WEIGHT DECAY AND NEURON CONDENSATION: A THREE-STAGE ANALYSIS OF TWO-LAYER RELU NETWORKS

Anonymous authors
Paper under double-blind review

ABSTRACT

Weight decay is widely used as a regularization technique in neural network training, yet its role in neuron condensation (parameter direction alignment) remains unclear. Starting from a parameter initialization in the neural tangent kernel regime, we characterize training dynamics under weight decay through three stages: rapid fitting, amplitude compression, and neuron condensation. Using a two-layer ReLU network, we analyze a residual correlation field that governs both neuron amplitudes and directions. During rapid fitting, the residual approaches a quasi-static equilibrium maintained by weight decay while the tangent kernel remains nearly unchanged. In the early stage of amplitude compression, kernel decay amplifies the residual correlation field, whose isolated attracting extrema provide candidates for condensation directions. As neuron amplitudes stabilize, we bound the drift of attracting extrema and demonstrate contraction of neuron directions around them, leading to neuron condensation. This staged analysis provides a dynamical understanding of how weight decay promotes a condensed representation, beyond reducing parameter norms.

1 INTRODUCTION

An overparameterized network fits the same training data in many ways, and which solution gradient-based training finds depends on the parameterization, the initialization and the optimizer (Luo et al., 2021; Chizat et al., 2019; Chizat & Bach, 2020). Classical capacity control does not explain why such networks generalize (Zhang et al., 2017); one explanation is that training returns solutions simpler than the size of the network suggests (Xu et al., 2025b;a). For a neural network this simplicity has a concrete form: the learned function may need far fewer neurons than the network has.

Condensation is how this happens during training: the input weight vectors of many neurons, biases included, concentrate around a few directions (Luo et al., 2021; Zhou et al., 2022; Xu et al., 2025b). Neurons sharing a direction compute proportional features and act as one, so a wide network behaves like a much narrower one.

Small initialization induces condensation (Maennel et al., 2018; Zhou et al., 2022). The price is speed: the smaller the initialization, the longer training takes to leave the neighborhood of the origin before fitting begins (Zhang et al., 2021; 2022). Dropout (Srivastava et al., 2014) induces condensation as well even with large initialization (Zhang & Xu, 2024).

A large initialization places a sufficiently wide network in the lazy regime instead: training fits the data quickly, but the features move little relative to their initial scale, so no condensation occurs (Jacot et al., 2018; Chizat et al., 2019; Luo et al., 2021). We study whether weight decay induces condensation even with large initialization, in two-layer ReLU regression. Sparse minimizers are known to exist for related norm-regularized formulations (Savarese et al., 2019; Parhi & Nowak, 2021), but their existence says nothing about whether the gradient-based learning can lead the network to such solutions. In related homogeneous models, training with large initialization and weak weight decay can transition from kernel behavior to a parameter-space maximum-margin bias(Lyu et al., 2024). This transition provides one explanation of grokking, in which test performance improves long after the training loss has stopped decreasing (Power et al., 2022; Liu et al., 2023; Kumar et al., 2024; Lyu et al., 2024). Consistent with this separation of time scales, our experiments

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.

show rapid initial fitting governed by the neural tangent kernel, followed by slower representation dynamics under weight decay. During this later evolution, neuron amplitudes shrink and surviving directions concentrate into a small number of groups. We characterize this condensation mechanism through three stages: residual equilibration, amplification of the residual correlation field, and neuron condensation as amplitudes stabilize.

The key in this paper is not the proof of theorems but the connection between the weight decay and the weight directions, with reasonable assumptions motivated by the experiments and dynamical analysis. We complement the two-layer analysis with empirical observations in Transformers. Figure 1 shows condensation in the query and key projection weights of a Transformer model trained on an anchor function task (Zhang et al., 2024; 2026). Each input contains an ordered pair of anchor tokens whose composition determines the target. All examples generated from the anchor pairs $(3, 4)$ and $(4, 3)$ are excluded from training and used only for evaluation. These observations extend the empirical evidence beyond two-layer networks.

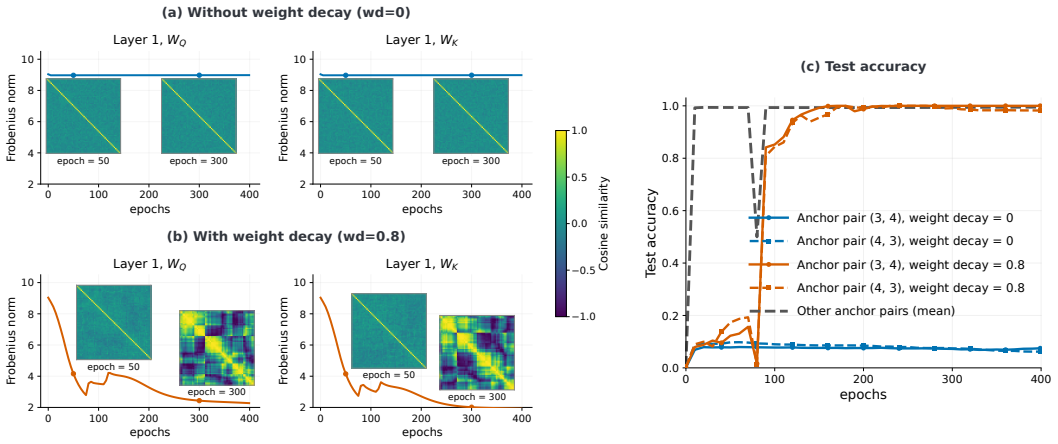

Figure 1: Parameter norm contraction, condensation and generalization in a Transformer. Panels (a) and (b) show the first layer W_Q and W_K weights for $\lambda = 0$ and $\lambda = 0.8$, respectively. Curves show their Frobenius norms, and insets show row-wise cosine similarities at epochs 50 and 300. Weight decay contracts the norms and induces row vectors alignment. Panel (c) shows that $\lambda = 0.8$ achieves substantially higher generalization accuracy on the held-out anchor pairs $(3, 4)$ and $(4, 3)$ than $\lambda = 0$. Experimental details are provided in Appendix G

Main contributions. We explain how weight decay induces neuron condensation in two-layer ReLU networks under large initialization by characterizing three stages of training: rapid fitting, amplitude compression, and neuron condensation. We show that the angular dynamics induced by the field admit only finitely many isolated attracting directions. During early amplitude compression, weight decay reduces the neural tangent kernel and amplifies this field. As amplitudes stabilize, we characterize the slow motion of attracting extrema and exponential angular contraction around them. Together, these results provide a dynamical explanation of how weight decay promotes neuron condensation under large initialization, distinct from previous analyses of condensation under small initialization and lazy training under large initialization.

2 Related Work

Condensation and early directional alignment. Small-initialization training can concentrate neurons along a limited set of directions, as studied through feature quantization (Maennel et al., 2018), initialization-dependent regimes (Luo et al., 2021), and initial condensation (Zhou et al., 2022; Chen et al., 2024). For shallow ReLU networks, Boursier et al. (2022) characterize alignment and subsequent training under orthogonal inputs while Boursier & Flammarion (2025) give a quantitative analysis of early alignment for more general data and investigate its sparsity bias phenomenon. Kumar & Haupt (2024) relate directional convergence near small initializations and certain saddles to stationary directions of a neural correlation function.

Weight decay and the transition beyond lazy training. The neural tangent kernel and lazy-training analyses describe regimes in which training is approximated by a linearized model (Jacot et al., 2018; Chizat et al., 2019). With L_2 regularization, Lewkowycz & Gur-Ari (2020) derive time-dependent kernel dynamics for homogeneous networks in an infinite-width regime and analyze the increase of the empirical loss. Lyu et al. (2024) prove a sharp transition from kernel to a parameter space maximum-margin bias for large initialization and vanishing weight decay; their general theory assumes smooth homogeneous models. Kumar et al. (2024) also connect delayed generalization to late feature learning, including examples without regularization. Focusing on convex neural network (Bach, 2017), Pilanci & Ergen (2020) establish convex reformulations of two-layer ReLU training with a block-sparsity penalty and L^2 regularization.

Other mechanisms leading to condensation. Dropout provides another route to neuron condensation (Zhang & Xu, 2024). Empirical work on Transformers also reports condensation under changes in initialization and weight decay (Zhang et al., 2024), while Chen & Luo (2025) analyze staged condensation and rank collapse from small initialization in a different architecture. Weight decay has additionally been linked to neural collapse in classification (Jacot et al., 2025) and low-rank structure at stationary points (Kuzborskij & Abbasi-Yadkori, 2026). These geometric properties differ from the within-group angular diameter studied here: concentration of sample representations or matrix rank alone does not characterize alignment of individual hidden-neuron directions.

3 PRELIMINARIES

3.1 BASIC NOTATION

We write $\|\cdot\|_2$ for the Euclidean norm and $\|\cdot\|_{\mathrm{op}}$ for its induced matrix norm. For $\boldsymbol{x}, \boldsymbol{y} \in \mathbb{R}^n$, the empirical inner product and norm are

$$\langle \boldsymbol{x}, \boldsymbol{y} \rangle_n = \frac{1}{n} \boldsymbol{x}^\top \boldsymbol{y}, \qquad \|\boldsymbol{x}\|_n = \sqrt{\langle \boldsymbol{x}, \boldsymbol{x} \rangle_n} = \frac{\|\boldsymbol{x}\|_2}{\sqrt{n}}.$$

The notation $O_p(c_m)$ and $o_p(c_m)$ means that division by $c_m > 0$ gives, respectively, a quantity bounded or converging to zero in probability as $m \to \infty$.

3.2 NETWORK AND TRAINING OBJECTIVE

Given samples $\{(x_i, y_i)\}_{i=1}^n \subset \mathbb{R} \times \mathbb{R}$, consider a two-layer ReLU network with m hidden neurons:

$$f_\Theta(x) = \sum_{k=1}^m a_k \,\mathrm{ReLU}(w_k x + b_k), \qquad \Theta = (a_k, w_k, b_k)_{k=1}^m. \tag{1}$$

For $\lambda > 0$, the training objective with weight decay is

$$\mathcal{L}(\Theta) = \frac{1}{n} \sum_{i=1}^n (f_\Theta(x_i) - y_i)^2, \qquad \Phi(\Theta) = \mathcal{L}(\Theta) + \frac{\lambda}{2} \sum_{k=1}^m (a_k^2 + w_k^2 + b_k^2). \tag{2}$$

Write $\boldsymbol{X} = (x_i)_{i=1}^n$, $\boldsymbol{y} = (y_i)_{i=1}^n$, and $\boldsymbol{e}(t) = f_{\Theta(t)}(\boldsymbol{X}) - \boldsymbol{y} \in \mathbb{R}^n$ for the training residual. The parameters $a_k(0), w_k(0), b_k(0)$ are initialized independently from $N(0, m^{-2\gamma})$, with large initialization $\gamma < 1/2$. We study gradient flow $\frac{\mathrm{d}\Theta(t)}{\mathrm{d}t} = -\nabla_\Theta \Phi(\Theta(t))$, with weight decay applied to all parameters. Our main experiment uses the target function $f_\star(x) = 0.2\,\mathrm{ReLU}(x - 1/3) + 0.2\,\mathrm{ReLU}(-x - 1/3)$ as shown in Figure 2(b). Detailed experimental settings are provided in Appendix E.

3.3 AMPLITUDE AND ANGULAR DYNAMICS

Write $\boldsymbol{v}_k = (w_k, b_k)^\top = r_k \boldsymbol{u}(\theta_k)$, where $r_k = \|\boldsymbol{v}_k\|_2$ and $\boldsymbol{u}(\theta) = (\cos\theta, \sin\theta)^\top$. For $r_k > 0$, define the direction $\theta_k \in [0, 2\pi)$, the amplitude $A_k = |a_k| r_k$, and $s_k = \mathrm{sgn}(a_k)$. Positive and negative neurons have $a_k > 0$ and $a_k < 0$ respectively. With the directional feature

$$\phi_\theta(x) = \mathrm{ReLU}(x\cos\theta + \sin\theta), \qquad \psi_\theta(x) = \mathbf{1}_{\{x\cos\theta + \sin\theta > 0\}}(-x\sin\theta + \cos\theta), \tag{3}$$

the network is $f_\Theta(x) = \sum_k s_k A_k \phi_{\theta_k}(x)$ by positive homogeneity of ReLU activation. Here ψ_θ is the angular derivative of ϕ_θ away from activation boundaries, where $x_i \cos\theta + \sin\theta = 0$ for some training input.

For a residual vector $\boldsymbol{e} \in \mathbb{R}^n$, define the *residual correlation field* as the scalar function q_e of the angle θ, with angular derivative p_e:

$$q_e(\theta) = \langle \boldsymbol{e}, \boldsymbol{\phi}_\theta(\boldsymbol{X}) \rangle_n , \qquad p_e(\theta) = \partial_\theta q_e(\theta) = \langle \boldsymbol{e}, \boldsymbol{\psi}_\theta(\boldsymbol{X}) \rangle_n . \tag{4}$$

During training, write $q_t = q_{e(t)}$, $p_t = p_{e(t)}$, $q_k = q_t(\theta_k)$, and $p_k = p_t(\theta_k)$. Let $J_t = \partial_\Theta f_{\Theta(t)}(\boldsymbol{X})$ and $K_t = J_t J_t^\top / n$ be the Jacobian and empirical neural tangent kernel (NTK). We have the following gradient flow.

Proposition 1 (Gradient flow)**.** *For $r_k > 0$, gradient flow for equation (2) satisfies*

$$\frac{\mathrm{d}a_k}{\mathrm{d}t} = -2 r_k q_k - \lambda a_k, \qquad \frac{\mathrm{d}r_k}{\mathrm{d}t} = -2 a_k q_k - \lambda r_k, \qquad \frac{\mathrm{d}\theta_k}{\mathrm{d}t} = -2 \frac{a_k}{r_k} p_k. \tag{5}$$

The training residual satisfies

$$\frac{\mathrm{d}\boldsymbol{e}}{\mathrm{d}t} = -2(K_t + \lambda I_n)\boldsymbol{e} - 2\lambda \boldsymbol{y}. \tag{6}$$

Moreover, the scale imbalance obeys

$$B_k(t) := a_k(t)^2 - r_k(t)^2 = \mathrm{e}^{-2\lambda t} B_k(0). \tag{7}$$

Weight decay acts directly on parameter scales, while angular motion is governed by p_k. Equation (7) shows that the imbalance $a_k^2 - r_k^2$ decays exponentially, motivating a balanced description of late training. For an initially strictly inactive neuron, $w_k x_i + b_k < 0$ for every training input, so $q_k = p_k = 0$: its amplitude decays as $A_k(t) = \mathrm{e}^{-2\lambda t} A_k(0)$ while its direction remains fixed. Proposition 1 is derived in Appendix A.

Holding K_t fixed in equation (6) gives the *quasi-static residual*

$$\boldsymbol{e}_{\mathrm{qs}}(t) := -\lambda (K_t + \lambda I_n)^{-1} \boldsymbol{y}. \tag{8}$$

It is an instantaneous reference for the actual residual $\boldsymbol{e}(t)$.

