Simplicity Bias and Optimization Threshold in Two-Layer ReLU Networks

Etienne Boursier [1] **Nicolas Flammarion** [2]

Abstract

Understanding generalization of overparametrized models remains a fundamental challenge in machine learning. The literature mostly studies generalization from an interpolation point of view, taking convergence towards a global minimum of the training loss for granted. This interpolation paradigm does not seem valid for complex tasks such as in-context learning or diffusion. It has instead been empirically observed that the trained models go from global minima to spurious local minima of the training loss as the number of training samples becomes larger than some level we call *optimization threshold*. This paper explores theoretically this phenomenon in the context of two-layer ReLU networks. We demonstrate that, despite overparametrization, networks might converge towards simpler solutions rather than interpolating training data, which leads to a drastic improvement on the test loss. Our analysis relies on the so called early alignment phase, during which neurons align toward specific directions. This directional alignment leads to a simplicity bias, wherein the network approximates the ground truth model without converging to the global minimum of the training loss. Our results suggest this bias, resulting in an optimization threshold from which interpolation is not reached anymore, is beneficial and enhances the generalization of trained models.

1. Introduction

Understanding the generalization capabilities of neural networks remains a fundamental open question in machine learning (Zhang et al., 2021; Neyshabur et al., 2017). Traditionally, research has focused on explaining why neural networks models can achieve zero training loss while still generalizing well to unseen data in supervised learning tasks (Chizat and Bach, 2018; Mei et al., 2018; Rotskoff and Vanden-Eijnden, 2022; Chizat and Bach, 2020; Boursier et al., 2022; Boursier and Flammarion, 2023). This phenomenon is often attributed to overparametrization enabling models to find solutions that interpolate the training data yet avoid overfitting (Belkin et al., 2019; Bartlett et al., 2021). However, the advent of generative AI paradigms—such as large language models (Vaswani et al., 2017) and diffusion models (Dhariwal and Nichol, 2021)—has introduced a paradigm shift in our understanding of generalization. In these settings, models can generate new data and perform novel tasks without necessarily interpolating the training data, raising fresh questions about how and why they generalize. This shift can be illustrated by two seemingly unrelated applications: in-context learning with transformers and generative modeling using diffusion methods.

Firstly, in-context learning (ICL) refers to the ability of large pretrained transformer models to learn new tasks from just a few examples, without any parameter updates (Brown et al., 2020; Min et al., 2022). A central question is whether ICL enables models to learn tasks significantly different from those encountered during pretraining. While prior work suggests that ICL leverages mechanisms akin to Bayesian inference (Xie et al., 2022; Garg et al., 2022; Bai et al., 2023), the limited diversity of tasks in pretraining datasets may constrain the model's ability to generalize. Raventós et al. (2024) investigated this effect by focusing on regression problems to quantify how increasing the variety of tasks during pretraining affects ICL's capacity to generalize to new, unseen tasks, in context.

Secondly, diffusion models have made remarkable strides in generating high-quality images from high-dimensional datasets (Sohl-Dickstein et al., 2015; Ho et al., 2020; Song et al., 2021). These models learn to generate new samples by training denoisers to estimate the score function—the gradient of the log probability density—of the noisy data distribution (Song and Ermon, 2019). A significant challenge in this context is approximating a continuous density from a relatively small training set without succumbing to the curse of dimensionality. Although deep neural networks may tend to memorize training data when the dataset is small relative to the network's capacity (Somepalli et al., 2023; Carlini et al., 2023), Yoon et al. (2023); Kadkhodaie

[1]INRIA, LMO, Université Paris-Saclay, Orsay, France [2]TML Lab, EPFL, Switzerland. Correspondence to: Etienne Boursier <etienne.boursier@inria.fr>.

Proceedings of the 42^{nd} International Conference on Machine Learning, Vancouver, Canada. PMLR 267, 2025.

et al. (2024) observed they generalize well when trained on sufficiently large datasets, rendering the model's behavior nearly independent of the specific training set.

The common thread connecting these examples is a fundamental change in how gradient descent behaves in overparametrized models when the number of data points exceeds a certain threshold. Rather than converging to the global minimum of the training loss, gradient descent converges to a simpler solution closely related to the true loss minimizer. In learning scenarios involving noisy data, the most effective solutions are often those that do not interpolate the data. Despite their capacity to overfit, these models exhibit a simplicity bias, generalizing well to the underlying ground truth instead of merely fitting the noise in the training data. While simplicity bias generally refers to the tendency of models to learn features of increasing complexity, until reaching data interpolation (Arpit et al., 2017; Rahaman et al., 2019; Kalimeris et al., 2019; Huh et al., 2023); this phenomenon seems to stop before full interpolation with modern architectures (even when training for a very long time, see e.g. Raventós et al., 2024, Figure 4). This observation underscores a significant shift in our understanding and approach to generalization in machine learning.

In this paper, we theoretically investigate this phenomenon in the toy setting of shallow ReLU networks applied to a regression problem. While multilayer perceptrons are foundational elements shared by the aforementioned models, focusing on shallow networks remains a significant simplification with respect to the architectures and algorithms used for training transformers and diffusion models. Despite this simplification, we aim to gain theoretical insights that shed light on similar behaviors observed in more complex models. Some recent theoretical works argued that overparametrized networks do not necessarily converge to global minima. In particular, Qiao et al. (2024) showed this effect for unidimensional data by illustrating the instability of global minima. Boursier and Flammarion (2024) advanced a different reason for this effect, given by the *early alignment* phenomenon: when initialized with sufficiently small weights, neurons primarily adjust their directions rather than their magnitudes in the early phase of training, aligning along specific directions determined by the stationary points of a certain known function.

Contributions. Our first contribution is to show that this function driving the early alignment phase concentrates around its expectation, which corresponds to the true loss function, as the number of training samples grows large. For simple teacher architectures, this expected function possesses only a few critical points. As a result, after the early alignment phase, the neurons become concentrated in a few key directions associated with the ground truth model. This behavior reveals a simplicity bias at the initial stages of training. Moreover, this directional concentration is believed to contribute to the non-convergence to the global minimizer of the training loss. However, this characterization only pertains to the initial stage of training. Therefore we extend our analysis to provide, under a restricted data model, a comprehensive characterization of the training dynamics, demonstrating that the simplicity bias persists until the end of training when the number of training samples exceeds some optimization threshold. We here provide an informal version of our main theorem, corresponding to Theorem 4.1.

Theorem 1.1 (Informal)**.** *Consider a specific regression setting with n data samples of dimension d. Using a two-layer ReLU neural network trained with gradient flow and small random initialization, we show that if $n \gtrsim d^3 \log d$, then regardless of the network's width, the learned function closely approximates the ordinary least squares solution.*

In particular, Theorem 1.1 shows that in the overparametrized regime, the final estimator does not minimize the training loss globally, yet it achieves near-optimal performance on the test data. We then confirm empirically our predictions.

2. Preliminaries

This section introduces the setting and the early alignment phenomenon, following the notations and definitions of Boursier and Flammarion (2024).

2.1. Notations

We denote by $\mathbb{S}_{d-1}$ the unit sphere of $\mathbb{R}^d$ and $B(\mathbf{0}, 1)$ the unit ball. We note $f(t) = \mathcal{O}_p\left(g(t)\right)$, if there exists a constant C_p, that only depends on p such that for any t, $|f(t)| \leq C_p g(t)$. We drop the p index, if the constant C_p is universal and does not depend on any parameter. Similarly we note $f(t) = \Omega_p(g(t))$, if there exists a constant $C_p > 0$, that only depends on p such that $f(t) \geq C_p g(t)$ and we write $f(t) = \Theta_p(g(t))$ if both $f(t) = \mathcal{O}_p\left(g(t)\right)$ and $f(t) = \Omega_p(g(t))$. For any bounded set A, $\mathcal{U}(A)$ denotes the uniform probability distribution on the set A.

2.2. Setting

We consider n data points $(x_k, y_k)_{k\in[n]} \in \mathbb{R}^{d+1}$ drawn i.i.d. from a distribution $\mu \in \mathcal{P}(\mathbb{R}^{d+1})$. We also denote by $\mathbf{X} = [x_1^\top, \dots, x_n^\top] \in \mathbb{R}^{d\times n}$ and $\mathbf{y} = (y_1, \dots, y_n) \in \mathbb{R}^n$ respectively the matrix whose columns are given by the input vectors x_k and the vector with coordinates given by the labels y_k. A two layer ReLU network is parametrized by $\theta = (w_j, a_j)_{j\in[m]} \in \mathbb{R}^{m\times(d+1)}$, corresponding to the prediction function

$$h_\theta : x \mapsto \textstyle\sum_{j=1}^m a_j \sigma(w_j^\top x),$$

where σ is the ReLU activation given by $\sigma(z) = \max(0, z)$. While training, we aim at minimizing the empirical square loss over the training data, defined as

$$L(\theta; \mathbf{X}, \mathbf{y}) = \frac{1}{2n} \sum_{k=1}^{n} (h_\theta(x_k) - y_k)^2.$$

As the limiting dynamics of (stochastic) gradient descent with vanishing learning rates, we study a subgradient flow of the training loss, which satisfies for almost any $t \in \mathbb{R}_+$,

$$\dot{\theta}(t) \in -\partial_\theta L(\theta(t); \mathbf{X}, \mathbf{y}), \tag{1}$$

where $\partial_\theta L$ stands for the Clarke subdifferential of L w.r.t. θ.

2.3. Early alignment dynamics

Initialization. In accordance to the feature learning regime (Chizat et al., 2019), the m neurons of the neural network are initialized as

$$(a_j(0), w_j(0)) = \lambda\, m^{-1/2}\, (\tilde{a}_j, \tilde{w}_j), \tag{2}$$

where $\lambda > 0$ is the scale of initialization and $(\tilde{a}_j, \tilde{w}_j)$ are vectors drawn i.i.d. from some distribution, satisfying the following domination property for any $m \in \mathbb{N}$:

$$|\tilde{a}_j| \geq \|\tilde{w}_j\| \text{ for any } j \in [m] \text{ and } \frac{1}{m} \sum_{j=1}^{m} \tilde{a}_j^2 \leq 1. \tag{3}$$

This property is common and allows for a simpler analysis, as it ensures that the signs of the output neurons $a_j(t)$ remain unchanged while training (Boursier et al., 2022).

Neuron dynamics. In the case of two layer neural networks with square loss and ReLU activation, Equation (1) can be written for each neuron $i \in [m]$ as

$$\begin{aligned} \dot{w}_i(t) &\in a_i(t) \mathfrak{D}_n(w_i(t), \theta(t)) \\ \dot{a}_i(t) &= w_i(t)^\top D_n(w_i(t), \theta(t))\rangle, \end{aligned} \tag{4}$$

where the vector $D_n(w_i(t), \theta(t))$ and set $\mathfrak{D}_n(w_i(t), \theta(t))$ are defined as follows, with $\partial\sigma$ the subdifferential of the ReLU activation σ:

$$D_n(w, \theta) = \frac{1}{n} \sum_{k=1}^{n} \mathbb{1}_{x_k^\top w > 0} (y_k - h_\theta(x_k)) x_k,$$

$$\mathfrak{D}_n(w, \theta) = \left\{ \frac{1}{n} \sum_{k=1}^{n} \eta_k (y_k - h_{\theta(t)}(x_k)) x_k \,\middle|\, \eta_k \in \partial\sigma(x_k^\top w) \right\}.$$

These derivations directly follow from the subdifferential of the training loss. In particular, $D_n(w, \theta)$ corresponds to a specific vector (subgradient) in the subdifferential $\mathfrak{D}_n(w, \theta)$. Also observe that the set $\mathfrak{D}_n(w, \theta)$ depends on w only through its activations $A_n(w)$, defined as

$$A_n : \begin{array}{c} \mathbb{R}^d \to \{-1, 0, 1\}^n \\ w \mapsto \operatorname{sign}(w^\top x_k)_{k \in [n]} \end{array}.$$

Furthermore, $\mathfrak{D}_n(w, \theta)$ only depends on θ via the prediction function h_θ. This observation is crucial to the early alignment phenomenon.

Early alignment. In the small initialization regime described by Equation (2), numerous works highlight an early alignment phase in the initial stage of training (Maennel et al., 2018; Atanasov et al., 2022; Boursier and Flammarion, 2024; Kumar and Haupt, 2024; Tsoy and Konstantinov, 2024). During this phase, the neurons exhibit minimal changes in norm, while undergoing significant changes in direction. This phenomenon is due to a discrepancy in the derivatives of the neurons' norms (which scale with λ) and of their directions (which scale in $\Theta(1)$). Specifically, for a sufficiently small initialization scale λ, the neurons align towards the critical directions of the following function G_n defined as

$$G_n : w \mapsto w^\top D_n(w, \mathbf{0}). \tag{5}$$

G_n is continuous, piecewise linear and can be interpreted as the correlation between the gradient information around the origin (given by $D_n(w, \mathbf{0})$) and the neuron w. The network neurons thus align with the critical directions on the sphere of G_n during the early training dynamics. These critical directions are called *extremal vectors*, defined as follows.

Definition 2.1. A vector $D \in \mathbb{R}^d$ is **extremal** with respect to G_n if there exists $w \in \mathbb{S}_{d-1}$ such that both hold

1. $D \in \mathfrak{D}_n(w, \mathbf{0})$;
2. $D = \mathbf{0}$ or $A_n(D) \in \{A_n(w), -A_n(w)\}$.

This definition directly follows from the KKT conditions of the maximization (or minimization) problem, constrained on the sphere, of the function G_n.

Implications of early alignment. By the end of the early alignment, most if not all neurons are nearly aligned with some extremal vector D. Maennel et al. (2018); Boursier and Flammarion (2024) argue that only a few extremal vectors exist in typical learning models. We further explore this claim in Section 3. As a consequence, only a few directions are represented by the network's weights at the end of the early dynamics, even though the neurons cover all possible directions at initialization. Boursier and Flammarion (2024) even show that this *quantization of directions* can prevent the network from interpolating the training set at convergence despite the overparametrization of the network.

Although this *failure of interpolation* has been considered a drawback by Boursier and Flammarion (2024), we show in Section 4 that it can also lead to a beneficial phenomenon of simplicity bias. Specifically, Section 4 illustrates on a simple linear example that for a large number of training samples, the model does not converge to interpolation. Instead, it converges towards the ordinary least square (OLS) estimator of the data. As a consequence, the model fits the true signal of the data, while effectively ignoring label noise. Before studying this example, we must first understand how extremal vectors behave as the number of training samples increases.

3. Geometry of alignment with many samples

We here aim to describe the geometry of the function G_n, with a specific focus on the extremal vectors, as the number of training samples n becomes large. These vectors are key in driving the early alignment phase of the training, making them essential to understanding the initial dynamics of the parameters.

This section provides a general result on the concentration of gradient information D_n of the train loss and support that the early alignment behavior in the infinite-data setting does not differ significantly from that in the large but finite n case. While the tail bound version of Theorem 3.1 is central to our analysis in Section 4, the results of Section 3 are not only stated in a general form for broader applicability, but also constitute standalone contributions that may be useful in future work.

Despite non-smoothness of the loss (due to ReLU activations), we can leverage the piecewise constant structure of the vector function $D_n(w)$, along with typical Rademacher complexity arguments, to derive uniform concentration bounds on the random function $w \mapsto D_n(w)$.

Theorem 3.1. *If the marginal law of x_1 is continuous with respect to the Lebesgue measure, then for any $n \in \mathbb{N}$,*

$$\mathbb{E}_{\mathbf{X},\mathbf{y}}\Big[\sup_{w\in\mathbb{S}_{d-1}}\sup_{D_n\in\mathfrak{D}_n(w,\mathbf{0})}\|D_n - D(w)\|_2\Big] = \mathcal{O}\Big(\sqrt{\frac{d\log n}{n}\mathbb{E}[\|y_1x_1\|_2^2]}\Big),$$

where for any $w \in \mathbb{S}_{d-1}$, $D(w) = \mathbb{E}[\mathbb{1}_{w^\top x_1>0}y_1x_1]$.

Theorem 3.1 indicates that as n grows large, the sets $\mathfrak{D}_n(w,\mathbf{0})$ converge to the corresponding vectors for the true loss, given by $D(w)$, at a rate $\sqrt{\frac{d\log n}{n}}$. Moreover this rate holds uniformly across all possible directions of $\mathbb{R}^d$ in expectation. A probability tail bound version of Theorem 3.1, which bounds this deviation with high probability, can also be derived (see Theorem D.1 in Appendix D). A complete proof of Theorem 3.1 is provided in Appendix C.

When $n \to \infty$, the alignment dynamics are thus driven by vectors D_n which are close to their expected value $D(w)$. Furthermore, when $n \to \infty$, the activations of a weight $A_n(w)$ exactly determine the direction of this weight, as every possible direction is then covered by the training inputs x_k. Specifically, for an infinite dataset indexed by $\mathbb{N}$, whose support covers all directions of $\mathbb{R}^d$, and defining the infinite activation function A as

$$A: \begin{array}{c}\mathbb{S}_{d-1}\to\{-1,0,1\}^{\mathbb{N}}\\ w\mapsto \operatorname{sign}(w^\top x_k)_{k\in\mathbb{N}}\end{array};$$

then A is injective. In this infinite data limit, the functions G_n converge to the function $G: w \mapsto w^\top D(w)$, which is differentiable in this limit, and a vector $D \in \mathbb{R}^d$ is **extremal** with respect to G if there exists $w \in \mathbb{S}_{d-1}$ such that both

$$1.\ D = D(w) \qquad 2.\ D = \mathbf{0} \text{ or } \frac{D}{\|D\|_2} \in \{w, -w\}. \quad (6)$$

When n becomes large, the extremal vectors of the data then concentrate toward the vectors satisfying Equation (6). This is precisely quantified by Proposition 3.1 below.

Proposition 3.1. *Assume the marginal law of x_1 is continuous w.r.t. the Lebesgue measure and that $\mathbb{E}[\|x_1y_1\|^4] < \infty$. Then for any $\varepsilon > 0$, there is $n^\star(\varepsilon) = \mathcal{O}_{\varepsilon,\mu}(d\log d)$ such that for any $n \geq n^\star(\varepsilon)$, with probability at least $1 - \mathcal{O}_\mu\left(\frac{1}{n}\right)$: for any extremal vector D_n of the finite data $(\mathbf{X},\mathbf{y}) \in \mathbb{R}^{n\times(d+1)}$, there exists a vector $D^\star \in \mathbb{R}^d$ satisfying Equation (6), such that*

$$\|D_n - D^\star\|_2 \leq \varepsilon.$$

Proposition 3.1 states that for large n, the extremal vectors concentrate towards the vectors satisfying Equation (6). Note that Proposition 3.1 is not a direct corollary of Theorem 3.1, but its proof relies on the tail bound version of Theorem 3.1 and continuity arguments. A complete proof is given in Appendix E.

Early alignment towards a few directions. Besides laying the ground for Theorem 4.1, Proposition 3.1 aims at describing the geometry of the early alignment when the number of training samples grows large. In particular, Proposition 3.1 shows that all extremal vectors concentrate towards the directions satisfying Equation (6). Although such a description remains abstract, we believe it is satisfied by only a few directions for many data distributions. As an example, for symmetric data distributions, it is respectively satisfied by a single or two directions, when considering a one neuron or linear teacher. More generally, we conjecture it should be satisfied by a small number of directions as soon as the labels are given by a small teacher network. Proving such a result is yet left for future work.

The early alignment phenomenon has been described in many works, to show that after the early training dynamics, only a few directions (given by the extremal vectors) are represented by the neurons (Bui Thi Mai and Lampert, 2021; Lyu et al., 2021; Boursier et al., 2022; Chistikov et al., 2023; Min et al., 2024; Boursier and Flammarion, 2024; Tsoy and Konstantinov, 2024). However, these works all rely on specific data examples, where extremal vectors can be easily expressed for a finite number of samples. Proposition 3.1 aims at providing a more general result, showing that for large n, it is sufficient to consider the directions satisfying Equation (6), which is easier to characterize from a statistical perspective. We thus believe that Proposition 3.1 advances our understanding of how sparse is the network representation (in directions) at the end of early alignment.

Proposition 3.1 implies that for large values of n ($\gtrsim d$), the early alignment phase results in the formation of a small number of neuron clusters, effectively making the neural network equivalent to a small-width network. Empirically, these clusters appear to be mostly preserved throughout training. The neural network then remains equivalent to a small-width network along its entire training trajectory.

In contrast, when the number of data is limited ($n \lesssim d$), this guarantee no longer holds and a large number of extremal vectors may exist. For example in the case of orthogonal data (which only holds for $n \leq d$), there are $\Theta(2^n)$ extremal vectors (Boursier et al., 2022). In such cases, there would still be a large number of neuron clusters at the end of the early alignment phase, maintaining a large *effective width* of the network. Studying how this effective width is maintained until the end of training in the orthogonal case remains an open problem. We conjecture that for a mild overparametrization ($n \lesssim m \ll 2^n$),[1] we would still have a relatively large effective width (increasing with n) at the end of training.

4. Optimization threshold and simplicity bias

The goal of this section is to illustrate the transition from interpolating the training data to a nearly optimal estimator (with respect to the true loss) that can arise when increasing the size of training data. Toward this end, this section proves on a toy data example, that for a large enough number of training samples, an overparametrized network will not converge to a global minimum of the training loss, but will instead be close to the minimizer of the true loss. This is done by analyzing the complete training dynamics, whose first phase–the so-called early alignment phase–is controlled using the tail bound version of Theorem 3.1. To this end, we consider the specific case of a linear data model:

$$y_k = x_k^\top \beta^\star + \eta_k \quad \text{for any } k \in [n], \tag{7}$$

where η_k is some noise, drawn i.i.d. from a centered distribution. We also introduce a specific set of assumptions regarding the data distribution.

