ALIGNMENT AND CONVERGENCE ANALYSES OF GRADIENT FLOW IN TWO-LAYER CONVOLUTIONAL NEURAL NETWORKS

Anonymous authors
Paper under double-blind review

ABSTRACT

Understanding why gradient-based training of neural networks learns to extract signal and ignore noise remains a central open problem in deep learning theory. We take a step toward this goal by analyzing the gradient-flow dynamics of two-layer ReLU convolutional networks with sum-pooling, trained on the exponential or logistic loss for binary classification from a small, balanced initialization. Each input consists of one signal patch and some noise patches; the signal patches of the two classes lie in two well-separated, nearly antipodal cones, so that the label-signed signal patches share a common axis $\boldsymbol{u}_\star$, and the noise patches are orthogonal to the signal subspace. Our main result indicates that during the early phase of the training, every filter that activates a signal patch of its own class at initialization aligns with either $\boldsymbol{u}_\star$ or $-\boldsymbol{u}_\star$, depending on the sign of its second-layer weight, to a prescribed cosine $\boldsymbol{c}^\star$ and within an explicit time determined by the signal geometry; such a filter thereafter activates every signal patch of its class and none of the other class. The alignment is driven by a self-reinforcing feedback mechanism. Since same-class signal patches lie in a narrow cone, every activated signal patch pushes the filter toward all the others, thus toward the axis. The noise patches, being orthogonal to the signal subspace, enter the analysis only through a single scalar that bounds their aggregate effect. In the second phase of the training ,the projection of the aligned filter onto $\pm\boldsymbol{u}_\star$ grows monotonically, and the training loss decays at rate $\mathcal{O}(1/t)$. Experiments on synthetic and MNIST data confirm the two phases.

1 INTRODUCTION

Neural networks trained by gradient descent from a small random initialization, without explicit regularization, routinely generalize well despite being heavily overparametrized. A leading explanation is the implicit bias of the training algorithm: first-order methods started near the origin favor sparse or low-rank solutions, a phenomenon established for matrix factorization (Gunasekar et al., 2017; Li et al., 2018; Arora et al., 2019; Stöger & Soltanolkotabi, 2021) and for linear networks, whose weights are learned one singular direction at a time (Saxe et al., 2014; Gidel et al., 2019; Jacot et al., 2021; Pesme & Flammarion, 2023). For a network of fixed width, the scale of the initialization controls the regime of training: a large initialization places it in the kernel regime, where the parameters barely move and the network behaves as its linearization, and a small one in the rich regime, where the features change substantially and the implicit bias departs from that of the kernel (Chizat et al., 2019; Woodworth et al., 2020).

For two-layer ReLU networks the rich regime has a concrete geometric description. Maennel et al. (2018) observed that, in the limit of a vanishing initialization, first-layer neurons first rotate toward one of finitely many directions determined by the data while their norms stay small, and only afterwards grow to fit the labels. This "align, then fit" picture underlies a line of results on shallow ReLU networks: on orthogonally separable data the neurons converge to the positive or the negative data center and the first-layer weights become low-rank (Phuong & Lampert, 2021); on orthogonal inputs the whole trajectory can be described (Boursier et al., 2022), and on correlated inputs generated by a single neuron as well (Chistikov et al., 2023); on two nearly parallel data points that are linearly but not orthogonally separable, the trajectory passes through four phases in which the

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.

activation pattern first simplifies and then grows again (Wang & Ma, 2023); the alignment phase explains the simplicity bias of the learned classifier (Lyu et al., 2021; Tsoy & Konstantinov, 2024) and, through the margin-maximization results for homogeneous networks (Lyu & Li, 2020; Ji & Telgarsky, 2020; Chizat & Bach, 2020), the solution reached at the end of training. Min et al. (2024) made the early phase quantitative for a finite initialization scale ϵ: the directional dynamics of every neuron follow, up to a controlled error, the signed sum of the inputs it activates during a window of length $\Theta(\frac{1}{n}\log\frac{1}{\epsilon})$, and on this window a neuron can only gain activations on inputs of its own class and lose activations on inputs of the other class, so that alignment is reached in time $O(\log n/\sqrt{\mu})$, where μ measures the separation of the data. Related alignment mechanisms have been analyzed at non-small initialization (Wang & Ma, 2022; Frei et al., 2023; Kou et al., 2023b) and for general homogeneous networks (Kumar & Haupt, 2024; 2025; Bantzis et al., 2026). All of these results concern fully connected networks, in which a neuron sees each input as a single vector.

On the other hand, the theory of convolutional networks has developed along a different path. For linear convolutional networks, gradient descent is biased toward sparsity in the frequency domain (Gunasekar et al., 2018), the geometry of the function space is understood (Kohn et al., 2022), and convergence of the gradient flow has recently been established (Diederen et al., 2026); hierarchical tensor factorizations give a related picture for deep convolutional architectures (Razin et al., 2022). For nonlinear one-hidden-layer convolutional networks, early results establish convergence of gradient descent for a single filter on Gaussian inputs (Brutzkus & Globerson, 2017; Du et al., 2018). The most developed line studies two-layer convolutional networks with sum-pooling on patch data, in which each input carries one label-dependent signal patch and otherwise noise patches: with polynomial ReLU activations (Cao et al., 2022), with ReLU activations and label noise (Kou et al., 2023a), with XOR-type signals (Meng et al., 2024), with several signal "views" (Allen-Zhu & Li, 2023), and under momentum or adaptive methods (Jelassi & Li, 2022; Zou et al., 2023). These works decompose every filter into a signal coefficient and noise coefficients and track these scalars over the entire trajectory, which determines whether the trained network generalizes. What the current literature do not provide, however, is a description of the filter directions: whether, when, and toward which direction the filters align before their norms grow, and which filters take part. In the fully connected case, this is precisely the content of the early alignment phase; for convolutional networks it has not been well-understood.

The quantitative theory of small initialization rests on a two-step structure. Near the origin the loss is approximately linear in the network output, so for a window of time of order $\log(1/\epsilon)$ the weights stay small while their directions follow a flow that depends only on the data (Maennel et al., 2018; Min et al., 2024; Kumar & Haupt, 2024); once the directions have settled, the norms grow and the trajectory passes through a sequence of saddles at which the effective rank of the weights increases (Jacot et al., 2021; Pesme & Flammarion, 2023; Bantzis et al., 2026). Turning the first step into a theorem requires two estimates: the error of the directional approximation over the whole window, and the time the directions need to converge, which must be shorter than the window. For fully connected networks both are available (Min et al., 2024). The resulting alignment is not unconditionally beneficial, since it produces sparse features but can also trap the gradient flow at spurious stationary points (Boursier & Flammarion, 2025). Carrying this program to convolutional networks meets three obstacles. First, the object that drives a filter changes: it is the signed sum of the patches the filter activates, typically a few signal patches together with many noise patches. Noise patches carry no label information, so the pairwise-correlation condition on which the fully connected analysis relies, and which ensures that every change of the activation pattern goes in the right direction, fails once patches are the unit of activation. Second, the noise has to be controlled uniformly over all filter directions, because the filters are data-dependent; the resulting bound is linear in P, since the $P-1$ noise patches of one input share its label and add coherently. Third, sum-pooling shortens the window: each filter contributes P terms to the output, the directional approximation is valid for a window shorter by a factor of P, and the initialization scale has to compensate for this.

Contributions. We provide a complete, non-asymptotic analysis of the gradient flow for a two-layer ReLU convolutional network with sum-pooling trained on patch data for binary classification with a balanced initialization of small scale. Our data model is the signal–noise patch model similar to the setting of Cao et al. (2022); Kou et al. (2023a), with the distributional assumptions on the signal replaced by a geometric one: the signal patches of the two classes lie in two narrow, nearly antipodal cones. Under this assumption the label-signed signal patches lie within a fixed angle of

a single direction $\boldsymbol{u}_\star$, the bisector of the two class axes, and the analysis tracks every filter through one scalar, its cosine with $\boldsymbol{u}_\star$.

- *Early alignment* (Theorem 1, Part 1). For an explicit upper bound on ϵ, every filter that activates a signal patch of its own class at initialization aligns with $\pm\boldsymbol{u}_\star$ to a prescribed cosine within an explicit time t_1, and from then on activates every signal patch of its class and none of the other class. The time t_1 depends on the cone geometry, the prescribed cosine, the initialization margin and the class signal masses, and contains no logarithmic factor in the number of samples.
- *Mechanism.* Two facts drive the proof. The set of signal patches a filter activates can change in one direction only, although the filter may activate and deactivate noise patches freely; and the rate at which a filter turns toward $\pm\boldsymbol{u}_\star$ admits an explicit lower bound in every region of the sphere, including the directions nearly orthogonal to $\boldsymbol{u}_\star$, where a filter may activate only a few of the signal patches, possibly a single one. A potential function that rescales time by the available drift then yields t_1 directly, without counting activation transitions.
- *Noise.* The noise enters the analysis through a single deterministic quantity, G, the largest norm of a signed sum of noise patches that one direction can activate, and Theorem 1 holds for any noise, random or not, for which G satisfies Assumption 4. Independent sub-Gaussian noise is one sufficient way to meet this condition: a uniform argument over all activation patterns bounds G by a quantity that grows linearly in P up to a logarithmic factor and as $\sqrt{n}$ in the sample size.
- *Final convergence* (Theorem 1, Part 2). After alignment the filters never leave their signal cone, the projection of each filter onto $\pm\boldsymbol{u}_\star$ grows monotonically, and the loss decays at rate $O(1/t)$ after an explicit transition time.

Experiments on synthetic patch data and on MNIST confirm that alignment precedes fitting and identify the initialization scale and noise level at which it degrades (Section 4).

2 Preliminaries

2.1 Problem setting

An input datum is a collection of P patches $\boldsymbol{X} = [\boldsymbol{x}^{(1)}, \dots, \boldsymbol{x}^{(P)}] \in \mathbb{R}^{D\times P}$. We consider a two-layer ReLU convolutional network with a single shared filter bank $W = [\boldsymbol{w}_1, \dots, \boldsymbol{w}_h] \in \mathbb{R}^{D\times h}$, second-layer weights $\boldsymbol{v} = [v_1, \dots, v_h]^\top \in \mathbb{R}^h$, and sum-pooling:

$$f(\boldsymbol{X}; W, \boldsymbol{v}) \;=\; \sum_{j=1}^{h} v_j \sum_{p=1}^{P} \sigma\left(\langle \boldsymbol{w}_j, \boldsymbol{x}^{(p)}\rangle\right), \qquad \sigma(z) = \max\{z, 0\}. \tag{1}$$

Given a training set $\{(\boldsymbol{X}_i, y_i)\}_{i=1}^n$ with $y_i \in \{\pm 1\}$, we train with the exponential loss $\ell(y, \hat{y}) = \exp(-y\hat{y})$ or the logistic loss $\ell(y, \hat{y}) = 2\log(1 + \exp(-y\hat{y}))$ on $L(W, \boldsymbol{v}) = \sum_{i=1}^n \ell(y_i, f(\boldsymbol{X}_i; W, \boldsymbol{v}))$, via the gradient flow (GF)

$$\dot{W} = -\nabla_W L(W, \boldsymbol{v}), \qquad \dot{\boldsymbol{v}} = -\nabla_{\boldsymbol{v}} L(W, \boldsymbol{v}), \tag{2}$$

where the ReLU subgradient is fixed at $\sigma'(z) = \mathbf{1}_{z>0}$. Here, the factor of two in the logistic loss is introduced to facilitate a unified analysis of the two loss functions under a common notation. Since σ' is discontinuous, (2) is understood in the sense of Carathéodory: a solution is an absolutely continuous trajectory satisfying (2) for almost every t, and all our statements concern an arbitrary such solution. Throughout, $\mathcal{I}_+ = \{i : y_i = +1\}$, $\mathcal{I}_- = \{i : y_i = -1\}$, and $X_{\max} := \max_{i\in[n],\, p\in[P]} \left\|\boldsymbol{x}_i^{(p)}\right\|$.

Balanced small initialization. Given a matrix $W_0 \in \mathbb{R}^{D\times h}$ with nonzero columns and a scale $\epsilon > 0$, we initialize

$$W(0) = \epsilon W_0, \qquad v_j(0) \in \{\|\boldsymbol{w}_j(0)\|, -\|\boldsymbol{w}_j(0)\|\}, \quad j \in [h], \tag{3}$$

and write $W_{\max} := \max_j \|[W_0]_{:,j}\|$. Thus W_0 fixes the initial shape of the filters, ϵ their scale, and only the sign of $v_j(0)$ remains to be chosen; we are interested in the regime where ϵ is small. The initialization is balanced, $v_j(0)^2 = \|\boldsymbol{w}_j(0)\|^2$, and the gradient flow preserves both the balance and the sign of v_j (Lemma 3), so the sign of $v_j(0)$ splits the neurons once and for all into $\mathcal{V}_+ := \{j : v_j(0) > 0\}$ and $\mathcal{V}_- := \{j : v_j(0) < 0\}$. Balancedness is a standard technical device to simplify the analysis (Phuong & Lampert, 2021; Boursier et al., 2022; Min et al., 2024); our experiments use an unbalanced Gaussian initialization of small variance.

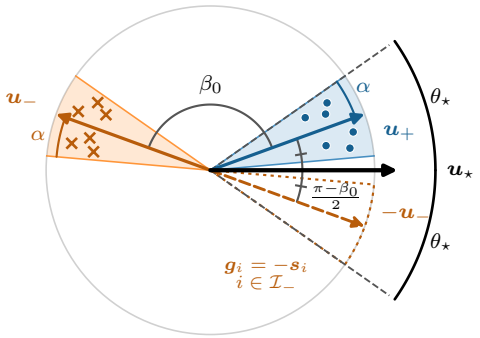

Figure 1: The two-cone signal model in the signal plane (Definition 1; angles not to scale). Positive signal patches (blue dots) lie within angle α of the class axis $\boldsymbol{u}_+$, negative ones (orange crosses) within α of $\boldsymbol{u}_-$, and the class axes make an angle $\beta_0 > \pi/2$. Signing the negative patches by their labels reflects their cone through the origin (dotted), and the common axis $\boldsymbol{u}_\star$ bisects the angle between $\boldsymbol{u}_+$ and $-\boldsymbol{u}_-$, so that every signed generator $\boldsymbol{g}_i = y_i \boldsymbol{s}_i$ lies within $\theta_\star = (\pi - \beta_0)/2 + \alpha$ of $\boldsymbol{u}_\star$ (dashed rays).

Notation. For nonzero $\boldsymbol{x}, \boldsymbol{y}$ we write $\cos(\boldsymbol{x}, \boldsymbol{y}) = \langle \boldsymbol{x}, \boldsymbol{y} \rangle / (\|\boldsymbol{x}\| \, \|\boldsymbol{y}\|)$ and $\angle(\boldsymbol{x}, \boldsymbol{y}) := \arccos \cos(\boldsymbol{x}, \boldsymbol{y}) \in [0, \pi]$; the angle is a metric on the unit sphere. For nonzero $\boldsymbol{w}$, $\mathcal{P}_{\boldsymbol{w}} = I - \boldsymbol{w}\boldsymbol{w}^\top / \|\boldsymbol{w}\|^2$ is the orthogonal projection onto $\boldsymbol{w}^\perp$. The *signed activation field* of a filter $\boldsymbol{w}$ is the signed sum of the patches it activates, $x_a(\boldsymbol{w}) := \sum_{(i,p):\, \langle \boldsymbol{w}, \boldsymbol{x}_i^{(p)} \rangle > 0} y_i \, \boldsymbol{x}_i^{(p)}$.

2.2 Data model and assumptions

Each input consists of one patch carrying the class information and $P - 1$ patches carrying noise, and the signal patches of the two classes are confined to two cones around two class axes (Figure 1).

Definition 1 (Two-cone signal–noise patch model)**.** Let $\mathcal{I}_+ := \{i \in [n] : y_i = +1\}$ and $\mathcal{I}_- := \{i \in [n] : y_i = -1\}$, both nonempty. Each input consists of one *signal* patch and $P - 1$ *noise* patches,

$$\boldsymbol{x}_i^{(1)} = \boldsymbol{s}_i \in \mathbb{R}^D \setminus \{0\}, \qquad \boldsymbol{x}_i^{(p)} = \boldsymbol{\xi}_i^{(p)}, \quad 2 \le p \le P,$$

and $V_{\text{sig}} := \operatorname{span}\{\boldsymbol{s}_i : i \in [n]\}$ denotes the signal subspace. The model is parametrized by two unit vectors $\boldsymbol{u}_+, \boldsymbol{u}_- \in V_{\text{sig}}$, the *class axes*, and a half-angle $\alpha \in [0, \pi/4)$. For $\boldsymbol{u}_+$ and $\boldsymbol{u}_-$, we can set them as $\boldsymbol{u}_+ = \sum_{i \in I+} s_i / \left\| \sum_{i \in I+} s_i \right\|, \boldsymbol{u}_- = \sum_{i \in I-} s_i / \left\| \sum_{i \in I-} s_i \right\|$.

We write $\boldsymbol{g}_i := y_i \boldsymbol{s}_i$ for the *signed generators*, $\rho_i := \|\boldsymbol{s}_i\|$ for the signal norms, and $\rho_{\min} := \min_i \rho_i$, $R_\pm := \sum_{i \in \mathcal{I}_\pm} \rho_i$, $R_{\min} := \min\{R_+, R_-\}$, $R := R_+ + R_-$. For $\mathcal{A} \subseteq [n]$ we write $R_{\mathcal{A}} := \sum_{i \in \mathcal{A}} \rho_i$ and $\boldsymbol{z}_{\mathcal{A}} := \sum_{i \in \mathcal{A}} \boldsymbol{g}_i$. The *signal cones* are

$$\mathcal{S}_+ := \left\{ \boldsymbol{w} \in \mathbb{R}^D : \langle \boldsymbol{w}, \boldsymbol{g}_i \rangle > 0 \ \ \forall i \in [n] \right\}, \qquad \mathcal{S}_- := -\mathcal{S}_+; \tag{4}$$

since $\langle \boldsymbol{w}, \boldsymbol{g}_i \rangle = y_i \langle \boldsymbol{w}, \boldsymbol{s}_i \rangle$, a filter in $\mathcal{S}_+$ activates every positive signal patch and no negative one, and a filter in $\mathcal{S}_-$ the opposite. Let $\nu_0 := \langle \boldsymbol{u}_+, \boldsymbol{u}_- \rangle$ and $\beta_0 := \arccos \nu_0$. Whenever $\boldsymbol{u}_+ \neq \boldsymbol{u}_-$, we define the *common signal axis* and the associated angle and cosines

$$\boldsymbol{u}_\star := \frac{\boldsymbol{u}_+ - \boldsymbol{u}_-}{\|\boldsymbol{u}_+ - \boldsymbol{u}_-\|} \in V_{\text{sig}}, \qquad \theta_\star := \frac{\pi - \beta_0}{2} + \alpha, \qquad \mu_\star := \cos \theta_\star, \qquad \mu := \cos 2\theta_\star. \tag{5}$$

In words, label signing folds the two classes onto one side of the sphere: the positive generators lie within α of $\boldsymbol{u}_+$, the negated negative generators within α of $-\boldsymbol{u}_-$, and these two axes are only $\pi - \beta_0$ apart, a small angle when the cones are nearly antipodal. The axis $\boldsymbol{u}_\star$ is the direction halfway between $\boldsymbol{u}_+$ and $-\boldsymbol{u}_-$, the natural common reference for the signed generators of both classes, and $\theta_\star$ is how far a signed generator can stray from it: half the gap between the two axes plus the cone half-angle. The two cosines measure the resulting coherence, $\mu_\star$ between a signed generator and $\boldsymbol{u}_\star$, and μ between two signed generators, which can be up to $2\theta_\star$ apart. A small $\theta_\star$ thus means that, once signed, all signal patches point roughly the same way, and this is what the analysis exploits.

Figure 1 shows these quantities in the signal plane; that $\theta_\star$ is indeed the largest angle between a signed generator and $\boldsymbol{u}_\star$ is Lemma 1 below.

Assumption 1 (Noise–signal orthogonality)**.** Every noise patch is orthogonal to the signal subspace: $\langle \boldsymbol{\xi}_i^{(p)}, \boldsymbol{s}_k \rangle = 0$ for all $i, k \in [n]$ and $2 \le p \le P$.

This is the deterministic core of the noise model of Cao et al. (2022); Kou et al. (2023a), where the noise is Gaussian in the orthogonal complement of the signal. It makes the signed activation field $x_a(\boldsymbol{w}) := \sum_{(i,p):\langle \boldsymbol{w}, \boldsymbol{x}_i^{(p)}\rangle>0} y_i \boldsymbol{x}_i^{(p)}$ split exactly into a signal part and a noise part (Lemma 7): the former is $\boldsymbol{z}_{\mathcal{A}(\boldsymbol{w})}$ for the set $\mathcal{A}(\boldsymbol{w}) := \{i : \langle \boldsymbol{w}, \boldsymbol{s}_i\rangle > 0\}$ of activated signal patches, and the latter is the *noise field*

$$\boldsymbol{g}(\boldsymbol{w}) := \sum_{i=1}^{n}\sum_{p=2}^{P} \mathbf{1}_{\langle \boldsymbol{w}, \boldsymbol{\xi}_i^{(p)}\rangle>0}\, y_i\, \boldsymbol{\xi}_i^{(p)} \;\in\; V_{\mathrm{sig}}^{\perp}, \tag{6}$$

and its size $G := \sup_{\boldsymbol{w}\in\mathbb{S}^{D-1}} \|\boldsymbol{g}(\boldsymbol{w})\|$. The noise thus contributes nothing along any signal direction, and enters the analysis only through G, the largest signed sum of noise patches that a single direction can activate.

Assumption 2 (Initialization)**.** Every column of W_0 is nonzero, and there is $\zeta_0 \in (0, 1]$ such that $\cos(\boldsymbol{w}_j(0), \boldsymbol{u}_\star) \geq -1 + \zeta_0$ for $j \in \mathcal{V}_+$ and $\cos(\boldsymbol{w}_j(0), -\boldsymbol{u}_\star) \geq -1 + \zeta_0$ for $j \in \mathcal{V}_-$.

