A BOUNDARY-LAYER MECHANISM FOR ONE-THIRD SCALING IN ONLINE SOFTMAX CLASSIFICATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Hard-label classification is usually trained with smooth surrogate losses such as softmax cross-entropy. We isolate a mechanism by which this mismatch between smooth surrogate and discrete labels produces power-law learning curves for a one-layer softmax student trained by online SGD on Gaussian inputs with argmax teacher labels. In the thermodynamic limit, the dynamics reduce to a growing student-teacher alignment and a residual student variance sustained by SGD noise. At late times, only shrinking layers of examples around teacher decision boundaries remain active. This yields an $\alpha^{-1/3}$ law in training time α not only for the test loss but also for the generalization error, one minus test accuracy. The error law is an online-noise effect. The loss law persists without gradient noise. Learning-rate annealing separates the two exponents, improving the error exponent from 1/3 toward 1/2 while slowing the loss. Simulations support the predictions. Correlated Gaussian inputs and whitened pretrained features change transients, not the late-time decay. Experiments with deep neural networks can show compatible error scaling. The mechanism is asymptotic and complementary to spectral explanations and data-statistics explanations of neural scaling laws.

1 INTRODUCTION

Scaling laws make learning dynamics predictable: they summarize how error decreases with data, compute, model size, or optimization time. Empirical work has made such laws central to modern machine learning (Hestness et al., 2017; Kaplan et al., 2020; Rae et al., 2021; Hoffmann et al., 2022). A major theoretical route explains these laws from structure in the data distribution, target function, or feature map. In kernel, random-feature, and linear models, power-law spectra of feature or data covariance matrices can be converted into power-law learning curves (Bordelon et al., 2020; Canatar et al., 2021; Maloney et al., 2022; Bahri et al., 2024; Lin et al., 2024); related geometric considerations connect exponents to the effective dimension of the data manifold (Nakada & Imaizumi, 2020; Sharma & Kaplan, 2022). Feature-learning analyses show how these spectral mechanisms can change once the representation evolves during training (Bordelon et al., 2024; 2025; Worschech & Rosenow, 2025).

Power laws can also arise from mechanisms that are not primarily spectral. Simplified sequence-modeling models show how scaling behavior can emerge without explicit power-law correlations in the data (Barkeshli et al., 2026), while recent work shows that softmax and cross-entropy can intrinsically produce one-third time scaling when learning peaked probability distributions (Liu et al., 2026). High-dimensional teacher–student theories provide another complementary viewpoint, with precise learning curves for multiclass classification under Bayes-optimal and empirical-risk minimization procedures, including cross-entropy loss (Cornacchia et al., 2023). Here we isolate a different late-time mechanism: a leading-order boundary-layer realization of the softmax/cross-entropy bottleneck in online multiclass classification, related to earlier observations that late-time updates concentrate near decision boundaries (Paccolat et al., 2021; Sclocchi & Wyart, 2024).

This motivates a focused question: what controls the late-time learning dynamics of online multiclass classification when the labels are discrete but the student is trained through differentiable probabilities? We study this question in a one-layer K-class teacher-student model with Gaussian inputs and online stochastic gradient descent (SGD), using the order-parameter methodology of the statistical physics of learning (Opper & Haussler, 1991; Biehl & Riegler, 1994; Saad & Solla,

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.

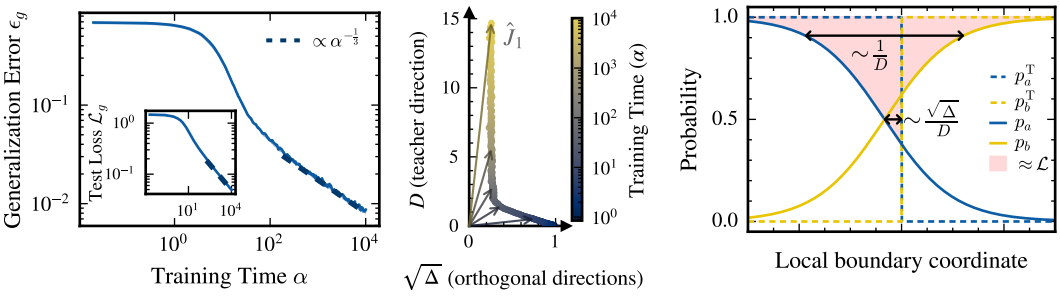

Figure 1: Left: The $1/3$ law appears not only in the test loss but also in the generalization error ϵ_g, i.e., one minus test accuracy. Middle: The model captures both the growth of the centered student-teacher alignment D and the rotational alignment to the teacher. Right: Near a teacher decision boundary, the late-time loss is controlled by the student boundary layer of width $O(D^{-1})$. The generalization error is controlled by residual boundary fluctuations only of order $\sqrt{\Delta}/D$.

1995). The model is simple enough to solve in the thermodynamic limit, but rich enough to expose an asymptotic mechanism that is easy to miss in the raw order parameters.

Our contributions. We derive an exact centered macroscopic closure for symmetric online K-class softmax learning. We then give a boundary-layer derivation of the softmax/cross-entropy bottleneck in hard-label classification, compute the leading classification-error asymptotics, and show how a learning-rate schedule can improve them. The results are stated formally under explicit assumptions (A1)–(A7), as Lemmas 1 and 2, Theorem 3, and Corollaries 4 and 5. See Fig. 1 for an overview. We test the mechanism with finite-dimensional simulations and controlled departures from isotropic Gaussian inputs, analyze exponent masking and recovery in a Zipf model, and examine Pythia token top-1 error, finding a fitted exponent compatible with $1/3$ for the 12B model.

The late-time dynamics becomes transparent only in centered coordinates, which remove the softmax redundancy under a common shift of all logits (Section 3): the centered student-teacher overlap D, the centered student squared norm Q_{eff}, and the residual student variance $\Delta = Q_{\mathrm{eff}} - D^2$, which comes from fluctuations of the student orthogonal to the teacher. In these variables, perfect learning is not convergence to a finite-weight fixed point. Instead, the centered overlap D grows without bound, while at fixed learning rate the residual variance Δ approaches a finite noise floor. The only examples that remain active lie within $O(D^{-1})$ of pairwise teacher decision boundaries. The shrinking measure of these active boundary layers gives $\dot{D} \propto D^{-2}$, and therefore $D \propto \alpha^{1/3}$. The classification error is controlled by the angle between student and teacher, $\epsilon_g \propto \sqrt{\Delta}/D$, giving

$$\epsilon_g(\alpha) \propto \alpha^{-1/3} \,. \tag{1}$$

Annealing the learning rate changes the residual noise floor. For a slowly decaying learning rate $\eta(\alpha) \propto \alpha^{-\gamma}$ with $0 \le \gamma < 1$, this gives $\epsilon_g(\alpha) \propto \alpha^{-(2+\gamma)/6}$, improving toward $\epsilon_g \propto \alpha^{-1/2}$ as $\gamma \to 1$. This rate remains slower than Bayes-optimal rates in theories, but mirrors reports for cross-entropy minimization at fixed regularization on static datasets (Cornacchia et al., 2023).

These results should be read in the context of Liu et al. (2026), who show that softmax and cross-entropy can generate one-third scaling of the test loss when learning peaked distributions. Our solvable model of online hard-label classification yields classification-error asymptotics, identifying the roles of decision-boundary geometry and the residual variance maintained by online gradient noise.

Throughout, $x \cdot y$ is the Euclidean inner product; $f \sim g$ means $f/g \to 1$ and $f \propto g$ means $c\, g \le f \le C\, g$ for constants $c, C > 0$, as $\alpha \to \infty$.

2 Related Work

Empirical scaling laws have motivated a broad set of theoretical explanations, including spectral, geometric, and feature-learning mechanisms (Bahri et al., 2024; Bordelon et al., 2020; Canatar et al.,

2021; Maloney et al., 2022; Sharma & Kaplan, 2022; Bordelon et al., 2024; 2025; Worschech & Rosenow, 2025). Our work is complementary: it gives a sharply scoped classification example in which decision-boundary geometry and online optimization noise produce a power law even without a structured input spectrum.

The closest recent result is the one-third softmax and cross-entropy bottleneck of Liu et al. (2026). They analyze peaked target distributions under gradient flow using a solvable ansatz in which the student is exactly aligned with the teacher, so that learning changes only the overall logit scale. This yields a $1/3$ loss law, supported by evidence from large-language-model training. However, mere rescaling preserves the student logit order, so the reduced theory has identical student and teacher decision boundaries and zero classification error. It therefore does not predict how classification error decreases during learning.

For classification tasks such as image recognition, generalization error measures the decision performance of interest more directly than cross-entropy, and the two metrics need not share a scaling exponent (Hestness et al., 2017). Exponents vary, but for ImageNet (Deng et al., 2009) top-1 error scaling with an exponent near one-third has been reported both for training-data scaling (Hestness et al., 2017) and saturating compute-frontier scaling (Zhai et al., 2022). This distinction also matters for LLMs: scaling laws have been developed for average downstream top-1 error (Gadre et al., 2025), and multiple-choice accuracy depends on the correct answer's logit relative to specific competitors, not its loss alone (Schaeffer et al., 2025). Moreover, decomposing cross-entropy loss reveals that a rank-entropy component based on the distribution of the correct token's logit rank can scale more robustly than total cross-entropy (Yan et al., 2026). Although token top-1 error requires caution because several next tokens may be plausible, it directly probes logit ordering rather than probability magnitude.

The present hard-label setting, a limiting case of peaked targets, resolves the gradient noise and parameter rotation explicitly left open by Liu et al. (2026) and yields the classification-error predictions of Sections 4 and 5; for the loss alone, our schedule result reduces to their time reparametrization $\tau = \int \eta$. Where their analytic derivation emphasizes softmax and cross-entropy, our boundary-layer calculation makes the geometric source of the slow drift explicit. A companion calculation for a smooth perceptron classifier trained on hard labels with mean-squared-error loss (Section B) illustrates the same local asymptotic mechanism beyond that surrogate. Table 1 summarizes the comparison.

Our analysis uses the classic teacher–student and thermodynamic-limit methodology of the statistical physics of learning (Biehl & Riegler, 1994; Seung et al., 1992; Opper & Haussler, 1991; Saad & Solla, 1995), complementary to physics-informed machine-learning methods that build physical constraints into models (Raissi et al., 2019). In this tradition, high-dimensional stochastic learning dynamics reduce to deterministic flows for self-averaging order parameters. One-third exponents have been observed before, as already the standard online perceptron algorithm exhibits one-third scaling of the generalization error (Biehl & Riegler, 1994). The present study distinguishes itself by analyzing gradient descent and multiclass classification. Power-law late-time behavior can also arise from dynamical degeneracies such as soft modes in over-realizable soft committee machines (Richert et al., 2022). Modern work has extended these ideas to richer neural-network and high-dimensional classification settings (Advani et al., 2020; Goldt et al., 2019; Mignacco et al., 2020).

High-dimensional teacher–student analyses provide reference points for Bayes-optimal classification, convex optimization, and learning curves with generic feature maps (Aubin et al., 2020; Loureiro et al., 2021; Cornacchia et al., 2023). These works address the asymptotic performance of estimators trained on a static set of examples, whereas the present paper addresses late-time online dynamics for hard-label classification under softmax and cross-entropy updates.

Closely related are analyses of hinge-loss training: for networks with one hidden layer, Paccolat et al. (2021) showed that examples far from the class boundaries cease to contribute to the gradient as training proceeds, so that the updates become concentrated on the examples near those boundaries. For a binary perceptron with Gaussian inputs, Sclocchi & Wyart (2024) derived a test error that decreases with a one-third power of training time, in an online regime in which SGD noise dominates. Here, by contrast, we consider a multiclass classifier trained online with the standard softmax cross-entropy loss. We show that the loss and the misclassification error are controlled by

two distinct boundary-layer widths, and that this distinction explains how decaying the learning rate can make the error fall faster, and the loss more slowly, than they would at a fixed learning rate.

Our results also connect to smooth surrogate losses and implicit bias. Calibration and consistency theory relates surrogate risk to classification risk (Zhang, 2004; Bartlett et al., 2006; Wang & Scott, 2024), and recent work gives universal square-root rates for smooth surrogate-loss H-consistency bounds (Mao et al., 2024). The limiting behavior of our annealing schedule, improving toward $\epsilon_g(\alpha) \propto \alpha^{-1/2}$, echoes this square-root behavior but concerns online optimization time rather than excess-risk transforms. For separable data, gradient descent on logistic or cross-entropy losses drives weights toward max-margin directions while the norm diverges (Soudry et al., 2018). Related work studies fixed-learning-rate SGD and multiclass extensions of this implicit-bias picture (Nacson et al., 2019; Ravi et al., 2024). Our population online setting differs because fresh examples continually inject boundary noise, leaving a residual uncertainty at fixed learning rate that annealing can reduce.

3 Centered online dynamics for the K-class softmax student

3.1 Teacher-student model assumptions

(A1) Single-layer student. We study a single-layer student that maps N-dimensional inputs to K logits, one for each class. It has weights $J_1, \dots, J_K \in \mathbb{R}^N$, with logits and softmax probabilities

$$t_a = \frac{J_a \cdot \xi}{\sqrt{N}}, \qquad p_a = \frac{e^{t_a}}{\sum_{b=1}^{K} e^{t_b}} \,. \tag{2}$$

(A2) Orthonormal single-layer teacher. Let $T_1, \dots, T_K \in \mathbb{R}^N$ be orthogonal teacher vectors, normalized as $T_a \cdot T_b / N = \delta_{ab}$, where δ_{ab} is the Kronecker delta.

(A3) Hard labels. Labels are generated by the teacher. The teacher fields and one-hot labels are

$$u_a = \frac{T_a \cdot \xi}{\sqrt{N}} \,, \qquad p_a^{\mathrm{T}} = \mathbf{1}\{u_a = \max_b u_b\} \,, \tag{3}$$

where $\mathbf{1}\{\cdot\}$ is the indicator function. Ties have probability zero under the Gaussian input distribution. Hard labels refer to the fact that the labels are deterministic and one-hot. They form the zero-temperature limit of the peaked soft targets of Liu et al. (2026).

(A4) Standard Gaussian inputs. Inputs follow $\xi \sim \mathcal{N}(0, I_N)$.

(A5) Online SGD on cross-entropy. For a single example, cross-entropy loss is $\mathcal{L} = -\log p_y$, where $y = \arg\max_a u_a$ is the teacher label. Online gradient descent with learning rate η and fixed finite batch size b gives the update

$$J_a^{\mu+1} = J_a^{\mu} + \frac{\eta}{b\sqrt{N}} \sum_{i=1}^{b} \left(p_a^{\mathrm{T},\mu,i} - p_a^{\mu,i}\right) \xi^{\mu,i}, \qquad \alpha = \frac{\mu}{N} \,. \tag{4}$$

(A6) Thermodynamic limit. We analyze the dynamics in the limit $N \to \infty$ with b and $\alpha = \mu/N$ fixed, taken before a late-time limit $\alpha \to \infty$.

(A7) Permutation-symmetric ansatz. Under permutation symmetry between classes, the standard statistical mechanics order parameters are

$$R = \frac{J_1 \cdot T_1}{N} \,, \qquad S = \frac{J_1 \cdot T_2}{N} \,, \qquad Q = \frac{J_1 \cdot J_1}{N} \,, \qquad C = \frac{J_1 \cdot J_2}{N} \,. \tag{5}$$

By symmetry, the choice of indices is arbitrary, with 1 and 2 denoting distinct classes. Here R is the overlap with the matching teacher, S is the overlap with a non-matching teacher, Q is the student squared norm, and C is the overlap between two distinct student weight vectors.

Assumption scope. (A1), (A2), (A6), and (A7) are standard in this model class (Biehl & Riegler, 1994; Saad & Solla, 1995; Goldt et al., 2019; Mignacco et al., 2020): (A2) is asymptotically satisfied by independent isotropic random teachers at fixed K, whose normalized off-diagonal overlaps are $O(N^{-1/2})$, finite-N runs at $N = 200$–500 already match the limiting theory with (A6) at prefactor

level (Section 6), and the i.i.d. student initialization in the Gaussian teacher–student simulations self-averages to the symmetric ansatz of (A7). (A5) is standard but load-bearing: it matches the single-pass regime of large-scale training and is essential for the generalization-error results though not for the loss. (A3) and (A4) are restrictive: (A3) is essential, because soft targets or label noise can mask the boundary-layer asymptotics; (A4) is relaxed empirically in Section 6.