Assumption 4.1. The samples x_k and the noise η_k are drawn i.i.d. from distributions μ_X, and μ_η satisfying, for some $c > 0$:

1. μ_X is symmetric, i.e., x_k and $-x_k$ follow the same distribution;
2. μ is continuous with respect to the Lebesgue measure;
3. $\mathbb{P}_{x\sim\mu_X}\left(|x^\top \beta^\star| \leq c\frac{\|x\|_2}{\sqrt{d}}\right) = 0$;
4. $\|\mathbb{E}_{x\sim\mu_X}[xx^\top] - \mathbf{I}_d\|_{\text{op}} < \min\left(\frac{c}{2\sqrt{d}\|\beta^\star\|_2}, \frac{3}{5}\right)$;
5. The random vector x_k is 1 sub-Gaussian and the noise satisfies $\mathbb{E}[\eta^4] < \infty$.

Conditions 1, 2 and 5 in Assumption 4.1 are relatively mild. However, item 3 is quite restrictive: it is needed to ensure that the volume of the activation cone containing $\beta^\star$ does not vanish when $n \to \infty$. A similar assumption is considered by Chistikov et al. (2023); Tsoy and Konstantinov (2024), for similar reasons. Additionally, Condition 4 ensures that $\mathbb{E}_x[xx^\top]\beta^\star$ and $\beta^\star$ are in the same activation cone. This assumption allows the training dynamics to remain within a single cone after the early alignment phase, significantly simplifying our analysis.

As an example, if the samples x_k are distributed i.i.d. as

$$x_k = \mathsf{s}_k \frac{\beta^\star}{\|\beta^\star\|} + \sqrt{d-1}\mathsf{v}_k \text{ with } \mathsf{v}_k \sim \mathcal{U}(\mathbb{S}_{d-1} \cap \{\beta^\star\}^\perp)$$
$$\text{and } \mathsf{s}_k \sim \mathcal{U}\left([-1-\varepsilon, -1+\varepsilon] \cup [1-\varepsilon, 1+\varepsilon]\right),$$

for a small enough $\varepsilon > 0$ and μ_η a standard Gaussian distribution, then Assumption 4.1 is satisfied. In this section, we also consider the following specific initialization scheme for any $i \in [m]$:

$$\begin{aligned} w_i(0) &\sim 0.5\lambda\, m^{-1/2}\,\mathcal{U}(B(\mathbf{0}, 1)) \\ \text{and } a_i(0) &\sim \lambda\, m^{-1/2}\,\mathcal{U}(\{-1, 1\}). \end{aligned} \tag{8}$$

In addition to the regime considered in Equations (2) and (3), this initialization introduces a stronger domination condition, as $|a_i(0)| \geq 2\|w_i(0)\|$. This condition reinforces the early alignment phase, ensuring that **all** neurons are nearly aligned with extremal vectors by the end of this phase. Assumption 4.1 and Equation (8) are primarily introduced to enable a tractable analysis and are discussed further in Section 4.2.

This set of assumptions allows to study the training dynamics separately on the following partition of the data:

$$\mathcal{S}_+ = \{k \in [n] \mid x_k^\top \beta^\star \geq 0\} \quad \text{and} \quad \mathcal{S}_- = [n] \setminus \mathcal{S}_+.$$

Hereafter, we denote by $\mathbf{X}_+ \in \mathbb{R}^{d\times|\mathcal{S}_+|}$ (resp. $\mathbf{X}_-$), the matrix with columns given by the vectors x_k for $k \in \mathcal{S}_+$ (resp. $k \in \mathcal{S}_-$). Similarly, we denote by $\mathbf{Y}_+ \in \mathbb{R}^{|\mathcal{S}_+|}$ (resp. $\mathbf{Y}_-$) the vector with coordinates given by the labels y_k for $k \in \mathcal{S}_+$ (resp. $k \in \mathcal{S}_-$).

Studying separately positive ($a_i > 0$) and negative ($a_i < 0$) neurons, we prove Theorem 4.1 below, which states that at convergence for a large enough number of training samples, the sum of the positive (resp. negative) neurons correspond to the OLS estimator on the subset $\mathcal{S}_+$ (resp. $\mathcal{S}_-$).

Theorem 4.1. *If Assumption 4.1 holds and the initialization scheme follows Equation (8), then there exists $\lambda^\star = \Theta(\frac{1}{d})$ and $n^\star = \Theta(d^3 \log d)$ such that for any $\lambda \leq \lambda^\star$, any $m \in \mathbb{N}$ and $n \geq n^\star$, with probability $1 - \mathcal{O}\left(\frac{d^2}{n} + \frac{1}{2^m}\right)$, the*

[1] Boursier et al. (2022) proved an effective width of 2 at the end of training when $m \gtrsim 2^n$.

parameters $\theta(t)$ converge to some θ_∞ such that

$$h_{\theta_\infty}(x) = (\beta_{n,+}^\top x)_+ - (-\beta_{n,-}^\top x)_+$$

for any $x \in \mathrm{Supp}(\mu_X)$, where $\mathrm{Supp}(\mu_X)$ is the support of the distribution μ_X, $\beta_{n,+} = (\mathbf{X}_+\mathbf{X}_+^\top)^{-1}\mathbf{X}_+\mathbf{Y}_+$ and $\beta_{n,-} = (\mathbf{X}_-\mathbf{X}_-^\top)^{-1}\mathbf{X}_-\mathbf{Y}_-$ are the OLS estimator respectively on the data in $\mathcal{S}_+$ and $\mathcal{S}_-$.

Precisely, the estimator learnt at convergence for a large enough n behaves μ_X-everywhere as the difference of two ReLU neurons, with nearly opposite directions (thanks to the distribution symmetry), resulting in a nearly linear estimator. These directions correspond to the OLS estimator of the data in $\mathcal{S}_+$ and in $\mathcal{S}_-$, respectively. The complete proof of Theorem 4.1 is deferred to Appendix F. We provide a detailed sketch in Section 4.1 below and discuss further Theorem 4.1 in Section 4.2.

4.1. Sketch of Proof of Theorem 4.1

The proof of Theorem 4.1 examines the complete training dynamics of positive neurons ($a_i(0) > 0$) and negative ones ($a_i(0) < 0$) separately. This decoupling is possible at the end of the early phase, due to Assumption 4.1, and is handled thanks to Lemma F.3 in the Appendix.

First note that for the given model, there are only two vectors satisfying Equation (6), corresponding to $\frac{1}{2}\Sigma\beta^\star$ and $-\frac{1}{2}\Sigma\beta^\star$ respectively, for $\Sigma = \mathbb{E}_{x\sim\mu_X}[xx^\top]$. From then and thanks to the third point of Assumption 4.1, the results from Section 3 imply that, for a large value of n and with high probability, there are only two extremal vectors, both of which are close to the expected ones mentioned above. By analyzing the early alignment phase similarly to Boursier and Flammarion (2024), we show that by the end of this early phase, (i) all neurons have small norms; (ii) positive (resp. negative) neurons are aligned with $\Sigma\beta^\star$ (resp. $-\Sigma\beta^\star$). More specifically, at time τ, defined as the end of the early alignment phase, we show that

$$\forall i \in [m],\ \frac{w_i(\tau)}{a_i(\tau)}^\top \Sigma\beta^\star = \|\Sigma\beta^\star\| - \mathcal{O}\big(\lambda^\varepsilon + \sqrt{\frac{d^2\log n}{n}}\big).$$

From that point onward, all positive neurons are nearly aligned and behave as a single neuron until the end of training. Moreover, they remain in the same activation cone until the end of training. Namely for any $i \in [m]$ and $t \geq \tau$,

$$a_i(t)\, x_k^\top w_i(t) > 0 \quad \text{for any } k \in \mathcal{S}_+,$$
$$a_i(t)\, x_k^\top w_i(t) < 0 \quad \text{for any } k \in \mathcal{S}_-.$$

We then show that during a second phase, all positive neurons grow until they reach the OLS estimator on the data in $\mathcal{S}_+$. Mathematically, for some time $\tau_{2,+} > \tau$,

$$\textstyle\sum_{i,a_i(0)>0} a_i(\tau_{2,+})w_i(\tau_{2,+}) \approx \beta_{n,+}.$$

Similarly, negative neurons end up close to $\beta_{n,-}$ after a different time $\tau_{2,-}$. Proving this second phase is quite technical and is actually decomposed into a slow growth and fast growth phases, following a similar approach to Lyu et al. (2021); Tsoy and Konstantinov (2024).

At the end of the second phase, the estimation function is already close to the one described in Theorem 4.1. From then, we control the neurons using a local Polyak-Łojasiewicz inequality (see Equation (45)) to show that they remain close to their value at the end of the second phase, and actually converge to a local minimum corresponding to the estimation function h_{θ_∞} described in Theorem 4.1.

4.2. Discussion

Theorem 4.1 shows that, under a specific linear data model, when the number of training samples exceeds a certain *optimization threshold*, the learned function converges to the OLS estimator—even in highly overparametrized settings where $m \gg n$. This result highlights two key insights:

- Despite overparametrization, the network can converge to a suboptimal solution of the training loss when initialized at a small scale.
- This *training failure* can in fact be beneficial: although suboptimal for the training loss, the resulting estimator is optimal for the test loss.

We now discuss further on Theorem 4.1 and its limitations.

Absence of interpolation. For many years, the literature has argued in favor of the fact that, if overparametrized enough, neural networks do converge towards interpolation of the training set, i.e., to a global minimum of the loss (Jacot et al., 2018; Du et al., 2019; Chizat and Bach, 2018; Wojtowytsch, 2020).

Yet, some recent works argued in the opposite direction that convergence towards global minima might not be achieved for regression tasks, even with infinitely overparametrized networks (Qiao et al., 2024; Boursier and Flammarion, 2024). Indeed, Theorem 4.1 still holds as $m \to \infty$: although interpolation of the data is possible from a statistical aspect[2], interpolation does not occur for optimization reasons. In this direction, Qiao et al. (2024) claim that for large values of n and univariate data, interpolation cannot happen because of the large (i.e., finite) stepsizes used for gradient descent. Following Boursier and Flammarion (2024), we here provide a complementary reason, which is due to the early alignment phenomenon and loss of omnidirectionality of the weights (i.e., the fact that the weights represent all directions in $\mathbb{R}^d$). Note that this loss of omnidirectionality is specific to the (leaky) ReLU activation and does not hold for smooth activations (see e.g. Chizat and Bach, 2018, Lemma

[2] Although the absence of bias term in the parametrization limits the expressivity of the neural network, interpolation is still possible as long as the data x_i are pairwise non-proportional (Carvalho et al., 2025, Theorem 2).

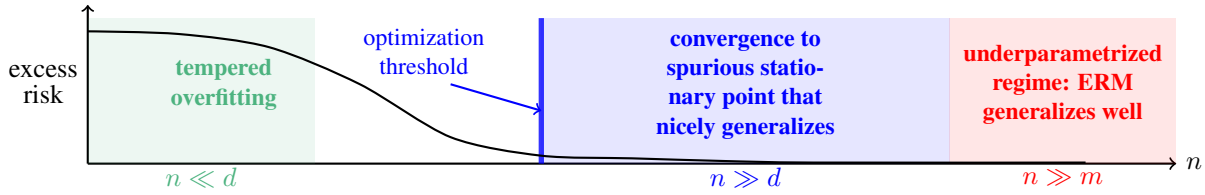


Figure 1: Different regimes of generalization: in green ($n \ll d$), the trained estimator interpolates the data and leads to tempered overfitting; after the optimization threshold in blue ($n \gg d$), we converge to a spurious stationary point of the training loss which generalizes well despite overparametrization, this regime is our main focus; in the underparametrized regime in red ($m \gg n$), the global minima do not interpolate anymore and generalize well.

C.10). We experimentally confirm in Appendix A.4 that both visions are complementary, as interpolation still does not happen for arbitrarily small learning rates.

Simplicity bias. Simplicity bias has been extensively studied in the literature (Arpit et al., 2017; Rahaman et al., 2019; Kalimeris et al., 2019; Huh et al., 2023). It is often described as the fact that networks learn features of increasing complexity while learning. In other words, simpler features are first learnt (e.g., a linear estimator), and more complex features might be learnt later. This has been observed in many empirical studies, leading to improved performance in generalization, except from a few nuanced cases (Shah et al., 2020). Yet in all these studies, the network interpolates the training set after being trained for a long enough time. In consequence, simplicity bias has been characterized by a first *feature learning phase*; and is then followed by an *interpolating phase*, where the remaining noise is fitted (Kalimeris et al., 2019).

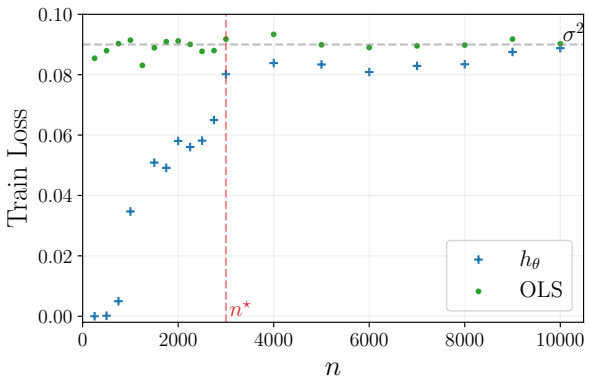
We here go further by showing that this last interpolating phase does not even happen in some cases. Theorem 4.1 indeed claims that after the first feature learning phase, where the network learns a linear estimator, nothing happens in training. The interpolating phase never starts, no matter how long we wait for. While interpolation is often observed for classification problems in practice, it is generally much harder to reach for regression problems (Stewart et al., 2023; Yoon et al., 2023; Kadkhodaie et al., 2024; Raventós et al., 2024). Theorem 4.1 confirms this tendency by illustrating a regression example where interpolation does not happen at convergence. Notably, we here focus on the blue regime in Figure 1 and show that while the global minima poorly generalize in this regime, the optimization scheme only converges to a spurious local minimum of the training loss, which has much better generalization properties. This in stark contrast to the underparametrized regime – $n \gg m$, in red in Figure 1 – where the global minimum has good guarantees, thanks to classical generalization bounds arguments (Bartlett and Mendelson, 2002).

Although implicit bias and simplicity bias often refer to the same behavior in the literature, we here distinguish the two terms: implicit bias is generally considered in the regime of interpolation (Soudry et al., 2018; Lyu and Li, 2019; Chizat and Bach, 2020; Ji and Telgarsky, 2019), while simplicity bias still exists in absence of interpolation.

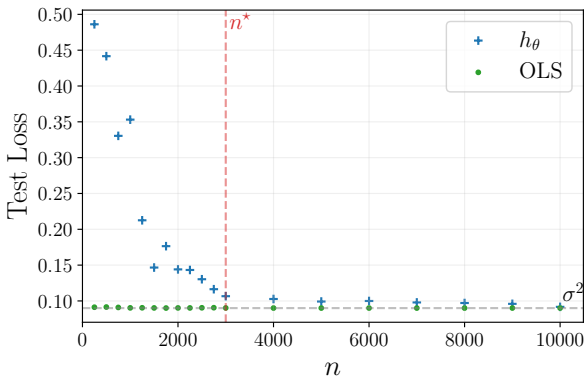
Improved test loss, due to overparametrization threshold. Theorem 4.1 states that for a large enough number of training samples, the interpolating phase does not happen during training, and the estimator then resembles the OLS estimator of the training set. In that regime, the excess risk scales as $\mathcal{O}(d/n)$ (Hsu et al., 2011) and thus quickly decreases to 0 as the number of training samples grows. In contrast when interpolation happens, we either observe a *tempered overfitting*, where the excess risk does not go down to 0 as the number of samples grows (Mallinar et al., 2022); or even a *catastrophic overfitting*, where the excess risk instead diverges to infinity as the size of the training set increases (Joshi et al., 2024).

The fact that the excess risk goes down to 0 as n grows in our example of Section 4 could not be due to a benign overfitting (Belkin et al., 2018; Bartlett et al., 2020), as benign overfitting occurs when the dimension d also grows to infinity. We here consider a fixed dimension instead, and this reduced risk is then solely due to the *optimization threshold*, i.e., the fact that for a large enough n, the interpolating phase does not happen anymore. While some works rely on early stopping before this interpolating phase to guarantee such an improved excess risk (Ji et al., 2021; Mallinar et al., 2022; Frei et al., 2023), it can be guaranteed without any early stopping after this optimization threshold. A similar threshold has been empirically observed in diffusion and in-context learning (Yoon et al., 2023; Kadkhodaie et al., 2024; Raventós et al., 2024), where the trained model goes from interpolation to generalization as the number of training samples increases.

Limitations and generality. While Theorem 4.1 considers a very specific setting, it describes a more general behavior. Although condition 3 of Assumption 4.1 and the initialization scheme of Equation (8) are quite artificial, they are merely required to allow a tractable analysis. The experiments of Section 5 are indeed run without these conditions and yield results similar to the predictions of Theorem 4.1 for large enough n.



(a) Evolution of train loss.



(b) Evolution of test loss.

Figure 2: Evolution of both train and test losses at convergence with respect to the number of training samples. σ^2 corresponds to the noise variance $\mathbb{E}[\eta^2]$.

More particularly, condition 3 of Assumption 4.1 is required to ensure that only two extremal vectors exist. Without this condition, there could be additional extremal vectors, but all concentrated around these two main extremal ones. On the other hand, Equation (8) is required to enforce the early alignment phase, so that all neurons are aligned towards extremal vectors at its end. With a more general initialization, some neurons could move arbitrarily slowly in the early alignment dynamics, ending unaligned at the end of early phase. Yet, such neurons would be very rare. Relaxing these two assumptions would make the final convergence point slightly more complex than the one in Theorem 4.1. Besides the two main ReLU components described in Theorem 4.1, a few small components could also be added to the final estimator, without significantly changing the reached excess risk, as observed in Section 5. This is observed in Figure 2, where the training loss is only slightly smaller than the training loss of OLS, with a comparable test loss.

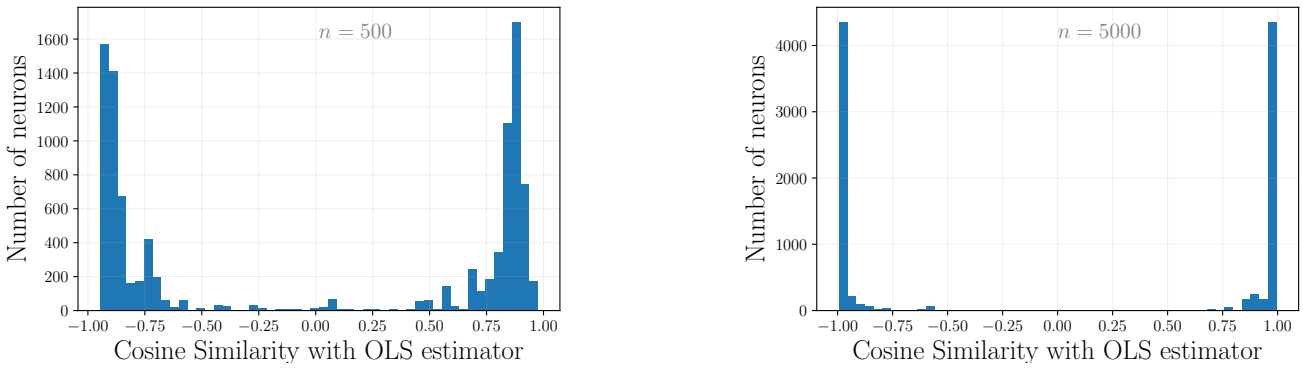
From a higher level, Theorem 4.1 is restricted to a linear teacher and a simple network architecture. It remains hard to assess how well the considered setting reflects the behavior of more complex architectures encountered in practice. We believe that the different conclusions of our work remain valid in more complex setups. In particular, additional experiments in Appendix A run with a more complex teacher, GeLU activations or with Adam optimizer yield similar behaviors: the obtained estimator does not interpolate for a large number of training samples, but instead accurately approximates the minimizer of the test loss. Similar behaviors have also been observed on more complex tasks as generative modeling or in-context learning (Yoon et al., 2023; Kadkhodaie et al., 2024; Raventós et al., 2024). Despite overparametrization, the trained model goes from perfect interpolation to generalization, as it fails at interpolating for a large number of training samples. In these works as well, this absence of interpolation does not seem due to an early stopping, but rather to convergence to a local minimum (see e.g., Raventós et al., 2024, Figure 4).

Lastly, Theorem 4.1 requires a very large number of samples with respect to the dimension, i.e., $n \gtrsim d^3 \log d$. Our experiments confirm that the optimization threshold only appears for a very large number of training samples with respect to the dimension. However, similar behaviors seem to occur for smaller orders of magnitude for n in more complex learning problems, such as the training of diffusion models (Yoon et al., 2023; Kadkhodaie et al., 2024). This dependency in d might indeed be different for more complex architectures (e.g., with attention) and is worth investigating for future work.

5. Experiments

This section illustrates our results on experiments on a toy model close to the setting of Section 4. More precisely, we train overparametrized two-layer neural networks ($m = 10\,000$) until convergence, on data from the linear model of Equation (7). The network is trained via stochastic gradient descent and the dimension is fixed to $d = 5$ to allow reasonable running times. The setup here is more general than Section 4, since i) the data input x_k are drawn from a standard Gaussian distribution (which does not satisfy Assumption 4.1); ii) the neurons are initialized as centered Gaussian of variance $10^{-5}/m$ (which does not satisfy Equations (3) and (8)). We refer to Appendix A for details on the considered experiments and additional experiments.

Figure 2 illustrates the behavior of both train loss and test loss at convergence, when the size of the training set n varies. As predicted by Theorem 4.1, when n exceeds some optimization threshold, the estimator at convergence does not interpolate the training set. Instead, it resembles the optimal OLS estimator, which yields a test loss close to the noise level $\mathbb{E}[\eta^2]$. In contrast for smaller training sets, the final estimator interpolates the data at convergence, which



(a) $n = 500$. (b) $n = 5\,000$.