A filter in $\mathcal{V}_+$ pointing exactly along $-\boldsymbol{u}_\star$ sits at a critical point of the directional dynamics, so a quantitative alignment time needs a margin ζ_0 away from it; when W_0 has a density the assumption holds with probability one. Fix a target cosine $c^* \in (\sin\theta_\star, 1)$ and set $\kappa := \min\left\{\frac{\zeta_0}{2},\ 1 - 2\sin\theta_\star,\ 1 - c^{*2}\right\}$.

Assumption 3 (Signal geometry)**.** (a) *(Two cones.)* Every signal patch lies within angle α of its class axis, and the class axes are nearly antipodal: $\cos(\boldsymbol{s}_i, \boldsymbol{u}_+) \geq \cos\alpha$ for $i \in \mathcal{I}_+$, $\cos(\boldsymbol{s}_i, \boldsymbol{u}_-) \geq \cos\alpha$ for $i \in \mathcal{I}_-$, and $\nu_0 < -\sin 2\alpha$. (b) *(Cone width.)* $\tan\theta_\star \leq \kappa/4$.

Under (a), $\nu_0 < 0$, so $\boldsymbol{u}_+ \neq \boldsymbol{u}_-$ and (5) is well defined, and the two cones collapse to a single one once the patches are signed by their labels:

Lemma 1 (Common signal axis)**.** *Under Assumption 3(a), $\theta_\star < \pi/4$, so that $\mu_\star > 1/\sqrt{2}$ and $\mu > 0$, and $\cos(\boldsymbol{g}_i, \boldsymbol{u}_\star) \geq \mu_\star$ for all $i \in [n]$.*

The proof is in Appendix C.1. The lemma says that, after label signing, all signal patches point roughly the same way: every $\boldsymbol{g}_i$ lies within $\theta_\star < \pi/4$ of $\boldsymbol{u}_\star$. Two things follow. Because any two signed generators are less than $\pi/2$ apart, they are positively correlated (Lemma 6), so the pull of one activated signal patch also increases the pre-activation of every other one. And because each $\boldsymbol{g}_i$ is nearly parallel to $\boldsymbol{u}_\star$, the total signal pull on a filter is mostly along $\boldsymbol{u}_\star$, with a sideways component of at most $\tan\theta_\star$ times the axial one; Assumption 3(b) makes this ratio small enough that the sideways pull never overturns the alignment toward $\boldsymbol{u}_\star$.

Assumption 4 (Signal-to-noise ratio)**.** $G \leq \frac{\kappa}{8} R_{\min}\mu_\star \quad$ and $\quad G\sin\theta_\star \leq \frac{\rho_{\min}}{48}$.

The two conditions compare the noise with the signal in two situations. A filter that leans clearly toward $\pm\boldsymbol{u}_\star$ sees the signal patches of a whole class (Lemma 8), so the first condition asks the noise to be a small fraction of a class mass. A filter nearly orthogonal to $\boldsymbol{u}_\star$ may see a single patch of mass $\rho_{\min}$, but there only a fraction $\sin\theta_\star$ of the noise can turn the filter toward or away from $\boldsymbol{u}_\star$; the second condition keeps that part below the pull of one patch, the price of assuming nothing about how the patches are spread inside their cones.

Assumptions 1–4 are conditions on the realized data, and so is our main theorem. Distributional assumptions enter only as a sufficient condition to bound G in Lemma 10: if the labels are independent Rademacher variables independent of the noise, and the noise patches are independent and σ_ξ-sub-Gaussian, then with probability at least $1 - \delta$

$$G \leq C\,\sigma_\xi\,(P-1)\sqrt{n\left(D\log\tfrac{5enP}{D} + \log\tfrac{2}{\delta}\right)} \tag{7}$$

for an absolute constant C, provided $n(P-1) \geq D + \log(2/\delta)$. The bound grows linearly in P up to a logarithmic factor; the linear factor comes from bounding the $P-1$ noise patches of one input, which share its label, by their number, and we do not claim it is tight. The alignment result itself holds for any noise, dependent or not, whenever the realized G satisfies Assumption 4.

3 Main Results

3.1 Early alignment and final convergence

Under a small initialization, training from (3) proceeds in two phases (Maennel et al., 2018; Boursier et al., 2022; Min et al., 2024): during an early phase every filter keeps a small norm while its direction changes substantially, and only afterwards do some filters grow their norms to fit the labels. During the early phase the outputs are small, the loss gradient $-\nabla_{\hat{y}}\ell(y_i, \hat{y}_i)$ is close to y_i, and by balancedness the direction of each filter obeys, up to a residual controlled by ϵ (Corollary 1),

$$\frac{d}{dt}\frac{\boldsymbol{w}_j}{\|\boldsymbol{w}_j\|} \;\simeq\; \operatorname{sign}(v_j(0))\,\mathcal{P}_{\boldsymbol{w}_j}\, x_a(\boldsymbol{w}_j)\,: \tag{8}$$

a filter in $\mathcal{V}_+$ is attracted by the positive patches it activates and repelled by the negative ones, and a filter in $\mathcal{V}_-$ the opposite. Our main result describes the resulting two phases: an early phase during which every filter that activates a signal patch of its own class aligns with the common axis to a prescribed cosine, and a later phase during which the aligned filters fit the data and the loss decays at rate $O(1/t)$. The alignment time and the admissible residual in (8) are

$$t_1 := \frac{4}{\kappa R_{\min}\mu_\star} + \frac{6\sin\theta_\star}{\rho_{\min}}, \qquad \varrho_1 := \min\Big\{\frac{\rho_{\min}}{48},\ \frac{\mu\rho_{\min}}{2},\ \frac{\kappa R_{\min}\mu_\star}{8}\Big\}, \tag{9}$$

and the length of the window on which (8) is valid is $T := \frac{1}{4nPX_{\max}}\log\frac{1}{\sqrt{h}\,\epsilon}$. For the second phase we also need the total patch norm of each input and the cosine of each signed generator with the common axis,

$$d_i := \sum_{p=1}^{P}\Big\|\boldsymbol{x}_i^{(p)}\Big\|, \qquad \mu_i := \cos(\boldsymbol{g}_i, \boldsymbol{u}_\star)\ \ (\geq \mu_\star \text{ by Lemma 1}), \qquad \Gamma := \max_{i\in[n]}\frac{d_i}{\rho_i\mu_i}\ \ (\geq 1). \tag{10}$$

Theorem 1. *Suppose Assumptions 1–4 hold, and let*

$$\epsilon \;\leq\; \min\Big\{\frac{1}{4\sqrt{h}\,PX_{\max}W_{\max}^2},\ \frac{\varrho_1}{4\sqrt{h}\,nP^2X_{\max}^2W_{\max}^2},\ \frac{1}{\sqrt{h}}\exp\big(-4nPX_{\max}\,t_1\big)\Big\}. \tag{11}$$

Then every solution to the gradient flow (2) starting from the initialization (3) has the following properties.

1. (Early alignment.) *We have $T \geq t_1$. Every neuron $j \in \mathcal{V}_+$ whose filter activates some positive signal patch at initialization, $\max_{i\in\mathcal{I}_+}\langle \boldsymbol{w}_j(0),\, \boldsymbol{s}_i\rangle > 0$, satisfies*

   $$\cos(\boldsymbol{w}_j(t), \boldsymbol{u}_\star) \geq c^* \qquad \text{for all } t \in [t_1, T],$$

   and in particular $\boldsymbol{w}_j(t) \in \mathcal{S}_+$ for all $t \in [t_1, T]$. Symmetrically, every $j \in \mathcal{V}_-$ with $\max_{i\in\mathcal{I}_-}\langle \boldsymbol{w}_j(0),\, \boldsymbol{s}_i\rangle > 0$ satisfies $\cos(\boldsymbol{w}_j(t), -\boldsymbol{u}_\star) \geq c^*$ and $\boldsymbol{w}_j(t) \in \mathcal{S}_-$ for all $t \in [t_1, T]$.

2. (Final convergence.) *Let $\widetilde{\mathcal{V}_+} := \{j \in \mathcal{V}_+ : \max_{i\in\mathcal{I}_+}\langle \boldsymbol{w}_j(0),\, \boldsymbol{s}_i\rangle > 0\}$ and $\widetilde{\mathcal{V}_-} := \{j \in \mathcal{V}_- : \max_{i\in\mathcal{I}_-}\langle \boldsymbol{w}_j(0),\, \boldsymbol{s}_i\rangle > 0\}$, and suppose both are nonempty. Then:*

   (a) *for every $j \in \widetilde{\mathcal{V}_+}$, $\boldsymbol{w}_j(t) \in \mathcal{S}_+$ for all $t \geq t_1$, and $t \mapsto \langle \boldsymbol{u}_\star, \boldsymbol{w}_j(t)\rangle$ is positive and nondecreasing on $[t_1, \infty)$; symmetrically, for every $j \in \widetilde{\mathcal{V}_-}$, $\boldsymbol{w}_j(t) \in \mathcal{S}_-$ for all $t \geq t_1$ and $t \mapsto \langle -\boldsymbol{u}_\star, \boldsymbol{w}_j(t)\rangle$ is positive and nondecreasing;*

   (b) *for every $b_0 > 0$ and $\epsilon \leq 1$ there is a time $t_2 \geq t_1$, bounded explicitly in (50), such that*

   $$L(t) \;\leq\; \frac{L(t_2)}{1 + \alpha\, L(t_2)\,(t - t_2)} \qquad \text{for all } t \geq t_2, \qquad \alpha := \tfrac{1}{2}\, b_0\,(\mu_\star\rho_{\min}\bar{\chi})^2 > 0,$$

   with $\bar{\chi} = 1$ for the exponential loss and $\bar{\chi} = \bar{\chi}(n) \in (0, 1]$ given in Lemma 17 for the logistic loss. In particular, $L(t) = O(1/t)$ and $L(t) \to 0$.

The proof is given in Appendices B–E and sketched in Section 3.2. Our main theorem results in the following inference.

Early alignment. Theorem 1, Part 1. describes the filters at the end of the alignment phase: every filter in $\mathcal{V}_+$ with some activation on a positive signal patch at initialization points, by time t_1, within angle $\arccos c^*$ of $\boldsymbol{u}_\star$, and therefore activates every positive signal patch and no negative one; filters in $\mathcal{V}_-$ do the same with $-\boldsymbol{u}_\star$. The target $c^* < 1$ is arbitrary and enters only through κ. The alignment is not a snapshot: $\{c_j \geq c^*\}$ is forward invariant on the whole window, and afterwards the exact dynamics keep the filter in $\mathcal{S}_+$ (Part 2.(a)). The hypothesis $\max_{i\in\mathcal{I}_+}\langle \boldsymbol{w}_j(0),\, \boldsymbol{s}_i\rangle > 0$ holds with probability at least $1/2$ for each filter in $\mathcal{V}_+$ under a symmetric random W_0, so with m neurons in $\mathcal{V}_+$ at least one aligns with probability at least $1-2^{-m}$; mild overparametrization makes $\widetilde{\mathcal{V}_+}$ and $\widetilde{\mathcal{V}_-}$ nonempty, as Part 2. requires. A filter that activates no positive signal patch at initialization aligns within time t_1 after its first such activation, and before that it can activate signal patches only during a bounded total time (Proposition 1, Appendix D.3).

The alignment time and the choice of ϵ. The approximation (8) is valid only on the window $[0, T]$, so alignment can be concluded only if it is achieved before the window closes; the third condition in (11) is exactly $T \geq t_1$, and without an explicit t_1 one would not know how small ϵ must be. The two terms of t_1 scale differently: the first is the time spent outside the band, where a whole class pushes the filter, and it decreases with the class masses $R_\pm$; the second is the time spent inside the band, where in the worst case a single patch of mass $\rho_{\min}$ pushes, and it is proportional to the band width $\sin\theta_\star$. When the cones degenerate to a single direction, $\theta_\star = 0$, the band disappears and $t_1 = 4/(\kappa R_{\min}\mu_\star)$. Since $R_{\min}$ grows linearly with the class sizes, the exponent $4nPX_{\max}t_1$ in (11) is $O(PX_{\max}/\rho_{\min})$ from the first term and $O(nPX_{\max}\sin\theta_\star/\rho_{\min})$ from the second.

Final convergence. Once a filter is in $\mathcal{S}_+$ it never leaves, and its component along the axis can only grow (Lemma 15); in particular the aligned filters cannot be trapped at a spurious stationary point, a failure mode of early alignment identified by Boursier & Flammarion (2025). Once the aligned mass $\sum_{j\in\widetilde{\mathcal{V}_+}}\langle \boldsymbol{u}_\star,\, \boldsymbol{w}_j\rangle^2$ exceeds b_0, the components of the gradient along $\pm\boldsymbol{u}_\star$ alone give $\|\nabla_W L\|_F^2 \geq \alpha L^2$, hence the rate. The bound (50) on t_2 is of the form $t_1 + O(\frac{\log(1/\epsilon)}{\mu_\star R_{\min}})\exp(C\epsilon^{-2(\Gamma-1)})$: it compares the growth of the aligned mass with that of the total mass of $\mathcal{V}_+$, and the exponent Γ in (10) equals 1 only in the noiseless case with $\theta_\star = 0$. For $\Gamma = 1$ the bound has the expected order $\frac{1}{n}\log\frac{1}{\epsilon}$; for $\Gamma > 1$ it grows exponentially in $\epsilon^{-2(\Gamma-1)}$. The exponent comes from a crude bound on the total norm of the filters of a class, which is our only control on the network outputs before t_2, rather than from a mechanism that slows the dynamics; we expect it can be removed with a noise bound that tolerates unequal contributions from different inputs.

Comparison to prior work. For fully connected networks on inputs with $\min_{i,k}\cos(y_i\boldsymbol{x}_i, y_k\boldsymbol{x}_k) = \mu > 0$, Min et al. (2024) obtain alignment within $O(\log n/\sqrt{\mu})$ by counting activation transitions: each activation pattern is left within a time inversely proportional to the number of active inputs, and the harmonic sum over at most n transitions produces the $\log n$. In the patch model a filter also activates noise patches, which are orthogonal to every signal patch, so the transition argument does not apply to the full activation pattern; we track the cosine with the common axis instead and obtain t_1 from a potential function, at the price of a geometric assumption on the signal and a separate control of the noise through G. Their $t_2 = \Theta(\frac{1}{n}\log\frac{1}{\sqrt{h}\epsilon})$ relies on each class being fitted by a linear subnetwork after alignment; here the filters in $\mathcal{V}_-$ keep activating noise patches of positive inputs, which breaks this decoupling and is the source of our weaker bound on t_2. The network (1) and the data of Definition 1 are those of Cao et al. (2022); Kou et al. (2023a) up to a fixed second layer and the special case $\alpha = 0$, $\boldsymbol{u}_- = -\boldsymbol{u}_+$; those works track the coefficients $\langle \boldsymbol{w}_j,\, \boldsymbol{\mu}\rangle$, $\langle \boldsymbol{w}_j,\, \boldsymbol{\xi}_i^{(p)}\rangle$ along the whole trajectory to characterize benign overfitting, and say nothing about the direction of the filters or the time at which a filter activates a whole class of signal patches. Our result is complementary and does not address test error; with (7), the condition $G \leq \kappa R_{\min}\mu_\star/8$ reads $n\cdot \mathrm{SNR}^2 \gtrsim P^2$ up to logarithms, with $\mathrm{SNR} = \rho_{\min}/(\sigma_\xi\sqrt{D})$, the same form as their benign regime.

3.2 Proof sketch

The proofs are given in Appendices B–E. For the early phase, we first show that for ϵ small the network outputs stay small on a window $[0, T]$ with T of order $\frac{1}{nP}\log\frac{1}{\epsilon}$, and that on this window the direction of every filter $\boldsymbol{w}_j$ evolves, up to a residual of size at most ϱ_1, as if it were driven by its signed activation field $x_a(\boldsymbol{w}_j)$ (Appendix B). By Assumption 1, $x_a(\boldsymbol{w}_j)$ splits into the sum of the activated signed generators, which lies in V_{sig}, and the activated-noise sum, which is orthogonal to

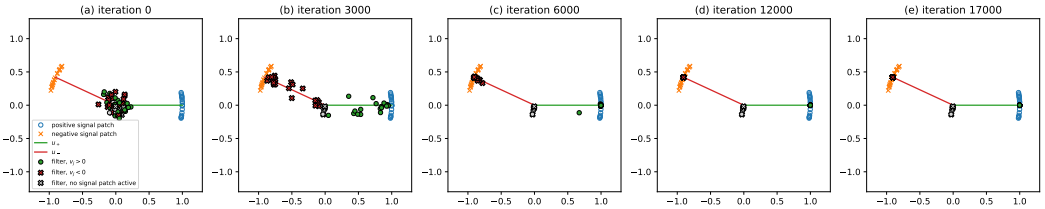

Figure 2: Filters projected onto the signal plane and normalized to unit norm, at five iterations of the base run ($\theta_\star = 24°$, $n = 40$, $P = 5$, $D = 100$, SNR $= 10$, $\epsilon = 10^{-6}$).

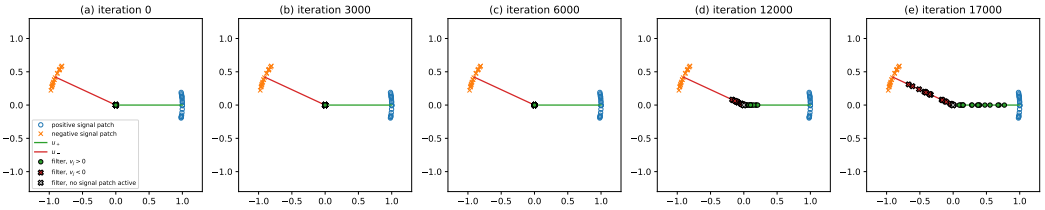

Figure 3: Same run and iterations as Figure 2 without normalization: the projections $\mathcal{P}_{V_{\text{sig}}} \boldsymbol{w}_j$ show that the norms stay at their initial scale until the directions have converged, and then grow along the class axes.

V_{sig}; the former pushes $\boldsymbol{w}_j$ toward $\boldsymbol{u}_\star$ with a strength governed by Lemma 1, and the latter enters only through G (Appendix C). We then track $c_j = \cos(\boldsymbol{w}_j, \boldsymbol{u}_\star)$ for a neuron $j \in \mathcal{V}_+$. Two facts drive the argument. First, the activation pattern of the signal patches can only change in one direction: once $\boldsymbol{w}_j$ activates a positive signal patch it never deactivates it, and as long as it activates some signal patch it never activates a negative one. Second, $\dot{c}_j$ admits explicit lower bounds in each region of the sphere: of order $\kappa R_{\min} \mu_\star$ when $|c_j| > \sin\theta_\star$, where a whole class is activated, and of order $\rho_{\min}$ inside the band $|c_j| \le \sin\theta_\star$, where at least one signal patch is activated. Integrating these bounds through a potential function that rescales time by the available drift gives the alignment time t_1 in (9), and the last condition in (11) ensures $T \ge t_1$ (Appendix D).

For the later phase, the small-initialization approximation is no longer available, and we use exact identities instead: once $\boldsymbol{w}_j \in \mathcal{S}_+$, the derivative of each $\langle \boldsymbol{w}_j, \boldsymbol{g}_i \rangle$ is a positive combination of the generators, so $\mathcal{S}_+$ is never left and $\langle \boldsymbol{u}_\star, \boldsymbol{w}_j \rangle$ can only grow. The aligned mass $\sum_{j \in \widetilde{\mathcal{V}_+}} \langle \boldsymbol{u}_\star, \boldsymbol{w}_j \rangle^2$ grows exponentially at a rate that stays bounded below as long as the network outputs are small, and we keep the outputs small by bounding the total mass of $\mathcal{V}_+$; this yields the transition time t_2. After t_2, the components of the gradient along $\pm \boldsymbol{u}_\star$ alone give $\|\nabla_W L\|_F^2 \ge \alpha L^2$, and integrating $\dot{L} \le -\alpha L^2$ gives the $O(1/t)$ rate (Appendix E).

4 Experiments

We first train on synthetic data drawn from our model: a base run that exhibits the two phases of Theorem 1, followed by ablation studies over different parameters. We then check whether similar pattern in our theorem appears on real data, in a binary classification task on two MNIST digits.

Base experiment. We train the network of Section 2 by full-batch gradient descent on the mean exponential loss, from the unbalanced Gaussian initialization $W_{ij}, v_j \sim \mathcal{N}(0, \epsilon^2)$; the sign of v_j splits the filters into $\mathcal{V}_+$ and $\mathcal{V}_-$ as in Section 2. Synthetic data follow Definition 1 with $n = 40$, $P = 5$, $D = 100$, $h = 50$, $\theta_\star = 24°$ and Gaussian noise patches of norm about 0.1ρ (SNR $= 10$); all settings and details are in Appendix A.

Figures 2 and 3 show the filters of the base experiment projected onto the signal plane, normalized and unnormalized. Figure 4 gives a curve of the training loss for the base experiment. The two phases of Theorem 1 are visible directly: the filters first turn toward the class axes at essentially constant norm of 0, and by $\hat{t}_1 = 5780$ almost every filter that activates a signal patch lies inside the

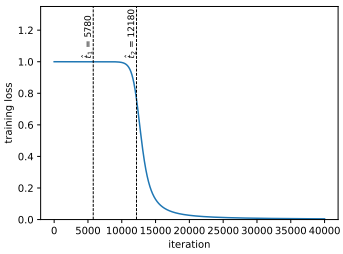

Figure 4: Training loss of the base experiment. The dashed lines mark the empirical value $\hat{t}_1 = 5780$ and $\hat{t}_2 = 12180$; the loss does not move during the alignment phase.

