Gradient flow dynamics of shallow ReLU networks for square loss and orthogonal inputs

Etienne Boursier, Loucas Pillaud-Vivien, Nicolas Flammarion

Gradient flow dynamics of shallow ReLU networks for square loss and orthogonal inputs

Etienne Boursier
TML, EPFL, Switzerland
etienne.boursier@epfl.ch

Loucas Pillaud-Vivien
TML, EPFL, Switzerland
loucas.pillaud-vivien@epfl.ch

Nicolas Flammarion
TML, EPFL, Switzerland
nicolas.flammarion@epfl.ch

Abstract

The training of neural networks by gradient descent methods is a cornerstone of the deep learning revolution. Yet, despite some recent progress, a complete theory explaining its success is still missing. This article presents, for orthogonal input vectors, a precise description of the gradient flow dynamics of training one-hidden layer ReLU neural networks for the mean squared error at small initialisation. In this setting, despite non-convexity, we show that the gradient flow converges to zero loss and characterise its implicit bias towards minimum variation norm. Furthermore, some interesting phenomena are highlighted: a quantitative description of the initial alignment phenomenon and a proof that the process follows a specific saddle to saddle dynamics.

1 Introduction

Artificial neural networks are nowadays trained successfully to solve a large variety of learning tasks. However, a large number of fundamental questions surround their impressive success. Among them, the convergence to global minima of their non-convex training dynamics and their ability to generalise well despite fitting perfectly the dataset have challenged traditional machine learning belief. While a complete theory is still lacking, the machine learning community has recently come up with key steps that allow to tame the complexity of the problem: proving the convergence of gradient flow to zero loss [Mei et al., 2018, Chizat and Bach, 2018, Sirignano and Spiliopoulos, 2020, Rotskoff and Vanden-Eijnden, 2022], investigating the algorithmic selection of a specific global minimum, often referred as the *implicit bias* of an algorithm [Neyshabur et al., 2014, Zhang et al., 2021]; while paying attention to the importance of the initialisation [Woodworth et al., 2020, Chizat et al., 2019]. The aim of this article is to analyse precisely these three points for regression problems. This is done in a specific setting: for orthogonal inputs, we provide a complete characterisation of the gradient flows dynamics of training one-hidden layer ReLU neural networks with the square loss at small initialisation. We show that this non-convex optimisation dynamics captures most of the complexity mentioned above and thus could be a first step towards analysing more general setups.

Global convergence of training loss for neural networks. Showing convergence of the gradient flow to a global minimum is an open and important question. Beyond the lazy regime (see next paragraph), only a few results were proven in the regression setting. The most promising route might be the link with Wassertein gradient flows for infinite neural networks. In that case, global convergence happens under mild conditions [Chizat and Bach, 2018, Wojtowytsch, 2020]. Other works focus on local convergence [Zhou et al., 2021, Safran et al., 2021], or general criteria that eventually fail to encompass practical setups [Chatterjee, 2022, Chen et al., 2022]. These latter works

36th Conference on Neural Information Processing Systems (NeurIPS 2022).

rest on Polyak-Łojasiewicz inequalities that in fact cannot be satisfied through the whole process if the dynamics travels near saddle points [Liu et al., 2022], as empirically observed [Dauphin et al., 2014]. On the contrary, the present paper proves global convergence without resorting to large overparameterisation, dealing carefully with saddles.

Feature learning and small initialisation. The scale of initialisation plays an essential role in the behavior of the training dynamics. Indeed, an important example is that, at large initialisation, known as the *lazy regime* [Chizat et al., 2019], the neurons move relatively slightly implying that the dynamics is nearly convex and described by an effective kernel method with respect to the *Neural Tangent Kernel* [Jacot et al., 2018, Allen-Zhu et al., 2019, Arora et al., 2019]. Instead, we are interested in another regime where the initialisation scale is small. This regime is known to be richer as it performs *feature learning* [Yang and Hu, 2021] but is also more challenging to analyse as it follows a truly non-convex dynamics (see details in Section 4).

Implicit bias of gradient methods training. There are many global minima to the mean squared error, i.e. ReLU neural networks that perfectly interpolate the dataset. An important question is to understand which one is selected by the gradient flow for a given initialisation [Neyshabur et al., 2014]. For linear neural networks, this question has been answered thoroughly [Arora et al., 2019, Yun et al., 2021, Min et al., 2021] with a discussion on the role of initialisation [Woodworth et al., 2020] and noise [Pesme et al., 2021]. For non-linear activations such as ReLU, no clear implicit bias criteria have been ever exhibited for the square loss besides a conjecture of a quantisation effect [Maennel et al., 2018]. Finally, note that in the classification setting, the favorable behavior of iterates going to infinity simplifies the analysis to prove implicit biases such as: max-margin for the ℓ_2 norm in case of linear models [Soudry et al., 2018, Ji and Telgarsky, 2019b], alignment of inner layers for linear neural networks [Ji and Telgarsky, 2019a] and max-margin for the variation norm induced by neural networks [Kurková and Sanguineti, 2001] in the case of one-hidden layer neural networks [Lyu and Li, 2019, Chizat and Bach, 2020].
Beyond the convergence results, the implicit bias characterisation anticipates the generalisation properties of the returned estimate as discussed in Section 3.2.

Dynamics of training for neural network. In the regression case, the starting point governs where the flow converges. This observation suggests that a complete analysis of the trajectory may be required when one wants to understand the implicit bias in this case. Such descriptions have been undertaken by Maennel et al. [2018], who describe the initial alignment phase at small initialisation, and Li et al. [2020], Jacot et al. [2021] who conjecture that the dynamics travels from saddle to saddle. These papers provide intuitive content that we prove rigorously in the orthogonal setup.
Finally, closest to our work are the following results on the classification of orthogonally separable data [Phuong and Lampert, 2020, Wang and Pilanci, 2021] and linearly separable, symmetric data [Lyu et al., 2021]. The classification setup provides easier tools to analyse the problem: indeed, after the initial alignment phase, the network has already perfectly classified the data points in these settings. From there, it is known that the training loss converges to zero and that the parameters direction is biased towards KKT points of the max-margin problem [Lyu and Li, 2019, Ji and Telgarsky, 2020]. Such tools cannot be applied after the alignment phase for regression, and we resort to a refined analysis of the trajectory to show both global convergence and implicit bias. On the other hand, Lyu et al. [2021] require a precise description of the dynamics to ensure convergence towards specific KKT points of the max-margin problem. Yet, the analysis of the dynamics is simplified by their symmetry assumption: the trajectory does not go through intermediate saddles and all the labels are simultaneously fitted. On the contrary, the dynamics we describe travels near an intermediate saddle point which separates two distinct fitting phases. This behaviour largely complicates the analysis, besides being more representative of the saddle to saddle dynamics observed in general settings.

1.1 Main contributions

We make the following contributions.

- We prove the convergence of the gradient flow towards a global minimum of the non-convex training loss for small enough initialisation and finite width.
- We characterise the global optimum retrieved for infinitesimal initialisation as a minimum ℓ_2 norm interpolator, which implies a minimum variation norm in terms of prediction function.
- As important as the convergence result, the dynamics is portrayed in Section 4: we quantitatively detail its different phases (alignment and fitting) and show it follows a saddle to saddle dynamics.

1.2 Notations

We denote by $\mathbb{1}_A$ the function equal to 1 if A is true and 0 otherwise. $\mathcal{U}(S)$ is the uniform distribution over the set S and $\mathcal{N}(\mu, \Sigma)$ is a Gaussian of mean μ and covariance Σ. We denote $\nabla_\theta h_\theta(x)$ the gradient of $\theta \mapsto h_\theta(x)$ at fixed x. For any $n \in \mathbb{N}^*$, $[\![n]\!]$ denotes the tuple of integers between 1 and n. The scalar product between $x, y \in \mathbb{R}^d$ is denoted by $\langle x, y \rangle$ and the Euclidean norm is denoted by $\|\cdot\|$ and called ℓ_2. $\mathcal{S}_{d-1}$ denotes the sphere of $\mathbb{R}^d$ for the Euclidean norm. $B(\theta, r)$ is the Euclidean ball of center θ and radius r. All the detailed proofs of the claimed results are deferred to the Appendix.

2 Setup and preliminaries

2.1 One-hidden layer neural network and training loss

Model. Let us fix an integer $n \in \mathbb{N}^*$ as well as input data $(x_1, \dots, x_n) \in (\mathbb{R}^d)^n$ and outputs $(y_1, \dots, y_n) \in \mathbb{R}^n$. We are interested in the minimisation of the mean squared error:

$$L(\theta) := \frac{1}{2n} \sum_{k=1}^{n} \left(h_\theta(x_k) - y_k\right)^2, \qquad \text{where } h_\theta(x) := \sum_{j=1}^{m} a_j \sigma(\langle w_j, x \rangle) \tag{1}$$

is a one-hidden layer neural network of width m defined with parameters $\theta = (a, W) \in \mathbb{R}^m \times \mathbb{R}^{m \times d}$. The vector $a \in \mathbb{R}^m$ stands for the weights of the last layer and $W^\top = [w_1 \cdots w_m] \in \mathbb{R}^{d \times m}$, where each $w_j \in \mathbb{R}^d$ represents a hidden neuron. To encompass the effect of the bias, an additional component can be added to the inputs $x^\top \leftarrow [x^\top, 1]$ without changing our results. Finally, the activation function σ is the ReLU: $\sigma(x) := \max\{0, x\}$.

We introduce here the main assumptions on the data inputs.

Assumption 1. *The input points form an orthonormal family, i.e.* $\forall k, k' \in [\![n]\!]$, $\langle x_k, x_{k'} \rangle = \mathbb{1}_{k=k'}$.

The data are assumed to be normalized only for convenience—the real limitation being that they are pairwise orthogonal. This assumption is exhaustively discussed in Section 3.2.

Assumption 2. *For all* $k \in [\![n]\!]$, $y_k \neq 0$ *and* $\sum_{k|y_k>0} y_k^2 \neq \sum_{k|y_k<0} y_k^2$.

This assumption on the data output is mild, e.g. has zero Lebesgue measure, and only permits to exclude degenerate situations.

Gradient flow. As the limiting dynamics of the (stochastic) gradient descent with infinitesimal step-sizes [Li et al., 2019], we study the following gradient flow

$$\frac{\mathrm{d}\theta^t}{\mathrm{d}t} = -\nabla L(\theta^t) = -\frac{1}{n} \sum_{k=1}^{n} \left(h_{\theta^t}(x_k) - y_k\right) \nabla_\theta h_{\theta^t}(x_k), \tag{2}$$

initialised at $\theta^0 := (a^0, W^0)$. Since the ReLU is not differentiable at 0, the dynamics should be defined as a subgradient inclusion flow [Bolte et al., 2010]. However, we show in Appendix D that the *only* ReLU subgradient that guarantees the existence of a global solution is $\sigma'(x) = \mathbb{1}_{x>0}$. Hence, we stick with this choice throughout the paper. Another important difficulty of this non-differentiability is that Cauchy-Lipschitz theorem does not apply and uniqueness is not ensured. There have been attempts to define the solution of this Ordinary Diffential Equation (ODE) unequivocally [Eberle et al., 2021] as well as ways to circumvent this difficulty by resorting to smooth activations or additional data assumptions [Wojtowytsch, 2020, Chizat and Bach, 2020]. Yet, we do not follow this line and demonstrate our results *for all the gradient flows* satisfying Equation (2).

