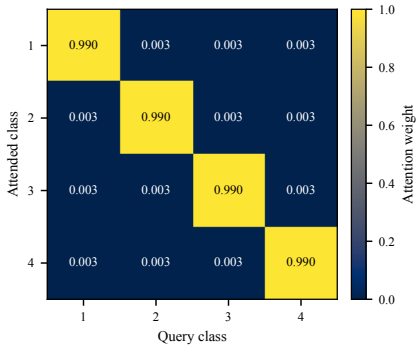


Figure 3: Query-to-class attention at $T_{\mathrm{conc}}(0.01)$ for $q = 24$. Continued training sends the diagonal entries to one and the off-target entries to zero.

Table 3: Learning-rate policy and the two time scales ($K = 32$).

Policy	q	T_{I}	$T_{\mathrm{acc}} - T_{\mathrm{I}}$	T_{acc}
Common step	32	210,002	69,282	279,284
	128	52,501	60,110	112,611
	256	26,251	54,663	80,914
Curvature scaled	32	13,126	4,330	17,456
	128	13,126	15,027	28,153
	256	13,126	27,331	40,457

Non-equiangular Gram matrices inside the basin. Theorem 1 applies beyond the equiangular family after basin entry. We generate five seeded, non-equiangular, full-rank Gram matrices with $K = 6$, whose spectral gains range from $\omega_G = 0.1104$ to 0.7112. Starting from target attention 0.9, every run remains in the basin and reduces the loss from 0.006 to 0.00075, with no loss increase or violation of (17) to numerical precision.

Reproducibility. The script `experiments/run_intrinsic_accuracy_regimes.py` regenerates Figures 1, 3 and 4 and the tables above. It writes ten comma-separated result files, a run manifest, and a machine-readable record of every predeclared integrity check. The reported run uses Python 3.12.13, NumPy 2.5.2, Matplotlib 3.11.1, and seed 20260811, and it completes on a CPU.

H.2 Finite-Sample SGD on the Token-Level Model

Setup. The prototypes are equiangular ($\rho = 0.3$, $\gamma = 0.7$) and embedded in $\mathbb{R}^{K+6}$. Each prompt has $N = 64$ context tokens with i.i.d. uniform classes and an independent uniform query class. An orthogonal nuisance $\xi \sim \mathcal{N}(0, 0.25\, I_6)$ occupies six extra coordinates. Tasks are fiber-constant and nonlinear, $f(x) = A \tanh(3w^\top P_{\mathcal{U}} x + b)$, with w uniform on the unit sphere of $\mathcal{U} = \mathrm{span}(V)$ and $b \sim \mathrm{Unif}[-1, 1]$. This prior is exchangeable on the prototypes; Monte Carlo with 2×10^6 draws gives a maximum relative deviation from $\sigma_f^2 P_\perp$ of 0.11% for $K = 4$ and 0.19% for $K = 8$. We train Q from zero by minibatch SGD with 1024 fresh prompts per step, and we evaluate the loss and the mean target attention $\bar{\alpha}_t$ on 10^4 fixed prompts at 120 checkpoints, with $\epsilon = 0.01$ and three seeds. The measured $\widehat{T}_{\mathrm{acc}}$ is the first checkpoint with loss at most ϵ, and $\widehat{T}_{\mathrm{I}}$ is the first checkpoint with $\bar{\alpha}_t \geq 1/2$. The population reference averages the exact class-level update over 2000 sampled count vectors. The curvature-scaled step is $\eta = 0.1/(\sigma_f^2 \gamma^2)$; the fixed step uses the value at the largest q.

Results. All nine measured regimes match the prediction (Table 5). The SGD stopping times lie within 1.9% of the random-count population dynamics (for the seven configurations whose step matches the reference) and within 13% of the balanced recurrence; the systematic excess over the balanced recurrence is the finite-N count effect quantified by Proposition 11. The measured initial