3.4 Neuron condensation

For a fixed group G of at least two neurons with $A_k(t) > 0$, define its angular diameter by

$$\mathrm{diam}_\theta\, G(t) := \max_{i,j \in G} |\theta_i(t) - \theta_j(t)|. \tag{9}$$

We say that G undergoes *neuron condensation* if $\mathrm{diam}_\theta\, G(t) \to 0$ as $t \to \infty$. Thus condensation means that distinct neurons approach a common direction. Different groups may condense around different directions.

3.5 Three-stage overview

Figure 2 illustrates the evolution from rapid data fitting to neuron condensation. The delayed decrease in test loss in panel (a) reveals a grokking phenomenon. Panel (b) shows the learned function gradually approaching the target as the network evolves toward a condensed solution that closely approximates the target function. The neuron feature snapshots in panel (c) show a complex representation with widely scattered features (blue dots) after rapid fitting. As amplitudes stabilize, the features develop a clearer directional structure (orange dots), and eventually form tight angular groups (green dots) corresponding to the condensed solution.

Figure 3 resolves this evolution into three stages:

1. **Stage I: Fitting.** The data loss decreases rapidly (a), and the residual approaches its quasi-static reference (b), while neuron directions remain broadly distributed (f).

2. **Stage II: Compression.** Neuron amplitudes continue to decrease until they stabilize (c), while the attracting minima of the residual correlation field deepen substantially (d), strengthening directional attraction. Neuron directions begin to narrow but remain dispersed at the end of this stage (f).

3. **Stage III: Condensation.** After amplitudes have nearly stabilized (c), the angular diameters of the two neuron groups continue to decrease approximately exponentially (e), with neurons condensing to specific directions (f).

Section 4 characterizes the residual left after fitting, the attracting directions and amplification of the residual correlation field, and how amplitude stabilization controls extrema drift and enables angular contraction.

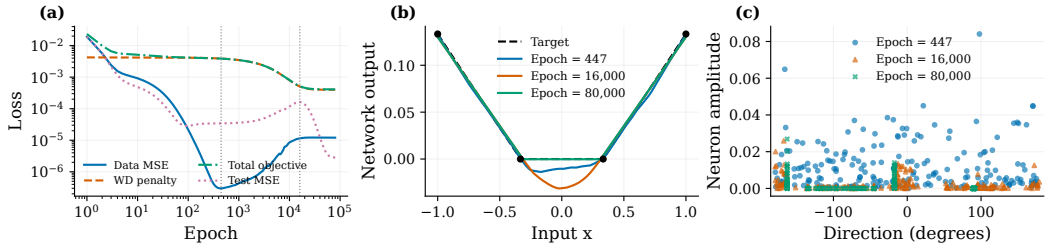

Figure 2: Training overview. (a) Data MSE, weight decay penalty, total objective, and test MSE; vertical markers indicate the loss minimum (epoch 447) and an epoch near the amplitude plateau (epoch 16,000). (b) Learned functions at epochs 447, 16,000, and 80,000. The four black dots are training samples. (c) Directions and amplitudes of all 200 neurons at different epochs.

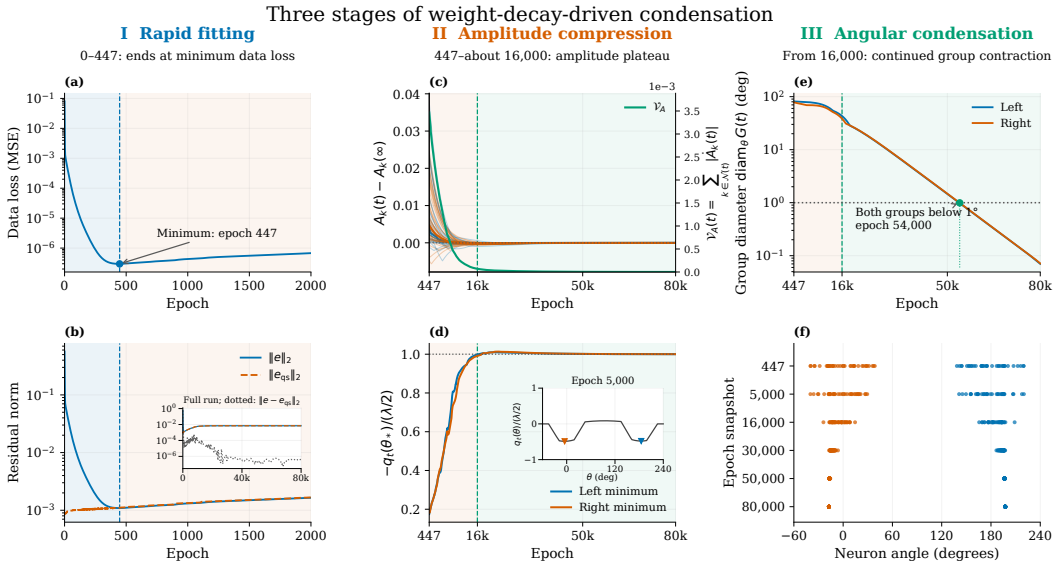

Figure 3: Diagnostics of the three stages. (a,b) Data MSE and residual norms $\|\boldsymbol{e}\|_2$, $\|\boldsymbol{e}_{\mathrm{qs}}\|_2$; the inset in (b) shows the full run, including $\|\boldsymbol{e} - \boldsymbol{e}_{\mathrm{qs}}\|_2$ (dotted). (c) Individual amplitude differences $A_k(t) - A_k(\infty)$ (left axis), with $A_k(\infty)$ approximated by the value at epoch 80,000, and total amplitude variation $\mathcal{V}_A(t) = \sum_{k\in\mathcal{N}(t)} |\dot{A}_k(t)|$ over all non-dead neurons (green curve, right axis). (d) Normalized depths $-q_t(\theta_*(t))/(\lambda/2)$ of the smooth negative minima associated with the two neuron groups. The inset shows $q_t(\theta)/(\lambda/2)$ at epoch 5,000. The normalized depth plateau in Stage III can be explained by the balanced amplitude dynamics in Appendix A.2. (e,f) Angular diameters and angle snapshots of two neuron groups.

4 MAIN RESULTS

We analyze how the residual correlation field governs neuron dynamics across the three training stages. We first identify the residual left after fast fitting, then establish the discreteness of the field's attracting extrema and characterize how amplitude compression amplifies the field. The main condensation theorem connects small amplitude variation to exponential contraction of a group's angular diameter.

4.1 STAGE I: FAST RESIDUAL EQUILIBRATION

Under the initialization in Section 3, let $\rho_m = m^{1-2\gamma}$ and $\varepsilon_m = \lambda/\rho_m$. The initial kernel satisfies $K_0/\rho_m \to \bar{K} \succ 0$ in probability as $m \to \infty$; see Appendix B.1 for the proof. The initial kernel has eigenvalues of order $\rho_m = m^{1-2\gamma}$, which diverges to infinity since $\gamma < 1/2$. The following theorem shows that the residual rapidly approaches its quasi-static reference while the kernel remains close to its initial value.

Theorem 2 (Residual equilibration)**.** *Assume $x_i \neq x_j$ for all $i \neq j$. Fix $\lambda > 0$ and $1/4 \leq \gamma < 1/2$. Let $\kappa_m = \lambda_{\min}(K_0)$ and, on the event $\kappa_m > 0$, define*

$$T_{\mathrm{I}} = \frac{\log(\log m/\varepsilon_m)}{2(\kappa_m + \lambda)}.$$

As $m \to \infty$, the nonlinear gradient flow satisfies

$$T_{\mathrm{I}} = O_p(\rho_m^{-1} \log(\rho_m/\lambda)) = o_p(1), \tag{10}$$

$$\sup_{t \leq T_{\mathrm{I}}} \frac{\|K_t - K_0\|_{\mathrm{op}}}{\rho_m} = O_p(\rho_m^{-1}(\log m)^{3/2}) = o_p(1), \tag{11}$$

$$\boldsymbol{e}(T_{\mathrm{I}}) = -\varepsilon_m \bar{K}^{-1}\boldsymbol{y} + o_p(\varepsilon_m), \qquad \|\boldsymbol{e}(T_{\mathrm{I}}) - \boldsymbol{e}_{\mathrm{qs}}(T_{\mathrm{I}})\|_2 = o_p(\varepsilon_m). \tag{12}$$

The proof is given in Appendix B. Figure 5 shows the evolution of the NTK eigenvalues at finite width. The bound $\gamma \geq 1/4$ is imposed to ensure that the initial residual satisfies $\|\boldsymbol{e}(0)\|_2 = O_p(1)$ in the infinite width limit. For any finite width network, the initial residual is finite.

Equation (11) shows that the kernel changes by $o_p(\rho_m)$ during initial fitting, a vanishing fraction of its initial scale. Throughout this stage, its eigenvalues therefore remain much larger than the fixed weight decay coefficient λ. Thus, the residual relaxes on a time scale $T_{\mathrm{I}} = o_p(1)$ is much faster than the effect of weight decay. For $\lambda > 0$ and $\boldsymbol{y} \neq \boldsymbol{0}$, the forcing term $-2\lambda\boldsymbol{y}$ sustains the quasi-static residual $\boldsymbol{e}_{\mathrm{qs}}(t) = -\lambda(K_t + \lambda I_n)^{-1}\boldsymbol{y}$. At the end of initial fitting, equation (12) shows that the tracking error $o_p(\varepsilon_m)$ can be neglected. Thus the actual residual retains this nonzero leading term, which generates the correlation field governing subsequent neuron motion.

4.2 STAGE II: DISCRETE EXTREMA AND AMPLIFICATION OF q_e

Let $\tilde{\boldsymbol{x}}_i = (x_i, 1)^\top$ and define the angular boundary set

$$\mathcal{B} = \{\theta \in [0, 2\pi) : \tilde{\boldsymbol{x}}_i^\top \boldsymbol{u}(\theta) = 0 \text{ for some } i\}.$$

An *activation cell* I is a connected component of $\mathbb{S}^1 \setminus \mathcal{B}$. Its active sample set $\mathcal{S}_I = \{i : \tilde{\boldsymbol{x}}_i^\top \boldsymbol{u}(\theta) > 0,\ \forall \theta \in I\}$ is the same for every $\theta \in I$; a cell is dead if this set is empty. Define the residual moment on each cell by

$$\boldsymbol{M}_I(\boldsymbol{e}) = \frac{1}{n} \sum_{i \in \mathcal{S}_I} e_i \tilde{\boldsymbol{x}}_i.$$

Within I, the residual correlation field has the sinusoidal form

$$q_e(\theta) = M_{I,1}(\boldsymbol{e}) \cos\theta + M_{I,2}(\boldsymbol{e}) \sin\theta, \tag{13}$$

where $M_{I,i}(\boldsymbol{e})$ denotes the ith component of $\boldsymbol{M}_I(\boldsymbol{e})$.

Recall that $\dot{\theta}_k = -2\beta_k p_e(\theta_k)$, where $\beta_k = a_k/r_k$ and $p_e = \partial_\theta q_e$. From equation (13) we have $q_e''(\theta) = -q_e(\theta)$. For fixed residual $\boldsymbol{e}$ and ratio β_k, the angular dynamics have equilibrium points

where $p_e(\theta_*) = 0$. Within a cell with $\boldsymbol{M}_I(\boldsymbol{e}) \neq \mathbf{0}$, these points are the smooth extrema of q_e. An equilibrium is locally attracting if and only if $a_k q_e(\theta_*) < 0$. Thus positive neurons are attracted to negative minima, and negative neurons to positive maxima. The following proposition characterizes the discrete set of extrema that can serve as condensation directions.

Proposition 3 (Discrete extrema)**.** *Fix a residual $\boldsymbol{e}$ and suppose $\boldsymbol{M}_I(\boldsymbol{e}) \neq \mathbf{0}$ on every non-dead activation cell. Let $\mathcal{E}(\boldsymbol{e})$ denote the set of strict local extrema of q_e, and define*

$$\mathcal{E}_{\mathrm{int}}(\boldsymbol{e}) := \mathcal{E}(\boldsymbol{e}) \setminus \mathcal{B}, \qquad \mathcal{E}_{\mathrm{bd}}(\boldsymbol{e}) := \mathcal{E}(\boldsymbol{e}) \cap \mathcal{B}.$$

These are, respectively, the smooth extrema inside activation cells and the extrema at activation boundaries, and satisfy

$$1 \leq |\mathcal{E}(\boldsymbol{e})| \leq 4n, \qquad |\mathcal{E}_{\mathrm{int}}(\boldsymbol{e})|,\ |\mathcal{E}_{\mathrm{bd}}(\boldsymbol{e})| \leq 2n.$$

The proof and a discussion of boundary extrema are given in Appendix C. We now focus on smooth extrema and their attraction strength. Consider a smooth negative minimum $\theta_*(t)$ and a positive neuron in the same activation cell. Writing $\delta_k(t) = \theta_k(t) - \theta_*(t)$, the sinusoidal form of the field equation (13) gives

$$\frac{\mathrm{d}\delta_k}{\mathrm{d}t} = 2\beta_k(t) q_t(\theta_*(t)) \sin \delta_k - \frac{\mathrm{d}\theta_*}{\mathrm{d}t}.$$

Since $\beta_k(t) q_t(\theta_*(t)) < 0$, the first term drives the neuron toward the minimum. For a fixed scale ratio, a more negative value of $q_t(\theta_*(t))$ produces stronger attraction.

We next examine how weight decay changes the field's strength during amplitude compression. Theorem 2 provides initial tracking of $\boldsymbol{e}_{\mathrm{qs}}$. The following result shows that slow kernel decay maintains this tracking and amplifies the residual correlation field during early amplitude compression. Set $C_\phi := \sup_{\theta \notin \mathcal{B}} \|\boldsymbol{\phi}_\theta(\boldsymbol{X})\|_n$.

Theorem 4 (Kernel decay and field amplification)**.** *Let $\boldsymbol{y} \neq \mathbf{0}$, $t_0 = T_{\mathrm{I}}$ and consider $[t_0, t_1]$ with $\lambda(t_1 - t_0) \leq L$, where L is fixed. Suppose the activation masks remain fixed and every non-dead neuron satisfies*

$$B^{-1} \leq \left| \frac{a_k(t)}{r_k(t)} \right| \leq B, \qquad |q_t(\theta_k(t))| + |p_t(\theta_k(t))| \leq \varepsilon\lambda.$$

Assume $K_{t_0} \succ 0$ and

$$\frac{\|K_{t_0}\|_{\mathrm{op}}}{\lambda_{\min}(K_{t_0})} \leq \chi, \qquad \frac{\lambda}{\lambda_{\min}(K_{t_0})} \leq \varepsilon,$$

where B, χ are fixed and ε is sufficiently small. Set $\delta_0 = \|\boldsymbol{e}(t_0) - \boldsymbol{e}_{\mathrm{qs}}(t_0)\|_n \,/\, \|\boldsymbol{e}_{\mathrm{qs}}(t_0)\|_n$. Then, writing $\tau = t - t_0$, there exists $C = C(B, L, \chi)$ such that

$$\left\| K_t - \mathrm{e}^{-2\lambda\tau} K_{t_0} \right\|_{\mathrm{op}} \leq C\varepsilon \mathrm{e}^{-2\lambda\tau} \|K_{t_0}\|_{\mathrm{op}}, \tag{14}$$

$$\sup_{s \in [t_0, t_1]} \frac{\|\boldsymbol{e}(s) - \boldsymbol{e}_{\mathrm{qs}}(s)\|_n}{\|\boldsymbol{e}_{\mathrm{qs}}(s)\|_n} \leq C(\delta_0 + \varepsilon), \tag{15}$$

$$\sup_{\theta} \left| q_t(\theta) - \mathrm{e}^{2\lambda\tau} q_{t_0}(\theta) \right| \leq C(\varepsilon + \delta_0) \mathrm{e}^{2\lambda\tau} C_\phi \|\boldsymbol{e}_{\mathrm{qs}}(t_0)\|_n. \tag{16}$$

The proof is given in Appendix C.2. Stage I gives $\delta_0 = o_p(1)$ at T_{I}, and equation (15) controls residual tracking throughout the stated compression interval. Consistent with this result, Figure 3(b) shows that $\boldsymbol{e}(t)$ remains close to $\boldsymbol{e}_{\mathrm{qs}}(t)$ after rapid fitting. When the correlations are small relative to λ, weight decay dominates parameter evolution and approximately rescales the kernel by $\mathrm{e}^{-2\lambda\tau}$. As weight decay reduces the kernel, $\boldsymbol{e}_{\mathrm{qs}}(t) \approx -\lambda K_t^{-1} \boldsymbol{y}$ grows, amplifying the correlation field.