2.2 Preliminary properties and initialisation

Let us derive here some preliminary properties of the gradient flows. If we rewrite explicitly the dynamics of Equation (2) on each layer separately, we have straightforwardly that for all $j \in [\![m]\!]$,

$$\frac{\mathrm{d}a_j^t}{\mathrm{d}t} = \langle D_j^{\theta^t}, w_j^t \rangle \qquad \text{and} \qquad \frac{\mathrm{d}w_j^t}{\mathrm{d}t} = D_j^{\theta^t} a_j^t, \tag{3}$$

where $D_j^{\theta^t} := -\frac{1}{n}\sum_{k=1}^n \mathbb{1}_{\langle w_j^t, x_k\rangle>0}\,(h_{\theta^t}(x_k)-y_k)\,x_k$ is a vector of $\mathbb{R}^d$. This vector solely depends on θ^t through the prediction function h_{θ^t} and on the neuron j through its activation vector $\mathsf{A}(w_j^t)$; where, for a vector $w\in\mathbb{R}^d$, $\mathsf{A}(w) := (\mathbb{1}_{\langle w, x_1\rangle>0}, \ldots, \mathbb{1}_{\langle w, x_n\rangle>0}) \in \{0,1\}^n$. From Equation (3), we deduce the following balancedness property [Arora et al., 2019].

Lemma 1. *For all $t\geq 0$ and all $j\in[\![m]\!]$, $(a_j^t)^2 - \|w_j^t\|^2 = (a_j^0)^2 - \|w_j^0\|^2$. Assume furthermore that for all $j\in[\![m]\!]$, the initialisation is balanced and non-zero: $|a_j^0| = \|w_j^0\| > 0$. Then $|a_j^t| = \|w_j^t\| > 0$ and letting $\mathsf{s} = \operatorname{sign}(a^0)\in\{1,-1\}^m$, for all $t\geq 0$, we have that $a_j^t = \mathsf{s}_j\|w_j^t\|$.*

Importantly, by Lemma 1, the study of Equation (3) reduces to the hidden layer W solely. We consider the following balanced initialisation:

$$\theta^0 = (a^0, W^0) \quad \text{with} \quad \begin{cases} w_j^0 = \lambda\, g_j \text{ where } g_j \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, I_d), \\ a_j^0 = \mathsf{s}_j\|w_j^0\| \text{ where } \mathsf{s}_j \overset{\text{i.i.d.}}{\sim} \mathcal{U}(\{-1,1\}). \end{cases} \tag{4}$$

As already stated, we are interested in the regime where the initialisation scale $\lambda>0$ is small. We also introduce the following sets of neurons that are crucial in the fitting process

$$S_{+,1} := \left\{ j\in[\![m]\!] \mid \mathsf{s}_j = +1 \quad \text{and} \quad \text{for all } k \text{ such that } y_k>0, \langle w_j^0, x_k\rangle \geq 0 \right\}, \tag{5}$$

$$S_{-,1} := \left\{ j\in[\![m]\!] \mid \mathsf{s}_j = -1 \quad \text{and} \quad \text{for all } k \text{ such that } y_k<0, \langle w_j^0, x_k\rangle \geq 0 \right\}. \tag{6}$$

Assumption 3. *The sets $S_{+,1}$ and $S_{-,1}$ are both non-empty.*

Assumption 3 states that there are some neurons in two given cones at initialisation. It holds with probability 1 when the support of initialisation covers all directions and the width m of the network goes to infinity. This is thus a weaker condition than the *omni-directionality* of neurons at initialisation [Wojtowytsch, 2020], which is instrumental to show convergence in the mean field regime [Chizat and Bach, 2018]. On the other hand, it is stronger than the alignment condition of Abbe et al. [2022], which is known to be necessary for weak learning but might not lead to the implicit bias described in the next section.

3 Convergence and implicit bias characterisation

3.1 Main result

Theorem 1 below states our main result on the convergence and implicit bias of one-hidden layer ReLU networks for regression tasks with orthogonal data.

Theorem 1. *Under Assumptions 1 to 3, there exists $\lambda^*>0$ depending only on the data and the width such that, if $\lambda\leq\lambda^*$, the gradient flow initialised according to Equation* (4) *converges almost surely to some θ_λ^∞ of zero training loss, i.e. $L(\theta_\lambda^\infty)=0$. Furthermore, there exists θ^* such that*

$$\lim_{\lambda\to 0}\lim_{t\to\infty}\theta^t = \theta^* \in \operatorname*{argmin}_{L(\theta)=0}\|\theta\|^2. \tag{7}$$

The significance of this result is thoroughly discussed in Section 3.2. Note that a quantitative and non-asymptotic version of Theorem 1, both in time and λ, is stated in Lemma 12 (Appendix B). Roughly, it states that the dynamics has already nearly converged after a time of order $-\ln(\lambda)$ and then that the convergence happens at exponential speed. Note also that the neural network need not be overparametrised for the result to hold: the only sufficient and necessary requirement on the width m stems from Assumption 3.

Sketch of proof. The proof of Theorem 1 rests on a precise description of the training dynamics, which is divided into four different phases. We here only sketch it at a very high level and a more thorough description, with quantitative intermediate lemmas, is given in Section 4.

During the first phase, hidden neurons align to a few representative directions, while remaining close to 0 in norm. In particular, all hidden neurons in $S_{+,1}$ (resp. $S_{-,1}$) align with some key vector D_+ (resp. $-D_-$) defined in Section 4.1. During the second phase, the neurons aligned with D_+ grow in norm, while staying aligned with D_+, until fitting all the positive labels of the dataset (up to some error scaling with λ). Meanwhile, all the other neurons stay idle. Then similarly, the neurons aligned

with $-D_-$ grow in norm during the third phase, until nearly fitting all the negative labels. Meanwhile, these neurons remain aligned with $-D_-$ and all other neurons remain idle. The precise description of these three phases is obtained by analysing the solutions of the limit ODEs when $\lambda = 0$. The approximation errors that occur from dealing with non-zero λ are then carefully handled via Grönwall comparison arguments. Due to the large time scales (of order $-\ln(\lambda)$), the error can propagate on such large time spans. Handling these error terms is the main challenge of our proof and remains intricate despite the orthogonality assumption.

After these three phases (which last a time $-\ln(\lambda)/\|D_-\|$), the parameters vector is close to some minimal ℓ_2-norm interpolator. From there we show, exploiting a local Polyak-Łojasiewicz condition, that the dynamics converges at exponential speed to a global minimum close to this interpolator.

3.2 Discussion

Even if the orthogonal setting we consider is quite restrictive, it carries several characteristics that may be generic, either because they have been observed empirically, shown in related contexts or simply conjectured. We discuss these important points below.

Convergence to zero loss. Theorem 1 states that the gradient flow converges to zero loss. Such a result is simple to show when the loss satisfies a Polyak-Łojasiewicz (PL) inequality [Bolte et al., 2007]: $\|\nabla L\|^2 \geq cL$ for $c > 0$. However, here, as the dynamics travels near saddles, this inequality is not verified through all the process. Circumventing this global argument, it is yet possible to formulate a refined analysis and show convergence if the dynamics arrives in a region where a local PL stands with a large enough constant. This refined analysis, inspired by the recent work of Chatterjee [2022][1], allows to characterise properly the last phase of the dynamics. We believe that this approach may help in showing convergence in other non-convex gradient flow/descent.

On the implicit bias. Additionally, Theorem 1 states that the gradient flow at infinitesimally small initialisation *selects* global minimisers with the smallest ℓ_2 parameter norm. To our knowledge, this is the first characterisation of the implicit bias for regression with non-linear neural networks. Although it might not hold for some degenerate situations [Vardi and Shamir, 2021], we believe it to be true beyond the orthogonal case.

This regularisation is *implicit*, meaning that this effect does not result from any explicit regularisation (e.g. weight decay) performed during training [Shevchenko et al., 2021, Parhi and Nowak, 2022]. This is only a consequence of the inner structure of the gradient flow and the scale of initialisation.

Furthermore, the implicit bias in parameter space can be translated in function space. Indeed, if we introduce formally the space of (infinite) neural networks, i.e. functions written as $f(x) := \int \sigma(\langle \theta, x \rangle) \mathrm{d}\mu(\theta)$, where μ is a positive finite measure on $\mathcal{S}_{d-1}$. Then, we can define the *variation norm*, $\|f\|_{\mathcal{F}_1}$, as the infimum of $\mu(\mathcal{S}_{d-1})$ over such representations [Kurková and Sanguineti, 2001, Bach, 2017]. We have the following link between the two formulations

$$\min_{L(\theta)=0} \frac{1}{2}\|\theta\|_2^2 = \min_{L(f)=0} \|f\|_{\mathcal{F}_1}, \tag{8}$$

with a slight abuse of notation when defining $L(f)$. Note that the result in terms of the ℓ_2-norm of the parameters is strictly stronger than that of the $\mathcal{F}_1$-norm of the function [Neyshabur et al., 2014].

Following Equation (8), note the striking parallel between the inductive bias of infinitesimally small initialisation for regression and that of the classification problem with the logistic loss as a max-margin problem with respect to the $\mathcal{F}_1$-norm [Chizat and Bach, 2020]. As already observed in the linear case [Woodworth et al., 2020], in contrast with classification, infinitesimally small initialisation is instrumental in regression to be biased towards small $\mathcal{F}_1$-norm functions. The role of initialisation is illustrated empirically in Appendix A.

Finally, let us stress that we did not address the question of what functions solve Equation (8), nor the question of the generalisation implied by such a bias. Related works on the first point come from a functional description of norms related to $\mathcal{F}_1$ [Savarese et al., 2019, Ongie et al., 2019, Debarre et al., 2022]. For the generalisation properties of small $\mathcal{F}_1$ norm functions, we refer to Kurková and Sanguineti [2001], Bach [2017]. Importantly, we recall that the question of how well low $\mathcal{F}_1$-norm functions generalise depends heavily on the *a priori* we have on the ground-truth [Petrini et al., 2022].

[1]Note that the argument is certainly not new, but the cited article has the benefit of clearly presenting it.

The initial alignment phenomenon. An important characteristics of the loss landscape is that the origin is a saddle point. Hence, as the dynamics is initialised at small scale λ, the radial movement is slow and neurons move out of the saddle after time scale $-\ln \lambda$. Meanwhile, the tangential movement of the neurons rules the dynamics and aligns their directions towards specific vectors. This has been first explained by Maennel et al. [2018] and referred as the *quantisation* phenomenon, because neural networks weights collapse to a small finite number of directions. We emphasise that this phase happens generically when initialisation is near the origin and that this part of our analysis can be directly extended to the general (i.e. non-orthognal) case. Phuong and Lampert [2020], Lyu et al. [2021] analysed a similar early alignment for classification with specific data structures.

The saddle to saddle dynamics. When initialising the dynamics of a gradient flow near a saddle point of the loss, it is expected (but hard to prove generically) that the dynamics will alternate slow movements near saddles and rapid junctions between them. Such a behavior has been conjectured for linear neural networks [Li et al., 2020, Jacot et al., 2021] initialised near the origin. We precisely prove that such a phenomenon occurs: after initialisation, the dynamics visits one strict saddle. See Section 4, Fact 1 for more details.