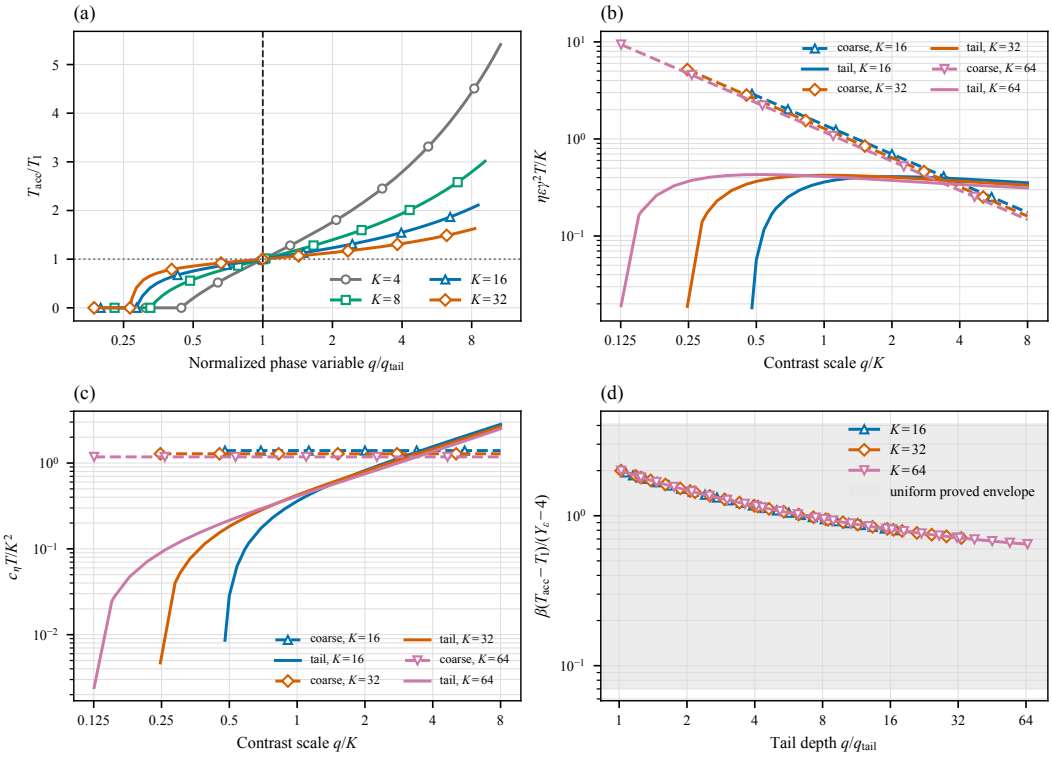


Figure 4: Deterministic checks of Theorem 5. (a) The ordering of T_{acc} and T_{I} reverses at $q/q_{\mathrm{tail}} = 1$ for four values of K. (b) With a common step, routing shortens with q while refinement approaches a contrast-independent scale. (c) With a curvature-scaled step, routing time is constant and refinement grows with $q - q_{\mathrm{tail}}$. (d) Every separated tail run lies inside the envelope (27). Panels (b) and (c) use q/K to expose the bottleneck crossover at constant order.

Table 4: Predicted and measured scaling exponents.

Quantity varied	Predicted	Measured	R^2
First-step gap increment versus K	-2	-2.000	1.0000
Routing time versus K	2	2.060	0.9965
Refinement time versus ϵ	-1	-0.974	1.0000
Refinement time versus η	-1	-1.000	1.0000
Refinement time versus Δ	-4	-4.000	1.0000
Routing time versus Δ	-6	-6.000	1.0000
Routing time versus L	-2	-2.000	1.0000

losses agree with the exact random-count value $\sigma_f^2(K-1)(N+1)/(2KN)$; for example, 0.00812 is measured for $K = 4$ and $q = 2.13$, against 0.008125 predicted and 0.008 under exact balance. At $q = 4q_{\mathrm{tail}}(K)$ with the curvature-scaled step, $\widehat{T}_{\mathrm{I}} = 184$, 980, and 3,979 for $K = 4$, 8, and 16, a fitted exponent of 2.22, compared with 2.18 for the recurrence on the same grid and 2 asymptotically. At fixed $\sigma_f^2 = 0.8$ and $\eta = 0.383$ with $K = 4$, the refinement times $\widehat{T}_{\mathrm{acc}} - \widehat{T}_{\mathrm{I}} = 356$, 668, and 1,187 for $\epsilon = 0.02$, 0.01, and 0.005 give an exponent of -0.87, the same as for the recurrence. The departure from -1 reflects the factor $1 - q_{\mathrm{tail}}/q$ at $q \in \{40, 80, 160\}$. Figures 5 and 6 show the trajectories and fits. The prompt length $N = 64$ is far below what condition (105) requires at these (K, ϵ); that condition is pessimistic, and the random-count population reference integrates the count-averaged update of Proposition 1 directly rather than through Theorem 4.

Table 5: Token-level SGD versus theory. Predictions are listed as (balanced recurrence / random-count population); a dash marks a configuration without a population reference.

K	q	Step	Regime (predicted = measured)	$\widehat{T}_{\mathrm{I}}$ (pred.)	$\widehat{T}_{\mathrm{acc}}$ (pred.)
4	2.13	curvature	initial	181 (175 / 181)	0 (0 / 0)
4	4.33	curvature	coarse	180 (175 / 181)	122 (115 / 124)
4	24	curvature	tail	184 (175 / 181)	524 (488 / 519)
4	2.13	fixed	initial	2,050 (1,960 / –)	0 (0 / –)
4	4.33	fixed	coarse	1,009 (965 / –)	689 (634 / –)
4	24	fixed	tail	184 (175 / 181)	520 (488 / 519)
8	1.83	curvature	initial	971 (913 / 968)	0 (0 / 0)
8	4.64	curvature	coarse	974 (913 / 968)	763 (705 / 765)
8	28	curvature	tail	980 (913 / 968)	1,990 (1,768 / 1,957)

