RELATIVE WEIGHT SCALES CONTROL FEATURE SPECIALIZATION IN RELU NETWORKS

Anonymous authors
Paper under double-blind review

ABSTRACT

The success of deep learning depends on learning useful representations, yet predicting how training organizes these representations across neurons remains difficult. In this work, we show that changing the scale of initial weights can determine which neurons learn a feature without altering any neuron's initial contribution. We construct ReLU networks with identical initial features and predictions that reach the same final predictions with different roles for their neurons. In one, all neurons share the learned feature. In the other, one neuron acquires it while every other neuron's contribution vanishes. The only change is the relative scale of each neuron's input and output weights. Our analysis explains how an initial learning advantage persists through convergence: as one neuron learns the target, it reduces the error driving the others and limits their subsequent adaptation. We prove this outcome in a nonlinear model under gradient flow and small-step gradient descent, and quantify how scale changes the speed and path of feature learning. Experiments verify the predicted dynamics and show that scale also changes feature assignment when two features compete. These results reveal how initialization can control the organization of a learned representation without changing what the network initially represents.

1 INTRODUCTION

Deep learning has made the acquisition of useful representations a central part of artificial intelligence (LeCun et al., 2015). These representations emerge from the joint adaptation of many neurons, although much of a trained network can often be removed with little loss in predictive accuracy (Frankle & Carbin, 2019; Liu et al., 2019). This raises a natural question about how learning uses a network's capacity: what makes some neurons acquire useful features while others become redundant?

Empirical studies of pruning offer seemingly different answers about the importance of the starting weights. Frankle & Carbin (2019) find sparse subnetworks that train successfully when reset to their original initial weights, whereas random reinitialization degrades their performance. In structured pruning experiments, however, Liu et al. (2019) find that training the retained architecture from random initialization can match or outperform fine-tuning the inherited weights. The usefulness of a retained architecture and that of its particular weights need not coincide.

These observations motivate an important distinction between recovering a network's accuracy and explaining how its components became useful. A reinitialized network may learn the task through a different internal representation. To understand which neurons carry that representation, we need to follow how initialization affects their learning. Starting weights determine both the features already represented and how those features adapt as the neurons reduce a shared prediction error. A neuron's eventual importance could therefore reflect an advantage in its initial feature or in its ability to adapt. To isolate the role of adaptation, we hold the architecture, data, and training rule fixed and ask whether neurons with identical initial features and contributions can still acquire different roles.

ReLU networks admit exactly this comparison. Multiplying a neuron's output weight by a positive factor and dividing its input weight by the same factor preserves its feature direction and its contribution to the network. This positive rescaling is a symmetry of the network function. Gradient descent updates the two weights separately, so their relative scales affect how the neuron learns. Kunin et al. (2024) show that this imbalance can accelerate feature learning.

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.

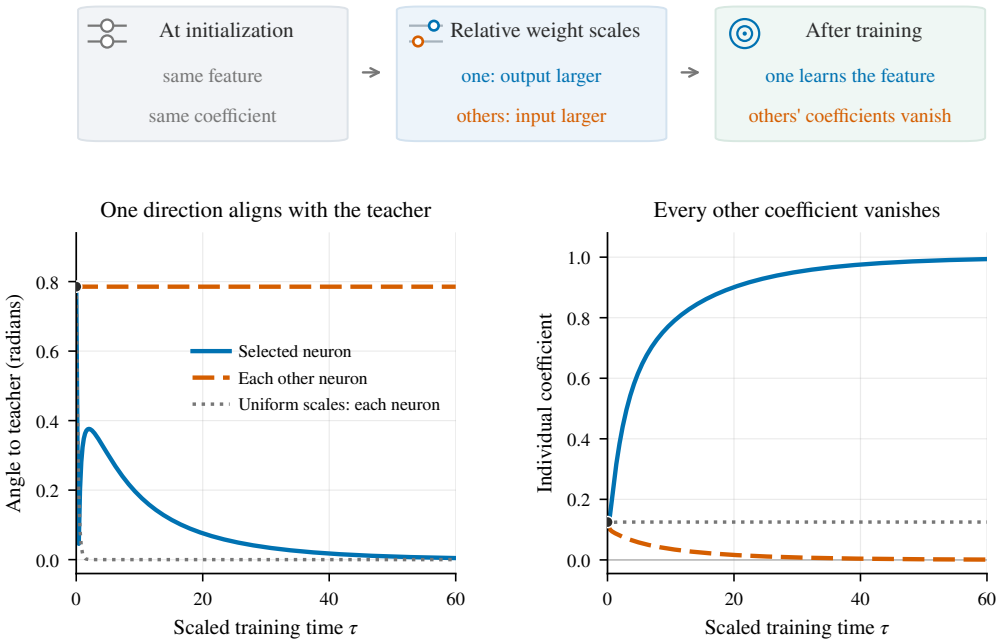

Figure 1: Identical initial features, different roles after training. Eight students start with the same direction and coefficient $1/8$. With the same relative scale across all students (gray dotted), all learn the teacher and retain equal coefficients. With a larger output weight relative to the input weight in one student (blue), and the opposite imbalance in the other seven (orange dashed), only the blue direction aligns and every orange coefficient vanishes. Neurons within each group follow identical trajectories by symmetry. These are population gradient flows with imbalance magnitude $D = 32$, where $D = |a^2 - \|w\|^2|$ for output weight a and input vector w. The coefficient is $q = a\|w\|$ and scaled time is $\tau = Dt$, where t is training time.

Our main result is that changing only the relative weight scales can determine whether a feature is shared or learned by a single neuron. Our construction starts all neurons with identical features and contributions. When their relative scales also agree, they remain identical and share the learned feature equally. Giving one neuron a sufficiently large output weight relative to its input weight, and the others the opposite imbalance, instead makes that neuron learn the feature while every other coefficient tends to zero (figure 1). Interestingly, both networks converge to the same prediction function. Thus, even identical initial and final predictions can conceal a different allocation of the learned feature across neurons.

We establish this result in a Gaussian model with one teacher feature and duplicate student initializations, under the angle condition of theorem 3.3. An early learning advantage alone does not guarantee this outcome, since slower neurons could continue to align and eventually share the feature. The proof shows why their remaining adaptation is limited. In the large-imbalance limit, the designated direction rotates toward the teacher while the other directions remain fixed. We prove convergence of this limiting system and then follow the trajectories at finite imbalance. As the selected student fits the target, the prediction error decays, limiting the entire subsequent rotation of the others. Their separation from the teacher forces their coefficients to vanish individually. Redundancy in this construction is therefore produced by the learning dynamics even though the neurons begin with the same feature.

Single-neuron analysis gives a quadratic separation in specialization times and inequivalent learning paths. Stability estimates give finite-time selection near duplicate initialization and convergence under gradient descent on the population loss with sufficiently small steps. Experiments verify specialization and coefficient decay. With two competing features, they show changes in assignment among nonduplicate students when only their scales change. Our results connect function-preserving initialization choices to the learned representation.

2 Related work

Scale and feature learning. Kunin et al. (2024) solve a minimal linear model and analyze how relative layer scales affect nonlinear feature learning. Their signed radial dynamics in Appendix C.2, Eq. (131), give our coefficient and direction equations when both learning rates equal one.

Teacher–student networks. Teacher–student analyses include population gradients and critical points (Tian, 2017), specialization at finite width and dimension (Tian, 2020), local convergence near over-realized solutions (Zhou et al., 2021), and regularized recovery (Akiyama & Suzuki, 2021). Overparameterization also affects the loss landscape and convergence speed (Safran et al., 2021; Xu & Du, 2023). ReLU identifiability results characterize equivalent parameters up to permutation and positive rescaling (Bona-Pellissier et al., 2022; 2023). Our local analysis uses a quantitative Gram-matrix condition to identify individual coefficients.

Symmetry and conservation. Gradient flow in homogeneous networks preserves differences of squared layer norms (Du et al., 2018; Marcotte et al., 2023), and initialization dependence also appears in adaptive-knot models (Williams et al., 2019). Path-lifting gives an intrinsic description of arbitrary-depth ReLU gradient flow for a dense set of initializations (Marcotte et al., 2026). Related optimization methods choose a better-conditioned parameterization (Lebeurrier et al., 2026), encourage balance (Terin, 2026), or use a geometry invariant to node rescaling (Neyshabur et al., 2015). We keep Euclidean training fixed and use the conserved imbalance to compare trajectories from the same predictor.

Kernel and distributional dynamics. Kernel analyses describe training near a fixed linearization (Jacot et al., 2018; Chizat et al., 2019) and the evolution of features beyond that regime (Atanasov et al., 2022). Infinite-width parameterizations (Yang & Hu, 2021), mean-field dynamics (Mei et al., 2019), and wide and deep linear models (Bordelon & Pehlevan, 2025) further relate initialization to learning. Our analysis follows individual neurons through convergence at finite width.

3 Theory

Our main construction uses a two-layer ReLU network without biases, Gaussian inputs, and a single ReLU target. We preserve each neuron's initial contribution while changing its relative weight scales. We first derive the coefficient and direction dynamics for general inputs, then show how their relative speeds and interaction through a shared prediction error determine specialization.

Notation. Throughout, $d \geq 2$ is the input dimension, $\|\cdot\|$ is the Euclidean norm (the operator norm for matrices), and $\mathbb{S}^{d-1}$ is the unit sphere. We write $[v]_+ = \max\{v, 0\}$ and $\angle(s, v) = \arccos(s^\top v)$ for unit vectors. A dot denotes differentiation in training time t. We use $O(\cdot)$ for an upper bound and $\Theta(\cdot)$ for upper and lower bounds of the same order, with fixed parameters unless specified otherwise.

3.1 Exact coefficient–direction dynamics

Let $\Theta = ((a_i, w_i))_{i=1}^m$ collect the output weights $a_i \in \mathbb{R}$ and input vectors $w_i \in \mathbb{R}^d$ of m neurons. For inputs x with finite second moment and a square-integrable target $y(x)$, consider the predictor and population square loss

$$f_\Theta(x) = \sum_{i=1}^m a_i [w_i^\top x]_+, \qquad \mathcal{L}(\Theta) = \frac{1}{2}\mathbb{E}\left[(f_\Theta(x) - y(x))^2\right]. \tag{1}$$

Expectations are over the input distribution, which is assumed to give zero mass to the relevant ReLU switching hyperplanes. To describe the geometry of learning, we separate each neuron's coefficient, direction, and scale. For $w_i \neq 0$, write

$$r_i = \|w_i\|, \qquad s_i = \frac{w_i}{r_i}, \qquad q_i = a_i r_i, \qquad \delta_i = a_i^2 - r_i^2. \tag{2}$$

The signed coefficient q_i and direction s_i determine the contribution $q_i[s_i^\top x]_+$. The imbalance δ_i records the relative squared weight magnitudes: $\delta_i > 0$ means $|a_i| > r_i$, while $\delta_i < 0$ means

$|a_i| < r_i$. Changing δ_i at fixed (q_i, s_i) preserves this contribution. Let $R_\Theta = f_\Theta - y$ be the prediction residual, $I = I_d$ the identity matrix, and $\mathbf{1}_A$ the indicator of an event A. Define the residual averages and rate factors

$$B_\Theta(s) = \mathbb{E}\big[R_\Theta(x)[s^\top x]_+\big], \quad \mathcal{T}_\Theta(s) = \mathbb{E}\big[R_\Theta(x)\mathbf{1}_{\{s^\top x>0\}}x\big], \\ \mu(q,\delta) = \sqrt{\delta^2+4q^2}, \qquad \chi(q,\delta) = \frac{2q}{\mu(q,\delta)-\delta}. \tag{3}$$

Under Euclidean gradient flow $\dot\Theta = -\nabla_\Theta \mathcal{L}$, and as long as $w_i \neq 0$, these variables obey the exact system

$$\dot q_i = -\mu_i B_\Theta(s_i), \qquad \dot s_i = -\chi_i(I - s_i s_i^\top)\mathcal{T}_\Theta(s_i), \qquad \dot\delta_i = 0, \tag{4}$$

Here $\mu_i = \mu(q_i,\delta_i) = a_i^2 + r_i^2$ and $\chi_i = \chi(q_i,\delta_i) = a_i/r_i$. For $q_i \neq 0$, this is the signed radial–directional system of Kunin et al. (2024, Appendix C.2, Eq. (131)): their $(\mu_i, \hat\beta_i)$ correspond to our (q_i, s_i) when both layer learning rates equal one. The same coordinates separate the loss decrease into changes in coefficients and directions:

$$-\dot{\mathcal{L}} = \sum_i \left[\underbrace{\mu_i B_\Theta(s_i)^2}_{\text{coefficient change}} + \underbrace{\frac{\mu_i+\delta_i}{2}\big\|(I - s_i s_i^\top)\mathcal{T}_\Theta(s_i)\big\|^2}_{\text{direction change}} \right]. \tag{5}$$

This expression separates fitting through coefficient adjustment from fitting through feature rotation. The conserved imbalance sets their relative rates, as derived in Appendix A.

Notice that the coefficient rate is unchanged when the sign of the imbalance is reversed. At a shared (q, s) with $q \neq 0$ and the same residual, the two parameterizations with $\delta = \pm D$ have identical coefficient velocities, but their direction-update factors satisfy

$$\frac{\chi(q,+D)}{\chi(q,-D)} = \left(\frac{\sqrt{D^2+4q^2}+D}{2q}\right)^2 = \Theta(D^2/q^2).$$

Thus the same functional state can evolve at very different angular speeds. To understand the resulting trajectories, we must also follow the changing coefficients and residual.

For the Gaussian model below, let $\gamma_d = \mathcal{N}(0, I_d)$, with $\|f\|^2_{L^2(\gamma_d)} = \mathbb{E}_{x\sim\gamma_d}[f(x)^2]$, and fix a target direction $v_\star \in \mathbb{S}^{d-1}$. We say that unit i specializes to $v_\star$ if $s_i(t) \to v_\star$ and $\liminf_{t\to\infty} |q_i(t)| > 0$. For redundant units, we study whether each coefficient converges to zero: $q_j(t) \to 0$ for every j.

3.2 Function-preserving scaling and specialization time

We begin with a single student, for which the residual is determined by its coefficient and alignment with the teacher. This gives a closed nonlinear system that lets us compare the full trajectories from a common initial function. We use Euclidean gradient flow throughout; proposition A.3 and remarks A.2 and A.6 discuss rescaling in deeper networks, scale-equivariant updates, and weight decay.

Let $x \sim \mathcal{N}(0, I_d)$, $q_\star > 0$, $\|s_\star\| = 1$, $f_\star(x) = q_\star[s_\star^\top x]_+$, and $c = s^\top s_\star$. Here c is the alignment of the student direction with the teacher. With the arc-cosine kernel

$$\kappa(c) = \frac{\sqrt{1-c^2} + (\pi - \arccos c)c}{2\pi},$$

the residual averages depend only on q and c, giving the closed system

$$\dot q = -\mu(q,\delta)\left(\frac{q}{2} - q_\star\kappa(c)\right), \qquad \dot c = \chi(q,\delta)q_\star\kappa'(c)(1-c^2), \qquad \dot\delta = 0. \tag{6}$$

The coefficient nullcline $q = 2q_\star\kappa(c)$ is the set on which $\dot q = 0$. We compare trajectories in the coefficient–alignment plane (q, c). For a common initial state (q_0, c_0) with $q_0 > 0$ and $-1 < c_0 < 1$, let (q_δ, c_δ) denote the solution with imbalance δ. Given an alignment tolerance $0 < \varepsilon < 1 - c_0$, define the specialization time

$$T_\delta(\varepsilon) = \inf\{t \geq 0 : c_\delta(t) \geq 1 - \varepsilon\}.$$

The next theorem gives matching orders in both imbalance and accuracy. Explicit upper and lower constants are given in theorems A.9 and A.10.

Theorem 3.1. *For every fixed $\delta \in \mathbb{R}$, $(q_\delta(t), c_\delta(t)) \to (q_\star, 1)$. For all sufficiently large D, uniformly over $0 < \varepsilon \le (1 - c_0)/2$,*

$$T_{+D}(\varepsilon) = \Theta\big(D^{-1}\log(1/\varepsilon)\big)\,, \qquad T_{-D}(\varepsilon) = \Theta(D\log(1/\varepsilon))\,. \tag{7}$$

Moreover, uniformly over the full range $0 < \varepsilon < 1 - c_0$,

$$T_{-D}(\varepsilon)/T_{+D}(\varepsilon) = \Theta(D^2). \tag{8}$$

The threshold on D and the constants in these two-sided bounds depend only on $q_0, c_0, q_\star$.

Corollary 3.2. *For every $D > 0$, $q_0 > 0$, and $-1 < c_0 < 1$, the trajectories of (6) with $\delta = +D$ and $\delta = -D$ and common initial state (q_0, c_0) cannot agree on any nontrivial interval after a strictly increasing C^1 reparameterization of time. Away from $q/2 = q_\star\kappa(c)$ the two vector fields are non-collinear. On this nullcline they have the same tangent direction, but their phase curves have different curvature.*

The two imbalances therefore yield the same learned feature through different sequences of learning, as illustrated in figure D.3. With negative imbalance, the coefficient first relaxes while the direction remains nearly fixed; positive imbalance allows rapid alignment. This difference persists on the coefficient nullcline: a common phase curve would have to remain on the nullcline with constant coefficient, which is impossible (corollary A.11).

3.3 Feature assignment among duplicate neurons

The single-student result establishes a difference in adaptation rates. We now show how this difference produces the allocation in figure 1 when students interact through the same residual. We initialize them with identical functional contributions and change only their parameter scales. Let the teacher be $f_\star(x) = [u^\top x]_+$ with $\|u\| = 1$, and initialize $m \ge 2$ students with the same coefficient and feature direction:

$$q_i(0) = \frac{1}{m}, \qquad s_i(0) = s_0, \qquad 0 < \phi = \angle(s_0, u) \le \pi/3.$$

For uniform constants, fix any compact interval $I_\phi \subset (0, \pi/3]$ containing ϕ. Set $\delta_k = +D$ for one designated index k and $\delta_j = -D$ for every $j \ne k$. For comparison, if all imbalances equal a common δ, symmetry keeps the students identical. Their total coefficient $Q = mq_i$ and alignment $c = s_i^\top u$ obey (6) with imbalance $m\delta$ and teacher coefficient one. Thus theorem 3.1 gives $q_i \to 1/m$ and $s_i \to u$ for every student. The mixed scale assignment changes this shared representation into the following one.

Write the contribution of the redundant units as

$$f_{\mathrm{red}}(x, t) := \sum_{j \ne k} q_j(t)[s_j(t)^\top x]_+.$$

Theorem 3.3. *Consider Euclidean gradient flow on the population loss (1) with $x \sim \mathcal{N}(0, I_d)$, $y = f_\star$, and the initialization above. For every fixed $m \ge 2$ and $\phi \in I_\phi$, there are $D_0(m), C_m, c_m > 0$, independent of D, such that for every $D \ge D_0(m)$,*

$$q_k(t) \to 1, \qquad s_k(t) \to u, \qquad q_j(t) \to 0 \quad (j \ne k).$$

There is a time $t_{\mathrm{cap}} \le C_m D^{-1}$ after which the loss decays exponentially:

$$\mathcal{L}(t) \le \mathcal{L}(t_{\mathrm{cap}})e^{-c_m D(t - t_{\mathrm{cap}})}, \qquad t \ge t_{\mathrm{cap}},$$

and every redundant direction satisfies

$$\int_{t_{\mathrm{cap}}}^{\infty} \|\dot{s}_j(t)\|\,\mathrm{d}t \le C_m D^{-2} \qquad (j \ne k).$$

Thus unit k is the unique student that specializes to the teacher, and every redundant coefficient converges to zero. Permuting the unique $+D$ imbalance permutes the selected index while leaving the initial predictor unchanged.