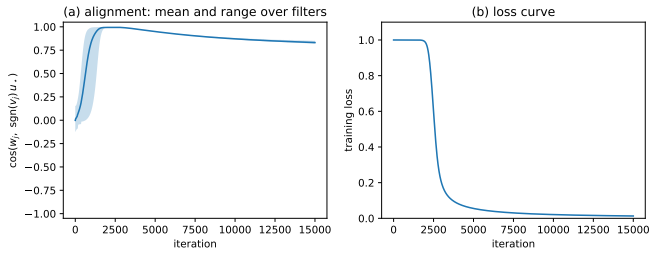

Figure 5: MNIST digits 1 vs 0, 14×14 patches ($P = 4$, $D = 196$, $n = 2000$). (a) $\cos(\boldsymbol{w}_j, \mathrm{sign}(v_j)\boldsymbol{u}_\star)$, mean (line) and range (band) over the $h = 50$ filters; (b) training loss. Alignment completes before the loss decreases and recedes slowly during fitting.

cone of its class while the loss is still 1.0; only afterwards do the norms grow along the class axes, the aligned masses exceed a certain value at $\hat{t}_2 = 12180$, and the loss decreases.

Ablation studies. Additional ablation studies are provided in Appendix A.2 where we vary the initialization scale, the noise level, the number of patches and the cone width (five seeds each), analyzing the mean alignment of the filters that activate a signal patch at four different iterations of training.

Real data. We take 1000 images each of the digits 1 ($y = +1$) and 0 ($y = -1$) from the MNIST training set, scale the pixels to $[0, 1]$, subtract the mean image, which is the same as centring every patch position, and cut each image into 16 non-overlapping 14×14 patches, so that $P = 4$ and $D = 196$; no patch is designated as signal or noise. We use the unbalanced Gaussian initialization with $\epsilon = 10^{-6}$, and step size $\eta = 2 \times 10^{-3}$; the network has $h = 50$ filters and is trained for 15000 iterations of full-batch gradient descent. Since there is no notion of a signal patch in the real-data experiment, no filter is excluded, and Figure 5 reports the mean and the range of c_j over all $h = 50$ filters. All filters align with $\pm\boldsymbol{u}_\star$ within about 2500 iterations, while the loss is still 1.0; the loss then decreases rapidly, and during this fitting phase the alignment recedes slowly. This is expected: once some inputs are fitted, the gradient is dominated by those that are not, and the filters move away from the class axes to fit them. The early phase, which is what Theorem 1 describes, is unaffected. Here the signed patches are spread over a wide range of angles rather than confined to a narrow cone, and the early alignment nevertheless takes place.

5 Conclusion

We analyzed the gradient flow of a two-layer ReLU convolutional network with sum-pooling, trained for binary classification on patch data from a small balanced initialization. When the signal patches of the two classes lie in two narrow, nearly antipodal cones, the label-signed patches share a common axis $\boldsymbol{u}_\star$, and every filter that activates a signal patch of its own class at initialization aligns with $\pm\boldsymbol{u}_\star$ by an explicit time t_1 that contains no logarithmic factor in n; the noise enters only through the largest signed sum of noise patches that a single direction can activate. After alignment the filters remain in their signal cones, and after a transition time t_2 the loss decays at rate $O(1/t)$. Experiments confirm this order of events and show that the quantitative conditions are conservative. Three limitations point to future work. For $\Gamma > 1$ the bound on t_2 is exponential in $\epsilon^{-2(\Gamma-1)}$; the exponent comes from a crude control of the total mass of a class of filters, and removing it requires a noise bound that holds under data-dependent loss weights. The second signal-to-noise condition compares the noise to the mass of a single patch and becomes restrictive when the cones are wide; an assumption on the spread of the patches inside their cones would relax it. Finally, the one-way switching of the signal activation pattern relies on sum-pooling; whether a comparable early phase exists under max-pooling, or in deeper convolutional networks, remains open.

DISCLOSURE OF AI USE

We used generative AI tools to assist with writing and editing experimental code, checking mathematical proofs developed by the authors, and polishing the manuscript for clarity and readability. The research ideas and original mathematical proofs were developed by the authors without generative AI assistance. The authors reviewed the AI-assisted writing, checked the experimental code, and independently verified the final mathematical proofs. We take full responsibility for the final content of this paper, including its text, mathematical claims and proofs, experimental results, and accompanying code.

REPRODUCIBILITY STATEMENT

All assumptions are stated in Section 2, and complete proofs of all results are given in Appendices B–E. Codes for the experiments are provided with the corresponding seeds in the supplement.

REFERENCES

Zeyuan Allen-Zhu and Yuanzhi Li. Towards understanding ensemble, knowledge distillation and self-distillation in deep learning. In *International Conference on Learning Representations*, 2023.

Sanjeev Arora, Nadav Cohen, Wei Hu, and Yuping Luo. Implicit regularization in deep matrix factorization. In *Advances in Neural Information Processing Systems*, volume 32, 2019.

Ioannis Bantzis, James B. Simon, and Arthur Jacot. Saddle-to-saddle dynamics in deep ReLU networks: Low-rank bias in the first saddle escape. In *International Conference on Learning Representations*, 2026.

Etienne Boursier and Nicolas Flammarion. Early alignment in two-layer networks training is a two-edged sword. *Journal of Machine Learning Research*, 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, 2022.

Alon Brutzkus and Amir Globerson. Globally optimal gradient descent for a ConvNet with Gaussian inputs. In *International Conference on Machine Learning*, 2017.

Yuan Cao, Zixiang Chen, Mikhail Belkin, and Quanquan Gu. Benign overfitting in two-layer convolutional neural networks. In *Advances in Neural Information Processing Systems*, volume 35, 2022.

Dmitry Chistikov, Matthias Englert, and Ranko Lazic. Learning a neuron by a shallow ReLU network: Dynamics and implicit bias for correlated inputs. In *Advances in Neural Information Processing Systems*, volume 36, 2023.

Lénaïc Chizat and Francis Bach. Implicit bias of gradient descent for wide two-layer neural networks trained with the logistic loss. In *Conference on Learning Theory*, 2020.

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

Jona-Maria Diederen, Holger Rauhut, and Ulrich Terstiege. Convergence of gradient flow for learning convolutional neural networks. *arXiv preprint arXiv:2601.08547*, 2026.

Simon S. Du, Jason D. Lee, Yuandong Tian, Aarti Singh, and Barnabás Póczos. Gradient descent learns one-hidden-layer CNN: Don't be afraid of spurious local minima. In *International Conference on Machine Learning*, 2018.

Spencer Frei, Gal Vardi, Peter L. Bartlett, Nathan Srebro, and Wei Hu. Implicit bias in leaky ReLU networks trained on high-dimensional data. In *International Conference on Learning Representations*, 2023.

Gauthier Gidel, Francis Bach, and Simon Lacoste-Julien. Implicit regularization of discrete gradient dynamics in linear neural networks. In *Advances in Neural Information Processing Systems*, volume 32, 2019.

Suriya Gunasekar, Blake Woodworth, Srinadh Bhojanapalli, Behnam Neyshabur, and Nathan Srebro. Implicit regularization in matrix factorization. In *Advances in Neural Information Processing Systems*, volume 30, 2017.

Suriya Gunasekar, Jason D. Lee, Daniel Soudry, and Nathan Srebro. Implicit bias of gradient descent on linear convolutional networks. In *Advances in Neural Information Processing Systems*, volume 31, 2018.

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. *arXiv preprint arXiv:2106.15933*, 2021.

Samy Jelassi and Yuanzhi Li. Towards understanding how momentum improves generalization in deep learning. In *International Conference on Machine Learning*, 2022.

Ziwei Ji and Matus Telgarsky. Directional convergence and alignment in deep learning. In *Advances in Neural Information Processing Systems*, volume 33, 2020.

Kathlén Kohn, Thomas Merkh, Guido Montúfar, and Matthew Trager. Geometry of linear convolutional networks. *SIAM Journal on Applied Algebra and Geometry*, 6(3):368–406, 2022.

Yiwen Kou, Zixiang Chen, Yuanzhou Chen, and Quanquan Gu. Benign overfitting in two-layer ReLU convolutional neural networks. In *International Conference on Machine Learning*, 2023a.

Yiwen Kou, Zixiang Chen, and Quanquan Gu. Implicit bias of gradient descent for two-layer ReLU and leaky ReLU networks on nearly-orthogonal data. In *Advances in Neural Information Processing Systems*, volume 36, 2023b.

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. Early directional convergence in deep homogeneous neural networks for small initializations. *Transactions on Machine Learning Research*, 2025.

Yuanzhi Li, Tengyu Ma, and Hongyang Zhang. Algorithmic regularization in over-parameterized matrix sensing and neural networks with quadratic activations. In *Conference on Learning Theory*, 2018.

Kaifeng Lyu and Jian Li. Gradient descent maximizes the margin of homogeneous neural networks. In *International Conference on Learning Representations*, 2020.

Kaifeng Lyu, Zhiyuan Li, Runzhe Wang, and Sanjeev Arora. Gradient descent on two-layer nets: Margin maximization and simplicity bias. In *Advances in Neural Information Processing Systems*, volume 34, 2021.

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

Xuran Meng, Difan Zou, and Yuan Cao. Benign overfitting in two-layer ReLU convolutional neural networks for XOR data. In *International Conference on Machine Learning*, 2024.

Hancheng Min, Enrique Mallada, and René Vidal. Early neuron alignment in two-layer ReLU networks with small initialization. In *International Conference on Learning Representations*, 2024.

Scott Pesme and Nicolas Flammarion. Saddle-to-saddle dynamics in diagonal linear networks. In *Advances in Neural Information Processing Systems*, volume 36, 2023.

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

Noam Razin, Asaf Maman, and Nadav Cohen. Implicit regularization in hierarchical tensor factorization and deep convolutional neural networks. In *International Conference on Machine Learning*, 2022.

Andrew M. Saxe, James L. McClelland, and Surya Ganguli. Exact solutions to the nonlinear dynamics of learning in deep linear neural networks. In *International Conference on Learning Representations*, 2014.

Dominik Stöger and Mahdi Soltanolkotabi. Small random initialization is akin to spectral learning: Optimization and generalization guarantees for overparameterized low-rank matrix reconstruction. In *Advances in Neural Information Processing Systems*, volume 34, 2021.

Nikita Tsoy and Nikola Konstantinov. Simplicity bias of two-layer networks beyond linearly separable data. In *International Conference on Machine Learning*, 2024.

Roman Vershynin. *High-Dimensional Probability: An Introduction with Applications in Data Science*. Cambridge University Press, 2018.

Mingze Wang and Chao Ma. Early stage convergence and global convergence of training mildly parameterized neural networks. In *Advances in Neural Information Processing Systems*, volume 35, 2022.

Mingze Wang and Chao Ma. Understanding multi-phase optimization dynamics and rich nonlinear behaviors of ReLU networks. In *Advances in Neural Information Processing Systems*, volume 36, 2023.

Blake Woodworth, Suriya Gunasekar, Jason D. Lee, Edward Moroshko, Pedro Savarese, Itay Golan, Daniel Soudry, and Nathan Srebro. Kernel and rich regimes in overparametrized models. In *Conference on Learning Theory*, 2020.

Difan Zou, Yuan Cao, Yuanzhi Li, and Quanquan Gu. Understanding the generalization of Adam in learning neural networks with proper regularization. In *International Conference on Learning Representations*, 2023.

APPENDICES

A Experimental Details and Ablation Studies **13**

A.1 Setup 13

A.2 Ablation Studies 13

B Gradient Flow under Balanced Small Initialization **14**

C Geometry of the Signal Cone and Uniform Control of the Noise **18**

C.1 Proof of Lemma 1 18

C.2 The Signal Cone 19

C.3 Uniform Bound on the Noise Field 20

D Proof of Theorem 1: Early Alignment Phase **22**

D.1 Auxiliary Lemmata 23

D.2 Proof of the First Part of Theorem 1 25

D.3 Neurons that Activate No Positive Signal Patch at Initialization 26

E Proof of Theorem 1: Final Convergence **27**

E.1 Auxiliary Lemmata 27

E.2 Proof of the Second Part of Theorem 1 29

A Experimental Details and Ablation Studies

This appendix gives the full setting of the experiments of Section 4 (Appendix A.1) and the four ablations (Appendix A.2). All experiments are implemented in NumPy; the code is provided as supplementary material.

A.1 Setup

Data generation. The signal subspace is $V_{\mathrm{sig}} = \mathrm{span}\{\boldsymbol{e}_1, \boldsymbol{e}_2\}$. The class axes are $\boldsymbol{u}_+ = \boldsymbol{e}_1$ and $\boldsymbol{u}_- = (\cos\beta_0, \sin\beta_0, 0, \dots, 0)$, and the signal patch of input i is $\boldsymbol{s}_i = \rho(\cos\vartheta_i, \sin\vartheta_i, 0, \dots, 0)$, where ϑ_i is the angle of the class axis of i plus $\pm\phi_i$ with $\phi_i \sim \mathrm{U}[0, \alpha]$ and either sign with probability $1/2$. The noise patches are $\boldsymbol{\xi}_i^{(p)} \sim \mathcal{N}(0, \sigma_\xi^2 I)$ on the coordinates $3, \dots, D$, so that Assumption 1 holds by construction. The classes are balanced, $n_+ = n_-$, all signal norms are $\rho = 1$, and the signal patch is placed at position $p = 1$, which is immaterial under sum-pooling. We parametrize the geometry by $\theta_\star$ through $\alpha = \theta_\star/2$ and $\beta_0 = \pi - \theta_\star$, so that $\theta_\star = (\pi - \beta_0)/2 + \alpha$ as in (5) and the class axes make an angle $\theta_\star/2$ with $\pm\boldsymbol{u}_\star$. The noise level is parametrized by $\mathrm{SNR} := \rho/(\sigma_\xi\sqrt{D})$, so that a noise patch has norm about ρ/SNR.

Base configuration.

D	P	n	h	$\theta_\star$ (α, β_0)	SNR	ϵ	η	iterations
100	5	40	50	24° (12°, 156°)	10	10^{-6}	2×10^{-3}	4×10^4

Training. Full-batch gradient descent with step size η on the mean exponential loss $\frac{1}{n}\sum_{i=1}^n e^{-y_i f(\boldsymbol{X}_i; W, \boldsymbol{v})}$, updating both layers, from $W_{ij}, v_j \sim \mathcal{N}(0, \epsilon^2)$ independently. The initialization is not balanced, so $\mathrm{sign}\, v_j$ may in principle change; the sets $\mathcal{V}_+(k)$ and $\mathcal{V}_-(k)$ at iteration k are defined by the current sign of $v_j(k)$.

Reported quantities. For filter j at iteration k we record $c_j(k) := \cos(\boldsymbol{w}_j(k), \mathrm{sign}(v_j(k))\, \boldsymbol{u}_\star)$, and $\bar{c}(k)$ denotes the average of $c_j(k)$ over the *live* filters, those that activate at least one signal patch at the end of training. Whether a filter is live can be read off at the end of training: the gradient from the noise patches lies in $V_{\mathrm{sig}}^\perp$, so once $\mathcal{A}(\boldsymbol{w}_j) = \emptyset$ the projection of $\boldsymbol{w}_j$ onto the signal plane no longer changes. The alignment time $\hat{t}_{\mathrm{align}}$ is the first iteration at which the signal-plane projection of every live filter makes an angle at most α with the axis of its class; this is stricter than entering $\mathcal{S}_+$ or $\mathcal{S}_-$, which in the base geometry are wedges of opening angle 132° and which every live filter has entered by iteration 2400. The transition time $\hat{t}_2$ is the first iteration at which both aligned masses $Q_\pm$ of Theorem 1(2) exceed $b_0 = 1/(4\rho)$. The empirical noise level $\widehat{G}$ is a lower estimate of $G = \sup_{\boldsymbol{w}} \|\boldsymbol{g}(\boldsymbol{w})\|$, obtained by iterating $\boldsymbol{w} \leftarrow \boldsymbol{g}(\boldsymbol{w})/\|\boldsymbol{g}(\boldsymbol{w})\|$ from random directions and from $\pm\sum_{i,p} y_i \boldsymbol{\xi}_i^{(p)}$ and taking the largest norm encountered; for Gaussian noise it is well described by $\widehat{G} \approx 0.9\, \sigma_\xi \sqrt{n(P-1)(D-2)}$. Snapshot iterations in the figures are rounded to the nearest 1000.

A.2 Ablation Studies

Each ablation varies one parameter and keeps the others at the base configuration. We run five seeds and report the median of $\bar{c}$ with the range over seeds as a band, read at the four landmark iterations of the base run, $\hat{t}_{\mathrm{align}}/2$, $\hat{t}_{\mathrm{align}}$, $\hat{t}_2$ and the end of training, rounded to 3000, 6000, 12000 and 40000. The fraction of filters that activate no signal patch is between 14% and 20% in all configurations and does not depend on the parameter varied. The training loss is almost insensitive to all four parameters, except for the shift of the fitting phase with ϵ noted below, and is not shown.

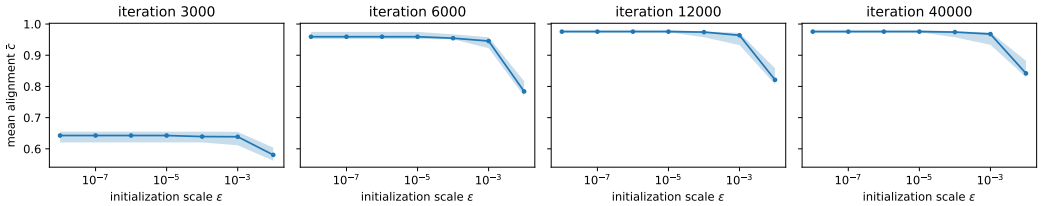

Figure 6: Mean alignment $\bar{c}$ of the live filters against the initialization scale ϵ at four iterations of training (median and range over five seeds; other parameters as in the base run).

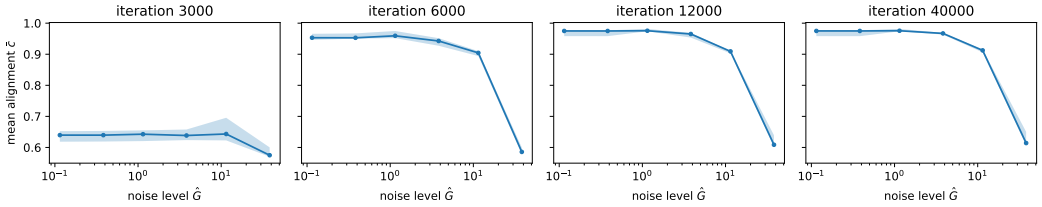

Figure 7: Mean alignment $\bar{c}$ of the live filters against the empirical noise level $\widehat{G}$ of Appendix A.1, at four iterations of training (median and range over five seeds; other parameters as in the base run).

Initialization scale (Figure 6). For $\epsilon \in \{10^{-2}, \dots, 10^{-8}\}$, the four columns are flat for $\epsilon \le 10^{-4}$, with final $\bar{c} = 0.976$, and alignment degrades from $\epsilon = 10^{-3}$ on, reaching a final $\bar{c} = 0.84$ at $\epsilon = 10^{-2}$. The onset of fitting moves later by 2320 iterations for every factor of 10 by which ϵ decreases, that is, linearly in $\log(1/\epsilon)$, which is the scaling of the window $T := \frac{1}{4nPX_{\max}} \log \frac{1}{\sqrt{h}\,\epsilon}$. For large ϵ the two phases overlap, and fitting begins before the directions have converged; this is the cause of the degradation.

Noise level (Figure 7). Varying $\mathrm{SNR} \in \{100, 30, 10, 3, 1, 0.3\}$ gives $\widehat{G} \approx 0.11, 0.38, 1.15, 3.8, 11.5, 38$, against class masses $R_{\pm} = 20$. Alignment is unchanged up to $\widehat{G} \approx 3.8$, about a fifth of $R_{\pm}$, with final $\bar{c} \in \{0.975, 0.975, 0.976, 0.967\}$; it drops to 0.912 at $\widehat{G} \approx 11.5$ and to 0.614 at $\widehat{G} \approx 38$, where the noise exceeds the class mass. At these noise levels the loss decreases earlier and further than in the base run: the network fits the noise patches, whose norms now exceed those of the signal patches, and the training loss does not distinguish the two. This is why alignment, not loss, is the quantity we track, and it is the regime excluded by the first condition of Assumption 4.

Number of patches (Figure 8). For $P \in \{2, 3, 5, 9, 17\}$ at $\mathrm{SNR} = 10$ the noise level stays small ($\widehat{G} \approx 2.3$ even at $P = 17$) and it is hard to observe the trend of mean alignment, so we sweep P at $\mathrm{SNR} = 1$, where $\widehat{G} \approx 6, 9, 13, 18, 25$ grows like $\sqrt{P-1}$ and the final $\bar{c}$ decreases monotonically from 0.96 to 0.77. The number of patches acts on alignment only through the noise level, as in the analysis, where P enters the alignment phase only through G and through the length of the window T.

Cone width (Figure 9). For $\theta_\star \in \{5, 10, 15, 20, 24, 30, 35, 40, 44\}$ degrees the final $\bar{c}$ decreases from 0.998 to 0.917. The different columns show that wider cones align more slowly and finally achieve a worse alignment, indicating a lower c^*.

B Gradient Flow under Balanced Small Initialization

This appendix collects the properties of the gradient flow (2) that are used throughout the proofs. None of them depends on the data model beyond the bound $\left\|\boldsymbol{x}_i^{(p)}\right\| \le X_{\max}$; in particular, they hold for arbitrary patches and labels. We first record a property of the loss (Lemma 2) and the

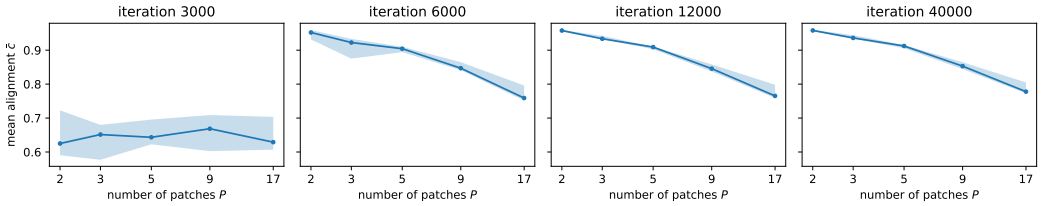

Figure 8: Mean alignment $\bar{c}$ of the live filters against the number of patches P at $\mathrm{SNR} = 1$, at four iterations of training (median and range over five seeds; other parameters as in the base run).