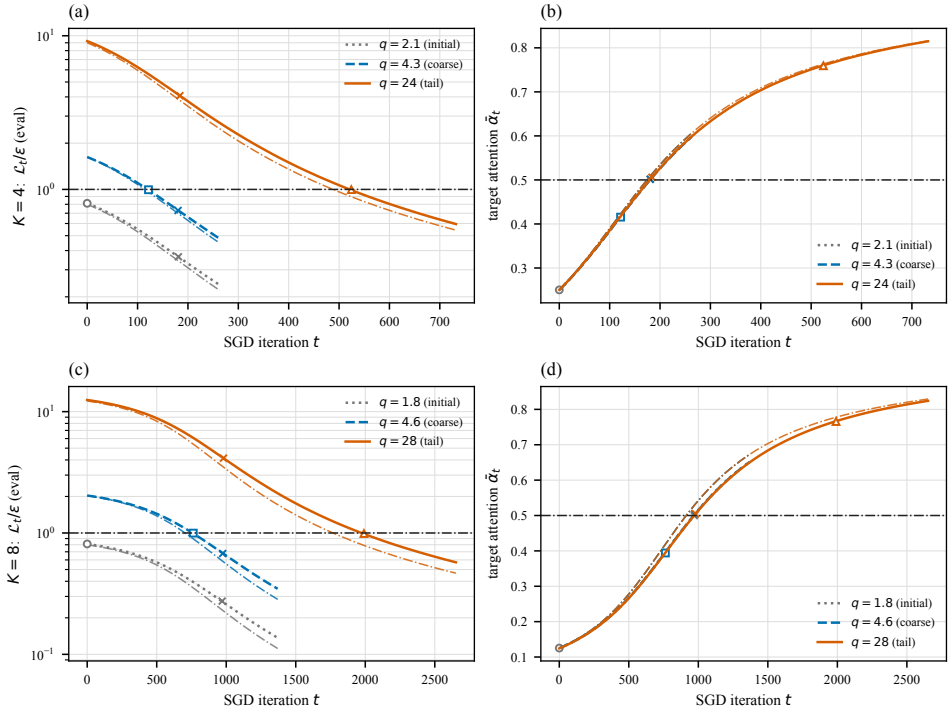


Figure 5: Token-level SGD in the three regimes for $K = 4$ (top) and $K = 8$ (bottom), curvature-scaled step. Left: evaluation loss relative to ϵ; right: mean target attention. Solid lines are SGD means over three seeds, thin dash-dotted lines are balanced-recurrence references, squares and triangles mark $\widehat{T}_{\mathrm{acc}}$, and crosses mark $\widehat{T}_{\mathrm{I}}$.

H.3 General Gram Matrices from Zero Initialization

Theorem 1 controls the tail after basin entry but does not assert entry. We integrate the exact population update $S \leftarrow S - \eta G H(S) G$ from $S = 0$ at the theorem step $\eta = 1/(6\Lambda_f \|G\|_{\mathrm{op}}^2)$, with balanced profiles and uniform queries, for $K \in \{4, 6, 8\}$. For each K we draw 30 random Gram matrices $G = V^\top V$, where $V \in \mathbb{R}^{(K+2)\times K}$ has i.i.d. Gaussian entries and unit-normalized columns (resampled when $\lambda_{\min}(G) < 0.05$). We add three equiangular matrices with $\rho \in \{-0.5/(K-1), 0.3, 0.6\}$ and two near-equiangular perturbations of operator norm 0.05 and 0.2, renormalized to unit diagonal. Each Gram matrix is paired with $\Sigma_f = P_\perp$ and with an anisotropic covariance whose eigenvalues on $\mathbf{1}^\perp$ are drawn uniformly from $[0.3, 1]$, giving 210 runs with a cap of 1.5×10^6 iterations.

All 210 runs entered the basin (15). No run showed a loss increase, a violation of (17) (minimum normalized slack 3.3×10^{-4}), a decrease of a diagonal score S_{kk}, or a drop of $\min_k \pi_{k|k}$ below $1/K$.

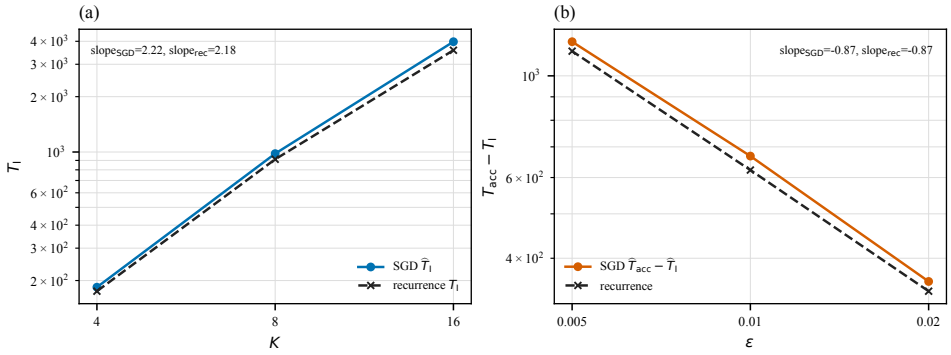


Figure 6: Token-level SGD scaling. (a) Routing time versus K at $q = 4q_{\text{tail}}(K)$. (b) Refinement time versus ϵ at fixed σ_f^2 and step. Dashed lines show the recurrence on the same grids.