Limitations and possible relaxations. As its main limitation, Theorem 1 assumes orthogonal data points x_k. The orthogonality assumption disentangles the analysis as the different phases, where either the neurons align towards some direction or grow in norm, are well separated in that case. More precisely, the neurons do not change in direction once they have a non-zero norm in the case of orthogonal data. This separation between alignment and norm growth does not hold in the general case, as observed empirically in Appendix A. Extending our result to more general data thus remains a major challenge and requires additional theoretical tools. Nonetheless, as it can be the case in high dimension, our analysis can easily be extended to nearly orthogonal data where $|\langle x_k, x_{k'}\rangle| \leq \delta$, with δ of order λ. If however δ is much larger than the initialisation scale, the dynamics is drastically different and becomes as hard as the general case to analyse. In Appendix A, we observe similar dynamics for high dimensional data, where the loss converges towards 0, goes through an intermediate saddle point and the final solution is close to a 2 neurons network.

A minor assumption is the balanced initialisation, i.e. $\|w_j^0\| = |a_j^0|$. If instead we initialise a^0 as a Gaussian scaling with λ, the initialisation would be nearly balanced for small λ. This assumption is thus mostly used for simplicity and our analysis can be extended to unbalanced initialisations.

It is unclear whether our analysis can be extended to any homogeneous activation function. The training trajectory might indeed not be biased towards minimal ℓ_2-norm for *leaky ReLU* activations [Lyu et al., 2021, Theorem 6.2]. Contrary to some beliefs, it suggests that the ℓ_2 implicit bias phenomenon does not occur for any homogeneous activation function, but might instead be specific to the ReLU.

The overparameterisation regime. Assumption 3 states a deterministic condition to guarantee convergence towards a minimal norm interpolator. This condition is not only sufficient, but also *necessary* for implicit bias towards minimal ℓ_2 norm. For isotropic initialisations, the width m needs to be exponential in the number of data points n for Assumption 3 to hold with high probability.

With a smaller (e.g. polynomial in n) number of neurons, Assumption 3 does not hold anymore. In that case, the training loss should still converge to 0, but the ℓ_2-norm of the parameters will not be minimal. More precisely, the estimated function will have more than two kinks. However, an adapted analysis might still show some sparsity in the number of kinks (and thus a weak bias) of the final solution. The training trajectory would then go through multiple saddle points (one saddle per kink).

Scale of initialisation. The exact value of λ^* is omitted for exposition's clarity. Roughly, it can be inferred from the analysis that λ^* scales as $\frac{\Theta(1)}{\sqrt{m}} e^{-\Theta(n)}$. Interestingly, the $\frac{1}{\sqrt{m}}$ term is reminiscent of the mean field regime, which is known to induce implicit bias [Chizat and Bach, 2020, Lyu et al., 2021]. On the other hand, the exponential dependency in n is common in the implicit bias literature [Woodworth et al., 2020]. For larger values of λ (but still in the mean field regime), the parameters empirically seem to also converge towards a minimal norm interpolator. The analysis yet becomes more intricate and we do not observe any separation between Phase 2 and Phase 3, i.e. there is no intermediate saddle in the trajectory.

4 Fine dynamics description: alignments and saddles

This section describes thoroughly the training dynamics of the gradient flow. It presents and discusses quantitative lemmas on the state of the neural network at the end of each different phase. In particular, mathematical formulations of the early alignment and saddle to saddle phenomena are provided.

4.1 Additional notations

First, we need to introduce additional notations for this section. We define vectors D_+ and D_- that are the two directions towards which the neurons align

$$D_+ \coloneqq \frac{1}{n}\sum_{k|y_k>0} y_k x_k \qquad \text{and} \qquad D_- \coloneqq \frac{1}{n}\sum_{k|y_k<0} y_k x_k.$$

We also need to define $c \coloneqq \max_{j\in[\![m]\!]} \|w_j^0\|/\lambda$ and $r \coloneqq \|D_+\|/\|D_-\|$. Assumption 2 implies that $r \neq 1$, and by symmetry we can assume $r > 1$ without any loss of generality. We additionally fix constants $\lambda_*, \varepsilon > 0$, small enough and depending only on the dataset and the width m.

Spherical coordinates. As radial and tangential movements are almost decoupled during the dynamics, it is natural to introduce the spherical coordinates of the neurons: for all $j \in [\![m]\!]$, denote $w_j = \mathrm{e}^{\rho_j} \cdot \mathsf{w}_j$, where $\rho_j = \ln \|w_j\| \in \mathbb{R}$ and $\mathsf{w}_j = w_j/\|w_j\| \in \mathcal{S}_{d-1}$. In these adapted coordinates, the system of ODEs (3) reduces to:

$$\frac{\mathrm{d}\rho_j^t}{\mathrm{d}t} = \mathsf{s}_j \langle D_j^{\theta^t}, \mathsf{w}_j^t\rangle \qquad \text{and} \qquad \frac{\mathrm{d}\mathsf{w}_j^t}{\mathrm{d}t} = \mathsf{s}_j \left(D_j^{\theta^t} - \langle D_j^{\theta^t}, \mathsf{w}_j^t\rangle \mathsf{w}_j^t\right). \tag{9}$$

4.2 Training dynamics

This section precisely describes the phases of the dynamics, summarised in Figure 1.

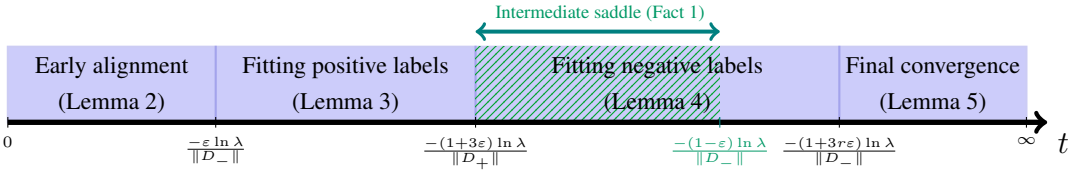


Figure 1: Timeline of the training dynamics.

Neuron alignment phase. During the first phase, all the neurons remain small in norm, while moving tangentially (i.e. in directions). The neurons align according to several key directions: an initial clustering of neurons' directions happens in this early phase, as observed by Maennel et al. [2018]. As the neurons have small norm, $h_{\theta^t} \approx 0$ for this phase and Equation (9) approximates

$$\frac{\mathrm{d}\mathsf{w}_j^t}{\mathrm{d}t} \approx \mathsf{s}_j \left(D_j^0 - \langle D_j^0, \mathsf{w}_j^t\rangle \mathsf{w}_j^t\right). \tag{10}$$

This ODE corresponds to the descent/ascent gradient flow (depending on the sign of s_j) on the sphere with objective $\langle D_j^0, \mathsf{w}_j\rangle$. All neurons end up minimizing or maximizing their scalar product with D_j^0, which only depends on the activation of w_j. As a consequence, neurons with similar activations align towards the same vector, leading to some quantisation of the neurons' directions. This alignment happens in a relatively short time, so that the neurons cannot largely grow in norm. Lemma 2 below quantifies this effect for neurons in $S_{+1,}$ and $S_{-,1}$, which are crucial to the training dynamics. Since all other neurons remain small in norm during the whole process, we do not focus on their direction.

Lemma 2 (First phase)**.** *For $\lambda \leq \lambda^*$, we have the following inequalities for $t_1 = \frac{-\varepsilon \ln(\lambda)}{\|D_-\|}$:*

(i) neurons in $S_{+,1}$ are aligned with D_+: $\quad \forall j \in S_{+,1}, \langle \mathsf{w}_j^{t_1}, D_+\rangle \geq (1-2\lambda^\varepsilon)\|D_+\|$,

(ii) neurons in $S_{-,1}$ are aligned with $-D_-$: $\quad \forall j \in S_{-,1}, \langle \mathsf{w}_j^{t_1}, -D_-\rangle \geq (1-2\lambda^\varepsilon)\|D_-\|$,

(iii) all neurons have small norm: $\quad \forall j \in [\![m]\!], \|w_j^{t_1}\| \leq 2c\lambda^{1-r\varepsilon}$.

Fitting positive labels. During the second phase, the norm of the neurons in $S_{+,1}$ (which are aligned with D_+) grows until fitting all positive labels. Meanwhile, all the other neurons do not move significantly. The key approximate ODE of this phase is given for $u_+(t) := \sum_{j \in S_{+,1}} \|w_j^t\|^2$ by

$$\frac{\mathrm{d}u_+(t)}{\mathrm{d}t} \approx 2\|D_+\| \left(1 - \frac{u_+(t)}{n\|D_+\|}\right) u_+(t).$$

This equation implies that $u_+(t)$, the sum of the squared norms of neurons in $S_{+,1}$, eventually converges to $n\|D_+\|$ within a time $-\ln(\lambda)/\|D_+\|$. Meanwhile, it needs to be shown that these neurons remain aligned with D_+ and that the other neurons remain small in norm. This fine control is technical and relies on the orthogonality assumption. If data were not orthogonal, neurons could indeed realign while growing in norm as illustrated by Figure 4d in Appendix A. Lemma 3 below describes the state of the network at the end of the second phase.

Lemma 3 (Second phase)**.** *If $\lambda \leq \lambda^*$, then for some time $t_2 \leq -\frac{1+3\varepsilon}{\|D_+\|}\ln(\lambda)$:*

(i) neurons in $S_{+,1}$ are aligned with D_+: $\quad \forall j \in S_{+,1}, \langle \mathsf{w}_j^{t_2}, D_+ \rangle \geq \|D_+\| - \lambda^{\frac{\varepsilon}{2}}$,

(ii) neurons in $S_{+,1}$ have a large norm: $\quad \sum_{j \in S_{+,1}} \|w_j^{t_2}\|^2 = n\|D_+\| - \lambda^{\frac{\varepsilon}{5}}$,

(iii) other neurons have small norm: $\quad \forall j \in [\![m]\!] \setminus S_{+,1}, \|w_j^{t_2}\| \leq 2c\lambda^{\varepsilon}$.

These three points directly imply that the loss is of order $\lambda^{\frac{\varepsilon}{5}}$ on the positive labels at time t_2.

Saddle to saddle dynamics. As explained above, the positive labels are almost fitted by the action of the neurons belonging to $S_{+,1}$ at the end of the second phase, whereas the other neurons still have infinitesimally small norm. At this point, the dynamics has reached the vicinity of a strict saddle point and requires a long time to escape it. The analysis actually leads to the following fact:

Fact 1. *There exists a (strict) saddle point $\theta_S \neq 0$ of L such that if $\lambda \leq \lambda^*$:*

$$\forall t \in \left[-\frac{1+3\varepsilon}{\|D_+\|}\ln(\lambda), -\frac{1-\varepsilon}{\|D_-\|}\ln(\lambda)\right], \quad \textit{we have } \|\theta^t - \theta_S\| \leq \lambda^{\frac{\varepsilon}{5}}.$$

The training trajectory thus starts at the saddle point 0 and passes through a second non-trivial saddle point at the end of the second phase. This lemma illustrates the phenomenon of *saddle to saddle dynamics* discussed in Section 3.2 and conjectured for linear models by Li et al. [2020], Jacot et al. [2021]. This intermediate saddle point is escaped when the norms of the neurons in $S_{-,1}$ have significantly grown (i.e. become non-zero), which happens during a third phase described below.