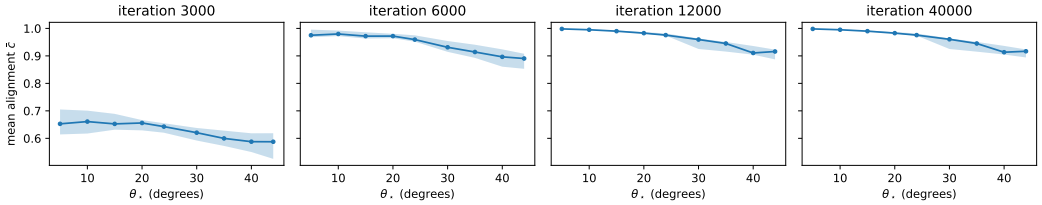

Figure 9: Mean alignment $\bar{c}$ of the live filters against the cone width $\theta_\star$ at four iterations of training (median and range over five seeds; other parameters as in the base run).

conservation law behind the balanced initialization (Lemma 3), then bound the filter norms and the network outputs on an initial time window (Lemma 4), and finally show that on this window the direction of every filter evolves, up to a small residual, according to its signed activation field which is defined as,

$$x_a(\boldsymbol{w}) \;:=\; \sum_{(i,p):\, \langle \boldsymbol{w},\, \boldsymbol{x}_i^{(p)} \rangle > 0} y_i\, \boldsymbol{x}_i^{(p)}, \tag{12}$$

following Lemma 5 and Corollary 1).

Solutions. Since σ' is discontinuous, (2) is understood in the sense of Carathéodory: a solution is an absolutely continuous trajectory $\{W(t), \boldsymbol{v}(t)\}_{t\geq 0}$ that satisfies (2) for almost every t, with $\sigma'(z) = \mathbf{1}_{z>0}$. All differential relations below hold for almost every t, and we omit this qualifier. All our statements concern an arbitrary such solution; we do not address the question of existence here.

Lemma 2 (Loss-gradient envelope)**.** *For both the exponential loss $\ell(y,\hat{y}) = \exp(-y\hat{y})$ and the logistic loss $\ell(y,\hat{y}) = 2\log(1+\exp(-y\hat{y}))$,*

$$|-\nabla_{\hat{y}}\ell(y,\hat{y}) - y| \;\leq\; 2\,|\hat{y}|\,, \qquad \forall y \in \{+1,-1\},\;\; \forall\, |\hat{y}| \leq 1.$$

Proof. For the exponential loss, $-\nabla_{\hat{y}}\ell(y,\hat{y}) = y\exp(-y\hat{y})$, so

$$|-\nabla_{\hat{y}}\ell(y,\hat{y}) - y| = |y\exp(-y\hat{y}) - y| \leq |y|\,|\exp(-y\hat{y}) - 1| \leq |\exp(-y\hat{y}) - 1| \leq 2\,|\hat{y}|\,,$$

where the last inequality is due to the fact that $2x \geq \max\{1-\exp(-x), \exp(x)-1\}$ for all $x \in [0,1]$.

For the logistic loss, $-\nabla_{\hat{y}}\ell(y,\hat{y}) = 2y\exp(-y\hat{y})/(1+\exp(-y\hat{y}))$, so

$$|-\nabla_{\hat{y}}\ell(y,\hat{y}) - y| = \left|\frac{2y\exp(-y\hat{y})}{1+\exp(-y\hat{y})} - y\right| = \left|\frac{y\big(\exp(-y\hat{y})-1\big)}{1+\exp(-y\hat{y})}\right| \leq |\exp(-y\hat{y})-1| \leq 2\,|\hat{y}|\,. \;\square$$

Lemma 3 (Balancedness and sign preservation)**.** *Consider any solution to the gradient flow (2) starting from the initialization (3). Then for every $j \in [h]$ and every $t \geq 0$,*

$$|v_j(t)| = \|\boldsymbol{w}_j(t)\| > 0 \qquad \textit{and} \qquad \operatorname{sign}(v_j(t)) = \operatorname{sign}(v_j(0)),$$

so that $v_j(t) = \operatorname{sign}(v_j(0))\,\|\boldsymbol{w}_j(t)\|$. Consequently, the filters evolve according to

$$\dot{\boldsymbol{w}}_j \;=\; -\operatorname{sign}(v_j(0))\;\|\boldsymbol{w}_j\| \sum_{i=1}^{n}\sum_{p=1}^{P} \mathbf{1}_{\langle \boldsymbol{w}_j,\, \boldsymbol{x}_i^{(p)} \rangle > 0}\, \nabla_{\hat{y}}\ell\big(y_i, f(\boldsymbol{X}_i; W, \boldsymbol{v})\big)\, \boldsymbol{x}_i^{(p)}. \tag{13}$$

Proof. Write $g_i := \nabla_{\hat{y}} \ell(y_i, f(\boldsymbol{X}_i; W, \boldsymbol{v}))$ and let $\Phi_{ij} := \sum_{p=1}^{P} \sigma(\langle \boldsymbol{w}_j, \boldsymbol{x}_i^{(p)} \rangle)$ denote the pooled response of filter j to input i, so that $f(\boldsymbol{X}_i; W, \boldsymbol{v}) = \sum_j v_j \Phi_{ij}$. The gradient flow (2) reads

$$\dot{v}_j = -\sum_{i=1}^{n} g_i \, \Phi_{ij}, \qquad \dot{\boldsymbol{w}}_j = -v_j \sum_{i=1}^{n} g_i \sum_{p=1}^{P} \mathbf{1}_{\langle \boldsymbol{w}_j, \boldsymbol{x}_i^{(p)} \rangle > 0} \, \boldsymbol{x}_i^{(p)}. \tag{14}$$

Since $z \, \mathbf{1}_{z>0} = \sigma(z)$ for every $z \in \mathbb{R}$, taking the inner product of the second equation with $\boldsymbol{w}_j$ gives $\langle \boldsymbol{w}_j, \dot{\boldsymbol{w}}_j \rangle = -v_j \sum_i g_i \Phi_{ij} = v_j \dot{v}_j$, hence

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

As $t \mapsto v_j(t)^2 - \|\boldsymbol{w}_j(t)\|^2$ is absolutely continuous, it is constant, and it vanishes at $t = 0$ by (3). Thus $|v_j(t)| = \|\boldsymbol{w}_j(t)\|$ for all $t \geq 0$.

It remains to show that v_j never vanishes; then $\operatorname{sign}(v_j(t))$ is constant by continuity, and (13) follows from (14) upon substituting $v_j = \operatorname{sign}(v_j(0)) \, \|\boldsymbol{w}_j\|$. By positive homogeneity of σ and Cauchy–Schwarz, $0 \leq \Phi_{ij} \leq \|\boldsymbol{w}_j\| \sum_p \left\| \boldsymbol{x}_i^{(p)} \right\| \leq P X_{\max} \|\boldsymbol{w}_j\|$. Moreover, $|g_i| \leq \ell(y_i, f(\boldsymbol{X}_i; W, \boldsymbol{v}))$ for both losses: for the exponential loss this holds with equality, and for the logistic loss, writing $r = \exp(-y_i f(\boldsymbol{X}_i; W, \boldsymbol{v}))$, $|g_i| = \frac{2r}{1+r} \leq 2 \log(1+r)$ because $\log(1+r) \geq \frac{r}{1+r}$ for $r \geq 0$. Since the loss is nonincreasing along the gradient flow, we obtain from (14) and $\|\boldsymbol{w}_j\| = |v_j|$,

$$|\dot{v}_j| \leq \sum_{i=1}^{n} |g_i| \, \Phi_{ij} \leq P X_{\max} \|\boldsymbol{w}_j\| \sum_{i=1}^{n} \ell\big(y_i, f(\boldsymbol{X}_i; W, \boldsymbol{v})\big) \leq P X_{\max} L(W(0), \boldsymbol{v}(0)) \, |v_j| \, .$$

Hence $\frac{d}{dt} |v_j| \geq -P X_{\max} L(W(0), \boldsymbol{v}(0)) \, |v_j|$, and by Grönwall's inequality $|v_j(t)| \geq |v_j(0)| \exp\big(- P X_{\max} L(W(0), \boldsymbol{v}(0)) \, t \big) > 0$ for every $t \geq 0$, where $|v_j(0)| = \|[W_0]_{:,j}\| \, \epsilon > 0$ because the columns of W_0 are nonzero. This finishes the proof. □

In particular, the index sets $\mathcal{V}_+ := \{j : v_j(0) > 0\}$ and $\mathcal{V}_- := \{j : v_j(0) < 0\}$ of the two types of neurons are time-invariant, and every filter $\boldsymbol{w}_j(t)$ is nonzero, so that $\boldsymbol{w}_j(t) / \|\boldsymbol{w}_j(t)\|$ is well defined for all $t \geq 0$.

Lemma 4 (Bounds on filter norms and network outputs)**.** *Consider any solution to the gradient flow* (2) *starting from the initialization* (3) *with*

$$\epsilon \leq \frac{1}{4\sqrt{h} \, P X_{\max} W_{\max}^2}. \tag{15}$$

Then for all $t \leq T$, with $T := \frac{1}{4nPX_{\max}} \log \frac{1}{\sqrt{h}\,\epsilon}$, we have

$$\max_{j \in [h]} \|\boldsymbol{w}_j(t)\|^2 \leq \frac{2\epsilon W_{\max}^2}{\sqrt{h}}, \qquad \max_{i \in [n]} |f(\boldsymbol{X}_i; W(t), \boldsymbol{v}(t))| \leq 2\epsilon \sqrt{h} \, P X_{\max} W_{\max}^2. \tag{16}$$

Proof. We first relate the network outputs to the filter norms. By Lemma 3, $|v_j| = \|\boldsymbol{w}_j\|$, hence for every $i \in [n]$,

$$\begin{aligned} |f(\boldsymbol{X}_i; W, \boldsymbol{v})| = \Big| \sum_{j=1}^{h} v_j \sum_{p=1}^{P} \sigma\big(\langle \boldsymbol{w}_j, \boldsymbol{x}_i^{(p)} \rangle\big) \Big| &\leq \sum_{j=1}^{h} \|\boldsymbol{w}_j\| \sum_{p=1}^{P} \|\boldsymbol{w}_j\| \left\| \boldsymbol{x}_i^{(p)} \right\| \\ &\leq P X_{\max} \sum_{j=1}^{h} \|\boldsymbol{w}_j\|^2 \leq h P X_{\max} \max_{j \in [h]} \|\boldsymbol{w}_j\|^2 . \end{aligned} \tag{17}$$

Let $T_0 := \inf\{t \geq 0 : \max_i |f(\boldsymbol{X}_i; W(t), \boldsymbol{v}(t))| > 2\epsilon \sqrt{h} \, P X_{\max} W_{\max}^2\}$. For $t \leq T_0$ we have $\max_i |f(\boldsymbol{X}_i; W, \boldsymbol{v})| \leq 2\epsilon \sqrt{h} P X_{\max} W_{\max}^2 \leq \frac{1}{2}$ by (15), so Lemma 2 applies, and (13) gives, for

every $j \in [h]$,

$$\begin{aligned}
\frac{d}{dt}\|\boldsymbol{w}_j\|^2 &= 2\langle \boldsymbol{w}_j,\, \dot{\boldsymbol{w}}_j\rangle \\
&= -2\,\mathrm{sign}(v_j(0))\,\|\boldsymbol{w}_j\| \sum_{i,p} \mathbf{1}_{\langle \boldsymbol{w}_j,\, \boldsymbol{x}_i^{(p)}\rangle > 0}\, \nabla_{\hat{y}}\ell\big(y_i, f(\boldsymbol{X}_i; W, \boldsymbol{v})\big)\langle \boldsymbol{x}_i^{(p)},\, \boldsymbol{w}_j\rangle \\
&\le 2\,\|\boldsymbol{w}_j\| \sum_{i,p} \big|\nabla_{\hat{y}}\ell\big(y_i, f(\boldsymbol{X}_i; W, \boldsymbol{v})\big)\big| \,\Big|\langle \boldsymbol{x}_i^{(p)},\, \boldsymbol{w}_j\rangle\Big| \\
&\le 2\,\|\boldsymbol{w}_j\| \sum_{i,p} \big(1 + 2\,|f(\boldsymbol{X}_i; W, \boldsymbol{v})|\,\big) \,\Big|\langle \boldsymbol{x}_i^{(p)},\, \boldsymbol{w}_j\rangle\Big| \qquad \text{(by Lemma 2)} \\
&\le 2\big(1 + 4\epsilon\sqrt{h}PX_{\max}W_{\max}^2\big) \sum_{i,p} \Big\|\boldsymbol{x}_i^{(p)}\Big\| \,\|\boldsymbol{w}_j\|^2 \qquad \text{(since } t \le T_0\text{)} \\
&\le 2nP\big(X_{\max} + 4\epsilon\sqrt{h}PX_{\max}^2W_{\max}^2\big)\,\|\boldsymbol{w}_j\|^2 . \qquad (18)
\end{aligned}$$

Let $\tau_j := \inf\{t \ge 0 : \|\boldsymbol{w}_j(t)\|^2 > 2\epsilon W_{\max}^2/\sqrt{h}\}$ and $j^* := \arg\min_j \tau_j$. Then $\tau_{j^*} \le T_0$: indeed, by (17), $\max_i |f(\boldsymbol{X}_i; W(t), \boldsymbol{v}(t))| > 2\epsilon\sqrt{h}PX_{\max}W_{\max}^2$ implies $\|\boldsymbol{w}_j(t)\|^2 > 2\epsilon W_{\max}^2/\sqrt{h}$ for some j, i.e. $t \ge \tau_j \ge \tau_{j^*}$. Therefore (18) holds for j^* on $[0, \tau_{j^*}]$, and Grönwall's inequality yields, for all $t \le \tau_{j^*}$,

$$\begin{aligned}
\|\boldsymbol{w}_{j^*}(t)\|^2 &\le \exp\Big(2nP\big(X_{\max} + 4\epsilon\sqrt{h}PX_{\max}^2W_{\max}^2\big)t\Big)\,\|\boldsymbol{w}_{j^*}(0)\|^2 \\
&= \exp\Big(2nP\big(X_{\max} + 4\epsilon\sqrt{h}PX_{\max}^2W_{\max}^2\big)t\Big)\epsilon^2\,\|[W_0]_{:,j^*}\|^2 \\
&\le \exp\Big(2nP\big(X_{\max} + 4\epsilon\sqrt{h}PX_{\max}^2W_{\max}^2\big)t\Big)\epsilon^2 W_{\max}^2 .
\end{aligned}$$

Suppose $\tau_{j^*} < T$. Then, by the continuity of $\|\boldsymbol{w}_{j^*}(t)\|^2$,

$$\begin{aligned}
\frac{2\epsilon W_{\max}^2}{\sqrt{h}} = \|\boldsymbol{w}_{j^*}(\tau_{j^*})\|^2 &\le \exp\Big(2nP\big(X_{\max} + 4\epsilon\sqrt{h}PX_{\max}^2W_{\max}^2\big)\tau_{j^*}\Big)\epsilon^2 W_{\max}^2 \\
&\le \exp\Big(2nP\big(X_{\max} + 4\epsilon\sqrt{h}PX_{\max}^2W_{\max}^2\big)\frac{1}{4nPX_{\max}}\log\frac{1}{\sqrt{h}\epsilon}\Big)\epsilon^2 W_{\max}^2 \\
&= \exp\Big(\frac{1 + 4\epsilon\sqrt{h}PX_{\max}W_{\max}^2}{2}\log\frac{1}{\sqrt{h}\epsilon}\Big)\epsilon^2 W_{\max}^2 \\
&\le \exp\Big(\log\frac{1}{\sqrt{h}\epsilon}\Big)\epsilon^2 W_{\max}^2 = \frac{\epsilon W_{\max}^2}{\sqrt{h}},
\end{aligned}$$

where the last inequality uses (15). This is the contradiction $2\epsilon \le \epsilon$. Hence $\tau_j \ge \tau_{j^*} \ge T$ for every j, which gives the first bound in (16) for all $t \le T$, and the second bound then follows from (17). □

Lemma 5 (Directional dynamics of the filters)**.** *Consider any solution to the gradient flow* (2) *starting from the initialization* (3)*. Whenever* $\max_i |f(\boldsymbol{X}_i; W, \boldsymbol{v})| \le 1$*, we have for every* $j \in [h]$*,*

$$\Big\|\frac{d}{dt}\frac{\boldsymbol{w}_j}{\|\boldsymbol{w}_j\|} - \mathrm{sign}(v_j(0))\Big(I - \frac{\boldsymbol{w}_j\boldsymbol{w}_j^\top}{\|\boldsymbol{w}_j\|^2}\Big)x_a(\boldsymbol{w}_j)\Big\| \;\le\; 2nPX_{\max}\max_{i\in[n]}|f(\boldsymbol{X}_i; W, \boldsymbol{v})|\,, \qquad (19)$$

where $x_a(\boldsymbol{w}_j) = \sum_{(i,p):\langle \boldsymbol{w}_j,\, \boldsymbol{x}_i^{(p)}\rangle > 0} y_i\boldsymbol{x}_i^{(p)}$ *is the signed activation field* (12)*.*

Proof. Since $\boldsymbol{w}_j \ne 0$ (Lemma 3), $\frac{d}{dt}\frac{\boldsymbol{w}_j}{\|\boldsymbol{w}_j\|} = \frac{\dot{\boldsymbol{w}}_j}{\|\boldsymbol{w}_j\|} - \frac{\boldsymbol{w}_j\langle \boldsymbol{w}_j, \dot{\boldsymbol{w}}_j\rangle}{\|\boldsymbol{w}_j\|^3} = \frac{1}{\|\boldsymbol{w}_j\|}\mathcal{P}_{\boldsymbol{w}_j}\dot{\boldsymbol{w}}_j$, and (13) gives

$$\frac{d}{dt}\frac{\boldsymbol{w}_j}{\|\boldsymbol{w}_j\|} = \mathrm{sign}(v_j(0))\,\mathcal{P}_{\boldsymbol{w}_j}\sum_{(i,p):\langle \boldsymbol{w}_j,\, \boldsymbol{x}_i^{(p)}\rangle > 0}\Big(-\nabla_{\hat{y}}\ell\big(y_i, f(\boldsymbol{X}_i; W, \boldsymbol{v})\big)\Big)\boldsymbol{x}_i^{(p)}.$$

Subtracting $\operatorname{sign}(v_j(0))\mathcal{P}_{\boldsymbol{w}_j} x_a(\boldsymbol{w}_j)$ and using that $\mathcal{P}_{\boldsymbol{w}_j}$ is a contraction, we obtain, whenever $\max_i |f(\boldsymbol{X}_i; W, \boldsymbol{v})| \le 1$,

$$\begin{aligned}
&\left\| \frac{d}{dt}\frac{\boldsymbol{w}_j}{\|\boldsymbol{w}_j\|} - \operatorname{sign}(v_j(0))\mathcal{P}_{\boldsymbol{w}_j} x_a(\boldsymbol{w}_j) \right\| \\
&\quad = \left\| \mathcal{P}_{\boldsymbol{w}_j} \sum_{(i,p):\langle \boldsymbol{w}_j, \boldsymbol{x}_i^{(p)}\rangle > 0} \Big(-\nabla_{\hat{y}} \ell\big(y_i, f(\boldsymbol{X}_i; W, \boldsymbol{v})\big) - y_i \Big) \boldsymbol{x}_i^{(p)} \right\| \\
&\quad \le \sum_{i=1}^{n}\sum_{p=1}^{P} \left| -\nabla_{\hat{y}} \ell\big(y_i, f(\boldsymbol{X}_i; W, \boldsymbol{v})\big) - y_i \right| \left\| \boldsymbol{x}_i^{(p)} \right\| \\
&\quad \le \sum_{i=1}^{n}\sum_{p=1}^{P} 2\,|f(\boldsymbol{X}_i; W, \boldsymbol{v})| \left\| \boldsymbol{x}_i^{(p)} \right\| \qquad \text{(by Lemma 2)} \\
&\quad \le 2nPX_{\max} \max_{i\in[n]} |f(\boldsymbol{X}_i; W, \boldsymbol{v})| \,. \qquad \square
\end{aligned}$$

Combining Lemmas 4 and 5 shows that on the window $[0, T]$ the directional residual in (19) can be made smaller than any prescribed threshold by taking ϵ small. We record this in the form in which it is used in Appendices D and E.

Corollary 1 (Directional approximation on the small-output window)**.** *Let $\varrho > 0$. Consider any solution to the gradient flow (2) starting from the initialization (3) with*

$$\epsilon \;\le\; \min\left\{ \frac{1}{4\sqrt{h}\,PX_{\max}W_{\max}^2},\; \frac{\varrho}{4\sqrt{h}\,nP^2X_{\max}^2W_{\max}^2} \right\}. \tag{20}$$

Then for all $t \le T$ and all $j \in [h]$,

$$\left\| \frac{d}{dt}\frac{\boldsymbol{w}_j}{\|\boldsymbol{w}_j\|} - \operatorname{sign}(v_j(0))\,\mathcal{P}_{\boldsymbol{w}_j} x_a(\boldsymbol{w}_j) \right\| \;\le\; 4\epsilon\sqrt{h}\,nP^2X_{\max}^2W_{\max}^2 \;\le\; \varrho.$$

Proof. The first condition in (20) is (15), so by Lemma 4 we have $\max_i |f(\boldsymbol{X}_i; W(t), \boldsymbol{v}(t))| \le 2\epsilon\sqrt{h}PX_{\max}W_{\max}^2 \le \frac{1}{2}$ for all $t \le T$. Lemma 5 then bounds the residual by $2nPX_{\max} \cdot 2\epsilon\sqrt{h}PX_{\max}W_{\max}^2 = 4\epsilon\sqrt{h}\,nP^2X_{\max}^2W_{\max}^2$, which is at most ϱ by the second condition in (20). $\square$

C Geometry of the Signal Cone and Uniform Control of the Noise

This appendix collects the facts about the data on which the proofs in Appendices D and E rely. Appendix C.1 proves Lemma 1. Appendix C.2 derives from it the inequalities that quantify how strongly the signal patches push a filter toward $\boldsymbol{u}_\star$ (Lemma 6), shows that the signed activation field (12) splits into a signal part, determined by which signal patches the filter activates, and a noise part orthogonal to V_{sig} (Lemma 7), and records which signal patches a filter activates depending on its angle with $\boldsymbol{u}_\star$ (Lemma 8). All of this is deterministic. Appendix C.3 bounds the size G of the noise part under a probabilistic model of the data (Lemma 10, prepared by Lemma 9); this is the only place where such a model enters.