Thus the initial scale determines the limiting representation, including the disappearance of every redundant contribution. To understand how this occurs, we study the fast time $\tau = Dt$ (theorem C.3). Let $q = q_k$, let $P = \sum_{j \neq k} q_j$ be the total redundant coefficient, and let θ and ψ be the signed angles of the selected and common redundant directions relative to u. The teacher plane is spanned by u and s_0, with orientation chosen so that s_0 has angle ϕ. In the limit, $\psi = \phi$ remains fixed. If s_θ is the unit vector at angle θ in this plane, the limiting predictor is $q[s_\theta^\top x]_+ + P[s_0^\top x]_+$. Its population loss is

$$E(q, P, \theta) = \frac{1}{2} \left\| q[s_\theta^\top \cdot]_+ + P[s_0^\top \cdot]_+ - f_\star \right\|^2_{L^2(\gamma_d)}.$$

Writing primes for τ derivatives, the limiting dynamics are

$$q' = -\partial_q E, \qquad P' = -(m-1)\partial_P E, \qquad \theta' = -q^{-2}\partial_\theta E, \tag{9}$$

from $(q, P, \theta) = (1/m, (m-1)/m, \phi)$. Notice that all redundant students enter through their total coefficient P. Their number changes the rate of its adjustment, while the loss landscape remains the same.

We first show that the coefficient-weighted direction $z = qs_\theta$ develops a component toward the teacher, keeping z away from zero. Kernel inequalities then exclude every stationary state except $(1, 0, 0)$ in an invariant region containing the initialization. The limiting flow converges to this state, leaving the redundant direction at angle $\phi > 0$.

At finite imbalance, the redundant directions also move. The restoring force in the equation for P limits the accumulated coefficient error, while the lower bound on $\|z\|$ prevents the direction equation from becoming singular. Together they keep the trajectory close to the limit even when the selected coefficient is initially small. This places the finite-D trajectory near $(1, 0, 0)$ for $D \geq D_0(m) = Am^{\nu/2}$, where $A > 0$ and $\nu > 3/2$ depend only on I_ϕ (proposition C.2).

Near the selected solution, the loss bounds the errors in q, θ, and P. The entry loss is below its minimum on the boundary for these coordinates, so decreasing loss prevents them from leaving the neighborhood. The total-rotation bound keeps the redundant directions inside their allowed angular region, away from the teacher. Exponential loss decay then gives $q \to 1$, $\theta \to 0$, and $P \to 0$. By symmetry, $q_j = P/(m-1) \to 0$ for every $j \neq k$.

3.4 Robustness and discrete gradient descent

For nearby but nonidentical initializations, the designated neuron still reaches the target neighborhood first. Subsequent convergence requires the local conditions of theorem C.7. With one positive scale imbalance and $m - 1$ negative ones, each with magnitude near D, measure departure from duplicate initialization by

$$\Delta_0 = \max_i \left(\left| q_i(0) - \frac{1}{m} \right| + \|s_i(0) - s_0\| \right) + \max_i \left| \frac{|\delta_i(0)|}{D} - 1 \right|. \tag{10}$$

Choose a small selection tolerance $r_\star > 0$ relative to the initial angle and the local convergence neighborhood. Let $T_\star$ bound the fast time needed by the limiting flow to enter a smaller neighborhood, and let $D_0(m) = Am^{\nu/2}$ be the sufficient imbalance threshold above. Appendix C fixes these constants before proposition C.2. For the state $Z = ((q_\ell, s_\ell))_{\ell=1}^m$, write $f_{-i}(x; Z) = \sum_{j \neq i} q_j [s_j^\top x]_+$ for the output of all other students. Unit i is selected when its state enters

$$\mathcal{W}_{\text{sel}}^{(i)} := \left\{ Z : q_i \geq 1 - r_\star, \quad \angle(s_i, u) \leq r_\star, \quad \|f_{-i}(\cdot; Z)\|_{L^2(\gamma_d)} \leq r_\star \right\}.$$

Theorem 3.4. *There is $a > 0$, depending only on I_ϕ, such that, with ν as above, for every $m \geq 2$ and $\phi \in I_\phi$, if $\Delta_0 \leq \varepsilon_m := am^{-\nu}$ and $D \geq D_0(m)$, then the first entry time*

$$t_{\text{sel}} := \inf\{t \geq 0 : Z(t) \in \mathcal{W}_{\text{sel}}^{(k)}\}$$

satisfies $t_{\text{sel}} \leq (T_\star + 1)/D$, while $Z(t) \notin \mathcal{W}_{\text{sel}}^{(i)}$ for every $i \neq k$ and $0 \leq t \leq t_{\text{sel}}$. At t_{sel}, every unit with negative imbalance satisfies $\angle(s_j(t_{\text{sel}}), u) > r_\star$.

The proof controls the total redundant coefficient and each neuron's deviation from the mean separately (theorem C.4). The local conditions for subsequent convergence are a Gram-matrix lower bound and an entry-loss bound.

For gradient descent, the full convergence result persists even though the imbalance changes at each step. We return to the duplicate initialization of section 3.3, with $\phi \in I_\phi$, and use learning rate $\eta = h/D$, where $h > 0$ is the step in fast time. Write $Y_n = (q_n, P_n, \theta_n, \psi_n)$ for the symmetric state after n iterations with $\mathcal{L}_n = \mathcal{L}(\Theta_n)$ and $\delta_{i,n} = a_{i,n}^2 - \|w_{i,n}\|^2$. Let $T_c = T_\star + 1$ and $Y_\infty(\tau) = (q_\infty(\tau), P_\infty(\tau), \theta_\infty(\tau), \phi)$, where the first three coordinates solve (B.13). Let $\mathcal{V}_{\text{in}}$ be an inner neighborhood of the selected solution, with a strict loss gap to the boundary of a larger neighborhood. The construction in (C.37) also bounds the original parameter scales away from zero.

Theorem 3.5. *For every fixed $m \geq 2$ and $\phi \in I_\phi$, there are $h_0(m), \widetilde{D}_0(m), C_m, c_m > 0$ such that, for all $0 < h \leq h_0(m)$ and $D \geq \widetilde{D}_0(m)$, full-batch gradient descent on the population loss converges to the selected representation: $q_n \to 1$, $P_n \to 0$, and $\theta_n \to 0$. In particular, each redundant coefficient $q_{j,n} = P_n/(m-1)$ tends to zero. Up to fast time T_c, the trajectory satisfies*

$$\max_{nh \leq T_c} \|Y_n - Y_\infty(nh)\| \leq C_m(h + D^{-2}).$$

Let n_{sel} be the first index at which the iterate enters the symmetric neighborhood $\mathcal{V}_{\text{in}}$ in (C.37). Then $n_{\text{sel}} \leq \lceil T_c/h \rceil$ and, for every $n \geq n_{\text{sel}}$,

$$\mathcal{L}_{n+1} \leq (1 - c_m h)\mathcal{L}_n, \qquad \sup_{n \geq 0} |\delta_{i,n} - \delta_{i,0}| \leq C_m \frac{h}{D} \quad \text{for every } i,$$

and, writing s_n for the common redundant direction at iteration n, we have

$$\sum_{n \geq n_{\text{sel}}} \|s_{n+1} - s_n\| \leq \frac{C_m}{D^2} \mathcal{L}_{n_{\text{sel}}}.$$

The imbalance drift is small enough to preserve the separation of learning rates. To show this, we express the original update as an exact Euler step in linearly scaled parameters and compare its interpolation with the limiting flow. Near the selected representation, a Hessian bound over each complete step gives loss decrease and prevents escape. Summing these decreases controls both direction motion and imbalance drift (theorem C.11).

4 Experiments

Our analysis shows that parameter scales can select a unique specialized neuron while redundant coefficients decay. Here we test these dynamics and their dependence on imbalance, sample size, and step size, then examine feature assignment with two competing teachers. We compare the large-D limit, the exact finite-D equations, gradient flow in the original parameters on the population loss, and gradient flow for the empirical loss on N Gaussian samples. Appendix D gives the protocols and complete results.

4.1 Specialization and coefficient decay

Figure 1 compares two scale assignments at the same initial functional state. With all imbalances positive, the students learn the teacher together and retain equal coefficients. With one positive and seven negative imbalances, the selected student aligns with the teacher and its coefficient approaches one. Each redundant coefficient tends to zero while its direction remains separated from the teacher. Figure 2 follows this process in all four dynamical descriptions. The reduced and original-parameter population flows agree to solver precision, checking two implementations of the same dynamics. Comparisons with the limiting and empirical flows test the large-D approximation and finite-sample effects. Panel (d) separates loss decrease through coefficient adjustment and feature rotation.

To isolate the initial difference in adaptation, figure D.3 compares single students with the same initial coefficient and alignment. Positive imbalance produces rapid alignment, while negative imbalance gives coefficient adjustment followed by slower rotation. Their specialization times differ by the predicted quadratic factor, and the paths themselves differ. The limiting approximations capture both phases (figure D.1(a)): slow-time alignment is compared from time zero, while the coefficient approximation applies after its initial relaxation.

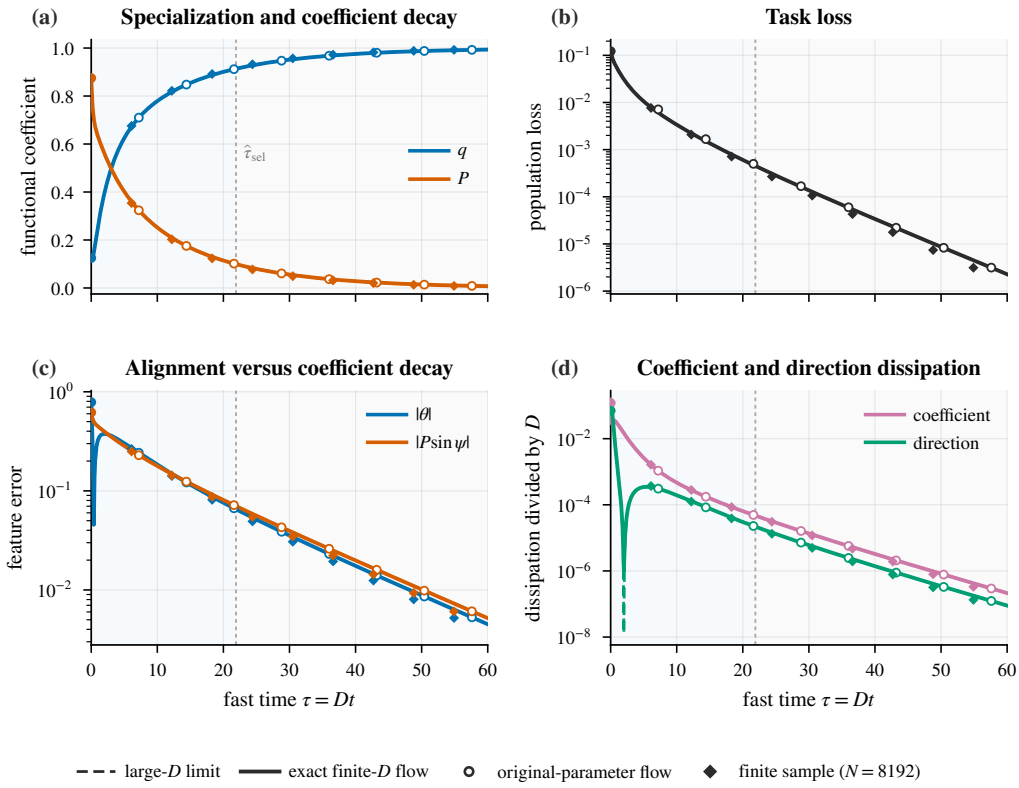

Figure 2: A designated neuron learns the teacher as redundant coefficients decay. The four dynamical descriptions agree along a representative trajectory with $m = 8$, $D = 32$, and $N = 8192$. (a) Selected coefficient q and total redundant coefficient P. The vertical line marks $\widehat{\tau}_{\text{sel}} = D\widehat{t}_{\text{sel}}$, defined by the thresholds in (D.3). (b) Population loss. (c) Alignment of the selected direction, $|\theta|$, and decay of the redundant contribution, $|P \sin \psi|$. (d) Loss decrease due to coefficient and direction changes, divided by D.

The two direction curves in Figure 2(c) reflect different mechanisms. The selected angle approaches zero, but $|P \sin \psi|$ decreases through coefficient decay: the redundant angle remains nearly fixed (Figure 1). Since the redundant students remain identical, each coefficient vanishes individually.

4.2 Scaling and Stability

We next vary the imbalance to test how the separation of learning rates controls selection. Figures 3(a,b) show faster selection and smaller subsequent direction motion with the predicted D^{-1} and D^{-2} orders across widths. As the redundant directions become more nearly fixed, the limiting-trajectory error also decreases quadratically (figure D.1(b), table D.2).

Measuring direction motion after selection tests whether the remaining neurons continue to change their features. The observed decrease supports the total-motion bound in theorem 3.3, connecting rapid feature acquisition to a persistent separation of selected and redundant directions.

The same selected neuron is recovered in every finite-sample run, with trajectory error decreasing approximately as $N^{-1/2}$ (figure 3(c), table D.3). Selection persists under the tested perturbations of the initial coefficients, directions, and imbalances. Figure D.1(c) confirms selection across the tested angles and widths. Panel (d) shows the weak dependence of rescaled selection time on width. Under gradient descent, the trajectory error decreases linearly with $h = \eta D$ (figure 3(d)), while subsequent redundant-direction motion retains its D^{-2} dependence (table D.5).

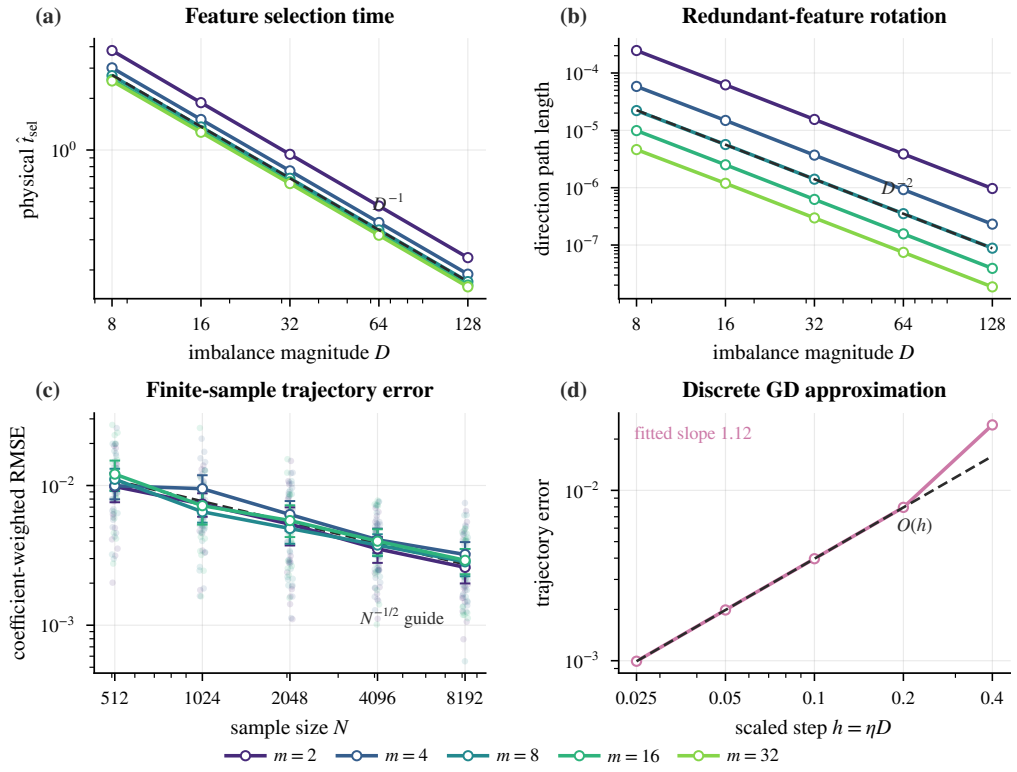

Figure 3: Scaling with imbalance, sample size, and step size. (a) Measured selection time $\widehat{t}_{\mathrm{sel}}$ from (D.3) follows D^{-1}. (b) Subsequent redundant-direction motion follows D^{-2}. (c) Empirical trajectory error follows approximately $N^{-1/2}$, with paired-bootstrap 95% confidence intervals. (d) Gradient-descent error is approximately linear in $h = \eta D$. Dashed lines mark reference orders.

4.3 Competition between two teacher features

We next train eight students with nonduplicate initial directions on two teacher features in ten dimensions. Every student's initial coefficient and direction remain fixed as the imbalance assignment changes. Configurations A and B give positive imbalance to units $(1, 4)$ and $(2, 3)$, respectively, with one designated student near each teacher; a third configuration uses $(1, 2)$, both nearer the first teacher. Success requires the designated students to recover their teachers' directions and coefficients, with all remaining coefficients small ((D.7)).

Moving the positive imbalances from A to B changes which neurons acquire the two features (figure D.2(a,b)). The fixed initial features isolate the effect of scale from differences in alignment. Both configurations succeed for every seed at each tested imbalance (table D.6). When both designated students begin nearer the first teacher, other students can retain part of the second feature. These experiments show how parameter scale and initial direction jointly determine feature assignment beyond the duplicate setting.

5 Conclusion

Parameter scale can determine how neurons specialize even when their initial features and functional contributions are fixed. In a Gaussian ReLU model, we prove that a designated student learns the teacher while every redundant coefficient vanishes. The result holds at finite imbalance under gradient flow and small-step population gradient descent. Two-teacher experiments further show that scale can change feature assignment among nonduplicate students, linking parameterization to the roles neurons acquire during learning.

AI USE STATEMENT

AI tools assisted literature searches, proof drafting and revision, experimental design, numerical implementation and code editing. The authors are responsible for the correctness, originality, and presentation of the submitted work.

REPRODUCIBILITY STATEMENT

Appendices A–C contain complete proofs of the theoretical results. Appendix D gives the numerical protocols and the results underlying Figures 1–3 and the two-teacher competition experiment.

REFERENCES

Shunta Akiyama and Taiji Suzuki. On learnability via gradient method for two-layer ReLU neural networks in teacher-student setting. In *International Conference on Machine Learning*, pp. 152–162, 2021.

Alexander Atanasov, Blake Bordelon, and Cengiz Pehlevan. Neural networks as kernel learners: The silent alignment effect. In *International Conference on Learning Representations*, 2022.

Joachim Bona-Pellissier, François Malgouyres, and François Bachoc. Local identifiability of deep ReLU neural networks: The theory. In *Advances in Neural Information Processing Systems*, volume 35, pp. 27549–27562, 2022.

Joachim Bona-Pellissier, François Bachoc, and François Malgouyres. Parameter identifiability of a deep feedforward ReLU neural network. *Machine Learning*, 112(11):4431–4493, 2023. doi: 10.1007/s10994-023-06355-4.

Blake Bordelon and Cengiz Pehlevan. Deep linear network training dynamics from random initialization: Data, width, depth, and hyperparameter transfer. In *International Conference on Machine Learning*, volume 267, pp. 4968–4997, 2025.

Lénaïc Chizat, Edouard Oyallon, and Francis Bach. On lazy training in differentiable programming. In *Advances in Neural Information Processing Systems*, volume 32, pp. 2933–2943, 2019.

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

Jonathan Frankle and Michael Carbin. The lottery ticket hypothesis: Finding sparse, trainable neural networks. In *International Conference on Learning Representations*, 2019. URL https://openreview.net/forum?id=rJl-b3RcF7.

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

Daniel Kunin, Allan Raventós, Clémentine Dominé, Feng Chen, David Klindt, Andrew Saxe, and Surya Ganguli. Get rich quick: Exact solutions reveal how unbalanced initializations promote rapid feature learning. In *Advances in Neural Information Processing Systems*, volume 37, 2024.

Arthur Lebeurrier, Titouan Vayer, and Rémi Gribonval. Path-conditioned training: A principled way to rescale ReLU neural networks. In *Proceedings of the 43rd International Conference on Machine Learning*, volume 306 of *Proceedings of Machine Learning Research*, 2026. arXiv:2602.19799.

Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. Deep learning. *Nature*, 521(7553):436–444, 2015. doi: 10.1038/nature14539.

Zhuang Liu, Mingjie Sun, Tinghui Zhou, Gao Huang, and Trevor Darrell. Rethinking the value of network pruning. In *International Conference on Learning Representations*, 2019. URL https://openreview.net/forum?id=rJlnB3C5Ym.

Sibylle Marcotte, Rémi Gribonval, and Gabriel Peyré. Abide by the law and follow the flow: Conservation laws for gradient flows. In *Advances in Neural Information Processing Systems*, volume 36, 2023.

Sibylle Marcotte, Gabriel Peyré, and Rémi Gribonval. Intrinsic training dynamics of deep neural networks. In *International Conference on Learning Representations*, 2026.

Song Mei, Theodor Misiakiewicz, and Andrea Montanari. Mean-field theory of two-layers neural networks: Dimension-free bounds and kernel limit. In *Conference on Learning Theory*, pp. 2388–2464, 2019.