Figure 3: Histogram of the cosine similarities of the neurons with the true OLS estimator $\hat{\beta}$, at the end of training.

yields a much larger test loss than OLS, corresponding to the tempered overfitting regime (Mallinar et al., 2022).

This optimization threshold is here located around $n^\star = 3000$, which suggests that the large dependency of this threshold in the dimension (which is here 5) in Theorem 4.1 seems necessary–see Appendix A.5 for experiments with larger dimensions. We still observe a few differences here with the predictions of Theorem 4.1, which are due to the two differences in the setups mentioned above. Indeed, even after this optimization threshold, the test loss of the obtained network is slightly larger than the one of OLS, while Theorem 4.1 predicts they should coincide. This is because in the experimental setup, a few neurons remain disaligned with the extremal ones at the end of the early alignment phase. These neurons will then later in training grow in norm, trying to fit a few data points. However there are only a few of such neurons, whose impact thus becomes limited–see Figure 3. As a consequence, they only manage to slightly improve the train loss and have little impact on the test loss.

5.1. Cosine similarity with OLS estimator

To illustrate Theorem 4.1 and the fact that neurons end up aligned with the OLS estimator beyond the optimization threshold $n^\star$, Figure 3 shows histograms of the cosine similarities[3] between all the neurons w_i of the network at the end of training and the true OLS estimator $\hat{\beta} = (\mathbf{X}\mathbf{X}^\top)^{-1}\mathbf{X}\mathbf{Y}$, for different sample complexities. This experiment follows the same setup as the one of Figure 2. In particular, Figure 3(a) shows this histogram for $n = 500$, where interpolation of the training data happens (see Figure 2(a)); and Figure 3(b) shows this histogram for $n = 5\,000$, where interpolation of the training data does not happen anymore, but the network generalizes well to unseen data.

[3]The cosine similarity between two vectors $u, v \in \mathbb{R}^d$ is defined as $\cos(u, v) = \frac{u^\top v}{\|u\|\,\|v\|}$.

While a majority of the neurons is already nicely aligned with the true OLS estimator in the $n = 500$ case, an important fraction of them are not aligned with this estimator (69% of them have a cosine similarity smaller than 0.9 in absolute value). These unaligned neurons contribute to a prediction function that significantly differs from the OLS one. On the other hand, nearly all neurons are aligned with this true estimator as n grows larger (91% of them have a cosine similarity larger than 0.9 in absolute value), confirming the predictions of Theorem 4.1. As explained above, there are still a few vectors that are disaligned with the OLS estimator here, but they are only a small fraction and thus have almost no impact on the estimated function.

6. Conclusion

This work illustrates on a simple linear example the phenomenon of non-convergence of the parameters towards a global minimum of the training loss, despite overparametrization. This non-convergence actually yields a simplicity bias on the final estimator, which can lead to an optimal fit of the true data distribution. A similar phenomenon has been observed on more complex and realistic settings (Yoon et al., 2023; Kadkhodaie et al., 2024; Raventós et al., 2024). However, a theoretical analysis remains out of reach in these cases. It is still unclear whether the observed non-convergence arises from the early alignment mechanism proposed in our work, from stability issues as suggested by Qiao et al. (2024), from other factors, or from a combination of these effects.

Our result is proven via the description of the early alignment phase. Besides the specific data example considered in Section 4, we also provide concentration bounds on the extremal vectors driving this early alignment. We believe these bounds (Theorem 3.1) can be used in subsequent works to better understand this early phase of the training dynamics, and how it yields biases towards simple estimators.

Acknowledgements

This work was supported by the Swiss National Science Foundation (grant number 212111) and by an unrestricted gift from Google.

Impact Statement

This paper presents work whose goal is to advance the field of Machine Learning. There are many potential societal consequences of our work, none which we feel must be specifically highlighted here.

References

Sanjeev Arora, Simon Du, Wei Hu, Zhiyuan Li, and Ruosong Wang. Fine-grained analysis of optimization and generalization for overparameterized two-layer neural networks. In *International Conference on Machine Learning*, pages 322–332. PMLR, 2019.

Devansh Arpit, Stanisław Jastrzębski, Nicolas Ballas, David Krueger, Emmanuel Bengio, Maxinder S Kanwal, Tegan Maharaj, Asja Fischer, Aaron Courville, Yoshua Bengio, et al. A closer look at memorization in deep networks. In *International conference on machine learning*, pages 233–242. PMLR, 2017.

Alexander Atanasov, Blake Bordelon, and Cengiz Pehlevan. Neural networks as kernel learners: The silent alignment effect. In *International Conference on Learning Representations*, 2022.

Yu Bai, Fan Chen, Huan Wang, Caiming Xiong, and Song Mei. Transformers as statisticians: Provable in-context learning with in-context algorithm selection. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023. URL https://openreview.net/forum?id=liMSqUuVg9.

Peter L Bartlett and Shahar Mendelson. Rademacher and gaussian complexities: Risk bounds and structural results. *Journal of Machine Learning Research*, 3(Nov):463–482, 2002.

Peter L Bartlett, Philip M Long, Gábor Lugosi, and Alexander Tsigler. Benign overfitting in linear regression. *Proceedings of the National Academy of Sciences*, 117(48): 30063–30070, 2020.

Peter L. Bartlett, Andrea Montanari, and Alexander Rakhlin. Deep learning: a statistical viewpoint. *Acta Numerica*, 30:87–201, 2021. doi: 10.1017/S0962492921000027.

Mikhail Belkin, Daniel J Hsu, and Partha Mitra. Overfitting or perfect fitting? risk bounds for classification and regression rules that interpolate. *Advances in neural information processing systems*, 31, 2018.

Mikhail Belkin, Daniel Hsu, Siyuan Ma, and Soumik Mandal. Reconciling modern machine-learning practice and the classical bias–variance trade-off. *Proceedings of the National Academy of Sciences*, 116(32):15849–15854, 2019.

Etienne Boursier and Nicolas Flammarion. Penalising the biases in norm regularisation enforces sparsity. *Advances in Neural Information Processing Systems*, 36:57795–57824, 2023.

Etienne Boursier and Nicolas Flammarion. Early alignment in two-layer networks training is a two-edged sword. *arXiv preprint arXiv:2401.10791*, 2024.

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

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901, 2020.

Phuong Bui Thi Mai and Christoph Lampert. The inductive bias of relu networks on orthogonally separable data. In *9th International Conference on Learning Representations*, 2021.

Nicolas Carlini, Jamie Hayes, Milad Nasr, Matthew Jagielski, Vikash Sehwag, Florian Tramer, Borja Balle, Daphne Ippolito, and Eric Wallace. Extracting training data from diffusion models. In *32nd USENIX Security Symposium (USENIX Security 23)*, pages 5253–5270, 2023.

Luís Carvalho, João L Costa, José Mourão, and Gonçalo Oliveira. The positivity of the neural tangent kernel. *SIAM Journal on Mathematics of Data Science*, 7(2):495–515, 2025.

Dmitry Chistikov, Matthias Englert, and Ranko Lazic. Learning a neuron by a shallow relu network: Dynamics and implicit bias for correlated inputs. *Advances in Neural Information Processing Systems*, 36:23748–23760, 2023.

Lenaic Chizat and Francis Bach. On the global convergence of gradient descent for over-parameterized models using optimal transport. *Advances in neural information processing systems*, 31, 2018.

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

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

Thomas M Cover. Geometrical and statistical properties of systems of linear inequalities with applications in pattern recognition. *IEEE transactions on electronic computers*, (3):326–334, 1965.

Prafulla Dhariwal and Alexander Quinn Nichol. Diffusion models beat GANs on image synthesis. In A. Beygelzimer, Y. Dauphin, P. Liang, and J. Wortman Vaughan, editors, *Advances in Neural Information Processing Systems*, 2021.

Simon Du, Jason Lee, Haochuan Li, Liwei Wang, and Xiyu Zhai. Gradient descent finds global minima of deep neural networks. In *International conference on machine learning*, pages 1675–1685. PMLR, 2019.

Spencer Frei, Niladri S Chatterji, and Peter L Bartlett. Random feature amplification: Feature learning and generalization in neural networks. *Journal of Machine Learning Research*, 24(303):1–49, 2023.

Shivam Garg, Dimitris Tsipras, Percy S Liang, and Gregory Valiant. What can transformers learn in-context? a case study of simple function classes. *Advances in Neural Information Processing Systems*, 35:30583–30598, 2022.

Dan Hendrycks and Kevin Gimpel. Gaussian error linear units (gelus). *arXiv preprint arXiv:1606.08415*, 2016.

Tom Henighan, Shan Carter, Tristan Hume, Nelson Elhage, Robert Lasenby, Stanislav Fort, Nicholas Schiefer, and Christopher Olah. Superposition, memorization, and double descent. *Transformer Circuits Thread*, 6:24, 2023.

Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 6840–6851. Curran Associates, Inc., 2020.

Daniel Hsu, Sham M Kakade, and Tong Zhang. An analysis of random design linear regression. *arXiv preprint arXiv:1106.2363*, 2011.

Minyoung Huh, Hossein Mobahi, Richard Zhang, Brian Cheung, Pulkit Agrawal, and Phillip Isola. The low-rank simplicity bias in deep networks. *Transactions on Machine Learning Research*, 2023. ISSN 2835-8856.

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

Ziwei Ji and Matus Telgarsky. The implicit bias of gradient descent on nonseparable data. In *Conference on Learning Theory*, pages 1772–1798. PMLR, 2019.

Ziwei Ji, Justin Li, and Matus Telgarsky. Early-stopped neural networks are consistent. *Advances in Neural Information Processing Systems*, 34:1805–1817, 2021.

Nirmit Joshi, Gal Vardi, and Nathan Srebro. Noisy interpolation learning with shallow univariate reLU networks. In *The Twelfth International Conference on Learning Representations*, 2024.

Zahra Kadkhodaie, Florentin Guth, Eero P Simoncelli, and Stéphane Mallat. Generalization in diffusion models arises from geometry-adaptive harmonic representations. In *The Twelfth International Conference on Learning Representations*, 2024.

Dimitris Kalimeris, Gal Kaplun, Preetum Nakkiran, Benjamin Edelman, Tristan Yang, Boaz Barak, and Haofeng Zhang. Sgd on neural networks learns functions of increasing complexity. *Advances in neural information processing systems*, 32, 2019.

Diederik P Kingma. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

Akshay Kumar and Jarvis Haupt. Directional convergence near small initializations and saddles in two-homogeneous neural networks. *Transactions on Machine Learning Research*, 2024. ISSN 2835-8856.

Qianxiao Li, Cheng Tai, and E Weinan. Stochastic modified equations and dynamics of stochastic gradient algorithms i: Mathematical foundations. *Journal of Machine Learning Research*, 20(40):1–47, 2019.

Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *International Conference on Learning Representations*, 2019.

Kaifeng Lyu and Jian Li. Gradient descent maximizes the margin of homogeneous neural networks. In *International Conference on Learning Representations*, 2019.

Kaifeng Lyu, Zhiyuan Li, Runzhe Wang, and Sanjeev Arora. Gradient descent on two-layer nets: Margin maximization and simplicity bias. *Advances in Neural Information Processing Systems*, 34:12978–12991, 2021.

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

Neil Mallinar, James Simon, Amirhesam Abedsoltan, Parthe Pandit, Misha Belkin, and Preetum Nakkiran. Benign, tempered, or catastrophic: Toward a refined taxonomy of

overfitting. *Advances in Neural Information Processing Systems*, 35:1182–1195, 2022.

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

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

Sewon Min, Xinxi Lyu, Ari Holtzman, Mikel Artetxe, Mike Lewis, Hannaneh Hajishirzi, and Luke Zettlemoyer. Rethinking the role of demonstrations: What makes in-context learning work? In Yoav Goldberg, Zornitsa Kozareva, and Yue Zhang, editors, *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 11048–11064, Abu Dhabi, United Arab Emirates, December 2022. Association for Computational Linguistics.

Preetum Nakkiran, Gal Kaplun, Yamini Bansal, Tristan Yang, Boaz Barak, and Ilya Sutskever. Deep double descent: Where bigger models and more data hurt. *Journal of Statistical Mechanics: Theory and Experiment*, 2021 (12):124003, 2021.

Behnam Neyshabur, Srinadh Bhojanapalli, David McAllester, and Nathan Srebro. Exploring generalization in deep learning. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, NIPS'17, page 5949–5958, Red Hook, NY, USA, 2017. Curran Associates Inc. ISBN 9781510860964.

Dan Qiao, Kaiqi Zhang, Esha Singh, Daniel Soudry, and Yu-Xiang Wang. Stable minima cannot overfit in univariate relu networks: Generalization by large step sizes. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024.

Nasim Rahaman, Aristide Baratin, Devansh Arpit, Felix Draxler, Min Lin, Fred Hamprecht, Yoshua Bengio, and Aaron Courville. On the spectral bias of neural networks. In *International conference on machine learning*, pages 5301–5310. PMLR, 2019.

Allan Raventós, Mansheej Paul, Feng Chen, and Surya Ganguli. Pretraining task diversity and the emergence of non-bayesian in-context learning for regression. *Advances in Neural Information Processing Systems*, 36, 2024.

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

Mark Rudelson and Roman Vershynin. Hanson-wright inequality and sub-gaussian concentration. 2013.

Harshay Shah, Kaustav Tamuly, Aditi Raghunathan, Prateek Jain, and Praneeth Netrapalli. The pitfalls of simplicity bias in neural networks. *Advances in Neural Information Processing Systems*, 33:9573–9585, 2020.

Jascha Sohl-Dickstein, Eric Weiss, Niru Maheswaranathan, and Surya Ganguli. Deep unsupervised learning using nonequilibrium thermodynamics. In Francis Bach and David Blei, editors, *Proceedings of the 32nd International Conference on Machine Learning*, volume 37 of *Proceedings of Machine Learning Research*, pages 2256–2265, Lille, France, 07–09 Jul 2015. PMLR.

Gowthami Somepalli, Vasu Singla, Micah Goldblum, Jonas Geiping, and Tom Goldstein. Diffusion art or digital forgery? investigating data replication in diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6048–6058, 2023.

Yang Song and Stefano Ermon. Generative modeling by estimating gradients of the data distribution. *Advances in neural information processing systems*, 32, 2019.

Yang Song, Jascha Sohl-Dickstein, Diederik P Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. Score-based generative modeling through stochastic differential equations. In *International Conference on Learning Representations*, 2021.

Daniel Soudry, Elad Hoffer, Mor Shpigel Nacson, Suriya Gunasekar, and Nathan Srebro. The implicit bias of gradient descent on separable data. *The Journal of Machine Learning Research*, 19(1):2822–2878, 2018.

Lawrence Stewart, Francis Bach, Quentin Berthet, and Jean-Philippe Vert. Regression as classification: Influence of task formulation on neural network features. In *International Conference on Artificial Intelligence and Statistics*, pages 11563–11582. PMLR, 2023.

Nikita Tsoy and Nikola Konstantinov. Simplicity bias of two-layer networks beyond linearly separable data. In *Proceedings of the 41st International Conference on Machine Learning*, pages 48728–48767, 2024.

Aad W Van Der Vaart and Jon A Wellner. *Weak convergence and empirical processes: with applications to statistics*. Springer Cham, 2023.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

Roman Vershynin. *High-dimensional probability: An introduction with applications in data science*, volume 47. Cambridge university press, 2018.

Stephan Wojtowytsch. On the convergence of gradient descent training for two-layer relu-networks in the mean field regime. *arXiv preprint arXiv:2005.13530*, 2020.

Sang Michael Xie, Aditi Raghunathan, Percy Liang, and Tengyu Ma. An explanation of in-context learning as implicit bayesian inference. In *International Conference on Learning Representations*, 2022.

TaeHo Yoon, Joo Young Choi, Sehyun Kwon, and Ernest K Ryu. Diffusion probabilistic models generalize when they fail to memorize. In *ICML 2023 Workshop on Structured Probabilistic Inference $\{\backslash\&\}$ Generative Modeling*, 2023.

Chiyuan Zhang, Samy Bengio, Moritz Hardt, Benjamin Recht, and Oriol Vinyals. Understanding deep learning (still) requires rethinking generalization. *Communications of the ACM*, 64(3):107–115, 2021.

Appendix

Table of Contents

A. Additional experiments

A.1. Experimental details

In the experiments of Figure 2, we initialized two-layer ReLU networks (without bias term) with $m = 10\,000$ neurons, initialized i.i.d. for each component as a Gaussian of variance $\frac{10^{-5}}{\sqrt{m}}$. We then generated training samples as

$$y_k = {\beta^\star}^\top x_k + \eta_k,$$

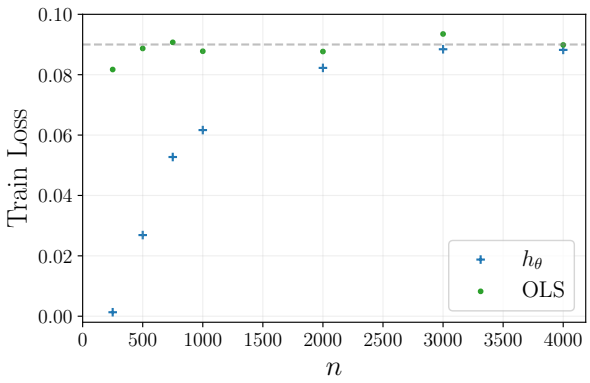
where η_k are drawn i.i.d. as centered Gaussian of variance $\sigma^2 = 0.09$, x_k are drawn i.i.d. as centered Gaussian variables and $\beta^\star$ is fixed, without loss of generality, to $\beta^\star = (1, 0, \dots, 0)$. The dimension is fixed to $d = 5$. We then train these networks on training datasets of different sizes (each dataset is resampled from scratch).

The neural networks are trained via stochastic gradient descent (SGD), with batch size 32 and learning rate 0.01. To ensure that we reached convergence of the parameters, we train the networks for 8×10^6 iterations of SGD, where the training seems stabilized.

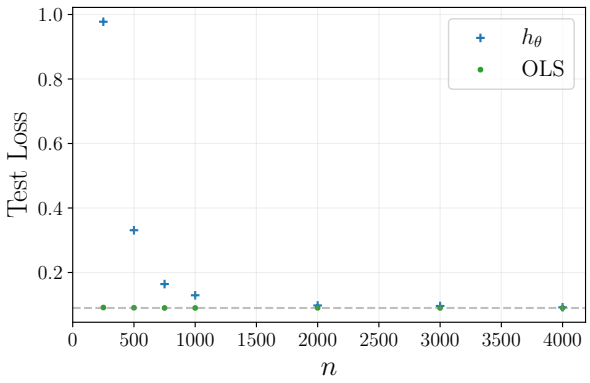
All the experiments were run on a personal MacBook Pro, for a total compute time of approximately 100 hours. The code can be found at `github.com/eboursier/simplicity_bias`.

A.2. GeLU activation

Our theoretical results can be directly extended to any homogeneous activation function, i.e., leaky ReLU activation. Yet, the theory draws different conclusions for differentiable activations functions and claims that for infinitely wide neural networks, the parameters should interpolate the data at convergence (Chizat and Bach, 2018). This result yet only holds for infinitely wide networks, and it remains unknown how wide a network should be to actually reach such an interpolation in practice. Figure 4 below presents experiments similar to Section 5, replacing the ReLU activation by the differentiable GeLU activation (Hendrycks and Gimpel, 2016). This activation is standard in modern large language models. Notably, it is used in the GPT2 architecture, which was used in the experiments of Raventós et al. (2024).



(a) Evolution of train loss.



(b) Evolution of test loss.

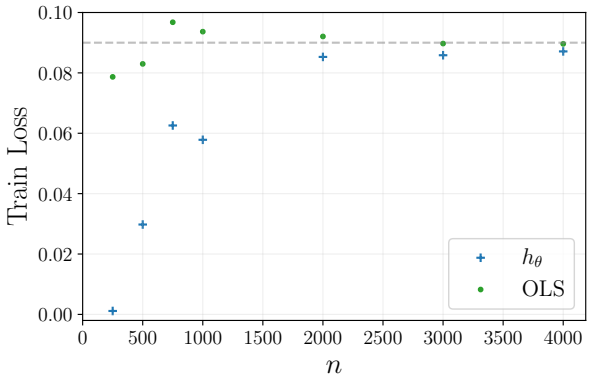
Figure 4: Evolution of both train and test losses at convergence with respect to the number of training samples, with GeLU activation.

While infinitely wide GeLU networks should overfit, even very wide networks ($m = 10\,000$) are far from this behavior in practice. In particular, we observe a phenomenon similar to Section 5 in Figure 4. Surprisingly, it even seems that interpolation is harder to reach with GeLU activation, as the network is already unable to interpolate for $n = 500$ training samples. We believe this is due to the fact that GeLU is close to a linear function around the origin (corresponding to our small initialization regime), making it harder to overfit noisy labels.

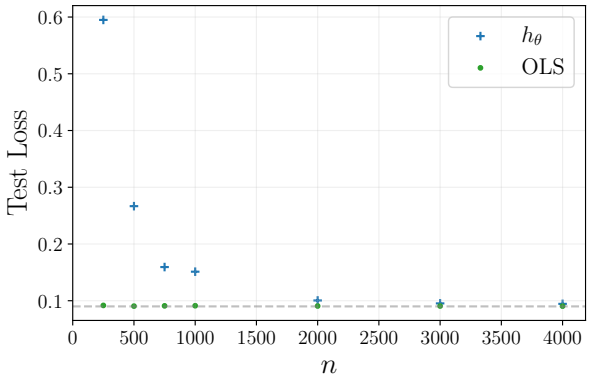
A.3. Momentum based optimizers

Our theoretical results hold for Gradient Flow, which is a first order approximation of typical gradient methods such as Gradient Descent (GD) or Stochastic Gradient Descent (SGD) (Li et al., 2019). Yet, recent large models implementations typically use different, momentum based algorithms, such as Adam (Kingma, 2014) or AdamW (Loshchilov and Hutter, 2019). To illustrate the generality of the optimization threshold we proved in a specific theoretical setting, we consider in Figure 5 below the same experiments as in Section 5, with the exception that i) we used GeLU activation functions (as in Appendix A.2) and ii) we minimized the training loss through the Adam optimizer, with pytorch default hyperparameters.