Throughout this appendix, the data follow Definition 1 and Assumption 3(a) holds, so that $\boldsymbol{u}_\star, \theta_\star, \mu_\star, \mu$ in (5) are well defined and Lemma 1 applies. Recall the notation $\boldsymbol{g}_i = y_i\boldsymbol{s}_i$, $\rho_i = \|\boldsymbol{g}_i\|$, $R_{\mathcal{A}} = \sum_{i\in\mathcal{A}} \rho_i$ and $\boldsymbol{z}_{\mathcal{A}} = \sum_{i\in\mathcal{A}} \boldsymbol{g}_i$ for $\mathcal{A} \subseteq [n]$, and $\angle(\boldsymbol{x}, \boldsymbol{y}) = \arccos\cos(\boldsymbol{x}, \boldsymbol{y})$.

C.1 Proof of Lemma 1

Lemma 1 (Common signal axis)**.** *Under Assumption 3(a), $\theta_\star < \pi/4$, so that $\mu_\star > 1/\sqrt{2}$ and $\mu > 0$, and $\cos(\boldsymbol{g}_i, \boldsymbol{u}_\star) \ge \mu_\star$ for all $i \in [n]$.*

Proof of Lemma 1. Since $\nu_0 < -\sin 2\alpha \le 0$, the vectors $\boldsymbol{u}_+$ and $-\boldsymbol{u}_-$ are not antipodal, and $\boldsymbol{u}_\star$ is the normalized sum of the two unit vectors $\boldsymbol{u}_+$ and $-\boldsymbol{u}_-$. Hence $\boldsymbol{u}_\star$ bisects the angle between them: it is a positive combination of $\boldsymbol{u}_+$ and $-\boldsymbol{u}_-$, so $\angle(\boldsymbol{u}_+, \boldsymbol{u}_\star)+\angle(\boldsymbol{u}_\star, -\boldsymbol{u}_-) = \angle(\boldsymbol{u}_+, -\boldsymbol{u}_-) = \pi - \beta_0$, and $\langle \boldsymbol{u}_+,\, \boldsymbol{u}_+ - \boldsymbol{u}_-\rangle = 1 - \nu_0 = \langle -\boldsymbol{u}_-,\, \boldsymbol{u}_+ - \boldsymbol{u}_-\rangle$ shows that the two angles are equal. Therefore

$$\angle(\boldsymbol{u}_+, \boldsymbol{u}_\star) = \angle(-\boldsymbol{u}_-, \boldsymbol{u}_\star) = \frac{\pi - \beta_0}{2}. \tag{21}$$

Let $i \in \mathcal{I}_+$. Then $\boldsymbol{g}_i = \boldsymbol{s}_i$, and Assumption 3(a) gives $\angle(\boldsymbol{g}_i, \boldsymbol{u}_+) \le \alpha$. By the triangle inequality and (21),

$$\angle(\boldsymbol{g}_i, \boldsymbol{u}_\star) \le \angle(\boldsymbol{g}_i, \boldsymbol{u}_+) + \angle(\boldsymbol{u}_+, \boldsymbol{u}_\star) \le \alpha + \frac{\pi - \beta_0}{2} = \theta_\star.$$

Let $i \in \mathcal{I}_-$. Then $\boldsymbol{g}_i = -\boldsymbol{s}_i$, and Assumption 3(a) gives $\angle(\boldsymbol{g}_i, -\boldsymbol{u}_-) = \angle(\boldsymbol{s}_i, \boldsymbol{u}_-) \le \alpha$, so the same argument with $-\boldsymbol{u}_-$ in place of $\boldsymbol{u}_+$ gives $\angle(\boldsymbol{g}_i, \boldsymbol{u}_\star) \le \theta_\star$.

Finally, since $\cos$ is decreasing on $[0, \pi]$ and $\frac{\pi}{2} + 2\alpha \in [\frac{\pi}{2}, \pi)$,

$$\theta_\star < \frac{\pi}{4} \iff \frac{\pi - \beta_0}{2} < \frac{\pi}{4} - \alpha \iff \beta_0 > \frac{\pi}{2} + 2\alpha \iff \nu_0 = \cos\beta_0 < \cos\left(\frac{\pi}{2} + 2\alpha\right) = -\sin 2\alpha,$$

and the last inequality is part of Assumption 3(a). Hence $\theta_\star < \pi/4$, so $\mu_\star = \cos\theta_\star > 1/\sqrt{2}$ and $\mu = \cos 2\theta_\star > 0$, and $\cos(\boldsymbol{g}_i, \boldsymbol{u}_\star) = \cos\angle(\boldsymbol{g}_i, \boldsymbol{u}_\star) \ge \cos\theta_\star = \mu_\star$ for every i. □

C.2 The Signal Cone

The first lemma converts the angular statement of Lemma 1 into the inner-product bounds used in the dynamics. For $\mathcal{A} \subseteq [n]$ we decompose $\boldsymbol{z}_\mathcal{A}$ along and orthogonally to $\boldsymbol{u}_\star$,

$$\boldsymbol{z}_\mathcal{A} = s_\mathcal{A}\, \boldsymbol{u}_\star + \boldsymbol{p}_\mathcal{A}, \qquad s_\mathcal{A} := \langle \boldsymbol{u}_\star,\, \boldsymbol{z}_\mathcal{A}\rangle, \qquad \boldsymbol{p}_\mathcal{A} := (I - \boldsymbol{u}_\star\boldsymbol{u}_\star^\top)\boldsymbol{z}_\mathcal{A} \perp \boldsymbol{u}_\star. \tag{22}$$

Lemma 6 (Inner-product bounds on the signal cone)**.** *For all $i, k \in [n]$ and all $\mathcal{A} \subseteq [n]$,*

(i) $\langle \boldsymbol{g}_i,\, \boldsymbol{g}_k\rangle \ge \mu\, \rho_i\rho_k$;

(ii) $\langle \boldsymbol{u}_\star,\, \boldsymbol{g}_i\rangle \ge \mu_\star\rho_i$ *and* $\big\|(I - \boldsymbol{u}_\star\boldsymbol{u}_\star^\top)\boldsymbol{g}_i\big\| \le \rho_i \sin\theta_\star$;

(iii) $s_\mathcal{A} \ge \mu_\star R_\mathcal{A}$, $\|\boldsymbol{p}_\mathcal{A}\| \le R_\mathcal{A}\sin\theta_\star$, *and* $\|\boldsymbol{z}_\mathcal{A}\| \le R_\mathcal{A}$.

Proof. By Lemma 1, $\angle(\boldsymbol{g}_i, \boldsymbol{u}_\star) \le \theta_\star < \pi/4$ for every i.

(ii) Since $\sin$ is increasing and $\cos$ is decreasing on $[0, \pi/2]$,

$$\langle \boldsymbol{u}_\star,\, \boldsymbol{g}_i\rangle = \rho_i\cos\angle(\boldsymbol{g}_i, \boldsymbol{u}_\star) \ge \rho_i\cos\theta_\star = \mu_\star\rho_i, \qquad \big\|(I - \boldsymbol{u}_\star\boldsymbol{u}_\star^\top)\boldsymbol{g}_i\big\| = \rho_i\sin\angle(\boldsymbol{g}_i, \boldsymbol{u}_\star) \le \rho_i\sin\theta_\star.$$

(i) By the triangle inequality, $\angle(\boldsymbol{g}_i, \boldsymbol{g}_k) \le \angle(\boldsymbol{g}_i, \boldsymbol{u}_\star) + \angle(\boldsymbol{u}_\star, \boldsymbol{g}_k) \le 2\theta_\star < \pi/2$, hence

$$\langle \boldsymbol{g}_i,\, \boldsymbol{g}_k\rangle = \rho_i\rho_k\cos\angle(\boldsymbol{g}_i, \boldsymbol{g}_k) \ge \rho_i\rho_k\cos 2\theta_\star = \mu\rho_i\rho_k.$$

(iii) Summing *(ii)* over $i \in \mathcal{A}$ and using the triangle inequality,

$$s_\mathcal{A} = \sum_{i\in\mathcal{A}}\langle \boldsymbol{u}_\star,\, \boldsymbol{g}_i\rangle \ge \mu_\star R_\mathcal{A}, \quad \|\boldsymbol{p}_\mathcal{A}\| \le \sum_{i\in\mathcal{A}}\big\|(I - \boldsymbol{u}_\star\boldsymbol{u}_\star^\top)\boldsymbol{g}_i\big\| \le R_\mathcal{A}\sin\theta_\star, \quad \|\boldsymbol{z}_\mathcal{A}\| \le \sum_{i\in\mathcal{A}}\rho_i = R_\mathcal{A}.$$

□

Note that in *(iii)* both the on-axis component $s_\mathcal{A}$ and the off-axis component $\|\boldsymbol{p}_\mathcal{A}\|$ scale with the same quantity $R_\mathcal{A}$; this is why no ratio $\max_i \rho_i/\rho_{\min}$ appears in our conditions.

Next we split the signed activation field (12) of a filter into its signal and noise parts. For $\boldsymbol{w} \in \mathbb{R}^D$, let

$$\mathcal{A}(\boldsymbol{w}) := \{i \in [n] : \langle \boldsymbol{w},\, \boldsymbol{s}_i\rangle > 0\} \tag{23}$$

be the set of inputs whose signal patch is activated by $\boldsymbol{w}$, and recall the noise field $\boldsymbol{g}(\boldsymbol{w})$ from (6).

Lemma 7 (Signal and noise parts of the activation field)**.** *Under Assumption 1, for every* $\boldsymbol{w} \in \mathbb{R}^D$*,*

$$x_a(\boldsymbol{w}) = \boldsymbol{z}_{\mathcal{A}(\boldsymbol{w})} + \boldsymbol{g}(\boldsymbol{w}), \qquad \boldsymbol{g}(\boldsymbol{w}) \in V_{\mathrm{sig}}^{\perp}.$$

In particular, $\langle \boldsymbol{u}_\star,\, x_a(\boldsymbol{w})\rangle = s_{\mathcal{A}(\boldsymbol{w})}$ *and* $\langle \boldsymbol{g}_i,\, x_a(\boldsymbol{w})\rangle = \langle \boldsymbol{g}_i,\, \boldsymbol{z}_{\mathcal{A}(\boldsymbol{w})}\rangle$ *for every* $i \in [n]$*.*

Proof. Split the sum in (12) into signal patches ($p = 1$) and noise patches ($p \geq 2$). The signal patch of input i is activated if and only if $i \in \mathcal{A}(\boldsymbol{w})$, and it contributes $y_i \boldsymbol{x}_i^{(1)} = y_i \boldsymbol{s}_i = \boldsymbol{g}_i$; hence the signal patches contribute $\boldsymbol{z}_{\mathcal{A}(\boldsymbol{w})}$. The noise patches contribute $\boldsymbol{g}(\boldsymbol{w})$ by (6), and $\boldsymbol{g}(\boldsymbol{w}) \in V_{\mathrm{sig}}^{\perp}$ because every $\boldsymbol{\xi}_i^{(p)}$ is orthogonal to V_{sig} by Assumption 1. The two identities follow since $\boldsymbol{u}_\star, \boldsymbol{g}_i \in V_{\mathrm{sig}}$. □

Finally, we record how the set $\mathcal{A}(\boldsymbol{w})$ of activated signal patches is constrained by the angle between $\boldsymbol{w}$ and $\boldsymbol{u}_\star$. Recall the signal cones $\mathcal{S}_+, \mathcal{S}_-$ from (4): $\boldsymbol{w} \in \mathcal{S}_+$ means $\mathcal{A}(\boldsymbol{w}) = \mathcal{I}_+$ with $\langle \boldsymbol{w},\, \boldsymbol{s}_i\rangle < 0$ for all $i \in \mathcal{I}_-$, and $\boldsymbol{w} \in \mathcal{S}_-$ means $\mathcal{A}(\boldsymbol{w}) = \mathcal{I}_-$ with $\langle \boldsymbol{w},\, \boldsymbol{s}_i\rangle < 0$ for all $i \in \mathcal{I}_+$. Both sets are open convex cones, $\mathcal{S}_+$ being the intersection of the open halfspaces $\{\langle \cdot,\, \boldsymbol{g}_i\rangle > 0\}$.

Lemma 8 (Activation pattern of the signal patches)**.** *Let* $\boldsymbol{w} \in \mathbb{R}^D \setminus \{0\}$ *and* $c := \cos(\boldsymbol{w}, \boldsymbol{u}_\star)$*.*

(a) *If* $c > \sin\theta_\star$*, then* $\boldsymbol{w} \in \mathcal{S}_+$*; in particular* $\mathcal{A}(\boldsymbol{w}) = \mathcal{I}_+$*.*

(b) *If* $c < -\sin\theta_\star$*, then* $\boldsymbol{w} \in \mathcal{S}_-$*; in particular* $\mathcal{A}(\boldsymbol{w}) = \mathcal{I}_-$*.*

(c) *If* $\mathcal{A}(\boldsymbol{w}) = \emptyset$*, then* $|c| \leq \sin\theta_\star$*.*

Proof. (a) If $c > \sin\theta_\star = \cos(\frac{\pi}{2} - \theta_\star)$, then $\angle(\boldsymbol{w}, \boldsymbol{u}_\star) < \frac{\pi}{2} - \theta_\star$, and by the triangle inequality and Lemma 1, for every $i \in [n]$,

$$\angle(\boldsymbol{w}, \boldsymbol{g}_i) \leq \angle(\boldsymbol{w}, \boldsymbol{u}_\star) + \angle(\boldsymbol{u}_\star, \boldsymbol{g}_i) < \frac{\pi}{2} - \theta_\star + \theta_\star = \frac{\pi}{2},$$

so $\langle \boldsymbol{w},\, \boldsymbol{g}_i\rangle = \|\boldsymbol{w}\|\, \rho_i \cos\angle(\boldsymbol{w}, \boldsymbol{g}_i) > 0$. *(b)* Apply *(a)* to $-\boldsymbol{w}$, using $\cos(-\boldsymbol{w}, \boldsymbol{u}_\star) = -c > \sin\theta_\star$. *(c)* Suppose $\mathcal{A}(\boldsymbol{w}) = \emptyset$, i.e. $\langle \boldsymbol{w},\, \boldsymbol{s}_i\rangle \leq 0$ for every i. If $c > \sin\theta_\star$, then *(a)* gives $\langle \boldsymbol{w},\, \boldsymbol{s}_i\rangle = \langle \boldsymbol{w},\, \boldsymbol{g}_i\rangle > 0$ for $i \in \mathcal{I}_+$, which is nonempty, a contradiction. If $c < -\sin\theta_\star$, then *(b)* gives $\langle \boldsymbol{w},\, \boldsymbol{s}_i\rangle = -\langle \boldsymbol{w},\, \boldsymbol{g}_i\rangle > 0$ for $i \in \mathcal{I}_-$, which is nonempty, again a contradiction. Hence $|c| \leq \sin\theta_\star$. □

C.3 Uniform Bound on the Noise Field

Since the filters are data-dependent, the noise field (6) has to be controlled uniformly over all directions, which is what the quantity G does. In this section we bound G under a probabilistic model of the data. The bound has two layers. The first uses only the randomness of the labels and expresses G through the spectral norm of the empirical second-moment matrix of the noise patches; the second bounds that spectral norm when the noise patches are independent and sub-Gaussian. We call a random vector $\boldsymbol{\xi} \in \mathbb{R}^D$ *σ-sub-Gaussian* if $\|\langle \boldsymbol{\xi},\, \boldsymbol{v}\rangle\|_{\psi_2} \leq \sigma$ for every unit vector $\boldsymbol{v}$, where $\|X\|_{\psi_2} := \inf\{t > 0 : \mathbb{E}\exp(X^2/t^2) \leq 2\}$ is the sub-Gaussian norm of a real random variable X.

Lemma 9 (Second-moment matrix of independent sub-Gaussian vectors)**.** *Let* $\zeta_1, \dots, \zeta_N \in \mathbb{R}^D$ *be independent σ-sub-Gaussian random vectors. There is an absolute constant* C_1 *such that for every* $\delta \in (0, 1)$*, with probability at least* $1 - \delta$*,*

$$\Big\| \sum_{k=1}^{N} \zeta_k \zeta_k^{\top} \Big\|_2 \;\leq\; C_1 \sigma^2 \big(N + D + \log(1/\delta)\big),$$

where $\|\cdot\|_2$ *denotes the spectral norm.*

Proof. Let $M := \sum_{k=1}^{N} \zeta_k \zeta_k^{\top}$ and let $\mathcal{N}$ be a $\frac{1}{4}$-net of $\mathbb{S}^{D-1}$ with $|\mathcal{N}| \leq 9^D$. Since M is symmetric, $\|M\|_2 \leq 2\max_{\boldsymbol{v} \in \mathcal{N}} \big|\boldsymbol{v}^{\top} M \boldsymbol{v}\big|$ (Vershynin, 2018, Exercise 4.4.3). Fix $\boldsymbol{v} \in \mathcal{N}$ and let $X_k := \langle \zeta_k,\, \boldsymbol{v}\rangle^2$, so that $\boldsymbol{v}^{\top} M \boldsymbol{v} = \sum_k X_k$. The X_k are independent, nonnegative, and sub-exponential with $\|X_k\|_{\psi_1} = \|\langle \zeta_k,\, \boldsymbol{v}\rangle\|_{\psi_2}^2 \leq \sigma^2$ (Vershynin, 2018, Lemma 2.7.6), and $\mathbb{E}X_k \leq C_2\sigma^2$

for an absolute constant C_2 (Vershynin, 2018, Proposition 2.7.1). By Bernstein's inequality (Vershynin, 2018, Theorem 2.8.1), there is an absolute constant $c > 0$ such that for every $u \geq 0$,

$$\mathbb{P}\Big(\sum_{k=1}^{N} (X_k - \mathbb{E}X_k) \geq u \Big) \leq \exp\Big(- c \min\Big\{ \frac{u^2}{N\sigma^4}, \frac{u}{\sigma^2} \Big\} \Big).$$

Let $s := D \log 9 + \log(1/\delta)$ and $u := \sigma^2(\sqrt{Ns/c} + s/c)$; then the right-hand side is at most $e^{-s} = 9^{-D}\delta$. A union bound over the net gives, with probability at least $1 - \delta$, simultaneously for all $\boldsymbol{v} \in \mathcal{N}$,

$$\boldsymbol{v}^\top M \boldsymbol{v} \leq C_2\sigma^2 N + \sigma^2\Big(\sqrt{Ns/c} + s/c\Big) \leq C_3\sigma^2(N + s),$$

using $\sqrt{Ns} \leq (N+s)/2$. Hence $\|M\|_2 \leq 2C_3\sigma^2(N + D\log 9 + \log(1/\delta))$, which is the claim. □

Lemma 10 (Uniform bound on the noise field)**.** *Let $N' := n(P-1)$, and let $\lambda_{\max}$ denote the spectral norm of the empirical second-moment matrix $\sum_{i\in[n]} \sum_{p=2}^{P} \boldsymbol{\xi}_i^{(p)} \boldsymbol{\xi}_i^{(p)\top}$ of the noise patches. Suppose that $N' \geq D$ and that the labels $y_1, \ldots, y_n$ are independent Rademacher variables, independent of the noise patches.*

(a) *For every $\delta \in (0,1)$, with probability at least $1-\delta$ over the labels (for any fixed realization of the noise patches),*

$$G \;\leq\; 2\sqrt{2(P-1)\lambda_{\max}\Big(D\log\frac{5eN'}{D} + \log\frac{1}{\delta}\Big)}.$$

(b) *If, in addition, the noise patches are independent and σ_ξ-sub-Gaussian and $N' \geq D + \log(2/\delta)$, then there is an absolute constant $C > 0$ such that with probability at least $1-\delta$ over the noise and the labels,*

$$G \;\leq\; C\,\sigma_\xi\,(P-1)\sqrt{n\Big(D\log\frac{5enP}{D} + \log\frac{2}{\delta}\Big)}.$$

Proof. Relabel the noise patches by $k = (i,p)$ with $p \geq 2$, so that k runs over $[N']$; write $\zeta_k := \boldsymbol{\xi}_i^{(p)}$ and $y(k) := y_i$ for the label of the input to which patch k belongs. Then $\boldsymbol{g}(\boldsymbol{w}) = \sum_{k=1}^{N'} \mathbf{1}_{\langle \boldsymbol{w},\, \zeta_k\rangle > 0}\, y(k)\, \zeta_k$, and $\lambda_{\max}$ is the spectral norm of $\sum_{k=1}^{N'} \zeta_k\zeta_k^\top$. We first prove *(a)*.