3.2 Exact closure in centered order parameters

The softmax probabilities are invariant under a common shift of the logits, $t_a \mapsto t_a + c$. The mean logit is therefore dynamically irrelevant. This motivates the centered order parameters

$$D := R - S\,, \qquad Q_{\text{eff}} := Q - C\,, \qquad \Delta := Q_{\text{eff}} - D^2\,. \tag{6}$$

Their geometric meaning is transparent in centered coordinates: define the centered student $\hat{J}_a := \sqrt{\frac{K}{K-1}}\Big(J_a - \bar{J}\Big)$ and centered teacher $\hat{T}_a := \sqrt{\frac{K}{K-1}}\Big(T_a - \bar{T}\Big)$ with $\bar{T} := \frac{1}{K}\sum_a T_a$ and $\bar{J} := \frac{1}{K}\sum_a J_a$. Then, for any class a,

$$D = \frac{\hat{J}_a \cdot \hat{T}_a}{N}\,, \qquad Q_{\text{eff}} = \frac{\hat{J}_a \cdot \hat{J}_a}{N}\,. \tag{7}$$

Thus D is the centered student-teacher overlap and Q_{eff} is the centered student norm, both independent of the class index a by permutation symmetry, and Δ is the residual student variance due to fluctuations of the student orthogonal to the teacher.

For the macroscopic dynamics, only the centered student logits and teacher fields matter. In the thermodynamic limit, their joint Gaussian law can be represented as

$$h_a := t_a - \bar{t} = D(u_a - \bar{u}) + \sqrt{\Delta}\,(z_a - \bar{z})\,, \tag{8}$$

where $\bar{t} = K^{-1}\sum_a t_a$, $\bar{u} = K^{-1}\sum_a u_a$, $\bar{z} = K^{-1}\sum_a z_a$, and the z_a are auxiliary i.i.d. standard Gaussians independent of the teacher fields. The first term is the centered teacher-aligned signal, with scale D. The second term is residual centered noise, with scale $\sqrt{\Delta}$. Since the softmax probabilities depend only on the centered logits h_a, this representation is used throughout the paper, simplifying the analysis by considering uncorrelated Gaussian variables u_a and z_a.

Let $g_a = p_a^{\mathrm{T}} - p_a$. In the thermodynamic limit, the permutation-symmetric online learning process closes exactly on the centered variables. With dots denoting derivatives with respect to α:

Lemma 1 (Centered closure)**.** *Under (A1)–(A7), the joint Gaussian law of the centered logits in the thermodynamic limit admits the representation Eq.* (8)*, and for fresh minibatches of size $b \geq 1$ the online dynamics closes exactly on the centered variables:*

$$\dot{D} = \frac{K}{K-1}\,\eta\,\langle g_1(u_1 - \bar{u})\rangle\,, \qquad \dot{Q}_{\text{eff}} = \frac{K}{K-1}\Big(2\eta\langle g_1 h_1\rangle + \frac{\eta^2}{b}\langle g_1^2\rangle\Big)\,, \qquad \dot{\Delta} = \dot{Q}_{\text{eff}} - 2D\dot{D}\,, \tag{9}$$

where the averages are over the centered Gaussian representation Eq. (8)*.*

No late-time $\alpha \to \infty$ approximation has been made yet. The derivation from the microscopic update is given in Section A.1.

4 Boundary layers control the late-time regime

Almost-perfect learning. The late-time regime we consider is characterized by $D \to \infty$ with bounded Δ. For large α the dynamics of Lemma 1 will enter this regime:

Lemma 2 (Global entry into the late-time regime)**.** *Let $0 \leq \eta(\alpha) \leq \eta_{\max} < \infty$ with $\int_0^\alpha \eta(s)\,ds \to \infty$. Under (A1)–(A7), for every admissible ($\Delta \geq 0$) finite permutation-symmetric initial state and every fixed finite b,*

$$\sup_{\alpha \geq 0} \Delta(\alpha) < \infty\,, \qquad D(\alpha) \to \infty\,. \tag{10}$$

A proof is given in Section A.2. In this regime, the signal term in Eq. (8) is large except near teacher decision boundaries: away from them, teacher and student agree exponentially well and examples

contribute exponentially little to the macroscopic flow. The leading contributions eventually come from pairwise boundary layers of width $O(D^{-1})$ around $u_a = u_b$. Higher-order ties are lower-dimensional sets and are subleading. While we derive the boundary layer results below for the limit $\alpha \to \infty$ and $D \to \infty$, in Section A.2.3 we provide a heuristic entry scale for α after which the boundary layer description becomes reasonable.

For a fixed pair $a \neq b$, the boundary is locally relevant only when the other $K-2$ teacher fields lie below the common value. This gives the geometric boundary density

$$c_K := \int_{-\infty}^{\infty} \varphi(s)^2 \Phi(s)^{K-2}\, ds, \tag{11}$$

where φ and Φ are the standard normal density and distribution function. All dependence on the number of classes enters the asymptotic prefactors through this boundary density and the number of pairwise boundaries.

Local binary reduction. Near one active boundary, scale the teacher gap as $u_a - u_b = \frac{x}{D}$. Then the corresponding centered student-logit gap has the form

$$h_a - h_b = x + \delta\,, \qquad \delta = \sqrt{2\Delta}\, z\,, \qquad z \sim \mathcal{N}(0,1)\,. \tag{12}$$

All other classes are exponentially suppressed at leading order, so the local softmax comparison is binary. The universal local update is $\Theta(x) - \sigma(x+\delta)$ with $\sigma(y) = 1/(1+e^{-y})$. The only non-elementary scalar function that remains is

$$\mathcal{B}(\Delta) = \int Dz \left[2\log\left(2\cosh\left(\sqrt{\frac{\Delta}{2}}\, z \right) \right) - 1 \right], \qquad Dz = \frac{e^{-z^2/2}}{\sqrt{2\pi}}\, dz\,. \tag{13}$$

For small Δ, $\mathcal{B}(\Delta) = 2\log 2 - 1 + \Delta/2 + O(\Delta^2)$.

Uniform boundary-layer expansion. The classification generalization error is

$$\epsilon_g := \Pr\left[\arg\max_a h_a \neq \arg\max_a u_a \right]. \tag{14}$$

It is governed by the same boundary layers: locally, a disagreement occurs when the signs of x and $x+\delta$ differ. The local reduction then gives the leading late-time behavior of all four macroscopic observables:

Theorem 3 (Boundary-layer expansion)**.** *Let c_K and $\mathcal{B}$ be as in Eqs.* (11) *and* (13)*. Under (A1)–(A7), as $D \to \infty$, uniformly for $\Delta \in [0, M]$ with fixed $M < \infty$,*

$$\dot{D} = \frac{K c_K\, \eta}{2D^2} \left(\frac{\pi^2}{6} + \Delta \right) + \eta\, o(D^{-2})\,, \tag{15}$$

$$\dot{\Delta} = \frac{K c_K}{D} \left[\frac{\eta^2}{b}\, \mathcal{B}(\Delta) - 2\eta\Delta \right] + \eta(\Delta + \eta/b)\, o(D^{-1})\,, \tag{16}$$

$$\epsilon_g = \Gamma_K \frac{\sqrt{\Delta}}{D} + O\left(\frac{\Delta}{D^2} \right), \qquad \Gamma_K = \frac{K(K-1)c_K}{\sqrt{\pi}}\,, \tag{17}$$

and $\mathcal{L}_g := \langle -\log p_y \rangle = \frac{K(K-1)c_K}{2D} \left(\frac{\pi^2}{6} + \Delta \right) + o(D^{-1})$.

The derivation is given in Section A.3 with explicit remainders. The loss and the error live on the same boundary layers but measure different aspects of them: $\mathcal{L}_g \propto D^{-1}(\pi^2/6 + \Delta)$ is dominated by the deterministic boundary-layer width $\propto 1/D$, while $\epsilon_g \propto \sqrt{\Delta}/D$ is a noise-to-signal angle of the student against the teacher. The factor D^{-2} in $\dot{D}$ connects to the derivative of the scale-D^{-1} loss, while the $\eta^2 \mathcal{B}(\Delta)/b$ term in $\dot{\Delta}$ is the surviving online-SGD noise. Fixed-learning-rate dynamics therefore does not drive Δ to zero; it drives Δ to a noise floor while D continues to grow:

Corollary 4 (Fixed learning rate: one-third law)**.** *For constant $\eta > 0$ and fixed b, Eq.* (16) *has the fixed point $2\Delta_* = \frac{\eta}{b}\, \mathcal{B}(\Delta_*)$ and the limiting trajectory satisfies*

$$\Delta(\alpha) \to \Delta_*\,, \qquad D(\alpha) \sim \left[\frac{3 K c_K}{2}\, \eta \left(\frac{\pi^2}{6} + \Delta_* \right) \right]^{1/3} \alpha^{1/3}\,. \tag{18}$$

Consequently, $\mathcal{L}_g(\alpha) \propto \alpha^{-1/3}$ and $\epsilon_g(\alpha) \propto \alpha^{-1/3}$.

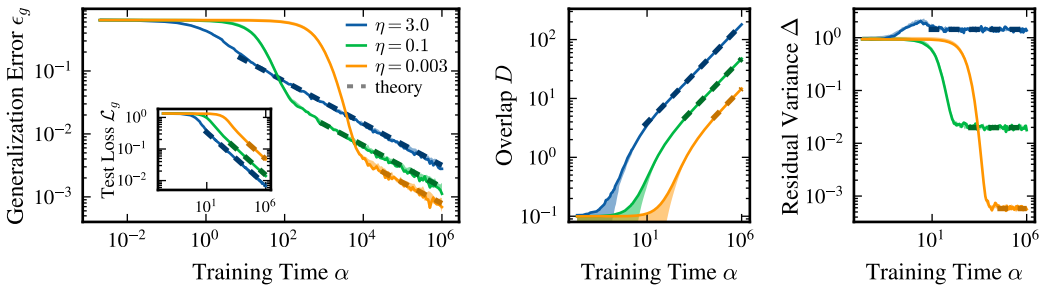

Figure 2: Finite-N validation for fixed learning rates in the $K = 3$ online teacher–student model. The panels show the generalization error, centered overlap D, and residual variance Δ as functions of macroscopic time $\alpha = \mu/N$. The curves show representative seed trajectories, with envelopes indicating fluctuations across six simulation seeds. Within these fluctuations, the trajectories agree with the predicted power-law prefactors and exponents: $D \propto \alpha^{1/3}$, Δ approaches a learning-rate-dependent floor, and $\epsilon_g \propto \sqrt{\Delta}/D \propto \alpha^{-1/3}$; see Eqs. (17) and (18).

The derivation is given in Section A.4. Within this solvable model, the one-third exponent is therefore not a consequence of a data spectrum. It follows from the boundary-layer drift $\dot{D} \propto D^{-2}$ together with a finite residual online-noise scale. Setting the online-noise term $\eta^2\mathcal{B}(\Delta)/b$ in Eq. (16) to zero (population gradient flow) would let Δ vanish faster than any power while $D \propto \alpha^{1/3}$ persists: the loss keeps its $\alpha^{-1/3}$ decay, consistent with Liu et al. (2026), while the error $\epsilon_g \propto \sqrt{\Delta}/D$ decays faster than any power. The one-third loss law is deterministic, the one-third error law is an online-SGD effect, invisible to gradient-flow analyses. A direct decomposition confirming that alignment growth and errors localize in the predicted boundary layers is reported in Section C.3.

5 Exponents with learning rate schedules

Taking a smaller fixed learning rate lowers the asymptotic noise floor, but it also delays entry into the asymptotic regime; see Fig. 2. Learning-rate schedules can improve the asymptotic error exponent because they allow the residual variance Δ to decrease while the accumulated learning time continues to grow. Suppose $\eta(\alpha) \to 0$ slowly enough that Δ adiabatically tracks the instantaneous fixed point $\Delta_*(\eta(\alpha))$ of Eq. (16). Adiabatic means that the relaxation rate $\lambda \simeq 2Kc_K\eta/D$ toward this fixed point dominates the schedule's relative rate of change $|\eta'|/\eta$. A small-Δ expansion of $\mathcal{B}$ gives $\Delta(\alpha) \sim \frac{\kappa}{b}\,\eta(\alpha)$ with $\kappa = (2\log 2 - 1)/2$. Then Eq. (15) reduces to $\dot{D} \sim \frac{Kc_K\pi^2}{12}\frac{\eta(\alpha)}{D^2}$. With

$$H(\alpha) = \int_0^\alpha \eta(\alpha')\,d\alpha', \tag{19}$$

one obtains

$$D(\alpha) \sim \left(\frac{Kc_K\pi^2}{4}H(\alpha)\right)^{1/3}, \qquad \Delta \sim \frac{\kappa}{b}\,\eta(\alpha)\,. \tag{20}$$

Together with Eq. (17), this establishes the results for power-law schedules:

Corollary 5 (Power-law schedules)**.** *For $\eta(\alpha) = \eta_0(1+\alpha/\alpha_0)^{-\gamma}$ with $0 < \gamma < 1$, the adiabatic tracking criterion holds, Eq.* (20) *applies, and*

$$\epsilon_g(\alpha) \sim A_{K,b}\,\frac{\sqrt{\eta(\alpha)}}{H(\alpha)^{1/3}} \propto \alpha^{-(2+\gamma)/6}, \qquad A_{K,b} = \Gamma_K\sqrt{\frac{\kappa}{b}}\left(\frac{4}{Kc_K\pi^2}\right)^{1/3}, \tag{21}$$

while by Theorem 3 the loss slows to $\mathcal{L}_g \propto \alpha^{-(1-\gamma)/3}$.

The error exponent therefore interpolates continuously from $1/3$ at $\gamma = 0$ toward $1/2$ as $\gamma \uparrow 1$. The endpoint is genuinely singular:

Proposition 6 (Borderline schedule)**.** *For $\gamma = 1$, Δ does not generically track the instantaneous fixed point: $\Delta(\alpha)/\eta(\alpha) \not\to \kappa/b$. Hence Corollary 5 does not apply at this endpoint.*

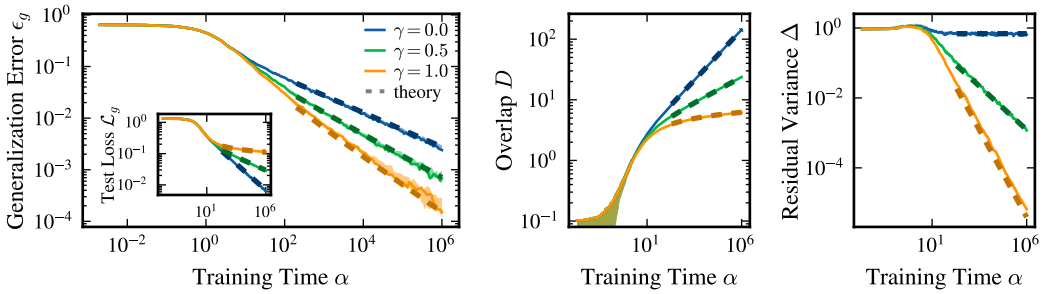

Figure 3: Schedule dependence in the $K = 3$ online teacher–student model. For $\eta(\alpha) \propto \alpha^{-\gamma}$, the theory predicts $\epsilon_g(\alpha) \propto \alpha^{-(2+\gamma)/6}$ for $0 \leq \gamma < 1$. Increasing γ slows the growth of the centered overlap, $D \propto \alpha^{(1-\gamma)/3}$ for $\gamma < 1$, but decreases the residual variance, $\Delta \propto \eta(\alpha)$; the latter effect improves the classification-error exponent. The $\gamma = 1$ curve is a borderline case, where the adiabatic approximation for Δ breaks down. For reference, the adiabatic predictions of Eqs. (20) and (21) are also shown.

Thus the exponent $1/2$ is approached from below within the adiabatic family, but is not attained by the borderline schedule itself (see Section A.5.2). If $\gamma > 1$, the integral $H(\alpha)$ converges and the centered overlap saturates at a finite value, so both the test loss and the classification error approach nonzero constants and power-law decay ceases. The derivations are collected in Section A.5.

6 Numerical validation and controlled departures

The simulations test the mechanism, not a universal empirical scaling law, first within the solvable model and then with correlated Gaussian inputs and whitened pretrained features as controlled departures from the theoretical assumptions. Numerical details, additional fixed-learning-rate sweeps, and the real-label whitened-feature comparison are reported in Section C.

