CLASSIFYING SUBSPACES WITH TWO-LAYER RELU NETWORKS: ALIGN, DEPART, AND REALIGN

Anonymous authors
Paper under double-blind review

ABSTRACT

High-dimensional data and the representations neural networks learn from it typically concentrate near low-dimensional structures, of which a union of linear subspaces is the canonical model. We give the first analysis of feature learning in this setting, for two-layer ReLU networks trained by gradient flow from small initialization. Existing results characterize feature learning under data assumptions—Gaussian, orthogonal, or μ-separable—that a union of subspaces may not satisfy. We treat the two extremes of subspace dimension, classes supported on two lines and on two hyperplanes, and establish an align, depart, and realign pattern: neurons first select directions orthogonal to one of the two subspaces, may temporarily depart from them during fitting, and ultimately return to them as the loss converges to zero. These directions define features that vanish on one of the classes, revealing an implicit bias toward representations based on distance to subspaces. We characterize how class balance and subspace separation govern early alignment, prove geometric constraints on subsequent feature evolution, and establish convergence guarantees. For hyperplanes, we additionally identify a mechanism that can make early alignment arbitrarily slow. Numerical experiments illustrate the predicted dynamics. Together, our results explain how data geometry shapes both early feature selection and the final learned representation.

1 INTRODUCTION

A defining strength of neural networks is their ability to learn useful representations directly from data rather than operate on a fixed collection of hand-crafted features (Rumelhart et al., 1986; Bengio et al., 2013; Zeiler & Fergus, 2014; Mallat, 2016; Tenney et al., 2019). Understanding how such representations emerge from gradient-based training is therefore important for determining which features optimization selects, how they affect generalization and robustness, and how they are influenced by optimization and architectural choices. This process of adapting a network's internal representations to the data is referred to as *feature learning*, and stands in contrast with the *lazy/NTK regime* in which the features remain essentially fixed during training (Chizat et al., 2019).

A growing line of theoretical work has identified *neuron alignment* as one mechanism by which feature learning takes place. From a small initialization, the norms of first-layer neurons remain small while their directions evolve rapidly, so that many randomly initialized neurons concentrate around a much smaller collection of data-dependent directions. This phenomenon, described as feature quantization or *align-then-fit* dynamics, has been analyzed for shallow ReLU networks and, more recently, for broader classes of homogeneous networks (Maennel et al., 2018; Phuong & Lampert, 2021; Boursier et al., 2022; Min et al., 2024; Kumar & Haupt, 2024; 2025b; Kunin et al., 2025).

Existing analyses take two complementary approaches. One exploits strong assumptions on the data—orthogonal inputs, linear separability, weakly correlated classes, mean-separated Gaussian mixtures—to identify the directions selected by individual neurons and characterize their subsequent evolution (Boursier et al., 2022; Phuong & Lampert, 2021; Chistikov et al., 2023; Min et al., 2024). The other characterizes candidate features through extrema, critical points, or KKT points of auxiliary data-dependent objectives (Maennel et al., 2018; Boursier & Flammarion, 2025; Tsoy & Konstantinov, 2024; Kumar & Haupt, 2024; 2025b), which enlarges the scope of the theory but leaves the geometry and dynamical relevance of these points to be resolved for the distribution at hand. In neither case does identifying a direction reveal why it is a meaningful feature: that de-

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.

pends on which variations it responds to or suppresses, and what invariances it encodes. Nor need an early-selected feature be the final one—early and late representations may coincide, refine one another, or differ substantially (Glasgow, 2024; Tsoy & Konstantinov, 2024; Phuong & Lampert, 2021). Understanding feature learning therefore requires relating the structure of the data to both the *meaning* of the selected features and their *evolution* throughout training.

Motivated by this, we study binary classification in which the two class-conditional distributions are supported on distinct linear subspaces, a setting outside the regimes for which explicit dynamics are currently understood. Unions of subspaces are a canonical model of structured high-dimensional data, with a long history in subspace classification and clustering (Vidal et al., 2005; Vidal, 2011); high-dimensional data and learned representations often exhibit substantially lower-dimensional geometric structure (Cohen et al., 2020; Rangamani et al., 2023), and several modern representation-learning frameworks explicitly encourage discriminative subspace structure (Yu et al., 2020; Chan et al., 2022; Yu et al., 2023). For feature learning theory, the model has a further advantage: it separates the geometry of the class supports from how data are distributed within them, allowing the support geometry to remain discriminative even when simpler directional signals that dominated prior analyses (Min et al., 2024), such as differences in class means, are weak or absent.

The classification problem itself points to the geometric features one would hope to learn. When the class subspaces S_+ and S_- are known and each class assigns zero probability to their intersection, the nearest-subspace classifier $h_B(x) = \operatorname{sign}\left(d(x, S_-) - d(x, S_+)\right)$ where $d(x, S) = \inf_{s \in S} \|x - s\|_2$, achieves zero population error and is Bayes optimal. We focus on two complementary endpoints where it has an especially simple representation: classes supported on two lines, the smallest nontrivial linear subspaces, and on two hyperplanes, the largest proper ones. A hyperplane has a unique normal $n_\pm$ up to sign, and a line has a unique normal within the two-dimensional span of the data, so on that span $d(x, S_\pm) = |\langle n_\pm, x\rangle| = (\langle n_\pm, x\rangle)_+ + (\langle -n_\pm, x\rangle)_+$ and h_B can be represented exactly by a bias-free two-layer ReLU network of four neurons (input weights) supported entirely on the class normals,

$$f_{\mathrm{NS}}(x) = (\langle n_-, x\rangle)_+ + (\langle -n_-, x\rangle)_+ - (\langle n_+, x\rangle)_+ - (\langle -n_+, x\rangle)_+.$$

Whether gradient flow from random small initialization discovers these normals on its own, and how the representation evolves as training proceeds, is our question. The two settings differ sharply in their intersection structure: distinct lines meet only at the origin, whereas generic hyperplanes share a high-dimensional subspace. We show that this difference produces qualitatively different alignment dynamics but parallel conclusions about feature selection and late-time recovery. The high-level takeaway is:

> *Gradient flow implicitly constructs a **subspace-distance classifier** in an **align, depart, realign** pattern: when the class distributions are sufficiently balanced, neurons first align with **normals** to one of the two subspaces, may depart from them during fitting but only within an invariant cone assigned at alignment, and eventually realign with the same normals.*

We separate training into an early *alignment phase*, in which neuron directions are selected while the network remains close to initialization, and a *fitting phase*, in which the network drives the loss to zero and the learned representation continues to evolve. Our main contributions are:

1. We explicitly characterize the early alignment dynamics and alignment time in both subspace models. Under a data balance condition, we show that gradient flow aligns neurons with normals to one of the two subspaces. In the two-hyperplane model, we additionally identify a normally hyperbolic saddle manifold that can make alignment time arbitrarily slow for some neurons.
2. Assuming that all neurons successfully align, we show that a neuron may leave its normal, but remains within an invariant cone. In both models, training drives the loss to zero while the neuron directions return to the same normals selected during the early alignment phase.
3. We show empirically that gradient descent exhibits the predicted training dynamics in controlled instances of both models and test their persistence beyond the exact theoretical assumptions.

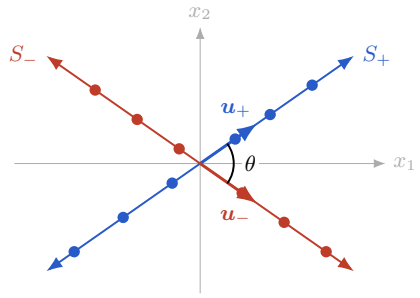
Notation. We write $[m] := \{1, \dots, m\}$ and denote the Euclidean unit sphere and closed unit ball by $\mathbb{S}^{d-1}$ and $\mathbb{B}$, respectively. For a linear subspace $V \subseteq \mathbb{R}^d$, let Π_V denote the orthogonal projection onto V. For a unit vector $\boldsymbol{z} \in \mathbb{S}^{d-1}$, we write $\Pi_{\boldsymbol{z}} := I - \boldsymbol{z}\boldsymbol{z}^\top$ for the orthogonal projection onto $\boldsymbol{z}^\perp = T_{\boldsymbol{z}}\mathbb{S}^{d-1}$. For any nonzero vector $\boldsymbol{z}$, let $\widehat{\boldsymbol{z}} := \boldsymbol{z}/\|\boldsymbol{z}\|$.

2 SETUP AND DATA GEOMETRY

This section introduces the two data models, the network and training dynamics we analyze.

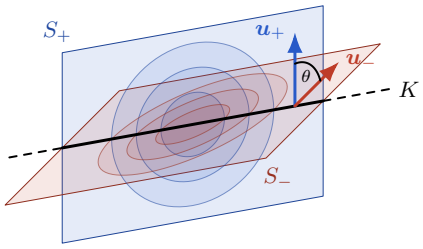
Data models and shared structure. Let $\mathbb{P}$ be the joint distribution of $(X, Y) \in \mathbb{R}^d \times \{-1, +1\}$. For each class label $c \in \{-, +\}$, let μ_c be the class-conditional distribution supported in a linear subspace $S_c \subset \mathbb{R}^d$. We assume $\mu_c(\{0\}) = 0$ for both classes. The two subspaces $S_\pm$ have the same dimension and we write their intersection as $K := S_+ \cap S_-$. We study two complementary realizations of this geometry:

- In the **two-line model**, $K = \{0\}$ and each μ_c is a finite atomic measure supported on the line S_c such that every half-line is populated.
- In the **two-hyperplane model** for $d \geq 3$, $S_\pm$ are $(d-1)$-dimensional with $\dim K = d - 2$, and each μ_c is spherically symmetric within S_c: under μ_c, $X = \rho_c \hat{X}_c$, where $\rho_+, \rho_-, \hat{X}_+, \hat{X}_-$ are all independent random variables with $\rho_c > 0$, $\mathbb{E}[\rho_c^{-1}] < \infty$, and $\hat{X}_c$ uniform on the unit sphere of S_c.



μ_+, μ_-: atomic

(a) Two-line model.



μ_+, μ_-: spherically symmetric

(b) Two-hyperplane model.

Figure 1: **Data geometry.**

Note that in both cases, $S_c \cap K^\perp$ is one-dimensional. Let $\boldsymbol{u}_c$ be a unit vector spanning this line, so that $S_c = K \oplus \operatorname{span} \boldsymbol{u}_c$. Choosing the signs of $\boldsymbol{u}_+$ and $\boldsymbol{u}_-$ appropriately, we write $\langle \boldsymbol{u}_+, \boldsymbol{u}_- \rangle = \cos\theta$, where $\theta \in (0, \pi/2]$ is the unique nonzero principal angle between the two supports. Thus, factoring out the shared intersection K yields a two-dimensional quotient plane $U := \operatorname{span}\{\boldsymbol{u}_+, \boldsymbol{u}_-\}$; see Figure 1. We use $Z_c := \langle X, \boldsymbol{u}_c \rangle$ to denote the coordinate of X along $\boldsymbol{u}_c$, so that a class-c example decomposes as $X = X_K + Z_c \boldsymbol{u}_c$ with $X_K \in K$. Finally, let $\boldsymbol{n}_c \in U$ be the unit vector orthogonal to $\boldsymbol{u}_c$ within U, with signs chosen so that $\langle \boldsymbol{n}_-, \boldsymbol{u}_+ \rangle = \sin\theta = -\langle \boldsymbol{n}_+, \boldsymbol{u}_- \rangle$. We will also use the four planar cones

$$\mathcal{C}(\tau \boldsymbol{n}_-) := \operatorname{cone}\{\tau \boldsymbol{n}_-, \tau \boldsymbol{u}_+\},\ \mathcal{C}(\tau \boldsymbol{n}_+) := \operatorname{cone}\{\tau \boldsymbol{n}_+, -\tau \boldsymbol{u}_-\}, \qquad \tau \in \{\pm 1\}, \tag{1}$$

where $\operatorname{cone}\{a, b\} := \{\alpha a + \beta b : \alpha, \beta \geq 0\}$. Each $\mathcal{C}(\nu)$ is therefore the region in U between the normal direction ν and the acute orientation of the other subspace direction.

The two models have different data assumptions for a structural reason. On two lines, all samples lie on four rays, and a neuron's activation pattern is described by finitely many ray masses making a direct finite-sample analysis possible. On two hyperplanes, this would induce a high-dimensional, sample-dependent activation arrangement with no comparable low-dimensional description; at the population level, spherical symmetry makes the dynamics explicit.

Network and training dynamics. We train a bias-free two-layer ReLU network of width m on the population exponential loss,

$$f(\boldsymbol{x}; W, \boldsymbol{v}) := \sum_{j=1}^{m} v_j \sigma(\langle \boldsymbol{w}_j, \boldsymbol{x} \rangle), \qquad \mathcal{L}(W, \boldsymbol{v}) := \mathbb{E}_{(X,Y)\sim\mathbb{P}}\left[e^{-Y f(X; W, \boldsymbol{v})}\right],$$

where $\sigma(a) := \max(0, a)$, $W = [\boldsymbol{w}_1, \ldots, \boldsymbol{w}_m] \in \mathbb{R}^{d \times m}$ and $\boldsymbol{v} \in \mathbb{R}^m$. Since $|f(\boldsymbol{x})| \leq \left(\sum_j |v_j| \, \|\boldsymbol{w}_j\|\right) \|\boldsymbol{x}\|$, the loss has envelope $e^{C\|\boldsymbol{x}\|}$ on bounded parameter sets, so $\mathcal{L}$ can be infinite. To avoid this, we assume throughout:

Assumption 1 (Radial integrability). *For each $c \in \{+, -\}$, let ρ_c denote the class-conditional radius, $\rho_c \overset{d}{=} \|X\| \mid Y = c$. We assume $\mathbb{E}\left[\rho_c^2 e^{q\rho_c}\right] < \infty$ for all $q > 0$.*

In the two-line model the support is finite and Assumption 1 is automatic; in the two-hyperplane model it holds for bounded radii and for Gaussian radial tails.

Because ReLU is nonsmooth at 0, we formulate continuous-time training as a Clarke subgradient flow on the parameter pair $\xi := (W, \boldsymbol{v})$:

$$\dot{\xi}(t) \in -\partial \mathcal{L}(\xi(t)) \qquad \text{for almost every } t \geq 0, \tag{2}$$

with ∂ the Clarke subdifferential and trajectories absolutely continuous. Prior analyses commonly fix $\sigma'(0) = 0$ to ensure global existence of the resulting gradient-flow ODE (Boursier et al., 2022; Min et al., 2024), the same convention used by automatic differentiation, and for discrete gradient descent in exact arithmetic the choice is largely immaterial: different values of $\sigma'(0)$ affect back-propagation only on a measure-zero set and, under generic initialization and step size, yield the same SGD trajectory almost surely (Bertoin et al., 2021). In continuous time, however, trajectories can reach activation boundaries, and in our setting the directions selected by gradient flow lie on these boundaries, where the convention $\sigma'(0) = 0$ yields a flow that cannot be continued. We therefore study all admissible Clarke continuations, and Section 6 shows that gradient descent with small step size exhibits the predicted training dynamics.

Small balanced initialization. Our analysis focuses on the small-initialization regime. Fix unit vectors $\bar{\boldsymbol{w}}_1, \ldots, \bar{\boldsymbol{w}}_m \in \mathbb{S}^{d-1}$ and signs $s_1, \ldots, s_m \in \{-1, +1\}$. For $\varepsilon > 0$, we initialize

$$\boldsymbol{w}_j(0) = \varepsilon\, \bar{\boldsymbol{w}}_j, \qquad v_j(0) = s_j \|\boldsymbol{w}_j(0)\|, \qquad j \in [m]. \tag{3}$$

Thus ε sets the initialization scale, each neuron is initially balanced, $v_j(0)^2 = \|\boldsymbol{w}_j(0)\|^2$, and s_j marks it as a positive- or negative-output neuron. The small scale is what separates the early alignment phase from the fitting phase. Balancedness is imposed to exploit the homogeneity of the two-layer parametrization and reduce each neuron to a norm, a direction, and an output sign; the balancedness of the weights and output signs are preserved along every solution of the flow, as shown in Appendix A.3:

$$v_j(t)^2 = \|\boldsymbol{w}_j(t)\|^2, \qquad \operatorname{sign}(v_j(t)) = \operatorname{sign}(v_j(0)) = s_j, \qquad \forall j \in [m], \qquad \forall t \geq 0. \tag{4}$$

Our analysis and results hold for any fixed choice of $\bar{\boldsymbol{w}}_j$ and s_j. Randomness enters only through their choice. In particular, the exceptional initial directions excluded from some of our theorems have probability zero under any absolutely continuous initialization distribution.

Active moment balance. Our alignment results require the two classes to be comparable in a sense that depends on the support geometry. Both phases of training are governed by the same family of quantities, which we introduce together. For a class c and a side $\tau \in \{-1, +1\}$, define the *one-sided active moment* and its residual-weighted counterpart, the *residual moment*,

$$M_c^\tau := \mathbb{E}\left[|Z_c|\, \mathbf{1}_{\{Y=c,\ \tau Z_c > 0\}}\right], \qquad \mathcal{R}_c^\tau(t) := \mathbb{E}\left[|Z_c|\, e^{-Yf(X;t)}\, \mathbf{1}_{\{Y=c,\ \tau Z_c > 0\}}\right]. \tag{5}$$

Both are a first moment along $\boldsymbol{u}_c$ of the class-c data on the side $\{\tau Z_c > 0\}$ of its support, which is a ray in the two-line model and a half-hyperplane in the two-hyperplane model; they differ only in that $\mathcal{R}_c^\tau$ weights each sample by how poorly it is currently fit, and they coincide when the output vanishes. Thus M_c^τ is a property of the data alone and governs the early phase, where small initialization keeps $e^{-Yf} \approx 1$, while $\mathcal{R}_c^\tau(t)$ governs the late phase once the loss is no longer negligible (Section 5).

Assumption 2 (Active moment balance)**.** *On each side $\tau \in \{-1, +1\}$, the two class-wise active moments are nonzero, finite, and comparable after accounting for the support angle:*

$$M_+^\tau \cos\theta < M_-^\tau, \qquad M_-^\tau \cos\theta < M_+^\tau. \tag{6}$$

For $\theta \neq \pi/2$, each ratio M_-^τ / M_+^τ lies in the *balance window* $(\cos\theta, 1/\cos\theta)$. This balances data moments against support geometry so that the normals are attracting in the small-initialization dynamics; Section 4 explains the condition, while Appendix B.5 shows that violating it can produce interior attractors. As $\theta \to 0$ the window shrinks to $\{1\}$, whereas at $\theta = \pi/2$ it becomes $(0, \infty)$.

3 Main Results

In this section, we state one main theorem for each data model, each describing the training trajectory through two stages: an early alignment phase in which neurons with positive output weights align with the normals to the negative class and neurons with negative output weights align with the

Property	Two lines	Two hyperplanes
Support intersection	$K = \{0\}$	$\dim K = d - 2$
Class distributions	Finite atomic measures	Spherically symmetric population measures
Early-alignment time	Uniform off a measure-zero set	Nonuniform; arbitrarily slow near a measure-zero set
Late-stage loss decay	$\mathcal{O}(t^{-1})$	$\Theta(t^{-1/2})$

Table 1: Comparison of the two data models under the assumptions of Section 2. Loss rates reflect the choice of finite vs population measures as well as support geometry.

normals to the positive class, and a fitting phase in which their norms grow and the loss decreases, the neurons may temporarily leave these normals, and the late dynamics eventually restore normal alignment. In both models, the result excludes exceptional initial directions and requires sufficiently small initialization. We begin with the two-line model, where all inputs lie in U and the orthogonal components $\Pi_{U^\perp}\boldsymbol{w}_j$ remain frozen, reducing the directional analysis to the two-dimensional data plane. In this setting, the alignment time and initialization threshold hold uniformly over all non-exceptional directions. We then explain how the shared intersection of the two hyperplanes creates a saddle that breaks this uniformity.

Theorem 3.1. *Let Assumption 2 hold, and set* $\Delta := \min_{\tau\in\{\pm1\}} \min\{M_+^\tau - \cos\theta\, M_-^\tau,\ M_-^\tau - \cos\theta\, M_+^\tau\} > 0$. *Assume* $\Pi_U \bar{\boldsymbol{w}}_j \neq 0$, *and* $\frac{\Pi_U \bar{\boldsymbol{w}}_j}{\|\Pi_U \bar{\boldsymbol{w}}_j\|} \notin \{\pm \boldsymbol{n}_{s_j}\}$ *for all* j. *Then there exist an alignment time* $T_{\text{align}} = \mathcal{O}\big((\pi-\theta)/\Delta\big)$ *and threshold* $\varepsilon_0 > 0$ *that depend on the data, such that, for every* $\varepsilon < \varepsilon_0$, *every Clarke gradient-flow trajectory satisfies:*

1. ***(Early alignment.)*** *There is a time* $t_0 \le T_{\text{align}}$ *at which every projected neuron reaches an opposite-class normal,*
$$\frac{\Pi_U \boldsymbol{w}_j(t_0)}{\|\Pi_U \boldsymbol{w}_j(t_0)\|} = \nu_j \in \{\pm \boldsymbol{n}_{-s_j}\}.$$
The orthogonal component $\Pi_{U^\perp}\boldsymbol{w}_j$ *is frozen at its initial value for all* $t \ge 0$.
2. ***(Departure, realignment, and convergence of the loss.)*** *For all* $t \ge t_0$, *the projected direction* $\Pi_U \boldsymbol{w}_j(t)$ *remains in the invariant cone* $\mathcal{C}(\nu_j)$. *If, moreover, each of the four normals is selected by at least one neuron, then* $\mathcal{L}(t) = \mathcal{O}(t^{-1})$, $\|\Pi_U \boldsymbol{w}_j(t)\| \longrightarrow \infty$, *and* $\widehat{\boldsymbol{w}}_j(t) \longrightarrow \nu_j$.

Remark 1. *Frozen orthogonal components become negligible as projected norms diverge; at* $\theta = \pi/2$, *projected directions remain normal. For* $M_c^\tau = M$, *Appendix B gives* $T_{\text{align}} = O(\log(1/\theta)/(M\theta))$ *and a conservative threshold* $\varepsilon_0 = \frac{\theta}{\sqrt{m}} \exp[-\Theta(\log(1/\theta)/\theta)]$ *as* $\theta \downarrow 0$. *This sufficient bound is far below the initialization used in our experiments.*

In the two-hyperplane model, the shared intersection carries no discriminative signal. Its directions form a stationary saddle manifold for the leading-order dynamics, with a measure-zero stable set. We therefore define an exceptional set $\mathcal{E}_s$ consisting of this stable set together with the same-class normals $\pm\boldsymbol{n}_s$. Although random initialization avoids $\mathcal{E}_s$ almost surely, alignment time grows logarithmically as initialization approaches it, while the small-output approximation is valid only for $\Theta(\log(1/\varepsilon))$ time. Hence no uniform initialization threshold applies over all generic directions. We instead state the theorem on compact sets $\mathcal{Z}_s \subset \mathbb{S}^{d-1} \setminus \mathcal{E}_s$, allowing T_{align} and ε_0 to depend on $\mathcal{Z}_s$; see Section 4 and Appendix C.

Theorem 3.2. *Let Assumption 2 hold, and write* $M_c := M_c^+ = M_c^-$. *For* $s \in \{1, -1\}$, *fix compact sets* $\mathcal{Z}_s \subset \mathbb{S}^{d-1} \setminus \mathcal{E}_s$ *with* $\bar{\boldsymbol{w}}_j \in \mathcal{Z}_{s_j}$ *for every* j. *Then there exist* $T_{\text{align}} < \infty$ *and* $\varepsilon_0 > 0$, *depending on the data and on* $\mathcal{Z}_\pm$, *such that for every* $\varepsilon < \varepsilon_0$, *every Clarke gradient-flow trajectory satisfies:*

1. ***(Early alignment.)*** *There exists* $t_0 \le T_{\text{align}}$ *such that for all* j, $\widehat{\boldsymbol{w}}_j(t_0) = \nu_j \in \{\pm \boldsymbol{n}_{-s_j}\}$.
2. ***(Departure, realignment, and convergence of the loss.)*** *For all* $t \ge t_0$, *the neurons remain in the invariant cone* $\mathcal{C}(\nu_j) \subset U$. *If, moreover, each of the four normals is selected by at least one neuron, then* $\mathcal{L}(t) = \Theta(t^{-1/2})$, $\|\boldsymbol{w}_j(t)\| \to \infty$, *and* $\widehat{\boldsymbol{w}}_j(t) \to \nu_j$.

What the fitting phase needs. Part 2 of Theorems 3.1 and 3.2 holds from any perfectly aligned configuration, without Assumption 2 or small initialization. Its remaining hypothesis, that each of the four normals is selected by at least one neuron, holds with high probability. If the initial directions are drawn uniformly on the sphere (e.g., normalized Gaussians) with independent fair output signs, and the initialization scale is sufficiently small for the given width, all four groups are populated with probability at least $1 - 5(3/4)^m$ (Appendix C.4).

Normal alignment as subspace-distance classification. Despite their different geometries, the two theorems show the same qualitative trajectory: early alignment with subspace normals is not merely a transient small-initialization effect. Although it need not persist throughout training, the normal directions identified in the early phase are recovered asymptotically. This alignment shows an implicit bias toward *annihilating the opposite class support*. Moreover, Theorem C.8 in the appendix makes the limiting network for the two-hyperplane model explicit showing that it can be viewed as a weighted subspace-distance classifier,

$$\frac{f(\boldsymbol{x};t)}{\sqrt{t}} \longrightarrow \alpha_+ \operatorname{dist}(\boldsymbol{x}, S_-) - \alpha_- \operatorname{dist}(\boldsymbol{x}, S_+), \tag{7}$$

with $\alpha_\pm$ determined by the learned group masses. Thus gradient flow selects features that suppress the opposite class support rather than align with a class-center direction.

Connections with prior work. The closest analyses of comparable non-linearly-separable geometries are the XOR studies of Tsoy & Konstantinov (2024) and Glasgow (2024). Their settings are orthogonal ($\theta = \pi/2$) instances of our geometries, with different within-class distributions. At $\theta = \pi/2$, the normal to one class support lies within the other, so annihilating the opposite class and aligning with one's own class select the same direction; only $\theta < \pi/2$ separates the two readings and exposes the effects of angle and imbalance. Moreover, Tsoy & Konstantinov (2024) use discrete well-separated clusters and a smooth homogeneous approximation to ReLU to avoid non-smoothness challenges, and their late XOR result assumes convergence of an auxiliary four-neuron network; Glasgow (2024) studies four equally weighted Boolean clusters and requires polylogarithmic width in the ambient dimension. In contrast, our framework allows general line-supported distributions and continuous hyperplane populations, highlights the roles of angle, within-subspace distribution, and class balance, and requires no width lower bound for alignment or cone invariance. An extended discussion of related work is in Appendix D.

4 Phase I: Early Alignment

This section shows that both data models share a single early-time mechanism. We reduce the coupled network dynamics to a set-valued flow on neuron directions, approximate it by a decoupled idealized field, classify the equilibria of that field, and transfer the conclusions to the exact flow by controlling the approximation error over a window long enough to contain the alignment.

Our early-time analysis is closely related to that of Maennel et al. (2018) and (Min et al., 2024). We extend it to the two subspace models, including population distributions, while retaining the full Clarke dynamics at the activation boundaries that contain the target directions. Differentiating the normalized direction $\widehat{\boldsymbol{w}}_j$ away from activation boundaries yields

$$\dot{\widehat{\boldsymbol{w}}}_j = s_j \Pi_{\widehat{\boldsymbol{w}}_j} \mathbb{E}\Big[Y e^{-Y f(X;W,\boldsymbol{v})} X \, \mathbf{1}_{\{\langle \widehat{\boldsymbol{w}}_j, X\rangle > 0\}} \Big]. \tag{8}$$

Apart from the exponential weight $e^{-Y f(X;W,\boldsymbol{v})}$, the right-hand side depends on neuron j only through its direction and output sign. Small initialization keeps the network prediction close to zero throughout the alignment phase, so $e^{-Y f(X;W,\boldsymbol{v})} \approx 1$. Away from ReLU boundaries, the neurons therefore evolve independently according to

$$\dot{\widehat{\boldsymbol{w}}}_j \approx \Pi_{\widehat{\boldsymbol{w}}_j} G^0_{s_j}(\widehat{\boldsymbol{w}}_j), \qquad G^0_s(\boldsymbol{z}) := s\, \mathbb{E}\big[Y X \, \mathbf{1}_{\{\langle \boldsymbol{z}, X\rangle > 0\}} \big]. \tag{9}$$

For an empirical distribution, (9) is simply the normalized signed sum of the currently activated samples, and hence recovers the early directional field used in prior finite-data analyses. The population formulation replaces this finite sum by the corresponding active first moment. Thus the familiar small-initialization approximation has the same interpretation in both settings: a positive-output neuron is attracted by the positively labeled data in its active halfspace and repelled by the negatively labeled data, while a negative-output neuron follows the opposite field.

Clarke completion at activation boundaries. The field in (9) can be discontinuous where a non-negligible part of the data lies on $\langle \boldsymbol{z}, X\rangle = 0$. These configurations cannot be discarded in our problem: the subspace normals that emerge as attractors are themselves activation-boundary directions. We therefore complete the smooth early field in the same way as the exact Clarke dynamics: by taking limits from regular directions and convexifying. Let Ω denote the set of directions at which the population field is smooth. The idealized directional dynamics are

$$\dot{\boldsymbol{z}} \in \Pi_{\boldsymbol{z}} \mathcal{G}_s^0(\boldsymbol{z}), \qquad \mathcal{G}_s^0(\boldsymbol{z}) := \overline{\mathrm{co}} \left\{ \lim_{k\to\infty} G_s^0(\boldsymbol{z}_k) : \boldsymbol{z}_k \to \boldsymbol{z},\ \boldsymbol{z}_k \in \Omega \right\}. \tag{10}$$

The approximation results in Appendix A.4 show that, throughout the alignment horizon, the exact directional dynamics remain uniformly close to (10): for every admissible Clarke trajectory,

$$\dot{\widehat{\boldsymbol{w}}}_j(t) \in \Pi_{\widehat{\boldsymbol{w}}_j(t)} \mathcal{G}_{s_j}^0(\widehat{\boldsymbol{w}}_j(t)) + \delta_\varepsilon(t)\mathbb{B}, \qquad \sup_{t \le T_{\mathrm{align}}} \delta_\varepsilon(t) \to 0 \quad \text{as } \varepsilon \to 0. \tag{11}$$

As in prior small-initialization analyses, the approximation is valid on a time window growing logarithmically as initialization decreases. For fixed compact generic sets of initial directions, the alignment horizon is independent of the initialization scale, and sufficiently small initialization ensures that this horizon lies within the approximation window. The next step is to compute $\mathcal{G}_s^0$, classify its attractors, then transfer the resulting alignment behavior to the exact Clarke dynamics.

Why the normals attract. The idealized field is determined by the classwise active moments

$$A_c(\boldsymbol{z}) := \mathbb{E}\big[X\, \mathbf{1}_{\{Y=c\}} \mathbf{1}_{\{\langle \boldsymbol{z}, X\rangle > 0\}}\big], \qquad G_s^0(\boldsymbol{z}) = s\big(A_+(\boldsymbol{z}) - A_-(\boldsymbol{z})\big), \qquad \forall \boldsymbol{z} \in \mathbb{S}^{d-1}.$$

At a subspace normal, the opposite class lies on the activation boundary, and its Clarke completion can cancel the attracting class's tangential contribution. For a positive-output neuron at $\tau \boldsymbol{n}_-$, this requires $M_-^\tau > \cos\theta\, M_+^\tau$; the other inequality in Assumption 2 applies at $-\tau \boldsymbol{n}_+$ for a negative-output neuron. The neighboring field then points toward the target with a strictly positive restoring bound, yielding finite-time alignment in the idealized flow.

The field in the two models. On two lines, $A_c(\boldsymbol{z}) = \tau M_c^\tau \boldsymbol{u}_c$ for $\tau = \mathrm{sign}\langle \boldsymbol{z}, \boldsymbol{u}_c\rangle$, so G_s^0 is constant on each of the four activation regions. Under Assumption 2 the only equilibria are the four normals, opposite-class attracting and same-class repelling; if an inequality is reversed, a stable interior direction appears in the corresponding wide region, which we classify in Appendix B.5. On two hyperplanes, spherical symmetry gives $A_c(\boldsymbol{z}) = M_c \Pi_{S_c}\boldsymbol{z}/\|\Pi_{S_c}\boldsymbol{z}\|$ away from subspace normals. The flow has the same four normal equilibria and a new stationary set, $\mathcal{K} = K \cap \mathbb{S}^{d-1}$: for $\boldsymbol{z} \in \mathcal{K}$, $\Pi_{S_c}\boldsymbol{z} = \boldsymbol{z}$, so both active moments are parallel to $\boldsymbol{z}$ and the tangential field vanishes. We show that this stationary sphere is in fact a normally hyperbolic saddle manifold, with a codimension-one stable manifold. This is the source of the nonuniform alignment time in Theorem 3.2: trajectories initialized close to the stable manifold can remain near $\mathcal{K}$ for a long time before escaping toward a normal. The full characterization of the saddle and proofs are given in Appendix C.

5 Phase II: Cone Invariance, Fitting, and Realignment

Once the network output is no longer negligible, the residual moments $\mathcal{R}_c^\tau(t)$ defined in equation (5) of Section 2 replace M_c^τ in the local analysis of Section 4, and a normal is attracting when

$$\mathcal{R}_+^\tau(t)\cos\theta < \mathcal{R}_-^\tau(t), \qquad \mathcal{R}_-^\tau(t)\cos\theta < \mathcal{R}_+^\tau(t). \tag{12}$$

This is Assumption 2 with time-varying moments, and it holds throughout the early phase, where $\mathcal{R}_c^\tau(t) \approx M_c^\tau$. As the norms grow, however, the two residual moments evolve, typically at different rates as the moment of the class currently fit better shrinks faster, so (12) can fail; the corresponding normal then stops attracting, and its neurons may leave it. We show that such departures are confined to a cone, that the loss still converges to zero, and that the neurons return to their early normals.

The two endpoints of $\mathcal{C}(\nu_j)$ are inward for different reasons. At the data direction, the opposite class is fully activated and exerts a strictly inward force irrespective of the residual weights. At the normal, only the cooperative side of the neuron's own class can drive it outward, and that side points into $\mathcal{C}(\nu_j)$. A residual imbalance can therefore displace a neuron from its normal, but only along the arc already assigned to it: the neuron's feature assignment persists even when its exact alignment does not, which is why cone invariance requires no analogue of (12).

Cone invariance has a second consequence, which makes the remaining analysis tractable. Every neuron now activates each class on a fixed side of its support, so the loss depends on the network only through its four outputs $f(\tau \boldsymbol{u}_c; t)$ on the data rays, and $\mathcal{R}_c^\tau$ is exactly the sensitivity of the loss to the corresponding output (Appendix C). The late phase therefore reduces to a competition between four scalars: the loss decreases precisely when these outputs grow, and (12) is a comparison between the rates at which they do so. What remains is to determine whether the neurons return to their normals. We use different approaches to answer this question in each model.

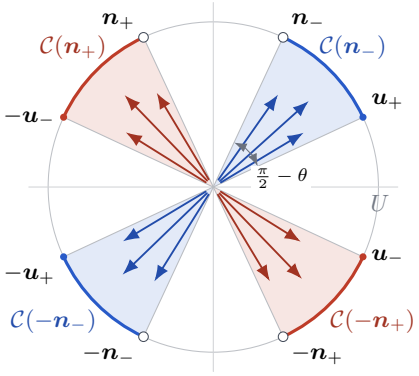


Figure 2: **Cone invariance in U.**

Two lines. In this model, the ratio in (12) need not return to the balance window, and Theorem 3.1 is proved without establishing that it does. Since all four groups are populated and every atom is bounded away from the origin, the corresponding margins diverge, giving $\mathcal{L} = \mathcal{O}(t^{-1})$ and divergence of all norms. The normalized margin is bounded below along the trajectory and the implicit bias of homogeneous networks applies (Lyu & Li, 2020): every normalized limit point is a KKT point of the parameter max-margin problem, and the only such points compatible with invariance to the cones $\mathcal{C}(\nu_j)$ are supported on the four normals.

Two hyperplanes. In this population setting, the class margins vanish, and the implicit-bias results invoked above do not apply. We therefore analyze the competition between the two classes directly, tracking $\mathcal{R}_+^\tau$ and $\mathcal{R}_-^\tau$ along the trajectory. Loss dissipation forces all four margin scores to diverge, and the resulting asymptotics are controlled by the data near the intersection. We show that the dynamics in this phase implicitly minimize an objective whose unique minimizer requires a residual ratio of 1 under spherical symmetry. As a result, $\mathcal{R}_-^\tau(t)/\mathcal{R}_+^\tau(t) \to 1$ and (12) is restored. Every normal regains a strict inward force which attracts the neurons in the corresponding cone. The remaining dynamics reduce to four scalar norm-growth equations and give $\mathcal{L} = \Theta(t^{-1/2})$.

