SGD SELECTS INITIALIZATION-DEPENDENT REPRESENTATION IN OVERPARAMETERIZED RELU NETWORKS

Anonymous authors
Paper under double-blind review

ABSTRACT

Overparameterized neural networks can represent the same function through many different hidden configurations, raising the question of which representation is favored by a given training algorithm. We study how stochastic gradient descent (SGD), jointly training both layers, resolves this ambiguity in finite-width, bias-free, two-layer ReLU networks trained to learn a linear teacher under Gaussian inputs. At width two, exact fitting leaves no choice: the neurons must align with the teacher directions. At every width $m \geq 4$, however, infinitely many non-aligned exact fits are possible. Despite this ambiguity, we prove that joint batch-size-one SGD from sufficiently small initialization drives every neuron arbitrarily close to one of the two teacher directions while reaching arbitrarily small population error in finite time. This occurs without explicit regularization, layerwise training, gradient clipping, or a learning-rate schedule. The dynamics additionally select an initialization-dependent allocation of the fitted function across neurons, providing a finer characterization of the representation selected by SGD. Numerical experiments reproduce the predicted alignment, fitting, and initialization-dependent allocation. Our results identify sufficient conditions under which standard SGD exhibits a strong implicit bias toward a highly structured internal representation, even when overparameterization creates infinitely many qualitatively different exact representations.

1 INTRODUCTION

Overparameterized neural networks (NNs) admit many solutions that fit the data, yet gradient-based training often exhibits a strong preference for particular solutions and representations (Brutzkus et al., 2017; Gunasekar et al., 2018; Chizat & Bach, 2020). Existing theoretical frameworks explain this phenomenon by introducing simplifying assumptions, such as explicit regularization, modified first-order optimization updates, or infinite-width limits (Jacot et al., 2018; Mei et al., 2018; Chizat et al., 2019; Rotskoff & Vanden-Eijnden, 2022). While these assumptions make the evolution of learned representations tractable to analyze, standard SGD dynamics can produce similarly structured representations even in the absence of such constraints.

In this work, we study the representations of SGD that are favored without the aforementioned assumptions. We take a finite, overparameterized two-layer ReLU network and train both layers at the same time with plain stochastic gradient descent (SGD) without regularization. We attempt to answer the following question

When many representations fit perfectly, which one does SGD prefer?

To isolate this representation-selection question, we deliberately choose a setting in which the statistical problem is simple but the optimization dynamics are not. Following the teacher–student paradigm used to study neural-network training dynamics (Goldt et al., 2019; Akiyama & Suzuki, 2021; Ba et al., 2022; Damian et al., 2022), we consider Gaussian inputs and a linear teacher $y(x) = \beta\theta_\star^\top x$ learned by an overparameterized two-layer ReLU student network

$$f(x) = \frac{1}{\sqrt{m}} \sum_{j=1}^{m} a_j \, \sigma(w_j^\top x), \qquad \sigma(t) = \mathrm{ReLU}(t). \tag{1}$$

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.

Least squares could recover the target directly; our question is which of the network's many hidden representations the training dynamics select. The difficulty lies in the coupled nonlinear dynamics created by training both layers simultaneously. Each hidden weight w_j determines which inputs activate its ReLU through the gate $\mathbf{1}\{w_j^\top x > 0\}$, while its output weight a_j controls how strongly that neuron influences the loss and hence the update of w_j. As the hidden weights move, these activation regions change as well, so the geometry of the updates evolves along the optimization trajectory itself. This difficulty is reflected in recent feature-learning analyses, which typically use structured proof strategies such as staged training, different gradient algorithms, or explicit regularization (Mousavi-Hosseini et al., 2022; Tsikouras et al., 2026; Zhang et al., 2026), or establish alignment during an early phase before the later interactions are studied (Maennel et al., 2018; Min et al., 2024; Boursier & Flammarion, 2025a;b).

The closest result to our setting is the recent work of (Tinaz et al., 2026), who analyze the same Gaussian linear teacher under joint training of both layers, but only for the minimal-width case $m = 2$. They leave open the overparameterized regime, where exact representations are no longer unique. Here overparameterization refers to width beyond the two neurons needed to represent the teacher. Moving to this regime changes the question itself, as convergence to zero loss no longer determines which representation the dynamics will select.

Width changes the question. With exactly two hidden units, the representation of a linear function is rigid: zero loss forces one neuron to point along $+\theta_\star$ and the other along $-\theta_\star$, with uniquely determined effective masses, modulo permutation and the positive-homogeneity rescaling of ReLU (see left side of Panel B in Figure 1). This is precisely the regime analyzed by (Tinaz et al., 2026).

At every integer width $m \geq 4$, this rigidity disappears. Additional neurons can form antipodal pairs away from the teacher: their nonlinear parts cancel, while their linear contributions combine to represent the same target (see right side of Panel B in Figure 1). We make this precise by characterizing the complete nondegenerate zero-loss set (Proposition 1). It contains a continuum of exact representations whose neurons are not aligned with the teacher poles. These solutions are not saddles that SGD must escape. Their residual vanishes on every input, so every gradient is exactly zero and SGD initialized at any one of them never moves.

From this, we understand that the loss is indifferent between the aligned representation and a continuum of non-aligned alternatives. This distinguishes our question from low-dimensional representation-learning results that use explicit weight decay as part of the training dynamics (Damian et al., 2022; Mousavi-Hosseini et al., 2022; Zhang et al., 2026); for instance (Mousavi-Hosseini et al., 2022) use regularization to rule out critical points away from the teacher subspace, and report that removing it can change the alignment behavior. Here no penalty resolves the ambiguity among exact fits, so any preference must arise from the trajectory. Our main theorem shows that it does.

Informal version of Theorem 1. Consider the two-layer ReLU network in Equation (1), trained to learn a linear teacher. For every integer width $m \geq 2$, joint SGD with batch size one, sampling with replacement from a fixed training set, brings every neuron close to either $+\theta_\star$ or $-\theta_\star$, provided the network is initialized at a sufficiently small scale. At the same time, the network learns the linear teacher to arbitrarily small population error.

Figure 1 summarizes this contrast: panel A shows the tiny, initially unaligned network, panel B shows the non-uniqueness of perfect representations created by overparameterization, and panel C shows the narrow two-pole region selected by SGD.

The surprising part is not that the network can learn a linear teacher. It is that it does so through a representation that the loss does not force. At width $m \geq 4$, non-aligned exact representations are abundant and stationary, yet small-initialization SGD drives *every hidden neuron* toward one of only two directions and preserves this alignment while the network grows to the fitting scale. The conclusion concerns every neuron, rather than only their span, their average, or a subset. It is therefore stronger than recovery of the teacher subspace: the entire hidden layer concentrates neuronwise around the two teacher poles.

Alignment is only the first level of selection. When a sign class contains more than one neuron, exact alignment still leaves its mass free to be divided among those neurons. Our second theorem shows that SGD remembers enough of the initialization to choose among them. While the network output

is still negligible, the teacher amplifies a growing component of each neuron depending on both its initial size and how favorably it points at its selected pole, converting small initial differences into different shares. Once aligned, same-class neurons grow at nearly the same multiplicative rate, so those shares stay almost fixed while the total mass rises. Theorem 2 makes this precise: SGD selects not only the two-pole geometry but an explicit, initialization-dependent division of labor within it.

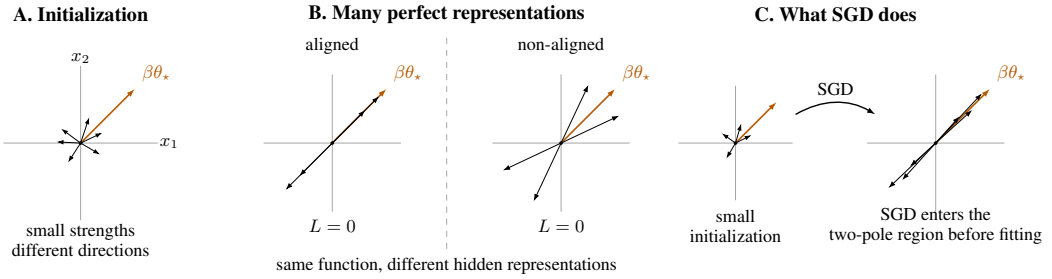


Figure 1: Overparameterization creates many exact hidden representations, while small-initialization SGD selects the aligned family. (A) At initialization, neurons have small strength and unrelated directions. (B) At width two, exact fitting is rigid, but at width $m \geq 4$ there are many perfect representations, including non-aligned ones. (C) From small initialization, SGD drives all neurons toward the two teacher poles $\pm\theta_\star$.

Why does SGD align? Think of each neuron as a vector pointing in some direction, with an associated strength. At initialization, the neurons point in different directions and can be only weakly aligned, or even nearly orthogonal, to the teacher. Their strengths are tiny, however, so that $f \approx 0$. This creates an early regime in which the teacher y dominates the dynamics and can organize the representation before the network is strong enough to fit it, a mechanism that also appears in small-initialization analyses of feature learning (Min et al., 2024; Boursier & Flammarion, 2025b; Zhang et al., 2026). For a linear teacher under Gaussian inputs, a key identity implies that the teacher term pulls each neuron toward its sign-selected teacher pole, independently of its current angle (Lemma B.2). As a result, the part of each neuron pointing toward a teacher pole grows, while the component orthogonal to the teacher changes little, causing the neurons to align while the network is still weak. In the second phase, their strengths grow toward the fitting scale. The main difficulty is showing that this later growth does not undo the alignment: we prove that interactions among neurons near the same pole are stabilizing, while interactions across the two poles remain weak.

Why is the aligned family special? The family selected by SGD is not just visually simpler than the non-aligned alternatives. It is exactly extremal among all perfect representations: an exact fit uses the least possible total neuron strength precisely when every active neuron sits at its corresponding teacher pole; among parameterizations achieving this minimum strength, the squared parameter norm is minimized precisely when each neuron is balanced, $|a_j| = \|w_j\|$. The variational result (Proposition 2) singles out the aligned family from the much larger zero-loss set; the dynamics resolve the ambiguity remaining inside it. More specifically, minimum norm still permits many allocations, from two-neuron solutions to dense ones. Theorem 2 identifies a dense balanced exact representation from the initialization and bounds the finite-time SGD endpoint's distance to it.

Summary of our contributions. Our key contributions can be summarized as follows:

- *Width creates genuine representation ambiguity.* Width two is rigid, whereas every integer width $m \geq 4$ admits infinitely many non-aligned exact fits, all of which are stationary under SGD on any dataset.
- *Joint SGD selects the aligned family.* From small initialization, vanilla SGD on one fixed Gaussian dataset aligns every neuron with its associated teacher pole, preserves this alignment through growth, and fits the teacher in *population* at every integer width. Among exact fits, aligned solutions minimize effective mass, while balanced aligned solutions minimize parameter norm.

- *The selected family is variationally distinguished.* Among exact fits, aligned solutions minimize effective mass, and balanced aligned solutions minimize parameter norm.
- *Initialization selects within that family.* The initial growing modes determine the relative neuron masses, which are set during alignment and remain nearly fixed through the fitting phase. This identifies an explicit dense representation, rather than only its two-pole geometry.

2 SETUP

Teacher, model, and loss. Let data be generated from $X \sim \gamma_d = N(0, I_d)$, of dimension $d \geq 2$, and let the teacher model be $y(X) = \beta\theta_\star^\top X$, where $\|\theta_\star\| = 1$ and $\beta > 0$. We train both layers of a bias-free ReLU network and evaluate its population square loss:

$$f_\Psi(x) = \frac{1}{\sqrt{m}} \sum_{j=1}^{m} a_j \sigma(w_j^\top x), \qquad \sigma(t) = \mathrm{ReLU}(t) \qquad \mathcal{L}(\Psi) = \frac{1}{2}\mathbb{E}\left[(f_\Psi(X) - y(X))^2\right]$$

Here $\Psi = ((a_j, w_j))_{j=1}^m$ and $m \geq 2$ is an integer. Our guarantees concern population loss; SGD updates use the empirical loss on the fixed dataset.

The training set. Fix a failure probability $\delta \in (0, 1/2)$ and an accuracy parameter ε. Before training, we draw once

$$\mathcal{D} = \{X_1, \ldots, X_M\}, \qquad X_1, \ldots, X_M \overset{\text{i.i.d.}}{\sim} \gamma_d \qquad M = \widetilde{O}\left(\frac{d + \log(1/\delta)}{\varepsilon^2}\right),$$

with labels $y_i = y(X_i)$. Throughout, $\widetilde{O}$ hides factors polylogarithmic in $1/\varepsilon$, and $a \asymp b$ means $c \leq a/b \leq C$ for universal constants $c, C > 0$. Write $\ell_i(\Psi) = \frac{1}{2}(f_\Psi(X_i) - y_i)^2$ and $\widehat{\mathcal{L}}(\Psi) = M^{-1}\sum_{i \leq M} \ell_i(\Psi)$, for the single-example and empirical losses, respectively. No fresh data is ever drawn again.

The algorithm. At every iteration $n \geq 0$ we draw an index $I_n \sim \mathrm{Unif}\{1, \ldots, M\}$, independently of everything else, and update *both layers simultaneously* on that single example:

$$a_{j,n+1} = a_{j,n} - \eta \nabla_{a_j} \ell_{I_n}(\Psi_n), \qquad w_{j,n+1} = w_{j,n} - \eta \nabla_{w_j} \ell_{I_n}(\Psi_n). \tag{2}$$

Both layers are updated after every sampled example, using a constant learning rate and no explicit regularization. We use the convention $\sigma'(0) = 0$.

Polar notation and initialization scale. Let $s_j := \mathrm{sign}(a_j(0)) \in \{-1, +1\}$ denote the initial sign of the output weight of neuron j; this sign selects the corresponding teacher pole $s_j\theta_\star$. Write $S_s = \{j : s_j = s\}$ and any neuron with $a_j \neq 0$ and $w_j \neq 0$, put $p_j = |a_j||w_j|$ (effective mass), $u_j = w_j/|w_j|$ (direction), $P_s = m^{-1/2}\sum_{j \in S_s} p_j$ (normalized sign-class mass), and $\theta_j = \angle(u_j, s_j\theta_\star)$ (angle to the selected pole). The parameter ε sets the initialization scale and the accuracy guarantee; it is not a tolerance chosen after training.

Assumption 1 (Initialization)**.** *Let $p_0 = c_0\beta\varepsilon^4/\sqrt{m}$, where the universal constant $c_0 > 0$ is fixed in Theorem 1. Fix $s_j \in \{-1, +1\}$ and $S_s = \{j : s_j = s\}$ with $m/4 \leq |S_s| \leq 3m/4$. Initialize*

$$a_j(0) = s_j b_j^0, \quad w_j(0) = r_j^0 u_j^0, \quad b_j^0, r_j^0 \in [\sqrt{p_0}, \tfrac{5}{4}\sqrt{p_0}], \quad \|u_j^0\| = 1, \quad s_j\langle u_j^0, \theta_\star\rangle \geq -\tfrac{1}{2}.$$

The two layer magnitudes may vary independently, so the initialization need not be balanced. The initialization is independent of the dataset and sampled indices. Corollary C.1 gives a sampling rule that does not use the teacher direction and satisfies Assumption 1 with probability at least $1 - \gamma$ (for any $\gamma \in (0, 1)$) when $d \gtrsim \log(m/\gamma)$ and $m \gtrsim \log(1/\gamma)$.

3 JOINT ORDINARY SGD ALIGNS THEN FITS

We now present our main finite-time guarantee for joint SGD. The statement uses four scales:

$$L = \log(1/\varepsilon), \qquad \bar{\varepsilon} = \varepsilon L, \qquad \tau = \frac{\eta\beta}{\sqrt{m}}, \qquad D = d + \log\frac{16m}{\delta\varepsilon}.$$

Here L is the logarithmic training time scale, $\bar{\varepsilon}$ is the target scale for the angles θ_j, τ is the normalized learning rate, and D is an effective dimension parameter.

Theorem 1 (SGD aligns then fits)**.** *There are universal constants $c_0, C_M, \varepsilon_0, C > 0$ such that the following holds. Let $0 < \varepsilon \le \varepsilon_0$, $0 < \delta < 1/2$, and let Assumption 1 hold. Draw*

$$M = \left\lceil C_M\big(d + \log(16/\delta)\big)\frac{L^2}{\varepsilon^2}\log^2(eL/\varepsilon)\right\rceil$$

i.i.d. Gaussian examples once, and run the SGD update (2) *with the constant learning rate $\eta = \tau\sqrt{m}/\beta$, where $\tau \asymp \varepsilon^3/(D^2L^3)$. Let $T_1 < T_\star$ be the iteration counts given in Appendix F; they depend only on $(\varepsilon, \delta, d, m)$ and satisfy $T_\star = O(D^2L^4/\varepsilon^3)$. Then, with probability at least $1-\delta$ over the dataset and the sampled indices, every output sign remains fixed at its initial value and every neuron stays nonzero through $T_\star$, and*

$$\max_{T_1 \le n \le T_\star} \max_{1 \le j \le m} \angle(u_{j,n}, s_j\theta_\star) \le C\bar{\varepsilon}, \qquad \|f_{\Psi_{T_\star}} - y\|_{L^2(\gamma_d)} \le C\beta\varepsilon.$$

Here T_1 marks the end of the alignment phase (Phase I) and $T_\star$ the fitting time. The full proof can be found in Appendices B– F.

Directional alignment holds for every neuron at every iteration between T_1 and $T_\star$ (Phase II), while the fitting guarantee bounds the true population $L^2(\gamma_d)$ error even though SGD only sees the M training examples. We next sketch the two-phase dynamics underlying Theorem 1.

3.1 Proof sketch

The dynamics have two phases: the neurons first turn toward their poles while the network is still negligible, and then grow to the fitting scale without ruining alignment. We first describe the population-gradient reference dynamics. The final part of the sketch explains how blocks of single-example SGD satisfy the same estimates with controlled errors.

Phase I: why the neurons align. Until the end of Phase I the network output satisfies $\|f_{\Psi_n}\|_{L^2(\gamma_d)} \le \beta\varepsilon^2$, so the residual is essentially $-y$ and each neuron is driven by the teacher alone. One might expect this force to favor only the neurons that already point the right way, since the ReLU gate depends on w_j. It does not: the gate integrates out, because $\mathbb{E}[(\theta_\star^\top X)\mathbf{1}\{w^\top X > 0\}X] = \frac{1}{2}\theta_\star$ for every $w \neq 0$ (Lemma B.2). The teacher therefore pulls every hidden weight w_j along its sign selected direction $s_j\theta_\star$, with a strength proportional to $|a_j|$ and independent of the angle that w_j currently makes.

Writing $A_j = |a_j|$ and $B_j = s_j\theta_\star^\top w_j$, for the second-layer magnitude and the component of w_j toward its selected pole, the teacher-dominated dynamics take the approximate form

$$A_j^+ \approx A_j + \frac{\tau}{2}B_j, \qquad B_j^+ \approx B_j + \frac{\tau}{2}A_j.$$

Thus the update of each quantity is proportional to the other: a larger second-layer weight strengthens the pull of w_j toward the teacher, while a larger teacher-aligned component strengthens the growth of $|a_j|$. The two therefore reinforce one another. Equivalently, in the coordinates

$$G_j = \frac{A_j + B_j}{2}, \qquad H_j = \frac{A_j - B_j}{2},$$

the dynamics approximately decouple: G_j grows like $(1+\tau/2)^n$, while H_j decays like $(1-\tau/2)^n$. Thus the growing mode is amplified while the difference H_j becomes negligible.

At the same time, the component of w_j orthogonal to $\theta_\star$ (nuisance component) receives no teacher force and stays of the order of its initial size $\sqrt{p_0}$. Hence the growing teacher-aligned component eventually dominates: w_j aligns with $s_j\theta_\star$, while $|a_j|$ becomes comparable to $s_j\theta_\star^\top w_j$.

The growth rate is common to all neurons; what differs between them is the initial growing mode

$$G_j^0 = \frac{|a_j(0)| + s_j\theta_\star^\top w_j(0)}{2}.$$

By the end of Phase I, $p_j = |a_j|\|w_j\|$ is proportional, up to small relative error, to $(G_j^0)^2$, so the same mechanism that aligns the neurons also determines their relative masses.

Phase II: why growth does not undo alignment. At the end of Phase I, the neurons point close to their sign-selected poles, but the network is still too small to fit the teacher. The class masses satisfy the approximate logistic recursion $P_s^+ \approx P_s + \frac{\tau}{\beta} P_s(\beta - P_s)$ (Lemma B.18), so both class masses approach β within $O(L/\tau)$ further steps.

The main difficulty is showing that this growth does not destroy alignment. Three facts control the angular dynamics. First, the net restoring coefficient is governed by the class deficit $\beta - P_s$. Second, the leading same-class interaction has a favorable sign when the angular dynamics are averaged with the neuron masses. Third, opposite-class neurons are nearly antipodal, so the overlap of their active regions, and hence their interaction, is small.

While $P_s \le \beta/2$, the restoring coefficient is uniformly positive and keeps every neuron within $O(\varepsilon)$ of its selected pole. After P_s crosses $\beta/2$, that coefficient weakens and may become slightly negative if the class overshoots the target, so the proof no longer relies on contraction. Instead, it shows directly that the accumulated adverse drift and error terms remain small, keeping the mass-weighted angular error $O(\varepsilon)$. The influence of the class on any individual neuron is controlled by this weighted error. Accumulating the resulting $O(\varepsilon)$ forcing over $O(L)$ normalized time gives the worst-neuron bound $O(\varepsilon L)$. Hence every neuron remains within $O(\varepsilon \log(1/\varepsilon))$ of its pole throughout Phase II. At $T_\star$, $|P_s - \beta| = O(\beta\varepsilon)$ and the mass-weighted angular error is $O(\varepsilon)$, which together yield $\|f_{\Psi_{T_\star}} - y\|_{L^2(\gamma_d)} = O(\beta\varepsilon)$.

Why one example at a time is enough. The preceding two phases describe the population dynamics, whereas SGD uses one training example per update. A single stochastic gradient need not approximate the population gradient, so the proof groups consecutive updates into blocks of K steps. These blocks are used only in the analysis; the algorithm updates both layers after every sampled example. Each block must contain enough steps, $K \asymp \frac{D^2 L^2}{\varepsilon^2}$, for the sampling fluctuations to average out, while its normalized duration, $K\tau \asymp \frac{\varepsilon}{L}$, must be small enough that the parameters move little within the block. These requirements give $\tau \asymp \frac{\varepsilon^3}{D^2 L^3}$. The total update over a block can then be compared with one controlled population-gradient step.

The estimates must hold relative to each neuron's current size: each effective mass is an $O(\varepsilon^4)$ fraction of its fitting-scale value at initialization and still an $O(\varepsilon^2)$ fraction at T_1, so an absolute error bound would not protect every neuron. A neuron-relative regularity estimate and uniform concentration of the Gaussian gated covariances provide this control throughout the trajectory, even though the same dataset is reused (Lemmas F.1 and F.3).

3.2 What function-level geometry misses

A common way to identify the features learned by a predictor is through its expected gradient outer product (EGOP)

$$G_f = \mathbb{E}[\nabla f(X) \nabla f(X)^\top].$$

Its leading eigenspace records the input directions along which the predictor varies most. This viewpoint underlies active-subspace methods and recent work on feature learning in NNs (Constantine, 2015; Beaglehole et al., 2024b; Radhakrishnan et al., 2024).

Theorem 1 implies that the input-gradient energy outside the teacher direction is small. Let $Q_\star = I - \theta_\star \theta_\star^\top$.

Corollary 1 (Near-rank-one gradient geometry)**.** *Under the assumptions of Theorem 1, on the same event of probability at least $1 - \delta$, and with $\lambda_1 \ge \cdots \ge \lambda_d$ the eigenvalues of G_f for $f = f_{\Psi_{T_\star}}$,*

$$\mathbb{E}\|Q_\star \nabla f(X)\|^2 \le C\beta^2\varepsilon^2, \qquad \sum_{i \ge 2} \lambda_i(G_f) \le C\beta^2\varepsilon^2.$$

Proof. See Appendices D and F.

This function-level description does not identify the hidden representation. Every exact fit, including the non-aligned representations constructed in Section 4, has $G_f = \beta^2 \theta_\star \theta_\star^\top$. Neuronwise alignment therefore reveals a preference of the dynamics that the EGOP cannot distinguish.

4 EXACT ZERO-SET GEOMETRY

Theorem 1 shows that SGD reaches the fitting scale with every neuron close to a teacher pole. We now ask whether this geometry is forced by exact fitting. At width two it is; from width four onward, even the hidden lines are no longer determined by exact fitting.

The starting point is the decomposition $\sigma(t) = \frac{t+|t|}{2}$. Each ReLU neuron therefore contributes a linear part and an even nonlinear part. Since the teacher $y(x) = \beta\theta_\star^\top x$ is linear and odd, an exact representation must reproduce its linear part while canceling all even contributions.

Proposition 1 (Exact zero set)**.** *Assume $a_j \neq 0 \neq w_j$ for every j, and set $s_j = \operatorname{sign}(a_j)$. Then $f_\Psi(x) = \beta\theta_\star^\top x$ in $L^2(\gamma_d)$ if and only if*

$$\frac{1}{\sqrt{m}} \sum_{j=1}^{m} s_j p_j u_j = 2\beta\theta_\star \tag{Z1}$$

and, for every unit vector u,

$$\sum_{j:\, u_j \in \{u,-u\}} s_j p_j = 0. \tag{Z2}$$

Proof. See Appendix G.

The two conditions are the linear and the nonlinear parts. (Z1) says the weighted directions must sum to the teacher direction with the correct magnitude; individual neurons may point elsewhere, only their sum is fixed. (Z2) says the even terms must cancel in every direction; u and $-u$ are grouped because $|u^\top x| = |(-u)^\top x|$, so among neurons on that line the positive and negative output contributions must have equal total mass. The simplest example is an antipodal pair, $p\,\sigma(u^\top x) - p\,\sigma(-u^\top x) = p\,u^\top x$: the nonlinear parts cancel and a linear contribution in direction u remains. A linear teacher can therefore be represented by combining several such pairs whose linear contributions sum to the teacher. Using more than one hidden line requires at least four neurons.

Corollary 2 (Width dichotomy)**.** *Let $m = 2$. Then $f_\Psi = y$ if and only if, after relabeling the neurons, $a_1 > 0 > a_2$, $u_1 = \theta_\star$, $u_2 = -\theta_\star$, and $p_1 = p_2 = \sqrt{2}\,\beta$. In particular both neurons are active, and only the split of each p_j between $|a_j|$ and $\|w_j\|$ is free.*

For every integer $m \geq 4$ and every prescribed angle $\gamma \in (0, \pi/2)$, there exist parameters with $f_\Psi = y$, $|a_j| = \|w_j\| > 0$ for every j, $\max_j \angle(u_j, s_j\theta_\star) = \gamma$, and at least two distinct hidden lines.

Proof. See Appendix G.

Corollary 2 shows when alignment is forced by fitting and when it is not. At width two, every exact fit is aligned. From width four on, exact fits can place every neuron at any angle $\gamma \in (0, \pi/2)$ from its pole, even with perfectly balanced layers, so neither zero loss nor balance forces alignment.

At width three, exact fitting still forces the teacher line, but need not place each neuron at its output-sign-selected pole; Appendix G gives an example.

The non-aligned exact fits are not states that SGD could later escape: their residual vanishes on every input, so every single-example gradient is zero. Thus every exact representation, aligned or not, is a fixed point of the algorithm (2), for every dataset and every sequence of sampled indices. Training on a finite dataset does not resolve this ambiguity: every exact representation satisfies $f_\Psi(X_i) = y_i$ for all i, and therefore has zero training loss as well. Nevertheless, Theorem 1 shows that, from small initialization, SGD reaches the fitting scale near the aligned family. This preference is therefore determined by the training trajectory rather than by zero loss or stationarity.

All exact representations are also indistinguishable at the level of the learned function: each represents the same linear teacher and hence, as noted in Section 3.2, has $G_f = \beta^2\theta_\star\theta_\star^\top$. The EGOP recovers the teacher direction but cannot reveal whether the hidden neurons themselves are aligned. Theorem 1 therefore identifies a property of the learned representation that is invisible both from the loss and from the input-gradient geometry.

The aligned representations are nevertheless distinguished inside the zero set by a simple size criterion. Define the effective mass and the raw squared parameter norm by

$$\Pi(\Psi) = \frac{1}{\sqrt{m}} \sum_{j=1}^{m} |a_j| \|w_j\|, \qquad S(\Psi) = \sum_{j=1}^{m} \big(a_j^2 + \|w_j\|^2\big).$$

Proposition 2 (Variational characterization of aligned representations)**.** *Among all parameterizations Ψ that represent the teacher exactly, the minimum effective mass and minimum raw squared parameter norm are*

$$\min_{f_\Psi = y} \Pi(\Psi) = 2\beta, \qquad \min_{f_\Psi = y} S(\Psi) = 4\beta\sqrt{m},$$

respectively. A representation minimizes Π if and only if every active neuron lies at its output-sign teacher pole and the two sign classes each have effective mass β. Among these representations, it minimizes S if and only if $w_j = a_j \theta_\star$ for every j.

Proof. See Appendix E.