At fixed learning rate, the finite-N simulations in Fig. 2 agree with Theorem 3 and Corollary 4, including test-loss decay, larger residual floors at larger learning rates, and later asymptotic onset at smaller ones. Additional simulations confirm the predicted K-dependence (Section C.1), batch-size dependence for $b > 1$ (Section C.2), and increasing localization of active examples and errors in the predicted boundary layers as D grows (Section C.3). The schedule sweep (Fig. 3) confirms both components of Eq. (20): annealing decreases Δ but slows the growth of accumulated learning time $H(\alpha)$ and D, improving the error exponent as in Eq. (21) while slowing loss decay. For $\gamma = 1$, adiabatic tracking fails asymptotically, yet the adiabatic prediction remains a good finite-time reference over the displayed range, where the error shows a near-$1/2$ power law.

The correlated-Gaussian experiment (Fig. 5) suggests that structured covariance mainly changes the crossover: larger β lengthens transients, while later error decay remains consistent with the boundary-layer prediction. This supports the interpretation that covariance structure and boundary-layer dynamics can control different time windows.

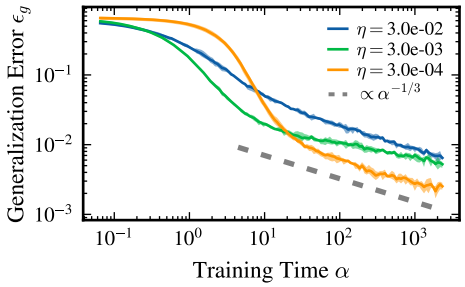

Figure 4: Non-Gaussian robustness test using whitened pretrained features. A linear softmax readout is trained online on whitened pretrained vision-transformer (ViT) features (Dosovitskiy et al., 2021) of CIFAR-5M images (Nakkiran et al., 2021) with teacher-generated labels. Whitening controls the covariance, while the feature distribution remains non-Gaussian. The teacher-generated labels avoid an early real-label performance floor and allow the late-time regime to be observed.

Fig. 4 relaxes the Gaussian assumption while controlling the covariance. The teacher-label run is consistent with the same qualitative scaling picture, suggesting that the mechanism is not immediately destroyed by non-Gaussian feature statistics. The

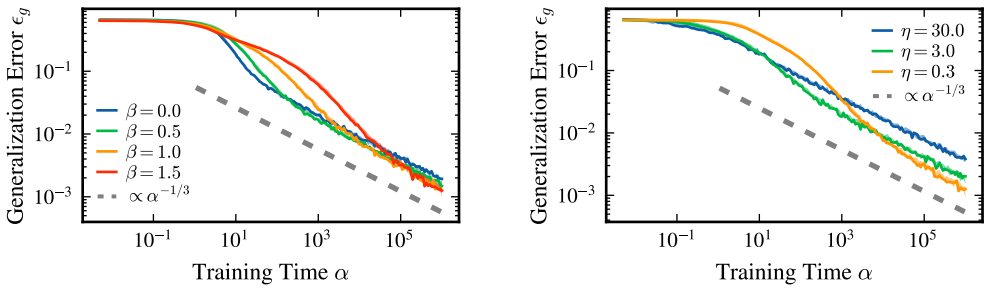

Figure 5: Controlled departure from isotropic inputs. Inputs are Gaussian with diagonal covariance spectrum $\langle \xi_i^2 \rangle \propto i^{-\beta}$, normalized so that the largest variance is one. Left: $\eta = 0.5$ is fixed and β is varied. Right: $\beta = 1.0$ is fixed and the learning rate is varied. Increasing β changes and lengthens the transient regime. The late-time decay remains consistent with the same boundary-layer asymptote, while smaller learning rates reduce asymptotic prefactors but can delay entry into the asymptotic regime.

corresponding real-label run (Fig. 9 in Section C.5) reaches a floor much earlier. We interpret this as a limitation of the model. For real labels, aspects such as model misspecification, label noise, or irreducible Bayes error can dominate the late-time error before the noiseless boundary-layer asymptote is cleanly visible.

Section C.6 adds two practical comparisons. Batch-size-one SGD on MNIST is consistent with one-third scaling of both test loss and generalization error. Further, we analyze the Pythia Transformer family (Biderman et al., 2023) similar to Liu et al. (2026), but concentrate on token top-1 error instead of loss. Token top-1 error probes relative-logit dynamics. Its excess above a fitted floor is well described by a power law over the selected window, with an exponent compatible with one-third for the 12-billion-parameter network.

7 Discussion and Scope

We have identified a boundary-layer mechanism for late-time online hard-label classification trained with softmax and cross-entropy in a solvable fixed-feature teacher-student model. For fixed learning rate, online gradient noise keeps the residual variance Δ finite, while only $O(D^{-1})$-thin decision-boundary layers remain active, which produces $D \propto \alpha^{1/3}$ and $\epsilon_g \propto \alpha^{-1/3}$.

Real-world learning problems can involve sub-tasks that occur at very different frequencies Michaud et al. (2023). Section D aggregates the fixed-rate law over Zipf-distributed sub-tasks (frequency $\propto k^{-\nu}$) in a minimal routed-expert model. The total error decays as $\alpha^{-\min\{1/3,\, c_\nu\}}$ with $c_\nu = 1 - 1/\nu$, so heavy frequency tails can mask the $1/3$ exponent in aggregate observables, whereas a student that represents only the most frequent sub-tasks approaches its coverage floor as $\alpha^{-1/3}$, and frequency-resolved errors show the $1/3$ for every sub-task. This motivates the power-law-plus-floor fits in Section C.6: Pythia's extensive training at fixed model size could predominantly refine already represented patterns, as in the late-time regime of the Zipf model.

Our analytical results are derived under assumptions (A1)–(A7): a one-layer readout with fixed features, online SGD on fresh examples, the thermodynamic limit, noiseless hard labels, and permutation symmetry. The theory does not claim that data structure is irrelevant, nor that feature learning cannot change the asymptotic class. Indeed, the correlated-input experiments show that structured covariance can indeed affect transients.

These limitations suggest concrete extensions. A soft-teacher or noisy-label version of the centered closure would connect the theory to Bayes-error floors in real data. Momentum should modify the online-noise term. Richer optimizers may change the range of useful learning-rate schedules, and feature-learning architectures could couple the boundary-layer dynamics of the readout to changes in the representation. The present work provides a baseline for those questions: a solvable classification scaling law whose exponent is generated by decision-boundary geometry and stochastic optimization noise.

AI USE STATEMENT

In this work, we used generative AI tools for the following tasks with required disclosure: help develop theoretical models, formulate mathematical claims, provide critical ingredients for proving mathematical claims, assist in the writing of proofs, provide feedback on research methodology and experiments, implement methods. We have not used generative AI tools for the following tasks: propose or refine hypotheses, design research methodology or experiments, assist with translation, support qualitative and thematic data analysis, interpret results. The following tasks are not applicable to this work: generate synthetic data sets, clean and reformat dataset. Additionally, we used generative AI tools for the following tasks with recommended disclosure: create or edit software code, draft parts of a research paper, summarize or analyse existing literature, searching for information, edit a research paper to improve readability, identify relevant literature, propose a title or keywords for a research paper. We have reviewed all AI-assisted work. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

REPRODUCIBILITY STATEMENT

We are committed to the reproducibility of our research. Our theoretical claims, including all assumptions, are presented in Sections 3 to 5 with derivations provided in Section A. Comprehensive details for reproducing our empirical results, including model architectures, data processing, and hyperparameter settings, are well documented in Section C. Source code is provided as supplementary material.

REFERENCES

Madhu S Advani, Andrew M Saxe, and Haim Sompolinsky. High-dimensional dynamics of generalization error in neural networks. *Neural Networks*, 132:428–446, 2020.

Benjamin Aubin, Florent Krzakala, Yue Lu, and Lenka Zdeborová. Generalization error in high-dimensional perceptrons: Approaching bayes error with convex optimization. *Advances in Neural Information Processing Systems*, 33:12199–12210, 2020.

Yasaman Bahri, Ethan Dyer, Jared Kaplan, Jaehoon Lee, and Utkarsh Sharma. Explaining neural scaling laws. *Proceedings of the National Academy of Sciences*, 121(27):e2311878121, 2024.

Maissam Barkeshli, Alberto Alfarano, and Andrey Gromov. On the Origin of Neural Scaling Laws: from Random Graphs to Natural Language. In *Forty-third International Conference on Machine Learning*, 2026.

Peter L Bartlett, Michael I Jordan, and Jon D McAuliffe. Convexity, classification, and risk bounds. *Journal of the American Statistical Association*, 101(473):138–156, 2006.

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

Michael Biehl and Peter Riegler. On-Line Learning with a Perceptron. *Europhysics Letters*, 28(7): 525–530, 1994.

Blake Bordelon, Abdulkadir Canatar, and Cengiz Pehlevan. Spectrum dependent learning curves in kernel regression and wide neural networks. In *International Conference on Machine Learning*, pp. 1024–1034. PMLR, 2020.

Blake Bordelon, Alexander Atanasov, and Cengiz Pehlevan. A Dynamical Model of Neural Scaling Laws. In *Forty-first International Conference on Machine Learning*, 2024.

Blake Bordelon, Alexander Atanasov, and Cengiz Pehlevan. How feature learning can improve neural scaling laws. *Journal of Statistical Mechanics: Theory and Experiment*, 2025(8):084002, 2025.

Abdulkadir Canatar, Blake Bordelon, and Cengiz Pehlevan. Spectral bias and task-model alignment explain generalization in kernel regression and infinitely wide neural networks. *Nature communications*, 12(1):2914, 2021.

Elisabetta Cornacchia, Francesca Mignacco, Rodrigo Veiga, Cédric Gerbelot, Bruno Loureiro, and Lenka Zdeborová. Learning Curves for the Multi-Class Teacher–Student Perceptron. *Machine Learning: Science and Technology*, 4(1):015019, 2023.

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pp. 248–255. Ieee, 2009.

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

Samir Yitzhak Gadre, Georgios Smyrnis, Vaishaal Shankar, Suchin Gururangan, Mitchell Wortsman, Rulin Shao, Jean Mercat, Alex Fang, Jeffrey Li, Sedrick Keh, et al. Language models scale reliably with over-training and on downstream tasks. In *International Conference on Learning Representations*, volume 2025, pp. 67661–67682, 2025.

Sebastian Goldt, Madhu Advani, Andrew M Saxe, Florent Krzakala, and Lenka Zdeborová. Dynamics of stochastic gradient descent for two-layer neural networks in the teacher-student setup. *Advances in neural information processing systems*, 32, 2019.

Joel Hestness, Sharan Narang, Newsha Ardalani, Gregory Diamos, Heewoo Jun, Hassan Kianinejad, Md Mostofa Ali Patwary, Yang Yang, and Yanqi Zhou. Deep learning scaling is predictable, empirically. *arXiv preprint arXiv:1712.00409*, 2017.

Jordan Hoffmann, Sebastian Borgeaud, Arthur Mensch, Elena Buchatskaya, Trevor Cai, Eliza Rutherford, Diego de Las Casas, Lisa Anne Hendricks, Johannes Welbl, Aidan Clark, et al. An empirical analysis of compute-optimal large language model training. *Advances in neural information processing systems*, 35:30016–30030, 2022.

Marcus Hutter. Learning curve theory. *arXiv preprint arXiv:2102.04074*, 2021.

Jared Kaplan, Sam McCandlish, Tom Henighan, Tom B Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeffrey Wu, and Dario Amodei. Scaling laws for neural language models. *arXiv preprint arXiv:2001.08361*, 2020.

Licong Lin, Jingfeng Wu, Sham M Kakade, Peter L Bartlett, and Jason D Lee. Scaling laws in linear regression: Compute, parameters, and data. *Advances in Neural Information Processing Systems*, 37:60556–60606, 2024.

Yizhou Liu, Ziming Liu, Cengiz Pehlevan, and Jeff Gore. Universal One-third Time Scaling in Learning Peaked Distributions. In *Forty-third International Conference on Machine Learning*, 2026.

Bruno Loureiro, Cedric Gerbelot, Hugo Cui, Sebastian Goldt, Florent Krzakala, Marc Mezard, and Lenka Zdeborová. Learning curves of generic features maps for realistic datasets with a teacher-student model. *Advances in Neural Information Processing Systems*, 34:18137–18151, 2021.

Alexander Maloney, Daniel A Roberts, and James Sully. A solvable model of neural scaling laws. *arXiv preprint arXiv:2210.16859*, 2022.

Anqi Mao, Mehryar Mohri, and Yutao Zhong. A universal growth rate for learning with smooth surrogate losses. *Advances in neural information processing systems*, 37:41670–41708, 2024.

Eric Michaud, Ziming Liu, Uzay Girit, and Max Tegmark. The Quantization Model of Neural Scaling. *Advances in Neural Information Processing Systems*, 36:28699–28722, 2023.

Francesca Mignacco, Florent Krzakala, Pierfrancesco Urbani, and Lenka Zdeborová. Dynamical mean-field theory for stochastic gradient descent in gaussian mixture classification. *Advances in Neural Information Processing Systems*, 33:9540–9550, 2020.

Mor Shpigel Nacson, Nathan Srebro, and Daniel Soudry. Stochastic gradient descent on separable data: Exact convergence with a fixed learning rate. In *The 22nd International Conference on Artificial Intelligence and Statistics*, pp. 3051–3059. PMLR, 2019.

Ryumei Nakada and Masaaki Imaizumi. Adaptive approximation and generalization of deep neural network with intrinsic dimensionality. *Journal of Machine Learning Research*, 21(174):1–38, 2020.

Preetum Nakkiran, Behnam Neyshabur, and Hanie Sedghi. The Deep Bootstrap Framework: Good Online Learners are Good Offline Generalizers. In *International Conference on Learning Representations*, 2021.

Yoonsoo Nam, Nayara Fonseca, Seok H Lee, Chris Mingard, and Ard A Louis. An exactly solvable model for emergence and scaling laws in the multitask sparse parity problem. *Advances in Neural Information Processing Systems*, 37:39632–39693, 2024.

Manfred Opper and David Haussler. Calculation of the learning curve of Bayes optimal classification algorithm for learning a perceptron with noise. In *Colt*, volume 91, pp. 75–87, 1991.

Jonas Paccolat, Leonardo Petrini, Mario Geiger, Kevin Tyloo, and Matthieu Wyart. Geometric compression of invariant manifolds in neural networks. *Journal of Statistical Mechanics: Theory and Experiment*, 2021(4):044001, 2021.

Guilherme Penedo, Hynek Kydlíček, Anton Lozhkov, Margaret Mitchell, Colin Raffel, Leandro Von Werra, Thomas Wolf, et al. The Fineweb Datasets: Decanting the Web for the Finest Text Data at Scale. *Advances in Neural Information Processing Systems*, 37:30811–30849, 2024.

Jack W Rae, Sebastian Borgeaud, Trevor Cai, Katie Millican, Jordan Hoffmann, Francis Song, John Aslanides, Sarah Henderson, Roman Ring, Susannah Young, et al. Scaling language models: Methods, analysis & insights from training gopher. *arXiv preprint arXiv:2112.11446*, 2021.

Maziar Raissi, Paris Perdikaris, and George E Karniadakis. Physics-informed neural networks: A deep learning framework for solving forward and inverse problems involving nonlinear partial differential equations. *Journal of Computational physics*, 378:686–707, 2019.

Hrithik Ravi, Clayton Scott, Daniel Soudry, and Yutong Wang. The implicit bias of gradient descent on separable multiclass data. *Advances in neural information processing systems*, 37:81324–81359, 2024.

Frederieke Richert, Roman Worschech, and Bernd Rosenow. Soft mode in the dynamics of overrealizable online learning for soft committee machines. *Physical Review E*, 105(5):L052302, 2022.

David Saad and Sara A Solla. Exact solution for on-line learning in multilayer neural networks. *Physical Review Letters*, 74(21):4337, 1995.

Rylan Schaeffer, Hailey Schoelkopf, Brando Miranda, Gabriel Mukobi, Varun Madan, Adam Ibrahim, Herbie Bradley, Stella Biderman, and Sanmi Koyejo. Why Has Predicting Downstream Capabilities of Frontier AI Models with Scale Remained Elusive? In *Forty-second International Conference on Machine Learning*, 2025.

Antonio Sclocchi and Matthieu Wyart. On the different regimes of stochastic gradient descent. *Proceedings of the National Academy of Sciences*, 121(9):e2316301121, 2024.

Hyunjune Sebastian Seung, Haim Sompolinsky, and Naftali Tishby. Statistical mechanics of learning from examples. *Physical review A*, 45(8):6056, 1992.