The time $T_{1/2}$ at which $\min_k \pi_{k|k} \geq 1/2$ ranged from 105 to 215,475 iterations, and basin entry occurred within 624,678 iterations. For the equiangular runs with isotropic covariance, $T_{1/2} = T_{\text{I}}$ lies inside (24). Across all runs, $T_{1/2}$ is best predicted by $K^2/(\eta\lambda_f\bar{\lambda}_\perp(G)^2)$, where $\bar{\lambda}_\perp(G)$ is the mean eigenvalue of G on $\mathbf{1}^\perp$: the pooled log–log slope is 0.91 with $R^2 = 0.84$, compared with $R^2 = 0.57$ for $\lambda_\perp(G)^2$ and 0.51 for ω_G (Figure 7). For the random Gram matrices, $\bar{\lambda}_\perp(G)$ exceeds $\lambda_\perp(G)$ by a median factor of 5.1, so the worst-case constants of Theorem 1 are pessimistic for basin entry. These unit-norm geometries satisfy self-similarity automatically, in contrast with Proposition 2. The ensemble does not probe nearly singular Gram matrices, strongly imbalanced queries, or non-unit norms.

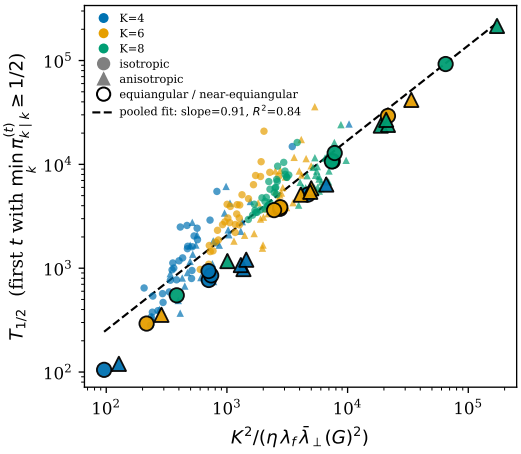


Figure 7: Time to majority target attention from zero initialization for 210 exact population runs with non-equiangular, equiangular, and near-equiangular unit-norm Gram matrices and isotropic or anisotropic task covariances. The dashed line is the pooled log–log fit against $K^2/(\eta\lambda_f\bar{\lambda}_\perp(G)^2)$.

H.4 Correlated Asymmetric Geometry from Zero Initialization

Theorem 3 certifies an absolute correlation budget under transitive symmetry, and Appendix D.7 shows that no argument covering the whole region Ω can exceed $4/K^2$. This experiment asks what happens from zero initialization for *arbitrary* asymmetric perturbations. We integrate the exact population update for $G = \gamma I + \rho\mathbf{1}\mathbf{1}^\top + E$ with uniform queries, balanced profile, and $\Sigma_f = P_\perp$, where E is symmetric with zero diagonal and prescribed correlation mass $c\gamma = \max_k \sum_{m\neq k} |E_{km}|$, either dense (Rademacher signs of equal magnitude) or sparse (three sampled neighbors per row, then symmetrized); $K \in \{4, 8, 16, 32, 64\}$, $\rho \in \{0, 0.3, 0.6, 0.8\}$, $c\gamma \in \{0.05, \ldots, 1\}$, six seeds

(at $K = 64$ only $\rho \in \{0, 0.6\}$, $c\gamma \in \{0.1, 0.35, 1\}$, and two seeds), step $\eta = 1/(6\lambda_{\perp,\max}(G)^2)$, and 1,153 runs in total. Each run records whether every gap $d_{k,j}$ stayed nonnegative, whether the loss was monotone, the time $T_{1/2}$ to majority attention against the exact symmetric value $T_{1/2}^{\mathrm{sym}}$ of Proposition 3, the largest within-column attention ratio, and whether majority persisted.