6 Numerical Validation

We next examine whether the predicted alignment dynamics appear under gradient descent and how robust they are beyond the settings covered by our theory. We train width-32 two-layer ReLU networks with balanced initialization at scale $\varepsilon = 10^{-5}$ using full-batch gradient descent on the exponential loss. The ambient dimension is $d = 5$, the principal angle between the class subspaces is $\theta = 30^\circ$, and each class is sampled isotropically within its support. Figure 3 shows the median and interquartile range over five seeds. Additional implementation details are given in Appendix E. To track the learned geometry, for each neuron we measure its *normal error* $\left\|\Pi_{S_{-s_j}} \widehat{\boldsymbol{w}}_j\right\|$, which vanishes exactly when the neuron is orthogonal to the opposite-class subspace (determined by the output sign at initialization); for two lines we evaluate the normalized component of $\boldsymbol{w}_j$ in the data plane. The top row reports the worst normal error across neurons, while the bottom row shows the corresponding training loss. The three columns test the phenomenon in different regimes: the noiseless line and hyperplane models analyzed theoretically, perturbations away from exact subspace support, and representative intermediate subspace dimensions not covered by our analysis.

The clean experiments reproduce the predicted *align–depart–realign* pattern. For hyperplanes, the behavior is stable across sample sizes and closely matches a quasi-Monte Carlo approximation of the population dynamics. Appendix E.3 tests the early-phase predictions quantitatively and the alignement time bound of Theorem 3.1 across angles. Appendix E.4 shows that late realignment slows markedly as the subspace angle decreases. The same qualitative evolution persists under small off-subspace perturbations, although visible degradation appears by $\sigma = 10^{-2}$ (Figure 3). Finally, the phenomenon is not restricted to the two endpoint geometries treated theoretically: several intermediate, including unequal, subspace dimensions display the same characteristic alignment dynamics. These experiments therefore suggest that the normal-space feature selection identified by the theory persists beyond exact lines and hyperplanes, while becoming progressively less exact as the subspace model is perturbed. Full noise and dimension sweeps, including regimes with weaker realignment, are reported in Appendix E.

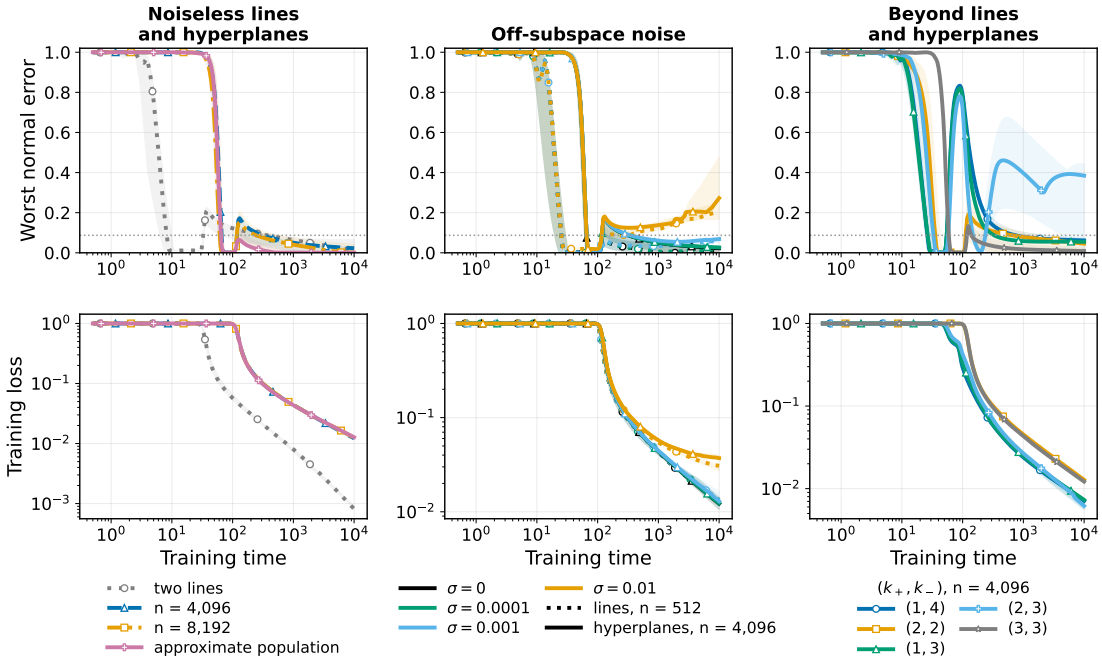


Figure 3: **Neuron alignment across geometries and perturbations.** Worst normal error (top) and training loss (bottom) for noiseless lines and hyperplanes (left), off-subspace noise (middle), and representative intermediate subspace dimensions (right). Curves show medians over five seeds with interquartile ranges.

7 Conclusion and Discussion

We studied how small-initialization gradient flow selects features when data is supported on a union of subspaces. Across two complementary subspace models, neurons align with subspace normals under suitable balance conditions, may depart from them during fitting, and ultimately recover the same directions as the loss vanishes. Experiments suggest that parts of this mechanism persist beyond our exact assumptions, and show where the theory begins to break down.

Early versus final features. In our models, the directions selected during early alignment can be lost and recovered, but this align–depart–realign pattern can break under perturbations such as data imbalance or noise. This raises the question of which properties of the data and training dynamics make early-selected features agree with those learned at convergence, and whether this agreement can be encouraged. If so, the learned directions could be locked in after alignment, with training continuing only to grow their magnitudes or optimize the readout. This would recover the final classifier without traversing the full fitting dynamics.

First-order versus geometric features. Our models show how gradient flow can recover discriminative structure beyond class-mean directions: when the first-order information is weak or absent, learned features instead reflect the geometry of the class supports, which in these symmetric models is closely tied to their covariance. Such features also have a different robustness profile: once a normal is selected, it no longer sees the opposite support pointwise and is thus insensitive to reweighting and distribution shifts within that support. More general data may contain both kinds of signal. This raises the question of whether training or preprocessing can suppress nuisance first-order cues in favor of geometric features that are more stable under distribution shift, and whether modern networks already exhibit such a transition implicitly.

Limitations. Our theoretical analysis is limited to binary classification with data exactly supported on two linear subspaces, small balanced initialization, and continuous-time gradient flow. Our results further rely on either atomic data on two lines or spherical symmetry on two hyperplanes, with balance conditions ensuring normal alignment. Extending the analysis to intermediate subspace dimensions, noisy or approximately subspace-supported data, and discrete gradient descent would clarify how robust the align–depart–realign mechanism is beyond these tractable models.

AI USE STATEMENT

We used generative AI tools to assist with brainstorming, mathematical proof checking, code development, and editing and restructuring the manuscript. All AI-assisted mathematical arguments, experimental code, and written material were independently reviewed and verified by the authors. We did not rely on generative AI as a source of scientific evidence or citations. The authors take full responsibility for the final content of this work, including all claims, proofs, experiments, and text produced with the aid of generative AI.

REFERENCES

Amirhesam Abedsoltan, Enric Boix-Adsera, Fivos Kalogiannis, and Mikhail Belkin. Revenge of monosemanticity: Neuron specialization as a new form of feature learning in MLPs, 2026. URL https://arxiv.org/abs/2608.24007.

Gerard Ben Arous, Reza Gheissari, and Aukosh Jagannath. High-dimensional limit theorems for SGD: Effective dynamics and critical scaling. In *Advances in Neural Information Processing Systems*, volume 35, 2022. URL https://proceedings.neurips.cc/paper_files/paper/2022/hash/a224ff18cc99a71751aa2b79118604da-Abstract-Conference.html.

Yoshua Bengio, Aaron Courville, and Pascal Vincent. Representation learning: A review and new perspectives. *IEEE transactions on pattern analysis and machine intelligence*, 35(8):1798–1828, 2013.

Raphaël Berthier. Incremental learning in diagonal linear networks. *Journal of Machine Learning Research*, 24(171):1–26, 2023. URL https://www.jmlr.org/papers/v24/22-1395.html.

David Bertoin, Jérôme Bolte, Sébastien Gerchinovitz, and Edouard Pauwels. Numerical influence of relu'(0) on backpropagation. In *Advances in Neural Information Processing Systems*, volume 34, pp. 468–479, 2021. URL https://proceedings.neurips.cc/paper/2021/hash/043ab21fc5a1607b381ac3896176dac6-Abstract.html.

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, 2025.

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

Kwan Ho Ryan Chan, Yaodong Yu, Chong You, Haozhi Qi, John Wright, and Yi Ma. Redunet: A white-box deep network from the principle of maximizing rate reduction. *Journal of Machine Learning Research*, 23(114):1–103, 2022.

Dmitry Chistikov, Matthias Englert, and Ranko Lazic. Learning a neuron by a shallow relu network: Dynamics and implicit bias for correlated inputs. In A. Oh, T. Neumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine (eds.), *Advances in Neural Information Processing Systems*, volume 36, pp. 23748–23760. Curran Associates, Inc., 2023. URL https://proceedings.neurips.cc/paper_files/paper/2023/file/4af24e6ce753c181e703f3f0be3b5e20-Paper-Conference.pdf.

Lénaïc Chizat, Edouard Oyallon, and Francis Bach. On lazy training in differentiable programming. In *Advances in Neural Information Processing Systems*, volume 32, pp. 2937–2947, 2019. URL https://papers.nips.cc/paper/8559-on-lazy-training-in-differentiable-programming.

Uri Cohen, SueYeon Chung, Daniel D Lee, and Haim Sompolinsky. Separability and geometry of object manifolds in deep neural networks. *Nature communications*, 11(1):1–13, 2020.

Samet Demir and Zafer Dogan. Asymptotic analysis of two-layer neural networks after one gradient step under gaussian mixtures data with structure. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=tNn6Hskmti.

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, 2024.

Morris W. Hirsch, Charles C. Pugh, and Michael Shub. *Invariant Manifolds*, volume 583 of *Lecture Notes in Mathematics*. Springer-Verlag, Berlin, Heidelberg, 1977. ISBN 978-3-540-08148-7. doi: 10.1007/BFb0092042.

Arthur Jacot, François Ged, Berfin Şimşek, Clément Hongler, and Franck Gabriel. Saddle-to-saddle dynamics in deep linear networks: Small initialization training, symmetry, and sparsity, 2021. URL https://arxiv.org/abs/2106.15933.

Akshay Kumar and Jarvis Haupt. Directional convergence near small initializations and saddles in two-homogeneous neural networks. *Transactions on Machine Learning Research*, 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, 2025a. URL https://www.jmlr.org/papers/v26/25-1089.html.

Akshay Kumar and Jarvis Haupt. Early directional convergence in deep homogeneous neural networks for small initializations. *Transactions on Machine Learning Research*, 2025b.

Daniel Kunin, Giovanni Luca Marchetti, Feng Chen, Dhruva Karkada, James B. Simon, Michael R. DeWeese, Surya Ganguli, and Nina Miolane. Alternating gradient flows: A theory of feature learning in two-layer neural networks. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 38, 2025. URL https://arxiv.org/abs/2506.06489.

Binghui Li, Zhixuan Pan, Kaifeng Lyu, and Jian Li. Feature averaging: An implicit bias of gradient descent leading to non-robustness in neural networks. In *The Thirteenth International Conference on Learning Representations*, 2025.

Kaifeng Lyu and Jian Li. Gradient descent maximizes the margin of homogeneous neural networks. In *International Conference on Learning Representations (ICLR)*, 2020. URL https://openreview.net/forum?id=SJeLIgBKPS.

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

Stéphane Mallat. Understanding deep convolutional networks. *Philosophical transactions. Series A, Mathematical, physical, and engineering sciences*, 374(2065):20150203, 2016.

Francesca Mignacco, Florent Krzakala, Pierfrancesco Urbani, and Lenka Zdeborová. Dynamical mean-field theory for stochastic gradient descent in gaussian mixture classification. In *Advances in Neural Information Processing Systems*, volume 33, pp. 9540–9550, 2020. URL https://proceedings.neurips.cc/paper/2020/hash/6c81c83c4bd0b58850495f603ab45a93-Abstract.html.

Hancheng Min and Rene Vidal. Gradient flow provably learns robust classifiers for orthonormal GMMs. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 44292–44350. PMLR, 2025. URL https://proceedings.mlr.press/v267/min25c.html.

Hancheng Min, Enrique Mallada, and René Vidal. Early neuron alignment in two-layer ReLU networks with small initialization. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=QibPzdVrRu.

Hancheng Min, Zhihui Zhu, and René Vidal. Neural collapse under gradient flow on shallow ReLU networks for orthogonally separable data. In *Advances in Neural Information Processing Systems*, volume 38, pp. 105551–105590, 2025. doi: 10.52202/085713-3184.

Mary Phuong and Christoph H. Lampert. The inductive bias of ReLU networks on orthogonally separable data. In *International Conference on Learning Representations*, 2021.

Akshay Rangamani, Marius Lindegaard, Tomer Galanti, and Tomaso A Poggio. Feature learning in deep classifiers through intermediate neural collapse. In *International Conference on Machine Learning*, pp. 28729–28745. PMLR, 2023.

Maria Refinetti, Sebastian Goldt, Florent Krzakala, and Lenka Zdeborová. Classifying high-dimensional gaussian mixtures: Where kernel methods fail and neural networks succeed. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 8936–8947. PMLR, 2021. URL https://proceedings.mlr.press/v139/refinetti21b.html.

David E Rumelhart, Geoffrey E Hinton, and Ronald J Williams. Learning representations by back-propagating errors. *nature*, 323(6088):533–536, 1986.

Ian Tenney, Dipanjan Das, and Ellie Pavlick. Bert rediscovers the classical nlp pipeline. In *Proceedings of the 57th annual meeting of the association for computational linguistics*, pp. 4593–4601, 2019.

Nikita Tsoy and Nikola Konstantinov. Simplicity bias of two-layer networks beyond linearly separable data. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*. PMLR, 2024. URL https://proceedings.mlr.press/v235/tsoy24a.html.

René Vidal. Subspace clustering. *IEEE Signal Processing Magazine*, 28(2):52–68, 2011. doi: 10.1109/MSP.2010.939739.

René Vidal, Yi Ma, and Shankar Sastry. Generalized principal component analysis (GPCA). *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 27(12):1945–1959, 2005. doi: 10.1109/TPAMI.2005.244.

Taesun Yeom, Taehyeok Ha, and Jaeho Lee. Over-alignment vs over-fitting: The role of feature learning strength in generalization. In *Proceedings of the 43rd International Conference on Machine Learning*, 2026. URL https://arxiv.org/abs/2602.00827. arXiv:2602.00827.

Yaodong Yu, Kwan Ho Ryan Chan, Chong You, Chaobing Song, and Yi Ma. Learning diverse and discriminative representations via the principle of maximal coding rate reduction. In *Advances in Neural Information Processing Systems*, volume 33. Curran Associates, Inc., 2020.

Yaodong Yu, Sam Buchanan, Druv Pai, Tianzhe Chu, Ziyang Wu, Shengbang Tong, Benjamin D. Haeffele, and Yi Ma. White-box transformers via sparse rate reduction. In *Advances in Neural Information Processing Systems*, volume 36. Curran Associates, Inc., 2023.

Matthew D Zeiler and Rob Fergus. Visualizing and understanding convolutional networks. In *European conference on computer vision*, pp. 818–833. Springer, 2014.

Hanxu Zhou, Qixuan Zhou, Tao Luo, Yaoyu Zhang, and Zhi-Qin John Xu. Towards understanding the condensation of neural networks at initial training. In *Advances in Neural Information Processing Systems*, volume 35, pp. 2184–2196, 2022. doi: 10.52202/068431-0159.

CONTENTS

1 Introduction **1**

2 Setup and data geometry **3**

3 Main results **4**

4 Phase I: early alignment **6**

5 Phase II: cone invariance, fitting, and realignment **7**

6 Numerical validation **8**

7 Conclusion and discussion **9**

A Results and proofs that apply to both models **15**

A.1 Notation 15

A.2 Clarke dynamics, the chain rule, and global existence 15

A.3 Conservation laws and exact neuron equations 18

A.4 Small-initialization control and the ideal early field 19

A.5 Planar dynamics, cone invariance, and score reduction 22

B Proofs specific to the two-line model **27**

B.1 Exact projected dynamics and the idealized field 28

B.2 Finite time alignment 29

B.3 Fitting rate and divergence of every projected neuron 32

B.4 Within-group agreement and KKT characterization of the limiting directions 34

B.5 Imbalanced regimes and conditional finite-time realignment 38

C Proofs for the two-hyperplane case **42**

C.1 Rotational moments and the ideal potential 43

C.2 The intersection saddle and global ideal basins 44

C.3 Robust transfer to exact early alignment 46

C.4 Random initialization and four-group coverage 47

C.5 Strong quotient-plane invariance after exact alignment 50

C.6 Spherical projection asymptotics and normalized-margin barriers 52

C.7 The tail variational problem and convergence 55

C.8 Residual balance, permanent realignment, and final rates 58

D Extended related work **60**

E Additional experimental details and results **61**

E.1 Full alignment, loss, and classification results 64
E.2 Residual balance during fitting 64
E.3 Quantitative early alignment time validation 65
E.4 Dependence on the subspace angle 66

A RESULTS AND PROOFS THAT APPLY TO BOTH MODELS

The two models share the analytic structure of training and, after the appropriate geometric reduction, the mechanism that drives the loss to zero. We establish these facts first. Early alignment and late directional recovery depend on the support geometry and are proved separately in Sections B and C. In particular, the late results below are conditional on a precise planar reduction; they do not assume that this reduction follows from symmetry alone.

A.1 NOTATION

Recall $\xi = (W, \boldsymbol{v})$ and

$$f(\boldsymbol{x};\xi) = \sum_{j=1}^{m} v_j \sigma(\langle \boldsymbol{w}_j, \boldsymbol{x}\rangle), \qquad \sigma(a) = \max\{a, 0\}, \qquad \mathcal{L}(\xi) = \mathbb{E}[e^{-Yf(X;\xi)}].$$

The parameter norm is $\|\xi\|^2 = \|W\|_F^2 + \|\boldsymbol{v}\|^2$. All expectations are under the joint probability law of (X, Y), with $\mathbb{P}(Y = c) > 0$ for $c \in \{+1, -1\}$. For a finite atomic law this means $\mathcal{L} = \sum_i p_i e^{-y_i f(\boldsymbol{x}_i)}$, where $p_i > 0$ and $\sum_i p_i = 1$. Any conversion to an unnormalized empirical loss must also rescale time and the constants in the gradient-flow estimates.

Integrability. Put $R = \|X\|$, whose class-c conditional law is that of ρ_c, and distinguish

$$\mathbb{E}[R^2 e^{q_* R}] < \infty \quad \text{for some } q_* > 0, \qquad \text{(local integrability)}$$
$$\mathbb{E}[R^2 e^{qR}] < \infty \quad \text{for every } q > 0. \qquad \text{(global integrability)}$$

These joint conditions are equivalent to their classwise versions. Finite support satisfies both. Local integrability suffices for the controlled early interval. Global integrability is assumed for the unconditional global and late-stage statements.

Neuron variables and balance. Set

$$Q(\xi) = \sum_j |v_j| \, \|\boldsymbol{w}_j\|, \qquad r_j = \|\boldsymbol{w}_j\|, \qquad \widehat{\boldsymbol{w}}_j = \boldsymbol{w}_j / r_j,$$
$$a_j = \Pi_U \boldsymbol{w}_j, \qquad \varrho_j = \|a_j\|. \tag{13}$$

The directions $\widehat{\boldsymbol{w}}_j$ and $\widehat{a}_j = a_j/\varrho_j$ are used only when their denominators are positive. Nonzero balanced initialization means $\boldsymbol{w}_j(0) \neq 0$ and $v_j(0) = s_j \|\boldsymbol{w}_j(0)\|$ with $s_j \in \{+1, -1\}$. Small initialization further specifies $\boldsymbol{w}_j(0) = \varepsilon \bar{\boldsymbol{w}}_j$ for fixed nonzero shapes $\bar{\boldsymbol{w}}_j$. This parameter balance is distinct from balance of the data's active moments. Recall the unweighted moments and their residual-weighted counterparts:

$$M_c^\tau = \mathbb{E}[|Z_c| \mathbf{1}_{\{Y=c,\, \tau Z_c > 0\}}],$$
$$\mathcal{R}_c^\tau(t) = \mathbb{E}[|Z_c| e^{-Yf(X;\xi(t))} \mathbf{1}_{\{Y=c,\, \tau Z_c > 0\}}]$$

No active-moment balance is assumed in this section.

A.2 CLARKE DYNAMICS, THE CHAIN RULE, AND GLOBAL EXISTENCE

We study locally absolutely continuous trajectories satisfying

$$\dot{\xi}(t) \in -\partial \mathcal{L}(\xi(t)) \qquad \text{for almost every } t. \tag{14}$$

All statements below hold for every such trajectory.

To describe the admissible ReLU subgradient at an activation boundary, we define

$$H(a) := \begin{cases} \{1\}, & a > 0, \\ [0, 1], & a = 0, \\ \{0\}, & a < 0. \end{cases}$$

Thus $H(a)$ equals the usual ReLU subgradient away from the kink and allows any value in $[0, 1]$ at $a = 0$.

Proposition 1. *Under local integrability, $\mathcal{L}$ is finite and locally Lipschitz on $\Omega_* := \{\xi : Q(\xi) < q_*\}$. Moreover, every $V \in -\partial\mathcal{L}(\xi)$, $\xi \in \Omega_*$, can be written as*

$$V_{\boldsymbol{w}_j} = v_j \mathbb{E}\Big[Y e^{-Yf(X;\xi)} \alpha_j(X,Y) X\Big],$$
$$V_{v_j} = \mathbb{E}\Big[Y e^{-Yf(X;\xi)} \sigma(\langle \boldsymbol{w}_j, X\rangle)\Big], \qquad \alpha_j(X,Y) \in H(\langle \boldsymbol{w}_j, X\rangle), \tag{15}$$

for measurable samplewise choices α_j. Under global integrability, the same conclusions hold globally.

Proof. Fix $\xi_0 \in \Omega_*$. Since $|f(\boldsymbol{x};\xi)| \le Q(\xi)\|\boldsymbol{x}\|$, there is a bounded neighborhood U of ξ_0 and $b < q_*$ such that $Q(\xi) \le b$ on U. ReLU Lipschitzness and boundedness of the parameters on U give

$$|f(\boldsymbol{x};\xi') - f(\boldsymbol{x};\xi'')| \le C_U \|\boldsymbol{x}\| \, \|\xi' - \xi''\|, \qquad \xi', \xi'' \in U.$$

As a result, the sample loss $\ell_{\boldsymbol{x},y}(\xi) := e^{-yf(\boldsymbol{x};\xi)}$ satisfies

$$|\ell_{\boldsymbol{x},y}(\xi') - \ell_{\boldsymbol{x},y}(\xi'')| \le C_U \|\boldsymbol{x}\| e^{b\|\boldsymbol{x}\|} \|\xi' - \xi''\|.$$

The envelope $\|\boldsymbol{x}\|e^{b\|\boldsymbol{x}\|}$, as well as $e^{b\|\boldsymbol{x}\|}$, is integrable under local integrability: on $\{\|X\| \le 1\}$ it is bounded, while on $\{\|X\| > 1\}$ it is dominated up to a constant by $\|X\|^2 e^{q_*\|X\|}$. Thus $\mathcal{L} = \mathbb{E}[\ell_{X,Y}]$ is finite and locally Lipschitz on Ω_*.

The standard Clarke subdifferentiation rule for integral functionals therefore gives

$$\partial\mathcal{L}(\xi) \subseteq \mathbb{E}[\partial \ell_{X,Y}(\xi)], \tag{16}$$

where the expectation denotes the set of expectations of measurable samplewise selections.

Next, we consider any fixed $(\boldsymbol{x}, y)$, the jth admissible ReLU subgradient depends only on $\boldsymbol{w}_j$, and the Clarke chain rule gives that every element of $-\partial\ell_{\boldsymbol{x},y}(\xi)$ has blocks

$$y e^{-yf(\boldsymbol{x};\xi)} v_j \alpha_j \boldsymbol{x}, \qquad y e^{-yf(\boldsymbol{x};\xi)} \sigma(\langle \boldsymbol{w}_j, \boldsymbol{x}\rangle), \qquad \alpha_j \in H(\langle \boldsymbol{w}_j, \boldsymbol{x}\rangle).$$

Combining this samplewise description with (16) yields (15). The inclusion in (16) need not be an equality, and samplewise choices of the α_j need not produce an element of the joint Clarke subdifferential.

Under global integrability, the same argument applies on every bounded parameter neighborhood. □

Proposition 2 (Chain rule and energy dissipation)**.** *For every locally absolutely continuous parameter path $\xi(\cdot)$ in Ω_*, the map $t \mapsto \mathcal{L}(\xi(t))$ is locally absolutely continuous and, for almost every t,*

$$\frac{d}{dt}\mathcal{L}(\xi(t)) = \langle g, \dot\xi(t)\rangle \qquad \text{for every } g \in \partial\mathcal{L}(\xi(t)). \tag{17}$$

Consequently, every training trajectory satisfies

$$\mathcal{L}(\xi(t)) + \int_a^t \|\dot\xi(u)\|^2 \, du = \mathcal{L}(\xi(a)) \tag{18}$$

on every compact interval $[a, t]$ in its domain. Under global integrability, the same statements hold without restricting the path to Ω_*.

Proof. Fix a compact time interval I on which $\xi(t) \in \Omega_*$. Since $\xi(I)$ is compact and $Q(\xi(t)) < q_*$, there is $b < q_*$ such that $Q(\xi(t)) \le b$ on I. The local Lipschitz bound from Proposition 1 then implies that $\mathcal{L} \circ \xi$ is absolutely continuous on I.

To identify its derivative, consider a sample $(\boldsymbol{x}, y)$ and write

$$b_j(t, \boldsymbol{x}) := \langle \boldsymbol{w}_j(t), \boldsymbol{x}\rangle.$$

The function $t \mapsto b_j(t, \boldsymbol{x})$ is absolutely continuous. Therefore, we have

$$\dot b_j(t, \boldsymbol{x}) = 0 \qquad \text{for almost every } t \text{ such that } b_j(t, \boldsymbol{x}) = 0.$$

And for almost every t,

$$\frac{d}{dt}\sigma(b_j(t,\boldsymbol{x})) = \alpha_j\,\dot{b}_j(t,\boldsymbol{x}) \qquad \text{for every } \alpha_j \in H(b_j(t,\boldsymbol{x})). \tag{19}$$

Away from the activation boundary this is the ordinary ReLU chain rule, while on the boundary both sides vanish almost everywhere.

Using the product rule followed by the ordinary chain rule for the exponential, shows that for almost every t,

$$\frac{d}{dt}\ell_{\boldsymbol{x},y}(\xi(t)) = \langle s, \dot{\xi}(t)\rangle \qquad \text{for every } s \in \partial\ell_{\boldsymbol{x},y}(\xi(t)). \tag{20}$$

The key point is that at a zero preactivation the right-hand side is independent of the admissible subgradient value $\alpha_j \in [0,1]$, because $\dot{b}_j(t,\boldsymbol{x}) = 0$ there almost everywhere.

On I, the sample subgradient is bounded by the local Lipschitz envelope from Proposition 1:

$$\left|\frac{d}{dt}\ell_{X,Y}(\xi(t))\right| \le CRe^{bR}\|\dot{\xi}(t)\|.$$

The right-hand side is integrable jointly in time and sample. Fubini's theorem therefore yields a full-measure set of times on which (20) holds almost surely and may be integrated over (X,Y). In particular,

$$\frac{d}{dt}\mathcal{L}(\xi(t)) = \mathbb{E}\Big[\langle s(X,Y), \dot{\xi}(t)\rangle\Big]$$

for every measurable samplewise selection representing a population Clarke subgradient. By Proposition 1, every $g \in \partial\mathcal{L}(\xi(t))$ admits such a representation. Therefore

$$\frac{d}{dt}\mathcal{L}(\xi(t)) = \langle g, \dot{\xi}(t)\rangle \qquad \text{for every } g \in \partial\mathcal{L}(\xi(t)),$$

which proves (17).

Finally, along a training trajectory, $-\dot{\xi}(t) \in \partial\mathcal{L}(\xi(t))$ almost everywhere. Taking $g = -\dot{\xi}(t)$ in (17) gives

$$\frac{d}{dt}\mathcal{L}(\xi(t)) = -\|\dot{\xi}(t)\|^2 \qquad \text{a.e.}$$

Integrating over $[a,t]$ yields (18). Under global integrability, the same argument applies on every compact parameter region. □

Theorem A.1 (Global existence)**.** *Under global integrability, a solution of* (14) *exists from every initial condition, and every maximal solution is defined for all* $t \ge 0$*. Moreover, for* $0 \le a \le t$*,*

$$\|\xi(t) - \xi(a)\| \le \sqrt{(t-a)\,[\mathcal{L}(\xi(a)) - \mathcal{L}(\xi(t))]}, \tag{21}$$

$$\|\xi(t)\| \le \|\xi(0)\| + \sqrt{t\,\mathcal{L}(\xi(0))}. \tag{22}$$

Proof. Under global integrability, Proposition 1 shows that $\mathcal{L}$ is locally Lipschitz on the whole parameter space. Its negative Clarke subdifferential is therefore upper semicontinuous and locally bounded, with nonempty compact convex values. The standard existence theorem for finite-dimensional differential inclusions gives a local solution from every initial condition.

Let ξ be any such solution. By the energy identity,

$$\int_a^t \|\dot{\xi}(u)\|^2\,du = \mathcal{L}(\xi(a)) - \mathcal{L}(\xi(t)).$$

Using absolute continuity and Cauchy–Schwarz, we get

$$\begin{aligned}\|\xi(t) - \xi(a)\| &\le \int_a^t \|\dot{\xi}(u)\|\,du \\ &\le \sqrt{t-a}\left(\int_a^t \|\dot{\xi}(u)\|^2\,du\right)^{1/2},\end{aligned}$$

which gives (21). We obtain (22) by taking $a = 0$ and using $\mathcal{L} \geq 0$.

It remains to rule out finite-time blow-up. Suppose a maximal solution were defined only on $[0, T)$ with $T < \infty$. For $a < t < T$, (21) and monotonicity of the loss give

$$\|\xi(t) - \xi(a)\| \leq \sqrt{(t-a)\,\mathcal{L}(\xi(a))} \leq \sqrt{(t-a)\,\mathcal{L}(\xi(0))}.$$

Thus $\xi(t)$ is Cauchy as $t \uparrow T$ and converges to some finite ξ_T. Moreover,

$$\int_0^T \|\dot{\xi}(u)\|^2\, du \leq \mathcal{L}(\xi(0)) < \infty,$$

so the trajectory extends absolutely continuously to $t = T$ by setting $\xi(T) = \xi_T$.

Since the Clarke field is locally well defined at ξ_T, a local solution starting from ξ_T exists. Concatenating it with the original trajectory extends the solution beyond T, contradicting maximality. Hence every maximal solution is global. □

A.3 Conservation laws and exact neuron equations

Proposition 3 (Balancedness, nonvanishing, and fixed output signs)**.** *Along every training trajectory, $v_j(t)^2 - \|\boldsymbol{w}_j(t)\|^2$ is constant in time. In particular, under nonzero balanced initialization,*

$$v_j(t) = s_j r_j(t), \qquad r_j(t) > 0, \tag{23}$$

at every finite admissible time.

Moreover, on any compact interval $I = [a, b]$ on which $Q \leq B$ and $M_B := \mathbb{E}[Re^{BR}] < \infty$, one has

$$r_j(a)e^{-M_B(t-a)} \leq r_j(t) \leq r_j(a)e^{M_B(t-a)}, \qquad t \in I. \tag{24}$$

Proof. For every $\alpha \in H(u)$, we have $\alpha u = \sigma(u)$. Using Proposition 1, we therefore obtain almost everywhere

$$\langle \boldsymbol{w}_j, \dot{\boldsymbol{w}}_j \rangle = v_j \dot{v}_j.$$

Hence

$$\frac{d}{dt}\left(v_j^2 - \|\boldsymbol{w}_j\|^2\right) = 2v_j\dot{v}_j - 2\langle \boldsymbol{w}_j, \dot{\boldsymbol{w}}_j \rangle = 0$$

almost everywhere, and absolute continuity gives conservation for all times.

Under balanced initialization this implies $|v_j(t)| = r_j(t)$ as long as the trajectory exists. On an interval where $Q \leq B$, we have using (15)

$$\|\dot{\boldsymbol{w}}_j\| \leq M_B |v_j|, \qquad |\dot{v}_j| \leq M_B r_j.$$

Thus, wherever $r_j > 0$,

$$|\dot{r}_j| = \frac{|\langle \boldsymbol{w}_j, \dot{\boldsymbol{w}}_j \rangle|}{r_j} \leq M_B r_j.$$

Integrating the differential inequalities $-M_B r_j \leq \dot{r}_j \leq M_B r_j$ gives (24) up to any possible first time at which r_j vanishes. The lower bound remains strictly positive at such a time, so no finite-time zero can occur. Since $|v_j| = r_j > 0$, continuity preserves the initial sign of v_j, proving (23). □

Proposition 4 (Frozen components orthogonal to the data span)**.** *Let $D \subseteq \mathbb{R}^d$ be a linear subspace such that $\mathbb{P}(X \in D) = 1$. Then, along every training trajectory,*

$$\Pi_{D^\perp} \boldsymbol{w}_j(t) = \Pi_{D^\perp} \boldsymbol{w}_j(0). \tag{25}$$

Proof. By (15), every first-layer velocity is an expectation of scalar multiples of X. Hence $\dot{\boldsymbol{w}}_j(t) \in D$ almost everywhere. Therefore

$$\frac{d}{dt}\Pi_{D^\perp} \boldsymbol{w}_j(t) = 0$$

almost everywhere, and integration gives (25). □

For the two-line model, the data span is $D = U$. We therefore write

$$\boldsymbol{w}_j = a_j + b_j, \qquad a_j := \Pi_U \boldsymbol{w}_j, \qquad b_j := \Pi_{U^\perp} \boldsymbol{w}_j,$$

where b_j is constant along training. If $\varrho_j := \|a_j\|$, then balancedness gives the exact identities

$$r_j^2 = \varrho_j^2 + \|b_j\|^2, \qquad v_j^2 - \|a_j\|^2 = \|b_j\|^2. \tag{26}$$

Thus the projected neuron is generally not itself balanced unless $b_j = 0$.

For the two-hyperplane model, $S_+ + S_- = \mathbb{R}^d$, so Proposition 4 gives no nontrivial frozen component. In particular, it does not imply that the components in $K = S_+ \cap S_-$ remain zero after alignment. That stronger invariance property is proved separately in Section C.5.

Proposition 5 (Exact radial and directional dynamics)**.** *Along a balanced trajectory, define* $V_j := \frac{\dot{\boldsymbol{w}}_j}{r_j}$ *almost everywhere. Then*

$$\begin{aligned} \dot{r}_j &= r_j \langle \widehat{\boldsymbol{w}}_j, V_j \rangle, & \dot{\widehat{\boldsymbol{w}}}_j &= (I - \widehat{\boldsymbol{w}}_j \widehat{\boldsymbol{w}}_j^\top) V_j, \\ \dot{a}_j &= r_j \Pi_U V_j, & \dot{\varrho}_j &= r_j \langle \widehat{a}_j, V_j \rangle, \end{aligned} \tag{27}$$

where

$$\widehat{\boldsymbol{w}}_j := \frac{\boldsymbol{w}_j}{r_j}, \qquad \widehat{a}_j := \frac{a_j}{\varrho_j},$$

and the formulas involving $\widehat{a}_j$ *are used only when* $\varrho_j > 0$*. On this set,*

$$\dot{\widehat{a}}_j = \frac{r_j}{\varrho_j} \big(\Pi_U - \widehat{a}_j \widehat{a}_j^\top\big) V_j, \qquad \frac{r_j}{\varrho_j} \geq 1. \tag{28}$$

Proof. Since $\dot{\boldsymbol{w}}_j = r_j V_j$,

$$\dot{r}_j = \left\langle \frac{\boldsymbol{w}_j}{r_j}, \dot{\boldsymbol{w}}_j \right\rangle = r_j \langle \widehat{\boldsymbol{w}}_j, V_j \rangle.$$

Differentiating $\widehat{\boldsymbol{w}}_j = \boldsymbol{w}_j / r_j$ then gives

$$\dot{\widehat{\boldsymbol{w}}}_j = (I - \widehat{\boldsymbol{w}}_j \widehat{\boldsymbol{w}}_j^\top) V_j.$$

For the projected component,

$$\dot{a}_j = \Pi_U \dot{\boldsymbol{w}}_j = r_j \Pi_U V_j.$$

Whenever $\varrho_j > 0$,

$$\dot{\varrho}_j = \langle \widehat{a}_j, \dot{a}_j \rangle = r_j \langle \widehat{a}_j, V_j \rangle,$$

because $\widehat{a}_j \in U$. Finally, differentiating $\widehat{a}_j = a_j / \varrho_j$ yields

$$\dot{\widehat{a}}_j = \frac{1}{\varrho_j} (I - \widehat{a}_j \widehat{a}_j^\top) \dot{a}_j = \frac{r_j}{\varrho_j} (\Pi_U - \widehat{a}_j \widehat{a}_j^\top) V_j.$$

Since $\varrho_j = \|\Pi_U \boldsymbol{w}_j\| \leq \|\boldsymbol{w}_j\| = r_j$, the final inequality follows. □

A.4 Small-initialization control and the ideal early field

Fix $0 < q_0 < q_*$ under local integrability and define

$$M_0 := \mathbb{E}[R e^{q_0 R}], \qquad C_2 := \mathbb{E}[R^2 e^{q_0 R}]. \tag{29}$$

Both quantities are finite and positive for the data models considered here.

