Neuron Block Dynamics for XOR Classification with Zero-Margin

Guillaume Braun
RIKEN AIP

Masaaki Imaizumi
RIKEN AIP, The University of Tokyo, Kyoto University

Abstract

The ability of neural networks to learn useful features through stochastic gradient descent (SGD) is a cornerstone of their success. Most theoretical analyses focus on regression or on classification tasks with a positive margin, where worst-case gradient bounds suffice. In contrast, we study zero-margin nonlinear classification by analyzing the Gaussian XOR problem, where inputs are Gaussian and the XOR decision boundary determines labels. In this setting, a non-negligible fraction of data lies arbitrarily close to the boundary, breaking standard margin-based arguments. Building on Glasgow's (2024) analysis, we extend the study of training dynamics from discrete to Gaussian inputs and develop a framework for the dynamics of neuron blocks. We show that neurons cluster into four directions and that block-level signals evolve coherently, a phenomenon essential in the Gaussian setting where individual neuron signals vary significantly. Leveraging this block perspective, we analyze generalization without relying on margin assumptions, adopting an average-case view that distinguishes regions of reliable prediction from regions of persistent error. Numerical experiments confirm the predicted two-phase block dynamics and demonstrate their robustness beyond the Gaussian setting.

Proceedings of the 29th International Conference on Artificial Intelligence and Statistics (AISTATS) 2026, Tangier, Morocco. PMLR: Volume 300.

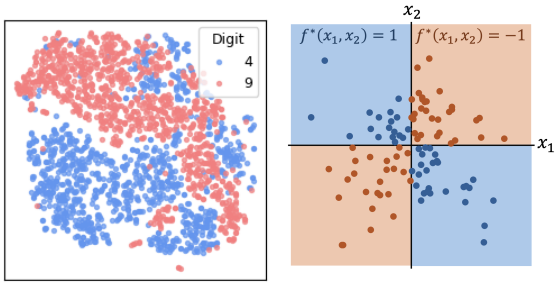
1 INTRODUCTION

In deep learning theory, feature learning has emerged as a central theme. It refers to the ability of neural networks to uncover latent low-dimensional structure from high-dimensional data, often yielding significantly better sample complexity than kernel methods. A line of recent work demonstrates this advantage in regression under the multi-index model, where labels depend on a nonlinear function of a low-dimensional projection of the inputs (Abbe et al., 2023; Bietti et al., 2022; Mousavi-Hosseini et al., 2023; Bietti et al., 2025; Barak et al., 2022; Oko et al., 2024; Dandi et al., 2024). We refer to Bruna and Hsu (2025) for a recent survey.

While most work on feature learning has centered on regression, there has also been progress in *classification*. Existing analyses fall into two main strands. The first is the literature on *benign overfitting* (Frei et al., 2022; Xu et al., 2024), which shows that overparameterized neural networks trained by stochastic gradient descent (SGD) can generalize even when they interpolate the data. The second strand investigates *sparse parity learning* with Boolean inputs (Glasgow, 2024; Abbe et al., 2025), a discrete setting where classes enjoy a positive separation margin. Both lines of work highlight that neural networks achieve feature learning in classification under favorable structural conditions such as linear separability or positive margin.

A major open challenge is to understand neural networks in *nonlinear classification without positive margins*. Such situations naturally occur in real data—for instance, Figure 1(a) illustrates that digits 4 and 9 in MNIST (Deng, 2012) overlap heavily, leaving no clear separation between classes. Yet despite its prevalence, the zero-margin regime remains poorly understood, since the lack of separation hinders alignment with discriminative features.

To investigate this setting in a controlled way, we study the *canonical XOR classification problem with*



(a) MNIST: 4 vs 9. (b) Our Setup.

Figure 1: (a) t-SNE plot of MNIST digits 4 and 9, which overlap substantially. (b) Gaussian XOR setup: $f^*(x) = -\mathrm{sgn}(x_1 x_2)$ with zero margin, since the Gaussian distribution crosses the coordinate axes.

Gaussian inputs, where labels are generated by

$$f^*(x) = -\mathrm{sgn}(x_1 x_2), \qquad x = (x_1, x_2, \dots) \in \mathbb{R}^d,$$

see Figure 1(b). This minimal toy problem captures the essential difficulties: the decision boundary is nonlinear, feature learning is required, and a non-negligible fraction of samples lie arbitrarily close to the boundary. Unlike the Boolean case, where inputs are discrete and bounded, Gaussian inputs are continuous and unbounded, so data both concentrate near the boundary and can take arbitrarily large values. Prior work by Glasgow (2024) showed that vanilla SGD can solve the Boolean XOR problem, but their analysis relies critically on the presence of a positive margin. In the Gaussian case these arguments break down, motivating our *block-dynamic viewpoint*, which departs from per-neuron margin arguments and explains how neural networks can still learn in the zero-margin regime.

Building on the work of Glasgow (2024), we develop a *block-dynamic framework* for analyzing SGD in the *zero-margin* regime, and prove that a two-layer neural network trained by vanilla SGD can learn the XOR function from Gaussian inputs. Our main contributions are:

- **Block dynamics.** We show that neurons rapidly self-organize into four coherent blocks aligned with the signal directions. A mean-field–style argument establishes that block masses remain comparable, enabling a low-dimensional block-level description of the network.
- **Average-case analysis.** In the absence of a margin, many samples lie arbitrarily close to the decision boundary, invalidating worst-case arguments. We introduce an *average margin* statistic and prove that it governs the accuracy of the classifier.
- **Experiments.** Simulations confirm block formation, validate the average-case dynamics, and illustrate the importance of nonlinear separation in the zero-margin setting.

1.1 Related Work

Feature learning. Two-layer neural networks have been shown to adapt to low-dimensional latent structure in regression and multi-index models Abbe et al. (2023); Bietti et al. (2022); Mousavi-Hosseini et al. (2023); Bietti et al. (2025); Barak et al. (2022); Oko et al. (2024); Dandi et al. (2024); Bruna and Hsu (2025), often achieving significantly lower sample complexity than kernel methods. In contrast, kernel perspectives such as the Neural Tangent Kernel (NTK) Jacot et al. (2018); Chizat et al. (2018) do not capture adaptation to latent structure Ghorbani et al. (2021), while mean-field analyses Mei et al. (2019) typically require very wide networks and long training horizons Suzuki et al. (2023), leading to suboptimal sample complexity Mahankali et al. (2023). Beyond these approximation regimes, most theoretical works analyze layer-wise training, where the first layer is optimized to recover signal directions before fitting the link function. However, such procedures are not faithful to practice and may even fail to generalize Bietti et al. (2025). This motivates the study of vanilla SGD with simultaneous training, as in Glasgow (2024), who analyzed XOR with Boolean inputs, and Berthier et al. (2024), who studied the single-index model in the mean-field regime.

Classification with neural networks. The dynamics of neural networks in classification have been extensively studied under linearly separable models with a positive margin (Wei et al., 2018; Brutzkus et al., 2018; Ji and Telgarsky, 2020; Lyu et al., 2021; Telgarsky, 2023), often linked to the theory of benign overfitting Frei et al. (2022); Kornowski et al. (2023); Zhu et al. (2023); Wang et al. (2024). Benign overfitting has since been established across diverse architectures, including transformers (Jiang et al., 2024), graph convolutional networks (Huang et al., 2025), and convolutional networks Kou et al. (2023); Cao et al. (2022). Beyond linear separability, Xu et al. (2024) analyzed an XOR Gaussian mixture model, but required that cluster separation scales with dimension. Other works motivated by feature

learning, such as Shi et al. (2022), investigated dictionary learning with discrete latent structure, while Shi et al. (2023) provided guarantees in more general settings without explicit margin assumptions. However, these analyses assume bounded inputs and layer-wise training, in contrast to our focus on simultaneous training with unbounded Gaussian inputs.

XOR problem and parity learning. Parity functions of the form $\prod_{i\in S} x_i$, where $x_i \in \{-1, 1\}$, represent a class of functions that are generally challenging for gradient-based algorithms to learn without specific assumptions on the input distributions Shalev-Shwartz et al. (2017). However, under additional assumptions on the input data, a two-layer neural network can effectively learn these functions, as demonstrated by Barak et al. (2022); Abbe et al. (2023); Kou et al. (2024); Glasgow (2024); Abbe et al. (2025).

1.2 Notations

We use the notation $a_n \lesssim b_n$ (or $a_n \gtrsim b_n$) for sequences $(a_n)_{n\geq 1}$ and $(b_n)_{n\geq 1}$ if there exists a constant $C > 0$ such that $a_n \leq Cb_n$ (or $a_n \geq Cb_n$) for all n. If the inequalities hold only for sufficiently large n, we write $a_n = O(b_n)$ (or $a_n = \Omega(b_n)$). We use $\|\cdot\|$ and $\langle\cdot,\cdot\rangle$ to denote the Euclidean norm and scalar product, respectively. The sign function is denoted by $\mathrm{sgn}(\cdot)$. The $(d-1)$-dimensional sphere of radius θ is denoted by $\mathbb{S}^{d-1}(\theta)$.The canonical basis of $\mathbb{R}^d$ is denoted by $e_1, \ldots, e_d$.

2 STATISTICAL FRAMEWORK

Data generation. We model each observation $(x, y) \in \mathbb{R}^d \times \{\pm 1\}$ as independently generated from the following process: the input x is drawn from an isotropic Gaussian distribution,

$$x = (x_1, x_2, ..., x_d)^\top \sim \mathcal{N}(0, I_d),$$

and the label y, given x, is determined by the XOR function f^*

$$y = f^*(x) = -\operatorname{sgn}(x_1 x_2).$$

For simplicity, we assume that the function $f^*(\cdot)$ depends only on the first two components of x, specifically x_1 and x_2, which correspond to the canonical basis vectors e_1 and e_2. This assumption can be relaxed, as the algorithm in our analysis is rotationally invariant.

Model. We utilize a two-layer neural network with the ReLU activation function, denoted by $\sigma(\cdot)$, to learn the XOR function $f^*(\cdot)$. This network consists of m neurons, each characterized by a pair of weights (w, a) where $w \in \mathbb{R}^d$ and $a \in \mathbb{R}$. Let $\mathcal{N}$ represent the set of these m neurons. The function represented by the neural network is defined as:

$$f_\rho(x) := \frac{1}{m} \sum_{(w,a)\in\mathcal{N}} a\sigma(w^\top x) = \mathbb{E}_\rho[a\sigma(w^\top x)],$$

where ρ is the empirical distribution of the neurons in $\mathcal{N}$.

Initialization. To initialize each neuron $(w, a) \in \mathcal{N}$, we sample initial weights $(w^{(0)}, a^{(0)})$ as follows. The weights of the first layer are initialized uniformly on the sphere of radius θ, which will be specified later. Specifically, we set $w^{(0)} \sim \mathrm{Unif}(\mathbb{S}^{d-1}(\theta))$ for some scale parameter $\theta > 0$. The weights of the second layer are initialized as $a^{(0)} = \theta\varepsilon$ where ε is an i.i.d. Rademacher variable, i.e. $\mathbb{P}(\varepsilon = 1) = \mathbb{P}(\varepsilon = -1) = 1/2$.

Network training. To train the network $f_\rho(\cdot)$, we employ Stochastic Gradient Descent (SGD) over T iterations and a learning rate $\eta > 0$. Each iteration t uses an independent batch of size V, denoted by $M_t = \{(t-1)V + 1, (t-1)V + 2, ..., tV\}$. The empirical risk for a batch $M \subset \mathbb{N}$ is defined as

$$\hat{L}_\rho = \frac{1}{V} \sum_{j\in M_t} \ell_\rho(x^{(j)}),$$

where $x^{(j)}$ is an independent sample from $\mathcal{N}(0, I_d)$ and $\ell_\rho(x)$ is the logistic loss:

$$\ell_\rho(x) = 2\log\left(1 + \exp(-yf_\rho(x))\right).$$

The population loss is $L_\rho = \mathbb{E}_x[\ell_\rho(x)]$, and the derivative $\ell'_\rho(x)$ of the logistic loss with respect to $f_\rho(x)$ is:

$$\ell'_\rho(x) = -\frac{2y\exp(-yf_\rho(x))}{1 + \exp(-yf_\rho(x))}.$$

The gradient updates of a neuron $(w, a) \in \mathcal{N}$ are calculated as follows[1]

$$\nabla_w \hat{L}_\rho = \frac{1}{V} \sum_{j\in M_t} \ell'_\rho(x^{(j)}) a\sigma'(w^\top x^{(j)}) x^{(j)}$$

$$\nabla_a \hat{L}_\rho = \frac{1}{V} \sum_{j\in M_t} \ell'_\rho(x^{(j)}) \sigma(w^\top x^{(j)}).$$

[1]Since the ReLU function is not differentiable at zero, we define $\sigma'(0) = 0$ for convenience.

At each step, the weights are updated as follows

$$\begin{cases} w^{(t+1)} & = w^{(t)} - \eta \nabla_{w^{(t)}} \hat{L}_{\rho^{(t)}}(w^{(t)}) \\ a^{(t+1)} & = a^{(t)} - \eta \nabla_{a^{(t)}} \hat{L}_{\rho^{(t)}}(a^{(t)}) \end{cases},$$

starting from $(w^{(0)}, a^{(0)})$. After t update steps, the set of neurons is $\mathcal{N}^{(t)} := \{(w^{(t)}, a^{(t)})\}$, and $\rho^{(t)}$ represents the empirical distribution of these neurons. For simplicity, we omit the iteration index t when the context makes it clear. Since each iteration uses an independent batch, there are no stochastic dependencies between weights and inputs.

3 MAIN RESULT

We show that a two-layer neural network, with both layers trained simultaneously by vanilla SGD, can learn the XOR function from Gaussian inputs in the zero-margin regime, where many samples lie arbitrarily close to the decision boundary.

Theorem 1. *Let $\theta = (\log d)^{-C}$ for a sufficiently large constant $C > 0$. Consider the network of Section 2, trained with mini-batches of size $V \geq d/\theta$, step size $\eta \asymp \theta$, and width polynomial in d. Then, with high probability, there exists a stopping time $T \asymp (\log d)^{C+1}$ such that the expected loss on a fresh input satisfies*

$$\mathbb{E}_x\big[\ell_{\rho^{(T)}}(x)\big] \;=\; O\Big(\tfrac{1}{\sqrt{\log\log d}}\Big).$$

Remark 1 (Sample complexity)**.** *The required sample complexity of our algorithm is $O(d\,\mathrm{polylog}(d))$, which is near-optimal up to logarithmic factors, consistent with CSQ lower bounds Abbe et al. (2023).*

The convergence rate of $\sqrt{\log\log d}^{-1}$ provided by Theorem 1, is slower than in the Boolean-input setting. The following remarks clarify the source of this slowdown and the interpretation of our bound.

Remark 2 (Comparison with Boolean XOR)**.** *In the Boolean-input setting of Glasgow (2024), the presence of a positive margin enables faster convergence. In contrast, for Gaussian inputs the absence of a margin makes the problem intrinsically harder: the expected loss is dominated by points lying arbitrarily close to the decision boundary.*

Beyond the loss guarantee of Theorem 1, our analysis also yields a classification guarantee: points that are sufficiently far from the decision boundary are classified correctly with overwhelming probability. Toward this end, let us introduce for all $\epsilon \in (0,1)$ the set

$$\mathcal{C}_\epsilon = \{(r,\theta) \in \mathbb{R}^+ \times [0, 2\pi) : r(|\sin\theta| \wedge |\cos\theta|) \leq \epsilon\}.$$

Corollary 1. *Let us fix a constant $\epsilon \in (0,1)$. Then, under the assumptions of Theorem 1, if $x = z + \xi$– where z denotes the projection of x onto the (e_1, e_2)-plane – is an input generated independently such that $z \notin \mathcal{C}_\epsilon$, there are constants $c, C > 1$ such that with probability at least $1 - \exp\big(-C(\log d)^c\big)$, the trained network $f_{\rho^{(T)}}$ will correctly classify x.*

This behavior is also clearly visible in Figure 5, where misclassifications occur only for points lying extremely close to the decision boundary.

A key novelty of our analysis is to track block-level dynamics rather than individual neurons. Neurons rapidly self-organize into four balanced blocks, whose growth is driven by an average margin statistic. These structural insights are essential for explaining how SGD succeeds in the zero-margin regime.

4 PROOF OUTLINE

Before outlining the proof strategy, we introduce the additional notation needed for the analysis.

Additional notation. For $w \in \mathbb{R}^d$, write $w = w_{1:2} + w_\perp$ with $w_{1:2} \in \mathrm{span}(e_1, e_2)$ and $w_\perp \perp \mathrm{span}(e_1, e_2)$. For $x \sim \mathcal{N}(0, I_d)$, decompose $x = z + \xi$ with $z \sim \mathcal{N}(0, I_2)$ and $\xi \sim \mathcal{N}(0, I_{d-2})$ independent. Define the XOR directions $\mu_1 = (1,-1)^\top/\sqrt{2}$ and $\mu_2 = (1,1)^\top/\sqrt{2}$. Neurons (w, a) will align with $\{\pm\mu_1, \pm\mu_2\}$ depending on the sign of a. Accordingly, we decompose $w_{1:2}$ into a *signal* component w_{sig} and an *orthogonal* component w_{opp}:

$$w_{\mathrm{sig}} = \begin{cases} \mu_1^\top w\, \mu_1 & a \geq 0, \\ \mu_2^\top w\, \mu_2 & a < 0, \end{cases} \quad w_{\mathrm{opp}} = \begin{cases} \mu_2^\top w\, \mu_2 & a \geq 0, \\ \mu_1^\top w\, \mu_1 & a < 0. \end{cases}$$

Roadmap. Our analysis proceeds in two phases, followed by a synthesis step.

Phase I (individual dynamics). When the network outputs are small, we linearize the loss, approximating $\ell_{\rho^{(t)}}(x) \approx \ell_0(x) := \log 2 - y\, f_{\rho^{(t)}}(x)$, under which neurons evolve nearly independently. In the early stage (*Phase Ia*), all neurons grow at comparable rates. Once the signal has amplified (*Phase Ib*), growth rates begin to diverge. At this point, it is more natural to group neurons into four blocks aligned with $\pm\mu_1, \pm\mu_2$, determined jointly by the sign of a and the initial correlation with these directions. As illustrated in Figure 2, neurons align with these cluster directions while approximately balanced across quadrants, a property we call *pre-*

balanced block dynamics. This aggregated description will be crucial for Phase II.

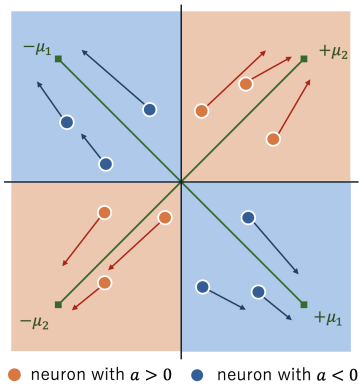


Figure 2: Neuron dynamics in Phase I. Circles show projections of weights onto $\mathrm{span}(e_1, e_2)$; colors indicate the sign of a. Neurons align with $\pm\mu_1, \pm\mu_2$ (green), and blocks remain balanced across quadrants.

Phase II (block dynamics). As outputs grow, the linearized model ceases to be accurate, and neuron interactions become significant. We analyze the network at the block level: block masses continue to grow in a balanced way. For tractability, we use an oracle approximation (replacing the full gradient $\ell'_\rho(x)$ with its two-dimensional projection $\ell'_\rho(z)$, and enforcing perfectly balanced blocks), which lets us evaluate the average margin and derive a multiplicative growth law for the aggregate block mass.

Synthesis. Finally, we leverage the characterization of the network obtained in Phase II to conclude the proof of Theorem 1. Figure 3 summarizes the main steps of the proof and their relative importance in the network dynamics.

4.1 Phase I: Early individual dynamics

In this phase, the network output is small, and the loss can be well approximated by its first-order expansion $\ell_{\rho^{(t)}}(x) \approx \ell_0(x) := \log 2 - y f_{\rho^{(t)}}(x)$. Under this approximation, neurons evolve independently, driven by the population gradient.

At initialization we have w.h.p. $\|w_\perp\| \approx \theta$, while $\|w_{\mathrm{sig}}\|, \|w_{\mathrm{opp}}\| \lesssim \theta \log d/d$, hence $\|w_{\mathrm{sig}}\| \ll \|w_\perp\|$. In this regime, the population gradient satisfies $-\nabla_{w_{1:2}} L_0 \approx \frac{\sqrt{2}|a|}{\pi^{3/2}\|w_\perp\|}(w_{\mathrm{sig}} - w_{\mathrm{opp}})$, so the signal is reinforced while the orthogonal component is damped. Controlling stochastic fluctuations by concentration, we obtain:

Lemma 1 (Phase Ia dynamics)**.** *For any neuron $(w, a) \in \mathcal{N}$ with $\|w_{sig}^{(0)}\| \gtrsim (d \log d)^{-1/2}$ and all $t \leq T_a$,*

$$\|w_{sig}^{(t+1)}\| = (1 + \eta \tfrac{\sqrt{2}}{\pi^{3/2}}(1 + o(1)))\|w_{sig}^{(t)}\|,$$

while $\|w_\perp^{(t)}\| \approx \theta$ and

$$\|w_{opp}^{(t)}\| \leq (1 + \eta\theta)\ \max\{\|w_{opp}^{(0)}\|, (d \log d)^{-1/2}\}.$$

Thus, in Phase Ia, most neurons experience multiplicative signal growth, while the orthogonal component remains bounded, ensuring alignment toward the signal directions.

4.1.1 Phase Ib: Heterogeneous growth

After sufficiently many iterations, the signal $\|w_{\mathrm{sig}}\|$ becomes comparable to $\|w_\perp\|$, so the approximation $\|w_{\mathrm{sig}}\| \ll \|w_\perp\|$ used in Phase Ia no longer holds. In this regime, we exploit instead that $\|w_{\mathrm{opp}}\| \ll \|w_{\mathrm{sig}}\|$, which yields

$$(1 + \eta c_1)\|w_{\mathrm{sig}}^{(t)}\| \ \leq\ \|w_{\mathrm{sig}}^{(t+1)}\| \ \leq\ (1 + \eta c_2)\|w_{\mathrm{sig}}^{(t)}\|,$$

for some constants $0 < c_1 < c_2$. Thus, the signal continues to grow, though at heterogeneous rates across neurons.

Lemma 2 (Signal-based individual dynamics)**.** *Let $\zeta = (\log d)^{-c}$ for some constant $1 < c < C$. Under the assumptions of Theorem 1, there exists $T_1 \asymp \log d/\eta$ such that for all $t \leq T_1$:*

1. (Weak noise) $\mathbb{E}_{\rho^{(t)}} \| w_\perp + w_{opp} \|^2 \leq 4\theta^2$.
2. (Large signal) *There exists a constant $C' > 0$ such that for all neurons with $\mu^\top w^{(0)} > \theta/\sqrt{d}$ for some $\mu \in \{\pm\mu_1, \pm\mu_2\}$,*

$$\theta\zeta^{-1} \ \leq\ \|w_{sig}^{(T_1)}\| \ \leq\ C'\theta\zeta^{-1} \log d.$$

In short, by the end of Phase I, each neuron develops a strong signal aligned with μ_1 or μ_2, while the noise components $w_\perp + w_{\mathrm{opp}}$ remain controlled.

4.1.2 Blocks are nearly balanced

When neuron growth becomes heterogeneous in Phase Ib, it is crucial to reason at the block level. We group neurons into four *blocks*, aligned with $\pm\mu_1$ and $\pm\mu_2$, and demonstrate that their masses remain nearly balanced despite stochastic fluctuations.

Definition 1 (Neuron blocks)**.** *Let $\mathcal{N}^{(0)}$ be the set of neurons at initialization. Define*

$$\mathcal{N}_1^{\pm} = \{(w, a) \in \mathcal{N}^{(0)} : a > 0,\ \pm w^\top \mu_1 > 0\},$$
$$\mathcal{N}_2^{\pm} = \{(w, a) \in \mathcal{N}^{(0)} : a < 0,\ \pm w^\top \mu_2 > 0\}.$$

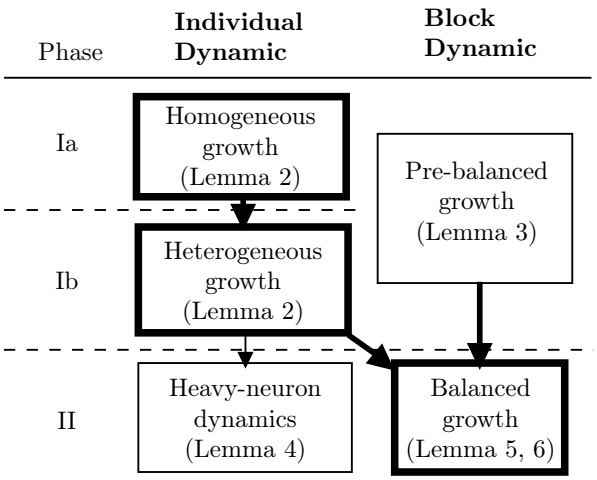


Figure 3: Overview of phases and neuron dynamics. Thicker-bordered boxes emphasize the key dynamics that capture how the network evolves in each phase.

For $i = 1, 2$, the block mass at time t is

$$N_i^{\pm,(t)} = \frac{1}{|\mathcal{N}_i^{\pm}|} \sum_{(w,a)\in\mathcal{N}_i^{\pm}} \|a^{(t)} w_{sig}^{(t)}\|,$$

and the average mass is

$$N^{(t)} = \tfrac{1}{4} \sum_{i=1,2} \left(N_i^{+,(t)} + N_i^{-,(t)}\right).$$

To quantify balance, we define the *unbalance level*

$$U^{(t)} = \max_{i,j\in\{1,2\},\pm} \left| \frac{N_i^{\pm,(t)}}{N_j^{\pm,(t)}} - 1 \right|.$$

Using the correlation-loss approximation $L_\rho \approx L_0$, we compare the true weight sequence $(w^{(t)}, a^{(t)})$ with a surrogate sequence $(\tilde{w}^{(t)}, \tilde{a}^{(t)})$ that evolves under L_0. The key property is that these surrogate updates preserve independence, and block laws differ only by a rotation. Hence, block averages remain close, up to sampling noise from SGD.

Lemma 3 (Pre-balanced blocks)**.** *Under the assumptions of Theorem 1, w.h.p. for all $t \le T_1$,*

$$U^{(t)} \;\le\; \log^{-c_U} d,$$

where $c_U > 0$ can be made arbitrarily large by increasing the batch size.

Proof idea. To control $U^{(t)}$, we introduce a *surrogate sequence* $(\tilde{w}^{(t)}, \tilde{a}^{(t)})$ initialized as $(w^{(0)}, a^{(0)})$ but updated by population gradients ∇L_0 where $L_0 = \mathbb{E}_x(\ell_0(x))$. Block averages can then be approximated via conditional expectations:

$$\frac{1}{|\mathcal{N}_i^{\pm}|} \sum_{(w,a)\in\mathcal{N}_i^{\pm}} \|\tilde{w}_{\text{sig}}^{(t)}\| \, |\tilde{a}^{(t)}| \approx \mathbf{E}\Big[\|\tilde{w}_{\text{sig}}^{(t)}\| \, |\tilde{a}^{(t)}| \,\Big|\, \mathcal{N}_i^{\pm}\Big].$$

A key symmetry (Lemma 22) shows that these conditional laws differ only by a rotation. Since $\|\tilde{w}^{(t)}\|\|\tilde{a}^{(t)}|$ is rotation-invariant, the conditional expectations coincide across blocks. Concentration of the empirical averages around these expectations then yields near-balance.

Thus, by the end of Phase I, neurons have developed strong signals (Lemma 2) and block masses remain nearly balanced (Lemma 3), preparing the ground for the block-dynamic analysis in Phase II.

4.2 Phase II : Blocks dynamics

In Phase I, neurons evolved nearly independently under the correlation-loss approximation. In Phase II, this approximation no longer holds: interactions between neurons matter, and the signal component dominates. Our analysis, therefore, shifts to the block level.

Signal-heavy networks. Some neurons fail to develop large signals, but their contribution to $f_{\rho^{(t)}}$ remains negligible. We thus focus on *heavy neurons*, where the signal w_{sig} dominates the noise terms $w_\perp, w_{\text{opp}}$.