The leading order scaling of weight decay changes the depths of the extrema but leaves their locations unchanged. However, in the actual dynamics, Figure 4 shows that the smooth extrema also move during Stage II. The following section controls this motion, which results in neuron condensation.

4.3 Stage III: Slow drift of extrema and neuron condensation

Let $\mathcal{N}(t) = \{k : \boldsymbol{\phi}_{\theta_k(t)}(\boldsymbol{X}) \neq \mathbf{0}\}$ be the set of non-dead neurons. Define their total amplitude, total amplitude variation rate, and angular energy by

$$\mathcal{A}(t) = \sum_{k \in \mathcal{N}(t)} A_k(t), \qquad \mathcal{V}_A(t) = \sum_{k \in \mathcal{N}(t)} \left| \frac{\mathrm{d}A_k(t)}{\mathrm{d}t} \right|, \qquad P(t) = \sum_{k \in \mathcal{N}(t)} A_k(t) p_k(t)^2. \tag{17}$$

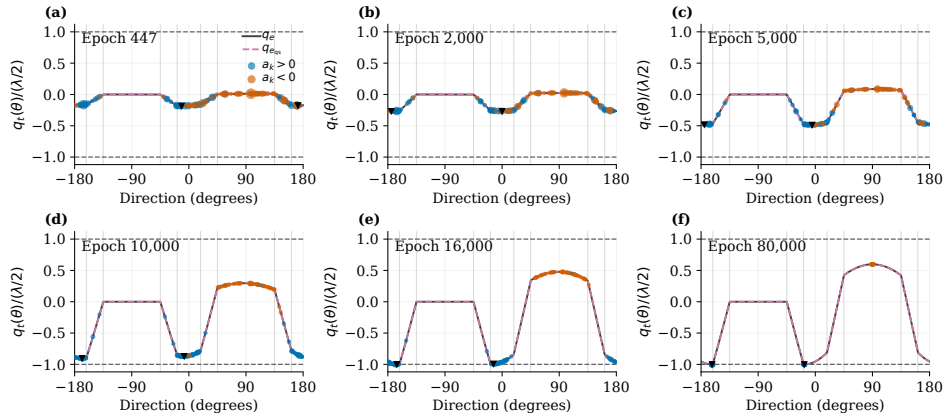

Figure 4: Residual correlation field evolution and surviving neurons. Curves show normalized $q_t(\theta)/(\lambda/2)$, computed from the actual residual $\boldsymbol{e}$ (black solid) and its quasi-stationary approximation $\boldsymbol{e}_{\text{qs}}$ (pink dashed), which nearly overlap. Blue and orange dots represent positive and negative neurons respectively. Black triangles mark smooth local minima. During Stage II (up to epoch 16,000), the field deepens substantially, while neuron directions remain dispersed around its extrema. During Stage III, neurons condense at these extrema.

We also use $C_\psi := \sup_{\theta \notin \mathcal{B}} \|\boldsymbol{\psi}_\theta(\boldsymbol{X})\|_n$ and $C_\phi = \sup_{\theta \notin \mathcal{B}} \|\boldsymbol{\phi}_\theta(\boldsymbol{X})\|_n$.

Attracting extrema of the residual correlation field evolve during training. To understand whether neurons can track these extrema and condense around them, we first bound their drift. The following theorem shows how amplitude variation $\mathcal{V}_A$ controls the angular energy P, and how these quantities jointly bound the velocity of the residual correlation field extrema.

Theorem 5 (Field drift)**.** *Suppose that, for all $t \geq T$, the gradient flow remains in fixed open activation cells and every non-dead neuron satisfies $a_k^2 = r_k^2 = A_k > 0$. Set $\overline{\mathcal{A}}_T := \Phi(\Theta(T))/\lambda$. Then, for all $t \geq T$,*

$$P(t) \leq \mathrm{e}^{-2\lambda(t-T)} P(T) + \frac{C_\phi^2}{4} \int_T^t \mathrm{e}^{-2\lambda(t-s)} \mathcal{V}_A(s)^2 \, ds, \tag{18}$$

$$\left\| \frac{\mathrm{d}\boldsymbol{e}(t)}{\mathrm{d}t} \right\|_n \leq C_\phi \mathcal{V}_A(t) + 2C_\psi \sqrt{\overline{\mathcal{A}}_T P(t)}. \tag{19}$$

If $\theta_*(t)$ is a differentiable branch of smooth negative minima in a fixed cell with depth $R(t) = -q_t(\theta_*(t)) \geq R_0 > 0$ for all $t \geq T$, then

$$\left| \frac{\mathrm{d}\theta_*(t)}{\mathrm{d}t} \right| \leq \frac{C_\psi}{R_0} \left[C_\phi \mathcal{V}_A(t) + 2C_\psi \sqrt{\overline{\mathcal{A}}_T P(t)} \right]. \tag{20}$$

The proof is given in Appendix D.2. Theorem 5 bounds the drift of attracting minima by the amplitude variation $\mathcal{V}_A(t)$ and the angular energy $P(t)$. The latter is itself controlled by amplitude variation: if $\mathcal{V}_A(t) \leq \nu_A$ for all $t \geq T$, then equation (18) gives

$$P(t) \leq \mathrm{e}^{-2\lambda(t-T)} P(T) + \frac{C_\phi^2 \nu_A^2}{8\lambda}. \tag{21}$$

The first term in equation (21) captures the influence of the initial angular energy and decays exponentially. Once it falls below fixed ε^2, substituting the resulting bound on $P(t)$ and $\mathcal{V}_A(t) \leq \nu_A$ into equation (20) yields

$$\begin{aligned} \left| \frac{\mathrm{d}\theta_*(t)}{\mathrm{d}t} \right| &\leq \frac{C_\psi}{R_0} \left[C_\phi \nu_A + 2C_\psi \sqrt{\overline{\mathcal{A}}_T \left(\varepsilon^2 + \frac{C_\phi^2 \nu_A^2}{8\lambda} \right)} \right] \\ &\leq \frac{C_\psi C_\phi}{R_0} \left(1 + C_\psi \sqrt{\frac{\overline{\mathcal{A}}_T}{2\lambda}} \right) \nu_A + \frac{2C_\psi^2 \sqrt{\overline{\mathcal{A}}_T}}{R_0} \varepsilon. \end{aligned} \tag{22}$$

Thus, the drift speed is bounded by a term proportional to ν_A. Figure 3(c) shows that $\mathcal{V}_A(t)$ remains small during Stage III, supporting a small upper bound ν_A over this period and motivating the slow extrema drift.

Weight decay also reduces the absolute imbalance $a_k^2 - r_k^2$ through equation (7), motivating the balanced model for surviving neurons. The next theorem shows that sufficiently slow drift keeps nearby neurons in an attracting band and yields angular contraction.

Theorem 6 (Angular contraction)**.** *Let G be a fixed group of at least two positive neurons and $\theta_*(t)$ a differentiable branch of smooth negative minima. For all $t \geq T$, assume that $a_k(t) = r_k(t) > 0$ for every $k \in G$, the minimum depth satisfies $R(t) \geq R_0 > 0$, and the band $[\theta_*(t) - d, \theta_*(t) + d]$ lies inside one fixed open activation cell, where $0 < d < \pi/2$. If*

$$|\theta_k(T) - \theta_*(T)| \leq d \quad (k \in G), \qquad \sup_{t \geq T} \left| \frac{\mathrm{d}\theta_*(t)}{\mathrm{d}t} \right| < 2R_0 \sin d,$$

then every neuron in G remains in the band and, with $c_d = 2R_0 \cos d > 0$,

$$\operatorname{diam}_\theta G(t) \leq \mathrm{e}^{-c_d(t-T)} \operatorname{diam}_\theta G(T), \qquad t \geq T. \tag{23}$$

Consequently, G undergoes neuron condensation. If additionally $\frac{\mathrm{d}\theta_*(t)}{\mathrm{d}t} \to 0$, then $|\theta_k(t) - \theta_*(t)| \to 0$ for every $k \in G$.

The proof is given in Appendix D.6. The theorem shows that a group can condense around a moving minimum: sufficiently slow drift retains the neurons in the attracting band, where their angular diameter decreases exponentially.

Together with Theorem 5, this gives the connection between amplitude stabilization and condensation. Small amplitude variation and a decayed initial angular-energy contribution yield a small drift bound. Once this bound is below $2R_0 \sin d$, the contraction theorem applies to neurons then inside the band, as long as the stated conditions persist. This describes how angular concentration can continue during Stage III after amplitudes vary little. The analogous result holds for exactly balanced negative neurons near a smooth positive maximum.

5 Discussion

Experiments with additional target functions in Appendix F.2 show the same progression from rapid fitting to late condensation. Our analysis characterizes neuron condensation around attracting extrema of the residual correlation field, but does not establish that these directions coincide with those in the target representation. The additional experiments also exhibit more condensation directions than those in the target representation. Appendix F.1 compares large initialization with and without weight decay against small initialization without weight decay. Both large initialization with weight decay and small initialization produce condensed solutions, but the former exhibits a distinct rapid fitting phase before condensation. Our dynamical analysis focuses on two-layer ReLU networks, while a corresponding dynamical theory for Transformers remains open.

6 Conclusion

Our work explains a mechanism for neuron condensation in two-layer ReLU networks trained with weight decay under large initialization. This mechanism differs from persistent lazy training at large initialization and condensation at small initialization. Our experiments show that rapid fitting precedes substantial directional alignment: the network first fits the data, then develops a condensed representation during continued training with weight decay.

By analyzing amplitude and angular dynamics, we characterize three stages: rapid fitting, amplitude compression, and neuron condensation. During rapid fitting, the residual approaches a quasi-static equilibrium maintained by weight decay. In the early stage of amplitude compression, we show that the residual correlation field admits only finitely many isolated attracting extrema and that early amplitude compression amplifies the field. As neuron amplitudes stabilize, we prove that neuron directions condense around slowly moving attracting directions. This analysis links weight decay to the formation of a condensed representation, clarifying its role beyond parameter norm reduction.

AI USE STATEMENT

Large Language Models (LLMs) were used as supporting tools in the writing and polishing of the manuscript. Specifically, LLMs were used to improve the readability of the manuscript, and to assist with the retrieval and discovery of potentially relevant literature and related work. Also, LLMs were used to help identify possible omissions in mathematical proofs. All literature identified with the assistance of LLMs was independently reviewed and verified by the authors before being cited.

LLMs were not used to formulate the research questions, design and conduct the experiments. All experimental procedures, results, mathematical arguments, and scientific conclusions were developed and validated by the authors. The authors take full responsibility for the accuracy of the manuscript.

ETHICS STATEMENT

This work studies how weight decay affects neural network training dynamics through mathematical analysis and experiments on synthetic data. It involves no human participants or personal data, and we identify no specific ethical concerns arising from the study.

REPRODUCIBILITY STATEMENT

The experiments can be reproduced using the synthetic data construction, model specifications, and training settings provided in Appendices E–G. They use standard neural network architectures and optimization methods, with no external datasets. The assumptions underlying the theoretical results are stated alongside each result, and the derivations and proofs are provided in Appendices A–D.

REFERENCES

Francis Bach. Breaking the curse of dimensionality with convex neural networks. *Journal of Machine Learning Research*, 18(19):1–53, 2017. URL https://jmlr.org/papers/v18/14-546.html.

Etienne Boursier and Nicolas Flammarion. Early alignment in two-layer networks training is a two-edged sword. *Journal of Machine Learning Research*, 26(183):1–75, 2025. URL https://jmlr.org/papers/v26/24-1523.html.

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. URL https://proceedings.neurips.cc/paper_files/paper/2022/hash/7eeb9af3eb1f48e29c05e8dd3342b286-Abstract-Conference.html.

Zheng-An Chen and Tao Luo. From condensation to rank collapse: A two-stage analysis of transformer training dynamics. In *Advances in Neural Information Processing Systems*, volume 38, 2025. URL https://proceedings.neurips.cc/paper_files/paper/2025/hash/3b576711b12ab036b45130fc8eb78504-Abstract-Conference.html.

Zheng-An Chen, Yuqing Li, Tao Luo, Zhangchen Zhou, and Zhi-Qin John Xu. Phase diagram of initial condensation for two-layer neural networks. *CSIAM Transactions on Applied Mathematics*, 5(3):448–514, 2024. doi: 10.4208/csiam-am.SO-2023-0016. URL https://doi.org/10.4208/csiam-am.SO-2023-0016.

Lénaïc Chizat and Francis Bach. Implicit bias of gradient descent for wide two-layer neural networks trained with the logistic loss. In *Proceedings of the Conference on Learning Theory*, volume 125 of *Proceedings of Machine Learning Research*, pp. 1305–1338. PMLR, 2020. URL https://proceedings.mlr.press/v125/chizat20a.html.

Lénaïc Chizat, Edouard Oyallon, and Francis Bach. On lazy training in differentiable programming. In *Advances in Neural Information Processing Systems*, volume 32, 2019. URL https://proceedings.neurips.cc/paper_files/paper/2019/hash/ae614c557843b1df326cb29c57225459-Abstract.html.

Arthur Jacot, Franck Gabriel, and Clément Hongler. Neural tangent kernel: Convergence and generalization in neural networks. In *Advances in Neural Information Processing Systems*, volume 31, 2018. URL https://proceedings.neurips.cc/paper_files/paper/2018/hash/5a4be1fa34e62bb8a6ec6b91d2462f5a-Abstract.html.

Arthur Jacot, Peter Súkeník, Zihan Wang, and Marco Mondelli. Wide neural networks trained with weight decay provably exhibit neural collapse. In *International Conference on Learning Representations*, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/hash/d33d31a87eb276ffe47fa7324648098b-Abstract-Conference.html.

Akshay Kumar and Jarvis Haupt. Directional convergence near small initializations and saddles in two-homogeneous neural networks. *Transactions on Machine Learning Research*, 2024. URL https://openreview.net/forum?id=hfrPag75Y0.

Tanishq Kumar, Blake Bordelon, Samuel J. Gershman, and Cengiz Pehlevan. Grokking as the transition from lazy to rich training dynamics. In *International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=vt5mnLVIVo.