Theorem A.2. *If* $Q(0) < q_0$*, then every trajectory can be continued at least through*

$$T_\varepsilon := \frac{1}{2M_0} \log \frac{q_0}{Q(0)}, \tag{30}$$

and, for every $0 \leq t \leq T_\varepsilon$*,*

$$r_j(0) e^{-M_0 t} \leq r_j(t) \leq r_j(0) e^{M_0 t}, \qquad Q(t) \leq Q(0) e^{2M_0 t} \leq q_0. \tag{31}$$

Proof. As long as $Q(t) \le q_0$, the samplewise velocity representation and $|f(X;\xi)| \le Q(t)R$ give

$$\|V_j(t)\| = \left\| s_j \mathbb{E}\Big[Y e^{-Y f(X;\xi(t))} \alpha_j X \Big] \right\| \le \mathbb{E}[R e^{q_0 R}] = M_0.$$

Proposition 5 therefore gives

$$|\dot{r}_j(t)| \le M_0 r_j(t).$$

Integrating,

$$r_j(0) e^{-M_0 t} \le r_j(t) \le r_j(0) e^{M_0 t}$$

while $Q \le q_0$. Since balancedness gives $Q = \sum_j r_j^2$, it follows that

$$Q(t) \le Q(0) e^{2M_0 t}.$$

Suppose the trajectory reached $Q = q_0$ for the first time at some $t < T_\varepsilon$. Then

$$Q(t) \le Q(0) e^{2M_0 t} < Q(0) e^{2M_0 T_\varepsilon} = q_0,$$

leading to a contradiction. Hence $Q(t) < q_0$ for $t < T_\varepsilon$, and continuity gives the bound at $t = T_\varepsilon$ as well.

The trajectory remains bounded and inside the region where the loss is locally Lipschitz, so the local continuation theorem rules out an earlier maximal endpoint. □

The ideal field. Recall the ideal directional dynamics

$$\dot{\boldsymbol{z}} \in \Pi_{\boldsymbol{z}} \mathcal{G}_s^0(\boldsymbol{z}), \qquad \mathcal{G}_s^0(\boldsymbol{z}) := \overline{\mathrm{co}} \left\{ \lim_{k\to\infty} G_s^0(\boldsymbol{z}_k) : \boldsymbol{z}_k \in \Omega,\ \boldsymbol{z}_k \to \boldsymbol{z} \right\}, \qquad \|\boldsymbol{z}\| = 1. \tag{32}$$

where $G_s^0(\boldsymbol{z}) := s\mathbb{E}\big[Y X \mathbf{1}_{\{\langle \boldsymbol{z}, X\rangle > 0\}} \big]$, and Ω denotes the corresponding set of regular directions.

The next lemma gives an equivalent and more convenient description of this completed field in terms of the admissible ReLU subgradient values introduced above.

Lemma 1. *For either data model and every unit vector $\boldsymbol{z}$,*

$$\mathcal{G}_s^0(\boldsymbol{z}) = \{ s\mathbb{E}[Y\alpha X] : \alpha = \alpha(X,Y) \textit{ measurable}, \quad \alpha \in H(\langle \boldsymbol{z}, X\rangle) \textit{ a.s.} \}. \tag{33}$$

Proof. We treat the two models separately.

Two lines. For a class c with $\langle \boldsymbol{z}, \boldsymbol{u}_c \rangle \neq 0$, the sign of the preactivation is fixed on each ray, so its contribution is uniquely determined. If instead $\langle \boldsymbol{z}, \boldsymbol{u}_c \rangle = 0$, then every point of that class lies on the activation boundary. $\alpha \in [0,1]$ produces exactly the interval $[-M_c^-, M_c^+]\,\boldsymbol{u}_c$ of possible unweighted class moments. Its two endpoints are precisely the limits obtained by perturbing $\boldsymbol{z}$ to the two sides of the boundary.

If $\boldsymbol{z} \in U^\perp$, both class projections vanish simultaneously. Since $\boldsymbol{u}_+$ and $\boldsymbol{u}_-$ are linearly independent, arbitrarily small perturbations inside U can realize all four sign combinations of the two projections. The convex hull of these four neighboring values is exactly the sum of the two class intervals, with the appropriate label signs. This proves (33) for the line model.

Two hyperplanes. Fix a class c. If $\Pi_{S_c} \boldsymbol{z} \neq 0$, the activation boundary has class-conditional probability zero and the subgradient is fixed almost surely. Spherical symmetry gives

$$\mathbb{E}\big[X \mathbf{1}_{\{Y=c,\ \langle \boldsymbol{z}, X\rangle > 0\}} \big] = M_c \frac{\Pi_{S_c} \boldsymbol{z}}{\|\Pi_{S_c} \boldsymbol{z}\|}.$$

Indeed, rotational invariance inside S_c forces the expectation to be parallel to $\Pi_{S_c} \boldsymbol{z}$, and its component in that direction determines the magnitude.

Now suppose $\Pi_{S_c} \boldsymbol{z} = 0$, so that $\boldsymbol{z}$ is a normal to S_c. The possible class-c selector moments form

$$\mathcal{C}_c := \big\{ \mathbb{E}[\alpha X \mathbf{1}_{\{Y=c\}}] : 0 \le \alpha \le 1 \big\}.$$

For any $h \in \mathbb{R}^d$, its support function is

$$\begin{aligned}\sup_{u \in \mathcal{C}_c} \langle h, u \rangle &= \sup_{0 \le \alpha \le 1} \mathbb{E}\big[\alpha \langle h, X \rangle \mathbf{1}_{\{Y=c\}}\big] \\ &= \mathbb{E}\big[(\langle h, X \rangle)_+ \mathbf{1}_{\{Y=c\}}\big] \\ &= M_c \|\Pi_{S_c} h\|.\end{aligned}$$

Therefore

$$\mathcal{C}_c = M_c \mathbb{B}_{S_c}.$$

Perturbing the normal $\boldsymbol{z}$ toward an arbitrary unit direction in S_c makes the regular class moment converge to the corresponding point on $M_c \mathbb{S}_{S_c}$. Thus the nearby regular traces generate the entire boundary of this ball, and convexification generates the ball itself. Because the two hyperplanes are distinct, a unit vector cannot be normal to both simultaneously, so the contribution of the other class varies continuously at such a point. This proves (33). □

Theorem A.3 (Approximation of the early field)**.** *At almost every time of a balanced trajectory satisfying $Q(t) \le q_0$, there exists $V_j^0(t) \in \mathcal{G}_{s_j}^0(\widehat{\boldsymbol{w}}_j(t))$ such that*

$$\|V_j(t) - V_j^0(t)\| \le \delta_{\mathrm{fld}}(Q(t)), \tag{34}$$

or, equivalently

$$\dot{\widehat{\boldsymbol{w}}}_j(t) = \Pi_{\widehat{\boldsymbol{w}}_j(t)} V_j^0(t) + e_j(t), \qquad \|e_j(t)\| \le \delta_{\mathrm{fld}}(Q(t)), \tag{35}$$

where

$$\delta_{\mathrm{fld}}(q) := \mathbb{E}\big[R(e^{qR} - 1)\big] \le C_2 q, \qquad 0 \le q \le q_0. \tag{36}$$

In particular, if $Q(0)e^{2M_0 T} \le q_0$, then $\delta_{\mathrm{fld}}(Q(t)) \le C_2 Q(0) e^{2M_0 T}$ for almost every $t \in [0, T]$.

Proof. At almost every time, Proposition 1 represents the first-layer velocity using subgradients $\alpha_j(X, Y) \in H(\langle \widehat{\boldsymbol{w}}_j, X \rangle)$. Using $v_j = s_j r_j$ and dividing by r_j gives

$$V_j = s_j \mathbb{E}\Big[Y e^{-Y f(X;\xi)} \alpha_j X\Big].$$

Define the corresponding unweighted velocity using the same subgradient: $V_j^0 := s_j \mathbb{E}[Y \alpha_j X]$. By Lemma 1 we have $V_j^0 \in \mathcal{G}_{s_j}^0(\widehat{\boldsymbol{w}}_j)$.

Since $|f(X;\xi)| \le QR$,

$$\begin{aligned}\|V_j - V_j^0\| &\le \mathbb{E}\Big[R\,|e^{-Y f(X;\xi)} - 1|\Big] \\ &\le \mathbb{E}\big[R(e^{QR} - 1)\big] = \delta_{\mathrm{fld}}(Q).\end{aligned}$$

For $0 \le q \le q_0$,

$$e^{qR} - 1 \le qR e^{q_0 R},$$

and therefore

$$\delta_{\mathrm{fld}}(q) \le q \mathbb{E}[R^2 e^{q_0 R}] = C_2 q.$$

Finally,

$$\dot{\widehat{\boldsymbol{w}}}_j = \Pi_{\widehat{\boldsymbol{w}}_j} V_j = \Pi_{\widehat{\boldsymbol{w}}_j} V_j^0 + \Pi_{\widehat{\boldsymbol{w}}_j}(V_j - V_j^0).$$

Thus one may take

$$e_j := \Pi_{\widehat{\boldsymbol{w}}_j}(V_j - V_j^0),$$

for which $\|e_j\| \le \delta_{\mathrm{fld}}(Q)$. The uniform estimate on $[0, T]$ follows from Theorem A.2. □

The same comparison controls radial and tangential motion simultaneously. Indeed, orthogonal decomposition of $V_j - V_j^0$ into its radial and tangential components gives

$$\left|\frac{\dot{r}_j}{r_j} - \langle \widehat{\boldsymbol{w}}_j, V_j^0 \rangle\right|^2 + \left\|\dot{\widehat{\boldsymbol{w}}}_j - \Pi_{\widehat{\boldsymbol{w}}_j} V_j^0\right\|^2 \le \delta_{\mathrm{fld}}(Q)^2.$$

Thus the ideal field approximates the full scaled neuron dynamics, not only its direction.

For the projected direction in the two-line model, however, the exact identity (28) gives

$$\dot{\widehat{a}}_j = \frac{r_j}{\varrho_j}\left[(\Pi_U - \widehat{a}_j\widehat{a}_j^\top)V_j^0 + e_{j,U}\right], \qquad \|e_{j,U}\| \le \delta_{\mathrm{fld}}(Q), \tag{37}$$

where

$$e_{j,U} := (\Pi_U - \widehat{a}_j\widehat{a}_j^\top)(V_j - V_j^0).$$

The factor r_j/ϱ_j therefore multiplies both the ideal drift and its error. Since this factor is neuron dependent and may become large when ϱ_j is small, the line analysis below works directly with the sign of the full bracket in (37); it does not replace the perturbation by a uniform additive error in projected angular velocity.

Corollary A.1 (Initialization scale for a prescribed early horizon)**.** *Fix $T < \infty$ and $\delta > 0$. For fixed-shape initialization, the condition*

$$\varepsilon^2 \sum_j \|\bar{\boldsymbol{w}}_j\|^2 \le e^{-2M_0 T} \min\left\{q_0, \frac{\delta}{C_2}\right\} \tag{38}$$

guarantees $Q(t) \le q_0$, and $\delta_{\mathrm{fld}}(Q(t)) \le \delta$ for almost every $t \in [0,T]$.

Proof. By Theorem A.2, $Q(t) \le Q(0)e^{2M_0 T}$ on $[0,T]$. Using $Q(0) = \varepsilon^2 \sum_j \|\bar{\boldsymbol{w}}_j\|^2$ and (36) gives both conclusions. □

We finally introduce a simple barrier argument that will be used repeatedly to turn inward-pointing differential inequalities into invariance statements.

Lemma 2. *Let g be absolutely continuous on an interval and suppose $g(a) \ge 0$. If, for some $\eta > 0$,$\dot{g}(t) \ge 0$ for almost every t such that $-\eta < g(t) < 0$, then g cannot cross from nonnegative to negative values.*

Likewise, let $d \ge 0$ be absolutely continuous with $d(a) = 0$. If $\dot{d}(t) \le 0$ for almost every t such that $0 < d(t) < \eta$, then d cannot depart from zero.

Both conclusions remain valid up to any stopping time on which the corresponding differential inequality holds.

Proof. For the first claim, suppose g becomes negative. A connected component (t_0, t_1) of the open set $\{g < 0\}$ then has $g(t_0) = 0$. By continuity, after possibly shortening the interval,

$$-\eta < g(t) < 0 \qquad (t_0 < t < t_1).$$

Absolute continuity and the assumed differential inequality give

$$g(t) - g(t_0) = \int_{t_0}^{t} \dot{g}(u)\, du \ge 0,$$

contradicting $g(t) < 0$.

The second claim is identical, applied to a connected component of $\{d > 0\}$ beginning at a point where $d = 0$. □

A.5 Planar dynamics, cone invariance, and score reduction

We next isolate a common reduced description of the two models. The only model-specific input is that the network preactivations are determined by their projection onto the two-dimensional plane U. For the two-line model this is automatic, since the data lie in U. For the two-hyperplane model it will hold after exact normal alignment, once strong invariance shows that the weights remain in U.

Definition A.1 (Planar reduction on an interval)**.** *A trajectory has the* planar reduction *on an interval I if, for every $t \in I$ and every neuron j,*

$$\langle \boldsymbol{w}_j(t), X\rangle = \langle a_j(t), X\rangle \qquad \textit{almost surely}, \qquad a_j(t) := \Pi_U \boldsymbol{w}_j(t). \tag{39}$$

On class c, the projected data have the form

$$\Pi_U X = Z_c \boldsymbol{u}_c.$$

Hence, under the planar reduction, the network output on class c depends on the sample only through the scalar coordinate Z_c.

For $c, \tau \in \{+1, -1\}$ define the signed scores

$$m_c^\tau(t) := c\, f(\tau \boldsymbol{u}_c; \xi(t)), \tag{40}$$

and the scalar loss and residual functions

$$\Lambda_c^\tau(q) := \mathbb{E}\Big[e^{-q|Z_c|}\mathbf{1}_{\{Y=c,\ \tau Z_c>0\}}\Big],$$
$$A_c^\tau(q) := \mathbb{E}\Big[|Z_c| e^{-q|Z_c|}\mathbf{1}_{\{Y=c,\ \tau Z_c>0\}}\Big]. \tag{41}$$

Under global integrability these quantities are finite for every $q \in \mathbb{R}$, and differentiation under the expectation gives

$$A_c^\tau(q) = -(\Lambda_c^\tau)'(q).$$

Finite-interval hypotheses. The arguments below are local in time and do not require global integrability. Fix a compact interval

$$I = [t_0, T]$$

on which the trajectory satisfies (15) almost everywhere, with

$$v_j = s_j r_j, \qquad r_j > 0.$$

Assume additionally that

$$B_I := \sup_{t\in I} Q(\xi(t)) < \infty, \qquad \mathbb{E}[R^2 e^{B_I R}] < \infty. \tag{42}$$

Under global integrability these conditions hold automatically on every finite interval by Theorem A.1 and Proposition 3. We state them explicitly because the reduction itself only needs this finite-interval control.

Proposition 6. *Suppose the finite-interval hypotheses hold and the trajectory has the planar reduction on I. Then, on $\{Y = c,\ \tau Z_c > 0\}$,*

$$Y f(X; \xi(t)) = |Z_c|\, m_c^\tau(t). \tag{43}$$

Consequently,

$$\mathcal{L}(t) = \sum_{c,\tau} \Lambda_c^\tau(m_c^\tau(t)), \qquad \mathcal{R}_c^\tau(t) = A_c^\tau(m_c^\tau(t)) > 0, \tag{44}$$

where $\mathcal{R}_c^\tau$ is the residual moment defined in the main text.

Whenever $\langle a_j, \boldsymbol{u}_+\rangle$ and $\langle a_j, \boldsymbol{u}_-\rangle$ are both nonzero, define

$$\eta_{j,c} := \operatorname{sign}\langle a_j, \boldsymbol{u}_c\rangle.$$

Then, almost everywhere at such times,

$$\dot a_j = s_j r_j \left(\eta_{j,+}\mathcal{R}_+^{\eta_{j,+}}\boldsymbol{u}_+ - \eta_{j,-}\mathcal{R}_-^{\eta_{j,-}}\boldsymbol{u}_-\right). \tag{45}$$

Proof. Fix a sample with $Y = c$ and $\tau Z_c > 0$. By the planar reduction, positive homogeneity of the ReLU network gives

$$f(X; \xi) = f(Z_c \boldsymbol{u}_c; \xi) = |Z_c| f(\tau \boldsymbol{u}_c; \xi).$$

Multiplying by $Y = c$ proves (43). The four events

$$\{Y = c,\ \tau Z_c > 0\}, \qquad c, \tau \in \{\pm 1\},$$

partition the data up to a null set. Summing their contributions gives the loss identity in (44), and inserting one factor of $|Z_c|$ gives the residual identity.

These quantities are finite at all realized scores. Indeed,

$$|m_c^\tau(t)| \le Q(\xi(t)) \le B_I, \qquad |Z_c| \le R,$$

so the relevant integrands are bounded by $e^{B_I R}$ and $Re^{B_I R}$. Both are integrable under (42). The same domination also gives continuity of the residual moments in time.

For the neuron velocity, project (15) onto U. If $\langle a_j, \boldsymbol{u}_c\rangle \neq 0$, then on class c the ReLU subgradient is fixed almost surely and equals $\mathbf{1}_{\{\eta_{j,c} Z_c > 0\}}$. Using $\Pi_U X = Z_c \boldsymbol{u}_c$ and separating the two classes gives (45). □

The four scores determine the loss and all residual moments, but their dynamics are not closed: the time derivatives of the scores still depend on the individual neuron configuration. The purpose of the reduction is instead to isolate the four scalar quantities through which the data distribution enters the subsequent geometric dynamics.

Assigned cones. For $c, \tau \in \{+1, -1\}$ recall

$$\nu_+^\tau := \tau \boldsymbol{n}_-, \qquad \nu_-^\tau := -\tau \boldsymbol{n}_+, \qquad e_c^\tau := \tau \boldsymbol{u}_{-c}, \qquad \mathcal{C}(\nu_c^\tau) := \mathrm{cone}\{\nu_c^\tau, \tau \boldsymbol{u}_c\}. \tag{46}$$

Thus $\mathcal{C}(\nu_c^\tau)$ is the closed cone between the target normal ν_c^τ and the corresponding oriented class direction $\tau \boldsymbol{u}_c$. Its unit-direction arc is

$$\mathcal{W}_c^\tau := \mathcal{C}(\nu_c^\tau) \cap \mathbb{S}^{d-1},$$

which has angular width $\pi/2 - \theta$.

The pair (ν_c^τ, e_c^τ) is orthonormal, and

$$\tau \boldsymbol{u}_c = \sin\theta\, \nu_c^\tau + \cos\theta\, e_c^\tau. \tag{47}$$

Hence, for $0 < \theta < \pi/2$, writing

$$a = h\nu_c^\tau + k e_c^\tau,$$

the condition $a \in \mathcal{C}(\nu_c^\tau) \setminus \{0\}$ is equivalent to

$$h > 0, \qquad k \geq 0, \qquad \ell := \cos\theta\, h - \sin\theta\, k \geq 0. \tag{48}$$

The two inequalities $k \geq 0$ and $\ell \geq 0$ describe respectively the target-normal boundary and the class-direction boundary of the cone. When $\theta = \pi/2$, the two rays coincide and the cone degenerates to the single oriented ray $h \geq 0$, $k = 0$.

Proposition 7 (Conditional cone invariance and target-coordinate growth)**.** *Under the hypotheses of Proposition 6, let $I = [t_0, T]$ and assign each neuron to a group G_c^τ such that*

$$s_j = c, \qquad a_j(t_0) \in \mathcal{C}(\nu_c^\tau) \setminus \{0\}.$$

Then, for every $t \in I$ and $j \in G_c^\tau$,

$$a_j(t) \in \mathcal{C}(\nu_c^\tau) \setminus \{0\}, \qquad \widehat{a}_j(t) \in \mathcal{W}_c^\tau. \tag{49}$$

Moreover, if

$$h_j(t) := \langle a_j(t), \nu_c^\tau \rangle,$$

then almost everywhere on I,

$$\dot{h}_j = \sin\theta\, r_j \mathcal{R}_c^\tau, \qquad \sin\theta\, \varrho_j \leq h_j \leq \varrho_j \leq r_j, \qquad h_j(t) \geq h_j(t_0) > 0. \tag{50}$$

Proof. Fix $j \in G_c^\tau$ and omit the neuron index. Write

$$a = h\nu_c^\tau + k e_c^\tau.$$

We first compute the target-normal coordinate. Throughout a neighborhood of the assigned cone away from 0, the neuron activates on the τ-oriented half of class c. Since $\nu_c^\tau \perp S_{-c}$, the opposite-class contribution has zero projection onto ν_c^τ. The own-class contribution therefore gives

$$\dot{h} = \sin\theta\, r\, \mathcal{R}_c^\tau > 0. \tag{51}$$

Consequently, up to any hypothetical first exit from the assigned cone,

$$h(t) \geq h(t_0) > 0.$$

By continuity, the same lower bound holds at the exit time. Thus a first exit cannot occur through 0. We next exclude exits through either nonzero boundary ray using the local barrier inequalities.

Assume first that $0 < \theta < \pi/2$. Inside the cone, where $k > 0$ and $\ell > 0$, both class derivatives have orientation τ. Projecting (45) onto (ν_c^τ, e_c^τ) gives

$$\dot{h} = \sin\theta\, r \mathcal{R}_c^\tau, \qquad \dot{k} = r\left(\cos\theta\, \mathcal{R}_c^\tau - \mathcal{R}_{-c}^\tau\right). \tag{52}$$

We now check the two cone boundaries separately.

At the target-normal boundary $k = 0$, consider the adjacent forbidden side $k < 0$. There the own-class derivative still has orientation τ, whereas the opposite-class derivative switches to $-\tau$. Hence

$$\dot{k} = r\left(\cos\theta\,\mathcal{R}_c^{\tau} + \mathcal{R}_{-c}^{-\tau}\right) > 0.$$

Thus the vector field points back toward $k \geq 0$ whenever the trajectory tries to enter the bad side. By Lemma 2, the boundary $k = 0$ cannot be crossed.

For the class-direction boundary, use the barrier coordinate

$$\ell := \cos\theta\, h - \sin\theta\, k.$$

The condition $\ell = 0$ is precisely the ray $\mathbb{R}_+ \tau \boldsymbol{u}_c$. On a neighborhood of this boundary the two relevant derivative orientations remain τ, and (52) gives

$$\dot{\ell} = \sin\theta\, r\mathcal{R}_{-c}^{\tau} > 0.$$

Hence ℓ cannot cross from nonnegative to negative values.

Any first exit from the cone would therefore have to occur either through $k = 0$, through $\ell = 0$, or through 0. The first two are excluded by the barrier argument, and 0 is excluded by the strict growth of h. This proves invariance.

When $\theta = \pi/2$, the cone reduces to the target ray. In the coordinate k, the two sides satisfy

$$\dot{k} = -r\mathcal{R}_{-c}^{\tau} < 0 \quad \text{for } k > 0, \qquad \dot{k} = r\mathcal{R}_{-c}^{-\tau} > 0 \quad \text{for } k < 0.$$

Thus neither side can be entered from $k = 0$. Together with $\dot{h} > 0$, this keeps the neuron on the same oriented ray.

Invariance now ensures that the own-class derivative retains orientation τ throughout I. Hence the target-coordinate identity (51) holds almost everywhere on I, and h is strictly increasing there. In particular,

$$h(t) \geq h(t_0) > 0 \qquad \text{for every } t \in I.$$

Next, we establish the norm comparisons. Since every $\widehat{a} \in \mathcal{W}_c^{\tau}$ makes angle at most $\pi/2 - \theta$ with ν_c^{τ},

$$h = \langle a, \nu_c^{\tau} \rangle \geq \sin\theta\, \|a\| = \sin\theta\, \varrho.$$

The inequalities $h \leq \varrho \leq r$ follow from Cauchy–Schwarz and orthogonal projection. This proves (50). □

For the two-line model, Proposition 7 confines the *projected* neuron directions; the components orthogonal to U may remain nonzero and are frozen by Proposition 4. For the two-hyperplane model, once strong invariance establishes $\boldsymbol{w}_j \in U$, the same proposition confines the full neuron directions. This invariance is a separate ingredient and cannot be inferred from the reduced planar field alone.

Corollary A.2. *Under Proposition 7, suppose a neuron in G_c^{τ} satisfies $\widehat{a}_j = \nu_c^{\tau}$, and define*

$$d_c^{\tau} := \cos\theta\, \mathcal{R}_c^{\tau} - \mathcal{R}_{-c}^{\tau}. \tag{53}$$

If $d_c^{\tau} < 0$ at that time, then the neuron remains at its projected target on a subsequent time neighborhood within I. If $d_c^{\tau} > 0$, no continuation can remain at that target throughout an interval on which the inequality persists.

Proof. At the target normal, $k = 0$. By Proposition 6, the residual moments are continuous on I.

Suppose first that $d_c^{\tau} < 0$. By continuity, the same strict inequality holds on a short subsequent interval. For $k > 0$ sufficiently close to zero, the interior equation (52) gives

$$\dot{k} = r\, d_c^{\tau} < 0.$$

Cone invariance already gives $k \geq 0$. The barrier principle therefore prevents k from departing from zero, so the projected direction remains at ν_c^{τ}. Now suppose $d_c^{\tau} > 0$. At $k = 0$, the own-class

contribution to the e_c^τ component is fixed, while the opposite-class activation is set-valued. The possible transverse velocities satisfy

$$\frac{\dot{k}}{r_j} \in \left[\cos\theta\, \mathcal{R}_c^\tau - \mathcal{R}_{-c}^\tau,\, \cos\theta\, \mathcal{R}_c^\tau + \mathcal{R}_{-c}^{-\tau}\right].$$

If $d_c^\tau > 0$, the left endpoint is strictly positive. Hence every admissible velocity has $\dot{k} > 0$. An absolutely continuous trajectory that remained at $k = 0$ on a nontrivial interval would instead satisfy $\dot{k} = 0$ almost everywhere on that interval, a contradiction. □

Theorem A.4 (Vanishing loss and divergent signed scores)**.** *Assume global integrability and nonzero balanced initialization. Suppose a global training trajectory has the planar reduction for all $t \geq t_0$, and that at time t_0 every neuron lies in its assigned nonzero cone as in Proposition 7. If each group G_c^τ is nonempty, then*

$$\mathcal{L}(t) \longrightarrow 0, \qquad m_c^\tau(t) \longrightarrow +\infty \quad \textit{for every } c, \tau. \tag{54}$$

More precisely, define

$$B_c^\tau := \sum_{j \in G_c^\tau} h_j(t_0)^2 > 0.$$

Then, for almost every $t \geq t_0$,

$$-\dot{\mathcal{L}}(t) \geq \sin^2\theta \sum_{c,\tau} B_c^\tau \left[A_c^\tau(m_c^\tau(t))\right]^2. \tag{55}$$

Proof. Apply Proposition 7 on every finite interval $[t_0, T]$. For $j \in G_c^\tau$,

$$\dot{h}_j = \sin\theta\, r_j \mathcal{R}_c^\tau, \qquad r_j(t) \geq h_j(t) \geq h_j(t_0) > 0.$$

The energy identity therefore gives

$$\begin{aligned}
-\dot{\mathcal{L}}(t) &= \|\dot{\xi}(t)\|^2 \\
&\geq \sum_j \|\dot{\boldsymbol{w}}_j(t)\|^2 \\
&\geq \sum_j \dot{h}_j(t)^2 \\
&= \sin^2\theta \sum_{c,\tau} (\mathcal{R}_c^\tau(t))^2 \sum_{j \in G_c^\tau} r_j(t)^2 \\
&\geq \sin^2\theta \sum_{c,\tau} B_c^\tau (\mathcal{R}_c^\tau(t))^2.
\end{aligned}$$

Using $\mathcal{R}_c^\tau(t) = A_c^\tau(m_c^\tau(t))$ proves (55).

Since $\mathcal{L}$ is nonnegative and nonincreasing, it has a limit

$$L_\infty := \lim_{t \to \infty} \mathcal{L}(t) \geq 0.$$

We show that $L_\infty = 0$.

For each (c, τ), $\Lambda_c^\tau(q) \longrightarrow 0$ as $q \to +\infty$ by dominated convergence, since $|Z_c| > 0$ almost surely on the corresponding side. Suppose, for contradiction, that $L_\infty > 0$. Because there are only four functions Λ_c^τ, we may choose $M < \infty$ such that

$$\Lambda_c^\tau(M) < \frac{L_\infty}{4} \qquad \text{for every } c, \tau. \tag{56}$$

At every time $t \geq t_0$,

$$\sum_{c,\tau} \Lambda_c^\tau(m_c^\tau(t)) = \mathcal{L}(t) \geq L_\infty.$$

Hence at least one pair (c, τ) satisfies

$$\Lambda_c^\tau(m_c^\tau(t)) \geq \frac{L_\infty}{4}.$$

By (56) and monotonicity of Λ_c^τ, the corresponding score must satisfy $m_c^\tau(t) < M$. Since

$$A_c^\tau(q) = -(\Lambda_c^\tau)'(q) > 0$$

and A_c^τ is nonincreasing in q, we then have $A_c^\tau(m_c^\tau(t)) \geq A_c^\tau(M)$. Therefore (55) implies the uniform bound

$$-\dot{\mathcal{L}}(t) \geq \sin^2\theta \min_{c,\tau} B_c^\tau \min_{c,\tau}[A_c^\tau(M)]^2 =: \kappa > 0$$

for almost every $t \geq t_0$. Integrating gives

$$\mathcal{L}(t) \leq \mathcal{L}(t_0) - \kappa(t - t_0),$$

which becomes negative for sufficiently large t, a contradiction. Thus $L_\infty = 0$.

Finally, the four-score representation writes the loss as a sum of four nonnegative terms:

$$\mathcal{L}(t) = \sum_{c,\tau} \Lambda_c^\tau(m_c^\tau(t)).$$

Hence $\Lambda_c^\tau(m_c^\tau(t)) \longrightarrow 0$ for every c, τ. If some score $m_c^\tau(t)$ failed to converge to $+\infty$, there would exist a finite M and a sequence $t_k \to \infty$ such that $m_c^\tau(t_k) \leq M$. Monotonicity of Λ_c^τ would then give $\Lambda_c^\tau(m_c^\tau(t_k)) \geq \Lambda_c^\tau(M) > 0$, contradicting convergence of that loss term to zero. Therefore $m_c^\tau(t) \to +\infty$ for all four pairs (c, τ). □

Theorem A.4 separates *fitting* from the more delicate problem of identifying the eventual neuron directions. Once the planar reduction and four populated cones are available, the loss converges to zero without any further balance or tail assumption. The theorem does not, however, determine a universal convergence rate or show that individual neurons return to their target normals.

For finite two-line data, the positive minimum sample radius strengthens (55) to a quadratic differential inequality for the loss; Section B.3 then gives an explicit $O(1/t)$ rate and proves divergence of the individual projected norms. For the hyperplane model, arbitrarily small projected coordinates occur, so the late rate depends on the behavior of their distribution near zero. Sections C.6–C.8 analyze this regime and establish normal realignment.

B Proofs specific to the two-line model

Throughout this section, the data distribution is

$$\mathcal{D} = \sum_{i=1}^{n} p_i \delta_{(\boldsymbol{x}_i, y_i)}, \qquad p_i > 0, \quad \sum_i p_i = 1, \qquad \boldsymbol{x}_i = t_i \boldsymbol{u}_{y_i}, \quad t_i \neq 0.$$

Every signed ray has positive mass. Write $d_i = |t_i|$ and $I_c^\tau = \{i : y_i = c,\ \tau t_i > 0\}$, where $c, \tau \in \{\pm 1\}$. The angle between $\boldsymbol{u}_+$ and $\boldsymbol{u}_-$ is $\theta \in (0, \pi/2]$, and $U = \operatorname{span}\{\boldsymbol{u}_+, \boldsymbol{u}_-\}$. We use

$$\begin{gathered} c_\theta = \cos\theta, \qquad M_c^\tau = \sum_{i \in I_c^\tau} p_i d_i, \qquad P_\tau = M_+^\tau, \quad N_\tau = M_-^\tau, \\ \underline{M} = \min_{c,\tau} M_c^\tau, \qquad X_1 = \sum_i p_i d_i, \qquad X_{\max} = \max_i d_i, \qquad d_{\min} = \min_i d_i > 0. \end{gathered} \tag{57}$$

The strict balance condition can be written as

$$\Delta = \min_{\tau \in \{\pm 1\}} \{P_\tau - c_\theta N_\tau,\ N_\tau - c_\theta P_\tau\} > 0. \tag{58}$$

B.1 Exact projected dynamics and the idealized field

Propositions 3 and 4 give, for every neuron,

$$\begin{aligned} \boldsymbol{w}_j = a_j + b_j, \quad a_j = \Pi_U \boldsymbol{w}_j, \quad b_j = \Pi_{U^\perp} \boldsymbol{w}_j(0), \quad c_j = \|b_j\|, \\ \varrho_j = \|a_j\|, \quad r_j = \|\boldsymbol{w}_j\| = \sqrt{\varrho_j^2 + c_j^2}, \quad v_j = s_j r_j. \end{aligned} \tag{59}$$

The network sees a_j but not b_j: every data point lies in U. Nevertheless, b_j still affects the speed through $|v_j| = r_j = \sqrt{\varrho_j^2 + c_j^2}$. In particular, $r_j^2 - \varrho_j^2 = c_j^2$ need not vanish, so one cannot replace the full-vector balance relation by $|v_j| = \varrho_j$. For $a_j \neq 0$, put $\widehat{a}_j = a_j/\varrho_j$ and $V_j = \dot{\boldsymbol{w}}_j/r_j \in U$. Then

$$\dot{a}_j = r_j V_j, \qquad \dot{\widehat{a}}_j = \frac{r_j}{\varrho_j}(I_U - \widehat{a}_j \widehat{a}_j^\top) V_j.$$

The multiplier $r_j/\varrho_j \geq 1$ changes the speed of projected angular motion, but not its sign. This distinction lets us use the planar geometry even for weights initialized outside U.

Lemma 3. *Fix a parameter value with $r_j > 0$. Let $\mathcal{A}(a_j)$ be the finite collection of active sample sets realized by open activation regions whose closures contain a_j. The projection of the Clarke velocity set onto its jth first-layer block, divided by r_j, is*

$$\mathcal{V}_j^{\mathrm{ex}} = \mathrm{co}\left\{ s_j \sum_{i \in A} p_i y_i e^{-y_i f(\boldsymbol{x}_i)} \boldsymbol{x}_i : A \in \mathcal{A}(a_j) \right\}. \tag{60}$$

If $a_j \neq 0$ and $\max_i |f(\boldsymbol{x}_i)| \leq \rho$, then

$$d_{\mathrm{H}}\big(\mathcal{V}_j^{\mathrm{ex}}, \mathcal{G}_{s_j}^0(\widehat{a}_j)\big) \leq \eta_\rho, \qquad \eta_\rho = (e^\rho - 1) X_1, \tag{61}$$

where $\mathcal{G}_s^0$ is the ideal field from (32).

Proof. **Block projection of the joint Clarke set.** On every open joint activation region the loss is smooth, and its negative $\boldsymbol{w}_j$-gradient is $v_j \sum_{i \in A} p_i y_i e^{-y_i f(\boldsymbol{x}_i)} \boldsymbol{x}_i$. The Clarke set is the convex hull of the limiting joint gradients. Any realizable region for neuron j can be combined with independently chosen adjacent regions for the other neurons. Projection of this convex hull onto the jth block therefore gives exactly (60).

Comparison with the ideal field. The ideal set uses the same active sets with the exponential factors replaced by one. Corresponding region vectors differ by at most $\sum_i p_i d_i |e^{-y_i f(\boldsymbol{x}_i)} - 1| \leq \eta_\rho$. Matching convex coefficients in either direction proves the Hausdorff bound. Subgradients for samples on the same ray remain coupled throughout. □

We now turn this vector description into a scalar angular equation. Choose planar coordinates with $\alpha = \theta/2$ and

$$\boldsymbol{u}_+ = (\cos\alpha, \sin\alpha), \quad \boldsymbol{u}_- = (\cos\alpha, -\sin\alpha), \quad \boldsymbol{n}_+ = (-\sin\alpha, \cos\alpha), \quad \boldsymbol{n}_- = (\sin\alpha, \cos\alpha).$$

Let $\widehat{a} = (\cos\phi, \sin\phi)$ and $e_\phi = (-\sin\phi, \cos\phi)$. The switching angles are

$$\phi_1 = \frac{\pi}{2} - \alpha, \quad \phi_2 = \frac{\pi}{2} + \alpha, \quad \phi_3 = \frac{3\pi}{2} - \alpha, \quad \phi_4 = \frac{3\pi}{2} + \alpha, \quad \phi_5 = \phi_1 + 2\pi.$$

On $J_k = (\phi_k, \phi_{k+1})$, denote the positive-sign ideal field by G_k and its angular component by $g_+(\phi) = \langle e_\phi, G_k \rangle$. The negative-sign field and angular component are their negatives. For example, on J_4 both class projections are positive. The active positive-class ray contributes $P_+ \boldsymbol{u}_+$ and the active negative-class ray contributes $-N_+ \boldsymbol{u}_-$. Thus $G_4 = P_+ \boldsymbol{u}_+ - N_+ \boldsymbol{u}_-$. Taking its inner product with e_ϕ gives the angular force. At the left endpoint $\phi = \phi_4$ this is $P_+ - c_\theta N_+$, and at the right endpoint $\phi = \phi_5$ it is $N_+ - c_\theta P_+$. The other rows follow by the same subgradient and label calculation. The endpoint columns in Table 2 are one-sided limits from within the indicated region.