We focus on Adam rather than AdamW here to follow the experimental setup of Raventós et al. (2024) and because our focus is on implicit regularization, thus avoiding explicit regularization techniques.



(a) Evolution of train loss.



(b) Evolution of test loss.

Figure 5: Evolution of both train and test losses at convergence with respect to the number of training samples, with GeLU activation and Adam optimizer.

The observed results are very similar to the ones of Figure 4, leading to similar conclusions than Appendix A.2 and the fact that considering Adam rather than SGD does not significantly change the final results.

A.4. Stability of minima

Qiao et al. (2024) argue that the non-convergence of the estimator towards interpolation is due to the instability of global

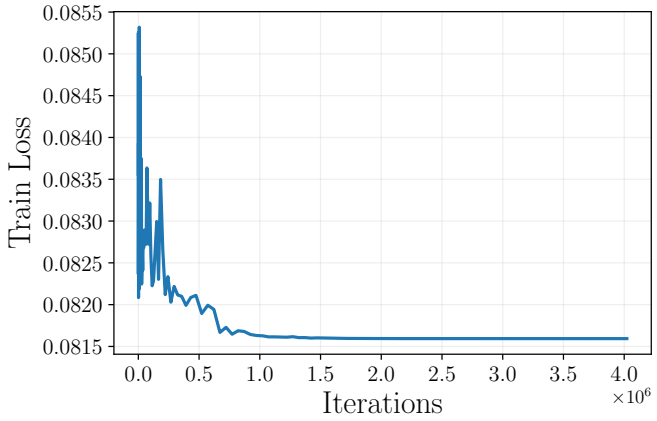


Figure 6: Evolution of training loss from warm restart with a decaying learning rate schedule ($n = 8000, d = 5$).

minima. More precisely they claim that for large stepsizes, gradient descent (GD) cannot stabilize around global minima of the loss for large values of n. We present an additional experiment in this section, illustrating that this non-convergence is not due to an instability of the convergence point of (S)GD, but to it being a stationary point of the loss as predicted by our theory.

For that, we consider a neural network initialized from the final point (warm restart) of training for $8\,000$ samples in the experiment of Figure 2.[4] We then continue training this network on the same training dataset, with a decaying learning rate schedule. Precisely, we start with a learning rate of 0.01 as in the main experiment, and multiply the learning rate by 0.85 every $50\,000$ iterations of SGD, so that after 4×10^6 iterations, the final learning rate is of order 10^{-8}.

We observe on Figure 6 that the training loss does not change much from the point reached at the end of training with the large learning rate 0.01. Indeed, the training loss was around 0.082 at the end of this initial training, which is slightly less than the noise level (0.09). While there seems to be some stabilization happening at the beginning of this decaying schedule, the training loss seems to converge to slightly more than 0.0815, confirming that the absence of interpolation is not due to an instability reason, but rather to a convergence towards a spurious stationary point of the loss.

A.5. Influence of dimensionality

Theorem 4.1 predicted an optimization threshold scaling in $\mathcal{O}\big(d^3 \log d\big)$. However, the experiments of Section 5 consider a fixed dimension ($d = 5$), making it unclear how tight is this theoretical optimization threshold and whether a similar dependency in the dimension is observed in practice. To investigate further this dependency in the dimension, we present in this section experiments in the same setup described in Appendix A.1, with the sole exception that the dimension is larger, fixed to $d = 10$.

Figure 7 illustrates the evolution of both the train and test losses as the number of training samples increases in this larger dimension setting. In that case, the optimization threshold seems much larger: interpolation stops happening around $n = 10\,000$ samples, and an estimation close to the OLS estimator really starts happening at much larger values of n, around $n = 80\,000$.

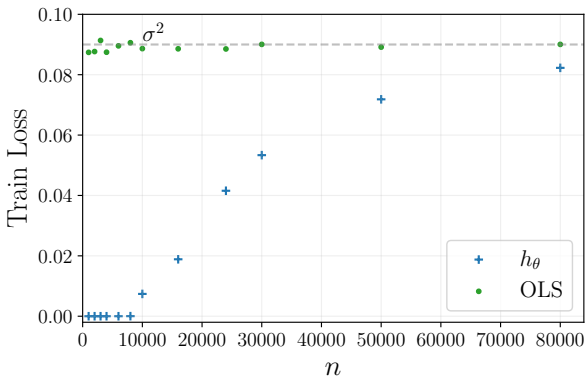
Comparing with the $d = 5$ case, it thus seems that the point at which interpolation stops indeed seems to roughly scale in d^3. However, this scaling seems even larger for the point where the estimator corresponds to the OLS one. We believe that this discrepancy is due to the differences between our theoretical and experimental setups, and in particular to the fact that multiple intermediate neurons can grow in our experimental setup (see *Limitations and generality* paragraph in Section 4.2).

A.6. 5 ReLU teacher network

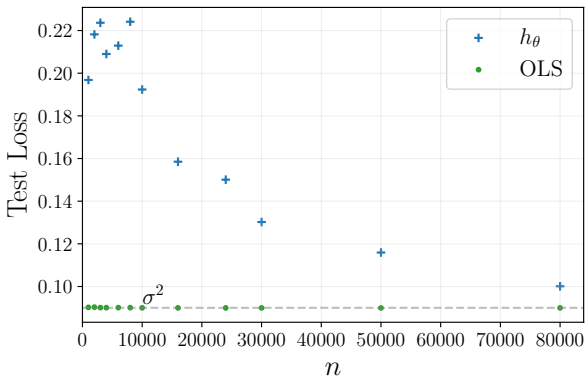
This section presents an additional experiment with a more complex data model. More precisely, we consider the exact same setup than Section 5 (described in Appendix A.1), with the difference that the labels y_k are given by

$$y_k = f^\star(x_k) + \eta_k,$$

[4] Another relevant experiment is to train from scratch (no warm restart) with a smaller learning rate. When running the experiment of Appendix A.1 with a smaller learning rate 0.001, we observe again that the parameters at convergence correspond to the OLS estimator.



(a) Evolution of train loss.



(b) Evolution of test loss.

Figure 7: Evolution of both train and test losses at convergence with respect to the number of training samples, with dimension $d = 10$.

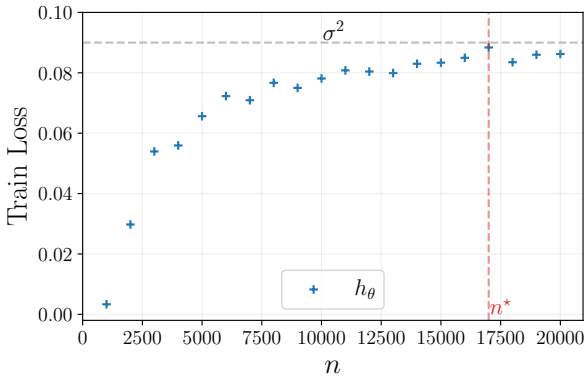
where $f^\star$ is a 5 ReLU network:

$$f^\star(x_k) = \frac{1}{5}\sum_{i=1}^{5}(x_k^\top \beta_i^\star)_+.$$

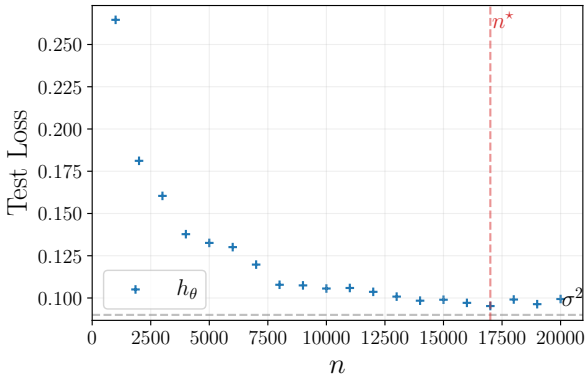
The parameters $\beta_i^\star$ are drawn i.i.d. at random following a standard Gaussian distribution. We use the exact same $\beta_i^\star$ across all the runs for different values of n. Also, x_k and η_k are generated in the same way as described in in Appendix A.1.

Figure 8 also presents the evolution of the train and test losses as the number of training samples varies. We observe a behavior similar to Figure 2, where interpolation is reached for small values of n, and is not reached anymore after some threshold $n^\star$. While the test loss is far from the optimal noise variance before this threshold, it then becomes close to it afterwards.

Yet, this transition from interpolation to generalization is slower in the 5 ReLU teacher case than in the linear one. Indeed, while interpolation does not happen anymore around $n = 2000$ in both cases, much more samples (around $n^\star = 17000$) are needed to have a simultaneously a training and testing loss close to the noise variance. These experiments suggest that the behavior predicted by Theorem 4.1 for a linear model also applies in more complex models such as the 5 ReLU teacher, but that the transition from interpolation to generalization can happen more slowly or with more training samples depending on the setting.



(a) Evolution of train loss.



(b) Evolution of test loss.

Figure 8: Evolution of both train and test losses at convergence with respect to the number of training samples with a 5 ReLU teacher. σ^2 corresponds to the noise variance $\mathbb{E}[\eta^2]$.

The slight difference with Figure 2 is that this optimization threshold here seems to appear for larger values of n.

B. Additional Discussions

Double descent. Double descent originally refers to the fact that the test loss obtained at convergence does not behave monotonically with the number of model parameters. Recently, different types of double descent have been proposed (Nakkiran et al., 2021). Notably, Henighan et al. (2023) study a data double descent, where the test loss follows a "double descent" shape when plotted against the number of training examples. The phenomenon we highlight here is different, as our test loss monotonically decreases with respect to the number of training points, as illustrated in both our experiments and Figure 1.

However, the toy experiments of Henighan et al. (2023) illustrate a similar phenomenon: for a sufficiently large number of training points, the training loss remains high while the model learns optimal features. It remains unclear though whether this high training loss stems from an underparametrized regime (i.e., the model lacks sufficient capacity to memorize the data) or if optimization fails to reach the empirical risk minimizer in their setup.

Feature learning and NTK regimes. We distinguish in this work between feature learning and the NTK/lazy regime, as they involve fundamentally different training dynamics (see Chizat et al., 2019, for an in-depth discussion). Our work specifically focuses on the feature learning regime with small initialization, as indicated by our initialization choice (Equation (2)), where both the inner and outer layers scale as $\frac{1}{\sqrt{m}}$.

In contrast, in the NTK/lazy regime (corresponding to large initialization scales), theory predicts that interpolation should occur at convergence, which is contrary to our main result. However, empirically demonstrating this interpolation in our toy model (with large n) is computationally challenging, as it would require an extremely large number of parameters.

C. Proof of Theorem 3.1

We recall Theorem 3.1 below.

Theorem 3.1. *If the marginal law of x_1 is continuous with respect to the Lebesgue measure, then for any $n \in \mathbb{N}$,*

$$\mathbb{E}_{\mathbf{X},\mathbf{y}}\Big[\sup_{w\in\mathbb{S}_{d-1}}\sup_{D_n\in\mathfrak{D}_n(w,\mathbf{0})}\|D_n - D(w)\|_2\Big] = \mathcal{O}\Big(\sqrt{\frac{d\log n}{n}\mathbb{E}[\|y_1x_1\|_2^2]}\Big),$$

where for any $w \in \mathbb{S}_{d-1}$, $D(w) = \mathbb{E}[\mathbb{1}_{w^\top x_1>0}y_1x_1]$.

Proof. We first show a similar result on the following expectation

$$\mathbb{E}_{\mathbf{X},\mathbf{y}}\left[\sup_{w\in\mathbb{S}_{d-1}}\|D_n(w) - D(w)\|_2\right] = \mathcal{O}\left(\sqrt{\frac{d\log n}{n}\mathbb{E}[\|yx\|_2^2]}\right), \tag{9}$$

where we recall $D_n(w) = \frac{1}{n}\sum_{k=1}^n \mathbb{1}_{w^\top x_k>0}y_kx_k$. We bound this expectation using typical uniform bound techniques for empirical processes.

A symmetrization argument allows to show, for i.i.d. Rademacher random variables $\varepsilon_k \in \{-1, 1\}$ (see Van Der Vaart and Wellner, 2023, Lemma 2.3.1.):

$$\mathbb{E}_{\mathbf{X},\mathbf{y}}\left[\sup_{w\in\mathbb{S}_{d-1}}\|D_n(w) - D(w)\|_2\right] \le 2\,\mathbb{E}_{\mathbf{X},\mathbf{y}}\left[\mathbb{E}_{\boldsymbol{\varepsilon}}\left[\sup_{w\in\mathbb{S}_{d-1}}\left\|\frac{1}{n}\sum_{k=1}^n \mathbb{1}_{w^\top x_k>0}\varepsilon_k y_k x_k\right\|_2 \mid \mathbf{X},\mathbf{y}\right]\right]. \tag{10}$$

From there, it remains to bound for any value of $\mathbf{X}, \mathbf{y}$ the conditioned expectation $\mathbb{E}_{\boldsymbol{\varepsilon}}[\cdot \mid \mathbf{X}, \mathbf{y}]$. We consider in the following a fixed value of $\mathbf{X}, \mathbf{y}$. Note that the vector $\sum_{k=1}^n \mathbb{1}_{w^\top x_k>0}\varepsilon_k y_k x_k$, actually only depends on w in the value of the vector $(\mathbb{1}_{w^\top x_k>0})_{k\in[n]}$. Define

$$\mathcal{A}(\mathbf{X},\mathbf{y}) = \left\{(\mathbb{1}_{w^\top x_k>0})_{k\in[n]} \mid w \in \mathbb{R}^d\right\}. \tag{11}$$

We thus have the equality:

$$\sup_{w\in\mathbb{S}_{d-1}}\left\|\frac{1}{n}\sum_{k=1}^n \mathbb{1}_{w^\top x_k>0}\varepsilon_k y_k x_k\right\|_2 = \sup_{\mathbf{u}\in\mathcal{A}(\mathbf{X},\mathbf{y})}\left\|\frac{1}{n}\sum_{k=1}^n u_k\varepsilon_k y_k x_k\right\|_2.$$

Moreover, classical geometric arguments (see e.g. Cover, 1965, Theorem 1) allow to bound $\operatorname{card}(\mathcal{A}(\mathbf{X},\mathbf{y}))$ for any $\mathbf{X},\mathbf{y}$:

$$\begin{aligned}\operatorname{card}(\mathcal{A}(\mathbf{X},\mathbf{y})) &\leq 2\sum_{k=0}^{d-1}\binom{n-1}{k}\\ &= \mathcal{O}\big(n^d\big).\end{aligned} \tag{12}$$

From there, we will bound individually for each $\mathbf{u}\in\mathcal{A}(\mathbf{X},\mathbf{y})$ the norm of $\frac{1}{n}\sum_{k=1}^n u_k\varepsilon_k y_k x_k$ and use a union bound argument.

Let $\mathbf{u}\in\mathcal{A}(\mathbf{X},\mathbf{y})$. Define $Z\in\mathbb{R}^{d\times n}$ the matrix whose column k is given by $Z^{(k)}=\frac{1}{n}u_k y_k x_k$. Then note that $\frac{1}{n}\sum_{k=1}^n u_k\varepsilon_k y_k x_k = Z\boldsymbol{\varepsilon}$. Hanson-Wright inequality then allows to bound the following probability (see Rudelson and Vershynin, 2013, Theorem 2.1) for some universal constant $c>0$ and any $t\geq 0$:

$$\mathbb{P}_{\boldsymbol{\varepsilon}}\left(\big|\|Z\boldsymbol{\varepsilon}\|_2-\|Z\|_{\mathrm{F}}\big|>t\;\middle|\;\mathbf{X},\mathbf{y}\right)\leq 2e^{-\frac{ct^2}{\|Z\|_{\mathrm{op}}^2}},$$

where $\|Z\|_{\mathrm{F}}$ and $\|Z\|_{\mathrm{op}}$ respectively denote the Frobenius and operator norm of Z. In particular, noting that $\|Z\|_{\mathrm{op}}\leq\|Z\|_{\mathrm{F}}$, this last equation implies that for any $t>0$

$$\mathbb{P}_{\boldsymbol{\varepsilon}}\left(\|Z\boldsymbol{\varepsilon}\|_2>(1+t)\|Z\|_{\mathrm{F}}\;\middle|\;\mathbf{X},\mathbf{y}\right)\leq 2e^{-ct^2}. \tag{13}$$

Moreover, note that

$$\begin{aligned}\|Z\|_{\mathrm{F}} &= \sqrt{\sum_{k=1}^n\|Z^{(k)}\|_2^2}\\ &= \sqrt{\sum_{k=1}^n\frac{1}{n^2}\|u_k y_k x_k\|_2^2}\leq\sqrt{\frac{1}{n}C(Z)},\end{aligned}$$

where $C(Z)=\frac{1}{n}\sum_{k=1}^n\|y_k x_k\|_2^2$ does not depend on $\mathbf{u}$.

Rewriting Equation (13) with this last inequality, and with $\delta=2e^{-ct^2}$, we finally have for each $\mathbf{u}\in\mathcal{A}(\mathbf{X},\mathbf{y})$:

$$\mathbb{P}_{\boldsymbol{\varepsilon}}\left(\|Z\boldsymbol{\varepsilon}\|_2>(1+\sqrt{\frac{1}{c}\ln(2/\delta)})\sqrt{\frac{C(Z)}{n}}\;\middle|\;\mathbf{X},\mathbf{y}\right)\leq\delta.$$

Considering a union bound over all the $\mathbf{u}\in\mathcal{A}(\mathbf{X},\mathbf{y})$, we have for some universal constant $c'>0$, thanks to Equation (12):

$$\mathbb{P}_{\boldsymbol{\varepsilon}}\left(\exists\mathbf{u}\in\mathcal{A}(\mathbf{X},\mathbf{y}),\|Z\boldsymbol{\varepsilon}\|_2>\left(1+\sqrt{c'(\log(2/\delta)+d\log(n)+1)}\right)\sqrt{\frac{C(Z)}{n}}\;\middle|\;\mathbf{X},\mathbf{y}\right)\leq\delta. \tag{14}$$

Moreover, conditioned on $\mathbf{X},\mathbf{y}$, $\|Z\boldsymbol{\varepsilon}\|_2$ is almost surely bounded by $\sqrt{n}\|Z\|_{\mathrm{op}}$, and so by $\sqrt{C(Z)}$. A direct bound on the expectation can then be derived using Equation (14) with $\delta=n^{-d}$:

$$\mathbb{E}_{\boldsymbol{\varepsilon}}\left[\sup_{\mathbf{u}\in\mathcal{A}(\mathbf{X},\mathbf{y})}\left\|\frac{1}{n}\sum_{k=1}^n u_k\varepsilon_k y_k x_k\right\|_2\right]=\mathcal{O}\left(\sqrt{\frac{d\log n}{n}+n^{-d}}\right)\sqrt{C(Z)}.$$

Wrapping up with Equation (11) and Equation (10) then allows to derive Equation (9),

$$\begin{aligned}\mathbb{E}_{\mathbf{X},\mathbf{y}}\left[\sup_{w\in\mathbb{S}_{d-1}}\|D_n(w)-D(w)\|_2\right] &= \mathcal{O}\left(\sqrt{\frac{d\log n}{n}}\right)\mathbb{E}_{\mathbf{X},\mathbf{y}}\left[\sqrt{C(Z)}\right]\\ &\leq \mathcal{O}\left(\sqrt{\frac{d\log n}{n}}\right)\sqrt{\mathbb{E}_{\mathbf{X},\mathbf{y}}\left[C(Z)\right]}\\ &= \mathcal{O}\left(\sqrt{\frac{d\log n}{n}}\right)\sqrt{\mathbb{E}\left[\|yx\|_2^2\right]}.\end{aligned}$$

Lemma C.1 below then allows to conclude.

Lemma C.1. *If the marginal law of x is continuous with respect to the Lebesgue measure, then almost surely:*

$$\sup_{w\in\mathbb{S}_{d-1}} \sup_{D_n\in\mathfrak{D}_n(w)} \|D_n - D(w)\|_2 \leq \sup_{w\in\mathbb{S}_{d-1}} \|D_n(w) - D(w)\|_2,$$

where $D_n(w) = \frac{1}{n}\sum_{k=1}^n \mathbb{1}_{w^\top x_k>0} y_k x_k$.

□

C.1. Proof of Lemma C.1

First observe that if the marginal law of x is continuous, then D is continuous with respect to w.