We first reduce the supremum over $\boldsymbol{w}$ to a maximum over a finite set of directions. Let $\mathcal{N}$ be a $\frac{1}{2}$-net of $\mathbb{S}^{D-1}$ with $|\mathcal{N}| \leq 5^D$. For every $\boldsymbol{x} \in \mathbb{R}^D$ one has $\|\boldsymbol{x}\| \leq 2\max_{\boldsymbol{v}\in\mathcal{N}} \langle \boldsymbol{v},\, \boldsymbol{x}\rangle$, hence

$$G \leq 2\max_{\boldsymbol{v}\in\mathcal{N}} \sup_{\boldsymbol{w}\in\mathbb{S}^{D-1}} \langle \boldsymbol{v},\, \boldsymbol{g}(\boldsymbol{w})\rangle = 2\max_{\boldsymbol{v}\in\mathcal{N}} \sup_{\boldsymbol{w}\in\mathbb{S}^{D-1}} \sum_{k=1}^{N'} \mathbf{1}_{\langle \boldsymbol{w},\, \zeta_k\rangle > 0}\, y(k) \langle \boldsymbol{v},\, \zeta_k\rangle. \tag{24}$$

Next we condition on the noise and count the possible activation patterns. Condition on the noise realization $\Xi := \{\zeta_k\}_{k\in[N']}$; by assumption, the labels remain independent Rademacher variables. As $\boldsymbol{w}$ ranges over $\mathbb{S}^{D-1}$, the activation pattern $s(\boldsymbol{w}) := (\mathbf{1}_{\langle \boldsymbol{w},\, \zeta_k\rangle > 0})_{k\in[N']}$, an element of $\{0,1\}^{N'}$, takes at most $(eN'/D)^D$ distinct values: homogeneous halfspaces in $\mathbb{R}^D$ have VC dimension at most D, so by the Sauer–Shelah lemma the number of patterns is at most $\sum_{j=0}^{D} \binom{N'}{j} \leq (eN'/D)^D$ for $N' \geq D$. Let $\Sigma = \Sigma(\Xi) \subseteq \{0,1\}^{N'}$ be the set of realized patterns. For fixed $\boldsymbol{v} \in \mathcal{N}$ and $s \in \Sigma$, grouping the patches by input (the label $y(k) = y_i$ is the same for all patches of input i) gives

$$\sum_{k=1}^{N'} s_k\, y(k) \langle \boldsymbol{v},\, \zeta_k\rangle = \sum_{i=1}^{n} y_i\, a_i(s, \boldsymbol{v}), \qquad a_i(s,\boldsymbol{v}) := \sum_{p\geq 2:\, s_{(i,p)}=1} \langle \boldsymbol{v},\, \boldsymbol{\xi}_i^{(p)}\rangle. \tag{25}$$

Conditionally on Ξ, the right-hand side is a Rademacher sum with fixed coefficients, hence sub-Gaussian with variance proxy $A(s,\boldsymbol{v})^2 := \sum_{i=1}^{n} a_i(s,\boldsymbol{v})^2$: by Hoeffding's inequality,

$$\mathbb{P}\Big(\sum_{i=1}^{n} y_i a_i(s,\boldsymbol{v}) \geq A(s,\boldsymbol{v})\,\tau \;\Big|\; \Xi \Big) \leq e^{-\tau^2/2}, \qquad \tau \geq 0. \tag{26}$$

We now take a union bound over the net and the patterns. Given Ξ, consider the family $\{\sum_i y_i a_i(s,\boldsymbol{v}) : \boldsymbol{v} \in \mathcal{N},\ s \in \Sigma\}$. It is a fixed finite collection of Rademacher sums, at most $K := |\mathcal{N}|\,|\Sigma| \le 5^D (eN'/D)^D$ of them, each with variance proxy at most $\bar{A} := \max_{\boldsymbol{v}\in\mathcal{N},\, s\in\Sigma} A(s,\boldsymbol{v})$. A union bound over this collection in (26), combined with (24), gives $\mathbb{P}(G \ge 2\bar{A}\tau \mid \Xi) \le K e^{-\tau^2/2}$. Choosing $\tau = \sqrt{2\log(K/\delta)}$, we obtain that conditionally on Ξ, with probability at least $1-\delta$,

$$G \le 2\bar{A}\sqrt{2\log(K/\delta)}, \qquad \log\frac{K}{\delta} \le D\log\frac{5eN'}{D} + \log\frac{1}{\delta}. \tag{27}$$

We stress that the union bound is over a collection that is finite and fixed once Ξ is given; no deterministic superset of the data-dependent family $\Sigma(\Xi)$ is needed.

It remains to bound the variance proxy, which we do deterministically. For every s, $\boldsymbol{v}$ and i, the sum defining $a_i(s,\boldsymbol{v})$ in (25) has at most $P-1$ terms, so by Cauchy–Schwarz $a_i(s,\boldsymbol{v})^2 \le (P-1)\sum_{p=2}^{P}\langle \boldsymbol{v},\, \boldsymbol{\xi}_i^{(p)}\rangle^2$. Summing over i,

$$A(s,\boldsymbol{v})^2 \le (P-1)\sum_{k=1}^{N'}\langle \boldsymbol{v},\, \zeta_k\rangle^2 = (P-1)\,\boldsymbol{v}^\top\Big(\sum_{k=1}^{N'}\zeta_k\zeta_k^\top\Big)\boldsymbol{v} \le (P-1)\,\lambda_{\max}.$$

This holds for all s and $\boldsymbol{v}$, hence $\bar{A} \le \sqrt{(P-1)\lambda_{\max}}$. Substituting into (27) proves *(a)*.

We turn to *(b)*. The vectors $\zeta_1,\dots,\zeta_{N'}$ are now independent and σ_ξ-sub-Gaussian (only the labels, not the noise vectors, are shared within an input), so Lemma 9, applied with $\delta/2$ in place of δ, gives with probability at least $1-\delta/2$ over the noise

$$\lambda_{\max} \le C_1\sigma_\xi^2\big(N' + D + \log(2/\delta)\big) \le 3C_1\sigma_\xi^2 N', \tag{28}$$

where the last inequality uses $N' \ge D + \log(2/\delta)$. On the intersection of this event with the event of *(a)* (applied with $\delta/2$), which has probability at least $1-\delta$, we combine *(a)* with (28), $N' = n(P-1)$ and $N' \le nP$:

$$\begin{aligned} G &\le 2\sqrt{2(P-1)\cdot 3C_1\sigma_\xi^2 N'\Big(D\log\frac{5eN'}{D} + \log\frac{2}{\delta}\Big)} \\ &\le 2\sqrt{6C_1}\,\sigma_\xi\,(P-1)\sqrt{n\Big(D\log\frac{5enP}{D} + \log\frac{2}{\delta}\Big)}, \end{aligned}$$

which is *(b)* with $C := 2\sqrt{6C_1}$. □

D Proof of Theorem 1: Early Alignment Phase

We break the proof of Theorem 1 into two parts. In this appendix we prove the first part, on directional convergence during the early phase of training; in Appendix E we prove the second part, on final convergence. Appendix D.1 collects the auxiliary lemmas: a lower bound on the derivative of $\cos(\boldsymbol{w}_j,\boldsymbol{u}_\star)$ (Lemma 11), the dynamics of the signal coordinates $\langle \boldsymbol{w}_j/\|\boldsymbol{w}_j\|, \boldsymbol{g}_i\rangle$ (Lemma 12), the fact that these coordinates can cross zero in one direction only (Lemma 13), and lower bounds on the drift of the cosine in each region of the sphere (Lemma 14). Appendix D.2 proves the first part of Theorem 1, and Appendix D.3 records what the same argument gives for neurons that do not activate a positive signal patch at initialization.

Standing conventions. Throughout this appendix, Assumptions 1–4 hold, and ϵ satisfies (11). In particular ϵ satisfies (20) with $\varrho = \varrho_1$, so Corollary 1 gives, for all $t \le T$ and all $j \in [h]$,

$$\frac{d}{dt}\frac{\boldsymbol{w}_j}{\|\boldsymbol{w}_j\|} = \operatorname{sign}(v_j(0))\,\mathcal{P}_{\boldsymbol{w}_j}x_a(\boldsymbol{w}_j) + \boldsymbol{r}_j(t), \qquad \|\boldsymbol{r}_j(t)\| \le \varrho_1. \tag{29}$$

We fix a neuron $j \in \mathcal{V}_+$, so that $\operatorname{sign}(v_j(0)) = +1$; the case $j \in \mathcal{V}_-$ is treated at the end of Appendix D.2. To lighten notation we drop the index j and write $\boldsymbol{w}(t) = \boldsymbol{w}_j(t)$, $\boldsymbol{r}(t) = \boldsymbol{r}_j(t)$, and

$$c(t) := \cos(\boldsymbol{w}(t),\boldsymbol{u}_\star), \qquad a_i(t) := \Big\langle \frac{\boldsymbol{w}(t)}{\|\boldsymbol{w}(t)\|}, \boldsymbol{g}_i\Big\rangle = \rho_i\cos(\boldsymbol{w}(t),\boldsymbol{g}_i), \qquad \mathcal{A}(t) := \mathcal{A}(\boldsymbol{w}(t)), \tag{30}$$

with $\mathcal{A}(\cdot)$ from (23). Since $\langle \boldsymbol{w},\, \boldsymbol{s}_i\rangle = y_i\langle \boldsymbol{w},\, \boldsymbol{g}_i\rangle$,

$$\mathcal{A}(t) = \{i \in [n] : y_i a_i(t) > 0\} = \{i \in \mathcal{I}_+ : a_i(t) > 0\} \cup \{i \in \mathcal{I}_- : a_i(t) < 0\}, \tag{31}$$

that is, a positive signal patch is activated when its coordinate is positive, and a negative signal patch is activated when its coordinate is negative. Recall that $\boldsymbol{w}(t) \neq 0$ for all t (Lemma 3), so c and a_i are well defined and absolutely continuous, and that $s_{\mathcal{A}}, \boldsymbol{p}_{\mathcal{A}}$ denote the components of $\boldsymbol{z}_{\mathcal{A}}$ along and orthogonal to $\boldsymbol{u}_\star$, see (22).

D.1 Auxiliary Lemmata

Lemma 11 (Lower bound on the derivative of the cosine)**.** *For all $t \le T$,*

$$\dot{c} \;\ge\; s_{\mathcal{A}(t)}\big(1 - c^2\big) - |c|\,\big(\,\big\|\boldsymbol{p}_{\mathcal{A}(t)}\big\| + G\big) - \varrho_1. \tag{32}$$

Proof. Write $\mathcal{A} = \mathcal{A}(t)$. Taking the inner product of (29) with $\boldsymbol{u}_\star$ and using $\langle \boldsymbol{u}_\star,\, \boldsymbol{w}/\,\|\boldsymbol{w}\|\rangle = c$,

$$\dot{c} = \langle \boldsymbol{u}_\star,\, \mathcal{P}_{\boldsymbol{w}} x_a(\boldsymbol{w})\rangle + \langle \boldsymbol{u}_\star,\, \boldsymbol{r}\rangle = \langle \boldsymbol{u}_\star,\, x_a(\boldsymbol{w})\rangle - c\,\Big\langle \frac{\boldsymbol{w}}{\|\boldsymbol{w}\|}, x_a(\boldsymbol{w})\Big\rangle + \langle \boldsymbol{u}_\star,\, \boldsymbol{r}\rangle.$$

By Lemma 7 and (22), $\langle \boldsymbol{u}_\star,\, x_a(\boldsymbol{w})\rangle = s_{\mathcal{A}}$ and

$$\Big\langle \frac{\boldsymbol{w}}{\|\boldsymbol{w}\|}, x_a(\boldsymbol{w})\Big\rangle = \Big\langle \frac{\boldsymbol{w}}{\|\boldsymbol{w}\|}, s_{\mathcal{A}}\boldsymbol{u}_\star + \boldsymbol{p}_{\mathcal{A}} + \boldsymbol{g}(\boldsymbol{w})\Big\rangle = s_{\mathcal{A}} c + \Big\langle \frac{\boldsymbol{w}}{\|\boldsymbol{w}\|}, \boldsymbol{p}_{\mathcal{A}}\Big\rangle + \Big\langle \frac{\boldsymbol{w}}{\|\boldsymbol{w}\|}, \boldsymbol{g}(\boldsymbol{w})\Big\rangle.$$

Hence, by Cauchy–Schwarz, $\|\boldsymbol{g}(\boldsymbol{w})\| \le G$ and $\|\boldsymbol{r}\| \le \varrho_1$,

$$\dot{c} = s_{\mathcal{A}}(1-c^2) - c\,\Big\langle \frac{\boldsymbol{w}}{\|\boldsymbol{w}\|}, \boldsymbol{p}_{\mathcal{A}}\Big\rangle - c\,\Big\langle \frac{\boldsymbol{w}}{\|\boldsymbol{w}\|}, \boldsymbol{g}(\boldsymbol{w})\Big\rangle + \langle \boldsymbol{u}_\star,\, \boldsymbol{r}\rangle \ge s_{\mathcal{A}}(1-c^2) - |c|\,\big(\,\|\boldsymbol{p}_{\mathcal{A}}\| + G\big) - \varrho_1. \;\square$$

Lemma 12 (Dynamics of the signal coordinates)**.** *For every $i \in [n]$ and all $t \le T$,*

$$\dot{a}_i = \langle \boldsymbol{z}_{\mathcal{A}(t)},\, \boldsymbol{g}_i\rangle - a_i\,\Big\langle \frac{\boldsymbol{w}}{\|\boldsymbol{w}\|}, x_a(\boldsymbol{w})\Big\rangle + \langle \boldsymbol{r},\, \boldsymbol{g}_i\rangle, \qquad |\langle \boldsymbol{r},\, \boldsymbol{g}_i\rangle| \le \varrho_1 \rho_i, \quad \Big|\Big\langle \frac{\boldsymbol{w}}{\|\boldsymbol{w}\|}, x_a(\boldsymbol{w})\Big\rangle\Big| \le R + G. \tag{33}$$

Moreover, if $\mathcal{A}(t) \neq \emptyset$, then for every $i \in [n]$,

$$\langle \boldsymbol{z}_{\mathcal{A}(t)},\, \boldsymbol{g}_i\rangle \;\ge\; \mu\,\rho_i\,R_{\mathcal{A}(t)} \;\ge\; \mu\,\rho_i\,\rho_{\min}. \tag{34}$$

Proof. Write $\mathcal{A} = \mathcal{A}(t)$. Taking the inner product of (29) with $\boldsymbol{g}_i$,

$$\dot{a}_i = \langle \mathcal{P}_{\boldsymbol{w}} x_a(\boldsymbol{w}),\, \boldsymbol{g}_i\rangle + \langle \boldsymbol{r},\, \boldsymbol{g}_i\rangle = \langle x_a(\boldsymbol{w}),\, \boldsymbol{g}_i\rangle - \Big\langle \frac{\boldsymbol{w}}{\|\boldsymbol{w}\|}, x_a(\boldsymbol{w})\Big\rangle\Big\langle \frac{\boldsymbol{w}}{\|\boldsymbol{w}\|}, \boldsymbol{g}_i\Big\rangle + \langle \boldsymbol{r},\, \boldsymbol{g}_i\rangle,$$

and $\langle x_a(\boldsymbol{w}),\, \boldsymbol{g}_i\rangle = \langle \boldsymbol{z}_{\mathcal{A}},\, \boldsymbol{g}_i\rangle$ by Lemma 7; this is (33). The two bounds follow from Cauchy–Schwarz, since $\|\boldsymbol{g}_i\| = \rho_i$ and, by Lemma 7 and Lemma 6(iii), $\|x_a(\boldsymbol{w})\| \le \|\boldsymbol{z}_{\mathcal{A}}\| + \|\boldsymbol{g}(\boldsymbol{w})\| \le R_{\mathcal{A}} + G \le R + G$. Finally, if $\mathcal{A} \neq \emptyset$, then by Lemma 6(i),

$$\langle \boldsymbol{z}_{\mathcal{A}},\, \boldsymbol{g}_i\rangle = \sum_{k\in\mathcal{A}} \langle \boldsymbol{g}_k,\, \boldsymbol{g}_i\rangle \ge \mu\rho_i \sum_{k\in\mathcal{A}} \rho_k = \mu\rho_i R_{\mathcal{A}} \ge \mu\rho_i\rho_{\min},$$

where the last step uses that $\mathcal{A}$ contains at least one index. $\square$

The next lemma shows that, as long as the neuron activates at least one signal patch, its signal activation pattern can change in one direction only.

Lemma 13 (One-way switching of the signal activation pattern)**.** *Let $0 \le t_a \le t_b \le T$.*

(a) *If $i \in \mathcal{I}_+$ and $a_i(t_a) > 0$, then $a_i(t) > 0$ for all $t \in [t_a, T]$.*

(b) *If $\mathcal{A}(t) \neq \emptyset$ for all $t \in [t_a, t_b]$, then for every $i \in [n]$, $a_i(t_a) \ge 0$ implies $a_i(t) \ge 0$ for all $t \in [t_a, t_b]$.*

In words: a positive signal patch, once activated, stays activated (a)*; and as long as the neuron activates at least one signal patch, a positive signal patch cannot be deactivated and a negative signal patch cannot be activated* (b)*.*

Proof. Both parts rest on the following observation. Let $\varsigma_i := \mu\rho_i\rho_{\min}/(4(R+G)) > 0$. If $t \le T$, $\mathcal{A}(t) \neq \emptyset$ and $|a_i(t)| \le \varsigma_i$, then by (33), (34) and $\varrho_1 \le \mu\rho_{\min}/2$,

$$\dot{a}_i \ge \mu\rho_i\rho_{\min} - \varsigma_i(R+G) - \varrho_1\rho_i \ge \mu\rho_i\rho_{\min}\Big(1 - \frac{1}{4} - \frac{1}{2}\Big) = \frac{\mu\rho_i\rho_{\min}}{4} > 0. \tag{35}$$

(a) Suppose, for contradiction, that $a_i(t) \le 0$ for some $t \in [t_a, T]$, and let $\tau := \inf\{t \in [t_a, T] : a_i(t) \le 0\}$. By continuity, $a_i(\tau) = 0$, $a_i > 0$ on $[t_a, \tau)$, and $\tau > t_a$. Choose $\delta \in (0, \tau - t_a]$ such that $a_i \le \varsigma_i$ on $[\tau-\delta, \tau]$. For $t \in [\tau-\delta, \tau)$ we have $a_i(t) \in (0, \varsigma_i]$, hence $i \in \mathcal{A}(t)$ by (31) (recall $i \in \mathcal{I}_+$), so $\mathcal{A}(t) \neq \emptyset$ and (35) applies. Integrating over $[\tau-\delta, \tau]$ gives $a_i(\tau) \ge a_i(\tau-\delta) + \delta\,\mu\rho_i\rho_{\min}/4 > 0$, which contradicts $a_i(\tau) = 0$.

(b) Suppose $a_i(t_a) \ge 0$ and, for contradiction, that $a_i(t) < 0$ for some $t \in [t_a, t_b]$; let $\tau := \inf\{t \in [t_a, t_b] : a_i(t) < 0\}$. Then $a_i(\tau) \ge 0$ (either $\tau = t_a$, or $a_i \ge 0$ on $[t_a, \tau)$ and a_i is continuous), while there are $t_k \downarrow \tau$ with $a_i(t_k) < 0$, so $a_i(\tau) = 0$. Choose $\delta > 0$ with $\tau + \delta \le t_b$ and $|a_i| \le \varsigma_i$ on $[\tau, \tau+\delta]$. On this interval $\mathcal{A}(t) \neq \emptyset$ by hypothesis, so (35) applies and $a_i(t) \ge a_i(\tau) + (t-\tau)\mu\rho_i\rho_{\min}/4 \ge 0$ for all $t \in [\tau, \tau+\delta]$, which contradicts $a_i(t_k) < 0$ for $t_k \in (\tau, \tau+\delta]$. □

Note where the heterogeneity of the signal norms went: by (34), the coordinate of a light generator (ρ_i small) is pushed more slowly, but it also has proportionally less to overcome; the only requirement is $\varrho_1 \le \mu\rho_{\min}/2$, i.e. the residual must not overwhelm the coherence of the lightest generator with the activated ones.

For the fixed neuron, define the first time at which it activates a positive signal patch,

$$\tau^{\mathrm{act}} := \inf\Big\{t \in [0,T] : \max_{i \in \mathcal{I}_+} a_i(t) > 0\Big\} \qquad (\inf\emptyset := +\infty). \tag{36}$$

Lemma 13(a) implies that for every $t \in (\tau^{\mathrm{act}}, T]$ some positive signal patch is activated: indeed, there is $s \in [\tau^{\mathrm{act}}, t)$ with $a_i(s) > 0$ for some $i \in \mathcal{I}_+$, and then $a_i(t) > 0$. Combining this with Lemma 8(a),(b) (a filter with $c > \sin\theta_\star$ activates every positive signal patch; one with $c < -\sin\theta_\star$ activates none), we obtain

$$\mathcal{A}(t) \neq \emptyset \ \text{ and } \ c(t) \ge -\sin\theta_\star \quad \text{for } t \in (\tau^{\mathrm{act}}, T], \qquad c(t) \le \sin\theta_\star \quad \text{for } t \in [0, \tau^{\mathrm{act}}) \cap [0, T]. \tag{37}$$

We now turn the general bound (32) into explicit drift bounds in each region of the sphere. Set

$$\underline{c} := \min\Big\{-1 + \frac{\zeta_0}{2},\ -2\sin\theta_\star\Big\}, \qquad \beta_1 := \frac{\kappa}{2} R_{\min}\mu_\star, \qquad \beta_2 := \frac{\rho_{\min}}{3}, \qquad \beta_0 := \sin\theta_\star\, G + \varrho_1.$$

Note that $\underline{c} < -\sin\theta_\star$, that $1 + \underline{c} = \min\{\zeta_0/2,\ 1 - 2\sin\theta_\star\} \ge \kappa$ by $\kappa := \min\Big\{\frac{\zeta_0}{2},\ 1 - 2\sin\theta_\star,\ 1 - c^{*2}\Big\}$, and that $c(0) \ge -1 + \zeta_0 > \underline{c}$ by Assumption 2.

Lemma 14 (Drift of the cosine in each region)**.** *For all $t \le T$:*

(i) *if $c(t) \in [\underline{c}, -\sin\theta_\star)$, then $\dot{c} \ge \beta_1$;*

(ii) *if $|c(t)| \le \sin\theta_\star$ and $\mathcal{A}(t) \neq \emptyset$, then $\dot{c} \ge \beta_2$;*

(iii) *if $c(t) \in (\sin\theta_\star, c^*]$, then $\dot{c} \ge \beta_1$;*

(iv) *if $\mathcal{A}(t) = \emptyset$, then $\dot{c} \ge -\beta_0$, where $\beta_0 \le \rho_{\min}/24 = \beta_2/8$.*

Moreover, $c(t) > \underline{c}$ for all $t \in [0, T]$, and if $c(t_a) \ge c^*$ for some $t_a \le T$, then $c(t) \ge c^*$ for all $t \in [t_a, T]$.

Proof. All four bounds instantiate (32) with Lemma 6(iii), using $\varrho_1 \le \min\{\rho_{\min}/48,\ \kappa R_{\min}\mu_\star/8\}$ from (9) and $G \le \kappa R_{\min}\mu_\star/8$, $G\sin\theta_\star \le \rho_{\min}/48$ from Assumption 4.