Utkarsh Sharma and Jared Kaplan. Scaling laws from the data manifold dimension. *Journal of Machine Learning Research*, 23(9):1–34, 2022.

Daniel Soudry, Elad Hoffer, Mor Shpigel Nacson, Suriya Gunasekar, and Nathan Srebro. The implicit bias of gradient descent on separable data. *Journal of Machine Learning Research*, 19(70): 1–57, 2018.

Yutong Wang and Clayton Scott. Unified binary and multiclass margin-based classification. *Journal of Machine Learning Research*, 25(143):1–51, 2024.

Roman Worschech and Bernd Rosenow. Analyzing neural scaling laws in two-layer networks with power-law data spectra. In *International Conference on Learning Representations*, volume 2025, pp. 89098–89145, 2025.

Junxi Yan, Zixi Wei, Qingyao Ai, Yiqun Liu, and Jingtao Zhan. What scales in cross-entropy scaling law? In *International Conference on Learning Representations*, volume 2026, pp. 79741–79758, 2026.

Xiaohua Zhai, Alexander Kolesnikov, Neil Houlsby, and Lucas Beyer. Scaling vision transformers. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1204–1213. IEEE, 2022.

Tong Zhang. Statistical behavior and consistency of classification methods based on convex risk minimization. *The Annals of Statistics*, 32(1):56–85, 2004.

A Proofs of the theoretical results

This section proves Lemmas 1 and 2, Theorem 3, and Corollaries 4 and 5 in order. Proposition 6 is treated together with the last corollary. Throughout, $K \geq 2$ and the fresh batch size b are fixed. We first take the thermodynamic limit at fixed $\alpha = \mu/N$, as in (A6), and then study the late-time behavior of the resulting ODE.

A.1 Lemma 1: centered closure

The teacher fields $u_a = T_a \cdot \xi/\sqrt{N}$ are i.i.d. standard Gaussians, and $t_a = J_a \cdot \xi/\sqrt{N}$ are the student logits. Under the permutation-symmetric ansatz, the macroscopic state is described by

$$R = \frac{J_1 \cdot T_1}{N}\,, \qquad S = \frac{J_1 \cdot T_2}{N}\,, \qquad Q = \frac{J_1 \cdot J_1}{N}\,, \qquad C = \frac{J_1 \cdot J_2}{N}\,. \tag{22}$$

For a fresh batch of b independent examples, the update is

$$J_a^{\mu+1} = J_a^{\mu} + \frac{\eta}{b\sqrt{N}} \sum_{i=1}^{b} g_a^i \xi^i\,, \qquad g_a^i = p_a^{\mathrm{T},i} - p_a^i\,. \tag{23}$$

Here $p_a^{\mathrm{T}} = \mathbf{1}\{a = y\}$, $y = \arg\max_c u_c$, and $p_a = e^{t_a}/\sum_c e^{t_c}$. We use the standard thermodynamic-limit overlap equations of online teacher–student learning (Biehl & Riegler, 1994; Saad & Solla, 1995). Rigorous fixed-time convergence to such ODEs in closely related Gaussian teacher–student models is proven in Goldt et al. (2019). For a perceptron derivation retaining batch-dependent noise, see Sclocchi & Wyart (2024). With our batch-averaged update and optimizer-step time, the corresponding overlap equations are

$$\dot{R} = \eta \langle g_1 u_1 \rangle\,, \tag{24}$$

$$\dot{S} = \eta \langle g_1 u_2 \rangle\,, \tag{25}$$

$$\dot{Q} = 2\eta \langle g_1 t_1 \rangle + \frac{\eta^2}{b} \langle g_1^2 \rangle\,, \tag{26}$$

$$\dot{C} = \eta \langle g_1 t_2 + g_2 t_1 \rangle + \frac{\eta^2}{b} \langle g_1 g_2 \rangle\,. \tag{27}$$

Brackets denote averages over the jointly Gaussian teacher and student fields at fixed order parameters. These are the thermodynamic-limit, $N \to \infty$, overlap equations, not connected to a late-time assumption. We do not repeat their stochastic-process convergence proof here.

For orientation, the single-step increments of R and Q are

$$\delta R = \frac{\eta}{Nb} \sum_i g_1^i u_1^i \,,$$
$$\delta Q = \frac{2\eta}{Nb} \sum_i g_1^i t_1^i + \frac{\eta^2}{N^2 b^2} \sum_{i,j} g_1^i g_1^j \, \xi^i \cdot \xi^j \,.$$

Multiplication by N converts these increments to derivatives in α. In the quadratic term, the b diagonal contributions have $\|\xi^i\|^2/N \to 1$ and give the factor $1/b$. The cross-example terms have conditional mean $O(N^{-2})$ per update and vanish on this time scale. Thus batching changes the quadratic noise term, not the linear drift. Counting processed examples instead of optimizer steps would also rescale time, whereas we define it as $\alpha = \mu/N$ for all values of b.

A.1.1 Centered Gaussian representation

Let $\bar{u} = K^{-1} \sum_a u_a$, $\bar{t} = K^{-1} \sum_a t_a$, and $h_a = t_a - \bar{t}$. Teacher labels and softmax probabilities are invariant under common shifts of their fields. A covariance calculation from Eq. (22) gives

$$\langle h_a (u_c - \bar{u}) \rangle = D(\delta_{ac} - 1/K) \,, \qquad \langle h_a h_c \rangle = Q_{\text{eff}} (\delta_{ac} - 1/K) \,. \tag{28}$$

Consequently, $\langle h_a \mid u_1, \ldots, u_K \rangle = D(u_a - \bar{u})$. The residual is independent of the teacher fields, is Gaussian, and has covariance $\Delta(\delta_{ac} - 1/K)$. This establishes $\Delta \geq 0$ and the representation

$$h_a = D(u_a - \bar{u}) + \sqrt{\Delta}\,(z_a - \bar{z}) \,, \qquad \Delta = Q_{\text{eff}} - D^2 \,, \tag{29}$$

where the z_a are independent standard Gaussians, independent of the u_a, and $\bar{z} = K^{-1} \sum_a z_a$.

A.1.2 Exact centered closure

Subtracting Eq. (25) from Eq. (24) and using class symmetry gives

$$\dot{D} = \eta \langle g_1 (u_1 - u_2) \rangle = \frac{K\eta}{K-1} \langle g_1 (u_1 - \bar{u}) \rangle \,. \tag{30}$$

Likewise, $\sum_a g_a = 0$ implies $\langle g_1 g_2 \rangle = -\langle g_1^2 \rangle / (K-1)$. Subtracting Eq. (27) from Eq. (26) therefore yields

$$\dot{Q}_{\text{eff}} = \frac{K}{K-1} \left[2\eta \langle g_1 h_1 \rangle + \frac{\eta^2}{b} \langle g_1^2 \rangle \right] , \tag{31}$$

and by definition of Δ

$$\dot{\Delta} = \dot{Q}_{\text{eff}} - 2D\dot{D} \,. \tag{32}$$

This proves Lemma 1. Two equivalent forms will be useful below. Gaussian integration by parts in the independent z_a gives $\langle g_1 (z_1 - \bar{z}) \rangle = -\sqrt{\Delta} \langle p_1 (1 - p_1) \rangle$: the common-mode derivative cancels because softmax is shift invariant, and $\partial_{h_1} g_1 = -p_1 (1 - p_1)$. Thus

$$\dot{\Delta} = \frac{K\eta}{K-1} \left[-2\Delta \langle p_1 (1 - p_1) \rangle + \frac{\eta}{b} \langle g_1^2 \rangle \right] . \tag{33}$$

The contraction is proportional to Δ exactly, before any asymptotic expansion. Also, summing $g_a u_a$ over classes gives

$$\begin{aligned} \dot{D} &= \frac{\eta}{K-1} \left\langle u_y - \sum_a p_a u_a \right\rangle \\ &= \frac{\eta}{K-1} \left\langle \sum_a p_a (u_y - u_a) \right\rangle \\ &\geq 0 \,. \end{aligned} \tag{34}$$

A.2 LEMMA 2: GLOBAL ENTRY INTO THE LATE-TIME REGIME

The derivation first bounds the positive noise source term relative to the negative contraction term proportional to Δ in Eq. (33), without assuming large D. This bounds Δ; positivity of Eq. (34) then forces D to diverge.

A.2.1 COMPARING THE SOURCE AND CONTRACTION

For $D \geq 0$ and $\Delta \geq 0$, we claim that

$$\langle g_1^2 \rangle \leq C_K(1+\sqrt{\Delta})\langle p_1(1-p_1)\rangle \tag{35}$$

for some constant C_K. To see this, use the uncentered logits $Du_a + \sqrt{\Delta}z_a$, which are independent Gaussians of variance $Q_{\text{eff}} = D^2 + \Delta$. Let G denote the top gap of K independent standard Gaussians. Choosing their ordered winner and runner-up gives its density

$$f_G(g) = K(K-1)\int_{-\infty}^{\infty} \varphi(v+g)\varphi(v)\Phi(v)^{K-2}\,dv, \qquad g \geq 0\,. \tag{36}$$

This density is bounded, is positive near zero, and has a bounded derivative. On $\sqrt{Q_{\text{eff}}}G \leq 1$, the two largest softmax probabilities are at least $1/K$ and e^{-1}/K. Since $K\langle p_1(1-p_1)\rangle = \langle 1 - \sum_a p_a^2\rangle$, integration over this event gives a lower bound. Conversely, the mass outside the student winner $\hat{y}$ is at most $(K-1)e^{-\sqrt{Q_{\text{eff}}}G}$. Integrating this bound gives

$$\langle p_1(1-p_1)\rangle \geq \frac{c}{1+\sqrt{Q_{\text{eff}}}}, \qquad \langle 1 - p_{\hat{y}}\rangle \leq \frac{C}{1+\sqrt{Q_{\text{eff}}}}, \tag{37}$$

with positive constants depending only on K. These bounds include $\sqrt{Q_{\text{eff}}} = 0$, when softmax is uniform.

For $D > 0$, a teacher–student disagreement requires the teacher top gap to be no larger than $(\sqrt{\Delta}/D)\max_{a,c}|z_a - z_c|$. The bounded gap density and finite Gaussian moments therefore imply

$$\Pr(\hat{y} \neq y) \leq \min\{1, C_K\sqrt{\Delta}/D\}. \tag{38}$$

Finally, $\sum_a g_a^2 \leq 2(1-p_y)$ and $1 - p_y \leq \mathbf{1}\{\hat{y} \neq y\} + 1 - p_{\hat{y}}$ give

$$\frac{\langle g_1^2\rangle}{\langle p_1(1-p_1)\rangle} \leq C_K\big[1 + (1+\sqrt{Q_{\text{eff}}})\Pr(\hat{y} \neq y)\big] \leq C_K'(1+\sqrt{\Delta}).$$

The last inequality follows by separating $D \leq \max\{1, \sqrt{\Delta}\}$ from its complement; at $D = 0$ use the trivial disagreement bound. This proves Eq. (35).

A.2.2 BOUNDED RESIDUAL VARIANCE AND DIVERGING ALIGNMENT

Write $\mu_K = \langle \max_a u_a \rangle > 0$. Integration by parts in u_a gives the additional identity

$$\dot{D} = \frac{\eta}{K-1}\left[\mu_K - KD\langle p_1(1-p_1)\rangle\right]. \tag{39}$$

While $D \leq 0$, this is at least $\eta\mu_K/(K-1)$. Thus any negative initial alignment reaches zero after finite accumulated learning time. During this interval, Eq. (33) gives $\dot{\Delta} \leq K\eta_{\max}\eta/[b(K-1)]$, so at the crossing

$$\Delta \leq \Delta(0) + \frac{K\eta_{\max}}{b\mu_K}(-D(0))_+.$$

For $D \geq 0$, Eq. (35) instead gives

$$\dot{\Delta} \leq \frac{K\eta}{K-1}\langle p_1(1-p_1)\rangle\left[-2\Delta + \frac{\eta}{b}C_K(1+\sqrt{\Delta})\right]. \tag{40}$$

Choose M above the initial/transient bound and large enough that $2M > (\eta_{\max}/b)C_K(1+\sqrt{M})$. At $\Delta = M$ the drift cannot point outward, so $0 \leq \Delta(\alpha) \leq M$ for all times.

By Eq. (34), D is nondecreasing. If it had a finite limit, (D, Δ) would remain in a compact set. At every finite state the average in Eq. (34) is strictly positive: teacher ties have probability zero, and every softmax probability is positive. This average is continuous in (D, Δ), so it has a positive minimum on that compact set. Hence $\dot{D} \geq c\eta$, contradicting $\int_0^\alpha \eta \to \infty$ and boundedness of D. This proves Lemma 2 without using the boundary-layer expansion.

A.2.3 HEURISTIC LATE-TIME REGIME ENTRY SCALE

Even though our theoretical analysis considers the limit $\alpha \to \infty$ and therefore $D \to \infty$, for finite-time numerics we can give a heuristic entry scale for α such that the boundary layer description becomes reasonable. Consider that the two largest Gaussian teacher fields are typically separated by only $O(1/\sqrt{2\log K})$, so the corresponding student-logit gap is $O(D/\sqrt{2\log K})$. Thus the classes are determined exponentially well away from decision boundaries under the condition $D \gg \sqrt{2\log K}$. For bounded moderate residual variance, the crossing $D(\alpha_*) \approx \sqrt{2\log K}$ serves as a heuristic entry scale. Combining it with Eq. (18) and the leading large-K estimate $c_K \approx \sqrt{2\log K}/K^2$ gives

$$\alpha_* \approx \frac{4K\log K}{3\eta\,(\pi^2/6 + \Delta_*)}, \tag{41}$$

predicting later crossover for smaller η and larger K.

A.3 THEOREM 3: UNIFORM BOUNDARY-LAYER EXPANSION

Lemma 2 places every trajectory, eventually, in the regime

$$D \to \infty\,, \qquad 0 \le \Delta \le M < \infty\,. \tag{42}$$

The task is now to evaluate the four Gaussian averages that the macroscopic dynamics needs, namely $\dot{D}$, the contraction $\langle p_1(1-p_1)\rangle$, the noise source $\langle g_1^2\rangle$ and the test loss $\mathcal{L}_g = \langle -\log p_y\rangle$, to leading order in this regime, with remainders that hold uniformly for Δ in $[0, M]$, so the expansion applies also when Δ varies with time or tends to zero.

We first consider geometry: how much input probability lies within a teacher boundary layer of width $O(D^{-1})$, the answer being $O(D^{-1})$ with a prefactor set by the boundary density c_K. Then we consider the local problem: inside such a layer only two classes compete, and after rescaling the teacher gap by D all four averages reduce to elementary one-dimensional integrals of explicit binary kernels. We show that everything else, meaning both the rest of the input space and the difference between the true multiclass integrand and the idealized binary one, contributes less. That last part produces the remainders, and we identify which of them dominates.

A.3.1 BOUNDARY DENSITY AND LOCAL BINARY INTEGRALS

Fix an unordered pair $\{a, c\}$. On $u_a = u_c = s$, all other fields must lie below s for this pair to compete for the label. The single-pair boundary density is therefore

$$c_K = \int_{-\infty}^{\infty} \varphi(s)^2 \Phi(s)^{K-2}\, ds. \tag{43}$$

For a nonnegative gap q, let $w_K(q) = f_G(q)/[K(K-1)]$ be the density for one specified ordered top pair. Then $w_K(0) = c_K$, and Eq. (36) gives

$$|w_K(q) - c_K| \le C_K q \quad (0 \le q \le 1), \qquad 0 \le w_K(q) \le \frac{e^{-q^2/4}}{2\sqrt{\pi}} \quad (q \ge 0). \tag{44}$$

In particular, $\Pr(G \le \lambda/D) = K(K-1)c_K\lambda/D + O_{K,\lambda}(D^{-2})$ for fixed λ. This is the $O(D^{-1})$ measure of the active layer.