Consider any $w \in \mathbb{S}_{d-1}$. We recall that the set $\mathfrak{D}_n(w)$ is defined as

$$\mathfrak{D}_n(w) = \Big\{ -\frac{1}{n}\sum_{k=1}^n \eta_k y_k x_k \;\Big|\; \forall k \in [n], \eta_k \begin{cases} \in [0,1] \text{ if } \langle w_j^t, x_k\rangle = 0 \\ = 1 \text{ if } \langle w_j^t, x_k\rangle > 0 \\ = 0 \text{ otherwise} \end{cases} \Big\}.$$

If all the values $w^\top x_k$ are non-zero, then $\mathfrak{D}_n(w)$ is the singleton given by $D_n(w)$ and thus

$$\sup_{D_n\in\mathfrak{D}_n(w)} \|D_n - D(w)\|_2 = \|D_n(w) - D(w)\|_2.$$

Otherwise, if $w^\top x_k = 0$ for at least one k, observe that[5]

$$\mathfrak{D}_n(w) = \liminf_{\substack{\varepsilon\to 0\\ \varepsilon>0}} \mathrm{Conv}(\{D_n(w') \mid w' \in \mathcal{S} \text{ and } \|w - w'\|_2 \leq \varepsilon\}),$$

where

$$\mathcal{S} = \{w' \in \mathbb{S}_{d-1} \mid {w'}^\top x_k \neq 0 \text{ for all } k\}.$$

In other words, for any $D_n \in \mathfrak{D}_n(w)$, w can be approached arbitrarily closed by vectors $w_i \in \mathcal{S}$ such that for some convex combination $\boldsymbol{\eta}$,

$$D_n = \sum_i \eta_i D_n(w_i).$$

From then, it comes that

$$\begin{aligned}
\|D_n - D(w)\| &\leq \sum_i \eta_i \|D_n(w_i) - D(w)\| \\
&\leq \sum_i \eta_i (\|D_n(w_i) - D(w_i)\| + \|D(w_i) - D(w)\|) \\
&\leq \sup_{w'\in\mathcal{S}} \|D_n(w') - D(w')\| + \sum_i \eta_i \|D(w_i) - D(w)\|.
\end{aligned}$$

Since D is continuous and the w_i can be chosen arbitrarily close to w, the right sum can be chosen arbitrarily close to 0.

In particular, we have shown that for any $D_n \in \mathfrak{D}_n(w)$,

$$\|D_n - D(w)\| \leq \sup_{w'\in\mathcal{S}} \|D_n(w') - D(w')\|.$$

This concludes the proof of Lemma C.1. □

D. Probability tail bound version of Theorem 3.1

While Theorem 3.1 bounds the maximal deviation of $D_n - D(w)$ in expectation, a high probability tail bound is also possible, as given by Theorem D.1 below.

[5] This observation directly follows from the definition of the Clarke subdifferential.

Theorem D.1. *If the marginal law of x is continuous with respect to the Lebesgue measure, then for any $n \in \mathbb{N}$ and $M \geq \mathbb{E}\left[\|yx\|^2\right]$,*

$$\mathbb{P}_{\mathbf{X},\mathbf{y}}\left[\sup_{w\in\mathbb{S}_{d-1}}\sup_{D_n\in\mathfrak{D}_n(w)}\|D_n - D(w)\|_2 > 4\left(1+\sqrt{c'(\log(2/\delta)+d\log(n)+1)}\right)\sqrt{\frac{M}{n}}\right] \leq \frac{4}{3}\delta + \frac{4}{3}\mathbb{P}_{\mathbf{X},\mathbf{y}}\left[\frac{1}{n}\sum_{k=1}^{n}\|y_k x_k\|^2 > M\right].$$

Proof. The proof follows the same lines as the proof of Theorem 3.1 in Appendix C. In particular, we first want to bound in probability the term $\sup_{w\in\mathbb{S}_{d-1}}\|D_n(w) - D(w)\|_2$. To this end, a probabilistic symmetrization argument (Van Der Vaart and Wellner, 2023, Lemma 2.3.7.) yields for any $t > 0$

$$\beta_n(y)\mathbb{P}_{\mathbf{X},\mathbf{y}}\left[\sup_{w\in\mathbb{S}_{d-1}}\|D_n(w) - D(w)\|_2 > t\right] \leq 2\mathbb{P}_{\mathbf{X},\mathbf{y}}\left[\mathbb{P}_{\boldsymbol{\varepsilon}}\left[\sup_{w\in\mathbb{S}_{d-1}}\left\|\frac{1}{n}\sum_{k=1}^{n}\varepsilon_k \mathbb{1}_{w^\top x_k>0} y_k x_k\right\| > \frac{t}{4} \mid \mathbf{X},\mathbf{y}\right]\right], \quad (15)$$

where $\beta_n(t) = 1 - \frac{4n}{t^2}\sup_{w,w'\in\mathbb{S}_{d-1}}\mathrm{Var}(\mathbb{1}_{w^\top x>0} y w'^\top x)$. In particular here, $\beta_n(t) \geq 1 - \frac{4n}{t^2}\mathbb{E}\left[\|yx\|^2\right]$. Moreover, we already showed Equation (14) in the proof of Theorem 3.1, which states

$$\mathbb{P}_{\boldsymbol{\varepsilon}}\left(\sup_{w\in\mathbb{S}_{d-1}}\left\|\frac{1}{n}\sum_{k=1}^{n}\varepsilon_k \mathbb{1}_{w^\top x_k>0} y_k x_k\right\| > \left(1+\sqrt{c'(\log(2/\delta)+d\log(n)+1)}\right)\sqrt{\frac{C(\mathbf{X},\mathbf{y})}{n}} \mid \mathbf{X},\mathbf{y}\right) \leq \delta,$$

where $C(\mathbf{X},\mathbf{y}) = \frac{1}{n}\sum_{k=1}^{n}\|y_k x_k\|^2$.

Equation (15) then rewrites for any $M > 0$:

$$\mathbb{P}_{\mathbf{X},\mathbf{y}}\left[\sup_{w\in\mathbb{S}_{d-1}}\|D_n(w) - D(w)\|_2 > 4\left(1+\sqrt{c'(\log(2/\delta)+d\log(n)+1)}\right)\sqrt{\frac{M}{n}}\right] \leq \beta_n(t)^{-1}\left(\delta + \mathbb{P}_{\mathbf{X},\mathbf{y}}\left[\frac{1}{n}\sum_{k=1}^{n}\|y_k x_k\|^2 > M\right]\right),$$

with

$$t = 4\left(1+\sqrt{c'(\log(2/\delta)+d\log(n)+1)}\right)\sqrt{\frac{M}{n}}.$$

Note that for any $M \geq \mathbb{E}\left[\|yx\|^2\right]$, $\beta_n(t) \geq \frac{3}{4}$, which implies

$$\mathbb{P}_{\mathbf{X},\mathbf{y}}\left[\sup_{w\in\mathbb{S}_{d-1}}\|D_n(w) - D(w)\|_2 > 4\left(1+\sqrt{c'(\log(2/\delta)+d\log(n)+1)}\right)\sqrt{\frac{M}{n}}\right] \leq \frac{4}{3}\delta + \frac{4}{3}\mathbb{P}_{\mathbf{X},\mathbf{y}}\left[\frac{1}{n}\sum_{k=1}^{n}\|y_k x_k\|^2 > M\right].$$

Theorem D.1 then follows, thanks to Lemma C.1. □

Corollary D.1 below provides a simpler tail bound, directly applying Lemma C.1 with Chebyshev's inequality to bound $\mathbb{P}_{\mathbf{X},\mathbf{y}}\left[\frac{1}{n}\sum_{k=1}^{n}\|y_k x_k\|^2 > M\right]$. Stronger tail bounds can be provided with specific conditions on the random variables x_k and y_k, but the one of Corollary D.1 is enough for our use in Section 4.

Corollary D.1. *Assume the marginal law of x is continuous with respect to the Lebesgue measure. Moreover, assume $\|xy\|$ admits a fourth moment. Then*

$$\mathbb{P}_{\mathbf{X},\mathbf{y}}\left[\sup_{w\in\mathbb{S}_{d-1}}\sup_{D_n\in\mathfrak{D}_n(w)}\|D_n - D(w)\|_2 > 4\left(1+\sqrt{c'(\log(2/\delta)+d\log(n)+1)}\right)\sqrt{\frac{2\mathbb{E}[\|yx\|^2]}{n}}\right]$$

$$\leq \frac{4}{3}\delta + \frac{4}{3}\frac{\mathbb{E}[\|yx\|^4]}{n\mathbb{E}[\|yx\|^2]^2}.$$

Proof. This is a direct consequence of Theorem D.1, using Chebyshev's inequality to bound $\mathbb{P}_{\mathbf{X},\mathbf{y}}\left[\frac{1}{n}\sum_{k=1}^n \|y_k x_k\|^2 > M\right]$. □

E. Proof of Proposition 3.1

In the following proof, we define the following subsets of the unit sphere in dimension d for any $\delta > 0$:

$$\mathcal{H} = \{w \in \mathbb{S}_{d-1} \mid D(w) \text{ satisfies Equation (6)}\},$$

$$\mathcal{H}(\delta) = D^{-1}\left(\bigcup_{w\in\mathcal{H}} B(D(w),\delta)\right) \cap \mathbb{S}_{d-1},$$

$$\Delta(\delta) = \min(1, \inf_{w\in\mathbb{S}_{d-1}\setminus\mathcal{H}(\delta)} \min\left(\|\frac{D(w)}{\|D(w)\|} - w\|, \|\frac{D(w)}{\|D(w)\|} + w\|, \|D(w)\|\right)).$$

Here, $B(D(w),\delta)$ denotes the open ball of radius δ, centered in $D(w)$.

Proof. Since the marginal distribution of X is continuous, the function $D : w \mapsto D(w)$ is continuous. In particular for any $\delta > 0$, the infimum defining $\Delta(\delta)$ is reached, so that $\Delta(\delta) > 0$ by definition of $\mathcal{H}$. In the following, we let $\delta = \frac{\varepsilon}{2}$. Thanks to Corollary D.1, with probability at least $1 - \mathcal{O}_\mu\left(\frac{1}{n}\right)$,

$$\sup_{w\in\mathbb{S}_{d-1}} \sup_{D_n\in\mathfrak{D}_n(w)} \|D_n - D(w)\|_2 = \mathcal{O}_\mu\left(\sqrt{\frac{d\log n}{n}}\right).$$

In particular, we can choose $n^\star(\varepsilon) = \mathcal{O}_\mu\left(\frac{d\log\left(\frac{d}{\min(\Delta(\frac{\varepsilon}{2})^4,\varepsilon^2)}\right)}{\min}(\Delta(\frac{\varepsilon}{2})^2,\varepsilon)\right)$ large enough so that for any $n \geq n^\star(\varepsilon)$, with probability at least $1 - \mathcal{O}_\mu\left(\frac{1}{n}\right)$:

$$\sup_{w\in\mathbb{S}_{d-1}} \sup_{D_n\in\mathfrak{D}_n(w)} \|D_n - D(w)\|_2 \leq \frac{1}{2}\min\left(\Delta(\frac{\varepsilon}{2})^2, \varepsilon\right). \tag{16}$$

We assume in the following of the proof that Equation (16) holds.

Consider an extremal vector D_n of the finite data $(\mathbf{X},\mathbf{y})$. By definition, there is some $w \in \mathbb{S}_{d-1}$ such that $D \in \mathcal{D}_n(w)$ and either

1. $D_n = \mathbf{0}$,
2. or $A_n(D_n) = A_n(w)$,
3. or $A_n(D_n) = -A_n(w)$.

In the first case, Equation (16) yields that $\|D(w)\|_2 < \Delta(\frac{\varepsilon}{2})$. Necessarily, by definition of $\Delta(\frac{\varepsilon}{2})$, $w \in \mathcal{H}(\frac{\varepsilon}{2})$. This means by definition of $\mathcal{H}(\frac{\varepsilon}{2})$ that there exists $D^\star \in \mathbb{R}^d$ satisfying Equation (6), such that

$$\|D(w) - D^\star\|_2 \leq \frac{\varepsilon}{2}.$$

In particular, using Equation (16) again yields $\|D_n - D^\star\|_2 \leq \varepsilon$.

In the second case ($A_n(D_n) = A_n(w)$), we can assume $D_n \neq 0$. In that case, as $\frac{D_n}{\|D_n\|}$ have the same activations, $\mathcal{D}_n(w) = \mathcal{D}_n(\frac{D_n}{\|D_n\|})$, i.e., we can assume without loss of generality that $w = \frac{D_n}{\|D_n\|}$ here. Similarly to the first case, if $w \in \mathcal{H}(\frac{\varepsilon}{2})$, then there exists $D^\star \in \mathbb{R}^d$ satisfying Equation (6), such that $\|D_n - D^\star\|_2 \leq \varepsilon$.

Let us show by contradiction that indeed $w \in \mathcal{H}(\frac{\varepsilon}{2})$. Assume $w \notin \mathcal{H}(\frac{\varepsilon}{2})$. In particular, $\|D(w)\|_2 \geq \Delta(\frac{\varepsilon}{2})$. We can now bound the norm of $\frac{D(w)}{\|D(w)\|} - w$:

$$\begin{aligned}
\left\| \frac{D(w)}{\|D(w)\|} - w \right\|_2 &= \left\| \frac{D(w)}{\|D(w)\|} - \frac{D_n}{\|D_n\|} \right\|_2 \\
&= \left\| \frac{D(w) - D_n}{\|D(w)\|} + \frac{D_n}{\|D_n\|} \left(\frac{\|D_n\| - \|D(w)\|}{\|D(w)\|} \right) \right\|_2 \\
&\leq \frac{\|D(w) - D_n\|}{\|D(w)\|} + \frac{\big|\|D_n\| - \|D(w)\|\big|}{\|D(w)\|} \\
&\leq 2 \frac{\|D(w) - D_n\|}{\|D(w)\|} \\
&\leq 2 \frac{\Delta(\frac{\varepsilon}{2})^2}{2\Delta(\frac{\varepsilon}{2})} = \Delta(\frac{\varepsilon}{2}).
\end{aligned}$$

By definition of $\Delta(\frac{\varepsilon}{2})$, this actually implies that $w \in \mathcal{H}(\frac{\varepsilon}{2})$, which contradicts the initial assumption. We thus indeed have $w \in \mathcal{H}(\frac{\varepsilon}{2})$, leading to the existence of a $D^\star$ with the wanted properties such that $\|D_n - D^\star\|_2 \leq \varepsilon$. In the third case ($A_n(D_n) = -A_n(w)$), symmetric arguments lead to the same conclusion, which concludes the proof of Proposition 3.1. □

F. Proof of Theorem 4.1

F.1. Notations and first classical results

In the whole Appendix F, we define $\Sigma = \mathbb{E}[xx^\top]$ and

$$\Sigma_{n,+} = \frac{1}{n} \sum_{k \in \mathcal{S}_+} x_k x_k^\top \quad , \quad \Sigma_{n,-} = \frac{1}{n} \sum_{k \in \mathcal{S}_-} x_k x_k^\top$$

and the following set of neurons:

$$\mathcal{I}_+ = \{i \in [m] \mid a_i(0) \geq 0\} \quad \text{and} \quad \mathcal{I}_- = \{i \in [m] \mid a_i(0) < 0\}.$$

We first start by stating the following, known balancedness lemma (see, e.g., Arora et al., 2019; Boursier et al., 2022).

Lemma F.1. *For any $i \in [m]$ and $t \in \mathbb{R}_+$, $a_i(t)^2 - \|w_i(t)\|^2 = a_i(0)^2 - \|w_i(0)\|^2$.*

Lemma F.1 can be simply proved by a direct computation of the derivative of $a_i(t)^2 - \|w_i(t)\|^2$. Thanks to Equation (8), this yields that the sign $a_i(t)$ is constant over time, and thus partitioned by the sets of neurons $\mathcal{I}_+$ and $\mathcal{I}_-$.

Also, note that with probability $1 - \frac{1}{2^{m-1}}$, the sets $\mathcal{I}_+$ and $\mathcal{I}_-$ are both non empty, which is assumed to hold in the following of the section.

In this section, all the $\mathcal{O}, \Theta$ and Ω notations hide constants depending on the fourth moment of η, the norm of $\beta^\star$ and the constant c of Assumption 4.1. Note that due to the sub-Gaussian property of x, its k-th moment can be bounded as $\mathbb{E}[\|x\|^k]\mathcal{O}\Big(d^{\frac{k}{2}}\Big)$ for any k.

F.2. Phase 1: early alignment

Lemma F.2. *If Assumption 4.1 holds, there exists $\lambda^\star = \Theta(\frac{1}{d})$ and $n^\star = \Theta(d^3 \log d)$ such that for any $\lambda \leq \lambda^\star$ and $\varepsilon \in (0, \frac{1}{4})$, $n \geq n^\star$ and for $\tau = \frac{\varepsilon \ln(1/\lambda)}{\|\Sigma\beta^\star\|}$, with probability $1 - \mathcal{O}(\frac{1}{n})$:*

1. *output weights do not change until τ:*

$$\forall t \leq \tau, \forall j \in [m], |a_j(0)|\lambda^{2\varepsilon} \leq |a_j(t)| \leq |a_j(0)|\lambda^{-2\varepsilon};$$

2. *all neurons align with $\pm\Sigma\beta^\star$:*

$$\forall i \in [m], \quad \langle \frac{w_i(\tau)}{a_i(\tau)}, \Sigma\beta^\star \rangle = \|\Sigma\beta^\star\| - \mathcal{O}\left(\lambda^\varepsilon + \sqrt{\frac{d^2 \log n}{n}} \right).$$

Proof. We start the proof by computing $D(w)$ for any $w \in \mathbb{S}_{d-1}$:

$$\begin{aligned} D(w) &= \mathbb{E}[\mathbb{1}_{w^\top x>0} yx] \\ &= \mathbb{E}[\mathbb{1}_{w^\top x>0} xx^\top \beta^\star] \\ &= \frac{1}{2}\left(\mathbb{E}[\mathbb{1}_{w^\top x>0} xx^\top] + \mathbb{E}[\mathbb{1}_{w^\top x<0} xx^\top]\right)\beta^\star \\ &= \frac{\Sigma\beta^\star}{2}. \end{aligned}$$

The second inequality comes from the independence between x and η, the third one comes from the symmetry of the distribution of x and the last one by continuity of this distribution.

Corollary D.1 additionally implies that for some $n^\star = \Theta(d^3 \log d)$ and any $n \geq n^\star$, with probability at least $1 - \mathcal{O}(\frac{1}{n})$,

$$\sup_{w\in\mathbb{S}_{d-1}} \sup_{D_n\in\mathfrak{D}_n(w)} \|D_n - D(w)\|_2 \leq \mathcal{O}\left(\sqrt{\frac{d^2 \log n}{n}}\right) \leq \frac{\alpha}{\sqrt{d}}, \tag{17}$$

with $\alpha = \frac{1}{4}\min(c - \sqrt{d}\|\Sigma\beta^\star - \beta^\star\|, \|\Sigma\beta^\star\|)$. Note[6] that $\alpha > 0$ thanks to the fourth point of Assumption 4.1. Moreover, using typical concentration inequality for sub-Gaussian vectors, we also have with probability $1 - \mathcal{O}(\frac{1}{n})$:

$$\frac{\sum_{k=1}^n \|x_k\|^2}{n} \leq 2\mathbb{E}_{\mu_X}[\|x\|^2] = \mathcal{O}(d)\,. \tag{18}$$

We assume in the following of the proof that both Equations (17) and (18) hold.

Since $D(w) = \frac{\Sigma\beta^\star}{2}$ for any w, we have for any $w \in \mathbb{S}_{d-1}$, $D_n \in \mathfrak{D}_n(w)$ and $k \in \mathcal{S}_+$:

$$\begin{aligned} x_k^\top D_n &= x_k^\top (D_n - D(w)) + \frac{1}{2}x_k^\top(\Sigma\beta^\star - \beta^\star) + \frac{1}{2}x_k^\top\beta^\star \\ &\geq \|x_k\|\left(-\|D_n - D(w)\| - \frac{1}{2}\|\Sigma\beta^\star - \beta^\star\| + \frac{c}{2\sqrt{d}}\right) \\ &\geq \frac{\alpha}{\sqrt{d}}\|x_k\| > 0. \end{aligned}$$

Similarly for any $k \in \mathcal{S}_-$, $x_k^\top D_n < 0$. This directly implies here that there are only two extremal vectors here:

$$\begin{aligned} D_n(\beta^\star) &= \Sigma_{n,+}\beta^\star + \frac{1}{n}\sum_{k\in\mathcal{S}_+}\eta_k x_k, \\ D_n(-\beta^\star) &= \Sigma_{n,-}\beta^\star + \frac{1}{n}\sum_{k\in\mathcal{S}_-}\eta_k x_k. \end{aligned} \tag{19}$$

We can now show, similarly to Boursier and Flammarion (2024), the early alignment phenomenon in the first phase.[7]

1. First note that Equation (17) and the definition of α imply that for any w:

$$\|D_n(w)\| \leq \|\Sigma\beta^\star\|. \tag{20}$$

We define $t_1 = \min\{t \geq 0 \mid \sum_{j=1}^m a_j(t)^2 \geq \lambda^{2-4\varepsilon}\}$.

For any $i \in [m]$ and $t \in [0, t_1]$, Equation (4) rewrites:

$$\begin{aligned} \left|\frac{\mathrm{d}a_i(t)}{\mathrm{d}t}\right| &= \left|w_i(t)^\top D_n^i(t)\right| \\ &\leq |a_i(t)|\left(\max_{w\in\mathbb{S}_{d-1}}\|D_n(w)\| + \frac{\sum_{k=1}^n\|x_k\|^2}{n}\lambda^{2-4\varepsilon}\right) \end{aligned}$$

[6] The additional d dependence comes from the expectation of $\|yx\|^2$ in the square root. Additionally, the probability bound comes from the fact that $\frac{\mathbb{E}_\mu[\|yx\|^4]}{\mathbb{E}_\mu[\|yx\|^2]^2} = \mathcal{O}(1)$ here.