The characterization has a simple interpretation. Any mass placed away from the two teacher poles is partly wasted: only its projection onto $\theta_\star$ contributes to the teacher, while the remaining components must cancel elsewhere. Alignment removes this excess. For a fixed effective mass, the raw parameter norm is then smallest when the two layers are balanced.

Proposition 2 identifies the aligned minimum-mass family and its balanced minimum-norm members. It leaves the allocation of mass within each sign class undetermined; the next section shows that the initialization determines it.

5 Initialization selects the neuronwise masses

Within the aligned family, exact fitting determines each sign-class mass but leaves its allocation among neurons free; the initialization predicts that allocation. Recall the initial growing mode $G_j^0 = \frac{1}{2}\left(|a_j(0)| + s_j \theta_\star^\top w_j(0)\right)$. It combines the second-layer weight magnitude with the first-layer component toward the sign-selected teacher pole. Define the predicted and observed shares by

$$\lambda_j = \frac{(G_j^0)^2}{\sum_{i \in S_{s_j}} (G_i^0)^2}, \qquad v_{j,n} = \frac{p_{j,n}}{\sum_{i \in S_{s_j}} p_{i,n}}.$$

Both sum to one within each sign class. For the fixed teacher, λ_j is determined from the initialization and $\theta_\star$ before training, while $v_{j,n}$ measures the allocation along training. A larger initial growing mode predicts a larger final share. The corresponding balanced aligned representation is

$$p_j^{\mathrm{sel}} = \beta\sqrt{m}\,\lambda_j, \qquad a_j^{\mathrm{sel}} = s_j \sqrt{p_j^{\mathrm{sel}}}, \qquad w_j^{\mathrm{sel}} = a_j^{\mathrm{sel}} \theta_\star.$$

We write $\Psi^{\mathrm{sel}} = ((a_j^{\mathrm{sel}}, w_j^{\mathrm{sel}}))_{j=1}^m$.

Theorem 2 (Initialization-dependent mass selection)**.** *Under the assumptions of Theorem 1, on the same event of probability at least $1 - \delta$, for every neuron j,*

$$e^{-C\varepsilon}\lambda_j \le v_{j,n} \le e^{C\varepsilon}\lambda_j \quad (T_1 \le n \le T_\star), \qquad \left| \frac{p_{j,T_\star}}{\beta\sqrt{m}\,\lambda_j} - 1 \right| \le C\varepsilon,$$

and

$$\frac{\|\Psi_{T_\star} - \Psi^{\mathrm{sel}}\|}{\|\Psi^{\mathrm{sel}}\|} \le C\varepsilon.$$

Proof. See Appendices E and F.

Thus the initialization predicts both the relative share and the fitting-time effective mass of every neuron. It also identifies the balanced aligned representation approached at the fitting time. Phase I establishes these shares up to the stated error, and they change little during Phase II.

6 Experiments

We test the two predictions of our theory: that alignment precedes fitting, and that the initialization determines the neuronwise allocation. In all experiments, SGD updates both layers after each example,

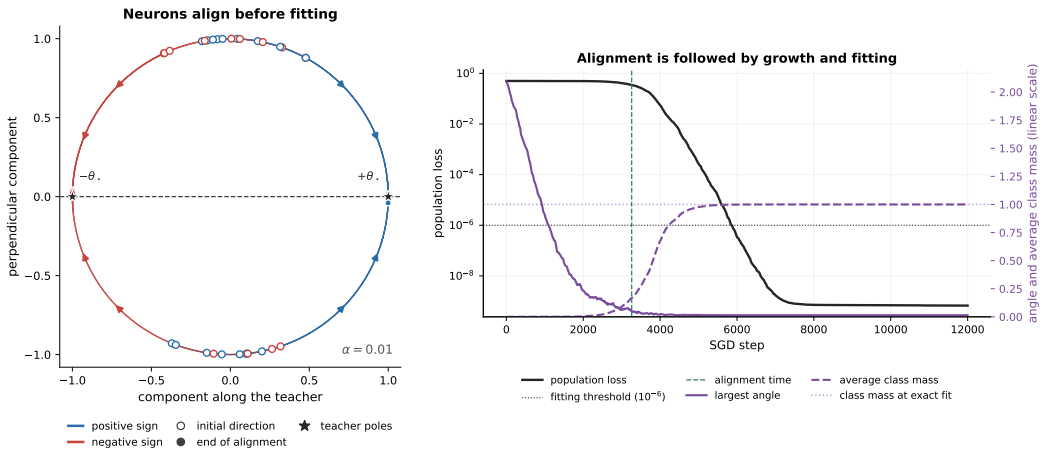


Figure 2: **The two phases of joint SGD. Left:** Normalized hidden directions for a random 3% of neurons initialized away from $\pm\theta_\star$. Open circles mark initialization, filled circles the end of the alignment phase, and arrows the direction of time. Each neuron moves toward the teacher pole selected by the sign of its output weight. **Right:** Population loss (black, logarithmic left axis), maximum angle to the selected pole (solid purple), and average sign-class mass (dashed purple). The directions align while the class mass and loss change little. After the alignment time (green), the class mass grows toward its exact-fit value and the loss falls below 10^{-8}.

without regularization. Performance is measured by the Gaussian population loss, computed in closed form. Full details are in Appendix H.

Figure 2 shows the two-phase dynamics in a run with $d = 2$, $m = 1000$, and $\beta = 1$. The hidden directions first move toward the teacher pole selected by the sign of their second-layer weight. During this period, the class masses and population loss change little. Once the directions are aligned, the class masses grow toward their exact-fit values and the loss rapidly decreases.

The two panels separate directional learning from growth. In the left panel, each point is $u_{j,n} = w_{j,n}/\|w_{j,n}\|$, showing how the neuron directions evolve during training. Movement around the unit circle therefore reflects actual changes in direction, and not changes in magnitude. In the right panel, the maximum angle reaches 0.05 radians at step $2{,}800$, before the network has reduced its loss considerably. The average class mass then rises toward $\beta = 1$, and the population loss crosses 10^{-6} at step $5{,}960$. This is the ordering used in the proof: SGD first selects the hidden directions and then increases their mass while preserving alignment.

We next test the neuronwise prediction of Theorem 2. In dimension $d = 20$, we run 20 trials at each width $m \in \{4, 16, 64\}$ and compare the final neuronwise shares with the shares λ_j predicted from initialization. Although all neurons start with the same effective mass, their final shares are nonuniform and closely match the growing-mode prediction: across all three widths, the median log-share error is at most 0.0065 and the log-share correlation is at least 0.9997. Figure 3 in the appendix shows one representative run per width, chosen to have the median final prediction error, and illustrates how individual shares move from their common initial value toward the allocation predicted before training. Additional experimental details are given in Appendix H.

7 Conclusion and Limitations

Joint SGD from small initialization approaches an aligned exact representation whose neuronwise masses are predicted by the initialization. The aligned exact family minimizes effective mass, and its balanced members minimize raw parameter norm; the dynamics identify a particular member without explicit regularization. Our experiments illustrate alignment before fitting and the predicted mass allocation. The guarantee concerns a finite fitting time on a fixed dataset, not asymptotic convergence. Proposition G.1 (Appendix G) exhibits cubic flatness of the population loss near an aligned fit; it does not determine the convergence rate of the SGD trajectory. Convergence of the same run, larger initialization scales, sharper learning-rate bounds, and richer targets remain open.

REPRODUCIBILITY STATEMENT

The appendices give the assumptions and proofs for alignment, population fitting, gradient geometry, variational characterization, and neuronwise mass selection. Appendix F establishes the fixed-dataset, single-example SGD guarantees, and Appendix G proves the exact zero-set characterization and cubic local obstruction. The experimental appendix specifies the initialization, training procedure, and evaluation formulas.

ETHICS STATEMENT

This paper is theoretical and studies optimization mechanisms in stylized NN models. It uses no human-subject data and introduces no direct high-risk application.

AI USE STATEMENT

AI tools were used during the development and preparation of this work to assist with the writing of the text and improving the organization. Additionally, we used AI tools to improve certain math statements.

All mathematical statements and proofs were independently checked by the authors, and cited literature was verified against the original sources. The authors take full responsibility for the correctness of the mathematical results, the accuracy of the citations, the novelty claims, and the final content of the paper.

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. PMLR, 2021.

Jimmy Ba, Murat A Erdogdu, Taiji Suzuki, Zhichao Wang, Denny Wu, and Greg Yang. High-dimensional asymptotics of feature learning: How one gradient step improves the representation. *Advances in Neural Information Processing Systems*, 35:37932–37946, 2022.

Daniel Beaglehole, Ioannis Mitliagkas, and Atish Agarwala. Feature learning as alignment: a structural property of gradient descent in non-linear neural networks. *arXiv preprint arXiv:2402.05271*, 2024a.

Daniel Beaglehole, Peter Súkeník, Marco Mondelli, and Mikhail Belkin. Average gradient outer product as a mechanism for deep neural collapse. *Advances in Neural Information Processing Systems*, 37:130764–130796, 2024b.

Alberto Bietti, Joan Bruna, Clayton Sanford, and Min Jae Song. Learning single-index models with shallow neural networks. *Advances in neural information processing systems*, 35:9768–9783, 2022.

Etienne Boursier and Nicolas Flammarion. Simplicity bias and optimization threshold in two-layer relu networks. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 5241–5275, 2025a.

Etienne Boursier and Nicolas Flammarion. Early alignment in two-layer networks training is a two-edged sword. *Journal of Machine Learning Research*, 26(183):1–75, 2025b.

Etienne Boursier, Loucas Pillaud-Vivien, and Nicolas Flammarion. Gradient flow dynamics of shallow relu networks for square loss and orthogonal inputs. *Advances in Neural Information Processing Systems*, 35:20105–20118, 2022.

Alon Brutzkus, Amir Globerson, Eran Malach, and Shai Shalev-Shwartz. Sgd learns over-parameterized networks that provably generalize on linearly separable data. *arXiv preprint arXiv:1710.10174*, 2017.

Lenaic Chizat and Francis Bach. Implicit bias of gradient descent for wide two-layer neural networks trained with the logistic loss. In *Conference on learning theory*, pp. 1305–1338. PMLR, 2020.

Lenaic Chizat, Edouard Oyallon, and Francis Bach. On lazy training in differentiable programming. *Advances in neural information processing systems*, 32, 2019.

Paul G Constantine. *Active subspaces: Emerging ideas for dimension reduction in parameter studies*. SIAM, 2015.

Paul G Constantine, Armin Eftekhari, Jeffrey Hokanson, and Rachel A Ward. A near-stationary subspace for ridge approximation. *Computer Methods in Applied Mechanics and Engineering*, 326:402–421, 2017.

Alexandru Damian, Jason Lee, and Mahdi Soltanolkotabi. Neural networks can learn representations with gradient descent. In *Conference on learning theory*, pp. 5413–5452. PMLR, 2022.

Simon S Du, Wei Hu, and Jason D Lee. Algorithmic regularization in learning deep homogeneous models: Layers are automatically balanced. volume 31, 2018.

Margalit Glasgow. Sgd finds then tunes features in two-layer neural networks with near-optimal sample complexity: A case study in the xor problem. In *International Conference on Learning Representations*, volume 2024, pp. 52419–52430, 2024.

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

Suriya Gunasekar, Jason D Lee, Daniel Soudry, and Nati Srebro. Implicit bias of gradient descent on linear convolutional networks. *Advances in neural information processing systems*, 31, 2018.

Arthur Jacot, Franck Gabriel, and Clément Hongler. Neural tangent kernel: Convergence and generalization in neural networks. *Advances in neural information processing systems*, 31, 2018.

Akshay Kumar and Jarvis Haupt. Directional convergence near small initializations and saddles in two-homogeneous neural networks. *arXiv preprint arXiv:2402.09226*, 2024.

Akshay Kumar and Jarvis Haupt. Towards understanding gradient flow dynamics of homogeneous neural networks beyond the origin. *Journal of Machine Learning Research*, 26(269):1–83, 2025.

Chaoyue Liu, Libin Zhu, and Mikhail Belkin. Loss landscapes and optimization in over-parameterized non-linear systems and neural networks. *Applied and Computational Harmonic Analysis*, 59: 85–116, 2022.

Hartmut Maennel, Olivier Bousquet, and Sylvain Gelly. Gradient descent quantizes relu network features. *arXiv preprint arXiv:1803.08367*, 2018.

Song Mei, Andrea Montanari, and Phan-Minh Nguyen. A mean field view of the landscape of two-layer neural networks. *Proceedings of the National Academy of Sciences*, 115(33):E7665–E7671, 2018.

Hancheng Min, Enrique Mallada, and Rene Vidal. Early neuron alignment in two-layer relu networks with small initialization. In *International Conference on Learning Representations*, volume 2024, pp. 3431–3473, 2024.

Alireza Mousavi-Hosseini, Sejun Park, Manuela Girotti, Ioannis Mitliagkas, and Murat A Erdogdu. Neural networks efficiently learn low-dimensional representations with sgd. *arXiv preprint arXiv:2209.14863*, 2022.

Greg Ongie, Rebecca Willett, Daniel Soudry, and Nathan Srebro. A function space view of bounded norm infinite width relu nets: The multivariate case. 2019.

Mary Phuong and Christoph H Lampert. The inductive bias of relu networks on orthogonally separable data. In *International Conference on Learning Representations*, 2020.

Adityanarayanan Radhakrishnan, Daniel Beaglehole, Parthe Pandit, and Mikhail Belkin. Mechanism for feature learning in neural networks and backpropagation-free machine learning models. *Science*, 383(6690):1461–1467, 2024.

Adityanarayanan Radhakrishnan, Mikhail Belkin, and Dmitriy Drusvyatskiy. Linear recursive feature machines provably recover low-rank matrices. *Proceedings of the National Academy of Sciences*, 122(13):e2411325122, 2025.

Basri Ronen, David Jacobs, Yoni Kasten, and Shira Kritchman. The convergence rate of neural networks for learned functions of different frequencies. *Advances in Neural Information Processing Systems*, 32, 2019.

Grant Rotskoff and Eric Vanden-Eijnden. Trainability and accuracy of artificial neural networks: An interacting particle system approach. *Communications on Pure and Applied Mathematics*, 75(9): 1889–1935, 2022.

Pedro Savarese, Itay Evron, Daniel Soudry, and Nathan Srebro. How do infinite width bounded norm networks look in function space? In *Conference on Learning Theory*, pp. 2667–2690. PMLR, 2019.

Nadav Timor, Gal Vardi, and Ohad Shamir. Implicit regularization towards rank minimization in relu networks. In *International Conference on Algorithmic Learning Theory*, pp. 1429–1459. PMLR, 2023.

Berk Tinaz, Changzhi Xie, and Mahdi Soltanolkotabi. When both layers learn: Training dynamics of representing linear models via relu networks. *arXiv preprint arXiv:2606.04476*, 2026.

Nikos Tsikouras, Yorgos Pantis, Ioannis Mitliagkas, and Christos Tzamos. A derandomization framework for structure discovery: Applications in neural networks and beyond. In *Proceedings of the Fourteenth Annual International Conference on Learning Representations*, 2026.

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

Olivier Zahm, Paul G Constantine, Clémentine Prieur, and Youssef M Marzouk. Gradient-based dimension reduction of multivariate vector-valued functions. *SIAM Journal on Scientific Computing*, 42(1):A534–A558, 2020.

Bohan Zhang, Zihao Wang, Hengyu Fu, and Jason Lee. Neural networks learn generic multi-index models near information-theoretic limit. In *International Conference on Learning Representations*, volume 2026, pp. 70833–70895, 2026.

Yedi Zhang, Andrew Saxe, and Peter E Latham. When are bias-free relu networks effectively linear networks? *arXiv preprint arXiv:2406.12615*, 2024.

Libin Zhu, Damek Davis, Dmitriy Drusvyatskiy, and Maryam Fazel. Average gradient outer product in kernel regression provably recovers the central subspace for multi-index models. *arXiv preprint arXiv:2605.15082*, 2026.

Zhenyu Zhu, Fanghui Liu, and Volkan Cevher. How gradient descent balances features: A dynamical analysis for two-layer neural networks. In *International Conference on Learning Representations*, volume 2025, pp. 9342–9385, 2025.

A Related work

Feature learning and low-dimensional targets. A broad literature proves that neural training can recover low-dimensional structure beyond kernel regimes. (Damian et al., 2022) show representation learning for low-dimensional polynomial targets, while (Ba et al., 2022) identify a rank-one spike aligned with the linear teacher component after one gradient step. (Bietti et al., 2022) study single-index learning with a shallow architecture, and (Mousavi-Hosseini et al., 2022) prove arbitrary-width

subspace learning for Gaussian multi-index targets using online SGD with weight decay. Those results cover substantially richer target classes than our linear-teacher theorem. Our focus is simultaneous training of both layers without explicit weight decay, and all-neuron representation selection through fitting.

Joint training and alignment dynamics. The closest result to our end-to-end dynamics is (Tinaz et al., 2026), who study the Gaussian linear teacher with exactly two hidden ReLU units and jointly train both layers by finite-data gradient descent. They prove global linear convergence from moderately small random initialization with order-wise optimal sample complexity, through alignment, growth, and local-refinement phases. Their result is substantially stronger on initialization, asymptotic convergence, and sample efficiency. The width-two restriction, however, removes the within-class interactions and non-identifiable wide zero set that are central here. We instead analyze every integer width $m \geq 2$ using single-example SGD sampled with replacement from one fixed Gaussian dataset, without an explicit regularizer, at the price of a more structured small initialization and conservative sample complexity.

Joint-layer dynamics have also been analyzed for XOR (Glasgow, 2024). In a different parameterization, (Zhu et al., 2025) study feature grouping under gradient descent with fixed positive output coefficients; their theorem does not jointly train both layers. General results for two-homogeneous models show that near-origin gradient flow drives nonvanishing parameter blocks toward stationary directions of a neural correlation function (Kumar & Haupt, 2024), and subsequent work follows related dynamics beyond the origin (Kumar & Haupt, 2025). For orthogonal regression data, (Boursier et al., 2022) prove convergence to zero loss and an implicit bias toward minimum variation norm. Related alignment results appear for orthogonally separable classification (Phuong & Lampert, 2020; Min et al., 2024). Our result instead treats single-example SGD on a fixed dataset, preserves all-neuron alignment through fitting, and tracks the growing-mode magnitudes needed for neuronwise mass selection. Earlier and complementary perspectives include directional quantization (Maennel et al., 2018) and the simplifying and obstructive effects of early alignment (Boursier & Flammarion, 2025b;a). The balancing phenomenon for homogeneous models goes back to (Du et al., 2018); here homogeneity is used samplewise to show that the first-order balance drift cancels exactly under SGD.

Expressivity of bias-free ReLU networks. The linear teacher is a natural realizable odd target in our model. (Ronen et al., 2019) identified restrictions on the odd components learnable by shallow bias-free ReLU networks on spherical data, and (Zhang et al., 2024) proved that every odd function represented by a two-layer bias-free ReLU network is linear. These results concern the functions representable by the architecture. Our result instead concerns which hidden representation joint SGD selects among the many parameterizations of the same linear function.

AGOP/EGOP and active subspaces. The gradient second-moment matrix is classical in active-subspace and ridge-approximation theory (Constantine, 2015; Constantine et al., 2017; Zahm et al., 2020); in particular, its spectral tail controls approximation by functions of the leading subspace under suitable input measures. In neural feature learning, average gradient outer product (AGOP) / EGOP has emerged as a descriptor and proposed mechanism for learned feature geometry (Radhakrishnan et al., 2024; Beaglehole et al., 2024a;b), and (Radhakrishnan et al., 2025) prove low-rank recovery properties for linear Recursive Feature Machines. Very recently, (Zhu et al., 2026) show that the AGOP of a kernel-ridge predictor can recover the central subspace of multi-index polynomial targets even before accurate prediction.

The generic implication from a small EGOP tail to ridge approximation is established theory. Our principal statement concerns hidden-neuron directions; the EGOP bound records a function-level consequence and cannot distinguish aligned from non-aligned exact representations.

Function-space versus parameter-space geometry. Function-space descriptions of two-layer ReLU networks and non-uniqueness of parameterizations are classical themes (Savarese et al., 2019; Ongie et al., 2019). (Timor et al., 2023) give positive and negative results on implicit rank minimization in ReLU networks. Our exact zero-set characterization isolates this distinction in the present linear-teacher model and shows a sharp $m = 2$ versus $m \geq 4$ change. Overparameterization-induced slow convergence also appears for a single ReLU teacher with fixed output coefficients (Xu & Du, 2023). For local optimization geometry, (Liu et al., 2022) develop PL-type conditions for over-

parameterized nonlinear systems; our cubic family gives an explicit failure of any state-independent local PL constant around the two-pole solution at width $m \geq 3$.

B WARM-UP TRAJECTORY ANALYSIS

We first develop the trajectory analysis in a simplified setting with exactly balanced sign classes, equal-magnitude initialization, and fresh batches. This lets us focus on the main dynamics without carrying the technical complications of a more general initialization and reused data. The phase analysis itself only requires small neuron-relative gradient errors and controlled balance drift, as stated after Lemma B.15. The preceding concentration bounds provide these controls for fresh batches, while Appendix F establishes their blockwise counterparts for fixed-dataset SGD.

Appendix C extends the balanced, even-width, equal-magnitude initialization used here to Assumption 1 used in the actual SGD algorithm, and Appendix D sharpens the final error bound to the rate stated in the main theorem.

B.1 SETTING, NOTATION, STANDING ASSUMPTIONS

B.1.1 MODEL AND DYNAMICS

$d \geq 2$, $m \geq 2$ even, $\beta > 0$, $\theta_\star \in \mathbb{S}^{d-1}$, $\gamma_d = \mathcal{N}(0, I_d)$, $X \sim \gamma_d$. Teacher $y(x) = \beta\theta_\star^\top x$. Parameter state $\Psi = (a_j, w_j)_{j \leq m} \in \mathbb{R}^m \times (\mathbb{R}^d)^m$;

$$f_\Psi(x) = \frac{1}{\sqrt{m}} \sum_{j=1}^{m} a_j \sigma(w_j^\top x), \qquad \sigma(z) = z_+; \qquad \mathcal{L}(\Psi) = \tfrac{1}{2}\mathbb{E}\big[(f_\Psi(X) - y(X))^2\big].$$

At iteration $n \geq 0$ draw $X_{n,1}, \ldots, X_{n,B} \sim \gamma_d$ i.i.d., independent of the past. With $R_n := f_{\Psi_n} - y$,

$$\begin{aligned}
\widehat{g}^a_{j,n} &= \frac{1}{B\sqrt{m}} \sum_{b=1}^{B} R_n(X_{n,b})\, \sigma(w_{j,n}^\top X_{n,b}), \\
\widehat{g}^w_{j,n} &= \frac{1}{B\sqrt{m}} \sum_{b=1}^{B} R_n(X_{n,b})\, a_{j,n} \mathbf{1}_{\{w_{j,n}^\top X_{n,b} > 0\}} X_{n,b}. \\
a_{j,n+1} &= a_{j,n} - \eta \widehat{g}^a_{j,n}, \qquad w_{j,n+1} = w_{j,n} - \eta \widehat{g}^w_{j,n}.
\end{aligned} \tag{1.1}$$

These are the gradients of $\widehat{\mathcal{L}}_n(\Psi) = \frac{1}{2B}\sum_b (f_\Psi(X_{n,b}) - y(X_{n,b}))^2$ with the convention $\sigma'(0) = 0$. Away from gate boundaries they coincide with the classical sample gradients. Conditional on the past, every nonzero fixed hidden weight has zero probability of a gate-boundary event under the fresh Gaussian batch. For reused data, Appendix F handles strict gates, including zero preactivations, directly.

$\mathcal{F}_n := \sigma(\text{initialization, batches } 0, \ldots, n-1)$, so Ψ_n is $\mathcal{F}_n$-measurable and batch n is independent of $\mathcal{F}_n$. Population gradients $g^a_{j,n} := \mathbb{E}[\widehat{g}^a_{j,n} \mid \mathcal{F}_n]$, $g^w_{j,n} := \mathbb{E}[\widehat{g}^w_{j,n} \mid \mathcal{F}_n]$.

B.1.2 INITIALIZATION

Signs $s_j \in \{\pm 1\}$, $S_\pm := \{j : s_j = \pm 1\}$, $|S_+| = |S_-| = m/2$;

$$p_0 := c_0 \frac{\beta\varepsilon^4}{\sqrt{m}}, \qquad a_{j,0} = s_j\sqrt{p_0}, \qquad w_{j,0} = \sqrt{p_0}\, u_j^0, \;\; u_j^0 \in \mathbb{S}^{d-1}, \qquad \max_j \big|\langle u_j^0, \theta_\star\rangle\big| \leq \tfrac{1}{2}. \tag{1.2}$$

So $|a_{j,0}| = \|w_{j,0}\| = \sqrt{p_0}$: the initialization is exactly balanced, $p_{j,0} = p_0$.

B.1.3 POLAR VARIABLES

The original parameters (a_j, w_j) contain two radial magnitudes, even though ReLU is positively homogeneous. It is therefore useful to separate each neuron into its sign, its effective magnitude,

and its direction. When $a_{j,n} \neq 0 \neq w_{j,n}$: $b_{j,n} := |a_{j,n}|$, $r_{j,n} := \|w_{j,n}\|$, $p_{j,n} := b_{j,n} r_{j,n}$, $u_{j,n} := w_{j,n}/r_{j,n}$. Identically $a_j \sigma(w_j^\top x) = \operatorname{sign}(a_j)\, p_j\, \sigma(u_j^\top x)$, so under sign stability

$$f_n(x) = \frac{1}{\sqrt{m}} \sum_{j=1}^{m} s_j p_{j,n} \sigma(u_{j,n}^\top x). \tag{1.3}$$

For each sign class, define the total effective mass $P_{s,n} := m^{-1/2} \sum_{j \in S_s} p_{j,n}$ and its unnormalized version $D_{s,n} := \sum_{j \in S_s} p_{j,n} = \sqrt{m}\, P_{s,n}$. We also write $\Pi_n := P_{+,n} + P_{-,n} = m^{-1/2} \sum_j p_{j,n}$.

The angular error of neuron j is $\theta_{j,n} := \angle(u_{j,n}, s_j \theta_\star)$. Its component orthogonal to the teacher direction is $\xi_{j,n} := (I - \theta_\star \theta_\star^\top) u_{j,n}$, so that $u_{j,n} = s_j \cos\theta_{j,n}\, \theta_\star + \xi_{j,n}$ and $\|\xi_{j,n}\| = \sin\theta_{j,n}$.

We use three summaries of the angular error. The worst angle is $\Theta_{s,n} := \max_{j \in S_s} \theta_{j,n}$, with $\Theta_n := \max_j \theta_{j,n}$. The mass-weighted mean-square transverse error is $V_{s,n} := D_{s,n}^{-1} \sum_{j \in S_s} p_{j,n} \|\xi_{j,n}\|^2$, and we write $W_{s,n} := \sqrt{V_{s,n}}$. Finally, $Z_{s,n} := \max_{j \in S_s} \|\xi_{j,n}\|$ controls the worst transverse component. The balance defect is $\Delta_{j,n} := a_{j,n}^2 - \|w_{j,n}\|^2$.

B.1.4 Scales, kernels, working region

$$L := \log(1/\varepsilon), \quad \bar\varepsilon := \varepsilon L, \quad \kappa := \frac{\beta}{2\sqrt{m}}, \quad \tau := \frac{\eta\beta}{\sqrt{m}} = 2\eta\kappa, \quad h := \eta\kappa = \tau/2;$$

$$K(\psi) := \frac{\sin\psi + (\pi - \psi)\cos\psi}{2\pi}, \qquad q(\psi) := \frac{\pi - \psi}{2\pi};$$

$$K_{ij,n} := K(\psi_{ij,n}), \quad q_{ij,n} := q(\psi_{ij,n}), \quad \psi_{ij,n} := \angle(u_{i,n}, u_{j,n});$$

$$\mathcal{R} := \big\{\Psi :\ \Pi \le 6\beta,\ \ \tfrac{1}{2} \le |a_j|/\|w_j\| \le 2\ \ \forall j\big\}.$$

B.1.5 Standing assumptions

- **(P1)** $d \ge 2$; $m \ge 2$ even; $\beta > 0$; $\theta_\star \in \mathbb{S}^{d-1}$; $\delta \in (0, \frac{1}{2})$.
- **(P2)** $0 < \varepsilon \le \varepsilon_0$; $\Theta_\star := A\bar\varepsilon$.
- **(P3)** Initialization 1.2.
- **(P4)** $\tau L \le c_\eta \varepsilon$, i.e. $\eta \le c_\eta \sqrt{m}\, \varepsilon/(\beta L)$.
- **(P5)** $N_1 := \lceil L/\log(1 + \tau/2) \rceil$, $N_2 := N_1 + \lceil C_L \tau^{-1} L \rceil$, $N_\star := N_2$.
- **(P6)** $B \ge C_B\, \varepsilon^{-2} L^2 D_\star$ with $D_\star := d + \log \frac{16 m N_\star}{\delta}$.