Near a boundary, write $u_a - u_c = x/D$. The student gap is then $h_a - h_c = x + \delta$, where $\delta = \sqrt{2\Delta}\, z$ and $z \sim \mathcal{N}(0,1)$. The local residual is $F_\delta(x) := \Theta(x) - \sigma(x+\delta)$, with $\sigma(y) = (1+e^{-y})^{-1}$. We use the elementary identities

$$A_0(\delta) := \int_{\mathbb{R}} F_\delta(x)\, dx = -\delta, \tag{45}$$

$$A_1(\delta) := \int_{\mathbb{R}} x F_\delta(x)\, dx = \frac{\pi^2}{6} + \frac{\delta^2}{2}, \tag{46}$$

$$A_2(\delta) := \int_{\mathbb{R}} (x+\delta) F_\delta(x)\, dx = \frac{\pi^2}{6} - \frac{\delta^2}{2}, \tag{47}$$

$$B_0(\delta) := \int_{\mathbb{R}} F_\delta(x)^2\, dx = 2\log\left(2\cosh\frac{\delta}{2}\right) - 1. \tag{48}$$

For example, differentiating A_1 gives $A_1'(\delta) = \delta$, while $A_1(0) = 2\int_0^\infty x/(1+e^x)\,dx = \pi^2/6$; also $A_2 = A_1 + \delta A_0$. The local loss satisfies

$$\mathcal{L}_0(\delta) := \int_0^\infty \left[\log(1+e^{-x-\delta}) + \log(1+e^{-x+\delta})\right] dx = \frac{\pi^2}{6} + \frac{\delta^2}{2}. \tag{49}$$

Here differentiation twice in δ gives one, and the value at zero is $\pi^2/6$. Averaging with $Dz = (2\pi)^{-1/2}e^{-z^2/2}\,dz$ yields

$$\langle A_1(\delta)\rangle_\delta = \frac{\pi^2}{6} + \Delta, \tag{50}$$

$$\langle A_2(\delta)\rangle_\delta = \frac{\pi^2}{6} - \Delta, \tag{51}$$

$$\mathcal{B}(\Delta) := \langle B_0(\delta)\rangle_\delta = \int Dz \left[2\log\left(2\cosh\left(\sqrt{\frac{\Delta}{2}}z\right)\right) - 1\right]. \tag{52}$$

In addition, symmetry of δ and $\int_{\mathbb{R}} \sigma'(x+\delta)\,dx = 1$ give $\int_0^\infty \langle \sigma'(x+\delta)\rangle_\delta\,dx = 1/2$.

A.3.2 Localization and the sources of the remainders

A binary calculation alone does not bound the multiclass error. We separate four regions before extending the binary integrals to the whole line. Let I, J be the teacher winner and runner-up, $G = u_I - u_J$, and $G_2 = u_{(2)} - u_{(3)}$; set $G_2 = \infty$ for $K = 2$. Choose $L = 8\log D$ and first restrict to

$$\sqrt{\Delta}\max_{a,c}|z_a - z_c| \le L/4. \tag{53}$$

On this event, partition teacher configurations into the bulk $G > L/D$, an isolated pair $G \le L/D$, $G_2 > 4L/D$, and near-triple ties $G \le L/D$, $G_2 \le 4L/D$.

Bulk and isolated pairs. In the bulk under Eq. (53), every student gap from the teacher winner is at least $3L/4$. $p_1(1-p_1)$, g_1^2, and the loss are therefore $O_K(e^{-3L/4})$. For the alignment drift use Eq. (34): a competitor with teacher gap $t \ge G$ contributes at most

$$te^{-Dt+L/4} \le \frac{L}{D}e^{-3L/4},$$

since $Dt \ge L > 1$. Thus its bulk contribution is smaller still.

For an isolated pair, put $x = DG \in [0, L]$ and $\delta = \sqrt{\Delta}(z_I - z_J)$. Every other class is at least $3L$ below the runner-up in student-logit units, giving

$$p_I = \sigma(x+\delta) + O_K(e^{-3L}), \qquad p_J = 1 - \sigma(x+\delta) + O_K(e^{-3L}). \tag{54}$$

In the binary limit $g_J = -g_I$, so the common boundary coordinate cancels in $\sum_a g_a u_a = g_I(u_I - u_J)$. The alignment contribution is consequently

$$\frac{p_J G}{K-1} = \frac{1}{(K-1)D}\frac{x}{1+e^{x+\delta}}. \tag{55}$$

This exact cancellation before integration supplies one factor D^{-1}; the layer measure supplies the other. The corresponding kernels for contraction, source, and loss are $\sigma'(x+\delta)$, $(1-\sigma(x+\delta))^2$, and $\log(1+e^{-x-\delta})$. The exponentially small errors in Eq. (54), including the teacher-gap-weighted third-class terms, are below the remainder orders stated below.

Near-triple ties. For $K \ge 3$, the joint density of (G, G_2) is bounded. For a specified ordered top triple it is obtained by integrating $\varphi(s+g+h)\varphi(s+h)\varphi(s)\Phi(s)^{K-3}$ over s. Hence

$$\Pr(G \le L/D,\ G_2 \le 4L/D) = O_K(L^2/D^2). \tag{56}$$

The contraction and source integrands are bounded, giving $O_K(L^2/D^2)$ contributions. Under the noise bound the loss is at most $L + \log K$, giving $O_K(L^3/D^2)$. For alignment, each gap $t \le 2L/D$ contributes at most $2L/D$; for $t > 2L/D$, the bound $te^{-Dt+L} \le (2L/D)e^{-L}$ applies. Thus the whole alignment integrand is $O_K(L/D)$, giving $O_K(L^3/D^3)$ after multiplication by the near-triple mass. The binary terms added across this set obey the same bounds. For $K = 2$ this region is absent.

Noise and binary tails. For $0 \le \Delta \le M$, a Gaussian union bound makes the probability of violating Eq. (53) at most $C_K e^{-c_M L^2}$ (zero when $M = 0$). The loss is bounded by $\log K + \sqrt{M}\max_{a,c}|z_a - z_c|$, and the alignment integrand by the teacher-field range. Thus the relevant Gaussian moments are bounded uniformly in D, and Cauchy–Schwarz bounds the discarded contributions by a quantity smaller than every inverse power of D. To remove the gap cutoff, consider any of the four nonnegative binary kernels

$$k(x,\delta) \in \left\{ \frac{x}{1+e^{x+\delta}},\ \sigma'(x+\delta),\ (1-\sigma(x+\delta))^2,\ \log(1+e^{-x-\delta}) \right\}.$$

Splitting the noise average into $\delta \ge -x/2$ and its complement gives, uniformly for $0 \le \Delta \le M$,

$$\langle k(x,\delta)\rangle_\delta \le C_M(1+x)\big(e^{-x/2} + e^{-c_M x^2}\big), \qquad x \ge 0. \tag{57}$$

On the first event this follows from logistic tails; on the second, from the Gaussian tail and, for the loss, $\log(1+e^{-x-\delta}) \le \log 2 + |\delta|$. In particular all polynomially weighted x moments are uniformly finite. The global density bound in Eq. (44) then gives

$$\int_L^\infty w_K(x/D)\langle k(x,\delta)\rangle_\delta\, dx \le C_{K,M}(1+L)\big(e^{-L/2} + e^{-c'_M L^2}\big). \tag{58}$$

With $L = 8\log D$, this is $O_{K,M}((1+\log D)D^{-4})$ plus a superpolynomially small term. The outside prefactor is D^{-2} for alignment and D^{-1} for the other averages. Thus cutoff removal costs $o(D^{-3})$ and $o(D^{-2})$, respectively, and introduces no additional logarithm.

Finally, replacing $w_K(x/D)$ by c_K costs $O_{K,M}(D^{-1})$ inside any of these integrals: for $x \le D$, use $|w_K(x/D) - c_K| \le C_K x/D$ and the first weighted kernel moment; for $x > D$, use Eq. (57). The resulting density corrections are $O(D^{-3})$ for alignment and $O(D^{-2})$ otherwise. For $K \ge 3$, the largest retained bounds are therefore the near-triple bounds, not the bulk, noise, density correction, or added tails. The logarithms reflect the chosen localization estimates; we do not claim that these bounds are sharp asymptotic corrections.

A.3.3 Asymptotic flow and test loss

There are $K(K-1)$ ordered top pairs, each with gap density $w_K(q)\,dq$. Using Eq. (55) and $q = x/D$ therefore gives

$$\begin{aligned}\dot D &= \frac{K\eta}{D^2}\int_0^\infty w_K(x/D)\left\langle \frac{x}{1+e^{x+\delta}}\right\rangle_\delta dx + \eta O_{K,M}\left(\frac{(\log D)^3}{D^3}\right)\\ &= \frac{Kc_K\eta}{2D^2}\left(\frac{\pi^2}{6}+\Delta\right) + \eta O_{K,M}\left(\frac{(\log D)^3}{D^3}\right).\end{aligned} \tag{59}$$

The half-line integral is $\langle A_1\rangle_\delta/2$, by the symmetry of δ. For a fixed class 1, either member of the top pair contributes to contraction and source, giving $2(K-1)$ ordered roles. Their half-line kernel integrals are $1/2$ and $\mathcal{B}(\Delta)/2$, respectively. Hence

$$\langle p_1(1-p_1)\rangle = \frac{(K-1)c_K}{D} + O_{K,M}\left(\frac{(\log D)^2}{D^2}\right), \tag{60}$$

$$\langle g_1^2\rangle = \frac{(K-1)c_K}{D}\mathcal{B}(\Delta) + O_{K,M}\left(\frac{(\log D)^2}{D^2}\right). \tag{61}$$

Inserting these into the exact source–contraction equation Eq. (33) yields

$$\begin{aligned}\dot\Delta = \frac{Kc_K}{D}\left[\frac{\eta^2}{b}\mathcal{B}(\Delta) - 2\eta\Delta\right]\\ + O_{K,M}\left(\frac{\eta\Delta(\log D)^2}{D^2} + \frac{\eta^2(\log D)^2}{bD^2}\right).\end{aligned} \tag{62}$$

The factors $\eta\Delta$ and η^2/b in this remainder are essential under annealing. Estimating $\dot Q_{\text{eff}}$ and $2D\dot D$ separately and then subtracting would obscure the exact factor Δ in the deterministic part. For

completeness, adding $2D\dot{D}$ back to $\dot{\Delta}$ gives the centered-norm expansion,

$$\dot{Q}_{\text{eff}} = \frac{Kc_K}{D}\left[\eta\left(\frac{\pi^2}{6}-\Delta\right)+\frac{\eta^2}{b}\mathcal{B}(\Delta)\right] + \eta O_{K,M}\left(\frac{(\log D)^3}{D^2}\right)+\frac{\eta^2}{b}O_{K,M}\left(\frac{(\log D)^2}{D^2}\right). \tag{63}$$

Its drift coefficient also follows from the A_2 integral.

For the test loss, each ordered pair contributes the half-line loss kernel, whose integral is $\langle \mathcal{L}_0\rangle_\delta/2$. Thus

$$\mathcal{L}_g = \langle -\log p_y\rangle = \frac{K(K-1)c_K}{2D}\left(\frac{\pi^2}{6}+\Delta\right)+O_{K,M}\left(\frac{(\log D)^3}{D^2}\right). \tag{64}$$

These estimates prove the flow and loss statements of Theorem 3, with explicit remainders. In particular, they remain usable for vanishing learning rates and vanishing Δ.

A.3.4 Classification error

Misclassification requires a separate small-noise estimate, since its layer has width $\sqrt{\Delta}/D$, rather than $1/D$. For $D>0$, division of all logits by D leaves their ordering unchanged. Writing $\sigma_\perp = \sqrt{\Delta}/D$,

$$\epsilon_g = \Pr\left[\arg\max_a (u_a+\sigma_\perp z_a) \neq \arg\max_a u_a\right].$$

The event that the teacher runner-up overtakes the teacher winner is contained in the error event. Its probability is

$$\int_0^\infty f_G(g)\Phi\left(-\frac{g}{\sqrt{2}\sigma_\perp}\right)dg = \sigma_\perp f_G(0)\int_0^\infty \Phi(-x/\sqrt{2})\,dx + O_K(\sigma_\perp^2) = \frac{K(K-1)c_K}{\sqrt{\pi}}\,\sigma_\perp + O_K(\sigma_\perp^2).$$

The remainder follows from the bounded derivative of f_G and the finite first moment of the Gaussian tail kernel. If an error occurs without this runner-up crossing, some third class must overtake the winner. This requires $G+G_2 \le \sigma_\perp \max_{a,c}|z_a - z_c|$. The joint gap density is bounded, so the probability is at most $C_K\sigma_\perp^2\langle \max_{a,c}|z_a-z_c|^2\rangle = O_K(\sigma_\perp^2)$. For $K=2$ there is no such correction. We obtain

$$\epsilon_g = \Gamma_K\frac{\sqrt{\Delta}}{D}+O_K\left(\frac{\Delta}{D^2}\right), \qquad \Gamma_K = \frac{K(K-1)c_K}{\sqrt{\pi}}. \tag{65}$$

Equivalently, for an unordered boundary the local disagreement interval has length $|\delta|$, and $\langle|\delta|\rangle = 2\sqrt{\Delta/\pi}$; summing $K(K-1)/2$ boundaries gives the same coefficient. Both the leading term and the remainder vanish at $\Delta=0$. This completes the proof of Theorem 3.

A.4 Corollary 4: fixed learning rate

For constant $\eta>0$, the limiting noise floor solves

$$2\Delta_* = \frac{\eta}{b}\mathcal{B}(\Delta_*). \tag{66}$$

To establish uniqueness and attraction, note that

$$\mathcal{B}(0) = 2\log 2 - 1 > 0, \qquad \mathcal{B}'(\Delta) = \frac{1}{2}\int Dz\,\operatorname{sech}^2\left(\sqrt{\frac{\Delta}{2}}z\right). \tag{67}$$

The derivative is positive and strictly decreasing, and $\mathcal{B}(\Delta) = O(1+\sqrt{\Delta})$. Thus $(\eta/b)\mathcal{B}(\Delta) - 2\Delta$ is strictly concave, starts positive, and tends to $-\infty$. It has exactly one positive zero, with positive drift below and negative drift above it.

This sign information also controls the full trajectory, not just the reduced equation. By Lemma 2, Δ is bounded and $D \to \infty$. Eq. (59) gives $\partial(D^3)/\partial\alpha = O(1)$, so $D^3 = O(1+\alpha)$ and the slow time

$$\tau(\alpha) = Kc_K\eta \int_{\alpha_1}^{\alpha} \frac{d\alpha'}{D(\alpha')}$$

diverges, with α_1 chosen after D becomes positive. In this time,

$$\frac{d\Delta}{d\tau} = \frac{\eta}{b}\mathcal{B}(\Delta) - 2\Delta + O_{K,M}\left((\Delta + \eta/b)\frac{(\log D)^2}{D}\right). \tag{68}$$

The perturbation vanishes uniformly on the bounded Δ interval. Outside any neighborhood of Δ_* it eventually cannot reverse the drift's sign. The trajectory therefore enters and remains in every such neighborhood, proving $\Delta \to \Delta_*$.

Multiplying Eq. (59) by $3D^2$ gives

$$\frac{\partial(D^3)}{\partial\alpha} = \frac{3Kc_K\eta}{2}\left(\frac{\pi^2}{6} + \Delta(\alpha)\right) + \eta O_{K,M}\left(\frac{(\log D)^3}{D}\right).$$

The error tends to zero. Integration and $\Delta \to \Delta_*$ imply

$$D^3(\alpha) \sim \frac{3Kc_K\eta}{2}\left(\frac{\pi^2}{6} + \Delta_*\right)\alpha. \tag{69}$$

Consequently,

$$D \propto \alpha^{1/3}, \qquad Q_{\text{eff}} = D^2 + \Delta \propto \alpha^{2/3}, \qquad \epsilon_g \sim \frac{\Gamma_K\sqrt{\Delta_*}}{D} \propto \alpha^{-1/3}, \tag{70}$$

and Eq. (64) gives $\mathcal{L}_g \propto \alpha^{-1/3}$. The relative loss remainder is $O((\log D)^3/D)$ and the relative error remainder is $O(\sqrt{\Delta}/D)$, so neither changes these laws. This proves Corollary 4.

The small-rate expansion is

$$\mathcal{B}(\Delta) = 2\log 2 - 1 + \frac{\Delta}{2} + O(\Delta^2), \qquad \Delta_* \sim \kappa\frac{\eta}{b} \quad \text{as } \eta/b \to 0, \tag{71}$$

where $\kappa = (2\log 2 - 1)/2$. For a fixed non-small η/b, the root equation Eq. (66), rather than this proportionality, determines the floor. For completeness, if the online source is removed (population gradient flow), Eq. (33) instead gives $\Delta \to 0$ and, for $\Delta > 0$, $(\log\Delta)' \sim -2Kc_K\eta/D$. Since $D \propto \alpha^{1/3}$ still holds, $\log\Delta = -c\alpha^{2/3} + o(\alpha^{2/3})$ for some $c > 0$. The error then decays faster than any power, whereas the loss retains one-third scaling.