Behnam Neyshabur, Ruslan Salakhutdinov, and Nathan Srebro. Path-SGD: Path-normalized optimization in deep neural networks. In *Advances in Neural Information Processing Systems*, volume 28, pp. 2422–2430, 2015.

Itay M. Safran, Gilad Yehudai, and Ohad Shamir. The effects of mild over-parameterization on the optimization landscape of shallow ReLU neural networks. In *Conference on Learning Theory*, pp. 3889–3934, 2021.

Rodrigo Carmo Terin. Scale redundancy and soft gauge fixing in positively homogeneous neural networks. *arXiv preprint arXiv:2602.14729*, 2026.

Yuandong Tian. An analytical formula of population gradient for two-layered ReLU network and its applications in convergence and critical point analysis. In *International Conference on Machine Learning*, pp. 3404–3413, 2017.

Yuandong Tian. Student specialization in deep rectified networks with finite width and input dimension. In *International Conference on Machine Learning*, pp. 9470–9480, 2020.

Francis Williams, Matthew Trager, Claudio Silva, Daniele Panozzo, Denis Zorin, and Joan Bruna. Gradient dynamics of shallow univariate ReLU networks. In *Advances in Neural Information Processing Systems*, volume 32, 2019.

Weihang Xu and Simon Du. Over-parameterization exponentially slows down gradient descent for learning a single neuron. In *Conference on Learning Theory*, pp. 1155–1198, 2023.

Greg Yang and Edward J. Hu. Tensor Programs IV: Feature learning in infinite-width neural networks. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 11727–11737, 2021.

Mo Zhou, Rong Ge, and Chi Jin. A local convergence theory for mildly over-parameterized two-layer neural network. In *Conference on Learning Theory*, pp. 4577–4632, 2021.

APPENDICES

Appendix A derives the coefficient and direction dynamics and proves the single-neuron results. Appendix B establishes global convergence of the limiting system and the local loss bounds. Appendix C uses these estimates to prove selection and convergence at finite imbalance, under initialization perturbations, and for gradient descent. Appendix D gives the experimental protocols and additional results.

We use $O(\cdot)$ for upper bounds and $\Theta(\cdot)$ for two-sided bounds, with other parameters fixed. Subscripts record allowed dependence of the constants. Generic positive constants c, C may change between estimates.

A EXACT COEFFICIENT AND DIRECTION DYNAMICS AND ONE-NEURON SPECIALIZATION

A.1 POPULATION GRADIENT FLOW IN COEFFICIENT AND DIRECTION VARIABLES

The same ReLU feature can be represented by different input and output norms. We derive equations that keep this scale information while tracking the coefficient and direction seen by the predictor. The particle notation below includes the finite network as a sum of point masses.

Let π_x be a probability measure on $\mathbb{R}^d$ with finite second moment and assume that it assigns zero mass to every hyperplane used below. Let $y \in L^2(\pi_x)$ and $\sigma(z) = z_+ = \max\{z, 0\}$. For a positive measure ϱ with finitely many particles on

$$\mathcal{P} := \mathbb{R} \times (\mathbb{R}^d \setminus \{0\}),$$

define

$$f_\varrho(x) = \int a\, \sigma(w^\top x)\, \varrho(\mathrm{d}a,\, \mathrm{d}w), \qquad \mathcal{L}(\varrho) = \frac{1}{2} \int \big(f_\varrho(x) - y(x)\big)^2\, \pi_x(\mathrm{d}x). \tag{A.1}$$

For the finite network in the main text, ϱ is the sum of unit point masses at the neuron parameters, so the integral is exactly the network sum. Write $R_\varrho = f_\varrho - y$. Along a characteristic of the population gradient flow in its measure formulation,

$$\dot{a} = -\int R_\varrho(x)\sigma(w^\top x)\, \pi_x(\mathrm{d}x), \qquad \dot{w} = -a \int R_\varrho(x)\mathbf{1}_{\{w^\top x > 0\}} x\, \pi_x(\mathrm{d}x). \tag{A.2}$$

This is the standard two-layer mean-field characteristic system (Mei et al., 2019). The signed radial–directional transformation below agrees, for nonzero coefficients, with Kunin et al. (2024, Appendix C.2, Eq. (131)) at equal unit learning rates. We include its derivation to fix conventions and to retain the regular extension through $q = 0$ when $\delta < 0$.

For $w \neq 0$, set

$$r = \|w\|, \qquad s = \frac{w}{r} \in \mathbb{S}^{d-1}, \qquad q = ar, \qquad \delta = a^2 - r^2. \tag{A.3}$$

Define

$$\mu(q, \delta) := \sqrt{\delta^2 + 4q^2}, \qquad r^2(q, \delta) = \frac{\mu(q, \delta) - \delta}{2}, \qquad \chi(q, \delta) := \frac{2q}{\mu(q, \delta) - \delta}. \tag{A.4}$$

The coordinate domain is

$$\mathcal{X} := \big\{(q, s, \delta) : s \in \mathbb{S}^{d-1},\ \mu(q, \delta) > \delta\big\}. \tag{A.5}$$

This condition is equivalent to $r > 0$. The formula for χ is regular at $q = 0$ whenever $\delta < 0$ and equals a/r. The inverse transformation is

$$r = \sqrt{\frac{\mu - \delta}{2}}, \qquad a = \frac{q}{r}, \qquad w = rs. \tag{A.6}$$

Hence the map $(a, w) \mapsto (q, s, \delta)$ is a smooth bijection from $\mathcal{P}$ onto $\mathcal{X}$.

Let ρ be the pushforward of ϱ under this coordinate map, namely the distribution obtained by applying the map to each particle. Positive homogeneity gives

$$f_\rho(x) = \int q\, \sigma(s^\top x)\, \rho(\mathrm{d}q,\, \mathrm{d}s,\, \mathrm{d}\delta). \tag{A.7}$$

With $R_\rho = f_\rho - y$, define the residual moments

$$B_\rho(s) := \int R_\rho(x)\sigma(s^\top x)\, \pi_x(\mathrm{d}x), \qquad \mathcal{T}_\rho(s) := \int R_\rho(x)\mathbf{1}_{\{s^\top x > 0\}} x\, \pi_x(\mathrm{d}x). \tag{A.8}$$

By homogeneity,

$$B_\rho(s) = s^\top \mathcal{T}_\rho(s). \tag{A.9}$$

Let $\mathsf{P}_s^\perp = I - ss^\top$.

Theorem A.1. *Let $(\varrho_t)_{t\in[0,T]}$ be a characteristic solution of (A.2) whose support remains in $\mathcal{P}$. Its pushforward ρ_t under the map $(a, w) \mapsto (q, s, \delta)$ satisfies*

$$\partial_t \rho + \partial_q(V_q \rho) + \mathrm{div}_{\mathbb{S}^{d-1}}(V_s \rho) = 0, \qquad V_\delta = 0, \tag{A.10}$$

where $\mathrm{div}_{\mathbb{S}^{d-1}}$ is spherical divergence and the coordinate velocities are

$$V_q = -\mu(q, \delta) B_\rho(s), \qquad V_s = -\chi(q, \delta)\mathsf{P}_s^\perp \mathcal{T}_\rho(s), \qquad V_\delta = 0. \tag{A.11}$$

Conversely, every characteristic solution of (A.10) that remains in $\mathcal{X}$ pulls back through (A.6) to a solution of (A.2). The two systems are therefore equivalent up to the first exit from the regular coordinate domain.

Moreover, the loss decay separates exactly into coefficient and directional contributions:

$$\frac{\mathrm{d}}{\mathrm{d}t}\mathcal{L}(\rho_t) = -\int \left[\underbrace{\mu(q,\delta)B_{\rho_t}(s)^2}_{\text{coefficient change}} + \underbrace{\frac{\mu(q,\delta)+\delta}{2}\left\|\mathsf{P}_s^{\perp}\mathcal{T}_{\rho_t}(s)\right\|^2}_{\text{direction change}} \right] \rho_t(\,\mathrm{d}q,\,\mathrm{d}s,\,\mathrm{d}\delta). \tag{A.12}$$

In particular, the entire δ-marginal is conserved.

Proof. Since $\sigma(w^\top x) = r\sigma(s^\top x)$, (A.2) becomes

$$\dot{a} = -rB_\rho(s), \qquad \dot{w} = -a\mathcal{T}_\rho(s). \tag{A.13}$$

Taking radial and tangential components gives

$$\dot{r} = s^\top \dot{w} = -aB_\rho(s), \qquad \dot{s} = \frac{1}{r}\mathsf{P}_s^{\perp}\dot{w} = -\frac{a}{r}\mathsf{P}_s^{\perp}\mathcal{T}_\rho(s). \tag{A.14}$$

Therefore

$$\dot{q} = r\dot{a} + a\dot{r} = -(r^2+a^2)B_\rho(s) = -\mu(q,\delta)B_\rho(s), \tag{A.15}$$

$$\dot{\delta} = 2a\dot{a} - 2r\dot{r} = -2arB_\rho(s) + 2arB_\rho(s) = 0. \tag{A.16}$$

The identity $a/r = \chi(q,\delta)$ yields the stated s-velocity. Applying the chain rule to test functions proves the pushforward continuity equation. The inverse calculation gives the converse.

For dissipation, the original Euclidean gradient flow gives

$$\begin{aligned}\frac{\mathrm{d}}{\mathrm{d}t}\mathcal{L} &= -\int \left[r^2 B_\rho(s)^2 + a^2\|\mathcal{T}_\rho(s)\|^2\right]\,\mathrm{d}\rho \\ &= -\int \left[(r^2+a^2)B_\rho(s)^2 + a^2\|\mathsf{P}_s^{\perp}\mathcal{T}_\rho(s)\|^2\right]\,\mathrm{d}\rho,\end{aligned} \tag{A.17}$$

where (A.9) was used to decompose $\mathcal{T} = Bs + \mathsf{P}_s^{\perp}\mathcal{T}$. Using

$$r^2 + a^2 = \mu(q,\delta), \qquad a^2 = \frac{\mu(q,\delta)+\delta}{2},$$

proves (A.12). □

For fixed $q \neq 0$, the relative directional weight $a^2/(a^2+r^2) = \frac{1}{2}(1+\delta/\sqrt{\delta^2+4q^2})$ increases from 0 to 1 as δ ranges from $-\infty$ to $+\infty$. The coefficient rate depends on $|\delta|$, while the directional rate also depends on its sign.

Remark A.2. Consider the regularized objective

$$\mathcal{L}_{\lambda_{\mathrm{wd}}}(a,w) = \mathcal{L}(a,w) + \frac{\lambda_{\mathrm{wd}}}{2}\left(a^2 + \|w\|^2\right), \qquad \lambda_{\mathrm{wd}} \geq 0.$$

The original-parameter flow becomes

$$\dot{a} = -rB - \lambda_{\mathrm{wd}}a, \qquad \dot{w} = -a\mathcal{T} - \lambda_{\mathrm{wd}}w.$$

On every regular neuron, the variables (q,s,δ) satisfy the exact identities

$$\dot{q} = -\mu(q,\delta)B - 2\lambda_{\mathrm{wd}}q, \qquad \dot{s} = -\chi(q,\delta)\mathsf{P}_s^{\perp}\mathcal{T}, \qquad \dot{\delta} = -2\lambda_{\mathrm{wd}}\delta. \tag{A.18}$$

Hence $\delta(t) = \delta(0)e^{-2\lambda_{\mathrm{wd}}t}$. Symmetric weight decay preserves its sign and reduces its magnitude. On the $O(D^{-1})$ time scale for $\delta = +D$, fixed λ_{wd} changes δ by a relative $O(\lambda_{\mathrm{wd}}/D)$ amount. Retaining the unregularized $O(D)$ specialization scale for $\delta = -D$ requires $\lambda_{\mathrm{wd}}D$ to be small.

A.2 Effect of node rescaling beyond two layers

The following proposition gives the effect of node rescaling on the gradient at a fixed parameter state.

Proposition A.3. *Consider a finite feedforward ReLU computation graph with independently parameterized edges and no normalization layers. Fix a hidden node with preactivation*

$$z = u^\top h + b,$$

where h is the vector of activations entering the node, with outgoing parameter vector v and all remaining parameters collected in ξ. Write the incoming affine parameter as $\bar{u} := (u, b)$. The bias-free case omits b from the parameter block. Assume $\bar{u} \neq 0$ and define, for $\rho > 0$,

$$\bar{u}^{(\rho)} := \bar{u}/\rho, \qquad v^{(\rho)} := \rho v, \qquad \Theta^{(\rho)} := g_\rho \Theta := (\bar{u}^{(\rho)}, v^{(\rho)}, \xi).$$

Then the realized network function is invariant:

$$f_{g_\rho \Theta} = f_\Theta. \tag{A.19}$$

Let $\mathcal{L}$ be any differentiable loss that depends on the parameters only through the realized network outputs. At every differentiability point,

$$\nabla_{\bar{u}^{(\rho)}} \mathcal{L}(\Theta^{(\rho)}) = \rho \nabla_{\bar{u}} \mathcal{L}(\Theta), \qquad \nabla_{v^{(\rho)}} \mathcal{L}(\Theta^{(\rho)}) = \rho^{-1} \nabla_v \mathcal{L}(\Theta), \qquad \nabla_\xi \mathcal{L}(\Theta^{(\rho)}) = \nabla_\xi \mathcal{L}(\Theta). \tag{A.20}$$

Under Euclidean gradient flow, write $\widehat{\bar{u}} = \bar{u}/\|\bar{u}\|$ and, when $v \neq 0$, $\widehat{v} = v/\|v\|$. The instantaneous normalized-direction velocities at the two rescaled states satisfy

$$\frac{\mathrm{d}}{\mathrm{d}t} \widehat{\bar{u}^{(\rho)}} \bigg|_{\Theta^{(\rho)}} = \rho^2 \, \frac{\mathrm{d}}{\mathrm{d}t} \widehat{\bar{u}} \bigg|_{\Theta}, \qquad \frac{\mathrm{d}}{\mathrm{d}t} \widehat{v^{(\rho)}} \bigg|_{\Theta^{(\rho)}} = \rho^{-2} \, \frac{\mathrm{d}}{\mathrm{d}t} \widehat{v} \bigg|_{\Theta}. \tag{A.21}$$

Node rescaling changes the relative speeds of the incoming and outgoing directions. If both velocities are nonzero and $\rho \neq 1$, their joint velocity is not proportional to the original velocity.

Proof. Positive homogeneity gives

$$\sigma\big((u/\rho)^\top h + b/\rho\big) = \rho^{-1} \sigma(u^\top h + b).$$

Multiplying every outgoing edge by ρ leaves each downstream preactivation unchanged. The complete network output is therefore unchanged, which proves (A.19).

The identity

$$\mathcal{L}(\bar{u}/\rho, \rho v, \xi) = \mathcal{L}(\bar{u}, v, \xi)$$

holds for all $(\bar{u}, v, \xi)$ in every differentiability region. Differentiating it with respect to the three parameter blocks yields (A.20). Under Euclidean gradient flow,

$$\dot{\bar{u}}^{(\rho)} = -\nabla_{\bar{u}^{(\rho)}} \mathcal{L}(\Theta^{(\rho)}) = \rho \dot{\bar{u}}, \qquad \dot{v}^{(\rho)} = -\nabla_{v^{(\rho)}} \mathcal{L}(\Theta^{(\rho)}) = \rho^{-1} \dot{v}.$$

For any nonzero vector z, writing $\widehat{z} = z/\|z\|$,

$$\frac{\mathrm{d}}{\mathrm{d}t} \frac{z}{\|z\|} = \frac{1}{\|z\|} \left(I - \widehat{z}\widehat{z}^\top\right) \dot{z}.$$

The normalized directions are unchanged by positive scaling, while $\|\bar{u}^{(\rho)}\| = \|\bar{u}\|/\rho$ and $\|v^{(\rho)}\| = \rho\|v\|$. Equation (A.21) follows. The two nonzero blocks are multiplied by distinct factors ρ^2 and ρ^{-2}, which proves the final claim. □

Corollary A.4. *Consider a two-layer scalar-output ReLU neuron with fixed $q = ar > 0$. Start from the balanced representation $a = r = \sqrt{q}$ and apply the reciprocal rescalings ρ and ρ^{-1} with $\rho \geq 1$. The resulting imbalance variables are*

$$\delta_+ = q(\rho^2 - \rho^{-2}), \qquad \delta_- = -q(\rho^2 - \rho^{-2}),$$

and their input-direction speeds are ρ^2 and ρ^{-2} times the balanced speed. Their ratio is therefore ρ^4. If $D = |\delta_\pm|$ and q is fixed, then $\rho^2 = \Theta(D)$ and $\rho^4 = \Theta(D^2)$ as $D \to \infty$.

Remark A.5. Proposition A.3 applies at any differentiability point for any data distribution, loss, and network depth. The global results use additional control along the trajectory and the Gaussian teacher–student structure. They are proved for bias-free two-layer networks.

Remark A.6. Let $\mathcal{U}$ be a deterministic one-step update map and suppose it is exactly equivariant under node rescaling:

$$\mathcal{U}(g_\rho \Theta) = g_\rho \mathcal{U}(\Theta) \qquad \text{for every admissible } (\rho, \Theta).$$

If $\widetilde{\Theta}_0 = g_\rho \Theta_0$, induction gives $\widetilde{\Theta}_n = g_\rho \Theta_n$ for every n. Hence $f_{\widetilde{\Theta}_n} = f_{\Theta_n}$ throughout training. The continuous-time analogue is a vector field satisfying $V(g_\rho \Theta) = Dg_\rho(\Theta) V(\Theta)$, where Dg_ρ denotes the derivative of the rescaling map. An update with this equivariance preserves the common function trajectory of rescaled initializations. Path-SGD is motivated by a rescaling-invariant path geometry (Neyshabur et al., 2015).

A.3 Gaussian ReLU kernel and closed ODE

Let $d \geq 2$ and $x \sim \mathcal{N}(0, I_d)$. Fix a teacher

$$f_\star(x) = q_\star \sigma(s_\star^\top x), \qquad q_\star > 0, \qquad \|s_\star\| = 1, \tag{A.22}$$

and a student

$$f(x) = q\sigma(s^\top x), \qquad q > 0, \qquad \|s\| = 1.$$

Set $c = s^\top s_\star$. Define the arc-cosine kernel

$$\kappa(c) := \mathbb{E}[\sigma(s^\top x)\sigma(s_\star^\top x)] = \frac{\sqrt{1-c^2} + (\pi - \arccos c)c}{2\pi}. \tag{A.23}$$

It satisfies

$$\kappa(-1) = 0, \qquad \kappa(1) = \frac{1}{2}, \qquad \kappa'(c) = \frac{\pi - \arccos c}{2\pi} > 0 \quad (-1 < c \leq 1). \tag{A.24}$$

Closed-form population gradients for Gaussian ReLU teacher–student models are classical (Tian, 2017). Here we combine them with the conserved scale-imbalance variable δ.

Lemma A.7. *For the one-teacher model,*

$$B(s) = \frac{q}{2} - q_\star \kappa(c), \tag{A.25}$$

and

$$\mathsf{P}_s^\perp \mathcal{T}(s) = -q_\star \kappa'(c) \mathsf{P}_s^\perp s_\star. \tag{A.26}$$

The population loss is

$$\mathcal{L}(q, c) = \frac{q^2}{4} + \frac{q_\star^2}{4} - q q_\star \kappa(c). \tag{A.27}$$

Proof. The identities $\mathbb{E}[\sigma(s^\top x)^2] = 1/2$ and (A.23) give (A.25) and (A.27). For a tangent perturbation $v \perp s$,

$$v^\top \mathbb{E}\left[x \mathbf{1}_{\{s^\top x > 0\}} \sigma(s_\star^\top x)\right] = \frac{\mathrm{d}}{\mathrm{d}\epsilon}\bigg|_{\epsilon=0} \kappa\left(\frac{s + \epsilon v}{\|s + \epsilon v\|} \cdot s_\star\right) = \kappa'(c) v^\top s_\star.$$

The self term is parallel to s, so tangential projection yields (A.26). □

Theorem A.8. *Fix a scale imbalance $\delta \in \mathbb{R}$. On the regular region $q > 0$ and $-1 < c < 1$, population gradient flow is exactly*

$$\dot{q} = -\mu(q, \delta)\left(\frac{q}{2} - q_\star \kappa(c)\right), \qquad \dot{c} = \chi(q, \delta) q_\star \kappa'(c)(1 - c^2), \qquad \dot{\delta} = 0. \tag{A.28}$$

The loss dissipates according to

$$\dot{\mathcal{L}} = -\mu(q, \delta)\left(\frac{q}{2} - q_\star \kappa(c)\right)^2 - \frac{\mu(q, \delta) + \delta}{2} q_\star^2 \kappa'(c)^2 (1 - c^2). \tag{A.29}$$

Proof. Insert (A.25)–(A.26) into theorem A.1 and use

$$\|\mathsf{P}_s^\perp s_\star\|^2 = 1 - c^2.$$

□

A.4 Invariant region and global convergence

To compare specialization times, we first need to know that each trajectory reaches the teacher. The coefficient bounds below keep the alignment equation regular and allow us to follow the solution for all time.