Results (Figure 8). Within the normalized budget $c \le \frac{1}{2}$, all 684 runs kept the region invariant with monotone loss and majority persisting through the horizon, and $T_{1/2}/T_{1/2}^{\mathrm{sym}} \le 1.83$ overall, ≤ 1.10 for $K \ge 16$, and ≤ 1.02 at $K = 64$: the symmetric routing law survives with a K-independent budget. All 16 violations occurred at $c \ge 1$, fourteen of them at $c > 1$, where the common-correlation coupling $\rho(\gamma I + E)(H\mathbf{1})\mathbf{1}^\top$ dominates. Within-column attention ratios grow with the gaps (up to 8.4 within budget), which is harmless for the region, and cross-column odds ratios stay below 3.32. *Negative control.* For the star perturbation of Proposition 7, the adversarial state (one column uniform, the others half on that class) has a negative gap velocity exactly when the per-entry correlation $c/(K-1)$ exceeds the threshold $4/K^2$ of Proposition 7, that is, when $c > 4(K-1)/K^2$ (0.75, 0.44, 0.23, 0.12, 0.06 for $K = 4, \ldots, 64$): on the grid the first failing masses are 1, 0.5, 0.35, 0.2, and 0.1, and at $K = 4$, $c = 0.75$ the velocity is 10^{-5}; the same star geometries started from zero keep the region for every tested $c \le 1$ at every K. *One correlated pair.* With $E_{jk} = E_{kj} = e\gamma$ and no other entries, the first-step relative velocity of the affected gap is exactly $(1-e)^2$ (Figure 8c), so the first-order limit $e < \frac{1}{2}$ is not a threshold; the region held for $e \le 0.7$. On transitive geometries (circulant E, $K \le 32$, $\rho \le 0.95$, $c \le \frac{1}{2}$, 300 runs at the step $K/(6\sigma_f^2\lambda_{\perp,\max}(G)^2)$; script `e7b_transitive_certificate.py`) the row sums $H^{(t)}\mathbf{1}$ vanish to machine precision and every gap increases at every step; within the certified budget $c \le \frac{1}{5}$ every conclusion of Theorem 3 holds with margin, the routing time staying below 0.12 of its bound.

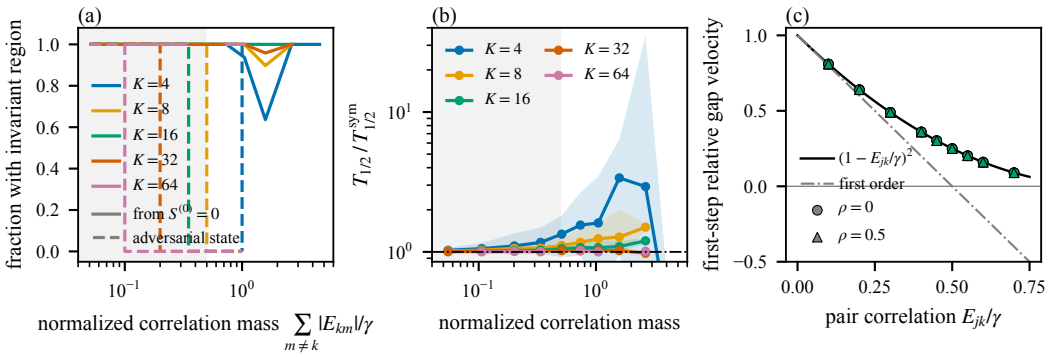


Figure 8: Correlated asymmetric geometry from zero initialization. (a) Fraction of runs whose region stays invariant against the normalized correlation mass; solid from $S^{(0)} = 0$, dashed from the adversarial state of Appendix D.7; the shaded band is $c \le \frac{1}{2}$. (b) Routing time relative to the symmetric law (median and range over ρ, E, and seeds). (c) First-step relative velocity of a single correlated pair against E_{jk}/γ, with the exact value $(1 - E_{jk}/\gamma)^2$.

H.5 Verification of Theorem 2, its Extensions, and the Trap

Theorem 2 with imbalanced queries. We integrate the exact population update from $S^{(0)} = 0$ for $G = I + E$ with dense Rademacher or star-shaped E, balanced profile, $\Sigma_f = P_\perp$, the certified step $\eta = 1/(6\,\|G\|_{\mathrm{op}}^2)$, $K \in \{4, 8, 16\}$, and query distributions that are uniform, Dirichlet(0.3) (clipped to $p_{\min} \ge 0.02/K$ and renormalized), or concentrated on one rare class with $p_{\min} \in \{0.5, 0.1, 0.02\}/K$; the correlation level $\bar{\rho}$ ranges from the certified $p_{\min}/(9K)$ through $2p_{\min}/K$ and $4/K^2$ up to 0.4 (378 positive-definite runs). At the certified budget and at $2p_{\min}/K$, all 126 runs satisfy every claim: the region is invariant, $\pi_{k|k} \ge 1/K$, the loss is monotone, the recurrence $\mathcal{L}_{t+1} \le \mathcal{L}_t - c_R\mathcal{L}_t^2$ holds at every step (its slack is smallest at $t = 0$, where the decrease is about $2.0\,c_R\mathcal{L}_0^2$), majority attention is reached within 0.11 of the bound (0.07 at the certified budget) $\lceil 32(K-2)/(\eta\sigma_f^2 p_{\min})\rceil + 1$ and kept, and, in the 96 runs where majority precedes the target, the tail time is within 0.15 of $2(K-1)/(\eta\omega_G\epsilon)$ (Figure 9a,c). Beyond the budgets the region fails in

147 of 252 runs, only for imbalanced queries (never for uniform queries up to $\bar{\rho} = 0.4$), never below $3.8 \times (2p_{\min}/K)$, and in 135 cases already at $t = 1$ (Figure 9b); the floor $1/K$ then drops to $0.52/K$ at worst, yet the loss stays monotone and every run converges, the four slowest (star E, $K = 16$, $\bar{\rho} = 0.2$) after a transient in which the rare class's attention falls to 0.034 before recovering.