(i) Since $c < -\sin\theta_\star$, Lemma 8(b) gives $\mathcal{A}(t) = \mathcal{I}_-$, so Lemma 6(iii) gives $s_{\mathcal{A}(t)} \ge \mu_\star R_-$ and $\|\boldsymbol{p}_{\mathcal{A}(t)}\| \le R_- \sin\theta_\star$. On $[\underline{c}, -\sin\theta_\star)$ we have $1 - c^2 = (1-c)(1+c) \ge 1 + \underline{c} \ge \kappa$ and $|c| \le 1$,

so

$$\begin{aligned}\dot c &\ge \kappa\mu_\star R_- - R_-\sin\theta_\star - G - \varrho_1 \\ &\ge \kappa\mu_\star R_- - \frac{\kappa}{4}\mu_\star R_- - \frac{\kappa}{4}R_{\min}\mu_\star \qquad \text{(by Assumption 3(b): } \sin\theta_\star = \mu_\star\tan\theta_\star \le \tfrac{\kappa}{4}\mu_\star\text{)} \\ &\ge \frac{\kappa}{2}R_{\min}\mu_\star = \beta_1. \qquad \text{(since } R_- \ge R_{\min}\text{)}\end{aligned}$$

(ii) Here $1-c^2 \ge 1-\sin^2\theta_\star = \mu_\star^2$ and $|c| \le \sin\theta_\star$, and Lemma 6(iii) gives $s_{\mathcal{A}(t)} \ge \mu_\star R_{\mathcal{A}(t)}$ and $\|\boldsymbol{p}_{\mathcal{A}(t)}\| \le R_{\mathcal{A}(t)}\sin\theta_\star$. Hence

$$\dot c \ge \mu_\star^3 R_{\mathcal{A}(t)} - \sin\theta_\star\big(R_{\mathcal{A}(t)}\sin\theta_\star + G\big) - \varrho_1 = R_{\mathcal{A}(t)}\big(\cos^3\theta_\star - \sin^2\theta_\star\big) - \sin\theta_\star\, G - \varrho_1.$$

Assumption 3(b) gives $\tan\theta_\star \le \frac{1}{4}$, hence $\theta_\star < \pi/6$ and $\cos^3\theta_\star - \sin^2\theta_\star \ge (\sqrt{3}/2)^3 - \frac{1}{4} > 0.399$. Since $\mathcal{A}(t) \ne \emptyset$ we have $R_{\mathcal{A}(t)} \ge \rho_{\min}$, and $\sin\theta_\star\, G \le \rho_{\min}/48$, $\varrho_1 \le \rho_{\min}/48$; therefore $\dot c \ge 0.399\,\rho_{\min} - \rho_{\min}/24 \ge \rho_{\min}/3 = \beta_2$.

(iii) Since $c > \sin\theta_\star$, Lemma 8(a) gives $\mathcal{A}(t) = \mathcal{I}_+$, and on $(\sin\theta_\star, c^*]$ we have $1-c^2 \ge 1-c^{*2} \ge \kappa$. The computation of *(i)* with R_+ in place of R_- gives $\dot c \ge \beta_1$.

(iv) If $\mathcal{A}(t) = \emptyset$, then $\boldsymbol{z}_{\mathcal{A}(t)} = 0$, so $s_{\mathcal{A}(t)} = 0$ and $\boldsymbol{p}_{\mathcal{A}(t)} = 0$, and (32) reads $\dot c \ge -|c|\,G - \varrho_1$. By Lemma 8(c), $|c| \le \sin\theta_\star$, hence $\dot c \ge -\sin\theta_\star\, G - \varrho_1 = -\beta_0$, and $\beta_0 \le \rho_{\min}/48 + \rho_{\min}/48 = \rho_{\min}/24 = \beta_2/8$.

We next show that the cosine never reaches $\underline{c}$. Suppose, for contradiction, that $c(t) < \underline{c}$ for some $t \le T$, and let $\tau := \inf\{t \in [0,T] : c(t) < \underline{c}\}$. Since $c(0) > \underline{c}$, we have $\tau > 0$, $c(\tau) = \underline{c}$ and $c \ge \underline{c}$ on $[0,\tau]$. As $\underline{c} < -\sin\theta_\star$, by continuity there is $\delta \in (0,\tau]$ with $c(t) \in [\underline{c}, -\sin\theta_\star)$ for all $t \in [\tau-\delta,\tau]$, so $\dot c \ge \beta_1$ on this interval by *(i)*, and $c(\tau) \ge c(\tau-\delta) + \beta_1\delta \ge \underline{c} + \beta_1\delta > \underline{c}$, a contradiction.

Finally, we show that the set $\{c \ge c^*\}$ is forward invariant. Suppose $c(t_a) \ge c^*$ and, for contradiction, $c(t_b) < c^*$ for some $t_b \in (t_a, T]$. Let $\sigma := \sup\{t \in [t_a,t_b] : c(t) \ge c^*\}$; then $c(\sigma) = c^*$, $\sigma < t_b$, and $c < c^*$ on $(\sigma, t_b]$. By continuity there is $\delta \in (0, t_b-\sigma]$ with $c(t) \in (\sin\theta_\star, c^*)$ for all $t \in (\sigma,\sigma+\delta]$, so $\dot c \ge \beta_1$ on $[\sigma,\sigma+\delta]$ by *(iii)*, and $c(\sigma+\delta) \ge c(\sigma) + \beta_1\delta > c^*$, contradicting $c(\sigma+\delta) < c^*$. □

D.2 Proof of the First Part of Theorem 1

Proof of Theorem 1, Part 1. We first verify that the choice of ϵ in (11) provides everything the argument needs. The first condition in (11) is (15), and the first two conditions together are (20) with $\varrho = \varrho_1$; hence Lemma 4 and Corollary 1 apply on $[0,T]$, and (29) holds. The third condition in (11) is equivalent to $\log\frac{1}{\sqrt{h}\epsilon} \ge 4nPX_{\max}t_1$, i.e. to $T \ge t_1$ by $T := \frac{1}{4nPX_{\max}}\log\frac{1}{\sqrt{h}\,\epsilon}$. Thus the whole argument below takes place inside the window $[0,T]$ on which (29) is valid.

We now fix one neuron. Let $j \in \mathcal{V}_+$ with $\max_{i\in\mathcal{I}_+}\langle \boldsymbol{w}_j(0),\, \boldsymbol{s}_i\rangle > 0$ and use the notation (30)–(36). The hypothesis says $\max_{i\in\mathcal{I}_+} a_i(0) > 0$, i.e. $\tau^{\mathrm{act}} = 0$, so by (37) the neuron activates at least one signal patch at every $t \in (0,T]$. We must show $c(t) \ge c^*$ for all $t \in [t_1,T]$; membership in $\mathcal{S}_+$ then follows from Lemma 8(a), since $c^* > \sin\theta_\star$.

The argument tracks c through a potential function that rescales time by the drift available in each region. Let $\varphi : [\underline{c}, c^*] \to [0,\infty)$ be the continuous, piecewise linear function with $\varphi(\underline{c}) = 0$ and

$$\varphi'(c) = \begin{cases} 1/\beta_1, & c \in [\underline{c}, -\sin\theta_\star) \cup (\sin\theta_\star, c^*], \\ 1/\beta_2, & c \in [-\sin\theta_\star, \sin\theta_\star]. \end{cases}$$

Thus φ is strictly increasing, and since the two outer intervals have total length at most 2,

$$\varphi(c^*) \le \frac{2}{\beta_1} + \frac{2\sin\theta_\star}{\beta_2} = \frac{4}{\kappa R_{\min}\mu_\star} + \frac{6\sin\theta_\star}{\rho_{\min}} = t_1. \tag{38}$$

By Lemma 14, $c(t) > \underline{c}$ for all $t \le T$, so $\varphi(c(t))$ is defined whenever $c(t) \le c^*$; as φ is Lipschitz and c is absolutely continuous, $t \mapsto \varphi(c(t))$ is absolutely continuous on any interval on which $c \le c^*$,

and $\frac{d}{dt}\varphi(c(t)) = \varphi'(c(t))\,\dot{c}(t)$ for almost every t at which $c(t) \notin \{\pm\sin\theta_\star\}$, while $\frac{d}{dt}\varphi(c(t)) = \dot{c}(t) = 0$ for almost every t at which $c(t) \in \{\pm\sin\theta_\star\}$. Lemma 14 now gives the two facts we use, valid for almost every $t \le T$ with $c(t) < c^*$:

$$\text{if } \mathcal{A}(t) \neq \emptyset, \ \text{ then } \ \frac{d}{dt}\varphi(c(t)) \ge 1, \tag{39}$$

because in the outer intervals $\dot{c} \ge \beta_1$ and $\varphi' = 1/\beta_1$ (parts (i),(iii)), in $[-\sin\theta_\star, \sin\theta_\star]$ we have $\dot{c} \ge \beta_2$ and $\varphi' = 1/\beta_2$ (part (ii)), and the levels $c = \pm\sin\theta_\star$ carry no time since $\dot{c} \ge \beta_2 > 0$ there; and

$$\text{if } \mathcal{A}(t) = \emptyset, \ \text{ then } \ \frac{d}{dt}\varphi(c(t)) \ge -\frac{1}{8}, \tag{40}$$

because then $|c(t)| \le \sin\theta_\star$ by Lemma 8(c), so $\varphi' = 1/\beta_2$ and $\frac{d}{dt}\varphi(c(t)) \ge -\beta_0/\beta_2 \ge -\frac{1}{8}$ (part (iv)).

We can now show that the neuron aligns within time t_1. Suppose, for contradiction, that $c(t) < c^*$ for some $t \in [t_1, T]$. By the forward invariance of $\{c \ge c^*\}$ (Lemma 14), $c(s) < c^*$ for all $s \in [0, t]$. Since the neuron activates a signal patch at every $s \in (0, t]$, (39) applies on $(0, t]$, and integrating gives

$$t \le \varphi(c(t)) - \varphi(c(0)) \le \varphi(c(t)) < \varphi(c^*) \le t_1 \le t,$$

where we used $\varphi \ge 0$, the strict monotonicity of φ and (38). This is a contradiction. Hence $c(t) \ge c^*$ for all $t \in [t_1, T]$, and $\boldsymbol{w}(t) \in \mathcal{S}_+$ on this interval by Lemma 8(a).

Finally we treat the neurons in $\mathcal{V}_-$. For $j \in \mathcal{V}_-$ we have $\operatorname{sign}(v_j(0)) = -1$ in (29). Flipping all labels, $y_i \mapsto -y_i$, exchanges $\mathcal{I}_+$ and $\mathcal{I}_-$, replaces $\boldsymbol{g}_i$ by $-\boldsymbol{g}_i$, $\boldsymbol{u}_\pm$ by $\boldsymbol{u}_\mp$ and hence $\boldsymbol{u}_\star$ by $-\boldsymbol{u}_\star$, and replaces $x_a(\boldsymbol{w})$ by $-x_a(\boldsymbol{w})$; thus (29) for a neuron in $\mathcal{V}_-$ is (29) for a neuron in $\mathcal{V}_+$ under the flipped labels. Definition 1 and Assumptions 1–4 are invariant under the flip (the margin condition of Assumption 2 for $\mathcal{V}_-$ becomes the one for $\mathcal{V}_+$), and $\mathcal{S}_+$ becomes $\mathcal{S}_-$. The argument above therefore applies verbatim and gives $\cos(\boldsymbol{w}_j(t), -\boldsymbol{u}_\star) \ge c^*$ and $\boldsymbol{w}_j(t) \in \mathcal{S}_-$ for all $t \in [t_1, T]$ whenever $\max_{i\in\mathcal{I}_-}\langle \boldsymbol{w}_j(0),\, \boldsymbol{s}_i\rangle > 0$. This finishes the proof of Part 1.. □

D.3 Neurons that Activate No Positive Signal Patch at Initialization

The proof above used the hypothesis $\max_{i\in\mathcal{I}_+}\langle \boldsymbol{w}_j(0),\, \boldsymbol{s}_i\rangle > 0$ only through $\tau^{\mathrm{act}} = 0$. The same argument describes an arbitrary neuron in $\mathcal{V}_+$: it aligns within time t_1 after it first activates a positive signal patch, and before that time it can activate signal patches only during a bounded amount of time. For a neuron $j \in \mathcal{V}_+$ let

$$\mathrm{act}_j(t) := \mathrm{Leb}\left\{s \in [0, t] : \mathcal{A}_j(s) \neq \emptyset\right\}$$

denote the total time in $[0, t]$ during which it activates at least one signal patch, where $\mathcal{A}_j(s) = \mathcal{A}(\boldsymbol{w}_j(s))$, and let τ_j^{act} be its first activation time (36).

Proposition 1. *Under the hypotheses of Theorem 1, every neuron $j \in \mathcal{V}_+$ satisfies the following, with $c_j(t) = \cos(\boldsymbol{w}_j(t), \boldsymbol{u}_\star)$.*

(a) *If $\tau_j^{\mathrm{act}} < T$, then $c_j(t) \ge c^*$, and hence $\boldsymbol{w}_j(t) \in \mathcal{S}_+$, for all $t \in [\tau_j^{\mathrm{act}} + t_1, T]$.*

(b) *For all $t \le \min\{\tau_j^{\mathrm{act}}, T\}$, $\mathrm{act}_j(t) \le t_1 + t/8$.*

(c) *For all $t \in [t_1, T]$, either $c_j(t) \ge c^*$, or $\tau_j^{\mathrm{act}} > t - t_1$ and the neuron activates no signal patch during a subset of $[0, t - t_1]$ of measure at least $\frac{7}{8}(t - t_1) - t_1$.*

The symmetric statements hold for $j \in \mathcal{V}_-$, with $\boldsymbol{u}_\star$ replaced by $-\boldsymbol{u}_\star$ and the roles of $\mathcal{I}_+$ and $\mathcal{I}_-$ exchanged.

Proof. Fix $j \in \mathcal{V}_+$, drop the index j, and use the notation and the potential φ of Appendix D.2; in particular (39) and (40) hold for almost every $t \le T$ with $c(t) < c^*$.

(a) Write $\tau = \tau^{\mathrm{act}}$ and suppose, for contradiction, that $c(t) < c^*$ for some $t \in [\tau + t_1, T]$. By the forward invariance of $\{c \ge c^*\}$, $c(s) < c^*$ for all $s \in [0, t]$. By (37) the neuron activates a signal patch at every $s \in (\tau, t]$, so (39) applies on $(\tau, t]$ and

$$t - \tau \le \varphi(c(t)) - \varphi(c(\tau)) \le \varphi(c(t)) < \varphi(c^*) \le t_1 \le t - \tau,$$

a contradiction. Membership in $\mathcal{S}_+$ follows from Lemma 8(a).

(b) Fix $t \le \min\{\tau^{\mathrm{act}}, T\}$. By (37) and continuity, $c(s) \le \sin\theta_\star < c^*$ for all $s \in [0,t]$, so (39) and (40) apply on $[0,t]$. Integrating, and using $\varphi \ge 0$ and $\varphi(\sin\theta_\star) \le \varphi(c^*) \le t_1$,

$$t_1 \ge \varphi(c(t)) - \varphi(c(0)) \ge \mathrm{act}(t) - \frac{1}{8}\big(t - \mathrm{act}(t)\big) \ge \mathrm{act}(t) - \frac{t}{8}.$$

(c) Let $t \in [t_1, T]$ and suppose $c(t) < c^*$. By *(a)*, $\tau^{\mathrm{act}} > t - t_1$ (otherwise $t \in [\tau^{\mathrm{act}} + t_1, T]$ and $c(t) \ge c^*$). Applying *(b)* at $t' := t - t_1 \le \min\{\tau^{\mathrm{act}}, T\}$, the time in $[0,t']$ during which no signal patch is activated is $t' - \mathrm{act}(t') \ge t' - t_1 - t'/8 = \frac{7}{8}t' - t_1$.

The statements for $j \in \mathcal{V}_-$ follow by the symmetry described at the end of Appendix D.2. □

E Proof of Theorem 1: Final Convergence

Since we have proved the first part of Theorem 1 in Appendix D, we use it as a fact and prove the second part. The small-output approximation of Appendix B is no longer available after time T; instead, the argument rests on exact identities for the gradient flow that hold for all $t \ge 0$ once a filter lies in a signal cone. Appendix E.1 collects the auxiliary lemmas: the exact dynamics of the signal coordinates and the fact that $\mathcal{S}_+, \mathcal{S}_-$ are trapping regions (Lemma 15), bounds on the filter norms up to time t_1 (Lemma 16), a comparison between the loss weights and the loss values (Lemma 17), and a lower bound on the gradient norm (Lemma 18). Appendix E.2 proves the second part of Theorem 1.

Standing conventions. Throughout this appendix the hypotheses of Theorem 1 hold and $\widetilde{\mathcal{V}_+}, \widetilde{\mathcal{V}_-}$ are as in Part 2.. For $i \in [n]$ let

$$\lambda_i(t) := -y_i \nabla_{\hat{y}} \ell\big(y_i, f(\boldsymbol{X}_i; W(t), \boldsymbol{v}(t))\big) = \begin{cases} e^{-y_i f(\boldsymbol{X}_i; W(t), \boldsymbol{v}(t))}, & \text{exponential loss,} \\ 2/\big(1 + e^{y_i f(\boldsymbol{X}_i; W(t), \boldsymbol{v}(t))}\big), & \text{logistic loss,} \end{cases} \tag{41}$$

which is positive in both cases; we call λ_i the *loss weight* of input i, and write $\ell_i(t) := \ell(y_i, f(\boldsymbol{X}_i; W(t), \boldsymbol{v}(t)))$ and $\Phi_{ij}(t) := \sum_{p=1}^{P} \sigma(\langle \boldsymbol{w}_j(t), \boldsymbol{x}_i^{(p)} \rangle)$ as in the proof of Lemma 3. Since $\nabla_{\hat{y}} \ell(y_i, \cdot) = -y_i \lambda_i$, the gradient flow (14)–(13) reads, for every $j \in [h]$ and all $t \ge 0$,

$$\dot{v}_j = \sum_{i=1}^{n} y_i \lambda_i \Phi_{ij}, \qquad \dot{\boldsymbol{w}}_j = \mathrm{sign}(v_j(0))\, \|\boldsymbol{w}_j\| \sum_{i=1}^{n} y_i \lambda_i \sum_{p:\, \langle \boldsymbol{w}_j, \boldsymbol{x}_i^{(p)} \rangle > 0} \boldsymbol{x}_i^{(p)}. \tag{42}$$

Recall d_i, μ_i, Γ from (10), and set $d_\Sigma := \sum_{i=1}^{n} d_i$, $d_{\max} := \max_i d_i$. Finally, let

$$\Lambda_\pm(t) := \sum_{i \in \mathcal{I}_\pm} \rho_i \mu_i \lambda_i(t), \qquad Q_+(t) := \sum_{j \in \widetilde{\mathcal{V}_+}} \langle \boldsymbol{u}_\star, \boldsymbol{w}_j(t) \rangle^2, \qquad Q_-(t) := \sum_{j \in \widetilde{\mathcal{V}_-}} \langle -\boldsymbol{u}_\star, \boldsymbol{w}_j(t) \rangle^2, \tag{43}$$

so that $\Lambda_\pm$ is the loss-weighted signal mass of each class projected on the axis, and $Q_\pm$ is the squared mass of each aligned group of neurons along the axis. Throughout, the case of $\widetilde{\mathcal{V}_-}$ follows from that of $\widetilde{\mathcal{V}_+}$ by the label flip described at the end of Appendix D.2, and we only write the arguments for $\widetilde{\mathcal{V}_+}$.

E.1 Auxiliary Lemmata

First, we show that $\mathcal{S}_+$ and $\mathcal{S}_-$ are trapping regions for the exact dynamics, and that inside them the mass along the axis can only grow.

Lemma 15 (Exact dynamics in the signal cones)**.** *Let $j \in \mathcal{V}_+$. For every $i \in [n]$ and all $t \ge 0$,*

$$\frac{d}{dt}\langle \boldsymbol{w}_j, \boldsymbol{g}_i \rangle = \|\boldsymbol{w}_j\| \sum_{k \in \mathcal{A}(\boldsymbol{w}_j)} \lambda_k \langle \boldsymbol{g}_k, \boldsymbol{g}_i \rangle, \qquad \frac{d}{dt}\langle \boldsymbol{u}_\star, \boldsymbol{w}_j \rangle = \|\boldsymbol{w}_j\| \sum_{k \in \mathcal{A}(\boldsymbol{w}_j)} \lambda_k \rho_k \mu_k. \tag{44}$$

Consequently, if $\boldsymbol{w}_j(t_a) \in \mathcal{S}_+$ for some $t_a \geq 0$, then $\boldsymbol{w}_j(t) \in \mathcal{S}_+$ for all $t \geq t_a$, and on $[t_a, \infty)$,

$$\frac{d}{dt}\langle \boldsymbol{u}_\star,\, \boldsymbol{w}_j\rangle = \|\boldsymbol{w}_j\|\, \Lambda_+ \;\geq\; 0. \tag{45}$$

The symmetric statements hold for $j \in \mathcal{V}_-$ with $\mathcal{S}_-$, $-\boldsymbol{u}_\star$ and Λ_-.

Proof. By (42) with $\operatorname{sign}(v_j(0)) = +1$, the signal patch of input k enters $\dot{\boldsymbol{w}}_j$ if and only if $k \in \mathcal{A}(\boldsymbol{w}_j)$ (see (23)), and it contributes $\|\boldsymbol{w}_j\|\, y_k \lambda_k \boldsymbol{s}_k = \|\boldsymbol{w}_j\|\, \lambda_k \boldsymbol{g}_k$; the noise patches contribute vectors in $V_{\mathrm{sig}}^{\perp}$ by Assumption 1. Taking the inner product with $\boldsymbol{g}_i \in V_{\mathrm{sig}}$ and with $\boldsymbol{u}_\star \in V_{\mathrm{sig}}$, and using $\langle \boldsymbol{u}_\star,\, \boldsymbol{g}_k\rangle = \rho_k \mu_k$, gives (44).