Define

$$h(c) := 2q_\star \kappa(c). \tag{A.30}$$

Then the coefficient equation is

$$\dot{q} = -\frac{\mu(q,\delta)}{2}\big(q - h(c)\big). \tag{A.31}$$

Theorem A.9. *Assume*

$$q(0) = q_0 > 0, \qquad -1 < c(0) = c_0 < 1.$$

Let

$$\underline{q} := \min\{q_0, h(c_0)\} > 0, \qquad \overline{q} := \max\{q_0, q_\star\}. \tag{A.32}$$

Then the solution of (A.28) *exists for all* $t \geq 0$ *and satisfies*

$$\underline{q} \leq q(t) \leq \overline{q}, \qquad c_0 \leq c(t) < 1. \tag{A.33}$$

Moreover,

$$c(t) \nearrow 1, \qquad q(t) \longrightarrow q_\star, \qquad \mathcal{L}(q(t), c(t)) \longrightarrow 0. \tag{A.34}$$

Proof. Since h is increasing and $0 < h(c) \leq q_\star$ on $c \in [c_0, 1]$, the vector field at $q = \underline{q}$ points inward:

$$q = \underline{q} \Longrightarrow q - h(c) \leq \underline{q} - h(c_0) \leq 0 \Longrightarrow \dot{q} \geq 0.$$

At $q = \overline{q}$,

$$q = \overline{q} \Longrightarrow q - h(c) \geq \overline{q} - q_\star \geq 0 \Longrightarrow \dot{q} \leq 0.$$

Thus $q \in [\underline{q}, \overline{q}]$. On this compact interval $\chi(q,\delta) > 0$. Since $\kappa'(c) > 0$ on $(-1, 1)$, c is strictly increasing until it reaches 1, and $c = 1$ is invariant.

Let $c_\infty = \lim_t c(t)$. If $c_\infty < 1$, then on the compact rectangle $[\underline{q}, \overline{q}] \times [c_0, c_\infty]$ the coefficient

$$\chi(q,\delta) q_\star \kappa'(c)(1 - c^2)$$

has a strictly positive lower bound, contradicting convergence of $c(t)$. Hence $c_\infty = 1$.

For any $\eta > 0$, eventually $|h(c(t)) - q_\star| < \eta$. Outside $[q_\star - 2\eta, q_\star + 2\eta]$, equation (A.31) then gives a drift of magnitude at least $\eta \min_{[\underline{q},\overline{q}]} \mu/2$ toward this interval. The trajectory enters it in finite time and cannot subsequently leave. Letting $\eta \downarrow 0$ proves $q(t) \to q_\star$, and (A.27) gives $\mathcal{L} \to 0$. □

A.5 Function-preserving rescaling and quadratic timescale separation

Fix $D > 0$ and a common coefficient and direction (q, s). The two scale imbalances $\delta = \pm D$ are realized by

$$a_\pm^2 = \frac{\sqrt{D^2 + 4q^2} \pm D}{2}, \quad r_\pm^2 = \frac{\sqrt{D^2 + 4q^2} \mp D}{2}, \quad a_\pm r_\pm = q. \tag{A.35}$$

They have identical predictors and losses. At the same state (q, c), their coefficient rate factor is identical:

$$\mu(q, +D) = \mu(q, -D) = \sqrt{D^2 + 4q^2}. \tag{A.36}$$

Their directional rate factors are

$$\chi_+(q, D) = \frac{\sqrt{D^2 + 4q^2} + D}{2q}, \qquad \chi_-(q, D) = \frac{\sqrt{D^2 + 4q^2} - D}{2q}, \qquad \chi_+\chi_- = 1. \tag{A.37}$$

For $0 < \varepsilon < 1 - c_0$, define the specialization hitting time

$$T_\delta(\varepsilon) := \inf\{t \geq 0 : c_\delta(t) \geq 1 - \varepsilon\}. \tag{A.38}$$

Let

$$I_{c_0}(\varepsilon) := \int_{c_0}^{1-\varepsilon} \frac{\mathrm{d}c}{1 - c^2} = \frac{1}{2} \log\left(\frac{(2 - \varepsilon)(1 - c_0)}{\varepsilon(1 + c_0)}\right), \tag{A.39}$$

and $k_0 := \kappa'(c_0) > 0$.

Theorem A.10. *Under the assumptions of theorem A.9, let $D \geq \overline{q}$. For the two trajectories with the same initial (q_0, c_0) and scale imbalances $\delta = \pm D$,*

$$\frac{\underline{q}}{Dq_\star} I_{c_0}(\varepsilon) \leq T_{+D}(\varepsilon) \leq \frac{\overline{q}}{Dq_\star k_0} I_{c_0}(\varepsilon), \tag{A.40}$$

$$\frac{2D}{\overline{q}q_\star} I_{c_0}(\varepsilon) \leq T_{-D}(\varepsilon) \leq \frac{2D}{\underline{q}q_\star k_0} I_{c_0}(\varepsilon). \tag{A.41}$$

It follows that

$$\frac{2k_0}{\overline{q}^2} D^2 \leq \frac{T_{-D}(\varepsilon)}{T_{+D}(\varepsilon)} \leq \frac{2}{\underline{q}^2 k_0} D^2. \tag{A.42}$$

In particular, uniformly for $\varepsilon \downarrow 0$,

$$T_{+D}(\varepsilon) = \Theta\left(D^{-1} \log \frac{1}{\varepsilon}\right), \qquad T_{-D}(\varepsilon) = \Theta\left(D \log \frac{1}{\varepsilon}\right). \tag{A.43}$$

Proof. Because c is strictly increasing, we may write $q_\delta(c)$ for the coefficient evaluated when the alignment equals c. Thus

$$T_\delta(\varepsilon) = \int_{c_0}^{1-\varepsilon} \frac{\mathrm{d}c}{\chi(q_\delta(c), \delta) q_\star \kappa'(c)(1 - c^2)}. \tag{A.44}$$

By theorem A.9, $\underline{q} \leq q_\delta(c) \leq \overline{q}$. For $D \geq \overline{q}$,

$$\frac{D}{\overline{q}} \leq \chi_+(q, D) \leq \frac{2D}{\underline{q}}, \tag{A.45}$$

where the upper bound uses

$$\sqrt{D^2 + 4q^2} \leq D + \frac{2q^2}{D}.$$

For $\delta = -D$,

$$\frac{\underline{q}}{2D} \leq \chi_-(q, D) \leq \frac{\overline{q}}{D}, \tag{A.46}$$

using $\chi_- = 2q/(\sqrt{D^2 + 4q^2} + D)$. Also,

$$k_0 \leq \kappa'(c) \leq \frac{1}{2} \qquad (c_0 \leq c \leq 1).$$

Insert these bounds into (A.44) to obtain (A.40)–(A.41). Taking ratios gives (A.42). On $0 < \varepsilon \leq (1 - c_0)/2$, the ratio $I_{c_0}(\varepsilon)/\log(1/\varepsilon)$ is positive and continuous, and extends to $\varepsilon = 0$ with value $1/2$. It therefore has positive lower and finite upper bounds, giving the uniform orders in theorem 3.1. □

Corollary A.11. *Fix $D > 0$, $q_0 > 0$, and $-1 < c_0 < 1$. Let $(q_\pm, c_\pm)$ solve (A.28) with $\delta = \pm D$ and common initial state (q_0, c_0). For any $t_1 > 0$ and strictly increasing C^1 map $\tau : [0, t_1] \to [0, \infty)$ with $\tau(0) = 0$, the identity*

$$(q_+(t), c_+(t)) = (q_-(\tau(t)), c_-(\tau(t))) \qquad (0 \leq t \leq t_1)$$

cannot hold throughout $[0, t_1]$. At a common state off the nullcline $q/2 = q_\star \kappa(c)$ the two vector fields are non-collinear. On the nullcline their phase curves have different second derivatives.

Proof. Put

$$R(q, c) = \frac{q}{2} - q_\star \kappa(c), \qquad A(c) = q_\star \kappa'(c)(1 - c^2) > 0.$$

Because $c_\pm$ are strictly increasing while $-1 < c_\pm < 1$, each trajectory has a local phase representation $q = Q_\pm(c)$. Equation (A.28) gives

$$Q'_\pm(c) = -\frac{\mu(Q_\pm(c), D)}{\chi(Q_\pm(c), \pm D) A(c)} R(Q_\pm(c), c). \tag{A.47}$$

A time reparameterization would give a common phase curve Q on a nontrivial c-interval. Subtracting its two equations in (A.47), and using $\chi(q, +D) \neq \chi(q, -D)$, forces $R(Q(c), c) = 0$

throughout that interval. The phase equation then gives $Q' = 0$, whereas $Q(c) = 2q_\star\kappa(c)$ gives $Q' = 2q_\star\kappa'(c) > 0$, a contradiction.

The local geometry follows from the same equation. Off the nullcline the slopes differ. On it, $Q'_\pm(c_0) = 0$ and differentiation gives

$$Q''_-(c_0) - Q''_+(c_0) = \frac{\mu(q_0, D)D}{q_0(1 - c_0^2)} > 0.$$

□

A.6 Fast and slow large-imbalance limits

The time bounds suggest different limiting descriptions for the two signs of imbalance. Positive imbalance permits alignment on the fast time, whereas negative imbalance produces an initial coefficient relaxation followed by slow alignment. We keep these two stages separate when approximating the coefficient.

The bounds above hold at finite D. The following limits describe the trajectories on the two time scales.

Proposition A.12. *Let*

$$Q_D^+(\tau) = q_{+D}(\tau/D), \qquad C_D^+(\tau) = c_{+D}(\tau/D).$$

For every $0 < \tau_{\max} < \infty$, there exist constants $D_0 < \infty$ and $C_{\tau_{\max}} < \infty$, depending only on $(q_0, c_0, q_\star, \tau_{\max})$, such that for every $D \geq D_0$,

$$\sup_{0\leq\tau\leq\tau_{\max}} \left\|(Q_D^+(\tau), C_D^+(\tau)) - (Q_+(\tau), C_+(\tau))\right\| \leq \frac{C_{\tau_{\max}}}{D^2}, \tag{A.48}$$

where

$$\begin{aligned} \frac{\mathrm{d}Q_+}{\mathrm{d}\tau} &= -\left(\frac{Q_+}{2} - q_\star\kappa(C_+)\right), \quad & \frac{\mathrm{d}C_+}{\mathrm{d}\tau} &= \frac{q_\star}{Q_+}\kappa'(C_+)(1 - C_+^2), \\ Q_+(0) &= q_0, & C_+(0) &= c_0. \end{aligned} \tag{A.49}$$

Proof. On the invariant rectangle $[\underline{q}, \overline{q}] \times [c_0, 1]$,

$$\frac{\mu(q, +D)}{D} = 1 + O(D^{-2}), \qquad \frac{\chi(q, +D)}{D} = \frac{1}{q} + O(D^{-2}).$$

The expansions and the Lipschitz constants of the vector fields are uniform on this rectangle. Continuous dependence and Gronwall's inequality give (A.48). □

Proposition A.13. *Let $h(c) = 2q_\star\kappa(c)$ and $e_D(t) = q_{-D}(t) - h(c_{-D}(t))$. There is a constant $C_{\mathrm{tr}} < \infty$ that depends only on $(q_0, c_0, q_\star)$. For every $D > 0$,*

$$|e_D(t)| \leq |e_D(0)|e^{-Dt/2} + \frac{C_{\mathrm{tr}}}{D^2} \qquad (t \geq 0). \tag{A.50}$$

On the fast scale $\tau = Dt$, $(q_{-D}(\tau/D), c_{-D}(\tau/D))$ converges on compact intervals to

$$\begin{aligned} \frac{\mathrm{d}Q_f}{\mathrm{d}\tau} &= -\left(\frac{Q_f}{2} - q_\star\kappa(c_0)\right), \quad & \frac{\mathrm{d}C_f}{\mathrm{d}\tau} &= 0, \\ Q_f(0) &= q_0, & C_f(0) &= c_0. \end{aligned} \tag{A.51}$$

The limiting solution is therefore

$$Q_f(\tau) = h(c_0) + \big(q_0 - h(c_0)\big)e^{-\tau/2}, \qquad C_f(\tau) = c_0. \tag{A.52}$$

On the slow scale $\sigma = t/D$, define $C_D^-(\sigma) := c_{-D}(D\sigma)$. This trajectory converges uniformly on compact intervals to the solution of

$$\frac{\mathrm{d}C_-}{\mathrm{d}\sigma} = 2q_\star^2\kappa(C_-)\kappa'(C_-)(1 - C_-^2), \qquad C_-(0) = c_0. \tag{A.53}$$

For every $0 < \sigma_0 < \sigma_{\max} < \infty$, there exist constants $D_0 < \infty$ and $C_{\sigma_0,\sigma_{\max}} < \infty$. They depend only on $(q_0, c_0, q_\star, \sigma_0, \sigma_{\max})$, and for every $D \geq D_0$,

$$\sup_{\sigma_0\leq\sigma\leq\sigma_{\max}} \big|q_{-D}(D\sigma) - 2q_\star\kappa(C_-(\sigma))\big| \leq \frac{C_{\sigma_0,\sigma_{\max}}}{D^2}. \tag{A.54}$$

Proof. Set $A(c) = q_\star \kappa'(c)(1-c^2)$. Since $q \in [\underline{q}, \overline{q}]$,

$$|h'(c)| \le q_\star, \qquad 0 \le A(c) \le \frac{q_\star}{2}, \qquad 0 \le \chi(q,-D) \le \frac{\overline{q}}{D}.$$

Differentiating $e_D = q - h(c)$ gives

$$\dot{e}_D = -\frac{\mu(q,-D)}{2} e_D - h'(c)\chi(q,-D)A(c).$$

Since $\mu \ge D$, variation of constants yields (A.50). The fast-scale limit follows from

$$\frac{\mu(q,-D)}{D} = 1 + O(D^{-2}), \qquad \frac{\chi(q,-D)}{D} = O(D^{-2}).$$

For the slow scale, the exact identity

$$|D\chi(q,-D) - q| = \frac{4|q|^3}{\left(\sqrt{D^2+4q^2}+D\right)^2} \le \frac{\overline{q}^3}{D^2} \tag{A.55}$$

shows

$$\frac{\mathrm{d}}{\mathrm{d}\sigma} C_D^-(\sigma) = q_{-D}(D\sigma) A(C_D^-(\sigma)) + O(D^{-2}).$$

The transient term in (A.50) has slow-time integral $O(D^{-2})$, and the remaining tracking error is also $O(D^{-2})$. Gronwall's inequality gives uniform convergence of C_D^- to (A.53). The coefficient statement follows from (A.50). □

B GLOBAL FAST DYNAMICS AND LOCAL GEOMETRY

Let $d \ge 2$, $X \sim \mathcal{N}(0, I_d)$, and let the teacher be

$$f_\star(x) = \sigma(u^\top x), \qquad \|u\| = 1.$$

The student network is

$$f(x) = \sum_{i=1}^{m} q_i(t)\sigma(s_i(t)^\top x), \qquad \|s_i(t)\| = 1.$$

We use population square loss

$$\mathcal{L} = \frac{1}{2}\mathbb{E}\big[(f(X) - f_\star(X))^2\big].$$

All students begin with the same coefficient and feature direction,

$$q_i(0) = \frac{1}{m}, \qquad s_i(0) = s_0, \qquad 0 < \phi = \angle(s_0, u) \le \frac{\pi}{3}. \tag{B.1}$$

Choose one index k and assign

$$\delta_k = +D, \qquad \delta_j = -D \quad (j \ne k),$$

where $\delta_i = a_i^2 - \|w_i\|^2$ is the conserved scale imbalance and $q_i = a_i\|w_i\|$. The underlying positive factors are uniquely reconstructed from

$$a_i^2 = \frac{\sqrt{D^2+4q_i^2}+\delta_i}{2}, \qquad \|w_i\|^2 = \frac{\sqrt{D^2+4q_i^2}-\delta_i}{2}.$$

Thus every choice of the selected index gives the same initial function and loss.

Set

$$M := m - 1.$$

By permutation symmetry, the M students with $\delta = -D$ remain identical. Write

$$q = q_k, \qquad p_{\mathrm{red}} = q_j \ (j \ne k), \qquad P := M p_{\mathrm{red}},$$

Let $e = (s_0 - \cos\phi\, u)/\sin\phi$ and $s_\alpha = \cos\alpha\, u + \sin\alpha\, e$. Define θ and ψ as the signed angles of the direction of unit k and the common redundant direction relative to u in the invariant plane $\mathrm{span}\{u, s_0\}$. The population function is exactly

$$f(x) = q\sigma(s_\theta^\top x) + P\sigma(s_\psi^\top x).$$

Define the angular ReLU kernel

$$g(\alpha) = \frac{\sin\alpha + (\pi - \alpha)\cos\alpha}{2\pi}, \qquad \omega(\alpha) := -g'(\alpha) = \frac{(\pi - \alpha)\sin\alpha}{2\pi}.$$

This is the angular form of the kernel in Appendix A: $g(\alpha) = \kappa(\cos\alpha)$ for $\alpha \in [0, \pi]$. In angular formulas, g is extended evenly and periodically with period 2π. The reduced loss is

$$\mathscr{E}(q, P, \theta, \psi) = \frac{1}{4}(q^2 + P^2 + 1) + qPg(|\psi - \theta|) - qg(|\theta|) - Pg(|\psi|). \tag{B.2}$$

The self and pair interactions of the redundant units sum to $P^2/4$ because their directions coincide.

Let

$$\mu_+(q) = \sqrt{D^2 + 4q^2}, \qquad \mu_-(p_{\mathrm{red}}) = \sqrt{D^2 + 4p_{\mathrm{red}}^2},$$
$$\chi_+(q) = \frac{\mu_+(q) + D}{2q}, \qquad \chi_-(p_{\mathrm{red}}) = \frac{2p_{\mathrm{red}}}{\mu_-(p_{\mathrm{red}}) + D}.$$

Proposition B.1. *When all redundant neurons remain identical, population gradient flow is exactly*

$$\dot{q} = -\mu_+(q)\,\partial_q\mathscr{E}, \qquad \dot{P} = -M\mu_-(P/M)\,\partial_P\mathscr{E}, \tag{B.3}$$

$$\dot{\theta} = -\frac{\chi_+(q)}{q}\,\partial_\theta\mathscr{E}, \qquad \dot{\psi} = -\frac{\chi_-(P/M)}{P}\,\partial_\psi\mathscr{E}, \tag{B.4}$$

where the last expression has its continuous extension at $P = 0$. Moreover,

$$\frac{\mathrm{d}}{\mathrm{d}t}\mathscr{E} = -\mu_+(q)(\partial_q\mathscr{E})^2 - M\mu_-(P/M)(\partial_P\mathscr{E})^2 - \frac{\chi_+(q)}{q}(\partial_\theta\mathscr{E})^2 - \frac{\chi_-(P/M)}{P}(\partial_\psi\mathscr{E})^2. \tag{B.5}$$

Proof. For one redundant unit, the coefficient gradient is

$$\partial_{q_j}\mathcal{L} = \frac{P}{2} + qg(|\psi - \theta|) - g(|\psi|) = \partial_P\mathscr{E}.$$

Therefore $\dot{P} = M\dot{p}_{\mathrm{red}}$ gives the second equation in (B.3). The common angular derivative satisfies

$$\partial_\psi\mathscr{E} = \sum_{j \neq k} \partial_{\psi_j}\mathcal{L},$$

and each summand is $1/M$ of the total. Dividing the individual angular gradient by the individual functional coefficient $p_{\mathrm{red}} = P/M$ gives the final equation in (B.4). The equations for unit k follow from the exact coordinate reduction in theorem A.1. Pairing the four state velocities with the four derivatives of (B.2) yields (B.5). □

For nonzero vectors, let $\angle(z, v)$ be the angle between their normalized directions. Define the Gaussian ReLU covariance

$$K(z, v) = \mathbb{E}_{x\sim\gamma_d}[[z^\top x]_+[v^\top x]_+] = \|z\|\|v\|g(\angle(z, v)).$$

When $\|v\| = 1$ and $\gamma = \angle(z, v)$,

$$\nabla_z K(z, v) = \frac{(\pi - \gamma)v + \sin\gamma\, z/\|z\|}{2\pi}. \tag{B.6}$$

The kernel is jointly C^2 when both vectors are nonzero. At the angular endpoints, $g(t) = \frac{1}{2}\cos t + t^3/(6\pi) + O(t^5)$ and $g(\pi - t) = t^3/(6\pi) + O(t^5)$. Near parallel or antiparallel vectors, the nonsmooth angular contribution is of the form $\|h_\perp\|^3$, where $h_\perp$ is a transverse displacement. This function is C^2 at zero. Differentiation and homogeneity give the uniform bounds

$$\|\nabla_{zz}^2 K(z, v)\| \le C\frac{\|v\|}{\|z\|}, \qquad \|\nabla_{zv}^2 K(z, v)\| \le C. \tag{B.7}$$

B.1 Scaled Cartesian dynamics

The angular equation contains a factor that becomes singular when the selected coefficient vanishes. Cartesian coordinates retain the same feature information without dividing by that coefficient. The following linear scaling also identifies which variables move on the fast time.