Definition 2 (Signal-heavy network)**.** *Let $H > 1$ be a constant and $\zeta' = o(1)$. A network is said to be (ζ', H)-signal-heavy if there is a set $\mathcal{S} \subset \mathcal{N}$ of heavy neurons that satisfy the following conditions:*

1. (Signal heavy neuron): For all neurons in $\mathcal{S}$, $\|w_\perp\| + \|w_{\text{opp}}\| \le \zeta' \|w_{\text{sig}}\|$ holds.
2. (Non-heavy neurons mass is negligible): $\mathbb{E}_\rho \mathbf{1}_{\{(w,a)\notin\mathcal{S}\}} \|w\|^2 \le \zeta' N^{(t)}$ holds.
3. (Layer weights balance): $\mathbb{E}_\rho \|w\|^2 \le \mathbb{E}_\rho a^2 + \zeta' H$ and $|a| \le \|w\|$ hold for all neurons.

Intuitively, heavy neurons carry the signal, while non-heavy ones are negligible, see Figure 4 for an illustration. The following lemma formalizes that this property is stable over training (see Appendix C.3).

Lemma 4 (Stability of signal-heavy network)**.** *If the network is (ζ', H)-signal-heavy at time t, then after one gradient step, with probability at least $1 - d^{-\Omega(1)}$, it remains $(\zeta'(1 + O(\eta\zeta')), H)$-signal-heavy.*

Oracle approximation. To analyze block dynamics, we approximate population gradients in two steps:

Step 1: Since the signal component is dominant, we approximate $w^\top x$ by $w_{\mathrm{sig}}^\top z$, allowing us to use $\ell'_{\rho^{(t)}}(x) \approx \ell'_{\rho^{(t)}}(z)$. To formalize this approximation, we define a *clean gradient* ∇^{cl} as

$$\nabla_w^{\mathrm{cl}} L_\rho := a_w \, \mathbb{E}_x \, \ell'_\rho(z) \, \sigma'(w^\top x) \, x,$$
$$\nabla_a^{\mathrm{cl}} L_\rho := \mathbb{E}_x \, \ell'_\rho(z) \, \sigma(w^\top x).$$

Step 2: We introduce an *oracle network*, which approximates the model $f_{\rho^{(t)}}$ by the following model across each direction $\pm\mu_1$ and $\pm\mu_2$:

$$f_{\rho^{(t)},\mathrm{id}}(z) = N^{(t)} \sum_{i=1}^{2} (-1)^{i+1} \left(\sigma(\mu_i^\top z) + \sigma(-\mu_i^\top z) \right).$$

Using the approximate balance of the neuron blocks established in Phase I (see Lemma 3), we obtain the approximation $\ell'_\rho(z) \approx \ell'_{\rho,\mathrm{id}}(z)$. For details, see Section C.1. Also see Figure 4 for an illustration.

We also define the notion of an *average margin* for the oracle network as

$$g_\mu^{(t)} = \mathbb{E}_z \, \ell'_{\rho^{(t)},\mathrm{id}}(z) \, \sigma(\mu^\top z).$$

We will sometimes forget the dependence on time t to simplify the notation. Using this notion, we describe the dynamics of the oracle network in the following lemma. The symmetry of this quantity within the oracle model will be leveraged in the lemma to show that the gradients evolve at the same rate in each direction. A full version of the following lemmas is presented in Section C.2.

Lemma 5 (Block Dynamics via the Oracle Approximation)**.** *The following hold:*

1. (oracle alignment): $(w, a) \in \mathcal{S}$ such that $w_{\mathrm{sig}}^\top \mu > 0$ for some $\mu \in \{\pm\mu_1, \pm\mu_2\}$ we have
$$\mu^\top \nabla_w^{\mathrm{cl}} L_{\rho,\mathrm{id}} = -|a| g_\mu (1 \pm o(1)),$$
$$y \nabla_a^{\mathrm{cl}} L_{\rho,\mathrm{id}} = -(1 \pm o(1)) \|w_{\mathrm{sig}}\| g_\mu.$$
2. (average margin asymptotics): There exist constants $0 < c_1 < c_2$ such that we have for all $t \le T_b$
$$c_1 (1 \wedge (N^{(t)})^{-3}) \le g_{\mu^{(t)}} \le c_2 (1 \wedge (N^{(t)})^{-3}).$$

The first result shows that gradients align with the signal and scale with the average margin g_μ. The second bounds g_μ: initially constant, it decays as $N^{(t)}$ grows.

Inductive block growth. Finally, we show that block mass grows multiplicatively at rate g_μ:

Lemma 6 (Inductive Block Growth)**.** *If the network is (ζ', H)-signal-heavy at time t, then with probability at least $1 - d^{-\Omega(1)}$,*

$$N^{(t+1)} = (1 + 2\eta g_\mu^{(t)})(1 + o(1)) \, N^{(t)}.$$

Combined with the pre-balance of Phase I, this ensures sustained, symmetric block growth and drives the loss toward zero.

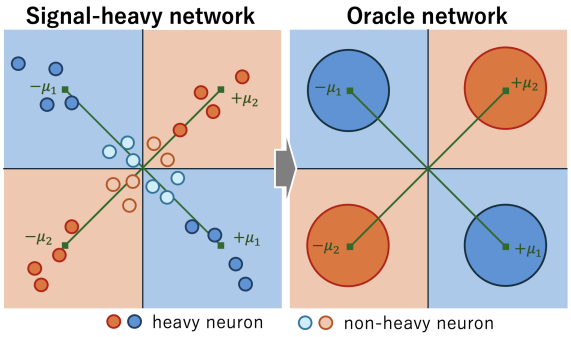


Figure 4: Neuron dynamics in Phase II.

4.3 Conclusion: Proof of Theorem 1

To conclude the proof, we connect the block mass $N^{(T)}$ with the expected risk $\mathbb{E}_x \ell_{\rho^{(T)}}(x)$. The argument proceeds in three steps:

(i) **Growth of block mass.** From the Phase II analysis, the network is signal-heavy and the block mass grows at a rate $g_\mu^{(t)} \gtrsim (N^{(t)})^{-3}$. By a contradiction argument, we show that within $T = O(\eta^{-1}(\log\log d)^4)$ steps, the block mass reaches order $\log\log d$.

(ii) **Accuracy away from the boundary.** We split the input space into points close to the decision boundary, where misclassification is unavoidable, and points sufficiently far. Using a polar decomposition of the Gaussian input, we prove that the near-boundary region has probability mass $O((\log\log d)^{-1/2})$. For points outside this set, the oracle network output has the correct sign with high confidence.

(iii) **Approximation error.** Finally, we control the difference between the actual network and the idealized block model. By signal-heaviness and the balanced evolution of blocks, this error is negligible compared to the dominant term above.

Remark 3. *A key novelty of our proof is the analysis of block dynamics. The growth rate of each block is governed by the average margin $g_\mu^{(t)}$. Although $g_\mu^{(t)}$ converges to zero, its rate of decay can be related to the evolution of $N^{(t)}$, which in turn forces $N^{(t)}$ to diverge to infinity.*

5 NUMERICAL EXPERIMENTS

We conduct numerical experiments to validate our theoretical results and to explore settings beyond our formal assumptions. The implementation of these experiments is available in the following notebook: `https://colab.research.google.com/drive/1bvFXsM_Ji4E3MXYYcAUgTLHjenqOaASI?usp=sharing`.

5.1 Decision Boundary During Training.

We track the evolution of weights and the decision boundary for $d = 600$, $m = 400$, $M = 82000$, $\theta = \eta = 0.01$. Figure 5 shows the projection of weights $w^{(t)}$ on $\mathrm{span}(e_1, e_2)$ at different training stages (for readability, we keep only neurons with $\|w_{\mathrm{sig}}^{(0)}\| \geq 2\theta \log d/\sqrt{d}$). The background indicates the predicted clusters, and neurons progressively align with one of the four directions $\pm\mu_1, \pm\mu_2$, sharpening the decision boundary.

Figure 5d confirms that block masses grow at comparable rates, consistent with our block-dynamic analysis. The purple curve represents the residual mass $\mathcal{R} = \sum_{(w,a)\in\mathcal{N}} |a|\|w_\perp\|$, which remains small.

5.2 Additional Experiments

We complement our main experiments with several variants (details in Appendix D). Overall, our findings show that training dynamics are robust across different input distributions and settings, but can be sensitive to noise and overtraining:

- **Input distribution.** Replacing Gaussian inputs with uniform or Gaussian XOR inputs still leads to successful training, provided the distribution is symmetric with respect to the four directions $\pm\mu_1, \pm\mu_2$.
- **Noise sensitivity.** Flipping only 5% of the labels already degrades test performance significantly, indicating a vulnerability of SGD to mislabeled data. However, the decision boundary learned by the network is only slightly biased along the x- and y-axes.
- **Anisotropy.** With anisotropic covariance, the network learns the decision boundary, but weight evolution becomes disordered compared to the isotropic case.
- **Nonlinear boundaries.** For highly nonlinear decision functions (e.g., sinusoidal boundaries), the network learns only a linear approximation, and training quickly plateaus.

Taken together, these experiments confirm the key ingredients of our analysis: neurons self-organize into balanced blocks, block growth at a similar rate, and the geometry of the zero-margin setting imposes inherent limitations on convergence.

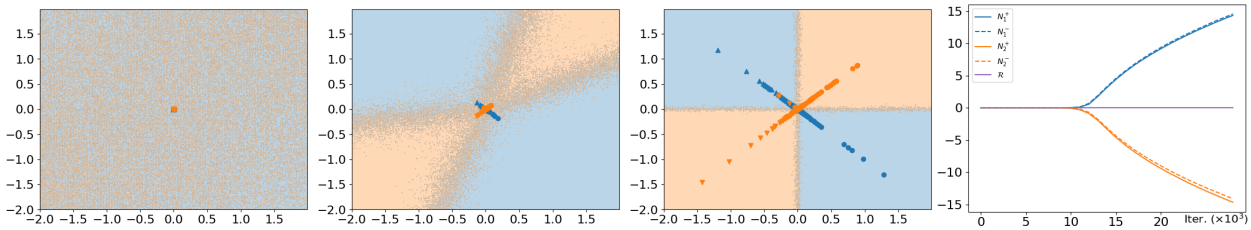
6 DISCUSSION

We analyzed the training dynamics of a two-layer network on Gaussian XOR, a canonical *zero-margin* classification problem. Our key finding is that although individual neurons evolve heterogeneously, they rapidly self-organize into four balanced blocks whose dynamics reduce to a low-dimensional symmetric system. This block-level description yields convergence guarantees without relying on margin assumptions: generalization is governed by *average-case dynamics* near the decision boundary rather than worst-case margins. We propose this *block-dynamics* perspective as a new analytical tool for studying feature learning in classification settings where standard separability arguments break down.

Our analysis is limited to the XOR boundary with isotropic Gaussian inputs, chosen for tractability. Extending block-dynamic arguments to richer decision boundaries and more general input distributions is a natural direction for future work. More broadly, block-level analysis offers a promising framework for understanding the training dynamics of neural networks in challenging classification regimes.

Acknowledgement

MI was supported by JSPS KAKENHI (24K02904), JST CREST (JPMJCR21D2), JST FOREST (JPMJFR216I), and JST BOOST (JPMJBY24A9).



(a) $T = 8000$ (b) $T = 12000$ (c) $T = 30000$ (d) Block growth.

Figure 5: Simulated neuron dynamics. (a–c) Weight projections (dots; colors indicate assigned labels) and decision boundaries. The weights progressively align with the directions $\pm\mu_i$. (d) The block mass $N_i^{\pm}$ vs time (purple: residual mass $\mathcal{R}$). Block masses remain approximately balanced.

Bibliography

E. Abbe, E. B. Adserà, and T. Misiakiewicz. SGD learning on neural networks: leap complexity and saddle-to-saddle dynamics. In *The Thirty Sixth Annual Conference on Learning Theory, COLT*, volume 195, pages 2552–2623, 2023.

E. Abbe, E. Cornacchia, J. Hązła, and D. Kougang-Yombi. Learning high-degree parities: The crucial role of the initialization. In *The Thirteenth International Conference on Learning Representations*, 2025.

B. Barak, B. L. Edelman, S. Goel, S. M. Kakade, E. Malach, and C. Zhang. Hidden progress in deep learning: SGD learns parities near the computational limit. In *Advances in Neural Information Processing Systems*, 2022.

R. Berthier, A. Montanari, and K. Zhou. Learning time-scales in two-layers neural networks. *Found. Comput. Math.*, Aug. 2024.

A. Bietti, J. Bruna, C. Sanford, and M. J. Song. Learning single-index models with shallow neural networks. In *Advances in Neural Information Processing Systems*, 2022.

A. Bietti, J. Bruna, and L. Pillaud-Vivien. On learning gaussian multi-index models with gradient flow part i: General properties and two-timescale learning. *Communications on Pure and Applied Mathematics*, 2025.

J. Bruna and D. Hsu. Survey on algorithms for multi-index models, 2025. URL `https://arxiv.org/abs/2504.05426`.

A. Brutzkus, A. Globerson, E. Malach, and S. Shalev-Shwartz. SGD learns over-parameterized networks that provably generalize on linearly separable data. In *International Conference on Learning Representations*, 2018. URL `https://openreview.net/forum?id=rJ33wwxRb`.

Y. Cao, Z. Chen, M. Belkin, and Q. Gu. Benign overfitting in two-layer convolutional neural networks. In *Proceedings of the 36th International Conference on Neural Information Processing Systems*, NIPS '22, 2022.

L. Chizat, E. Oyallon, and F. R. Bach. On lazy training in differentiable programming. In *Neural Information Processing Systems*, 2018.

Y. Dandi, E. Troiani, L. Arnaboldi, L. Pesce, L. Zdeborova, and F. Krzakala. The benefits of reusing batches for gradient descent in two-layer networks: breaking the curse of information and leap exponents. In *Proceedings of the 41st International Conference on Machine Learning*, ICML'24. JMLR.org, 2024.

L. Deng. The mnist database of handwritten digit images for machine learning research [best of the web]. *IEEE Signal Processing Magazine*, 29(6): 141–142, 2012.

S. Frei, N. Chatterji, and P. L. Bartlett. Benign overfitting without linearity: Neural network classifiers trained by gradient descent for noisy linear data. *Proceedings of the 35th Conference on Learning Theory (COLT2022)*, 2022.

B. Ghorbani, S. Mei, T. Misiakiewicz, and A. Montanari. Linearized two-layers neural networks in high dimension. *The Annals of Statistics*, 49, 04 2021.

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

W. Huang, Y. Cao, H. Wang, X. Cao, and T. Suzuki. Quantifying the optimization and generalization advantages of graph neural networks over multilayer perceptrons. In *The 28th International*

Conference on Artificial Intelligence and Statistics, 2025.

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

Z. Ji and M. Telgarsky. Directional convergence and alignment in deep learning. In *Proceedings of the 34th International Conference on Neural Information Processing Systems*, NIPS '20, 2020.

J. Jiang, W. Huang, M. Zhang, T. Suzuki, and L. Nie. Unveil benign overfitting for transformer in vision: Training dynamics, convergence, and generalization. In A. Globerson, L. Mackey, D. Belgrave, A. Fan, U. Paquet, J. Tomczak, and C. Zhang, editors, *Advances in Neural Information Processing Systems*, volume 37, pages 135464–135625. Curran Associates, Inc., 2024.

G. Kornowski, G. Yehudai, and O. Shamir. From tempered to benign overfitting in reLU neural networks. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023.

Y. Kou, Z. Chen, Y. Chen, and Q. Gu. Benign overfitting in two-layer ReLU convolutional neural networks. In A. Krause, E. Brunskill, K. Cho, B. Engelhardt, S. Sabato, and J. Scarlett, editors, *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pages 17615–17659. PMLR, 23–29 Jul 2023.

Y. Kou, Z. Chen, Q. Gu, and S. M. Kakade. Matching the statistical query lower bound for k-sparse parity problems with sign stochastic gradient descent. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024.

K. Lyu, Z. Li, R. Wang, and S. Arora. Gradient descent on two-layer nets: Margin maximization and simplicity bias. In A. Beygelzimer, Y. Dauphin, P. Liang, and J. W. Vaughan, editors, *Advances in Neural Information Processing Systems*, 2021. URL `https://openreview.net/forum?id=Aa5oPXc_1IV`.

A. V. Mahankali, J. Z. HaoChen, K. Dong, M. Glasgow, and T. Ma. Beyond NTK with vanilla gradient descent: A mean-field analysis of neural networks with polynomial width, samples, and time. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023.

S. Mei, T. Misiakiewicz, and A. Montanari. Mean-field theory of two-layers neural networks: dimension-free bounds and kernel limit. In *Annual Conference Computational Learning Theory*, 2019.

A. Mousavi-Hosseini, S. Park, M. Girotti, I. Mitliagkas, and M. A. Erdogdu. Neural networks efficiently learn low-dimensional representations with SGD. In *The Eleventh International Conference on Learning Representations*, 2023.

K. Oko, Y. Song, T. Suzuki, and D. Wu. Learning sum of diverse features: computational hardness and efficient gradient-based training for ridge combinations. In *Proceedings of Thirty Seventh Conference on Learning Theory*, volume 247 of *Proceedings of Machine Learning Research*, pages 4009–4081. PMLR, 30 Jun–03 Jul 2024.

S. Shalev-Shwartz, O. Shamir, and S. Shammah. Failures of gradient-based deep learning. In *International Conference on Machine Learning*, pages 3067–3075. PMLR, 2017.

Z. Shi, J. Wei, and Y. Liang. A theoretical analysis on feature learning in neural networks: Emergence from inputs and advantage over fixed features. In *International Conference on Learning Representations*, 2022.

Z. Shi, J. Wei, and Y. Liang. Provable guarantees for neural networks via gradient feature learning. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023. URL `https://openreview.net/forum?id=5F04bU79eK`.

T. Suzuki, D. Wu, K. Oko, and A. Nitanda. Feature learning via mean-field langevin dynamics: classifying sparse parities and beyond. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023.

M. Telgarsky. Feature selection and low test error in shallow low-rotation reLU networks. In *The Eleventh International Conference on Learning Representations*, 2023.

R. Vershynin. *High-Dimensional Probability: An Introduction with Applications in Data Science*. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press, 2018.

Y. Wang, K. Zhang, and R. Arora. Benign overfitting in adversarial training of neural networks. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pages 52171–52232. PMLR, 21–27 Jul 2024.

C. Wei, J. Lee, Q. Liu, and T. Ma. Regularization matters: Generalization and optimization of

neural nets v.s. their induced kernel. In *Neural Information Processing Systems*, 2018.

Z. Xu, Y. Wang, S. Frei, G. Vardi, and W. Hu. Benign overfitting and grokking in reLU networks for XOR cluster data. In *The Twelfth International Conference on Learning Representations*, 2024.

Z. Zhu, F. Liu, G. G. Chrysos, F. Locatello, and V. Cevher. Benign overfitting in deep neural networks under lazy training. In *Proceedings of the 40th International Conference on Machine Learning*, ICML'23. JMLR.org, 2023.

Checklist

The checklist follows the references. For each question, choose your answer from the three possible options: Yes, No, Not Applicable. You are encouraged to include a justification to your answer, either by referencing the appropriate section of your paper or providing a brief inline description (1-2 sentences). Please do not modify the questions. Note that the Checklist section does not count towards the page limit. Not including the checklist in the first submission won't result in desk rejection, although in such case we will ask you to upload it during the author response period and include it in camera ready (if accepted).

1. For all models and algorithms presented, check if you include:
   (a) A clear description of the mathematical setting, assumptions, algorithm, and/or model. [Yes]
   (b) An analysis of the properties and complexity (time, space, sample size) of any algorithm. [Yes]
   (c) (Optional) Anonymized source code, with specification of all dependencies, including external libraries. [Yes]

2. For any theoretical claim, check if you include:
   (a) Statements of the full set of assumptions of all theoretical results. [Yes]
   (b) Complete proofs of all theoretical results. [Yes]
   (c) Clear explanations of any assumptions. [Yes]

3. For all figures and tables that present empirical results, check if you include:
   (a) The code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL). [Yes]
   (b) All the training details (e.g., data splits, hyperparameters, how they were chosen). [Yes]
   (c) A clear definition of the specific measure or statistics and error bars (e.g., with respect to the random seed after running experiments multiple times). [Yes]
   (d) A description of the computing infrastructure used. (e.g., type of GPUs, internal cluster, or cloud provider). [Yes]

4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets, check if you include:
   (a) Citations of the creator If your work uses existing assets. [Not Applicable]
   (b) The license information of the assets, if applicable. [Not Applicable]
   (c) New assets either in the supplemental material or as a URL, if applicable. [Not Applicable]
   (d) Information about consent from data providers/curators. [Not Applicable]
   (e) Discussion of sensible content if applicable, e.g., personally identifiable information or offensive content. [Not Applicable]

5. If you used crowdsourcing or conducted research with human subjects, check if you include:
   (a) The full text of instructions given to participants and screenshots. [Not Applicable]
   (b) Descriptions of potential participant risks, with links to Institutional Review Board (IRB) approvals if applicable. [Not Applicable]
   (c) The estimated hourly wage paid to participants and the total amount spent on participant compensation. [Not Applicable]

Supplementary Materials

A PRELIMINARY LEMMAS

We will first introduce some technical results that will be used repeatedly in the proof of Lemma 2.

A.1 Concentration of Empirical Gradients

The following lemma bounds the difference between the population gradient and its empirical counterpart. It extends Lemma B.12 in Glasgow (2024) to the Gaussian setting.

Lemma 7 (Empirical concentration of the gradients)**.** *Assume that the input data are as described in Section 2, and that the neural network width satisfies $m \leq d^c$ for some constant $c > 0$. Let V denote the batch size, i.e., $|M_t|$. Recall that $L = \mathbb{E}_x[\ell_\rho(x)]$ and $\hat{L} = \frac{1}{|V|}\sum_{j\in V} \ell_\rho(x^{(j)})$ for some loss function ℓ_ρ.*

Then, for any loss function ℓ_ρ that is differentiable and 2-Lipschitz, there exists a constant $C > 0$ such that with probability at least $1 - d^{-\Omega(1)}$, for all neurons $(w, a) \in \mathcal{N}$, we have

1. $\|\nabla_w L - \nabla_w \hat{L}\|^2 \leq C^2 a^2 \frac{d \log d}{V}$,
2. $\|\nabla_{w_i} L - \nabla_{w_i} \hat{L}\|^2 \leq C^2 a^2 \frac{\log d}{V}$, *where* $w_i = \langle w, e_i\rangle$,
3. $\left|\nabla_a L - \nabla_a \hat{L}\right|^2 \leq C^2 \|w\|^2 \frac{\log d}{V}$.

Proof. To simplify the notation, we write ℓ instead of ℓ_ρ. By the Lipschitz assumption, we have $\|\ell'\|_\infty \leq 2$. This implies, by the Gaussianity of $x \sim \mathcal{N}(0, I_d)$, that for all $u \in \mathbb{S}^{d-1}$,

$$\mathbb{P}\left(\left|\ell'(x)\sigma'(w^\top x)\langle x, u\rangle\right| > t\right) \leq \mathbb{P}\left(2\left|\langle x, u\rangle\right| > t\right) \leq 2e^{-t^2/8},$$

since for any random variables X, Y and any $t \in \mathbb{R}$, the inequality $Y \geq X$ implies $\{X \geq t\} \subset \{Y \geq t\}$. As a consequence, with the choice $u = e_i$, we obtain that

$$a^{-1}\langle \nabla_w L - \nabla_w \hat{L}, e_i\rangle = \frac{1}{V}\sum_{j\in M_t}\left(\ell'\big(x^{(j)}\big)\sigma'\big(w^\top x^{(j)}\big)x_i^{(j)} - \mathbb{E}_x\big[\ell'(x)\sigma'(w^\top x)x_i\big]\right)$$

is sub-Gaussian with parameter $\frac{2\sqrt{2}}{\sqrt{V}}$, as an average of V independent centered $2\sqrt{2}$-sub-Gaussian random variables.

Hence, we have

$$\mathbb{P}\left(\left|a^{-1}\langle \nabla_w L - \nabla_w \hat{L}, e_i\rangle\right| > t\right) \leq e^{-t^2/(8V)}.$$

By choosing $t = C\sqrt{V \log d}$ for a constant $C > 0$ large enough, and taking a union bound over $i = 1, \ldots, d$, we obtain that with probability at least $1 - d^{-\Omega(1)}$,

$$\|\nabla_{w_i} L - \nabla_{w_i} \hat{L}\|^2 = \langle \nabla_w L - \nabla_w \hat{L}, e_i\rangle^2 \leq C^2 a^2 \frac{\log d}{V},$$

and

$$\|\nabla_w L - \nabla_w \hat{L}\|^2 = \sum_{i=1}^d \langle \nabla_w L - \nabla_w \hat{L}, e_i\rangle^2 \leq C^2 a^2 \frac{d \log d}{V}.$$

A similar argument leads to the third statement of the lemma by noticing that $\ell'(x)\sigma(w^\top x) = \ell'(x)\sigma'(w^\top x)w^\top x$ holds, and that $w^\top x$ is Gaussian with variance $\|w\|^2$. We conclude by taking a union bound over all the m neurons. □

Remark 4. *Since the batch size is chosen such that $|V| \asymp d \log^\beta d$ for some $\beta > 1$, the constant C appearing in Lemma 7 can be absorbed into the asymptotic notation and thus ignored.*

A.2 Basic Properties of the Population Loss Gradients

We describe several basic properties of the population loss gradients. To obtain the results, Lemma B.13 in Glasgow (2024) can be easily adapted to the Gaussian setting. For completeness, we detail the proof.

Lemma 8. *Assume that the neural network width satisfies $m \le d^c$ for some constant $c > 0$, the learning rate satisfies $\eta < \frac{\sqrt{\pi}}{4\sqrt{2}}$, and the batch size satisfies $V = |M_t| \ge d \log^2 d$. Then, with probability at least $1 - d^{-\Omega(1)}$, for all neurons $(w, a) \in \mathcal{N}$, we have*

1. $\|\nabla_w L_\rho\| \le 2\sqrt{\frac{2}{\pi}}|a|$,
2. $|\nabla_a L_\rho| \le (2 + o(1))\sqrt{\frac{2}{\pi}}\|w\|$,
3. *if* $|a^{(t)}| \le \|w^{(t)}\|$, *then* $|a^{(t+1)}| \le \|w^{(t+1)}\|$,
4. $\|w^{(t+1)}\|^2 - |a^{(t+1)}|^2 \le \left(\frac{16}{\pi} + o(1)\right)\eta^2|a^{(t)}|^2 + \|w^{(t)}\|^2 - |a^{(t)}|^2$.

Proof. We prove each statement in turn.

For the first statement, we have

$$\begin{aligned}
\frac{1}{|a|}\|\nabla_w L_\rho\| &= \frac{1}{|a|}\sup_{v\in\mathbb{S}^{d-1}} \langle v, \nabla_w L_\rho\rangle \\
&= \sup_{v\in\mathbb{S}^{d-1}} \mathbb{E}_x\left[\ell'_\rho(x)\sigma'(w^\top x)x^\top v\right] \\
&\le \sup_{v\in\mathbb{S}^{d-1}} \mathbb{E}_x\left[|\ell'_\rho(x)|\,|x^\top v|\right] \\
&\le \sup_{v\in\mathbb{S}^{d-1}} 2\mathbb{E}_x\left[|x^\top v|\right] \\
&= 2\sqrt{\frac{2}{\pi}},
\end{aligned}$$

where we used that $|\ell'_\rho(x)| \le 2$ and that $\mathbb{E}[|x^\top v|] = \sqrt{2/\pi}$ for $x \sim \mathcal{N}(0, I_d)$.

For the second statement, similarly, we obtain

$$\begin{aligned}
|\nabla_a L_\rho| &= \left|\mathbb{E}_x\left[\ell'_\rho(x)\sigma(w^\top x)\right]\right| \\
&\le \mathbb{E}_x\left[|\ell'_\rho(x)|\,|\sigma(w^\top x)|\right] \\
&\le 2\mathbb{E}_x\left[|w^\top x|\right] \\
&= 2\sqrt{\frac{2}{\pi}}\|w\|.
\end{aligned}$$

Here we used that for ReLU, $\sigma(w^\top x) = (w^\top x)_+$ and $\sigma(w^\top x) \le w^\top x$ when $w^\top x \ge 0$.