The universal constants $c_0, c_1, c_\eta, C_B, C_L, C_T, A, \varepsilon_0$ are fixed in the balanced-case proof below and, finally, in Appendix C. $C \ge 1$ and $c \le 1$ denote universal constants that may change between occurrences. Error parameters:

$$\iota := \varepsilon^2 + \rho_B \ \ \text{(Phase I)}, \qquad \zeta := \Theta_\star^2 + \rho_B + \tau \ \ \text{(Phase II)},$$

with ρ_B defined in Lemma B.12.

Role of (P6). Condition P6 is specific to the fresh-batch realization of the deterministic error bounds used below: together with Lemmas B.12–B.13, it ensures that ρ_B is sufficiently small. The subsequent phase analysis is deterministic conditional on these error bounds. Appendix F establishes the corresponding error bounds blockwise for SGD on a fixed dataset, allowing the same deterministic phase analysis to be reused in that setting.

Lemma B.1 (Smallness budget)**.** *Let $A \ge 1$. Suppose $\varepsilon_0 \le 10^{-2}$ and $A^2 \varepsilon_0 \log^3(1/\varepsilon_0) \le 1$. Then for every $0 < \varepsilon \le \varepsilon_0$,*

$$L \ge \log 100 > 4, \qquad A^2 \varepsilon L^3 \le 1, \qquad \Theta_\star^2 L \le \varepsilon,$$
$$\Theta_\star^2 \le \varepsilon \le 10^{-2}, \qquad \Theta_\star \le \tfrac{1}{10}, \qquad \varepsilon L \le 1.$$

If moreover $\rho_B \le c_1 \varepsilon/L$ and $\tau \le c_\eta \varepsilon/L$ with $c_1, c_\eta \le 1$ (Lemma B.13, (P4)), then

$$\zeta \le 3\varepsilon, \qquad \zeta L \le 3\varepsilon, \qquad \iota L \le 2\varepsilon, \qquad \zeta \le 3\bar\varepsilon. \tag{1.4}$$

Proof. $g(\varepsilon) := \varepsilon \log^3(1/\varepsilon)$ has $g'(\varepsilon) = L^2(L-3) > 0$ for $L > 3$, i.e. on $(0, e^{-3})$. Since $10^{-2} < e^{-3} \approx 0.0498$, g is increasing on $(0, \varepsilon_0]$, so $A^2 g(\varepsilon) \le A^2 g(\varepsilon_0) \le 1$. Then $\Theta_\star^2 L = A^2\varepsilon^2 L^3 = \varepsilon \cdot A^2 \varepsilon L^3 \le \varepsilon$; and since $L \ge 1$, $\Theta_\star^2 = A^2\varepsilon^2 L^2 \le A^2\varepsilon^2 L^3 \le \varepsilon \le 10^{-2}$, hence $\Theta_\star \le 10^{-1}$. Also $\varepsilon L \le \varepsilon L^3 \le A^{-2} \le 1$. For 1.4: $\zeta = \Theta_\star^2 + \rho_B + \tau \le \varepsilon + c_1\varepsilon + c_\eta\varepsilon \le 3\varepsilon$; $\zeta L \le \Theta_\star^2 L + c_1\varepsilon + c_\eta\varepsilon \le 3\varepsilon$; $\iota L = \varepsilon^2 L + \rho_B L \le \varepsilon + c_1\varepsilon \le 2\varepsilon$ (using $\varepsilon L \le 1$); $\zeta \le 3\varepsilon \le 3\bar{\varepsilon}$ (as $L \ge 1$). □

B.2 Preliminaries

B.2.1 Standard facts

This appendix uses the following standard result:

- **(E1)** Gaussian concentration: $\|h(X) - \mathbb{E}h(X)\|_{\psi_2} \le C\,\mathrm{Lip}(h)$ for $X \sim \gamma_d$.
- **(E2)** Bernstein's inequality for sums of i.i.d. centered sub-exponential random variables.
- **(E3)** Volumetric bound: $\mathbb{S}^{d-1}$ has a $\frac{1}{2}$-net of cardinality $\le 5^d$.
- **(E4)** Gaussian integration by parts (Stein's identity): $\mathbb{E}[(\theta^\top X)h(X)] = \mathbb{E}[\theta^\top \nabla h(X)]$ for $h \in C^1$ of at most polynomial growth, $X \sim \gamma_d$, $\theta \in \mathbb{R}^d$.

B.2.2 Gaussian identities

Lemma B.2 (Gaussian identities)**.** *Let $u, v, \theta \in \mathbb{S}^{d-1}$, $\psi := \angle(u, v)$, $X \sim \gamma_d$. Then*

$$\textit{(i)}\ \mathbb{E}\big[(\theta^\top X)\sigma(v^\top X)\big] = \tfrac{1}{2}\,\theta^\top v; \qquad \textit{(ii)}\ \mathbb{E}\big[(\theta^\top X)\mathbf{1}_{\{u^\top X>0\}}X\big] = \tfrac{1}{2}\,\theta;$$

$$\textit{(iii)}\ \mathbb{E}\big[\sigma(v^\top X)\mathbf{1}_{\{u^\top X>0\}}X\big] = q(\psi)\,v + \frac{\sin\psi}{2\pi}\,u;$$

$$\textit{(iv)}\ \mathbb{E}\big[\sigma(u^\top X)\sigma(v^\top X)\big] = K(\psi), \qquad \mathbb{P}(u^\top X > 0,\ v^\top X > 0) = q(\psi).$$

Proof. Fix a unit $u^\perp \perp u$ such that $\mathrm{span}(u, u^\perp)$ contains v (for (iii),(iv)) resp. θ (for (ii)); write $X = t\,u + s\,u^\perp + X_0$ with $t, s \sim \mathcal{N}(0,1)$ independent and X_0 independent of (t, s), mean zero, supported on $\mathrm{span}(u, u^\perp)^\perp$. Put $v = \cos\psi\,u + \sin\psi\,u^\perp$, so $v^\top X = \cos\psi\,t + \sin\psi\,s$.

(iv) Write $(t, s) = R(\cos\Phi, \sin\Phi)$. By rotational invariance of $\mathcal{N}(0, I_2)$, $\Phi \sim \mathrm{Unif}[0, 2\pi)$ and R (Rayleigh, density $re^{-r^2/2}$ on $(0,\infty)$, $\mathbb{E}R^2 = 2$) are independent. Then $u^\top X = R\cos\Phi$ and $v^\top X = R\cos(\Phi - \psi)$. Both are positive iff $\Phi \in (-\frac{\pi}{2}, \frac{\pi}{2}) \cap (\psi - \frac{\pi}{2}, \psi + \frac{\pi}{2}) = (\psi - \frac{\pi}{2}, \frac{\pi}{2})$, an arc of length $\pi - \psi$; hence the probability equals $\frac{\pi-\psi}{2\pi} = q(\psi)$. Next,

$$\begin{aligned}\mathbb{E}[\sigma(u^\top X)\sigma(v^\top X)] &= \mathbb{E}[R^2]\frac{1}{2\pi}\int_{\psi-\pi/2}^{\pi/2} \cos\phi\,\cos(\phi-\psi)\,d\phi \\ &= \frac{1}{2\pi}\int_{\psi-\pi/2}^{\pi/2} \big[\cos\psi + \cos(2\phi-\psi)\big]\,d\phi.\end{aligned}$$

using $2\cos\phi\cos(\phi-\psi) = \cos\psi + \cos(2\phi-\psi)$ and $\mathbb{E}R^2 = 2$. Now $\int_{\psi-\pi/2}^{\pi/2}\cos\psi\,d\phi = (\pi - \psi)\cos\psi$, and

$$\int_{\psi-\pi/2}^{\pi/2}\cos(2\phi-\psi)\,d\phi = \tfrac{1}{2}\big[\sin(2\phi-\psi)\big]_{\psi-\pi/2}^{\pi/2} = \tfrac{1}{2}\big[\sin(\pi-\psi) - \sin(\psi-\pi)\big] = \sin\psi.$$

Hence $K(\psi) = \frac{(\pi-\psi)\cos\psi + \sin\psi}{2\pi}$.

(i) Apply (E4) to the mollifications $h_\epsilon := \sigma(v^\top\cdot) * \varphi_\epsilon \in C^\infty$: $\mathbb{E}[(\theta^\top X)h_\epsilon(X)] = \mathbb{E}[\theta^\top\nabla h_\epsilon(X)]$. As $\epsilon \downarrow 0$, $h_\epsilon \to \sigma(v^\top\cdot)$ pointwise with $|h_\epsilon(x)| \le |v^\top x| + \epsilon$, and $\nabla h_\epsilon \to v\,\mathbf{1}_{\{v^\top x>0\}}$ for a.e. x with $\|\nabla h_\epsilon\|_\infty \le 1$; dominated convergence gives $\mathbb{E}[(\theta^\top X)\sigma(v^\top X)] = (\theta^\top v)\mathbb{P}(v^\top X > 0) = \frac{1}{2}\theta^\top v$.

(ii) For any $e \perp \mathrm{span}(u, \theta)$, the random variable $e^\top X$ is independent of $(\theta^\top X, t)$ and has mean 0, so the e-component of the left side vanishes; and $\theta^\top e = 0$. So it suffices to compute the components along u and $u^\perp$, with $\theta = \cos\phi_0\, u + \sin\phi_0\, u^\perp$:

$$u\text{-component} = \mathbb{E}[(\cos\phi_0\, t + \sin\phi_0\, s)\mathbf{1}_{\{t>0\}} t] = \cos\phi_0\, \mathbb{E}[t^2 \mathbf{1}_{\{t>0\}}] = \tfrac{1}{2}\cos\phi_0,$$

$$u^\perp\text{-component} = \mathbb{E}[(\cos\phi_0\, t + \sin\phi_0\, s)\mathbf{1}_{\{t>0\}} s] = \sin\phi_0\, \mathbb{E}[s^2]\, \mathbb{P}(t>0) = \tfrac{1}{2}\sin\phi_0.$$

(iii) Components orthogonal to $\mathrm{span}(u, v)$ vanish by the argument in (ii), and the right-hand side lies in that span. The u-component of the left side is $\mathbb{E}[\sigma(v^\top X)\mathbf{1}_{\{t>0\}} t] = \mathbb{E}[\sigma(v^\top X)\sigma(u^\top X)] = K(\psi)$, and the u-component of the right side is $q(\psi)\cos\psi + \frac{\sin\psi}{2\pi} = \frac{(\pi-\psi)\cos\psi + \sin\psi}{2\pi} = K(\psi)$. For the $u^\perp$-component, condition on t and apply one-dimensional (E4) in s (again via mollification): $\mathbb{E}_s[\sigma(\cos\psi\, t + \sin\psi\, s)\, s] = \sin\psi\, \mathbb{P}_s(v^\top X > 0)$, so

$$\mathbb{E}[\sigma(v^\top X)\mathbf{1}_{\{t>0\}} s] = \sin\psi\, \mathbb{P}(t>0,\ v^\top X > 0) = \sin\psi\, q(\psi),$$

matching the $u^\perp$-component $q(\psi)\sin\psi$ of the right side. □

Remark B.1 (Boundary term). $\frac{\sin\psi}{2\pi} u = \phi(0)\, \mathbb{E}[\sigma(v^\top X) \mid u^\top X = 0]\, u$ with $\phi(0) = (2\pi)^{-1/2}$. This term is parallel to u, hence annihilated by $I - uu^\top$. We record for later: $K(\psi) - q(\psi)\cos\psi = \frac{\sin\psi}{2\pi}$.

Lemma B.3 (Kernel bounds)**.** *For* $\psi \in [0, \pi]$*:* $0 \le \frac{1}{2} - K(\psi) \le \frac{\psi^2}{4}$*;* $0 \le K(\pi - \psi) \le \frac{\psi^3}{6\pi}$*;* $0 \le q(\psi) \le \frac{1}{2}$*;* $q(\pi - \psi) = \frac{\psi}{2\pi}$.

Proof. $K'(\psi) = -\frac{(\pi-\psi)\sin\psi}{2\pi}$, so $K' \le 0$ and $|K'(\psi)| \le \frac{\pi\psi}{2\pi} = \frac{\psi}{2}$; integrating from 0 and using $K(0) = \frac{1}{2}$ gives the first claim. Next $K(\pi-\psi) = \frac{\sin\psi - \psi\cos\psi}{2\pi}$ and $\sin\psi - \psi\cos\psi = \int_0^\psi t\sin t\, dt \in [0, \psi^3/3]$. The statements for q are immediate. □

B.2.3 Elementary discrete lemmas

Lemma B.4 (Normalization remainder)**.** *For* $u \in \mathbb{S}^{d-1}$ *and* $\|\varpi\| \le \frac{1}{5}$,

$$\Big\| \frac{u + \varpi}{\|u + \varpi\|} - u - (I - uu^\top)\varpi \Big\| \ \le\ 14\|\varpi\|^2.$$

Proof. Let $\sigma_0 := u^\top\varpi$, $\varpi^\perp := (I - uu^\top)\varpi$, $\|u+\varpi\| =: 1+\mu$. Then $(1+\mu)^2 = 1 + 2\sigma_0 + \|\varpi\|^2 =: 1 + x$ with $|x| \le 2\|\varpi\| + \|\varpi\|^2 \le 0.44 \le \frac{1}{2}$, so $\mu = \sqrt{1+x} - 1$. Since $|\sqrt{1+y} - 1 - \frac{y}{2}| \le \frac{y^2}{2}$ for $|y| \le \frac{1}{2}$,

$$|\mu - \sigma_0| \le \big|\mu - \tfrac{x}{2}\big| + \big|\tfrac{x}{2} - \sigma_0\big| \le \tfrac{x^2}{2} + \tfrac{\|\varpi\|^2}{2} \le \tfrac{(2.2\|\varpi\|)^2}{2} + \tfrac{\|\varpi\|^2}{2} \le 3\|\varpi\|^2.$$

Hence $|\mu| \le |\sigma_0| + 3\|\varpi\|^2 \le 1.6\|\varpi\| \le 0.32$ and $1 + \mu \ge \frac{1}{2}$. Finally, using $\varpi = \sigma_0 u + \varpi^\perp$,

$$\frac{u+\varpi}{1+\mu} - u - \varpi^\perp = \frac{(u+\varpi) - (1+\mu)(u+\varpi^\perp)}{1+\mu} = \frac{(\sigma_0 - \mu)u - \mu\varpi^\perp}{1+\mu},$$

whose norm is $\le 2\big(3\|\varpi\|^2 + 1.6\|\varpi\| \cdot \|\varpi\|\big) \le 14\|\varpi\|^2$.

□

Lemma B.5 (Norm remainder)**.** *If* $w \ne 0$*,* $r = \|w\|$*,* $u = w/r$ *and* $\eta\|V\| \le r/5$*, then*

$$\big|\, \|w + \eta V\| - r - \eta\, u^\top V \,\big| \ \le\ \frac{3\eta^2\|V\|^2}{r}.$$

Proof. $\|w + \eta V\| = r\sqrt{1+x}$ with $x = 2\eta u^\top V/r + \eta^2\|V\|^2/r^2$. With $\epsilon := \eta\|V\|/r \le \frac{1}{5}$, $|x| \le 2\epsilon + \epsilon^2 \le 0.44 \le \frac{1}{2}$, so $r\sqrt{1+x} = r + \frac{rx}{2} + r\vartheta$ with $|\vartheta| \le \frac{x^2}{2}$. Now $\frac{rx}{2} = \eta u^\top V + \frac{\eta^2\|V\|^2}{2r}$ and $\frac{rx^2}{2} \le \frac{r(2.2\epsilon)^2}{2} = 2.42\,\frac{\eta^2\|V\|^2}{r}$. Summing, the error is $\le 3\eta^2\|V\|^2/r$. □

Lemma B.6 (From squared to unsquared error). *Let $x, x_+ \geq 0$, $a \in \mathbb{R}$ with $\eta|a| \leq \frac{1}{2}$, $b \geq 0$, and suppose $x_+^2 \leq x^2 - 2\eta a x^2 + 2\eta b x + 4\eta^2 b^2$. Then $x_+ \leq (1-\eta a)x + 3\eta b$.*

Proof. Put $y := (1-\eta a)x + 3\eta b \geq 0$. Using $(1-\eta a)^2 \geq 1 - 2\eta a$ and $1 - \eta a \geq \frac{1}{2}$,

$$y^2 \geq (1-2\eta a)x^2 + 6\eta b(1-\eta a)x + 9\eta^2 b^2 \geq x^2 - 2\eta a x^2 + 3\eta b x + 9\eta^2 b^2 \geq x_+^2.$$

□

Lemma B.7 (Scalar recursions). *Let $(x_n)_{n \geq n_0}$ be nonnegative. 1. If $x_{n+1} \leq (1-\lambda_n)x_n + g$ for $n \in [n_0, N)$, with $\lambda_n \in [\lambda, 1]$, $\lambda > 0$, $g \geq 0$, then $x_n \leq \max\{x_{n_0}, g/\lambda\}$ for all $n \in [n_0, N]$. 2. If $x_{n+1} \leq (1+\mu)x_n + g$ for $n \in [n_0, N)$, with $\mu, g \geq 0$, then $x_n \leq e^{\mu(n-n_0)}\big(x_{n_0} + g\,(n-n_0)\big)$.*

Proof. (1) If $x_n \leq g/\lambda$ then $x_{n+1} \leq (1-\lambda_n)\frac{g}{\lambda} + g = \frac{g}{\lambda} + g\big(1 - \frac{\lambda_n}{\lambda}\big) \leq \frac{g}{\lambda}$. If $x_n > g/\lambda$ then $x_{n+1} \leq x_n - \lambda_n x_n + g < x_n - \lambda_n \frac{g}{\lambda} + g \leq x_n$. Induct. (2) Unrolling, $x_n \leq (1+\mu)^{n-n_0} x_{n_0} + g\sum_{k<n-n_0}(1+\mu)^k \leq (1+\mu)^{n-n_0}\big(x_{n_0} + g(n-n_0)\big)$ and $(1+\mu)^k \leq e^{\mu k}$. □

Lemma B.8 (Balance ratio). *If $b, r > 0$ and $|b^2 - r^2| \leq \varrho\, br$ with $0 \leq \varrho \leq \frac{1}{2}$, then $\big|\frac{b}{r} - 1\big| \leq \varrho$ and $\big|\frac{b}{r} + \frac{r}{b} - 2\big| \leq 2\varrho^2$.*

Proof. $x := b/r > 0$ satisfies $|x^2 - 1| \leq \varrho x$. If $x \geq 1$: $x - 1 = \frac{x^2-1}{x+1} \leq \frac{\varrho x}{x+1} \leq \varrho$. If $x < 1$: $1 - x = \frac{1-x^2}{1+x} \leq \frac{\varrho x}{1+x} \leq \varrho$. Then $x \geq 1 - \varrho \geq \frac{1}{2}$ and $x + \frac{1}{x} - 2 = \frac{(x-1)^2}{x} \leq 2\varrho^2$. □

B.3 Sub-Gaussian bounds on the working region

We use the standard facts $\|YZ\|_{\psi_1} \leq \|Y\|_{\psi_2}\|Z\|_{\psi_2}$ and $\|v^\top X\|_{\psi_2} \leq C$ for $v \in \mathbb{S}^{d-1}$, plus:

Lemma B.9 (Centering). $\|Y - \mathbb{E}Y\|_{\psi_1} \leq C\|Y\|_{\psi_1}$.

Proof. $\|\cdot\|_{\psi_1}$ is a norm and $\|c\|_{\psi_1} = |c|/\log 2$ for constants. From $t \leq e^t$, $|\mathbb{E}Y| \leq \mathbb{E}|Y| \leq \|Y\|_{\psi_1}\mathbb{E}\big[e^{|Y|/\|Y\|_{\psi_1}}\big] \leq 2\|Y\|_{\psi_1}$. Hence $\|Y - \mathbb{E}Y\|_{\psi_1} \leq (1 + 2/\log 2)\|Y\|_{\psi_1}$.

□

Lemma B.10 (Residual bounds). *If $\Psi \in \mathcal{R}$ and $R = f_\Psi - y$, then*

$$\mathrm{Lip}(f_\Psi) \leq \Pi \leq 6\beta, \qquad \mathrm{Lip}(R) \leq 7\beta,$$
$$|\mathbb{E}R(X)| \leq \frac{\Pi}{\sqrt{2\pi}} \leq 3\beta, \qquad \|R(X)\|_{\psi_2} \leq C\beta,$$
$$\|R\|_{L^2} \leq \frac{\Pi}{\sqrt{2}} + \beta \leq 6\beta.$$

Proof. $\nabla f_\Psi(x) = \frac{1}{\sqrt{m}}\sum_j a_j \mathbf{1}_{\{w_j^\top x > 0\}} w_j$ a.e., so $\|\nabla f_\Psi\| \leq \frac{1}{\sqrt{m}}\sum_j |a_j|\|w_j\| = \Pi$; f_Ψ is Lipschitz as a finite sum of Lipschitz functions, and $\mathrm{Lip}(y) = \beta$. Since $\mathbb{E}y(X) = 0$ and $\mathbb{E}\sigma(w^\top X) = \|w\|/\sqrt{2\pi}$, $|\mathbb{E}R| = |\mathbb{E}f_\Psi| \leq \Pi/\sqrt{2\pi}$. By (E1), $\|R - \mathbb{E}R\|_{\psi_2} \leq C\,\mathrm{Lip}(R) \leq 7C\beta$, and $\|c\|_{\psi_2} \leq C|c|$ for constants. Finally $\|\sigma(u^\top X)\|_{L^2} = 1/\sqrt{2}$ for $u \in \mathbb{S}^{d-1}$.

□

Lemma B.11 (Population-gradient bounds). *On $\mathcal{R}$: $|g_j^a| \leq C\beta r_j/\sqrt{m}$ and $\|g_j^w\| \leq C\beta b_j/\sqrt{m}$.*

Proof. By Cauchy–Schwarz,

$$|g_j^a| \leq \frac{r_j}{\sqrt{m}}\|R\|_2\|\sigma(u_j^\top X)\|_2, \qquad |v^\top g_j^w| \leq \frac{b_j}{\sqrt{m}}\|R\|_2\|v^\top X\|_2$$

for every unit vector v. Apply Lemma B.10. □

B.4 ADAPTIVE MINI-BATCH CONCENTRATION

Lemma B.12 (Uniform adaptive mini-batch concentration)**.** *There is a universal $C_\star \geq 1$ such that for every integer $N \geq 1$ and every $\delta \in (0,1)$, setting*

$$D_N := d + \log \frac{8mN}{\delta}, \qquad \rho_B := C_\star\Big(\sqrt{\tfrac{D_N}{B}} + \tfrac{D_N}{B}\Big),$$

the event $\mathcal{E}$ that for every $0 \leq n < N$ with $\Psi_n \in \mathcal{R}$ and every $j \leq m$

$$\big|\widehat{g}^a_{j,n} - g^a_{j,n}\big| \leq \frac{\beta r_{j,n}}{\sqrt{m}}\rho_B, \qquad \big\|\widehat{g}^w_{j,n} - g^w_{j,n}\big\| \leq \frac{\beta b_{j,n}}{\sqrt{m}}\rho_B \tag{4.1}$$

satisfies $\mathbb{P}(\mathcal{E}) \geq 1 - \delta$.

Proof. Fix $n < N$ and $j \leq m$; condition on $\mathcal{F}_n$ and work on the $\mathcal{F}_n$-measurable event $\{\Psi_n \in \mathcal{R}\}$, on which $R_n, a_{j,n}, w_{j,n}, r_{j,n}, b_{j,n}$ are constants and the batch $(X_{n,b})_b$ is i.i.d. γ_d independent of $\mathcal{F}_n$.

Second layer. Let $Y_b := \frac{1}{\sqrt{m}} R_n(X_{n,b})\sigma(w_{j,n}^\top X_{n,b})$, so $\widehat{g}^a_{j,n} = \frac{1}{B}\sum_b Y_b$ and $\mathbb{E}[Y_b \mid \mathcal{F}_n] = g^a_{j,n}$. By Lemma B.10 and $\|\sigma(w_j^\top X)\|_{\psi_2} \leq r_j \|u_j^\top X\|_{\psi_2} \leq C r_j$,

$$\|Y_b\|_{\psi_1} \leq \frac{1}{\sqrt{m}}\|R_n(X)\|_{\psi_2}\|\sigma(w_{j,n}^\top X)\|_{\psi_2} \leq \frac{C_1'\beta r_{j,n}}{\sqrt{m}},$$

and by Lemma B.9 the centered summand satisfies $\|Y_b - \mathbb{E}Y_b\|_{\psi_1} \leq \frac{C_1\beta r_{j,n}}{\sqrt{m}} =: K_a$. By (E2) there is a universal $c_2 \in (0,1]$ with

$$\mathbb{P}\Big(\big|\tfrac{1}{B}\textstyle\sum_b (Y_b - \mathbb{E}Y_b)\big| > t\Big) \leq 2\exp\Big[-c_2 B \min\big(\tfrac{t^2}{K_a^2}, \tfrac{t}{K_a}\big)\Big].$$

Set $t := K_a\, c_2^{-1}\big(\sqrt{u/B} + u/B\big)$ and let $\theta_0 := t/K_a \geq c_2^{-1}\max\{\sqrt{u/B},\, u/B\}$. Case $u \leq B$: $\theta_0^2 \geq c_2^{-2}u/B \geq c_2^{-1}u/B$ and $\theta_0 \geq c_2^{-1}u/B$, so $c_2 B\min(\theta_0^2, \theta_0) \geq u$. Case $u > B$: $\theta_0 \geq c_2^{-1}u/B > c_2^{-1} \geq 1$, hence $\theta_0^2 \geq \theta_0$ and $\min(\theta_0^2, \theta_0) = \theta_0 \geq c_2^{-1}u/B$, so again $c_2 B\min(\cdot) \geq u$. Therefore

$$\mathbb{P}\Big(\big|\widehat{g}^a_{j,n} - g^a_{j,n}\big| > \tfrac{C_3\beta r_{j,n}}{\sqrt{m}}\big(\sqrt{u/B} + u/B\big)\;\Big|\;\mathcal{F}_n\Big) \leq 2e^{-u}, \qquad C_3 := C_1/c_2. \tag{4.2}$$

First layer. For fixed $v \in \mathbb{S}^{d-1}$ put $Y_b^v := \frac{1}{\sqrt{m}} R_n(X_{n,b}) a_{j,n}\mathbf{1}_{\{w_{j,n}^\top X_{n,b} > 0\}}(v^\top X_{n,b})$. Since $|\mathbf{1}_{\{\cdot\}}v^\top X| \leq |v^\top X|$ pointwise, $\|Y_b^v\|_{\psi_1} \leq \frac{C_1'\beta b_{j,n}}{\sqrt{m}}$ uniformly in v, so 4.2 holds for $|v^\top(\widehat{g}^w_{j,n} - g^w_{j,n})|$ with $r_{j,n}$ replaced by $b_{j,n}$. Take a $\frac{1}{2}$-net $\mathcal{N}$ of $\mathbb{S}^{d-1}$ with $|\mathcal{N}| \leq 5^d$ (E3); then $\|z\| \leq 2\max_{v\in\mathcal{N}} v^\top z$ for every $z \in \mathbb{R}^d$. Applying the scalar bound with u replaced by $u + d\log 5$ at each $v \in \mathcal{N}$ and union bounding, the failure probability is $\leq 2\cdot 5^d e^{-u-d\log 5} = 2e^{-u}$, and on the complement

$$\|\widehat{g}^w_{j,n} - g^w_{j,n}\| \leq \frac{2C_3\beta b_{j,n}}{\sqrt{m}}\Big(\sqrt{\tfrac{u+d\log 5}{B}} + \tfrac{u+d\log 5}{B}\Big). \tag{4.3}$$

Union bound. Take $u := \log\frac{8mN}{\delta}$; since $\log 5 \leq 2$, $u + d\log 5 \leq 2D_N$, so with $C_\star := 4C_3$ the right sides of 4.2, 4.3 are at most $\frac{\beta r_{j,n}}{\sqrt{m}}\rho_B$ and $\frac{\beta b_{j,n}}{\sqrt{m}}\rho_B$. Let $\mathcal{B}_{n,j}$ be the event that 4.1 fails at (n,j) **and** $\Psi_n \in \mathcal{R}$. As $\{\Psi_n \in \mathcal{R}\} \in \mathcal{F}_n$,

$$\mathbb{P}(\mathcal{B}_{n,j}) = \mathbb{E}\big[\mathbf{1}_{\{\Psi_n\in\mathcal{R}\}}\,\mathbb{P}(4.1 \text{ fails at } (n,j) \mid \mathcal{F}_n)\big] \leq 4e^{-u} = \frac{\delta}{2mN}.$$

Union bounding over $j \leq m$ and $n < N$ gives $\mathbb{P}(\mathcal{E}^c) \leq \delta/2 \leq \delta$. □

Remark B.2 (Adaptivity). Adaptivity is handled entirely by the $\mathcal{F}_n$-measurability of $\{\Psi_n \in \mathcal{R}\}$; no stopping-time device is needed. On $\mathcal{E}$, the bounds 4.1 hold at every iteration whose state lies in $\mathcal{R}$, however the trajectory arrived there. The subsequent phase estimates are deterministic on the event $\mathcal{E}$.