A.5 Corollary 5: power-law schedules

A.5.1 Tracking for $0 < \gamma < 1$

Consider the shifted power law used in the experiments,

$$\eta(\alpha) = \eta_0\left(1 + \frac{\alpha}{\alpha_0}\right)^{-\gamma}, \qquad \eta_0, \alpha_0 > 0, \qquad 0 < \gamma < 1. \tag{72}$$

Its accumulated learning time is

$$H(\alpha) = \frac{\eta_0\alpha_0}{1-\gamma}\left[\left(1 + \frac{\alpha}{\alpha_0}\right)^{1-\gamma} - 1\right]. \tag{73}$$

First determine the scales without assuming tracking. Bounded Δ and Eq. (59) give $c_1\eta \le \partial(D^3)/\partial\alpha \le c_2\eta$ at late times, hence $D^3 \propto H$. Define a slow time by $d\tau/d\alpha = Kc_K\eta/D$. Then $\tau \propto H^{2/3} \to \infty$. For any fixed $\varepsilon > 0$, the drift in Eq. (62), divided by η/D, is uniformly negative on $\Delta \in [\varepsilon, M]$ once η is small. Its remainder tends to zero on that interval. The same trapping argument as above proves $\Delta \to 0$. Returning to the alignment equation and integrating yields

$$D^3(\alpha) \sim \frac{Kc_K\pi^2}{4}H(\alpha). \tag{74}$$

The relaxation rate near the small-Δ root is $\lambda \sim 2Kc_K\eta/D$. For this schedule,

$$\frac{|\eta'|/\eta}{\lambda} \propto \frac{D}{\eta(\alpha+\alpha_0)} \propto \alpha^{-2(1-\gamma)/3} \to 0. \tag{75}$$

To show that this time-scale comparison actually implies tracking, set $v = b\Delta/\eta$. The factorized remainder in Eq. (62) gives

$$\frac{dv}{d\tau} = \mathcal{B}(\eta v/b) - 2v - \frac{D\eta'}{Kc_K\eta^2}v + O_{K,M}\left((1+v)\frac{(\log D)^2}{D}\right). \tag{76}$$

The last two coefficients vanish by Eq. (75) and $D \to \infty$. Keep the factor $1+v$ until boundedness is established: since $\mathcal{B}(\eta v/b) = \mathcal{B}(\Delta) \le \mathcal{B}(M)$, the right side is negative for all sufficiently large v, uniformly at late times. Thus v eventually stays bounded. On this bounded interval the equation becomes $dv/d\tau = 2\log 2 - 1 - 2v + o(1)$, and the sign comparison gives

$$\Delta(\alpha) \sim \frac{\kappa}{b}\eta(\alpha), \qquad \kappa = \frac{2\log 2 - 1}{2}. \tag{77}$$

An unfactorized absolute error in $\dot{\Delta}$ would not justify this step as $\eta \to 0$; Eq. (76) shows why the retained remainders do.

Combining Eqs. (65), (74) and (77) gives

$$\epsilon_g(\alpha) \sim A_{K,b}\frac{\sqrt{\eta(\alpha)}}{H(\alpha)^{1/3}}, \qquad A_{K,b} = \Gamma_K\sqrt{\frac{\kappa}{b}}\left(\frac{4}{Kc_K\pi^2}\right)^{1/3}, \tag{78}$$

or, with the shifted power law inserted,

$$\epsilon_g(\alpha) \sim A_{K,b}(1-\gamma)^{1/3}\eta_0^{1/6}\alpha_0^{\gamma/6}\alpha^{-(2+\gamma)/6}. \tag{79}$$

The relative classification-error remainder is $O(\sqrt{\Delta}/D) = o(1)$. Meanwhile,

$$\mathcal{L}_g(\alpha) \sim \frac{K(K-1)c_K\pi^2}{12D(\alpha)} \propto H(\alpha)^{-1/3} \propto \alpha^{-(1-\gamma)/3}, \tag{80}$$

with relative remainder $O((\log D)^3/D) = o(1)$. This proves Corollary 5. The endpoint $\gamma = 0$ has the one-third exponent by Corollary 4, but its floor must be obtained from Eq. (66); the small-rate prefactor in Eq. (78) does not apply to arbitrary constant η.

A.5.2 Borderline schedule: Proposition 6

For $\gamma = 1$, write $\eta = A/(\alpha+\alpha_0)$ with $A = \eta_0\alpha_0$. The preceding argument up to Eq. (74) uses only $H \to \infty$ and $\eta \to 0$, not tracking. It therefore still gives $\Delta \to 0$ and

$$D^3(\alpha) \sim \frac{Kc_K\pi^2 A}{4}\log\alpha, \qquad \lambda(\alpha)\alpha \propto (\log\alpha)^{-1/3} \to 0. \tag{81}$$

The variance now relaxes more slowly than the learning rate changes. To identify its actual behavior, set $t = \log(\alpha+\alpha_0)$. The exact equation Eq. (33) becomes

$$\begin{aligned} \frac{d\Delta}{dt} &= -a(t)\Delta + q(t), \qquad a(t) = \frac{2KA}{K-1}\langle p_1(1-p_1)\rangle, \\ 0 \le q(t) &= \frac{KA^2e^{-t}}{b(K-1)}\langle g_1^2\rangle \le Ce^{-t}. \end{aligned} \tag{82}$$

By Eqs. (60) and (81),

$$\int_{t_0}^{t} a(s)\,ds = b_\Delta t^{2/3} + o(t^{2/3}), \qquad b_\Delta = \frac{3Kc_KA}{(Kc_K\pi^2A/4)^{1/3}} > 0. \tag{83}$$

Variation of constants gives $\Delta(t) = e^{-\int_{t_0}^{t} a}\left[\Delta(t_0) + \int_{t_0}^{t} e^{\int_{t_0}^{s} a}q(s)\,ds\right]$. The bracket converges to a finite positive limit: its integrand is at most $\exp(Cs^{2/3} - s)$, and finite-batch SGD generates positive variance even from $\Delta(0) = 0$. Consequently,

$$\log\Delta(\alpha) = -b_\Delta(\log\alpha)^{2/3} + o\big((\log\alpha)^{2/3}\big), \qquad \frac{\Delta(\alpha)}{\eta(\alpha)} \to \infty. \tag{84}$$

This proves Proposition 6. The remainder controls the logarithmic asymptotic, not a constant-prefactor equivalent for Δ. Eq. (65) then shows that the error decays more slowly than α^{-p} for every $p > 0$, whereas the loss decays as $(\log \alpha)^{-1/3}$. Thus $1/2$ is approached only as the limiting adiabatic exponent for $\gamma \uparrow 1$, not at $\gamma = 1$ itself.

For $\gamma > 1$, $H(\infty) < \infty$. The exact bound $\dot{D} \le \eta \langle \max_a u_a - \min_a u_a \rangle / (K-1)$ makes D converge to a finite limit. Since $\dot{\Delta} \le C_K \eta^2 / b$, Δ remains bounded. The source and contraction in Eq. (33) are then integrable, so Δ also converges. The total contraction is finite, and hence cannot fully remove the positive variance generated by finite-batch SGD. Both loss and error consequently retain nonzero limiting values in this case.

B BINARY WARMUP: SMOOTH STUDENT FOR A HARD TEACHER

This appendix records the simpler binary mechanism used as a warmup. It is not needed for the K-class proof, but it shows that the same qualitative ingredients–a diverging norm, a shrinking angle, and an online-noise floor–also appear outside the softmax model.

Let $T, J \in \mathbb{R}^N$ with $T \cdot T/N = 1$ and $\xi \sim \mathcal{N}(0, I_N)$, and define

$$u = \frac{T \cdot \xi}{\sqrt{N}}, \qquad t = \frac{J \cdot \xi}{\sqrt{N}}. \tag{85}$$

The teacher label is $\tau(u) = \mathrm{sgn}(u)$ and the student output is

$$g(t) = \mathrm{erf}\left(\frac{t}{\sqrt{2}}\right). \tag{86}$$

For squared loss $\mathcal{L} = (\tau(u) - g(t))^2/2$, online gradient descent gives

$$J^{\mu+1} = J^\mu + \frac{\eta}{\sqrt{N}} [\tau - g(t)] g'(t) \xi^\mu. \tag{87}$$

With

$$Q = \frac{J \cdot J}{N}, \qquad \rho = \frac{J \cdot T}{N}, \qquad R = \frac{\rho}{\sqrt{Q}}, \tag{88}$$

the fields $(u, t/\sqrt{Q})$ are standard correlated Gaussians with correlation R. The thermodynamic-limit flow is

$$\frac{d\rho}{d\alpha} = \frac{2\eta}{\pi(Q+1)} \left[\sqrt{Q - \rho^2 + 1} - \frac{\rho}{\sqrt{2Q+1}} \right], \tag{89}$$

$$\frac{dQ}{d\alpha} = \frac{4\eta}{\pi(Q+1)} \left[\frac{\rho}{\sqrt{Q-\rho^2+1}} - \frac{Q}{\sqrt{2Q+1}} \right] + \frac{2\eta^2}{\pi^2 \sqrt{2Q+1}} \left[\pi + 2 \arcsin\left(\frac{Q}{3Q+1}\right) - 4 \arcsin\left(\frac{\rho}{\sqrt{(3Q+1)(2(Q-\rho^2)+1)}}\right) \right]. \tag{90}$$

The explicit η^2 term is the variance of the online update. Introducing $r = 1 - R$ gives

$$\frac{dr}{d\alpha} = \frac{1-r}{2Q} \frac{dQ}{d\alpha} - \frac{1}{\sqrt{Q}} \frac{d\rho}{d\alpha}. \tag{91}$$

For $Q \gg 1$, $r \ll 1$, and $s = Qr = O(1)$, the large-Q expansion has the form

$$\frac{dQ}{d\alpha} = c(s, \eta) Q^{-1/2} + O(Q^{-3/2}), \tag{92}$$

$$\frac{dr}{d\alpha} = r_3(s, \eta) Q^{-3/2} + O(Q^{-5/2}), \tag{93}$$

Table 1: Comparison with the aligned-student theory of Liu et al. (2026).

	Liu et al. (2026)	this work
targets	peaked soft distributions	one-hot labels (zero-temperature limit)
dynamics	gradient flow, aligned-student ansatz	online SGD, rotational dynamics in (D, Δ)
one-third law for	test cross-entropy loss (calibration)	classification error *and* test loss
classification error	zero identically	$\epsilon_g \sim \Gamma_K \sqrt{\Delta}/D \propto \alpha^{-1/3}$, exact prefactor
gradient noise	left open	noise floor $2\Delta_* = (\eta/b)\,\mathcal{B}(\Delta_*)$
learning-rate schedules	loss-level reparametrization $\tau = \int \eta$	error exponent $1/3 \to 1/2$ while loss slows

where

$$c(s,\eta) = \frac{4\eta}{\pi}\left(\frac{1}{\sqrt{1+2s}} - \frac{1}{\sqrt{2}}\right) + \frac{\sqrt{2}}{\pi^2}\eta^2 J(s), \tag{94}$$

$$r_3(s,\eta) = -\frac{4\eta s}{\pi\sqrt{1+2s}} + \frac{\eta^2}{\pi^2\sqrt{2}}J(s), \tag{95}$$

$$J(s) = \pi + 2\arcsin\left(\frac{1}{3}\right) - 4\arcsin\left(\frac{1}{\sqrt{3(1+4s)}}\right). \tag{96}$$

A consistent power law requires $r_3(s_*, \eta) = 0$, which fixes $s_* = \lim_{\alpha\to\infty} Qr$. Then

$$Q(\alpha) \sim \left[\frac{3}{2}c(s_*,\eta)\alpha\right]^{2/3}, \qquad r(\alpha) \sim \frac{s_*}{Q(\alpha)} \propto \alpha^{-2/3}. \tag{97}$$

The binary classification error is

$$\epsilon_g = \frac{1}{\pi}\arccos R = \frac{1}{\pi}\arccos(1-r) \sim \frac{\sqrt{2r}}{\pi}, \tag{98}$$

and hence $\epsilon_g \propto \alpha^{-1/3}$. If the learning rate decays adiabatically, the same reduced structure gives a noise-floor relation $r \sim k\eta/Q$ and leads to $\epsilon_g \propto \alpha^{-(2+\gamma)/6}$ when $\eta \propto \alpha^{-\gamma}$ with $\gamma < 1$. This binary calculation is a useful sanity check and demonstrates similar behavior beyond softmax and cross-entropy loss, but the main paper's results are the multiclass softmax boundary-layer formulas derived above.

C Numerical protocols and additional robustness checks

This appendix summarizes the numerical settings underlying the main figures and records additional robustness checks. The reported observables are the test misclassification rate ϵ_g, the centered margin $D = R - S$, the residual variance $\Delta = Q_{\text{eff}} - D^2$, as well as the test loss $\mathcal{L}_g$. The total computational cost for all simulations reported in the manuscript was on the order of 10^3 CPU-hours.

For the two main teacher-student validation figures, Figs. 2 and 3, we used the online $K = 3$ softmax teacher-student model at dimension $N = 500$. The teacher vectors were chosen orthonormal with $T_a \cdot T_b/N = \delta_{ab}$, and the student weights were initialized with independent entries $J_{ai} \sim \mathcal{N}(0,1)$. Each SGD update used a fresh Gaussian input example, and time is reported in macroscopic units $\alpha = \mu/N$. The curves were generated from six independent random seeds. The plotted envelopes show the corresponding seed-to-seed fluctuations around a representative trajectory. For the fixed-learning-rate comparison in Fig. 2, the theoretical prediction uses the full asymptotic prefactor: the boundary density c_K was evaluated numerically, and for each fixed learning rate the residual variance floor Δ_* was obtained by numerically solving $2\Delta_* = \eta\,\mathcal{B}(\Delta_*)$, with $\mathcal{B}$ defined in Eq. (13). For the schedule experiment in Fig. 3, the learning-rate family was $\eta(\alpha) = \eta_0(1 + \alpha/\alpha_0)^{-\gamma}$, with $\eta_0 = 2.0$, $\alpha_0 = 10.0$, and the plotted exponents $\gamma \in \{0, 0.5, 1\}$. The generalization error ϵ_g was estimated by Monte Carlo evaluation on $M = 10^5$ fresh test examples. Figures showing numerical values for the order parameters D or Δ use the centered student and teacher vectors of class one to determine them.

C.1 NUMBER OF CLASSES: K-DEPENDENCE

The main-text simulations focus on $K = 3$, while the asymptotic theory is stated for fixed but arbitrary K. We therefore include an explicit K-sweep as a check that the predicted fixed-K boundary-layer mechanism is not unique to the three-class case. For fixed learning rate and batch size $b = 1$, the theory predicts that the exponents are independent of K,

$$D \propto \alpha^{1/3}\,, \qquad \epsilon_g \propto \alpha^{-1/3}\,, \tag{99}$$

while the prefactors depend on K through the boundary density c_K and the classification-error constant

$$\Gamma_K = \frac{K(K-1)c_K}{\sqrt{\pi}}. \tag{100}$$

More explicitly,

$$D^3(\alpha) \sim \frac{3Kc_K}{2}\,\eta\left(\frac{\pi^2}{6} + \Delta_*\right)\alpha, \qquad \epsilon_g \sim \Gamma_K \frac{\sqrt{\Delta_*}}{D}, \tag{101}$$

where the fixed-learning-rate noise floor Δ_* is determined by $2\Delta_* = \eta\mathcal{B}(\Delta_*)$ and is independent of K at this leading order.

Fig. 6 shows simulations for $K = 5, 20, 50, 100$ at $N = 200$, with fixed learning rates $\eta = 1.0, 0.1, 0.01$ and $\gamma = 0$. The fluctuation envelopes are computed over six seeds. Across all tested values of K, the late-time trajectories remain in excellent agreement with the predicted asymptotic structure. The centered overlap follows the $D \propto \alpha^{1/3}$ law, the residual variance approaches the learning-rate-dependent floor, and the generalization error follows the resulting $\epsilon_g \propto \sqrt{\Delta}/D$ decay. Increasing K induces a later crossover into the asymptotic regime. This is expected since for larger K, the top teacher fields are closer and the pairwise boundary-layer approximation becomes valid only once the centered margin given by D is large compared to the typical extreme-value scale, $D \gg \sqrt{2\log K}$. Thus larger K produces longer transients before the fixed-K late-time theory becomes visible.