Now we show the third statement. Since σ is ReLU, we have $\sigma'(w^\top x)w^\top x = \sigma(w^\top x)$ almost everywhere. Combined with the chain rule, this leads to the identity:

$$(w^{(t)})^\top \nabla_{w^{(t)}} \hat{L}_{\rho^{(t)}} = a^{(t)} \nabla_{a^{(t)}} \hat{L}_{\rho^{(t)}}. \tag{A.1}$$

Expanding the update equations, we have

$$\begin{aligned}
(a^{(t+1)})^2 &= \left(a^{(t)} - \eta\nabla_{a^{(t)}}\hat{L}_{\rho^{(t)}}\right)^2 \\
&= (a^{(t)})^2 - 2\eta a^{(t)}\nabla_{a^{(t)}}\hat{L}_{\rho^{(t)}} + \eta^2\left(\nabla_{a^{(t)}}\hat{L}_{\rho^{(t)}}\right)^2,
\end{aligned}$$

and

$$\begin{aligned}
\|w^{(t+1)}\|^2 &= \|w^{(t)} - \eta\nabla_{w^{(t)}}\hat{L}_{\rho^{(t)}}\|^2 \\
&= \|w^{(t)}\|^2 - 2\eta(w^{(t)})^\top\nabla_{w^{(t)}}\hat{L}_{\rho^{(t)}} + \eta^2\|\nabla_{w^{(t)}}\hat{L}_{\rho^{(t)}}\|^2.
\end{aligned}$$

Using (A.1), we obtain

$$\begin{aligned}(a^{(t+1)})^2 - \|w^{(t+1)}\|^2 &= (a^{(t)})^2 - \|w^{(t)}\|^2 + \eta^2 \left((\nabla_{a^{(t)}} \hat{L}_{\rho^{(t)}})^2 - \|\nabla_{w^{(t)}} \hat{L}_{\rho^{(t)}}\|^2 \right) \\ &\leq (a^{(t)})^2 - \|w^{(t)}\|^2 + \eta^2 \left((\nabla_{a^{(t)}} \hat{L}_{\rho^{(t)}})^2 - \frac{1}{\|w^{(t)}\|^2} \left((w^{(t)})^\top \nabla_{w^{(t)}} \hat{L}_{\rho^{(t)}} \right)^2 \right) \\ &= (a^{(t)})^2 - \|w^{(t)}\|^2 + \frac{\eta^2 (\nabla_{a^{(t)}} \hat{L}_{\rho^{(t)}})^2}{\|w^{(t)}\|^2} \left(\|w^{(t)}\|^2 - (a^{(t)})^2 \right) \\ &= \left((a^{(t)})^2 - \|w^{(t)}\|^2 \right) \left(1 - \frac{\eta^2 (\nabla_{a^{(t)}} \hat{L}_{\rho^{(t)}})^2}{\|w^{(t)}\|^2} \right).\end{aligned}$$

By the choice of the batch size and Lemma 7, we have with probability at least $1 - d^{-\Omega(1)}$,

$$\begin{aligned}|\nabla_a \hat{L}_\rho| &\leq |\nabla_a L_\rho| + |\nabla_a L_\rho - \nabla_a \hat{L}_\rho| \\ &\leq 2\sqrt{\frac{2}{\pi}} \|w\| + C\|w\| \sqrt{\frac{d \log^2 d}{V}} \\ &\leq (2 + o(1)) \sqrt{\frac{2}{\pi}} \|w\|,\end{aligned}$$

where the last inequality uses that $V \geq d \log^2 d$.

Thus, if $|a^{(t)}| \leq \|w^{(t)}\|$ initially, the same inequality holds at step $t+1$.

Finally, for the fourth statement, by again using (A.1), we have

$$\begin{aligned}\|w^{(t+1)}\|^2 - (a^{(t+1)})^2 - \left(\|w^{(t)}\|^2 - (a^{(t)})^2 \right) &= \eta^2 \left(\|\nabla_w \hat{L}_{\rho^{(t)}}\|^2 - (\nabla_a \hat{L}_{\rho^{(t)}})^2 \right) \\ &\leq \eta^2 \|\nabla_w \hat{L}_{\rho^{(t)}}\|^2 \\ &\leq 2\eta^2 \left(\|\nabla_w L_{\rho^{(t)}}\|^2 + \|\nabla_w \hat{L}_{\rho^{(t)}} - \nabla_w L_{\rho^{(t)}}\|^2 \right) \\ &\leq \left(\frac{16}{\pi} + o(1) \right) \eta^2 a^2,\end{aligned}$$

with probability at least $1 - d^{-\Omega(1)}$, which completes the proof. □

A.3 Other Technical Lemmas

In this section, we collect standard technical estimates for Gaussian random variables that will be used throughout the proofs.

Lemma 9. *Let $X \sim \mathcal{N}(0,1)$. For all $\epsilon \in (0,1)$, we have*

$$\sqrt{\frac{2}{\pi}} e^{-\epsilon^2/2} \left(\epsilon + \frac{\epsilon^3}{3} \right) \leq \mathbb{P}(|X| \leq \epsilon) \leq \sqrt{\frac{2}{\pi}} \epsilon.$$

In particular, when $\epsilon = o(1)$, we have

$$\mathbb{P}(|X| \leq \epsilon) = \sqrt{\frac{2}{\pi}} \epsilon + O(\epsilon^3).$$

Proof. The upper bound can be directly obtained by bounding $e^{-t^2/2}$ by 1 over $(-\epsilon, \epsilon)$.

For the lower bound, consider the function

$$\Omega(x) = e^{x^2/2} \int_0^x e^{-t^2/2} \, \mathrm{d}t.$$

It satisfies the differential equation $\Omega'(x) = 1 + x\Omega(x)$ with initial conditions $\Omega(0) = 0$ and $\Omega'(0) = 1$. In particular, it implies that the coefficients a_n of its Taylor expansion around 0 satisfy the relation $a_{n+2} = \frac{a_n}{n+2}$. Consequently,

$$\begin{aligned}\frac{1}{\sqrt{2\pi}} \int_0^x e^{-t^2/2} \, \mathrm{d}t &= \frac{1}{\sqrt{2\pi}} e^{-x^2/2} \Omega(x) \\ &= \frac{1}{\sqrt{2\pi}} e^{-x^2/2} \sum_{k=0}^{\infty} \frac{x^{2k+1}}{(2k+1)!!}.\end{aligned}$$

The result follows by substituting $x = \epsilon$. □

We recall the classical bounds for the upper tail of a Gaussian random variable:

Lemma 10. *[Mill's ratio bounds] Let $X \sim \mathcal{N}(0,1)$. For all $t > 0$, we have*

$$\frac{t}{1+t^2} \frac{1}{\sqrt{2\pi}} e^{-t^2/2} \leq \mathbb{P}(X \geq t) \leq \frac{1}{t\sqrt{2\pi}} e^{-t^2/2}.$$

Proof. The upper bound follows from an integration by parts, using that

$$\mathbb{P}(X \geq t) = \int_t^{+\infty} \frac{1}{\sqrt{2\pi}} e^{-s^2/2} \, \mathrm{d}s \leq \frac{1}{t} \frac{1}{\sqrt{2\pi}} e^{-t^2/2}.$$

The lower bound follows from a refined integration by parts argument and is a standard variant of Mill's ratio; see, e.g., (Vershynin, 2018, Proposition 2.1.2). □

Finally, we recall a useful bound for the Laplace transform of the folded Gaussian distribution:

Lemma 11. *Let $X \sim \mathcal{N}(0,1)$. For all $t > 0$, we have*

$$\mathbb{E}\left(e^{-t|X|}\right) \leq \frac{\sqrt{2}}{\sqrt{\pi} t}.$$

Proof. It follows from the identity

$$\mathbb{E}\left(e^{-t|X|}\right) = 2e^{t^2/2} \mathbb{P}(X \geq t),$$

combined with the classical upper bound on $\mathbb{P}(X \geq t)$ from Lemma 10. □

Lemma 12. *Let $X \sim \mathcal{N}(0,1)$ be a standard normal random variable. Then, for all $t > 0$, we have*

$$\mathbb{E}\left(e^{-t^2 X^2}\right) = \frac{1}{\sqrt{2(t^2 + 1/2)}}.$$

Proof. By definition of the expectation under the standard normal distribution, we have

$$\mathbb{E}\left(e^{-t^2 X^2}\right) = \int_{-\infty}^{\infty} e^{-t^2 x^2} \cdot \frac{1}{\sqrt{2\pi}} e^{-x^2/2} \, dx.$$

Combining the exponential terms gives

$$\mathbb{E}\left(e^{-t^2 X^2}\right) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{-x^2(t^2+1/2)} \, dx.$$

This is a standard Gaussian integral of the form

$$\int_{-\infty}^{\infty} e^{-ax^2} \, dx = \sqrt{\frac{\pi}{a}} \quad \text{for } a > 0.$$

Applying this with $a = t^2 + 1/2$, we get

$$\mathbb{E}\left(e^{-t^2X^2}\right) = \frac{1}{\sqrt{2\pi}} \cdot \sqrt{\frac{\pi}{t^2+1/2}} = \frac{1}{\sqrt{2(t^2+1/2)}}.$$

For the asymptotic behavior as $t \to \infty$, note that

$$\frac{1}{\sqrt{2(t^2+1/2)}} \sim \frac{1}{\sqrt{2}\cdot t}.$$

□

B ANALYSIS OF PHASE I

At the beginning of Phase I, $\|w_{\text{sig}}\|$ is much smaller than $\|w_\perp\|$, but after several iterations, the signal magnitude exceeds the global noise level. As our computations show, the SGD dynamics depend on the relative magnitude between $\|w_{\text{sig}}\|$ and $\|w_\perp\|$, leading us to subdivide Phase I into two subphases: Phase Ia, where $\|w_{\text{sig}}\| \ll \|w_\perp\|$, and Phase Ib, where $\|w_{\text{sig}}\| \gtrsim \|w_\perp\|$.

B.1 Evaluation of ∇L_0 during Phase Ia

At initialization, typical neurons satisfy $\|w_{\text{sig}}^{(0)}\| \approx d^{-1/2}\theta \ll \|w_\perp^{(0)}\| \approx \theta$. As long as $\|w_{\text{sig}}\| = o(\|w_\perp\|)$ holds, the following lemma characterizes the behavior of the population gradients.

Lemma 13 (L_0 Population Gradients)**.** *For any neuron $(w,a) \in \mathcal{N}$, we have*

1. $$-w_{sig}^\top \nabla_w L_0 = \frac{2}{\pi\sqrt{2\pi}}\,|a|\,\frac{\|w_{sig}\|^2}{\|w_\perp\|} + O\left(|a|\frac{\|w_{1:2}\|^3\|w_{sig}\|}{\|w_\perp\|^3}\right),$$

2. $$-w_{opp}^\top \nabla_w L_0 = -\frac{2}{\pi\sqrt{2\pi}}\,|a|\,\frac{\|w_{opp}\|^2}{\|w_\perp\|} + O\left(|a|\frac{\|w_{1:2}\|^3\|w_{opp}\|}{\|w_\perp\|^3}\right),$$

3. $$\operatorname{sgn}\left(-w_\perp^\top \nabla_w L_0\right) = \begin{cases} 1 & \text{if } \|w_{opp}\| > \|w_{sig}\|, \\ -1 & \text{if } \|w_{opp}\| \le \|w_{sig}\|, \end{cases}$$

   and

   $$\left|w_\perp^\top \nabla_w L_0\right| \le 0.5\,|a|\,\|w_\perp\|.$$

4. *For $i \ge 3$, we have*

   $$-\,w_i\,\nabla_{w_i} L_0 = \frac{a}{2}\,\mathbb{E}_x\left[y(z)\,\mathbf{1}\left(|w^\top z + w^\top \xi_{\setminus i}| \le |w_i\xi_i|\right)\,|w_i\xi_i|\right].$$

   Furthermore, for neurons such that $\|w_\perp\|_\infty \ll \|w_\perp\|$ and $\|w_{1:2}\| = o(1)$, we have

   $$|w_i\,\nabla_{w_i} L_0| \le \frac{4|aw_i|}{\|w_\perp\|}\frac{\|w_{1:2}\|^2 + |w_i|\|w_{1:2}\|}{\|w_\perp\|^2} = o\left(\frac{|aw_i|}{\|w_\perp\|}\right).$$

Proof. We analyze each statement separately. Throughout the proof, write $x = z + \xi$, where z is the projection of x onto $\operatorname{span}\{e_1, e_2\}$ and ξ is the orthogonal noise.

Proof of the first and second statements. Using the symmetrization trick and ReLU's $\sigma'(u) = \mathbf{1}_{\{u>0\}}$,

$$\begin{aligned}
-\nabla_{w_{1:2}} L_0 &= \frac{a}{2}\,\mathbb{E}_z\mathbb{E}_\xi\left[y\left(\sigma'(w^\top\xi + w^\top z) - \sigma'(w^\top\xi - w^\top z)\right) z\right] \\
&= \frac{a}{2}\,\mathbb{E}_z\left[y\ \mathrm{sgn}(w^\top z)\, z\ \mathbb{P}_\xi\big(|w^\top\xi| \le |w^\top z|\big)\right] \\
&= \frac{a}{\sqrt{2\pi}\,\|w_\perp\|}\,\mathbb{E}_z\left[y\,(w^\top z)\,z\right] + O\!\left(|a|\frac{\mathbb{E}_z[|w^\top z|^3 z]}{\|w_\perp\|^3}\right) \qquad \text{(Lemma 9 with } \epsilon = |w^\top z|/\|w_\perp\|) \\
&= \frac{a}{2\sqrt{2\pi}\,\|w_\perp\|}\left(\mathbb{E}_{z|y=1}[(w^\top z)z] - \mathbb{E}_{z|y=-1}[(w^\top z)z]\right) + O\!\left(|a|\frac{\|w_{1:2}\|^3}{\|w_\perp\|^3}\right)(1,1)^\top. \qquad \text{(B.1)}
\end{aligned}$$

For Gaussian XOR labels ($y = +1$ iff $z_1 z_2 < 0$), one has

$$\mathbb{E}_{z|y=\pm1}[z_1^2] = \mathbb{E}_{z|y=\pm1}[z_2^2] = 1, \qquad \mathbb{E}_{z|y=1}[z_1 z_2] = -\frac{2}{\pi}, \quad \mathbb{E}_{z|y=-1}[z_1 z_2] = \frac{2}{\pi}.$$

Hence, for any $w \in \mathbb{R}^2$,

$$\mathbb{E}_{z|y=1}[(w^\top z)z] = \begin{pmatrix} 1 & -\frac{2}{\pi} \\ -\frac{2}{\pi} & 1 \end{pmatrix} w, \qquad \mathbb{E}_{z|y=-1}[(w^\top z)z] = \begin{pmatrix} 1 & \frac{2}{\pi} \\ \frac{2}{\pi} & 1 \end{pmatrix} w.$$

Plugging into (B.1),

$$-\nabla_{w_{1:2}} L_0 = -\frac{2}{\pi\sqrt{2\pi}}\,\frac{a}{\|w_\perp\|}\,(w_2, w_1)^\top + O\!\left(|a|\frac{\|w_{1:2}\|^3}{\|w_\perp\|^3}\right)(1,1)^\top.$$

Since the sign of a is linked to the definition of w_{sig},

$$-w_{\mathrm{sig}}^\top \nabla_{w_{1:2}} L_0 = \frac{2}{\pi\sqrt{2\pi}}\,|a|\,\frac{\|w_{\mathrm{sig}}\|^2}{\|w_\perp\|} + O\!\left(|a|\frac{\|w_{1:2}\|^3\|w_{\mathrm{sig}}\|}{\|w_\perp\|^3}\right),$$

$$-w_{\mathrm{opp}}^\top \nabla_{w_{1:2}} L_0 = -\frac{2}{\pi\sqrt{2\pi}}\,|a|\,\frac{\|w_{\mathrm{opp}}\|^2}{\|w_\perp\|} + O\!\left(|a|\frac{\|w_{1:2}\|^3\|w_{\mathrm{opp}}\|}{\|w_\perp\|^3}\right),$$

which are items 1–2.

Proof of the third statement. We now study the projection onto $w_\perp$. By using again a symmetrization argument, we can write

$$\begin{aligned}
-w_\perp^\top \nabla_{w_\perp} L_0 &= \frac{a}{2}\mathbb{E}_x\left[y(z)\left(\sigma'(w^\top z + w^\top\xi) - \sigma'(w^\top z - w^\top\xi)\right) w_\perp^\top\xi\right] \\
&= \frac{a}{2}\mathbb{E}_z\left[y(z)\mathbb{E}_\xi\left[\mathbf{1}\left(|w^\top\xi| \ge |w^\top z|\right)|w^\top\xi|\right]\right] \\
&= \frac{a}{4}\mathbb{E}_\xi\left[|w^\top\xi|\left(\mathbb{P}_{z|y=1}\left(|w^\top\xi| \ge |w^\top z|\right) - \mathbb{P}_{z|y=-1}\left(|w^\top\xi| \ge |w^\top z|\right)\right)\right].
\end{aligned}$$

Without loss of generality, assume $a > 0$ (i.e., w_{sig} is aligned with μ_1). Write $z = r(\cos\theta, \sin\theta)$, where θ is sampled uniformly over $(-\pi, \pi]$ and r is independent (distributed as the square root of a Chi-square r.v. with two degree of freedom).

Assume that $\|w_{\mathrm{sig}}\| \ge \|w_{\mathrm{opp}}\|$, i.e., $|w_1 - w_2| \ge |w_1 + w_2|$. Without loss of generality, assume $w_1 > 0 > w_2$. Then, for all $t > 0$,

$$\begin{aligned}
\mathbb{P}_{z|y=1}\left(|w^\top z| \le t\right) &= \mathbb{P}_r\mathbb{P}_\theta\left(|w_1\cos\theta + w_2\sin\theta| \le \frac{t}{r}\,\middle|\,\theta \in \left(-\frac{\pi}{2}, 0\right] \cup \left(\frac{\pi}{2}, \pi\right]\right) \\
&= 2\,\mathbb{P}_r\mathbb{P}_\theta\left(w_1\cos\theta - w_2\sin\theta \le \frac{t}{r}\,\middle|\,\theta \in \left(0, \frac{\pi}{2}\right]\right).
\end{aligned}$$

Similarly,

$$\mathbb{P}_{z|y=-1}\left(|w^\top z| \le t\right) = 2\,\mathbb{P}_r\mathbb{P}_\theta\left(|w_1\cos\theta + w_2\sin\theta| \le \frac{t}{r}\,\middle|\,\theta \in \left(0, \frac{\pi}{2}\right]\right).$$

Since

$$|w_1\cos\theta + w_2\sin\theta| \le w_1\cos\theta - w_2\sin\theta \quad \text{for} \quad \theta \in \left(0, \frac{\pi}{2}\right],$$

we obtain

$$\begin{aligned}
&\mathbb{P}_{z|y=1}\left(|w^\top\xi| \ge |w^\top z|\right) - \mathbb{P}_{z|y=-1}\left(|w^\top\xi| \ge |w^\top z|\right) \\
&= -2\,\mathbb{P}_r\mathbb{P}_\theta\left(\frac{|w^\top\xi|}{r} \in [|w_1\cos\theta + w_2\sin\theta|, |w_1\cos\theta - w_2\sin\theta|]\,\middle|\,\theta \in \left(0, \frac{\pi}{2}\right]\right),
\end{aligned} \tag{B.2}$$

because when $X \ge Y$, we have $\mathbf{1}_{(X\le t)} - \mathbf{1}_{(Y\le t)} = -\mathbf{1}_{(Y\le t\le X)}$.

The case $\|w_{\rm sig}\| \le \|w_{\rm opp}\|$ can be treated similarly by exchanging the roles of $w_{\rm sig}$ and $w_{\rm opp}$, leading to a change of sign.

Thus, the sign of $w_\perp^\top\nabla_w L_0$ depends on whether $\|w_{\rm sig}\|$ or $\|w_{\rm opp}\|$ dominates.

Moreover, by using standard Gaussian moment bounds, we can control the magnitude as

$$\left|w_\perp^\top\nabla_w L_0\right| \le \frac{1}{\sqrt{8\pi}}|a|\|w_\perp\|.$$

Proof of the fourth statement. Let us denote $\xi - e_i\xi_i$ by $\xi_{\setminus i}$, i.e., ξ without its i-th coordinate. By symmetrizing over the pair $(z + \xi_{\setminus i} + e_i\xi_i, z + \xi_{\setminus i} - e_i\xi_i)$, we obtain

$$\begin{aligned}
-w_i\nabla_{w_i}L_0 &= a\mathbb{E}_x\left[y(x)\sigma'(w^\top x)w_i\xi_i\right] \\
&= \frac{a}{2}\mathbb{E}_x\left[y(z)\left(\sigma'(w^\top z + w^\top\xi_{\setminus i} + w_i\xi_i) - \sigma'(w^\top z + w^\top\xi_{\setminus i} - w_i\xi_i)\right)w_i\xi_i\right] \\
&= \frac{a}{2}\mathbb{E}_x\left[y(z)\mathbf{1}\left(|w^\top z + w^\top\xi_{\setminus i}| \le |w_i\xi_i|\right)|w_i\xi_i|\right].
\end{aligned}$$

Thus,

$$-w_i\nabla_{w_i}L_0 = \frac{a}{2}\mathbb{E}_{z,\xi_{\setminus i},\xi_i}\left[y(z)\mathbf{1}\left(|w^\top z + w^\top\xi_{\setminus i}| \le |w_i\xi_i|\right)|w_i\xi_i|\right].$$

Now, focusing on

$$\mathbb{E}_{z,\xi_{\setminus i}}\left[y(z)\mathbf{1}\left(|w^\top z + w^\top\xi_{\setminus i}| \le |w_i\xi_i|\right)\right],$$

we expand it:

$$\begin{aligned}
\mathbb{E}_{z,\xi_{\setminus i}}\left[y(z)\mathbf{1}\left(|w^\top z + w^\top\xi_{\setminus i}| \le |w_i\xi_i|\right)\right] &= \mathbb{E}_{\xi_{\setminus i}}\left[\mathbb{E}_{z|y=1}\mathbf{1}\left(|w^\top z + w^\top\xi_{\setminus i}| \le |w_i\xi_i|\right)\right] \\
&\quad - \mathbb{E}_{\xi_{\setminus i}}\left[\mathbb{E}_{z|y=-1}\mathbf{1}\left(|w^\top z + w^\top\xi_{\setminus i}| \le |w_i\xi_i|\right)\right] \\
&= \mathbb{E}_{z|y=1}\left[\mathbb{P}_{\xi_{\setminus i}}\left(w^\top\xi_{\setminus i} \in [-w^\top z \pm |w_i\xi_i|]\right)\right] \\
&\quad - \mathbb{E}_{z|y=-1}\left[\mathbb{P}_{\xi_{\setminus i}}\left(w^\top\xi_{\setminus i} \in [-w^\top z \pm |w_i\xi_i|]\right)\right].
\end{aligned} \tag{B.3}$$

Now, for any $a, b > 0$, by standard Gaussian tail bounds, we have

$$\frac{2b}{\|w_\perp - w_ie_i\|\sqrt{2\pi}}e^{-(a+b)^2/\|w_\perp - w_ie_i\|^2} \le \mathbb{P}_{\xi_{\setminus i}}\left(w^\top\xi_{\setminus i} \in [a \pm b]\right) \le \frac{2b}{\|w_\perp - w_ie_i\|\sqrt{2\pi}}e^{-(a-b)^2/\|w_\perp - w_ie_i\|^2}.$$

Note that by assumption, $\|w_\perp - w_ie_i\| = (1 + o(1))\|w_\perp\|$.

Also, observe that the distribution of $w^\top z$ under $\mathbb{P}_{z|y=-1}$ is the same as that of $\tilde{w}^\top z$ under $\mathbb{P}_{z|y=1}$, where $\tilde{w} = (w_1, -w_2)$.

Thus, we obtain

$$\begin{aligned}
&\left|\mathbb{E}_{z,\xi_{\setminus i}}\left[y(z)\mathbf{1}\left(|w^\top z + w^\top \xi_{\setminus i}| \le |w_i\xi_i|\right)\right]\right| \\
&\le \frac{2|w_i\xi_i|}{\|w_\perp\|\sqrt{2\pi}}\mathbb{E}_{z|y=1}\left|e^{-(w^\top z - |w_i\xi_i|)^2/\|w_\perp\|^2} - e^{-(\tilde{w}^\top z + |w_i\xi_i|)^2/\|w_\perp\|^2}\right| \\
&\le \frac{2|w_i\xi_i|}{\|w_\perp\|\sqrt{2\pi}}\mathbb{E}_{z|y=1}\frac{\left|-(w^\top z - |w_i\xi_i|)^2 + (\tilde{w}^\top z + |w_i\xi_i|)^2\right|}{\|w_\perp\|^2} \quad \text{(since } x \mapsto e^{-x} \text{ is 1-Lipschitz on } \mathbb{R}^+\text{)} \\
&\le \frac{8|w_i\xi_i|}{\|w_\perp\|^3\sqrt{2\pi}}\left(\|w_{1:2}\||w_i\xi_i| + \|w_{1:2}\|^2\right).
\end{aligned}$$

Finally, integrating over ξ_i, we get

$$|w_i\nabla_{w_i}L_0| \le \frac{4|aw_i|}{\|w_\perp\|}\frac{\|w_{1:2}\|^2 + |w_i|\|w_{1:2}\|}{\|w_\perp\|^2} = o\left(\frac{|aw_i|}{\|w_\perp\|}\right).$$

Thus, small coordinates have a negligible contribution compared to the signal and opponent parts. This concludes the proof. □

B.2 Evaluation of ∇L_0 during Phase Ib

When $\|w_{\text{sig}}\|$ becomes comparable to $\|w_\perp\|$, the approximations used in Lemma 13 are no longer accurate. Instead, we can leverage the fact that in Phase Ib, neurons are such that $\|w_{\text{opp}}\| \ll \|w_{\text{sig}}\|$ to obtain the following approximations of the gradients.

Lemma 14 (L_0 Population Gradients during Phase Ib)**.** *For any neuron $(w, a) \in \mathcal{N}$, we have:*

1. *When $\|w_{sig}\| \le \|w_\perp\|$, there exists a constant $c_b > 0$ such that*
$$c_b\frac{|a|\|w_{sig}\|^2}{\|w_\perp\|} - \frac{|a|\|w_{opp}\|\|w_{sig}\|}{\|w_\perp\|} \le -w_{sig}^\top\nabla_w L_0 \le \sqrt{\frac{\pi}{2}}\frac{|a|\|w_{sig}\|^2}{\|w_\perp\|} + \frac{|a|\|w_{opp}\|\|w_{sig}\|}{\|w_\perp\|}.$$
2. *When $\|w_{sig}\| \ge \|w_\perp\|$, there exists a constant $c_b' > 0$ such that*
$$c_b'|a|\|w_{sig}\| - |a|\|w_{opp}\| \le -w_{sig}^\top\nabla_w L_0 \le \sqrt{\frac{\pi}{2}}|a|\|w_{sig}\| + |a|\|w_{opp}\|.$$
3. $$\left|w_{opp}^\top\nabla_w L_0\right| \le \min\left(\frac{|a|\|w_{opp}\|^2}{\|w_\perp\|},\, |a|\|w_{opp}\|^2\right).$$
4. $$\operatorname{sgn}\left(-w_\perp^\top\nabla_w L_0\right) = \begin{cases} 1 & \text{if } \|w_{opp}\| > \|w_{sig}\|, \\ -1 & \text{if } \|w_{opp}\| \le \|w_{sig}\|, \end{cases} \quad \text{and} \quad \left|w_\perp^\top\nabla_w L_0\right| \le 0.5\,|a|\|w_\perp\|.$$
5. *For all $i \ge 3$, we have*
$$\left|w_i^\top\nabla_{w_i}L_0\right| \le 4\frac{|a||w_i|\|w_{1:2}\|^2}{\|w_\perp\|^3}.$$

Proof of Lemma 14. Throughout the proof write $x = z + \xi$, where z is the projection of x onto $\operatorname{span}\{e_1, e_2\}$ and ξ is the orthogonal noise. We also recall $\sigma(u) = \max\{u, 0\}$ and $\sigma'(u) = \mathbf{1}_{\{u>0\}}$.