Propositions 8 and 9. With anisotropy $\kappa \in \{1, 1.5, 2, 3, 4\}$ or profile imbalance $\zeta \in \{0, 0.1, 0.25, 0.4, 0.5\}$, all 48 runs satisfying the hypotheses pass every claim, as do the 32 runs at the certified budget with $\kappa \in \{3, 4\}$ or $\zeta \in \{0.4, 0.5\}$; violations of the region (56 of 320) occur only at $\bar{\rho} \geq 0.05$, at least 48 times the certified budget, and only for the rare-class distribution, while monotonicity, the rate, and convergence hold in all 320 runs.

Proposition 2. For $K = 2$ we parametrize G by the norm ratio $s = \|v_2\|^2 / \|v_1\|^2 \in \{1, 2, 4, 10, 25\}$ and the correlation $r = G_{12}/\sqrt{s} \in [-0.9, 0.98]$, with $p_2 \in \{0.5, 0.2, 0.1, 0.05, 0.02, 0.01\}$ (270 cells, 2×10^6 certified steps; Figure 10). The example of Proposition 2 traps ($\pi_{2|2} \to 2.5 \times 10^{-9}$ after 2×10^7 steps, with $d_2/d_1 \to -3$), while its uniform-query variant reaches $\mathcal{L} \leq 3 \times 10^{-3}$ at $t \approx 2.6 \times 10^5$. No cell traps at $p_2 = \frac{1}{2}$, none with a nonnegative initial push $p_2 G_{22} - p_1 G_{12} \geq 0$, and none where self-similarity holds ($r\sqrt{s} < 1$). Empirically d_2/d_1 approaches $-G_{12}$ along a collapse, and each of the six followed collapsed cells with $G_{12} < 2$ turns around in a 2×10^7-step run, whereas the example ($G_{12} = 3$) keeps collapsing.

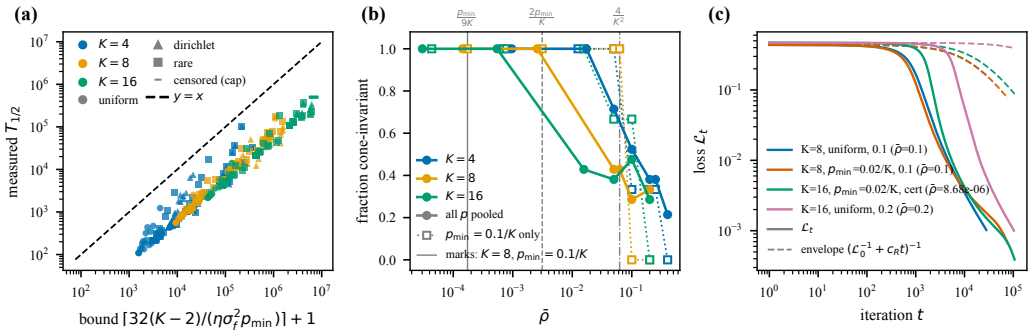


Figure 9: Verification of Theorem 2 with imbalanced queries (378 runs). (a) Measured time to majority attention against the bound $\lceil 32(K-2)/(\eta\sigma_f^2 p_{\min})\rceil + 1$ (dashed: equality); markers distinguish uniform, Dirichlet, and rare-class query distributions. (b) Fraction of runs with an invariant region against $\bar{\rho}$, pooled over query distributions (solid) and for $p_{\min} = 0.1/K$ only (dotted); vertical lines mark $p_{\min}/(9K)$, $2p_{\min}/K$, and $4/K^2$ for $K = 8$ and $p_{\min} = 0.1/K$. (c) Loss against the envelope $(\mathcal{L}_0^{-1} + c_R t)^{-1}$ for four runs.

H.6 Token-Level SGD with Random Geometry and Imbalanced Queries

A natural question is whether the routing law and the regime classification survive on the token-level model (1) when the geometry lies outside every theorem's budget and the queries are imbalanced. We repeat the protocol of Appendix H.2 (random nonlinear single-index prior, $N = 64$ multinomial counts, orthogonal nuisance, minibatch 1,024, three seeds, and the step $\eta = 0.1/(\sigma_f^2 \bar{\lambda}_\perp(G)^2)$ with $\bar{\lambda}_\perp(G)$ the mean eigenvalue of G on $\mathbf{1}^\perp$, a heuristic rather than a certified choice) with the random unit-norm Gram matrices of Appendix H.3 (two per K, off-diagonal correlations of typical size $K^{-1/2}$) and query distributions that are either Dirichlet(0.5) draws conditioned on $p_{\min} \geq 0.1/K$ or concentrated on one rare class with $p_{\min} = 0.1/K$, keeping context classes uniform; $K \in \{4, 8\}$ and the three values of q of Appendix H.2, 24 configurations in all. The population reference integrates the count-averaged update of Proposition 1 for the same G and p. Because the prior is no longer exchangeable on these prototypes, Σ_f is anisotropic; it is estimated by Monte Carlo, $\sigma_f^2 := \operatorname{tr}\Sigma_f/(K-1)$ defines q, and the population reference uses the full Σ_f.