Ilja Kuzborskij and Yasin Abbasi-Yadkori. Low-rank bias, weight decay, and model merging in neural networks. In *Proceedings of The 29th International Conference on Artificial Intelligence and Statistics*, volume 300 of *Proceedings of Machine Learning Research*, pp. 460–468. PMLR, 2026. URL https://proceedings.mlr.press/v300/kuzborskij26a.html.

Aitor Lewkowycz and Guy Gur-Ari. On the training dynamics of deep networks with L_2 regularization. In *Advances in Neural Information Processing Systems*, volume 33, 2020. URL https://proceedings.neurips.cc/paper_files/paper/2020/hash/32fcc8cfe1fa4c77b5c58dafd36d1a98-Abstract.html.

Ziming Liu, Eric J. Michaud, and Max Tegmark. Omnigrok: Grokking beyond algorithmic data. In *International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=zDiHoIWa0q1.

Tao Luo, Zhi-Qin John Xu, Zheng Ma, and Yaoyu Zhang. Phase diagram for two-layer ReLU neural networks at infinite-width limit. *Journal of Machine Learning Research*, 22(71):1–47, 2021. URL https://jmlr.org/papers/v22/20-1123.html.

Kaifeng Lyu, Jikai Jin, Zhiyuan Li, Simon S. Du, Jason D. Lee, and Wei Hu. Dichotomy of early and late phase implicit biases can provably induce grokking. In *International Conference on Learning Representations*, 2024. URL https://proceedings.iclr.cc/paper_files/paper/2024/hash/909c8fef63e1cede406ce9e6794f99a2-Abstract-Conference.html.

Hartmut Maennel, Olivier Bousquet, and Sylvain Gelly. Gradient descent quantizes ReLU network features. *arXiv preprint arXiv:1803.08367*, 2018. URL https://arxiv.org/abs/1803.08367.

Rahul Parhi and Robert D. Nowak. Banach space representer theorems for neural networks and ridge splines. *Journal of Machine Learning Research*, 22(43):1–40, 2021. URL https://jmlr.org/papers/v22/20-583.html.

Mert Pilanci and Tolga Ergen. Neural networks are convex regularizers: Exact polynomial-time convex optimization formulations for two-layer networks. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pp. 7695–7705. PMLR, 2020. URL https://proceedings.mlr.press/v119/pilanci20a.html.

Alethea Power, Yuri Burda, Harri Edwards, Igor Babuschkin, and Vedant Misra. Grokking: Generalization beyond overfitting on small algorithmic datasets. *arXiv preprint arXiv:2201.02177*, 2022. URL https://arxiv.org/abs/2201.02177.

Pedro Savarese, Itay Evron, Daniel Soudry, and Nathan Srebro. How do infinite width bounded norm networks look in function space? In *Proceedings of the Conference on Learning Theory*, volume 99 of *Proceedings of Machine Learning Research*, pp. 2667–2690. PMLR, 2019. URL https://proceedings.mlr.press/v99/savarese19a.html.

Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. Dropout: A simple way to prevent neural networks from overfitting. *Journal of Machine Learning Research*, 15(56):1929–1958, 2014.

Zhi-Qin John Xu, Yaoyu Zhang, and Tao Luo. Overview frequency principle/spectral bias in deep learning. *Communications on Applied Mathematics and Computation*, 7(3):827–864, 2025a. URL https://link.springer.com/article/10.1007/s42967-024-00398-7.

Zhi-Qin John Xu, Yaoyu Zhang, and Zhangchen Zhou. An overview of condensation phenomenon in deep learning. *arXiv preprint arXiv:2504.09484*, 2025b. URL https://arxiv.org/abs/2504.09484.

Chiyuan Zhang, Samy Bengio, Moritz Hardt, Benjamin Recht, and Oriol Vinyals. Understanding deep learning requires rethinking generalization. In *International Conference on Learning Representations*, 2017.

Yaoyu Zhang, Zhongwang Zhang, Tao Luo, and Zhi-Qin John Xu. Embedding principle of loss landscape of deep neural networks. In *NeurIPS 2021 spotlight*, volume 34, 2021. URL https://papers.nips.cc/paper/2021/hash/7cc532d783a7461f227a5da8ea80bfe1-Abstract.html.

Yaoyu Zhang, Yuqing Li, Zhongwang Zhang, Tao Luo, and Zhi-Qin John Xu. Embedding principle: a hierarchical structure of loss landscape of deep neural networks. *Journal of Machine Learning*, pp. 60–113, 2022. doi: 10.4208/jml.220108. URL https://www.global-sci.org/intro/article_detail/jml/20372.html.

Zhongwang Zhang and Zhi-Qin John Xu. Implicit regularization of dropout. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 46(6):4206–4217, 2024. URL https://arxiv.org/abs/2207.05952.

Zhongwang Zhang, Pengxiao Lin, Zhiwei Wang, Yaoyu Zhang, and Zhi-Qin John Xu. Initialization is critical to whether transformers fit composite functions by reasoning or memorizing. In *Advances in Neural Information Processing Systems*, volume 37, 2024. URL https://proceedings.neurips.cc/paper_files/paper/2024/hash/19c145aaad40927c51f4d10eaa339c20-Abstract-Conference.html.

Zhongwang Zhang, Zhiwei Wang, Junjie Yao, Zhangchen Zhou, Xiaolong Li, Weinan E, and Zhi-Qin John Xu. Anchor function: A type of benchmark functions for studying language models. *Journal of Machine Learning*, 5(2):97–134, 2026. doi: 10.4208/jml.250723. URL https://doi.org/10.4208/jml.250723.

Hanxu Zhou, Qixuan Zhou, Tao Luo, Yaoyu Zhang, and Zhi-Qin John Xu. Towards understanding the condensation of neural networks at initial training. In *Advances in Neural Information Processing Systems*, volume 35, 2022. URL https://proceedings.neurips.cc/paper_files/paper/2022/hash/0f4d1fc085b7504c140e66bb26ed8842-Abstract-Conference.html.

A NEURON DYNAMICS

A.1 NEURON DYNAMICS AND BALANCE

We derive Proposition 1 using the loss normalization $\mathcal{L} = \|\boldsymbol{e}\|_n^2$ in equation (2). Write $\boldsymbol{u}_k = \boldsymbol{u}(\theta_k)$, $\boldsymbol{u}_{\perp,k} = \boldsymbol{u}_\perp(\theta_k) = (-\sin\theta_k, \cos\theta_k)^\top$, let I_k be the activation cell containing θ_k, and set $\chi_{ki} = \mathbf{1}_{\{\boldsymbol{v}_k^\top \tilde{\boldsymbol{x}}_i > 0\}}$ inside an open cell. Define

$$\boldsymbol{\phi}_k(t) := \boldsymbol{\phi}_{\theta_k(t)}(\boldsymbol{X}), \qquad \boldsymbol{\psi}_k(t) := \boldsymbol{\psi}_{\theta_k(t)}(\boldsymbol{X}).$$

The derivatives of a training prediction are

$$\partial_{a_k} f_\Theta(x_i) = r_k(\boldsymbol{\phi}_k)_i, \qquad \nabla_{\boldsymbol{v}_k} f_\Theta(x_i) = a_k \chi_{ki} \tilde{\boldsymbol{x}}_i.$$

Consequently, with $\boldsymbol{M}_k = n^{-1}\sum_i e_i \chi_{ki} \tilde{\boldsymbol{x}}_i = \boldsymbol{M}_{I_k}(\boldsymbol{e})$,

$$\begin{aligned}
\partial_{a_k}\mathcal{L} &= \frac{2}{n}\sum_i e_i r_k(\boldsymbol{\phi}_k)_i = 2 r_k q_k, \\
\nabla_{\boldsymbol{v}_k}\mathcal{L} &= \frac{2}{n}\sum_i e_i a_k \chi_{ki} \tilde{\boldsymbol{x}}_i = 2 a_k \boldsymbol{M}_k, \\
\frac{\mathrm{d}a_k}{\mathrm{d}t} &= -2 r_k q_k - \lambda a_k, \qquad \frac{\mathrm{d}\boldsymbol{v}_k}{\mathrm{d}t} = -2 a_k \boldsymbol{M}_k - \lambda \boldsymbol{v}_k.
\end{aligned} \tag{24}$$

In particular, the two components of the last equation are

$$\frac{\mathrm{d}w_k}{\mathrm{d}t} = -\frac{2a_k}{n}\sum_i e_i \chi_{ki} x_i - \lambda w_k, \qquad \frac{\mathrm{d}b_k}{\mathrm{d}t} = -\frac{2a_k}{n}\sum_i e_i \chi_{ki} - \lambda b_k.$$

Radial and angular projections. For $r_k > 0$, differentiating $\boldsymbol{v}_k = r_k \boldsymbol{u}_k$ gives $\frac{\mathrm{d}\boldsymbol{v}_k}{\mathrm{d}t} = \frac{\mathrm{d}r_k}{\mathrm{d}t}\boldsymbol{u}_k + r_k \frac{\mathrm{d}\theta_k}{\mathrm{d}t}\boldsymbol{u}_{\perp,k}$. The orthonormality of $\boldsymbol{u}_k, \boldsymbol{u}_{\perp,k}$ and the identities $\boldsymbol{u}_k^\top \boldsymbol{M}_k = q_k$, $\boldsymbol{u}_{\perp,k}^\top \boldsymbol{M}_k = p_k$ imply

$$\begin{aligned}
\frac{\mathrm{d}r_k}{\mathrm{d}t} &= \boldsymbol{u}_k^\top \frac{\mathrm{d}\boldsymbol{v}_k}{\mathrm{d}t} = -2 a_k q_k - \lambda r_k, \\
r_k \frac{\mathrm{d}\theta_k}{\mathrm{d}t} &= \boldsymbol{u}_{\perp,k}^\top \frac{\mathrm{d}\boldsymbol{v}_k}{\mathrm{d}t} = -2 a_k p_k.
\end{aligned} \tag{25}$$

Dividing the second equation by r_k gives the angular equation in equation (5). The decay term has zero tangential projection because $\boldsymbol{u}_{\perp,k}^\top \boldsymbol{v}_k = 0$.

Balance and inactive neurons. The mixed data-gradient terms cancel exactly:

$$\begin{aligned}
\frac{\mathrm{d}B_k}{\mathrm{d}t} &= 2a_k(-2 r_k q_k - \lambda a_k) - 2 r_k(-2 a_k q_k - \lambda r_k) \\
&= -2\lambda(a_k^2 - r_k^2) = -2\lambda B_k.
\end{aligned} \tag{26}$$

Multiplication by $\mathrm{e}^{2\lambda t}$ proves equation (7). It also gives the absolute scale bound

$$\big||a_k(t)| - r_k(t)\big|^2 \le |a_k(t)^2 - r_k(t)^2| = \mathrm{e}^{-2\lambda t}|B_k(0)|.$$

If $\boldsymbol{v}_k(0)^\top \tilde{\boldsymbol{x}}_i < 0$ for every i, all data gradients vanish. The solution is $a_k(t) = \mathrm{e}^{-\lambda t} a_k(0)$ and $\boldsymbol{v}_k(t) = \mathrm{e}^{-\lambda t}\boldsymbol{v}_k(0)$. Every strict inactive inequality therefore persists, θ_k stays fixed when defined, and $A_k(t) = \mathrm{e}^{-2\lambda t} A_k(0)$.

Residual dynamics. The parameter equation is $\frac{\mathrm{d}\Theta}{\mathrm{d}t} = -2J_t^\top \boldsymbol{e}/n - \lambda\Theta$. For each sample, the contribution of neuron k to $J_t\Theta$ is

$$a_k \partial_{a_k} f_\Theta(x_i) + \boldsymbol{v}_k^\top \nabla_{\boldsymbol{v}_k} f_\Theta(x_i) = 2 a_k \,\mathrm{ReLU}(\boldsymbol{v}_k^\top \tilde{\boldsymbol{x}}_i).$$

Thus $J_t\Theta = 2 f_\Theta(\boldsymbol{X})$, and the chain rule gives

$$\begin{aligned}
\frac{\mathrm{d}\boldsymbol{e}}{\mathrm{d}t} = J_t \frac{\mathrm{d}\Theta}{\mathrm{d}t} &= -\frac{2}{n} J_t J_t^\top \boldsymbol{e} - \lambda J_t \Theta \\
&= -2K_t \boldsymbol{e} - 2\lambda f_\Theta(\boldsymbol{X}) = -2(K_t + \lambda I_n)\boldsymbol{e} - 2\lambda \boldsymbol{y}.
\end{aligned} \tag{27}$$

For fixed $K \succeq 0$, setting the residual velocity to zero gives $\boldsymbol{e}_{\mathrm{eq}}(K) = -\lambda(K + \lambda I_n)^{-1}\boldsymbol{y}$.

At zero preactivation, use the same selection $\chi_{ki} \in [0, 1]$ in the parameter flow and Jacobian, with p_k denoting the selected angular gradient. Along an absolutely continuous trajectory, preactivation derivatives vanish almost everywhere on their zero sets; the ReLU chain rule and the identities above therefore hold almost everywhere, including boundary contacts. In particular,

$$\frac{\mathrm{d}}{\mathrm{d}t}\Phi(\Theta(t)) = -\left\|\frac{\mathrm{d}\Theta}{\mathrm{d}t}\right\|_2^2 \leq 0.$$

A.2 General and balanced amplitude dynamics

For $r_k > 0$, the quotient rule and equation (5) give

$$\begin{aligned}\frac{\mathrm{d}\beta_k}{\mathrm{d}t} &= \frac{\frac{\mathrm{d}a_k}{\mathrm{d}t}r_k - a_k\frac{\mathrm{d}r_k}{\mathrm{d}t}}{r_k^2} \\ &= \frac{(-2r_kq_k - \lambda a_k)r_k - a_k(-2a_kq_k - \lambda r_k)}{r_k^2} = 2q_k(\beta_k^2 - 1).\end{aligned} \tag{28}$$

On any interval where $a_k \neq 0$, its sign s_k is constant and $A_k = |a_k|r_k > 0$. The logarithmic derivative is

$$\begin{aligned}\frac{\frac{\mathrm{d}A_k}{\mathrm{d}t}}{A_k} &= \frac{\frac{\mathrm{d}a_k}{\mathrm{d}t}}{a_k} + \frac{\frac{\mathrm{d}r_k}{\mathrm{d}t}}{r_k} \\ &= -2\frac{r_k}{a_k}q_k - \lambda - 2\frac{a_k}{r_k}q_k - \lambda = -2\lambda - 2q_k(\beta_k + \beta_k^{-1}).\end{aligned} \tag{29}$$

Equivalently, $\frac{\mathrm{d}A_k}{\mathrm{d}t} = -2\lambda A_k - 2s_k(a_k^2 + r_k^2)q_k$. If $B_k(T) = 0$, the balance identity preserves $|a_k| = r_k$; for a nonzero neuron this means $\beta_k = s_k$ and $A_k = a_k^2 = r_k^2$. Substitution gives the balanced dynamics

$$\frac{\mathrm{d}A_k}{\mathrm{d}t} = -2A_k(\lambda + 2s_kq_k), \qquad \frac{\mathrm{d}\theta_k}{\mathrm{d}t} = -2s_kp_k. \tag{30}$$

In this regime,

$$A_k(t) = A_k(T)\exp\left[-2\lambda(t - T) - 4s_k\int_T^t q_k(s)\,ds\right],$$

so a positive amplitude remains positive on every finite interval with bounded correlations. The general equations retain the evolving ratio β_k; under exact balance, $\beta_k = s_k$ gives the reduced system equation (30). For surviving positive neurons with nearly stationary amplitudes and directions close to attracting minima, these dynamics give $-2q_t(\theta_*)/\lambda \approx 1$, as shown in Figure 3(d).