[7] We could directly reuse Theorem 1 from Boursier and Flammarion (2024) here, but it would not allow us to choose an initialization scale $\lambda^\star$ that does not depend on n.

$$\leq |a_i(t)| \left(\|\Sigma\beta^\star\| + 2\mathbb{E}[\|x\|^2]\lambda^{2-4\varepsilon} \right).$$

As a consequence, a simple Grönwall argument yields that for any $t \in [0, t_1]$:

$$|a_i(t)| \leq |a_i(0)| \exp(t\|\Sigma\beta^\star\| + 2t\mathbb{E}[\|x\|^2]\lambda^{2-4\varepsilon}).$$

In particular, for our choice of τ, for a small enough $\lambda^\star = \mathcal{O}\left(d^{-\frac{1}{2-4\varepsilon}}\right)$, for any $t \leq \min(\tau, t_1)$:

$$|a_i(t)| < |a_i(0)|\lambda^{-\varepsilon}. \tag{21}$$

Note that this implies that $t < t_1$, i.e., $\tau < t_1$. As a consequence. Moreover, we can also show that $|a_i(t)| > |a_i(0)|\lambda^\varepsilon$ for any $t \leq \tau$, which implies the first point of Lemma F.2.

2. For the second point, let $i \in \mathcal{I}_+$ and denote $\overline{w}_i(t) = \frac{w_i(t)}{a_i(t)}$. Thanks to Lemma F.1, $\overline{w}_i(t) \in B(0,1)$ and $a_i(t)$ is of constant sign. Also, for almost any $t \in [0, \tau]$:

$$\frac{\mathrm{d}\overline{w}_i(t)}{\mathrm{d}t} \in \mathcal{D}_n(w_i(t), \theta(t)) - \langle \overline{w}_i(t), \mathcal{D}_n(w_i(t), \theta(t)) \rangle \overline{w}_i(t).$$

Since $a_i(t) > 0$ for $i \in \mathcal{I}_+$,

$$\begin{aligned}
\frac{\mathrm{d}\langle \overline{w}_i(t), \Sigma\beta^\star \rangle}{\mathrm{d}t} &\in \langle \mathcal{D}_n(w_i(t), \theta(t)), \Sigma\beta^\star \rangle - \langle \overline{w}_i(t), \mathcal{D}_n(w_i(t), \theta(t)) \rangle \langle \overline{w}_i(t), \Sigma\beta^\star \rangle \\
&\geq \inf_{D_n \in \mathfrak{D}_n(w_i(t), \theta(t))} \langle D_n, \Sigma\beta^\star \rangle - \langle \overline{w}_i(t), D_n \rangle \langle \overline{w}_i(t), \Sigma\beta^\star \rangle \\
&\geq \inf_{D_n \in \mathfrak{D}_n(w_i(t))} \langle D_n, \Sigma\beta^\star \rangle - \langle \overline{w}_i(t), D_n \rangle \langle \overline{w}_i(t), \Sigma\beta^\star \rangle - 2\|\Sigma\beta^\star\|\lambda^{2-4\varepsilon} \\
&\geq \langle D(w_i(t)), \Sigma\beta^\star \rangle - \langle \overline{w}_i(t), D(w_i(t)) \rangle \langle \overline{w}_i(t), \Sigma\beta^\star \rangle \\
&\quad - 2\|\Sigma\beta^\star\| \left(\lambda^{2-4\varepsilon} + \sup_{D_n \in \mathcal{D}_n(w_i(t)} \|D_n - D(w_i(t))\| \right) \\
&\geq \frac{1}{2} \left(\|\Sigma\beta^\star\|^2 - \langle \overline{w}_i(t), \Sigma\beta^\star \rangle^2 \right) - \mathcal{O}\left(\lambda^{2-4\varepsilon} + \sqrt{\frac{d^2 \log n}{n}} \right).
\end{aligned}$$

Solutions of the ODE $f'(t) = a^2 - f(t)^2$ with $f(0) \in (-a, a)$ are of the form $f(t) = a\tanh(a(t + t_0))$ for some $t_0 \in \mathbb{R}$. By Grönwall comparison, we thus have

$$\langle \overline{w}_i(t), \Sigma\beta^\star \rangle \geq a\tanh(\frac{a}{2}(t + t_j)), \tag{22}$$

$$\text{where} \quad a = \|\Sigma\beta^\star\| - \mathcal{O}\left(\lambda^{2-4\varepsilon} + \sqrt{\frac{d^2 \log n}{n}} \right)$$

$$\text{and} \quad \langle \overline{w}_i(0), \Sigma\beta^\star \rangle = a\tanh(\frac{a}{2}t_j).$$

Thanks to the choice of initialization given by Equation (8), $\|\overline{w}_i(0)\| \leq \frac{1}{2}$ and so $\langle \overline{w}_i(0), \Sigma\beta^\star \rangle \geq -\frac{1}{2}\|\Sigma\beta^\star\|_2$. Moreover, $\tanh(x) \leq -1 + 2e^{2x}$, so that

$$-\frac{1}{2}\|\Sigma\beta^\star\| \leq a(-1 + 2e^{at_j}).$$

Since $a = \|\Sigma\beta^\star\| - \mathcal{O}\left(\lambda^{2-4\varepsilon} + \sqrt{\frac{d^2 \log n}{n}} \right)$, this yields

$$2ae^{at_j} \geq \frac{1}{2}\|\Sigma\beta^\star\| + \mathcal{O}\left(\lambda^{2-4\varepsilon} + \sqrt{\frac{d^2 \log n}{n}} \right).$$

The previous inequality can be rewritten as

$$-2ae^{-t_j} \geq \frac{-4a^2}{\frac{1}{2}\|\Sigma\beta^\star\| + \mathcal{O}\left(\lambda^{2-4\varepsilon} + \sqrt{\frac{d^2 \log n}{n}} \right)}$$

$$\geq \frac{-8a^2}{\|\Sigma\beta^\star\|}(1+\mathcal{O}\left(\lambda^{2-4\varepsilon}+\sqrt{\frac{d^2\log n}{n}}\right))$$
$$\geq 8\|\Sigma\beta^\star\|+\mathcal{O}\left(\lambda^{2-4\varepsilon}+\sqrt{\frac{d^2\log n}{n}}\right).$$

Using that $\tanh(x)\geq 1-2e^{-2x}$, Equation (22) becomes at time τ and $\frac{n}{\log(n)}=\Omega(d^2)$,

$$\begin{aligned}\langle \overline{w}_i(\tau),\Sigma\beta^\star\rangle &\geq a-2ae^{-at_j}e^{-a\tau}\\ &\geq \|\Sigma\beta^\star\|-\|\Sigma\beta^\star\|(8\|\Sigma\beta^\star\|+\mathcal{O}\left(\lambda^{2-4\varepsilon}+\sqrt{\frac{d^2\log n}{n}}\right))e^{\frac{a\varepsilon\log\lambda}{\|\Sigma\beta^\star\|}}\\ &\quad-\mathcal{O}\left(\lambda^{2-4\varepsilon}+\sqrt{\frac{d^2\log n}{n}}\right)\\ &\geq \|\Sigma\beta^\star\|-\mathcal{O}\left(\lambda^{\varepsilon}+\sqrt{\frac{d^2\log n}{n}}\right).\end{aligned}$$

3. The same arguments can be done with negative neurons. □

F.3. Decoupled autonomous systems

In the remaining of the proof, we will focus on an alternative solution (w,a), which is solution of the following differential equations for any $t\geq\tau$

$$\begin{aligned}\frac{\mathrm{d}\mathsf{w}_i(t)}{\mathrm{d}t}&=\mathsf{a}_i(t)D_+(t) \quad\text{and}\quad \frac{\mathrm{d}\mathsf{a}_i(t)}{\mathrm{d}t}=\langle \mathsf{w}_i(t),D_+(t)\rangle \quad\text{for any } i\in\mathcal{I}_+,\\ \frac{\mathrm{d}\mathsf{w}_i(t)}{\mathrm{d}t}&=\mathsf{a}_i(t)D_-(t) \quad\text{and}\quad \frac{\mathrm{d}\mathsf{a}_i(t)}{\mathrm{d}t}=\langle \mathsf{w}_i(t),D_-(t)\rangle \quad\text{for any } i\in\mathcal{I}_-,\end{aligned} \tag{23}$$

where

$$D_+(t)=\frac{1}{n}\sum_{k\in\mathcal{S}_+}\left(\sum_{i\in\mathcal{I}_+}\mathsf{a}_i(t)\langle\mathsf{w}_i(t),x_k\rangle-y_k\right)x_k,$$
$$D_-(t)=\frac{1}{n}\sum_{k\in\mathcal{S}_-}\left(\sum_{i\in\mathcal{I}_-}\mathsf{a}_i(t)\langle\mathsf{w}_i(t),x_k\rangle-y_k\right)x_k$$

and with the initial condition $\mathsf{w}_i(\tau),\mathsf{a}_i(\tau)=w_i(\tau),a_i(\tau)$ for any $i\in[m]$. We also note in the following $\overline{\mathsf{w}}_i=\frac{\mathsf{w}_i}{\mathsf{a}_i}$ and the estimations of the training data x_k for any $k\in[n]I$ as:

$$h_\vartheta(x_k)=\begin{cases}\sum_{i\in\mathcal{I}_+}\mathsf{a}_i\langle\mathsf{w}_i,x_k\rangle & \text{if } k\in\mathcal{S}_+\\ \sum_{i\in\mathcal{I}_-}\mathsf{a}_i\langle\mathsf{w}_i,x_k\rangle & \text{if } k\in\mathcal{S}_-\end{cases}.$$

This construction allows to study separately the dynamics of both sets of neurons $\mathcal{I}_+$ and $\mathcal{I}_-$, without any interaction between each other. As precised by Lemma F.3 below, $\mathsf{w}_i,\mathsf{a}_i$ coincide with w_i,a_i as long as the neurons all remain in the sector they are at the end of the early alignment phase.

Lemma F.3. *Define* $T_+=\inf\{t\geq\tau\mid\exists(i,k)\in\mathcal{I}_+\times[n],\mathrm{sign}(x_k^\top\mathsf{w}_i(t))\neq\mathrm{sign}(x_k^\top\beta^\star)\}$ *and* $T_-=\inf\{t\geq\tau\mid\exists(i,k)\in\mathcal{I}_-\times[n],\mathrm{sign}(x_k^\top\mathsf{w}_i(t))\neq-\mathrm{sign}(x_k^\top\beta^\star)\}$.

Then for any $i\in[m]$ *and any* $t\in[\tau,\min(T_+,T_-)]$*:* $(\mathsf{w}_i(t),\mathsf{a}_i(t))=(w_i(t,a_i(t))$. *Moreover, for any* $t\in[\tau,\min(T_+,T_-)]$ *and* $k\in[n]$, $h_{\vartheta(t)}(x_k)=h_{\theta(t)}(x_k)$.

While analyzing the complete dynamics of (w,a), we will see that both T_+ and T_- are infinite in the considered range of parameters, thus leading to a complete description of the dynamics of (w,a).

Proof. Thanks to the definition of T_+ and T_-, the evolution of $(w_i(t), a_i(t))$ given by 4 coincides with the evolution of $(\mathsf{w}_i(t), \mathsf{a}_i(t))$ given by Equation (23) for $t \in [\tau, \min(T_+, T_-)]$. The associated ODE is Lipschitz on the considered time interval and thus admits a unique solution, hence leading to $(\mathsf{w}_i(t), \mathsf{a}_i(t)) = (w_i(t, a_i(t))$ on the considered interval. The equality $h_{\vartheta(t)}(x_k) = h_{\theta(t)}(x_k)$ directly derives from the ReLU activations and definitions of T_+ and T_-. □

F.4. Phase 2: neurons slow growth

For some $\varepsilon_2 > 0$, we define the following stopping time for any $\circ \in \{+, -\}$:

$$\tau_{2,\circ} = \inf\{t \geq \tau \mid \sum_{i \in \mathcal{I}_\circ} \mathsf{a}_i(t)^2 \geq \varepsilon_2\}.$$

Lemma F.4. *If Assumption 4.1 holds, for any $\varepsilon \in (0, \frac{1}{4})$, there exist $\lambda^\star = \Theta(\frac{1}{d})$, $\varepsilon_2^\star = \Theta(d^{-\frac{3}{2}})$ and $n^\star = \Theta(d^3 \log d)$ such that for any $\lambda \leq \lambda^\star$, $n \geq n^\star$, $\circ \in \{+, -\}$, $\varepsilon_2 \in [\lambda^{2-4\varepsilon}, \varepsilon_2^\star]$, with probability $1 - \mathcal{O}(\frac{1}{n} + \frac{1}{2^m})$, $\tau_{2,\circ} < +\infty$ and at this time,*

1. *neurons in $\mathcal{I}_\circ$ are aligned with each other*

$$\forall i, j \in \mathcal{I}_\circ, \quad \langle \overline{\mathsf{w}}_j(\tau_{2,\circ}), \overline{\mathsf{w}}_i(\tau_{2,\circ}) \rangle = 1 - \mathcal{O}\left(\frac{\lambda^{\frac{1}{2}}}{\varepsilon_2}\right);$$

2. *neurons in $\mathcal{I}_\circ$ are in the same cone as $\circ\beta^\star$ for any $t \in [\tau, \tau_{2,\circ}]$:*

$$\forall i \in \mathcal{I}_\circ, \quad \min_{k \in \mathcal{S}_\circ} \langle \overline{\mathsf{w}}_i(\tau_{2,\circ}), \frac{x_k}{\|x_k\|} \rangle = \Omega(\frac{1}{\sqrt{d}}) \quad \textit{and} \quad \max_{k \in \mathcal{S}_{-\circ}} \langle \overline{\mathsf{w}}_i(\tau_{2,\circ\circ}), \frac{x_k}{\|x_k\|} \rangle = -\Omega(\frac{1}{\sqrt{d}}).$$

Proof. In the following, we assume without loss of generality that $\circ = +$. Additionally, we assume that the random event $\mathcal{I}_+ \neq \emptyset$ and Equations (17) and (18) hold. First, by definition of $\tau_{2,+}$, for any $t \in [\tau, \tau_{2,+}]$:

$$\begin{aligned}\|D_+(t) - D_n(\beta^\star)\|_2 &\leq \frac{1}{n}\left(\sum_{i \in \mathcal{I}_+} \mathsf{a}_i(t)^2\right) \sum_{k \in \mathcal{S}_+} \|x_k\|^2 \\ &\leq 2\varepsilon_2 \mathbb{E}_{\mu_X}[\|x\|^2].\end{aligned}$$

This also implies with Equation (20) that $\|D_+(t)\| \leq \|\Sigma\beta^\star\| + 2\varepsilon_2 \mathbb{E}_{\mu_X}[\|x\|^2]$. Additionally, we have with Equation (17) that

$$\begin{aligned}\|D_+(t) - \frac{\Sigma\beta^\star}{2}\|_2 &\leq \|D_+(t) - D_n(\beta^\star)\|_2 + \|D_n(\beta^\star) - D(\beta^\star)\|_2 \\ &\leq \mathcal{O}\left(d\varepsilon_2 + \sqrt{\frac{d^2 \log n}{n}}\right). \end{aligned} \tag{24}$$

Then for any $k \in \mathcal{S}_+, i \in \mathcal{I}_+$ and $t \in [\tau, \tau_{2,+}]$, as long as $\langle \overline{\mathsf{w}}_i(t), x_k \rangle \geq 0$,

$$\begin{aligned}\frac{\mathrm{d}\langle \overline{\mathsf{w}}_i(t), \frac{x_k}{\|x_k\|} \rangle}{\mathrm{d}t} &= \langle D_+(t), \frac{x_k}{\|x_k\|} \rangle - \langle D_+(t), \overline{\mathsf{w}}_i(t) \rangle \langle \overline{\mathsf{w}}_i(t), \frac{x_k}{\|x_k\|} \rangle \\ &\geq \langle D_n(\beta^\star), \frac{x_k}{\|x_k\|} \rangle - 2\varepsilon_2 \mathbb{E}_{\mu_X}[\|x\|^2] - \|D_+(t)\| \langle \overline{\mathsf{w}}_i(t), \frac{x_k}{\|x_k\|} \rangle \\ &\geq \langle D_n(\beta^\star), \frac{x_k}{\|x_k\|} \rangle - 2\varepsilon_2 \mathbb{E}_{\mu_X}[\|x\|^2] - (\|D_n(\beta^\star)\| + 2\varepsilon_2 \mathbb{E}_{\mu_X}[\|x\|^2]) \langle \overline{\mathsf{w}}_i(t), \frac{x_k}{\|x_k\|} \rangle. \end{aligned} \tag{25}$$

As $\langle D_n(\beta^\star), \frac{x_k}{\|x_k\|} \rangle \geq \frac{\alpha}{\sqrt{d}}$ (Equation 19), thanks to Lemma F.2 and the third point in Assumption 4.1, and $\|D_n(\beta^\star)\| = \mathcal{O}(1)$, for a small enough $\varepsilon_2^\star = \Theta(d^{-\frac{3}{2}})$, $\min_{k \in \mathcal{S}_+} \langle \overline{\mathsf{w}}_i(\tau), \frac{x_k}{\|x_k\|} \rangle = \Omega(\frac{1}{\sqrt{d}})$. Equation (27) then implies for a small enough choice of $\varepsilon_2^\star = \Theta(d^{-\frac{3}{2}})$ and $\varepsilon_2 \leq \varepsilon_2^\star$:

$$\min_{t \in [\tau, \tau_{2,+}} \min_{k \in \mathcal{S}_+} \langle \overline{\mathsf{w}}_i(\tau), \frac{x_k}{\|x_k\|} \rangle = \Omega(\frac{1}{\sqrt{d}}). \tag{26}$$

Similarly, we can also show

$$\max_{t \in [\tau, \tau_{2,+}} \max_{k \in \mathcal{S}_-} \langle \overline{\mathsf{w}}_i(\tau), \frac{x_k}{\|x_k\|} \rangle = -\Omega(\frac{1}{\sqrt{d}}),$$

which implies the second point of Lemma F.4. Actually, we even have for this choice of parameters the more precise inequality (for the same reasons) that for any $k \in \mathcal{S}_+, i \in \mathcal{I}_+$ and $t \in [\tau, \tau_{2,+}]$,

$$\langle \overline{\mathsf{w}}_i(\tau), \frac{x_k}{\|x_k\|} \rangle \geq \langle \frac{D_n(\beta^\star)}{\|D_n(\beta^\star)\|}, \frac{x_k}{\|x_k\|} \rangle - \mathcal{O}(d\varepsilon_2)\,. \tag{27}$$

We now simultaneously lower and upper bound the duration of the second phase $\tau_{2,+} - \tau_2$. For any $t \in [\tau, \tau_{2,+}]$:

$$\begin{aligned}
\frac{1}{2}\frac{\mathrm{d}\sum_{i\in\mathcal{I}_+}\mathsf{a}_i(t)^2}{\mathrm{d}t} &= \sum_{i\in\mathcal{I}_+}\mathsf{a}_i(t)^2\langle\overline{\mathsf{w}}_i(t), D_+(t)\rangle \\
&= \frac{1}{n}\sum_{i\in\mathcal{I}_+}\mathsf{a}_i(t)^2\sum_{k\in\mathcal{S}_+}(y_k - h_{\vartheta(t)}(x_k))\langle\overline{\mathsf{w}}_i(t), x_k\rangle \\
&\geq \sum_{i\in\mathcal{I}_+}\mathsf{a}_i(t)^2\left(\frac{\sum_{k\in\mathcal{S}_+} y_k\|x_k\|\langle\overline{\mathsf{w}}_i(t), \frac{x_k}{\|x_k\|}\rangle}{n} - \frac{\varepsilon_2\sum_{k\in\mathcal{S}_+}\|x_k\|^2}{n}\right).
\end{aligned} \tag{28}$$

Note that $\mathbb{E}[\mathbb{1}_{k\in\mathcal{S}_+} y_k\|x_k\|] \geq \frac{c\mathbb{E}[\|x\|^2]}{2\sqrt{d}}$. Using Chebyshev inequality, we thus have for a small enough choice of $\varepsilon_2^\star = \Theta(d^{-\frac{3}{2}})$, for any $t \in [\tau, \tau_{2,+}]$:

$$\frac{\mathrm{d}\sum_{i\in\mathcal{I}_+}\mathsf{a}_i(t)^2}{\mathrm{d}t} \geq \Omega(1)\sum_{i\in\mathcal{I}_+}\mathsf{a}_i(t)^2.$$

A Grönwall comparison then directly yields $\tau_{2,+} < \infty$.