Results (Figure 11). All 24 configurations land in the predicted regime, and so do their population references. The SGD stopping times agree with the population reference to within 3.9% for $\widehat{T}_{\mathrm{I}}$ (mean 1.0%) and 4.8% for $\widehat{T}_{\mathrm{acc}}$ (mean 1.8%), the initial loss matches the count-averaged prediction to

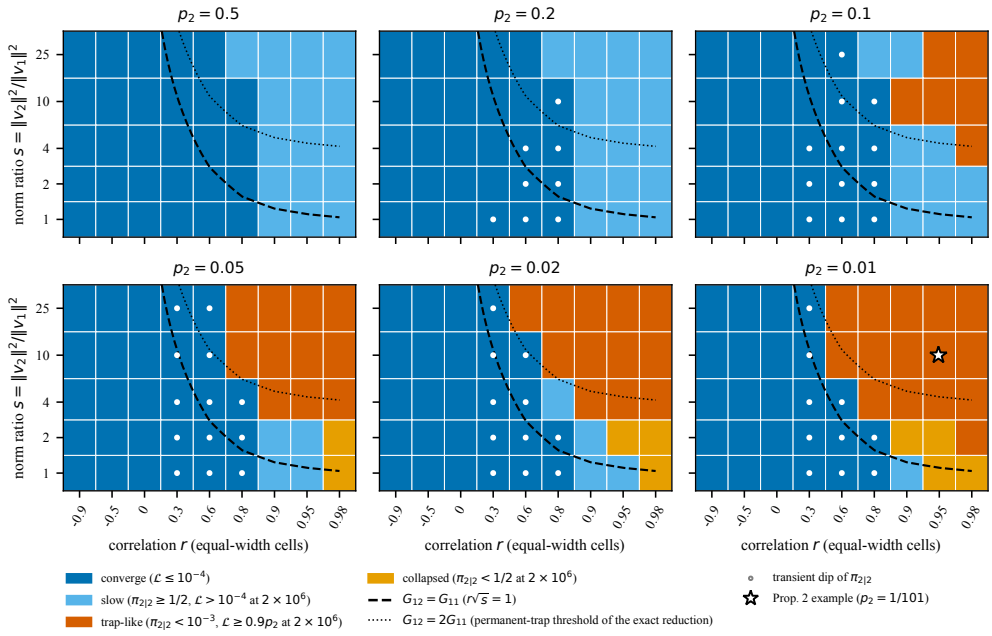


Figure 10: Trap map for $K = 2$ (Proposition 2): outcome after 2×10^6 certified steps as a function of the norm ratio s and the correlation r, one panel per query imbalance p_2. Dashed: the self-similarity boundary $G_{12} = G_{11}$; dotted: $G_{12} = 2G_{11}$; dots mark a transient dip of $\pi_{2|2}$; the star is the example of Proposition 2.

within 2.1%, and no run was censored. The K-scaling of $\widehat{T}_{\mathrm{I}}$ between $K = 4$ and $K = 8$ is geometry-dependent, with ratios from 3.0 to 10.5 across the four geometry–query pairs (SGD and population within 3.1% of each other), while the median over an ensemble of 16 random geometries per K is 5.6, close to the symmetric value 5.2 at these K; at these two sizes the ensemble median follows the symmetric law, while individual Gram matrices spread around it according to their contrast spectra.

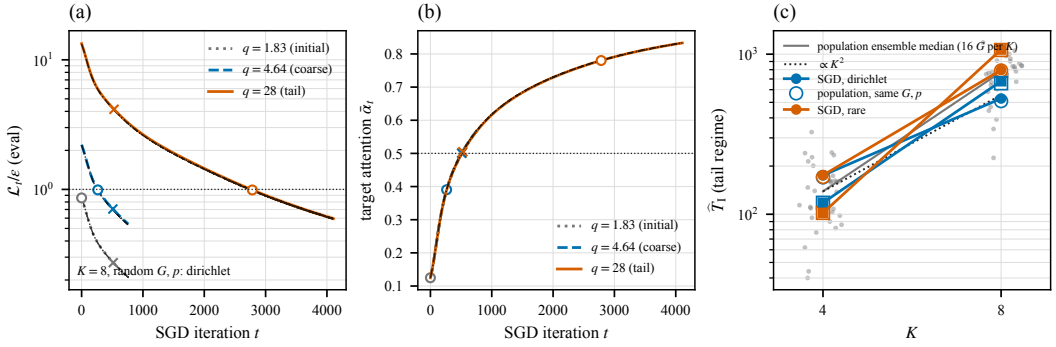


Figure 11: Token-level SGD with a random unit-norm Gram matrix and Dirichlet query frequencies ($K = 8$). (a) Evaluation loss relative to ϵ in the three regimes (mean of three seeds) against the population reference for the same G and p (dash-dotted); circles mark $\widehat{T}_{\mathrm{acc}}$ and crosses $\widehat{T}_{\mathrm{I}}$. (b) Mean target attention. (c) $\widehat{T}_{\mathrm{I}}$ in the tail regime against K: SGD (filled) and population (open) for two geometries and two query distributions, the population ensemble median over 16 geometries per K (gray), and the K^2 law (dotted).