B Fast residual equilibration

B.1 Proof of Theorem 2: initialization

The proof has three steps: initialization estimates, short-time kernel stability, and transfer to the residual scale.

Writing α_{ij} for the angle between $\tilde{\boldsymbol{x}}_i$ and $\tilde{\boldsymbol{x}}_j$, the limiting kernel is

$$\bar{K}_{ij} = \frac{\|\tilde{\boldsymbol{x}}_i\|_2\,\|\tilde{\boldsymbol{x}}_j\|_2}{2\pi n}\left[\sin\alpha_{ij} + 2(\pi - \alpha_{ij})\cos\alpha_{ij}\right]. \tag{31}$$

Let $\sigma_m = m^{-\gamma}$, and write $a_k(0) = \sigma_m\xi_k$, $\boldsymbol{v}_k(0) = \sigma_m\boldsymbol{g}_k$ with independent standard Gaussian ξ_k and $\boldsymbol{g}_k \in \mathbb{R}^2$. Each output summand has mean zero and variance

$$\mathbb{E}\left[a_k(0)^2\,\mathrm{ReLU}(\boldsymbol{v}_k(0)^\top\tilde{\boldsymbol{x}}_i)^2\right] = \sigma_m^4\mathbb{E}\xi_k^2\,\mathbb{E}\,\mathrm{ReLU}(\boldsymbol{g}_k^\top\tilde{\boldsymbol{x}}_i)^2 = \frac{\sigma_m^4}{2}\,\|\tilde{\boldsymbol{x}}_i\|_2^2\,.$$

Independence across neurons gives $\mathrm{Var}(f_{\Theta(0)}(x_i)) = (1+x_i^2)m^{1-4\gamma}/2$. For fixed n, $\boldsymbol{y}$ and $\gamma \geq 1/4$, summing these second moments gives $\|\boldsymbol{e}(0)\|_2 = O_p(1)$.

With $\chi_{ki}(t) = \mathbf{1}_{\{\boldsymbol{v}_k(t)^\top \tilde{\boldsymbol{x}}_i > 0\}}$, the empirical kernel entries are

$$(K_t)_{ij} = \frac{1}{n} \sum_k \left[\mathrm{ReLU}(\boldsymbol{v}_k^\top \tilde{\boldsymbol{x}}_i)\, \mathrm{ReLU}(\boldsymbol{v}_k^\top \tilde{\boldsymbol{x}}_j) + a_k^2 \chi_{ki}\chi_{kj} \tilde{\boldsymbol{x}}_i^\top \tilde{\boldsymbol{x}}_j \right].$$

To compute the expectation, rotate coordinates so that the directions of $\tilde{\boldsymbol{x}}_i, \tilde{\boldsymbol{x}}_j$ have angles $0, \alpha$, where $\alpha = \alpha_{ij} \in [0, \pi]$. Write a standard Gaussian as $\boldsymbol{g} = R_g(\cos\varphi, \sin\varphi)$, with φ uniform on $[0, 2\pi)$, independent of R_g, and $\mathbb{E}R_g^2 = 2$. The common positive half-planes overlap on $[\alpha - \pi/2, \pi/2]$, modulo 2π. Hence

$$\begin{aligned}
\mathbb{E}[\mathrm{ReLU}(\boldsymbol{g}^\top \tilde{\boldsymbol{x}}_i)\, \mathrm{ReLU}(\boldsymbol{g}^\top \tilde{\boldsymbol{x}}_j)] &= \frac{\|\tilde{\boldsymbol{x}}_i\|_2 \|\tilde{\boldsymbol{x}}_j\|_2}{\pi} \int_{\alpha-\pi/2}^{\pi/2} \cos\varphi \cos(\varphi - \alpha)\, d\varphi \\
&= \frac{\|\tilde{\boldsymbol{x}}_i\|_2 \|\tilde{\boldsymbol{x}}_j\|_2}{2\pi} [\sin\alpha + (\pi - \alpha)\cos\alpha], \\
\mathbb{P}(\boldsymbol{g}^\top \tilde{\boldsymbol{x}}_i > 0,\ \boldsymbol{g}^\top \tilde{\boldsymbol{x}}_j > 0) &= \frac{\pi - \alpha}{2\pi}. \qquad (32)
\end{aligned}$$

The integral follows from $2\cos\varphi\cos(\varphi - \alpha) = \cos\alpha + \cos(2\varphi - \alpha)$. Adding the second kernel term, using $\mathbb{E}\xi_k^2 = 1$ and $\tilde{\boldsymbol{x}}_i^\top \tilde{\boldsymbol{x}}_j = \|\tilde{\boldsymbol{x}}_i\|_2 \|\tilde{\boldsymbol{x}}_j\|_2 \cos\alpha$, proves equation (31).

Since $\rho_m = m\sigma_m^2$, the law of large numbers for the independent kernel summands gives $K_0/\rho_m \to \bar{K}$ in operator norm in probability, with n fixed. For positivity, a zero quadratic form of the random-feature contribution would imply

$$\sum_i z_i\, \mathrm{ReLU}(\boldsymbol{g}^\top \tilde{\boldsymbol{x}}_i) = 0$$

for almost every Gaussian $\boldsymbol{g}$, hence for every $\boldsymbol{g}$ by continuity. At a nonzero point on $\boldsymbol{g}^\top \tilde{\boldsymbol{x}}_j = 0$, the other terms are smooth because the inputs are distinct. The gradient jump is therefore $z_j \tilde{\boldsymbol{x}}_j = 0$, so every $z_j = 0$ and $\bar{K} \succ 0$.

B.2 Proof of Theorem 2: kernel stability

Residual and parameter bounds. Let h_m be a deterministic horizon with $h_m\sqrt{\log m} \to 0$. Positivity of K_t gives

$$\frac{1}{2}\frac{\mathrm{d}}{\mathrm{d}t}\|\boldsymbol{e}\|_2^2 = -2\boldsymbol{e}^\top K_t \boldsymbol{e} - 2\lambda\|\boldsymbol{e}\|_2^2 - 2\lambda\boldsymbol{e}^\top \boldsymbol{y} \le 2\lambda\|\boldsymbol{e}\|_2\|\boldsymbol{y}\|_2.$$

Thus $\mathrm{d}\|\boldsymbol{e}\|_2/\mathrm{d}t \le 2\lambda\|\boldsymbol{y}\|_2$ (in the upper Dini sense at $\boldsymbol{e} = 0$). Since $\|\boldsymbol{e}(0)\|_2 = O_p(1)$, the residual is uniformly bounded in probability on $[0, h_m]$. Localize to an arbitrarily high-probability event on which this bound and the Gaussian initial-norm bounds below have fixed constants. With $C_x = (n^{-1}\sum_i \|\tilde{\boldsymbol{x}}_i\|_2^2)^{1/2}$,

$$\|\boldsymbol{M}_k\|_2 \le C_x \|\boldsymbol{e}\|_n, \qquad |q_k| \le C_x \|\boldsymbol{e}\|_n.$$

For $\boldsymbol{\eta}_k = (a_k, \boldsymbol{v}_k)$, equation (24) consequently gives $\left\|\frac{\mathrm{d}\boldsymbol{\eta}_k}{\mathrm{d}t}\right\|_2 \le C\|\boldsymbol{\eta}_k\|_2$, with C independent of k, m. Gronwall's inequality yields

$$\begin{aligned}
\|\boldsymbol{\eta}_k(t)\|_2 &\le \mathrm{e}^{Ct}\|\boldsymbol{\eta}_k(0)\|_2, \\
\|\boldsymbol{\eta}_k(t) - \boldsymbol{\eta}_k(0)\|_2 &\le (\mathrm{e}^{Ct} - 1)\|\boldsymbol{\eta}_k(0)\|_2.
\end{aligned}$$

In particular,

$$\sup_{t \le h_m} \|\boldsymbol{\eta}_k(t) - \boldsymbol{\eta}_k(0)\|_2 \le C h_m \sqrt{a_k(0)^2 + r_k(0)^2}.$$

The Gaussian initialization also gives

$$\max_k \|\boldsymbol{\eta}_k(0)\|_2 = O_p(m^{-\gamma}\sqrt{\log m}), \qquad \frac{1}{m}\sum_k m^{2\gamma}\|\boldsymbol{\eta}_k(0)\|_2^2 = O_p(1).$$

Boundary contacts and kernel variation. A contact with the boundary of sample i can occur only if

$$|\boldsymbol{v}_k(0)^\top \tilde{\boldsymbol{x}}_i| \le C m^{-\gamma} h_m \sqrt{\log m}$$

on the localized event. Lipschitz continuity controls the feature products and changes in a_k^2; changes in activation masks are bounded by the initial squared output weights on these contact events. Hence

$$\sup_{t \le h_m} \frac{\|K_t - K_0\|_{\mathrm{op}}}{\rho_m} \le \frac{C h_m}{m} \sum_k m^{2\gamma}(a_k(0)^2 + r_k(0)^2) + \frac{C}{m} \sum_k m^{2\gamma} a_k(0)^2 \sum_i \mathbf{1}_{\{|\boldsymbol{v}_k(0)^\top \tilde{\boldsymbol{x}}_i| \le C m^{-\gamma} h_m \sqrt{\log m}\}}.$$

Independence of $a_k(0)$ and $\boldsymbol{v}_k(0)$ and the bounded Gaussian density of $\boldsymbol{g}_k^\top \tilde{\boldsymbol{x}}_i$ imply

$$\mathbb{E}\left[\xi_k^2 \mathbf{1}_{\{|\boldsymbol{g}_k^\top \tilde{\boldsymbol{x}}_i| \le C h_m \sqrt{\log m}\}}\right] = \mathbb{E}\xi_k^2 \, \mathbb{P}(|\boldsymbol{g}_k^\top \tilde{\boldsymbol{x}}_i| \le C h_m \sqrt{\log m}) \le C' h_m \sqrt{\log m}.$$

Since n is fixed, Markov's inequality gives

$$\sup_{t \le h_m} \frac{\|K_t - K_0\|_{\mathrm{op}}}{\rho_m} = O_p(h_m \sqrt{\log m}).$$

Choosing

$$h_m = \frac{\rho_m^{-1}}{\lambda_{\min}(\bar{K})} \log(\log m / \varepsilon_m)$$

ensures $T_{\mathrm{I}} \le h_m$ with probability tending to one and proves equation (11).

B.3 Proof of Theorem 2: residual equilibration

Recall $\boldsymbol{e}_{\mathrm{qs}}(0) = -\lambda(K_0 + \lambda I_n)^{-1}\boldsymbol{y}$, and set $\Delta_m = \sup_{t \le T_{\mathrm{I}}} \|K_t - K_0\|_{\mathrm{op}}$. Then $\Delta_m/\kappa_m = o_p(1)$ and $\Delta_m T_{\mathrm{I}} = O_p(\rho_m^{-1}(\log m)^{5/2}) = o_p(1)$. For $\boldsymbol{z}_0(t) = \boldsymbol{e}(t) - \boldsymbol{e}_{\mathrm{qs}}(0)$, subtract the frozen equilibrium relation to obtain

$$\frac{\mathrm{d}\boldsymbol{z}_0}{\mathrm{d}t} = -2(K_t + \lambda I_n)\boldsymbol{z}_0 - 2(K_t - K_0)\boldsymbol{e}_{\mathrm{qs}}(0).$$

On $\Delta_m < \kappa_m$, the homogeneous evolution operator $U(t,s)$ satisfies $\|U(t,s)\|_{\mathrm{op}} \le \mathrm{e}^{-2(\kappa_m + \lambda - \Delta_m)(t-s)}$. This follows by differentiating the squared norm of a homogeneous solution and using $K_t \succeq (\kappa_m - \Delta_m) I_n$. Variation of constants gives

$$\boldsymbol{z}_0(t) = U(t,0)\boldsymbol{z}_0(0) - 2\int_0^t U(t,s)(K_s - K_0)\boldsymbol{e}_{\mathrm{qs}}(0)\, ds.$$

Bounding the integral by its exponential envelope yields

$$\|\boldsymbol{z}_0(T_{\mathrm{I}})\|_2 \le \mathrm{e}^{-2(\kappa_m + \lambda - \Delta_m)T_{\mathrm{I}}} \|\boldsymbol{z}_0(0)\|_2 + \frac{\Delta_m}{\kappa_m + \lambda - \Delta_m} \|\boldsymbol{e}_{\mathrm{qs}}(0)\|_2. \tag{33}$$

The first term is $O_p(\varepsilon_m/\log m)\mathrm{e}^{2\Delta_m T_{\mathrm{I}}}$ and the second is $o_p(\varepsilon_m)$. Since $\boldsymbol{e}_{\mathrm{qs}}(0) = -\varepsilon_m \bar{K}^{-1}\boldsymbol{y} + o_p(\varepsilon_m)$, the first expansion follows. The resolvent identity gives

$$\boldsymbol{e}_{\mathrm{qs}}(t) - \boldsymbol{e}_{\mathrm{qs}}(0) = \lambda(K_t + \lambda I_n)^{-1}(K_t - K_0)(K_0 + \lambda I_n)^{-1}\boldsymbol{y},$$
$$\|\boldsymbol{e}_{\mathrm{qs}}(t) - \boldsymbol{e}_{\mathrm{qs}}(0)\|_2 \le \frac{\lambda \Delta_m \|\boldsymbol{y}\|_2}{(\kappa_m + \lambda - \Delta_m)(\kappa_m + \lambda)} = o_p(\varepsilon_m), \qquad t \le T_{\mathrm{I}}. \tag{34}$$

Here $\kappa_m/\rho_m \to \lambda_{\min}(\bar{K}) > 0$ and $\Delta_m/\rho_m = o_p(1)$. Combining this bound with the expansion of $\boldsymbol{e}(T_{\mathrm{I}})$ proves its claimed tracking error relative to $\boldsymbol{e}_{\mathrm{qs}}(T_{\mathrm{I}})$ and completes Theorem 2.

C Discrete extrema and amplification of q_e

C.1 Proof of Proposition 3

Proof. Each equation $\tilde{\boldsymbol{x}}_i^\top \boldsymbol{u}(\theta) = 0$ contributes two antipodal boundary angles. Thus there are at most $2n$ boundaries and at most $2n$ open activation cells, each of length at most π.

On a non-dead cell I, the assumption $\boldsymbol{M}_I(\boldsymbol{e}) \neq \boldsymbol{0}$ makes $q_e(\theta) = \boldsymbol{M}_I(\boldsymbol{e})^\top \boldsymbol{u}(\theta)$ a nonconstant sinusoid. Its critical points are spaced by π, so each open cell contains at most one smooth extremum. A dead cell has $q_e = 0$ and no strict interior extremum. Consequently,

$$|\mathcal{E}_{\mathrm{int}}(\boldsymbol{e})| \leq 2n, \qquad |\mathcal{E}_{\mathrm{bd}}(\boldsymbol{e})| \leq |\mathcal{B}| \leq 2n.$$

Every strict local extremum is either inside a cell or at an activation boundary, giving $|\mathcal{E}(\boldsymbol{e})| \leq 4n$.