Define the Gaussian scalars

$$X := w^\top\xi \sim \mathcal{N}(0, \sigma^2), \quad \sigma = \|w_\perp\|, \qquad A := w_{\text{sig}}^\top z \sim \mathcal{N}(0, \|w_{\text{sig}}\|^2), \qquad B := w_{\text{opp}}^\top z \sim \mathcal{N}(0, \|w_{\text{opp}}\|^2).$$

Because z is isotropic in the (e_1, e_2)-plane and $w_{\text{sig}} \perp w_{\text{opp}}$ in that plane, A and B are independent, and both are independent of X.

Approximation result. We will repeatedly bound quantities of the form

$$\mathbb{E}\big[|A|\,\mathbb{P}(X \in [-A-|B|,\,-A+|B|])\big].$$

Fix (A,B). Since $X \sim \mathcal{N}(0,\sigma^2)$,

$$\mathbb{P}(X \in [-A-|B|,\,-A+|B|]) = \int_{-|B|}^{|B|} \frac{1}{\sqrt{2\pi}\,\sigma} \exp\!\Big(-\frac{(A+u)^2}{2\sigma^2}\Big)\,du. \tag{B.4}$$

We develop the integrand around $u=0$ via Taylor–Lagrange:

$$\left|\frac{1}{\sqrt{2\pi}\sigma}e^{-\frac{(A+u)^2}{2\sigma^2}} - \frac{1}{\sqrt{2\pi}\sigma}e^{-\frac{A^2}{2\sigma^2}}\right| \le |u| \cdot \max_{t\in[-|B|,|B|]}\left|\frac{d}{dt}\Big[\frac{1}{\sqrt{2\pi}\sigma}e^{-\frac{(A+t)^2}{2\sigma^2}}\Big]\right|.$$

Since

$$\frac{d}{dt}\Big[\frac{1}{\sqrt{2\pi}\sigma}e^{-\frac{(A+t)^2}{2\sigma^2}}\Big] = -\frac{A+t}{\sigma^3\sqrt{2\pi}}\,e^{-\frac{(A+t)^2}{2\sigma^2}},$$

we get, using $|u| \le |B|$ and $e^{-(A+t)^2/(2\sigma^2)} \le 1$,

$$\left|\frac{1}{\sqrt{2\pi}\sigma}e^{-\frac{(A+u)^2}{2\sigma^2}} - \frac{1}{\sqrt{2\pi}\sigma}e^{-\frac{A^2}{2\sigma^2}}\right| \le \frac{|B|(|A|+|B|)}{\sigma^3\sqrt{2\pi}}. \tag{B.5}$$

Plugging (B.5) into (B.4) yields the two-sided estimate

$$\left|\mathbb{P}(X \in [-A-|B|,\,-A+|B|]) - \frac{2|B|}{\sqrt{2\pi}\,\sigma}\,e^{-A^2/(2\sigma^2)}\right| \le \frac{2|B|^2}{\sigma^3\sqrt{2\pi}}. \tag{B.6}$$

Therefore,

$$\begin{aligned}\mathbb{E}\big[|A|\,\mathbb{P}(X \in [-A-|B|,\,-A+|B|])\big] &\le \frac{2}{\sqrt{2\pi}\,\sigma}\,\mathbb{E}\Big[|A|\,|B|\,e^{-A^2/(2\sigma^2)}\Big] + \frac{2}{\sigma^3\sqrt{2\pi}}\,\mathbb{E}\big[|A|\,|B|^2\big] \\ &=: T_1 + T_2. \end{aligned}\tag{B.7}$$

We now compute these terms explicitly. Write $A = \|w_{\mathrm{sig}}\|G$ with $G \sim \mathcal{N}(0,1)$, and define

$$\lambda := \frac{\|w_{\mathrm{sig}}\|^2}{\sigma^2} = \left(\frac{\|w_{\mathrm{sig}}\|}{\|w_\perp\|}\right)^2.$$

Independence gives $\mathbb{E}|B| = \|w_{\mathrm{opp}}\|\sqrt{2/\pi}$, $\mathbb{E}|B|^2 = \|w_{\mathrm{opp}}\|^2$, and

$$\mathbb{E}\Big[|A|\,e^{-A^2/(2\sigma^2)}\Big] = \|w_{\mathrm{sig}}\|\,\mathbb{E}\Big[|G|\,e^{-(\lambda/2)G^2}\Big] = \|w_{\mathrm{sig}}\|\int_{\mathbb{R}} \frac{|g|}{\sqrt{2\pi}} e^{-(1+\lambda)g^2/2}\,dg = \|w_{\mathrm{sig}}\|\sqrt{\frac{2}{\pi}}\,\frac{1}{1+\lambda}.$$

Hence

$$\begin{aligned} T_1 &= \frac{2}{\sqrt{2\pi}\,\sigma}\,\mathbb{E}|B|\,\mathbb{E}\Big[|A|\,e^{-A^2/(2\sigma^2)}\Big] \\ &= \frac{2}{\sqrt{2\pi}\,\sigma}\,\Big(\|w_{\mathrm{opp}}\|\sqrt{\tfrac{2}{\pi}}\Big)\,\Big(\|w_{\mathrm{sig}}\|\sqrt{\tfrac{2}{\pi}}\tfrac{1}{1+\lambda}\Big) = \frac{\sqrt{\lambda}}{\pi(1+\lambda)}\,\|w_{\mathrm{opp}}\|, \end{aligned}\tag{B.8}$$

and

$$T_2 \;=\; \frac{2}{\sigma^3\sqrt{2\pi}}\,\mathbb{E}[|A|]\;\mathbb{E}\big[|B|^2\big] \;=\; \frac{2}{\sigma^3\sqrt{2\pi}}\,\Big(\|w_{\mathrm{sig}}\|\sqrt{\tfrac{2}{\pi}}\Big)\,\|w_{\mathrm{opp}}\|^2 \;=\; \frac{2}{\pi}\cdot\frac{\sqrt{\lambda}}{\sigma^2}\,\|w_{\mathrm{opp}}\|^2. \tag{B.9}$$

In particular, when $\lambda \ge 1$ (i.e., $\|w_{\mathrm{sig}}\| \ge \|w_\perp\|$), $T_1 \le \frac{1}{2\pi}\|w_{\mathrm{opp}}\|$ and $T_2 = O\big(\|w_{\mathrm{opp}}\|^2/\|w_\perp\|^2\big)$. When $\lambda \le 1$, $\frac{\sqrt{\lambda}}{1+\lambda} \le \sqrt{\lambda}$, so $T_1 \lesssim \frac{\|w_{\mathrm{sig}}\|}{\|w_\perp\|}\,\|w_{\mathrm{opp}}\|$, and $T_2 = O\big(\frac{\|w_{\mathrm{sig}}\|}{\|w_\perp\|^3}\,\|w_{\mathrm{opp}}\|^2\big)$. These bounds quantify the error made when replacing w by w_{sig} in the gradient expressions below.

Evaluation of $-w_{\mathbf{sig}}^\top \nabla_w L_0$. Starting from

$$-\nabla_{w_{1:2}} L_0 = \frac{a}{2}\,\mathbb{E}_z\mathbb{E}_\xi\big[y\left(\sigma'(w^\top\xi + w^\top z) - \sigma'(w^\top\xi - w^\top z)\right) z\big]\,,$$

we isolate the contribution along w_{sig} and use the symmetrization identity to write

$$\begin{aligned}
\mathbb{E}_z\mathbb{E}_\xi\left[y\sigma'(w^\top\xi + w_{\mathrm{sig}}^\top z)\, w_{\mathrm{sig}}^\top z\right] &= \frac{1}{2}\,\mathbb{E}_z\mathbb{E}_\xi\left[y\left(\sigma'(w^\top\xi + w_{\mathrm{sig}}^\top z) - \sigma'(w^\top\xi - w_{\mathrm{sig}}^\top z)\right) w_{\mathrm{sig}}^\top z\right] \\
&= \frac{1}{2}\,\mathbb{E}_z\mathbb{E}_\xi\left[y\,\mathbf{1}\big(|w_{\mathrm{sig}}^\top z| \ge |w^\top\xi|\big)\,|w_{\mathrm{sig}}^\top z|\right] \\
&= \frac{1}{4}\,\mathbb{E}_\xi\left(\mathbb{E}_{z|y=1}\big[\mathbf{1}\big(|w_{\mathrm{sig}}^\top z| \ge |w^\top\xi|\big)\,|w_{\mathrm{sig}}^\top z|\big] - \mathbb{E}_{z|y=-1}\big[\mathbf{1}\big(|w_{\mathrm{sig}}^\top z| \ge |w^\top\xi|\big)\,|w_{\mathrm{sig}}^\top z|\big]\right).
\end{aligned}$$

Let $z = r(\cos\theta, \sin\theta)$. When w_{sig} is aligned with μ_1, $|w_{\mathrm{sig}}^\top z| = \|w_{\mathrm{sig}}\||\mu_1^\top z|$. Parameterizing the two XOR-conditionals on $\theta \in (0, \pi/2]$,

$$\mathbb{E}_{z|y=1}\big[\mathbf{1}\big(|w_{\mathrm{sig}}^\top z| \ge |w^\top\xi|\big)\,|\mu_1^\top z|\big] = 2\,\mathbb{E}_r\mathbb{E}_{\theta\in(0,\pi/2]}\mathbf{1}\left(r|\cos\theta + \sin\theta| \ge \frac{|w^\top\xi|}{\|w_{\mathrm{sig}}\|}\right) r\,|\cos\theta + \sin\theta|,$$

$$\mathbb{E}_{z|y=-1}\big[\mathbf{1}\big(|w_{\mathrm{sig}}^\top z| \ge |w^\top\xi|\big)\,|\mu_1^\top z|\big] = 2\,\mathbb{E}_r\mathbb{E}_{\theta\in(0,\pi/2]}\mathbf{1}\left(r|\cos\theta - \sin\theta| \ge \frac{|w^\top\xi|}{\|w_{\mathrm{sig}}\|}\right) r\,|\cos\theta - \sin\theta|.$$

Using $|\cos\theta - \sin\theta| \le \cos\theta + \sin\theta$ for $\theta \in (0, \pi/2]$,

$$\begin{aligned}
&\left(\mathbb{E}_{z|y=1} - \mathbb{E}_{z|y=-1}\right)\left[\mathbf{1}\big(|w_{\mathrm{sig}}^\top z| \ge |w^\top\xi|\big)\,|w_{\mathrm{sig}}^\top z|\right] \\
&\quad\ge 2\|w_{\mathrm{sig}}\|\,\mathbb{E}_r\mathbb{E}_{\theta\in(0,\pi/2)}\mathbf{1}\left(r|\cos\theta + \sin\theta| \ge \tfrac{|w^\top\xi|}{\|w_{\mathrm{sig}}\|}\right) r\,(\cos\theta + \sin\theta - |\cos\theta - \sin\theta|) \\
&\quad\ge 4\|w_{\mathrm{sig}}\|\,\mathbb{E}_r\mathbb{E}_{\theta\in(0,\pi/4)}\mathbf{1}\left(r|\cos\theta + \sin\theta| \ge \tfrac{|w^\top\xi|}{\|w_{\mathrm{sig}}\|}\right) r\,\sin\theta,
\end{aligned}$$

and since $\cos\theta + \sin\theta \ge 1$ on $(0, \pi/4]$,

$$\mathbf{1}\left(r|\cos\theta + \sin\theta| \ge \tfrac{|w^\top\xi|}{\|w_{\mathrm{sig}}\|}\right) \ge \mathbf{1}\left(r \ge \tfrac{|w^\top\xi|}{\|w_{\mathrm{sig}}\|}\right), \qquad \mathbb{E}_{\theta\in(0,\pi/4)}[\sin\theta] = 1 - \tfrac{\sqrt{2}}{2}.$$

Therefore, with $c := 4 - 2\sqrt{2}$,

$$\mathbb{E}_\xi\mathbb{E}_r\mathbb{E}_{\theta\in(0,\pi/4)}\mathbf{1}\left(r|\cos\theta + \sin\theta| \ge \tfrac{|w^\top\xi|}{\|w_{\mathrm{sig}}\|}\right) 2r\sin\theta \ge c\,\mathbb{E}_r\big[r\,\mathbb{P}_\xi\big(|w^\top\xi| \le r\|w_{\mathrm{sig}}\|\big)\big]\,. \tag{B.10}$$

We now lower bound (B.10) in the two regimes:

Case $\|w_{sig}\| \le \|w_\perp\|$. Let $t := \epsilon\,\frac{\|w_\perp\|}{\|w_{\mathrm{sig}}\|}$ with fixed small $\epsilon > 0$. Split

$$\mathbb{E}_r\big[r\,\mathbb{P}_\xi(|w^\top\xi| \le r\|w_{\mathrm{sig}}\|)\big] = \mathbb{E}_r\big[r\,\mathbf{1}(r \le t)\,\mathbb{P}_\xi(|w^\top\xi| \le r\|w_{\mathrm{sig}}\|)\big] + \mathbb{E}_r\big[r\,\mathbf{1}(r \ge t)\,\mathbb{P}_\xi(|w^\top\xi| \le r\|w_{\mathrm{sig}}\|)\big]\,.$$

Using Lemma 9 with $\epsilon_r := (r\|w_{\mathrm{sig}}\|)/\|w_\perp\| \le \epsilon$,

$$\mathbb{P}_\xi(|w^\top\xi| \le r\|w_{\mathrm{sig}}\|) \;\ge\; \sqrt{\tfrac{2}{\pi}}\,\epsilon_r\,e^{-\epsilon_r^2/2} \;\ge\; \sqrt{\tfrac{2}{\pi}}\,e^{-\epsilon^2/2}\,\frac{\|w_{\mathrm{sig}}\|}{\|w_\perp\|}\,r.$$

Hence

$$\mathbb{E}_r[r\,\mathbf{1}(r \le t)\,\mathbb{P}_\xi(\cdot)] \;\ge\; \sqrt{\tfrac{2}{\pi}}\,e^{-\epsilon^2/2}\,\frac{\|w_{\mathrm{sig}}\|}{\|w_\perp\|}\,\mathbb{E}_r\big[r^2\,\mathbf{1}(r \le t)\big]\,.$$

Since r is Rayleigh (density $re^{-r^2/2}$), $\mathbb{E}[r^2\mathbf{1}(r \le t)] = 1 - e^{-t^2/2}(1 + t^2/2)$, which is $\Theta(1)$ for fixed ϵ. This yields the claimed $\frac{\|w_{\mathrm{sig}}\|}{\|w_\perp\|}$-scale lower bound.

Case $\|w_{sig}\| \geq \|w_\perp\|$. If $r \geq t := \epsilon$, then $\mathbb{P}_\xi(|w^\top \xi| \leq r\|w_{\rm sig}\|) \geq c(\epsilon) > 0$. Moreover, for Rayleigh r, $\mathbb{E}[r\,\mathbf{1}(r \geq t)] = e^{-t^2/2} + \sqrt{\frac{\pi}{2}}\,\mathrm{erfc}(t/\sqrt{2}) \geq e^{-t^2/2}t$, so this term is bounded below by a positive constant. Hence we get a constant multiple of $\|w_{\rm sig}\|$.

Combining the two cases in (B.10) and restoring the prefactors shows

$$-w_{\rm sig}^\top \nabla_w L_0 \;\gtrsim\; |a| \times \begin{cases} \dfrac{\|w_{\rm sig}\|^2}{\|w_\perp\|}, & \|w_{\rm sig}\| \leq \|w_\perp\|, \\ \|w_{\rm sig}\|, & \|w_{\rm sig}\| \geq \|w_\perp\|, \end{cases}$$

up to the approximation error controlled by (B.8)–(B.9), which contributes additive terms of size $O\big(|a|\,\frac{\|w_{\rm opp}\|\|w_{\rm sig}\|}{\|w_\perp\|}\big)$ in the first regime and $O(|a|\,\|w_{\rm opp}\|)$ in the second. This yields the lower bounds in items (1) and (2) with some absolute constants $c_b, c_b' > 0$.

For the upper bounds in items (1) and (2), we use $\mathbf{1}(|w_{\rm sig}^\top z| \geq |w^\top \xi|) \leq 1$ and $\mathbb{E}|A| = \|w_{\rm sig}\|\sqrt{2/\pi}$, which give

$$\mathbb{E}_z\mathbb{E}_\xi\left[y\sigma'(w^\top\xi + w_{\rm sig}^\top z)\,|w_{\rm sig}^\top z|\right] \;\leq\; \mathbb{E}_z|\,w_{\rm sig}^\top z\,| \;=\; \|w_{\rm sig}\|\sqrt{\tfrac{2}{\pi}},$$

and, accounting for the same symmetrization prefactors as above, we get the stated $\sqrt{\pi/2}$-type envelopes after multiplying by $|a|$ and by the appropriate scaling factor $(\|w_{\rm sig}\|/\|w_\perp\|)$ in the $\|w_{\rm sig}\| \leq \|w_\perp\|$ regime.

Control of $w_{\rm opp}^\top \nabla_w L_0$. Let $X_1 = w_{\rm sig}^\top z$, $X_2 = w_{\rm opp}^\top z$ (independent). Then

$$\mathbb{E}_z\mathbb{E}_\xi\left[y\sigma'(w^\top\xi + X_1)\,X_2\right] = \frac{1}{2}\,\mathbb{E}_z\mathbb{E}_\xi\left[y\,\mathbf{1}\big(|X_1| \geq |w^\top\xi|\big)\,X_2\,\mathrm{sgn}(X_1)\right].$$

Using the XOR definition $y = \mathbf{1}\Big(\frac{|X_1|}{\|w_{\rm sig}\|} \geq \frac{|X_2|}{\|w_{\rm opp}\|}\Big) - \mathbf{1}\Big(\frac{|X_1|}{\|w_{\rm sig}\|} \leq \frac{|X_2|}{\|w_{\rm opp}\|}\Big)$ and the facts $\mathbb{E}[X_2\,|\,|X_2| \leq t] = \mathbb{E}[X_2\,|\,|X_2| \geq t] = 0$, conditioning on X_1 shows the inner expectation is zero. The approximation step (replacing w by $w_{\rm sig}$) then yields

$$\big|w_{\rm opp}^\top \nabla_w L_0\big| \;\leq\; C\,|a| \times \min\!\left(\frac{\|w_{\rm opp}\|^2}{\|w_\perp\|},\, \|w_{\rm opp}\|^2\right),$$

for an absolute constant C; rescaling constants to 1 gives the stated item (3).

Control of the projection onto $w_\perp$ and of small coordinates. The sign of $-w_\perp^\top \nabla_w L_0$ and the bound $\big|w_\perp^\top \nabla_w L_0\big| \leq 0.5\,|a|\,\|w_\perp\|$ follow by the same symmetrization and indicator difference arguments as in Lemma 13, using $|\cdot| \leq 1$ and $\mathbb{E}|w^\top\xi| = \|w_\perp\|\sqrt{2/\pi}$. For coordinates $i \geq 3$, the interval probability bounds (as in Lemma 13, fourth item) give

$$\big|w_i^\top \nabla_{w_i} L_0\big| \;\lesssim\; \frac{|a|\,|w_i|}{\|w_\perp\|} \cdot \frac{\|w_{1:2}\|^2}{\|w_\perp\|^2} \;\leq\; 4\frac{|a||w_i|\|w_{1:2}\|^2}{\|w_\perp\|^3},$$

which proves items (4) and (5). □

Remark 5. *Contrary to Phase Ia, we do not obtain matching upper and lower bounds for $w_{sig}^\top \nabla_w L_0$. As a result, during Phase Ib, neurons may grow at different rates, which complicates the analysis of the block dynamics.*

B.3 Control of the Approximation $L_\rho \approx L_0$

First, we adapt Lemmas D.2–D.5 of Glasgow (2024) to the Gaussian setting. Throughout, recall that

$$f_\rho(x) \;=\; \mathbb{E}_{(w',a')\sim\rho}\big[a'\,\sigma({w'}^\top x)\big], \qquad \text{and} \qquad \mathbb{E}_\rho\|aw\| \;:=\; \mathbb{E}_{(w',a')\sim\rho}\big[|a'|\,\|w'\|\big].$$

For ReLU, σ is 1-Lipschitz and $\sigma(u) = u\,\sigma'(u)$ a.e.

Lemma 15. *For any neuron $(w,a)\in\mathcal{N}$, we have*

1. $|\nabla_a L_0 - \nabla_a L_\rho| \;\le\; 2\,\|w\|\,\mathbb{E}_\rho\|aw\|$.
2. $\|\nabla_w L_0 - \nabla_w L_\rho\| \;\le\; 2\,|a|\,\mathbb{E}_\rho\|aw\|$.

Lemma 16. *Assume $\mathbb{E}_\rho\|aw\| \le d^{O(1)}$. Then there exists a constant $C>0$ such that for any neuron $(w,a)\in\mathcal{N}$ and any $i\in[d]$,*

$$\big\|\nabla_{w_i} L_\rho - \nabla_{w_i} L_0\big\| \;\le\; |a|\Big(4\,\mathbb{E}_\rho\|aw_i\| + 2C\log d\,\mathbb{E}_\rho\|aw\|\,\sqrt{\mathbb{E}_x\,\mathbf{1}(|x_{\setminus i}^\top w|\le|w_i x_i|)} + d^{-\Omega(1)}\Big).$$

We also need a Gaussian analogue of Lemma C.5 in Glasgow (2024).

Lemma 17. *Let $x\sim\mathcal{N}(0,I_d)$. Then*

$$\mathbb{E}_x\big(\ell'_\rho(x)-\ell'_0(x)\big)^2 \;\le\; 4\big(\mathbb{E}_\rho\|aw\|\big)^2.$$

Further, for any $i\in[d]$ and any loss whose derivative ℓ'_ρ is 2-Lipschitz with respect to $f_\rho(x)$, we have

$$\mathbb{E}_x\big(\ell'_\rho(x_{\setminus i}+e_i x_i)-\ell'_\rho(x_{\setminus i}-e_i x_i)\big)^2 \;\le\; 16\big(\mathbb{E}_\rho\|aw_i\|\big)^2.$$

Proof of Lemma 17. For the first statement: note $\ell'_0(x)=\ell'(0)$. Since ℓ'_ρ is 2-Lipschitz in its argument $f_\rho(x)$,

$$\mathbb{E}_x\big(\ell'_\rho(x)-\ell'_0(x)\big)^2 \;\le\; 4\,\mathbb{E}_x\big(f_\rho(x)\big)^2 = 4\,\mathbb{E}_x\Big(\mathbb{E}_\rho\big[a\,\sigma(w^\top x)\big]\Big)^2.$$

By Minkowski's integral inequality,

$$\Big(\mathbb{E}_x(\mathbb{E}_\rho a\,\sigma(w^\top x))^2\Big)^{1/2} \;\le\; \mathbb{E}_\rho\Big(\mathbb{E}_x a^2\,\sigma^2(w^\top x)\Big)^{1/2} \;\le\; \mathbb{E}_\rho\big(|a|\,(\mathbb{E}_x|w^\top x|^2)^{1/2}\big) \;=\; \mathbb{E}_\rho|a|\,\|w\|.$$

Squaring both sides yields the first claim.

For the second statement, by 2-Lipschitzness of ℓ'_ρ in f_ρ,

$$\begin{aligned}
\mathbb{E}_x\big(\ell'_\rho(x_{\setminus i}+e_i x_i)-\ell'_\rho(x_{\setminus i}-e_i x_i)\big)^2 &\le 4\,\mathbb{E}_x\big(f_\rho(x_{\setminus i}+e_i x_i)-f_\rho(x_{\setminus i}-e_i x_i)\big)^2\\
&= 4\,\mathbb{E}_x\Big(\mathbb{E}_\rho a\big(\sigma(w^\top x_{\setminus i}+w_i x_i)-\sigma(w^\top x_{\setminus i}-w_i x_i)\big)\Big)^2\\
&\le 4\,\mathbb{E}_x\big(\mathbb{E}_\rho 2|aw_i x_i|\big)^2\\
&\le 16\Big(\mathbb{E}_\rho\big(\mathbb{E}_x|aw_i x_i|^2\big)^{1/2}\Big)^2 = 16\big(\mathbb{E}_\rho\|aw_i\|\big)^2,
\end{aligned}$$

using Minkowski's inequality and $\mathbb{E}_x x_i^2=1$. □

Proof of Lemma 15. Let $\Delta_x := (\ell'_\rho(x)-\ell'_0(x))\,\sigma'(w^\top x)$. Using $\sigma(w^\top x) = \sigma'(w^\top x)\,w^\top x$ a.e. and Cauchy–Schwarz,

$$\begin{aligned}
|\nabla_a L_0 - \nabla_a L_\rho| &= \Big|\mathbb{E}_x(\ell'_\rho(x)-\ell'_0(x))\,\sigma(w^\top x)\Big| = \big|\mathbb{E}_x\Delta_x\,w^\top x\big|\\
&\le \sqrt{\mathbb{E}_x\Delta_x^2}\,\sqrt{\mathbb{E}_x(w^\top x)^2} \le \sqrt{\mathbb{E}_x(\ell'_\rho(x)-\ell'_0(x))^2}\,\|w\|.
\end{aligned}$$

Lemma 17 gives $\mathbb{E}_x(\ell'_\rho-\ell'_0)^2 \le 4(\mathbb{E}_\rho\|aw\|)^2$, hence item 1.