We now want to show that the neurons $\overline{\mathsf{w}}_i$ are almost aligned at the end of the second phase. For that, we first need to lower bound the duration of the phase. Note that Equation (28), with Equation (26), also leads for any $t \in [\tau, \tau_{2,+}]$ to

$$\begin{aligned}
\frac{1}{2}\frac{\mathrm{d}\sum_{i\in\mathcal{I}_+}\mathsf{a}_i(t)^2}{\mathrm{d}t} &\leq \frac{1}{n}\sum_{i\in\mathcal{I}_+}\mathsf{a}_i(t)^2\sum_{k\in\mathcal{S}_+} y_k\langle\overline{\mathsf{w}}_i(t), x_k\rangle \\
&= \sum_{i\in\mathcal{I}_+}\mathsf{a}_i(t)^2\langle\overline{\mathsf{w}}_i(t), D_n(\beta^\star)\rangle \\
&\leq \sum_{i\in\mathcal{I}_+}\mathsf{a}_i(t)^2\|D_n(\beta^\star)\|.
\end{aligned}$$

Note that by continuity, $\sum_{i\in\mathcal{I}_+}\mathsf{a}_i(\tau_{2,+})^2 = \varepsilon_2$. As $\sum_{i\in\mathcal{I}_+}\mathsf{a}_i(\tau)^2 \leq \lambda^{2-4\varepsilon}$, thanks to Lemma F.2, a Grönwall inequality argument leads to the following as $\varepsilon_2 \geq \lambda^{2-4\varepsilon}$,

$$\tau_{2,+} - \tau \geq \frac{1}{2\|D_n(\beta^\star)\|}\ln\left(\frac{\varepsilon_2}{\lambda^{2-4\varepsilon}}\right). \tag{29}$$

For any pair of neurons $i, j \in \mathcal{I}_+$, we consider the evolution of the mutual alignment:

$$\begin{aligned}
\frac{\mathrm{d}\langle\overline{\mathsf{w}}_i(t), \overline{\mathsf{w}}_j(t)\rangle}{\mathrm{d}t} &= \langle D_+(t), \overline{\mathsf{w}}_i(t) + \overline{\mathsf{w}}_j(t)\rangle(1 - \langle\overline{\mathsf{w}}_i(t), \overline{\mathsf{w}}_j(t)\rangle) \\
&= (\langle D_n(\beta^\star), \overline{\mathsf{w}}_i(t) + \overline{\mathsf{w}}_j(t)\rangle - \mathcal{O}(\varepsilon_2))\,(1 - \langle\overline{\mathsf{w}}_i(t), \overline{\mathsf{w}}_j(t)\rangle).
\end{aligned} \tag{30}$$

Moreover, Equation (27) leads to the following alignment between $\overline{\mathsf{w}}_i(t)$ and $D_n(\beta^\star)$ for any $t \in [\tau, \tau_{2,+}]$:

$$\begin{aligned}
\langle D_n(\beta^\star), \overline{\mathsf{w}}_i(t)\rangle &= \frac{1}{n}\sum_{k\in\mathcal{S}_+} y_k\langle x_k, \overline{\mathsf{w}}_i(t)\rangle \\
&= \frac{1}{n}\sum_{k\in\mathcal{S}_+} y_k\langle x_k, \frac{D_n(\beta^\star)}{\|D_n(\beta^\star)\|}\rangle - \|x_k\|\mathcal{O}(d\varepsilon_2) \\
&= \langle D_n(\beta^\star), \frac{D_n(\beta^\star)}{\|D_n(\beta^\star)\|}\rangle - \mathcal{O}(d\varepsilon_2) \\
&= \|D_n(\beta^\star)\| - \mathcal{O}(d\varepsilon_2)\,.
\end{aligned}$$

Equation (30) then rewrites for any $t \in [\tau, \tau_{2,+}]$ as

$$\frac{\mathrm{d}\langle \overline{\mathsf{w}}_i(t), \overline{\mathsf{w}}_j(t)\rangle}{\mathrm{d}t} \geq (2\|D_n(\beta^\star)\| - \mathcal{O}(d\varepsilon_2))\,(1 - \langle \overline{\mathsf{w}}_i(t), \overline{\mathsf{w}}_j(t)\rangle).$$

Moreover, thanks to Lemma F.2, a simple algebraic manipulation yields[8] $\langle \overline{\mathsf{w}}_i(\tau), \overline{\mathsf{w}}_j(\tau)\rangle \geq 1 - \mathcal{O}\left(\lambda^\varepsilon + \sqrt{\frac{d^2 \log n}{n}}\right)$. Grönwall inequality then yields, for the considered range of parameters,

$$\begin{aligned}
\langle \overline{\mathsf{w}}_i(\tau_{2,+}), \overline{\mathsf{w}}_j(\tau_{2,+})\rangle &\geq 1 - (1 - \langle \overline{\mathsf{w}}_i(\tau), \overline{\mathsf{w}}_j(\tau)\rangle)\, e^{-(2\|D_n(\beta^\star)\| - \mathcal{O}(d\varepsilon_2))(\tau_{2,+} - \tau)} \\
&\geq 1 - \mathcal{O}\left(\lambda^\varepsilon + \sqrt{\frac{d^2 \log n}{n}}\right) \frac{\lambda^{2-4\varepsilon}}{\varepsilon_2} e^{\mathcal{O}\left(d\varepsilon_2 \ln(\frac{\varepsilon_2}{\lambda^{2-4\varepsilon}})\right)} \\
&\geq 1 - \mathcal{O}\left(\frac{\lambda}{\varepsilon_2}\right) \lambda^{-(2-4\varepsilon)\mathcal{O}(d\varepsilon_2)}.
\end{aligned}$$

The second inequality comes from the bound on $\tau_{2,+} - \tau$ in Equation (29). The third one comes from the fact that $\varepsilon \leq \frac{1}{3}$ and $\varepsilon_2 \ln(\varepsilon_2) = \mathcal{O}(1)$. Noticing that $2 - 4\varepsilon \geq 1$ finally yields the first item of Lemma F.4 for a small enough $\varepsilon_2^\star = \Theta(d^{-\frac{3}{2}})$. □

F.5. Phase 3: neurons fast growth

The third phase is defined for some $\varepsilon_3 > 0$ and δ_3 by the following stopping time, for any $\circ \in \{+, -\}$:

$$\tau_{3,\circ} = \inf\{t \geq \tau_{2,\circ} \mid \|\hat{\beta}_\circ(t) - \beta_{n,\circ}\|_{\Sigma_{n,\circ}} \leq \varepsilon_3 \text{ or } \exists i \in \mathcal{I}_\circ, k \in \mathcal{S}_\circ, \langle \mathsf{w}_i(t), \frac{x_k}{\|x_k\|}\rangle \leq \delta_3\},$$
$$\text{where} \quad \hat{\beta}_\circ(t) = \sum_{i \in \mathcal{I}_\circ} \mathsf{a}_i(t)\mathsf{w}_i(t).$$

Lemma F.5. *If Assumption 4.1 holds, for any $\varepsilon \in (0, \frac{1}{4})$, there exist $\lambda^\star = \Theta(\frac{1}{d})$, $\varepsilon_2^\star = \Theta(d^{-\frac{3}{2}})$, $n^\star = \Theta(d^3 \log d)$, $\alpha_0 = \Theta(1)$, $\delta_3 = \Theta(\frac{1}{\sqrt{d}})$ and $\varepsilon_3^\star = \Theta(1)$ such that for any $\lambda \leq \lambda^\star$, $n \geq n^\star$, $\circ \in \{+, -\}$, $\varepsilon_2 \in [\lambda^{2-4\varepsilon}, \varepsilon_2^\star]$ and $\varepsilon_3 \in [\lambda^{\alpha_0 \varepsilon \varepsilon_2}, \varepsilon_3^\star]$, with probability $1 - \mathcal{O}\left(\frac{d^2}{n} + \frac{1}{2^m}\right)$, $\tau_{3,\circ} < +\infty$ and*

1. *neurons in $\mathcal{I}_\circ$ are in the same cone as $\circ\beta^\star$ for any $t \in [\tau, \tau_{2,\circ}]$:*

$$\forall i \in \mathcal{I}_\circ, \quad \min_{k \in \mathcal{S}_\circ} \langle \overline{\mathsf{w}}_i(t), \frac{x_k}{\|x_k\|}\rangle \geq 2\delta_3 \quad \text{and} \quad \max_{k \in \mathcal{S}_{-\circ}} \langle \overline{\mathsf{w}}_i(t), \frac{x_k}{\|x_k\|}\rangle \leq -2\delta_3.$$

In particular, $\|\hat{\beta}_\circ(\tau_{3,\circ}) - \beta_{n,\circ}\|_{\Sigma_{n,\circ}} = \varepsilon_3$ by continuity.

Proof. Similarly to the proof of Lemma F.4, we assume that $\circ = +$, that the random event $\mathcal{I}_+ \neq \emptyset$, Equations (17) and (18) and the first and second items states in Lemma F.4 all hold. We can first show that for any $t \in [\tau_{2,+}, \tau_{3,+}]$,

$$\sum_{i \in \mathcal{I}_+} \mathsf{a}_i(t)^2 \geq \varepsilon_2.$$

Indeed, recall that the output weights a_i evolve for any $t \in [\tau_{2,+}, \tau_{3,+}]$ as

$$\begin{aligned}
\frac{\mathrm{d}\mathsf{a}_i(t)}{\mathrm{d}t} &= \langle \mathsf{w}_i(t), D_+(t)\rangle \\
&= \langle \mathsf{w}_i(t), D_n(\beta^\star)\rangle - \frac{1}{n} \sum_{k \in \mathcal{S}_+} h_{\vartheta(t)}(x_k) \langle \mathsf{w}_i(t), x_k\rangle \\
&\geq \mathsf{a}_i(t) \left(\frac{1}{n} \sum_{k \in \mathcal{S}_+} \langle \overline{\mathsf{w}}_i(t), x_k\rangle \langle \beta^\star, x_k\rangle + \frac{1}{n} \sum_{k \in \mathcal{S}_+} \langle \overline{\mathsf{w}}_i(t), \eta_k x_k\rangle - \mathcal{O}\left(d \sum_{i \in \mathcal{I}_+} \mathsf{a}_i(t)^2 \right) \right). \qquad (31)
\end{aligned}$$

[8] A similar manipulation can be found in (Boursier and Flammarion, 2024, proof of Lemma 5).

The last inequality comes from the fact that $\frac{\sum_{k\in\mathcal{S}_+}\|x_k\|^2}{n} = \mathcal{O}(d)$. From then, note that $\frac{1}{n}\sum_{k\in\mathcal{S}_+}\langle \overline{\mathsf{w}}_i(t), x_k\rangle\langle\beta^\star, x_k\rangle \geq \Omega(\delta_3\sqrt{d})$ during this phase. Moreover, using Chebyshev inequality, we can show for any $z>0$ that with probability at least $1-\mathcal{O}\big(\frac{d}{z^2 n}\big)$

$$\frac{1}{n}\left\|\sum_{k\in\mathcal{S}_+}\eta_k x_k\right\|_2 \leq z. \tag{32}$$

Taking a small enough $z=\Theta(\delta_3)$, Equation (32) holds with probability $1-\mathcal{O}\Big(\frac{d}{\delta_3^2 n}\Big)$ and, along Equation (31), this implies that for any $t\in[\tau_{2,+},\tau_{3,+}]$ and $i\in\mathcal{I}_+$:

$$\frac{\mathrm{d}\mathsf{a}_i(t)}{\mathrm{d}t} \geq \mathsf{a}_i(t)\left(\Omega(\delta_3) - \mathcal{O}\left(d\sum_{i\in\mathcal{I}_+}\mathsf{a}_i(t)^2\right)\right).$$

In particular, there exists $r=\Theta(\frac{\delta_3}{d})$ such that if $\sum_{i\in\mathcal{I}_+}\mathsf{a}_i(t)^2\leq r$, all the $\mathsf{a}_i(t)$ are increasing. Moreover thanks to Lemma F.4, $\sum_{i\in\mathcal{I}_+}\mathsf{a}_i(\tau_{2,+})^2=\varepsilon_2$. As $\delta_3=\Theta(\frac{1}{\sqrt{d}})$, we can choose $\varepsilon_2^\star=\Theta(d^{-\frac{3}{2}})$ small enough so that during the third phase,

$$\sum_{i\in\mathcal{I}_+}\mathsf{a}_i(t)^2 \geq \varepsilon_2. \tag{33}$$

Now note that by definition of $\beta_{n,+}$,

$$\begin{aligned} D_+(t) &= -\frac{1}{n}\sum_{k\in\mathcal{S}_+} x_k x_k^\top \hat{\beta}_+(t) - x_k y_k \\ &= -\Sigma_{n,+}(\hat{\beta}_+(t)-\beta_{n,+}) \end{aligned} \tag{34}$$

As a consequence, $\hat{\beta}_+(t)$ evolves as follows:

$$\begin{aligned} \frac{\mathrm{d}\hat{\beta}_+(t)}{\mathrm{d}t} &= \sum_{i\in\mathcal{I}_+}\left(\mathsf{a}_i(t)^2\mathbf{I}_d + \mathsf{w}_i(t)\mathsf{w}_i(t)^\top\right)D_+(t) \\ &= -\left(\sum_{i\in\mathcal{I}_+}\mathsf{a}_i(t)^2\mathbf{I}_d + \sum_{i\in\mathcal{I}_+}\mathsf{w}_i(t)\mathsf{w}_i(t)^\top\right)\Sigma_{n,+}(\hat{\beta}_+(t)-\beta_{n,+}) \end{aligned}$$

In particular, this implies:

$$\frac{1}{2}\frac{\mathrm{d}\|\hat{\beta}_+(t)-\beta_{n,+}\|^2_{\Sigma_{n,+}}}{\mathrm{d}t} = \left\langle \frac{\mathrm{d}\hat{\beta}_+(t)}{\mathrm{d}t}, \Sigma_{n,+}(\hat{\beta}_+(t)-\beta_{n,+})\right\rangle \tag{35}$$

$$= -(\hat{\beta}_+(t)-\beta_{n,+})^\top\Sigma_{n,+}\left(\sum_{i\in\mathcal{I}_+}\mathsf{a}_i(t)^2\mathbf{I}_d + \sum_{i\in\mathcal{I}_+}\mathsf{w}_i(t)\mathsf{w}_i(t)^\top\right)\Sigma_{n,+}(\hat{\beta}_+(t)-\beta_{n,+}). \tag{36}$$

The matrix $\Sigma_{n,+}^{1/2}\left(\sum_{\mathcal{I}_+}\mathsf{a}_i(t)^2\mathbf{I}_d + \sum_{\mathcal{I}_+}\mathsf{w}_i(t)\mathsf{w}_i(t)^\top\right)\Sigma_{n,+}^{1/2}$ is symmetric, positive definite. Thanks to Equation (33), its smallest eigenvalue is larger than $\varepsilon_2\lambda_{\min}(\Sigma_{n,+})$, where $\lambda_{\min}(\cdot)$ denotes the smallest eigenvalue of a matrix. Using typical concentration inequalities on the empirical covariance (see e.g. Vershynin, 2018, Section 4.7), with probability $1-\mathcal{O}(\frac{1}{n})$, $\|\Sigma_{n,+}-\frac{\Sigma}{2}\|_{\mathrm{op}} = \mathcal{O}\left(\sqrt{\frac{d+\log n}{n}}\right)$. With the fourth point in Assumption 4.1, we then have for a large enough $n^\star=\Theta(d^3\log d)$ and with probability $1-\mathcal{O}(\frac{1}{n})$,

$$\begin{aligned} &\mathcal{O}(1) \geq \lambda_{\max}(\Sigma_{n,+}) \geq \lambda_{\min}(\Sigma_{n,+}) \geq \Omega(1) \\ \text{and}\quad &\frac{\lambda_{\max}(\Sigma_{n,+})}{\lambda_{\min}(\Sigma_{n,+})} = \frac{\lambda_{\max}(\Sigma)}{\lambda_{\min}(\Sigma)} + \mathcal{O}\left(\sqrt{\frac{d+\log n}{n}}\right), \end{aligned} \tag{37}$$

where $\lambda_{\max}(\cdot)$ denotes the largest eigenvalue.

Assume Equation (37) holds in the following, so that the smallest eigenvalue of $\Sigma_{n,+}^{1/2}\left(\sum_{\mathcal{I}_+}\mathsf{a}_i(t)^2\mathbf{I}_d+\sum_{\mathcal{I}_+}\mathsf{w}_i(t)\mathsf{w}_i(t)^\top\right)\Sigma_{n,+}^{1/2}$ is larger than a term of order ε_2. As a consequence, Equation (36) yields

$$\frac{1}{2}\frac{\mathrm{d}\|\hat{\beta}_+(t)-\beta_{n,+}\|^2_{\Sigma_{n,+}}}{\mathrm{d}t}\leq-\Omega(\varepsilon_2)\|\hat{\beta}_+(t)-\beta_{n,+}\|^2_{\Sigma_{n,+}}.$$

Since the third phase ends if $\|\hat{\beta}_+(t)-\beta_{n,+}\|^2_{\Sigma_{n,+}}$ becomes smaller than ε_3^2, this yields:

$$\tau_{3,+}-\tau_{2,+}=\mathcal{O}\left(\frac{1}{\varepsilon_2}\ln(\frac{1}{\varepsilon_3})\right). \tag{38}$$

Now recall that for any $i,j\in\mathcal{I}_+$,

$$\begin{aligned}\frac{\mathrm{d}(1-\langle\overline{\mathsf{w}}_i(t),\overline{\mathsf{w}}_j(t)\rangle)}{\mathrm{d}t}&=-\langle D_+(t),\overline{\mathsf{w}}_i(t)+\overline{\mathsf{w}}_j(t)\rangle(1-\langle\overline{\mathsf{w}}_i(t),\overline{\mathsf{w}}_j(t)\rangle)\\&\leq 2\|D_+(t)\|_2(1-\langle\overline{\mathsf{w}}_i(t),\overline{\mathsf{w}}_j(t)\rangle).\end{aligned}$$

Notice from Equation (34) and the previous discussion that $\|D_+(t)\|_2=\mathcal{O}(1)$. As a consequence, a simple Grönwall inequality with Equation (38) yields that for any $t\in[\tau_{2,+},\tau_{3,+}]$:

$$\begin{aligned}\langle\overline{\mathsf{w}}_i(t),\overline{\mathsf{w}}_j(t)\rangle&\geq 1-(1-\langle\overline{\mathsf{w}}_i(\tau_{2,+}),\overline{\mathsf{w}}_j(\tau_{2,+})\rangle)\exp((t-\tau_{2,+})\mathcal{O}(1))\\&\geq 1-\mathcal{O}\left(\frac{\lambda^{\frac{1}{2}}}{\varepsilon_2}\right)\exp\left(\mathcal{O}\left(\frac{1}{\varepsilon_2}\ln(\frac{1}{\varepsilon_3})\right)\right)\\&\geq 1-\mathcal{O}\left(\lambda^{\frac{1}{2}-\varepsilon}\right).\end{aligned}$$

The second inequality comes from the value of $(1-\langle\overline{\mathsf{w}}_i(\tau_{2,+}),\overline{\mathsf{w}}_j(\tau_{2,+})\rangle)$, thanks to Lemma F.4. The last one comes from our choice of ε_3 for a large enough $\alpha_0=\Theta(1)$.

In particular, this last inequality can be used to show[9] that for any $i,j\in\mathcal{I}_+$ and $t\in[\tau_{2,+},\tau_{3,+}]$, $\overline{\mathsf{w}}_i(t)=\overline{\mathsf{w}}_j(t)+\mathcal{O}\left(\lambda^{\frac{1-2\varepsilon}{4}}\right)$. In particular, this yields for any $i\in\mathcal{I}_+$ and $t\in[\tau_{2,+},\tau_{3,+}]$

$$\hat{\beta}_+(t)=\sum_{j\in\mathcal{I}_+}\mathsf{a}_j(t)^2\overline{\mathsf{w}}_j(t) \tag{39}$$

$$=\left(\sum_{j\in\mathcal{I}_+}\mathsf{a}_j(t)^2\right)\left(\overline{\mathsf{w}}_i(t)+\mathcal{O}\left(\lambda^{\frac{1-2\varepsilon}{4}}\right)\right). \tag{40}$$

Since $\|\overline{\mathsf{w}}_i(t)\|_2=1-\mathcal{O}\left(\lambda^{\frac{1}{2}-\varepsilon}\right)$, this last equality actually yields the following comparison for $t\in[\tau_{2,+},\tau_{3,+}]$:

$$\|\hat{\beta}_+(t)\|_2\leq\sum_{j\in\mathcal{I}_+}\mathsf{a}_j(t)^2\leq(1+\mathcal{O}\left(\lambda^{\frac{1-2\varepsilon}{4}}\right))\|\hat{\beta}_+(t)\|_2. \tag{41}$$

In particular, since $\|\hat{\beta}_+(t)\|_2=\mathcal{O}(1)$, this yields $\sum_{j\in\mathcal{I}_+}\mathsf{a}_j(t)^2=\mathcal{O}(1)$.

From there, for any $x_k\in\mathcal{S}_+$ and $i\in\mathcal{I}_+$, $\langle\overline{\mathsf{w}}_i(t),x_k\rangle$ evolves as follows during the third phase

$$\begin{aligned}\frac{\mathrm{d}\langle\overline{\mathsf{w}}_i(t),\frac{x_k}{\|x_k\|}\rangle}{\mathrm{d}t}&=\langle D_+(t),\frac{x_k}{\|x_k\|}\rangle-\langle D_+(t),\overline{\mathsf{w}}_i(t)\rangle\langle\overline{\mathsf{w}}_i(t),\frac{x_k}{\|x_k\|}\rangle\\&=\langle\beta_{n,+}-\hat{\beta}_+(t),\Sigma_{n,+}\frac{x_k}{\|x_k\|}\rangle-\mathcal{O}\left(\langle\overline{\mathsf{w}}_i(t),\frac{x_k}{\|x_k\|}\rangle\right)\\&=\frac{1}{2}\langle\beta_{n,+}-\hat{\beta}_+(t),\frac{x_k}{\|x_k\|}\rangle+\langle(\Sigma_{n,+}-\frac{\mathbf{I}_d}{2})(\beta_{n,+}-\hat{\beta}_+(t)),\frac{x_k}{\|x_k\|}\rangle-\mathcal{O}\left(\langle\overline{\mathsf{w}}_i(t),\frac{x_k}{\|x_k\|}\rangle\right).\end{aligned}$$

Note that

$$\beta_{n,+}=\beta^\star+\frac{\Sigma_{n,+}^{-1}}{n}\sum_{k\in\mathcal{S}_+}\eta_k x_k.$$

[9] For that, we decompose $\overline{\mathsf{w}}_i=\alpha_{ij}\overline{\mathsf{w}}_j+u_{ij}$ with $u_{ij}\perp\overline{\mathsf{w}}_j$ and show that $\alpha_{ij}=1-\mathcal{O}\left(\lambda^{\frac{1}{2}-\varepsilon}\right)$ and $\|u_{ij}\|^2=\mathcal{O}\left(\lambda^{\frac{1}{2}-\varepsilon}\right)$.