Fitting negative labels. The norm of the neurons in $S_{-,1}$ (which are aligned with $-D_-$) grows until fitting all negative labels during the third phase. Meanwhile, all other neurons do not move significantly. The additional difficulty in the analysis of this phase compared to the second one is that of controlling the possible movements of neurons in $S_{+,1}$. Their norm is indeed large during the whole phase, but they do not change consequently, because the positive labels are nearly perfectly fitted.

Lemma 4 (Third phase)**.** *If $\lambda \leq \lambda^*$, then for some time $t_3 \leq -\frac{1+3r\varepsilon}{\|D_-\|}\ln(\lambda)$:*

(i) neurons in $S_{-,1}$ are aligned with $-D_-$: $\quad \forall j \in S_{-,1}, \langle \mathsf{w}_j^{t_3}, -D_- \rangle \geq \|D_-\| - \lambda^{\frac{\varepsilon}{14}}$,

(ii) neurons in $S_{-,1}$ have a large norm: $\quad \sum_{j \in S_{-,1}} \|w_j^{t_3}\|^2 = n\|D_-\| - \lambda^{\frac{\varepsilon}{29}}$,

(iii) neurons in $S_{+,1}$ did not move since phase 2: $\quad \forall j \in S_{+,1}, \|w_j^{t_2} - w_j^{t_3}\| \leq \lambda^{\frac{\varepsilon}{15}}$,

(iv) other neurons have small norm: $\quad \forall j \in [\![m]\!] \setminus (S_{+,1} \cup S_{-,1}), \|w_j^{t_3}\| \leq 3c\lambda^{\varepsilon}$.

Thanks to the orthogonality assumption, the set of minimal ℓ_2-norm interpolators can be exactly described by Proposition 1 in Appendix C. The minimal interpolators are actually *equivalent* to a neural network of width 2: the first hidden neuron is collinear with D_+ and the second one is collinear with $-D_-$. Lemma 4 then ensures at the end of the third phase that the parameter vector is $\lambda^{\frac{\varepsilon}{29}}$-close to an interpolator θ^* of minimal ℓ_2-norm, that does not depend on λ (see Lemma 12). It remains to show that the training trajectory converges to a point close to this interpolator at infinity.

Convergence phase. To show this final convergence, we use a local PL condition given by Lemma 5.

Lemma 5 (Local PL condition)**.** *For $\lambda \leq \lambda^*$, we have the following lower bound on the PL constant*

$$\inf_{\theta \in B(\theta^*,\, \lambda^{\frac{\varepsilon}{240}}) \cap \Theta} \frac{\|\nabla L(\theta)\|^2}{L(\theta)} \geq \|D_-\|,$$

where Θ is the set of parameters verifying the balancedness property.

Adapting arguments from the recent work by Chatterjee [2022], this implies that the training trajectory converges to an interpolator and stays in the aforementioned ball. It thus converges exponentially at a rate $\|D_-\|$ to a point close to a minimal norm interpolator, and the distance to this point goes to 0 when λ goes to 0, hence implying Theorem 1. This exponential rate is only asymptotical: the dynamics still require a large time $-\ln(\lambda)/\|D_-\|$ to escape the two first saddles.

5 Experiments

This section confirms empirically the dynamics described in Section 4 on an orthogonal toy example. The code and animated versions of the figures are available in `github.com/eboursier/GFdynamics`. Additional experiments can be found in Appendix A; they illustrate the necessity of small initialisation for implicit bias and present similar experiments on non-orthogonal toy data. For the latter, we observe some similar training phenomena, but major differences appearing in the dynamics highlight the difficulty of dealing with non-orthogonality.

We consider the following two-point dataset: $x_1, y_1 = (-0.5, 1), -1$ and $x_2, y_2 = (2, 1), 1$. It corresponds to unidimensional data with a second 1 coordinate for the bias term. We choose unidimensional data for a simpler visualisation. However, it restricts the number of observations to $n = 2$ to maintain orthogonality. Also, the inputs' norms are not 1 here, but we recall that our analysis is not specific to this case. The width of the neural network is $m = 60$. We choose a balanced initialisation at scale $\lambda = 10^{-6}/\sqrt{m}$. We then run gradient descent with a step size 10^{-3} to approximate the gradient flow trajectory.

Figure 2 shows the training dynamics on this example. In particular, the state of the network is shown at different steps. In Figure 2a, all the neurons are close to 0 at initialisation. Figure 2b shows the end of the first phase, where the neurons are aligned towards two key directions. After the second phase, shown in Figure 2c, all the neurons aligned with D_+ have grown in norm and the positive label is perfectly fitted. Similarly at the end of third phase in Figure 2d, all neurons aligned with $-D_-$ have grown in norm and the negative label is fitted.

At the end of training, the loss is 0 and the estimated function is *simple*. In particular, it only has two kinks, which illustrates the sparsity induced by the implicit bias. Also, the final estimated function might be counter-intuitive. Previous works on implicit bias indeed conjectured that the learned estimator is linear if the data can be linearly fitted [Kalimeris et al., 2019, Lyu et al., 2021]. However, the learned function in Figure 2d has a smaller $\mathcal{F}_1$-norm than the linear interpolator.

Figure 3 shows the evolution of the loss during training. The saddle to saddle dynamics is well observed here: the parameters vector starts from the 0 saddle point at initialisation and needs 5000 iterations to leave this first saddle. A second saddle is then encountered at the end of the second phase and the trajectory only leaves this saddle around iteration 11000, once the norm of the neurons in $S_{-,1}$ start being significant during the third phase. All these different experiments confirm Theorem 1 and the precise dynamics described in Section 4. Moreover, such training phenomena are not specific to the orthogonal data case, as observed in Appendix A.

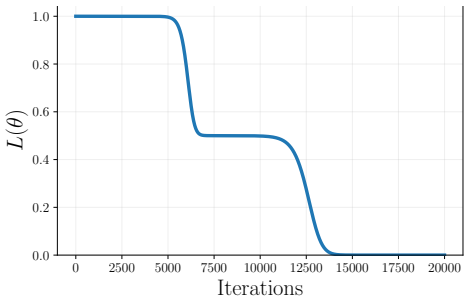
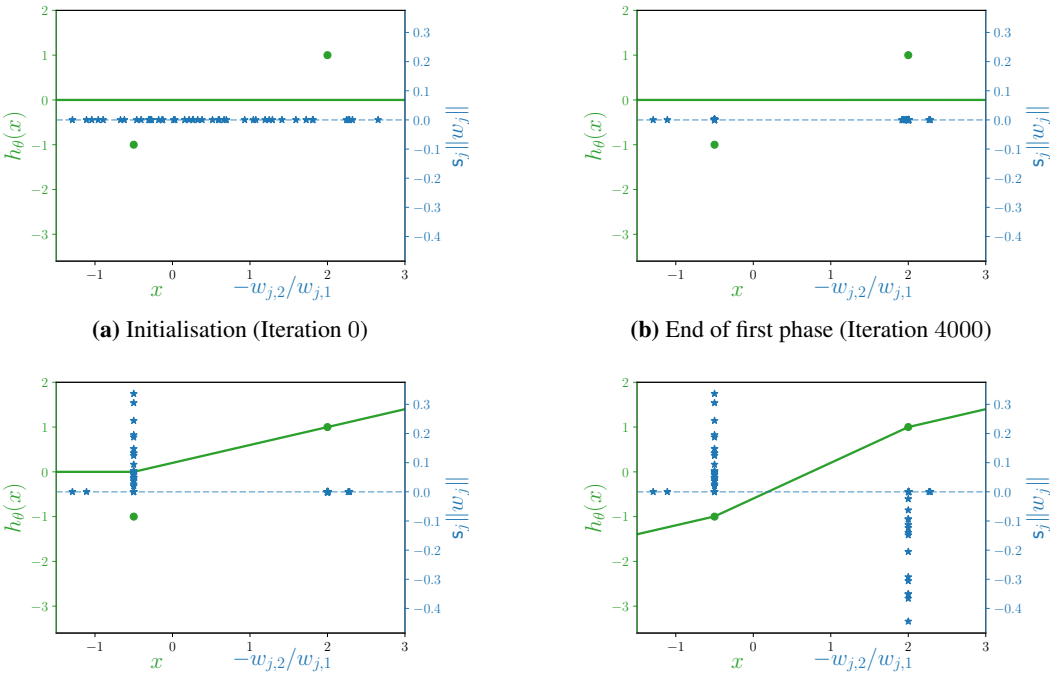


Figure 3: Evolution of the training loss.

6 Conclusion and perspectives

We have shown that the training of non-linear neural networks on orthogonal data presents a rich dynamics with a small and omnidirectional intialisation. Convergence holds generically despite



(a) Initialisation (Iteration 0)

(b) End of first phase (Iteration 4000)

(c) End of second phase (Iteration 8000)

(d) End of third phase (Iteration 15000)

Figure 2: State of training at different stages. The green dots correspond to the data, while the green line is the estimated function h_θ. Each blue star represents a neuron w_j: its x-axis value is given by $-w_{j,2}/w_{j,1}$, which coincides with the position of the kink of its associated ReLU; its y-axis value is given by $\mathsf{s}_j\|w_j\|$, which we recall is the associated value of the output layer.

a truly non-convex landscape and the limit enjoys an implicit bias as a minimum ℓ_2 parameter norm. Obviously, removing the orthogonal assumption on the inputs, while keeping a fine level of description is a major, but difficult, perspective for future work. Another key point to better understand the good generalisation of neural networks is to analyse the properties of the functions solving the minimum variation norm problem stated in Equation (8).

References

Emmanuel Abbe, Elisabetta Cornacchia, Jan Hazla, and Christopher Marquis. An initial alignment between neural network and target is needed for gradient descent to learn. In *International Conference on Machine Learning*, pages 33–52. PMLR, 2022.

Zeyuan Allen-Zhu, Yuanzhi Li, and Yingyu Liang. Learning and generalization in overparameterized neural networks, going beyond two layers. *Advances in neural information processing systems*, 32, 2019.

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.

Francis Bach. Breaking the curse of dimensionality with convex neural networks. *The Journal of Machine Learning Research*, 18(1):629–681, 2017.

David Bertoin, Jérôme Bolte, Sébastien Gerchinovitz, and Edouard Pauwels. Numerical influence of ReLu'(0) on backpropagation. *Advances in Neural Information Processing Systems*, 34, 2021.

Alberto Bietti and Julien Mairal. On the inductive bias of neural tangent kernels. *Advances in Neural Information Processing Systems*, 32, 2019.

Jérôme Bolte, Aris Daniilidis, and Adrian Lewis. The Łojasiewicz inequality for nonsmooth subanalytic functions with applications to subgradient dynamical systems. *SIAM Journal on Optimization*, 17(4):1205–1223, 2007.

Jérôme Bolte, Aris Daniilidis, Olivier Ley, and Laurent Mazet. Characterizations of Łojasiewicz inequalities: subgradient flows, talweg, convexity. *Transactions of the American Mathematical Society*, 362(6):3319–3363, 2010.

Sourav Chatterjee. Convergence of gradient descent for deep neural networks. *arXiv preprint arXiv:2203.16462*, 2022.

Zhengdao Chen, Eric Vanden-Eijnden, and Joan Bruna. On feature learning in neural networks with global convergence guarantees. *arXiv preprint arXiv:2204.10782*, 2022.

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.