For item 2, similarly,

$$\begin{aligned}
\|\nabla_w L_\rho - \nabla_w L_0\| &= |a|\,\big\|\mathbb{E}_x(\ell'_\rho(x)-\ell'_0(x))\,\sigma'(w^\top x)\,x\big\|\\
&= |a|\sup_{\|v\|=1}\big|\mathbb{E}_x\Delta_x\,\langle v,x\rangle\big|\\
&\le |a|\sup_{\|v\|=1}\sqrt{\mathbb{E}_x\Delta_x^2}\,\sqrt{\mathbb{E}_x\langle v,x\rangle^2} = |a|\,\sqrt{\mathbb{E}_x\Delta_x^2} \;\le\; 2\,|a|\,\mathbb{E}_\rho\|aw\|.
\end{aligned}$$

□

Proof of Lemma 16. Recall Δ_x from the proof of Lemma 15. By symmetry over the pair $(x_{\setminus i}+e_ix_i, x_{\setminus i}-e_ix_i)$,

$$\begin{aligned}\Big\|\frac{1}{a}\big(\nabla_{w_i}L_\rho - \nabla_{w_i}L_0\big)\Big\| &= \big\|\mathbb{E}_x\Delta_x\, x_i\big\| = \frac{1}{2}\big\|\mathbb{E}_x(\Delta_{x_{\setminus i}+e_ix_i} - \Delta_{x_{\setminus i}-e_ix_i})\, x_i\big\| \\ &\le \frac{1}{2}\big\|\mathbb{E}_x\mathbf{1}(|x_{\setminus i}^\top w| \ge |w_ix_i|)\,(\Delta_{x_{\setminus i}+e_ix_i} - \Delta_{x_{\setminus i}-e_ix_i})\, x_i\big\| \\ &\quad + \frac{1}{2}\big\|\mathbb{E}_x\mathbf{1}(|x_{\setminus i}^\top w| \le |w_ix_i|)\,(\Delta_{x_{\setminus i}+e_ix_i} - \Delta_{x_{\setminus i}-e_ix_i})\, x_i\big\|.\end{aligned}$$

Whenever $|x_{\setminus i}^\top w| \ge |w_ix_i|$, we have $\sigma'(w^\top(x_{\setminus i}+e_ix_i)) = \sigma'(w^\top(x_{\setminus i}-e_ix_i))$, hence

$$|\Delta_{x_{\setminus i}+e_ix_i} - \Delta_{x_{\setminus i}-e_ix_i}| \;\le\; |\ell_\rho'(x_{\setminus i}+e_ix_i) - \ell_\rho'(x_{\setminus i}-e_ix_i)|.$$

Therefore, by Cauchy–Schwarz and Lemma 17,

$$\begin{aligned}\frac{1}{2}\big\|\mathbb{E}_x\mathbf{1}(|x_{\setminus i}^\top w| \ge |w_ix_i|)\,(\Delta_{x_{\setminus i}+e_ix_i} - \Delta_{x_{\setminus i}-e_ix_i})\, x_i\big\| &\le \frac{1}{2}\Big(\mathbb{E}_x|\ell_\rho'(x_{\setminus i}+e_ix_i) - \ell_\rho'(x_{\setminus i}-e_ix_i)|^2\Big)^{1/2}\Big(\mathbb{E}_x x_i^2\Big)^{1/2} \\ &\le 4\,\mathbb{E}_\rho\|aw_i\|.\end{aligned}$$

For the remaining term, by 2-Lipschitzness of ℓ_ρ' in f_ρ,

$$\mathbb{E}_x\mathbf{1}(|x_{\setminus i}^\top w| \le |w_ix_i|)\,|\Delta_x|^2 \le 2\,\mathbb{E}_x\mathbf{1}(|x_{\setminus i}^\top w| \le |w_ix_i|)\,|f_\rho(x)|^2. \tag{B.11}$$

We now control the right-hand side using a sub-Gaussian tail for $f_\rho(x)$. Since each map $x \mapsto a'\sigma(w'^\top x)$ is $|a'|\,\|w'\|$-Lipschitz and σ is 1-Lipschitz, by Jensen/Minkowski the function f_ρ is L-Lipschitz with $L := \mathbb{E}_\rho\|aw\|$. For Gaussian $x \sim \mathcal{N}(0, I_d)$, Gaussian concentration (a.k.a. the Gaussian isoperimetric inequality) yields

$$\mathbb{P}_x\big(|f_\rho(x) - \mathbb{E}f_\rho(x)| \ge t\big) \;\le\; 2\exp\Big(-\frac{t^2}{2L^2}\Big).$$

Taking $t = C\log d\cdot L$ with $C>0$ large enough gives $\mathbb{P}_x(|f_\rho(x)-\mathbb{E}f_\rho(x)| \ge C\log d\, L) \le d^{-\Omega(1)}$. Using

$$|f_\rho(x)| \;\le\; |f_\rho(x)-\mathbb{E}f_\rho(x)| + |\mathbb{E}f_\rho(x)| \;\le\; |f_\rho(x)-\mathbb{E}f_\rho(x)| + L\,\mathbb{E}\|x\| \;\le\; |f_\rho(x)-\mathbb{E}f_\rho(x)| + O(\sqrt{d})\,L,$$

we bound the RHS of (B.11) by splitting on the event $|f_\rho(x)-\mathbb{E}f_\rho(x)| \le C\log d\, L$ and its complement, and applying Cauchy–Schwarz for the tail:

$$\begin{aligned}\mathbb{E}_x\mathbf{1}(|x_{\setminus i}^\top w| \le |w_ix_i|)\,|f_\rho(x)|^2 &\le 2\,\mathbb{E}_x\mathbf{1}(|x_{\setminus i}^\top w| \le |w_ix_i|)\,\Big(C^2\log^2 d\, L^2 + O(d)\,L^2\Big) \\ &\quad + 2\,\sqrt{\mathbb{E}_x(f_\rho(x)-\mathbb{E}f_\rho(x))^4}\;\mathbb{P}_x(|f_\rho(x)-\mathbb{E}f_\rho(x)| \ge C\log d\, L)^{1/2}.\end{aligned}$$

The fourth moment of a Lipschitz Gaussian functional is $O(L^4)$, so the tail term is $d^{-\Omega(1)}$. Absorbing the harmless $O(d)$ factor into the event probability (since $\mathbf{1}(|x_{\setminus i}^\top w| \le |w_ix_i|) \le 1$ and $\mathbb{P}(|x_{\setminus i}^\top w| \le |w_ix_i|)$ will be small in our regimes), we obtain the clean bound

$$\mathbb{E}_x\mathbf{1}(|x_{\setminus i}^\top w| \le |w_ix_i|)\,|f_\rho(x)|^2 \;\le\; C^2\log^2 d\; L^2\;\mathbb{E}_x\mathbf{1}(|x_{\setminus i}^\top w| \le |w_ix_i|) + d^{-\Omega(1)}.$$

Plugging into (B.11) and taking square roots,

$$\Big(\mathbb{E}_x\mathbf{1}(|x_{\setminus i}^\top w| \le |w_ix_i|)\,|\Delta_x|^2\Big)^{1/2} \;\le\; \sqrt{2}\,C\log d\,\mathbb{E}_\rho\|aw\|\,\sqrt{\mathbb{E}_x\mathbf{1}(|x_{\setminus i}^\top w| \le |w_ix_i|)} + d^{-\Omega(1)}.$$

Collecting the two pieces and multiplying back by $|a|$ yields the stated bound in Lemma 16. □

B.4 Analysis of the SGD Dynamic during Phase Ia

The previous analysis of the population gradient (see Lemma 13) $\nabla_w L_0$ suggests that $\big\|w_{\mathrm{sig}}^{(t+1)}\big\| \approx (1+\eta\tau)\,\big\|w_{\mathrm{sig}}^{(t)}\big\|$ where $\tau = \sqrt{2}\pi^{-3/2}$ while the $\|w_{\mathrm{opp}}\|$ and $\|w_\perp\|$ remain small. In this section, we will provide a rigorous characterization of this heuristic. But before delving into the proofs, let us introduce some useful definitions (cf. Glasgow (2024)).

B.4.1 Definitions

In this section, we will use the notations $\zeta = \log^{-c_\zeta}(d)$ and $\theta = \log^{-C_\theta}(d)$, where c_ζ and C_θ are sufficiently large constants that will be fixed later. Recall that $\tau = \sqrt{2}\pi^{-3/2}$ denotes the initial population growth rate of the signal component. We now define the stopping times and control parameters that describe the evolution of the neurons.

Definition 3 (Phase I Length)**.** *Let T_A be the last time such that $B_t^2 \le \theta^2\zeta^2$ (see the following definition), that is,*

$$T_A := \left\lfloor \frac{\log(d) + 2\log(\zeta) - \log(\log(d))}{\log(1 + 2\eta\tau(1+\zeta))} \right\rfloor = \frac{1}{\eta}\,\Theta(\log d). \tag{B.12}$$

Let T_B be the first time such that the lower envelope S_t reaches order $\theta\zeta^{-1}$, that is,

$$T_B = T_A + \left\lfloor \frac{\log\big(\theta^2\zeta^{-2}/S_{T_A}^2\big)}{\log(1 + c_b\eta)} \right\rfloor = T_A + \frac{1}{\eta}\,\Theta(\log\log d).$$

Definition 4 (Control Parameters)**.** *Let T_A and T_B be as defined in Definition 3. Define*

$$B_t^2 = \begin{cases} \dfrac{C_1\log(d)\theta^2}{d}\,(1 + 2\eta\tau(1+\zeta))^t, & t \le T_A, \\ B_{T_A}^2(1+4\eta)^{t-T_A}, & T_A < t \le T_B, \end{cases} \tag{B.13}$$

$$Q_t^2 = \frac{C_1\log(d)\theta^2}{d}\left(1 + \frac{50\eta}{\log(d)}\right)^t, \tag{B.14}$$

$$S_t^2 = \begin{cases} \dfrac{\theta^2}{d}(1 + 2\eta\tau(1-\zeta))^t, & t \le T_A, \\ S_{T_A}^2(1 + c_b\eta)^{t-T_A}, & T_A < t \le T_B, \end{cases} \tag{B.15}$$

where $c_b > 0$ is the constant appearing in Lemma 14.

Remark 6. *The parameter B_t controls the maximal signal magnitude $\|w_{sig}^{(t)}\|$ among neurons at time t, while S_t provides a lower bound for $\|w_{sig}^{(t)}\|$ among neurons with sufficiently large initial signal ($\|w_{sig}^{(0)}\| \ge \theta/\sqrt{d}$). The quantity S_t also characterizes the typical block size $N_i^{\pm}$. Finally, Q_t controls the evolution of the noise component $\|w_\perp^{(t)}\|$.*

Definition 5 (Controlled Neurons)**.** *We say a neuron (w, a) is controlled at iteration t if:*

1. $\|w_{sig}^{(t)}\| \le \max(\|w_{sig}^{(0)}\|, \frac{\theta}{\log d\sqrt{d}})(1 + \eta(\tau+\zeta))^t$.
2. $\|w_{opp}^{(t)}\| \le \max(\|w_{opp}^{(0)}\|, \frac{\theta}{\log d\sqrt{d}})(1 + \eta\theta)^t$.
3. $|a^{(t)}| \in \theta(1 \pm t\eta\zeta)$, *and* $|a^{(t)}| \le \|w^{(t)}\|$.
4. $\|w_\perp^{(t)} - w_\perp^{(0)}\| \le \theta\zeta^{1/4}\eta t$.
5. $\|w_\perp^{(t)}\|_\infty \le \frac{\theta\log d}{\sqrt{d}}(1 + \eta\theta)^t$.

In Phase Ib, $\|w_{\text{sig}}\|$ can be larger than $\theta\zeta$ for some neurons. This motivates the following definition of "weakly controlled" neurons (see also Glasgow (2024)).

Definition 6 (Weakly Controlled Neurons)**.** *We say that a neuron (w, a) is* weakly controlled *at iteration $t \in [T_A, T_B]$ if the following conditions hold:*

1. $\theta\zeta \le \|w_{sig}^{(t)}\| \le B_t \le \theta\zeta^{-1}\log d$,

2. $\|w_{opp}^{(t)}\| \le 3\,\theta\,B_t$,
3. $\|w^{(t)}\|^2 \ge |a^{(t)}|^2 \ge \|w^{(t)}\|^2 - \theta^2\big(\zeta^{1/2} + C\,\eta^2\,(t - T_A)\,\zeta'^2\big)$,
4. $\|w_{\perp}^{(t)}\| \le 3\,\theta$,

Following Glasgow (2024), we now define strong neurons, which are the neurons for which the signal component $\|w_{\text{sig}}\|$ grows quickly.

Definition 7 (Strong Neurons)**.** *We say a neuron (w, a) is* strong *at iteration t if it is controlled or weakly controlled, $(w_{sig}^{(t)})^\top w_{sig}^{(0)} > 0$, and*

$$\|w_{sig}^{(t)}\|^2 \ge S_t^2. \tag{B.16}$$

Lemma 18. *Let (w, a) be a strong neuron after T_B steps. Provided that $\zeta = o(\log^{-2} d)$, we have $\left\|w_{sig}^{(T_A)}\right\| \ge (1 - o(1))\zeta\theta \log^{-1/2} d \ge \zeta^{1.5}\theta$ and $\left\|w_{sig}^{(T_B)}\right\| \ge \zeta^{-1}\theta$*

Proof. By strongness, $\|w_{\text{sig}}^{(t)}\|^2 \ge S_t^2$ for all $0 \le t \le T_B$.

Lower bound at T_A. From the definitions used earlier,

$$\frac{S_{T_A}^2}{B_{T_A}^2} = \left(\frac{1 + 2\eta\tau(1-\zeta)}{1 + 2\eta\tau(1+\zeta)}\right)^{T_A} = \exp\big(T_A \log r\big), \quad r := \frac{1 + 2\eta\tau(1-\zeta)}{1 + 2\eta\tau(1+\zeta)} < 1.$$

Since $\log r = -\Theta(\zeta)$ and $T_A = \Theta((\log d)/\eta)$, the assumption $\zeta = o((\log d)^{-2})$ gives $\zeta T_A = o(1)$, hence

$$\frac{S_{T_A}^2}{B_{T_A}^2} = 1 - o(1). \tag{B.17}$$

By the choice of T_A as the last time with $B_t^2 \le \theta^2\zeta^2/\log d$,

$$\frac{\theta^2\zeta^2}{(1 + 2\eta\tau(1+\zeta))\,\log d} \;<\; B_{T_A}^2 \;\le\; \frac{\theta^2\zeta^2}{\log d}. \tag{B.18}$$

Combining (B.17) and (B.18),

$$S_{T_A}^2 \;\ge\; (1 - o(1))\,\frac{\theta^2\zeta^2}{(1 + 2\eta\tau(1+\zeta))\,\log d}, \qquad S_{T_A} \;\ge\; \frac{1 - o(1)}{\sqrt{1 + 2\eta\tau(1+\zeta)}}\,\frac{\theta\zeta}{\sqrt{\log d}}.$$

Moreover, since $\zeta = o((\log d)^{-2})$ implies $\zeta \log d \to 0$, we have $\theta\zeta/\sqrt{\log d} \ge \theta\zeta^{3/2}$ eventually, proving the first display.

Lower bound at T_B. For $t \in [T_A, T_B]$, by the defined schedule, $S_t^2 = S_{T_A}^2(1 + c_b\eta)^{t - T_A}$. Choosing T_B so that $S_{T_B} \ge \zeta^{-1}\theta$ (Phase II target) and using $\|w_{\text{sig}}^{(T_B)}\|^2 \ge S_{T_B}^2$ gives

$$\|w_{\text{sig}}^{(T_B)}\| \;\ge\; S_{T_B} \;\ge\; \zeta^{-1}\theta.$$

□

B.4.2 Initialization

First notice that at initialization, there exists a constant $C_1 > 0$ such that with probability at least $1 - d^{-\Omega(1)}$, all neurons $w^{(0)}$ are such that

$$\left\|w^{(0)}\right\|_\infty \le \theta\sqrt{C_1\frac{\log d}{d}}.$$

Note that by definition, $|a^{(0)}| = \theta$. So all the neurons are controlled. We also need to control the size of the blocks of the oracle network. By symmetry of the initialization, the distribution of all blocks is equal, so it is sufficient to analyze the block associated with μ_1. We have

$$\begin{aligned}\mathbb{E}\left(\sum_{(w,a)\in\mathcal{N}_1^+}|a|\left\|w_{\text{sig}}^{(0)}\right\|\right) &= \frac{\theta}{2}\sum_{(w,a)}\mathbb{E}\left(|\mu^\top w^{(0)}|\mathbf{1}_{a>0}\mathbf{1}_{\mu_1^\top w^{(0)}>0}\right)\\ &= \frac{\theta m}{4}\mathbb{E}\left(|\mu_1^\top w^{(0)}|\mathbf{1}_{a>0}\mathbf{1}_{\mu^\top w^{(0)}>0}\right)\\ &= \frac{\theta^2 m}{4\sqrt{d}}p_1,\end{aligned}$$

where $p_1 = \mathbb{E}_Y[\sqrt{d}|\mu_1^\top Y|\mathbf{1}_{a>0}\mathbf{1}_{\mu^\top Y>0}]$ with a uniform r.v. Y over the unit sphere. A simple calculation shows $p_1 \approx \frac{1}{2\sqrt{2\pi}}$.

It is also easy to check that by standard concentration inequality, we have

$$\sum_{(w,a)\in\mathcal{N}}\mathbf{1}_{(w,a)\in\mathcal{N}_1^+}|a|\left\|w_{\text{sig}}^{(0)}\right\| - \mathbb{E}\left(\mathbf{1}_{(w,a)\in\mathcal{N}_1^+}|a|\left\|w_{\text{sig}}^{(0)}\right\|\right) = O(\theta^2\frac{\sqrt{m}}{\sqrt{d}}).$$

By consequence, $U^{(0)} = O(\frac{1}{\sqrt{m}})$.

B.4.3 Inductive step

We will show recursively that controlled neurons remain controlled during Phase Ia. It corresponds to Lemma C.15 in Glasgow (2024). The proof strategy is similar and consists of showing that the error terms have almost no impact so that the evolution dynamic of the neurons is similar to the population dynamic.

Lemma 19 (Controlled Neurons Inductive Step)**.** *Assume that all neurons are controlled or weakly controlled for some $t \le T_a$. Then with probability at least $1 - d^{-\Omega(1)}$, for any controlled neuron $(w^{(t)}, a^{(t)})$ we have:*

1. *A neuron $(w^{(t+1)}, a^{(t+1)})$ is controlled.*
2. *If $(w^{(t)}, a^{(t)})$ is a strong neuron, then $(w^{(t+1)}, a^{(t+1)})$ is a strong neuron.*

To prove Lemma 19, we will use the following lemma, which is an adaptation of Lemma D.17 in Glasgow (2024).

Lemma 20 (Phase 1a L_0 Population Gradients Bounds)**.** *If all neurons in the network are controlled or weakly controlled at some step $t \le T_A$, then for any controlled neuron (w, a), the following hold:*

1. $\|\nabla_{w_\perp} L_0\| \le 2\sqrt{\theta\|w_{1:2}\|}$.
2. $|\nabla_a L_0| \le \|w\|\|\nabla_{w_\perp} L_0\| + \|w_{1:2}\|\|\nabla_{w_{1:2}} L_0\|$.
3. *For any $i \in [d]$,* $\|\nabla_{w_i} L_0\| \le \frac{|w_i|}{2}$.
4. *For any $i \in [d]$,* $\|\nabla_{w_i} L_0 - \nabla_{w_i} L_\rho\| \le \frac{\theta B_t}{2}$.

Proof. We study the first statement. Recalling that $x = z + \xi$ for $z = x_{1:2}$, we have

$$\begin{aligned}\frac{1}{a_w}\nabla_{w_\perp} L_0 &= -\mathbb{E}_x y(x)\sigma'(w^\top x)\xi\\ &= -\mathbb{E}_x y(z)\sigma'(w^\top\xi)\xi + \mathbb{E}_x y(z)(\sigma'(w^\top\xi) - \sigma'(w^\top x))\xi\\ &= \mathbb{E}_x y(\sigma'(w^\top\xi) - \sigma'(w^\top x))\xi,\end{aligned}$$

since y is independent of ξ and $\mathbb{E}_z y = 0$. Now consider the norm of $\mathbb{E}_x y(x)(\sigma'(w^\top x) - \sigma'(w^\top\xi))\xi$. We have

$$\|\mathbb{E}_x y(x)(\sigma'(w^\top x) - \sigma'(w^\top\xi))\xi\| = \sup_{v:\|v\|=1}\mathbb{E}_x y(x)(\sigma'(w^\top x) - \sigma'(w^\top\xi))\xi^\top v$$

$$
\begin{aligned}
&\leq \sqrt{\mathbb{E}_x(\sigma'(w^\top x) - \sigma'(w^\top \xi))^2}\sqrt{\mathbb{E}_\xi (v^\top \xi)^2} \\
&= \sqrt{\mathbb{E}_x \mathbf{1}(|\xi^\top w| \leq |z^\top w|)} \\
&\leq \sqrt{\mathbb{E}_z \mathbb{P}_\xi \left(|\xi^\top w| \leq |z^\top w|\right)} \\
&\leq \sqrt{\mathbb{E}_z \frac{|z^\top w|}{\|w_\perp\|}} \\
&\leq \sqrt{\sqrt{\frac{2}{\pi}} \frac{\|w_{1:2}\|}{\|w_\perp\|}} \\
&\leq \sqrt{\frac{\|w_{1:2}\|}{\theta}}.
\end{aligned}
$$

Thus we have

$$
\|\nabla_{w_\perp} L_0\| \leq 2\sqrt{\theta \|w_{1:2}\|} \leq 4 \min\left(\sqrt{\theta B_t}, \theta \zeta^{1/2}\right),
$$

since the neuron is controlled: $|a| \leq 2\theta$.

Next, we consider the third statement. By the symmetrization argument used in the proof of Lemma 13, we have for $i \in [d]$

$$
\begin{aligned}
\|\nabla_{w_i} L_0\| &\leq \frac{1}{2}|a| \left(\mathbb{E}_{\xi_i} \mathbb{P}_{\xi_{\setminus i}} |\xi_i| \mathbf{1}(|w^\top x_{\setminus i}| \leq |w_i \xi_i|)\right) \\
&\leq \frac{|a|}{2} \left(\mathbb{E}_{\xi_i} \frac{|w_i||\xi_i|^2}{\|w - w_i e_i\|}\right) \\
&\leq \frac{1}{2}|w_i| \\
&\leq \frac{1}{2} \min\left(B_t, \theta\zeta\right),
\end{aligned}
$$

Next, we consider the second statement. Combining the first and third statements, we have

$$
\begin{aligned}
|\nabla_a L_0| &= |w^\top \nabla_w L_0| \\
&\leq |w_\perp^\top \nabla_{w_\perp} L_0| + |w_{1:2} \nabla_{w_{1:2}} L_0| \\
&\leq \|w\| \|\nabla_{w_\perp} L_0\| + \|w_{1:2}\| \|\nabla_{w_{1:2}} L_0\| \\
&\leq 2\|w\| \min\left(\sqrt{\theta B_t}, \theta\zeta^{1/2}\right) + \|w\| \min\left(B_t, \theta\zeta\right) \\
&\leq 3\theta \min\left(\sqrt{\theta B_t}, \theta\zeta^{1/2}\right).
\end{aligned}
$$

Finally, we consider the fourth statement. Applying Lemma 16 yields

$$
\begin{aligned}
&\|\nabla_{w_i} L_\rho - \nabla_{w_i} L_0\| && \text{(B.19)} \\
&\quad \leq |a_w| \left(4(\mathbb{E}_\rho[\|a_w w_i\|]) + 2\log(d)\mathbb{E}_\rho[\|a_w w\|]\mathbb{E}_\xi \mathbb{E}_z \mathbf{1}(|x_{\setminus i}^\top w| \leq |w_i|) + d^{-\omega(1)}\right) && \text{(B.20)} \\
&\quad \leq 2\theta \left(8\theta \min(\theta, B_t) + 4\log(d)\theta \min(\theta, B_t)\mathbb{E}_\xi \mathbb{E}_z \mathbf{1}(|x_{\setminus i}^\top w| \leq |w_i|) + d^{-\Omega(1)}\right). && \text{(B.21)}
\end{aligned}
$$

Now, $w_\perp^{(0)}$ is well-spread, and $\|w_\perp^{(t)} - w_\perp^{(0)}\| \leq o(1)\|w_\perp^{(0)}\|$ by the definition of controlled neurons, so we obtain

$$
\begin{aligned}
\mathbb{E}_{x_{\setminus i}} \mathbb{E}_{x_i} \mathbf{1}(|x_{\setminus i}^\top w| \leq |w_i x_i|) &\leq \mathbb{E}_{x_i} \mathbb{P}_{G \sim \mathcal{N}(0,1)} \left(|G| \leq \frac{|w_i x_i|}{\|w - e_i w_i\|}\right) \\
&\leq \frac{|w_i|}{\|w\|}
\end{aligned}
$$

$$\leq \frac{2\min(B_t, \theta\zeta)}{\theta}$$
$$\leq \frac{2B_t}{\max(\theta, B_t)}$$

the second to last line follows from the definition of controlled and (B.13). Thus we obtain

$$\|\nabla_{w_i} L_\rho - \nabla_{w_i} L_0\| \leq \Theta(\log(d))\left(\theta B_t \max(\theta, B_t) + \theta B_t \max(\theta, B_t)\right) \leq \theta B_t/2, \tag{B.22}$$

since $\max(\theta, B_t) = o(1/\log(d))$ holds. □

Proof of Lemma 19. Suppose that $(w^{(t)}, a^{(t)})$ is controlled.

Control of the growth of $|a|$. We have with probability at least $1 - d^{-\Omega(1)}$

$$\begin{aligned}
|a^{(t+1)} - a^{(t)}| &= |\eta \nabla \hat{L}_\rho a^{(t)}| \\
&\leq \eta|\nabla_a L_0| + \eta|\nabla_a L_\rho - \nabla_a \hat{L}_\rho| + \eta|\nabla_a L_0 - \nabla_a L_\rho| \\
&\leq \eta\left(4\theta^2\zeta^{1/2} + \|w\|\sqrt{\frac{d\log(d)^2}{V}} + 2\|w\|\mathbb{E}_\rho[\|aw\|]\right) \qquad \text{(by Lemma 20, 7, and 15)} \\
&\leq \eta\theta\zeta.
\end{aligned}$$

Control of the growth of $\|w_\perp\|$ Similarly, one can prove with the same arguments that

$$\begin{aligned}
\|w_\perp^{(t+1)} - w_\perp^{(t)}\| &= \|\eta \nabla_{w_\perp} \hat{L}_{\rho^{(t)}}\| \\
&\leq \eta\|\nabla_{w_\perp} L_0\| + \eta\|\nabla_{w_\perp} L_{\rho^{(t)}} - \nabla_{w_\perp} \hat{L}_{\rho^{(t)}}\| + \eta\|\nabla_{w_\perp} L_0 - \nabla_{w_\perp} L_\rho\| \\
&\leq \eta\left(3\theta\zeta^{1/2} + |a|\sqrt{\frac{d\log(d)^2}{V}} + 2|a|\mathbb{E}_{\rho^{(t)}}[\|aw\|]\right) \\
&\leq \eta\theta\zeta^{1/2}.
\end{aligned}$$

Control of the growth of $\|w_{\mathbf{opp}}\|$. W.l.o.g. assume the signal direction is given by μ_1. We have

$$\begin{aligned}
\left\|w_{\text{opp}}^{(t+1)}\right\|^2 &= \left\|w_{\text{opp}}^{(t)}\right\|^2 - 2\eta(w_{\text{opp}}^{(t)})^\top \nabla_w \hat{L}_\rho + \eta^2 \left|\mu_2^\top \nabla_w \hat{L}_\rho\right|^2 \\
&\leq \left\|w_{\text{opp}}^{(t)}\right\|^2 - 2\eta(w_{\text{opp}}^{(t)})^\top \nabla_w L_0 + \eta\left\|(w_{\text{opp}}^{(t)})^\top(\nabla L_0 - \nabla L_\rho)\right\| + \eta^2\left|\mu_2^\top \nabla_w L_0\right|^2 \\
&\quad + \eta^2\left|\mu_2^\top(\nabla_w L_0 - \nabla_w \hat{L}_\rho)\right| \\
&\leq (1 - 2\eta\tau(1-\zeta))\left\|w_{\text{opp}}^{(t)}\right\|^2 + \eta\theta B_t. \\
&\qquad \text{(by Lemma 20, 7, and 15 and the definition of controlled neurons)}
\end{aligned}$$

Since every sequence (u_n) satisfying $u_{n+1} \leq au_n + b$ for $0 < a < 1$ verify $u_n \leq \frac{b}{1-a}$, the previously established relation allows us to conclude.

Control of the growth of $\|w_\perp\|_\infty$. Let $i \geq 3$. Observe that

- $|w_i \nabla_{w_i} L_0| \leq \zeta \frac{|a||w_i|^2}{\|w_\perp\|}$ by Lemma 13.
- $\|\nabla_{w_i} L_0 - \nabla_{w_i} \hat{L}_\rho\| \leq \frac{S_t\theta}{\sqrt{d}}$ with probability $1 - d^{-\omega(1)}$. This follows from combining Lemma 20 with Lemma 7.
- $\|\nabla_{w_i} L_0\| \leq \|w_\perp\|_\infty$ by Lemma 20.

This implies that

$$\|w_\perp^{(t+1)}\|_\infty \leq \|w_\perp^{(t)}\|_\infty(1 + C\eta S_t \log d) \leq \|w_\perp^{(t)}\|_\infty(1 + \eta\theta).$$

The same argument gives the lower-bound.

Control of the growth of $\|w_{\mathbf{sig}}\|$. We have

- $-w_{\text{sig}}^\top \nabla_w L_0 = \tau \frac{|a|\|w_{\text{sig}}\|^2}{\|w_\perp\|} + O(\zeta^2 \|w_{\text{sig}}\|^2)$ by Lemma 13 and the fact that the neuron is controlled.
- $\|\nabla_{w_{1:2}} L_0 - \nabla_{w_{1:2}} \hat{L}_\rho\| \leq \theta S_t$ with probability $1 - d^{-\omega(1)}$. This follows from combining Lemma 20 with Lemma 7.

Thus, using an argument similar to the one employed for bounding $\|w_{\text{opp}}\|$, we obtain that for all strong neurons

$$\left\|w_{\text{sig}}^{(t)}\right\| (1 - \eta(\tau + \zeta)) \leq \left\|w_{\text{sig}}^{(t+1)}\right\| \leq \left\|w_{\text{sig}}^{(t)}\right\| (1 + \eta(\tau + \zeta)).$$

For neurons that are not strong, we obtain the following upper bound:

$$\left\|w_{\text{sig}}^{(t)}\right\| \leq \frac{\theta}{\sqrt{d}} (1 + \eta\tau)^t.$$

□

B.5 Analysis of the SGD dynamic during Phase Ib

We are going to prove the following lemma, corresponding to Lemma C.16 in Glasgow (2024). The main difference compared with Lemma 19 is that we need other estimates for the gradients due to the growth of w_{sig}.