Lemma 4. *Under (58), $g_+ < 0$ on J_1, J_3, $g_+ > 0$ on J_2, J_4, and $|g_+| \geq \Delta$ on every closed one-sided region. For sign s, the attracting switching normals are $\mathcal{N}_s = \{\pm \boldsymbol{n}_{-s}\}$ and the repelling switching normals are $\mathcal{E}_s = \{\pm \boldsymbol{n}_s\}$.*

region	G_k	Left angular trace	Right angular trace
J_1	$P_+\boldsymbol{u}_+ + N_-\boldsymbol{u}_-$	$-(c_\theta P_+ + N_-)$	$-(P_+ + c_\theta N_-)$
J_2	$-P_-\boldsymbol{u}_+ + N_-\boldsymbol{u}_-$	$P_- - c_\theta N_-$	$N_- - c_\theta P_-$
J_3	$-P_-\boldsymbol{u}_+ - N_+\boldsymbol{u}_-$	$-(c_\theta P_- + N_+)$	$-(P_- + c_\theta N_+)$
J_4	$P_+\boldsymbol{u}_+ - N_+\boldsymbol{u}_-$	$P_+ - c_\theta N_+$	$N_+ - c_\theta P_+$

Table 2: Ideal region fields and angular traces for a positive neuron.

Proof. The table gives the stated signs and a magnitude at least Δ at each endpoint. For J_2, J_4, whose traces are differences of moments, this follows directly from the definition of Δ. For J_1, J_3, whose trace magnitudes are sums, positivity supplies the same lower bound. For example, $\Delta \le N_- - c_\theta P_- \le N_-$ and $\Delta \le P_+ - c_\theta N_+ \le P_+$ control the two endpoints of J_1. The bounds on J_3 follow in the same way.

Each angular component is a sinusoid on an interval of length strictly less than π. Equal nonzero endpoint signs preclude an interior zero: two zeros of a nonzero sinusoid are separated by π. On an interval where its sign is fixed, its absolute value is positive and has second derivative equal to its negative. It is therefore concave, and its minimum is attained at an endpoint. The directions of the two adjacent traces identify the attracting and repelling normals. The same calculation includes $\theta = \pi/2$. □

For a positive neuron, the angular force on J_4 points toward $\phi_5 = \phi_1 + 2\pi$, while the force on J_1 points back toward ϕ_1. These two regions therefore meet at the attracting normal $\boldsymbol{n}_-$. At ϕ_2, the two neighboring forces point away from the boundary, so $\boldsymbol{n}_+$ is repelling. The antipodal pair has the same classification. Reversing the output sign reverses every angular force and exchanges the attracting and repelling pairs.

Every direction outside $\mathcal{E}_s$ thus has a uniquely assigned attracting endpoint, denoted $\nu_s(\widehat{a}) \in \mathcal{N}_s$. A direction already at an attracting endpoint is assigned to itself. The exclusion of $\mathcal{E}_s$ is needed: the ideal completed field there admits a zero angular velocity.

B.2 Finite time alignment

For $0 \le \eta < \Delta$, define

$$T_{\mathrm{normal}}(\eta) = \max_{1\le k\le 4} \int_{\phi_k}^{\phi_{k+1}} \frac{d\phi}{|g_+(\phi)| - \eta} \le \frac{\pi - \theta}{\Delta - \eta}. \tag{62}$$

It applies to both neuron signs. The denominator is the ideal angular speed after allowing an error of size η in the least favorable direction. Integrating its reciprocal over an entire region bounds the travel time from any starting point in that region. The widest region has length $\pi - \theta$, which gives the simpler upper bound in (62).

We first prove alignment assuming that the network outputs stay small. We then choose the initialization scale so that this assumption lasts long enough for every neuron to be captured.

Proposition 8. *Suppose (58) holds and $a_j(0) \neq 0$, $\widehat{a}_j(0) \notin \mathcal{E}_{s_j}$ for every j. Assume $\max_i |f_{\xi(t)}(\boldsymbol{x}_i)| \le \rho$ on $[0,T]$, where $\eta_\rho < \Delta$ and $T \ge T_{\mathrm{normal}}(\eta_\rho)$. Then every $a_j(t)$ is nonzero on $[0,T]$. Each projected direction reaches $\nu_j = \nu_{s_j}(\widehat{a}_j(0))$ at a time $t_j \le T_{\mathrm{normal}}(\eta_\rho)$ and remains there until T.*

Proof. **Motion toward the assigned normal.** Before a switching boundary is reached, Lemma 3 gives

$$\dot{\phi}_j = \frac{r_j}{\varrho_j}\big(g_{s_j}(\phi_j) + e_j\big), \qquad |e_j| \le \eta_\rho.$$

Its sign is fixed and points toward the assigned attracting endpoint. Indeed, the ideal angular magnitude is at least Δ, whereas the perturbation has magnitude at most $\eta_\rho < \Delta$. Thus the projected direction cannot turn around inside a region.

Excluding a vanishing projection. An angular equation is meaningful only while $\varrho_j > 0$. To ensure that the direction reaches a boundary before the projection could disappear, compare radial change with angular change. The force bound $\|V_j\| \le e^{\rho} X_1$ and the angular margin give on any such interval

$$\left| \frac{d \log \varrho_j}{d\phi_j} \right| = \left| \frac{\langle \widehat{a}_j, V_j \rangle}{\langle e_{\phi_j}, V_j \rangle} \right| \le \frac{e^{\rho} X_1}{\Delta - \eta_\rho}.$$

The common factor r_j/ϱ_j cancels in this ratio. Integrating along the monotone angular path bounds the decrease in $\log \varrho_j$ by $e^{\rho} X_1 (\pi - \theta)/(\Delta - \eta_\rho)$. In particular, ϱ_j stays bounded away from zero before the first boundary hit. Since $r_j/\varrho_j \ge 1$, the angular motion is at least as fast as the comparison motion defining (62). The hit therefore occurs by that time.

Normal invariance after alignment. At the attracting endpoint both adjacent angular traces point strictly inward. Lemma 2 precludes departure on either side for every continuation. On this boundary, $h_j = \langle a_j, \nu_j \rangle$ obeys $\dot{h}_j = \sin\theta r_j \mathcal{R}_{s_j}^{\tau_j} > 0$, where $\nu_j = \nu_{s_j}^{\tau_j}$ in the notation of (46). Indeed, projecting (15) onto ν_j removes the opposite-class term and fixes the own-class subgradient to $\{\tau_j Z_{s_j} > 0\}$. This calculation requires no alignment assumption on the other neurons. Thus the projection remains nonzero. The same argument covers a neuron initialized at its attracting endpoint. □

Lemma 5 (Growth on a translated normal ray)**.** *Suppose $\widehat{a}_j = \nu_c^{\tau}$ on $[t_a, t_b]$ and $s_j = c$. Then $\boldsymbol{w}_j = b_j + h_j \nu_c^{\tau}$, $h_j > 0$, and*

$$\dot{h}_j = \sin\theta \mathcal{R}_c^{\tau}(t) \sqrt{h_j^2 + c_j^2}. \tag{63}$$

If outputs have magnitude at most ρ on this interval, set $\lambda_{\mathrm{amb}} = \sin\theta e^{-\rho} \underline{M} > 0$. Then

$$\sin \angle(\boldsymbol{w}_j(t), \nu_c^{\tau}) \le \operatorname{sech}\big(\lambda_{\mathrm{amb}}(t - t_a)\big) \le 2e^{-\lambda_{\mathrm{amb}}(t - t_a)}. \tag{64}$$

If $c_j = 0$, the angle is identically zero.

Proof. Since $a_j = h_j \nu_c^{\tau}$ with $h_j > 0$, the own-class subgradient is $\mathbf{1}_{\{\tau Z_c > 0\}}$. Projecting the selector equation (15) onto ν_c^{τ} eliminates the opposite-class contribution because $\nu_c^{\tau} \perp S_{-c}$. Using $v_j = c r_j$ and $\langle \nu_c^{\tau}, Z_c \boldsymbol{u}_c \rangle = \sin\theta \tau Z_c$ gives

$$\dot{h}_j = \sin\theta r_j \mathbb{E}\Big[|Z_c| e^{-Y f(X;\xi(t))} \mathbf{1}_{\{Y = c,\, \tau Z_c > 0\}} \Big] = \sin\theta r_j \mathcal{R}_c^{\tau}.$$

This identity holds regardless of the other neurons' directions. Substitute $r_j = \sqrt{h_j^2 + c_j^2}$ to obtain (63). The output bound implies $e^{-\rho} M_c^{\tau} \le \mathcal{R}_c^{\tau} \le e^{\rho} M_c^{\tau}$. For $c_j > 0$,

$$\frac{d}{dt} \operatorname{arsinh}(h_j/c_j) = \sin\theta \mathcal{R}_c^{\tau} \ge \lambda_{\mathrm{amb}}.$$

Since $h_j(t_a) > 0$ and $c_j/r_j = \operatorname{sech}(\operatorname{arsinh}(h_j/c_j))$, we obtain

$$\frac{c_j}{r_j(t)} \le \operatorname{sech}\big(\lambda_{\mathrm{amb}}(t - t_a)\big).$$

The left-hand side is the sine of the angle between $b_j + h_j \nu_c^{\tau}$ and ν_c^{τ}, because $b_j \perp \nu_c^{\tau}$. The bound $\operatorname{sech} u \le 2e^{-u}$ for $u \ge 0$ completes the proof. □

Thus the orthogonal component need not decrease: it remains exactly equal to b_j. Alignment occurs because the component along the normal grows and makes b_j negligible relative to the full norm. In the ideal dynamics the residual moments are constant, so this growth can be integrated explicitly.

Corollary B.1 (The ideal full-vector dynamics)**.** *For the ideal one-neuron inclusion $\dot{\boldsymbol{w}} \in \|\boldsymbol{w}\| \mathcal{G}_s^0(\widehat{a})$, where $\widehat{a} = \Pi_U \boldsymbol{w} / \|\Pi_U \boldsymbol{w}\|$, assume $\Pi_U \boldsymbol{w}(0) \ne 0$, a nonrepelling initial projection, and (58). There is a unique global trajectory. Its projection reaches its assigned normal by $T_{\mathrm{normal}}(0)$ and stays there. Denote its alignment time by t_{hit} and its assigned normal by ν_c^{τ}. Writing $\boldsymbol{w} = b + h\nu_c^{\tau}$ and $c_\perp = \|b\|$, the scalar solution is*

$$h(t) = \begin{cases} c_\perp \sinh\big(\operatorname{arsinh}(h(t_{\mathrm{hit}})/c_\perp) + \sin\theta M_c^{\tau}(t - t_{\mathrm{hit}})\big), & c_\perp > 0, \\ h(t_{\mathrm{hit}}) \exp\big(\sin\theta M_c^{\tau}(t - t_{\mathrm{hit}})\big), & c_\perp = 0. \end{cases}$$

Proof. The alignment proof applies with zero error and without an output horizon. Within a region, the vector field $\|\boldsymbol{w}\|G$ is locally Lipschitz away from zero and has at most linear growth. At an attracting boundary, the no-departure argument forces the projected direction to stay there; its remaining scalar equation is unique. These observations give uniqueness and global continuation, and integration of (63) with constant residual moment gives the formula. A nonzero b prevents full vector alignment at any finite time. □

We next return to the exact, coupled network. Its residual weights are close to one only while the outputs remain small. The following bootstrap estimates how long this lasts. It uses the total initial squared norm S_0, so it accounts for the effect of all neurons on the common outputs.

Lemma 6 (Finite-data output bootstrap)**.** *Write $\boldsymbol{w}_j(0) = \varepsilon\bar{\boldsymbol{w}}_j$ and $v_j(0) = s_j\varepsilon\|\bar{\boldsymbol{w}}_j\|$, and set $S_0 = \sum_j \|\bar{\boldsymbol{w}}_j\|^2 > 0$. If $\varepsilon^2 X_{\max} S_0 < \rho$, then*

$$T_{\text{out}} = \frac{1}{2e^{\rho}X_1}\log\frac{\rho}{\varepsilon^2 X_{\max}S_0} \tag{65}$$

is a controlled horizon: $\max_i |f_{\xi(t)}(\boldsymbol{x}_i)| \le \rho$ for $0 \le t \le T_{\text{out}}$. More precisely, $r_j(t) \le r_j(0)e^{e^{\rho}X_1 t}$ and $\max_i |f_{\xi(t)}(\boldsymbol{x}_i)| \le \varepsilon^2 X_{\max}S_0 e^{2e^{\rho}X_1 t}$ there.

Proof. As long as outputs are bounded by ρ, $\|\dot{\boldsymbol{w}}_j\| \le e^{\rho}X_1 r_j$, hence $r_j(t) \le r_j(0)e^{e^{\rho}X_1 t}$. Balancedness and the ReLU bound give $|f(\boldsymbol{x}_i)| \le d_i \sum_j r_j^2$. Therefore, throughout any such interval,

$$\max_i |f(\boldsymbol{x}_i)| \le X_{\max}\sum_j r_j(0)^2 e^{2e^{\rho}X_1 t} = \varepsilon^2 X_{\max}S_0 e^{2e^{\rho}X_1 t}.$$

The right-hand side equals ρ at T_{out} and is strictly smaller before that time. A first time at which an output reached magnitude ρ before T_{out} would contradict this estimate. The bound thus holds through T_{out}, with its endpoint included by continuity. Global continuation follows from Theorem A.1, whose moment assumption is automatic for finite data. □

There are now two time scales to compare. The alignment clock depends on the region geometry and the allowed error; the output horizon grows as $\log(1/\varepsilon)$. Making the initialization small enough places the entire alignment process inside the controlled output interval.

Theorem B.1 (Quantitative early alignment for two lines)**.** *Assume strict balance, balanced initialization, and nonzero, nonrepelling initial projections. Choose $\rho > 0$ with $\eta_\rho < \Delta$. If*

$$\varepsilon^2 X_{\max}S_0 \exp\bigl(2e^{\rho}X_1 T_{\text{normal}}(\eta_\rho)\bigr) < \rho, \tag{66}$$

then every projected direction reaches its assigned normal by $T_{\text{normal}}(\eta_\rho)$ and stays there through T_{out}. For $\delta \in (0,1)$, let

$$T_\delta = T_{\text{normal}}(\eta_\rho) + (\sin\theta e^{-\rho}\underline{M})^{-1}\log(2/\delta).$$

Replacing $T_{\text{normal}}(\eta_\rho)$ by T_δ in (66) ensures $\sin\angle(\boldsymbol{w}_j(t), \nu_j) \le \delta$ for every j and $t \in [T_\delta, T_{\text{out}}]$.

Proof. The smallness inequality makes the bootstrap horizon longer than the required alignment clock, so Proposition 8 applies. In detail, taking logarithms in (66) gives $T_{\text{out}} > T_{\text{normal}}(\eta_\rho)$. For the full-vector assertion, each neuron has spent at least $t - T_{\text{normal}}(\eta_\rho)$ on its translated normal ray at time t. Lemma 5 therefore gives

$$\sin\angle(\boldsymbol{w}_j(t), \nu_j) \le 2\exp\bigl[-\sin\theta e^{-\rho}\underline{M}\bigl(t - T_{\text{normal}}(\eta_\rho)\bigr)\bigr].$$

The definition of T_δ makes this at most δ. The strengthened smallness condition ensures $T_{\text{out}} > T_\delta$, so the interval on which this conclusion is used is nonempty. □

Corollary B.2 (A sharper clock for equal active moments)**.** *If $M_c^\tau = M$ for all four rays and $\eta \le M\sin^2\alpha = \Delta/2$, then*

$$T_{\text{normal}}(\eta) \le \max\left\{\frac{1}{M\sin\alpha}\log\frac{1+\cos\alpha}{1-\cos\alpha},\ \frac{1}{M\cos\alpha}\log\frac{1+\sin\alpha}{1-\sin\alpha}\right\}. \tag{67}$$

Consequently this clock is $O(\log(1/\theta)/(M\theta))$ as $\theta \downarrow 0$.

Proof. On the wide regions the angular magnitude is $2M \sin\alpha\,|\cos\phi|$, with $|\cos\phi| \geq \sin\alpha$. On the narrow regions it is $2M\cos\alpha\,|\sin\phi|$, with $|\sin\phi| \geq \cos\alpha$. In both cases $|g| - \eta \geq |g|/2$. For a wide region, shift the angle by π to integrate over $[-\pi/2+\alpha, \pi/2-\alpha]$. Its travel time is at most

$$\frac{1}{M\sin\alpha}\int_{-\pi/2+\alpha}^{\pi/2-\alpha}\frac{du}{\cos u} = \frac{1}{M\sin\alpha}\log\frac{1+\cos\alpha}{1-\cos\alpha}.$$

For a narrow region, the analogous integral is

$$\frac{1}{M\cos\alpha}\int_{\pi/2-\alpha}^{\pi/2+\alpha}\frac{du}{\sin u} = \frac{1}{M\cos\alpha}\log\frac{1+\sin\alpha}{1-\sin\alpha}.$$

Taking the larger bound proves (67). As $\alpha = \theta/2 \downarrow 0$, the first bound has order $\log(1/\theta)/(M\theta)$, while the second is smaller. □

For example, $\rho = \log(1+\Delta/(2X_1))$ gives $\eta_\rho = \Delta/2$, and an explicit sufficient initialization scale is

$$\varepsilon < \sqrt{\frac{\rho}{X_{\max}S_0}}\, e^{-e^{\rho}X_1 T_{\mathrm{normal}}(\eta_\rho)}.$$

For unit initialization shapes, S_0 equals the width. The sharper equal-moment clock is not a uniform estimate over arbitrarily near-critical moment ratios. All preceding results include $\theta = \pi/2$ directly. Finally, excluding repelling directions and zero projections is a genericity condition; it does not ensure that all four attracting targets receive a neuron.

B.3 Fitting rate and divergence of every projected neuron

From now on, moment balance and small initialization are unnecessary. The hypothesis that replaces them is geometric: every neuron must already lie in an assigned cone, and each of the four groups must contain at least one neuron. Early alignment is one way to reach such a configuration, but the following arguments apply to any trajectory that reaches it. Assume that at some finite time t_0 the neurons admit a partition into four nonempty groups $\mathcal{J}_c^\tau$ such that

$$s_j = c, \qquad 0 \neq a_j(t_0) \in \mathcal{C}(\nu_c^\tau) \quad (j \in \mathcal{J}_c^\tau), \qquad \mathcal{J}_c^\tau \neq \varnothing \quad (c,\tau \in \{\pm 1\}). \tag{68}$$

The cones and their orthonormal bases are defined in (46)–(47). For line data the planar reduction is automatic: $\langle \boldsymbol{w}_j, \boldsymbol{x}_i\rangle = \langle a_j, \boldsymbol{x}_i\rangle$. Proposition 7 therefore applies for all $t \geq t_0$. In particular, with $\nu = \nu_c^\tau$, $e = e_c^\tau$, and $a_j = h_j\nu + k_j e$ in group $\mathcal{J}_c^\tau$,

$$\dot{h}_j = \sin\theta\, r_j \mathcal{R}_c^\tau, \qquad h_j(t) \geq h_j(t_0) > 0, \qquad \sin\theta\,\varrho_j \leq h_j \leq \varrho_j \leq r_j. \tag{69}$$

The four ray scores and their loss and residual functions specialize to

$$m_c^\tau = cf(\tau \boldsymbol{u}_c), \qquad \Lambda_c^\tau(m) = \sum_{i\in I_c^\tau} p_i e^{-d_i m}, \qquad A_c^\tau(m) = \sum_{i\in I_c^\tau} p_i d_i e^{-d_i m}, \qquad \mathcal{R}_c^\tau = A_c^\tau(m_c^\tau).$$

Thus $\mathcal{L} = \sum_{c,\tau}\Lambda_c^\tau(m_c^\tau)$. The score m_c^τ is the signed output at a unit point on a ray. A sample at radius d_i has margin $d_i m_c^\tau$. The residual $\mathcal{R}_c^\tau$ is the same ray's loss weighted once more by radius; since $d_i \geq d_{\min}$, a ray carrying positive loss also exerts a proportionate residual force. Define

$$\mathsf{N}_c^\tau = \sum_{j\in\mathcal{J}_c^\tau} r_j^2, \qquad \mathsf{B}_c^\tau = \sum_{j\in\mathcal{J}_c^\tau}\big(h_j(t_0)^2 + c_j^2\big), \qquad \mathsf{B}_* = \min_{c,\tau}\mathsf{B}_c^\tau > 0.$$

Here N_c^τ is the current squared norm of a group. The quantity B_c^τ is a fixed positive lower bound for it, because h_j increases and b_j is fixed. The proof of fitting will use this persistent amount of group norm to obtain a lower bound on loss dissipation.

Lemma 7 (A score bound by its own group)**.** *Under* (68)*,* $m_c^\tau \leq \mathsf{N}_c^\tau$.

Proof. On the ray $\tau\boldsymbol{u}_c$, groups with orientation $-\tau$ are inactive. The group with sign $-c$ contributes nonpositively to the signed score. Consequently,

$$m_c^\tau = \sum_{j\in\mathcal{J}_c^\tau} r_j\langle a_j, \tau\boldsymbol{u}_c\rangle - \sum_{j\in\mathcal{J}_{-c}^\tau} r_j\langle a_j, \tau\boldsymbol{u}_c\rangle \leq \sum_{j\in\mathcal{J}_c^\tau} r_j\varrho_j \leq \mathsf{N}_c^\tau.$$

All displayed inner products are nonnegative by cone membership. □

The score bound has a second role beyond the fitting estimate: if a ray's score diverges, the group assigned to that ray must grow. We will use this observation to show that its residual force acts for an infinite amount of accumulated time, which in turn makes *every* neuron in the group grow.

Theorem B.2 (Quantitative fitting from populated cones)**.** *Under* (68)*, let* $C = \sin^2\theta d_{\min}^2 \mathsf{B}_*/4 > 0$*. Then, for* $t \geq t_0$*,*

$$-\dot{\mathcal{L}} \geq C\mathcal{L}^2 \quad a.e., \qquad \mathcal{L}(t) \leq \frac{\mathcal{L}(t_0)}{1 + C\mathcal{L}(t_0)(t - t_0)}. \tag{70}$$

Every ray score $m_c^\tau(t)$ *tends to* $+\infty$*.*

Proof. We estimate loss dissipation through the target-normal coordinates. Their growth is positive even when other coordinates or individual scores are not monotone. The energy identity and (69) imply

$$-\dot{\mathcal{L}} = \|\dot{\xi}\|^2 \geq \sum_j \dot{h}_j^2 = \sin^2\theta \sum_{c,\tau} (\mathcal{R}_c^\tau)^2 \sum_{j \in \mathcal{J}_c^\tau} r_j^2 \geq \sin^2\theta \mathsf{B}_* \sum_{c,\tau} (\mathcal{R}_c^\tau)^2.$$

Here $r_j^2 = \varrho_j^2 + c_j^2 \geq h_j(t_0)^2 + c_j^2$. Since $\mathcal{R}_c^\tau \geq d_{\min}\Lambda_c^\tau$ and there are four nonnegative ray losses, Cauchy–Schwarz gives

$$\sum_{c,\tau} (\mathcal{R}_c^\tau)^2 \geq d_{\min}^2 \sum_{c,\tau} (\Lambda_c^\tau)^2 \geq \frac{d_{\min}^2}{4} \left(\sum_{c,\tau} \Lambda_c^\tau \right)^2 = \frac{d_{\min}^2}{4} \mathcal{L}^2.$$

This proves $-\dot{\mathcal{L}} \geq C\mathcal{L}^2$. Since $\mathcal{L} > 0$ at every finite time, integration of $(1/\mathcal{L})' = -\dot{\mathcal{L}}/\mathcal{L}^2 \geq C$ gives the stated bound.

Each ray loss is nonnegative and bounded above by $\mathcal{L}$, so each tends to zero. For any $i \in I_c^\tau$, $p_i e^{-d_i m_c^\tau} \leq \Lambda_c^\tau(m_c^\tau) \to 0$. Since $p_i, d_i > 0$, this forces $m_c^\tau \to +\infty$. □

Lemma 8 (Divergent residual clocks and individual norms)**.** *Under* (68)*, for every group,*

$$\int_{t_0}^{\infty} \mathcal{R}_c^\tau(t)\, dt = \infty.$$

For every neuron, $h_j, \varrho_j, r_j \to \infty$ *and* $c_j/r_j \to 0$*.*

Proof. **Group growth forces a divergent residual integral.** By (69),

$$h_j \leq r_j \leq K_j h_j, \qquad K_j = \left(\sin^{-2}\theta + \frac{c_j^2}{h_j(t_0)^2} \right)^{1/2}.$$

Fix a group, set $H = \sum_{j \in \mathcal{J}_c^\tau} h_j^2$, and $K = \max_{j \in \mathcal{J}_c^\tau} K_j$. Lemma 7 and the fitting theorem give $K^2 H \geq \mathsf{N}_c^\tau \geq m_c^\tau \to \infty$. On the other hand,

$$\dot{H} = 2\sin\theta \mathcal{R}_c^\tau \sum_j h_j r_j \leq 2\sin\theta K \mathcal{R}_c^\tau H.$$

If the residual integral were finite, integration of this inequality would give

$$H(t) \leq H(t_0) \exp\left(2\sin\theta K \int_{t_0}^{t} \mathcal{R}_c^\tau(u)\, du \right),$$

uniformly bounding H. This contradicts the score lower bound.

The same residual integral drives every neuron. For each individual neuron in the group,

$$(\log h_j)' = \sin\theta \mathcal{R}_c^\tau \frac{r_j}{h_j} \geq \sin\theta \mathcal{R}_c^\tau.$$

Thus $h_j(t) \geq h_j(t_0) \exp(\sin\theta \int_{t_0}^t \mathcal{R}_c^\tau)$ tends to infinity. The inequalities $h_j \leq \varrho_j \leq r_j$ give the other norm divergences, and $c_j/r_j \to 0$ because c_j is constant. No lower bound on an individual neuron's fraction of its group's norm is used: all neurons are driven by the same divergent residual integral. □

B.4 WITHIN-GROUP AGREEMENT AND KKT CHARACTERIZATION OF THE LIMITING DIRECTIONS

Fitting and norm divergence do not yet identify the neuron directions: a neuron could, in principle, stay away from its normal inside its cone. We resolve this in three stages. First, a pairwise identity shows that directions within each group approach one another. Second, a homogeneous margin argument makes every normalized parameter limit a direction of a KKT point and ensures that each group is represented in that limit. Third, the geometry of the two cones forces those KKT directions to be the normals.

The distinction between a group and an individual neuron matters here. A group will retain positive norm after normalization, but an individual neuron might have vanishing normalized norm. Pairwise agreement is what allows us to transfer the limiting direction of a surviving group member to every other member.

We begin by expressing the boundary behavior in coordinates. In the interior, all neurons in a group see the same scaled force. At the normal boundary $k_j = 0$, the coordinate cannot become negative; a nonnegative correction records the cancellation required for normal invariance.

Lemma 9 (Reflected cone coordinates)**.** *Fix a group $\mathcal{J}_c^\tau$ and put $d_c^\tau = c_\theta \mathcal{R}_c^\tau - \mathcal{R}_{-c}^\tau$. For each of its neurons there is a measurable $\ell_j \geq 0$ such that, almost everywhere on $[t_0, \infty)$,*

$$\begin{aligned} \dot{h}_j &= \sin\theta \mathcal{R}_c^\tau r_j, \\ \dot{k}_j &= r_j(d_c^\tau + \ell_j), \\ \dot{r}_j &= \sin\theta \mathcal{R}_c^\tau h_j + d_c^\tau k_j, \qquad \ell_j k_j = 0. \end{aligned} \tag{71}$$

Here ℓ_j is a boundary correction, not the class-direction barrier coordinate ℓ used in Section A.5. One can take $\ell_j = 0$ when $k_j > 0$ and $\ell_j = (-d_c^\tau)_+$ when $k_j = 0$, up to null sets.

Proof. For $k_j > 0$, the two rays of orientation τ are active. Their vector field in the (ν, e) basis is $\sin\theta \mathcal{R}_c^\tau \nu + d_c^\tau e$, proving the first two equations with $\ell_j = 0$. This remains the same block field at the other cone boundary, which is not a switching direction. At $k_j = 0$, the target-coordinate equation is unchanged. An absolutely continuous nonnegative function has zero derivative almost everywhere on its zero set, so $\dot{k}_j = 0$ there a.e. The boundary velocity interval from Corollary A.2 has lower endpoint $r_j d_c^\tau$. Hence $d_c^\tau > 0$ can occur on this zero set only on a null set, and $\ell_j = -d_c^\tau$ gives the formula. Finally, differentiating $r_j^2 = h_j^2 + k_j^2 + c_j^2$ and using $\ell_j k_j = 0$ gives

$$r_j \dot{r}_j = h_j \dot{h}_j + k_j \dot{k}_j = r_j\big(\sin\theta \mathcal{R}_c^\tau h_j + d_c^\tau k_j\big),$$

which is the radial equation. The correction makes no contribution to radial motion because it acts only where $k_j = 0$. When $\theta = \pi/2$ the cone is a ray, $k_j = 0$, and the same identities hold with $d_c^\tau = -\mathcal{R}_{-c}^\tau$. □

The next quantity compares two neurons without dividing by their norms. It contains both their directional discrepancy and the effect of their frozen full-vector components:

$$H_{jk} = \big(r_j r_k - \varrho_j \varrho_k\big) + \varrho_j \varrho_k\big(1 - \langle \widehat{a}_j, \widehat{a}_k \rangle\big).$$

Both terms are nonnegative. If H_{jk} stays bounded while both projected norms diverge, the directional discrepancy must vanish.

Lemma 10 (Pairwise agreement)**.** *For neurons j, k in the same group, the quantity $H_{jk} = r_j r_k - \langle a_j, a_k \rangle$ is nonnegative and nonincreasing. Moreover,*

$$\|\widehat{a}_j(t) - \widehat{a}_k(t)\|^2 \leq \frac{2H_{jk}(t_0)}{\varrho_j(t)\varrho_k(t)} \longrightarrow 0. \tag{72}$$

Proof. Nonnegativity follows from $r_j r_k \geq \varrho_j \varrho_k$. To see the cancellation in its subgradient, abbreviate $A = \sin\theta \mathcal{R}_c^\tau$ and $D = d_c^\tau$ for this calculation. Equation (71) gives

$$\begin{aligned} \frac{d}{dt}(r_j r_k) &= A(h_j r_k + r_j h_k) + D(k_j r_k + r_j k_k), \\ \frac{d}{dt}\langle a_j, a_k \rangle &= A(r_j h_k + h_j r_k) + D(r_j k_k + k_j r_k) + r_j \ell_j k_k + r_k \ell_k k_j. \end{aligned}$$

The common interior force therefore cancels exactly, leaving

$$\dot{H}_{jk} = -r_j \ell_j k_k - r_k \ell_k k_j \le 0.$$

Also $H_{jk} \ge \varrho_j \varrho_k (1 - \langle \widehat{a}_j, \widehat{a}_k \rangle)$. Lemma 8 completes the proof. □

Agreement is asymptotic for arbitrary cone entry. Exact synchronization is a special case: if $c_j = c_k = 0$ and the two projected directions already agree, then $H_{jk} = 0$ and they continue to agree.

We now identify which common directions can occur at late times. Because the ReLU network is homogeneous of degree two in its joint parameters, scaling all parameters changes margins quadratically. The relevant object is therefore a margin divided by the squared parameter norm. For the weighted exponential loss, the next proposition constructs a loss-based lower bound on this normalized minimum margin and uses its increase to control directional motion.

Proposition 9 (Weighted homogeneous KKT limits)**.** *Consider the finite-data loss $\mathcal{L}(\xi) = \sum_i p_i e^{-Q_i(\xi)}$, where $Q_i(\xi) = y_i f_\xi(\boldsymbol{x}_i)$, along a global Clarke trajectory with $\mathcal{L}(\xi(t)) \to 0$. Let $p_{\min} = \min_i p_i$. After any time t_* with $\mathcal{L}(t_*) < p_{\min}$,*

$$\widetilde{\gamma}(t) = \frac{\log(p_{\min}/\mathcal{L}(\xi(t)))}{\|\xi(t)\|^2} \tag{73}$$

is positive and nondecreasing. Every accumulation point of $\xi(t)/\|\xi(t)\|$ is the direction of a Clarke KKT point of

$$\min_\xi \frac{1}{2}\|\xi\|^2 \quad \textit{subject to} \quad Q_i(\xi) \ge 1 \quad (1 \le i \le n). \tag{74}$$

Here KKT means that for some $\lambda_i \ge 0$ and $g_i \in \partial Q_i(\xi)$, $\xi = \sum_i \lambda_i g_i$ and $\lambda_i(Q_i(\xi) - 1) = 0$.

This is the degree-two, weighted-exponential specialization of the homogeneous margin argument of Lyu & Li (2020). We give the argument to make the normalization and the claim about *every* accumulation point explicit. In particular, $\mathcal{L} < 1$ alone does not imply positive margins for a probability-normalized loss.

Proof. **Step 1: homogeneity and a velocity representation.** Each Q_i is locally Lipschitz and positively homogeneous of degree two. Its Clarke subgradients satisfy, for $a > 0$,

$$\langle \xi, g_i \rangle = 2Q_i(\xi), \qquad \partial Q_i(a\xi) = a\, \partial Q_i(\xi). \tag{75}$$

These follow first on activation regions and then by limits and convex hulls. The velocity admits the necessary representation

$$\dot{\xi} = \sum_i p_i e^{-Q_i(\xi)} g_i, \qquad g_i \in \partial Q_i(\xi), \tag{76}$$

by taking the same convex combination of adjacent joint region gradients for each sample.

Step 2: a positive normalized margin. For $t \ge t_*$ write $R = \|\xi\|$, $\vartheta = \xi/R$, $b = \log(p_{\min}/\mathcal{L}) > 0$, $\omega_i = p_i e^{-Q_i}/\mathcal{L}$, and $\overline{Q} = \sum_i \omega_i Q_i$. Here R denotes the norm of the entire parameter vector. The weights ω_i are nonnegative and sum to one; they record each sample's fraction of the current loss. If $Q_{\min} = \min_i Q_i$, then $p_{\min} e^{-Q_{\min}} \le \mathcal{L} \le e^{-Q_{\min}}$, whence

$$Q_{\min} - \log(1/p_{\min}) \le b \le Q_{\min} \le \overline{Q}. \tag{77}$$

For the lower bound on the loss, take a sample attaining $Q_{\min}$; its probability is at least $p_{\min}$. For the upper bound, use $Q_i \ge Q_{\min}$ and $\sum_i p_i = 1$. In particular, $\mathcal{L} < p_{\min}$ gives $b > 0$ and forces every sample margin to be positive. Euler's identity, the energy identity, and the orthogonal decomposition of $\dot{\xi}$ give, almost everywhere,

$$\dot{R} = \frac{2\mathcal{L}\overline{Q}}{R} > 0, \qquad \dot{b} = \frac{\|\dot{\xi}\|^2}{\mathcal{L}}, \qquad \|\dot{\xi}\|^2 = \dot{R}^2 + R^2\|\dot{\vartheta}\|^2. \tag{78}$$

Consequently,

$$\dot{\widetilde{\gamma}} = 2\frac{\dot{R}}{R}\frac{\overline{Q} - b}{R^2} + \frac{\|\dot{\vartheta}\|^2}{\mathcal{L}} \ge 0. \tag{79}$$

The two terms have distinct roles. The first is nonnegative because the residual-weighted average margin $\overline{Q}$ is at least b. The second measures change in the parameter direction. Thus increasing the normalized margin also pays for directional motion. Set $\gamma_* = \widetilde{\gamma}(t_*) > 0$. Homogeneity and (77) imply $\gamma_* \le \widetilde{\gamma} \le \min_i Q_i(\vartheta)$. The latter function is bounded on the unit sphere, so $\widetilde{\gamma}$ has a finite limit. Also $\mathcal{L} \to 0$ forces every $Q_i \to \infty$, hence $R \to \infty$.