Yann N Dauphin, Razvan Pascanu, Caglar Gulcehre, Kyunghyun Cho, Surya Ganguli, and Yoshua Bengio. Identifying and attacking the saddle point problem in high-dimensional non-convex optimization. *Advances in neural information processing systems*, 27, 2014.

Thomas Debarre, Quentin Denoyelle, Michael Unser, and Julien Fageot. Sparsest piecewise-linear regression of one-dimensional data. *Journal of Computational and Applied Mathematics*, 406: 114044, 2022.

Simon Eberle, Arnulf Jentzen, Adrian Riekert, and Georg S Weiss. Existence, uniqueness, and convergence rates for gradient flows in the training of artificial neural networks with ReLU activation. *arXiv preprint arXiv:2108.08106*, 2021.

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

Arthur Jacot, François Ged, Franck Gabriel, Berfin Şimşek, and Clément Hongler. Saddle-to-saddle dynamics in deep linear networks: Small initialization training, symmetry, and sparsity. *arXiv preprint arXiv:2106.15933*, 2021.

Ziwei Ji and Matus Telgarsky. Gradient descent aligns the layers of deep linear networks. In *International Conference on Learning Representations*, 2019a.

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

Ziwei Ji and Matus Telgarsky. Directional convergence and alignment in deep learning. *Advances in Neural Information Processing Systems*, 33:17176–17186, 2020.

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.

Vera Kurková and Marcello Sanguineti. Bounds on rates of variable-basis and neural-network approximation. *IEEE Transactions on Information Theory*, 47(6):2659–2665, 2001.

Qianxiao Li, Cheng Tai, and E Weinan. Stochastic modified equations and dynamics of stochastic gradient algorithms i: Mathematical foundations. *The Journal of Machine Learning Research*, 20 (1):1474–1520, 2019.

Zhiyuan Li, Yuping Luo, and Kaifeng Lyu. Towards resolving the implicit bias of gradient descent for matrix factorization: Greedy low-rank learning. In *International Conference on Learning Representations*, 2020.

Chaoyue Liu, Libin Zhu, and Mikhail Belkin. Loss landscapes and optimization in over-parameterized non-linear systems and neural networks. *Applied and Computational Harmonic Analysis*, 2022.

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, 2021.

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

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):E7665–E7671, 2018.

Hancheng Min, Salma Tarmoun, Rene Vidal, and Enrique Mallada. On the explicit role of initialization on the convergence and implicit bias of overparametrized linear networks. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 7760–7768. PMLR, 18–24 Jul 2021.

Behnam Neyshabur, Ryota Tomioka, and Nathan Srebro. In search of the real inductive bias: On the role of implicit regularization in deep learning. *arXiv preprint arXiv:1412.6614*, 2014.

Greg Ongie, Rebecca Willett, Daniel Soudry, and Nathan Srebro. A function space view of bounded norm infinite width ReLu nets: The multivariate case. *arXiv preprint arXiv:1910.01635*, 2019.

Rahul Parhi and Robert D Nowak. What kinds of functions do deep neural networks learn? Insights from variational spline theory. *SIAM Journal on Mathematics of Data Science*, 4(2):464–489, 2022.

Scott Pesme, Loucas Pillaud-Vivien, and Nicolas Flammarion. Implicit bias of sgd for diagonal linear networks: A provable benefit of stochasticity. *Advances in Neural Information Processing Systems*, 34, 2021.

Leonardo Petrini, Francesco Cagnetta, Eric Vanden-Eijnden, and Matthieu Wyart. Learning sparse features can lead to overfitting in neural networks. *arXiv preprint arXiv:2206.12314*, 2022.

Mary Phuong and Christoph H Lampert. The inductive bias of ReLU networks on orthogonally separable data. In *International Conference on Learning Representations*, 2020.

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. doi: https://doi.org/10.1002/cpa.22074.

Itay M Safran, Gilad Yehudai, and Ohad Shamir. The effects of mild over-parameterization on the optimization landscape of shallow ReLU neural networks. In *Proceedings of Thirty Fourth Conference on Learning Theory*, volume 134 of *Proceedings of Machine Learning Research*, pages 3889–3934. PMLR, 15–19 Aug 2021.

Pedro Savarese, Itay Evron, Daniel Soudry, and Nathan Srebro. How do infinite width bounded norm networks look in function space? In *Conference on Learning Theory*, pages 2667–2690. PMLR, 2019.

Alexander Shevchenko, Vyacheslav Kungurtsev, and Marco Mondelli. Mean-field analysis of piecewise linear solutions for wide ReLu networks. *arXiv preprint arXiv:2111.02278*, 2021.

Justin Sirignano and Konstantinos Spiliopoulos. Mean field analysis of neural networks: A law of large numbers. *SIAM Journal on Applied Mathematics*, 80(2):725–752, 2020.

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.

Gal Vardi and Ohad Shamir. Implicit regularization in ReLu networks with the square loss. In *Conference on Learning Theory*, pages 4224–4258. PMLR, 2021.

Yifei Wang and Mert Pilanci. The convex geometry of backpropagation: Neural network gradient flows converge to extreme points of the dual convex program. *arXiv preprint arXiv:2110.06488*, 2021.

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.

Blake Woodworth, Suriya Gunasekar, Jason D Lee, Edward Moroshko, Pedro Savarese, Itay Golan, Daniel Soudry, and Nathan Srebro. Kernel and rich regimes in overparametrized models. In *Conference on Learning Theory*, pages 3635–3673. PMLR, 2020.

Greg Yang and Edward J Hu. Tensor programs iv: Feature learning in infinite-width neural networks. In *International Conference on Machine Learning*, pages 11727–11737. PMLR, 2021.

Chulhee Yun, Shankar Krishnan, and Hossein Mobahi. A unifying view on implicit bias in training linear neural networks. In *International Conference on Learning Representations*, 2021.

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.

Mo Zhou, Rong Ge, and Chi Jin. A local convergence theory for mildly over-parameterized two-layer neural network. In *Conference on Learning Theory*, pages 4577–4632. PMLR, 2021.

Appendix

Table of Contents

A Additional experiments

A.1 Unidimensional data

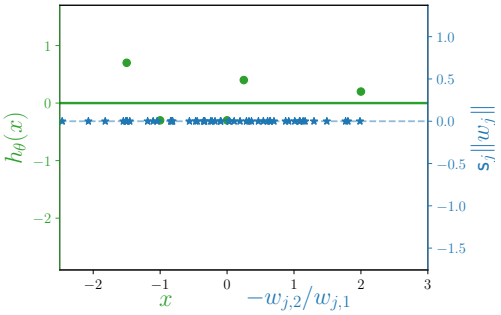
This section presents additional experiments in the general setting where data are non-orthogonal. To be able to visualize the results, similarly to Section 5, we consider unidimensional data with a bias term, but with 5 data points. Here again, we consider a neural network of width $m = 60$ and run gradient descent with step size 10^{-3}.

Similarly to Figure 2, Figure 4 presents the training dynamics with a small initialisation ($\lambda = \frac{10^{-4}}{\sqrt{m}}$). As in the orthogonal case, we first observe an early alignment phase in Figure 4b. Afterwards, two groups of neurons (against a single group in the orthogonal case) grow in norms until reaching an intermediate saddle point in Figure 4c. Note that during this norm-growth phase, the group of neurons does not remain aligned with a fixed direction, but changed in direction. A similar phenomenon happens when leaving the intermediate saddle point in Figure 4d: the norm of these neurons still grow but they also change their direction. This behavior is what makes the general case fundamentally harder to analyse than the orthogonal one, where such a behavior does not happen. We believe that controlling this type of behavior is the key towards dealing with the general set up.

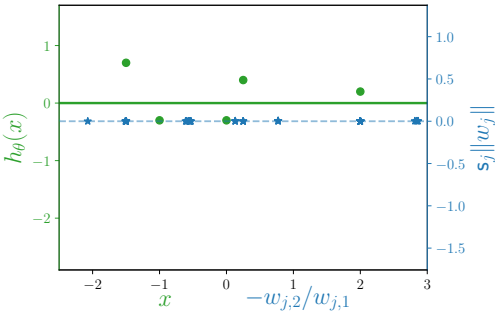
In Figure 4e, we have a zero training loss and as in the orthogonal case, the estimated function is sparse in its number of kinks. Figure 4f shows that a saddle to saddle dynamics also happens in this general setup.

Figure 5 on the other hand studies the impact of the initialisation scale. In particular, it shows the training dynamics for a large scale of initialisation ($\lambda = \frac{10}{\sqrt{m}}$). By comparing Figures 5a and 5b, we indeed observe the *lazy regime* [Chizat et al., 2019]: the neuron weights do not significantly move between the initialisation and the end of training.

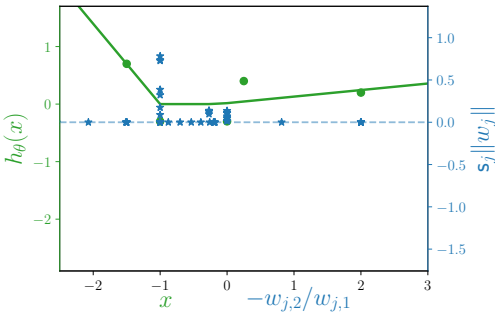
The final estimated function is not as simple as for small initialisation: it is approximately the interpolator coming from the Neural Tangent Kernel at initialisation, whose associated RKHS is described in Bietti and Mairal [2019]. The strong biased induced by initialisation and the poor generalising properties of this kernel interpolator illustrate the benefits of the *rich regime* obtained for small initialisation. However, this large initialisation has the advantage of converging towards



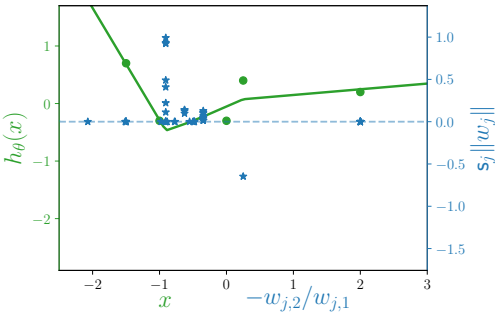
(a) Initialisation (Iteration 0)



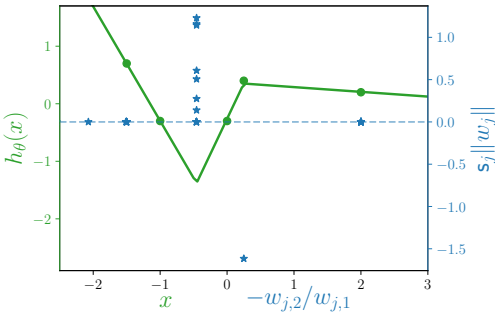
(b) End of early alignment (Iteration 15000)



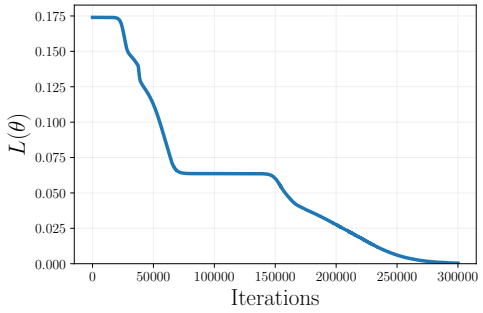
(c) Intermediate saddle (Iteration 94000)



(d) Simultaneous norm growth and realignment (Iteration 150000)

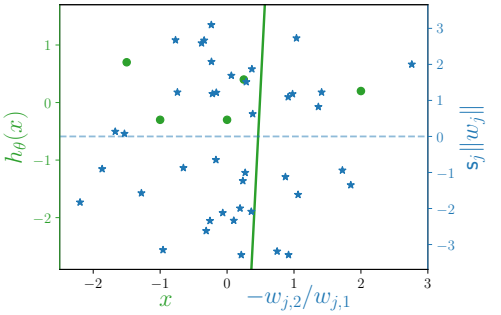


(e) Final state (Iteration 300000)

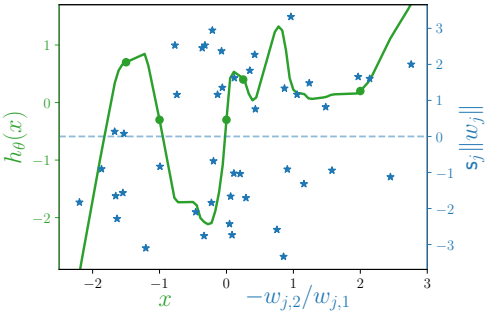


(f) Loss profile

Figure 4: State of training at different stages and loss profile.