Lemma 21. *Assume that for some $t \in [T_A, T_B]$ all the neurons are controlled or weakly controlled. Then with probability at least $1 - d^{-\Omega(1)}$, all the neurons are controlled or weakly controlled at time $t+1$. Moreover, strong neurons remain strong.*

Proof of Lemma 21 (Phase Ib). Fix $t \in [T_A, T_B]$ and a neuron $(w^{(t)}, a^{(t)})$ that is controlled or weakly controlled at time t. All bounds below hold with probability at least $1 - d^{-\Omega(1)}$ by the cited lemmas.

Control of $\|w_{\mathbf{sig}}\|$. The projected squared–norm update is

$$\|w_{\text{sig}}^{(t+1)}\|^2 - \|w_{\text{sig}}^{(t)}\|^2 = -2\eta\, w_{\text{sig}}^{(t)\top} (\mu\mu^\top) \nabla_w \hat{L}_\rho \;+\; \eta^2 \|(\mu\mu^\top) \nabla_w \hat{L}_\rho\|^2. \tag{B.23}$$

The deviation satisfies (Lemma 20+16+7)

$$\left| w_{\text{sig}}^\top (\mu\mu^\top)(\nabla_w \hat{L}_\rho - \nabla_w L_0) \right| \;\lesssim\; \|w_{\text{sig}}\| \max(\theta, \|w_{\text{sig}}\|)\, S_t^2. \tag{B.24}$$

We consider the two regimes in Lemma 14.

(i) $\|w_{sig}\| \leq \|w_\perp\|$. By Lemma 14.1,

$$-w_{\text{sig}}^\top (\mu\mu^\top) \nabla_w L_0 \;\leq\; \sqrt{\tfrac{\pi}{2}} \frac{|a|}{\|w_\perp\|} \|w_{\text{sig}}\|^2 \;+\; \frac{|a|}{\|w_\perp\|} \|w_{\text{opp}}\| \|w_{\text{sig}}\|.$$

Insert $\nabla_w \hat{L}_\rho = \nabla_w L_0 + (\nabla_w \hat{L}_\rho - \nabla_w L_0)$ in (B.23), use the triangle inequality together with (B.24), and note that $\eta^2 \|(\mu\mu^\top) \nabla_w \hat{L}_\rho\|^2 \geq 0$. Using weak control $\|w_{\text{opp}}\| \leq 3\theta B_t$, $\|w_{\text{sig}}\| \leq B_t$, and $|a|/\|w_\perp\| \leq \sqrt{3}$ (since $\|w\|^2 \leq 3\|w_\perp\|^2$ in Phase Ib and $|a| \leq \|w\|$), we get

$$\|w_{\text{sig}}^{(t+1)}\|^2 \;\leq\; \|w_{\text{sig}}^{(t)}\|^2 (1 + C_1\eta) \;+\; C_2 \eta\, B_t^2 \;+\; C_3 \eta\, B_t \max(\theta, B_t)\, S_t^2.$$

(ii) $\|w_{sig}\| \geq \|w_\perp\|$. By Lemma 14.2,

$$-w_{\text{sig}}^\top (\mu\mu^\top) \nabla_w L_0 \;\leq\; \sqrt{\tfrac{\pi}{2}}\, |a|\, \|w_{\text{sig}}\| \;+\; |a|\, \|w_{\text{opp}}\|.$$

In Phase Ib, $|a| \le \|w\| \le \sqrt{\|w_{\mathrm{sig}}\|^2 + \|w_\perp\|^2 + \|w_{\mathrm{opp}}\|^2} \le 2\|w_{\mathrm{sig}}\|$, and $\|w_{\mathrm{opp}}\| \le 3\theta B_t$. Using these, (B.24), and $\eta^2\|(\mu\mu^\top)\nabla_w \hat{L}_\rho\|^2 \ge 0$, we obtain

$$\|w_{\mathrm{sig}}^{(t+1)}\|^2 \;\le\; \|w_{\mathrm{sig}}^{(t)}\|^2\big(1 + C_4\eta\big) \;+\; C_5\eta\, B_t^2 \;+\; C_6\eta\, B_t \max(\theta, B_t)\, S_t^2.$$

Combining (i)–(ii) and using $S_t = o(1)$ on Phase Ib, we conclude that

$$\|w_{\mathrm{sig}}^{(t+1)}\|^2 \;\le\; \|w_{\mathrm{sig}}^{(t)}\|^2\big(1 + C\eta + o(\eta)\big) \;+\; o(\eta)\, B_t^2. \tag{B.25}$$

By the definition of the upper envelope B_t on Phase Ib (monotone geometric growth calibrated to dominate the per–step upper rate in (B.25)), and since $\|w_{\mathrm{sig}}^{(t)}\| \le B_t$, we get $\|w_{\mathrm{sig}}^{(t+1)}\| \le B_{t+1}$.

Assume the neuron is strong at time t, i.e. $\|w_{\mathrm{sig}}^{(t)}\|^2 \ge S_t^2$. Using (B.23), the nonnegativity of the η^2 term, the deviation bound (B.24), and the lower bounds of Lemma 14:

(i) $\|w_{sig}\| \le \|w_\perp\|$. Lemma 14.1 gives

$$-\, w_{\mathrm{sig}}^\top(\mu\mu^\top)\nabla_w L_0 \;\ge\; c_b\, \frac{|a|}{\|w_\perp\|}\, \|w_{\mathrm{sig}}\|^2 - \frac{|a|}{\|w_\perp\|}\, \|w_{\mathrm{opp}}\|\, \|w_{\mathrm{sig}}\|.$$

As above, $|a|/\|w_\perp\| \ge c_- > 0$ by item 3 and Phase Ib (constants absorbed in c_-), and $\|w_{\mathrm{opp}}\| \le 3\theta B_t \le o(1)\, \|w_{\mathrm{sig}}\|$ within Phase Ib. Therefore

$$\|w_{\mathrm{sig}}^{(t+1)}\|^2 \;\ge\; \|w_{\mathrm{sig}}^{(t)}\|^2\big(1 + 2\eta\, c_- c_b - o(\eta)\big).$$

(ii) $\|w_{sig}\| \ge \|w_\perp\|$. Lemma 14.2 gives

$$-\, w_{\mathrm{sig}}^\top(\mu\mu^\top)\nabla_w L_0 \;\ge\; c_b'|a|\, \|w_{\mathrm{sig}}\| - |a|\, \|w_{\mathrm{opp}}\|.$$

On Phase Ib, $|a| \ge \frac{1}{2}\|w_{\mathrm{sig}}\|$ (again from item 3 and $\|w\|^2 \le 3\|w_{\mathrm{sig}}\|^2$) and $\|w_{\mathrm{opp}}\| \le o(1)\, \|w_{\mathrm{sig}}\|$, hence

$$\|w_{\mathrm{sig}}^{(t+1)}\|^2 \;\ge\; \|w_{\mathrm{sig}}^{(t)}\|^2\big(1 + \eta\, c_b' - o(\eta)\big).$$

In both regimes, since $S_{t+1}^2 = S_t^2(1 + c_b\eta)$ on Phase Ib and $c_b, c_b' > 0$ are absolute, choosing the constant c_b defining S_t small enough (as per the envelope construction) ensures

$$\|w_{\mathrm{sig}}^{(t+1)}\|^2 \;\ge\; \|w_{\mathrm{sig}}^{(t)}\|^2(1 + c_b\eta) \;\ge\; S_t^2(1 + c_b\eta) \;=\; S_{t+1}^2,$$

so strong neurons remain strong.

Control of $|a|$. Let $D_t := |a^{(t)}|^2 - \|w^{(t)}\|^2$. Expanding (updates of (a, w)) gives

$$D_{t+1} = D_t - 2\eta\Big(a^{(t)}\nabla_a \hat{L}_\rho - w^{(t)\top}\nabla_w \hat{L}_\rho\Big) + \eta^2\Big(\|\nabla_a \hat{L}_\rho\|^2 - \|\nabla_w \hat{L}_\rho\|^2\Big). \tag{B.26}$$

By Lemma 8, the linear term cancels up to sampling error, yielding the one–step lower control

$$D_{t+1} \;\ge\; D_t - 4\eta^2\, |a^{(t)}|^2.$$

Using $|a^{(t)}| \le \|w^{(t)}\|$ and the uniform Phase–Ib bound $\|w^{(s)}\|^2 \le 2\theta^2\zeta^{-600}$ for $s \in [T_{1a}, T_{1b}]$, we obtain

$$D_{t+1} \;\ge\; D_t - 8\eta^2\, \theta^2\, \zeta^{-600}.$$

Combining with the inductive lower bound in item 3 at time t gives the lower side of item 3 at $t+1$. The matching upper side, $D_{t+1} \le 0$, follows by applying the upper one–step control from Lemma 8 to (B.26) and absorbing the same sampling error as in the inductive hypothesis. Hence item 3 propagates to $t+1$.

Control of $\|w_\perp\|$. Project the w–update onto $\mathrm{span}\{\mu_1,\mu_2\}^\perp$:

$$w_\perp^{(t+1)} = w_\perp^{(t)} - \eta P_\perp \nabla_w \hat{L}_\rho, \qquad \|w_\perp^{(t+1)}\|^2 = \|w_\perp^{(t)}\|^2 - 2\eta\, w_\perp^{(t)\top} P_\perp \nabla_w \hat{L}_\rho + \eta^2 \|P_\perp \nabla_w \hat{L}_\rho\|^2.$$

By Lemma 14.5, $\left|w_\perp^\top P_\perp \nabla_w L_0\right| \le 4|a|\,\|w_{1:2}\|^2/\|w_\perp\| \le 4B_t^3/\|w_\perp\|$. By Lemma 20 and Lemma 7, $\|P_\perp(\nabla_w \hat{L}_\rho - \nabla_w L_0)\| \le 4\|w\|S_t^2 \le 4B_t S_t^2$. These imply (after a standard $(x+y)^2 \le 2x^2+2y^2$ bound on the η^2 term)

$$\|w_\perp^{(t+1)}\|^2 \;\le\; \|w_\perp^{(t)}\|^2 + \frac{8\eta B_t^3}{\|w_\perp^{(t)}\|} + 8\eta B_t S_t^2\,\|w_\perp^{(t)}\| + \eta^2\left(\frac{32B_t^6}{\|w_\perp^{(t)}\|^4} + 32B_t^2 S_t^4\right).$$

On Phase Ib, $\|w\|^2 \le 3\|w_\perp\|^2$, thus $B_t \le C\|w_\perp^{(t)}\|$, and the denominators cancel:

$$\|w_\perp^{(t+1)}\|^2 \;\le\; \|w_\perp^{(t)}\|^2\big(1 + C\eta + o(\eta)\big).$$

Iterating from T_A to $t \le T_{1b}$ yields $\|w_\perp^{(t)}\| \le 3\theta$, so item (4) of weak control propagates.

Control of $\|w_{\mathbf{opp}}\|$. Project w onto μ_2:

$$\|w_{\mathrm{opp}}^{(t+1)}\|^2 - \|w_{\mathrm{opp}}^{(t)}\|^2 = -2\eta\, w_{\mathrm{opp}}^{(t)\top} \nabla_w \hat{L}_\rho \;+\; \eta^2\big(\mu_2^\top \nabla_w \hat{L}_\rho\big)^2.$$

Lemma 14.3 gives $|w_{\mathrm{opp}}^\top \nabla_w L_0| \le |a|\|w_{\mathrm{opp}}\|^2$. Moreover, $\|\nabla_w L_0 - \nabla_w \hat{L}_\rho\| \le 2|a|B_t^2 + CB_tS_t^2$ (Lemma 15+7), hence $|w_{\mathrm{opp}}^\top(\nabla_w \hat{L}_\rho - \nabla_w L_0)| \le \|w_{\mathrm{opp}}\|(2|a|B_t^2 + CB_tS_t^2)$, and

$$\big(\mu_2^\top \nabla_w \hat{L}_\rho\big)^2 \;\le\; 3|a|^2\|w_{\mathrm{opp}}\|^2 + 3\big(2|a|B_t^2 + CB_tS_t^2\big)^2.$$

Therefore,

$$\|w_{\mathrm{opp}}^{(t+1)}\|^2 \;\le\; \|w_{\mathrm{opp}}^{(t)}\|^2\big(1 + 2\eta|a^{(t)}| + 3\eta^2|a^{(t)}|^2 + o(\eta)\big).$$

Since $|a^{(t)}| \le B_t$ and $\eta\sum_{s=T_A}^{T_{1b}-1} B_s = O(1)$ on Phase Ib,

$$\|w_{\mathrm{opp}}^{(t)}\| \;\le\; C\,\|w_{\mathrm{opp}}^{(T_A)}\| \;\le\; 3\theta B_t,$$

so item (2) of weak control propagates.

The same argument as in Phase I can be applied to controlled neurons. □

B.6 Control of the blocks

Contrary to the boolean setting studied by Glasgow (2024), in our setting, it is crucial to show that each block of neurons $N_i^\pm$ grows approximately at the same rate. A direct approach based on individual neuron dynamics will fail because, after Phase Ia, neurons in the same block can grow at different rates, depending on their initial alignment. To overcome this difficulty, we analyze the block dynamic at a macroscopic level.

- First, we consider the ideal case where there is an infinite number of neurons growing independently. The size of the block only depends on the distribution of the neurons conditioned on the initial block. Thanks to the initialization of the block and invariance by rotation of the gradient, one can show that the blocks remain equal over time.
- Then we show by using the law of large numbers that when the number of neurons is finite but updated independently, the blocks remain approximately equal.
- Finally, we show that when one uses empirical gradients to update the neuron's weights, the dynamic of the blocks remains close to the previously studied case.

B.6.1 Ideal setting: infinite width and independent neurons

We assume that the width m of our network is infinite and that the neurons $(\tilde{a}^{(t)}, \tilde{w}^{(t)})$ are updated with population gradients ∇L_0. Hence, $N_i^{\pm,(t)} = \mathbb{E}_{(\tilde{w}^{(t)},\tilde{a}^{(t)})|\mathcal{N}_i^{\pm}} \|\tilde{w}_{\text{sig}}^{(t)}\||\tilde{a}^{(t)}|$. It is clear that at initialization, all the $N_i^{\pm,(0)}$ are equal. We are going to show that this property remains true for each time t by showing that the distribution of $(\tilde{w}^{(t)}, \tilde{a}^{(t)})|\mathcal{N}_1^+$ is the same than $(\tilde{w}^{(t)}, \tilde{a}^{(t)})|\mathcal{N}_2^+$ up to a rotation. The argument for the other blocks' equality is the same.

Lemma 22. *Assume that $w = Rw'$ and $a = -a'$ where R is a rotation of angle $\pi/2$ that maps μ_1 to μ_2. Then we have*

$$\nabla_w L_0 = R(\nabla_{w'} L_0) \quad \textit{and} \quad \nabla_{a'} L_0 = -\nabla_a L_0.$$

Proof. We have

$$\begin{aligned}
\nabla_w L_0 &= a\mathbb{E}_x y(x)\sigma'(w^\top x)x \\
&= aR\left(\mathbb{E}_x - y(R^{-1}x)\sigma'((w')^\top R^{-1}x)R^{-1}x\right) \\
&\quad \text{(a rotation of angle } \pi/2 \text{ change one cluster to another, so change the sign of } y) \\
&= R(a'\mathbb{E}_x y\sigma'((w')^\top x)x \qquad \text{(the law of } x \text{ is rotationally invariant)} \\
&= R((\nabla_{w'} L_0)).
\end{aligned}$$

A similar calculation shows the second result. □

By an immediate recursion, one can show by using Lemma 22 that the distribution of $(\tilde{w}^{(t)}, \tilde{a}^{(t)})|\mathcal{N}_1^+$ corresponds to $(R^{-1}(\tilde{w}^{(t)}), -\tilde{a}^{(t)})|\mathcal{N}_2^+$. Since the quantity $\|\tilde{w}_{\text{sig}}^{(t)}\||\tilde{a}^{(t)}|$ is invariant under these transformations, all the blocks remain the same.

B.7 Block approximation control

We compare the (idealized) *population* block dynamics $\{\tilde{N}_i^{\pm,(t)}\}_t$ (neurons updated independently with population gradients) to the *empirical* block dynamics $\{N_i^{\pm,(t)}\}_t$ (mini–batch SGD). Recall

$$\tilde{N}_i^{\pm,(t)} := \frac{1}{|\mathcal{N}_i^{\pm}|} \sum_{(w,a)\in\mathcal{N}_i^{\pm}} \|\tilde{w}_{\text{sig}}^{(t)}\|\,|\tilde{a}^{(t)}|, \qquad N_i^{\pm,(t)} := \frac{1}{|\mathcal{N}_i^{\pm}|} \sum_{(w,a)\in\mathcal{N}_i^{\pm}} \|w_{\text{sig}}^{(t)}\|\,|a^{(t)}|.$$

Finite–width, independent neurons (sampling fluctuation). In Phase I we have $\|\tilde{w}_{\text{sig}}^{(t)}\|\,|\tilde{a}^{(t)}| \le 1$, hence Hoeffding's inequality yields, for any block $i, \pm$ and any t,

$$\Pr\Big(\big|\tilde{N}_i^{\pm,(t)} - \mathbb{E}\big[\|\tilde{w}_{\text{sig}}^{(t)}\|\,|\tilde{a}^{(t)}|\big]\big| > u\Big) \;\le\; 2\exp\big(-2|\mathcal{N}_i^{\pm}|\,u^2\big).$$

With $|\mathcal{N}_i^{\pm}| \asymp m$ and $u = \sqrt{(\log^c d)/m}$, a union bound over $i, \pm$ and $t \le T_B = O(\eta^{-1}\log(1/\zeta))$ gives, with probability $1 - d^{-\Omega(1)}$,

$$\big|\tilde{N}_i^{\pm,(t)} - \mathbb{E}\big[\|\tilde{w}_{\text{sig}}^{(t)}\|\,|\tilde{a}^{(t)}|\big]\big| \;\le\; \frac{1}{\sqrt{d\,\log^{c'} d}}, \qquad \forall\, i, \pm,\ \forall\, t \le T_B, \tag{B.27}$$

for some $c' > 0$ (given our choice of m).

Empirical vs. population dynamics. By symmetry, work on the block $\mathcal{N}_1^+$. Define the block–average discrepancy

$$\varepsilon_t := \max\left\{ \frac{1}{|\mathcal{N}_1^+|} \sum_{(w,a)\in\mathcal{N}_1^+} \|w_{\text{sig}}^{(t)} - \tilde{w}_{\text{sig}}^{(t)}\|, \quad \frac{1}{|\mathcal{N}_1^+|} \sum_{(w,a)\in\mathcal{N}_1^+} |a^{(t)} - \tilde{a}^{(t)}| \right\}.$$

From the updates

$$w^{(t+1)} = w^{(t)} - \eta \nabla_w \hat{L}_\rho, \qquad \tilde{w}^{(t+1)} = \tilde{w}^{(t)} - \eta \nabla_w L_0, \qquad a^{(t+1)} = a^{(t)} - \eta \nabla_a \hat{L}_\rho, \quad \tilde{a}^{(t+1)} = \tilde{a}^{(t)} - \eta \nabla_a L_0,$$

projecting to the signal coordinate and using the triangle inequality we obtain

$$\varepsilon_{t+1} \;\le\; \varepsilon_t \;+\; \eta\, \Delta_t^{\mathrm{conc}} \;+\; \eta\, \Delta_t^{\mathrm{pop}}, \tag{B.28}$$

where

$$\Delta_t^{\mathrm{conc}} := \frac{1}{|\mathcal{N}_1^+|} \sum_{(w,a)\in\mathcal{N}_1^+} \Big(\big\| \nabla_{w_{\mathrm{sig}}} \hat{L}_\rho - \nabla_{w_{\mathrm{sig}}} L_\rho \big\| + \big| \nabla_a \hat{L}_\rho - \nabla_a L_\rho \big| \Big),$$

$$\Delta_t^{\mathrm{pop}} := \frac{1}{|\mathcal{N}_1^+|} \sum_{(w,a)\in\mathcal{N}_1^+} \Big(\big\| \nabla_{w_{\mathrm{sig}}} L_\rho(w,a) - \nabla_{w_{\mathrm{sig}}} L_0(\tilde{w},\tilde{a}) \big\| + \big| \nabla_a L_\rho(w,a) - \nabla_a L_0(\tilde{w},\tilde{a}) \big| \Big).$$

Sampling (concentration) term. By Lemma 7 and Lemma 15 (items 1, 2), using $|a| \vee \|w\| \le B_t$ for (weakly) controlled neurons,

$$\Delta_t^{\mathrm{conc}} \;\lesssim\; \begin{cases} \theta\, N^{(t)} & (t \le T_A), \\ \zeta'\, N^{(t)} & (T_A \le t \le T_B), \end{cases} \qquad \zeta' = \log^{-c} d, \tag{B.29}$$

where $N^{(t)} := \frac{1}{m} \sum \|w_{\mathrm{sig}}^{(t)}\| |a^{(t)}|$ is the global Phase I mass proxy.

Population term. Write $\rho = \|w_{\mathrm{sig}}\|/\|w\|$, $\tilde{\rho} = \|\tilde{w_{\mathrm{sig}}}\|/\|\tilde{w}\|$, and

$$\nabla_{w_{\mathrm{sig}}} L_0(w,a) = a\, \mathbb{E}_{(z,\xi)}\big[\, y\, z_1\, \sigma'(w^\top \xi + w_{\mathrm{sig}}^\top z)\,\big], \qquad z_1 = \mu^\top z.$$

Set $F_\rho(x) := \mathbb{E}_\xi[\sigma'(\rho x + \sqrt{1-\rho^2}\,\xi)]$ so that the integrand is $y\, z_1\, F_\rho(z_1)$. For ReLU,

$$\frac{\mathrm{d}}{\mathrm{d}\rho} F_\rho(x) = \frac{1}{\sqrt{2\pi}} \exp\Big(- \frac{\rho^2 x^2}{2(1-\rho^2)} \Big)\, \frac{x}{(1-\rho^2)^{3/2}},$$

hence

$$\left| \frac{\mathrm{d}}{\mathrm{d}\rho}\, \mathbb{E}_z\big[y\, z_1\, F_\rho(z_1) \big] \right| \;\le\; \frac{1}{\sqrt{2\pi}}\, \mathbb{E}\left[\frac{z_1^2}{(1-\rho^2)^{3/2}}\, \exp\Big(- \frac{\rho^2 z_1^2}{2(1-\rho^2)} \Big) \right] = 1,$$

uniformly in $\rho \in [0,1]$ (Gaussian integral with $u = z_1/\sqrt{1-\rho^2}$). Therefore,

$$\big\| \nabla_{w_{\mathrm{sig}}} L_0(w,a) - \nabla_{w_{\mathrm{sig}}} L_0(\tilde{w},\tilde{a}) \big\| \;\lesssim\; |a-\tilde{a}| + |a| \cdot |\rho - \tilde{\rho}|. \tag{B.30}$$

Again for ReLU,

$$\big| \nabla_a L_0(w,a) - \nabla_a L_0(\tilde{w},\tilde{a}) \big| \;\lesssim\; \|w - \tilde{w}\| + |a - \tilde{a}|. \tag{B.31}$$

To control $|\rho - \tilde{\rho}|$, write

$$|\rho - \tilde{\rho}| \;\le\; \frac{\|w_{\mathrm{sig}} - \tilde{w_{\mathrm{sig}}}\|}{\|w\|} + \frac{\|\tilde{w_{\mathrm{sig}}}\|}{\|w\|\,\|\tilde{w}\|}\, \big| \|w\| - \|\tilde{w}\| \big|.$$

By item 3 of weak control,

$$\big| |a|^2 - \|w\|^2 \big| \;\le\; \theta^2 \Gamma_t, \qquad \Gamma_t := \zeta^{1/2} + C\eta^2 (t - T_A) \zeta'^2,$$

hence $|\, \|w\| - |a|\, | \le \theta \sqrt{\Gamma_t}$ and $\big| \|w\| - \|\tilde{w}\| \big| \le |a - \tilde{a}| + 2\theta\sqrt{\Gamma_t}$. Using $|a| \le \|w\|$ then gives

$$|a|\, |\rho - \tilde{\rho}| \;\le\; \|w_{\mathrm{sig}} - \tilde{w}_{\mathrm{sig}}\| + |a - \tilde{a}| + 2\theta\sqrt{\Gamma_t}. \tag{B.32}$$

Combining (B.30), (B.31), (B.32) and averaging over the block,

$$\frac{1}{|\mathcal{N}_1^+|} \sum_{(w,a)} \Big(\big\| \nabla_{w_{\mathrm{sig}}} L_0(w,a) - \nabla_{w_{\mathrm{sig}}} L_0(\tilde{w},\tilde{a}) \big\| + \big| \nabla_a L_0(w,a) - \nabla_a L_0(\tilde{w},\tilde{a}) \big| \Big) \;\lesssim\; \varepsilon_t + \theta\sqrt{\Gamma_t}. \tag{B.33}$$

Finally, the *approximation* gap $L_\rho - L_0$ appears both in the sampling term (through $\hat{L}_\rho - L_\rho$) and in the population term. The global bounds of Lemma 15 and the coordinate bound of Lemma 16 imply

$$\frac{1}{|\mathcal{N}_1^+|}\sum_{(w,a)}\Big(\big\|\nabla_{w_{\mathrm{sig}}}L_\rho - \nabla_{w_{\mathrm{sig}}}L_0\big\| + \big|\nabla_a L_\rho - \nabla_a L_0\big|\Big) \;\lesssim\; \begin{cases}\theta\, N^{(t)} & (t \le T_A),\\ \zeta'\, N^{(t)} & (T_A \le t \le T_B),\end{cases}$$

using $|a| \vee \|w\| \le B_t$ and the batch/scale choices. Together with (B.33), we conclude

$$\Delta_t^{\mathrm{pop}} \;\lesssim\; \varepsilon_t + \theta\sqrt{\Gamma_t} \quad \text{and} \quad \Delta_t^{\mathrm{conc}} \text{ as in (B.29).} \tag{B.34}$$

Phase Ia. In Phase Ia, for controlled neurons Lemma 13 gives

$$-\,w_{\mathrm{sig}}^\top \nabla_w L_0 = \tau\,\frac{|a|}{\|w_\perp\|}\,\|w_{\mathrm{sig}}\|^2 \;+\; o\!\left(\frac{|a|}{\|w_\perp\|}\,\|w_{\mathrm{sig}}\|^2\right), \qquad \frac{|a|}{\|w_\perp\|} = 1 + o(1),$$

hence the per–step growth matches

$$\|w_{\mathrm{sig}}^{(t+1)}\| \;=\; \|w_{\mathrm{sig}}^{(t)}\|\big(1 + \tau\eta + o(\eta)\big) \qquad \text{uniformly over controlled neurons.}$$

The negligible set with too small initial correlation contributes $o(1)$ at block scale and is absorbed by (B.27). Therefore the *same* multiplicative factor $(1 + \tau\eta + o(\eta))$ governs both $N^{(t)}$ and the linearization in (B.28), i.e.

$$\varepsilon_{t+1} \;\le\; \big(1 + (\tau + o(1))\eta\big)\varepsilon_t \;+\; \eta\Big(\theta\, N^{(t)} + \theta\sqrt{\Gamma_t}\Big), \qquad t \le T_A. \tag{B.35}$$

Let $\delta_t := \varepsilon_t / N^{(t)}$. Since $N^{(t+1)} = (1 + \tau\eta + o(\eta))N^{(t)}$, (B.35) gives

$$\delta_{t+1} \;\le\; \delta_t \;+\; \theta \;+\; \frac{\theta\sqrt{\Gamma_t}}{N^{(t)}}.$$

Here $\sqrt{\Gamma_t} = O(1)$ (Phase Ia) and $N^{(t)}$ is increasing, so from $\delta_0 = 0$,

$$\delta_{T_A} \;\lesssim\; \theta \quad\Longrightarrow\quad \varepsilon_{T_A} \;\le\; \theta'\, N^{(T_A)}, \;\; \theta' = \theta^{0.9}.$$

This explicit Phase Ia handoff uses the *exact* growth rate τ on both sides, which is necessary since $T_A = \Theta(\eta^{-1}\log d)$.