Step 3: finite directional energy in radial time. Physical time is not the most useful clock for the remaining argument: the loss and hence the speed can become small simply because the parameter norm is large. Instead, measure time by the increase in that norm. Use the radial time $s = \log(R/R(t_*))$, which increases to infinity. The first equation in (78) has a continuous, strictly positive right-hand side along the trajectory, so this change of time is legitimate on every compact interval. Dividing (79) by $\dot{s}$ yields

$$\frac{d\widetilde{\gamma}}{ds} \ge 2\frac{\overline{Q}}{R^2}\|\vartheta_s\|^2 \ge 2\gamma_*\|\vartheta_s\|^2, \qquad \int_0^\infty \|\vartheta_s\|^2\, ds < \infty. \tag{80}$$

Put $\widetilde{g}_i = g_i/R \in \partial Q_i(\vartheta)$. The normalized vector equation is

$$\vartheta_s + \vartheta = \frac{\sum_i \omega_i \widetilde{g}_i}{2\sum_i \omega_i Q_i(\vartheta)}. \tag{81}$$

Subgradients are uniformly bounded on the unit sphere and the denominator is at least $2\gamma_*$. Thus $\|\vartheta_s\| \le K$ for some finite K. We have obtained both a finite total squared speed and a uniform speed bound in radial time. These facts will rule out repeated visits to a region where the normalized stationarity condition fails.

Step 4: stationarity at every accumulation point. Let ϑ_* be any accumulation point, and let $m_* = \min_i Q_i(\vartheta_*) \ge \gamma_* > 0$. Along any sequence $s_\ell \to \infty$ with $\vartheta(s_\ell) \to \vartheta_*$, the Gibbs weights ω_i vanish for indices with $Q_i(\vartheta_*) > m_*$: their normalized margin gap is eventually positive, while the exponent multiplies that gap by $R^2 \to \infty$. By compactness and the closed graph of the Clarke subdifferential, limits of the remaining weights and subgradients have $\omega_i^* \ge 0$, $\sum_i \omega_i^* = 1$, $g_i^* \in \partial Q_i(\vartheta_*)$, and weights supported on $\{i : Q_i(\vartheta_*) = m_*\}$. We claim that such limits can be chosen with

$$\sum_i \omega_i^* g_i^* = 2m_*\vartheta_*. \tag{82}$$

To make the concentration of the weights explicit, choose an index k attaining m_* at ϑ_*. If $Q_i(\vartheta_*) > m_*$, then along the convergent sequence

$$\frac{\omega_i}{\omega_k} = \frac{p_i}{p_k}\exp\!\big[-R^2\big(Q_i(\vartheta) - Q_k(\vartheta)\big)\big] \longrightarrow 0.$$

Thus only minimum-margin samples can carry weight in the limit. What remains is to show that their averaged subgradient is radial, as required by (82).

If this were impossible, compactness and (81) would provide a neighborhood of ϑ_*, a late time, and $\eta > 0$ such that $\|\vartheta_s\| \ge \eta$ almost everywhere when the trajectory is in that neighborhood after that time. Otherwise a sequence with vanishing speed would give (82). If the trajectory eventually stayed in the neighborhood, this lower bound would contradict (80). If it did not, take concentric balls of radii ϵ and 2ϵ around ϑ_* inside that neighborhood. Accumulation at ϑ_* would give infinitely many disjoint passages from the smaller ball to the boundary of the larger one. Each passage covers a distance at least ϵ and, since speed is at most K, takes radial time at least ϵ/K. The lower speed bound then makes each passage contribute at least $\eta^2\epsilon/K$ to $\int \|\vartheta_s\|^2 ds$. Infinitely many such passages contradict its finiteness. This proves (82).

Step 5: rescaling to the margin-one constraints. The unit vector ϑ_* has minimum margin m_* rather than one. Degree-two homogeneity tells us exactly how to normalize it. Finally let $\xi_* = \vartheta_*/\sqrt{m_*}$, $\lambda_i = \omega_i^*/(2m_*)$, and $g_i' = g_i^*/\sqrt{m_*} \in \partial Q_i(\xi_*)$. Then $Q_i(\xi_*) \ge 1$, equality holds on the support of λ, and (82) is exactly $\xi_* = \sum_i \lambda_i g_i'$. These are the required KKT conditions. □

Corollary B.3 (Sharp asymptotic loss rate for finite atomic data)**.** *Under* (68)*,*

$$\mathcal{L}(t) = \Theta\!\left(\frac{1}{t\log t}\right), \qquad \|\xi(t)\| = \Theta\!\left(\sqrt{\log t}\right).$$

Proof. Theorem B.2 gives $\mathcal{L}(t) \to 0$. For sufficiently large t, put $R = \|\xi\|$, $p_{\min} = \min_i p_i$, $b = \log(p_{\min}/\mathcal{L}) > 0$, and $Q_i = y_i f(\boldsymbol{x}_i; \xi)$. Proposition 9 gives $b/R^2 \geq c_0 > 0$. Finite support and degree-two homogeneity also give $b \leq \min_i Q_i \leq KR^2$ and $\|g_i\| \leq KR$ for $g_i \in \partial Q_i(\xi)$, with a fixed K. Thus $R^2 = \Theta(b)$. The velocity representation and Euler identity yield

$$\|\dot{\xi}\| \leq KR\mathcal{L}, \qquad \|\dot{\xi}\| \geq \frac{\langle \xi, \dot{\xi} \rangle}{R} = \frac{2\sum_i p_i e^{-Q_i} Q_i}{R} \geq \frac{2b\mathcal{L}}{R}.$$

The energy identity therefore gives $-\dot{\mathcal{L}} = \Theta(\mathcal{L}^2 \log(1/\mathcal{L}))$ almost everywhere. Writing $u = 1/\mathcal{L}$, this is $\dot{u} = \Theta(\log u)$. Integration, using $\int^u ds/\log s \sim u/\log u$, gives $u = \Theta(t \log t)$, and then $R^2 = \Theta(\log t)$. □

This specializes the degree-two rate of Lyu & Li (2020, Theorem 4.3) to the weighted loss and Clarke trajectories used here. The earlier bound (70) remains an explicit estimate valid from cone entry, whereas this result is asymptotic.

Lemma 11 (Every group survives normalization)**.** *Under (68), choose $t_* \geq t_0$ with $\mathcal{L}(t_*) < p_{\min}$ and set $\gamma_* = \widetilde{\gamma}(t_*) > 0$. For $d^\tau_{c,\max} = \max_{i \in I^\tau_c} d_i$ and $t \geq t_*$,*

$$\frac{\mathsf{N}^\tau_c(t)}{\|\xi(t)\|^2} \geq \frac{\gamma_*}{d^\tau_{c,\max}} > 0. \tag{83}$$

Every normalized parameter limit therefore contains a nonzero neuron from each group. Its nonzero neurons are balanced, lie in U, and retain their assigned signs and cones.

Proof. For each $i \in I^\tau_c$, positive homogeneity on the ray gives $Q_i = d_i m^\tau_c$. By Proposition 9, $d_i m^\tau_c \geq b \geq \gamma_* \|\xi\|^2$. Combine this with Lemma 7 and $d_i \leq d^\tau_{c,\max}$ to get (83). There are finitely many neurons, so a group with positive normalized squared norm has a nonzero neuron in every convergent normalized parameter subsequence. Balancedness and cone membership pass to the limit. The fixed vectors b_j disappear after division by $\|\xi(t)\| \to \infty$. □

We can now reduce normal recovery to a two-direction calculation. Fixing τ leaves one positive group and one negative group. Within each group the pairwise lemma supplies a common limiting direction. The following rigidity statement shows that KKT stationarity cannot support either direction away from its normal inside its cone.

Lemma 12 (Two-cone KKT rigidity)**.** *Consider a balanced KKT point of (74) whose nonzero neurons lie in their assigned planar cones. Fix an orientation τ. Suppose both signs have a nonzero neuron of that orientation and all nonzero neurons within each of these two groups have a common direction, denoted $\boldsymbol{z}_+, \boldsymbol{z}_-$. Then*

$$\boldsymbol{z}_+ = \tau \boldsymbol{n}_-, \qquad \boldsymbol{z}_- = -\tau \boldsymbol{n}_+, \qquad \sum_{i \in I^\tau_+} \lambda_i d_i = \sum_{i \in I^\tau_-} \lambda_i d_i = \frac{1}{\sin\theta}.$$

Proof. Let $\Lambda_+ = \sum_{i \in I^\tau_+} \lambda_i d_i$ and $\Lambda_- = \sum_{i \in I^\tau_-} \lambda_i d_i$. These are the dual-weighted radii on the two rays of orientation τ; they play the role of class-force magnitudes in the stationarity equations. Necessary first-layer KKT stationarity has the form $\boldsymbol{w}_j = v_j \sum_i \lambda_i y_i \alpha_{ij} \boldsymbol{x}_i$, where α_{ij} is one on an active sample, zero on an inactive sample, and belongs to $[0, 1]$ on a switching sample. This representation is necessary by the sample Clarke subdifferential; sufficiency of arbitrary choices is not needed. It also implies $\boldsymbol{w}_j \in U$. Divide by $r_j > 0$ and use $v_j = s_j r_j$.

Stationarity at a normal or away from it. Put $\widetilde{\boldsymbol{z}}_\pm = \tau \boldsymbol{z}_\pm$ and $G = \Lambda_+ \boldsymbol{u}_+ - \Lambda_- \boldsymbol{u}_-$. If the positive direction is not its normal endpoint, both rays of orientation τ are active and

$$\widetilde{\boldsymbol{z}}_+ = G, \qquad c_\theta \Lambda_+ - \Lambda_- > 0. \tag{84}$$

If it is the normal endpoint, projection onto that normal gives $1 = \sin\theta \Lambda_+$. Similarly, a negative direction away from its normal satisfies

$$-\widetilde{\boldsymbol{z}}_- = G, \qquad \Lambda_+ - c_\theta \Lambda_- < 0, \tag{85}$$

whereas a negative normal direction gives $1 = \sin\theta\Lambda_-$. These assertions remain valid if a nonnormal direction is at the other endpoint of its assigned cone. To interpret the strict inequalities, a positive nonnormal direction has a strictly positive component along $e_+^\tau = \tau \boldsymbol{u}_-$. Its stationarity equation therefore requires $c_\theta\Lambda_+ - \Lambda_- > 0$. For a negative nonnormal direction the analogous component along $e_-^\tau = \tau \boldsymbol{u}_+$ requires $c_\theta\Lambda_- - \Lambda_+ > 0$.

Excluding the nonnormal alternatives. For $c_\theta > 0$, both directions cannot be nonnormal: the two strict inequalities would give $\Lambda_- < c_\theta\Lambda_+ < c_\theta^2\Lambda_-$. If only the positive direction were nonnormal, then $\Lambda_- = 1/\sin\theta$, because the negative direction is normal. The positive stationarity equation also gives $\|G\| = 1$. Expanding,

$$1 = \|G\|^2 = \Lambda_+^2 + \frac{1}{\sin^2\theta} - \frac{2c_\theta\Lambda_+}{\sin\theta} = 1 + \left(\Lambda_+ - \frac{c_\theta}{\sin\theta}\right)^2 .$$

Thus $\Lambda_+ = c_\theta/\sin\theta$. But then $c_\theta\Lambda_+ - \Lambda_- = (c_\theta^2 - 1)/\sin\theta < 0$, contradicting the required strict inequality. Interchanging the two signs excludes the case where only the negative direction is nonnormal. Thus both directions are normal, and their normal projections give the stated dual sums. For $c_\theta = 0$ the cones are already rays, so only the normal-projection calculation is needed. □

Theorem B.3 (Asymptotic normal recovery from populated cones)**.** *Under* (68)*, $\mathcal{L}(t) = O(t^{-1})$, every projected neuron norm tends to infinity, and*

$$\widehat{a}_j(t) \longrightarrow \nu_c^\tau, \qquad \frac{\boldsymbol{w}_j(t)}{\|\boldsymbol{w}_j(t)\|} \longrightarrow \nu_c^\tau \quad \textit{for every } j \in \mathcal{J}_c^\tau . \tag{86}$$

These conclusions require neither moment balance nor a common lower bound on individual normalized neuron masses.

Proof. The fitting and divergence assertions are Theorem B.2 and Lemma 8.

Identify the directions of surviving neurons. Consider any sequence $t_\ell \to \infty$. After a subsequence, $\xi(t_\ell)/\|\xi(t_\ell)\|$ converges. By Proposition 9, a positive rescaling of this limit is a KKT point. Lemma 11 provides a nonzero neuron in each group of that limit. Pairwise agreement ensures that all nonzero neurons within each group share its direction. Lemma 12, applied separately to the two orientations, makes each such direction its assigned normal.

Transfer the direction to every neuron. For any neuron, including one whose normalized mass vanishes along this subsequence, compare its projected direction with a surviving neuron in its group using (72). Their directions differ by a quantity tending to zero, because both individual projected norms diverge. Consequently, the comparison works even when those norms grow at very different rates. Thus every projected direction converges to its assigned normal along the extracted subsequence.

Obtain convergence along the whole trajectory. If any neuron failed to converge globally, a sequence staying a fixed distance from its normal would contradict the preceding subsequence argument. Finally, $\boldsymbol{w}_j/r_j = (\varrho_j/r_j)\widehat{a}_j + b_j/r_j$ and $\varrho_j/r_j \to 1$, $b_j/r_j \to 0$, giving full vector recovery. □

Corollary B.4 (Assembly of the balanced two-line result)**.** *Under Theorem B.1, suppose the four assigned targets are all populated. At the last early alignment time,* (68) *holds. Hence every subsequent continuation remains in its assigned projected cones, has loss $O(t^{-1})$, and satisfies* (86)*.*

Proof. At the last alignment time all projected directions are at their assigned normal endpoints. The planar reduction is automatic, so the shared invariance theorem extends cone membership to all future times, without maintaining the early output bound. Theorem B.3 then applies. □

B.5 Imbalanced regimes and conditional finite-time realignment

The late result above is independent of moment balance. Imbalance changes the early attractors and, consequently, the argument that places a given initialization inside the assigned cones. This subsection gives the exact region classification, a quantitative window for neurons starting inside a stable interior region, and a separate sufficient condition for eventual finite-time normal realignment.

The ideal interior attractors describe an early phase in which residual weights are close to one. They need not be the final directions of the trained network. If the exact trajectory later satisfies the populated cone hypothesis, the preceding recovery theorem still forces convergence to the normals.

Interior equilibria of the ideal field. Assume first that $\theta < \pi/2$. For $\tau \in \{\pm 1\}$, define

$$\mathcal{S}_\tau = \{\boldsymbol{z} \in U : \|\boldsymbol{z}\| = 1,\ \tau\langle \boldsymbol{z}, \boldsymbol{u}_+\rangle > 0,\ \tau\langle \boldsymbol{z}, \boldsymbol{u}_-\rangle > 0\}, \qquad \mathbf{d}_\tau = P_\tau \boldsymbol{u}_+ - N_\tau \boldsymbol{u}_-, \qquad D_\tau = \|\mathbf{d}_\tau\| > 0.$$

These are the regions where both rays of the same orientation are active. The ideal force for sign s on $\mathcal{S}_\tau$ is $s\tau\mathbf{d}_\tau$. Thus the classification asks a simple geometric question: does the force point into the same region in which it was computed? If it does, its normalized direction is an interior attractor for that sign. Otherwise there is no such interior attracting direction in that region.

Proposition 10 (Imbalanced region classification)**.** *Write $\rho_\tau = N_\tau/P_\tau$. If $\rho_\tau < c_\theta$, the direction $\zeta_+^\tau = \tau\mathbf{d}_\tau/D_\tau$ belongs to $\mathcal{S}_\tau$; it attracts positive neurons and repels negative neurons. If $\rho_\tau > 1/c_\theta$, the direction $\zeta_-^\tau = -\tau\mathbf{d}_\tau/D_\tau$ belongs to $\mathcal{S}_\tau$; it attracts negative neurons and repels positive neurons. If $c_\theta < \rho_\tau < 1/c_\theta$, that region has no interior equilibrium for either sign. The other two activation regions have no interior equilibria for any positive active moments. At $\rho_\tau = c_\theta$ or $1/c_\theta$, the corresponding equilibrium is on a switching normal.*

Proof. The projections of $\mathbf{d}_\tau$ onto $\boldsymbol{u}_+, \boldsymbol{u}_-$ are $P_\tau - c_\theta N_\tau$ and $c_\theta P_\tau - N_\tau$. Both are positive exactly when $\rho_\tau < c_\theta$, and both are negative exactly when $\rho_\tau > 1/c_\theta$. For example, in the first case $\tau\langle\zeta_+^\tau, \boldsymbol{u}_c\rangle > 0$ for both classes, so ζ_+^τ lies in the region. For a positive neuron, the force is $D_\tau\zeta_+^\tau$; a signed angular deviation e from that direction satisfies $\dot{e} = -D_\tau \sin e$ in the ideal unit-direction flow. Small deviations therefore decay. Reversing the neuron sign reverses this angular equation and makes the same equilibrium repelling. The case $\rho_\tau > 1/c_\theta$ follows with the negative sign.

When $c_\theta < \rho_\tau < 1/c_\theta$, the two projections of $\mathbf{d}_\tau$ have opposite signs. Neither its direction nor its antipode belongs to $\mathcal{S}_\tau$, so neither neuron sign has an interior equilibrium there. At either critical ratio one of the projections vanishes, placing the equilibrium on a switching normal. For the other two regions, the endpoint angular traces in Table 2 have the same strict sign for all positive moments. Their length is less than π, so their angular field has no zero. □

For $\theta = \pi/2$, strict balance holds automatically and there are no interior imbalanced attractors. For $\theta < \pi/2$, genericity conditions in an imbalanced regime must also exclude the smooth repellers just identified; excluding repelling switching normals alone would not suffice.

Proposition 11 (Ideal contraction inside a stable region)**.** *Fix one of the stable interior cases above and let ζ be its attracting direction and $D = D_\tau$ its force magnitude. For an ideal neuron initialized with nonzero projection strictly inside this region and with its attracting sign, the projected trajectory remains in the region and never vanishes. If $e(t)$ is its signed angle from ζ, chosen with $|e(0)| < \pi$, then*

$$\tan\frac{|e(t)|}{2} \le \tan\frac{|e(0)|}{2}\, e^{-Dt}. \tag{87}$$

The projected norm diverges and the full vector direction converges to ζ. Unless initially aligned, projected alignment occurs only asymptotically.

Proof. Within the region, $\dot{a} = rD\zeta$. The vector component of a orthogonal to ζ is constant. If that component is nonzero, a cannot vanish; if it is zero, region membership forces $a = h\zeta$ with $h > 0$, and h increases. The conical region generated by the region is convex and contains ζ in its interior. Adding a positive multiple of ζ therefore keeps a in that region and $\widehat{a}$ in the region. More explicitly, while the solution is in the region,

$$a(t) = a(0) + D\zeta \int_0^t r(u)\, du.$$

This representation rules out exit through a region boundary. Linear growth of the field gives global continuation. The angular equation is $\dot{e} = -(r/\varrho)D\sin e$, so $r/\varrho \ge 1$ and the half-angle identity give, whenever $e \ne 0$,

$$\frac{d}{dt}\log\tan\frac{|e|}{2} = \frac{(|e|)'}{\sin|e|} = -\frac{r}{\varrho}D \le -D.$$

Integration proves (87); if $e(0) = 0$, the invariant ray gives the assertion directly. After a finite time, $|e| < \pi/3$ and $\dot{\varrho} = rD\cos e \geq Dr/2$. Integrating $\operatorname{arsinh}(\varrho/c)$ when $c > 0$, or $\log \varrho$ when $c = 0$, proves divergence and disappearance of the frozen full-vector component in direction. A nonzero transverse component within U precludes exact alignment at finite time. □

At a critical ratio the same constant-force calculation applies on the side containing the region: the attracting force points along its boundary, and an interior initial direction approaches that boundary asymptotically. There is no positive angular margin at the limiting boundary, so the strict finite-time alignment theorem cannot be extended to this case by setting $\Delta = 0$.

A robust early window around interior targets. The next statement applies to a nonempty finite set $\mathcal{J}_{\mathrm{int}}$ of neurons, each initialized strictly inside a region in which its sign has a stable interior target. Other neurons may be present; the output estimate controls their effect through the common loss. For $j \in \mathcal{J}_{\mathrm{int}}$ let ζ_j, D_j denote the target and magnitude, and let $e_j(0) \in (-\pi, \pi)$ be the initial angle. Set

$$D_* = \min_j D_j, \qquad \beta_* = \min\{\pi/2,\ \min_j(\pi - |e_j(0)|)\} > 0, \qquad d_* = \min_j \operatorname{dist}_{\mathbb{S}^1}(\zeta_j, \partial\mathcal{S}_{\tau_j}) > 0.$$

Let $\kappa_* > 0$ be the minimum, over these neurons and the two endpoints of their regions, of the inward angular trace $D_j \sin\angle(\zeta_j, \text{endpoint})$.

The quantity d_* keeps the target tube away from switching boundaries. The angular buffer β_* keeps the initial direction away from the antipode of its target, where contraction would become arbitrarily slow. The boundary margin κ_* ensures that a small perturbation cannot drive an interior trajectory out of its region. Choose

$$\begin{gathered} 0 < \delta < \min\{d_*/2, \beta_*/2, \pi/4\}, \\ \eta_\rho \leq \frac{1}{2}\min\{D_*\sin\delta,\ D_*\sin\beta_*,\ \kappa_*\}, \\ T_{\mathrm{int}} = \frac{2}{D_*}\left[\log\frac{\cot(\beta_*/2)}{\tan(\delta/2)}\right]_+, \qquad \lambda_{\mathrm{int}} = D_*\cos\delta - \eta_\rho > 0, \qquad T_{\mathrm{amb}} = \lambda_{\mathrm{int}}^{-1}\log\frac{2}{\sin\delta}. \end{gathered} \tag{88}$$

Here distances and angles are in radians. The cap $\beta_* \leq \pi/2$ ensures a positive perturbation allowance even if all neurons start at their interior targets. The three bounds on η_ρ preserve, respectively, attraction at the tube boundary, attraction at the antipodal buffer, and inward motion near the region endpoints. The time T_{int} is the comparison time needed to enter the projected tube. Once there, the lower radial-growth rate λ_{int} makes the frozen full-vector component relatively small; T_{amb} is the additional time allocated to that step.

Theorem B.4 (Quantitative interior tubes for the exact dynamics)**.** *For the preceding collection, use the balanced initialization and S_0 of Lemma 6. If*

$$\varepsilon^2 X_{\max} S_0 \exp\bigl(2e^\rho X_1(T_{\mathrm{int}} + T_{\mathrm{amb}})\bigr) < \rho, \tag{89}$$

then for every $j \in \mathcal{J}_{\mathrm{int}}$,

$$\angle(\widehat{a}_j(t), \zeta_j) \leq \delta \quad (T_{\mathrm{int}} \leq t \leq T_{\mathrm{out}}), \qquad \angle(\boldsymbol{w}_j(t), \zeta_j) \leq 2\delta \quad (T_{\mathrm{int}} + T_{\mathrm{amb}} \leq t \leq T_{\mathrm{out}}).$$

The projected δ-tube is invariant while the output bound holds.

Proof. **Preserve the region and the angular buffer.** The smallness assumption supplies the required output horizon. As long as a neuron is in its initial region, Lemma 3 writes its force as $D_j\zeta_j + E_j$, $\|E_j\| \leq \eta_\rho$. The trace approached from *inside* the region points inward at both endpoints with margin at least $\kappa_* - \eta_\rho > 0$. The same strict sign holds in a small neighborhood on the inside of each endpoint. Hence an interior trajectory cannot reach those endpoints while its projection is nonzero: its angular distance from the endpoint increases whenever that distance is sufficiently small. No assertion about the trace on the other side of an endpoint is required.

For $\delta \leq |e_j| \leq \pi - \beta_*$,

$$\frac{d}{dt}|e_j| \leq -\frac{r_j}{\varrho_j}(D_j\sin|e_j| - \eta_\rho) \leq -\frac{D_*}{2}\sin|e_j|.$$

The second inequality follows because the minimum of $\sin u$ on $[\delta, \pi - \beta_*]$ is $\min\{\sin\delta, \sin\beta_*\}$. The same inequality prevents exit through the upper angular buffer.

Reach the tube without losing the projection. Before tube entry, the force norm is at most $D_j + \eta_\rho$. Dividing radial motion by angular motion cancels the factor r_j/ϱ_j and gives

$$\left|\frac{d\log\varrho_j}{d|e_j|}\right| \le \frac{D_j + \eta_\rho}{D_j \min\{\sin\delta, \sin\beta_*\} - \eta_\rho} < \infty.$$

The denominator is positive by (88). Since the angular excursion is finite and monotone, $\log\varrho_j$ cannot tend to $-\infty$ before tube entry. Thus the projection remains nonzero. The preceding angular inequality also yields

$$\frac{d}{dt}\log\tan\frac{|e_j|}{2} \le -\frac{D_*}{2}.$$

Initially $\tan(|e_j(0)|/2) \le \cot(\beta_*/2)$, so integration gives entry by T_{int}.

Keep the tube invariant and align the full-vector direction. At $|e_j| = \delta$, the angular subgradient points strictly inward; the no-departure argument makes the tube invariant. Within it, $\dot\varrho_j \ge r_j(D_j\cos\delta - \eta_\rho) \ge \lambda_{\mathrm{int}} r_j > 0$. This also handles neurons that start inside the tube: their projections are nonzero initially and increase while they remain in it.

For $c_j > 0$, integration of $(\operatorname{arsinh}(\varrho_j/c_j))' \ge \lambda_{\mathrm{int}}$ from tube entry gives $\sin\angle(\boldsymbol{w}_j, \widehat{a}_j) \le 2e^{-\lambda_{\mathrm{int}}(t-T_{\mathrm{int}})}$. At $T_{\mathrm{int}} + T_{\mathrm{amb}}$ this is at most $\sin\delta$. The angle between $\boldsymbol{w}_j$ and $\widehat{a}_j$ lies in $[0, \pi/2)$, so it is at most δ. If $c_j = 0$ it is zero. The spherical triangle inequality proves the full-angle claim. □

To connect this early statement to the late recovery theorem, the tube must lie in the assigned cone, not merely in the activation region. The next lemma supplies that additional geometric fact and shows how to choose a sufficiently small tube.

Lemma 13 (Interior targets belong to the late cones)**.** *Every stable interior target ζ_c^τ lies strictly inside the assigned arc $\mathcal{W}_c^\tau$. Consequently a sufficiently small projected tube around that target is contained in $\mathcal{C}(\nu_c^\tau)$.*

Proof. For a positive stable target, its unnormalized force is

$$\tau\mathbf{d}_\tau = \sin\theta P_\tau \nu_+^\tau + (c_\theta P_\tau - N_\tau)e_+^\tau.$$

Both coefficients are positive, and if they are denoted h, k, then $c_\theta h - \sin\theta k = \sin\theta N_\tau > 0$. These are the strict cone inequalities in (48). For a negative stable target,

$$-\tau\mathbf{d}_\tau = \sin\theta N_\tau \nu_-^\tau + (c_\theta N_\tau - P_\tau)e_-^\tau,$$

and $c_\theta h - \sin\theta k = \sin\theta P_\tau > 0$. Strict membership on a compact arc gives a positive distance to its endpoints and hence the asserted tube. □

Corollary B.5 (Recovery after an imbalanced early phase)**.** *If an exact early trajectory places every neuron in an assigned cone at a finite time and populates all four groups, then Theorem B.3 applies, regardless of the moment ratios. For interior-target neurons, Theorem B.4 and Lemma 13 provide such an entry certificate when δ is smaller than the target's distance to its cone endpoints.*

Proof. These hypotheses are precisely (68). The late theorem uses no early moment-balance hypothesis. □

This corollary is conditional on cone entry and occupancy. The local interior-tube theorem does not by itself certify those properties for every generic initialization in an imbalanced regime.

Finally, asymptotic recovery and exact realignment are different claims. To force an angle to reach zero in finite time, it is enough to have a persistent strict inward bias when time is measured by the accumulated residual of its group. That residual time is infinite by Lemma 8. The following condition gives precisely such a bias.

Proposition 12 (A sufficient condition for finite-time realignment)**.** *Assume* (68)*. Fix a group* $\mathcal{J}_c^\tau$*. If there exist* $T_1 \geq t_0$ *and* $\zeta > 0$ *such that*

$$\frac{\mathcal{R}_{-c}^\tau(t)}{\mathcal{R}_c^\tau(t)} \geq c_\theta + \zeta \qquad (t \geq T_1), \tag{90}$$

then every neuron in that group reaches its projected normal in finite time and stays there. If this condition holds for all four groups, a finite common time exists after which every projected direction is exactly its assigned normal.

Proof. For $\theta = \pi/2$ cone membership already forces normality. Otherwise, let $\psi_j \in [0, \pi/2 - \theta]$ be the angle from ν_c^τ toward e_c^τ. Whenever $\psi_j > 0$, the interior equations give

$$\dot{\psi}_j = \frac{r_j}{\varrho_j}\mathcal{R}_c^\tau \left[\left(c_\theta - \frac{\mathcal{R}_{-c}^\tau}{\mathcal{R}_c^\tau} \right) \cos\psi_j - \sin\theta \sin\psi_j \right] \leq -\sin\theta\zeta\, \mathcal{R}_c^\tau.$$

Here $r_j/\varrho_j \geq 1$ and $\cos\psi_j \geq \sin\theta$. If the angle stayed positive forever, integration and Lemma 8 would force it below zero. More explicitly, up to a first hit,

$$\psi_j(t) \leq \psi_j(T_1) - \sin\theta\zeta \int_{T_1}^t \mathcal{R}_c^\tau(u)\, du.$$

The integral eventually exceeds the finite quantity $\psi_j(T_1)/(\sin\theta\zeta)$, so the angle must hit zero before that happens. The strict inward residual inequality and Lemma 2 prevent departure. There are finitely many neurons, so their alignment times have a finite maximum whenever the condition holds for every group. □

After this realignment, each neuron obeys the translated-ray equation (63). A frozen nonzero b_j still prevents finite-time *full* normal alignment. The unconditional normal convergence in Theorem B.3 does not establish the residual-ratio hypothesis (90); finite-time realignment is therefore stated separately.

C Proofs for the two-hyperplane case

For rotational hyperplane populations, the early dynamics must first remove components along the intersection $K = S_+ \cap S_-$. Unlike the two-line case, this intersection is part of the data span, so its components are not frozen. Rotational symmetry nevertheless makes the ideal directional field explicit. Its intersection sphere is a saddle. Under strict effective-moment balance, almost every initial direction escapes toward one of the two normals assigned to its output sign.

The proof has three stages. First, we analyze this ideal geometry and transfer finite-time alignment to the exact small-initialization dynamics. The transfer is uniform on compact sets avoiding the ideal exceptional set, which consists of the saddle's stable set and the same-class normals. Second, after all neurons have reached their normals, we prove that *every* continuation remains in the quotient plane $U = K^\perp$. This permits the shared cone-invariance and fitting results to be used. Third, spherical symmetry and the inverse-radius moment assumed in the main body identify the late residual balance and force every neuron to return to its normal in finite time and remain there permanently.

We use the hyperplane model and notation of Section A.1: $d \geq 3$, $S_c = \boldsymbol{n}_c^\perp = K \oplus \operatorname{span}\{\boldsymbol{u}_c\}$, $U = K^\perp = \operatorname{span}\{\boldsymbol{n}_+, \boldsymbol{n}_-\}$, and $X \mid \{Y = c\} = \rho_c V_c$, where $\rho_c > 0$ is independent of the uniform direction $V_c = \hat{X}_c$ in S_c. Put $\pi_c = \mathbb{P}(Y = c) > 0$, $c_\theta = \cos\theta \in [0, 1)$, and $s_\theta = \sin\theta > 0$. All neurons are initially nonzero and balanced; write $\boldsymbol{z}_j := \widehat{\boldsymbol{w}}_j = \boldsymbol{w}_j/r_j$ for their unit directions. As in the main body, throughout this section we assume

$$\mathbb{E}[\rho_c^{-1}] < \infty, \qquad \mathbb{E}[\rho_c^2 e^{b\rho_c}] < \infty \quad \text{for every } b > 0, \quad c \in \{+, -\}. \tag{91}$$

The second condition is (global integrability). The inverse-radius moment enters only the final-stage analysis: Section C.6 derives the required projection asymptotics from it and spherical symmetry. No separate projected-density hypothesis is imposed.

C.1 Rotational moments and the ideal potential

For a single rotational class, the active half of the distribution has a mean parallel to the projection of the neuron onto that class's hyperplane. Symmetry determines the direction of this mean; a one-dimensional integral determines its magnitude. For $\ell = d-1$, let

$$\beta_\ell = \frac{\Gamma(\ell/2)}{2\sqrt{\pi}\,\Gamma((\ell+1)/2)}, \qquad M_c = \pi_c \beta_{d-1} \mathbb{E}[\rho_c] = M_c^+ = M_c^-. \tag{92}$$

These effective moments, rather than the class probabilities alone, determine the early balance condition:

$$\gamma_+ = M_- - c_\theta M_+ > 0, \qquad \gamma_- = M_+ - c_\theta M_- > 0. \tag{93}$$

At $\theta = \pi/2$, both inequalities hold automatically. For a positive neuron near an opposite-class normal, γ_+ is the inward force supplied by the negative class minus the largest outward component supplied by the positive class. The quantity γ_- has the analogous meaning for a negative neuron. These margins will become lower bounds on the speed of exact alignment.

Lemma 14 (Rotational moments and the completed ideal field)**.** *Write $P_c = \Pi_{S_c}$ and let B_{S_c} be the closed unit ball of S_c. Define*

$$\mathcal{Q}_c(\boldsymbol{z}) = \begin{cases} \{P_c\boldsymbol{z}/\|P_c\boldsymbol{z}\|\}, & P_c\boldsymbol{z} \neq 0, \\ B_{S_c}, & P_c\boldsymbol{z} = 0. \end{cases}$$

Then the scaled ideal field and its directional inclusion are

$$\mathcal{G}_s^0(\boldsymbol{z}) = s\big(M_+\mathcal{Q}_+(\boldsymbol{z}) - M_-\mathcal{Q}_-(\boldsymbol{z})\big), \qquad \dot{\boldsymbol{z}} \in \Pi_{\boldsymbol{z}}\mathcal{G}_s^0(\boldsymbol{z}). \tag{94}$$

For a nonzero class projection, $\mathbb{E}[X\mathbf{1}_{\{Y=c,\langle \boldsymbol{z},X\rangle>0\}}] = M_c P_c\boldsymbol{z}/\|P_c\boldsymbol{z}\|$.

Proof. Rotations fixing $P_c\boldsymbol{z}$ force the active moment to lie in that direction. Its component there is $\pi_c\mathbb{E}[\rho_c]\mathbb{E}[(V_{c,1})_+]$. The first coordinate of a uniform sphere in $\mathbb{R}^\ell$ has density

$$\frac{\Gamma(\ell/2)}{\sqrt{\pi}\,\Gamma((\ell-1)/2)}(1-v^2)^{(\ell-3)/2}\mathbf{1}_{\{|v|<1\}}.$$

Writing the normalizing factor in this density as a_ℓ, we obtain

$$\mathbb{E}[(V_{c,1})_+] = a_\ell \int_0^1 v(1-v^2)^{(\ell-3)/2}\,dv = \frac{a_\ell}{\ell-1} = \beta_\ell.$$

The middle equality follows from the substitution $u = 1-v^2$; the last uses the Gamma recurrence. This proves (92), including the class-probability factor π_c. At a zero projection, the selector set has support function $M_c\|P_c h\|$, hence is $M_c B_{S_c}$, as proved in Lemma 1. Combining the two class contributions gives (94). □

Let

$$\Phi_0(\boldsymbol{z}) = M_+\|P_+\boldsymbol{z}\| - M_-\|P_-\boldsymbol{z}\|, \qquad \Phi = \Phi_0|_{\mathbb{S}^{d-1}}.$$

This potential records the difference between the lengths of the two class projections, weighted by their active moments. Positive neurons ascend Φ, and negative neurons ascend $-\Phi$. At most one class projection vanishes on the sphere, so the Clarke sum rule with a smooth summand gives $\mathcal{G}_s^0 = s\,\partial\Phi_0$ there. Away from the four normals $\mathcal{N} = \{\pm\boldsymbol{n}_+, \pm\boldsymbol{n}_-\}$, the ideal dynamics are $\dot{\boldsymbol{z}} = s\,\operatorname{grad}_{\mathbb{S}}\Phi(\boldsymbol{z})$, and $(s\Phi(\boldsymbol{z}))' = \|\operatorname{grad}_{\mathbb{S}}\Phi(\boldsymbol{z})\|^2$. For derivatives near a smooth point it is convenient to use a different ambient extension,

$$\widetilde{\Phi}(\boldsymbol{z}) = M_+\sqrt{1-\langle \boldsymbol{z},\boldsymbol{n}_+\rangle^2} - M_-\sqrt{1-\langle \boldsymbol{z},\boldsymbol{n}_-\rangle^2}.$$

It agrees with Φ on the sphere, and

$$\nabla\widetilde{\Phi}(\boldsymbol{z}) = -\frac{M_+\langle \boldsymbol{z},\boldsymbol{n}_+\rangle}{\|P_+\boldsymbol{z}\|}\boldsymbol{n}_+ + \frac{M_-\langle \boldsymbol{z},\boldsymbol{n}_-\rangle}{\|P_-\boldsymbol{z}\|}\boldsymbol{n}_- \quad (\boldsymbol{z}\in\mathbb{S}^{d-1}\setminus\mathcal{N}). \tag{95}$$

The gradient of Φ_0 additionally contains $(M_+/\|P_+\boldsymbol{z}\| - M_-/\|P_-\boldsymbol{z}\|)\boldsymbol{z}$; the two ambient gradients must not be identified. Their difference is radial, however, so projection onto the tangent space of the sphere gives the same spherical gradient. The extension $\widetilde{\Phi}$ is useful because its ambient gradient lies entirely in U, making the critical-point calculation transparent.