Lemma B.13 (Batch-size consequence)**.** *Under (P6) with $N = N_\star$ and $C_B \geq (2C_\star/c_1)^2$, we have $\rho_B \leq c_1\varepsilon/L \leq 1$.*

Proof. $D_{N_\star} \leq D_\star$. From $B \geq C_B\varepsilon^{-2}L^2 D_\star$: $\sqrt{D_\star/B} \leq \varepsilon/(L\sqrt{C_B})$, and $D_\star/B \leq \varepsilon^2/(C_B L^2) \leq \varepsilon/(L\sqrt{C_B})$ since $\varepsilon \leq 1 \leq L$ and $C_B \geq 1$. Hence $\rho_B \leq 2C_\star\varepsilon/(L\sqrt{C_B}) \leq c_1\varepsilon/L$.

□

B.5 BALANCE

Lemma B.14 (First-order balance cancellation)**.** *For 1.1, for every j, every n, and every realization of the batch,*

$$\Delta_{j,n+1} = \Delta_{j,n} + \eta^2 \Big[(\widehat{g}^a_{j,n})^2 - \|\widehat{g}^w_{j,n}\|^2 \Big]. \tag{5.1}$$

Proof. Samplewise, $w_j^\top \left[a_j \mathbf{1}_{\{w_j^\top x>0\}} x \right] = a_j \mathbf{1}_{\{w_j^\top x>0\}} (w_j^\top x) = a_j \sigma(w_j^\top x)$. Averaging over the batch with the common weight $\frac{1}{B\sqrt{m}} R_n(x)$ gives the exact identity

$$a_{j,n}\, \widehat{g}^a_{j,n} = w_{j,n}^\top \widehat{g}^w_{j,n}. \tag{5.2}$$

Expanding,

$$\Delta_{j,n+1} = (a_j - \eta \widehat{g}^a_j)^2 - \|w_j - \eta \widehat{g}^w_j\|^2 = \Delta_{j,n} - 2\eta \big(a_j \widehat{g}^a_j - w_j^\top \widehat{g}^w_j \big) + \eta^2 \big[(\widehat{g}^a_j)^2 - \|\widehat{g}^w_j\|^2 \big],$$

and the $O(\eta)$ term vanishes by 5.2. □

Corollary B.1 (One-step balance drift)**.** *On $\mathcal{E}$, if $\Psi_n \in \mathcal{R}$ then $|\Delta_{j,n+1} - \Delta_{j,n}| \le C_\sharp\, \tau^2 p_{j,n}$ for a universal $C_\sharp$.*

Proof. By Lemma B.11 and 4.1 with $\rho_B \le 1$: $|\widehat{g}^a_j| \le C\beta r_j/\sqrt{m}$, $\|\widehat{g}^w_j\| \le C\beta b_j/\sqrt{m}$. On $\mathcal{R}$, $r_j^2 + b_j^2 = p_j \big(\frac{r_j}{b_j} + \frac{b_j}{r_j} \big) \le \frac{5}{2} p_j$. Insert into 5.1 and use $\eta^2\beta^2/m = \tau^2$. □

Lemma B.15 (Accumulated balance drift)**.** *On $\mathcal{E}$, if $\Psi_k \in \mathcal{R}$ for all $k < n$, then $|\Delta_{j,n}| \le C_\sharp \tau^2 \sum_{k<n} p_{j,k}$.*

Proof. $\Delta_{j,0} = 0$ by 1.2; telescope Corollary B.1.

□

What the phase analysis needs. The phase analysis below only needs two properties of each update: the stochastic gradient must remain close to the population gradient, and the layer imbalance $\Delta_j = b_j^2 - r_j^2$ must change slowly . Specifically, for updates of the form (1.1), suppose that on $\mathcal{R}$,

$$|\widehat{g}^a_j - g^a_j| \le \frac{\beta r_j}{\sqrt{m}} \rho, \qquad \|\widehat{g}^w_j - g^w_j\| \le \frac{\beta b_j}{\sqrt{m}} \rho, \qquad \rho \le \varepsilon/L,$$

and

$$|\Delta_j^+ - \Delta_j| \le C_\sharp \tau^2 p_j.$$

Here g^a_j, g^w_j denote the population gradients at the current state. Under the same step-size and initialization conditions, the phase analysis below applies with ρ in place of ρ_B. The second bound controls the accumulated layer imbalance.

The preceding lemmas establish these bounds in the fresh-batch setting. For fixed-dataset SGD, Lemma F.4 establishes their blockwise counterparts, with normalized step $\omega = K\tau$ in place of τ.

B.6 PHASE I: PERTURBED HYPERBOLIC DYNAMICS

The purpose of Phase I is to show that the teacher creates a uniform longitudinal force before the network itself becomes large enough to interfere. The pair (a_j, α_j) then behaves like a perturbed two-dimensional hyperbolic system: one mode grows and the other decays. The transverse component receives only the small network and sampling errors.

Set $\alpha_{j,n} := \theta_\star^\top w_{j,n}$, $w^\perp_{j,n} := (I - \theta_\star \theta_\star^\top) w_{j,n}$, and $\tau_1 := \inf\{n \ge 0 : \|f_n\|_{L^2(\gamma_d)} > \beta\varepsilon^2\}$.

Lemma B.16 (Phase-I dynamics)**.** *Work on $\mathcal{E}$. If $n < \min(\tau_1, N_\star)$ and $\Psi_n \in \mathcal{R}$, then for every j*

$$a_{j,n+1} = a_{j,n} + \eta\kappa\alpha_{j,n} + \eta e^a_{j,n}, \qquad \alpha_{j,n+1} = \alpha_{j,n} + \eta\kappa a_{j,n} + \eta e^\alpha_{j,n}, \tag{6.1}$$

$$|e^a_{j,n}| \le C\kappa\iota\, r_{j,n}, \qquad |e^\alpha_{j,n}| \le C\kappa\iota\, b_{j,n}, \qquad \|w^\perp_{j,n+1} - w^\perp_{j,n}\| \le C\eta\kappa\iota\, b_{j,n}. \tag{6.2}$$

If in addition $\Psi_k \in \mathcal{R}$ and $k < \min(\tau_1, N_\star)$ for all $k \le n$, then

$$b_{j,n+1} + r_{j,n+1} \le 2\sqrt{p_0}\, (1+h)^{n+1} e^{C\iota h(n+1)}. \tag{6.3}$$

Proof. Second layer. $-g_j^a = \frac{1}{\sqrt{m}}\mathbb{E}[(y - f_n)\sigma(w_j^\top X)]$. By Lemma B.2(i),

$$\frac{1}{\sqrt{m}}\mathbb{E}[y\,\sigma(w_j^\top X)] = \frac{\beta r_j}{\sqrt{m}}\mathbb{E}[(\theta_\star^\top X)\sigma(u_j^\top X)] = \frac{\beta r_j}{2\sqrt{m}}\theta_\star^\top u_j = \kappa\alpha_j.$$

By Cauchy–Schwarz and $n < \tau_1$, $\frac{1}{\sqrt{m}}|\mathbb{E}[f_n\sigma(w_j^\top X)]| \le \frac{r_j}{\sqrt{m}}\|f_n\|_2\|\sigma(u_j^\top X)\|_2 \le \sqrt{2}\,\kappa\varepsilon^2 r_j$. On $\mathcal{E}$, $|\widehat{g}_j^a - g_j^a| \le \frac{\beta r_j}{\sqrt{m}}\rho_B = 2\kappa\rho_B r_j$. Since $a_{j,n+1} = a_{j,n} - \eta\widehat{g}_j^a$, this gives 6.1–6.2 for e^a.

Radial first layer. $-g_j^w = \frac{a_j}{\sqrt{m}}\mathbb{E}[(y-f_n)\mathbf{1}_{\{w_j^\top X>0\}}X]$. By Lemma B.2(ii), $\mathbb{E}[y\mathbf{1}_{\{u_j^\top X>0\}}X] = \frac{\beta}{2}\theta_\star$, so

$$-\theta_\star^\top g_j^w = \kappa a_j - \frac{a_j}{\sqrt{m}}\mathbb{E}\big[f_n\mathbf{1}_{\{u_j^\top X>0\}}\theta_\star^\top X\big], \qquad \big|\mathbb{E}[f_n\mathbf{1}\,\theta_\star^\top X]\big| \le \|f_n\|_2\|\theta_\star^\top X\|_2 \le \beta\varepsilon^2.$$

With $|\theta_\star^\top(\widehat{g}_j^w - g_j^w)| \le \frac{\beta b_j}{\sqrt{m}}\rho_B$ this gives 6.1–6.2 for e^α.

Transverse. $(I - \theta_\star\theta_\star^\top)\frac{\beta}{2}\theta_\star = 0$, and by duality

$$\|(I - \theta_\star\theta_\star^\top)g_j^w\| \le \frac{b_j}{\sqrt{m}}\sup_{v\perp\theta_\star,\|v\|=1}\big|\mathbb{E}[f_n\mathbf{1}_{\{u_j^\top X>0\}}v^\top X]\big| \le \frac{b_j}{\sqrt{m}}\|f_n\|_2 \le 2\kappa\varepsilon^2 b_j;$$

adding the noise $2\kappa\rho_B b_j$ gives the last bound in 6.2.

Growth. From 6.1–6.2 and $|\alpha_j| \le r_j$: $b_{j,k+1} \le b_{j,k} + h(1 + C\iota)r_{j,k}$. Also $\|g_j^w\| \le \frac{b_j}{\sqrt{m}}\big(\frac{\beta}{2} + \beta\varepsilon^2\big)$, so with the noise $\|\widehat{g}_j^w\| \le \kappa b_j(1 + C\iota)$ and $r_{j,k+1} \le r_{j,k} + h(1 + C\iota)b_{j,k}$. Adding, $b + r$ grows per step by a factor $\le 1 + h(1 + C\iota) \le (1 + h)e^{C\iota h}$; iterate from $b_{j,0} + r_{j,0} = 2\sqrt{p_0}$. □

B.7 End of Phase I

Lemma B.17 (Horizon arithmetic)**.** *For $0 < h \le \frac{1}{2}$ and $L \ge 1$:*

$$L \le hN_1 \le 2L + h \le 3L,$$
$$(1 + h)^{N_1} \in [\varepsilon^{-1}, 2\varepsilon^{-1}], \qquad (1 - h)^{N_1} \le \varepsilon,$$
$$N_1 \le \frac{2L}{h} + 1 = \frac{4L}{\tau} + 1.$$

Proof. By definition $N_1\log(1 + h) \in [L,\ L + \log(1 + h)]$. Since $\log(1 + h) \le h$, $hN_1 \ge N_1\log(1 + h) \ge L$. Since $\log(1 + h) \ge h/2$ for $h \in (0, 1]$, $N_1 \le \frac{L}{\log(1+h)} + 1 \le \frac{2L}{h} + 1$, whence $hN_1 \le 2L + h \le 3L$ (using $h \le 1 \le L$). Also $(1 + h)^{N_1} = e^{N_1\log(1+h)} \in [e^L,\ e^L(1 + h)] \subseteq [\varepsilon^{-1}, 2\varepsilon^{-1}]$, and $(1 - h)^{N_1} \le e^{-hN_1} \le e^{-L} = \varepsilon$. □

Proposition B.1 (End of Phase I)**.** *Under (P1)–(P6), on the event $\mathcal{E}$:*

1. *$N_1 < \tau_1$ and $\Psi_n \in \mathcal{R}$ for all $0 \le n \le N_1$;*
2. *$\operatorname{sign}(a_{j,n}) = s_j$ and $b_{j,n}, r_{j,n} > 0$ for all j and $0 \le n \le N_1$;*
3. *$|\Delta_{j,n}|/p_{j,n} \le C_4\tau$ for all j and $0 \le n \le N_1$;*
4. *$\theta_{j,N_1} \le C_1\varepsilon$ for all j; hence $W_{s,N_1} \le Z_{s,N_1} \le C_1\varepsilon$ for $s = \pm 1$;*
5. *$c_5\,p_0\varepsilon^{-2} \le p_{j,N_1} \le C_2'\,p_0\varepsilon^{-2}$;*
6. *$c_6\,c_0\beta\varepsilon^2 \le P_{s,N_1} \le C_6'\,c_0\beta\varepsilon^2$ for $s = \pm 1$.*

The constants $C_1, C_2', C_4, c_5, c_6, C_6'$ do not depend on c_0; item 1 requires $c_0 \le c_0^{\max}$ for a universal $c_0^{\max}$.

Proof. Strong induction on n with hypothesis

$$\mathcal{I}_n: \qquad \forall k \le n:\ \ \Psi_k \in \mathcal{R},\ \ \|f_k\|_2 \le \beta\varepsilon^2,\ \ \operatorname{sign}(a_{j,k}) = s_j,\ \ b_{j,k}, r_{j,k} > 0.$$

Base $n = 0$. $\Pi_0 = \frac{m}{\sqrt{m}} p_0 = \sqrt{m}\, p_0 = c_0 \beta \varepsilon^4 \le 6\beta$; $b_{j,0}/r_{j,0} = 1$; $\|f_0\|_2 \le \Pi_0/\sqrt{2} \le \beta\varepsilon^2$; signs and positivity by 1.2.

Induction step. Assume $\mathcal{I}_n$ for some $n < N_1$. Set $A_{j,k} := s_j a_{j,k}$, $B_{j,k} := s_j \alpha_{j,k}$, $G_{j,k} := \frac{1}{2}(A+B)$, $H_{j,k} := \frac{1}{2}(A-B)$. By 1.2, $A_{j,0} = \sqrt{p_0}$ and $|B_{j,0}| = \sqrt{p_0}|\langle u_j^0, \theta_\star\rangle| \le \frac{1}{2}\sqrt{p_0}$, so

$$G_{j,0}, H_{j,0} \in \left[\tfrac{1}{4}\sqrt{p_0},\ \tfrac{3}{4}\sqrt{p_0}\right]. \tag{7.1}$$

Lemma B.16 applies at every $k \le n$ and gives

$$G_{j,k+1} = (1+h)G_{j,k} + \eta E^G_{j,k}, \qquad H_{j,k+1} = (1-h)H_{j,k} + \eta E^H_{j,k},$$
$$|E^G_{j,k}| + |E^H_{j,k}| \le C\kappa\iota\,(b_{j,k} + r_{j,k}) \le C\kappa\iota\sqrt{p_0}(1+h)^k, \tag{7.2}$$

where we used 6.3 together with $hk \le hN_1 \le 3L$ (Lemma B.17) and $e^{3C\iota L} \le 2$, which holds by (S2) below.

(a) G grows. Dividing by $(1+h)^{k+1}$ and telescoping,

$$\begin{aligned}\left|(1+h)^{-k}G_{j,k} - G_{j,0}\right| &\le \eta \sum_{l<k} (1+h)^{-(l+1)} |E^G_{j,l}| \\ &\le C\eta\kappa\iota\sqrt{p_0}\, k = Ch\iota\sqrt{p_0}\, k \\ &\le 3C\iota L\sqrt{p_0} \le \tfrac{1}{8}\sqrt{p_0}.\end{aligned}$$

the last step by (S2). With 7.1, for all $k \le n+1$,

$$\tfrac{1}{8}\sqrt{p_0}(1+h)^k \ \le\ G_{j,k} \ \le\ \sqrt{p_0}(1+h)^k. \tag{7.3}$$

(b) H decays. $H_{j,k} = (1-h)^k H_{j,0} + \eta \sum_{l<k}(1-h)^{k-1-l} E^H_{j,l}$, and

$$\sum_{l<k}(1-h)^{k-1-l}(1+h)^l \le (1+h)^{k-1} \sum_{i\ge 0}\left(\tfrac{1-h}{1+h}\right)^i = (1+h)^{k-1}\cdot \tfrac{1+h}{2h} = \tfrac{(1+h)^k}{2h},$$

so, using $\eta\kappa = h$,

$$|H_{j,k} - (1-h)^k H_{j,0}| \le C\iota\sqrt{p_0}(1+h)^k, \qquad |H_{j,k}| \le \tfrac{3}{4}\sqrt{p_0}(1-h)^k + C\iota\sqrt{p_0}(1+h)^k. \tag{7.4}$$

(c) Sign stability and non-vanishing at $n+1$. Since $H_{j,0} > 0$, 7.4 gives $H_{j,k} \ge -C\iota\sqrt{p_0}(1+h)^k$, so with 7.3

$$b_{j,k} = A_{j,k} = G_{j,k} + H_{j,k} \ \ge\ \left(\tfrac{1}{8} - C\iota\right)\sqrt{p_0}(1+h)^k \ \ge\ \tfrac{1}{16}\sqrt{p_0}(1+h)^k > 0 \qquad (k \le n+1), \tag{7.5}$$

using (S2). Also $r_{j,n+1} \ge r_{j,n} - \eta\|\widehat{g}^w_{j,n}\| \ge r_{j,n}\left(1 - C\tau\,\frac{b_{j,n}}{r_{j,n}}\right) \ge r_{j,n}(1 - 2C\tau) > 0$ by (S1).

(d) Transverse freeze. By 6.2, 6.3 and $\sum_{l<k}(1+h)^l \le (1+h)^k/h$,

$$\|w^\perp_{j,k}\| \le \|w^\perp_{j,0}\| + C\eta\kappa\iota \sum_{l<k} b_{j,l} \le \sqrt{p_0}\left(1 + C\iota(1+h)^k\right) \qquad (k \le n+1). \tag{7.6}$$

(e) Upper bound on p; closing the lazy bound and $\Pi \le 6\beta$. From 7.3–7.4, $b_{j,k} \le 2\sqrt{p_0}(1+h)^k$; and

$$r^2_{j,k} = \alpha^2_{j,k} + \|w^\perp_{j,k}\|^2 \le (G_{j,k} + |H_{j,k}|)^2 + \|w^\perp_{j,k}\|^2 \le C_2\, p_0 (1+h)^{2k},$$

so, by Lemma B.17,

$$p_{j,k} = b_{j,k} r_{j,k} \le C_2\, p_0 (1+h)^{2k} \le 4C_2\, p_0 \varepsilon^{-2} \qquad (k \le n+1 \le N_1), \tag{7.7}$$

with C_2 universal and independent of c_0 (it is built only from the constants in 7.3, 7.4, 7.6). Hence

$$\Pi_k \le \frac{m}{\sqrt{m}} \cdot 4C_2 p_0 \varepsilon^{-2} = 4C_2 c_0 \beta\varepsilon^2, \qquad \|f_k\|_2 \le \Pi_k/\sqrt{2}.$$

Set $c_0^{\max} := 1/(8C_2)$. For $c_0 \le c_0^{\max}$ this gives $\Pi_{n+1} \le \frac{1}{2}\beta\varepsilon^2 \le \beta \le 6\beta$ and $\|f_{n+1}\|_2 \le \frac{1}{2\sqrt{2}}\beta\varepsilon^2 < \beta\varepsilon^2$.

(f) Lower bound on r; balance; closing $b/r \in [\frac{1}{2}, 2]$. By Lemma B.15 (applicable: $\Psi_k \in \mathcal{R}$ for $k \le n$), 7.7, and $\sum_{k\le n}(1+h)^{2k} \le \frac{(1+h)^{2(n+1)}}{(1+h)^2-1} \le \frac{(1+h)^{2(n+1)}}{2h}$, together with $\tau^2/h = 2\tau$,

$$|\Delta_{j,n+1}| \;\le\; C_\sharp \tau^2 \sum_{k\le n} C_2 p_0 (1+h)^{2k} \;\le\; C_3' \,\tau\, p_0 (1+h)^{2(n+1)}, \tag{7.8}$$

with C_3' universal and c_0-free. Since $r^2 = b^2 - \Delta$, 7.5 and 7.8 give

$$r_{j,n+1}^2 \;\ge\; \big(\tfrac{1}{256} - C_3'\tau\big) p_0 (1+h)^{2(n+1)} \;\ge\; \tfrac{1}{512}\, p_0 (1+h)^{2(n+1)}$$

by (S1). Hence $p_{j,n+1} \ge c_5\, p_0 (1+h)^{2(n+1)}$ with $c_5 := \frac{1}{16}\cdot\frac{1}{\sqrt{512}}$, and combining with 7.8,

$$\frac{|\Delta_{j,n+1}|}{p_{j,n+1}} \;\le\; \frac{C_3'}{c_5}\,\tau \;=:\; C_4\tau \;\le\; \tfrac{1}{2} \tag{7.9}$$

by (S1). Lemma B.8 now gives $b_{j,n+1}/r_{j,n+1} \in [\frac{1}{2}, 2]$. Together with (c) and (e), $\mathcal{I}_{n+1}$ holds.

This completes the induction and proves items 1, 2, 3. Item 5 follows from 7.7, the lower bound above, and Lemma B.17, with $C_2' := 4C_2$.

(g) Alignment at N_1 (item 4). By 7.3, 7.4, Lemma B.17 and $\iota \le 2\varepsilon$ (from 1.4, $\iota \le \iota L$),

$$B_{j,N_1} = G_{j,N_1} - H_{j,N_1} \;\ge\; \tfrac{1}{8}\tfrac{\sqrt{p_0}}{\varepsilon} - \tfrac{3}{4}\sqrt{p_0}\,\varepsilon - C\iota\cdot 2\tfrac{\sqrt{p_0}}{\varepsilon} \;\ge\; \tfrac{1}{16}\tfrac{\sqrt{p_0}}{\varepsilon} \;>\; 0,$$

using (S2). By 7.6, $\|w_{j,N_1}^\perp\| \le \sqrt{p_0}\big(1 + 2C\iota/\varepsilon\big) \le C\sqrt{p_0}$. Since $s_j\alpha_{j,N_1} = B_{j,N_1} > 0$, we have $\theta_{j,N_1} \in [0, \frac{\pi}{2})$ and

$$\theta_{j,N_1} \le \tan\theta_{j,N_1} = \frac{\|w_{j,N_1}^\perp\|}{B_{j,N_1}} \le 16C\varepsilon =: C_1\varepsilon.$$

p_0 cancels, so C_1 is c_0-free. Finally $Z_{s,N_1} = \max_{j\in S_s}\sin\theta_{j,N_1} \le C_1\varepsilon$ and $W_{s,N_1} \le Z_{s,N_1}$.

(h) Item 6. $P_{s,N_1} = \frac{1}{\sqrt{m}}\sum_{j\in S_s} p_{j,N_1}$ with $|S_s| = m/2$, and $\frac{m}{2\sqrt{m}}p_0\varepsilon^{-2} = \frac{1}{2}c_0\beta\varepsilon^2$; apply item 5. □

B.8 Phase II: mass dynamics

After Phase I, individual directions are already close to their target poles. The remaining macroscopic quantity is the mass carried by each sign class. The next lemma shows that, up to controlled errors, each class follows a logistic recursion with stable target β.

For the remainder of Phase II, we work on $\mathcal{E}$ and on iterations $n \in [N_1, N_2]$ satisfying the **Phase-II hypotheses**

$$\begin{aligned} &\textbf{(H1)}\quad \Theta_n \le \Theta_\star, && \textbf{(H2)}\quad P_{s,n} \le 2\beta \quad (s=\pm),\\ &\textbf{(H3)}\quad \operatorname{sign}(a_{j,n}) = s_j \quad \forall j, && \textbf{(H4)}\quad \frac{|\Delta_{j,n}|}{p_{j,n}} \le C_\Delta\tau L \le \varepsilon \quad \forall j. \end{aligned} \tag{H}$$

(H) implies $\Pi_n \le 4\beta$ and, by Lemma B.8, $b_{j,n}/r_{j,n} \in [1-\varepsilon, 1+\varepsilon] \subseteq [\frac{1}{2}, 2]$; hence $\Psi_n \in \mathcal{R}$ and 4.1 is available at n. Define

$$F_{j,n} := s_j\,\mathbb{E}\big[(y - f_n)\sigma(u_{j,n}^\top X)\big], \qquad \delta_{s,n} := \frac{\beta - P_{s,n}}{2\sqrt{m}}.$$

Lemma B.18 (Phase-II mass dynamics)**.** *Under (H) at iteration n (indices suppressed):*

$$F_j = \frac{\beta\cos\theta_j}{2} - \frac{1}{\sqrt{m}}\sum_{i=1}^m s_i s_j p_i K_{ij}, \qquad \Big|F_j - \frac{\beta - P_{s_j}}{2}\Big| \le C\beta\Theta_\star^2; \tag{8.1}$$

sign stability holds at $n+1$, with $b_j^+ \ge b_j(1 - C\tau) > 0$ and $r_j^+ \ge r_j(1-C\tau) > 0$; and

$$p_{j,n+1} = p_{j,n} + \frac{2\eta}{\sqrt{m}}F_{j,n}\,p_{j,n} + \eta\,p_{j,n}\,\omega_{j,n}, \qquad |\omega_{j,n}| \le C\frac{\beta}{\sqrt{m}}\big(\varepsilon^2 + \rho_B + \tau\big); \tag{8.2}$$

$$P_{s,n+1} = P_{s,n} + \frac{\eta}{\sqrt{m}}P_{s,n}(\beta - P_{s,n}) + \eta P_{s,n}\varrho_{s,n}, \qquad |\varrho_{s,n}| \le C_5\frac{\beta}{\sqrt{m}}\zeta. \tag{8.3}$$

Proof. 8.1. By Lemma B.2(i), $s_j\mathbb{E}[y\,\sigma(u_j^\top X)] = s_j\frac{\beta}{2}\theta_\star^\top u_j = \frac{\beta}{2}\cos\theta_j$; by 1.3 and Lemma B.2(iv), $s_j\mathbb{E}[f_n\sigma(u_j^\top X)] = \frac{1}{\sqrt{m}}\sum_i s_i s_j p_i K_{ij}$.

For i with $s_i = s_j$: both u_i, u_j lie within angle $\Theta_\star$ of $s_j\theta_\star$, so $\psi_{ij} \le \theta_i + \theta_j \le 2\Theta_\star$ by the triangle inequality for the angular metric; Lemma B.3 gives $|K_{ij} - \frac{1}{2}| \le \Theta_\star^2$, hence

$$\Big|\frac{1}{\sqrt{m}}\sum_{i\in S_{s_j}} p_i K_{ij} - \frac{P_{s_j}}{2}\Big| \le \Theta_\star^2\, P_{s_j} \le 2\beta\Theta_\star^2.$$

For i with $s_i = -s_j$: $\psi_{ij} \ge \pi - \theta_i - \theta_j \ge \pi - 2\Theta_\star$, so $0 \le K_{ij} \le \frac{(2\Theta_\star)^3}{6\pi} \le \Theta_\star^3$, and the whole opposite-class contribution is $\le \Theta_\star^3 P_{-s_j} \le 2\beta\Theta_\star^3 \le 2\beta\Theta_\star^2$. Finally $|\cos\theta_j - 1| \le \Theta_\star^2/2$. Summing gives 8.1.

Step 0 (sign stability) . Write $\widehat{V}_j^a := -\widehat{g}_j^a$, $\widehat{V}_j^w := -\widehat{g}_j^w$. By Lemma B.11 and 4.1, $\eta|\widehat{V}_j^a| \le C\tau r_j \le 2C\tau b_j$ and $\eta\|\widehat{V}_j^w\| \le C\tau b_j \le 2C\tau r_j$. By (S1) $2C\tau < 1$, so

$$s_j a_j^+ = s_j(a_j + \eta\widehat{V}_j^a) = b_j + \eta s_j\widehat{V}_j^a \;\ge\; b_j(1-2C\tau) > 0, \qquad r_j^+ \ge r_j(1-2C\tau) > 0.$$

Hence $\operatorname{sign}(a_j^+) = s_j$ and $b_j^+ = s_j a_j^+ = b_j + \eta s_j\widehat{V}_j^a$ exactly.

Step 1 (population radial velocities, exact). With V_j^a, V_j^w the population velocities,

$$\begin{aligned} s_j V_j^a &= \frac{s_j}{\sqrt{m}}\mathbb{E}[(y-f)\sigma(w_j^\top X)] = \frac{r_j}{\sqrt{m}}F_j, \\ u_j^\top V_j^w &= \frac{a_j}{\sqrt{m}}\mathbb{E}[(y-f)\mathbf{1}_{\{u_j^\top X>0\}}u_j^\top X] = \frac{a_j s_j}{\sqrt{m}}F_j = \frac{b_j}{\sqrt{m}}F_j. \end{aligned} \tag{8.4}$$

using $\mathbf{1}_{\{t>0\}}t = \sigma(t)$ and $a_j s_j = b_j$.