Finally, q_e is continuous on the circle and nonzero on some non-dead cell, so it attains a nonzero global maximum or minimum. The points attaining this value cannot lie in dead-cell interiors; they belong to the finite set of interior critical points and activation boundaries. Each such point is therefore isolated among the global optimizers and is a strict local extremum. Hence $|\mathcal{E}(\boldsymbol{e})| \geq 1$. □

Activation boundaries are fixed by the inputs. A boundary extremum therefore has fixed position while it persists; for positive neurons, the one-sided condition $p_t(\theta_b^-) < 0 < p_t(\theta_b^+)$ gives attraction toward the boundary.

C.2 Proof of Theorem 4

Proof. Write $K_* = K_{t_0}$, $\kappa_* := \lambda_{\min}(K_*) > 0$, $\tau = t - t_0$, and $\alpha(t) = \mathrm{e}^{-2\lambda\tau}$. For the kernel calculation, abbreviate $q_k = q_t(\theta_k(t))$ and $p_k = p_t(\theta_k(t))$.

Kernel derivative. For each neuron, let χ_{ki} be its fixed activation indicator on input i, and let

$$(S_k)_{ij} := \chi_{ki}\chi_{kj}\tilde{\boldsymbol{x}}_i^\top \tilde{\boldsymbol{x}}_j.$$

The polar identity $\chi_{ki}\tilde{\boldsymbol{x}}_i = (\boldsymbol{\phi}_k)_i \boldsymbol{u}_k + (\boldsymbol{\psi}_k)_i \boldsymbol{u}_{\perp,k}$ gives $S_k = \boldsymbol{\phi}_k\boldsymbol{\phi}_k^\top + \boldsymbol{\psi}_k\boldsymbol{\psi}_k^\top$. The matrix S_k is constant while the mask is fixed. The contribution of neuron k to the kernel is therefore

$$K_k = \frac{1}{n}(r_k^2 \boldsymbol{\phi}_k\boldsymbol{\phi}_k^\top + a_k^2 S_k), \qquad K_t = \sum_k K_k.$$

Using $\frac{\mathrm{d}\boldsymbol{\phi}_k}{\mathrm{d}t} = \boldsymbol{\psi}_k \frac{\mathrm{d}\theta_k}{\mathrm{d}t}$ and the parameter equations, we obtain

$$\begin{aligned} \frac{\mathrm{d}K_k}{\mathrm{d}t} &= -2\lambda K_k + E_k, \\ E_k &= -\frac{4a_k r_k q_k}{n}(\boldsymbol{\phi}_k\boldsymbol{\phi}_k^\top + S_k) - \frac{2a_k r_k p_k}{n}(\boldsymbol{\psi}_k\boldsymbol{\phi}_k^\top + \boldsymbol{\phi}_k\boldsymbol{\psi}_k^\top). \end{aligned} \tag{35}$$

For any real vectors $\boldsymbol{v}, \boldsymbol{w}$, $-(\boldsymbol{v}\boldsymbol{v}^\top + \boldsymbol{w}\boldsymbol{w}^\top) \preceq \boldsymbol{v}\boldsymbol{w}^\top + \boldsymbol{w}\boldsymbol{v}^\top \preceq \boldsymbol{v}\boldsymbol{v}^\top + \boldsymbol{w}\boldsymbol{w}^\top$. Also $|a_k r_k| \leq B r_k^2$ and $|a_k r_k| \leq B a_k^2$ by the ratio assumption. Thus the first term of E_k is bounded above and below by $\pm 4B|q_k|K_k$, and the second by $\pm 2B|p_k|K_k$. Dead neurons contribute zero to K_t. Summing and using $|q_k| + |p_k| \leq \varepsilon\lambda$ gives

$$\frac{\mathrm{d}K_t}{\mathrm{d}t} = -2\lambda K_t + E(t), \qquad -4B\varepsilon\lambda K_t \preceq E(t) \preceq 4B\varepsilon\lambda K_t. \tag{36}$$

Kernel decay. Let $F(t) = \alpha(t)^{-1} K_t$. For each fixed vector $\boldsymbol{v}$, equation (36) implies

$$-4B\varepsilon\lambda\, \boldsymbol{v}^\top F \boldsymbol{v} \leq \frac{\mathrm{d}}{\mathrm{d}t}(\boldsymbol{v}^\top F \boldsymbol{v}) \leq 4B\varepsilon\lambda\, \boldsymbol{v}^\top F \boldsymbol{v}.$$

Integrating these scalar inequalities yields

$$\mathrm{e}^{-h(t)} K_* \preceq F(t) \preceq \mathrm{e}^{h(t)} K_*, \qquad h(t) := 4B\varepsilon\lambda\tau \leq 4BL\varepsilon. \tag{37}$$

Consequently,

$$\|F(t) - K_*\|_{\mathrm{op}} \leq (\mathrm{e}^{h(t)} - 1)\, \|K_*\|_{\mathrm{op}} \leq C_1 \varepsilon\, \|K_*\|_{\mathrm{op}},$$

where C_1 depends only on B, L for sufficiently small ε. Multiplication by $\alpha(t)$ proves equation (14). In particular, the kernel remains positive definite on the interval.

Quasi-static scaling. Using the actual kernel just estimated, set $\boldsymbol{e}_{\mathrm{qs}}(t) = -\lambda(K_t + \lambda I_n)^{-1}\boldsymbol{y}$. Let $A_0 = K_* + \lambda I_n$ and

$$D(t) = F(t) - K_* + \lambda(\alpha(t)^{-1} - 1)I_n.$$

Then $K_t + \lambda I_n = \alpha(t)(A_0 + D(t))$. The inverse-matrix identity gives the exact comparison

$$\boldsymbol{e}_{\mathrm{qs}}(t) - \alpha(t)^{-1}\boldsymbol{e}_{\mathrm{qs}}(t_0) = -\alpha(t)^{-1}(A_0 + D(t))^{-1}D(t)\boldsymbol{e}_{\mathrm{qs}}(t_0). \tag{38}$$

By equation (37), $\lambda_{\min}(A_0 + D(t)) \geq \mathrm{e}^{-h(t)}\kappa_*$. Moreover, using the spectral assumptions and $\alpha^{-1} \leq \mathrm{e}^{2L}$,

$$\frac{\|D(t)\|_{\mathrm{op}}}{\kappa_*} \leq \chi(\mathrm{e}^{h(t)} - 1) + \varepsilon(\mathrm{e}^{2L} - 1) \leq C_2\varepsilon,$$

where C_2 depends only on B, L, χ. Taking empirical norms in equation (38) therefore gives

$$\left\|\boldsymbol{e}_{\mathrm{qs}}(t) - \alpha(t)^{-1}\boldsymbol{e}_{\mathrm{qs}}(t_0)\right\|_n \leq C\varepsilon\alpha(t)^{-1}\left\|\boldsymbol{e}_{\mathrm{qs}}(t_0)\right\|_n.$$

For sufficiently small ε, this also gives the uniform comparison

$$\tfrac{1}{2}\alpha(t)^{-1}\left\|\boldsymbol{e}_{\mathrm{qs}}(t_0)\right\|_n \leq \left\|\boldsymbol{e}_{\mathrm{qs}}(t)\right\|_n \leq 2\alpha(t)^{-1}\left\|\boldsymbol{e}_{\mathrm{qs}}(t_0)\right\|_n. \tag{39}$$

Residual tracking. Set $\underline{\kappa} = \mathrm{e}^{-2L-4BL\varepsilon}\kappa_*$. By equation (37), throughout the interval,

$$\lambda_{\min}(K_t) \geq \underline{\kappa}, \qquad \frac{\|K_t\|_{\mathrm{op}}}{\lambda_{\min}(K_t)} \leq \mathrm{e}^{8BL\varepsilon}\chi, \qquad \frac{\lambda}{\underline{\kappa}} \leq C\varepsilon.$$

From equation (36), $\left\|\dot{K}_t\right\|_{\mathrm{op}} \leq (2 + 4B\varepsilon)\lambda\|K_t\|_{\mathrm{op}}$. Differentiating $(K_t + \lambda I_n)\boldsymbol{e}_{\mathrm{qs}}(t) = -\lambda\boldsymbol{y}$ therefore yields

$$\dot{\boldsymbol{e}}_{\mathrm{qs}} = -(K_t + \lambda I_n)^{-1}\dot{K}_t\boldsymbol{e}_{\mathrm{qs}}, \qquad \|\dot{\boldsymbol{e}}_{\mathrm{qs}}\|_n \leq \frac{\left\|\dot{K}_t\right\|_{\mathrm{op}}}{\lambda_{\min}(K_t) + \lambda}\|\boldsymbol{e}_{\mathrm{qs}}\|_n \leq C\lambda\|\boldsymbol{e}_{\mathrm{qs}}\|_n.$$

Let $\boldsymbol{z} = \boldsymbol{e} - \boldsymbol{e}_{\mathrm{qs}}$. Subtracting the moving equilibrium from equation (6) gives

$$\dot{\boldsymbol{z}} = -2(K_t + \lambda I_n)\boldsymbol{z} - \dot{\boldsymbol{e}}_{\mathrm{qs}}. \tag{40}$$

The homogeneous evolution operator satisfies $\|U(t,s)\|_{\mathrm{op}} \leq \mathrm{e}^{-2(\underline{\kappa}+\lambda)(t-s)}$. Variation of constants therefore yields

$$\|\boldsymbol{z}(t)\|_n \leq \mathrm{e}^{-2(\underline{\kappa}+\lambda)\tau}\|\boldsymbol{z}(t_0)\|_n + \int_{t_0}^{t}\mathrm{e}^{-2(\underline{\kappa}+\lambda)(t-s)}\|\dot{\boldsymbol{e}}_{\mathrm{qs}}(s)\|_n\,ds. \tag{41}$$

Together with equation (39), this gives

$$\begin{aligned}\|\boldsymbol{z}(t)\|_n &\leq \mathrm{e}^{-2(\underline{\kappa}+\lambda)\tau}\|\boldsymbol{z}(t_0)\|_n + C\lambda\|\boldsymbol{e}_{\mathrm{qs}}(t_0)\|_n\int_{t_0}^{t}\mathrm{e}^{-2(\underline{\kappa}+\lambda)(t-s)}\mathrm{e}^{2\lambda(s-t_0)}\,ds \\ &\leq \mathrm{e}^{-2(\underline{\kappa}+\lambda)\tau}\|\boldsymbol{z}(t_0)\|_n + \frac{C\lambda}{\underline{\kappa}+2\lambda}\alpha(t)^{-1}\|\boldsymbol{e}_{\mathrm{qs}}(t_0)\|_n.\end{aligned}$$

Dividing by the lower bound in equation (39) gives

$$\frac{\|\boldsymbol{z}(t)\|_n}{\|\boldsymbol{e}_{\mathrm{qs}}(t)\|_n} \leq 2\mathrm{e}^{-2(\underline{\kappa}+2\lambda)\tau}\delta_0 + \frac{C\lambda}{\underline{\kappa}+2\lambda} \leq C(\delta_0 + \varepsilon), \tag{42}$$

which proves equation (15) uniformly on $[t_0, t_1]$. Thus the initial mismatch relaxes on the kernel time scale, while motion of the reference on the decay time scale produces only a small lag.

Amplification of the actual field. Using the exact decomposition

$$\boldsymbol{e}(t) - \alpha(t)^{-1}\boldsymbol{e}(t_0) = \boldsymbol{e}_{\mathrm{qs}}(t) - \alpha(t)^{-1}\boldsymbol{e}_{\mathrm{qs}}(t_0) + \boldsymbol{z}(t) - \alpha(t)^{-1}\boldsymbol{z}(t_0),$$

the quasi-static scaling estimate and the preceding bound on $\boldsymbol{z}$ give

$$\left\|\boldsymbol{e}(t) - \alpha(t)^{-1}\boldsymbol{e}(t_0)\right\|_n \leq C(\varepsilon + \delta_0)\alpha(t)^{-1}\left\|\boldsymbol{e}_{\mathrm{qs}}(t_0)\right\|_n.$$

Cauchy–Schwarz and the definition of C_ϕ now prove equation (16) for $q_t = q_{e(t)}$. The term δ_0 accounts for the initial tracking error; Theorem 2 gives $\delta_0 = o_p(1)$ for $\boldsymbol{y} \neq \boldsymbol{0}$. □

D SLOW DRIFT OF EXTREMA AND NEURON CONDENSATION

D.1 AMPLITUDE BUDGET AND ANGULAR ENERGY

Under the hypotheses of Theorem 5, $\mathcal{N}(t)$ is constant for $t \geq T$; write $\mathcal{N}$ for this set. Exact balance and energy dissipation give

$$\lambda \mathcal{A}(t) = \frac{\lambda}{2} \sum_{k \in \mathcal{N}} (a_k(t)^2 + r_k(t)^2) \leq \Phi(\Theta(t)) \leq \Phi(\Theta(T)), \qquad t \geq T. \tag{43}$$

Hence $\mathcal{A}(t) \leq \overline{\mathcal{A}}_T$.