Proposition 13 (Critical points and the fixed intersection direction)**.** *Every smooth critical point of Φ lies in $K \cup U$. Along a smooth ideal orbit, $k = \Pi_K \boldsymbol{z}$ satisfies*

$$\dot{k} = \left[s \left(\frac{M_+}{\|P_+ \boldsymbol{z}\|} - \frac{M_-}{\|P_- \boldsymbol{z}\|} \right) - \langle \boldsymbol{z}, G_s^0(\boldsymbol{z}) \rangle \right] k. \tag{96}$$

Thus U is invariant on smooth orbit segments. If $k(0) \neq 0$, then $k(t)/\|k(t)\|$ is fixed, and the orbit lies in $U + \operatorname{span}\{k(0)\}$.

Proof. At a smooth spherical critical point, $\nabla \widetilde{\Phi} = \lambda \boldsymbol{z}$. If $\Pi_K \boldsymbol{z} \neq 0$, projection onto K gives $\lambda = 0$, and independence of $\boldsymbol{n}_+, \boldsymbol{n}_-$ in (95) forces both normal coordinates to vanish. Hence $\boldsymbol{z} \in K$. Otherwise $\boldsymbol{z} \in U$. Thus a smooth equilibrium cannot mix a nonzero intersection component with a nonzero quotient component.

Projecting $\dot{\boldsymbol{z}} = G_s^0(\boldsymbol{z}) - \langle \boldsymbol{z}, G_s^0(\boldsymbol{z}) \rangle \boldsymbol{z}$ onto K, and using $\Pi_K P_c \boldsymbol{z} = k$, gives (96). Its scalar coefficient is bounded on compact smooth orbit segments. Consequently $k(t)$ is a scalar exponential multiple of its initial value on each such segment. It stays zero if initially zero, and otherwise never changes its direction. This proves the invariant-subspace and fixed-direction claims. □

Lemma 15 (No smooth quotient equilibrium under strict balance)**.** *Under (93), Φ has no smooth critical point on $U \cap \mathbb{S}^{d-1}$.*

Proof. On U, each class projection is one-dimensional: $P_c \boldsymbol{z} = \langle \boldsymbol{z}, \boldsymbol{u}_c \rangle \boldsymbol{u}_c$. The ideal field is therefore constant inside each activation sector, just as in the two-line calculation. In a quotient sector with signs $\varsigma_c = \operatorname{sgn}\langle \boldsymbol{z}, \boldsymbol{u}_c \rangle$, the positive full field is $F = M_+ \varsigma_+ \boldsymbol{u}_+ - M_- \varsigma_- \boldsymbol{u}_-$. A stationary unit direction must be $\delta F / \|F\|$ for $\delta \in \{\pm 1\}$. If $r = \varsigma_+ \varsigma_-$, its sector consistency requires both

$$\delta(M_+ - r c_\theta M_-) > 0, \qquad \delta(r c_\theta M_+ - M_-) > 0.$$

Under strict balance the first parenthesis is positive and the second negative for both $r = \pm 1$. No choice of δ is possible. The critical set is the same for either sign of the ideal flow. □

C.2 The intersection saddle and global ideal basins

Theorem C.1 (Normally hyperbolic intersection saddle)**.** *The sphere $\mathcal{K} = K \cap \mathbb{S}^{d-1}$ consists of ideal equilibria. Its transverse spherical Hessian is*

$$A = \left(M_- \boldsymbol{n}_- \boldsymbol{n}_-^\top - M_+ \boldsymbol{n}_+ \boldsymbol{n}_+^\top \right)\big|_U ,$$

with eigenvalues

$$\lambda_\pm = \frac{M_- - M_+ \pm \sqrt{(M_- - M_+)^2 + 4 M_+ M_- s_\theta^2}}{2}, \qquad \lambda_- < 0 < \lambda_+. \tag{97}$$

For either neuron sign, $\mathcal{K}$ has one stable and one unstable normal direction. Its local stable manifold has dimension $d-2$ and zero spherical measure. For $d = 3$, $\mathcal{K}$ consists of two isolated hyperbolic saddles.

Proof. **The transverse linearization.** At $k \in \mathcal{K}$, the full ideal field equals $s(M_+ - M_-)k$, so its tangent projection vanishes. The smooth extension has gradient zero and Hessian $M_- \boldsymbol{n}_- \boldsymbol{n}_-^\top - M_+ \boldsymbol{n}_+ \boldsymbol{n}_+^\top$ at k. It vanishes on $T_k \mathcal{K}$ and preserves $U \subset T_k \mathbb{S}^{d-1}$, with no curvature correction. On U its trace is $M_- - M_+$ and determinant $-M_+ M_- s_\theta^2 < 0$, giving the eigenvalues. The negative determinant is the key fact: one quotient direction is contracting and the other expanding. Changing the neuron sign interchanges their roles without removing the saddle.

Dimension and measure of the stable set. The linearized sign-s field is sA in the two normal directions and zero along $\mathcal{K}$. The normal rates are uniform on this compact equilibrium manifold and dominate the identity center flow. The stable-manifold theorem for normally hyperbolic compact manifolds (Hirsch et al., 1977) therefore gives a local stable manifold of dimension $(d-3)+1 = d-2$, which has zero measure in the $(d-1)$-dimensional sphere. The same argument at each isolated equilibrium covers $d = 3$. □

The stable-manifold theorem is used only in a smooth neighborhood of $\mathcal{K}$: both class projections have norm one on $\mathcal{K}$, so this neighborhood is separated from the nonsmooth normals. The manifold $\mathcal{K}$ itself has dimension $d-3$. Adding its one stable transverse direction produces dimension $d-2$, still one dimension below the ambient sphere. This dimension deficit explains the measure-zero exceptional set.

The normals require a different local argument because the field is nonsmooth there. We measure distance to a target by the norm of the projection onto its hyperplane. Within a fixed hemisphere, this quantity vanishes only at that target. A negative derivative bounded away from zero then gives finite-time alignment.

Lemma 16 (Local attraction, repulsion, and robust alignment)**.** *Assume* (93)*. For each sign s there are $\bar{\chi}_s, \bar{\delta}_s > 0$ such that, near either target $\boldsymbol{z}_* \in \{\pm \boldsymbol{n}_{-s}\}$, with $\chi_{-s}(\boldsymbol{z}) = \|P_{-s}\boldsymbol{z}\|$, the ideal flow satisfies $\dot{\chi}_{-s} \le -\gamma_s/2$ whenever $0 < \chi_{-s} < \bar{\chi}_s$ in that target's hemisphere. An exact early trajectory with $\delta_{\mathrm{fld}}(Q(t)) \le \bar{\delta}_s$ satisfies*

$$\dot{\chi}_{-s} \le -\gamma_s/4 \quad \textit{when } 0 < \chi_{-s} < \bar{\chi}_s. \tag{98}$$

Starting in this neighborhood, it reaches $\boldsymbol{z}_*$ within $4\chi_{-s}(\boldsymbol{z}(t_a))/\gamma_s$ if the error bound holds for that long. Every continuation then stays at $\boldsymbol{z}_*$ while the error bound holds. Ideal sticking is permanent.

Near a same-class normal $\pm\boldsymbol{n}_s$, the sign-s ideal flow instead satisfies $\dot{\chi}_s \ge \gamma_{-s}/2$ in a sufficiently small punctured neighborhood.

Proof. **A uniform inward bound near an opposite-class normal.** Consider $s = +1$ and $\boldsymbol{z}_* = \pm\boldsymbol{n}_-$. Off the target put $\eta = P_-\boldsymbol{z}/\chi_-$ and $q_+(\boldsymbol{z}) = P_+\boldsymbol{z}/\|P_+\boldsymbol{z}\|$. Then

$$\dot{\chi}_- = M_+\langle \eta, q_+(\boldsymbol{z})\rangle - M_- - \chi_-\langle \boldsymbol{z}, G_+^0(\boldsymbol{z})\rangle.$$

The term $-M_-$ is the direct pull toward the target. The other class can oppose this pull, but its limiting projection onto S_- has norm only c_θ. The map q_+ is smooth near $\boldsymbol{z}_*$, and $\|P_- q_+(\boldsymbol{z}_*)\| = c_\theta$. Thus, uniformly over η, $\dot{\chi}_- \le -\gamma_+ + C\chi_-$. The directional perturbation in Theorem A.3 changes this derivative by at most $\delta_{\mathrm{fld}}(Q)$ because χ_- is 1-Lipschitz. Choose a radius with $C\bar{\chi}_+ \le \gamma_+/4$, and then $\bar{\delta}_+ \le \gamma_+/4$, shrinking the neighborhood to avoid the equator and other normals. This gives the stated bounds.

Finite-time alignment and strong sticking. As long as $\chi_- > 0$, integration of the exact bound gives $\chi_-(t) \le \chi_-(t_a) - (\gamma_+/4)(t - t_a)$. It must therefore reach zero within the stated time. If χ_- left zero, a connected component of its positive set beginning at zero would have strictly negative derivative almost everywhere near its left endpoint, contradicting nonnegativity. This is the no-departure argument of Lemma 2; the hemisphere fixes which target was reached. For the ideal field a stationary continuation also exists: its assigned-class tangent component has norm $c_\theta M_+ < M_-$ and is canceled by the opposite-class selector ball.

Repulsion near a same-class normal. Near $\pm\boldsymbol{n}_+$ the same computation gives $\dot{\chi}_+ \ge M_+ - c_\theta M_- - C\chi_+$, proving repulsion. Exchanging the assigned and opposite classes proves the negative cases. □

At an ideal target the radial rate is $\dot{r}/r = s_\theta M_s$; every field value has that radial component. For an exact early neuron that stays there, $|\dot{r}/r - s_\theta M_s| \le \delta_{\mathrm{fld}}(Q)$. If a balance margin vanishes, the strict local alignment argument fails. At a same-class normal under strict balance, zero tangent velocity is an admissible ideal selection, so these initial directions must be excluded from an every-trajectory alignment statement. There is no contradiction between this stationary selection and repulsion: repulsion is a statement about nearby nonnormal directions, whereas a trajectory initialized exactly at the nonsmooth normal can select zero tangent velocity.

Let $\mathcal{W}_s^{\mathrm{st}}(\mathcal{K})$ denote smooth initial directions whose forward ideal orbits converge in distance to $\mathcal{K}$, and set

$$\mathcal{E}_s = \mathcal{W}_s^{\mathrm{st}}(\mathcal{K}) \cup \{\pm\boldsymbol{n}_s\}.$$

Theorem C.2 (Generic finite-time ideal alignment)**.** *Under strict balance, every sign-s ideal trajectory starting away from $\{\pm\boldsymbol{n}_s\}$ either reaches one of $\{\pm\boldsymbol{n}_{-s}\}$ in finite time and stays there, or stays smooth for all time and converges to $\Pi_K\boldsymbol{z}(0)/\|\Pi_K\boldsymbol{z}(0)\| \in \mathcal{K}$. The latter alternative requires $\Pi_K\boldsymbol{z}(0) \neq 0$ and is exactly membership in $\mathcal{W}_s^{\mathrm{st}}(\mathcal{K})$. The exceptional set $\mathcal{E}_s$ has spherical measure zero, and each of the two target basins is open.*

Proof. **A nonaligned trajectory remains in a smooth region.** Smooth uniqueness holds away from $\mathcal{N}$ and the local lemma forces unique directional continuation at a target. Repulsion precludes reaching or accumulating on a same-class normal: in a small punctured neighborhood its projection distance increases at a uniform positive rate. Such a neighborhood cannot be entered from outside and is exited in finite time if the initial direction lies inside it.

The ideal speed is bounded by $M_+ + M_-$. Thus a finite maximal smooth interval has a limiting direction. A smooth limit permits continuation, a same-class normal is excluded, and a target limit is finite-time alignment. Every nonaligned trajectory is consequently global and smooth. Its omega-limit set also avoids the targets, since entry into a local alignment neighborhood would force alignment. Its tail lies in a compact smooth region.

The potential identifies its only possible limit. The bounded potential $s\Phi$ is increasing, so $\int^\infty \|\operatorname{grad}_{\mathbb{S}} \Phi(\boldsymbol{z}(t))\|^2 dt < \infty$. On the compact smooth tail the integrand is uniformly continuous. It therefore tends to zero: otherwise disjoint intervals of a fixed positive length would contribute a fixed positive amount to the integral. Every omega-limit point is smooth critical, hence belongs to $\mathcal{K}$ by Proposition 13 and Lemma 15. If $\Pi_K \boldsymbol{z}(0) = 0$, invariance of U makes this impossible. Otherwise its fixed intersection direction leaves only $\Pi_K \boldsymbol{z}(0)/\|\Pi_K \boldsymbol{z}(0)\|$ as a possible limit, proving convergence. In particular, approaching the critical manifold in distance is enough here to obtain convergence to a single point: the direction of the intersection component has already been fixed by (96).

The exceptional set has measure zero. Every orbit in the global stable set eventually remains in the local stable neighborhood. At some integer time it lies on the local stable manifold. Before that time the smooth finite-time flow is a local diffeomorphism on its open domain. The global stable set is therefore contained in countably many smooth inverse images of the measure-zero local stable manifold. It has measure zero, as does $\mathcal{E}_s$. The use of integer times makes this a countable union; no uncountable-union measure argument is needed.

Aligned basins are open. Finally an aligned initial direction either lies already in a local target neighborhood, or has a compact smooth orbit segment entering one. Continuous dependence on that segment, followed by the local alignment lemma, proves openness of each basin. □

Slow passage near the saddle. In a smooth local coordinate η transverse to the local stable manifold, invariance of that manifold gives $\dot{\eta} = (\nu + o(1))\eta$, where $\nu > 0$ is the unstable eigenvalue of sA. In a sufficiently small neighborhood, $\nu/2 \le (d/dt) \log|\eta| \le 3\nu/2$. For trajectories remaining there until $|\eta|$ reaches a fixed small radius, the residence time is therefore of order $\nu^{-1} \log(1/|\eta(0)|)$. This is a local estimate, not a global alignment clock. Nor is the ideal stable set identified with an exceptional set of the exact coupled training flow.

C.3 Robust transfer to exact early alignment

The ideal basin theorem is qualitative: each generic direction reaches a target, but directions close to the saddle's stable set can take a long time. The exact network follows the ideal field only while its outputs remain small. We therefore combine two comparisons: ordinary continuous dependence along a smooth ideal orbit segment, followed by the robust local alignment estimate near its target. Compactness turns the resulting direction-dependent bounds into a common horizon and error tolerance.

Theorem C.3 (Uniform exact alignment on compact generic sets)**.** *Assume local integrability and strict effective-moment balance. Fix compact direction sets $\mathcal{Z}_s \subset \mathbb{S}^{d-1} \setminus \mathcal{E}_s$ for each sign that is present. There exist $T_{\text{align}} < \infty$ and $\delta_{\mathcal{Z}} > 0$ such that every exact trajectory with $\boldsymbol{z}_j(0) \in \mathcal{Z}_{s_j}$ and*

$$Q(0) = \varepsilon^2 \sum_j \|\bar{\boldsymbol{w}}_j\|^2 \le e^{-2M_0 T_{\text{align}}} \min\{q_0, \delta_{\mathcal{Z}}/C_2\} \tag{99}$$

aligns every neuron by T_{align} at the same normal selected by its ideal basin. Earlier-aligned neurons remain there through T_{align}, and afterwards for as long as the robust local error bound remains valid. Initial directions at target normals are allowed. The constants q_0, M_0, C_2 are those of (29)*.*

Proof. **Initial directions already near a target.** Fix the robust radii and tolerances from Lemma 16. If an initial direction lies in a target neighborhood with projection radius less than $\bar{\chi}_s/2$, choose a

small open neighborhood of that initial direction lying in radius $3\bar{\chi}_s/4$ and in the same hemisphere. Apply the local alignment lemma directly, including when the reference direction is exactly the target.

Comparison along a smooth ideal segment. For any other generic reference direction, the ideal basin theorem provides a finite time t_a strictly before alignment, at which the projection radius is less than $\bar{\chi}_s/4$. The reference orbit on $[0, t_a]$ is compact and smooth. Choose a tube around it whose closure avoids all four normals, and extend its smooth field to a Lipschitz field near the tube, with constant L. Until an exact direction exits that tube, the approximation in Section A.4 and Gronwall give

$$d(t) \le e^{Lt} d(0) + \delta \int_0^t e^{L(t-u)} du \quad \text{if } \delta_{\mathrm{fld}}(Q) \le \delta.$$

Choose the initial neighborhood and $\delta > 0$ so that this bound stays below the tube radius and below $\bar{\chi}_s/4$ on $[0, t_a]$. A first-exit argument keeps the exact direction in the tube. At t_a it is in the robust alignment neighborhood of the same target. The comparison stops before the ideal orbit reaches the nonsmooth normal. Thus no Lipschitz estimate across a switching normal is needed; the local alignment lemma handles the remaining part of the motion.

A common choice for all initial directions. These open initial neighborhoods cover each compact set $\mathcal{Z}_s$. Select finite subcovers. Take $\delta_{\mathcal{Z}}$ smaller than all tube and local tolerances, and T_{align} larger than all reference times plus their local alignment bounds $4\bar{\chi}_s/\gamma_s$. The initialization inequality and Corollary A.1 supply the required error bound through this common horizon, with continuation supplied by Theorem A.2. Indeed, they give $Q(t) \le Q(0)e^{2M_0 T_{\mathrm{align}}} \le q_0$ and $\delta_{\mathrm{fld}}(Q(t)) \le C_2 Q(0) e^{2M_0 T_{\mathrm{align}}} \le \delta_{\mathcal{Z}}$ throughout the comparison interval, in the almost-everywhere sense needed for the velocity estimates. Local strong sticking keeps each aligned neuron fixed until all have aligned. The neighborhoods may be chosen within the open ideal basins, so target assignments are preserved. □

Corollary C.1 (Fixed shapes, random directions, and coverage)**.** *For every fixed finite collection of generic directions and signs there is a positive initialization threshold satisfying* (99)*. If the direction law is absolutely continuous with respect to product spherical measure, such a threshold exists almost surely and may depend on the realized directions. If at least one neuron of each sign starts in each of its two ideal target basins, all four normal groups are populated at the common alignment time.*

Proof. A finite generic set is compact, so the preceding theorem applies. Absolute continuity and the measure-zero exceptional sets make all finitely many directions generic almost surely. The basin-preserving conclusion gives the coverage assertion. □

Genericity alone does not give four-group coverage at finite width. Likewise, the almost-sure statement above does not provide a fixed positive threshold independent of the realized directions. The smooth comparison estimates may require increasingly small errors on long reference orbits near the stable set.

C.4 Random initialization and four-group coverage

The deterministic recovery results require all four normal groups to be represented. We now replace this coverage hypothesis by a probability estimate under a standard random initialization. Throughout this subsection, let

$$\bar{\boldsymbol{w}}_j \overset{\text{iid}}{\sim} \mathrm{Unif}(\mathbb{S}^{d-1}), \qquad \mathbb{P}(s_j = +1) = \mathbb{P}(s_j = -1) = \tfrac{1}{2}, \qquad \boldsymbol{w}_j(0) = \varepsilon \bar{\boldsymbol{w}}_j, \quad v_j(0) = \varepsilon s_j, \tag{100}$$

where all directions and signs are independent. Equivalently, one may normalize independent standard Gaussian vectors to obtain the directions. This is a random choice of the unit shapes and signs in the main body's balanced initialization.

Lemma 17 (Equal probabilities for the four ideal groups)**.** *Assume the strict moment-balance condition of the relevant model. Assign each generic initialization to its ideal target ν_c^τ, with $c = s_j$, and let $L_j = (c, \tau)$ be this label. For hyperplanes the assignment uses the full direction; for lines*

it uses the normalized projection onto U. Then the L_j are independent and uniform on $\{\pm 1\}^2$. In particular, if $N_c^\tau = \#\{j : L_j = (c, \tau)\}$, then

$$(N_+^+, N_+^-, N_-^+, N_-^-) \sim \text{Multinomial}(m; \tfrac{1}{4}, \tfrac{1}{4}, \tfrac{1}{4}, \tfrac{1}{4}).$$

For $m \geq 1$, the event $\mathcal{O}_m$ that all four counts are positive has probability

$$\mathbb{P}(\mathcal{O}_m) = 1 - 4(3/4)^m + 6(1/2)^m - 4(1/4)^m \geq 1 - 4(3/4)^m. \tag{101}$$

Proof. **The two basins have equal mass for each output sign.** In the hyperplane model, write

$$\mathcal{B}_c^\tau = \{z : \text{the sign-}c \text{ ideal trajectory captures at } \nu_c^\tau\}.$$

The rotational ideal field satisfies $\mathcal{G}_c^0(-z) = -\mathcal{G}_c^0(z)$, including at the nonsmooth normals. Since $\Pi_{-z} = \Pi_z$, the antipode of an ideal trajectory is an ideal trajectory with the same output sign. Consequently $-\mathcal{B}_c^\tau = \mathcal{B}_c^{-\tau}$. The two basins are disjoint and cover the sphere outside the measure-zero exceptional set by Theorem C.2. Antipodal invariance of the uniform initialization therefore gives probability $1/2$ to each basin.

For lines, Lemma 4 shows that the repelling normals $\pm \boldsymbol{n}_c$ separate the two target basins. The basin of ν_c^τ is exactly the open projected semicircle on which $\tau\langle z, \boldsymbol{u}_c\rangle > 0$. Thus the label is $\tau = \operatorname{sgn}\langle \tilde{\boldsymbol{w}}_j, \boldsymbol{u}_c\rangle$. The separating set and the set of zero projections have spherical measure zero, and the two hemispheres have equal probability. This argument does not require antipodal symmetry of the line data.

Count the independent initial labels. Conditional on either sign, the two target labels have probability $1/2$. The independent fair output sign therefore makes each of the four labels have probability $1/4$. Labels of different neurons depend only on their own initial direction and sign, so they are independent. For a prescribed group, the probability of no assigned neuron is $(3/4)^m$. A union bound gives the inequality. Inclusion–exclusion over the four missing-group events gives the exact formula: excluding two groups leaves mass $1/2$, excluding three leaves mass $1/4$, and excluding all four is impossible when $m \geq 1$. □

The labels in this lemma are assigned by the ideal one-neuron geometry. The exact network is coupled, so independence of its trained neurons is not assumed. The next proposition transfers the initial labels to the groups using the early-alignment theorems.

Proposition 14 (Four-group coverage for the exact flow)**.** *Assume the data hypotheses and strict moment balance of the corresponding main theorem, and use* (100)*.*

For two lines, there is a deterministic $\varepsilon_{\text{line}}(m) > 0$ such that, for every $0 < \varepsilon \leq \varepsilon_{\text{line}}(m)$, with probability at least $1 - 4(3/4)^m$, every Clarke trajectory has all four projected normal groups populated at the common early-alignment time. The scale can be chosen as a data-dependent constant times $m^{-1/2}$.

For two hyperplanes, for every $m \geq 1$ and $\delta \in (0, 1)$ there are deterministic constants $T_{m,\delta} < \infty$ and $\varepsilon_{\text{hyp}}(m, \delta) > 0$, chosen before sampling the initialization, such that, with probability at least

$$1 - 4(3/4)^m - \delta, \tag{102}$$

every Clarke trajectory, simultaneously for every $0 < \varepsilon \leq \varepsilon_{\text{hyp}}(m, \delta)$, has all neurons at their assigned normals at $T_{m,\delta}$, with all four groups populated. The bound can be sharpened to $\mathbb{P}(\mathcal{O}_m) - \delta$.

Proof. **Lines have a uniform alignment estimate.** With probability one, all finitely many projected directions are nonzero and avoid their repelling normals. The unit shapes give $S_0 = m$ in Theorem B.1. Fix $\rho > 0$ such that $\eta_\rho < \Delta$, and choose, for example,

$$\varepsilon_{\text{line}}(m) = \frac{1}{2}\sqrt{\frac{\rho}{X_{\max} m}}\, \exp[-e^\rho X_1 T_{\text{normal}}(\eta_\rho)]\,.$$

This satisfies the strict early-smallness inequality for every ε in the stated range, independently of the realized directions. Exact alignment preserves the ideal target labels. Consequently the four-group coverage event has exactly the probability in (101).

Hyperplanes admit a deterministic threshold on a high-probability set. Let $\sigma_{\mathbb{S}}$ be uniform spherical probability. The ideal generic set $\mathbb{S}^{d-1} \setminus \mathcal{E}_c$ has full $\sigma_{\mathbb{S}}$-measure. By inner regularity, choose deterministic compact sets

$$\mathcal{Z}_c \subset \mathbb{S}^{d-1} \setminus \mathcal{E}_c, \qquad \sigma_{\mathbb{S}}(\mathcal{Z}_c) \geq 1 - \delta/m, \qquad c \in \{\pm 1\}.$$

Let $\mathcal{K}_{m,\delta}$ be the event that $\bar{\boldsymbol{w}}_j \in \mathcal{Z}_{s_j}$ for every neuron. Since the signs are independent of the directions,

$$\mathbb{P}(\mathcal{K}_{m,\delta}^c) \leq \sum_{j=1}^{m} \mathbb{P}(\bar{\boldsymbol{w}}_j \notin \mathcal{Z}_{s_j}) \leq \delta.$$

Apply Theorem C.3 to these fixed compact sets, obtaining $T_{m,\delta}$ and $\delta_{\mathcal{Z}} > 0$. Since $Q(0) = m\varepsilon^2$, one permissible choice is

$$\varepsilon_{\mathrm{hyp}}(m, \delta) = \frac{e^{-M_0 T_{m,\delta}}}{\sqrt{m}} \sqrt{\min\{q_0, \delta_{\mathcal{Z}}/C_2\}}. \tag{103}$$

On $\mathcal{K}_{m,\delta}$, every exact trajectory captures all neurons by this common horizon and preserves their ideal labels, for every scale in the stated range. Thus $\mathcal{K}_{m,\delta} \cap \mathcal{O}_m$ implies exact four-group coverage, and

$$\mathbb{P}(\mathcal{K}_{m,\delta} \cap \mathcal{O}_m) \geq \mathbb{P}(\mathcal{O}_m) - \mathbb{P}(\mathcal{K}_{m,\delta}^c) \geq \mathbb{P}(\mathcal{O}_m) - \delta.$$

No independence between $\mathcal{K}_{m,\delta}$ and $\mathcal{O}_m$ is needed. □

Corollary C.2 (Recovery with probability tending to one with width)**.** *Under the assumptions of Proposition 14, for every width $m \geq 1$ there exists a deterministic $\varepsilon_*(m) > 0$ such that, for $0 < \varepsilon \leq \varepsilon_*(m)$, with probability at least*

$$1 - 5(3/4)^m, \tag{104}$$

every Clarke trajectory populates all four groups during early alignment and satisfies the corresponding late recovery conclusions. For lines, $\mathcal{L}(t) = O(t^{-1})$, every projected norm diverges, and each full direction converges to its selected normal. For hyperplanes, every neuron is permanently normal after a finite time, $\mathcal{L}(t) = \Theta(t^{-1/2})$, and $r_j(t) = \Theta(t^{1/4})$. In particular, confidence $1 - \eta$, $\eta \in (0, 1)$, is ensured by

$$m \geq \left\lceil \frac{\log(5/\eta)}{\log(4/3)} \right\rceil$$

and the corresponding sufficiently small initialization scale.

Proof. In the hyperplane result set $\delta = (3/4)^m$. The line bound is already stronger than (104). On the resulting coverage event, apply Theorem B.3 or Theorem C.9, respectively. The displayed width condition is equivalent to $5(3/4)^m \leq \eta$. □

Dependence of the initialization scale. The hyperplane threshold in (103) is deterministic, but the compact sets and their alignment constants depend on m and δ. Increased width can place a neuron closer to the saddle's stable set, so the present argument does not provide a width-independent constant multiplying $m^{-1/2}$ for hyperplanes. The corollary removes the coverage condition under a specified small-initialization regime; it does not keep ε fixed as m increases.

Predetermined output-sign counts. If the signs are fixed with $m_+ \geq 1$ positive and $m_- \geq 1$ negative neurons, while the directions remain independent and uniform, the ideal coverage probability is instead

$$(1 - 2^{1-m_+})(1 - 2^{1-m_-}).$$

Indeed, within each sign, either of its two basins is missing with probability 2^{1-m_c}, and the two sign families use disjoint independent directions. The same exact-flow transfer applies, with a loss of at most δ in the hyperplane case. In particular, for even width and $m_+ = m_- = m/2$, the coverage failure is at most $4\,2^{-m/2}$ before the hyperplane transfer allowance.

C.5 STRONG QUOTIENT-PLANE INVARIANCE AFTER EXACT ALIGNMENT

Normal entry puts every weight in U, but this by itself does not establish that every subsequent nonsmooth continuation stays there. To prove that stronger statement, we bound the growth of the total squared component in K by a constant times that same quantity. Since it is zero at entry, Gronwall's inequality forces it to remain zero. Only after this step may we use the planar cone dynamics.

Suppose that at a finite time t_1 every neuron is exactly at an assigned normal:

$$\boldsymbol{z}_j(t_1) = \nu_{s_j}^{\tau_j}, \qquad \nu_+^\tau = \tau \boldsymbol{n}_-, \quad \nu_-^\tau = -\tau \boldsymbol{n}_+, \qquad \tau_j \in \{\pm 1\}. \tag{105}$$

Let $\mathcal{J}_c^\tau = \{j : s_j = c,\ \tau_j = \tau\}$ and use the cones $\mathcal{C}(\nu_c^\tau)$ from (46). On a finite exact-flow interval $I = [t_1, T]$, the estimates below require the selector representation, balanced nonzero neurons, and

$$B_I = \sup_{t \in I} Q(t) < \infty, \qquad \mathfrak{M}_I = \sum_{c=\pm 1} \pi_c \mathbb{E}[\rho_c^2 e^{B_I \rho_c}] < \infty. \tag{106}$$

Global integrability and Theorem A.1 give these bounds on every finite interval. The classwise moment bound in (106) is exactly the joint envelope (42).

Lemma 18 (Weighted spherical-slab estimate)**.** *Let $X = RV$ be rotationally symmetric in $K \oplus \operatorname{span}\{\boldsymbol{u}\}$, with $\|\boldsymbol{u}\| = 1$. For $a \neq 0$, $k \in K$, $\|k\|/|a| \leq 1/2$, and $B \geq 0$,*

$$\mathbb{E}\big[e^{BR}|\langle k, X\rangle| \mathbf{1}_{\{|a\langle X, \boldsymbol{u}\rangle| \leq |\langle k, X\rangle|\}}\big] \leq C_d\, \mathbb{E}[R e^{BR}]\, \frac{\|k\|^2}{|a|}. \tag{107}$$

Proof. The event in the indicator describes samples whose gate can change when the small intersection component k is added to a planar weight. Its probability is bounded by a constant times the thickness $\|k\|/|a|$ of a slab around zero projected coordinate. The coordinate $\langle V, \boldsymbol{u}\rangle$ has density proportional to $(1 - v^2)^{(d-4)/2}$, bounded near zero also when $d = 3$. Hence $\mathbb{P}(|\langle V, \boldsymbol{u}\rangle| \leq \delta) \leq C_d \delta$ for $0 \leq \delta \leq 1/2$. The indicated event implies $|\langle V, \boldsymbol{u}\rangle| \leq \|k\|/|a|$. Bound $|\langle k, X\rangle| \leq R\|k\|$ and use independence of R, V to obtain

$$\mathbb{E}[e^{BR}|\langle k, X\rangle| \mathbf{1}_{\text{slab}}] \leq \mathbb{E}[R e^{BR}] \|k\|\, \mathbb{P}\left(|\langle V, \boldsymbol{u}\rangle| \leq \frac{\|k\|}{|a|}\right) \leq C_d \mathbb{E}[R e^{BR}] \frac{\|k\|^2}{|a|}.$$

One factor of $\|k\|$ comes from the integrand and the other from the slab probability. This quadratic dependence is what the invariance argument needs. □

The next estimate must be collective because changing one neuron's intersection component changes the residual weights seen by all neurons. We separate this coupling from the gate change of each individual neuron by comparing the full network with its planar projection.

Lemma 19 (Collective control of transverse components)**.** *Write $\boldsymbol{w}_j = a_j + k_j$, where $a_j = \Pi_U \boldsymbol{w}_j$, $k_j = \Pi_K \boldsymbol{w}_j$, and put $D_K = \frac{1}{2}\sum_j \|k_j\|^2$. Enlarge each assigned cone to an open conic neighborhood, excluding its vertex, in which $|\langle a_j, \boldsymbol{u}_{s_j}\rangle| \geq (s_\theta/2)\|a_j\|$. On a subinterval of I where all a_j lie in these neighborhoods and $\|k_j\| \leq (s_\theta/4)\|a_j\|$, there is a finite constant C_I, independent of the selectors, such that*

$$\dot{D}_K \leq C_I D_K \qquad \text{a.e.} \tag{108}$$

Proof. **Separate the projected residual from the network error.** Let $\bar{f}(x) = \sum_j v_j \sigma(\langle a_j, x\rangle)$. On class c, write $X = X_{K,c} + Z_c \boldsymbol{u}_c$. Rotational symmetry gives $\mathbb{E}[X_{K,c} \mid Z_c] = 0$ and $\bar{f}(X) = \bar{f}_c(Z_c)$. The selector equation yields

$$\dot{D}_K = \sum_{j,c} v_j c\, \pi_c \mathbb{E}[e^{-cf(X)} \alpha_j S_{j,c} \mid Y = c], \qquad S_{j,c} = \langle k_j, X_{K,c}\rangle.$$

Here and below the class-conditional expectation uses the specified class radial and angular law. Split the residual into $e^{-c\bar{f}_c(Z_c)}$ and $e^{-cf(X)} - e^{-c\bar{f}_c(Z_c)}$. Set $a_{j,c} = \langle a_j, \boldsymbol{u}_c\rangle$. If $P_c \boldsymbol{w}_j \neq 0$, the gate is $\mathbf{1}_{\{a_{j,c} Z_c + S_{j,c} > 0\}}$ almost surely. If $P_c \boldsymbol{w}_j = 0$, then $k_j = 0$ and its transverse contribution is zero, regardless of the selector. There are now three contributions to bound. With the projected residual,

the opposite class has a favorable sign. The assigned class is controlled by the slab estimate. The difference between the actual and projected residuals is controlled only after summing over neurons.

The opposite-class term is nonpositive. For a mean-zero real variable S and any threshold b, $\mathbb{E}[S\mathbf{1}_{\{S>b\}}] \geq 0$: for $b \geq 0$ this is direct, and for $b < 0$ use $-\mathbb{E}[S\mathbf{1}_{\{S\leq b\}}] \geq 0$. Apply this conditionally on Z_c. For the opposite class $c = -s_j$, the coefficient $v_j c$ is negative. Its contribution with projected residual $e^{-c\bar{f}_c(Z_c)}$ is therefore nonpositive. This sign argument requires no lower bound on the opposite-class projection $a_{j,c}$, which may vanish at an assigned normal.

The assigned-class term is quadratic in the transverse component. For the assigned class $c = s_j$, subtract the gate $\mathbf{1}_{\{a_{j,c}Z_c>0\}}$, whose transverse conditional expectation is zero. The remaining gate difference is supported on $|a_{j,c}Z_c| \leq |S_{j,c}|$. Since $|\bar{f}_c(Z_c)| \leq B_I\rho_c$, the slab bound gives

$$|\text{assigned-class projected-residual contribution}| \leq C_d\pi_c|v_j|\mathbb{E}[\rho_c e^{B_I\rho_c}]\frac{\|k_j\|^2}{|a_{j,c}|} \leq C_I'\|k_j\|^2.$$

Indeed $|a_{j,c}| \geq (s_\theta/2)\|a_j\|$ and the transverse tube condition with balancedness bounds $|v_j|/\|a_j\|$. The first weighted moment is finite by (106), splitting $\rho_c \leq 1$ and $\rho_c > 1$.

The residual error is controlled by the collective norm. Finally,

$$|f(X) - \bar{f}(X)| \leq \|X\|\sum_\ell |v_\ell|\|k_\ell\|, \qquad |e^{-cf(X)} - e^{-c\bar{f}(X)}| \leq e^{B_I\|X\|}|f(X) - \bar{f}(X)|.$$

The residual-difference contribution, summed over all neurons and classes, is thus bounded by

$$\mathfrak{M}_I\left(\sum_j |v_j|\|k_j\|\right)^2 \leq 2\mathfrak{M}_I\left(\sum_j v_j^2\right)D_K \leq C_I''D_K.$$

The middle inequality is Cauchy–Schwarz: $(\sum_j |v_j|\|k_j\|)^2 \leq (\sum_j v_j^2)(\sum_j \|k_j\|^2)$. Balancedness gives $\sum_j v_j^2 = Q \leq B_I$, so the coefficient is finite on the prescribed interval. Every term is therefore either nonpositive or bounded by a constant times D_K. Combining the three bounds proves (108). □

Theorem C.4 (Joint strong invariance after exact normal entry)**.** *On every admissible interval $I = [t_1, T]$ satisfying (106), each continuation from (105) satisfies*

$$\Pi_K \boldsymbol{w}_j(t) = 0, \qquad 0 \neq \boldsymbol{w}_j(t) \in \mathcal{C}(\nu_{s_j}^{\tau_j}) \quad (t \in I). \tag{109}$$

Under global integrability this holds for all future time, without any residual-balance condition or restriction on boundary selectors. If $\theta = \pi/2$, the assigned cones are rays, so every direction remains at its assigned normal on the same interval.