(a) Initialisation (Iteration 0)

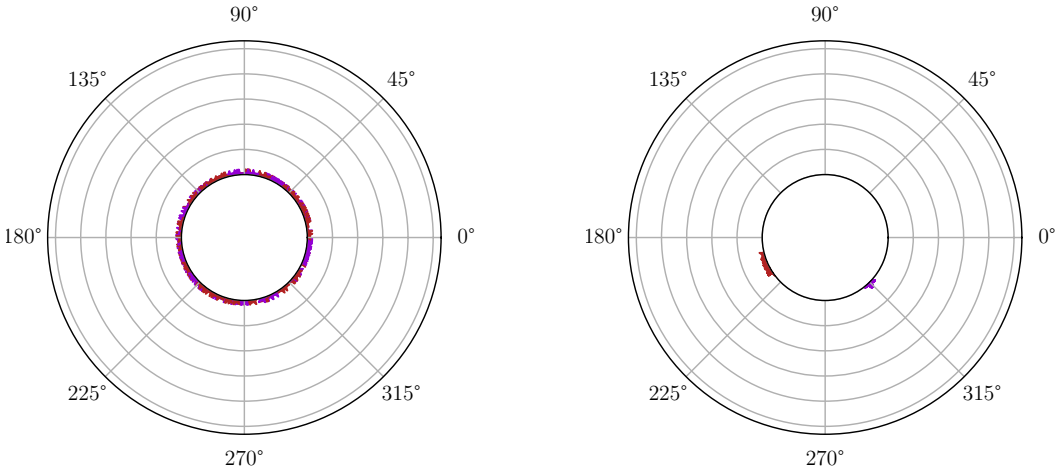


(b) End of training (Iteration 5000)

Figure 5: Training dynamics for large initialisation.

0 loss much faster: the trajectory indeed does not go through saddle points that might significantly slow down the learning process. Similar results are observed for large initialisations when either considering unbalanced initialisation or orthogonal data.

A.2 High dimensional data



(a) Initialisation (Iteration 0)

(b) End of alignment phase (Iteration 40000)

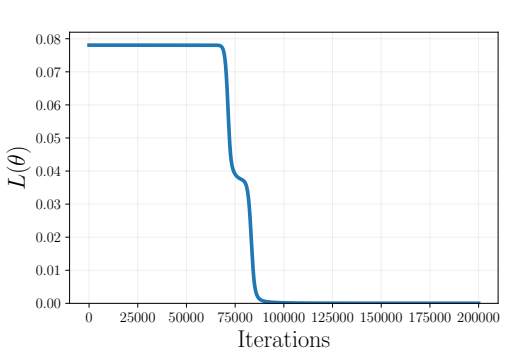
(c) Growth of first cluster (Iteration 80000)

(d) Final state (Iteration 200000)

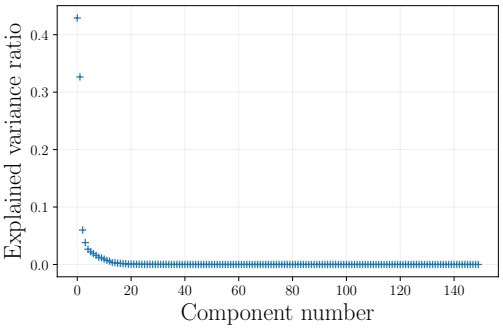
Figure 6: State of training at different stages. Each red (resp. purple) star represents a single neuron with $\mathsf{s}_j = -1$ (resp. $\mathsf{s}_j = 1$): it shows (in polar coordinates) the projection of the hidden layer weight onto the 2 dimensional space spanned by the two principal components of the hidden layer weights at the final state of training. The inner circle corresponds to 0 norm vectors, whose direction is given by the angle.

This section presents additional experiments for high dimensional, nearly orthogonal data. We generated $n = 75$ data points x_i drawn independently at random according to a standard Gaussian distribution of dimension $d = 150$. It is then known that such points are almost orthogonal with large probability. The labels y_i are then given by a 6-neurons teacher network, whose weights were drawn at random following a Gaussian distribution.

From there, we trained a neural network with width $m = 200$ (without bias terms), an initialisation scale $\lambda = \frac{10^{-20}}{\sqrt{md}}$ and a gradient step size of 10^{-3}. Figure 6 illustrates the training dynamics of the



(a) Loss profile



(b) Explained variance ratios of the principal components analysis of the final hidden weights

Figure 7: Additional information on the high dimensional experiment.

parameters when projected onto the 2 dimensional space of the two principal components of the 200×150 matrix associated to the hidden layer of the network.

The training loss profile is given by Figure 7a. Figure 7b finally shows the explained variance ratio of the principal components of the PCA used in Figure 6.

Behaviors close to the orthogonal case can easily be observed here. First, we can see in Figure 7a that the training loss converges to 0 and that the dynamics goes through an intermediate saddle point around iteration 75000. Also, thanks to Figure 6, we see the training dynamics (at least when projected onto the 2 dimensional space of the two principal components) follows similar phases. At first, we observe an early alignment phase where the neurons align towards two key directions. During a second phase, a first cluster of neurons grows in norm while keeping a fixed direction. During a third phase, the same happens for the other cluster of neurons. Only after these three phases, the neurons will slightly move from these two key directions and reach the final state of Figure 6d, for which we see that the neurons are not exactly aligned with two directions.

This last state is what differs from the exactly orthogonal case, where the final solution consists of a 2 neurons network. This is obviously not the case here, as can be seen from Figure 6d, but also from the explained variance ratios given in Figure 7b. Indeed, we can there see that 80% of the final state neurons are explained by the two principal components. This implies that the estimated interpolator is close to a 2 neurons network, but still far from being only represented by 2 direction (around 20% of its variance is explained by the remaining directions).

Note that we here chose a very small initialisation scale (of order 10^{-20}). As explained in Section 3.2, this confirms the exponential dependence of λ in the number of data points and is merely needed to observe a clear saddle point in the dynamics, but similar final states of training are observed for much larger values initialisation scales.

B Main proofs

In this section, we prove the main theorem. Appendices B.1 and B.2 provide additional notations and recall the assumption on the initialisation. Then, each subsection corresponds to the study of the different phases of the dynamics: Appendices B.3 to B.6 prove respectively the alignment phase, the positive, then negative label fitting phases and finally the convergence phase.

B.1 Additional notations

First, we introduce the following additional notations. We need to define the two dynamical vectors that encode respectively the fitting of positive and negative labels

$$D_+^t = -\frac{1}{n} \sum_{k|y_k>0} (h_{\theta^t}(x_k) - y_k) x_k \quad \text{and} \quad D_-^t = -\frac{1}{n} \sum_{k|y_k<0} (h_{\theta^t}(x_k) - y_k) x_k.$$

We also define the following open subsets of $\mathbb{R}^d$:

$$S_+ = \{w \in \mathbb{R}^d \mid \forall k, \mathbb{1}_{\langle w, x_k\rangle \geq 0} \geq \mathbb{1}_{y_k > 0}\} \quad \text{and} \quad S_- = \{w \in \mathbb{R}^d \mid \forall k, \mathbb{1}_{\langle w, x_k\rangle \geq 0} \geq \mathbb{1}_{y_k < 0}\}.$$

Note that $D_+ \in S_+$ and $-D_- \in S_-$ and that neurons defined in Equations (5) and (6) correspond to

$$S_{+,1} = \{j \in [\![m]\!] \mid w_j^0 \in S_+ \text{ and } \mathsf{s}_j = 1\} \quad \text{and} \quad S_{-,1} = \{j \in [\![m]\!] \mid w_j^0 \in S_- \text{ and } \mathsf{s}_j = -1\}.$$

B.2 Initialisation and assumptions on the dataset

We recall here, for the sake of completeness, the setup at initialisation. We initialise the dynamics on a balanced fashion: $a_j^0 = \mathsf{s}_j \|w_j^0\|$. We take $\lambda > 0$ and assume that $w_j^0 = \lambda g_j$, where each $g_j \sim \mathcal{N}(0, I_d)$ is an independent standard Gaussian. We assume that both $S_{+,1}$ and $S_{-,1}$ are non-empty and that for all k, $y_k \neq 0$. We also assume without loss of generality that $\|D_+\| > \|D_-\|$. This makes $r = \|D_+\|/\|D_-\| > 1$. We fix $\varepsilon > 0$ small enough so that

$$(1+3\varepsilon)\max\left(1-\alpha, \frac{\|D_-\|}{\|D_+\|}\right) \leq 1-\varepsilon \quad \text{and} \quad (1+3r\varepsilon)(1-\beta) \leq 1-\varepsilon$$

$$\text{where} \quad \alpha = \min_{k|y_k>0} \frac{y_k^2}{2\|D_+\|^2}, \quad \beta = \min_{k|y_k<0} \frac{y_k^2}{2\|D_-\|^2}.$$

We finally introduce an arbitrarily small $\lambda_* > 0$ that only depends on the training set $(x_k, y_k)_k$ and set $c = \max_{j\in[\![m]\!]} \|w_j^0\|/\lambda = \max_{j\in[\![m]\!]} \|g_j\|$.

Because of the non-differentiability of the ReLU activation, the gradient flow is not uniquely defined. In the orthogonal case, we have in particular the following ODE

$$\frac{\mathrm{d}\langle w_j^t, x_k\rangle}{\mathrm{d}t} = -\frac{h_{\theta^t}(x_k) - y_k}{n} \|w_j^t\| \mathbb{1}_{\langle w_j^t, x_k\rangle > 0}, \tag{11}$$

which leads to the following key lemma.

Lemma 6. *For any $t' \geq t$, $j \in [\![m]\!]$ and $k \in [n]$:*

$$\mathbb{1}_{\langle w_j^t, x_k\rangle < 0} \implies \mathbb{1}_{\langle w_j^{t'}, x_k\rangle < 0}.$$

Proof. This is a direct consequence of Equation (11). □

B.3 Phase 1: proof of Lemma 2

During the first phase, the neurons remain small in norm and they move tangentially to align with the vectors D_+ and D_-. The duration of this movement is typically sub-logarithmic in λ, the initialisation scale. As $\|D_+\| > \|D_-\|$, neurons in $S_{+,1}$ move slightly faster to align with D_+ than the ones in $S_{-,1}$ that align with $-D_-$. We begin by describing the dynamics of neurons in $S_{+,1}$ in Lemma 7 and derive similar results for the one of $S_{-,1}$ in Lemma 8. These two Lemmas constitute together Lemma 2 of the main text.