Let (a_+, w_+) denote the selected unit and (a_-, w_-) any of the M identical redundant units. Introduce the linear coordinates

$$b = \frac{a_+}{\sqrt{D}}, \qquad z = \sqrt{D}\, w_+, \qquad p = M\sqrt{D}\, a_-, \qquad v = \frac{w_-}{\sqrt{D}}, \qquad \Xi = (b, z, p, v). \tag{B.8}$$

Here p is the scaled total redundant amplitude. The functional coefficients and directions are recovered by

$$q = b\|z\|, \quad P = p\|v\|, \quad s_+ = z/\|z\|, \quad s_- = v/\|v\|. \tag{B.9}$$

These maps are smooth wherever $z, v \neq 0$, including at $p = P = 0$. In these coordinates the predictor and loss contain no D:

$$f_\Xi(x) = b[z^\top x]_+ + p[v^\top x]_+, \qquad \mathcal{E}(\Xi) = \tfrac{1}{2}\|f_\Xi - f_\star\|^2_{L^2(\gamma_d)}. \tag{B.10}$$

The raw gradient of each redundant unit is $1/M$ of the gradient obtained by differentiating their common parameters. The chain rule therefore gives, in fast time $\tau = Dt$ (primes denote τ derivatives), the exact equations

$$\Xi' = -\mathsf{M}_D \nabla \mathcal{E}(\Xi), \qquad \mathsf{M}_D = \mathrm{diag}(D^{-2}, I_d, M, (MD^2)^{-1} I_d). \tag{B.11}$$

The initial point is

$$\Xi_D(0) = (1, s_0/m, M/m, s_0) + O(D^{-2}),$$

with an error uniform in $m \geq 2$. At $D = \infty$, $b = 1$ and $v = s_0$ are frozen, while $z' = -\nabla_z \mathcal{E}$ and $p' = -M\partial_p \mathcal{E}$. Writing $z = q s_\theta$ recovers (B.13).

B.2 The limiting system

The limit of (B.11) has $b = 1$ and $v = s_0$. With $z = q s_\theta$ and $p = P$, define

$$E_M(q, P, \theta) := \frac{1}{4}(q^2 + P^2 + 1) + qPg(\phi - \theta) - qg(|\theta|) - Pg(\phi), \tag{B.12}$$

for $\phi - \pi \leq \theta \leq \phi$. The fast system is

$$q' = -\partial_q E_M, \qquad P' = -M\partial_P E_M, \qquad \theta' = -q^{-2}\partial_\theta E_M, \tag{B.13}$$

with initial condition

$$q(0) = q_0 := \frac{1}{M+1}, \qquad P(0) = P_0 := \frac{M}{M+1}, \qquad \theta(0) = \phi. \tag{B.14}$$

The energy landscape is independent of M. The subscript records the metric weight in the flow:

$$\frac{\mathrm{d}}{\mathrm{d}\tau} E_M = -(\partial_q E_M)^2 - M(\partial_P E_M)^2 - q^{-2}(\partial_\theta E_M)^2. \tag{B.15}$$

B.3 Arc-cosine kernel inequalities

Lemma B.2. *For $0 < \phi < \pi$,*

$$g(x)g(\phi - x) - \omega(x)\omega(\phi - x) \leq \tfrac{1}{2} g(\phi) \qquad (0 \leq x \leq \phi), \tag{B.16}$$

$$g(x)g(\phi + x) + \omega(x)\omega(\phi + x) \leq \tfrac{1}{2} g(\phi) \qquad (0 \leq x \leq \pi - \phi). \tag{B.17}$$

The function

$$F_\phi(x) = \big(\tfrac{1}{4} - g(\phi + x)^2\big)\omega(x) - \big(\tfrac{1}{2} g(\phi) - g(\phi + x)g(x)\big)\omega(\phi + x) \tag{B.18}$$

satisfies $F_\phi(0) = 0$ and $F_\phi(x) > 0$ for $0 < x \leq \pi - \phi$.

Proof. The difference between the right and left sides of (B.16) is symmetric about $\phi/2$, vanishes at zero, and has derivative $(\phi - 2x)\sin x \sin(\phi - x)/(2\pi^2) \geq 0$ on $[0, \phi/2]$. For (B.17), define

$$R_\phi(x) = \tfrac{1}{2}g(\phi) - g(x)g(\phi + x) - \omega(x)\omega(\phi + x).$$

Then $R_\phi(0) = 0$ and $R'_\phi(x) = (2\pi - \phi - 2x)\sin x \sin(\phi + x)/(2\pi^2) \geq 0$. Next put

$$H(t) = \tfrac{1}{4} - g(t)^2 - \omega(t)^2, \qquad H(0) = 0, \quad H'(t) = \frac{(\pi - t)\sin^2 t}{\pi^2} > 0 \quad (0 < t < \pi). \tag{B.19}$$

Since $F_\phi(x) = H(\phi + x)\omega(x) - R_\phi(x)\omega(\phi + x)$,

$$\frac{\mathrm{d}}{\mathrm{d}x}\frac{F_\phi(x)}{\omega(\phi + x)} = \frac{\phi \sin x \sin(\phi + x)}{2\pi^2} + H(\phi + x)\frac{\mathrm{d}}{\mathrm{d}x}\frac{\omega(x)}{\omega(\phi + x)} > 0. \tag{B.20}$$

Here $(\log \omega)''(t) = -(\pi - t)^{-2} - \sin^{-2} t < 0$, so the ratio $\omega(x)/\omega(\phi + x)$ is strictly increasing for $0 < x < \pi - \phi$. The value at zero and $F_\phi(\pi - \phi) = \omega(\pi - \phi)/4 > 0$ complete the proof. □

B.4 An invariant region for the limiting system

Before locating the limiting state, we show that the trajectory stays in a bounded region and that the selected vector cannot reach zero. A component of that vector toward the teacher supplies the needed lower bound.

Put $\alpha = \phi - \theta$ and $Z = qg(\alpha) - g(\phi)$. The duplicate initial point belongs to

$$\mathcal{C}_\phi = \{0 < q \leq 1,\ 0 \leq P \leq 1,\ \phi - \pi \leq \theta \leq \phi,\ Z \leq 0\}, \tag{B.21}$$

since $Z(0) = 1/(2m) - g(\phi) < 0$ for $0 < \phi \leq \pi/3$. In this range, $g(\phi) \geq g(\pi/3) = 1/6 + \sqrt{3}/(4\pi) > 1/4$. The selected vector acquires a positive component toward the teacher immediately, which prevents it from reaching zero.

Lemma B.3. *The region $\mathcal{C}_\phi$ is forward invariant. For the duplicate initialization, let*

$$e_\perp = \frac{u - \cos\phi\, s_0}{\sin\phi}, \qquad y = e_\perp^\top z = q\sin(\phi - \theta).$$

Then, for all $\tau \geq 0$,

$$q(\tau) \geq y(\tau) \geq \frac{\phi \sin\phi}{\pi}(1 - e^{-\tau/2}). \tag{B.22}$$

On any fixed compact interval $I_\phi \subset (0, \pi/3]$, a constant $c_I > 0$ independent of m satisfies

$$q(\tau) \geq c_I\big(m^{-1} + \min\{\tau, 1\}\big). \tag{B.23}$$

Proof. Until the first exit from (B.21), the Cartesian fast equation is

$$z' = -z/2 + \nabla_z K(z, u) - P\nabla_z K(z, s_0).$$

Using (B.6) and projecting onto $e_\perp$ gives

$$y' = -\frac{y}{2} + \frac{(\pi - |\theta|)\sin\phi + (\sin|\theta| - P\sin\alpha)\sin\alpha}{2\pi}. \tag{B.24}$$

The numerator is smallest at $P = 1$. For $0 \leq \theta \leq \phi$, its excess over $(\pi - \phi)\sin\phi$ is

$$\alpha\sin\phi + \sin\alpha(\sin\theta - \sin\alpha) \geq \sin\alpha(\sin\phi + \sin\theta - \sin\alpha) \geq 0.$$

For $\phi - \pi \leq \theta \leq 0$, put $t = \pi - \phi + \theta$. Its excess over $\phi\sin\phi$ is

$$t\sin\phi + \sin t(\sin(\phi + t) - \sin t) \geq \sin t(\sin\phi + \sin(\phi + t) - \sin t) \geq 0.$$

The final signs follow, respectively, from

$$\sin\phi + \sin\theta - \sin(\phi - \theta) = 4\sin(\theta/2)\cos(\phi/2)\cos((\phi - \theta)/2),$$

and

$$\sin\phi + \sin(\phi + t) - \sin t = 4\sin(\phi/2)\cos(t/2)\cos((\phi + t)/2).$$

Thus $y' \geq -y/2 + \phi \sin\phi/(2\pi)$. Since $y(0) = 0$, this proves (B.22) on the stopped interval and excludes $q = 0$ at any positive first-exit time. For a general point in $\mathcal{C}_\phi$, the same argument uses $y(0) \geq 0$.

All other faces are inward-pointing. At $q = 1$ and $P = 1$,

$$\partial_q E_M = \tfrac{1}{2} + Pg(\alpha) - g(|\theta|) \geq 0, \qquad \partial_P E_M = \tfrac{1}{2} + qg(\alpha) - g(\phi) > 0.$$

At $P = 0$, $P' = -MZ \geq 0$. The angular velocities at $\theta = \phi$ and $\theta = \phi - \pi$ are $-\omega(\phi)/q < 0$ and $\omega(\pi - \phi)/q > 0$. At $Z = 0$, direct differentiation gives

$$Z' = -\tfrac{1}{2}g(\phi) - P\{g(\alpha)^2 + \omega(\alpha)^2\} + g(|\theta|)g(\alpha) - \mathrm{sgn}(\theta)\omega(|\theta|)\omega(\alpha) \leq 0,$$

by (B.16) for $\theta \geq 0$ and (B.17) for $\theta < 0$. These signs close the first-exit argument.

Finally, $q' \geq -1$ on the region, so $q(\tau) \geq m^{-1} - \tau$. For $\tau \leq 1/(2m)$ this gives $q \geq 1/(2m)$. For later times use (B.22) and $1 - e^{-\tau/2} \geq (1 - e^{-1/2}) \min\{\tau, 1\}$. Reducing c_I proves (B.23). □

B.5 Local loss and gradient bounds

Near the selected representation, small loss must control the coefficients and selected direction. The Hessian gives this control through the three feature variations associated with these coordinates. We first state the elementary estimate used to turn Hessian positivity into loss and gradient bounds.

Lemma B.4. *Let $L(\xi;\zeta)$ be a nonnegative C^2 scalar loss near $\xi = 0$, with ζ in a compact parameter set. Suppose $L(0;\zeta) = 0$, $\nabla_\xi L(0;\zeta) = 0$, and $\nabla_\xi^2 L(0;\zeta) \succeq \lambda I$ uniformly, for $\lambda > 0$. On a uniform neighborhood of zero,*

$$\frac{\lambda}{4}\|\xi\|^2 \leq L(\xi;\zeta) \leq C\|\xi\|^2, \qquad \|\nabla_\xi L(\xi;\zeta)\|^2 \geq \lambda L(\xi;\zeta). \tag{B.25}$$

Proof. Hessian continuity gives $\nabla_\xi^2 L \succeq \lambda I/2$ on a smaller ball. Taylor's theorem gives the quadratic bounds. Strong convexity, compared with $\xi = 0$, gives $L(\xi) \leq \langle \nabla L(\xi), \xi\rangle - \lambda\|\xi\|^2/4 \leq \|\nabla L(\xi)\|^2/\lambda$. □

Near the selected solution, fix the common redundant angle ψ and write $\xi = (q - 1, \theta, P)$ and $\rho = \|\xi\|$. The transverse Hessian at zero is the explicit Gram matrix

$$G_\psi = \begin{pmatrix} 1/2 & 0 & g(\psi) \\ 0 & 1/2 & \omega(\psi) \\ g(\psi) & \omega(\psi) & 1/2 \end{pmatrix}, \qquad \lambda_{\min}(G_\psi) = \tfrac{1}{2} - \sqrt{g(\psi)^2 + \omega(\psi)^2} > 0. \tag{B.26}$$

Here $0 < \psi < \pi$, as holds in the neighborhoods considered below. Its columns correspond to the selected coefficient, the selected tangent direction, and the redundant coefficient. Positivity follows from the same $H(\psi) > 0$ used in the global kernel argument.

Lemma B.5. *There are $r_0, \sigma_0, c, C > 0$, uniformly in m and ϕ on any compact $I_\phi \subset (0, \pi/3]$, such that on $\mathcal{U}_0 = \{\rho < r_0,\ |\psi - \phi| < \sigma_0\}$,*

$$c\rho^2 \leq \mathscr{E} \leq C\rho^2, \tag{B.27}$$

$$(\partial_q \mathscr{E})^2 + (\partial_P \mathscr{E})^2 + q^{-2}(\partial_\theta \mathscr{E})^2 \geq c\mathscr{E}, \tag{B.28}$$

$$|\partial_\psi \mathscr{E}| \leq C|P|\sqrt{\mathscr{E}}. \tag{B.29}$$

Proof. The loss is C^2 in the regular coordinates, vanishes at $\xi = 0$, and has Hessian G_ψ there. Its least eigenvalue is uniformly positive for ψ in a small compact neighborhood of I_ϕ. Apply lemma B.4. The upper and lower bounds on q make the resulting gradient bound equivalent to (B.28). Finally,

$$\partial_\psi \mathscr{E} = P\langle f - f_\star, \mathbf{1}_{\{s_\psi^\top x > 0\}} (\partial_\psi s_\psi)^\top x \rangle,$$

where the brackets denote the $L^2(\gamma_d)$ inner product. Cauchy–Schwarz proves (B.29). □

B.6 Uniform fast convergence

We now combine the invariant region with the local estimates. The kernel inequalities exclude other stationary points, and compactness gives a positive gradient-to-loss ratio away from the selected state. The local bound extends this control to the state itself.

Theorem B.6. *For every $m \geq 2$ and $0 < \phi \leq \pi/3$, the duplicate-initialized fast flow exists globally and converges to $(1, 0, 0)$. On each compact $I_\phi \subset (0, \pi/3]$, there are $c_I, C_I > 0$, independent of m, such that for $\tau \geq 1$,*

$$E_M(\tau) \leq E_M(1)e^{-c_I(\tau-1)}, \qquad \rho(\tau) \leq C_I e^{-c_I(\tau-1)/2}. \tag{B.30}$$

For $0 < r < r_0$, the flow therefore reaches $\rho \leq r$ by fast time

$$T(r) = 1 + \frac{2}{c_I} \log \frac{C_I}{r}, \tag{B.31}$$

after enlarging C_I to at least one.

Proof. By lemma B.3, all trajectories for $\tau \geq 1$ lie in the joint compact set

$$\mathcal{K}_I = \{(\phi, q, P, \theta) : \phi \in I_\phi,\ (q, P, \theta) \in \mathcal{C}_\phi,\ q \geq \underline{q}_I\}, \quad \underline{q}_I = \min_{\phi \in I_\phi} \frac{\phi \sin \phi}{\pi}(1 - e^{-1/2}) > 0.$$

Write

$$S = P\omega(\phi - \theta) + \operatorname{sgn}(\theta)\omega(|\theta|), \qquad \Gamma = (\partial_q E_M)^2 + (\partial_P E_M)^2 + S^2.$$

Since $M \geq 1$, (B.15) gives $E_M' \leq -\Gamma$. The only zero of Γ is $(1, 0, 0)$: for $\theta > 0$, $S > 0$. At $\theta = 0$, $S = 0$ gives $P = 0$ and then $\partial_q E_M = 0$ gives $q = 1$. For $\theta = -x < 0$, coefficient stationarity requires

$$P = \frac{\frac{1}{2}g(\phi) - g(\phi + x)g(x)}{\frac{1}{4} - g(\phi + x)^2}.$$

The strict sign $F_\phi(x) > 0$ from lemma B.2 implies $P\omega(\phi + x) < \omega(x)$, so $S \neq 0$, including at $x = \pi - \phi$.

Near the target, lemma B.5 gives $\Gamma \geq cE_M$ and $E_M \geq c\rho^2$. Outside this neighborhood, Γ has a positive minimum on $\mathcal{K}_I$ and E_M is bounded above. Also, a zero-loss state is stationary, so the target is the only zero of E_M. Its positive minimum outside the neighborhood and the upper bound on ρ give the second inequality below. Combining the local and exterior bounds yields, throughout $\mathcal{K}_I$,

$$\Gamma \geq c_I E_M, \qquad E_M \geq c_I \rho^2. \tag{B.32}$$

Integrating $E_M' \leq -c_I E_M$ and using the uniformly bounded initial loss proves (B.30) and (B.31). □

B.7 Continuation through a loss barrier

The local estimates apply only while the trajectory remains in their neighborhood. The following lemma makes that continuation explicit: loss decrease prevents exit in the controlled coordinates, and an integrated velocity bound prevents exit in the remaining coordinates.

Lemma B.7. *Let a trajectory have local coordinates (ξ, ζ) in $\mathcal{U} = \{\|\xi\| < r,\ \zeta \in \operatorname{int} \mathcal{S}\}$, where $\mathcal{S}$ is compact. Assume the vector field is regular on the closure and that, for constants $c, \lambda, A > 0$, while the trajectory remains there,*

$$L \geq c\|\xi\|^2, \qquad -\dot{L} \geq \lambda D L, \qquad \|\dot{\zeta}\| \leq AL/D.$$

Put $\ell_\partial = \inf_{\|\xi\| = r,\, \zeta \in \mathcal{S}} L$ and $d_0 = \operatorname{dist}(\zeta(t_0), \partial\mathcal{S}) > 0$. If the initial state is in $\mathcal{U}$ and

$$L(t_0) < \ell_\partial, \qquad \frac{AL(t_0)}{\lambda D^2} < d_0, \tag{B.33}$$

then the trajectory stays in $\mathcal{U}$ for all $t \geq t_0$, with $L(t) \leq L(t_0)e^{-\lambda D(t-t_0)}$, $\xi(t) \to 0$, and $\int_{t_0}^{\infty} \|\dot{\zeta}\|\, \mathrm{d}t \leq AL(t_0)/(\lambda D^2)$.

Proof. Up to the first exit time, the loss is at most $L(t_0)$ and the total ζ motion is at most $AL(t_0)/(\lambda D^2)$. The two strict inequalities exclude the transverse and remaining boundary faces, respectively. Regularity excludes any other finite obstruction. Thus the first exit is infinite. Integration and coercivity give the asserted estimates. □

C FINITE IMBALANCE, PERTURBATIONS, AND GRADIENT DESCENT

Let $d \geq 2$ and fix a compact angle interval $I_\phi \subset (0, \pi/3]$. We retain the Gaussian teacher $f_\star = [u^\top x]_+$ and designate unit 1. All constants below may depend on I_ϕ; generic positive constants c, C may change between estimates. Unqualified function norms and inner products refer to $L^2(\gamma_d)$, and $X \sim \gamma_d$ denotes a Gaussian input. Sums indexed by j run over $j = 2, \dots, m$ unless stated otherwise. Appendix B shows that the limiting trajectory reaches the selected representation. Here we control how a finite imbalance, a perturbation of the initial state, or a discrete update changes that trajectory. The redundant amplitudes enter the selected unit's equation through their sum. Separating this sum from the differences between amplitudes avoids treating all interactions as independent sources of error.