Step 2 (one-step expansions). With $\varsigma_j^a := s_j(\widehat{V}_j^a - V_j^a)$, $|\varsigma_j^a| \le \frac{\beta r_j}{\sqrt{m}}\rho_B$:

$$b_j^+ = b_j + \frac{\eta r_j}{\sqrt{m}}F_j + \eta\varsigma_j^a.$$

By Step 0, $\eta\|\widehat{V}_j^w\| \le 2C\tau r_j \le r_j/5$ by (S1), so Lemma B.5 gives $r_j^+ = r_j + \eta u_j^\top\widehat{V}_j^w + \vartheta_j$ with $|\vartheta_j| \le 3\eta^2\|\widehat{V}_j^w\|^2/r_j \le C\tau^2 r_j$. With $\varsigma_j^w := u_j^\top(\widehat{V}_j^w - V_j^w)$, $|\varsigma_j^w| \le \frac{\beta b_j}{\sqrt{m}}\rho_B$:

$$r_j^+ = r_j + \frac{\eta b_j}{\sqrt{m}}F_j + \eta\varsigma_j^w + \vartheta_j.$$

Step 3 (product). Multiplying,

$$p_j^+ = b_j^+ r_j^+ = p_j + \eta\frac{r_j^2 + b_j^2}{\sqrt{m}}F_j + \eta\big(r_j\varsigma_j^a + b_j\varsigma_j^w\big) + b_j\vartheta_j + \Xi_j,$$

where Ξ_j is the sum of products of two increments, each of size $\le C\tau\max(b_j, r_j)$, so $|\Xi_j| \le C\tau^2 p_j$. By Lemma B.8 with $\varrho = \varepsilon$,

$$\frac{r_j^2 + b_j^2}{\sqrt{m}}F_j = \frac{p_j}{\sqrt{m}}\Big(\frac{r_j}{b_j} + \frac{b_j}{r_j}\Big)F_j = \frac{2p_j}{\sqrt{m}}F_j + \frac{p_j}{\sqrt{m}}O(\varepsilon^2)F_j,$$

and $|F_j| \le C\beta$ by 8.1 and (H2). Also $|r_j\varsigma_j^a + b_j\varsigma_j^w| \le \frac{C\beta\rho_B}{\sqrt{m}}p_j$ and $|b_j\vartheta_j| \le C\tau^2 p_j = C\eta\frac{\beta}{\sqrt{m}}\tau p_j$. Collecting gives 8.2.

8.3. Sum 8.2 over $j \in S_s$ and divide by $\sqrt{m}$; insert 8.1. The $O(\beta\Theta_\star^2)$ error contributes $\eta P_s \cdot C\frac{\beta\Theta_\star^2}{\sqrt{m}}$, and $\max_j|\omega_j| \le C\frac{\beta}{\sqrt{m}}(\varepsilon^2 + \rho_B + \tau)$; since $\varepsilon^2 \le \Theta_\star^2$ (Lemma B.1 gives $\Theta_\star \ge A\varepsilon \ge \varepsilon$), both are $\le C\frac{\beta}{\sqrt{m}}\zeta$. □

Remark B.3 (Role of the angular error). The imbalance enters 8.2 only quadratically, since $x + x^{-1} - 2 = (x-1)^2/x$; in the angular transverse recursion it enters linearly. This asymmetry is why the mass dynamics tolerates a much larger imbalance than the angular dynamics.

B.9 LOGISTIC COMPARISON

Lemma B.19 (Logistic comparison). *Assume (H) at every iteration of $[N_1, n]$ with $n \le N_2 - 1$, and $\zeta \le c_\zeta$ (see (S3)). Assume also, for a fixed universal $c_{\mathrm{in}} > 0$, that for each $s \in \{\pm 1\}$*

$$c_{\mathrm{in}} c_0 \beta \varepsilon^2 \le P_{s,N_1} \le \frac{\beta}{2}, \qquad \log \frac{1}{2c_{\mathrm{in}} c_0} \le L.$$

Put $q := \eta/\sqrt{m}$ (so $q\beta = \tau$), $T_B(p) := p + qp(B-p)$, and $\beta_\pm := \beta(1 \pm C_5\zeta)$ with C_5 from 8.3. Then for each $s \in \{\pm 1\}$: 1. (sandwich) $P_k^- \le P_{s,k} \le P_k^+$ for $N_1 \le k \le n+1$, where $P_{k+1}^\pm = T_{\beta_\pm}(P_k^\pm)$ and $P_{N_1}^\pm = P_{s,N_1}$; 2. (no overshoot) $P_{s,k} \le \beta_+ \le \frac{3}{2}\beta$ for $N_1 \le k \le n+1$; 3. (time to half) $P_k^- \ge \beta/2$ for some $k \le N_1 + C_6\tau^{-1}L$ with $C_6 := 25$, provided $n+1$ is at least that index; 4. (endpoint accuracy) if $C_L \ge C_7 := C_6 + 2$ and $n+1 = N_2$, then $|P_{s,N_2} - \beta| \le \beta(\bar\varepsilon + C_5\zeta) \le C\beta\bar\varepsilon$.

Proof. By 8.3, $P_{k+1} = T_{\kappa_k}(P_k)$ with $\kappa_k := \beta + \sqrt{m}\varrho_{s,k} \in [\beta_-, \beta_+]$.

(1)–(2). For $p \in [0, 3\beta]$, $\partial_p T_B(p) = 1 + q(B - 2p) \ge 1 - 6q\beta = 1 - 6\tau \ge \frac{1}{2}$ by (S1); so T_B is strictly increasing in p on $[0, 3\beta]$, and it is increasing in B for $p > 0$. If $0 \le P \le \beta_+$ then $\beta_+ - T_{\beta_+}(P) = (\beta_+ - P)(1 - qP) \ge 0$ since $qP \le 2\tau < 1$; so $[0, \beta_+]$ is forward invariant for T_{β_+}. As $P_{N_1}^+ = P_{s,N_1} \le \beta/2 \le \beta_+$, we get $P_k^+ \le \beta_+$ for all k. Now induct: if $P_k^- \le P_k \le P_k^+$ with all three in $[0, 3\beta]$, then by the two monotonicities $P_{k+1} = T_{\kappa_k}(P_k) \in [T_{\beta_-}(P_k^-), T_{\beta_+}(P_k^+)] = [P_{k+1}^-, P_{k+1}^+]$, and $P_{k+1}^+ \le \beta_+ \le \frac{3}{2}\beta$, $P_{k+1}^- \ge 0$. (Here $\beta_+ \le \frac{3}{2}\beta$ by (S3).)

(3). While $P_k^- \le \beta/2$, and $\beta_- \ge \frac{3}{4}\beta$ by (S3),

$$P_{k+1}^- \ge P_k^-\big(1 + q(\beta_- - \tfrac{\beta}{2})\big) \ge P_k^-\big(1 + \tfrac{\tau}{4}\big).$$

From $P_{N_1}^- = P_{s,N_1} \ge c_{\mathrm{in}} c_0 \beta \varepsilon^2$, the number of steps needed to exceed $\beta/2$ is at most

$$\frac{\log\big(\frac{1}{2c_{\mathrm{in}} c_0 \varepsilon^2}\big)}{\log(1 + \tau/4)} \le \frac{8}{\tau}\Big(2L + \log \tfrac{1}{2c_{\mathrm{in}} c_0}\Big) \le \frac{24L}{\tau},$$

using $\log(1 + x) \ge x/2$ on $[0, 1]$ and the displayed assumption. So $C_6 = 25$ suffices.

(4). Let k_0 be the first index $\ge N_1$ with $P_{k_0}^- \ge \beta/2$; by (3), $k_0 \le N_1 + C_6\tau^{-1}L$. If $k_0 = N_1$ then $P_{k_0}^- = \beta/2 \le \beta_-$. Otherwise $P_{k_0-1}^- < \beta/2$ and T_{β_-} is increasing in p,

$$P_{k_0}^- = T_{\beta_-}(P_{k_0-1}^-) \le T_{\beta_-}(\tfrac{\beta}{2}) = \tfrac{\beta}{2} + q\tfrac{\beta}{2}\big(\beta_- - \tfrac{\beta}{2}\big) \le \tfrac{\beta}{2}(1 + \tau) \le \beta_-,$$

by (S1) and (S3); hence $P_{k_0}^- \in [\frac{\beta}{2}, \beta_-]$. This interval is forward invariant for T_{β_-}: for P in it, $T_{\beta_-}(P) \ge P \ge \frac{\beta}{2}$ (since $\beta_- \ge P$) and $T_{\beta_-}(P) \le \beta_-$ (by the identity in (2)). Put $D_k := \beta_- - P_k^- \ge 0$ for $k \ge k_0$; then

$$D_{k+1} = \beta_- - P_k^- - qP_k^-(\beta_- - P_k^-) = D_k\big(1 - qP_k^-\big) \le D_k\big(1 - \tfrac{\tau}{2}\big),$$

so $D_k \le \beta e^{-\tau(k-k_0)/2}$, and $D_k \le \beta\bar\varepsilon$ as soon as $k - k_0 \ge \frac{2}{\tau}\log\frac{1}{\bar\varepsilon}$. Since $\log\frac{1}{\bar\varepsilon} = L - \log L \le L$, it suffices that $k - k_0 \ge 2\tau^{-1}L$, which holds at $k = N_2$ when $C_L \ge C_6 + 2$. Therefore

$$\beta(1 - C_5\zeta) - \beta\bar\varepsilon \;\le\; P_{N_2}^- \;\le\; P_{s,N_2} \;\le\; \beta(1 + C_5\zeta),$$

i.e. $|P_{s,N_2} - \beta| \le \beta(\bar\varepsilon + C_5\zeta)$, and $\zeta \le 3\bar\varepsilon$ by 1.4. □

Corollary B.2 (No mass collapse). *Under the same hypotheses, $p_{j,k} \le e^{C(C_L+1)\zeta L} p_{j,n+1} \le 2\,p_{j,n+1}$ for $N_1 \le k \le n+1$.*

Proof. By 8.1–8.2 and part (2) of Lemma B.19, $\beta - P_{s,k} \ge -C_5\beta\zeta$, so $p_{j,k+1} \ge p_{j,k}(1 - C\tau\zeta)$. By (S3), $C\tau\zeta \le 1/2$. Hence, with $x = C\tau\zeta$, the elementary inequality $(1 - x)^{-1} \le e^{2x}$ gives

$$p_{j,k} \le p_{j,n+1}\exp\big(C\tau\zeta\,(n + 1 - k)\big).$$

There are at most $N_2 - N_1 \le C_L\tau^{-1}L + 1$ Phase-II steps, so after enlarging the universal constant C we obtain $p_{j,k} \le e^{C(C_L+1)\zeta L} p_{j,n+1}$. The final bound by $2p_{j,n+1}$ is (S5). □

B.10 TRANSVERSE RECURSION

Controlling only the largest angle would discard the cancellation created by same-class interactions. We therefore first track the weighted mean-square transverse error V_s (equivalently $W_s = \sqrt{V_s}$), where the gate-overlap matrix is positive semidefinite, and then use this averaged control to bound the worst neuron through Z_s. The following exact recursion is the key input.

Fix s and $j \in S_s$; suppress n; write $P^\perp := I - \theta_\star\theta_\star^\top$.

Lemma B.20 (Transverse recursion)**.** *Under (H) at iteration n, on $\mathcal{E}$:*

$$\begin{aligned} \xi_j^+ - \xi_j &= \eta v_j + \eta \tilde{e}_j, \\ v_j &= -\Lambda_j \xi_j - \frac{1}{m}\sum_{i\in S_s} p_i q_{ij} \xi_i, \\ \Lambda_j &= \frac{\beta\cos\theta_j}{2\sqrt{m}} - \frac{1}{m}\sum_{i\in S_s} p_i q_{ij}\langle u_i, u_j\rangle. \end{aligned} \tag{10.1}$$

$$\|\tilde{e}_j\| \le e_\star := C_9 \frac{\beta}{\sqrt{m}}\zeta; \tag{10.2}$$

$$\delta_s - \frac{\beta\Theta_\star^2}{4\sqrt{m}} \;\le\; \Lambda_j \;\le\; \frac{\beta}{2\sqrt{m}}; \tag{10.3}$$

$$\|v_j\| \le C\frac{\beta}{\sqrt{m}}\big(\|\xi_j\| + W_s\big), \qquad \frac{1}{m}\sum_{i\in S_s} p_i q_{ij}\|\xi_i\| \le \frac{P_s}{2\sqrt{m}} W_s; \tag{10.4}$$

$$\sum_{i,j\in S_s} p_i p_j\, q_{ij}\,\langle \xi_i, \xi_j\rangle \;\ge\; 0. \tag{10.5}$$

Proof. **Step 1 (normalization, explicit remainder).** $w_j^+ = w_j + \eta\widehat{V}_j^w$ with $\widehat{V}_j^w = -\widehat{g}_j^w$. Put $\varpi_j := \eta\widehat{V}_j^w / r_j$; by Lemma B.11, 4.1 and $b_j \le 2r_j$, $\|\varpi_j\| \le C\tau \le \frac{1}{5}$ by (S1). Lemma B.4 gives

$$u_j^+ - u_j = (I - u_j u_j^\top)\varpi_j + E_j^{\mathrm{norm}}, \qquad \|E_j^{\mathrm{norm}}\| \le 14\|\varpi_j\|^2 \le C\eta\tfrac{\beta}{\sqrt{m}}\tau.$$

Splitting $\widehat{V}_j^w = V_j^w + (\widehat{V}_j^w - V_j^w)$ with $\|\widehat{V}_j^w - V_j^w\| \le \frac{\beta b_j}{\sqrt{m}}\rho_B \le \frac{2\beta r_j}{\sqrt{m}}\rho_B$,

$$u_j^+ - u_j = \frac{\eta}{r_j}(I - u_j u_j^\top)V_j^w + \eta\varrho_j, \qquad \|\varrho_j\| \le C\frac{\beta}{\sqrt{m}}(\rho_B + \tau). \tag{10.6}$$

Step 2 (population direction field). $V_j^w = \frac{a_j}{\sqrt{m}}B_j$ with $B_j := \frac{\beta}{2}\theta_\star - \mathbb{E}[f_n \mathbf{1}_{\{u_j^\top X>0\}}X]$, by Lemma B.2(ii). By 1.3 and Lemma B.2(iii),

$$\mathbb{E}\big[f_n \mathbf{1}_{\{u_j^\top X>0\}}X\big] = \frac{1}{\sqrt{m}}\sum_{i=1}^m s_i p_i\Big(q_{ij}u_i + \frac{\sin\psi_{ij}}{2\pi}u_j\Big).$$

The second family of terms is parallel to u_j and is therefore projected to zero by $I - u_j u_j^\top$:

$$(I - u_j u_j^\top)V_j^w = \frac{a_j}{\sqrt{m}}(I - u_j u_j^\top)\Big[\frac{\beta}{2}\theta_\star - \frac{1}{\sqrt{m}}\sum_i s_i p_i q_{ij} u_i\Big].$$

Step 3 (imbalance factor). $\frac{a_j}{r_j}\frac{b_j}{r_j} = s_j(1+\epsilon_j)$ with $|\epsilon_j| \le \varepsilon$, by (H4) and Lemma B.8.

Step 4 (projection onto $\theta_\star^\perp$). Since $\langle u_j, \theta_\star\rangle = s_j\cos\theta_j$,

$$P^\perp(I - u_j u_j^\top)\theta_\star = P^\perp\big(\theta_\star - s_j\cos\theta_j\, u_j\big) = -s_j\cos\theta_j\,\xi_j, \qquad P^\perp(I - u_j u_j^\top)u_i = \xi_i - \langle u_i, u_j\rangle\xi_j.$$

Hence, with $\sigma_{ij} := s_i s_j$ (the outer s_j from Step 3 combining with s_i inside the sum),

$$\xi_j^+ - \xi_j = \eta(1+\epsilon_j)\Big[-\frac{\beta\cos\theta_j}{2\sqrt{m}}\xi_j - \frac{1}{m}\sum_{i=1}^m \sigma_{ij} p_i q_{ij}\big(\xi_i - \langle u_i, u_j\rangle\xi_j\big)\Big] + \eta P^\perp\varrho_j. \tag{10.7}$$

Step 5 (class split). The terms with $i \in S_s$ ($\sigma_{ij} = 1$) are exactly v_j of 10.1. For $i \notin S_s$: $\pi - \psi_{ij} \le \theta_i + \theta_j \le 2\Theta_\star$, so $q_{ij} \le \Theta_\star/\pi$ by Lemma B.3, and using $\|\xi_i\|, \|\xi_j\| \le \Theta_\star$ and $|\langle u_i, u_j\rangle| \le 1$ the opposite-class vector has norm at most

$$\frac{1}{m}\sum_{i\notin S_s} p_i \frac{\Theta_\star}{\pi}\big(\|\xi_i\| + \|\xi_j\|\big) \le \frac{2\Theta_\star^2}{\pi}\cdot\frac{D_{-s}}{m} = \frac{2\Theta_\star^2}{\pi}\cdot\frac{P_{-s}}{\sqrt{m}} \le C\frac{\beta\Theta_\star^2}{\sqrt{m}}.$$

Step 6 (proof of 10.4). For $i, j \in S_s$ we have $\psi_{ij} \le 2\Theta_\star \le \frac{1}{5}$, so $\langle u_i, u_j\rangle = \cos\psi_{ij} \in [0, 1]$, hence $0 \le q_{ij}\langle u_i, u_j\rangle \le \frac{1}{2}$ and $|\Lambda_j| \le \frac{\beta}{2\sqrt{m}} + \frac{P_s}{2\sqrt{m}} \le C\frac{\beta}{\sqrt{m}}$. By weighted Cauchy–Schwarz,

$$\frac{1}{m}\sum_{i\in S_s} p_i q_{ij}\|\xi_i\| \le \frac{1}{2m}\sum_{i\in S_s} p_i\|\xi_i\| \le \frac{D_s}{2m}\Big(\frac{1}{D_s}\sum_{i\in S_s} p_i\|\xi_i\|^2\Big)^{1/2} = \frac{P_s}{2\sqrt{m}}W_s.$$

Combining the two displays gives $\|v_j\| \le C\frac{\beta}{\sqrt{m}}(\|\xi_j\| + W_s)$.

Step 7 (proof of 10.2). From 10.7, $\tilde{e}_j = \epsilon_j v_j + (1+\epsilon_j)\cdot(\text{opposite-class term}) + P^\perp \varrho_j$. By 10.4, $\|\xi_j\|, W_s \le \Theta_\star$, and $\varepsilon \le \Theta_\star$,

$$|\epsilon_j|\,\|v_j\| \le \varepsilon \cdot C\frac{\beta}{\sqrt{m}}\cdot 2\Theta_\star \le C\frac{\beta\Theta_\star^2}{\sqrt{m}}.$$

Together with Step 5 and 10.6, $\|\tilde{e}_j\| \le C_9\frac{\beta}{\sqrt{m}}(\Theta_\star^2 + \rho_B + \tau) = e_\star$.

Step 8 (proof of 10.3). Since $\frac{P_s}{2\sqrt{m}} = \frac{D_s}{2m} = \frac{1}{m}\sum_{i\in S_s} p_i \cdot \frac{1}{2}$,

$$\Lambda_j - \delta_s = \frac{\beta(\cos\theta_j - 1)}{2\sqrt{m}} + \frac{1}{m}\sum_{i\in S_s} p_i\Big(\tfrac{1}{2} - q_{ij}\langle u_i, u_j\rangle\Big). \tag{10.8}$$

The second sum is ≥ 0 by Step 6; the first is $\ge -\frac{\beta\Theta_\star^2}{4\sqrt{m}}$ since $1 - \cos\theta_j \le \theta_j^2/2 \le \Theta_\star^2/2$. This gives the lower bound. The upper bound follows from 10.1 by discarding the subtracted sum, which is nonnegative: $\Lambda_j \le \frac{\beta\cos\theta_j}{2\sqrt{m}} \le \frac{\beta}{2\sqrt{m}}$.

Step 9 (proof of 10.5).) $q_{ij} = \big\langle \mathbf{1}_{\{u_i^\top\cdot>0\}}, \mathbf{1}_{\{u_j^\top\cdot>0\}}\big\rangle_{L^2(\gamma_d)}$, so $(q_{ij})_{i,j\in S_s}$ is a Gram matrix, hence positive semidefinite. Applying this coordinatewise to $z_i := p_i\xi_i \in \mathbb{R}^d$,

$$\sum_{i,j\in S_s} q_{ij}\langle z_i, z_j\rangle = \sum_{k=1}^{d}\sum_{i,j\in S_s} q_{ij} z_i^{(k)} z_j^{(k)} \;\ge\; 0.$$

□

Remark B.4 (Sign of the transverse coefficient). In (10.8), the first term is nonpositive and bounded in magnitude by $\beta\Theta_\star^2/(4\sqrt{m})$, while the second term is nonnegative and can be of order $\beta\Theta_\star/\sqrt{m}$. Thus we cannot conclude that $\Lambda_j - \delta_s$ is nonnegative or that its absolute value is $O(\beta\Theta_\star^2/\sqrt{m})$. The proof uses only the lower bound

$$\Lambda_j \ge \delta_s - \frac{\beta\Theta_\star^2}{4\sqrt{m}},$$

established in (10.3).

B.11 Angular recursions

Lemma B.21 (Angular recursions)**.** *Assume (H) at iteration n and $\zeta \le c_\zeta$. There is a universal $C_D \ge 1$ (chosen after C_5, C_8, C_9) such that, with*

$$a_{s,n} := \delta_{s,n} - C_D\frac{\beta\zeta}{\sqrt{m}}, \qquad b_\star := C_D\frac{\beta}{\sqrt{m}}\zeta,$$

one has $\eta|a_{s,n}| \le C\tau \le \frac{1}{2}$ and

$$W_{s,n+1} \le \big(1 - \eta a_{s,n}\big)W_{s,n} + \eta b_\star, \tag{11.1}$$

$$Z_{s,n+1} \le \big(1 - \eta a_{s,n}\big)Z_{s,n} + \eta\frac{P_{s,n}}{2\sqrt{m}}W_{s,n} + \eta b_\star. \tag{11.2}$$

Proof. Write $N_s := \sum_{j\in S_s} p_j\|\xi_j\|^2$, $D_s := \sum_{j\in S_s} p_j$, $V_s = N_s/D_s$. By Lemma B.18, $p_j^+ = p_j(1+\eta\lambda_j)$ with

$$\lambda_j = \frac{2}{\sqrt{m}}F_j + \omega_j = \lambda_s + \mu_j, \qquad \lambda_s := \frac{\beta - P_s}{\sqrt{m}}, \qquad |\mu_j| \le C\frac{\beta\zeta}{\sqrt{m}}, \qquad |\lambda_j| \le C\frac{\beta}{\sqrt{m}}, \tag{11.3}$$

and by Lemma B.20, $\xi_j^+ = \xi_j + \eta v_j + \eta\tilde{e}_j$ with $\|\tilde{e}_j\| \le e_\star$.

Step 1 (expansion of N_s).

$$N_s^+ - N_s = \underbrace{2\eta\sum_j p_j\langle\xi_j, v_j\rangle}_{\text{(I)}} + \underbrace{2\eta\sum_j p_j\langle\xi_j, \tilde{e}_j\rangle}_{\text{(II)}} + \underbrace{\eta\sum_j p_j\lambda_j\|\xi_j\|^2}_{\text{(III)}} + \mathcal{E}_N,$$

$$\mathcal{E}_N = \eta^2\sum_j p_j(1+\eta\lambda_j)\|v_j + \tilde{e}_j\|^2 + 2\eta^2\sum_j p_j\lambda_j\langle\xi_j, v_j + \tilde{e}_j\rangle,$$

all sums over $j \in S_s$.

Step 2 (second-order terms). By 10.4 and $\sum_j p_j(\|\xi_j\|^2 + W_s^2) = 2D_sV_s$,

$$\sum_j p_j\|v_j\|^2 \le C\frac{\beta^2}{m}\sum_j p_j(\|\xi_j\| + W_s)^2 \le 2C\frac{\beta^2}{m}\sum_j p_j\big(\|\xi_j\|^2 + W_s^2\big) = 4C\frac{\beta^2}{m}D_sV_s.$$

Hence, using $1 + \eta\lambda_j \le 2$ and $\|x+y\|^2 \le 2\|x\|^2 + 2\|y\|^2$,

$$\eta^2\sum_j p_j(1+\eta\lambda_j)\|v_j + \tilde{e}_j\|^2 \le C\eta^2\frac{\beta^2}{m}D_sV_s + 4\eta^2D_se_\star^2.$$

For the cross term, $|\lambda_j| \le C\beta/\sqrt{m}$, $\sum_j p_j\|\xi_j\|(\|\xi_j\| + W_s) \le 2D_sV_s$ (Cauchy–Schwarz for the second piece) and $\sum_j p_j\|\xi_j\| \le D_sW_s$, so

$$2\eta^2\sum_j p_j|\lambda_j|\,\|\xi_j\|\big(\|v_j\| + \|\tilde{e}_j\|\big) \le C\eta^2\frac{\beta^2}{m}D_sV_s + C\eta^2\frac{\beta}{\sqrt{m}}e_\star D_sW_s.$$

Since $\eta^2\beta^2/m = \eta\tau\frac{\beta}{\sqrt{m}}$ and $\eta^2\beta/\sqrt{m} = \eta\tau$,

$$|\mathcal{E}_N| \;\le\; C_8\,\eta\tau\frac{\beta}{\sqrt{m}}D_sV_s \;+\; C_8\,\eta\tau\,e_\star D_sW_s \;+\; 4\eta^2D_se_\star^2. \tag{11.4}$$

The first term is **multiplicative in** V_s (it perturbs the rate by $C_8\tau\beta/\sqrt{m} \le C_8\zeta\beta/\sqrt{m}$); the second is $\le \eta e_\star D_sW_s$ once $C_8\tau \le 1$, which is (S1). The second-order contribution is multiplicative in the current squared angular error, apart from terms controlled by the small forcing. It can therefore be absorbed in the recursion below.

Step 3 (quotient rule). Exactly,

$$V_s^+ - V_s = \rho\cdot\frac{1}{D_s}\Big[(N_s^+ - N_s) - V_s(D_s^+ - D_s)\Big], \qquad \rho := \frac{D_s}{D_s^+}, \quad |\rho - 1| \le C\tau,$$

where $D_s^+ = D_s + \eta\sum_j p_j\lambda_j$. Since $\sum_j p_j(\|\xi_j\|^2 - V_s) = 0$ annihilates the class-common part λ_s,

$$\text{(III)} - V_s(D_s^+ - D_s) = \eta\sum_j p_j\mu_j\big(\|\xi_j\|^2 - V_s\big), \qquad \Big|\frac{1}{D_s}[\cdots]\Big| \le 2\eta\max_j|\mu_j|\,V_s \le C\eta\frac{\beta\zeta}{\sqrt{m}}V_s. \tag{11.5}$$

Step 4 (drift term). By 10.1, 10.3 and 10.5,

$$\frac{\text{(I)}}{D_s} = -\frac{2\eta}{D_s}\sum_j p_j\Lambda_j\|\xi_j\|^2 - \frac{2\eta}{mD_s}\sum_{i,j\in S_s} p_ip_jq_{ij}\langle\xi_i,\xi_j\rangle \;\le\; -2\eta\Big(\delta_s - \frac{\beta\Theta_\star^2}{4\sqrt{m}}\Big)V_s. \tag{11.6}$$

Step 5 (forcing term). $\dfrac{|\text{(II)}|}{D_s} \le 2\eta e_\star\dfrac{1}{D_s}\sum_j p_j\|\xi_j\| \le 2\eta e_\star W_s.$

Step 6 (assembling). By Steps 2–5, with $C_8\tau \le 1$,

$$\frac{1}{D_s}\big[(N_s^+ - N_s) - V_s(D_s^+ - D_s)\big] \le -2\eta a' V_s + 3\eta e_\star W_s + 4\eta^2 e_\star^2,$$
$$a' := \delta_s - \frac{\beta}{\sqrt{m}}\Big(\frac{\Theta_\star^2}{4} + C\zeta + \frac{C_8\tau}{2}\Big).$$

Multiplying by ρ: the nonnegative terms gain a factor $\le 1 + C\tau \le 2$, while

$$-2\eta\rho a' V_s = -2\eta a' V_s + 2\eta(1-\rho)a' V_s \le -2\eta\Big(a' - C\tau\tfrac{\beta}{\sqrt{m}}\Big)V_s,$$

using $|1-\rho| \le C\tau$ and $|a'| \le C\beta/\sqrt{m}$. Hence, since $\tau \le \zeta$ and $\Theta_\star^2 \le \zeta$, for C_D large enough,

$$V_s^+ \;\le\; V_s - 2\eta a_s V_s + 6\eta e_\star W_s + 8\eta^2 e_\star^2.$$

Apply Lemma B.6 with $x = W_s$, $x_+ = W_s^+$, $b := 3e_\star$ (so $6\eta e_\star W_s = 2\eta b W_s$ and $8\eta^2 e_\star^2 \le 4\eta^2 b^2$) and $\eta|a_s| \le C\tau \le \frac{1}{2}$ by (S1): $W_s^+ \le (1-\eta a_s)W_s + 9\eta e_\star$. Finally enlarge C_D so that $9e_\star \le b_\star$, i.e. $C_D \ge 9C_9$; this yields 11.1.