C.2 BATCH SIZE: b-DEPENDENCE

At fixed learning rate, increasing the batch size suppresses the online-noise term in Eq. (9), while leaving the drift of D at a given state unchanged. The residual variance floor satisfies

$$2\Delta_* = \frac{\eta}{b}\mathcal{B}(\Delta_*)\,, \qquad \Delta_* \sim \frac{2\log 2 - 1}{2}\frac{\eta}{b} \quad \text{for } \eta/b \to 0\,. \tag{102}$$

Thus, for fixed finite b, the one-third exponents remain unchanged. At small η/b, the error prefactor decreases approximately as $b^{-1/2}$, whereas D and the test loss depend only weakly on b through Δ_*.

Fig. 7 compares $b \in \{1, 16, 128\}$ at $N = 200$, $K = 3$, and $\eta = 0.1$, using the orthonormal teachers and Gaussian student initialization described above, with four seeds $0, 1, 2, 3$. Each update averages the gradients of b fresh Gaussian examples as in Eq. (4). Time $\alpha = \mu/N$ counts minibatch updates, so the number of training examples is $bN\alpha$. We train to $\alpha = 10^5$ and evaluate at logarithmically spaced checkpoints on 10^4 independent Gaussian test inputs per seed, held fixed across checkpoints and batch sizes. The theoretical curves use the full prefactors of Corollary 4 and Theorem 3, with Δ_* obtained by numerical root finding and $\mathcal{B}$ evaluated using 5×10^5 standard Gaussian samples. Increasing b lowers the observed residual variance and error, while the overlap and loss curves nearly coincide, in agreement with the small-η/b prediction.

C.3 BOUNDARY-LAYER LOCALIZATION

To localize the examples responsible for error and alignment growth, let $G = u_y - \max_{a\neq y} u_a$ be the top-two teacher-logit gap and $\hat{y} = \arg\max_a h_a$ the student prediction. For the per-example cross-entropy $\ell = -\log p_y$, the teacher-aligned negative loss gradient is

$$-\frac{\partial \ell}{\partial D} = u_y - \sum_a p_a u_a \geq 0\,, \qquad \dot{D} = -\frac{\eta}{K-1}\left\langle \frac{\partial \ell}{\partial D}\right\rangle\,. \tag{103}$$

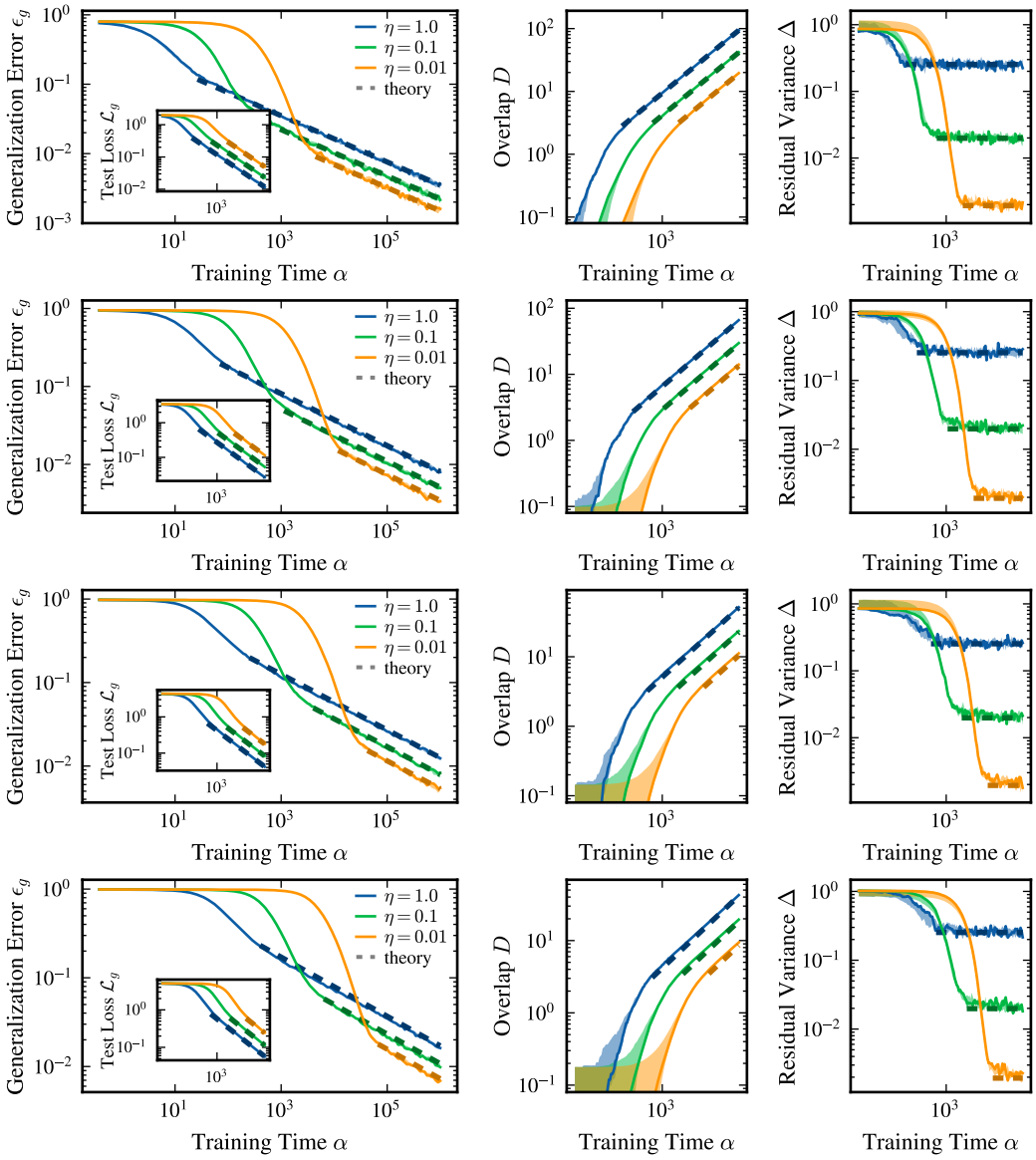

Figure 6: Dependence on the number of classes. Fixed-learning-rate simulations for $K = 5, 20, 50, 100$ (increasing from top to bottom) at $N = 200$, with $\eta = 1.0, 0.1, 0.01$, $\gamma = 0$, and envelopes over six seeds. For each value of K, the panels show the generalization error ϵ_g, the centered student-teacher overlap D, and the residual variance Δ. The late-time behavior agrees with the fixed-K boundary-layer prediction: $D \propto \alpha^{1/3}$, Δ approaches the learning-rate noise floor, and $\epsilon_g \propto \alpha^{-1/3}$. Increasing K delays the onset of the asymptotic regime.

The second identity follows from Lemma 1. This diagnostic measures the per-example contribution to alignment growth, not the loss itself or the full gradient norm.

We use the same isotropic teacher–student setup with $N = 200$, $K = 3$, $\eta = 0.5$, $b = 1$, and one trajectory, saved at 20 approximately logarithmically spaced checkpoints from $\alpha = 10^2$ to 10^5. At each checkpoint, 2×10^7 fresh joint Gaussian teacher–student fields are sampled. The conditional profiles use 400 uniform gap bins on $[0, 0.6]$ for error and $[0, 2.7]$ for the gradient. At each checkpoint, we define the median teacher gaps for error and the alignment-driving loss gradient,

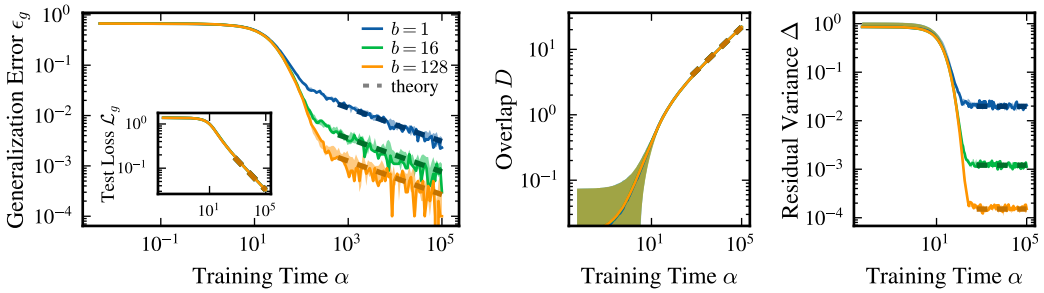

Figure 7: Batch-size dependence at fixed learning rate. Left: generalization error, with test loss in the inset. Middle: centered overlap D. Right: residual variance Δ. Solid curves show a single seed and shaded envelopes show the minimum–maximum over four seeds. Dashed curves are the asymptotic predictions. Larger batches reduce the noise floor and error prefactor without changing the late-time one-third exponents.

respectively, by

$$\Pr\Big(G \le G_{\epsilon_g}^{\text{median}}(D) \mid \hat{y} \neq y\Big) = \frac{1}{2}, \qquad \frac{\Big\langle -\frac{\partial \ell}{\partial D} \cdot \mathbf{1}\Big\{G \le G_{-\frac{\partial \ell}{\partial D}}^{\text{median}}(D)\Big\}\Big\rangle}{\langle -\frac{\partial \ell}{\partial D}\rangle} = \frac{1}{2}. \tag{104}$$

Thus, the corresponding gap thresholds contain half of all classification errors and half of the total alignment-driving gradient weight.

Fig. 8 shows the conditional error $\Pr(\hat{y} \neq y \mid G)$ and gradient $\big\langle -\frac{\partial \ell}{\partial D} \mid G\big\rangle$, together with the gaps below which half of all errors or half of the total alignment-driving weight is accumulated. Both concentrate at smaller teacher gaps as D grows. The median gaps are consistent with D^{-1} scaling: at fixed learning rate, Δ stays near a nonzero floor, so both the $O(D^{-1})$ learning-signal width and the narrower $O(\sqrt{\Delta}/D)$ error width shrink inversely with D.

C.4 Correlated Gaussian inputs

For the correlated-Gaussian robustness check at $N = 200$, the input covariance is diagonal with eigenvalues proportional to $i^{-\beta}$ and normalized so that the largest variance is one. Concretely, inputs were generated as $\xi_i = \sqrt{\lambda_i}\, x_i$, with $x_i \sim \mathcal{N}(0,1)$ independently and

$$\lambda_i = \left(\frac{a}{a+i}\right)^{\beta}, \qquad i = 0, \dots, N-1\,, \tag{105}$$

with $a = 10$ in the experiments, so that the largest variance has $\lambda_0 = 1$ and increasing β produces a stronger power-law anisotropy. The teacher vectors were sampled from independent standard-normal entries and then orthogonalized by Gram-Schmidt with respect to the normalized Euclidean inner product, followed by the normalization $T_a \cdot T_b/N = \delta_{ab}$. Thus the covariance spectrum is structured, but the teacher directions are otherwise random with respect to the covariance eigenbasis. The student weights were initialized with independent $J_{ai} \sim \mathcal{N}(0,1)$, as in the isotropic simulations. The labels are still generated by the teacher. Increasing β creates a structured spectrum and lengthens the transient. The observed late-time behavior remains consistent with the same boundary-layer exponent.

C.5 Whitened feature experiments

The feature experiments use pretrained vision-transformer features as a non-Gaussian test bed (Dosovitskiy et al., 2021). Whitening removes the leading covariance structure, so the experiment probes whether non-Gaussian feature statistics destroy the boundary-layer pattern. Teacher-generated labels are used in the main text because they avoid an early performance floor and allow the late-time regime to remain visible. Real-label runs can reach a floor too early for a clean asymptotic fit, but they are useful as a practical diagnostic.

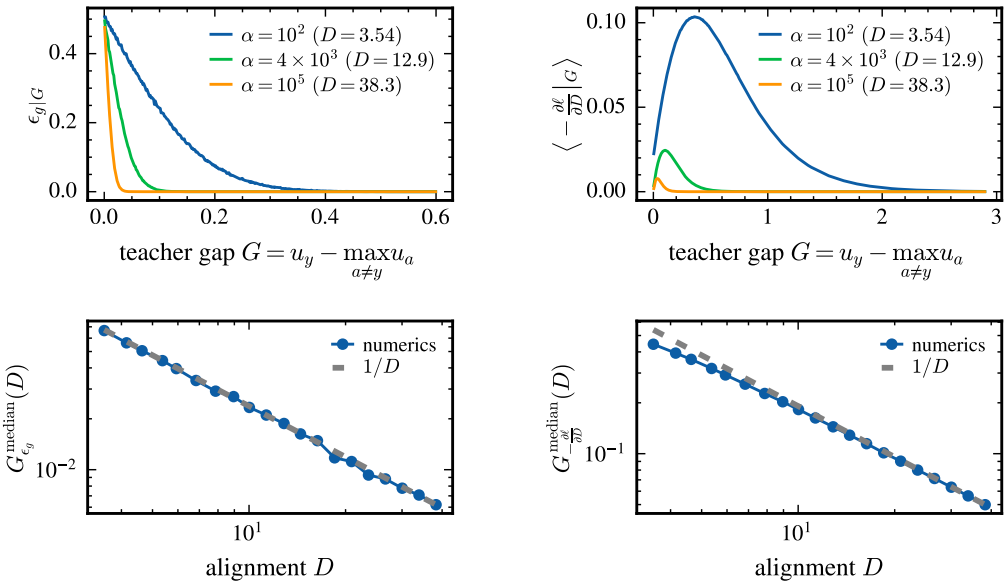

Figure 8: Direct boundary-layer diagnostics. Top: error probability (left) and mean alignment-driving loss gradient (right), conditioned on the teacher gap, at $\alpha = 10^2, 4 \times 10^3, 10^5$. Bottom: corresponding median teacher gaps as functions of D. The dashed $1/D$ guides are normalized at the final checkpoint. They test the scaling rather than a predicted prefactor. Errors and alignment growth become increasingly localized near teacher decision boundaries.

In these experiments we used CIFAR-5M images (Nakkiran et al., 2021) and extracted fixed features with the pretrained ViT-B/16 model `google/vit-base-patch16-224-in21k` (Dosovitskiy et al., 2021). Images were passed through the ViT feature extractor and the pooled `[CLS]` representation was used as the input feature vector. For comparability with the $K = 3$ teacher-student simulations, we restricted the dataset to three CIFAR classes: airplane, automobile, and horse. This class triple was chosen because, among the tested three-class subsets, it gave one of the lowest apparent irreducible errors when a linear classifier was trained on the ViT features, making it a favorable case for observing a long late-time regime before a real-label floor is reached. Before training the linear readout, we subtracted the empirical feature mean and applied ZCA whitening: if X denotes the matrix of centered ViT features with empirical covariance $C = X^\top X/(n-1)$, we used the eigendecomposition $C = U\Lambda U^\top$ and transformed the features by $X \mapsto XU(\Lambda + \varepsilon I)^{-1/2}U^\top$, with a small numerical regularizer $\varepsilon = 10^{-5}$. The resulting whitened features have approximately identity empirical covariance, but their distribution remains strongly non-Gaussian. Teacher labels are the argmax predictions of a frozen bias-free linear classifier trained for one epoch on these features and real labels using cross-entropy SGD with learning rate 5×10^{-4} and batch size 10. The figures show envelopes over four seeds.

The readout architecture differs slightly from the analytically normalized teacher-student model. In the ViT-feature experiments we trained a linear layer, $f(x) = Wx$, without the explicit $1/\sqrt{N}$ factor in the logit definition. Consequently the runs used the default smaller PyTorch linear-layer initialization and correspondingly smaller learning rates than in the Gaussian teacher-student simulations. Specifically, for input dimension N, `torch.nn.Linear` initializes weights from the uniform distribution $\mathcal{U}(-N^{-1/2}, N^{-1/2})$. The readout was trained with cross-entropy loss using online SGD without momentum, with batch size one. Time was again reported in normalized units as the number of SGD steps divided by the feature dimension.

The numerical evidence should be interpreted conservatively. The asymptotic theory is for isotropic Gaussian inputs in the thermodynamic limit. The correlated-input and whitened-feature experiments show that the mechanism can remain visible under controlled departures, but they do not constitute a theorem for arbitrary feature distributions or real-data classification.

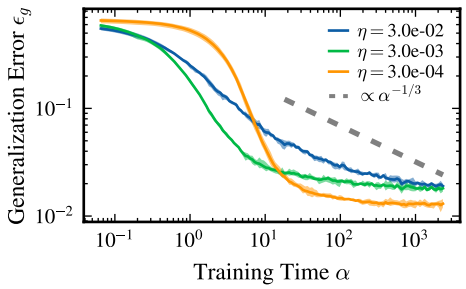

Figure 9: Whitened pretrained-feature experiment with real labels. This run is included as a qualitative practical comparison to the teacher-label experiment in Fig. 4; the real-label performance floor is reached earlier, making it less suitable as a clean asymptotic test.