C.1 A COMMON STABILITY ESTIMATE

Allow $\delta_1 > 0$ and $\delta_j < 0$ for $j \geq 2$, and measure initial departure from the duplicate state by

$$\Delta_0 = \max_i \big(|q_i(0) - m^{-1}| + \|s_i(0) - s_0\|\big) + \max_i \left| \frac{|\delta_i(0)|}{D} - 1 \right|. \tag{C.1}$$

Use the unaggregated linear coordinates

$$b = a_1/\sqrt{D}, \quad z = \sqrt{D}\, w_1, \qquad p_j = \sqrt{D}\, a_j, \quad v_j = w_j/\sqrt{D} \quad (j \geq 2). \tag{C.2}$$

They represent $f = b[z^\top x]_+ + \sum_{j\geq 2} p_j [v_j^\top x]_+$. With $\mathcal{E} = \frac{1}{2}\|f - f_\star\|^2_{L^2(\gamma_d)}$ and $\tau = Dt$, the exact flow is

$$b' = -D^{-2}\partial_b \mathcal{E}, \quad z' = -\nabla_z \mathcal{E}, \qquad p_j' = -\partial_{p_j}\mathcal{E}, \quad v_j' = -D^{-2}\nabla_{v_j}\mathcal{E}. \tag{C.3}$$

Imbalance perturbations enter only through the initial point of this system.

Put $M = m - 1$, $K_0(z) = K(z, s_0)$, and $K_\star(z) = K(z, u)$. Freezing $b = 1$ and $v_j = s_0$ gives the reference fast flow, whose redundant amplitudes equal P_∞/M. Write its solution as $x_\infty = (z_\infty, P_\infty/M, \dots, P_\infty/M)$, where $z_\infty = q_\infty s_{\theta_\infty}$ and $(q_\infty, P_\infty, \theta_\infty)$ solves (B.13). Its total amplitude couples to the selected vector, while differences between redundant amplitudes remain constant. The following estimate separates these two effects.

Choose $r_\star < \min_{I_\phi} \phi/8$ small enough for lemma B.5. Fix $0 < r_{\mathrm{cap}} < \min\{r_\star/4, 1/4\}$ so that the loss at $\rho \leq 2r_{\mathrm{cap}}$ is below the boundary loss at $\rho = r_\star/2$, and put $\mathcal{V}_{\mathrm{cap}} = \{\rho < r_{\mathrm{cap}}\}$. Using theorem B.6, set

$$T_\star = T(r_{\mathrm{cap}}), \qquad T_{\max} = T_\star + 2. \tag{C.4}$$

Let $\tau^\infty_{\mathrm{cap}}$ be the reference trajectory's first entry time into $\overline{\mathcal{V}_{\mathrm{cap}}}$. Then $\tau^\infty_{\mathrm{cap}} \leq T_\star$. The projection lower bound also gives

$$\int_0^{T_{\max}} \frac{\mathrm{d}\tau}{q_\infty(\tau)} \leq C_I(1 + \log m). \tag{C.5}$$

Lemma C.1. *Let $x = (z, p_2, \dots, p_m)$ be an absolutely continuous solution of*

$$z' = -z/2 - P\nabla K_0(z) + \nabla K_\star(z) + r_z, \qquad p_j' = -(P/2 + K_0(z) - g(\phi)) + r_j, \quad P = \sum_{j\geq 2} p_j. \tag{C.6}$$

Here r_z and r_j are additive forcing terms. Set $e_z = \|z - z_\infty\|$, $e_P = P - P_\infty$, and $c_j = p_j - P/M$. Suppose

$$\|r_z\| + \max_j |r_j| \leq C_0\epsilon, \qquad e_z(0) + |e_P(0)|/M + \max_j |c_j(0)| \leq C_0\epsilon.$$

Up to any stopping time $T \leq T_{\max}$ on which $e_z < q_\infty/2$, there are $C > 0$ and $\beta > 1/2$, independent of m, such that

$$\sup_{\tau\leq T} e_z(\tau) \leq Cm^\beta\epsilon, \qquad \sup_{\tau\leq T}\max_j |c_j(\tau)| \leq C\epsilon, \tag{C.7}$$

$$|e_P(\tau)| \leq CM\epsilon e^{-M\tau/2} + Cm^\beta\epsilon, \qquad \sup_{\tau\leq T}\|x - x_\infty\| \leq Cm^\beta\epsilon. \tag{C.8}$$

The constants depend only on I_ϕ and C_0.

Proof. Let $\bar{r} = M^{-1}\sum_j r_j$. Summing and subtracting the amplitude equations gives the exact scalar and zero-sum equations

$$e_P' = -\tfrac{M}{2}e_P - M(K_0(z) - K_0(z_\infty)) + M\bar{r}, \qquad c_j' = r_j - \bar{r}. \tag{C.9}$$

Gaussian Cauchy–Schwarz gives $\|\nabla K_0\| \le 1/2$: for any unit a, both $\mathbb{E}[(a^\top X)^2 \mathbf{1}_{\{z^\top X>0\}}]$ and $\mathbb{E}[[s_0^\top X]_+^2]$ equal $1/2$. Variation of constants therefore yields

$$|e_P(\tau)| \le e^{-M\tau/2}|e_P(0)| + \tfrac{M}{2}\int_0^\tau e^{-M(\tau-s)/2} e_z(s)\, \mathrm{d}s + 2C_0\epsilon. \tag{C.10}$$

To estimate the selected vector, split its interaction error as $P\nabla K_0(z) - P_\infty \nabla K_0(z_\infty) = P_\infty(\nabla K_0(z) - \nabla K_0(z_\infty)) + e_P \nabla K_0(z)$. The first term is nonnegative when paired with $z - z_\infty$, since K_0 is convex and $P_\infty \ge 0$. The segment joining z to z_∞ stays at norm at least $q_\infty/2$ under the stopping condition. The Hessian bound (B.7) controls the teacher term on this segment. Together with the scalar equation, these observations give, almost everywhere,

$$e_z' \le -\tfrac{1}{2}e_z + \frac{C}{q_\infty}e_z + \tfrac{1}{2}|e_P| + C_0\epsilon,$$
$$(|e_P|)' \le -\tfrac{M}{2}|e_P| + \tfrac{M}{2}e_z + MC_0\epsilon.$$

The weight $1/M$ cancels the coupling and decay terms directly:

$$\mathcal{V} = e_z + \frac{|e_P|}{M}, \qquad \mathcal{V}' \le \frac{C}{q_\infty}e_z + 2C_0\epsilon \le \frac{C}{q_\infty}\mathcal{V} + 2C_0\epsilon. \tag{C.11}$$

Since $\mathcal{V}(0) \le C_0\epsilon$, Gronwall and (C.5) yield

$$\sup_{\tau \le T} \mathcal{V}(\tau) \le Cm^\beta \epsilon. \tag{C.12}$$

This bounds e_z. The convolution kernel in (C.10) has mass at most one, giving the stated e_P bound. Direct integration of $c_j' = r_j - \bar{r}$ bounds the zero-sum part. Finally, the orthogonal decomposition

$$\|x - x_\infty\|^2 = e_z^2 + e_P^2/M + \sum_j c_j^2 \tag{C.13}$$

proves the last assertion after increasing β to exceed $1/2$. Only the reference total amplitude was required to be nonnegative. □

Proposition C.2. *There are $A, a, C > 0$ and $\nu > 3/2$, depending only on I_ϕ, such that, if*

$$D \ge D_0(m) := Am^{\nu/2}, \qquad \Delta_0 \le \varepsilon_m := am^{-\nu}, \tag{C.14}$$

the exact flow, with active state $x_D = (z, p_2, \dots, p_m)$, satisfies

$$\sup_{\tau \le T_{\max}} \|x_D - x_\infty\| \le Cm^{\nu-1}(\Delta_0 + D^{-2}), \tag{C.15}$$

$$\sup_{\tau \le T_{\max}} \left(|b-1| + \max_j \|v_j - s_0\|\right) \le C(\Delta_0 + D^{-2}). \tag{C.16}$$

At the reference capture time $\tau_{\mathrm{cap}}^\infty$, the redundant output error obeys

$$\left\| \sum_j p_j [v_j^\top \cdot]_+ - P_\infty [s_0^\top \cdot]_+ \right\|_{L^2} \le Cm^{\nu-1}(\Delta_0 + D^{-2}). \tag{C.17}$$

The exponent is not optimized.

Proof. Put $\epsilon = \Delta_0 + D^{-2}$ and use the initial reconstruction $a_i^2 = (\sqrt{\delta_i^2 + 4q_i^2} + \delta_i)/2$, $\|w_i\|^2 = (\sqrt{\delta_i^2 + 4q_i^2} - \delta_i)/2$. It gives the initial assumptions of lemma C.1 and slow-coordinate error $O(\epsilon)$. Choose $a \le 1/4$ to keep all initial coefficients positive. Stop when $e_z = q_\infty/2$ or at a boundary of

$$\|z\| < 2, \quad \sum_j |p_j| < 2, \quad 1/2 < b < 3/2, \quad 1/2 < \|v_j\| < 3/2. \tag{C.18}$$

The bounded residual gives $|b'| \le CD^{-2}$ and $\|v_j'\| \le CD^{-2}|p_j|$, proving (C.16) up to the stop. The kernel bounds then put the active equations in the form (C.6), with

$$\|r_z\| + \max_j |r_j| \le C\epsilon. \tag{C.19}$$

Apply lemma C.1 and set $\nu = \beta + 1$. Small a and large A make $e_z < c_I/(4m) < q_\infty/2$. Moreover,

$$\sum_j |p_j| \le P_\infty + |e_P| + M \max_j |c_j| \le 1 + C(M + m^\beta)\epsilon < 2.$$

The upper bound on z and the slow-coordinate bounds exclude all remaining faces, closing the stop on $[0, T_{\max}]$.

The reference needs at least fast time $1/2$ to reach $q_\infty \ge 3/4$, because $q_\infty(0) \le 1/2$ and $q_\infty' \le 1/2$ in the invariant region. Thus $\tau_{\mathrm{cap}}^\infty \ge 1/2$, and $Me^{-M\tau_{\mathrm{cap}}^\infty/2} \le C$. Equation (C.8) gives $|e_P(\tau_{\mathrm{cap}}^\infty)| \le Cm^\beta\epsilon$. The ReLU Lipschitz bound and $\sum |p_j| < 2$ give

$$\left\| \sum_j p_j [v_j^\top \cdot]_+ - P_\infty [s_0^\top \cdot]_+ \right\|_{L^2} \le C(|e_P| + \epsilon),$$

which proves (C.17). □

C.2 Specialization and coefficient decay

The comparison estimate brings the finite-imbalance trajectory into the neighborhood constructed in Appendix B. The remaining task is to show that it stays there. The loss barrier and redundant-direction motion bound then give convergence of the individual coefficients.

Theorem C.3. *Under exact duplicate initialization, $D \ge D_0(m)$ guarantees entry into the symmetric neighborhood $\mathcal{V}_m = \{\rho < r_\star/2,\ |\psi - \phi| < r_\star/2\}$ by time $t_{\mathrm{cap}} \le T_\star/D$. Thereafter,*

$$\mathcal{L}(t) \le \mathcal{L}(t_{\mathrm{cap}}) e^{-cD(t - t_{\mathrm{cap}})}, \qquad q_1(t) \to 1, \quad s_1(t) \to u, \quad q_j(t) \to 0\ (j \ge 2),$$

and each redundant direction satisfies

$$\int_{t_{\mathrm{cap}}}^\infty \|\dot{s}_j(t)\|\, \mathrm{d}t \le \frac{C\mathcal{L}(t_{\mathrm{cap}})}{(m-1)D^2}.$$

The constants are uniform in m and $\phi \in I_\phi$. Permuting the unique positive imbalance permutes the selected neuron without changing the initial predictor.

Proof. At limiting capture, q_∞ is near one and $P_\infty^2 + \theta_\infty^2 + (q_\infty - 1)^2 \le r_{\mathrm{cap}}^2$. The selected coefficient and direction are Lipschitz functions of the scaled coordinates there. Equations (C.8) and (C.16) control the total functional coefficient and redundant angle at capture. Thus proposition C.2, with adjusted a, A, gives entry into the inner symmetric neighborhood with a strict loss gap to the transverse boundary of $\mathcal{V}_m$. Exact dissipation and lemma B.5 give $-\dot{\mathcal{L}} \ge cD\mathcal{L}$ on the outer neighborhood. Moreover,

$$|\dot{\psi}| \le \frac{|\partial_\psi \mathscr{E}|}{(m-1)D} \le \frac{C\mathcal{L}}{(m-1)D}.$$

Increase A so that the resulting total motion is smaller than the directional margin. Apply lemma B.7 with $\xi = (q - 1, \theta, P)$ and $\zeta = \psi$. Regularity follows from $q > 1/2$ and the conserved imbalance signs, which keep all input norms nonzero. This proves invariance, contraction, and $P \to 0$. Symmetry gives $q_j = P/(m-1) \to 0$, and permutation equivariance proves the last assertion. □

Theorem C.4. *Under (C.14), define*

$$\mathcal{W}_{\mathrm{sel}}^{(i)} = \{q_i \ge 1 - r_\star,\ \angle(s_i, u) \le r_\star,\ \|\sum_{j \ne i} q_j [s_j^\top \cdot]_+\|_{L^2} \le r_\star\}.$$

The perturbed flow first enters $\mathcal{W}_{\mathrm{sel}}^{(1)}$ at a time $t_{\mathrm{sel}} \le T_\star/D$. No other unit enters its own selection region before that time, and $\angle(s_j(t_{\mathrm{sel}}), u) > r_\star$ for every $j \ge 2$.

Proof. At limiting capture, (C.15) controls the selected coefficient and direction errors, and (C.17) controls the redundant output directly in Gaussian L^2. The choices of a, A in proposition C.2 make these errors smaller than the fixed selection margins. This proves entry by $T_\star/D$. Throughout this interval, (C.16) keeps every redundant direction close to s_0, at teacher angle greater than $r_\star$. Such a unit cannot enter its own selection region. □

C.3 Local identifiability and asymptotic convergence

With nonduplicate students, a small sum of redundant features need not make each coefficient small. To identify the coefficients, we impose a quantitative independence condition on the local feature variations. Their Gram matrix gives both the condition and the resulting loss bound. On the selected branch $q_1 > 0$, set

$$\mathbf{h} = q_1 s_1 - u, \qquad \beta = (q_2, \dots, q_m), \qquad \mathbf{s} = (s_2, \dots, s_m), \qquad \beta_j = q_j \quad (j \geq 2).$$

For $\|\mathbf{h}\| < 1/2$, $q_1 = \|u + \mathbf{h}\| > 1/2$ and the selected contribution equals $[(u + \mathbf{h})^\top x]_+$. With $\mathbf{s}$ fixed, the transverse loss is

$$L_{\mathbf{s}}(\mathbf{h}, \beta) = \tfrac{1}{2} \left\| [(u + \mathbf{h})^\top \cdot]_+ - [u^\top \cdot]_+ + \sum_{j \geq 2} \beta_j [s_j^\top \cdot]_+ \right\|^2_{L^2(\gamma_d)}. \tag{C.20}$$

Its Hessian at $(\mathbf{h}, \beta) = 0$ is

$$\mathcal{G}_{\mathbf{s}} = \begin{pmatrix} \frac{1}{2} I_d & T_{\mathbf{s}} \\ T_{\mathbf{s}}^\top & K_{\mathbf{s}} \end{pmatrix}, \qquad (K_{\mathbf{s}})_{ij} = K(s_i, s_j), \tag{C.21}$$

Here $K_{\mathbf{s}}$ is indexed by the redundant units $i, j \in \{2, \dots, m\}$, and the column of $T_{\mathbf{s}}$ corresponding to unit j is

$$t_j = \mathbb{E}[X \mathbf{1}_{\{u^\top X > 0\}} [s_j^\top X]_+] = \frac{\sin \vartheta_j \, u + (\pi - \vartheta_j) s_j}{2\pi}, \qquad \vartheta_j = \angle(u, s_j).$$

We require a uniform bound

$$\mathcal{G}_{\mathbf{s}} \succeq \lambda_0 I, \qquad \lambda_0 > 0, \tag{C.22}$$

on a compact neighborhood $\mathcal{S}$ of redundant directions. Positive definiteness is equivalent to the Schur-complement condition $K_{\mathbf{s}} - 2T_{\mathbf{s}}^\top T_{\mathbf{s}} \succ 0$. Thus redundant features remain distinguishable after removing the part explained by first-order changes of the selected unit. The redundant coefficient map is then injective, so nonzero coefficients cannot cancel as functions.

Lemma C.5. *If the unoriented hyperplanes $u^\perp, s_2^\perp, \dots, s_m^\perp$ are pairwise distinct, then $\mathcal{G}_{\mathbf{s}} \succ 0$. For fixed m, d, a uniform positive separation between these hyperplanes gives a uniform positive lower eigenvalue on compact direction sets.*

Proof. A vector in the Gram nullspace would give

$$\mathbf{1}_{\{u^\top x > 0\}} \mathbf{h}^\top x + \sum_{j \geq 2} \beta_j [s_j^\top x]_+ = 0$$

almost everywhere. Gaussian full support makes this identity valid on every open cell cut out by the hyperplanes. At a generic point of $u^\perp$, away from the other hyperplanes, its jump is $\mathbf{h}^\top x$. Thus $\mathbf{h} = \alpha u$. The identity is now a continuous piecewise linear combination of ReLU features. Across a generic point of $s_j^\perp$, the gradient jump is $\beta_j s_j$, so $\beta_j = 0$. Across $u^\perp$ it is αu, so $\alpha = 0$. Continuity of the Gram matrix and compactness prove the uniform assertion. □

For unit directions, the separation in this lemma is measured by $\min\{\angle(s_i, s_j), \pi - \angle(s_i, s_j)\}$, also including u. Its quantitative size matters: two redundant directions at angle ϑ give, by the coefficient difference vector,

$$\lambda_{\min}(\mathcal{G}_{\mathbf{s}}) \leq \tfrac{1}{2} - g(\vartheta) = \tfrac{1}{4}\vartheta^2 + O(\vartheta^3). \tag{C.23}$$

Consequently the local radius, entry loss, and direction margin below must still be matched to λ_0 near duplicate directions. Under exact duplication, symmetry first combines the redundant coefficients into one effective feature. The resulting Gram matrix, restricted to the teacher plane, is precisely (B.26).

Lemma C.6. *Assume (C.22) uniformly on $\mathcal{S}$. For some $0 < r < 1/2$, uniformly on $\|(\mathbf{h}, \beta)\| \le r$ and $\mathbf{s} \in \mathcal{S}$,*

$$L_{\mathbf{s}} \ge \frac{\lambda_0}{4}(\|\mathbf{h}\|^2 + \|\beta\|^2), \qquad \|\nabla_{\mathbf{h},\beta} L_{\mathbf{s}}\|^2 \ge \lambda_0 L_{\mathbf{s}}. \tag{C.24}$$

If the conserved imbalances satisfy $\delta_1 \ge D/2$ and $\delta_j \le -D/2$ for $j \ge 2$, the population flow also obeys

$$-\dot{\mathcal{L}} \ge \tfrac{1}{2}\lambda_0 D\mathcal{L}, \qquad \|\dot{\mathbf{s}}\|_2 \le \frac{C}{\sqrt{\lambda_0} D}\mathcal{L} \tag{C.25}$$

while it remains in this neighborhood.

Proof. The scalar loss (C.20) is C^2 by the regularity of K on nonzero inputs. Its value and gradient vanish at the origin, and its Hessian there is (C.21). Apply lemma B.4 to obtain (C.24) uniformly on $\mathcal{S}$.