Suppose $\boldsymbol{w}_j(t_a) \in \mathcal{S}_+$, i.e. $\langle \boldsymbol{w}_j(t_a),\, \boldsymbol{g}_i\rangle > 0$ for all i, and, for contradiction, that $\boldsymbol{w}_j(t) \notin \mathcal{S}_+$ for some $t > t_a$. Let $\tau := \inf\{t \geq t_a : \min_i \langle \boldsymbol{w}_j(t),\, \boldsymbol{g}_i\rangle \leq 0\}$. By continuity, $\min_i \langle \boldsymbol{w}_j(\tau),\, \boldsymbol{g}_i\rangle = 0$, $\tau > t_a$, and $\boldsymbol{w}_j(t) \in \mathcal{S}_+$ for $t \in [t_a, \tau)$. On $[t_a, \tau)$ we have $\mathcal{A}(\boldsymbol{w}_j) = \mathcal{I}_+$, so by (44) and Lemma 6(i), for every i,

$$\frac{d}{dt}\langle \boldsymbol{w}_j,\, \boldsymbol{g}_i\rangle = \|\boldsymbol{w}_j\| \sum_{k \in \mathcal{I}_+} \lambda_k \langle \boldsymbol{g}_k,\, \boldsymbol{g}_i\rangle \geq \|\boldsymbol{w}_j\|\, \mu\rho_i \sum_{k \in \mathcal{I}_+} \lambda_k \rho_k \;>\; 0,$$

since $\|\boldsymbol{w}_j\| > 0$ (Lemma 3), $\lambda_k > 0$ and $\mathcal{I}_+ \neq \emptyset$. Hence every $\langle \boldsymbol{w}_j,\, \boldsymbol{g}_i\rangle$ is nondecreasing on $[t_a, \tau]$, and $\langle \boldsymbol{w}_j(\tau),\, \boldsymbol{g}_i\rangle \geq \langle \boldsymbol{w}_j(t_a),\, \boldsymbol{g}_i\rangle > 0$ for all i, contradicting $\min_i \langle \boldsymbol{w}_j(\tau),\, \boldsymbol{g}_i\rangle = 0$. Thus $\boldsymbol{w}_j(t) \in \mathcal{S}_+$ for all $t \geq t_a$, and (45) is the second identity in (44) with $\mathcal{A}(\boldsymbol{w}_j) = \mathcal{I}_+$. □

Next we bound the filter norms at time t_1 from below and above in terms of the initialization; the lower bound is what we need, and it does not follow from Lemma 4.

Lemma 16 (Filter norms up to time t_1)**.** *For every $j \in [h]$ and all $t \in [0, t_1]$,*

$$\epsilon^2 e^{-4d_\Sigma t}\, \|[W_0]_{:,j}\|^2 \;\leq\; \|\boldsymbol{w}_j(t)\|^2 \;\leq\; \epsilon^2 e^{4d_\Sigma t}\, \|[W_0]_{:,j}\|^2 .$$

Proof. Since $t_1 \leq T$ (Part 1.), Lemma 4 gives $\max_i |f(\boldsymbol{X}_i; W, \boldsymbol{v})| \leq 2\epsilon\sqrt{h} P X_{\max} W_{\max}^2 \leq \frac{1}{2}$ on $[0, t_1]$, so by Lemma 2, $\lambda_i = |\nabla_{\hat{y}} \ell(y_i, f(\boldsymbol{X}_i; W, \boldsymbol{v}))| \leq |y_i| + 2\,|f(\boldsymbol{X}_i; W, \boldsymbol{v})| \leq 2$. From (42) and Cauchy–Schwarz, for all $t \in [0, t_1]$,

$$\begin{aligned}\Big|\frac{d}{dt}\,\|\boldsymbol{w}_j\|^2\Big| &= 2\,|\langle \boldsymbol{w}_j,\, \dot{\boldsymbol{w}}_j\rangle| \leq 2\,\|\boldsymbol{w}_j\| \sum_{i=1}^{n} \lambda_i \sum_{p=1}^{P} \Big|\langle \boldsymbol{w}_j,\, \boldsymbol{x}_i^{(p)}\rangle\Big| \\ &\leq 4\,\|\boldsymbol{w}_j\|^2 \sum_{i=1}^{n}\sum_{p=1}^{P} \Big\|\boldsymbol{x}_i^{(p)}\Big\| = 4 d_\Sigma\, \|\boldsymbol{w}_j\|^2 .\end{aligned}$$

Grönwall's inequality in both directions, together with $\|\boldsymbol{w}_j(0)\|^2 = \epsilon^2\, \|[W_0]_{:,j}\|^2$, gives the claim. □

The last two lemmas concern the fitting phase. The first compares the loss weight of an input with its loss value; the second turns this comparison into a lower bound on the gradient norm.

Lemma 17 (Loss weights versus loss values)**.** *Let $\overline{L} := 2n \log(1 + e^{1/2})$ and let $\bar{\chi}$ be as in Theorem 1, i.e. $\bar{\chi} = 1$ for the exponential loss and $\bar{\chi} = 2(1 - e^{-\overline{L}/2})/\overline{L}$ for the logistic loss. Then for all $t \geq t_1$, $0 < L(W(t), \boldsymbol{v}(t)) \leq \overline{L}$ and*

$$\lambda_i(t) \;\geq\; \bar{\chi}\, \ell_i(t) \qquad \text{for all } i \in [n].$$

Proof. At time t_1 we have $|f(\boldsymbol{X}_i; W, \boldsymbol{v})| \leq \frac{1}{2}$ (Lemma 4, as $t_1 \leq T$), hence $\ell_i(t_1) \leq e^{1/2} \leq 2\log(1 + e^{1/2})$ for both losses and $L(W(t_1), \boldsymbol{v}(t_1)) \leq \overline{L}$. The loss is positive and nonincreasing along the gradient flow, which gives the first claim for all $t \geq t_1$. For the exponential loss, $\lambda_i = e^{-y_i f(\boldsymbol{X}_i; W, \boldsymbol{v})} = \ell_i$ and the second claim holds with $\bar{\chi} = 1$. For the logistic loss, solving $\ell_i = 2\log(1 + e^{-y_i f(\boldsymbol{X}_i; W, \boldsymbol{v})})$ for $e^{-y_i f(\boldsymbol{X}_i; W, \boldsymbol{v})}$ and substituting into (41) gives $\lambda_i = 2(1 - e^{-\ell_i/2})$. The function $x \mapsto 2(1 - e^{-x/2})$ is concave and vanishes at 0, so on $[0, \overline{L}]$ it lies above the chord through its endpoints, whose slope is $2(1 - e^{-\overline{L}/2})/\overline{L} = \bar{\chi}$; since $0 < \ell_i(t) \leq L(W(t), \boldsymbol{v}(t)) \leq \overline{L}$, this gives $\lambda_i \geq \bar{\chi}\ell_i$. □

Lemma 18 (Lower bound on the gradient norm)**.** *Let $t \geq t_1$ be such that $\boldsymbol{w}_j(t) \in \mathcal{S}_+$ for all $j \in \widetilde{\mathcal{V}_+}$ and $\boldsymbol{w}_j(t) \in \mathcal{S}_-$ for all $j \in \widetilde{\mathcal{V}_-}$. Then*

$$\|\nabla_W L(W(t), \boldsymbol{v}(t))\|_F^2 \;\geq\; Q_+(t)\Lambda_+(t)^2 + Q_-(t)\Lambda_-(t)^2.$$

If moreover $Q_+(t) \geq b_0$ and $Q_-(t) \geq b_0$ for some $b_0 > 0$, then

$$\|\nabla_W L(W(t), \boldsymbol{v}(t))\|_F^2 \;\geq\; \tfrac{1}{2}\, b_0\, (\mu_\star \rho_{\min} \bar{\chi})^2\, L(W(t), \boldsymbol{v}(t))^2.$$

Proof. By (2), $\nabla_{\boldsymbol{w}_j} L = -\dot{\boldsymbol{w}}_j$, so for $j \in \widetilde{\mathcal{V}_+}$, (45) gives $\langle \boldsymbol{u}_\star,\, \nabla_{\boldsymbol{w}_j} L\rangle = -\,\|\boldsymbol{w}_j\|\,\Lambda_+$, and symmetrically $\langle -\boldsymbol{u}_\star,\, \nabla_{\boldsymbol{w}_j} L\rangle = -\,\|\boldsymbol{w}_j\|\,\Lambda_-$ for $j \in \widetilde{\mathcal{V}_-}$. Keeping only these neurons and only the component of each gradient along $\pm\boldsymbol{u}_\star$,

$$\begin{aligned}\|\nabla_W L\|_F^2 &\geq \sum_{j\in\widetilde{\mathcal{V}_+}} \langle \boldsymbol{u}_\star,\, \nabla_{\boldsymbol{w}_j} L\rangle^2 + \sum_{j\in\widetilde{\mathcal{V}_-}} \langle -\boldsymbol{u}_\star,\, \nabla_{\boldsymbol{w}_j} L\rangle^2 \\ &= \Lambda_+^2 \sum_{j\in\widetilde{\mathcal{V}_+}} \|\boldsymbol{w}_j\|^2 + \Lambda_-^2 \sum_{j\in\widetilde{\mathcal{V}_-}} \|\boldsymbol{w}_j\|^2 \;\geq\; Q_+\Lambda_+^2 + Q_-\Lambda_-^2,\end{aligned}$$

where the last step uses $\|\boldsymbol{w}_j\|^2 \geq \langle \pm\boldsymbol{u}_\star,\, \boldsymbol{w}_j\rangle^2$. For the second claim, Lemma 17 and $\rho_i\mu_i \geq \rho_{\min}\mu_\star$ give $\Lambda_+ \geq \mu_\star\rho_{\min}\bar{\chi}\sum_{i\in\mathcal{I}_+}\ell_i =: \mu_\star\rho_{\min}\bar{\chi}L_+$ and likewise $\Lambda_- \geq \mu_\star\rho_{\min}\bar{\chi}L_-$ with $L_- := \sum_{i\in\mathcal{I}_-}\ell_i$; hence, using $Q_\pm \geq b_0$ and $L_+^2 + L_-^2 \geq \frac{1}{2}(L_+ + L_-)^2 = \frac{1}{2}L^2$,

$$\|\nabla_W L\|_F^2 \geq b_0(\mu_\star\rho_{\min}\bar{\chi})^2\big(L_+^2 + L_-^2\big) \geq \tfrac{1}{2}\, b_0\, (\mu_\star\rho_{\min}\bar{\chi})^2 L^2. \qquad \square$$

E.2 Proof of the Second Part of Theorem 1

Proof of Theorem 1, Part 2. We first show that the aligned neurons never leave the signal cones. By Part 1., every $j \in \widetilde{\mathcal{V}_+}$ satisfies $\cos(\boldsymbol{w}_j(t_1), \boldsymbol{u}_\star) \geq c^*$ and $\boldsymbol{w}_j(t_1) \in \mathcal{S}_+$. Lemma 15 with $t_a = t_1$ shows that $\boldsymbol{w}_j(t) \in \mathcal{S}_+$ for all $t \geq t_1$ and that $\langle \boldsymbol{u}_\star,\, \boldsymbol{w}_j(t)\rangle$ is nondecreasing on $[t_1, \infty)$; it is positive there because $\langle \boldsymbol{u}_\star,\, \boldsymbol{w}_j(t_1)\rangle \geq c^*\,\|\boldsymbol{w}_j(t_1)\| > 0$. The same holds for $\widetilde{\mathcal{V}_-}$ with $\mathcal{S}_-$ and $-\boldsymbol{u}_\star$. This proves *(a)*. In particular, Q_+ and Q_- are nondecreasing on $[t_1, \infty)$, and the hypothesis of Lemma 18 holds for every $t \geq t_1$.

Next we lower bound the growth of the aligned mass Q_+. For $t \geq t_1$, by (45) and $0 < \langle \boldsymbol{u}_\star,\, \boldsymbol{w}_j\rangle \leq \|\boldsymbol{w}_j\|$,

$$\dot{Q}_+ = 2\sum_{j\in\widetilde{\mathcal{V}_+}} \langle \boldsymbol{u}_\star,\, \boldsymbol{w}_j\rangle\, \|\boldsymbol{w}_j\|\, \Lambda_+ \;\geq\; 2\Lambda_+ Q_+,$$

hence, writing $I_+(t) := \int_{t_1}^t \Lambda_+(s)\, ds$ for the integrated loss-weighted signal mass,

$$Q_+(t) \;\geq\; Q_+(t_1)\, e^{2I_+(t)}, \qquad Q_+(t_1) \;\geq\; \underline{Q}_+ := c^{*2}\epsilon^2 e^{-4d_\Sigma t_1} \sum_{j\in\widetilde{\mathcal{V}_+}} \|[W_0]_{:,j}\|^2, \tag{46}$$

where the lower bound on $Q_+(t_1)$ uses $\langle \boldsymbol{u}_\star,\, \boldsymbol{w}_j(t_1)\rangle \geq c^*\,\|\boldsymbol{w}_j(t_1)\|$ and Lemma 16.

We also need an upper bound on the growth of the total mass of $\mathcal{V}_+$. Let $E_+(t) := \sum_{j\in\mathcal{V}_+} \|\boldsymbol{w}_j(t)\|^2$. For any $j \in \mathcal{V}_+$ and $t \geq 0$, by (42), $\frac{d}{dt}\|\boldsymbol{w}_j\|^2 = 2\,\|\boldsymbol{w}_j\|\sum_i y_i\lambda_i\Phi_{ij}$, and dropping the (nonpositive) terms with $i \in \mathcal{I}_-$ and using $0 \leq \Phi_{ij} \leq d_i\,\|\boldsymbol{w}_j\|$ and $d_i \leq \Gamma\rho_i\mu_i$,

$$\frac{d}{dt}\|\boldsymbol{w}_j\|^2 \leq 2\,\|\boldsymbol{w}_j\|^2 \sum_{i\in\mathcal{I}_+} \lambda_i d_i \leq 2\Gamma\Lambda_+\,\|\boldsymbol{w}_j\|^2.$$

Summing over $j \in \mathcal{V}_+$ and integrating from t_1, and bounding $E_+(t_1)$ by Lemma 16,

$$E_+(t) \;\leq\; \overline{E}_+\, e^{2\Gamma I_+(t)}, \qquad \overline{E}_+ := \epsilon^2 e^{4d_\Sigma t_1} \sum_{j\in\mathcal{V}_+} \|[W_0]_{:,j}\|^2. \tag{47}$$

The two growth bounds combine into a lower bound on the loss weights of the positive class. For $i \in \mathcal{I}_+$ and $t \geq t_1$, by (1) and $v_j = \operatorname{sign}(v_j(0))\,\|\boldsymbol{w}_j\|$, the neurons in $\mathcal{V}_-$ contribute nonpositively

to $f(\boldsymbol{X}_i; W, \boldsymbol{v})$, so $f(\boldsymbol{X}_i; W, \boldsymbol{v}) \le \sum_{j\in\mathcal{V}_+} \|\boldsymbol{w}_j\| \Phi_{ij} \le d_i E_+$. For the exponential loss this gives $\lambda_i = e^{-f(\boldsymbol{X}_i;W,\boldsymbol{v})} \ge e^{-d_i E_+}$; for the logistic loss, $f(\boldsymbol{X}_i; W, \boldsymbol{v}) \le M$ implies $\lambda_i = 2/(1 + e^{f(\boldsymbol{X}_i;W,\boldsymbol{v})}) \ge 2/(1+e^M) \ge e^{-M}$. In both cases, by (47) and $d_i \le d_{\max}$,

$$\lambda_i(t) \;\ge\; \exp\Big(- d_{\max}\overline{E}_+ \, e^{2\Gamma I_+(t)}\Big) \qquad \text{for all } i \in \mathcal{I}_+,\ t \ge t_1. \tag{48}$$

We can now bound the time at which the aligned mass reaches the threshold b_0. Fix $b_0 > 0$ and $\epsilon \le 1$, and let $t_2^+ := \inf\{t \ge t_1 : Q_+(t) \ge b_0\}$. Since Q_+ is nondecreasing, $Q_+(t) \ge b_0$ for all $t \ge t_2^+$. By (46), $Q_+(t) \ge b_0$ as soon as $I_+(t) \ge \bar{g}_+ := \frac{1}{2}\log_+(b_0/\underline{Q}_+)$, where $\log_+ x := \max\{0, \log x\}$. To bound the time needed for I_+ to reach $\bar{g}_+$, note that by (43), (48) and $\sum_{i\in\mathcal{I}_+}\rho_i\mu_i \ge \mu_\star R_+$,

$$\dot{I}_+ = \Lambda_+ \;\ge\; h(I_+), \qquad h(g) := \mu_\star R_+ \exp\Big(- d_{\max}\overline{E}_+ \, e^{2\Gamma g}\Big) > 0,$$

and h is decreasing. Let $H(g) := \int_0^g \frac{du}{h(u)}$, which is strictly increasing. For $t \ge t_1$, $\frac{d}{dt}H(I_+(t)) = \dot{I}_+/h(I_+) \ge 1$, so $H(I_+(t)) \ge t - t_1$. Hence if $t - t_1 > H(\bar{g}_+)$, then $H(I_+(t)) > H(\bar{g}_+)$, so $I_+(t) > \bar{g}_+$ and $Q_+(t) \ge b_0$. Therefore

$$t_2^+ \;\le\; t_1 + H(\bar{g}_+) = t_1 + \int_0^{\bar{g}_+} \frac{\exp\left(d_{\max}\overline{E}_+ e^{2\Gamma u}\right)}{\mu_\star R_+}\, du \;\le\; t_1 + \frac{\bar{g}_+}{\mu_\star R_+}\exp\Big(d_{\max}\overline{E}_+ e^{2\Gamma \bar{g}_+}\Big),$$

where we bounded the increasing integrand by its value at $\bar{g}_+$. It remains to make the two quantities involving $\bar{g}_+$ explicit. By (46) and $\epsilon \le 1$,

$$\bar{g}_+ = \tfrac{1}{2}\log_+\Big(\frac{b_0 e^{4d_\Sigma t_1}}{c^{*2}\sum_{j\in\widetilde{\mathcal{V}_+}}\|[W_0]_{:,j}\|^2}\cdot\frac{1}{\epsilon^2}\Big) \le \log\frac{1}{\epsilon} + C_0^+, \qquad C_0^+ := \tfrac{1}{2}\log_+\frac{b_0 e^{4d_\Sigma t_1}}{c^{*2}\sum_{j\in\widetilde{\mathcal{V}_+}}\|[W_0]_{:,j}\|^2},$$

and $e^{2\Gamma\bar{g}_+} = \max\{1, b_0/\underline{Q}_+\}^\Gamma \le (1 + b_0/\underline{Q}_+)^\Gamma$, so that, by (47), (46) and $\epsilon \le 1$,

$$d_{\max}\overline{E}_+ e^{2\Gamma\bar{g}_+} \le d_{\max}\,\epsilon^2 e^{4d_\Sigma t_1}\sum_{j\in\mathcal{V}_+}\|[W_0]_{:,j}\|^2\Big(1+\frac{b_0 e^{4d_\Sigma t_1}}{c^{*2}\sum_{j\in\widetilde{\mathcal{V}_+}}\|[W_0]_{:,j}\|^2}\Big)^\Gamma \epsilon^{-2\Gamma} = C_{\mathrm{tr}}^+\,\epsilon^{-2(\Gamma-1)},$$

with

$$C_{\mathrm{tr}}^\pm := d_{\max}e^{4d_\Sigma t_1}\sum_{j\in\mathcal{V}_\pm}\|[W_0]_{:,j}\|^2\Big(1+\frac{b_0 e^{4d_\Sigma t_1}}{c^{*2}\sum_{j\in\widetilde{\mathcal{V}}_\pm}\|[W_0]_{:,j}\|^2}\Big)^\Gamma, \qquad C_{\mathrm{tr}} := \max\{C_{\mathrm{tr}}^+, C_{\mathrm{tr}}^-\}, \qquad C_0 := \max\{C_0^+, C_0^-\}, \tag{49}$$

where C_0^- is defined as C_0^+ with $\widetilde{\mathcal{V}_-}$ in place of $\widetilde{\mathcal{V}_+}$. The same argument for $\widetilde{\mathcal{V}_-}$ gives the analogous bound on $t_2^- := \inf\{t \ge t_1 : Q_-(t) \ge b_0\}$, and with $t_2 := \max\{t_2^+, t_2^-\}$ and $R_\pm \ge R_{\min}$ we obtain

$$t_2 \le t_1 + \frac{\log(1/\epsilon) + C_0}{\mu_\star R_{\min}}\exp\Big(C_{\mathrm{tr}}\,\epsilon^{-2(\Gamma-1)}\Big), \qquad Q_+(t) \ge b_0 \;\text{ and }\; Q_-(t) \ge b_0 \quad \text{for all } t \ge t_2. \tag{50}$$

It remains to prove the $O(1/t)$ convergence after t_2. For $t \ge t_2$, Lemma 18 gives $\|\nabla_W L\|_F^2 \ge \alpha L^2$ with $\alpha = \frac{1}{2}b_0(\mu_\star\rho_{\min}\bar{\chi})^2$. Along the gradient flow (2), $\dot{L} = -\|\nabla_W L\|_F^2 - \|\nabla_{\boldsymbol{v}} L\|^2 \le -\alpha L^2$. Since L is positive and absolutely continuous, $\frac{d}{dt}\frac{1}{L} = -\frac{\dot{L}}{L^2} \ge \alpha$, and integrating from t_2 to t,

$$\frac{1}{L(W(t),\boldsymbol{v}(t))} \ge \frac{1}{L(W(t_2),\boldsymbol{v}(t_2))} + \alpha(t - t_2) = \frac{1 + \alpha L(W(t_2),\boldsymbol{v}(t_2))(t-t_2)}{L(W(t_2),\boldsymbol{v}(t_2))}.$$

Taking reciprocals proves *(b)*; the denominator grows linearly in t, so $L(W(t),\boldsymbol{v}(t)) = O(1/t)$ and $L(W(t),\boldsymbol{v}(t)) \to 0$. This finishes the proof. □