Step 7 (the max, valid also when $\Lambda_j < 0$). Fix $j \in S_s$. By 10.3, $\eta\Lambda_j \le \eta\frac{\beta}{2\sqrt{m}} = \tau/2 < 1$, so $1 - \eta\Lambda_j > 0$ and $\|\xi_j - \eta\Lambda_j\xi_j\| = (1-\eta\Lambda_j)\|\xi_j\|$. (No lower bound on $\eta\Lambda_j$ is claimed: Λ_j may be negative when $P_s > \beta$.) Since $\Lambda_j \ge \delta_s - \frac{\beta\Theta_\star^2}{4\sqrt{m}} \ge a_s$ and $1 - \eta a_s \ge 0$,

$$(1-\eta\Lambda_j)\|\xi_j\| \le (1-\eta a_s)\|\xi_j\| \le (1-\eta a_s)Z_s.$$

By the triangle inequality applied to 10.1, together with 10.4 and 10.2,

$$\|\xi_j^+\| \le (1-\eta a_s)Z_s + \eta\frac{P_s}{2\sqrt{m}}W_s + \eta e_\star.$$

Taking the maximum over $j \in S_s$ and using $e_\star \le b_\star$ gives 11.2. □

B.12 Closing the angular bootstrap

Proposition B.2 (Angular bootstrap). *There are universal constants $C_W, C_\Theta \ge 1$ (independent of A) such that the following holds. Work on $\mathcal{E}$, assume that (H) holds at every iteration of $[N_1, n]$, with $n \le N_2 - 1$, and assume the initial class-mass hypotheses of Lemma B.19. Suppose also that, for a universal C_1,*

$$W_{s,N_1} \le Z_{s,N_1} \le C_1\varepsilon, \qquad P_{s,N_1} < \frac{\beta}{2},$$

then for $s = \pm 1$

$$\max_{N_1 \le k \le n+1} W_{s,k} \le C_W\,\varepsilon, \qquad \max_{N_1 \le k \le n+1} Z_{s,k} \le C_\Theta\,\bar{\varepsilon}. \tag{12.1}$$

Proof. Fix s and use the assumed initial angular bound. Define

$$n_s := \min\{k \in [N_1, n+1] :\; P_{s,k} > \beta/2\}, \qquad \min\emptyset := \infty.$$

Since $P_{s,N_1} < \beta/2$, we have $n_s > N_1$.

Regime I: $N_1 \le k < \min(n_s, n+1)$. Here $P_{s,k} \le \beta/2$, so $\delta_{s,k} \ge \frac{\beta}{4\sqrt{m}}$ and, by (S3),

$$a_{s,k} = \delta_{s,k} - C_D\frac{\beta\zeta}{\sqrt{m}} \;\ge\; \frac{\beta}{8\sqrt{m}},$$

hence $\eta a_{s,k} \ge \tau/8 > 0$; and $\eta a_{s,k} \le \eta\delta_{s,k} \le \eta\frac{\beta}{2\sqrt{m}} = \tau/2 \le 1$. Part (1) of Lemma B.7 applied to 11.1 with $\lambda = \tau/8$, $g = \eta b_\star = \tau C_D\zeta$ gives, for $N_1 \le k \le \min(n_s, n+1)$,

$$W_{s,k} \;\le\; \max\big\{C_1\varepsilon,\; 8C_D\zeta\big\} \;\le\; (C_1 + 24C_D)\,\varepsilon \;=:\; C'_W\varepsilon, \tag{12.2}$$

using $\zeta \le 3\varepsilon$ from 1.4. Feeding (12.2) and $P_{s,k} \le \beta/2$ into 11.2,

$$Z_{s,k+1} \le (1-\eta a_{s,k})Z_{s,k} + \tau\Big(\tfrac{1}{4}C'_W\varepsilon + C_D\zeta\Big),$$

and part (1) of Lemma B.7 gives $Z_{s,k} \le \max\{C_1\varepsilon,\ 8(\frac{1}{4}C'_W + 3C_D)\varepsilon\} =: C'_Z\varepsilon$ on the same range. In particular, if $n_s \le n+1$ then $W_{s,n_s} \le C'_W\varepsilon$ and $Z_{s,n_s} \le C'_Z\varepsilon$.

If $n_s = \infty$, Regime II below is empty and (12.1) already follows from (12.2) and $C'_Z\varepsilon \le C'_Z\bar{\varepsilon}$.

Regime II: $n_s \le k \le n+1$**.** By part (2) of Lemma B.19, $P_{s,k} \le \beta(1 + C_5\zeta)$, so $\delta_{s,k} \ge -\frac{C_5\beta\zeta}{2\sqrt{m}}$ and

$$-\eta a_{s,k} \ \le\ \eta\frac{\beta}{\sqrt{m}}\Big(\tfrac{C_5}{2} + C_D\Big)\zeta \ =:\ \mu \ \ge 0.$$

Also $k - n_s \le N_2 - N_1 \le C_L\tau^{-1}L + 1$. part (2) of Lemma B.7 applied to 11.1:

$$\begin{aligned} W_{s,k} &\le e^{\mu(k-n_s)}\big(W_{s,n_s} + \eta b_\star(k-n_s)\big) \\ &\le e^{C(C_L+1)\zeta L}\Big(C'_W\varepsilon + C_D\zeta(C_L L + \tau)\Big) \\ &\le 2\big(C'_W + 3C_D(C_L+1)\big)\varepsilon =: C_W\varepsilon. \end{aligned} \tag{12.3}$$

using $\zeta L \le 3\varepsilon$ and $\zeta \le 3\varepsilon$ from 1.4, $\tau \le 1$, and $e^{C(C_L+1)\zeta L} \le 2$ from (S5). Since Regime I already gave $W_{s,k} \le C'_W\varepsilon \le C_W\varepsilon$, the first half of (12.1) holds throughout.

Now 11.2 in Regime II, with $P_{s,k} \le 2\beta$ and 12.3:

$$Z_{s,k+1} \le (1+\mu)Z_{s,k} + \tau\big(C_W\varepsilon + C_D\zeta\big).$$

Part (2) of Lemma B.7 gives

$$Z_{s,k} \le e^{\mu(k-n_s)}\Big(Z_{s,n_s} + \tau(C_W\varepsilon + C_D\zeta)(k-n_s)\Big) \le 2\Big(C'_Z\varepsilon + (C_W\varepsilon + 3C_D\varepsilon)(C_L L + \tau)\Big) \le C_\Theta\bar{\varepsilon},$$

$$C_\Theta := 2\big(C'_Z + (C_W + 3C_D)(C_L+1)\big), \tag{12.4}$$

using $\varepsilon L = \bar{\varepsilon}$, $\varepsilon \le \bar{\varepsilon}$ (as $L \ge 1$) and $\tau \le 1$. Regime I gave $Z_{s,k} \le C'_Z\varepsilon \le C_\Theta\bar{\varepsilon}$.

□

Lemma B.22 (Angle comparison)**.** *If $\theta \in [0, \frac{\pi}{2}]$ and $\sin\theta \le z$, then $\theta \le \frac{\pi}{2}z \le 2z$.*

Proof. $\sin\theta \ge \frac{2}{\pi}\theta$ on $[0, \frac{\pi}{2}]$ (concavity of sin). □

Lemma B.23 (No jump to the far hemisphere)**.** *On $\mathcal{E}$, if $\Psi_k \in \mathcal{R}$ then*

$$\|u_{j,k+1} - u_{j,k}\| \le C\tau, \qquad |\theta_{j,k+1} - \theta_{j,k}| \le \tfrac{\pi}{2}C\tau.$$

Consequently, under the smallness conditions $C\tau \le \frac{1}{100}$ (S1) and $\Theta_\star \le \frac{1}{10}$, if $\theta_{j,k} \le \Theta_\star$, then

$$\theta_{j,k+1} \le \Theta_\star + \frac{1}{50} < \frac{\pi}{2}.$$

Proof. Write $w_{j,k+1} = w_{j,k} + \eta\widehat{V}^w_{j,k}$, where $\widehat{V}^w_{j,k} := -\widehat{g}^w_{j,k}$, and put $\varpi_j := \eta\widehat{V}^w_{j,k}/r_{j,k}$. Because $\Psi_k \in \mathcal{R}$, Lemma B.11 and the concentration bound 4.1 give $\|\widehat{V}^w_{j,k}\| \le C\beta b_{j,k}/\sqrt{m} \le 2C\beta r_{j,k}/\sqrt{m}$. Therefore $\|\varpi_j\| \le C\tau \le 1/5$ by (S1). Lemma B.4 now yields

$$\|u_{j,k+1} - u_{j,k}\| \le \|\varpi_j\| + 14\|\varpi_j\|^2 \le C\tau.$$

For unit vectors, $\|u - u'\| = 2\sin(\angle(u,u')/2) \ge \frac{2}{\pi}\angle(u,u')$, so $\angle(u_{j,k+1}, u_{j,k}) \le \frac{\pi}{2}C\tau$. Since $u \mapsto \angle(u, s_j\theta_\star)$ is 1-Lipschitz in the angular metric, the same bound holds for $|\theta_{j,k+1} - \theta_{j,k}|$. At the first exit from $\theta \le \Theta_\star$, the previous angle is at most $\Theta_\star$, so the one-step bound gives $\theta_{j,k} \le \Theta_\star + 1/50 < \pi/2$. □

B.13 Closing the balanced baseline

B.13.1 Convention on constants

The named constants below are chosen sequentially, each in terms of constants fixed earlier; no choice depends on a later one. The generic symbols C and c continue to denote universal constants that may change between occurrences. Separately, the proof uses finitely many smallness inequalities, listed as (S1)–(S7); these are imposed only after the constants they involve have been fixed. Since $\tau \le c_\eta\varepsilon/L \le \varepsilon_0$, $\rho_B \le c_1\varepsilon/L \le \varepsilon_0$ and $\zeta \le 3\varepsilon \le 3\varepsilon_0$ (Lemma B.1, with $c_1, c_\eta \le 1$), every occurrence of "for τ small", "for ζ small", etc., in the preceding lemmas refers to one of these.

B.13.2 ORDER OF CHOICE

1. First fix the constants in Proposition B.1, Corollary B.1, and Lemma B.12. They are determined by the Phase-I, concentration, and balance estimates and do not depend on c_0.
2. Choose a universal $c_0 \in (0, \min\{c_0^{\max}, (2C_6')^{-1}\}]$, where $c_0^{\max} = 1/(8C_2)$. Appendix C may reduce this fixed constant and then reduce ε_0 accordingly.
3. $C_6 := 25,\ C_L := C_7 = C_6 + 2 = 27$ (Lemma B.19).
4. C_5 (Lemma B.18); C_9 (Lemma B.20); C_8 (Lemma B.21, Step 2).
5. C_D (Lemma B.21, Step 6): an explicit expression in C_5, C_8, C_9 and universal constants, satisfying in particular $C_D \geq 9C_9$.
6. C_Δ (Step 3 of the proof below): an explicit expression in $C_\sharp, C_4, C_L$.
7. $c_\eta := \min\{1,\ 1/C_\Delta\}$ and $c_1 := 1;\ C_B := (2C_\star/c_1)^2$.
8. $C_W' := C_1 + 24C_D,\ C_Z' := \max\{C_1,\ 2C_W' + 24C_D\},\ C_W := 2(C_W' + 3C_D(C_L+1)),$ $C_\Theta := 2(C_Z' + (C_W + 3C_D)(C_L+1))$ (Proposition B.2). Appendix C may enlarge C_1; in that case these four derived constants are recomputed in the displayed order before A is chosen. They remain independent of A.
9. Choose $A \geq 4C_\Theta + 2C_1 + 1$. Appendix C may enlarge this universal constant when passing to the full initialization class in Assumption 1.
10. ε_0: a sufficiently small positive value satisfying (S1)–(S7) below.
11. $C_T := C_L + 6$.

B.13.3 SMALLNESS CONDITIONS

With c_τ, c_ι, c_ζ denoting explicit constants determined by items 1–8:

- **(S1)** $\tau \leq c_\tau$, where c_τ is small enough for: $2C\tau < 1$ and $2C\tau \leq \frac{1}{5}$ (Lemma B.18, Step 0; Lemma B.20, Step 1; and Proposition B.1 (c)); $C_3'\tau \leq \frac{1}{512}$ and $C_4\tau \leq \frac{1}{2}$ (Proposition B.1 (f)); $\tau \leq \frac{1}{12}$ and $\tau \leq \frac{1}{2}$ (Lemma B.19); $C_8\tau \leq 1$ and $C\tau \leq \frac{1}{2}$ (Lemma B.21); $C\tau \leq \frac{1}{100}$ (Lemma B.23). Guaranteed by $\varepsilon_0 \leq c_\tau$ (since $\tau \leq c_\eta\varepsilon_0 \leq \varepsilon_0$).
- **(S2)** $\iota L \leq c_\iota$, where c_ι is small enough for: $e^{3C\iota L} \leq 2$ 7.2; $3C\iota L \leq \frac{1}{8}$ (Proposition B.1 (a)); $C\iota \leq \frac{1}{16}$ (Proposition B.1 (c), (g)). Guaranteed by $\varepsilon_0 \leq c_\iota/2$ (since $\iota L \leq 2\varepsilon \leq 2\varepsilon_0$ by 1.4).
- **(S3)** $\zeta \leq c_\zeta$, where c_ζ is small enough for: $C_5\zeta \leq \frac{1}{4}$ (so $\beta_- \geq \frac{3}{4}\beta$, $\beta_+ \leq \frac{3}{2}\beta$, Lemma B.19); $C_D\zeta \leq \frac{1}{8}$ (Proposition B.2 Regime I); and $C\zeta \leq \frac{1}{2}$, which together with $\tau \leq 1$ gives $C\tau\zeta \leq \frac{1}{2}$ in Corollary B.2. Guaranteed by $\varepsilon_0 \leq c_\zeta/3$.
- **(S4)** $\varepsilon_0 \leq 2c_6c_0$ (part (3) of Lemma B.19).
- **(S5)** $e^{3C(C_L+1)\varepsilon_0} \leq 2$ (Corollary B.2 and the Regime-II estimate in Proposition B.2, using $\zeta L \leq 3\varepsilon_0$).
- **(S6)** $\varepsilon_0 \leq 10^{-2}$ and $A^2\varepsilon_0 \log^3(1/\varepsilon_0) \leq 1$ (Lemma B.1).
- **(S7)** $c_\eta \leq 1/C_\Delta$ (so $C_\Delta\tau L \leq \varepsilon$, giving the second inequality in (H4)).

All of (S1)–(S6) are upper bounds on ε_0 in terms of constants defined in items 1–9; (S7) is item 7. Hence the choice is consistent and all constants are universal.

Proof for the balanced equal-magnitude initialization. Apply Lemma B.12 with $N = N_\star$; let $\mathcal{E}$ be the resulting event, $\mathbb{P}(\mathcal{E}) \geq 1 - \delta$. By Lemma B.13, $\rho_B \leq c_1\varepsilon/L$; by Lemma B.1, 1.4 holds. All statements below are deterministic consequences valid pointwise on $\mathcal{E}$.

Phase I. Proposition B.1 gives: $\Psi_n \in \mathcal{R}$ and $\|f_n\|_2 \leq \beta\varepsilon^2$ for $0 \leq n \leq N_1$; sign stability and non-vanishing; $|\Delta_{j,n}|/p_{j,n} \leq C_4\tau$; $\theta_{j,N_1} \leq C_1\varepsilon$; $p_{j,N_1} \asymp p_0\varepsilon^{-2}$; $P_{s,N_1} \asymp c_0\beta\varepsilon^2$. Since $A \geq 2C_1 + 1$ and $L \geq 1$,

$$\Theta_{N_1} \leq C_1\varepsilon < \tfrac{1}{2}A\varepsilon L = \tfrac{1}{2}\Theta_\star,$$

so Phase II starts strictly inside the bootstrap region. More precisely, Proposition B.1 item 6 supplies the hypotheses of Lemma B.19 with $c_{\text{in}} = c_6$: the upper bound is at most $\beta/2$ after the universal smallness choices, while (S4) gives $\log(1/(2c_6c_0)) \le L$.

Phase II: first-exit induction. Define

$$n^\dagger := \min\big\{n \in [N_1, N_2] : \text{ (H) fails at } n\big\}, \qquad \min \emptyset := \infty.$$

(H) holds at $n = N_1$: (H1) by the display above; (H2) since $P_{s,N_1} \le \beta$; (H3), (H4) by Proposition B.1, using $C_4\tau \le C_\Delta \tau L$ (as $C_\Delta \ge C_4$ and $L \ge 1$). Hence $n^\dagger > N_1$.

Suppose, for contradiction, that $n^\dagger \le N_2$. Then (H) holds at every $n \in [N_1, n^\dagger - 1]$, so Lemmas B.18, B.20, and B.21 apply at every such n, and Lemma B.19, Corollary B.2 and Proposition B.2 apply with $n = n^\dagger - 1 \le N_2 - 1$.

To rule out the first exit, it suffices to verify at $n^\dagger$ the four bootstrap conditions: sign stability and non-vanishing (H3), class-mass control (H2), layer-balance control (H4), and angular control (H1).

Step 1 – (H3) and non-vanishing. Lemma B.18 Step 0, applied at $n = n^\dagger - 1$, gives $\operatorname{sign}(a_{j,n^\dagger}) = s_j$ and $b_{j,n^\dagger}, r_{j,n^\dagger} > 0$ for every j.

Step 2 – (H2). Part (2) of Lemma B.19 with $n + 1 = n^\dagger$ gives $P_{s,n^\dagger} \le \frac{3}{2}\beta < 2\beta$.

Step 3 – (H4). By Corollary B.2, $p_{j,k} \le 2p_{j,n^\dagger}$ for $N_1 \le k \le n^\dagger$. Telescoping Corollary B.1 from N_1 (legitimate, since $\Psi_k \in \mathcal{R}$ for all $k < n^\dagger$) and using Proposition B.1 item 3 together with $n^\dagger - N_1 \le C_L\tau^{-1}L + 1$,

$$|\Delta_{j,n^\dagger}| \le |\Delta_{j,N_1}| + C_\sharp\tau^2 \sum_{N_1 \le k < n^\dagger} p_{j,k} \le 2C_4\tau\, p_{j,n^\dagger} + 2C_\sharp\tau^2\big(C_L\tau^{-1}L + 1\big)p_{j,n^\dagger} \le C_\Delta\, \tau L\, p_{j,n^\dagger},$$

with $C_\Delta := 2C_4 + 2C_\sharp(C_L + 1)$ (using $\tau \le 1 \le L$). By (S7), $C_\Delta\tau L \le \varepsilon$.

Step 4 – (H1); the contradiction. Proposition B.2 with $n = n^\dagger - 1$ gives $Z_{s,n^\dagger} \le C_\Theta\bar{\varepsilon}$ for both s. By Lemma B.23 (applicable since $\Psi_k \in \mathcal{R}$ for $k < n^\dagger$, $C\tau \le \frac{1}{100}$ by (S1), and $\Theta_\star \le \frac{1}{10}$ by Lemma B.1), every $\theta_{j,n^\dagger} < \frac{\pi}{2}$, so Lemma B.22 applies:

$$\Theta_{n^\dagger} = \max_s \Theta_{s,n^\dagger} \le 2 \max_s Z_{s,n^\dagger} \le 2C_\Theta\bar{\varepsilon} < (4C_\Theta + 2C_1 + 1)\bar{\varepsilon} \le A\bar{\varepsilon} = \Theta_\star.$$

Steps 1–4 show that (H) holds at $n^\dagger$, contradicting its definition. Hence $n^\dagger = \infty$, i.e. **(H) holds at every** $n \in [N_1, N_2]$, and therefore all conclusions of Lemmas B.18–B.21 and Propositions B.1 and B.2 are valid on $[N_1, N_2]$.

Conclusions. Sign stability, non-vanishing, and the balance bound follow on Phase I from Proposition B.1, and on Phase II from Steps 1 and 3 above. Proposition B.2, followed by Lemmas B.22 and B.23, gives $\Theta_{N_\star} \le 2C_\Theta\bar{\varepsilon}$. Part (4) of Lemma B.19 gives $|P_{s,N_\star} - \beta| \le C\beta\bar{\varepsilon}$. The sharper endpoint and population-fitting bounds stated in Theorem 1 follow from Proposition D.1.

Horizon. By Lemma B.17, $N_1 \le 4L/\tau + 1$, and $N_2 - N_1 \le C_L L/\tau + 1$; since $\tau \le 1 \le L$, $N_\star \le (C_L + 6)L/\tau = C_T\frac{\sqrt{m}}{\eta\beta}\log\frac{1}{\varepsilon}$.

All of the above holds on $\mathcal{E}$, of probability at least $1 - \delta$. □

C Independently varying initialization magnitudes

The technical proof in Appendix B treats the exactly balanced, equal-magnitude special case. We now extend the argument to the initialization in Assumption 1. The equal-magnitude proof used positivity of both mode coefficients, which is no longer guaranteed. We instead compare each neuron with an explicit teacher-only trajectory that stays uniformly inside the layer-ratio bounds.

No training update is changed. The symbols $p_0, L, \bar{\varepsilon}, \tau, h = \tau/2, N_1, N_\star$ retain their meanings from the previous proof; p_0 is now a reference scale rather than every neuron's exact initial mass.

Proposition C.1 (Initialization robustness)**.** *Under the step, horizon, and fresh-batch conditions of Appendix B, the finite-time alignment and fitting conclusions hold, after adjusting universal constants,*

for every integer $m \geq 2$ under

$$\frac{m}{4} \leq |S_s| \leq \frac{3m}{4} \quad (s = \pm 1), \qquad a_{j,0} = s_j b_{j,0}, \quad w_{j,0} = r_{j,0} u_j^0, \\ b_{j,0}, r_{j,0} \in [\sqrt{p_0}, \tfrac{5}{4}\sqrt{p_0}], \quad \|u_j^0\| = 1, \quad s_j \langle u_j^0, \theta_\star \rangle \geq -\tfrac{1}{2}. \tag{3}$$

In particular, neither $b_{j,0} = r_{j,0}$, equal neuron masses, nor equal sign-class cardinalities are required. The step and batch orders remain

$$\eta \leq c_\eta \frac{\sqrt{m}\,\varepsilon}{\beta L}, \qquad B \geq C_B \frac{L^2}{\varepsilon^2}\left(d + \log \frac{16 m N_\star}{\delta}\right), \qquad N_\star = N_1 + \lceil C_L L/\tau \rceil .$$

All output signs remain unchanged and every neuron is nonzero on $[0, N_\star]$. The balance conclusions become

$$\frac{|\Delta_{j,n}|}{p_{j,n}} \leq C\big((1+h)^{-2n} + \tau\big), \qquad 0 \leq n \leq N_1, \tag{4}$$

$$\frac{|\Delta_{j,n}|}{p_{j,n}} \leq C\big(\varepsilon^2 + \tau L\big) \leq \varepsilon, \qquad N_1 \leq n \leq N_\star. \tag{5}$$

The final inequality is secured by the universal smallness choices. There is no assertion of $O(\varepsilon)$ balance at initialization.

Proof. We give the changes to Phase I, then verify the inputs to Phase II. The conditional concentration event of Lemma B.12 depends only on the working region $\mathcal{R}$, not on initialization equalities. Its probability is unchanged. Work throughout on this event and put $\iota = \varepsilon^2 + \rho_B$, $q_n = (1+h)^n$ and $\ell_n = (1-h)^n$. Constants below are universal.

A teacher-only comparison. For a fixed neuron suppress j, and set $A_n = s_j a_{j,n}$, $B_n = s_j \theta_\star^\top w_{j,n}$, $G_0 = (b_0 + B_0)/2$, $H_0 = (b_0 - B_0)/2$. Define the exact unperturbed discrete dynamics by

$$\widetilde{A}_n = G_0 q_n + H_0 \ell_n, \qquad \widetilde{B}_n = G_0 q_n - H_0 \ell_n, \qquad \widetilde{w}_n = s_j \widetilde{B}_n \theta_\star + Q_\star w_0, \qquad \widetilde{r}_n = \|\widetilde{w}_n\|.$$

First suppose $B_0 < 0$. The one-sided condition then implies $|B_0| \leq r_0/2$, and the initialization gives

$$\frac{3}{16}\sqrt{p_0} \leq G_0, H_0 \leq \frac{15}{16}\sqrt{p_0}.$$

The elementary horizon bounds in Lemma B.17 give $hN_1 \leq 3L$, $q_{N_1} \in [\varepsilon^{-1}, 2\varepsilon^{-1}]$, and $\ell_n \leq q_n^{-1}$. Thus $\widetilde{A}_n \geq (3/16)\sqrt{p_0} q_n$. Also $\|Q_\star w_0\| \geq (\sqrt{3}/2)\sqrt{p_0}$. If $q_n \leq 8$ this implies $\widetilde{r}_n \geq (\sqrt{3}/16)\sqrt{p_0} q_n$. If $q_n > 8$, use $\widetilde{B}_n \geq \sqrt{p_0}[(3/16)q_n - (15/16)q_n^{-1}] \geq (1/10)\sqrt{p_0} q_n$. Consequently both magnitudes are bounded above and below by universal multiples of $\sqrt{p_0} q_n$ on $[0, N_1]$.

We also need a ratio bound with a strict margin inside $[1/2, 2]$. Let $T = \|Q_\star w_0\|^2$ and $D_0 = b_0^2 - B_0^2 > 0$. Writing $x = b_0/r_0 \in [4/5, 5/4]$ and $z = B_0/r_0 \in [-1/2, 1/2]$ gives

$$\frac{13}{25} \leq \frac{D_0}{T} = \frac{x^2 - z^2}{1 - z^2} \leq \frac{7}{4}.$$

Since $\widetilde{A}_n^2 - \widetilde{B}_n^2 = D_0(1-h^2)^n$ and $N_1 h^2 \leq 3hL$, reducing the universal smallness constants ensures $(1-h^2)^n \geq 25/26$. Therefore

$$\frac{1}{2} \leq \frac{D_0(1-h^2)^n}{T} \leq \frac{7}{4}, \qquad \frac{1}{\sqrt{2}} \leq \frac{\widetilde{A}_n}{\widetilde{r}_n} \leq \frac{\sqrt{7}}{2}.$$

The last ratio squared lies between 1 and $D_0(1-h^2)^n/T$, since its numerator and denominator have the same additional term $\widetilde{B}_n^2$.

If $B_0 \geq 0$, H_0 need not be positive, so we instead compare with

$$A_n^c = G_0 q_n + H_0/q_n, \qquad B_n^c = G_0 q_n - H_0/q_n, \qquad r_n^c = \sqrt{(B_n^c)^2 + \|Q_\star w_0\|^2}.$$

These equal $A_n^c = b_0 \cosh(\log q_n) + B_0 \sinh(\log q_n)$ and $B_n^c = B_0 \cosh(\log q_n) + b_0 \sinh(\log q_n)$. Both A_n^c, B_n^c are nonnegative and nondecreasing, and $(A_n^c)^2 - (r_n^c)^2 = b_0^2 - r_0^2$. Since $r_n^c \geq r_0$,

$$\min(1, b_0/r_0) \leq A_n^c/r_n^c \leq \max(1, b_0/r_0),$$

so the ratio is in $[4/5, 5/4]$. Also $A_n^c \geq (b_0/2)q_n$: if $H_0 \geq 0$ use $G_0 \geq b_0/2$, and otherwise use $H_0/q_n \geq H_0 q_n$. Thus both layer magnitudes are comparable to $\sqrt{p_0}q_n$. Since $\ell_n = q_n^{-1}(1-h^2)^n$,

$$|\ell_n - q_n^{-1}| \leq nh^2 q_n^{-1} \leq 3hLq_n^{-1}.$$

The discrete comparison differs from (A_n^c, B_n^c) by at most $ChL\sqrt{p_0}q_n^{-1}$. For sufficiently small hL, it therefore has the same magnitude bounds and a strict ratio margin inside $[1/2, 2]$. This also covers a direction exactly at its selected teacher pole, where the transverse component is zero. In both cases $G_0 \geq c\sqrt{p_0}$ and $|H_0| \leq C\sqrt{p_0}$; these are the mode bounds needed below.

Closing Phase I without initial balance. Use the same first-exit induction as Proposition B.1, on sign stability, nonzero magnitudes, $\mathcal{R}$, and $\|f_n\|_2 \leq \beta\varepsilon^2$. Before a possible exit, Lemma B.16 applies; its growth bound now starts from $b_0 + r_0 \leq (5/2)\sqrt{p_0}$. Hence $b_n + r_n \leq C\sqrt{p_0}q_n$. Variation of constants in the G, H recursions gives, including a potential first exiting step,

$$\begin{aligned} |G_n - G_0 q_n| &\leq C\iota hn\sqrt{p_0}q_n, \\ |H_n - H_0\ell_n| &\leq C\iota\sqrt{p_0}q_n, \\ \|Q_\star w_n - Q_\star w_0\| &\leq C\iota\sqrt{p_0}q_n. \end{aligned}$$

These follow respectively by dividing the growing-mode recursion by q_n, summing the geometric convolution with ℓ_n, and summing the transverse increments. Thus

$$|A_n - \widetilde{A}_n| + \|w_n - \widetilde{w}_n\| \leq C\iota L\sqrt{p_0}q_n.$$

Choose ε_0 small enough that $C\iota L$ is smaller than the fixed margins in the preceding magnitude and ratio estimates. Then $A_n, r_n > 0$, $b_n/r_n \in (1/2, 2)$, and $cp_0q_n^2 \leq p_n \leq Cp_0q_n^2$. Summing over all neurons gives $\Pi_n \leq C\sqrt{m}p_0\varepsilon^{-2} = Cc_0\beta\varepsilon^2$. Choosing c_0 small closes $\|f_n\|_2 \leq \Pi_n/\sqrt{2} < \beta\varepsilon^2$ and $\Pi_n < 6\beta$. This prevents the first exit and proves the estimates through N_1.