For the selected vector $q_1 s_1$, the coefficient–direction equations give the mobility matrix

$$\mathsf{A}_1 = \mu_1 s_1 s_1^\top + a_1^2(I - s_1 s_1^\top) \succeq (D/2)I.$$

Each redundant coefficient has mobility $\mu_j \ge D/2$. The remaining directional contributions to dissipation are nonnegative, so $-\dot{\mathcal{L}} \ge (D/2)\|\nabla_{\mathbf{h},\beta} L_{\mathbf{s}}\|^2$. This proves the first bound in (C.25). For a redundant unit,

$$|\chi_j| \le 2|q_j|/D, \qquad \|\mathcal{T}(s_j)\| \le \sqrt{2\mathcal{L}}.$$

Taking the Euclidean product norm and using $\sum_{j\ge 2} q_j^2 \le 4\mathcal{L}/\lambda_0$ proves the second bound. □

Theorem C.7. *Suppose a trajectory with $\delta_1 \ge D/2$ and $\delta_j \le -D/2$ enters a neighborhood of the selected solution at time t_0, with $q_1(t_0) > 0$. Let $\mathcal{S}$ be a compact neighborhood of $\mathbf{s}(t_0)$ satisfying (C.22), with $d_0 = \operatorname{dist}(\mathbf{s}(t_0), \partial\mathcal{S}) > 0$. Interiors and boundaries are relative to the product of direction spheres. Choose $r < 1/2$ as in lemma C.6 and assume that the entry state lies in $\mathcal{U} = \{\|(\mathbf{h}, \beta)\| < r,\ \mathbf{s} \in \operatorname{int}\mathcal{S}\}$. Define*

$$\partial_\perp \mathcal{U} = \{\|(\mathbf{h}, \beta)\| = r,\ \mathbf{s} \in \mathcal{S}\}, \qquad \ell_\partial = \inf_{\partial_\perp \mathcal{U}} \mathcal{L} \ge \lambda_0 r^2/4. \tag{C.26}$$

For a universal constant $C > 0$, if

$$\mathcal{L}(t_0) < \ell_\partial, \qquad \frac{C\mathcal{L}(t_0)}{\lambda_0^{3/2} D^2} \le d_0/2, \tag{C.27}$$

then the trajectory stays in $\mathcal{U}$ and

$$\mathcal{L}(t) \le \mathcal{L}(t_0) e^{-\lambda_0 D(t-t_0)/2}, \qquad q_1(t) \to 1, \quad s_1(t) \to u, \quad q_j(t) \to 0\ (j \ge 2).$$

In addition,

$$\int_{t_0}^{\infty} \|\dot{\mathbf{s}}(t)\|_2 \, \mathrm{d}t \le \frac{C\mathcal{L}(t_0)}{\lambda_0^{3/2} D^2}. \tag{C.28}$$

Proof. Apply lemma B.7 with $\xi = (\mathbf{h}, \beta)$ and $\zeta = \mathbf{s}$, using lemma C.6. The loss gap excludes transverse exit and the stopped motion bound excludes directional exit. The coordinate field is regular on the closure: $\|u + \mathbf{h}\| > 1/2$, and the negative imbalance keeps every redundant input norm positive, including when its coefficient is zero. Coercivity then gives $(\mathbf{h}, \beta) \to 0$: the selected neuron specializes to the teacher, and every redundant coefficient converges to zero. □

C.4 Discrete full-batch gradient descent

For gradient descent, the imbalance is no longer conserved. We use a linear scaling of the original parameters so that each discrete update remains an exact Euler step. This permits a trajectory comparison before selection and a direct bound on the accumulated imbalance drift afterward.

Consider population gradient descent in the original parameters,

$$\Theta_{n+1} = \Theta_n - \eta \nabla_\Theta \mathcal{L}(\Theta_n), \qquad \eta = h/D. \tag{C.29}$$

Write $\mathcal{L}_n = \mathcal{L}(\Theta_n)$ and $\delta_{i,n} = a_{i,n}^2 - \|w_{i,n}\|^2$. Duplicate symmetry is preserved at every step. Because the coordinate change (B.8) is linear, this update is exactly

$$\Xi_{n+1} = \Xi_n - h\mathsf{M}_D \nabla \mathcal{E}(\Xi_n). \tag{C.30}$$

Thus the same Cartesian vector field governs both gradient flow and gradient descent. For comparison with the fast limit, write

$$Y_n = (q_n, P_n, \theta_n, \psi_n), \qquad Y_\infty(\tau) = (q_\infty(\tau), P_\infty(\tau), \theta_\infty(\tau), \phi). \tag{C.31}$$

Lemma C.8. *Let $T_c = T_\star + 1$ and let $\nu > 3/2$ be as in proposition C.2. There are $A, a > 0$, depending only on I_ϕ, such that*

$$D \geq Am^{\nu/2}, \qquad 0 < h \leq am^{-(\nu+1)} \tag{C.32}$$

imply

$$\max_{nh \leq T_c} \|Y_n - Y_\infty(nh)\| \leq C_m(h + D^{-2}). \tag{C.33}$$

All these steps, including their full Euler segments, stay in a compact regular Cartesian neighborhood. On these steps,

$$\mathcal{L}_{n+1} \leq \mathcal{L}_n - \frac{\eta}{2}\|\nabla_\Theta \mathcal{L}(\Theta_n)\|^2. \tag{C.34}$$

Proof. Interpolate the exact update (C.30) linearly and unaggregate its redundant amplitudes. On each interval $[nh, (n+1)h]$, the interpolant has the constant derivative of the finite-D field at its left endpoint. Stop at the boundaries used in proposition C.2, including $e_z = q_\infty/2$. Up to the stop, the residual is bounded, so each active component moves by at most Ch per step. The slow derivatives still satisfy $|b'| \leq CD^{-2}$ and $\|v_j'\| \leq CD^{-2}|p_{j,n}|$. Their total changes are $O(D^{-2})$ on $[0, T_{\max}]$. Also, the total amplitude moves by at most Cmh within a step, and $\|z\| \geq c_I/(2m)$. The kernel Hessian bound (B.7) therefore gives the componentwise defect

$$\|r_z\| + \max_j |r_j| \leq C(D^{-2} + mh) \tag{C.35}$$

from (C.6), almost everywhere along the interpolant. The same bound applies to a partial stopped segment: its endpoints and joining lines remain in the enlarged regular neighborhood when mh is small compared with the $1/m$ radial margin.

Apply lemma C.1 with $\epsilon = D^{-2} + mh$. Condition (C.32) makes $\epsilon \leq cm^{-\nu}$ for a sufficiently small c, so the stopping argument in proposition C.2 closes for complete Euler segments as well. In particular, $h \leq 1$ ensures that segments starting before T_c end before $T_{\max}$. The change to coefficient and direction coordinates is Lipschitz on the resulting regular neighborhood for fixed m, proving (C.33). The aggregate estimate also gives the direct capture-output bound (C.17), with $\Delta_0 + D^{-2}$ replaced by $D^{-2} + mh$.

In the collapsed coordinates, the only growing kernel Hessian factor on this neighborhood is $1/\|z\|$, so $\|\nabla^2 \mathcal{E}\| \leq C_I m$. Put $Q_n = \langle \nabla \mathcal{E}(\Xi_n), \mathsf{M}_D \nabla \mathcal{E}(\Xi_n) \rangle$, where the brackets are the Euclidean inner product in scaled parameter space. Since $\|\mathsf{M}_D\| \leq m$ for $D \geq 1$, $\|\mathsf{M}_D \nabla \mathcal{E}\|^2 \leq mQ_n$. Taylor's theorem on the complete segment gives a decrease of at least $hQ_n/2$ when $h \leq c_I/m^2$, which follows from (C.32) after reducing a. The chain rule, with redundant multiplicity included, gives

$$DQ_n = \|\nabla_\Theta \mathcal{L}(\Theta_n)\|^2, \tag{C.36}$$

and hence (C.34). □

C.4.1 Local descent and direction motion

Choose nested neighborhoods of the selected zero-loss manifold,

$$\mathcal{V}_{\text{in}} \Subset \mathcal{V}_{\text{mid}} \Subset \mathcal{V}_{\text{out}}. \tag{C.37}$$

Here $\Subset$ means that the closure of each set lies in the interior of the next. In scaled coordinates these are product neighborhoods in the transverse variables $(q-1, P, \theta)$ and the remaining variables $(b, \|v\|, \psi)$, centered at $(0,0,0)$ and $(1,1,\phi)$, respectively. All three have $b, \|z\|, \|v\|$ bounded away from zero. Choose the transverse radii $r_{\text{in}} < r_{\text{mid}} < r_{\text{out}}$ so that

$$\sup_{\overline{\mathcal{V}_{\text{in}}}} \mathcal{E} \leq Cr_{\text{in}}^2 < cr_{\text{mid}}^2 \leq \inf_{\partial_\perp \mathcal{V}_{\text{mid}}} \mathcal{E}. \tag{C.38}$$

Here $\partial_\perp$ denotes the boundary in the transverse variables, and the last infimum allows all remaining coordinates in the closure of their middle neighborhood. The limiting path enters strictly inside $\mathcal{V}_{\text{in}}$ by time $T_\star$ after choosing the smaller entry set in proposition C.2. All radii remain independent of m, D.

Lemma C.9. *On $\mathcal{V}_{\text{out}}$, uniformly for $m \geq 2$ and $D \geq 1$,*

$$\|\nabla^2 \mathcal{E}\| \leq K_I, \qquad Q := \langle \nabla \mathcal{E}, \mathsf{M}_D \nabla \mathcal{E} \rangle \geq c_I \mathcal{E}. \tag{C.39}$$

Furthermore,

$$|\partial_b \mathcal{E}| \leq C_I \sqrt{\mathcal{E}}, \qquad \|\nabla_v \mathcal{E}\| \leq C_I |p| \sqrt{\mathcal{E}} \leq C_I \mathcal{E}. \tag{C.40}$$

Proof. The collapsed loss and the chosen neighborhood do not depend on m. The Gaussian Cartesian kernel has bounded second derivatives when its arguments are nonzero and bounded, giving the Hessian estimate. For the gradient bound, $q = b\|z\|$ and $P = p\|v\|$ give

$$\|\nabla_z \mathcal{E}\|^2 + M(\partial_p \mathcal{E})^2 = b^2(\partial_q \mathscr{E})^2 + \|z\|^{-2}(\partial_\theta \mathscr{E})^2 + M\|v\|^2(\partial_P \mathscr{E})^2 \geq c_I \mathcal{E}$$

by lemma B.5. These terms are contained in Q. If $R = f_\Xi - f_\star$, then

$$\partial_b \mathcal{E} = \langle R, [z^\top x]_+ \rangle, \qquad \nabla_v \mathcal{E} = p\, \mathbb{E}[R \mathbf{1}_{\{v^\top X > 0\}} X].$$

Cauchy–Schwarz and $|p| \leq C_I \sqrt{\mathcal{E}}$, from local coercivity and the lower bound on $\|v\|$, prove (C.40). □

Lemma C.10. *Suppose an iterate enters $\mathcal{V}_{\text{in}}$ at index n_{cap} and the preceding steps satisfy (C.34). For $D \geq A_I$ and $0 < h \leq a_I/m$, with fixed $A_I, a_I > 0$, all later iterates stay in $\mathcal{V}_{\text{mid}}$, and*

$$\mathcal{L}_{n+1} \leq (1 - c_I h)\mathcal{L}_n \qquad (n \geq n_{\text{cap}}). \tag{C.41}$$

The common redundant direction $s_{-,n} = v_n/\|v_n\|$ satisfies

$$\sum_{n \geq n_{\text{cap}}} \|s_{-,n+1} - s_{-,n}\| \leq \frac{C_I}{D^2} \mathcal{L}_{n_{\text{cap}}}. \tag{C.42}$$

For every original neuron,

$$\sum_{n \geq 0} |\delta_{i,n+1} - \delta_{i,n}| \leq 2\eta \mathcal{L}_0 = \frac{2h}{D} \mathcal{L}_0. \tag{C.43}$$

Proof. Work up to the first exit of a complete Euler segment from $\mathcal{V}_{\text{mid}}$. An iterate in its closure has a fixed positive distance from the complement of $\mathcal{V}_{\text{out}}$. The vector field has norm at most $C_I m$ there, so $h \leq c_I/m$ places its entire segment in $\mathcal{V}_{\text{out}}$. For $0 \leq t \leq 1$, Taylor's theorem and lemma C.9 give

$$\begin{aligned} \mathcal{E}(\Xi_n - th\mathsf{M}_D \nabla \mathcal{E}(\Xi_n)) &\leq \mathcal{L}_n - thQ_n + \tfrac{1}{2} K_I t^2 h^2 m Q_n \\ &\leq \mathcal{L}_n - \tfrac{1}{2} thQ_n \qquad (hK_I m \leq 1). \end{aligned} \tag{C.44}$$

At $t = 1$ this proves (C.41), with a smaller c_I. At every intermediate t the loss is at most $\mathcal{L}_n$, so the strict gap (C.38) excludes a transverse first exit.

The other coordinates have small total motion directly from (C.30) and (C.40):

$$|b_{n+1} - b_n| \leq \frac{C_I h}{D^2} \sqrt{\mathcal{L}_n}, \qquad \|v_{n+1} - v_n\| \leq \frac{C_I h}{MD^2} \mathcal{L}_n. \tag{C.45}$$

Summing the geometric loss bound, up to the stopped segment and including any fraction of its last step, yields

$$\sum |b_{n+1} - b_n| \leq C_I D^{-2} \sqrt{\mathcal{L}_{n_{\text{cap}}}}, \qquad \sum \|v_{n+1} - v_n\| \leq \frac{C_I}{MD^2} \mathcal{L}_{n_{\text{cap}}}.$$

For large D, these are smaller than the fixed margins from the inner neighborhood to the b, radial, and directional boundaries of the middle neighborhood. Thus no first exit is possible. Normalization of v is Lipschitz on this regular set, proving (C.42).

Finally, homogeneity gives $a_i \partial_{a_i} \mathcal{L} = w_i^\top \nabla_{w_i} \mathcal{L}$ and hence the exact identity

$$\delta_{i,n+1} - \delta_{i,n} = \eta^2 \left[(\partial_{a_i} \mathcal{L}_n)^2 - \|\nabla_{w_i} \mathcal{L}_n\|^2 \right]. \tag{C.46}$$

The pre-selection descent and (C.44), together with (C.36), give $\mathcal{L}_{n+1} \leq \mathcal{L}_n - \eta \|\nabla_\Theta \mathcal{L}_n\|^2/2$ at every step. Taking absolute values in (C.46) and telescoping the loss therefore proves (C.43). □

Theorem C.11. *For $m \geq 2$ and $\phi \in I_\phi$, duplicate-initialized gradient descent with $\eta = h/D$ satisfying (C.32) (with adjusted A, a) obeys (C.33) and enters $\mathcal{V}_{\text{in}}$ within $\lceil T_c/h \rceil$ iterations. All subsequent iterates remain in $\mathcal{V}_{\text{mid}}$ and satisfy (C.41), (C.42), and (C.43). Moreover, $q_n \to 1$, $P_n \to 0$, and $\theta_n \to 0$: the designated unit specializes and every redundant coefficient vanishes.*

Proof. At reference entry, the interpolation estimates in lemma C.8 bound the selected and total-amplitude errors by $Cm^{\nu-1}(D^{-2} + mh)$. Rounding this time up to the next iterate adds at most Cmh, since the active component velocities are bounded. Equation (C.32) makes these errors smaller than the fixed entry margins, uniformly in m, giving entry before T_c. The same condition implies $D \geq A_I$ and $h \leq a_I/m$. Apply lemma C.10. Geometric loss decay and local coercivity imply $(q_n, P_n, \theta_n) \to (1, 0, 0)$. Exact exchange symmetry gives $q_{j,n} = P_n/M \to 0$ for every redundant unit. All constants are uniform over the fixed compact interval I_ϕ. □

D Experimental and numerical validation

We use the four-coordinate states $Y_\infty = (q_\infty, P_\infty, \theta_\infty, \phi)$ for the limit and $Y_D = (q_D, P_D, \theta_D, \psi_D)$ for the exact symmetric flow. We compare these with training in the original parameters (a_i, w_i). Here q and θ describe the selected neuron, while P and ψ describe the total coefficient and common direction of the redundant neurons. Original-parameter training uses either the population loss or the empirical loss on independent samples $x_n \sim \mathcal{N}(0, I_2)$ with labels $y_n = [u^\top x_n]_+$.

Write s_{sel} and s_{red} for the selected and common redundant directions, and use fast time $\tau = Dt$. Let t_{cap} be the theoretical entry time in theorem 3.3, and let $T > 0$ be a fixed fast-time horizon. For every fixed multiplicity, the theoretical rates are

$$\begin{aligned} t_{\text{cap}} &= \Theta_m(D^{-1}), \\ \int_{t_{\text{cap}}}^{\infty} \|\dot{s}_{\text{red}}(t)\| \, \mathrm{d}t &= O_m(D^{-2}), \\ \sup_{0 \leq \tau \leq T} \|Y_D(\tau) - Y_\infty(\tau)\|_\infty &= O_{m,T}(D^{-2}). \end{aligned} \tag{D.1}$$

Here $\|\cdot\|_\infty$ is the maximum absolute coordinate. We test these rates using the measured selection time $\widehat{t}_{\text{sel}}$, defined as the first time (D.3) holds, with $\widehat{\tau}_{\text{sel}} = D\widehat{t}_{\text{sel}}$. This numerical threshold is distinct from the neighborhood used to define t_{cap}. Table D.1 lists the single-teacher configurations; the duplicate population and empirical studies use $\phi = \pi/4$. Population direction lengths are integrated from $\widehat{t}_{\text{sel}}$ to the stated endpoint, testing the predicted order of remaining motion. For each empirical width and seed, samples are nested across N. Paired-bootstrap intervals resample the seed rows jointly across sample sizes, using 10,000 replicates.

Table D.1: Single-teacher experimental settings. The population and empirical studies use duplicate initialization.

Study	Configuration
Single neuron	$q_\star = 1$, $q_0 = 0.6$, $c_0 = 0$, $\varepsilon = 0.05$; $D \in \{2, 4, 8, 16, 32, 64, 128\}$
Population	$m \in \{2, 4, 8, 16, 32\}$, $D \in \{8, 16, 32, 64, 128\}$; endpoint $\tau = 80$
Empirical	$m \in \{2, 4, 8, 16\}$, $D = 32$, $N \in \{512, 1024, 2048, 4096, 8192\}$; 20 coupled seeds, endpoint $\tau = 60$

For the exact trajectories $(q_D^\pm, c_D^\pm)$ in physical time, the single-unit errors in figure D.1(a) are

$$\begin{aligned} e_+(D) &= \sup_{0 \leq \tau \leq 4} \left\| (q_D^+(\tau/D), c_D^+(\tau/D)) - (Q^+(\tau), C^+(\tau)) \right\|_2, \\ e_-(D) &= \sup_{0 \leq \sigma \leq 2} \left| c_D^-(D\sigma) - C^-(\sigma) \right|. \end{aligned} \tag{D.2}$$

Here (Q^+, C^+) and C^- are the fast and slow limiting solutions of (A.49) and (A.53). Both intervals include zero. The slow error measures alignment only. Its coefficient approximation applies after

the initial layer, on $\sigma \geq \sigma_0 > 0$, as in Proposition A.13. At zero the slow coefficient is $2\kappa(0) = 1/\pi$, whereas $q_0 = 0.6$. Supremum errors are approximated on 1,001 equally spaced time points, using $D \in \{4, 8, 16, 32, 64, 128\}$. The population and finite-sample sweeps use the selection criterion below, where $q_{\mathrm{sel}} = q$, $P_{\mathrm{red}} = P$, and $\theta_{\mathrm{sel}} = \theta$:

$$q_{\mathrm{sel}} \geq 0.90, \qquad |P_{\mathrm{red}}| \leq 0.10, \qquad |\theta_{\mathrm{sel}}| \leq 0.10. \tag{D.3}$$

To compare directions in proportion to their contribution to the predictor, we use the coefficient-weighted state

$$\mathsf{X}(\tau) = \big(q(\tau),\, P(\tau),\, q(\tau)s_{\mathrm{sel}}(\tau),\, P(\tau)s_{\mathrm{red}}(\tau)\big) \in \mathbb{R}^6, \tag{D.4}$$

Let X_N and $\mathsf{X}_{\mathrm{pop}}$ be this state along the empirical and population flows, respectively. Their root-mean-square trajectory error is

$$\mathrm{Err}_N = \left[\frac{1}{6T}\int_0^T \|\mathsf{X}_N(\tau) - \mathsf{X}_{\mathrm{pop}}(\tau)\|_2^2 \,\mathrm{d}\tau\right]^{1/2}, \tag{D.5}$$

with $T = 60$. Weighting each direction by its coefficient prevents a nearly vanishing unit from dominating the trajectory metric.