H.7 Task-Relevant Continuous Perturbations

Outside Theorem 6, we train the token-level model (1) by minibatch SGD with $K = 4$ equiangular prototypes ($\rho = 0.3$) embedded in $\mathbb{R}^{10}$, $N = 64$ i.i.d. context tokens, and inputs $x = v_z + r\zeta$ with $\zeta \sim \mathcal{N}(0, I_d/d)$. Tasks are $f(x) = A\tanh(3w^\top x + b)$ with w uniform on the unit sphere of $\mathbb{R}^{10}$ and $b \sim \mathrm{Unif}[-1, 1]$, so labels vary within classes. On the prototypes this prior is exchangeable (measured maximum relative deviation from $\sigma_f^2 P_\perp$ of 0.24%). We set A so that $q = 24$ at $\epsilon = 0.01$ and use $\eta = 0.1/(\sigma_f^2\gamma^2)$. Each step uses 1024 fresh prompts, losses are evaluated on 10^4 fixed prompts, and results are averaged over three seeds.

With $r = 0$, the measured crossover $\widehat{T}_{\mathrm{I}} = 184$ and accuracy time $\widehat{T}_{\mathrm{acc}} = 529$ are within 5.2% and 8.4% of the balanced population predictions 175 and 488. For $r/\Delta \in \{0.1, 0.2, 0.4\}$, the target-attention trajectories remain close to the prediction ($\widehat{T}_{\mathrm{I}} = 184$, 184, 188), while the loss levels off (Figure 12). A natural reference level is the within-class label variance $\frac{1}{2}\mathbb{E}\,\mathrm{Var}_\zeta\, f(v_z + r\zeta)$, the loss of the predictor that outputs each class's mean label, equal to 0.25ϵ, 1.00ϵ, and 3.67ϵ. The losses at the end of the horizon are 0.46ϵ, 1.17ϵ, and 3.61ϵ. For $r/\Delta = 0.4$ the loss sits at this level and is still decreasing slowly, so the level is a reference rather than a proven lower bound: an attention model that observes the perturbed query can in principle exploit within-class structure. For $r/\Delta \in \{0.1, 0.2\}$ the loss is also still decreasing, consistent with the reference level plus the prototype-model loss at the same iteration (0.22ϵ). Accuracy ϵ is reached for $r/\Delta \le 0.1$ ($\widehat{T}_{\mathrm{acc}} = 529$ and 619), but not for $r/\Delta \ge 0.2$, where the reference level is close to or above ϵ (1.00ϵ and 3.67ϵ). Task-relevant variation thus leaves the routing law intact and moves the loss to the within-class label-variance level, suggesting that the regime classification applies approximately with ϵ measured relative to that level.

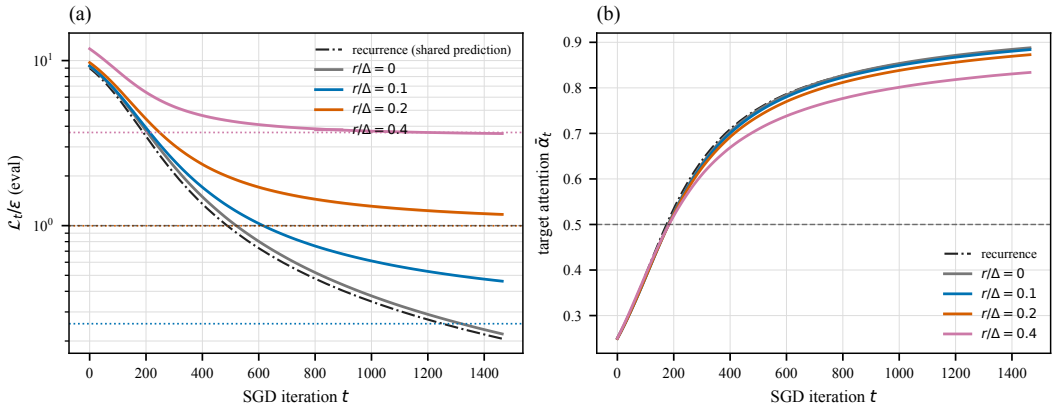


Figure 12: Token-level SGD with task-relevant input perturbations ($K = 4$, $q = 24$). (a) Evaluation loss relative to ϵ; dotted lines mark the predicted within-class floors and the dash-dotted curve is the population recurrence. (b) Mean target attention.