First, we define the following ending time of the phase $+1$:

$$t_{+1} = \frac{-\varepsilon \ln(\lambda)}{\|D_+\|}. \tag{12}$$

We have the following lemma.

Lemma 7 (Phase $+1$)**.** *If $\lambda \leq \lambda_*$, then we have the following inequalities,*

(i) $\forall j \in [\![m]\!], \|w_j^{t_{+1}}\| \leq 2c\lambda^{1-\varepsilon}$

(ii) $\forall j \in S_{+,1}, \langle \mathsf{w}_j^{t_{+1}}, D_+\rangle \geq (1-2\lambda^{\varepsilon})\|D_+\|$.

(iii) For all $j \in S_{+,1}$, let k such that $y_k < 0$, then $\langle w_j^{t_{+1}}, x_k\rangle = 0$.

The condition (i) of the lemma means that the neurons do not grow so much in this first phase, whereas (ii) states that they have aligned with vectors D_+. Finally (iii) shows that neurons in $S_{+,1}$ deactivate along negative labels during this first phase.

Proof. We divide the proof in three steps. First one is to control the growth of the neurons norm until t_{+1}. The second step shows that the tangential movement is faster, while the third one shows that neurons in $S_{+,1}$ "unalign" with the directions of negative labels.

First step: we show (i), i.e. that for $t \leq t_{+1}$, we have $\|w_j^t\| \leq 2c\lambda^{1-\varepsilon}$. Note first that, thanks to the balancedness, $a_j^t = \mathsf{s}_j \|w_j^t\|$. Let $\tau_\lambda := \inf\{t \geq 0 \mid \exists j \in [\![m]\!], \|w_j^t\| > 2c\lambda^{1-\varepsilon}\}$, then for all $t \leq \tau_\lambda$, we have $\|w_j^t\| \leq 2c\lambda^{1-\varepsilon}$ and hence,

$$|h_{\theta^t}(x_k)| \leq \sum_{j=1}^m \|w_j^t\| |\sigma(\langle w_j, x_k\rangle)| \leq \sum_{j=1}^m \|w_j^t\|^2 \leq 4mc^2\lambda^{2(1-\varepsilon)}.$$

Then, for λ_* such that $\lambda_*^{2(1-\varepsilon)} \leq \frac{\min_k |y_k|}{4mc^2}$, $y_k - h_{\theta^t}(x_k)$ and y_k have the same sign for any $t \leq \tau_\lambda$. As a consequence, for j such that $\mathsf{s}_j = 1$, Equation (3) yields

$$\frac{\mathrm{d}\|w_j^t\|}{\mathrm{d}t} = \frac{1}{n} \sum_{k|\langle x_k, w_j^t\rangle > 0} (y_k - h_{\theta^t}(x_k)) \langle x_k, w_j^t\rangle \leq \frac{1}{n} \sum_{k|\langle x_k, w_j^t\rangle > 0} (y_k - h_{\theta^t}(x_k)) \langle x_k, w_j^t\rangle \mathbb{1}_{y_k > 0}.$$

Now, denote $D_{+,j}^t := \dfrac{1}{n} \displaystyle\sum_{k|\langle x_k, w_j^t\rangle > 0} (y_k - h_{\theta^t}(x_k))\, x_k \mathbb{1}_{y_k > 0}$, we have

$$\begin{aligned}\langle D_{+,j}^t, w_j^t\rangle &\leq \|D_{+,j}^t\| \|w_j^t\| \leq \|D_+^t\| \|w_j^t\| \leq (\|D_+\| + \|\frac{1}{n} \sum_{k, y_k > 0} h_{\theta^t}(x_k) x_k\|) \|w_j^t\| \\ &\leq (\|D_+\| + 4mc^2\lambda^{2(1-\varepsilon)}) \|w_j^t\|.\end{aligned}$$

The same series of inequalities holds in the case $\mathsf{s}_j = -1$ for $\|D_-^t\|$. Overall,

$$\begin{aligned}\frac{\mathrm{d}\|w_j^t\|}{\mathrm{d}t} &\leq \left(\max\{\|D_+\|, \|D_-\|\} + 4mc^2\lambda^{2(1-\varepsilon)}\right) \|w_j^t\| \\ &\leq \left(\|D_+\| + 4mc^2\lambda^{2(1-\varepsilon)}\right) \|w_j^t\|.\end{aligned}$$

By Grönwall's lemma, this gives for any $t \leq \tau_\lambda$, $\|w_j^t\| \leq \|w_j^0\| e^{\left(\|D_+\| + 4mc^2\lambda^{2(1-\varepsilon)}\right)t}$. This shows that for $t \leq \min(\tau_\lambda, t_{+1})$,

$$\|w_j^t\| \leq c\lambda^{1-\varepsilon} e^{\frac{-4\varepsilon mc^2\lambda^{2(1-\varepsilon)}}{\|D_+\|} \ln(\lambda)} < 2c\lambda^{1-\varepsilon}, \tag{13}$$

where λ^* has been taken small enough. Hence $\tau_\lambda \geq t_{+1}$.

Second step: we show condition (ii). Indeed, let us now choose any $j \in S_{+,1}$. We have the following decomposition:

$$D_j^t = D_+ - \frac{1}{n} \sum_{k|\langle x_k, w_j^t\rangle > 0} h_{\theta^t}(x_k) x_k \mathbb{1}_{y_k > 0} + D_{-,j}^t,$$

so that as all vectors in the $D_{-,j}^t$ are different from the one of D_+, by orthogonality we have,

$$\langle D_j^t, D_+\rangle = \|D_+\|^2 - \frac{1}{n} \sum_{k|\langle x_k, w_j^t\rangle > 0} h_{\theta^t}(x_k) \langle x_k, D_+\rangle \mathbb{1}_{y_k > 0}.$$

Now, let us analyse the tangential movement for these j's: for all $t \leq t_{+1}$, Equation (9) leads to the following growth comparison

$$\begin{aligned}\frac{\mathrm{d}\langle \mathsf{w}_j^t, D_+\rangle}{\mathrm{d}t} &= \langle D_j^t, D_+\rangle - \langle D_j^t, \mathsf{w}_j^t\rangle \langle \mathsf{w}_j^t, D_+\rangle \\ &= \|D_+\|^2 - \langle \mathsf{w}_j^t, D_+\rangle^2 - \frac{1}{n} \sum_{k|\langle x_k, w_j^t\rangle > 0} h_{\theta^t}(x_k) \langle x_k, D_+ - \langle D_+, \mathsf{w}_j^t\rangle \mathsf{w}_j^t\rangle \mathbb{1}_{y_k > 0} - \langle D_{j,-}^t, \mathsf{w}_j^t\rangle \langle \mathsf{w}_j^t, D_+\rangle,\end{aligned}$$

and as we have, $-\langle D^t_{j,-}, \mathsf{w}^t_j\rangle\langle \mathsf{w}^t_j, D_+\rangle \geq 0$ and the third term lower bounded by $-4mc^2\|D_+\|\lambda^{2(1-\varepsilon)}$, it yields

$$\frac{\mathrm{d}\langle \mathsf{w}^t_j, D_+\rangle}{\mathrm{d}t} \geq \|D_+\|^2 - 4mc^2\|D_+\|\lambda^{2(1-\varepsilon)} - \langle \mathsf{w}^t_j, D_+\rangle^2. \tag{14}$$

Solutions of the ODE $f'(t) = a^2 - f^2(t)$ with value in $(-a, a)$ are of the form $f(t) = a\tanh(a(t + t_0))$ for some t_0. Note in the remaining of the proof $a^2 = \|D_+\|^2 - 4mc^2\|D_+\|\lambda^{2(1-\varepsilon)}$. In our case, define t_0 such that $\langle \mathsf{w}^0_j, D_+\rangle = a\tanh(at_0)$, then, by Grönwall comparison, it yields:

$$\langle \mathsf{w}^t_j, D_+\rangle \geq a\tanh(a(t + t_0)).$$

Note that $\langle \mathsf{w}^0_j, D_+\rangle \geq 0$, so that $t_0 \geq 0$. As a consequence, we simply have

$$\langle \mathsf{w}^t_j, D_+\rangle \geq a\tanh(at).$$

Now, using the inequality $\tanh(x) \geq 1 - 2e^{-2x}$, we have

$$\langle \mathsf{w}^{t+1}_j, D_+\rangle \geq a(1 - 2e^{2a\frac{\varepsilon}{\|D_+\|}\ln(\lambda)}).$$

We have the following inequalities on a when choosing λ_* small enough:

$$a = \sqrt{\|D_+\|^2 - 4mc^2\|D_+\|\lambda^{2(1-\varepsilon)}} \geq \frac{3\|D_+\|}{4}.$$

This leads for any $j \in S_{+,1}$ and λ^* small enough to

$$\begin{aligned}\langle \mathsf{w}^{t_{+1}}_j, D_+\rangle &> (\|D_+\| - 2c\sqrt{m\|D_+\|}\lambda^{1-\varepsilon})\left(1 - 2\lambda^{\frac{3\varepsilon}{2}}\right)\\ &> \|D_+\|\left(1 - 2\lambda^{\frac{3\varepsilon}{2}}\right) - 2c\sqrt{m\|D_+\|}\lambda^{1-\varepsilon}\\ &> \|D_+\|\left(1 - 2\lambda^{\varepsilon}\right),\end{aligned}$$

which proves condition (ii).

Third step: we show (iii). Consider $j \in S_{+,1}$ and k such that $y_k < 0$. If $\langle w^0_j, x_k\rangle < 0$, then by Lemma 6, there is nothing to prove. Otherwise, as for $t \leq t_{+1}$ we have $\|w^t_j\| \leq 2c\lambda^{1-\varepsilon}$, it yields (as long as $\langle \mathsf{w}^t_j, x_k\rangle > 0$)

$$\begin{aligned}\frac{\mathrm{d}}{\mathrm{d}t}\langle \mathsf{w}^t_j, x_k\rangle &= \langle D^t_j, x_k\rangle - \langle D^t_j, \mathsf{w}^t_j\rangle\langle \mathsf{w}^t_j, x_k\rangle\\ &= \frac{1}{n}(y_k - h_\theta(x_k)) - \frac{1}{n}\langle \mathsf{w}^t_j, x_k\rangle \sum_{l|\langle \mathsf{w}^t_j, x_l\rangle > 0}(y_l - h_\theta(x_l))\langle \mathsf{w}^t_j, x_l\rangle < 0.\end{aligned}$$