Proof. **First prove planarity on a stopped interval.** At t_1 the projected vectors lie in the enlarged conic neighborhoods of the preceding lemma, and $D_K(t_1) = 0$. Stop at the first exit of any projected vector from those neighborhoods or the first failure of a transverse tube inequality. Before that time, Gronwall and (108) give $D_K = 0$. Indeed, $0 \leq D_K(t) \leq D_K(t_1)e^{C_I(t-t_1)} = 0$.

Then apply conditional cone invariance. Thus the full network lies in U, and only now do we apply Proposition 7: the selector and balance hypotheses hold by admissibility, the envelope (42) follows from (106), and the planar reduction now holds on this stopped interval. Thus all neurons remain in their original assigned cones.

Exclude the stopping event. At a putative finite exit, continuity preserves both conclusions. Nonzero norms are bounded away from zero on the finite interval, excluding the cone vertices. The original cones stay strictly inside the enlarged neighborhoods, and zero transverse components cannot violate the tube inequalities. This contradicts exit. This proves invariance throughout I. Under global integrability, apply the argument on every finite interval. For $\theta = \pi/2$ the assigned cones are oriented rays, as treated separately in the shared invariance proposition, giving immediate permanent alignment. □

The order of this proof matters: the transverse estimate establishes planarity before the conditional planar invariance result is used. Symmetry of the field on U alone would not establish strong invariance for every nonsmooth continuation. The boundary criteria of Corollary A.2 now apply unconditionally along these trajectories: at a target, $\mathcal{R}_{-c}^{\tau} > c_\theta \mathcal{R}_c^{\tau}$ gives strict local sticking, whereas the reverse strict inequality precludes sticking while it persists.

Corollary C.3 (Population fitting after populated normal entry)**.** *Assume global integrability, (105), and four-group occupancy,*

$$\mathcal{J}_c^{\tau} \neq \varnothing \qquad (c, \tau \in \{\pm 1\}). \tag{110}$$

Then every continuation satisfies $\mathcal{L}(t) \to 0$ and $m_c^{\tau}(t) \to +\infty$ for all four scores. This fitting conclusion does not yet use the inverse-radius moment.

Proof. The strong invariance theorem verifies the planar reduction and cone hypotheses of Theorem A.4. The occupancy hypothesis verifies its remaining assumption. □

C.6 Spherical projection asymptotics and normalized-margin barriers

Throughout the next three subsections assume the hypotheses of Corollary C.3. The competitive-arc argument in this and the next subsection concerns $0 < c_\theta < 1$; the orthogonal case will follow directly from its already aligned scalar dynamics. Fitting supplies $m_c^{\tau} \to \infty$, but there is no positive minimum data radius along the projected rays: samples can lie arbitrarily close to K. Their contribution governs the late loss. We first separate the network's growing scale from its normalized configuration, then use the spherical projection law near zero to control that configuration.

Oriented neurons and exact normalized equations. For $\mathcal{J}^{\tau} = \mathcal{J}_+^{\tau} \cup \mathcal{J}_-^{\tau}$, define

$$\widetilde{\boldsymbol{w}}_j = \tau s_j \boldsymbol{w}_j, \qquad \zeta_j = \widetilde{\boldsymbol{w}}_j / r_j = \tau s_j \boldsymbol{z}_j, \qquad \lambda_j = r_j^2, \qquad b^{\tau} = \sum_{j \in \mathcal{J}^{\tau}} \lambda_j \zeta_j.$$

Multiplication by τs_j puts both neuron signs and both ray orientations into one common planar picture. For example, an assigned positive target becomes $\boldsymbol{n}_-$, whereas an assigned negative target $-\tau \boldsymbol{n}_+$ becomes $\boldsymbol{n}_+$. The coefficient $\lambda_j = r_j^2$ is the magnitude with which a balanced neuron contributes to its oriented linear score. The oriented unit directions lie in the fixed compact arcs

$$\mathcal{A}_+ = \operatorname{arc}(\boldsymbol{n}_-, \boldsymbol{u}_+), \qquad \mathcal{A}_- = \operatorname{arc}(\boldsymbol{n}_+, -\boldsymbol{u}_-), \qquad \mathcal{A} = \mathcal{A}_+ \cup \mathcal{A}_-,$$

where the arcs have length $\pi/2 - \theta$. The signs and orientations in this change of variables give

$$m_+^{\tau} = \langle b^{\tau}, \boldsymbol{u}_+ \rangle, \qquad m_-^{\tau} = -\langle b^{\tau}, \boldsymbol{u}_- \rangle.$$

Let $F^{\tau}(t) = \mathcal{R}_+^{\tau}(t)\boldsymbol{u}_+ - \mathcal{R}_-^{\tau}(t)\boldsymbol{u}_-$.

Proposition 15 (Exact oriented and normalized equations)**.** *Almost everywhere after t_1,*

$$\begin{aligned} \dot{\zeta}_j &= \Pi_{\zeta_j} F^{\tau} && \text{if } \zeta_j \notin \{\boldsymbol{n}_-, \boldsymbol{n}_+\}, \\ \frac{\dot{r}_j}{r_j} &= \langle \zeta_j, F^{\tau} \rangle, \qquad \dot{\lambda}_j = 2\lambda_j \langle \zeta_j, F^{\tau} \rangle && \text{for every } j \in \mathcal{J}^{\tau}. \end{aligned} \tag{111}$$

On the times at which ζ_j equals its target, $\dot{\zeta}_j = 0$ almost everywhere. For each orientation set

$$\begin{gathered} S^{\tau} = \sum_{j \in \mathcal{J}^{\tau}} \lambda_j, \quad \omega_j = \lambda_j / S^{\tau}, \quad \mu^{\tau} = \sum_{j \in \mathcal{J}^{\tau}} \omega_j \delta_{\zeta_j}, \quad \bar{b}^{\tau} = b^{\tau} / S^{\tau}, \\ h_+^{\tau} = \langle \bar{b}^{\tau}, \boldsymbol{u}_+ \rangle, \qquad h_-^{\tau} = -\langle \bar{b}^{\tau}, \boldsymbol{u}_- \rangle, \qquad m_c^{\tau} = S^{\tau} h_c^{\tau}. \end{gathered} \tag{112}$$

Thus S^{τ} is the total squared norm for one orientation, ω_j is a neuron's fraction of that norm, and μ^{τ} is the corresponding distribution of directions. Its barycenter $\bar{b}^{\tau}$ determines the normalized scores h_c^{τ}. Both the atom locations and their weights evolve. Writing $g_j = \langle \zeta_j, F^{\tau} \rangle$ and $\bar{g} = \sum_j \omega_j g_j$ gives

$$\dot{S}^{\tau} = 2S^{\tau}\bar{g} = 2(m_+^{\tau}\mathcal{R}_+^{\tau} + m_-^{\tau}\mathcal{R}_-^{\tau}), \qquad \dot{\omega}_j = 2\omega_j(g_j - \bar{g}). \tag{113}$$

Moreover, $S^{\tau} \to \infty$ for each orientation.

Proof. Planarity and cone membership give the four-score identity from Proposition 6. Away from the target the active hemispheres are fixed, and the conditional mean of X_K given Z_c vanishes. The exact selector equation becomes $\dot{\tilde{\boldsymbol{w}}}_j = r_j F^\tau$, yielding the radial and directional equations. At a target the zero-gate class is orthogonal to that target, so the radial equation is unchanged. The derivative of an absolutely continuous path is zero almost everywhere on each fixed point level set, giving the directional statement at targets. Summing $\dot{\lambda}_j$ and differentiating λ_j/S^τ proves (113). In particular,

$$\dot{\omega}_j = \frac{\dot{\lambda}_j}{S^\tau} - \omega_j \frac{\dot{S}^\tau}{S^\tau} = 2\omega_j(g_j - \bar{g}).$$

A direction with radial growth rate above the weighted average gains normalized mass; one below that average loses normalized mass. Finally $|h_c^\tau| \le 1$ and $m_c^\tau \to \infty$ by Corollary C.3, so $S^\tau \to \infty$. □

Suppress the orientation when no confusion can arise, and for $a \in \mathcal{A}$ put $x(a) = \langle a, \boldsymbol{u}_+ \rangle$ and $y(a) = -\langle a, \boldsymbol{u}_- \rangle$. In coordinates $\boldsymbol{u}_+ = (1,0)$, $\boldsymbol{n}_+ = (0,1)$, $\boldsymbol{u}_- = (c_\theta, -s_\theta)$, $\boldsymbol{n}_- = (s_\theta, c_\theta)$, the arcs satisfy

$$\begin{aligned} &x(\mathcal{A}_+) \subset [s_\theta, 1], \quad y(\mathcal{A}_+) \subset [-c_\theta, 0], \quad x(\mathcal{A}_-) \subset [-c_\theta, 0], \quad y(\mathcal{A}_-) \subset [s_\theta, 1], \\ &h_+ + h_- \ge 1 - c_\theta, \qquad \mathrm{Var}_\mu(x) \ge s_\theta h_+ - h_+^2, \qquad \mathrm{Var}_\mu(y) \ge s_\theta h_- - h_-^2. \end{aligned} \tag{114}$$

Indeed $x + y \ge 1 - c_\theta$ on both arcs: its angular maximum is in the gap and its minimum on each arc is at the data endpoint, where it equals $1 - c_\theta$. Also $x^2 \ge s_\theta x$ and $y^2 \ge s_\theta y$ pointwise, which gives the variance bounds. The negative coordinate ranges describe competition between the two families: a positive-family atom helps the positive score but can decrease the negative one, and conversely. Nevertheless, the sum $h_+ + h_-$ has a fixed positive lower bound. Thus if one normalized score is small, the other cannot also be small.

Projection asymptotics under the main-body assumptions. Let $\ell = d - 1 \ge 2$ and write $V_{c,1} = \langle V_c, \boldsymbol{u}_c \rangle$. The density of this signed spherical coordinate is

$$f_\ell(v) = a_\ell (1 - v^2)^{(\ell-3)/2} \mathbf{1}_{\{|v|<1\}}, \qquad a_\ell = \frac{\Gamma(\ell/2)}{\sqrt{\pi}\,\Gamma((\ell-1)/2)}.$$

In particular, f_ℓ is continuous at zero with value a_ℓ. Since $Z_c = \rho_c V_{c,1}$, the constants relevant to the late dynamics are

$$\kappa_c = a_\ell \mathbb{E}[\rho_c^{-1}], \qquad C_c = \pi_c \kappa_c \in (0, \infty). \tag{115}$$

These constants are determined by the data model. Rotational symmetry makes them independent of the orientation τ.

Lemma 20 (Spherical small-projection and Laplace asymptotics)**.** *Under the standing spherical model and (91), for either orientation,*

$$\mathbb{P}(0 < \tau Z_c \le r \mid Y = c) = \kappa_c r + o(r) \qquad (r \downarrow 0). \tag{116}$$

Moreover, as $m \to \infty$,

$$\begin{aligned} \Lambda_c^\tau(m) &= C_c m^{-1}(1 + o(1)), \\ A_c^\tau(m) &= C_c m^{-2}(1 + o(1)), \\ H_c^\tau(m) := \mathbb{E}[|Z_c|^2 e^{-m|Z_c|} \mathbf{1}_{\{Y=c, \tau Z_c > 0\}}] &= 2C_c m^{-3}(1 + o(1)). \end{aligned} \tag{117}$$

For $h \in [\delta, 1]$, $\delta > 0$, the expansions at $m = Sh$ are uniform as $S \to \infty$.

Proof. **Average the small-projection probability over the radius.** Define the one-sided angular distribution function

$$F_\ell(u) = \int_0^{\min\{u,1\}} f_\ell(v)\, dv \qquad (u \ge 0).$$

Continuity at zero gives $F_\ell(u)/u \to a_\ell$ as $u \downarrow 0$. Also $F_\ell(u) \le L_\ell u$ for every $u \ge 0$ and some finite L_ℓ: use boundedness of f_ℓ on $[0, 1/2]$ for small u, and $F_\ell(u) \le 1/2 \le u$ for $u \ge 1/2$. Independence of radius and direction gives

$$\frac{\mathbb{P}(0 < \tau Z_c \le r \mid Y = c)}{r} = \mathbb{E}\left[\frac{F_\ell(r/\rho_c)}{r}\right].$$

The integrand converges to a_ℓ/ρ_c and is bounded by L_ℓ/ρ_c. Dominated convergence using $\mathbb{E}[\rho_c^{-1}] < \infty$ proves (116).

Control all three Laplace moments with the same envelope. For $k = 0, 1, 2$ and $a > 0$, define

$$J_k(a) = \int_0^1 v^k e^{-av} f_\ell(v)\, dv, \qquad H_k(a) = a^{k+1} J_k(a).$$

We claim that H_k is bounded on $(0, \infty)$ and

$$H_k(a) \longrightarrow a_\ell \Gamma(k+1) \qquad (a \to \infty).$$

On $v \le 1/2$, substitute $u = av$. The resulting integrand is $u^k e^{-u} f_\ell(u/a)\mathbf{1}_{\{u \le a/2\}}$. It is bounded by a constant times $u^k e^{-u}$ and converges pointwise to $a_\ell u^k e^{-u}$. On $v \ge 1/2$, the contribution to $H_k(a)$ is at most $\frac{1}{2} a^{k+1} e^{-a/2}$, which is bounded and tends to zero. Finally, for $a \le 1$, $H_k(a) \le 1/2$. These bounds prove the claim, including when $\ell = 2$, where the angular density is unbounded at $v = 1$.

Conditioning on the radius now gives

$$\begin{aligned} &m^{k+1}\mathbb{E}\Big[|Z_c|^k e^{-m|Z_c|}\mathbf{1}_{\{Y=c, \tau Z_c > 0\}}\Big] \\ &\quad = \pi_c \mathbb{E}\big[\rho_c^{-1} H_k(m\rho_c)\big] \longrightarrow C_c \Gamma(k+1). \end{aligned}$$

The boundedness of H_k supplies an integrable envelope proportional to ρ_c^{-1}. Dominated convergence proves the three formulas, since $\Gamma(1) = \Gamma(2) = 1$ and $\Gamma(3) = 2$. Uniformity follows from $Sh \ge S\delta \to \infty$. □

Large scores $m_c^\tau = S^\tau h_c^\tau$ do not by themselves ensure that h_c^τ stays positive uniformly: the scale S^τ might grow faster than one of the scores. Establishing a lower bound on h_c^τ is therefore a separate step.

Lemma 21 (Direct normalized-margin inequalities)**.** *For each orientation, almost everywhere after normal entry,*

$$\dot h_+ \ge 2\mathcal{R}_+ h_+ (s_\theta - h_+) - 4\mathcal{R}_-, \qquad \dot h_- \ge 2\mathcal{R}_- h_- (s_\theta - h_-) - 4\mathcal{R}_+. \tag{118}$$

Proof. For this proof write $x_j = \langle \zeta_j, \boldsymbol{u}_+ \rangle$ and $z'_j = \langle \zeta_j, \boldsymbol{u}_- \rangle$. The weight equation gives

$$\dot h_+ = 2\mathcal{R}_+ \operatorname{Var}_\mu(x) - 2\mathcal{R}_- \operatorname{Cov}_\mu(x, z') + \sum_j \omega_j \langle \boldsymbol{u}_+, \dot\zeta_j \rangle.$$

The first two terms come from redistribution of the weights ω_j; the last comes from motion of the directions themselves. Both scalar coordinates lie in $[-1, 1]$, so the covariance has absolute value at most one. At a non-target atom,

$$\langle \boldsymbol{u}_+, \dot\zeta_j \rangle = \mathcal{R}_+(1 - x_j^2) - \mathcal{R}_-(c_\theta - x_j z'_j) \ge -2\mathcal{R}_-.$$

At a target its derivative is zero almost everywhere, which satisfies the same lower bound. Use (114) to get the first inequality. The same computation with $y = -\langle a, \boldsymbol{u}_- \rangle$ and interchanged classes proves the second. □

Theorem C.5 (Normalized margins stay away from zero)**.** *Under the standing assumptions and four-group occupancy, there exist $t_0 < \infty$ and $\delta > 0$ such that $h_c^\tau(t) \ge \delta$ for every c, τ and every $t \ge t_0$.*

Proof. **Compare residuals using already-divergent scores.** All four unnormalized margins tend to infinity. Choose t_0 so that they are positive and their scalar tail expansions imply, for a common $K_0 > 0$ and both orientations,

$$\frac{\mathcal{R}_+}{\mathcal{R}_-} \ge K_0 \left(\frac{h_-}{h_+}\right)^2, \qquad \frac{\mathcal{R}_-}{\mathcal{R}_+} \ge K_0 \left(\frac{h_+}{h_-}\right)^2 \quad (t \ge t_0).$$

Only large unnormalized margins are used here. The factor S^{-2} cancels from the ratio because both classes have the same m^{-2} residual asymptotic. We have not yet assumed that either normalized margin is bounded away from zero.

A small normalized margin has a dominant corrective force. Put $d_0 = 1 - c_\theta > 0$. If $0 < h_+ \le \min\{s_\theta/2, d_0/2\}$, then $h_- \ge d_0/2$ and

$$\frac{\mathcal{R}_+ h_+}{\mathcal{R}_-} \ge K_0 (d_0/2)^2 h_+^{-1}.$$

This lower bound diverges as $h_+ \downarrow 0$. It eventually outweighs the adverse term in (118). Choose $b > 0$ smaller than $s_\theta/2$ and $d_0/2$ with $s_\theta K_0 (d_0/2)^2 b^{-1} > 4$. Whenever $0 < h_+ \le b$, the first inequality in (118) gives $\dot{h}_+ \ge s_\theta \mathcal{R}_+ h_+ - 4\mathcal{R}_- > 0$. The same conclusion holds for $\dot{h}_-$ when $0 < h_- \le b$.

Turn the derivative bound into a barrier. Take $\delta > 0$ strictly smaller than b and the four positive values at t_0. A connected component below δ starting at that level would have positive derivative almost everywhere, an impossibility. Thus none of the four margins crosses below δ. □

C.7 The tail variational problem and convergence

Fix an orientation and omit its superscript. The margin barrier now justifies using the Laplace expansions uniformly along the trajectory. For this orientation, the leading loss is

$$\Lambda_+(Sh_+) + \Lambda_-(Sh_-) = S^{-1}\big(C_+ h_+^{-1} + C_- h_-^{-1} + o(1)\big).$$

The factor S^{-1} describes the overall growth of the network. The expression in parentheses describes how well the normalized configuration allocates that scale between the two classes. For a probability measure μ on $\mathcal{A}$ with positive normalized margins define

$$\Psi(\mu) = C_+ h_+^{-1} + C_- h_-^{-1}, \qquad F_{\mathrm{tail}}(\mu) = C_+ h_+^{-2} \boldsymbol{u}_+ - C_- h_-^{-2} \boldsymbol{u}_-. \tag{119}$$

The vector F_{tail} is the negative gradient of the objective with respect to its barycenter $\bar{b}$ and satisfies $\langle \bar{b}, F_{\mathrm{tail}} \rangle = \Psi$. The first task is to identify the unique minimizing measure. The second is to show that the exact trajectory converges to it, including when atoms lie at nonsmooth target endpoints. The optimization is over weighted directions, not individual neuron parameters. Several neurons at the same normal represent a single atom whose mass is their combined normalized squared norm.

The geometric ingredient is an exposed face: the set of points in the convex hull of the arcs that maximize a specified linear functional. The direction of that functional is determined by a residual ratio. We first compute all these faces, then use positivity of both margins to select the one compatible with an optimum.

Lemma 22 (Exposed faces of the two-arc hull)**.** *For $\rho > 0$ put $F_\rho = \boldsymbol{u}_+ - \rho \boldsymbol{u}_-$. The face of $\mathrm{conv}(\mathcal{A})$ maximizing $\langle F_\rho, a \rangle$ is*

$$\begin{cases} \{F_\rho/\|F_\rho\|\} \subset \mathcal{A}_+ \setminus \{\boldsymbol{n}_-, \boldsymbol{u}_+\}, & 0 < \rho < c_\theta, \\ \{\boldsymbol{n}_-\}, & c_\theta \le \rho < 1, \\ \mathrm{conv}\{\boldsymbol{n}_-, \boldsymbol{n}_+\}, & \rho = 1, \\ \{\boldsymbol{n}_+\}, & 1 < \rho \le c_\theta^{-1}, \\ \{F_\rho/\|F_\rho\|\} \subset \mathcal{A}_- \setminus \{\boldsymbol{n}_+, -\boldsymbol{u}_-\}, & \rho > c_\theta^{-1}. \end{cases} \tag{120}$$

Proof. In the coordinates used for (114), $F_\rho = (1 - c_\theta \rho, s_\theta \rho)$. Its angle increases from that of $\boldsymbol{u}_+$ toward that of $-\boldsymbol{u}_-$. Its unit direction lies in the positive arc when $\rho < c_\theta$, in the gap when $c_\theta < \rho < c_\theta^{-1}$, and in the negative arc when $\rho > c_\theta^{-1}$. In either arc its own direction uniquely maximizes the linear functional on the unit circle. In the gap, the only candidates are the adjacent targets, with scores $\langle F_\rho, \boldsymbol{n}_- \rangle = s_\theta$ and $\langle F_\rho, \boldsymbol{n}_+ \rangle = \rho s_\theta$. At the transition ratios the direction is the corresponding target. Taking the convex hull of the maximizing atoms proves the formula. In particular, equal target scores occur only at $\rho = 1$. This is the only exposed face containing atoms from both families, and it is the segment joining the two normals. □

Lemma 23 (The unique tail-optimal measure)**.** *Among probability measures on $\mathcal{A}$ with $h_\pm > 0$, the unique minimizer of Ψ is*

$$\mu_* = \omega_* \delta_{\boldsymbol{n}_-} + (1-\omega_*)\delta_{\boldsymbol{n}_+}, \qquad \omega_* = \frac{C_+^{1/2}}{C_+^{1/2} + C_-^{1/2}}. \tag{121}$$

Its margins are $h_{+,*} = s_\theta \omega_*$ and $h_{-,*} = s_\theta(1-\omega_*)$, and its tail residual ratio is one.

Proof. **Existence and uniqueness of the optimal barycenter.** A positive mixture of the two target atoms is feasible. A finite sublevel of Ψ bounds both margins away from zero; compactness of $\mathrm{conv}(\mathcal{A})$ therefore gives a minimizing barycenter. The Hessian of its objective on U is

$$2\left[C_+ h_+^{-3} \boldsymbol{u}_+ \boldsymbol{u}_+^\top + C_- h_-^{-3} \boldsymbol{u}_- \boldsymbol{u}_-^\top\right],$$

which is positive definite since $\boldsymbol{u}_+, \boldsymbol{u}_-$ are independent. Thus the minimizing barycenter is unique.

Its exposing force must balance the two targets. Write the objective as a function of $\bar{b}$, and let $\bar{b}_*$ be its minimizer. The convex first-order condition says that, for every feasible barycenter $\bar{b}$,

$$\langle \nabla_{\bar{b}} \Psi(\bar{b}_*), \bar{b} - \bar{b}_* \rangle \geq 0, \qquad \langle F_{\mathrm{tail}}(\bar{b}_*), \bar{b} \rangle \leq \langle F_{\mathrm{tail}}(\bar{b}_*), \bar{b}_* \rangle.$$

The second inequality also holds for every $\bar{b}$ in the full arc hull: apply it first to $\bar{b}_* + t(\bar{b} - \bar{b}_*)$, whose margins are positive for sufficiently small $t > 0$, and divide by t. Thus $\bar{b}_*$ belongs to the face exposed by F_ρ, where $\rho = C_- h_-^{-2} / (C_+ h_+^{-2})$. If $\rho \neq 1$, Lemma 22 puts this face in a single atomic family, which gives a nonpositive margin for the other class by (114). This is infeasible. Hence $\rho = 1$, and the barycenter is a mixture of the two targets.

For a mixture with mass ω at $\boldsymbol{n}_-$, the two margins are $s_\theta \omega$ and $s_\theta(1-\omega)$. The equation $\rho = 1$ becomes

$$C_+ \omega^{-2} = C_-(1-\omega)^{-2}, \qquad \frac{\omega}{1-\omega} = \left(\frac{C_+}{C_-}\right)^{1/2}.$$

Solving gives (121).

The barycenter determines the entire optimal measure. Only the target atoms maximize the exposing functional on $\mathcal{A}$. Any probability measure with the optimal barycenter must therefore be supported on those targets, and its barycenter determines their masses. This proves uniqueness of the measure. This last step is needed because a unique barycenter does not in general imply a unique measure representing it. □

Identifying the optimizer does not yet prove that training approaches it. We next construct a dissipation quantity with two nonnegative parts. The variance part measures unequal growth rates of atoms and therefore redistribution of their masses. The directional part measures motion of the atoms along their arcs. To handle the nonsmooth endpoints, define the following lower bound on directional dissipation:

$$d(a,F) = \begin{cases} 0, & a \in \{\boldsymbol{n}_-, \boldsymbol{n}_+\}, \\ \|\Pi_a F\|^2, & a \in \mathcal{A} \setminus \{\boldsymbol{n}_-, \boldsymbol{n}_+\}, \end{cases} \qquad \mathcal{D}(\mu) = 2\,\mathrm{Var}_\mu(\langle a, F_{\mathrm{tail}}(\mu)\rangle) + \int d(a, F_{\mathrm{tail}}(\mu))\, d\mu(a). \tag{122}$$

target.

Lemma 24 (Unique zero and compact dissipation gaps)**.** *On the feasible set $h_\pm > 0$, $\mathcal{D}(\mu) = 0$ holds exactly when $\mu = \mu_*$. For a fixed finite number of atoms, $\mathcal{D}$ is lower semicontinuous on every compact normalized state set with $h_\pm \geq \delta > 0$, allowing zero atom weights. In particular, on any nonempty such set with $\Psi \geq \Psi_* + \epsilon$, $\epsilon > 0$, its minimum is positive, where $\Psi_* = \Psi(\mu_*)$.*

Proof. **Classify all feasible zero-dissipation states.** Write $F_{\mathrm{tail}} = A F_\rho$, $A > 0$. Zero dissipation requires every positive-weight atom to be either a target or a smooth stationary direction, and all such atoms must have the same score. Feasibility requires positive mass in both families, by (114). After dividing scores by A, the target scores are s_θ and ρs_θ. A non-target stationary direction, if present, is $F_\rho / \|F_\rho\|$ with score $\|F_\rho\|$. Its antipode cannot belong to the arcs: F_ρ has strictly positive second coordinate and both arcs lie in the closed upper half-plane.

If $\rho < c_\theta$, the only non-target stationary candidate is in $\mathcal{A}_+$. Both its score and the positive target score exceed the negative target score, because $\rho < c_\theta < 1$ and $\|F_\rho\|^2 - \rho^2 s_\theta^2 = (1 - c_\theta\rho)^2 > 0$. Thus zero variance is impossible with mass in both families. For $c_\theta \le \rho \le c_\theta^{-1}$ there is no non-target stationary candidate, so the two target scores agree only at $\rho = 1$. For $\rho > c_\theta^{-1}$, both negative-family candidate scores exceed the positive target score: $\rho > 1$ and $\|F_\rho\|^2 - s_\theta^2 = (\rho - c_\theta)^2 > 0$. It follows that a feasible zero has $\rho = 1$ and support only on the targets. The ratio equation fixes its weights to those of μ_*. The converse is immediate.

Pass from a unique zero to a uniform positive gap. For compactness, use the state $(\omega_1, \dots, \omega_N, \zeta_1, \dots, \zeta_N)$ in the closed simplex times the assigned compact arcs. The margin constraints are closed, and F_{tail} is continuous and bounded on $h_\pm \ge \delta$. The variance term is continuous. The function $d(a, F)$ is jointly lower semicontinuous, and so is each nonnegative weighted term $\omega_j d(\zeta_j, F_{\text{tail}})$, including when $\omega_j \to 0$. Thus $\mathcal{D}$ attains its minimum on every compact subset. The indicated objective-separated subset contains no zero, so its minimum is strictly positive. Thus a normalized state whose objective stays a fixed distance above the optimum must dissipate at a uniformly positive rate in the rescaled dynamics. □

Lemma 25 (Exact asymptotic descent and an infinite residual clock)**.** *Along each orientation's exact trajectory, after the normalized margin barrier applies, there is $\eta(t) \to 0$ such that*

$$\frac{d}{dt}\Psi(\mu(t)) \le -S(t)^{-2}\big(\mathcal{D}(\mu(t)) - \eta(t)\big) \quad \text{a.e.} \tag{123}$$

Moreover,

$$S(t) = \Theta(t^{1/2}), \qquad \int^{\infty} S(t)^{-2}\, dt = \infty. \tag{124}$$

Proof. **Differentiate the exact normalized trajectory.** On the fixed compact normalized state set given by Theorem C.5, the Laplace asymptotics are uniform. They give $F(t) = \ell(t)(F_{\text{tail}}(\mu(t)) + e(t))$, where $\ell = S^{-2}$ and $\|e(t)\| \to 0$. Put $G_j = \langle \zeta_j, F_{\text{tail}} \rangle$. Differentiate the barycenter using (113) and then the objective:

$$\dot{\bar{b}} = 2\sum_j \omega_j (g_j - \bar{g})\zeta_j + \sum_j \omega_j \dot{\zeta}_j, \qquad \dot{\Psi} = -\langle F_{\text{tail}}, \dot{\bar{b}} \rangle.$$

Consequently,

$$\dot{\Psi} = -2\,\mathrm{Cov}_\mu(G, \langle a, F \rangle) - \sum_j \omega_j \langle F_{\text{tail}}, \dot{\zeta}_j \rangle.$$

The covariance equals $\ell\,\mathrm{Var}_\mu(G) + O(\ell\|e\|)$. At each non-target atom, $\langle F_{\text{tail}}, \dot{\zeta}_j \rangle = \ell \|\Pi_{\zeta_j} F_{\text{tail}}\|^2 + O(\ell\|e\|)$. At a target, its derivative is zero almost everywhere on the target-level set, as is $d(\zeta_j, F_{\text{tail}})$. All constants are uniform on the fixed compact state set. This gives (123), without assuming continuity of the endpoint velocity.

The available rescaled time is infinite. Finally the exact mass identity and the tail expansions yield

$$\dot{S} = 2\Psi(\mu) S^{-1}(1 + o(1)).$$

The normalized margin bound makes Ψ bounded above and bounded below by a positive constant. Consequently $(S^2)'$ lies between two positive constants for all large times. Integration gives $S^2 = \Theta(t)$ and the divergent clock in (124). Equivalently, the clock with differential $ds = S(t)^{-2} dt$ runs for an infinite amount of time. This matters because the physical-time force tends to zero: a positive dissipation gap can force convergence only if its accumulated effect does not become finite. □

Theorem C.6 (Tail implicit bias of every exact continuation)**.** *Under the preceding hypotheses, for both orientations,*

$$\mu^\tau(t) \rightharpoonup \mu_*, \qquad \Psi(\mu^\tau(t)) \longrightarrow \Psi_*. \tag{125}$$

Proof. **Every objective level above the optimum is eventually reached.** The margin barrier places each trajectory in a fixed compact normalized state set for all sufficiently large times. For any $\epsilon > 0$, its closed subset $\Psi \ge \Psi_* + \epsilon$ either is empty or has a positive minimum d_ϵ of $\mathcal{D}$. Take time sufficiently large that $\eta(t) < d_\epsilon/2$. As long as the trajectory remains in that subset,

$$\dot{\Psi} \le -\tfrac{1}{2} d_\epsilon S^{-2}.$$

It cannot remain there forever because the clock in (124) diverges. Otherwise integration would force Ψ below its lower bound Ψ_*.

Once reached late enough, that level cannot be crossed upward. After entry below the indicated objective level, a later crossing above it is impossible: on a connected component above that level, the derivative would be strictly negative almost everywhere, contradicting integration from its left endpoint. Thus $\limsup \Psi \le \Psi_* + \epsilon$. Since $\epsilon > 0$ is arbitrary and $\Psi \ge \Psi_*$, the objective converges to its minimum.

Identify the limiting measure. Every convergent subsequence of normalized states therefore represents an optimal measure. Uniqueness of that measure and compactness prove weak convergence of the full trajectory.

□

C.8 Residual balance, permanent realignment, and final rates

We now pass from the normalized population to individual neurons. The limiting measure determines the ratio of the two residual forces. Once this ratio is close to one, both forces point toward the assigned targets, with a uniform angular advantage. Their infinite accumulated strength then forces each neuron to reach its target in finite time, regardless of its normalized mass.

Theorem C.7 (Residual balance and finite-time permanent realignment)**.** *Assume the standing data hypotheses and exact normal entry (105). If $\theta = \pi/2$, all directions remain at their assigned normals from t_1 onward, without an occupancy hypothesis. If $0 < \theta < \pi/2$, assume additionally four-group occupancy. Then, for each orientation,*

$$\rho^\tau(t) := \frac{\mathcal{R}^\tau_-(t)}{\mathcal{R}^\tau_+(t)} \longrightarrow 1, \qquad \int^\infty \mathcal{R}^\tau_c(t)\, dt = \infty \quad (c = \pm 1). \tag{126}$$

There is a finite common time T_{rec} such that $\boldsymbol{z}_j(t) = \nu^{\tau_j}_{s_j}$ for every neuron and every $t \ge T_{\mathrm{rec}}$.

Proof. **Recover the residual ratio and its time scale.** In the nonorthogonal case, the tail expansions and Theorem C.6 give

$$\rho^\tau(t) = \frac{C_-}{C_+}\left(\frac{h^\tau_+(t)}{h^\tau_-(t)}\right)^2 (1 + o(1)) \longrightarrow 1.$$

The normalized margins are bounded below, so each residual is comparable to $(S^\tau)^{-2}$. Its time integral diverges by Lemma 25.

Obtain a strict angular drift for both signs. Choose T_{ratio} such that, for both orientations and all later times, $(1 + c_\theta)/2 \le \rho^\tau \le 2/(1 + c_\theta)$. For a neuron in $\mathcal{J}^\tau_c$, let $\psi_j \in [0, \pi/2 - \theta]$ be its angular distance from its assigned target along its cone. At $\psi_j > 0$ the exact planar equation gives

$$\dot\psi_j = \mathcal{R}^\tau_c \cos(\theta + \psi_j) - \mathcal{R}^\tau_{-c} \cos\psi_j \le -\frac{s_\theta(1 - c_\theta)}{2}\mathcal{R}^\tau_c. \tag{127}$$

Indeed both ρ^τ and its inverse are at least $(1 + c_\theta)/2$, and $\cos\psi_j \ge s_\theta$. More explicitly, expanding $\cos(\theta + \psi_j)$ gives

$$\dot\psi_j = \mathcal{R}^\tau_c \left[\left(c_\theta - \frac{\mathcal{R}^\tau_{-c}}{\mathcal{R}^\tau_c}\right)\cos\psi_j - s_\theta \sin\psi_j\right].$$

The coefficient of $\cos\psi_j$ is at most $-(1 - c_\theta)/2$, and the remaining term is nonpositive. This proves the displayed bound uniformly over each assigned arc.

Integrate the drift and rule out departure. If the angle never reached zero after T_{ratio}, integration of (127) against the divergent residual clock would give

$$\psi_j(t) \le \psi_j(T_{\mathrm{ratio}}) - \frac{s_\theta(1 - c_\theta)}{2}\int_{T_{\mathrm{ratio}}}^t \mathcal{R}^\tau_c(s)\, ds,$$

whose right-hand side eventually becomes negative. It therefore hits zero in finite time. The same inequality on any attempted positive component of ψ_j prevents subsequent departure. There are finitely many neurons, so the maximum of their alignment times is finite. This argument applies

even to neurons whose normalized weights tend to zero: the angular inequality contains no factor ω_j, so a small mass does not weaken the drift.

For $\theta = \pi/2$, Theorem C.4 already confines every neuron to its assigned ray, so $T_{\mathrm{rec}} = t_1$ works without the additional assumptions. □

Proposition 16 (Four scalar equations after permanent alignment)**.** *On any interval where all directions remain at their assigned normals, let* $\mathsf{N}_c^\tau = \sum_{j\in\mathcal{J}_c^\tau} r_j^2$. *Then*

$$\begin{aligned} m_c^\tau &= s_\theta \mathsf{N}_c^\tau, \qquad \frac{\dot r_j}{r_j} = s_\theta A_c^\tau(m_c^\tau) \quad (j\in\mathcal{J}_c^\tau), \\ \dot{\mathsf{N}}_c^\tau &= 2s_\theta \mathsf{N}_c^\tau A_c^\tau(s_\theta \mathsf{N}_c^\tau), \qquad \dot m_c^\tau = 2s_\theta m_c^\tau A_c^\tau(m_c^\tau). \end{aligned} \tag{128}$$

For a reference time t_* *in that interval,*

$$r_j(t)^2 = \omega_j^* \mathsf{N}_c^\tau(t), \qquad \omega_j^* = \frac{r_j(t_*)^2}{\mathsf{N}_c^\tau(t_*)} > 0, \qquad \dot{\mathcal{L}} = -2s_\theta \sum_{c,\tau} m_c^\tau \big(A_c^\tau(m_c^\tau)\big)^2. \tag{129}$$

The proportions are asserted only for populated groups.