The sharper transverse bound gives $\|Q_\star w_{N_1}\| \leq C\sqrt{p_0}$, using $\iota/\varepsilon \leq C$. The growing and decaying modes give $B_{N_1} \geq c\sqrt{p_0}/\varepsilon > 0$. Hence every endpoint angle is at most $C\varepsilon$. Since each sign class contains between $m/4$ and $3m/4$ neurons, $P_{s,N_1} \asymp c_0\beta\varepsilon^2$ for both signs, with universal constants. No equality of the two masses is used. Indeed, summing $cp_0\varepsilon^{-2} \leq p_{j,N_1} \leq Cp_0\varepsilon^{-2}$ over any class with $m/4 \leq |S_s| \leq 3m/4$ gives these bounds for every integer width. The later weighted recursions use class masses, not equal counts. The count condition is feasible for every $m \geq 2$ by choosing $\lfloor m/2 \rfloor$ and $\lceil m/2 \rceil$ neurons.

The inherited balance defect. The exact samplewise cancellation still holds, but Lemma B.15 must be replaced by

$$|\Delta_{j,n}| \leq |\Delta_{j,0}| + C_\sharp\tau^2 \sum_{k<n} p_{j,k}, \qquad |\Delta_{j,0}| \leq \frac{9}{16}p_0.$$

The geometric mass bounds and $\sum_{k<n} q_k^2 \leq q_n^2/(2h)$ prove (4). At N_1 this is $C(\varepsilon^2 + \tau)$.

Phase II and the full horizon. Let $c_{\mathrm{I}}, C_{\mathrm{I}}, C_{\mathrm{ang}} > 0$ be universal constants such that the Phase-I estimates above give

$$\begin{aligned} c_{\mathrm{I}}c_0\beta\varepsilon^2 \leq P_{s,N_1} \leq C_{\mathrm{I}}c_0\beta\varepsilon^2, \\ \Theta_{N_1} \leq C_{\mathrm{ang}}\varepsilon, \qquad \frac{|\Delta_{j,N_1}|}{p_{j,N_1}} \leq C_{\mathrm{I}}(\varepsilon^2 + \tau). \end{aligned} \tag{6}$$

Reduce the fixed universal constant c_0, if necessary, so that the upper class-mass bound in (6) is at most $\beta/2$. Reduce ε_0 so that $\log(1/(2c_{\mathrm{I}}c_0)) \leq L$. With these substitutions, Lemma B.19 applies with $c_{\mathrm{in}} = c_{\mathrm{I}}$ to each sign class separately: the sandwich and no-overshoot conclusions are unchanged, each lower comparison reaches $\beta/2$ within $25L/\tau$ steps, and the remaining $2L/\tau$ steps give the same endpoint accuracy. Unequal class sizes and unequal values of P_{+,N_1} and P_{-,N_1} are therefore harmless.

The Phase-II first-exit argument is unchanged except for the inherited layer imbalance. Replace (H4) by

$$\textbf{(H4}') \qquad \frac{|\Delta_{j,n}|}{p_{j,n}} \le C_\Delta^{\mathrm{rob}}(\varepsilon^2 + \tau L) \le \varepsilon \qquad \text{for every } j. \tag{7}$$

The proofs of Lemmas B.18, B.20, and B.21 use (H4) only through the consequence $|\Delta_j|/p_j \le \varepsilon$. Hence all Phase-II lemmas and Proposition B.2 remain valid with (H4′).

Enlarge C_1 so that $C_1 \ge C_{\mathrm{ang}}$ and recompute $C'_W, C'_Z, C_W, C_\Theta$ in the order of the master constant list. Then (6), together with the class-mass bounds above, supplies the same Phase-II initial conditions as in Appendix B. Unequal class sizes enter only through these class masses, so no further modification is needed.

It remains only to verify the modified balance condition at a possible first exit. Let $n^\dagger$ be the first failure time of (H1)–(H3) and (H4′). Before $n^\dagger$, Corollary B.2 gives $p_{j,k} \le 2p_{j,n^\dagger}$ for $N_1 \le k \le n^\dagger$. Therefore

$$\begin{aligned}
|\Delta_{j,n^\dagger}| &\le |\Delta_{j,N_1}| + C_\sharp \tau^2 \sum_{N_1 \le k < n^\dagger} p_{j,k} \\
&\le 2C_{\mathrm{I}}(\varepsilon^2 + \tau)p_{j,n^\dagger} + 2C_\sharp \tau^2 (C_L L/\tau + 1)p_{j,n^\dagger} \\
&\le C_\Delta^{\mathrm{rob}}(\varepsilon^2 + \tau L)p_{j,n^\dagger}.
\end{aligned}$$

Since $\tau L \le c_\eta \varepsilon$, choosing $c_\eta \le 1/(2C_\Delta^{\mathrm{rob}})$ and $\varepsilon_0 \le 1/(2C_\Delta^{\mathrm{rob}})$ gives $C_\Delta^{\mathrm{rob}}(\varepsilon^2 + \tau L) \le \varepsilon$. Thus (H4′) also holds at $n^\dagger$.

The other three exit conditions are closed exactly as in the balanced proof: Lemma B.18 gives sign stability and nonvanishing, Lemma B.19 gives the class-mass bound, and Proposition B.2 together with Lemmas B.23 and B.22 gives the strict angular bound. Hence the same first-exit contradiction shows that (H1)–(H3) and (H4′) hold through $N_\star$.

The logistic endpoint estimate and Proposition B.2 now give the class-mass and alignment conclusions. Sign stability, nonvanishing, and (5) follow from the argument above. The horizon bound is unchanged. Appendix D converts these estimates into the stated $O(\beta\varepsilon)$ population-fitting bound.

The constants are chosen in the same sequential order as in Appendix B. The robust Phase-I constants and C_Δ^{rob} are fixed before c_η; the angular constants are then recomputed before choosing A; finally, ε_0 is chosen sufficiently small to satisfy the resulting smallness conditions. These choices also provide the deterministic constants used in Appendix F. □

Corollary C.1 (A teacher-direction-independent random initialization)**.** *Draw iid Rademacher signs s_j, iid uniform directions u_j^0 on $\mathbb{S}^{d-1}$, and independently draw all $b_{j,0}, r_{j,0}$ uniformly on $[\sqrt{p_0}, (5/4)\sqrt{p_0}]$. The finite-time conclusions hold with probability at least*

$$1 - \delta - 2e^{-m/8} - m(5/4)^{-(d-1)/2}.$$

The lower bound is informative when m and d are sufficiently large; for every integer width $m \ge 2$ the deterministic assumptions remain applicable. For example, the initialization failure probability is at most γ if $m \ge 8\log(4/\gamma)$ and $d \ge 1 + 2\log(4m/\gamma)/\log(5/4)$.

Proof. Hoeffding's inequality bounds failure of the sign-count condition by $2e^{-m/8}$. For a uniform direction write $u = g/\|g\|$ with standard Gaussian g. The event $|u^\top \theta_\star| > 1/2$ is, after rotation, $3g_1^2 > \sum_{k=2}^d g_k^2$. Markov's inequality applied to the exponential with coefficient $1/8$ bounds its probability by $2(5/4)^{-(d-1)/2}$. By symmetry the one-sided event $s_j u_j^\top \theta_\star < -1/2$ has half this probability. Union bound over neurons, then combine with the conditional training guarantee: the fresh-batch result above, or Theorem F.1 for fixed-dataset SGD. In the latter case the initialization is drawn independently of the dataset and indices. No initialization step uses the teacher. □

Scope. The bounded interval is a convenient explicit sufficient condition, not an optimized range. Independent Gaussian output weights are not covered: some magnitudes can be arbitrarily small relative to the corresponding hidden norms, so the uniform growing-mode margin used above can fail. The small overall scale remains essential to this proof.

D SHARPER FITTING AND TRANSVERSE GRADIENT BOUNDS

Proposition D.1 (An $O(\varepsilon)$ fitting endpoint)**.** *For the controlled trajectory established in Appendices B–C,*

$$|P_{s,N_\star} - \beta| \le C\beta\varepsilon \quad (s = \pm 1), \qquad \|f_{\Psi_{N_\star}} - y\|_{L^2(\gamma_d)} \le C\beta\varepsilon, \qquad \mathcal{L}(\Psi_{N_\star}) \le C\beta^2\varepsilon^2.$$

Moreover, with $Q_\star = I - \theta_\star\theta_\star^\top$,

$$\mathbb{E}\|Q_\star \nabla f_{\Psi_{N_\star}}(X)\|^2 \le C\beta^2\varepsilon^2, \qquad \sum_{i\ge 2} \lambda_i(G_{f_{\Psi_{N_\star}}}) \le C\beta^2\varepsilon^2.$$

These claims hold both for the balanced special case and for Proposition C.1.

Proof. Lemma B.19 compares P_s with logistic recursions targeting $\beta_\pm = \beta(1 \pm C_5\zeta)$. The lower comparison reaches $\beta/2$ by $k_0 \le N_1 + C_6 L/\tau$ and subsequently has deficit at most $\beta\exp[-\tau(n - k_0)/2]$. Choosing $C_L \ge C_6 + 2$ gives deficit at most $C\beta\varepsilon$ at $N_\star$; the constant allows integer rounding. The smallness budget $\zeta L \le 3\varepsilon$ therefore yields $|P_{s,N_\star} - \beta| \le C\beta\varepsilon$. This retains the exponential estimate instead of weakening it to $\beta\bar{\varepsilon}$.

Write $f^\sharp(x) = P_+\sigma(z) - P_-\sigma(-z)$, where $z = \theta_\star^\top x$, and suppress $N_\star$. Proposition B.2 gives the stronger averaged estimate $W_s^2 = (\sum_{j\in S_s} p_j)^{-1} \sum_{j\in S_s} p_j \sin^2\theta_j \le C\varepsilon^2$. The angular bootstrap keeps every θ_j acute, so $\theta_j \le 2\sin\theta_j$. The ReLU Lipschitz inequality and weighted Cauchy–Schwarz give

$$\|f - f^\sharp\|_2 \le \frac{1}{\sqrt{m}} \sum_j p_j\theta_j \le 2\sum_s P_s W_s \le C\beta\varepsilon.$$

Also $\|f^\sharp - y\|_2 \le |P_+ - \beta| + |P_- - \beta|$, proving the fit bound. Pointwise off gate boundaries,

$$\|Q_\star \nabla f(x)\| \le \frac{1}{\sqrt{m}} \sum_j p_j \sin\theta_j \le \sum_s P_s W_s \le C\beta\varepsilon.$$

Squaring and taking expectations proves the transverse bound. The spectral tail is at most $\mathrm{tr}(Q_\star G_f)$ by the variational principle. No post-$N_\star$ training step is used. □

Appendix F transfers these conclusions to the actual SGD time $T_\star$.

E VARIATIONAL AND NEURONWISE SELECTION

E.1 VARIATIONAL CHARACTERIZATION

Proof of Proposition 2. For active neurons, let $s_j = \mathrm{sign}(a_j)$, $u_j = w_j/\|w_j\|$, and $\theta_j = \angle(u_j, s_j\theta_\star)$. Define

$$c(\Psi) = \mathbb{E}[f_\Psi(X)\theta_\star^\top X], \qquad D(\Psi) = \frac{1}{\sqrt{m}} \sum_{p_j>0} p_j(1 - \cos\theta_j),$$

and

$$H(\Psi) = \frac{1}{2\sqrt{m}} \sum_j (|a_j| - \|w_j\|)^2.$$

The Gaussian identity $\mathbb{E}[\sigma(w^\top X)\theta_\star^\top X] = \theta_\star^\top w/2$ gives

$$\Pi(\Psi) - 2c(\Psi) = D(\Psi)$$

and direct expansion gives

$$\frac{S(\Psi)}{2\sqrt{m}} - 2c(\Psi) = D(\Psi) + H(\Psi) = \frac{1}{2\sqrt{m}} \sum_j \|w_j - a_j\theta_\star\|^2.$$

At an exact fit, $c(\Psi) = \beta$. Since $D, H \ge 0$, these identities imply $\Pi(\Psi) \ge 2\beta$ and $S(\Psi) \ge 4\beta\sqrt{m}$.

Equality in the first bound holds exactly when every active neuron has $\theta_j = 0$, or equivalently $u_j = s_j\theta_\star$. The represented function is then $P_+\sigma(\theta_\star^\top x) - P_-\sigma(-\theta_\star^\top x)$, and exact fitting forces $P_+ = P_- = \beta$. Equality in the second bound also requires $H = 0$, including at inactive neurons, which is equivalent to $w_j = a_j\theta_\star$ for every j. Both lower bounds are attained, for example, by one positive and one negative active neuron with squared magnitudes $\beta\sqrt{m}$, with all remaining raw parameters zero. □

E.2 Neuronwise mass selection

Proof of Theorem 2. We work with the controlled trajectory from Appendices B–C; Appendix F transfers the conclusions to $T_\star$. All constants are independent of $m, d, \beta, \varepsilon, \delta$. We retain the error terms in Phase I rather than only its constant-factor growth conclusions. Write $h = \tau/2$, $q = (1+h)^{N_1}$ and $\iota = \varepsilon^2 + \rho_B$. The growing-mode bound in Proposition B.1, part (a) of its proof, gives

$$G_{j,N_1} = qG_j^0 + O(\iota L\sqrt{p_0}q).$$

Proof (b) gives $H_{j,N_1} = (1-h)^{N_1}H_{j,0} + O(\iota\sqrt{p_0}q)$, and part (d) of the same proof gives $\|Q_\star w_{j,N_1}\| \le C\sqrt{p_0}(1+\iota q)$. Proposition C.1 establishes the same estimates for the more general initialization used in the main theorem. In both cases $G_j^0 \asymp \sqrt{p_0}$, $|H_{j,0}| \le C\sqrt{p_0}$, $q \asymp \varepsilon^{-1}$, and $(1-h)^{N_1} \le q^{-1}$. Consequently

$$G_{j,N_1} = qG_j^0(1+O(\iota L)), \quad \frac{|H_{j,N_1}|}{qG_j^0} \le C(\varepsilon^2+\iota), \quad \frac{\|Q_\star w_{j,N_1}\|}{qG_j^0} \le C(\varepsilon+\iota).$$

Here $b_j = G_j + H_j$ and $s_j\theta_\star^\top w_j = G_j - H_j$. Expanding $r_j = \sqrt{(G_j-H_j)^2 + \|Q_\star w_j\|^2}$ thus yields

$$p_{j,N_1} = q^2(G_j^0)^2(1+e_j), \qquad |e_j| \le C(\iota L + \varepsilon^2) \le C\varepsilon.$$

The final bound uses $L \ge 1$, $\iota L \le 2\varepsilon$ and sufficiently small ε_0. Normalizing inside a class gives $|\log((p_{j,N_1}/\sum_{i\in S_{s_j}} p_{i,N_1})/\lambda_j)| \le C\varepsilon$.

For Phase II, Lemma B.18 gives

$$p_{j,n+1} = p_{j,n}(A_{s_j,n} + e_{j,n}), \qquad A_{s,n} = 1 + \tau(1 - P_{s,n}/\beta), \qquad |e_{j,n}| \le C\tau\zeta,$$

where $\zeta = \Theta_\star^2 + \rho_B + \tau$. Put $v_{j,n} = p_{j,n}/\sum_{i\in S_{s_j}} p_{i,n}$ and $\bar e_{s,n} = \sum_{i\in S_s} v_{i,n}e_{i,n}$. Then

$$\frac{v_{j,n+1}}{v_{j,n}} = \frac{A_{s_j,n} + e_{j,n}}{A_{s_j,n} + \bar e_{s_j,n}}.$$

The class-mass and small-step bounds ensure both numerator and denominator are at least $1/2$. Since $\log$ is 2-Lipschitz on $[1/2,\infty)$, $|\log(v_{j,n+1}/v_{j,n})| \le C\tau\zeta$. Summing at most $C_L L/\tau + 1$ steps and using $\zeta L \le 3\varepsilon$ proves the first bound of Theorem 2, uniformly over the entire phase. This accumulation is $O(\zeta L)$, not $O(L)$: the potentially large class-common growth cancels exactly. Appendix D gives $P_{s,N_\star} = \beta(1+O(\varepsilon))$, which proves the theorem's endpoint mass bound.

Because $\sum_{j\in S_s}\lambda_j = 1$, the selected template has $P_+ = P_- = \beta$ and is a balanced aligned exact representation. The initialization bounds imply $c/m \le \lambda_j \le C/m$; the endpoint mass bound proves the asserted individual mass bounds. For raw parameters, the balance estimate gives

$$\kappa_\star := \max_j \frac{|a_{j,N_\star}^2 - \|w_{j,N_\star}\|^2|}{p_{j,N_\star}} \le C\varepsilon,$$

and hence

$$|b_j - \sqrt{p_j}| + |r_j - \sqrt{p_j}| \le C\kappa_\star\sqrt{p_j}.$$

For completeness, write $x = b_j/r_j > 0$. Then $|x - x^{-1}| = |b_j^2 - r_j^2|/p_j \le \kappa_\star$, so $x = 1 + O(\kappa_\star)$ and $b_j/\sqrt{p_j} = \sqrt{x}$, $r_j/\sqrt{p_j} = x^{-1/2}$. Combining with the endpoint mass bound shows $|b_j - \sqrt{p_j^{\mathrm{sel}}}| + |r_j - \sqrt{p_j^{\mathrm{sel}}}| \le C\varepsilon\sqrt{p_j}$. Thus

$$|a_j - a_j^{\mathrm{sel}}|^2 + \|w_j - w_j^{\mathrm{sel}}\|^2 \le Cp_j\big(\varepsilon^2 + \|u_j - s_j\theta_\star\|^2\big).$$

All endpoint angles are acute and $\|u_j - s_j\theta_\star\|^2 = 2(1-\cos\theta_j) \le 2\sin^2\theta_j$. The existing estimate $W_s^2 \le C\varepsilon^2$ therefore gives

$$\sum_j p_j \|u_j - s_j\theta_\star\|^2 \le 2\sqrt{m}\sum_s P_s W_s^2 \le C\beta\sqrt{m}\varepsilon^2.$$

Since $\sum_j p_j = \sqrt{m}\Pi = O(\beta\sqrt{m})$, summing the preceding bounds proves

$$\|\Psi_{N_\star} - \Psi^{\mathrm{sel}}\|^2 \le C\beta\sqrt{m}\,\varepsilon^2,$$

which is the claimed relative-distance estimate. □

F SINGLE-EXAMPLE SGD ON A FIXED DATASET

This appendix proves Theorem 1. The phase analysis in Appendices B–E applies to an update whose gradient is uniformly close, relative to each neuron's size, to the population gradient. A single-example gradient does not satisfy such a bound. We therefore group consecutive iterates into short intervals and show that the total update over each interval has the required accuracy. These intervals are used only in the proof: the algorithm updates both layers after every sampled example.

F.1 ALGORITHM AND STATEMENT

Retain the Gaussian linear teacher and model normalization from the main text:

$$f_\Psi(x) = m^{-1/2}\sum_j a_j\sigma(w_j^\top x), \quad y(x) = \beta\theta_\star^\top x, \quad p_0 = c_0\beta\varepsilon^4/\sqrt{m}.$$

Set

$$L = \log(1/\varepsilon), \quad D = d + \log\frac{16m}{\delta\varepsilon}, \quad \omega = c_\omega\varepsilon/L.$$

Here ω is the normalized duration of an analysis interval. It is distinct from the half-step $h = \tau/2$ used in the phase analysis. For an integer K defined below, use the constant step

$$\tau = \omega/K, \qquad \eta = \frac{\sqrt{m}}{\beta}\tau.$$

Define analysis indices and actual SGD iteration indices by

$$J_1 = \left\lceil \frac{L}{\log(1+\omega/2)} \right\rceil, \qquad J_\star = J_1 + \lceil C_L L/\omega \rceil, \qquad T_1 = KJ_1, \quad T_\star = KJ_\star.$$

We first state the fixed-dataset SGD theorem, that is, Theorem 1 in the main body. Its proof is given at the end of this appendix, after the concentration and transfer estimates on which it relies.

Theorem F.1 (Single-example SGD on one fixed dataset)**.** *There are universal constants $c_\omega, C_K, C_M, C, \varepsilon_0 > 0$ such that the following holds. Let $0 < \varepsilon \le \varepsilon_0$ and $0 < \delta < 1/2$, and suppose that Assumption 1 holds. Draw once a dataset of iid Gaussian examples of size*

$$M = \left\lceil C_M(d + \log(16/\delta))\frac{L^2}{\varepsilon^2}\log^2(eL/\varepsilon) \right\rceil$$

and set

$$K = \left\lceil C_K\frac{D^2L^2}{\varepsilon^2} \right\rceil, \qquad \tau = \omega/K.$$

At every iteration draw a single index independently and uniformly from $\{1,\dots,M\}$, with replacement, and update both layers on that example's square loss. With probability at least $1-\delta$ over the dataset and sampled indices, all output signs remain unchanged and every neuron remains nonzero through $T_\star$. Moreover,

$$\max_{T_1 \le n \le T_\star}\max_j \angle(u_{j,n}, s_j\theta_\star) \le C\varepsilon L, \qquad \|f_{\Psi_{T_\star}} - y\|_{L^2(\gamma_d)} \le C\beta\varepsilon.$$

Define

$$G_j^0 = \tfrac{1}{2}\big(|a_j(0)| + s_j\theta_\star^\top w_j(0)\big), \qquad \lambda_j = \frac{(G_j^0)^2}{\sum_{i\in S_{s_j}}(G_i^0)^2}.$$

Then, for every j and every $T_1 \le n \le T_\star$,

$$\left|\log \frac{p_{j,n}/\sum_{i\in S_{s_j}} p_{i,n}}{\lambda_j}\right| \le C\varepsilon.$$

At $T_\star$, for $s \in \{-1,+1\}$,

$$|P_{s,T_\star} - \beta| \le C\beta\varepsilon, \qquad \left|\frac{p_{j,T_\star}}{\beta\sqrt{m}\,\lambda_j} - 1\right| \le C\varepsilon.$$

For the exact representation Ψ^{sel} defined by

$$a_j^{\mathrm{sel}} = s_j\sqrt{\beta\sqrt{m}\,\lambda_j}, \qquad w_j^{\mathrm{sel}} = a_j^{\mathrm{sel}}\theta_\star,$$

we also have

$$\frac{\|\Psi_{T_\star} - \Psi^{\mathrm{sel}}\|}{\|\Psi^{\mathrm{sel}}\|} \le C\varepsilon.$$

Finally,

$$\mathbb{E}\|Q_\star\nabla f_{\Psi_{T_\star}}(X)\|^2 \le C\beta^2\varepsilon^2, \qquad \sum_{i\ge 2}\lambda_i(G_{f_{\Psi_{T_\star}}}) \le C\beta^2\varepsilon^2.$$

The parameters satisfy

$$\tau \asymp \frac{\varepsilon^3}{D^2L^3}, \qquad T_\star = O(D^2L^4/\varepsilon^3).$$

Theorem F.1 contains Theorem 1 and the fixed-dataset conclusions used in Theorem 2 and Corollary 1. The integer K determines the learning rate and the deterministic reporting times; it is not an algorithmic batch size. In particular, the dataset size has no training-horizon factor and no explicit width dependence.

F.2 Normalized dynamics and local control

Write $s_0 = \sqrt{\beta/\sqrt{m}}$, $a_j = s_0A_j$, $w_j = s_0W_j$, and $z = ((A_j, W_j))_j$. Define

$$R_z(x) = m^{-1}\sum_i A_i\sigma(W_i^\top x) - \theta_\star^\top x,$$

$$g_j(z;x) = \big(R_z(x)\sigma(W_j^\top x),\ R_z(x)A_j\mathbf{1}_{W_j^\top x>0}x\big), \qquad F(z) = \mathbb{E}g(z;X).$$

Write $\mathcal{D} = (X_1,\dots,X_M)$ for the fixed dataset. An actual SGD step is exactly $z_{n+1} = z_n - \tau g(z_n; X_{I_n})$. Within a block, X_k in the sums below denotes the selected example $X_{I_{n_0+k}}$, not a fresh draw. Let $\mathsf{M}(z) = m^{-1}\sum_j |A_j|\|W_j\|$ and

$$\mathcal{R}_0 = \{\mathsf{M} \le 6,\ 1/2 \le |A_j|/\|W_j\| \le 2\ \forall j\}.$$

This is the manuscript's working region in the normalized variables. At the start of an interval, denote the state by $z^0 \in \mathcal{R}_0$ and put $t_j = |A_j^0| + \|W_j^0\|$. Define

$$e(z,z^0) = \max_j \frac{|A_j - A_j^0| + \|W_j - W_j^0\|}{t_j}.$$

On $e \le 1/32$, the output signs are fixed, all raw factors are nonzero, $\mathsf{M}(z) \le 12$, and the layer ratios are between $1/4$ and 4. Also t_j is comparable, by universal factors, to each initial layer norm.

Lemma F.1 (Neuron-relative population regularity)**.** *On this local region,*

$$|F_j^a(z)| + \|F_j^w(z)\| \le Ct_j, \qquad |F_j^a(z) - F_j^a(z^0)| + \|F_j^w(z) - F_j^w(z^0)\| \le Ct_je(z,z^0).$$

The constants do not depend on the smallest neuron norm, width, or dimension.

Proof. Gaussian Cauchy–Schwarz and $\|R_z\|_2 \le 1 + \mathsf{M}(z)$ give the first bound. For the second, use

$$\mathbb{E}[\sigma(W_i^\top X)\mathbf{1}_{u_j^\top X>0}X] = r_i\left(q(\psi_{ij})u_i + \frac{\sin\psi_{ij}}{2\pi}u_j\right), \quad r_i = \|W_i\|, \quad q(\psi) = \frac{\pi-\psi}{2\pi}.$$

Angular distance is Lipschitz in either direction, relative to Euclidean distance on the unit sphere, with constant $\pi/2$. Both q and $\sin$ are Lipschitz on $[0,\pi]$. Therefore the expression in parentheses is uniformly Lipschitz in (u_i, u_j). A relative parameter perturbation of size e changes each unit direction by at most Ce, each radius by at most Cer_i^0, and each output coefficient by at most $Ce|A_i^0|$. Summing changes in the displayed feature formula is bounded by $Ce\,m^{-1}\sum_i |A_i^0|r_i^0 \le 6Ce$. The prefactor A_j in F_j^w changes by $Ce|A_j^0|$. Its teacher contribution is $-A_j\theta_\star/2$, which does not depend on the gate direction. These facts give the hidden-gradient bound. For F_j^a, the analogous formula uses $r_jr_i(\sin\psi + (\pi-\psi)\cos\psi)/(2\pi)$ and teacher term $-\theta_\star^\top W_j/2$. The same calculation applies. □

Data reuse creates two errors: the empirical gradient differs from the population gradient, and the sampled gradient differs from the empirical gradient. A uniform covariance event controls the first for all parameter states. Conditional martingale bounds control the second along the adaptive trajectory.

F.3 Probability bounds on a fixed dataset

We first record the conditional Bernstein inequality used for the sampled indices. It is stated in scalar form; a standard net argument will give the corresponding vector estimate.

Lemma F.2 (Conditional Bernstein inequality)**.** *Suppose that the centered adapted increments Y_k satisfy*

$$\mathbb{E}[e^{sY_k} \mid \mathcal{F}_k] \le \exp\left(\frac{s^2v}{2(1-b|s|)}\right), \qquad |s| < 1/b.$$

Then, for every $q \ge 1$ and $u > 0$,

$$\Pr\left(\left|\sum_{k<q} Y_k\right| > C(\sqrt{qvu} + bu)\right) \le 2e^{-u}.$$

If $\|Y_k\|_{\psi_1|\mathcal{F}_k} \le C_0$, the hypothesis holds with v, b depending only on C_0. If instead $|Y_k| \le b_0$ and $\mathbb{E}[Y_k^2 \mid \mathcal{F}_k] \le v_0$, it holds with $v = Cv_0$ and $b = Cb_0$.

Proof. Multiplying the conditional moment generating functions and applying the Chernoff bound with $s \asymp \min\{\sqrt{u/(qv)}, 1/b\}$ proves the two tails. For a conditionally sub-exponential increment, expand the exponential and use $\mathbb{E}[|Y_k|^r \mid \mathcal{F}_k] \le r!C^r$. For a bounded increment, use $\mathbb{E}[|Y_k|^r \mid \mathcal{F}_k] \le b_0^{r-2}v_0$ for $r \ge 2$. □

We apply Lemma F.2 to an output coordinate and to a fixed $1/2$-net of the unit sphere for each hidden-gradient block. The net has at most 5^d elements and recovers vector norm up to a factor two. A union bound over neurons, intervals, and prefixes adds only $C[d + \log(16mKJ_\star/\delta)]$ to u. For the parameter choices above, after fixing the universal constants and reducing ε_0 if necessary,

$$d + \log(16mKJ_\star/\delta) \le CD, \qquad d + \log(16M/\delta) \le CD. \tag{8}$$

These inequalities follow by substituting the polynomial expressions for $K, M, J_\star$ and using $\log D \le D$ and $\log L \le L \le D$.

We next invoke the uniform covariance estimate that controls the difference between empirical and population gradients throughout the working region.