D.2 PROOF OF THEOREM 5

Proof. Fix any finite $t > T$. On $[T, t]$, write $\mathcal{N} = \mathcal{N}(t)$, which is constant. Under the theorem's balance assumption,

$$\frac{\mathrm{d}\boldsymbol{e}}{\mathrm{d}t} = \boldsymbol{v}_A + \boldsymbol{v}_\theta, \qquad \boldsymbol{v}_A = \sum_{k \in \mathcal{N}} s_k \frac{\mathrm{d}A_k}{\mathrm{d}t} \boldsymbol{\phi}_k, \qquad \boldsymbol{v}_\theta = -2 \sum_{k \in \mathcal{N}} A_k p_k \boldsymbol{\psi}_k.$$

Here $\boldsymbol{v}_A, \boldsymbol{v}_\theta \in \mathbb{R}^n$ are output-velocity vectors, whereas $\mathcal{V}_A$ is the scalar sum of absolute amplitude derivatives. To see the first identity, differentiate the entire output sum:

$$\frac{\mathrm{d}\boldsymbol{e}}{\mathrm{d}t} = \sum_{k \in \mathcal{N}} s_k \left(\frac{\mathrm{d}A_k}{\mathrm{d}t} \boldsymbol{\phi}_k + A_k \frac{\mathrm{d}\theta_k}{\mathrm{d}t} \boldsymbol{\psi}_k\right).$$

Strictly dead neurons have zero output throughout this interval. Substituting $\frac{\mathrm{d}\theta_k}{\mathrm{d}t} = -2 s_k p_k$ cancels the two factors of s_k and gives $\boldsymbol{v}_\theta$. Within a fixed cell, $\partial_\theta \boldsymbol{\psi}_\theta(\boldsymbol{X}) = -\boldsymbol{\phi}_\theta(\boldsymbol{X})$, so the full chain rule is

$$\begin{aligned} \frac{\mathrm{d}p_k}{\mathrm{d}t} &= \left\langle \frac{\mathrm{d}\boldsymbol{e}}{\mathrm{d}t}, \boldsymbol{\psi}_k \right\rangle_n + \left\langle \boldsymbol{e}, \partial_\theta \boldsymbol{\psi}_{\theta_k}(\boldsymbol{X}) \right\rangle_n \frac{\mathrm{d}\theta_k}{\mathrm{d}t} \\ &= \left\langle \frac{\mathrm{d}\boldsymbol{e}}{\mathrm{d}t}, \boldsymbol{\psi}_k \right\rangle_n - q_k \frac{\mathrm{d}\theta_k}{\mathrm{d}t} = \left\langle \frac{\mathrm{d}\boldsymbol{e}}{\mathrm{d}t}, \boldsymbol{\psi}_k \right\rangle_n + 2 s_k q_k p_k. \end{aligned} \tag{44}$$

In the derivative of P, the amplitude term $-4 s_k A_k q_k p_k^2$ therefore cancels the angular chain-rule term $+4 s_k A_k q_k p_k^2$. Summing gives

$$\begin{aligned} \frac{\mathrm{d}P}{\mathrm{d}t} &= \sum_{k \in \mathcal{N}} \left(\frac{\mathrm{d}A_k}{\mathrm{d}t} p_k^2 + 2 A_k p_k \frac{\mathrm{d}p_k}{\mathrm{d}t}\right) \\ &= -2\lambda P + 2 \left\langle \frac{\mathrm{d}\boldsymbol{e}}{\mathrm{d}t}, \sum_{k \in \mathcal{N}} A_k p_k \boldsymbol{\psi}_k \right\rangle_n \\ &= -2\lambda P - \|\boldsymbol{v}_\theta\|_n^2 - \langle \boldsymbol{v}_A, \boldsymbol{v}_\theta \rangle_n \\ &= -2\lambda P - \|\boldsymbol{v}_\theta + \boldsymbol{v}_A/2\|_n^2 + \tfrac{1}{4} \|\boldsymbol{v}_A\|_n^2 \leq -2\lambda P + \tfrac{1}{4} C_\phi^2 \mathcal{V}_A^2. \end{aligned} \tag{45}$$

Here the triangle inequality gives $\|\boldsymbol{v}_A\|_n \leq \sum_k \left|\frac{\mathrm{d}A_k}{\mathrm{d}t}\right| \|\boldsymbol{\phi}_k\|_n \leq C_\phi \mathcal{V}_A$. Multiplying the differential inequality by $\mathrm{e}^{2\lambda(t-T)}$ and integrating yields

$$P(t) \leq \mathrm{e}^{-2\lambda(t-T)} P(T) + \frac{C_\phi^2}{4} \int_T^t \mathrm{e}^{-2\lambda(t-s)} \mathcal{V}_A(s)^2 \, ds,$$

which is equation (18). Weighted Cauchy–Schwarz gives the remaining velocity estimate explicitly:

$$\begin{aligned} \|\boldsymbol{v}_\theta\|_n &\leq 2 \sum_k A_k |p_k| \, \|\boldsymbol{\psi}_k\|_n \leq 2 C_\psi \sum_k \sqrt{A_k} \sqrt{A_k p_k^2} \\ &\leq 2 C_\psi \left(\sum_k A_k\right)^{1/2} \left(\sum_k A_k p_k^2\right)^{1/2} = 2 C_\psi \sqrt{\mathcal{A} P}. \end{aligned} \tag{46}$$

Combining $\left\|\frac{\mathrm{d}\boldsymbol{e}}{\mathrm{d}t}\right\|_n \le \|\boldsymbol{v}_A\|_n + \|\boldsymbol{v}_\theta\|_n$ with equation (43) proves equation (19).

At a smooth minimum, $q_t''(\theta_*) = R > 0$ makes its branch locally differentiable by the implicit function theorem. Differentiating $p_t(\theta_*(t)) = 0$ yields

$$0 = \left\langle \frac{\mathrm{d}\boldsymbol{e}}{\mathrm{d}t}, \boldsymbol{\psi}_{\theta_*}(\boldsymbol{X}) \right\rangle_n + q_t''(\theta_*)\frac{\mathrm{d}\theta_*}{\mathrm{d}t}.$$

Solving for its velocity and applying Cauchy–Schwarz gives

$$\frac{\mathrm{d}\theta_*}{\mathrm{d}t} = -\frac{\left\langle \frac{\mathrm{d}\boldsymbol{e}}{\mathrm{d}t}, \boldsymbol{\psi}_{\theta_*}(\boldsymbol{X}) \right\rangle_n}{R}, \qquad \left|\frac{\mathrm{d}\theta_*}{\mathrm{d}t}\right| \le \frac{C_\psi}{R_0}\left\|\frac{\mathrm{d}\boldsymbol{e}}{\mathrm{d}t}\right\|_n,$$

Combining this bound with equation (19) proves equation (20). □

D.3 Consequences of small amplitude variation

If $\mathcal{V}_A(t) \le \nu_A$ for all $t \ge T$, integrating the exponential kernel in equation (18) gives

$$P(t) \le \mathrm{e}^{-2\lambda(t-T)}P(T) + \frac{C_\phi^2\nu_A^2}{8\lambda}(1 - \mathrm{e}^{-2\lambda(t-T)}) =: H_{\nu_A}(t)^2. \tag{47}$$

Combining this with equation (19) and equation (20) gives

$$\left|\frac{\mathrm{d}\theta_*(t)}{\mathrm{d}t}\right| \le \frac{C_\psi}{R_0}\left[C_\phi\nu_A + 2C_\psi\sqrt{\overline{\mathcal{A}}_T}H_{\nu_A}(t)\right]. \tag{48}$$

Under the hypotheses of Theorem 5, if $\mathcal{V}_A(t) \to 0$, its energy bound tends to zero. Indeed, for any $\eta > 0$, choose $S \ge T$ so that $\mathcal{V}_A(s)^2 \le \eta$ for $s \ge S$. For $t \ge S$ the convolution is bounded by

$$\mathrm{e}^{-2\lambda(t-S)}\int_T^S \mathrm{e}^{-2\lambda(S-s)}\mathcal{V}_A(s)^2\,ds + \frac{\eta}{2\lambda}.$$

Letting $t \to \infty$ and then $\eta \downarrow 0$ proves $P(t) \to 0$. The residual and minimum velocities then tend to zero by equation (19) and equation (20). A uniform bound $\mathcal{V}_A \le \nu_A$ alone instead gives $\limsup_{t\to\infty} P(t) \le C_\phi^2\nu_A^2/(8\lambda)$.

D.4 Angular tracking

Lemma 7 (Angular tracking)**.** *Consider an exactly balanced positive-output neuron $a_k = r_k > 0$ for all $t \ge T$. Let $\theta_*(t)$ be a differentiable smooth negative-minimum branch with depth $R(t) \ge R_0 > 0$ for all $t \ge T$. Suppose the closed band $[\theta_*(t) - d, \theta_*(t) + d]$, with $0 < d < \pi/2$, lies inside one fixed open activation cell for all $t \ge T$. If*

$$|\delta_k(T)| \le d, \qquad \sup_{t\ge T}\left|\frac{\mathrm{d}\theta_*(t)}{\mathrm{d}t}\right| \le \omega < 2R_0\sin d, \qquad \delta_k(t) = \theta_k(t) - \theta_*(t),$$

then the neuron remains in this band. With $c_d = 2R_0\cos d > 0$,

$$|\delta_k(t)| \le \mathrm{e}^{-c_d(t-T)}|\delta_k(T)| + \int_T^t \mathrm{e}^{-c_d(t-s)}\left|\frac{\mathrm{d}\theta_*(s)}{\mathrm{d}s}\right| ds \tag{49}$$

$$\le \mathrm{e}^{-c_d(t-T)}|\delta_k(T)| + \frac{\omega}{c_d}\big(1 - \mathrm{e}^{-c_d(t-T)}\big). \tag{50}$$

Proof. The cell representation $q_t(\theta) = -R(t)\cos(\theta - \theta_*(t))$ gives $p_t(\theta) = R(t)\sin(\theta - \theta_*(t))$. For an exactly balanced positive-output neuron, differentiating $\delta_k = \theta_k - \theta_*$ therefore gives

$$\frac{\mathrm{d}\delta_k}{\mathrm{d}t} = -2R(t)\sin\delta_k - \frac{\mathrm{d}\theta_*}{\mathrm{d}t}.$$

At $\delta_k = d$ the relative velocity is at most $-2R_0\sin d + \omega < 0$; at $\delta_k = -d$ it is at least $2R_0\sin d - \omega > 0$. A first-exit argument therefore makes the band forward invariant for all $t \ge T$. For $|\delta| \le d$, $\operatorname{sgn}(\delta)\sin\delta \ge \cos d\,|\delta|$. The upper right Dini derivative therefore satisfies

$$D^+|\delta_k| \le -c_d|\delta_k| + \left|\frac{\mathrm{d}\theta_*}{\mathrm{d}t}\right|,$$

including at $\delta_k = 0$. Multiplication by $\mathrm{e}^{c_d(t-T)}$ and integration proves equation (49)–equation (50). □

D.5 PAIRWISE ANGULAR CONTRACTION

Recall the angular diameter in equation (9). The following estimate proves equation (23) once Lemma 7 supplies the invariant band.

Lemma 8 (Pairwise contraction)**.** *Let G be a fixed group of exactly balanced positive-output neurons, $a_k = r_k > 0$ for all $t \geq T$. Suppose that, for all $t \geq T$, a smooth negative-minimum branch $\theta_*(t)$ has depth $R(t) \geq R_0 > 0$ and all group directions remain in a common band*

$$|\theta_k(t) - \theta_*(t)| \leq d < \pi/2$$

contained in one open activation cell. Use a common continuous angular coordinate on this band. Then for $i, j \in G$ and $t \geq T$,

$$|\theta_i(t) - \theta_j(t)| \leq \mathrm{e}^{-2R_0 \cos(d)(t-T)} |\theta_i(T) - \theta_j(T)|. \tag{51}$$

Consequently, $\operatorname{diam}_\theta G(t)$ obeys the same contraction bound and tends to zero, regardless of whether the well converges to a fixed angle.

Taking the difference of two angular equations gives

$$\frac{d}{dt}(\theta_i - \theta_j) = -2R\big(\sin\delta_i - \sin\delta_j\big). \tag{52}$$

Proof. Set $\Delta_{ij} = \theta_i - \theta_j = \delta_i - \delta_j$. The divided difference

$$c_{ij}(t) = \int_0^1 \cos\big(\delta_j(t) + s(\delta_i(t) - \delta_j(t))\big)\, ds$$

is well defined even when the two angles coincide. Since both offsets lie in $[-d, d]$, it satisfies $c_{ij} \geq \cos d > 0$. Equation (52) is therefore $\frac{\mathrm{d}\Delta_{ij}}{\mathrm{d}t} = -2R(t)c_{ij}(t)\Delta_{ij}$. Integration proves equation (51); taking the maximum over the finitely many pairs gives the diameter bound.

□

D.6 PROOF OF THEOREM 6

Proof. Set

$$\omega := \sup_{t \geq T} \left| \frac{\mathrm{d}\theta_*(t)}{\mathrm{d}t} \right| < 2R_0 \sin d.$$

Lemma 7, applied to every $k \in G$, shows that all group directions remain in the stated band. Lemma 8 then gives equation (23). The positive rate $c_d = 2R_0 \cos d$ therefore implies that the diameter tends to zero as $t \to \infty$.

If additionally $\frac{\mathrm{d}\theta_*}{\mathrm{d}t}(t) \to 0$, the tracking estimate equation (49) implies $|\theta_k(t) - \theta_*(t)| \to 0$ for each $k \in G$. To see this, split its convolution at a fixed late time S. The part over $[T, S]$ decays exponentially, whereas the part over $[S, t]$ is at most $c_d^{-1} \sup_{s \geq S} |\frac{\mathrm{d}\theta_*}{\mathrm{d}t}(s)|$. Letting $t \to \infty$ and then $S \to \infty$ proves the claim. For negative neurons near a positive maximum, $q_t = R\cos\delta$ gives $\frac{\mathrm{d}\theta_k}{\mathrm{d}t} = 2p_k = -2R\sin\delta_k$, so the same argument applies. □

E EXPERIMENTAL DETAILS FOR THE MAIN CONDENSATION EXPERIMENT

E.1 TARGET AND TRAINING DATA

The four-point experiment in Figures 2, 3, and 5 fits the target

$$f_\star(x) = 0.2\,\mathrm{ReLU}(x - 1/3) + 0.2\,\mathrm{ReLU}(-x - 1/3), \qquad x \in [-1, 1], \tag{53}$$

using noiseless samples

$$\boldsymbol{X} = (-1, -1/3, 1/3, 1)^\top, \qquad \boldsymbol{y} = (2/15, 0, 0, 2/15)^\top.$$

Test MSE is evaluated on 10,000 equally spaced points in $[-1, 1]$.

E.2 NETWORK AND OPTIMIZATION

The network in equation (1) has $m = 200$ hidden neurons. All parameters are initialized independently as $a_k, w_k, b_k \sim \mathcal{N}(0, m^{-2\gamma})$ with $\gamma = 0.4$. Training uses 80,000 full-batch SGD updates, learning rate $\eta = 0.1$, and weight decay $\lambda = 10^{-3}$:

$$\Theta_{E+1} = (1 - \eta\lambda)\Theta_E - \eta\nabla_\Theta \mathcal{L}(\Theta_E).$$

Figure 5 shows the evolution of the NTK eigenvalues throughout training. The large initialization kernel estimate underlying Theorem 2, established in Appendix B.1, gives $K_0/m^{1-2\gamma} \to \bar{K} \succ 0$ in probability. Thus, for $\gamma < 1/2$ and sufficiently large width, all initial NTK eigenvalues are much larger than the fixed weight decay coefficient λ.

Following rapid fitting, their early decline is consistent with Theorem 4, which predicts $K_t \approx \mathrm{e}^{-2\lambda(t-T_1)} K_{T_1}$ during early amplitude compression under its stated assumptions. The eigenvalues subsequently stabilize in Stage III and remain above λ throughout the observed trajectory.

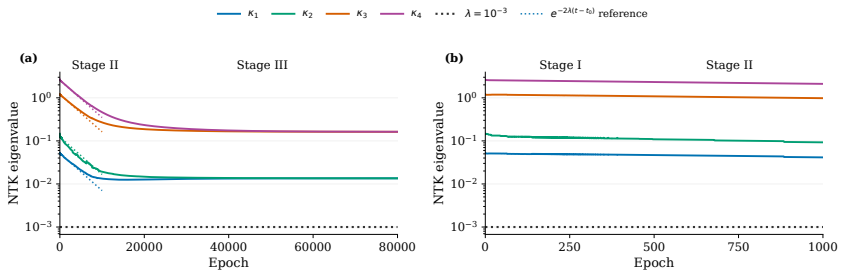

Figure 5: NTK spectrum under large initialization. (a) Full training trajectory: all four eigenvalues remain above $\lambda = 10^{-3}$, decrease during Stage II, and approach plateaus during Stage III. (b) Zoom into the first 1,000 epochs, showing the comparatively small changes in the eigenvalues during initial fitting. The horizontal dotted line indicates $\lambda = 10^{-3}$.

F Additional ReLU condensation experiments

F.1 Initialization scale and weight decay

Figure 6 compares large initialization ($\gamma = 0.4$, $\lambda = 0$ or 10^{-3}) with small initialization ($\gamma = 1$, $\lambda = 0$) on the four-point target. Both large initialization with weight decay and small initialization without decay learn condensed representations with two dominant angular groups. Their fitting dynamics differ: large initialization reaches a low training loss within a few hundred updates, well before late condensation, whereas small initialization fits more slowly.