Proof. At an assigned normal, a neuron is inactive on the opposite class and contributes $s_j s_\theta r_j^2(\tau Z_{s_j})_+$ on its assigned class. Summation gives the margin formula. The target radial component in (111) gives the logarithmic radial rate. Summing $2r_j\dot r_j$ gives both scalar equations. The logarithmic rate is common within a group, so all ratios r_j/r_k are constant on that interval, proving the proportion identity. Finally differentiate the four-score loss and use $(\Lambda_c^\tau)' = -A_c^\tau$:

$$\dot{\mathcal{L}} = -\sum_{c,\tau} A_c^\tau(m_c^\tau)\dot m_c^\tau = -2s_\theta \sum_{c,\tau} m_c^\tau \big(A_c^\tau(m_c^\tau)\big)^2.$$

□

Theorem C.8 (Aligned asymptotic rates)**.** *Under the standing assumptions and four-group occupancy, suppose permanent alignment has occurred, as guaranteed above. Set*

$$B_c = 4s_\theta C_c, \qquad D_c = C_c B_c^{-1/2}. \tag{130}$$

Then, for every populated group,

$$\begin{aligned} m_c^\tau(t) &\sim (B_c t)^{1/2}, & \mathsf{N}_c^\tau(t) &\sim s_\theta^{-1}(B_c t)^{1/2}, \\ \mathcal{R}_c^\tau(t) &\sim \frac{1}{4s_\theta t}, & \Lambda_c^\tau(m_c^\tau(t)) &\sim D_c t^{-1/2}, \\ r_j(t) &\sim \sqrt{\omega_j^*/s_\theta}\,(B_c t)^{1/4} & & (j\in\mathcal{J}_c^\tau). \end{aligned} \tag{131}$$

In particular, $\mathcal{L}(t) \sim 2(D_+ + D_-)t^{-1/2}$. *Consequently,* $\mathcal{L} = \Theta(t^{-1/2})$, $m_c^\tau, \mathsf{N}_c^\tau = \Theta(t^{1/2})$, *and* $r_j = \Theta(t^{1/4})$.

Proof. **Integrate the margin equation.** Each positive scalar margin increases to infinity: a finite limit would leave the right-hand side of its scalar equation strictly positive. Substituting $A_c^\tau(m) \sim C_c m^{-2}$ into (128) gives, writing $m = m_c^\tau$,

$$\frac{d}{dt}m^2 = 2m\dot m = 4s_\theta m^2 A_c^\tau(m) \longrightarrow 4s_\theta C_c = B_c.$$

Integration yields $m^2/t \to B_c$, hence the margin rate. The mass and neuron-norm rates follow from (128) and (129).

Substitute into the residual and loss expansions. The residual satisfies

$$\mathcal{R}_c^\tau(t) = A_c^\tau(m_c^\tau(t)) \sim \frac{C_c}{B_c t} = \frac{1}{4s_\theta t}.$$

Its leading coefficient is independent of the class, consistently with the residual-ratio limit proved before realignment. Likewise,

$$\Lambda_c^\tau(m_c^\tau(t)) \sim C_c(B_c t)^{-1/2} = D_c t^{-1/2}.$$

Both orientations have the same class constants, so summing the four terms gives the loss asymptotic. □

A fixed positive class radius satisfies the standing assumptions. So does a nondegenerate isotropic Gaussian population within a hyperplane: its radius has Gaussian tails, and its inverse first moment is finite in dimension $d - 1 \geq 2$. The preceding formulas therefore apply to both examples with the same exponents, with class dependence entering only through C_c and the group proportions.

The following statement collects the conclusions in their logical order.

Theorem C.9. *Assume the spherical hyperplane model, (91), nonzero balanced initialization, and (93). Fix compact generic direction sets and the smallness condition (99). Then every Clarke trajectory is global, and at $t_1 = T_{\mathrm{align}}$ all neurons are exactly at their assigned normals. Every subsequent continuation lies in U and its assigned cones.*

If $\theta = \pi/2$, the directions remain normal from t_1 onward, without an occupancy assumption. If all four groups are populated, the loss tends to zero and all four scores diverge for every $0 < \theta \leq \pi/2$. Every neuron is then permanently normal after a finite time, and the scalar equations and asymptotics (128)–(131) hold. For both orientations the limiting normalized measure is (121) and the residual ratio tends to one.

Proof. Global continuation and balancedness follow from Theorem A.1 and Proposition 3. Theorem C.3 gives simultaneous normal entry. The strong invariance theorem and Corollary C.3 give invariance and, under occupancy, fitting. For $0 < \theta < \pi/2$, the margin barrier, tail convergence, and realignment theorem give the remaining claims, followed by the scalar rate theorem. For $\theta = \pi/2$, invariance already gives permanent alignment. The scalar mass asymptotics make the positive-family fraction converge to $B_+^{1/2}/(B_+^{1/2} + B_-^{1/2}) = \omega_*$, and the residual asymptotics give ratio one. □

Scope and relation to the main-body theorem. The theorem proves more than asymptotic directional recovery: every neuron returns to its assigned normal in finite time and stays there thereafter. Its norm grows at rate $t^{1/4}$ and the loss has rate $\Theta(t^{-1/2})$, yielding both norm divergence and the loss rate asserted in the main body. These conclusions concern exact population Clarke trajectories from exact normal entry; no discrete-time or approximate-entry assertion is used.

D Extended related work

Early feature selection under small initialization. In the vanishing-initialization limit, gradient descent on two-layer ReLU networks *quantizes* first-layer features: neuron directions concentrate on a finite set of data-dependent directions (Maennel et al., 2018). Zhou et al. (2022) characterize this condensation during the initial stage of training and bound the number of orientations that can emerge, and Boursier & Flammarion (2025) quantify early alignment toward extremal directions of an effective neural-correlation objective. Kumar & Haupt (2024) show that, near sufficiently small initializations, normalized trajectories of two-homogeneous networks approach KKT points of a neural correlation function, and Kumar & Haupt (2025a) follow the dynamics beyond the origin to the first saddle encountered after this escape. These results identify which directions are selected early, but not whether they persist through fitting.

Early versus late features. Several works connect early alignment to later training. On orthogonally separable data, activation regions eventually stabilize, after which the network decomposes into effectively linear class-specific subnetworks whose directions refine toward class-wise max-margin solutions (Phuong & Lampert, 2021). For well-separated data, small-initialization gradient flow drives neurons into class-dependent activation cones or a dead region in bounded time; these regions remain invariant while the loss vanishes and the first-layer weights become approximately low-rank, although the early direction is not identified with the final one (Min et al., 2024). For XOR, Glasgow (2024) prove a find-then-tune mechanism: nearly decoupled neurons first discover informative features, which the later dynamics maintain, balance, and selectively amplify. Early and late features can also differ. On orthogonally separable data, early alignment can be followed by convergence to class-wise max-margin directions with neural-collapse structure (Min et al., 2025); neurons can rotate away from early extremal directions and, in regression, reach spurious stationary points (Boursier & Flammarion, 2025); and intermediate representations can retain information that

the later, stronger simplicity bias discards (Tsoy & Konstantinov, 2024). Sequential feature acquisition is described most completely under stronger algebraic structure: saddle-to-saddle dynamics ending at a minimum-variation-norm interpolant for shallow ReLU networks on orthogonal inputs under square loss (Boursier et al., 2022), a proposed saddle-to-saddle chain for deep linear networks with its first transition proved (Jacot et al., 2021), and incremental coordinate entry for diagonal linear networks (Berthier, 2023).

Gaussian-mixture and clustered data. A separate literature obtains explicit dynamics from probabilistic structure. In high-dimensional limits the dynamics close over finitely many order parameters: dynamical mean-field theory describes SGD for a single-layer Gaussian-mixture classifier (Mignacco et al., 2020); closed ODEs for weight–mixture overlaps exhibit a finite-width two-layer network exploiting information unavailable to its kernel limit (Refinetti et al., 2021); and rigorous ODE and SDE scaling limits for summary statistics of SGD cover two-layer networks on binary and XOR mixtures (Ben Arous et al., 2022). Demir & Dogan (2025) isolate the effect of a single gradient step on structured mixtures in a proportional regime, showing that it can introduce higher-order polynomial features. Closer to neuron alignment, Min & Vidal (2025) prove, for polynomial-ReLU networks on mixtures with orthonormal centers, an alignment phase toward individual cluster centers followed by a fitting phase that preserves this specialization and approaches a robust classifier. Yeom et al. (2026) show, for logistic gradient flow on a binary Gaussian mixture, that the initialization scale trades off over-alignment with noisy empirical class statistics against overfitting. Abedsoltan et al. (2026) show that neurons of a two-layer ReLU network can become monosemantic, specializing to cluster-specific predictive directions in clustered regression, whereas on orthogonal clusters gradient descent can instead average class-relevant cluster features (Li et al., 2025). In all of these settings, features are described through overlaps with mixture statistics or attraction to cluster-specific directions, rather than through the geometry of the class supports.

Relationship to our setting. Our early phase shares with Maennel et al. (2018), Kumar & Haupt (2024), and Boursier & Flammarion (2025) an approximately decoupled directional flow, but the selected directions here are interpretable as normals to class supports. Unlike the persistence results of Phuong & Lampert (2021) and Min et al. (2024), a neuron may leave its selected feature during fitting; unlike the results of Min et al. (2025), Boursier & Flammarion (2025), and Tsoy & Konstantinov (2024), it then returns to it, so the early and limiting features coincide even though the intermediate representation differs.

The closest analyses of comparable non-linearly-separable geometry are the XOR studies of Tsoy & Konstantinov (2024) and Glasgow (2024), which are orthogonal ($\theta = \pi/2$) instances of our two-line and two-hyperplane geometries. There the opposite-class normals coincide with the signal directions, which hides the normal geometry and removes the effects of subspace angle and imbalance. Tsoy & Konstantinov (2024) study discrete, well-separated XOR clusters with a smooth homogeneous approximation to ReLU, and their late-phase result assumes directional convergence of an auxiliary four-neuron network; Glasgow (2024) studies a symmetric hyperplane instance with four equally weighted Boolean clusters and a fixed nonzero margin. Our line model allows arbitrary finite atomic distributions on the lines, our hyperplane model covers continuous rotationally symmetric populations, including spherical and Gaussian distributions, and together they expose the roles of subspace angle, within-subspace distribution, and class balance. Our alignment and cone-invariance results impose no width lower bound; four-group coverage enters only when all selected normals must participate in the fitting dynamics.

E Additional experimental details and results

This section provides the full experimental protocol underlying Section 6, together with additional results omitted from the main text for readability. All experiments use the same training code and differ only in the data geometry, sample size, or off-subspace perturbation.

Model and optimization. We train a width-$m = 32$ two-layer ReLU network without biases,

$$f(x; W, v) = \sum_{j=1}^{m} v_j \sigma(\langle w_j, x \rangle).$$

The first-layer weights are initialized independently as

$$w_j(0) \sim \mathcal{N}\left(0, \frac{\varepsilon^2}{d} I_d\right), \qquad \varepsilon = 10^{-5},$$

and independent signs $s_j \in \{-1, +1\}$ are sampled uniformly. We set

$$v_j(0) = s_j \|w_j(0)\|,$$

so that every neuron is exactly balanced at initialization. We optimize the empirical exponential loss

$$\widehat{\mathcal{L}}(W, v) = \frac{1}{N} \sum_{i=1}^{N} \exp\big(-y_i f(x_i; W, v)\big)$$

using full-batch gradient descent with step size $\eta = 10^{-2}$, updating W and v simultaneously and taking $\sigma'(0) = 0$. Every run is trained to time $T = 10^4$, corresponding to 10^6 gradient steps. Exponents are clipped to $[-80, 80]$ for numerical stability, and all computations use 64-bit floating point arithmetic.

Snapshots are recorded every 50 steps, or 0.5 units of training time. We use the same initialization across experimental conditions within each seed, while the seed also determines the training samples. Results are aggregated over five seeds and are reported as the median with interquartile range.

Data geometry. The ambient dimension is $d = 5$ and the nontrivial principal angle between the two class subspaces is $\theta = 30^\circ$. Within a k-dimensional support S_c with orthonormal basis B_c, clean samples are generated as

$$x = B_c z, \qquad z \sim \mathcal{N}(0, I_k).$$

Thus every direction inside the support has unit variance. We pair every sample with its reflection $-x$, so that each empirical class distribution is centrally symmetric.

For the two-line model,

$$S_+ = \operatorname{span}(u_+), \qquad S_- = \operatorname{span}(u_-), \qquad \langle u_+, u_- \rangle = \cos\theta,$$

with 512 samples per class. In addition to Gaussian radii, we consider a bounded-radius variant in which $|g|$ is sampled uniformly from $[0.1, 5]$ and $x = g u_c$, again together with its reflection.

For the hyperplane model, the two $(d-1)$-dimensional supports share a $(d-2)$-dimensional intersection. In $d = 5$ we take

$$K = \operatorname{span}(e_1, e_2, e_3),$$

and let the remaining directions $u_+ \in S_+$ and $u_- \in S_-$ form angle θ in $\operatorname{span}(e_4, e_5)$. We use $4{,}096$ and $8{,}192$ samples per class. To approximate the corresponding population dynamics, we also use a scrambled-Sobol Gaussian quadrature consisting of $4{,}096$ points and their reflections per class.

To test geometries outside the two cases analyzed theoretically, we also consider pairs of dimensions (k_+, k_-). The two supports share $\min\{k_+, k_-\} - 1$ directions, differ by the prescribed angle θ in one two-dimensional plane, and, when $k_+ \neq k_-$, the larger support has the remaining private directions. All such experiments use $4{,}096$ samples per class.

Off-subspace perturbations. For a clean sample $x \in S_c$, the noisy experiments replace it by

$$x + \sigma_{\text{noise}} \boldsymbol{\eta}_{\text{noise}}, \qquad \boldsymbol{\eta}_{\text{noise}} \sim \mathcal{N}(0, \Pi_{S_c^\perp}),$$

independently before reflection. Hence the noise is isotropic in the orthogonal complement of the class support. We consider

$$\sigma_{\text{noise}} \in \{10^{-4}, 10^{-3}, 10^{-2}, 10^{-1}\}.$$

For hyperplanes this perturbation is one-dimensional, whereas for lines in $d = 5$ it acts in the four-dimensional orthogonal complement.

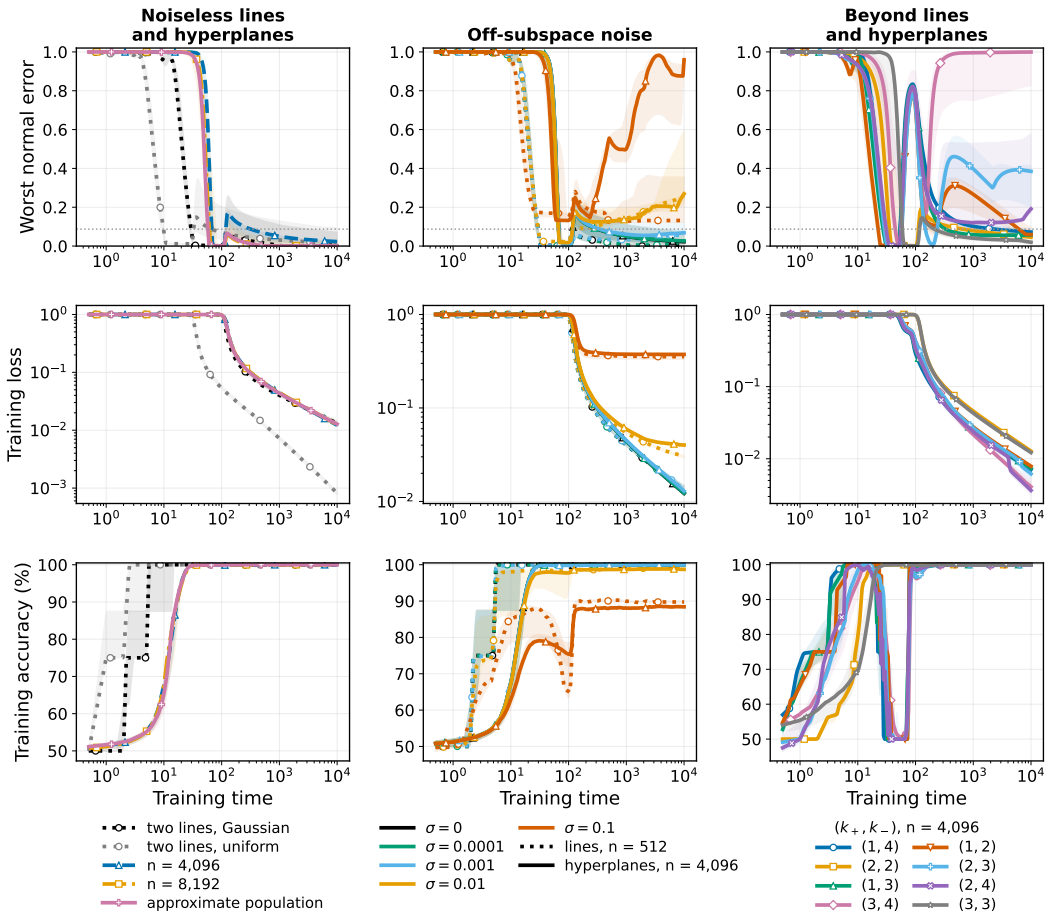


Figure 4: **Complete experimental sweep.** Worst normal error (top), exponential training loss (middle), and training accuracy (bottom) for the noiseless settings (left), off-subspace perturbations (middle), and all tested subspace dimensions (right). Unlike the main-text figure, this plot includes the largest noise level and the complete dimension sweep. Curves show medians over five seeds with interquartile ranges.

Recorded quantities. In addition to training classification error, we record the exponential loss and the geometry of every first-layer neuron. For a neuron whose output sign assigns it to the opposite-class support S_{opp}, we define its normal error by

$$e_j(t) := \left\| \Pi_{S_{\mathrm{opp}}} \widehat{\boldsymbol{w}}_j(t) \right\|.$$

This quantity is the sine of the angle between $\widehat{\boldsymbol{w}}_j$ and the normal space $S_{\mathrm{opp}}^{\perp}$, and vanishes exactly when the learned feature is orthogonal to the opposite-class support. We report the worst error $\max_j e_j(t)$. In the two-line experiments, components orthogonal to the data plane are dynamically inert, so the error is evaluated using the normalized projection of w_j onto $\mathrm{span}(u_+, u_-)$.

We also monitor the four residual moments described below. Finally, for comparisons across geometries we define the fitting-onset time

$$T_{\mathrm{fit}} := \inf \left\{ t : \widehat{\mathcal{L}}(t) \le 0.95\, \widehat{\mathcal{L}}(0) \right\}.$$

This normalization is used only in the residual-ratio plots below.

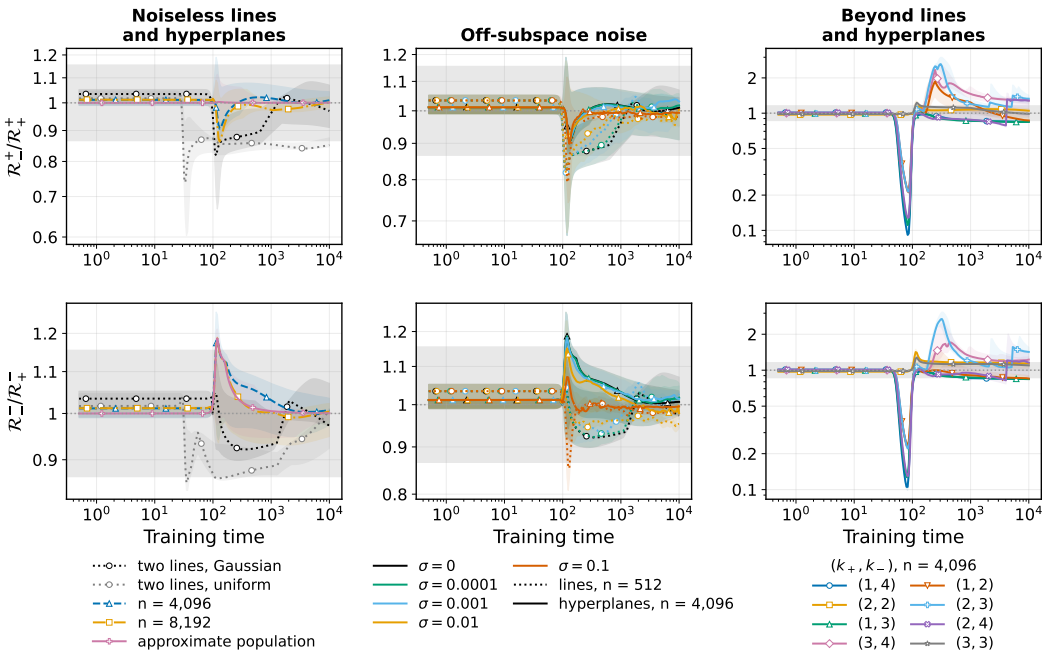


Figure 5: **Residual-moment ratios during fitting.** Ratios $\widehat{\mathcal{R}}_-^\tau/\widehat{\mathcal{R}}_+^\tau$ for $\tau = +1$ (top) and $\tau = -1$ (bottom), plotted against time normalized by the fitting onset T_{fit}. The shaded band denotes $(\cos\theta, 1/\cos\theta)$, the normal-attraction window appearing in the line and hyperplane analyses. Curves show medians over five seeds with interquartile ranges.

E.1 Full alignment, loss, and classification results

Figure 4 reports the complete sweep underlying the main-text experiment. In contrast to Figure 3, we retain the largest noise level $\sigma_{\text{noise}} = 10^{-1}$, all tested dimension pairs, and the classification-accuracy curves.

The additional cases clarify the limits of the phenomenon. In the clean line and hyperplane models, early normal alignment precedes the substantial decrease in loss, followed by departure around the onset of fitting and subsequent realignment. Small off-subspace perturbations preserve this qualitative trajectory, whereas the approximation degrades progressively as σ_{noise} increases. At $\sigma_{\text{noise}} = 10^{-1}$, normal alignment can break down substantially and the loss and classification performance also deteriorate, showing that the behavior is not insensitive to arbitrary violations of the subspace model.

The complete dimension sweep is similarly nonuniform. Several equal- and unequal-dimensional pairs exhibit the same align–depart–realign pattern as the two endpoint geometries, while more asymmetric cases show weaker or incomplete late realignment. Notably, successful fitting can still occur in some of these cases despite substantial final normal error. Thus normal-space feature recovery is a stronger property than eventual classification on the training sample.

E.2 Residual balance during fitting

The early analysis is governed by unweighted class moments, but once the network output becomes non-negligible the directional force is reweighted by the exponential residuals. To track this effect, let u_c denote the distinguished class direction in the principal two-dimensional plane and set $Z_i = \langle x_i, u_c \rangle$ for $y_i = c$. For each orientation $\tau \in \{-1, +1\}$ we record the empirical residual moment

$$\widehat{\mathcal{R}}_c^\tau(t) := \frac{1}{n_c} \sum_{\substack{i: y_i = c \\ \tau Z_i > 0}} |Z_i| \, e^{-y_i f(x_i;t)}.$$

These empirical moments use class-conditional normalization $1/n_c$; the joint-law convention is $(n_c/N)\widehat{\mathcal{R}}_c^\tau$. With equal class counts, the common factor cancels in the ratios below. We plot the ratios

$$\rho_\tau(t) := \frac{\widehat{\mathcal{R}}_-^\tau(t)}{\widehat{\mathcal{R}}_+^\tau(t)}, \qquad \tau \in \{-1, +1\},$$

against normalized time t/T_{fit}.

For the line and hyperplane geometries, the theoretical normal-attraction condition identifies the interval

$$\cos\theta < \rho_\tau < \frac{1}{\cos\theta}.$$

The gray region in Figure 5 denotes this window. Because the intermediate-dimensional experiments fall outside the settings for which this condition is proved, the same window is shown there only as a diagnostic rather than as a theoretical prediction.

Before fitting, the residual weights are close to one and all ratios remain near their initially balanced value. The largest excursions occur around $t \approx T_{\mathrm{fit}}$, precisely when the neuron directions begin to depart from the normal spaces. In the clean line and hyperplane experiments, the ratios subsequently move back toward one as fitting proceeds, broadly tracking the observed realignment.

The noisy experiments also show that residual balance is not by itself sufficient for normal recovery away from exact subspace support: for small and moderate noise the ratios can remain close to the balance window even while the normal error develops a nonzero late-time floor. In the intermediate-dimensional experiments, some of the more asymmetric pairs exhibit much larger transient residual imbalance around fitting onset, consistent with their stronger directional departures. We emphasize, however, that the balance window is only a sufficient normal-attraction criterion in the analyzed geometries and should not be interpreted as a necessary condition for eventual realignment in the broader experiments.

E.3 Quantitative early alignment time validation

We next test the quantitative predictions of the early alignment analysis in the two-line model. For each neuron that reaches its predicted opposite-class normal, let T_j^{hit} denote its first hitting time, and define

$$T_{\mathrm{align}} := \max_j T_j^{\mathrm{hit}}.$$

We vary the principal angle θ and consider both Gaussian radial laws and bounded uniform radii, in ambient dimensions $d = 2$ and $d = 5$. For the alignment-time experiments we use a very small initialization $\varepsilon = 10^{-12}$, so that the dynamics remain close to the idealized early field over the entire alignment interval.

Figure 6 compares the observed hitting times with the corresponding predictions of the idealized dynamics. Panel (a) shows the worst-neuron alignment time as a function of θ, together with the ideal worst-case time and the theoretical upper bounds. The measured times follow the same strong angle dependence predicted by the theory and remain below the alignment-time bound across the tested radial laws and ambient dimensions. In particular, alignment becomes progressively slower as the two class subspaces approach one another.

Panel (b) makes the comparison neuron-by-neuron in $d = 2$. The measured hitting times closely track those obtained by integrating the idealized field from the same initial directions, over several orders of magnitude in time. This provides a more stringent check than the worst-case comparison in panel (a): the ideal early dynamics predict not only the overall alignment scale but also the substantial variation in hitting times across individual neurons.

Finally, panel (c) compares alignment with the onset of fitting at the initialization scale $\varepsilon = 10^{-5}$ used in our main experiments. We plot $T_{\mathrm{fit}}/T_{\mathrm{align}}$, with T_{fit} defined as in Section 6. Whenever all neurons align, this ratio is larger than one and increases markedly with θ, showing that normal selection typically occurs before a substantial decrease in the loss. At the smallest angles, some neurons fail to reach a normal, consistent with the weakening of the angular attraction as the subspaces become nearly parallel.

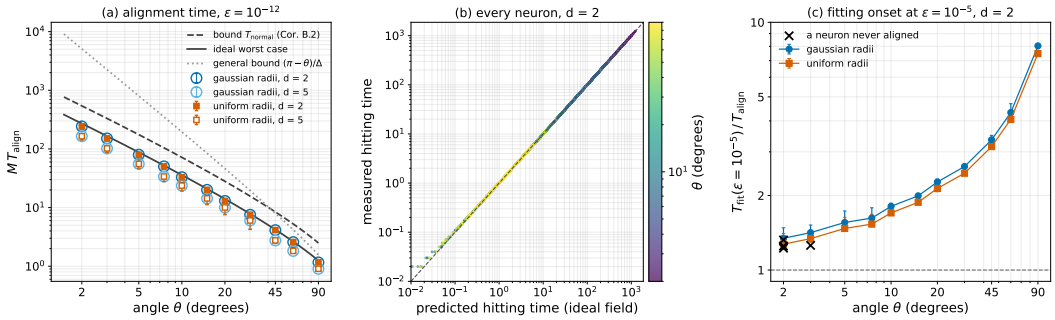


Figure 6: **Quantitative validation of early alignment.** (a) Worst-neuron alignment time versus principal angle, compared with the ideal-field prediction and theoretical upper bounds. (b) Measured versus ideal-field hitting time for individual neurons in $d = 2$. (c) Ratio of fitting onset to alignment time at $\varepsilon = 10^{-5}$; crosses indicate runs in which at least one neuron did not align.

E.4 Dependence on the subspace angle

We next examine how the alignment dynamics depend on the principal angle between the class subspaces. We repeat the clean hyperplane experiment for

$$\theta \in \{10^\circ, 30^\circ, 60^\circ\},$$

keeping the remaining training setup fixed. For each angle we record the worst-neuron normal error, the training loss, and the two residual-moment ratios

$$\rho_\tau(t) := \frac{\widehat{\mathcal{R}}_-^\tau(t)}{\widehat{\mathcal{R}}_+^\tau(t)}, \qquad \tau \in \{-1, +1\}.$$

The corresponding normal-attraction window is

$$\cos\theta < \rho_\tau < \frac{1}{\cos\theta}.$$

Thus the admissible interval narrows sharply as the two subspaces become nearly parallel: it is approximately

$$(0.985, 1.015), \qquad (0.866, 1.155), \qquad (0.500, 2.000)$$

for $\theta = 10^\circ, 30^\circ, 60^\circ$, respectively.

Figure 7 shows that the qualitative alignment dynamics become increasingly robust as the subspace angle grows. Figure 8 shows the individual seed trajectories. Before fitting, neurons approach the opposite-class normal spaces in all three regimes. Around T_{fit}, the residual ratios undergo their largest excursions and the neuron directions depart from the normals. The subsequent behavior depends strongly on the subspace angle. At $\theta = 60^\circ$, the wide balance window accommodates these residual fluctuations and every displayed seed rapidly returns to essentially exact normal alignment. At $\theta = 30^\circ$, realignment is slower and some worst neurons retain a visible error for a substantial portion of training. At $\theta = 10^\circ$, the balance window is extremely narrow and the seedwise behavior becomes markedly less uniform: although many neurons align, some worst-neuron trajectories remain far from a normal throughout the observed training horizon.

The distinction between the worst and median errors is useful here: in several runs the median neuron reaches a normal while the worst neuron does not, so averaging over neurons or seeds would hide incomplete realignment. These observations do not contradict Theorem C.9, which guarantees permanent realignment for every $\theta \in (0, \pi/2]$ in the population setting, but three features of the finite experiment can delay realignment beyond the observed horizon. First, at $\theta = 10^\circ$ the balance window has half-width of about 1.5%, comparable to the sampling fluctuation of the empirical moment ratio at the sample sizes used here, so the balance condition may fail for the empirical distribution in some seeds. Second, realignment is driven by the residual clock $\int \mathcal{R}_c^\tau \, dt$, which grows only logarithmically in time, while the clock length required by Theorem C.7 increases as θ decreases. The sweep therefore shows how quickly realignment degrades as the supports approach one another.

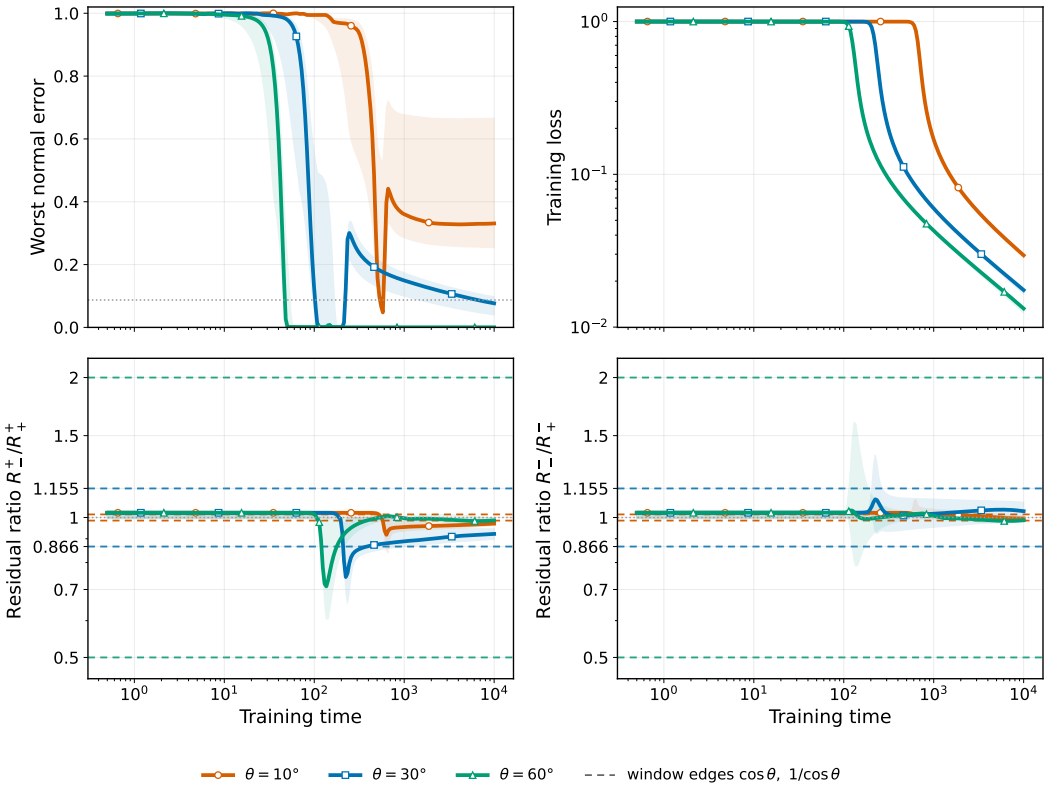


Figure 7: **Effect of subspace separation on alignment.** Worst normal error, training loss, and residual-moment ratios for $\theta \in \{10°, 30°, 60°\}$. Dashed horizontal lines mark the angle-dependent balance-window boundaries $\cos\theta$ and $1/\cos\theta$. As θ decreases, the window narrows and comparable residual fluctuations are accompanied by progressively weaker late realignment.

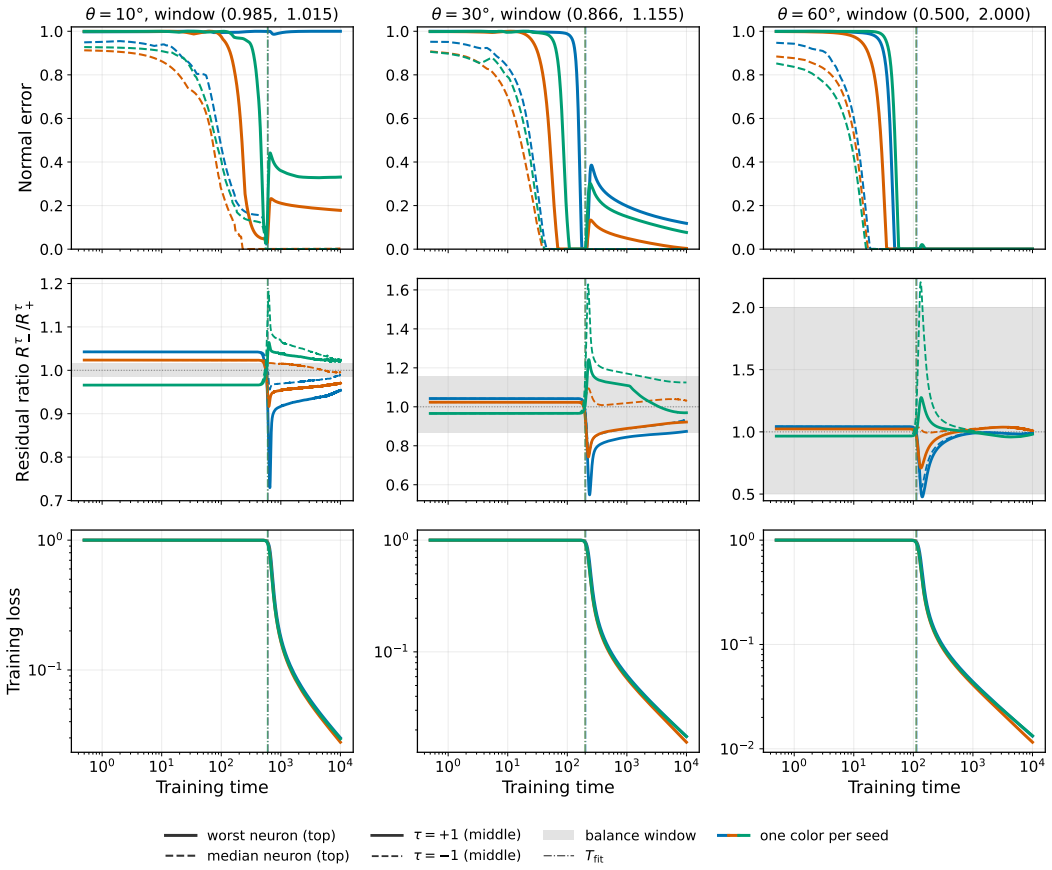


Figure 8: **Angle dependence and seedwise alignment.** Individual-seed trajectories for $\theta \in \{10°, 30°, 60°\}$. Top: worst (solid) and median (dashed) neuron normal error. Middle: residual-moment ratios for $\tau = +1$ (solid) and $\tau = -1$ (dashed), with the corresponding balance window shaded. Bottom: training loss; dash-dotted vertical lines mark T_{fit}.