Lemma F.3 (Uniform gated covariance; (Tinaz et al., 2026), Lemma 10)**.** *There are universal constants $C, c > 0$ such that, for $q \ge 2$, $0 < \nu < 1/2$, and iid $X_i \sim N(0, I_q)$, if $M \ge Cq\nu^{-2}\log^2(1/\nu)$, then with probability at least $1 - 3e^{-cq}$,*

$$\sup_{u,v\in\mathbb{S}^{q-1}} \left\| \frac{1}{M}\sum_{i=1}^{M} X_iX_i^\top \mathbf{1}_{u^\top X_i\ge 0}\mathbf{1}_{v^\top X_i\ge 0} - \mathbb{E}[XX^\top \mathbf{1}_{u^\top X\ge 0}\mathbf{1}_{v^\top X\ge 0}] \right\|_{\mathrm{op}} \le \nu.$$

Proof. This is Lemma 10 of (Tinaz et al., 2026); its proof is given in their Appendix C.6. □

Choose $\nu = \varepsilon/(1024L)$ and apply the lemma after the following dimension augmentation. To obtain failure $\delta/4$ in the observed dimension d, append independent Gaussian coordinates to dimension at least $\max\{d, c^{-1}\log(12/\delta)\}$. Restrict the event to directions with zero appended coordinates and the upper-left covariance block. The restricted event depends only on the observed coordinates. The full augmented event implies it, so its marginal probability is at least $1 - \delta/4$.

Let $\Sigma_M = M^{-1}\sum_i X_i X_i^\top$ and define

$$K_M^{\geq}(u,v) = M^{-1}\sum_i X_i X_i^\top \mathbf{1}_{u^\top X_i \geq 0}\mathbf{1}_{v^\top X_i \geq 0}.$$

Let $K_M(u,v)$ use strict inequalities in both indicators, and let $H_M(u) = M^{-1}\sum_i X_i X_i^\top \mathbf{1}_{u^\top X_i > 0}$. For one fixed direction e, almost surely $\Sigma_M = K_M^{\geq}(e,e) + K_M^{\geq}(-e,-e)$, hence $\|\Sigma_M - I\| \leq 2\nu$. For arbitrary, data-dependent directions, the exact identities

$$H_M(u) = \Sigma_M - K_M^{\geq}(-u,-u),$$
$$K_M(u,v) = \Sigma_M - K_M^{\geq}(-u,-u) - K_M^{\geq}(-v,-v) + K_M^{\geq}(-u,-v)$$

handle strict gates, including zero preactivations. Thus H_M and K_M have errors at most 3ν and 5ν. The arbitrary-width gradient formula

$$F_{M,j}^w = A_j\left(m^{-1}\sum_i A_i K_M(u_j,u_i)W_i - H_M(u_j)\theta_\star\right)$$

and its scalar version

$$F_{M,j}^a = (W_j)^\top\left(m^{-1}\sum_i A_i K_M(u_j,u_i)W_i - H_M(u_j)\theta_\star\right)$$

give, for every state with $\mathsf{M} \leq 12$, after subtracting their population counterparts and summing $m^{-1}\sum_i |A_i|\|W_i\| = \mathsf{M}$,

$$|F_{M,j}^a - F_j^a| \leq 63\nu\|W_j\|, \qquad \|F_{M,j}^w - F_j^w\| \leq 63\nu|A_j|. \tag{9}$$

The event is uniform in the parameters. We also use the Gaussian event $\max_{i\leq M}\|X_i\|^2 \leq CD$, whose failure probability is at most $\delta/4$.

Fix an interval starting at $z^0 \in \mathcal{R}_0$ and let

$$\kappa = \min\big(\{q \in \{1,\dots,K\} : e(z_q, z^0) > 1/32\} \cup \{K+1\}\big).$$

In the martingale sums below, multiply the increment at time k by the predictable indicator $\mathbf{1}_{\{k<\kappa\}}$ and freeze the stopped trajectory after time κ. For an interval whose initial state is outside $\mathcal{R}_0$, set all corresponding increments to zero. We suppress these indicators in the notation. Conditional on the dataset and the past sampled indices, the mean of $g(z;X_I)$ is $F_M(z)$. Since $\|\Sigma_M\| \leq 2$ and $\mathsf{M} \leq 12$, the empirical residual norm is bounded:

$$\left(M^{-1}\sum_i R_z(X_i)^2\right)^{1/2} \leq \sqrt{2}(1+\mathsf{M}) \leq 13\sqrt{2}.$$

Cauchy–Schwarz and $\|\Sigma_M\| \leq 2$ also give $|F_{M,j}^a| + \|F_{M,j}^w\| \leq Ct_j$ on the local region. The normalized scalar gradient components used in the net argument are bounded by CD and have conditional variance at most CD. Indeed their square is at most $CDR_z(X_I)^2$, and the last display bounds its empirical expectation. Centering changes only universal constants. Lemma F.2, followed by the net and union bounds described above, gives simultaneously for all neurons, intervals, and prefixes $q \leq K$,

$$\left|\sum_{k<q}(g_j^a - F_{M,j}^a)/t_j\right| + \left\|\sum_{k<q}(g_j^w - F_{M,j}^w)/t_j\right\| \leq C(D\sqrt{K} + D^2), \qquad q \leq K. \tag{10}$$

Indeed, the variance contribution is $\sqrt{K\cdot D\cdot(CD)} = O(D\sqrt{K})$, and the linear Bernstein term is $O(D^2)$. A union bound gives (10) with failure probability at most $\delta/2$. The conditional mean is $F_M(z)$ because the indices are sampled independently with replacement.

F.4 Closing each interval and transferring the phase proof

Lemma F.4 (Interval estimates)**.** *On the events just constructed, every interval starting in $\mathcal{R}_0$ satisfies the following estimates. Denote its endpoint by z^+ and write*

$$\overline{g} = K^{-1} \sum_{k<K} g(z_k; X_k), \qquad z^+ = z^0 - \omega \overline{g}.$$

Then, with universal constants,

$$\max_{q \le K} e(z_q, z^0) \le C\omega, \tag{11}$$

$$|\overline{g}_j^a - F_j^a(z^0)| \le \|W_j^0\|\, \rho,$$

$$\|\overline{g}_j^w - F_j^w(z^0)\| \le |A_j^0|\, \rho, \tag{12}$$

$$|\Delta_j^+ - \Delta_j^0| \le C_b \omega^2 p_j^0. \tag{13}$$

Here $\Delta_j = A_j^2 - \|W_j\|^2$, $p_j^0 = |A_j^0|\|W_j^0\|$, and

$$\rho \le C\left(\omega + D/\sqrt{K} + D^2/K\right) + 72\nu \le \frac{\varepsilon}{2L}.$$

The constants c_ω and C_K can be chosen so that the last inequality holds and C_b is no larger than the balance-drift constant used in Appendix B.

Proof. Before and including a possible local exiting update, the accumulated conditional-mean displacement is at most $CK\tau t_j = C\omega t_j$. The stopped noise sum in (10), multiplied by τ, is also at most $C\omega t_j$ because $K \ge C_K D^2 L^2/\varepsilon^2$. For sufficiently small c_ω and ε_0, a first exiting iterate would therefore satisfy $e \le C\omega < 1/32$, a contradiction. This proves (11); the stopped process includes the first candidate exiting update.

For the effective gradient, divide the terminal martingale sum by K. Lemma F.1 and (11) bound the difference between the average population mean and $F(z^0)$ by $C\omega t_j$. The uniform empirical bias (9) contributes at most 72ν times the relevant starting layer norm, because (11) keeps the layer magnitudes within a factor $9/8$ of their initial values. Converting t_j to either starting layer norm costs only a universal factor. This proves (12) and the displayed bounds on ρ. Choose c_ω to make the $C\omega$ contribution at most $\varepsilon/(8L)$; choose C_K so the sampling contribution is at most $\varepsilon/(4L)$. The fixed choice $\nu = \varepsilon/(1024L)$ gives $72\nu < \varepsilon/(8L)$.

Because the gradients within an analysis interval are evaluated at different parameter states, their average need not satisfy the samplewise balance identity. We therefore telescope the exact balance identity over the individual SGD updates.

$$\Delta_j^+ - \Delta_j^0 = \tau^2 \sum_{k<K} \left[(g_j^a(z_k; X_k))^2 - \|g_j^w(z_k; X_k)\|^2\right].$$

On the dataset event $\max_{i \le M} \|X_i\|^2 \le CD$, local layer comparability gives the pointwise estimate $(g_j^a)^2 + \|g_j^w\|^2 \le CD^2 p_j^0$. It follows that the balance defect is at most $C\omega^2(D^2/K)p_j^0$ after summing over the interval. Since $D^2/K \le \varepsilon^2/(C_K L^2)$, increasing C_K places this bound within the balance-drift budget of Appendix B. This proves (13). □

Proposition F.1 (Transfer to the phase analysis)**.** *Let $\widetilde{\Psi}_l = \Psi_{Kl}$ and regard ω as the normalized step of this endpoint sequence. On the event of Lemma F.4, the endpoint sequence satisfies the deterministic update bounds stated after Lemma B.15, with normalized step ω and relative gradient error at most $\varepsilon/(2L)$. Consequently, the conclusions of Proposition B.1, Proposition C.1, Proposition B.2, and Proposition D.1 hold at the corresponding endpoint indices J_1 and $J_\star$. The deterministic share and parameter-distance estimates derived in Appendix E hold at the same endpoints.*

Proof. Undoing the normalization in (12) gives the per-neuron relative gradient bound used in the Phase-I growing, decaying, and transverse recursions. Those recursions use only the decomposition of the effective update into the population gradient and a bounded error; they do not require the effective update to be unbiased.

Equation (13) gives the one-step balance-drift bound with ω in place of τ. In the balanced case it is telescoped from zero, and under Assumption 1 it is telescoped from the initial defect as in Proposition C.1. The Phase-II radial expansion uses the population radial velocities, the relative error bound, layer comparability, and the normalization remainder. The transverse recursion uses the population Gaussian feature formula and the current balance estimate. Thus the logistic and angular recursions, and the first-exit argument built from them, apply without changing their hypotheses.

In raw coordinates the effective step is $K\eta = \sqrt{m}\,\omega/\beta$; the scaling $a_j = s_0 A_j$, $w_j = s_0 W_j$ converts (12) into (4.1) with ρ_B replaced by ρ. Likewise, multiplying (13) by s_0^2 gives the raw balance bound $C_b \omega^2 p_j$. Thus both required bounds hold with the same normalization. The deterministic allowance is ε/L, since $c_1 = 1$ in Appendix B. Lemma F.4 gives $\rho \le \varepsilon/(2L)$, leaving room for the fixed constants in the normalization estimates. Finally, Proposition D.1 and the selection argument use only the endpoint recursions just obtained. This proves the stated transfer. □

Proof of Theorem F.1. The uniform covariance event and the input-norm event fail with probability at most $\delta/2$ in total. Conditional on these dataset events, the union bound leading to (10) fails with probability at most $\delta/2$. We henceforth work on their intersection.

Apply Proposition F.1 by induction over the intervals. At an endpoint already known to lie in $\mathcal{R}_0$, Lemma F.4 controls the entire next interval and its endpoint. The deterministic first-exit argument then keeps the next endpoint in $\mathcal{R}_0$. Starting from the admissible initialization, the induction closes Phase I through J_1 and Phase II through $J_\star$.

Within an interval, (11) changes each layer magnitude by a relative $O(\omega)$ and each unit direction by an angle $O(\omega)$. It therefore preserves signs and nonvanishing, and changes each effective mass and normalized within-class share by a relative $O(\omega)$. Since $\omega = c_\omega \varepsilon/L$, the Phase-II conclusions extend from the endpoints to every actual SGD iterate. The fitting-time conclusions hold directly because $T_\star = K J_\star$ is an endpoint.

Choose the constants in the following order: first retain the universal constants from the deterministic phase analysis; next choose c_ω small enough for the local tube and step-size conditions; then choose C_M and C_K large enough for the covariance, martingale, and balance bounds; finally reduce ε_0 to satisfy the remaining smallness conditions and (8). These choices are independent of m, d, β, ε, and δ.

Since $J_\star = O(L/\omega) = O(L^2/\varepsilon)$ and $K = O(D^2 L^2/\varepsilon^2)$,

$$T_\star = K J_\star = O(D^2 L^4/\varepsilon^3), \qquad \tau = \omega/K \asymp \varepsilon^3/(D^2 L^3).$$

□

G Exact representations and local degeneracy

This section proves the zero-set statements from Section 4 and the local obstruction cited in Section 7.

Using $\sigma(t) = (t + |t|)/2$, every nondegenerate parameter vector satisfies

$$\begin{aligned} f_\Psi(x) &= \tfrac{1}{2}\langle V_\Psi, x\rangle + \tfrac{1}{2} H_\Psi(x), \\ V_\Psi &= \frac{1}{\sqrt{m}} \sum_j s_j p_j u_j, \\ H_\Psi(x) &= \frac{1}{\sqrt{m}} \sum_j s_j p_j |u_j^\top x|. \end{aligned} \tag{14}$$

The first term is odd and the second is even, so they are orthogonal in $L^2(\gamma_d)$.

Lemma G.1 (Independence of absolute-value features)**.** *Let $\nu_1, \dots, \nu_K \in \mathbb{S}^{d-1}$, $d \ge 2$, span pairwise distinct lines. If $\sum_{k=1}^K c_k |\nu_k^\top x| = 0$ almost everywhere, then $c_1 = \cdots = c_K = 0$.*

Proof. The identity holds everywhere by continuity. The distributional Hessian of $|\nu^\top x|$ is a matrix-valued measure supported on $\nu^\perp$:

$$\left\langle \partial^2 |\nu^\top \cdot|, \varphi \right\rangle = 2\nu\nu^\top \int_{\nu^\perp} \varphi \, d\mathcal{H}^{d-1}.$$

For each k, the hyperplane $\nu_k^\perp$ contains a point outside the union of all other such hyperplanes. Testing in a small neighborhood of that point isolates c_k, so $c_k = 0$. Repeating this for every k proves the claim. □

Proof of Proposition 1. By (14), $f_\Psi = y$ if and only if its odd part equals the teacher and its even part vanishes. Equality of the odd parts is exactly (Z1). Since $|u^\top x| = |(-u)^\top x|$, group the neurons by the line spanned by u_j. The coefficient of the resulting absolute-value feature on a line ℓ is

$$\frac{1}{\sqrt{m}} \sum_{j:\, u_j \in \ell} s_j p_j.$$

Lemma G.1 shows that the even part vanishes if and only if every such coefficient is zero, which is (Z2). □

Proof of Corollary 2. For $m = 2$, condition (Z2) forces the two neurons to lie on the same line with equal masses. They cannot have the same orientation, because then the left side of (Z1) vanishes. Hence $u_2 = -u_1$, and (Z1) gives $u_1 = \theta_\star$ and $p_1 = p_2 = \sqrt{2}\,\beta$.

For any integer $m \geq 4$, put $\nu_\pm = \cos\gamma\,\theta_\star \pm \sin\gamma\, e$ and $p = \beta\sqrt{m}/(2\cos\gamma)$, with $e \perp \theta_\star$ a unit vector. Take a positive/negative antipodal pair on each line $\nu_\pm$, each member with mass p. Their combined output is $p(\nu_+ + \nu_-)^\top x/\sqrt{m} = y(x)$. If $m > 4$, repeatedly split a neuron into two with the same sign and direction and positive masses summing to its former mass. This gives exactly m active neurons without changing the function. Set $a_j = s_j\sqrt{p_j}$ and $w_j = \sqrt{p_j}u_j$. Every neuron is balanced, the maximum angle is γ, and the two lines are distinct since $0 < \gamma < \pi/2$. □

Because an $L^2(\gamma_d)$ equality between these continuous functions is pointwise, every exact representation has zero residual on every sample. Its mini-batch gradient is therefore identically zero, proving the fixed-point claim in the main text.

Width three. Every occupied line must contain at least two active neurons by (Z2). Thus a nondegenerate width-three exact fit has only one hidden line, which (Z1) forces to be the teacher line. This does not force output-sign alignment: for any $q > 0$, take

$$(s_j, u_j, p_j) = (+1, \theta_\star, \beta\sqrt{3}), \quad (+1, -\theta_\star, q), \quad (-1, -\theta_\star, \beta\sqrt{3} + q).$$

The output is exactly y, but the second neuron is at the wrong sign-selected pole. This distinguishes teacher-line recovery from the stronger neuronwise conclusion of Theorem 1.

Thus width three does not yet admit multiple exact hidden lines, but the next proposition shows that cubic local degeneracy already occurs at width three.

Proposition G.1 (Cubic degeneracy)**.** *For every integer $m \geq 3$ and $d \geq 2$, there is a balanced family $\Psi^\flat(t)$ approaching a balanced aligned two-pole fit, with $P_+ = P_- = \beta$ and maximum angle t, such that*

$$\mathcal{L}(\Psi^\flat(t)) = \beta^2\left(\frac{t^3}{6\pi} + \frac{t^4}{16} + O(t^5)\right), \qquad \frac{\|\nabla\mathcal{L}(\Psi^\flat(t))\|^2}{2\mathcal{L}(\Psi^\flat(t))} \leq C\frac{\beta t}{\sqrt{m}}.$$

Consequently, no neighborhood of $\Psi^\flat(0)$ admits a constant $\mu > 0$ such that $\|\nabla\mathcal{L}(\Psi)\|^2 \geq 2\mu\mathcal{L}(\Psi)$ throughout that neighborhood. Thus a uniform local Polyak–Łojasiewicz (PL) inequality fails for the population loss.

Proof. We first construct the family at width three and then obtain every $m \geq 3$ by scaling and splitting neurons.

Width three. Fix a unit vector $e \perp \theta_\star$ and write $\nu(\alpha) = \cos\alpha\,\theta_\star + \sin\alpha\, e$. Take $(s_1, u_1, p_1) = (+1, \nu(t), \beta\sqrt{3}/2)$, $(s_2, u_2, p_2) = (+1, \nu(-t), \beta\sqrt{3}/2)$, and $(s_3, u_3, p_3) = (-1, -\theta_\star, \beta\sqrt{3})$. Balance every neuron by setting $a_j = s_j\sqrt{p_j}$ and $w_j = \sqrt{p_j}\,u_j$.

Then $P_+ = (p_1 + p_2)/\sqrt{3} = \beta$ and $P_- = p_3/\sqrt{3} = \beta$, while the maximum angle to the sign-selected teacher pole is t. At $t = 0$ this is a balanced aligned exact representation.

Write $z = \theta_\star^\top x$ and $r = e^\top x$. The residual is

$$R_t(x) = \frac{\beta}{2}\left\{\sigma(\nu(t)^\top x) + \sigma(\nu(-t)^\top x)\right\} - \beta\sigma(z).$$

Using the kernel K from Lemma B.2, $\mathcal{L} = \frac{\beta^2}{2}[\frac{3}{4} + \frac{1}{2}K(2t) - 2K(t)]$, and hence

$$\mathcal{L} = \beta^2\left(\frac{t^3}{6\pi} + \frac{t^4}{16} + O(t^5)\right).$$

It remains to bound the gradient. Using $\sigma(q) = (q+|q|)/2$ and $(|a+b|+|a-b|)/2 = \max\{|a|,|b|\}$, we can write

$$R_t = \frac{\beta}{2}\left[(\cos t - 1)z + \max\{\cos t\,|z|, \sin t\,|r|\} - |z|\right].$$

For $0 < t \le 1/10$, define $W = \{|z| \le \tan t\,|r|\}$. Polar coordinates in the $(\theta_\star, e)$-plane give $\mathbb{P}(W) \le Ct$ and $\mathbb{E}[\mathbf{1}_W (v^\top X)^2] \le Ct$ for every unit vector v.

For directions outside the $(\theta_\star, e)$-plane, decompose v into its planar and orthogonal components. The latter Gaussian component is independent of the wedge event, so the same bound holds with a constant independent of d.

On W^c, $\max\{\cos t\,|z|, \sin t\,|r|\} = \cos t\,|z|$, and therefore $\|R_t \mathbf{1}_{W^c}\|_2 \le C\beta t^2$. On W, for each of the three neuron directions, $|\sigma(u_j^\top X)| \le Ct|r|$, so $\|\sigma(u_j^\top \cdot)\mathbf{1}_W\|_2 \le Ct^{3/2}$. Since the loss expansion gives $\|R_t\|_2 \asymp \beta t^{3/2}$, for every neuron j and unit vector v,

$$\left|\mathbb{E}\left[R_t \mathbf{1}_{\{u_j^\top X > 0\}} v^\top X\right]\right| \le C\beta t^2, \qquad \left|\mathbb{E}\left[R_t \sigma(u_j^\top X)\right]\right| \le C\beta t^2.$$

Therefore $\|\nabla\mathcal{L}(\Psi_3^\flat(t))\|^2 \le \frac{C\beta^2 t^4}{3}\sum_{j=1}^3 (a_j^2 + \|w_j\|^2)$. Since $\sum_{j=1}^3 (a_j^2 + \|w_j\|^2) = 2\sum_{j=1}^3 p_j = 4\beta\sqrt{3}$, we obtain $\|\nabla\mathcal{L}(\Psi_3^\flat(t))\|^2 \le C\beta^3 t^4/\sqrt{3}$.

General width. Fix any integer $m \ge 3$ and choose positive integers r_1, r_2, r_3 with $r_1+r_2+r_3 = m$. Starting from the width-three construction above, put $c = (m/3)^{1/4}$ and replace base neuron j by r_j identical copies with raw parameters $a_{j,\ell}^{(m)} = c\,a_j^{(3)}/\sqrt{r_j}$ and $w_{j,\ell}^{(m)} = c\,w_j^{(3)}/\sqrt{r_j}$, for $1 \le \ell \le r_j$.

Each copy remains balanced. By positive homogeneity of ReLU,

$$\frac{1}{\sqrt{m}}\sum_{\ell=1}^{r_j} a_{j,\ell}^{(m)}\sigma\left((w_{j,\ell}^{(m)})^\top x\right) = \frac{1}{\sqrt{3}} a_j^{(3)}\sigma\left((w_j^{(3)})^\top x\right).$$

Thus the width-m network represents exactly the same function as the width-three network. In particular, its loss is unchanged, its class masses remain $P_+ = P_- = \beta$, and its maximum angle remains t.

The corresponding raw gradient blocks satisfy $\nabla_{a_{j,\ell}^{(m)}}\mathcal{L}_m = (3/m)^{1/4}\nabla_{a_j^{(3)}}\mathcal{L}_3/\sqrt{r_j}$ and $\nabla_{w_{j,\ell}^{(m)}}\mathcal{L}_m = (3/m)^{1/4}\nabla_{w_j^{(3)}}\mathcal{L}_3/\sqrt{r_j}$. Summing their squared norms over all copies gives $\|\nabla\mathcal{L}_m\|^2 = \sqrt{3/m}\,\|\nabla\mathcal{L}_3\|^2 \le C\beta^3 t^4/\sqrt{m}$.

Since $2\mathcal{L}_m = \beta^2 t^3(1 + O(t))/(3\pi)$, we obtain

$$\frac{\|\nabla\mathcal{L}_m\|^2}{2\mathcal{L}_m} \le C\frac{\beta t}{\sqrt{m}}.$$

Letting $t \downarrow 0$ rules out any positive state-independent PL constant on a full neighborhood of the aligned fit. □

H Experimental details

SGD first aligns then fits. We use $d = 2$, $m = 1000$, $\beta = 1$, batch size $B = 1$, and random seed 17. Both layers are updated simultaneously after each freshly drawn Gaussian example. There are $m/2$ neurons of each initial output-weight sign. Writing $\alpha = 10^{-2}$, we set

$$p_0 = \frac{\beta\alpha^2}{\sqrt{m}}, \qquad |a_j(0)| = \|w_j(0)\| = \sqrt{p_0}.$$

The initial directions $u_j(0)$ are sampled independently and uniformly from the unit circle, conditional on $s_j\langle u_j(0), \theta_\star\rangle \geq -\frac{1}{2}$, as required by Assumption 1. The two layers are initialized with balanced magnitudes. For the directional-trajectory visualization, we display a random 3% subset of the neurons satisfying $|\langle u_j(0), \theta_\star\rangle| \leq \frac{1}{2}$, so that the plotted neurons start away from both teacher poles and their subsequent alignment is visually apparent.

The initial directions $u_j(0)$ are sampled independently from a symmetric random initialization on the unit circle, consistent with the initialization regime of the theorem. The two layers are initialized with balanced magnitudes. In the theorem's parameterization, $\alpha = \sqrt{c_0}\,\varepsilon^2$.

The numerical parameters illustrate the mechanism; we do not claim that they satisfy the theorem's unspecified smallness constants. Both layers are updated simultaneously using Gaussian samples. The dimensionless step size is $\tau = \eta\beta/\sqrt{m} = 0.003$, equivalently $\eta = 0.003\sqrt{m}/\beta$. We run $12{,}000$ SGD steps and record the trajectory every 20 steps. No weight decay, clipping, normalization, or explicit rebalancing is used. The alignment time is the first recorded iterate at which every neuron is within 0.05 radians of its sign-selected teacher pole. The population loss falls below 10^{-8}.

The reported loss is the exact population loss under the Gaussian input distribution, rather than an empirical estimate on a test set. At each recorded iterate, we evaluate

$$\mathcal{L} = \frac{1}{2}\,\mathbb{E}\left[\left(\frac{1}{\sqrt{m}}\sum_{j=1}^{m} a_j\sigma(w_j^\top X) - \beta\,\theta_\star^\top X\right)^2\right]$$

in closed form. Expanding the square leaves only pairwise ReLU expectations, ReLU–teacher cross terms, and the teacher second moment. For nonzero w_i, w_j, writing $\phi_{ij} = \angle(w_i, w_j)$, we use

$$\mathbb{E}\big[\sigma(w_i^\top X)\sigma(w_j^\top X)\big] = \frac{\|w_i\|\|w_j\|}{2\pi}\left(\sin\phi_{ij} + (\pi - \phi_{ij})\cos\phi_{ij}\right),$$

together with

$$\mathbb{E}\big[\sigma(w_j^\top X)\theta_\star^\top X\big] = \frac{1}{2}\,w_j^\top\theta_\star, \qquad \mathbb{E}\big[(\theta_\star^\top X)^2\big] = 1.$$

Neuronwise-allocation experiment. For Figure 3, we use the same values of α, β, B, τ, and the same fitting threshold. We set $d = 20$, vary $m \in \{4, 16, 64\}$, and run 20 independent seeds at each width. Every run is trained for at most $12{,}000$ steps. For each neuron, define its initial growing mode by

$$G_j^0 = \frac{1}{2}\left(|a_j(0)| + s_j\,\theta_\star^\top w_j(0)\right), \qquad s_j = \operatorname{sign}(a_j(0)).$$

At the first monitored fitting time, we record

$$\lambda_j = \frac{(G_j^0)^2}{\sum_{i\in S_{s_j}}(G_i^0)^2}, \qquad v_{j,n_{\mathrm{fit}}} = \frac{p_{j,n_{\mathrm{fit}}}}{\sum_{i\in S_{s_j}} p_{i,n_{\mathrm{fit}}}}.$$

Here n_{fit} denotes the first monitored iterate at which the exact population loss satisfies $\mathcal{L}(\Psi_{n_{\mathrm{fit}}}) \leq 10^{-6}$; the loss is evaluated every 20 SGD steps. The plotted coordinates are $(m\lambda_j/2, mv_{j,n_{\mathrm{fit}}}/2)$, so a value of one corresponds to uniform allocation within a sign class. A point on the diagonal $y = x$ corresponds to $v_{j,n_{\mathrm{fit}}} = \lambda_j$, meaning that the neuron's observed share of its sign-class mass at fitting exactly matches the share predicted from its initialization. Thus, closeness to the diagonal measures the accuracy of the initialization-dependent allocation prediction.

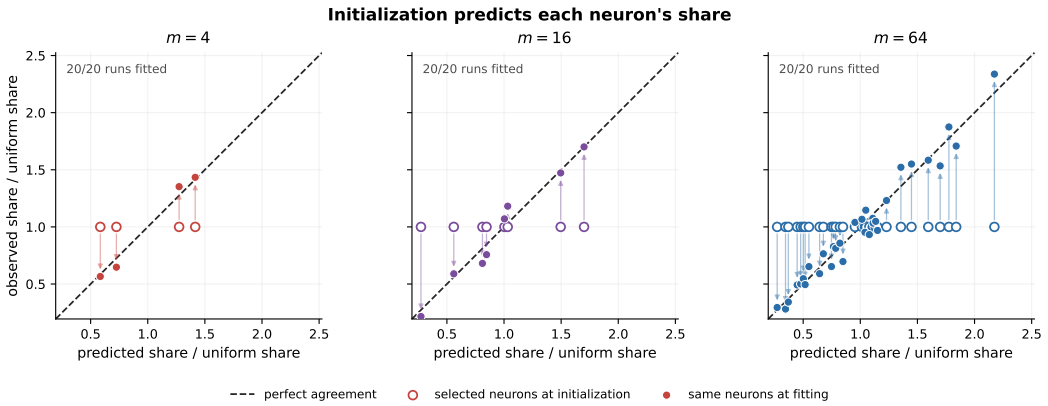


Figure 3: **Initialization predicts each neuron's share.** Share predicted before training against share observed during SGD, both measured relative to uniform allocation. Each panel shows the run with median final prediction error. We display half of its neurons, sampled evenly across the predicted-share range, with a minimum of four displayed neurons. Open circles show their initial shares, which all equal one after normalization by uniform allocation. Arrows connect each neuron's initial and fitting-time shares; they do not show intermediate iterates. Their movement from the horizontal line at one toward the dashed diagonal shows how SGD redistributes mass toward the neuron-specific allocation predicted before training.