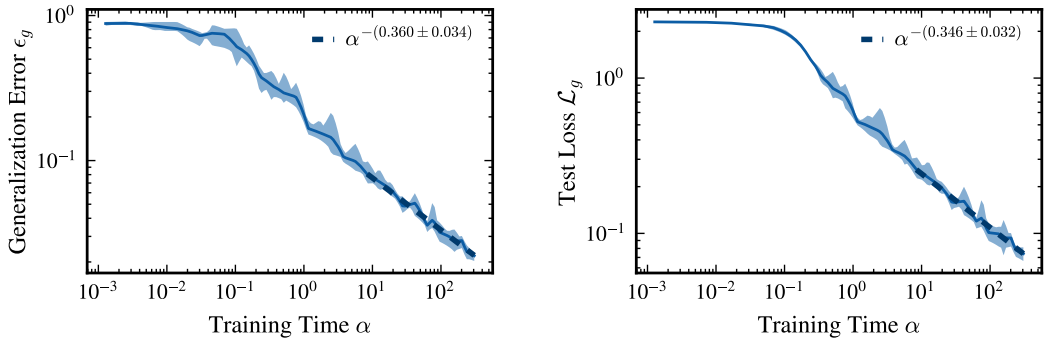

Figure 10: SGD on MNIST. The generalization error (left) and test loss (right) are compatible with one-third scaling. Curves are averaged over six seeds; the shaded envelopes span the seed trajectories. Dashed lines are fits over $\alpha \in [8.4, 306]$. The exponent uncertainties shown in the panels are twice the covariance-based standard errors of the fits.

C.6 MNIST and Pythia experiments

SGD on MNIST. We trained a 784–512–512–10 ReLU MLP with cross-entropy and SGD at learning rate 0.01, batch size one, and no momentum or weight decay for four epochs. The first epoch used the original images. The remaining three used random rotations up to 10° and translations up to one pixel, reducing exact example reuse and bringing this finite-dataset setup closer to online SGD. Liu et al. (2026) demonstrated a one-third loss law in an MNIST training run with full-batch Adam. We show that it persists for SGD. We report the mean and envelope over six seeds, with α equal to the number of updates divided by 784. Fits of the across-seed means over $\alpha \in [8.4, 306]$ show very good agreement of test loss with one-third scaling (Fig. 10). Crucially, we also analyze test accuracy, reported through its complement, generalization error. Accuracy is the more direct performance metric for image classification, and the theory of Liu et al. makes no prediction for it. The observed error decay over this window is compatible with the one-third scaling our theory predicts.

Pythia next-token prediction. Pythia is a controlled suite of autoregressive transformer language models ranging from 70M to 12B parameters, with 154 checkpoints per model released across training (Biderman et al., 2023). Following the checkpoint evaluation pipeline of Liu et al. (2026), we reproduced their Pythia loss curves and exponent estimates, confirming their results. However, changing either edge of the fit window by one checkpoint can change the exponent significantly, and considerably more than the covariance-based statistical fit error; we therefore report both contributions to the uncertainty below. Our contribution lies in extending this checkpoint analysis to token top-1 error. Evaluation uses the first 4,800 documents of a pinned FineWeb stream (Penedo et al., 2024), truncated to 1,024 tokens, in batches of 12.

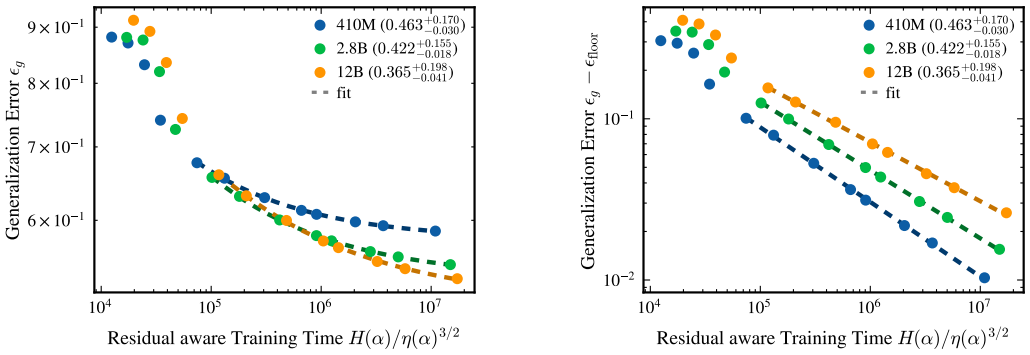

Figure 11: Pythia token top-1 error versus residual-aware training time. Left: fits to a power law plus a floor. Right: the same curves after subtracting the fitted floor. The central fit window spans checkpoints 2,000–79,000. The uncertainties reported in the legends are asymmetric because, on each side, they sum the covariance-based statistical fit error with the largest exponent change obtained by shifting either fit-window edge by one checkpoint.

Token top-1 error should be interpreted cautiously, because a text prefix can admit several plausible next tokens. It nevertheless probes a central distinction in LLM evaluation because raw probability magnitude and relative ordering of logits need not evolve together. Accuracy- and logit-rank-based scaling are established LLM questions: downstream top-1 error has been related to scaling laws (Gadre et al., 2025), while multiple-choice accuracy depends on margins against specific competitors and is not determined by correct-token loss alone (Schaeffer et al., 2025). Furthermore, decomposing cross-entropy loss reveals that a component based only on the rank of the correct logit compared to the largest one can scale more robustly than total cross-entropy (Yan et al., 2026). Thus, better loss need not imply better task performance, and top-1 error directly probes the relative-logit dynamics central to our theory. In the Zipf model of Section D, a student with fixed coverage recovers the boundary-layer $1/3$ exponent in its excess error once all represented sub-tasks have entered their late-time regime, despite their unequal frequencies. This motivates testing for the same exponent in Pythia, where extensive training at fixed model size could predominantly refine already represented patterns while limited coverage contributes an approximately constant error floor.

Pythia was trained with a learning-rate schedule, so raw training steps do not by themselves capture the varying update scale. Liu et al. (2026) analyze the loss against what we define as $H(\alpha) = \int_0^\alpha \eta(\alpha')\, d\alpha'$, as considered before for learning rate schedules in Section A.5. For the error plots we consider a residual-aware training time $T_{\text{res}}(\alpha) = H(\alpha)/\eta(\alpha)^{3/2}$. Under adiabatic tracking in our model, Eq. (78) reads $\epsilon_g \propto T_{\text{res}}^{-1/3}$: the factor $\eta^{-3/2}$ incorporates the schedule-induced reduction of the residual variance Δ. We fit $\epsilon_g = cT_{\text{res}}^{-\beta} + \epsilon_{\text{floor}}$, a power law plus a floor as in the loss analysis of Liu et al. (2026) and as motivated by the sub-task model of Section D, over Pythia checkpoints 2,000–79,000. For Pythia's adaptive optimizer, weight decay, and cosine schedule, this time coordinate is heuristic, not a consequence of Corollary 5 (Biderman et al., 2023).

The Pythia networks are deep transformers that differ substantially from a one-layer online teacher–student model, so exact exponent comparisons require cautious interpretation. The primary conclusion is that token top-1 error is well approximated by a power law over the selected fit window. The 12B result is compatible with our one-third prediction, while the two smaller transformers have slightly steeper central fits. The reason for the steeper fits is not established.

D Aggregation over Zipf-distributed sub-tasks: a minimal coverage-limited model

Real classification data can mix many sub-problems of very different frequency Michaud et al. (2023), and practical students represent only part of them. This appendix aggregates the fixed-learning-rate law of Corollary 4 over Zipf-distributed sub-tasks in a minimal routed-expert model.

Its status differs from that of Sections 3 to 5: conditioned on the routing, every expert follows the dynamics of Lemma 1 exactly, so the per-sub-task exponent is agrees with the $1/3$ exponent derived in the main manuscript. The aggregation below is a scaling-level calculation, not a theorem. It is assembled from standard ingredients of quanta and template models of scaling laws (Hutter, 2021; Michaud et al., 2023; Nam et al., 2024), which postulate the learning curve of each quantum; here that curve is the boundary-layer law of the main text. All exponents stated below refer to the regime in which every activated expert has reached its residual noise floor Δ_* (Corollary 4); before that, relaxation from above this floor can temporarily steepen an expert's error decay without changing the leading loss scaling (see the discussion following Corollary 4).

D.1 Model

Data. An infinite dictionary of templates $k = 1, 2, \dots$ is sampled with Zipf weights

$$p_k = \frac{k^{-\nu}}{\zeta(\nu)}, \qquad \nu > 1, \tag{106}$$

with normalization $\zeta(\nu)$. A data point is a pair $x = (k, \xi)$: a template index and a local Gaussian example $\xi \sim \mathcal{N}(0, I_N)$.

Task. Template k carries its own K-class sub-task with orthonormal teachers $T_k^{(1)}, \dots, T_k^{(K)} \in \mathbb{R}^N$, $T_k^{(a)} \cdot T_k^{(b)}/N = \delta_{ab}$, drawn independently across k; the label of (k, ξ) is $y = \arg\max_a T_k^{(a)} \cdot \xi/\sqrt{N}$. Nothing about T_k can be inferred from other templates.

Student. A hard-routed mixture of linear experts: the router reads k and selects expert weights $J_k^{(1)}, \dots, J_k^{(K)}$, which produce logits $t_a = J_k^{(a)} \cdot \xi/\sqrt{N}$, softmax probabilities, and the cross-entropy loss of the main text. Training is online SGD on the stream with fixed learning rate η and for simplicity a batch size of one; a sample (k, ξ, y) updates only the routed expert. Experts share no parameters, so, conditioned on the routing, expert k performs exactly the online dynamics of the main text, assumptions (A1)–(A7), in its own macroscopic time

$$\alpha_k(t) = \frac{n_k(t)}{N} \simeq \frac{p_k\, t}{N}\,. \tag{107}$$

Here, t counts stream samples and $n_k(t)$ the samples routed to k. Consequently its test error is the base-model error at that time, $\epsilon_k(t) = \epsilon_g(\alpha_k(t))$, and Theorem 3 and Corollary 4 apply expert by expert. Only the $\mathcal{O}(t^{1/\nu})$ experts of seen templates are ever updated.

Three facts about Zipf sampling. (a) After t draws, template k has appeared $n_k \simeq p_k t$ times. (b) Templates are seen only if $p_k t \gtrsim 1$, i.e. for ranks $k \lesssim (t/\zeta(\nu))^{1/\nu}$; the number of distinct templates grows as $t^{1/\nu} \ll t$ (Heaps' law). (c) The probability mass of unseen templates (missing mass) is $m(t) = \sum_{p_k t \lesssim 1} p_k \propto t^{-c_\nu}$ with

$$c_\nu := \frac{\nu - 1}{\nu} = 1 - \frac{1}{\nu} \in (0, 1)\,. \tag{108}$$

Activation threshold. Expert k enters the boundary-layer regime when $\alpha_k \approx \alpha_*$, the entry scale of Section A.2.3 (Eq. (41)), i.e. at stream time

$$t_*(k) = \frac{N\alpha_*}{p_k} = N\alpha_*\, \zeta(\nu)\, k^\nu, \qquad k_*(t) = \Big(\frac{t}{N\alpha_*\zeta(\nu)}\Big)^{1/\nu}, \tag{109}$$

where $k_*(t)$ is the rank of the activation frontier at time t. At scaling level we take $\epsilon_k(t) = \mathcal{O}(1)$ for $t \lesssim t_*(k)$, more precisely of order the chance error $\epsilon_0 = 1 - 1/K$, and $\epsilon_k(t) = \epsilon_g(p_k t/N) \propto (p_k t/N)^{-1/3}$ for $t \gg t_*(k)$.

D.2 Error decomposition and the one-pass law

Independence of the sub-tasks gives

$$\epsilon(t) = \sum_k p_k\, \epsilon_k(t) = \underbrace{\sum_{k > k_*(t)} p_k\, \epsilon_k(t)}_{\text{coverage}} + \underbrace{\sum_{k \le k_*(t)} p_k\, \epsilon_g\big(p_k t/N\big)}_{\text{refinement}}\,. \tag{110}$$

Coverage is a resource that only new data can buy; refinement is what training buys on templates already in hand. At the stated scaling level, the coverage term is of order $\epsilon_0 \sum_{k>k_*} p_k \propto k_*^{-(\nu-1)} \propto t^{-c_\nu}$. In the refinement term every activated expert contributes its own one-third law, and the sum $\sum_{k\le k_*} p_k^{2/3} \propto \sum_{k\le k_*} k^{-2\nu/3}$ converges if and only if $\nu > 3/2$. Hence, for one pass over the stream,

$$\epsilon(t) \;\propto\; \begin{cases} t^{-1/3}, & \nu > 3/2 \quad (c_\nu > 1/3), \\ t^{-1/3}\log t, & \nu = 3/2 \quad (c_\nu = 1/3), \\ t^{-c_\nu}, & 1 < \nu < 3/2 \quad (c_\nu < 1/3), \end{cases} \qquad \text{i.e.} \qquad \epsilon(t) \propto t^{-\min\{1/3,\, c_\nu\}} \tag{111}$$

up to the logarithm at the marginal point. For light tails ($\nu > 3/2$) the frequent templates dominate and the boundary-layer exponent is the visible law. For heavy tails ($\nu < 3/2$) the activation frontier and the missing mass contribute at the same order, and the $1/3$ is invisible in the total error even though every activated expert individually follows it. The masking in Eq. (111) is a property of the sum, not of its terms.

D.3 Coverage-limited students and class-restricted observables

Now let the student carry experts only for the $k_{\max}$ most frequent templates and predict the remaining ones at chance. Its error has a static floor set by coverage alone,

$$\epsilon_{\text{floor}}(k_{\max}) = \epsilon_0 \sum_{k>k_{\max}} p_k \propto k_{\max}^{-(\nu-1)}, \tag{112}$$

and the excess error above the floor is the class-restricted error $\epsilon_{\le k_{\max}}(t) = \sum_{k\le k_{\max}} p_k\, \epsilon_k(t)$, which crosses over at $t_*(k_{\max})$:

$$\epsilon_{\le k_{\max}}(t) \;\propto\; t^{-\min\{1/3,\, c_\nu\}}, \qquad t \ll t_*(k_{\max}), \tag{113}$$

$$\epsilon_{\le k_{\max}}(t) \;\sim\; \Gamma_K \sqrt{\Delta_*}\left[\tfrac{3Kc_K}{2}\,\eta\big(\tfrac{\pi^2}{6}+\Delta_*\big)\right]^{-1/3}\Big(\frac{t}{N}\Big)^{-1/3} \sum_{k\le k_{\max}} p_k^{2/3}, \quad t \gg t_*(k_{\max}), \tag{114}$$

where Eq. (114) is Corollary 4 summed over the represented experts, prefactor included, because all experts share η and hence the same noise floor Δ_*. Below $t_*(k_{\max})$ the cutoff is invisible: the sum is dominated by the frontier $k_*(t) < k_{\max}$ and by the not-yet-activated templates below the cutoff. Above it the frequency tail is gone and the bare boundary-layer exponent is the only time dependence left. For heavy tails ($\nu < 3/2$) both branches are of order $k_{\max}^{-(\nu-1)}$ at the crossover, comparable to the floor itself; for $\nu > 3/2$ there is nothing to unmask. Coverage limitation therefore does not remove the mechanism: it sets the last stage of the approach to the floor, and the total error of a coverage-limited student has the form "power law plus floor" used for the fits in Section C.6, with the power law becoming $t^{-1/3}$ once all represented templates are activated.

This provides a possible interpretation of the Pythia analysis in Section C.6. Each Pythia model is trained at fixed parameter count on a large text stream, with approximately 1.7×10^{11} tokens processed by the upper end of our fit window (Biderman et al., 2023). Common linguistic patterns may therefore be well sampled even while finite model capacity limits the range of patterns that can be represented. In the present model, the corresponding regime requires an approximately fixed coverage cutoff $k_{\max}$ and sufficient training of all represented sub-tasks, $t \gg t_*(k_{\max})$: the student is capacity limited in coverage but data rich within its represented subset. Further training then refines predictions within that subset, while the unrepresented tail contributes a constant floor. This offers a mechanism for the power-law-plus-floor behavior considered in Section C.6, with the schedule-dependent time coordinate used there providing a heuristic extension to Pythia. The aggregate fits alone do not establish that Pythia has reached this regime, and its fitted floor can also contain irreducible next-token uncertainty.