In the large initialization run with $\lambda = 10^{-3}$, all non-dead neurons become condensed, including a low-amplitude group near $90°$. Dead neurons remain dispersed while their amplitudes vanish. With small initialization and no decay, some low-amplitude non-dead neurons remain dispersed despite the two dominant groups. Large initialization without decay fits rapidly,but neurons don't develop a condensed representation.

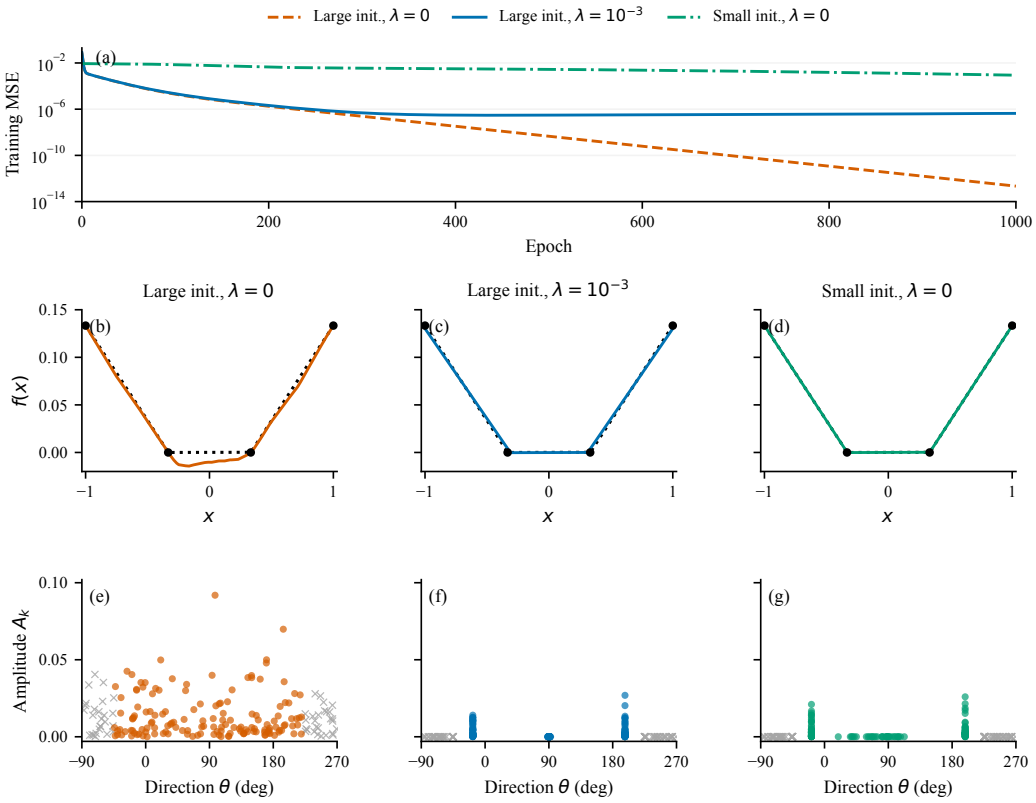

Figure 6: Fitting speed and final condensation at different initialization scales. (a) Training MSE during the first 1,000 updates. (b–d) Outputs at epoch 80,000, with targets (dotted) and training samples (black dots). (e–g) Final directions and unsigned amplitudes $A_k = |a_k| r_k$. Columns compare large initialization with $\lambda = 0$ and 10^{-3}, and small initialization with $\lambda = 0$. Colored circles show non-dead neurons; gray crosses show neurons inactive on every training input.

F.2 TARGETS WITH THREE RELU BREAKPOINTS

Figure 7 compares a convex target (A) with targets whose peak and trough occur in opposite orders (B, C):

$$
\begin{aligned}
f_{\mathrm{A}}(x) &= 0.06\,\mathrm{ReLU}(x+3/4) + 0.09\,\mathrm{ReLU}(-x) + 0.12\,\mathrm{ReLU}(x-3/4),\\
f_{\mathrm{B}}(x) &= 0.3\,\mathrm{ReLU}(x+3/4) - 0.5\,\mathrm{ReLU}(x) + 0.4\,\mathrm{ReLU}(x-3/4),\\
f_{\mathrm{C}}(x) &= 0.2\,\mathrm{ReLU}(x+3/4) + 0.4\,\mathrm{ReLU}(-x) - 0.4\,\mathrm{ReLU}(x-3/4),
\end{aligned}
\tag{54}
$$

where $x \in [-1, 1]$. After rapid fitting, the outputs already track the targets while neuron directions remain diffuse. Later, neurons with appreciable amplitudes form narrow angular groups in all three cases. The final outputs show a regularization bias.

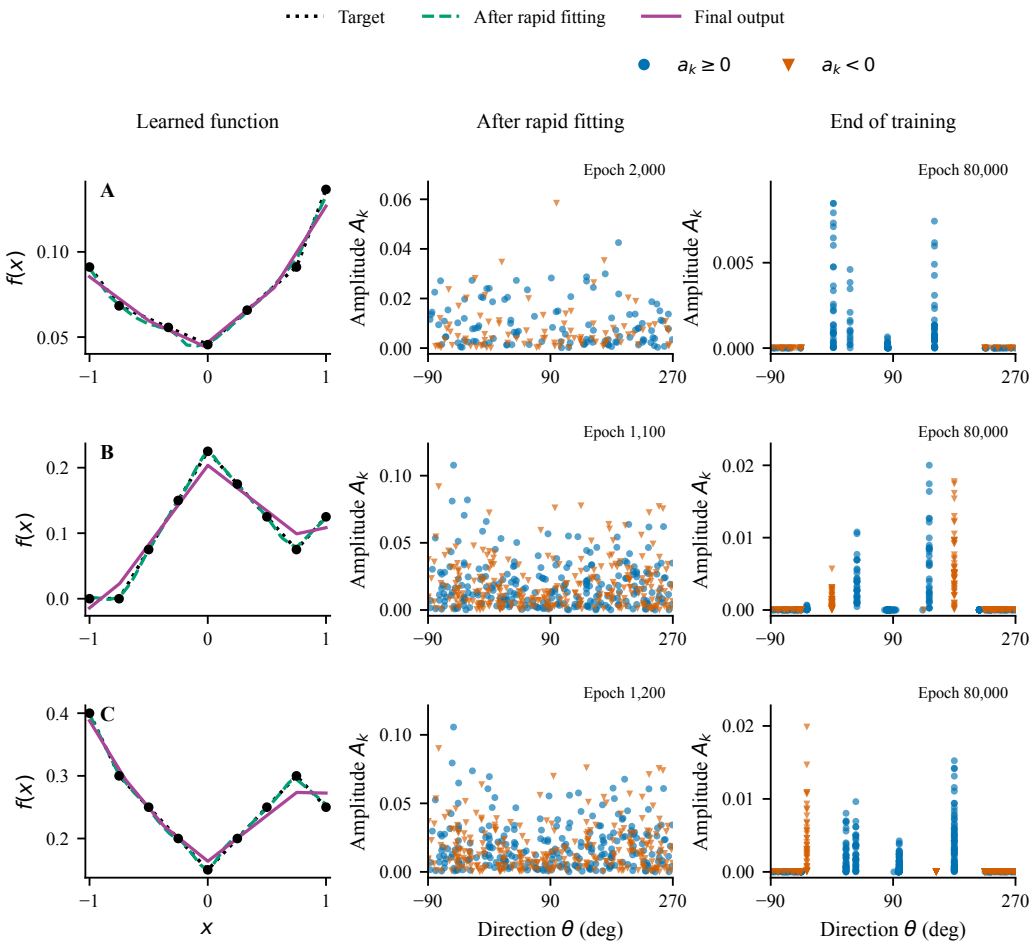

Figure 7: Fitting precedes condensation for the targets in equation (54) (rows A–C). Left: target, training samples, early output, and final output. Middle/right: neurons after rapid fitting/at epoch 80,000, with identical marker sizes and separately scaled amplitude axes. Circles and triangles denote nonnegative and negative output weights.

G TRANSFORMER EXPERIMENT ON ANCHOR FUNCTION COMPOSITION

This section specifies the experiment in Figure 1. We use the anchor function framework of Zhang et al. (2024), with the data split and training settings described below. The comparison consists of two runs with weight decay coefficients 0 and 0.8, identical saved data, and identical initial parameters throughout the network.

G.1 ANCHOR FUNCTIONS AND DATA CONSTRUCTION

Each input is a sequence of nine integer tokens containing exactly one consecutive anchor pair $(a, b) \in \{1, 2, 3, 4\}^2$. The token immediately preceding this pair is the key x. The four elementary functions are

$$g_1(x) = x + 5, \qquad g_2(x) = x + 1, \qquad g_3(x) = x - 2, \qquad g_4(x) = x - 8.$$

The target is the integer $y = g_a(g_b(x)) = x + d_a + d_b$, where $(d_1, d_2, d_3, d_4) = (5, 1, -2, -8)$. These translations commute, so their order does not affect the target. All pairs use this compositional rule, including $(3, 4)$ and $(4, 3)$, whose target is $x - 10$.

Using zero-based positions, we sample the key position $p \in \{0, \ldots, 6\}$ uniformly and place (x, a, b) at positions $(p, p + 1, p + 2)$. The other six input tokens are sampled independently and uniformly from $\{20, \ldots, 99\}$ and do not affect the target. The key is drawn uniformly from the same integer range, subject to

$$x \bmod 9 \neq p \quad \text{for training}, \qquad x \bmod 9 = p \quad \text{for testing}.$$

The modulus is the input length, 9, as implemented in this experiment. Training contains the 14 anchor pairs other than $(3, 4)$ and $(4, 3)$; both excluded pairs are held out entirely. Testing contains all 16 pairs. Thus the test groups for the 14 seen pairs assess generalization to held-out key–position combinations, while the other two groups also assess unseen anchor compositions. No non-compositional target is introduced in this experiment.

G.2 NETWORK AND INITIALIZATION

The model is a decoder-only Transformer with two layers and one causal self-attention head per layer. Given an input sequence represented by one-hot vectors X^{in}, the input to the Transformer block is

$$H^{(0)} = X^{\text{in}} W^{\text{em}} + X^{\text{pos}}, \tag{55}$$

where X^{pos} is a learned positional embedding. For layer l, we compute:

$$\begin{aligned} \widetilde{H}^{(l)} &= \mathrm{LN}\left(H^{(l-1)}\right), \\ Q^{(l)} &= \widetilde{H}^{(l)} W_Q^{(l)}, \\ K^{(l)} &= \widetilde{H}^{(l)} W_K^{(l)}, \\ V^{(l)} &= \widetilde{H}^{(l)} W_V^{(l)}. \end{aligned} \tag{56}$$

The causal attention matrix and the output of the attention layer are

$$\begin{aligned} Attn^{(l)} &= \mathrm{softmax}\left(\frac{Q^{(l)} K^{(l)\top}}{\sqrt{d_k}}\right) \text{(with mask)}, \\ U^{(l)} &= H^{(l-1)} + Attn^{(l)} V^{(l)} W_O^{(l)}, \end{aligned} \tag{57}$$

The feedforward layer is computed as:

$$\begin{aligned} \widetilde{U}^{(l)} &= \mathrm{LN}^{(l)}\left(U^{(l)}\right), \\ H^{(l)} &= \widetilde{U}^{(l)} + \mathrm{ReLU}\left(\widetilde{U}^{(l)} W_1^{(l)}\right) W_2^{(l)}. \end{aligned} \tag{58}$$

The final logits are

$$Z = H^{(N)} W^{\text{out}} + b^{\text{out}}. \tag{59}$$

LN refers to Layer Normalization.

Each linear weight matrix $W \in \mathbb{R}^{d_{\text{in}} \times d_{\text{out}}}$ is initialized entrywise as

$$W_{ij} \sim \mathcal{N}\left(0, d_{\text{in}}^{-2\gamma}\right) \tag{60}$$

Table 1: Settings shared by the two Transformer runs in Figure 1; only weight decay differs.

Setting	Value
Layers / heads per layer	2 / 1
Embedding / feedforward dimension	400 / 1,200
Query, key, and value dimension	64
Input length / vocabulary size	9 / 120
Initialization rate	0.48
Optimizer	AdamW
Batch size / training epochs	4,096 / 400
Learning rate: initial / peak / minimum	5×10^{-5} / 5×10^{-4} / 5×10^{-5}
Warmup / cosine schedule	40 / 360 epochs
Global gradient-norm clipping	1.0

G.3 Training objective and weight decay comparison

Training minimizes cross entropy between the logits at the last output position and the appended target y. AdamW applies decoupled weight decay to *all* trainable parameters, including embeddings, LayerNorm parameters, and the output bias; there are no parameter groups exempted from decay. The learning rate warms up from 5×10^{-5} to 5×10^{-4} over 40 epochs, then follows a cosine schedule with period parameter 360 epochs and minimum 5×10^{-5}.

G.4 Cosine similarities

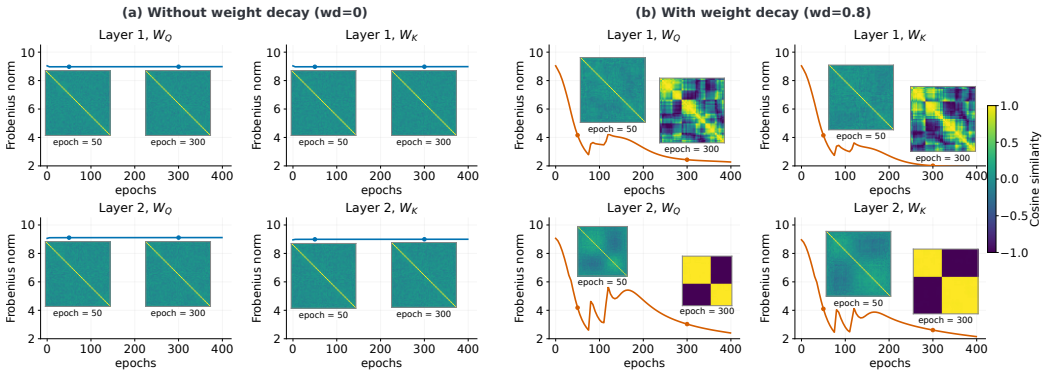

Figure 8: Weight contraction and condensation across Transformer layers. The top and bottom rows show layers 1 and 2, respectively. Panel (a) corresponds to training without weight decay ($\lambda = 0$), and Panel (b) corresponds to training with weight decay ($\lambda = 0.8$). Within each panel, the columns show W_Q and W_K. Curves show Frobenius norms during training, while the insets show row-wise cosine similarities at epochs 50 and 300. Weight decay induces norm contraction and structured directional alignment in both layers.

In our experiment, $W_Q, W_K \in \mathbb{R}^{64 \times 400}$. For rows $\boldsymbol{w}_i, \boldsymbol{w}_j \in \mathbb{R}^{400}$, define cosine similarity

$$C_{ij} = \frac{\boldsymbol{w}_i^\top \boldsymbol{w}_j}{\|\boldsymbol{w}_i\|_2 \, \|\boldsymbol{w}_j\|_2}.$$

Rows and columns are reordered by the same permutation and grouped according to cosine similarity. Figure 1 uses the query and key projection matrices from the first Transformer blocks. Figure 8 shows result on both Transformer layers.