This then yields, thanks to Equation (37)

$$\frac{\mathrm{d}\langle \overline{\mathsf{w}}_i(t), \frac{x_k}{\|x_k\|}\rangle}{\mathrm{d}t} \geq \frac{1}{2}\langle \beta^\star - \hat{\beta}_+(t), \frac{x_k}{\|x_k\|}\rangle + \langle (\Sigma_{n,+} - \frac{\mathbf{I}_d}{2})(\beta_{n,+} - \hat{\beta}_+(t)), \frac{x_k}{\|x_k\|}\rangle \\ - \mathcal{O}\left(\frac{1}{n}\|\sum_{k\in\mathcal{S}_+}\eta_k x_k\|_2\right) - \mathcal{O}\Big(\langle \overline{\mathsf{w}}_i(t), \frac{x_k}{\|x_k\|}\rangle\Big). \quad (42)$$

From there, thanks to the third point of Assumption 4.1 and Equation (40):

$$\langle \beta^\star - \hat{\beta}_+(t), \frac{x_k}{\|x_k\|}\rangle \geq \frac{c}{\sqrt{d}} - \mathcal{O}\Big(\langle \overline{\mathsf{w}}_i(t), \frac{x_k}{\|x_k\|}\rangle\Big) - \mathcal{O}\Big(\lambda^{\frac{1-2\varepsilon}{4}}\Big). \quad (43)$$

Additionally, using the fact that $\|\beta_{n,+} - \hat{\beta}_+(t)\|_{\Sigma_{n,+}}$ is decreasing over time and smaller than $\|\beta_{n,+}\|_{\Sigma_{n,+}} + \mathcal{O}(\lambda^{2-4\varepsilon})$ at the beginning of the second phase,

$$\begin{aligned}
\langle (\Sigma_{n,+} - \frac{\mathbf{I}_d}{2})(\beta_{n,+} - \hat{\beta}_+(t)), \frac{x_k}{\|x_k\|}\rangle &\geq -\left\|\Sigma_{n,+} - \frac{\mathbf{I}_d}{2}\right\|_{\mathrm{op}} \|\beta_{n,+} - \hat{\beta}_+(t)\|_2 \\
&\geq -\frac{1}{2}\|\Sigma - \mathbf{I}_d\|_{\mathrm{op}} \sqrt{\frac{1}{\lambda_{\min}(\Sigma_{n,+})}}\|\beta_{n,+} - \hat{\beta}_+(t)\|_{\Sigma_{n,+}} - \mathcal{O}\left(\sqrt{\frac{d+\log n}{n}}\right) \\
&\geq -\frac{1}{2}\|\Sigma - \mathbf{I}_d\|_{\mathrm{op}} \sqrt{\frac{1}{\lambda_{\min}(\Sigma_{n,+})}}\left(\|\beta_{n,+}\|_{\Sigma_{n,+}} + \mathcal{O}(\lambda^{2-4\varepsilon})\right) - \mathcal{O}\left(\sqrt{\frac{d+\log n}{n}}\right) \\
&\geq -\frac{1}{2}\|\Sigma - \mathbf{I}_d\|_{\mathrm{op}} \sqrt{\frac{\lambda_{\max}(\Sigma_{n,+})}{\lambda_{\min}(\Sigma_{n,+})}}\|\beta_{n,+}\|_2 - \mathcal{O}\left(\lambda^{2-4\varepsilon} + \sqrt{\frac{d+\log n}{n}}\right) \\
&\geq -\frac{1}{2}\sqrt{\frac{\lambda_{\max}(\Sigma_{n,+})}{\lambda_{\min}(\Sigma_{n,+})}}\|\Sigma - \mathbf{I}_d\|_{\mathrm{op}}\|\beta^\star\|_2 \\
&\quad - \mathcal{O}\left(\lambda^{2-4\varepsilon} + \frac{1}{n}\|\sum_{k\in\mathcal{S}_+}\eta_k x_k\|_2 + \sqrt{\frac{d+\log n}{n}}\right)
\end{aligned}$$

Now using Equation (37) and the fourth point of Assumption 4.1, note that

$$\sqrt{\frac{\lambda_{\max}(\Sigma_{n,+})}{\lambda_{\min}(\Sigma_{n,+})}} \leq 2 + \mathcal{O}\left(\sqrt{\frac{d+\log n}{n}}\right).$$

So that the previous inequality yields

$$\langle (\Sigma_{n,+} - \frac{\mathbf{I}_d}{2})(\beta_{n,+} - \hat{\beta}_+(t)), \frac{x_k}{\|x_k\|}\rangle \geq -\|\Sigma - \mathbf{I}_d\|_{\mathrm{op}}\|\beta^\star\|_2 - \mathcal{O}\left(\lambda^{2-4\varepsilon} + \frac{1}{n}\|\sum_{k\in\mathcal{S}_+}\eta_k x_k\|_2 + \sqrt{\frac{d+\log n}{n}}\right). \quad (44)$$

Finally, thanks to Equation (32), $\frac{1}{n}\|\sum_{k\in\mathcal{S}_+}\eta_k x_k\|_2 \leq z'$ with probability at least $1 - \mathcal{O}(\frac{d}{z'^2 n})$. Using Equations (43) and (44) in Equation (42) finally yields for the third phase:

$$\frac{\mathrm{d}\langle \overline{\mathsf{w}}_i(t), \frac{x_k}{\|x_k\|}\rangle}{\mathrm{d}t} \geq \frac{c}{2\sqrt{d}} - \|\Sigma - \mathbf{I}_d\|_{\mathrm{op}}\|\beta^\star\|_2 - \mathcal{O}\left(\lambda^{\frac{1-2\varepsilon}{4}} + z' + \sqrt{\frac{d+\log n}{n}}\right) - \mathcal{O}\Big(\langle \overline{\mathsf{w}}_i(t), \frac{x_k}{\|x_k\|}\rangle\Big).$$

Thanks to the fourth point of Assumption 4.1, $\frac{c}{2\sqrt{d}} - \|\Sigma - \mathbf{I}_d\|_{\mathrm{op}}\|\beta^\star\|_2 > 0$, so that we can choose $\lambda^\star, z' = \Theta(1)$ small enough and $n^\star = \Theta(d^3 \log d)$ large enough so that the previous inequality becomes, with probability at least $1 - \mathcal{O}(\frac{d}{n})$

$$\frac{\mathrm{d}\langle \overline{\mathsf{w}}_i(t), \frac{x_k}{\|x_k\|}\rangle}{\mathrm{d}t} \geq \Omega(\frac{1}{\sqrt{d}}) - \mathcal{O}\Big(\langle \overline{\mathsf{w}}_i(t), \frac{x_k}{\|x_k\|}\rangle\Big).$$

A simple Grönwall argument with the second point of Lemma F.4 then implies that for any $t \in [\tau_{2,+}, \tau_{3,+}]$, $i \in \mathcal{I}_+$ and $k \in \mathcal{S}_+$,

$$\langle \overline{\mathsf{w}}_i(t), \frac{x_k}{\|x_k\|}\rangle \geq \Omega(\frac{1}{\sqrt{d}}).$$

Since the term $\Omega(\frac{1}{\sqrt{d}})$ here does not depend on δ_3, we can choose $\delta_3 = \Theta(\frac{1}{\sqrt{d}})$ small enough so that

$$\langle \overline{\mathsf{w}}_i(t), \frac{x_k}{\|x_k\|} \rangle \geq 2\delta_3.$$

We can show similarly for $k \in \mathcal{S}_-$, so that point 1 in Lemma F.5 holds, which concludes the proof. □

F.6. Phase 4: final convergence

The last phase is defined for some $\varepsilon_4 > \varepsilon_3$ by the following stopping time, for any $\circ \in \{+, -\}$:

$$\tau_{4,\circ} = \inf\{t \geq \tau_{3,\circ} \mid \|\hat{\beta}_\circ(t) - \hat{\beta}_\circ(\tau_{3,\circ})\|_{\Sigma_{n,\circ}} \geq \varepsilon_4\}.$$

Lemma F.6. *If Assumption 4.1 holds, for any $\varepsilon \in (0, \frac{1}{4})$, there exist $\lambda^\star = \Theta(\frac{1}{d})$, $\varepsilon_2^\star = \Theta(d^{-\frac{3}{2}})$, $n^\star = \Theta(d^3 \log d)$, $\alpha_0 = \Theta(1)$, $\delta_3 = \Theta(\frac{1}{\sqrt{d}})$, $\varepsilon_3^\star = \Theta(\frac{1}{\sqrt{d}})$ and $\varepsilon_4 = \Theta(\varepsilon_3^\star)$ such that for any $\lambda \leq \lambda^\star$, $n \geq n^\star$, $\circ \in \{+, -\}$, $\varepsilon_2 \in [\lambda^{2-4\varepsilon}, \varepsilon_2^\star]$ and $\varepsilon_3 \in [\lambda^{\alpha_0 \varepsilon \varepsilon_2}, \varepsilon_3^\star]$, with probability $1 - \mathcal{O}\Big(\frac{d^2}{n} + \frac{1}{2^m}\Big)$, $\tau_{4,\circ} = +\infty$ and*

1. *neurons in $\mathcal{I}_\circ$ are in the same cone as $\circ\beta^\star$ for any $t \geq \tau_{3,\circ}$:*

$$\forall i \in \mathcal{I}_\circ, \quad \min_{k \in \mathcal{S}_\circ} \langle \overline{\mathsf{w}}_i(t), \frac{x_k}{\|x_k\|} \rangle > 0 \quad \text{and} \quad \max_{k \in \mathcal{S}_{-\circ}} \langle \overline{\mathsf{w}}_i(t), \frac{x_k}{\|x_k\|} \rangle < 0.$$

2. *$\lim_{t \to \infty} \vartheta(t)$ exists and $\lim_{t \to \infty} \hat{\beta}_\circ(t) = \beta_{n,\circ}$.*

Proof. Similarly to the previous phases, we assume that $\circ = +$, that the random event $\mathcal{I}_+ \neq \emptyset$, Equations (17), (18) and (37) and the statements of Lemma F.5 all hold.

Define in the following the positive loss L_+ for any $\vartheta_+ \in \mathbb{R}^{(d+1)\times \mathcal{I}_+}$ by

$$L_+(\vartheta) = \frac{1}{2n} \sum_{k \in \mathcal{S}_+} \left(\sum_{i \in \mathcal{I}+} \mathsf{a}_i \langle \mathsf{w}_i, x_k \rangle - y_k \right)^2.$$

Note that the autonomous system given by Equation (23) actually defines a gradient flow over L_+, i.e., for $\vartheta_+ = (\mathsf{a}_i, \mathsf{w}_i)_{i \in \mathcal{I}_+}$,

$$\frac{\mathrm{d}\vartheta_+(t)}{\mathrm{d}t} = -\nabla L_+(\vartheta_+(t)).$$

The main argument for this phase is to prove a local Polyak-Łojasiewicz inequality:

$$\begin{aligned} &\|\nabla L_+(\vartheta_+)\|_2^2 \geq \Omega(1)(L_+(\vartheta_+) - L_{n,+}) \\ &\text{for any } \vartheta_+ \text{ such that } \quad \|\sum_{i \in \mathcal{I}_+} \mathsf{a}_i \mathsf{w}_i - \hat{\beta}_+(\tau_{3,\circ})\|_{\Sigma_{n,+}} \leq \varepsilon_4, \\ &\text{where} \quad L_{n,+} = \frac{1}{2n} \sum_{k \in \mathcal{S}_+} \left(\langle \beta_{n,+}, x_k \rangle - y_k \right)^2. \end{aligned} \tag{45}$$

Indeed, we can lower bound $\|\nabla L_+(\vartheta_+)\|_2$ for any such ϑ_+ as follows

$$\begin{aligned} \|\nabla L_+(\vartheta_+)\|_2^2 &\geq \sum_{i \in \mathcal{I}_+} \left\| \frac{\partial L_+(\vartheta_+)}{\partial \mathsf{w}_i} \right\|^2 \\ &= \left(\sum_{i \in \mathcal{I}_+} \mathsf{a}_i(t)^2 \right) \|D_+(t)\|_2^2 \\ &\geq \lambda_{\min}(\Sigma_{n,+}) \left(\sum_{i \in \mathcal{I}_+} \mathsf{a}_i(t)^2 \right) \|\hat{\beta}_+ - \beta_{n,+}\|_{\Sigma_{n,+}}^2 \end{aligned}$$

where $\hat{\beta}_+ = \sum_{i\in\mathcal{I}_+} \mathsf{a}_i\mathsf{w}_i$. The last inequality comes from Equation (34). Note that for a small enough choice of $\varepsilon_3^\star = \mathcal{O}(1)$ and $\varepsilon_4 = \Theta(\varepsilon_3^\star)$, $\sum_{i\in\mathcal{I}_+} \mathsf{a}_i(t)^2 = \Omega(1)$ in the considered set. Moreover, Equation (37) implies $\lambda_{\min}(\Sigma_{n,+}) = \Omega(1)$, so that

$$\|\nabla L_+(\vartheta_+)\|_2^2 \geq \Omega(1)\|\hat{\beta}_+ - \beta_{n,+}\|_{\Sigma_{n,+}}^2. \tag{46}$$

On the other hand, a simple algebraic manipulation yields for any ϑ_+:

$$\begin{aligned}
L_+(\vartheta_+) - L_{n,+} &= \frac{1}{2n}\sum_{k\in\mathcal{S}_+}\left(\langle\hat{\beta}_+, x_k\rangle - y_k\right)^2 - \left(\langle\beta_{n,+}, x_k\rangle - y_k\right)^2 \\
&= \frac{1}{2n}\sum_{k\in\mathcal{S}_+}\left(\hat{\beta}_+ - \beta_{n,+}\right)^\top x_k - \left(x_k^\top(\hat{\beta}_+ - \beta_{n,+} + 2\beta_{n,+}) - 2y_k\right) \\
&= \frac{1}{2}\left(\hat{\beta}_+ - \beta_{n,+}\right)^\top \Sigma_{n,+}\left(\hat{\beta}_+ - \beta_{n,+}\right) + \frac{1}{n}\mathbf{X}_{n,+}(\mathbf{X}_{n,+}^\top\beta_{n,+} - \mathbf{y}),
\end{aligned}$$

where $\mathbf{X}_{n,+}$ is the $|\mathcal{S}_+| \times d$ matrix, whose rows are given by x_k for $k \in \mathcal{S}_+$. By definition of the OLS estimator $\beta_{n,+}$, $\mathbf{X}_{n,+}^\top\beta_{n,+} - \mathbf{y} = \mathbf{0}$, so that

$$L_+(\vartheta_+) - L_{n,+} = \frac{1}{2}\|\hat{\beta}_+ - \beta_{n,+}\|_{\Sigma_{n,+}}^2. \tag{47}$$

Combining Equation (46) with Equation (47) finally yields the Polyak-Łojasiewicz inequality given by Equation (45).

From there, this implies by chain rule for any $t \in [\tau_{3,+}, \tau_{4,+}]$

$$\begin{aligned}
\frac{\mathrm{d}L_+(\vartheta_+(t))}{\mathrm{d}t} &= -\|\nabla L_+(\vartheta_+)\|_2^2 \\
&\leq -\Omega(1)(L_+(\vartheta_+(t)) - L_{n,+}).
\end{aligned}$$

By Grönwall inequality, this implies for some $\nu = \Theta(1)$, for any $t \in [\tau_{3,+}, \tau_{4,+}]$

$$\begin{aligned}
L_+(\vartheta_+(t)) - L_{n,+} &\leq (L_+(\vartheta_+(\tau_{3,+})) - L_{n,+})e^{-\nu(t-\tau_{3,+})} \\
&\leq \frac{\varepsilon_3^2}{2}e^{-\nu(t-\tau_{3,+})}.
\end{aligned} \tag{48}$$

The last inequality comes from the fact that at the end of the third phase, $\|\hat{\beta}_+(t) - \beta_{n,+}\|_{\Sigma_{n,+}} = \varepsilon_3$.

We bounded by below the norm of $\nabla L_+(\vartheta_+(s))$, but it can also easily be bounded by above as

$$\begin{aligned}
\|\nabla L_+(\vartheta_+(s))\|_2^2 &\leq \left(\sum_{i\in\mathcal{I}_+}\mathsf{a}_i(t)^2 + \|\mathsf{w}_i(t)\|_2^2\right)\|D_+(t)\|_2^2 \\
&\leq 2\lambda_{\max}(\Sigma_{n,+})\left(\sum_{i\in\mathcal{I}_+}\mathsf{a}_i(t)^2\right)\|\hat{\beta}_+(t) - \beta_{n,+}\|_{\Sigma_{n,+}}^2 \\
&\leq \mathcal{O}(1)\left(L_+(\vartheta_+(t)) - L_{n,+}\right)
\end{aligned}$$

From there, the variation of $\vartheta_+(t)$ can easily be bounded for any $t \in [\tau_{3,+}, \tau_{4,+}]$ as

$$\begin{aligned}
\|\vartheta_+(t) - \vartheta(\tau_{3,+})\|_2 &\leq \int_{\tau_{3,+}}^t \|\nabla L_+(\vartheta_+(s))\|\mathrm{d}s \\
&\leq \mathcal{O}(1)\,\varepsilon_3\int_0^{t-\tau_{3,+}} e^{-\frac{\nu}{2}s}\mathrm{d}s \\
&\leq \mathcal{O}(\varepsilon_3)\,.
\end{aligned} \tag{49}$$

Moreover, note that

$$\hat{\beta}_+(t) - \hat{\beta}_\circ(\tau_{3,+}) = \sum_{\mathcal{I}_+}(\mathsf{a}_i(t) - \mathsf{a}_i(\tau_{3,+}))\mathsf{w}_i(\tau_{3,+}) + \sum_{\mathcal{I}_+}(\mathsf{w}_i(t) - \mathsf{w}_i(\tau_{3,+}))\mathsf{a}_i(\tau_{3,+}).$$

In particular,

$$\|\hat{\beta}_+(t) - \hat{\beta}_\circ(\tau_{3,+})\|_2 \leq \sum_{\mathcal{I}_+}|\mathsf{a}_i(t) - \mathsf{a}_i(\tau_{3,+})|\|\mathsf{w}_i(\tau_{3,+})\|_2 + \sum_{\mathcal{I}_+}\|\mathsf{w}_i(t) - \mathsf{w}_i(\tau_{3,+})\|_2\mathsf{a}_i(\tau_{3,+})$$

$$\leq \sqrt{\sum_{\mathcal{I}_+}(\mathsf{a}_i(t)-\mathsf{a}_i(\tau_{3,+}))^2}\sqrt{\sum_{\mathcal{I}_+}\|\mathsf{w}_i(\tau_{3,+})\|_2^2}+\sqrt{\sum_{\mathcal{I}_+}\|\mathsf{w}_i(t)-\mathsf{w}_i(\tau_{3,+})\|_2^2}\sqrt{\sum_{\mathcal{I}_+}\mathsf{a}_i(\tau_{3,+})^2}$$
$$\leq \mathcal{O}(1)\,\|\vartheta(t)-\vartheta(\tau_{3,+})\|_2$$
$$\leq \mathcal{O}(\varepsilon_3)\,.$$

We can thus choose $\varepsilon_3^\star = \mathcal{O}(1)$ and $\varepsilon_4 = \Theta(\varepsilon_3^\star)$ small enough such that Equation (46) still holds, but ε_4 large enough with respect to $\varepsilon_3^\star$ such that the previous inequality ensures for any $t \in [\tau_{3,+}, \tau_{4,+}]$:

$$\|\hat{\beta}_+(t) - \hat{\beta}_\circ(\tau_{3,+})\|_{\Sigma_{n,+}} \leq \frac{\varepsilon_4}{2}.$$

In particular, this implies that $\tau_{4,+} = +\infty$. Since $\vartheta_+(t)$ has finite variation (Equation 49), this also implies that $\lim_{t\to\infty}\vartheta_+(t)$ exists. The same holds for $\vartheta_-(t)$ by symmetric arguments, so that $\lim_{t\to\infty}\vartheta(t)$ exists. Moreover, Equations (47) and (48) imply that

$$\lim_{t\to\infty}\hat{\beta}_+(t) = \beta_{n,+}.$$

This yields the second point of Lemma F.6.

It now remains to prove the first point of Lemma F.6. Note that for any $t \geq \tau_{3,+}$ and $i \in \mathcal{I}_+$:

$$\|\overline{\mathsf{w}}_i(t) - \overline{\mathsf{w}}_i(\tau_{3,+})\|_2 \leq 2\int_{\tau_{3,+}}^t \|D_+(s)\|_2 \mathrm{d}s$$
$$\leq \mathcal{O}(\varepsilon_3)\,.$$

Thanks to the first point of Lemma F.5, we can choose $\varepsilon_3^\star = \Theta(\frac{1}{\sqrt{d}})$ small enough so thatfor any $t \geq \tau_{3,+}$ and $i \in \mathcal{I}_+$:

$$\min_{k\in\mathcal{S}_+}\langle \overline{\mathsf{w}}_i(t), \frac{x_k}{\|x_k\|}\rangle > 0 \quad \text{and} \quad \max_{k\in\mathcal{S}_-}\langle \overline{\mathsf{w}}_i(t), \frac{x_k}{\|x_k\|}\rangle < 0,$$

which concludes the proof of Lemma F.6. □

Proof of Theorem 4.1. We can conclude the proof of Theorem 4.1 by noticing that we can indeed choose $\varepsilon, \varepsilon_2, \varepsilon_3, \varepsilon_4$ such that for any $\lambda \leq \lambda^\star = \Theta(\frac{1}{d})$ and $n \geq n^\star = \Theta_\mu(d^3 \log d)$, with probability $1 - \mathcal{O}\Big(\frac{d^2}{n} + \frac{1}{2^m}\Big)$, the statements of Lemmas F.2 and F.4 to F.6 all simultaneously hold. In particular, the stopping times T_+ and T_- defined in Lemma F.3 are infinite. Lemma F.3 then implies that for any $t \geq \tau$, $\vartheta(t) = \theta(t)$. From then, Lemma F.6 implies Theorem 4.1. □