D.1 Agreement across dynamical descriptions

Figure 2 compares the four descriptions at $(m, D, N) = (8, 32, 8192)$. Panel (c) plots the selected angle $|\theta|$ and the magnitude $|P \sin\psi|$ of the redundant coefficient-weighted direction perpendicular to the teacher. Panel (d) uses the two terms in (5), divided by D; for empirical training, their expectations are sample averages. All four descriptions select the $+D$ unit and show the redundant coefficient decaying as the loss falls. The direction metrics distinguish this coefficient decay from alignment of the redundant feature. The large-D limit follows the coefficient-weighted coordinates with an $O(D^{-2})$ correction for redundant-direction motion. The reduced and original-parameter systems agree to numerical precision across the population grid. The full empirical system likewise agrees with its symmetry-reduced form, and permutation symmetry keeps the duplicate units identical.

D.2 Scaling with imbalance magnitude and multiplicity

Panels (a)–(b) of figure 3, panel (b) of figure D.1, and table D.2 test the three predictions in equation (D.1). The fitted exponents remain stable from $m = 2$ through $m = 32$.

Table D.2: Population scaling laws. All regressions use $D \in \{8, 16, 32, 64, 128\}$. Here ℓ_{red} denotes the post-selection path length of the common redundant direction, "limit slope" refers to the last error in (D.1), and R^2 is the coefficient of determination of the log–log regression.

m	$\widehat{t}_{\mathrm{sel}}$ slope	R^2	ℓ_{red} slope	R^2	limit slope	R^2
2	-1.0019	> 0.9999	-1.9991	> 0.9999	-1.9981	> 0.9999
4	-0.9982	> 0.9999	-1.9965	> 0.9999	-1.9983	> 0.9999
8	-0.9967	> 0.9999	-1.9945	> 0.9999	-1.9926	> 0.9999
16	-0.9951	> 0.9999	-1.9964	> 0.9999	-1.9941	> 0.9999
32	-0.9964	> 0.9999	-1.9907	> 0.9999	-1.9946	> 0.9999

m	$D\widehat{t}_{\mathrm{sel}}$	$D^2\ell_{\mathrm{red}}$
2	30.24	1.59×10^{-2}
4	24.16	3.78×10^{-3}
8	21.84	1.44×10^{-3}
16	20.82	6.40×10^{-4}
32	20.34	3.03×10^{-4}

Both rescaled quantities remain bounded over the tested multiplicities, consistent with the fixed-m predictions.

D.3 FINITE-SAMPLE TRAJECTORY SCALING

Table D.3 shows an approximately linear relation between log trajectory error and log sample size at each width. Every empirical run satisfies the selection criterion (D.3).

Table D.3: Finite-sample full-trajectory scaling at $D = 32$ using 20 coupled seeds. The error is defined in (D.5).

m	slope	R^2	mean $\sqrt{N}\ \mathrm{Err}_N$	selection rate
2	-0.4908	0.9970	0.2316	20/20 at every N
4	-0.4482	0.9589	0.2724	20/20 at every N
8	-0.4718	0.9753	0.2356	20/20 at every N
16	-0.4942	0.9858	0.2552	20/20 at every N

The slopes are consistent with $N^{-1/2}$ over the tested fixed-dimensional configurations. The weighting in (D.4) keeps the metric well defined as the redundant coefficient approaches zero.

D.4 ROBUSTNESS AND DISCRETE TRAINING

The angle, perturbation, multiplicity, and discrete-training studies use the selection criterion

$$q_1 > 0.90, \qquad s_1^\top u > \cos(0.10), \qquad \|f_{\mathrm{red}}\|_{L^2} < 0.10. \tag{D.6}$$

We denote the first time this criterion holds by $\widehat{t}_{\mathrm{sel}}^{(f)}$. Unlike (D.3), it thresholds the redundant output norm. For identical redundant directions, $\|f_{\mathrm{red}}\|_{L^2} = |P|/\sqrt{2}$, so the two numerical thresholds differ. Table D.4 lists the perturbation and discrete configurations. For the near-duplicate study, starting from $q_i = 1/m$, we draw a vector z with independent standard-Gaussian entries, center and rescale it so that $\max_i |z_i| = 1$, and set $q_i = (1/m)(1 + \epsilon z_i)$. Initial-angle perturbations and imbalance multipliers are sampled independently and uniformly from $[-\epsilon, \epsilon]$ and $[1 - \epsilon, 1 + \epsilon]$, respectively, with the signs of δ_i preserved. Each trajectory is integrated to $\tau = 50$.

The discrete comparison records the maximum componentwise error in $(q_i, q_i s_i)_{i=1}^m$ between discrete and continuous original-parameter trajectories.

Table D.4: Perturbation and discrete-training configurations. The first four studies end at $\tau = 50$, and the direction-motion study ends at $\tau = 80$.

Study	Configuration
Initial angle	$D = 32$, $m \in \{2, 8, 32\}$; eight equally spaced angles in $[0.77, 0.90]$
Width	$D = 64$, $\phi = \pi/4$, $m \in \{2, 4, 8, 16, 32, 64, 128, 256\}$
Near duplicates	$(m, D, \phi) = (8, 32, \pi/4)$; $\epsilon \in \{10^{-4}, 10^{-3}, 10^{-2}, 0.03, 0.05\}$, ten seeds each
Discrete error	$(m, D, \phi) = (8, 64, \pi/4)$; $h \in \{0.025, 0.05, 0.1, 0.2, 0.4\}$
Direction motion	$m = 8$, $\phi = \pi/4$, $D \in \{16, 24, 32, 48, 64, 96, 128\}$; $h \in \{0.2, 0.1, 0.05\}$, $\eta = h/D$

D.5 POST-SELECTION DIRECTION MOTION UNDER GRADIENT DESCENT

The discrete-time analysis predicts D^{-2} scaling of the redundant-direction path length. For the configurations in table D.4, we sum the direction increments after the selection criterion (D.6) is first met. Table D.5 reports the fitted exponents and rescaled path lengths.

Table D.5: Post-selection path length of the common redundant direction under original-parameter gradient descent. The exponent is fitted against D.

h	fitted D exponent	R^2	mean $D^2 \sum_n \|s_{n+1} - s_n\|$
0.20	-1.9953	> 0.9999	1.9029×10^{-3}
0.10	-1.9952	> 0.9999	1.9294×10^{-3}
0.05	-1.9975	> 0.9999	1.9305×10^{-3}

For each fixed h, the rescaled path is nearly constant across the tested values of D, consistent with the predicted D^{-2} scaling.

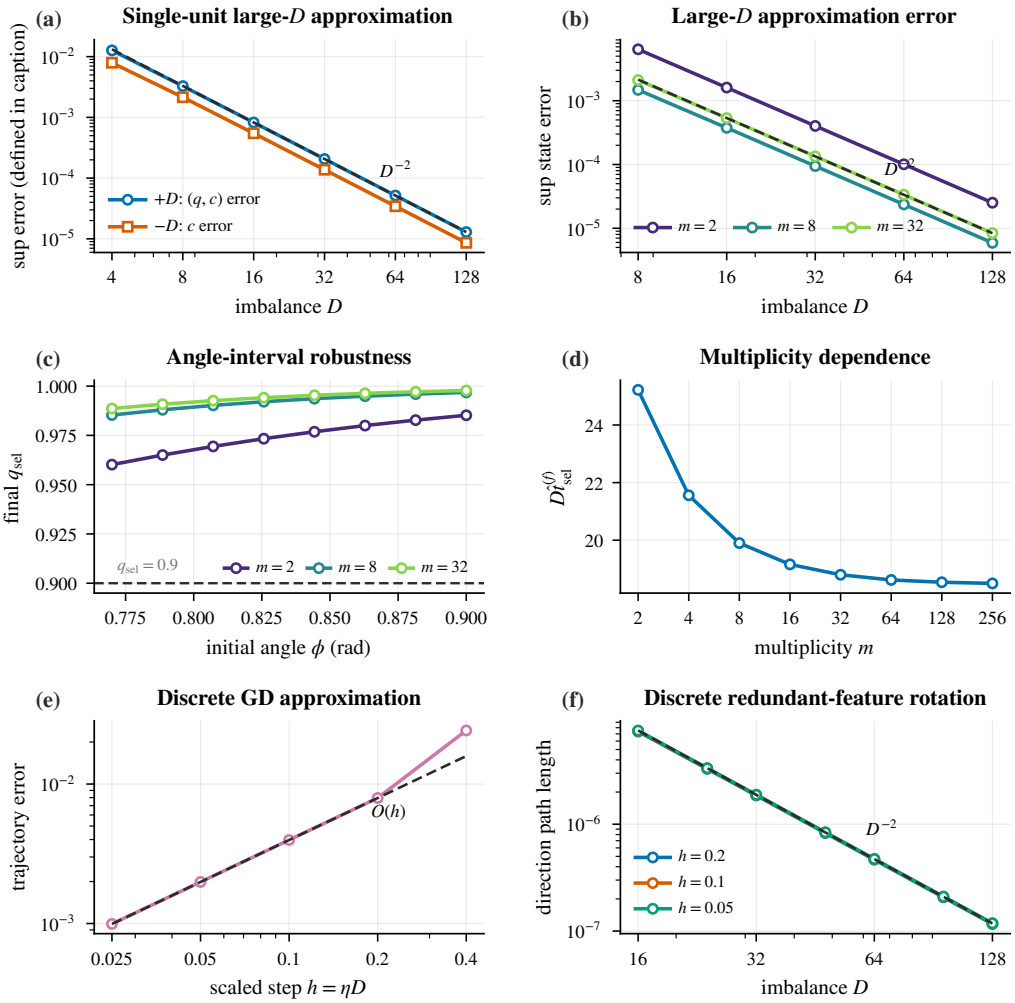

Figure D.1: Scaling and robustness results. (a) Single-unit errors (D.2): e_+ compares (q, c) on $\tau \in [0, 4]$, and e_- compares only c on $\sigma \in [0, 2]$, both including zero. (b) Convergence of the exact reduced flow to its large-D limit. (c) Selection across the tested angles and multiplicities. (d) Rescaled selection time $D\widehat{t}_{\mathrm{sel}}^{(f)}$ from (D.6) across multiplicities. (e) Discrete-to-continuous trajectory error as a function of $h = \eta D$. (f) Post-selection motion of the redundant direction. Dashed lines show the predicted orders.

D.6 TWO-TEACHER COMPETITION WITH NONDUPLICATE STUDENTS

Let $e_1, \dots, e_{10}$ be the standard basis of $\mathbb{R}^{10}$. We use $x \sim \mathcal{N}(0, I_{10})$ and

$$f_\star(x) = [u_1^\top x]_+ + 0.8[u_2^\top x]_+, \qquad u_{1,2} = \cos(0.7)e_1 \mp \sin(0.7)e_2.$$

Eight students have initial directions

$$s_i(0) = \frac{\cos\theta_i e_1 + \sin\theta_i e_2 + 0.12 z_i}{\|\cos\theta_i e_1 + \sin\theta_i e_2 + 0.12 z_i\|}, \qquad z_i \sim \mathcal{N}(0, I_8) \text{ in } \operatorname{span}(e_3, \dots, e_{10}),$$

where $(\theta_1, \dots, \theta_8) = (-0.45, -0.15, 0.15, 0.45, -1.1, 1.1, -1.5, 1.5)$. Independently, $q_i(0)$ is uniform on $[0.15, 0.25]$. The smallest pairwise initial student angle over all 20 seeds is 0.349 radians. Thus the students are distinct and do not lie in the teacher plane.

Configuration A designates units $(1, 4)$, configuration B designates $(2, 3)$, and the same-side configuration designates $(1, 2)$. All three share the same initial coefficients and directions for each seed, and each is tested at $D \in \{16, 32, 64\}$. The designated pair has imbalance $+D$, and all other neurons have imbalance $-D$. Reconstruction through (A.6) changes only the parameter scales and preserves the predictor for every input. Each neuron retains its own initial coefficient and direction across configurations.

We integrate the original-parameter population flow using the analytic Gaussian ReLU kernel, with DOP853 relative tolerance 10^{-9}, absolute tolerance 10^{-11}, and maximum fast-time step 2. Each of the 180 runs (20 seeds, 3 imbalance magnitudes, 3 assignments) ends at $\tau = Dt = 4000$. For the designated pair (i_1, i_2) and teacher coefficients $(b_1, b_2) = (1, 0.8)$, the endpoint success criterion is

$$\max_{k=1,2} \angle(s_{i_k}, u_k) < 0.05, \quad \max_{k=1,2} |q_{i_k} - b_k| < 0.05, \quad \max_{j \notin \{i_1, i_2\}} |q_j| < 0.01, \qquad (b_1, b_2) = (1, 0.8). \tag{D.7}$$

We evaluate this joint criterion at the common endpoint for all seeds.

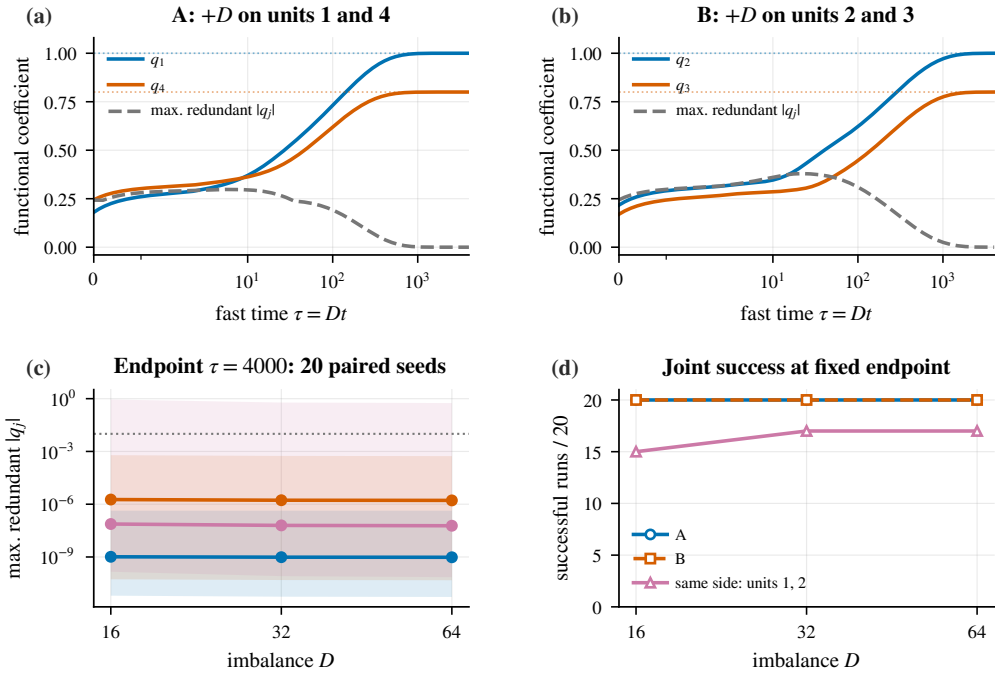

Figure D.2: Two-teacher competition at a fixed initial predictor. (a)–(b) At $D = 32$ and seed 0, moving the positive imbalances from units $(1, 4)$ to $(2, 3)$ changes which students learn the two teachers. Gray curves show the largest magnitude among coefficients of neurons with negative imbalance. (c) Endpoint redundant-coefficient magnitude: lines are medians and bands span the minimum to maximum over 20 paired seeds. The dotted line is 10^{-2}. (d) Counts satisfying (D.7). A and B overlap at 20/20. The same-side condition tests competition between two designated students initially nearer the first teacher.

Table D.6: Two-teacher endpoint results at $\tau = 4000$. The last column is the largest coefficient magnitude among neurons with negative imbalance over all students and seeds in the condition.

D	neurons with $+D$	success count	maximum redundant $\lvert q_j \rvert$
16	A: $(1, 4)$	20/20	4.29×10^{-7}
16	B: $(2, 3)$	20/20	6.14×10^{-4}
16	same side: $(1, 2)$	15/20	8.97×10^{-1}
32	A: $(1, 4)$	20/20	4.30×10^{-7}
32	B: $(2, 3)$	20/20	5.57×10^{-4}
32	same side: $(1, 2)$	17/20	6.03×10^{-1}
64	A: $(1, 4)$	20/20	4.30×10^{-7}
64	B: $(2, 3)$	20/20	5.44×10^{-4}
64	same side: $(1, 2)$	17/20	5.64×10^{-1}

Configurations A and B yield the specified teacher assignments across all seeds, as shown in figure D.2 and table D.6. The same-side configuration produces both assignments to distinct teachers and solutions that distribute the second feature among additional neurons. In the latter case, both designated neurons can remain near the first teacher. These outcomes exhibit the interaction between scale imbalance and initial direction in feature competition.

Table D.7 reports alignment and numerical accuracy. The tighter-tolerance comparison repeats all three configurations at $D = 32$ for seeds 0 and 19, using relative and absolute tolerances 10^{-11} and 10^{-13}. We compare trajectories on a common grid containing zero and 600 logarithmically spaced points from 0.01 to 4000.

Table D.7: Alignment and numerical checks for the two-teacher experiment. Values are maxima over the corresponding runs.

Quantity	Upper bound
Assigned-teacher angle in A and B (radians)	1.52×10^{-4}
Initial coefficient reconstruction error	5.6×10^{-17}
Final imbalance drift	1.2×10^{-12}
Coefficient and alignment change under tighter tolerances	2.2×10^{-8}

D.7 Equal-scale control and single-student comparison

The equal-scale control in figure 1 uses $m = 8$, $D = 32$, $q_i(0) = 1/8$, $s_i(0) = s_0$, and $\phi = \pi/4$, as in the mixed-imbalance trajectory. All students have $\delta_i = +D$. The initial functional contribution of every student is therefore identical across the two configurations. Exact symmetry permits integration of the original input and output weights of one representative student, with all eight contributions included in the residual. We use DOP853 with relative tolerance 10^{-11}, absolute tolerance 10^{-13}, and maximum fast-time step 0.1, recording the same grid on $0 \le \tau \le 60$ as the mixed configuration. Tightening both tolerances by a factor of ten changes the recorded parameters by less than 1.1×10^{-13}; the maximum imbalance drift is below 4.3×10^{-14}.

Figure D.3 displays the single-student comparison underlying theorem 3.1 and corollary 3.2. Both students start at $(q_0, c_0) = (0.6, 0)$ with teacher coefficient $q_\star = 1$. The phase curves use $D = 16$ and stop at alignment $c = 0.95$. The time ratio uses this same alignment threshold and the imbalance values in table D.1. The points are the archived hitting-time measurements used to test the quadratic separation.

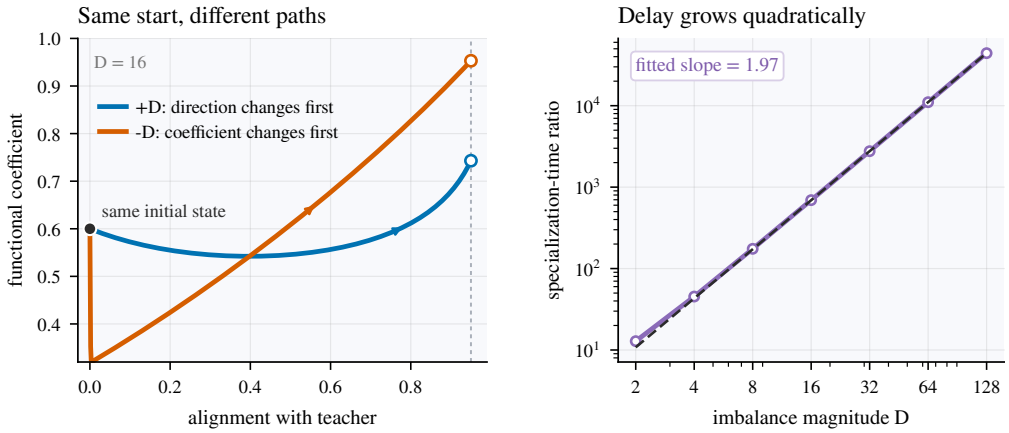

Figure D.3: Opposite imbalances change the path and speed of single-student learning. Left: coefficient–alignment trajectories from a common initial state, at $D = 16$. Right: the ratio of specialization times grows quadratically with D. The dashed line has slope two.