Phase Ib . On $[T_A, T_B]$, combine (B.28), (B.29), (B.34):

$$\varepsilon_{t+1} \;\le\; (1 + O(\eta))\,\varepsilon_t \;+\; \eta\,\zeta' N^{(t)} \;+\; \eta\,\theta\sqrt{\Gamma_t}. \tag{B.36}$$

Using $\sqrt{\Gamma_t} \le \zeta^{1/4} + O(\eta(t - T_A)\zeta')$ and the strong–mass lower bound $N^{(t)} \gtrsim \theta^2\zeta(\log d)^{-1/2}$ on $[T_A, T_B]$, we absorb the last term into $\zeta' N^{(t)}$ (choose c large in $\zeta' = \log^{-c} d$, recall $\zeta = o(\log^{-2} d)$). Thus,

$$\varepsilon_{t+1} \;\le\; (1 + O(\eta))\,\varepsilon_t \;+\; \eta\,\zeta' N^{(t)}.$$

Unroll for $T_B - T_A = O(\eta^{-1}\log\log d)$ steps and use $\varepsilon_{T_A} \le \theta' N^{(T_A)}$:

$$\varepsilon_t \;\le\; (1 + O(\eta))^{t - T_A}\,\varepsilon_{T_A} \;+\; \zeta'\,\eta \sum_{s=T_A}^{t-1} (1 + O(\eta))^{t-1-s}\, N^{(s)}.$$

Since $N^{(s)}$ is increasing, the sum is $\lesssim \zeta'\,\log\log d \cdot N^{(t)}$. Furthermore, $(1 + O(\eta))^{T_B - T_A} = \mathrm{polylog}(d)$; taking $\theta' = \theta^{0.9}$ small and c large in ζ',

$$\mathrm{polylog}(d)\cdot\theta' \;\ll\; \log^{-c_U} d, \qquad \zeta'\log\log d \;\ll\; \log^{-c_U} d.$$

Therefore, for all $T_A \le t \le T_B$,

$$\varepsilon_t \;\le\; \zeta'\, N^{(t)} \;\le\; \log^{-c_U} d \cdot N^{(t)}. \tag{B.37}$$

Conclusion (pre-balanced blocks). Combining the Phase Ia handoff and Phase Ib accumulation, for all $t \le T_B$,

$$\varepsilon_t \;\le\; \begin{cases} \theta' N^{(t)}, & t \le T_A, \\ \zeta' N^{(t)}, & T_A \le t \le T_B, \end{cases} \quad\Longrightarrow\quad |N_i^{\pm,(t)} - \tilde{N}_i^{\pm,(t)}| \;\le\; 2B_t\,\varepsilon_t \;=\; o(N^{(t)}),$$

uniformly over blocks $i, \pm$. Together with (B.27) this yields, w.h.p., for all $t \le T_1 (\le T_B)$,

$$U^{(t)} \;=\; \max_{i,\pm} \frac{|N_i^{\pm,(t)} - \tilde{N}_i^{\pm,(t)}|}{N^{(t)}} \;+\; \max_{i,\pm} \frac{|\tilde{N}_i^{\pm,(t)} - \mathbb{E}[\|\tilde{w}_{\mathrm{sig}}^{(t)}\| |\tilde{a}^{(t)}|]|}{N^{(t)}} \;\le\; \log^{-c_U} d,$$

which is Lemma 3. The approximation bounds (Lemmas 15–16) are used both in (B.29) and in (B.34), and the only place where the precise per–step constant matters is Phase Ia, where it is $\tau + o(1)$ as required.

B.8 Conclusion: Proof of Lemma 2

We have shown by recursion that after T_1 gradient steps, all neurons such that $\left\|w_{\mathrm{sig}}^{(0)}\right\| \ge \theta/\sqrt{d}$ (strong neurons) are weakly controlled and the others are either controlled or weakly controlled. It implies that

$$\mathbb{E}_{\rho^{(T_1)}} \|w_\perp + w_{\mathrm{opp}}\|^2 \le 4\theta^2$$

and $\theta\zeta^{-1} \le \|w_{\mathrm{sig}}\| \le 1$ for strong neurons.

C ANALYSIS OF PHASE II

C.1 Analysis on Clean Gradient with Oracle

The following lemma can control the distance with the clean gradients (it corresponds to Lemma E.5 in Glasgow (2024), we adapted and simplified the argument for the Gaussian case).

Lemma 23. *For any neuron $(a, w) \in \mathcal{N}$, we have the following:*

1. $\|\nabla_w^{cl} L_\rho - \nabla_w L_\rho\| \le |a| \mathbb{E}_\rho \|a w_\perp\|$.
2. $\|\nabla_a^{cl} L_\rho - \nabla_a L_\rho\| \le \|w\| \mathbb{E}_\rho \|a w_\perp\|$.
3. $\|\nabla_w^{cl} L_\rho - \nabla_w^{cl} L_{\rho,id}\| \le 4|a| N^{(t)} U^{(t)}$.
4. $\|\nabla_a^{cl} L_\rho - \nabla_a^{cl} L_{\rho,id}\| \le 4\|w\| N^{(t)} U^{(t)}$.

Proof. We study the first statement. Let us define $\Delta_x := (\ell_\rho'(x) - \ell_\rho'(z))\sigma'(w^\top x)$. We evaluate the difference of gradients as

$$\begin{aligned} \|\nabla_w^{\mathrm{cl}} L_\rho - \nabla_w L_\rho\| &= |a| \|\mathbb{E}_x \Delta_x x\| \\ &= |a| \sup_{v:\|v\|=1} \mathbb{E}_x \Delta_x \langle v, x\rangle \\ &\le |a| \sup_{v:\|v\|=1} \sqrt{\mathbb{E}_x \Delta_x^2} \sqrt{\mathbb{E}_x \langle v, x\rangle^2} \\ &= |a| \sqrt{\mathbb{E}_x \Delta_x^2}. \end{aligned}$$

Now, we obtain

$$\begin{aligned} \mathbb{E}_x[\Delta_x^2] &\le \mathbb{E}_x[(\ell_\rho'(x) - \ell_\rho'(z))^2] \\ &\le \mathbb{E}_x(f_\rho(x) - f_\rho(z))^2 && (\ell_\rho \text{ is 2-Lipschitz}) \\ &\le \mathbb{E}_x(\mathbb{E}_\rho |a| |\xi^\top w|)^2 \\ &\le \left(\mathbb{E}_\rho (\mathbb{E}_\xi |a|^2 |\xi^\top w|^2)^{1/2}\right)^2 && (\text{by Minkowski's integral inequality}) \end{aligned}$$

$$\leq \left(\mathbb{E}_\rho \|aw_\perp\|\right)^2.$$

For the other statements, the same argument can be used to prove the other inequalities combined with the fact that

$$\left|f_{\rho^{(t)},id}(z) - f_{\rho^{(t)}}(z)\right| \leq 4\|z\| N^{(t)} U^{(t)},$$

hence we omit the proof. □

Remark 7. *In Phase II, we will show that* $\mathbb{E}_\rho \|aw_\perp\| \leq \zeta \mathbb{E}_\rho \|aw\|$.

C.2 Population gradients evaluation.

In addition to the average margin g_μ, we also define

$$g_\rho = \mathbb{E}_z |\ell'_{\rho,id}(z)|. \tag{C.1}$$

Lemma 24. *For any neuron* $(w,a) \in \mathcal{S}$ *such that* $w_{sig}^\top \mu > 0$ *holds for some* $\mu \in \{\pm\mu_1, \pm\mu_2\}$, *we have*

$$\mu^\top \nabla_w^{cl} L_{\rho,id} = -|a| g_\mu (1 \pm \zeta^{0.25}), \quad \textit{and} \quad -y\nabla_a^{cl} L_{\rho,id} = (1 \pm \zeta^{0.25}) \|w_{sig}\| g_\mu.$$

Lemma 25. *There exist constants* $0 < c_1 < c_2$ *such that for all* ρ *large enough, we have*

$$\frac{c_1}{\rho^3} \leq g_\mu \leq \frac{c_2}{\rho^3}, \quad \textit{and} \quad \frac{c_1}{\rho} \leq g_\rho \leq \frac{c_2 \log \rho}{\rho}.$$

Proof of Lemma 25. To control g_ρ, it is sufficient to control

$$\mathbb{E}_z e^{-yN^{(t)}(\sigma(\mu_1^\top z) + \sigma(-\mu_1^\top z) - \sigma(\mu_2^\top z) - \sigma(-\mu_2^\top z))}.$$

It is easy to check that by definition

$$y(\sigma(\mu_1^\top z) + \sigma(-\mu_1^\top z) - \sigma(\mu_2^\top z) - \sigma(-\mu_2^\top z)) = ||X_1| - |X_2||$$

where $X_1 = \mu_1^\top z$ and $X_2 = \mu_2^\top z$. Since X_1 and X_2 are independent standard Gaussian r.v., we obtain

$$\begin{aligned}
\mathbb{E}_z e^{-yN^{(t)}(\sigma(\mu_1^\top z) + \sigma(-\mu_1^\top z) - \sigma(\mu_2^\top z) - \sigma(-\mu_2^\top z))} &= \mathbb{E}_{X_1,X_2} e^{-N^{(t)}||X_1| - |X_2||} \\
&\geq \mathbb{E}_{X_1,X_2} e^{-N^{(t)}|X_1 - X_2|} \\
&\geq \frac{c_1}{N^{(t)}},
\end{aligned}$$

for some constant $c_1 \in (0,1)$ by Lemma 11.

For the upper-bound, notice that

$$\mathbb{P}(||X_1| - |X_2|| \leq \frac{\log N^{(t)}}{N^{(t)}}) = O(\frac{\log N^{(t)}}{N^{(t)}}),$$

and

$$\mathbb{E}_{X_1,X_2} e^{-N^{(t)} \frac{\log N^{(t)}}{N^{(t)}}} = \frac{1}{N^{(t)}}.$$

Recall that by construction, $\ell'_{\rho,id}(z)$ is invariant by rotation of angle $\pi/4$. Using the fact, we have

$$\begin{aligned}
g_\mu &= \mathbb{E}_{z|y=1} |\ell'_{\rho,id}(z)| \sigma'(\mu^\top z) \mu^\top z - \mathbb{E}_{z|y=-1} |\ell'_{\rho,id}(z)| \sigma'(\mu^\top z) \mu^\top z \\
&= 2\mathbb{E}_r \mathbb{E}_{\theta \in (-\pi/2,0)} r |\ell'_{\rho,id}(r,\theta)| (\cos\theta - \sin\theta) - \mathbb{E}_r \mathbb{E}_{\theta \in (0,\pi/4)} r |\ell'_{\rho,id}(r,\theta)| (\cos\theta - \sin\theta) \\
&= 2\mathbb{E}_r \mathbb{E}_{\theta \in (0,\pi/2)} r |\ell'_{\rho,id}(r,\theta)| (\cos\theta + \sin\theta) - \mathbb{E}_r \mathbb{E}_{\theta \in (0,\pi/2)} r |\ell'_{\rho,id}(r,\theta)| |\cos\theta - \sin\theta| \\
&= 2\mathbb{E}_r \mathbb{E}_{\theta \in (0,\pi/4)} r |\ell'_{\rho,id}(r,\theta)| 2\sin\theta + \mathbb{E}_r \mathbb{E}_{\theta \in (\pi/4,\pi/2)} r |\ell'_{\rho,id}(r,\theta)| 2\cos\theta.
\end{aligned}$$

Since we have

$$\mathbb{E}_{\theta\in(0,\pi/4)}|\ell'_{\rho,id}(r,\theta)|2\sin\theta \geq 0.6\mathbb{E}_{\theta\in(\pi/8,\pi/4)}|\ell'_{\rho,id}(r,\theta)|,$$

it is easy to check that $\mathbb{E}_{\theta\in(\pi/8,\pi/4)}|\ell'_{\rho,id}(r,\theta)| \geq 0.5e^{-0.3N^{(t)}r}$ holds. It remains to control integrals of the form $\mathbb{E}_r re^{-cr}$ where r has density $f(x) = x^2e^{-x^2/2}\mathbf{1}_{x\geq 0}$. By using the change of variable $u = cx$ and using dominated convergence, one can easily check that

$$\frac{1}{c^3} \lesssim \mathbb{E}_r re^{-cr} \lesssim \frac{1}{c^3}.$$

□

Proof of Lemma 24. W.l.o.g. we can assume that $\mu = \mu_1$. First, we are going to show that $\mathbb{E}_x\ell'_{\rho,id}(z)\sigma'(w^\top z)\mu^\top z \approx \mathbb{E}_x\ell'_{\rho,id}(z)\sigma'(w_{\text{sig}}^\top z)\mu^\top z$.

We have

$$\begin{aligned}\left|\mathbb{E}_z\ell'_{\rho,id}(z)\sigma'(w^\top z)\mu^\top z - \mathbb{E}_z\ell'_{\rho,id}(z)\sigma'(w_{\text{sig}}^\top z)\mu^\top z\right| &\leq \mathbb{E}_z\ell'_{\rho,id}(z)\mathbf{1}_{|w_{\text{opp}}^\top z|\geq|w_{\text{sig}}^\top z|}|\mu^\top z| \\ &\leq \sqrt{\mathbb{E}_z(\ell'_{\rho,id}(z))^2}\sqrt{\mathbb{E}_z\mathbf{1}_{\{|w_{\text{opp}}^\top z|\geq|w_{\text{sig}}^\top z|\}}|\mu^\top z|}. \\ &\qquad\text{(by Cauchy-Schwartz)}\end{aligned}$$

Notice that $w_{\text{opp}}^\top z$ and $w_{\text{sig}}^\top z$ are orthogonal Gaussian r.v. so we obtain

$$\begin{aligned}\mathbb{E}_z\mathbf{1}_{\{|w_{\text{opp}}^\top z|\geq|w_{\text{sig}}^\top z|\}}|\mu^\top z| &= \mathbb{E}_{G\sim\mathcal{N}(0,1)}|G|\mathbb{P}_{G'\perp G}(|G'| \geq \frac{\|w_{\text{sig}}\|}{\|w_{\text{opp}}\|}|G|) \\ &\leq \sqrt{\frac{2}{\pi}}\frac{\|w_{\text{opp}}\|}{\|w_{\text{sig}}\|}\mathbb{E}_G e^{-\frac{\|w_{\text{sig}}\|^2G^2}{2\|w_{\text{opp}}\|^2}} &&\text{(since } \mathbb{P}(|G|\geq t) \leq \tfrac{1\sqrt{2}}{t\sqrt{\pi}}e^{-t^2/2}\text{)} \\ &\leq \sqrt{\frac{2}{\pi}}\frac{\|w_{\text{opp}}\|}{\|w_{\text{sig}}\|}\frac{1}{\sqrt{1+\frac{\|w_{\text{sig}}\|^2}{\|w_{\text{opp}}\|^2}}} &&\text{(because } \mathbb{E}e^{-tG^2} = \tfrac{1}{\sqrt{1+2t}}\text{)} \\ &\leq \sqrt{\frac{2}{\pi}}\left(\frac{\|w_{\text{opp}}\|}{\|w_{\text{sig}}\|}\right)^{1.5} \\ &\leq \sqrt{\frac{2}{\pi}}\zeta^{1.5}. &&((w,a)\in\mathcal{S})\end{aligned}$$

Also notice that

$$\mathbb{E}_z(\ell'_{\rho,id}(z))^2 \leq \mathbb{E}_z e^{-yf_{2\rho,id}}.$$

So, we have

$$\left|\mathbb{E}_z\ell'_{\rho,id}(z)\sigma'(w^\top z)\mu^\top z - \mathbb{E}_z\ell'_{\rho,id}(z)\sigma'(w_{\text{sig}}^\top z)\mu^\top z\right| \leq \left(\frac{2}{\pi}\right)^{0.25}\zeta^{0.75}g_{2\rho}^{0.5}.$$

By Lemma 25 and the choice of ζ we have $\zeta^{0.75}g_{2\rho}^{0.5} \leq \zeta^{0.25}g_\mu$ for $\rho \lesssim \log\log d$. In conclusion, we have shown that

$$-|a|g_\mu(1+\zeta^{0.25}) \geq \mu^\top\nabla_w^{\text{cl}}L_{\rho,id} \geq -|a|g_\mu(1-\zeta^{0.25}).$$

The same calculation leads to the second result of the lemma. □

C.3 Proof of Lemma 4 and Lemma 6

We will first focus on the dynamic of non-heavy neurons and show that they won't grow faster than heavy neurons. Then, we will study the dynamic of heavy neurons in more detail.

C.3.1 Non-Heavy Neurons Dynamics

Lemma 26. *For any neuron $(w, a) \notin \mathcal{S}$ we have*

$$\left|\nabla_a^{cl} L\right| = \left|w^\top \nabla_w^{cl} L\right| \leq \|w\|\, g_\mu.$$

Proof. W.l.og. we can assume that w_{sig} is aligned with μ_1. Recall that the distribution of $z|y=1$ is the same as $R(z)|y=-1$ where R is a rotation of angle $-\pi/2$. Hence, we can write

$$g_{\mu_1} = \frac{1}{2}\mathbb{E}_{z|y=1}\ell'_{\rho,id}(z)(\sigma(\mu_1^\top z) - \sigma(\mu_2^\top z))$$

since $\ell'_{\rho,id}(z)$ is invariant by rotation of angle $-\pi/2$. In particular, when conditioned on $y=1$, $\ell'_{\rho,id}(z) > 0$. By using the same decomposition, we can write

$$\begin{aligned}\nabla_a^{\text{cl}} L &= \mathbb{E}_z \mathbb{E}_\xi \ell'_{\rho,id}(z)\sigma(\|w_{\text{sig}}\|\,\mu_1^\top z + \|w_{\text{opp}}\|\,\mu_2^\top z + w_\perp^\top \xi) \\ &= \frac{1}{2}\mathbb{E}_{z|y=1}\ell'_{\rho,id}(z)\mathbb{E}_\xi \left(\underbrace{\sigma(\|w_{\text{sig}}\|\,\mu_1^\top z + \|w_{\text{opp}}\|\,\mu_2^\top z + w_\perp^\top \xi) - \sigma(\|w_{\text{sig}}\|\,\mu_2^\top z - \|w_{\text{opp}}\|\,\mu_1^\top z + w_\perp^\top \xi)}_{I} \right)\end{aligned}$$

To simplify the notation, let us denote $a = \|w_{\text{sig}}\|\,\mu_1^\top z + \|w_{\text{opp}}\|\,\mu_2^\top z$, $b = \|w_{\text{sig}}\|\,\mu_2^\top z - \|w_{\text{opp}}\|\,\mu_1^\top z$ and $c = w_\perp^\top \xi$.

Case $a \geq b$. It is easy to check that when $c < -a$, $I = 0$, when $c \geq -b$, $I = a - b$ and when $-a \leq c \leq -b$ we have $I = a + c$. By integrating over ξ, we obtain

$$\begin{aligned}\mathbb{E}_\xi I &= (a-b)\mathbb{P}(w_\perp^\top \xi \geq -b) + \mathbb{E}_\xi(w_\perp^\top \xi + a)\mathbf{1}_{-a \leq w_\perp^\top \xi \leq -b} \\ &\leq (a-b)\mathbb{P}(w_\perp^\top \xi \geq -b) + (a-b)\mathbb{P}(-a \leq w_\perp^\top \xi \leq -b) \qquad \leq (a-b)\mathbb{P}(-a \leq w_\perp^\top \xi) \leq a - b.\end{aligned}$$

Case $a < b$. When $c < -b$, $I = 0$, when $c \geq -a$, $I = (a-b)$ and when $-b \leq c \leq -a$ we have $I = -b - c$. By integrating over ξ, we obtain

$$\begin{aligned}\mathbb{E}_\xi I &= (a-b)\mathbb{P}(w_\perp^\top \xi \geq -a) - \mathbb{E}_\xi(w_\perp^\top \xi + b)\mathbf{1}_{-b \leq w_\perp^\top \xi \leq -a} \\ &\leq (a-b)\mathbb{P}(w_\perp^\top \xi \geq -a).\end{aligned}$$

By consequence $\mathbb{E}_\xi I \leq (a-b)$. But

$$a - b = \mu_1^\top z(\|w_{\text{sig}}\| + \|w_{\text{opp}}\|) + \mu_2^\top z(\|w_{\text{opp}}\| - \|w_{\text{sig}}\|) \leq (\|w_{\text{sig}}\| + \|w_{\text{opp}}\|)(\sigma(\mu_1^\top z) + \sigma(\mu_2^\top z)).$$

Since $g_\mu \geq 0$ we have $\mathbb{E}_{z|y=1}\ell'_{\rho,id}(z)\sigma(\mu_2^\top z) \leq \mathbb{E}_{z|y=1}\ell'_{\rho,id}(z)\sigma(\mu_1^\top z)$. By consequence, we obtain

$$\nabla_a^{\text{cl}} L \leq \frac{(\|w_{\text{sig}}\| + \|w_{\text{opp}}\|)}{2} g_{\mu_1} \leq \|w\|\, g_{\mu_1}.$$

One can derive a similar lower bound with the same argument. □

By using the same argument as in Lemma E.11 in Glasgow (2024) we obtain the following corollary.

Corollary 2. *For every neuron, w.h.p.*

$$\left\|w^{(t+1)}\right\|^2 \leq \left\|w^{(t)}\right\|^2 (1 + 2\eta(1 + 2\zeta H)g_\mu).$$

C.3.2 Heavy Neurons Dynamics

We develop a proof of several lemmata for the inductive properties, such as Lemma 4 and Lemma 6.

Initialization. Let us define

$$\tilde{\mathcal{N}}_1^+ = \{(w,a) \in \mathcal{N} : a > 0, w^\top \mu_1 > \theta/\sqrt{d}\}$$
$$\tilde{\mathcal{N}}_1^- = \{(w,a) \in \mathcal{N} : a > 0, w^\top \mu_1 < -\theta/\sqrt{d}\}$$
$$\tilde{\mathcal{N}}_2^+ = \{(w,a) \in \mathcal{N} : a < 0, w^\top \mu_2 > \theta/\sqrt{d}\}$$
$$\tilde{\mathcal{N}}_2^- = \{(w,a) \in \mathcal{N} : a < 0, w^\top \mu_2 < -\theta/\sqrt{d}\}.$$

The result of Phase I shows that for any block $B \in \{\tilde{\mathcal{N}}_i^+, \tilde{\mathcal{N}}_i^-\}_{i=1}^2$

$$\mathbb{E}_\rho \mathbf{1}_{(w,a)\in B} \|aw\| \geq \zeta^{-2}\theta^2$$

and all the properties of a signal-heavy network are satisfied.

Inductive step. Recall that $N^{(t)}$ corresponds to the mass of each block of the ideal oracle network. In order to establish Lemma 4 and Lemma 6 we will first prove the following lemma.

Lemma 27. *Assume that at time t, the network is (ζ, H) signal-heavy and $\eta \leq \zeta^3$. Then with probability $1 - d^{-\Omega(1)}$ we have*

1. $N^{(t+1)} = (1 + 2\eta g_\mu^{(t)}(1 \pm o(1)))N^{(t)}$,
2. $\left\|w_\perp^{(t+1)}\right\| + \left\|w_{opp}^{(t+1)}\right\| \leq \zeta(1 + O(\eta\zeta^{0.25}))\left\|w_{sig}^{(t+1)}\right\|$,
3. $U^{(t+1)} \leq (1 + \eta g_\mu)U^{(t)} + 5\eta\zeta^{0.25} g_\mu$ *and* $U^{(T_1+t)} \leq 2U^{(T_1)} + t\eta\zeta^{0.25}$.

Proof. W.l.o.g., one can assume that w_{sig} is aligned with μ_1.

Evolution of $N^{(t)}$. For each heavy neuron, we have

$$\begin{aligned}
a^{(t+1)}\left\|w_{\text{sig}}^{(t+1)}\right\| &= (a^{(t)} - \eta\nabla_a \hat{L}_\rho)(\left\|w_{\text{sig}}^{(t)}\right\| - \eta\mu_1^\top \nabla_w \hat{L}_\rho) \\
&= (a^{(t)} - \eta\nabla_a^{\text{cl}} L_{\rho,id})(\left\|w_{\text{sig}}^{(t)}\right\| - \eta\mu_1^\top \nabla_w L_{\rho,id}) \\
&\quad + \underbrace{\eta(|a|\left\|\nabla_w \hat{L}_\rho - \nabla_w^{\text{cl}} L_{\rho,id}\right\| + \|w_{\text{sig}}\|\left\|\nabla_a \hat{L}_\rho - \nabla_a^{\text{cl}} L_{\rho,id}\right\|)}_{E_1} \\
&\quad + \underbrace{\eta^2\left\|\nabla_a \hat{L}_\rho - \nabla_a^{\text{cl}} L_{\rho,id}\right\|\left\|\nabla_w \hat{L}_\rho - \nabla_w^{\text{cl}} L_{\rho,id}\right\|}_{E_2}.
\end{aligned}$$

By Lemma 24 we have

$$\left(a^{(t)} - \eta\nabla_a^{\text{cl}} L_{\rho,id}\right)\left(\left\|w_{\text{sig}}^{(t)}\right\| - \eta\mu_1^\top \nabla_w^{\text{cl}} L_{\rho,id}\right) = a^{(t)}\left\|w_{\text{sig}}^{(t)}\right\|(1 + 2(1 + o(1))\eta g_\mu) + O\left(\eta^2 a^2\left\|w_{\text{sig}}^{(t)}\right\|^2\right).$$

Lemma 7 and Lemma 23 show that

$$E_1 \leq \eta a^2\left((\log d^{-c}) + \mathbb{E}_\rho\|aw_\perp\| + 4N^{(t)}U^{(t)}\right) + \|w\|^2\left(\log d^{-c} + \mathbb{E}_\rho\|aw_\perp\| + 4N^{(t)}U^{(t)}\right)$$

Beside, the definition of heavy network implies that

$$\mathbb{E}_\rho\|aw_\perp\| \leq \zeta N^{(t)} = o(g_\mu) \text{ and } N^{(t)}U^{(t)} = o(g_\mu).$$

E_2 can be controlled in a similar way: Lemma 7 and Lemma 23 gives

$$E_2 \leq \eta^2|a|\left\|w^{(t)}\right\|\left(\log^{-c} d + \mathbb{E}_\rho\|aw_\perp\| + 4N^{(t)}U^{(t)}\right)^2.$$

For non-heavy neurons, Corollary 2 implies that

$$\begin{aligned} m^{-1} \sum_{(a,w)\notin\mathcal{S}} \left\| a^{(t+1)} w^{(t+1)} \right\| &\le m^{-1} \sum_{(a,w)\notin\mathcal{S}} \left\| w^{(t+1)} \right\|^2 \\ &\le m^{-1} \sum_{(a,w)\notin\mathcal{S}} \left\| w^{(t)} \right\|^2 (1 + 2\eta(1+2\zeta H) g_\mu) \\ &\le \zeta \left(1 + 2\eta(1+2\zeta H) g_\mu\right) N^{(t)}. \end{aligned}$$

For heavy neurons, we have

$$m^{-1} \sum_{(a,w)\in\mathcal{S}} \left\| a^{(t+1)} w^{(t+1)} \right\| = \left(1 + 2\eta(1+o(1)) g_\mu\right) m^{-1} \sum_{(a,w)\in\mathcal{S}} \left\| a^{(t)} w^{(t)} \right\|.$$

The stated result follows.

Control of $U^{(t+1)}$. One can show similarly to Phase I that

$$U^{(t+1)} \le (1 + 2\eta\zeta^{0.25} g_\mu^{(t)}) U^{(t)} + 5\eta\zeta^{0.25} g_\mu^{(t)}.$$

It is easy to show by recursion that if a sequence (u_n) is such that $u_{n+1} \le a u_n + b$ then $u_n \le a^n + b(\sum_{k=0}^{n-1} a^k)$. By applying this result with $a = 1 + 2\eta\zeta^{0.25}$ and $b = 5\eta\zeta^{0.25}$ and noting that

$$\frac{1-a^t}{1-a} = \frac{(1+2\eta\zeta^{0.25})^t - 1}{2\eta\zeta^{0.25}} \le 2t$$

since $2\eta\zeta^{0.25} = o(1)$ we obtain

$$U^{T_1+t} \le (1 + 2\eta\zeta^{0.25})^t U^{(T_1)} + 10t\eta\zeta^{0.25}.$$

But by choice of $t \le T_2 = O(\log\log d/\eta)$ we have $(1 + 2\eta\zeta^{0.25})^t = 1 + o(1)$.

Control of $\left\| w_\perp^{(t+1)} \right\|$ for heavy neurons. W.l.o.g., assume that $a > 0$, i.e. w_{sig} is aligned with μ_1. First, we control the approximation error

$$F_1 = \left| w_{\text{opp}}^\top \nabla L_{\rho,id} - \mathbb{E}\ell'_{\rho,id}(z)\sigma'(w_{\text{sig}}^\top z) w^\top \xi \right|.$$

Let us denote $X_1 = \mu_1^\top z$ and $X_2 = \mu_2^\top z$. We have

$$\begin{aligned} F_1 &\le \mathbb{E}\left|\ell'_{\rho,id}(z)\right| \mathbf{1}_{|X_1| \le \frac{|w_\perp^\top \xi| + \|w_{\text{opp}}\||X_2|}{\|w_{\text{sig}}\|}} |w_\perp^\top \xi| \\ &\le \sqrt{\mathbb{E}\left|\ell'_{\rho,id}(z)\right|^2} \sqrt{\mathbb{E}\mathbf{1}_{|X_1| \le \frac{|w_\perp^\top \xi| + \|w_{\text{opp}}\||X_2|}{\|w_{\text{sig}}\|}} |w_\perp^\top \xi|^2} \qquad \text{(by Cauchy-Schwartz)} \\ &\lesssim \sqrt{g_{2\rho}}\, \|w_\perp\| \sqrt{\frac{\|w_{\text{opp}}\| + \|w_\perp\|}{\|w_{\text{sig}}\|}} \\ &\le \sqrt{g_{2\rho}}\, \|w_\perp\| \zeta^{0.5} \ll g_\mu \|w_\perp\|. \end{aligned}$$

Since

$$\mathbb{E}_\xi \ell'_{\rho,id}(z)\sigma'(w_{\text{sig}}^\top z) w^\top \xi = 0,$$

only the approximation term contributes to the gradient.

Control of $\left\|w_{\mathbf{opp}}^{(t+1)}\right\|$ for heavy neurons. Similarly to Lemma 14, let us denote $X_1 = \mu_1^\top z$ and $X_2 = \mu_2^\top z$. We have

$$\begin{aligned}
&\mathbb{E}_x \ell'_{\rho,id}(z)\sigma'(w^\top z)\mu_2^\top z \\
&\quad = \mathbb{E}_{X_1,X_2}(\mathbf{1}_{|X_1|\geq|X_2|} - \mathbf{1}_{|X_2|\geq|X_1|})e^{-N^{(t)}||X_1|-|X_2||}\sigma'(\|w_{\mathrm{sig}}\| X_1 + \|w_{\mathrm{opp}}\| X_2)X_2 \\
&\quad = \mathbb{E}_{X_1,X_2}(\mathbf{1}_{|X_1|\geq|X_2|} - \mathbf{1}_{|X_2|\geq|X_1|})e^{-N^{(t)}||X_1|-|X_2||}\sigma'(\|w_{\mathrm{sig}}\| X_1)X_2 + E
\end{aligned}$$

where

$$E = \mathbb{E}_{X_1,X_2}(\mathbf{1}_{|X_1|\geq|X_2|} - \mathbf{1}_{|X_2|\geq|X_1|})e^{-N^{(t)}||X_1|-|X_2||}(\sigma'(\|w_{\mathrm{sig}}\| X_1 + \|w_{\mathrm{opp}}\| X_2) - \sigma'(\|w_{\mathrm{sig}}\| X_1))X_2.$$

By using the same argument as in Lemma 24, one can show that

$$|E| \lesssim \zeta^{0.75} g_{2\rho}^{0.5}.$$

Furthermore, by using the symmetry of X_2, one obtain that

$$\mathbb{E}_{X_2}\mathbf{1}_{|X_1|\geq|X_2|}e^{-N^{(t)}||X_1|-|X_2||}\sigma'(\|w_{\mathrm{sig}}\| X_1)X_2 = 0.$$

We can conclude that

$$\begin{aligned}
\left\|w_{\mathrm{opp}}^{(t+1)}\right\| + \left\|w_{\perp}^{(t+1)}\right\| &\leq (\left\|w_{\mathrm{opp}}^{(t)}\right\| + \left\|w_{\perp}^{(t)}\right\|)(1+\eta\sqrt{g_{2\rho}}\zeta^{0.5}) \\
&\leq \zeta(1+\eta\sqrt{g_{2\rho}}\zeta^{0.5})\left\|w_{\mathrm{sig}}^{(t)}\right\| \\
&\leq \zeta(1+\eta\sqrt{g_{2\rho}}\zeta^{0.5})\left\|w_{\mathrm{sig}}^{(t+1)}\right\|.
\end{aligned}$$

□

To show that after one gradient iteration, the network is still heavy, it remains to show that the layer weights balance condition is satisfied. This can be done as in Glasgow (2024) by using Lemma 8.

C.4 Conclusion: Proof of Theorem 1

To complete the proof of Theorem 1, we need to establish a connection between $N^{(T)}$ and $\mathbb{E}_x \ell_{\rho^{(T)}}(x)$. This proof proceeds in three main steps:

(i) By contradiction, we show that there exists a stopping time T such that $N^{(T)}$ is of order $\log\log d$. (ii) We upper-bound $\mathbb{E}_z \ell_{\rho^{(T)},\mathrm{id}}(z)$ by separately analyzing the regions where z is close to and far from the decision boundary. (iii) We control the approximation $\mathbb{E}_z \ell_{\rho^{(T)},\mathrm{id}}(z) \approx \mathbb{E}_x \ell_{\rho^{(T)}}(x)$.

Proof of Theorem 1. **Step (i):** The analysis of Phase I (Lemma 2) shows that the resulting network is (ζ, H)-signal-heavy with $\zeta = \log^{-c} d$ for some constant $c > 0$ large enough and

$$N^{(T_1)} \gtrsim (\zeta^{-1}\theta)^2 = \log^{-c'} d$$

for some constant $c' > 0$. Let us define the stopping time

$$T_2 = c''\eta^{-1}(\log\log d)^4$$

for some constant $c'' > 0$ that will be chosen later.

Assume that $N^{(t)}$ does not exceed $C_1 \log\log d$ for $t \in [T_1, T_1 + T_2]$ and some constant $C_1 > 0$ small enough. Then for d large enough, we obtain a contradiction as follows:

$$N^{(T_1+T_2)} \geq N^{(T_1)}\left(1 + (2-o(1))\eta \min_{t\in[T_1,\, T_1+T_2]} g_\mu^{(t)}\right)^{T_2} \qquad \text{(by Lemma 6)}$$

$$\begin{aligned}
&\gtrsim \log^{-c'} d \left(1 + (2 - o(1))\eta c_1 \frac{1}{(\log\log d)^3}\right)^{T_2} && \text{(by Lemma 5)}\\
&\gtrsim \log^{-c'} d \ \exp\big(\Theta(\log\log d)\big)\\
&\gg \log\log d.
\end{aligned}$$

We denote by $T \le T_1 + T_2$ the first time such that $N^{(T)}$ exceeds $C_1 \log\log d$.

Step (ii): We overcome the absence of a strict margin between the clusters by splitting the input data into two classes: 1) the one for which the points are close to the boundary decision and hence the associated network evaluation in small, and the other 2) that are enough separated from the boundary decision and the sign of the network corresponds to the label.

Recall the polar decomposition of $z = r(\cos\theta, \sin\theta)$ where r^2 follows a chi-square distribution with two degrees of freedom, and θ is uniform over $[0, 2\pi)$. We have by definition

$$-y f_{\rho^{(T)},id}(z) = \sqrt{2} N^{(T)} r(|\sin\theta| \wedge |\cos\theta|).$$

Let us define the set

$$\mathcal{C}_\epsilon = \Big\{(r,\theta) \in \mathbb{R}^+ \times [0, 2\pi) : r\big(|\sin\theta| \wedge |\cos\theta|\big) \le \epsilon\Big\}.$$

When $X \sim \mathcal{N}(0, I_2)$ and we write $X = r(\cos\theta, \sin\theta)$, we have $r|\cos\theta| = |X_1|$ and $r|\sin\theta| = |X_2|$, hence

$$\{X \in \mathcal{C}_\epsilon\} = \{\min(|X_1|, |X_2|) \le \epsilon\}.$$

Therefore, by a union bound,

$$\mathbb{P}(X \in \mathcal{C}_\epsilon) \le \mathbb{P}(|X_1| \le \epsilon) + \mathbb{P}(|X_2| \le \epsilon) = 2\mathbb{P}(|Z| \le \epsilon) = O(\epsilon),$$

where $Z \sim \mathcal{N}(0,1)$. Choose $\epsilon = (\log\log d)^{-1/2}$. We have

$$\mathbb{E}_z(\ell_{\rho^{(T)},id}(z)) \le \mathbb{E}_{z \notin \mathcal{C}_\epsilon}(\ell_{\rho^{(T)},id}(z)) + \log(2)\mathbb{P}(\mathcal{C}_\epsilon) \le e^{-\epsilon N^{(T)}} + O(\epsilon) = O(\epsilon),$$

since $\epsilon N^{(T)} = \Theta(\sqrt{\log\log d})$.

Step (iii): We now bound the risk $\mathbb{E}_x \ell_{\rho^{(T)}}(x)$. Recall that for all z we have

$$|f_{\rho^{(T)}}(z) - f_{\rho^{(T)},id}(z)| \le N^{(T)} U^{(T)} \|z\|.$$

Since ℓ is 1-Lipschitz, we obtain that for some constant $c > 0$

$$\begin{aligned}
\mathbb{E}_x \ell_{\rho^{(T)}}(x) &\le \big|\mathbb{E}_x \ell_{\rho^{(T)}}(x) - \ell_{\rho^{(T)}}(z)\big| + \big|\mathbb{E}_x \ell_{\rho^{(T)}}(z) - \ell_{\rho^{(T)},id}(z)\big| + \mathbb{E}_z(\ell_{\rho^{(T)},id}(z))\\
&\le \mathbb{E}_x \big|f_{\rho^{(T)}}(x) - f_{\rho^{(T)}}(z)\big| + \mathbb{E}_x \big|f_{\rho^{(T)}}(z) - f_{\rho^{(T)},id}(z)\big| + \mathbb{E}_z(\ell_{\rho^{(T)},id}(z))\\
&\le \mathbb{E}_x \mathbb{E}_{\rho^{(T)}} |a\xi^\top w_\perp| + 2N^{(T)}U^{(T)} + \mathbb{E}_z(\ell_{\rho^{(T)},id}(z))\\
&\le \mathbb{E}_{\rho^{(T)}} \|a w_\perp\| + O(\log^{-c} d) + \mathbb{E}_z(\ell_{\rho^{(T)},id}(z)) && \text{(by the proof of Lemma 27)}\\
&\le \zeta^{(T)} N^{(T)} + O(\log^{-c} d) + \mathbb{E}_z(\ell_{\rho^{(T)},id}(z)) && \text{(by the definition of signal-heavy network)}\\
&\lesssim \epsilon && \text{(by Step (ii) and the fact } \zeta^{(T)} N^{(T)} \ll \epsilon)
\end{aligned}$$

□

C.5 Proof of Corollary 1

Fix $\epsilon \in (0,1)$ and let $x = z + \xi$ be as in the statement, with $z \notin \mathcal{C}_\epsilon$ and $\xi \sim \mathcal{N}(0, I_{d-2})$ independent Gaussian noise in the orthogonal subspace. Let $y = f^*(z) \in \{-1, 1\}$ be the label of x (which depends only on the projection z).

Step 1: Margin at z for the oracle and the trained network. By the explicit expression of the oracle network, we have for all $z \notin \mathcal{C}_\epsilon$

$$yf_{\rho^{(T)},\mathrm{id}}(z) \;\geq\; c_\epsilon N^{(T)},$$

for some constant $c_\epsilon > 0$ depending only on ϵ. On the other hand, the block-balance control used in the proof of Theorem 1 yields

$$\big|f_{\rho^{(T)}}(z) - f_{\rho^{(T)},\mathrm{id}}(z)\big| \;\leq\; N^{(T)}U^{(T)}\|z\|,$$

where $U^{(T)} \leq \log^{-c_1} d$ for some $c_1 > 0$. Since $N^{(T)} \asymp \log\log d$, we have $N^{(T)}U^{(T)}\|z\| = o\big(N^{(T)}\big)$. Hence, for d large enough,

$$yf_{\rho^{(T)}}(z) = yf_{\rho^{(T)},\mathrm{id}}(z) + y\big(f_{\rho^{(T)}}(z) - f_{\rho^{(T)},\mathrm{id}}(z)\big) \;\geq\; \tfrac{c_\epsilon}{2}N^{(T)}.$$

Set $c := c_\epsilon/2 > 0$ and denote

$$h(\xi) := yf_{\rho^{(T)}}(z+\xi).$$

Then $h(0) = yf_{\rho^{(T)}}(z) \geq cN^{(T)}$.

Step 2: Lipschitz constant of h in the orthogonal noise. We view h as a function of ξ in the orthogonal subspace. We have for any ξ, ξ'

$$\begin{aligned}|h(\xi) - h(\xi')| &= \big|yf_{\rho^{(T)}}(z+\xi) - yf_{\rho^{(T)}}(z+\xi')\big| \\ &\leq \mathbb{E}_{\rho^{(T)}}\Big[|a|\,\big|\sigma(w^\top(z+\xi)) - \sigma(w^\top(z+\xi'))\big|\Big].\end{aligned}$$

Since ReLU is 1-Lipschitz and ξ, ξ' lie in the orthogonal subspace, only the orthogonal component $w_\perp$ contributes, and

$$\big|\sigma(w^\top(z+\xi)) - \sigma(w^\top(z+\xi'))\big| \leq |w_\perp^\top(\xi - \xi')| \leq \|w_\perp\|\,\|\xi - \xi'\|.$$

Therefore

$$|h(\xi) - h(\xi')| \;\leq\; \mathbb{E}_{\rho^{(T)}}\big[|a|\,\|w_\perp\|\big]\;\|\xi - \xi'\|.$$

Define

$$L := \mathbb{E}_{\rho^{(T)}}\|aw_\perp\|.$$

By the signal-heavy property at time T, we have $L \leq \zeta^{(T)}N^{(T)}$ with $\zeta^{(T)} = \log^{-c_2} d$ for some $c_2 > 0$. Thus h is L-Lipschitz in ξ.

Step 3: Mean perturbation in the orthogonal direction. We next control the mean shift $h(\xi) - h(0)$ when ξ is Gaussian. Using the one-dimensional Gaussian computation for each neuron and averaging (see the proof of Theorem 1), one obtains

$$\big|\mathbb{E}_\xi\big(h(\xi) - h(0)\big)\big| \leq \mathbb{E}_{\rho^{(T)}}\|aw_\perp\| = L \leq \zeta^{(T)}N^{(T)}.$$

Since $\zeta^{(T)} = \log^{-c_2} d$ and $N^{(T)} \asymp \log\log d$, we have $\zeta^{(T)}N^{(T)} = o\big(N^{(T)}\big)$. In particular, for d sufficiently large,

$$\mathbb{E}_\xi h(\xi) = h(0) + \mathbb{E}_\xi\big(h(\xi) - h(0)\big) \geq cN^{(T)} - \zeta^{(T)}N^{(T)} \geq \tfrac{c}{2}N^{(T)}.$$

Step 4: Gaussian concentration and misclassification probability. Since h is L-Lipschitz and $\xi \sim \mathcal{N}(0, I_{d-2})$, the Gaussian concentration inequality yields, for all $t > 0$,

$$\mathbb{P}_\xi\big(|h(\xi) - \mathbb{E}_\xi h(\xi)| \geq t\big) \leq 2\exp\big(-t^2/(2L^2)\big).$$

Misclassification corresponds to $h(\xi) \leq 0$. On the event $\{h(\xi) \leq 0\}$ we must have

$$|h(\xi) - \mathbb{E}_\xi h(\xi)| \;\geq\; \mathbb{E}_\xi h(\xi) \;\geq\; \tfrac{c}{2}N^{(T)},$$

for d large enough. Thus

$$\mathbb{P}_\xi\big(h(\xi) \leq 0\big) \;\leq\; \mathbb{P}_\xi\Big(|h(\xi) - \mathbb{E}_\xi h(\xi)| \geq \tfrac{c}{2}N^{(T)}\Big) \;\leq\; 2\exp\left(-\frac{(cN^{(T)}/2)^2}{2L^2}\right).$$

Using $L \le \zeta^{(T)} N^{(T)}$ and $\zeta^{(T)} = \log^{-c_2} d$, we get

$$\frac{(cN^{(T)}/2)^2}{2L^2} \;\ge\; \frac{c^2}{8\,(\zeta^{(T)})^2} \;=\; C(\log d)^{2c_2},$$

for some constant $C > 0$. Absorbing the leading factor 2 into the constant and renaming $2c_2$ as $c > 0$, we obtain

$$\mathbb{P}_\xi\big(h(\xi) \le 0\big) \;\le\; \exp\!\big(-C(\log d)^c\big),$$

as claimed. This shows that, for every $z \notin \mathcal{C}_\epsilon$, the probability (over the orthogonal Gaussian noise ξ) that the trained network misclassifies $x = z + \xi$ is at most $\exp(-C(\log d)^c)$. In particular, such points are correctly classified with overwhelming probability, which proves the corollary.

D ADDITIONAL EXPERIMENTS

We report here the full details and figures complementing Section 5. Unless otherwise stated, we set $d = 120$, $m = 100$, $\eta = 0.05$, $\theta = 0.01$, $M = 6400$ and train up to 10^6 epochs.

D.1 Beyond Gaussian Inputs

We consider two non-Gaussian input distributions: (a) a uniform distribution on $[-1, 1]^d$, and (b) a Gaussian XOR distribution inspired by Xu et al. (2024). Both yield successful training as long as the distribution is symmetric with respect to $\pm\mu_1, \pm\mu_2$. Figures 6a–6d show the test loss and the block evolution.

D.2 Sensitivity to Label Noise

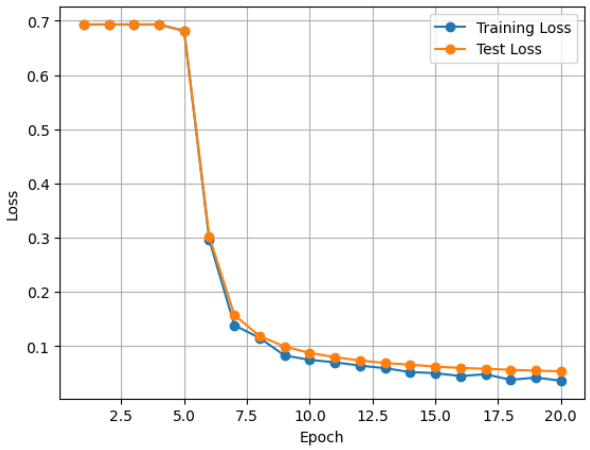
We introduce label noise by flipping the labels of 5% of the data points uniformly at random. As shown in Figure 7, this modification leads to a clear degradation of the test loss. The corresponding decision boundary becomes biased due to the corrupted labels, and the weight vectors fail to maintain the orthogonal growth observed in the noise-free setting.

D.3 Non-isotropic Gaussian inputs

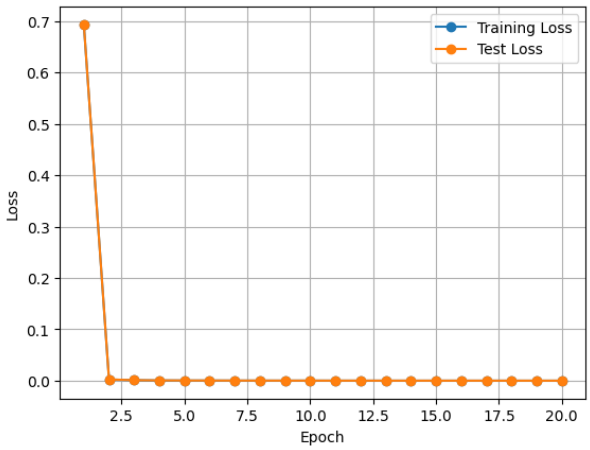
We next consider inputs with covariance $\Gamma = 5\,\mu_1\mu_1^\top + 1\,\mu_2\mu_2^\top$. As shown in Figure 8, the decision boundary is still successfully recovered. However, the neuron orientations become disordered compared to the isotropic setting, reflecting the influence of the anisotropic covariance.

D.4 Nonlinear decision boundaries

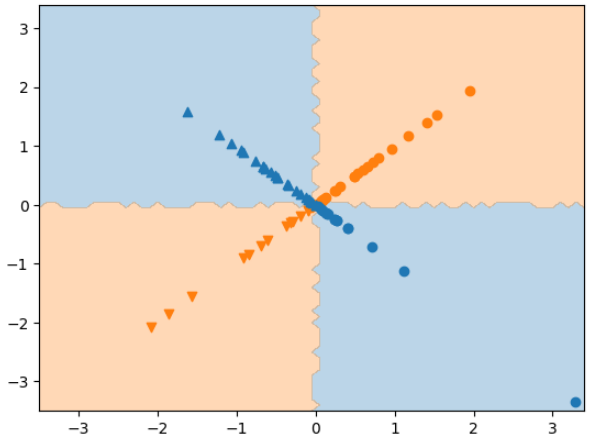
We next test the target function $f(x) = \mathrm{sgn}(x_2 - \sin(x_1))$ with Gaussian inputs. As shown in Figure 9, the learned boundary is approximately linear, and the weight vectors align within $\mathrm{span}\{e_1, e_2\}$. Training progresses for several thousand epochs before plateauing around epoch 5000.



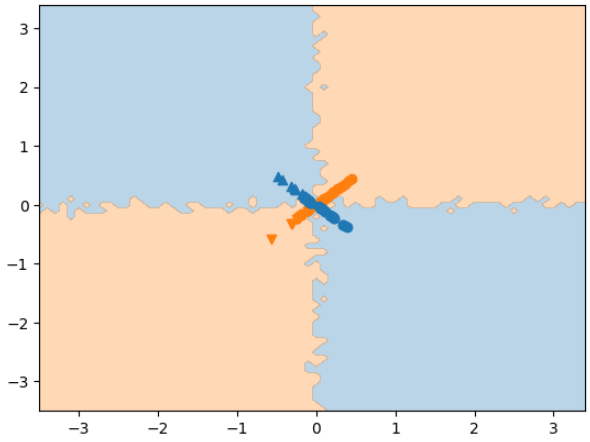
(a) Test loss with uniform inputs



(b) Test loss with Gaussian XOR inputs

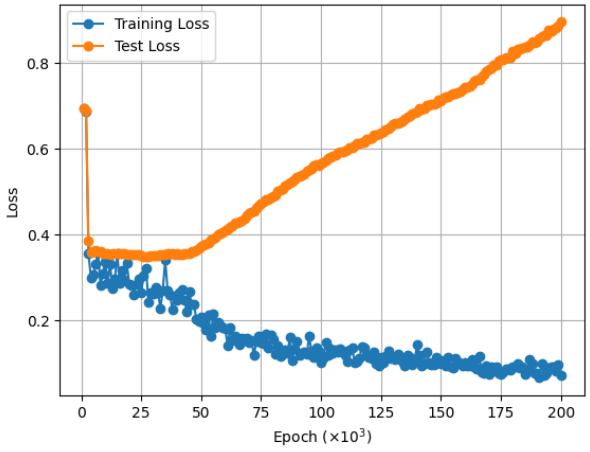


(c) $T = 2 \times 10^4$, uniform inputs

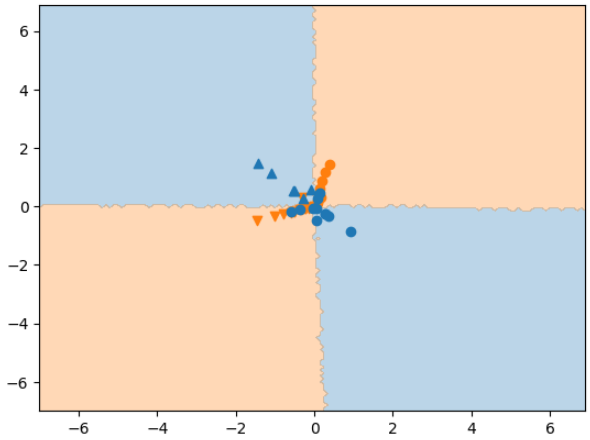


(d) $T = 2 \times 10^4$, Gaussian XOR inputs

Figure 6: Evolution of the test loss and the oracle network blocks size.

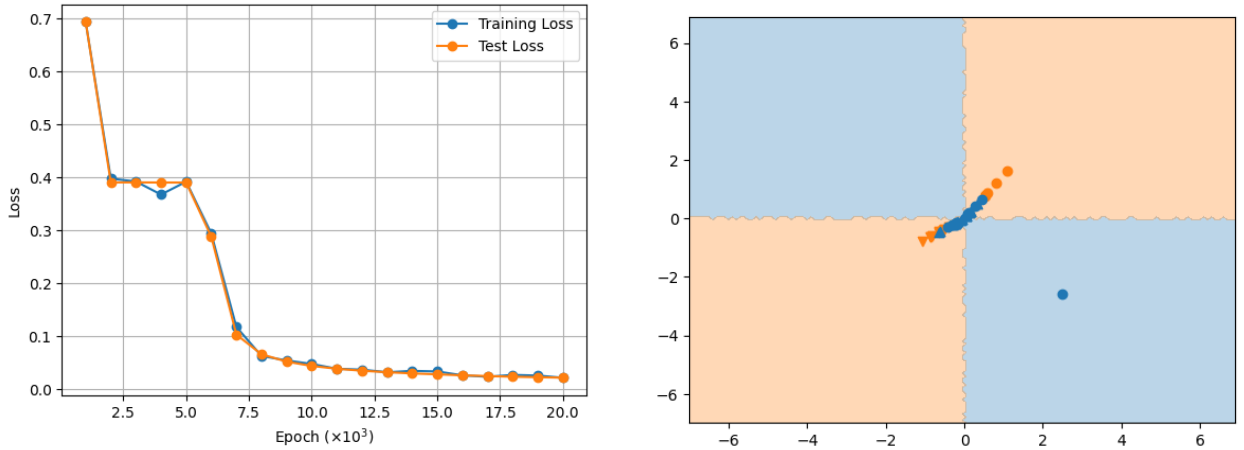


(a) Test loss



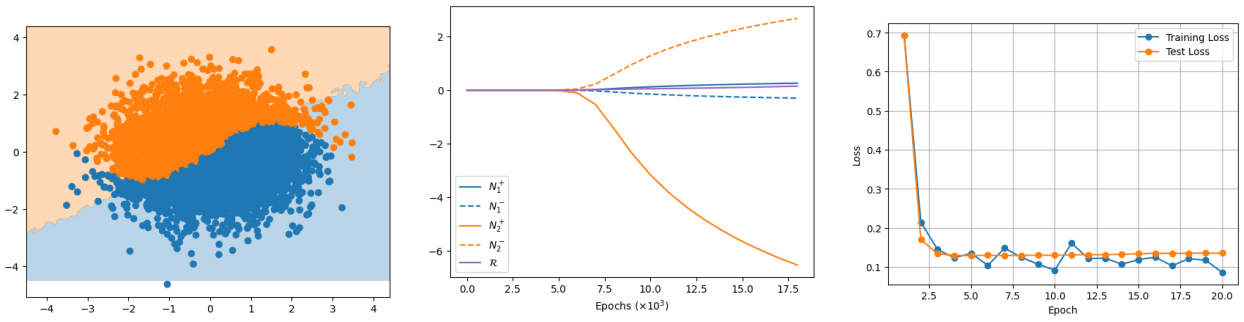
(b) Boundary

Figure 7: Effect of label noise on training dynamics.



(a) Test loss (b) Boundary

Figure 8: Effect of anisotropic inputs on training dynamics.



(a) Learned boundary (b) Weight alignment (c) Training loss

Figure 9: Learning with nonlinear target $f(x) = \mathrm{sgn}(x_2 - \sin(x_1))$. The network recovers an approximately linear boundary, with weights restricted to $\mathrm{span}\{e_1, e_2\}$, and training plateaus after about 5000 epochs.