Hence, if for some time $\tau \leq t_{+1}$, we have $\langle \mathsf{w}^\tau_j, x_k\rangle = 0$, then for all times $t \in [\tau, t_{+1}]$, this quantity remains 0. We now continue to upperbound the derivative of $\langle \mathsf{w}^\tau_j, x_k\rangle$:

$$\begin{aligned}\frac{\mathrm{d}}{\mathrm{d}t}\langle \mathsf{w}^t_j, x_k\rangle &\leq \frac{1}{n}(y_k + 4mc^2\lambda^{2(1-\varepsilon)}) - \frac{1}{n}\langle \mathsf{w}^t_j, x_k\rangle \sum_{l|\langle \mathsf{w}^t_j, x_l\rangle > 0}(y_l - 4mc^2\lambda^{2(1-\varepsilon)})\langle \mathsf{w}^t_j, x_l\rangle\\ &\leq \frac{y_k}{n} - \frac{1}{n}\langle \mathsf{w}^t_j, x_k\rangle \sum_{l|\langle \mathsf{w}^t_j, x_l\rangle > 0} y_l\langle \mathsf{w}^t_j, x_l\rangle \mathbb{1}_{y_l<0} + 5mc^2\lambda^{2(1-\varepsilon)}.\end{aligned}$$

From this, we sum all the $y_k < 0$, and noting $f(t) := -\frac{1}{n}\sum_{k|\langle \mathsf{w}^t_j, x_k\rangle > 0} y_k\langle \mathsf{w}^t_j, x_k\rangle \mathbb{1}_{y_k<0}$, we have

$$\frac{\mathrm{d}}{\mathrm{d}t}f(t) \leq -\frac{1}{n^2}\sum_{k|y_k<0} y_k^2 + f(t)^2 - 5mc^2\lambda^{2(1-\varepsilon)}\frac{1}{n}\sum_{k|y_k<0} y_k.$$

If we let $a^2 := \frac{1}{n^2}\sum_{k|y_k<0} y_k^2 + 5mc^2\lambda^{2(1-\varepsilon)}\frac{1}{n}\sum_{k|y_k<0} y_k > 0$, we have $f'(t) \leq -a^2 + f(t)^2$, and as by Cauchy-Schwarz

$$\begin{aligned}
f(0)^2 &\leq \frac{1}{n^2}\sum_{k|\langle \mathsf{w}_j^0, x_k\rangle>0} y_k^2 \mathbb{1}_{y_k<0} \sum_{k|\langle \mathsf{w}_j^0, x_k\rangle>0} \langle \mathsf{w}_j^0, x_k\rangle^2 \mathbb{1}_{y_k<0} \\
&= \frac{1}{n^2}\left(\sum_{k|\langle \mathsf{w}_j^0, x_k\rangle>0} y_k^2 \mathbb{1}_{y_k<0}\right)\left(\|\mathsf{w}_j^0\|^2 - \sum_k \langle \mathsf{w}_j^0, x_k\rangle^2 \left(\mathbb{1}_{y_k>0}\mathbb{1}_{\langle \mathsf{w}_j^0, x_k\rangle>0} + \mathbb{1}_{\langle \mathsf{w}_j^0, x_k\rangle<0}\right)\right) \\
&< \left(a^2 + \lambda^{1-2\varepsilon}\right)\left(1 - \sum_k \langle \mathsf{w}_j^0, x_k\rangle^2 \left(\mathbb{1}_{y_k>0}\mathbb{1}_{\langle \mathsf{w}_j^0, x_k\rangle>0} + \mathbb{1}_{\langle \mathsf{w}_j^0, x_k\rangle<0}\right)\right) \\
&< a^2,
\end{aligned}$$

where the two last inequalities are valid for λ small enough. Hence $t \mapsto f(t)$ is decreasing and $f'(t) \leq -a^2 + f(0)^2$, that is if we call $b := a^2 - f(0)^2 > 0$, we have $f(t) \leq -bt + f(0)$, and for $\tau = f(0)/b$, we have $f(\tau) = 0$. And as $\tau < t_{+1}$, we have $f(t_{+1}) = 0$. This concludes the proof of condition (iii) and of the lemma. □

Now, we show that the exact same conclusion is also valid for the neurons in $S_{-,1}$ i.e., after a sub-logarithmic time, they eventually align with $-D_-$ and deactivates with respect to positive outputs. We define here a ending time similar to (12):

$$t_{-1} = \frac{-\varepsilon r \ln(\lambda)}{\|D_+\|}. \tag{15}$$

We prove the following lemma analogous to Lemma 7.

Lemma 8 (Phase -1)**.** *If $\lambda \leq \lambda_*$, then we have the following inequalities,*

(i) $\forall j \in [\![m]\!], \|w_j^{t_{-1}}\| \leq 2c\lambda^{1-r\varepsilon}$

(ii) $\forall j \in S_{-,1}, \langle \mathsf{w}_j^{t_{-1}}, -D_-\rangle \geq (1-2\lambda^\varepsilon)\|D_-\|$.

(iii) For all $j \in S_{-,1}$, let k such that $y_k > 0$, then $\langle w_j^{t_{-1}}, x_k\rangle = 0$.

Proof. The proof is essentially the same than the proof of Lemma 7. We will be short and underline solely the main differences.

First step: we show condition (i), i.e. that for $t \leq t_{-1}$, we have $\|w_j^t\| \leq 2c\lambda^{1-r\varepsilon}$. Indeed, let $\tau_\lambda^r := \inf\{t \geq 0 \mid \exists j \in [\![m]\!], \|w_j^t\| > 2c\lambda^{1-r\varepsilon}\}$. For all $t \leq \tau_\lambda^r$, we have $|h_{\theta^t}(x_k)| \leq 4mc^2\lambda^{2(1-r\varepsilon)}$ and similarly to the proof of Lemma 7, we have that for $t \leq \tau_\lambda^r$, $\|w_j^t\| \leq \|w_j^0\| e^{\left(\|D_+\| + 4mc^2\lambda^{2(1-r\varepsilon)}\right)t}$. This shows that for $t \leq \min(\tau_\lambda^r,\, t_{-1})$,

$$\|w_j^t\| \leq c\lambda^{1-r\varepsilon} e^{\frac{-4\varepsilon mc^2\lambda^{2(1-r\varepsilon)}}{\|D_+\|}\ln(\lambda)} < 2c\lambda^{1-r\varepsilon}, \tag{16}$$

where λ^* has been taken small enough. Hence $\tau_\lambda^r \geq t_{-1}$.

Second step: we show (ii), i.e. that the neurons almost align after time t_{-1}. Indeed, for $j \in S_{-,1}$, similarly to Lemma 7, we have that for all $t \leq t_{-1}$:

$$\frac{\mathrm{d}\langle \mathsf{w}_j^t, -D_-\rangle}{\mathrm{d}t} \geq \|D_-\|^2 - 4mc^2\|D_-\|\lambda^{2(1-r\varepsilon)} - \langle \mathsf{w}_j^t, D_-\rangle^2.$$

Denoting $a_r^2 = \|D_-\|^2 - 4mc^2\|D_-\|\lambda^{2(1-r\varepsilon)}$, we have by Grönwall comparison

$$\langle \mathsf{w}_j^t, -D_-\rangle \geq a_r \tanh(a_r t).$$

Now, this gives $\langle \mathsf{w}_j^{t_{-1}}, -D_-\rangle > a_r(1 - 2e^{2a_r\frac{r\varepsilon}{\|D_+\|}\ln(\lambda)})$ and lower bounding a_r as before, we have

$$\langle \mathsf{w}_j^{t_{-1}}, -D_-\rangle > \|D_-\|\left(1 - 2\lambda^\varepsilon\right),$$

and this shows the condition (ii).

Third step: we show (iii). Consider $j \in S_{-,1}$ and k such that $y_k > 0$. If $\langle w_j^0, x_k\rangle < 0$, then by Lemma 6, there is nothing to prove. Otherwise, as for $t \le t_{-1}$ we have $\|w_j^t\| \le 2c\lambda^{1-r\varepsilon}$, it yields (as long as $\langle \mathsf{w}_j^t, x_k\rangle > 0$)

$$\begin{aligned}\frac{\mathrm{d}}{\mathrm{d}t}\langle \mathsf{w}_j^t, x_k\rangle &= -\langle D_j^t, x_k\rangle + \langle D_j^t, \mathsf{w}_j^t\rangle\langle \mathsf{w}_j^t, x_k\rangle \\ &= \frac{1}{n}(h_\theta(x_k) - y_k) - \frac{1}{n}\langle \mathsf{w}_j^t, x_k\rangle \sum_{l|\langle \mathsf{w}_j^t, x_l\rangle > 0} (h_\theta(x_l) - y_l)\langle \mathsf{w}_j^t, x_l\rangle < 0.\end{aligned}$$

Hence, if for some time $\tau \le t_{-1}$, we have $\langle \mathsf{w}_j^\tau, x_k\rangle = 0$, then for all times $t \in [\tau, t_{-1}]$, this quantity will remain 0. We now we continue to upperbound the derivative of $\langle \mathsf{w}_j^\tau, x_k\rangle$:

$$\begin{aligned}\frac{\mathrm{d}}{\mathrm{d}t}\langle \mathsf{w}_j^t, x_k\rangle &\le \frac{1}{n}(4mc^2\lambda^{2(1-r\varepsilon)} - y_k) - \frac{1}{n}\langle \mathsf{w}_j^t, x_k\rangle \sum_{l|\langle \mathsf{w}_j^t, x_l\rangle > 0} (-4mc^2\lambda^{2(1-r\varepsilon)} - y_l)\langle \mathsf{w}_j^t, x_l\rangle \\ &\le -\frac{y_k}{n} + \frac{1}{n}\langle \mathsf{w}_j^t, x_k\rangle \sum_{l|\langle \mathsf{w}_j^t, x_l\rangle > 0} y_l\langle \mathsf{w}_j^t, x_l\rangle \mathbb{1}_{y_l>0} + 5mc^2\lambda^{2(1-r\varepsilon)}.\end{aligned}$$

From this, we sum all the $y_k > 0$, and noting $f(t) := \frac{1}{n}\sum_{k|\langle \mathsf{w}_j^t, x_k\rangle > 0} y_k\langle \mathsf{w}_j^t, x_k\rangle \mathbb{1}_{y_k>0}$, we have

$$\frac{\mathrm{d}}{\mathrm{d}t} f(t) \le -\frac{1}{n^2}\sum_{k|y_k>0} y_k^2 + f(t)^2 + 5mc^2\lambda^{2(1-r\varepsilon)}\frac{1}{n}\sum_{k|y_k>0} y_k.$$

If we let $a^2 := \frac{1}{n^2}\sum_{k|y_k>0} y_k^2 - 5mc^2\lambda^{2(1-r\varepsilon)}\frac{1}{n}\sum_{k|y_k>0} y_k > 0$, we have $f'(t) \le -a^2 + f(t)^2$, and as by Cauchy-Schwarz

$$\begin{aligned}f(0)^2 &\le \frac{1}{n^2}\sum_{k|\langle \mathsf{w}_j^0, x_k\rangle > 0} y_k^2\mathbb{1}_{y_k>0} \sum_{k|\langle \mathsf{w}_j^0, x_k\rangle > 0} \langle \mathsf{w}_j^0, x_k\rangle^2\mathbb{1}_{y_k>0} \\ &< \frac{1}{n^2}\left(\sum_{k|\langle \mathsf{w}_j^0, x_k\rangle > 0} y_k^2\mathbb{1}_{y_k>0}\right)\left(\|\mathsf{w}_j^0\|^2 - \sum_k \langle \mathsf{w}_j^0, x_k\rangle^2\left(\mathbb{1}_{y_k<0}\mathbb{1}_{\langle \mathsf{w}_j^0, x_k\rangle>0} + \mathbb{1}_{\langle \mathsf{w}_j^0, x_k\rangle<0}\right)\right) \\ &< \left(a^2 + \lambda^{1-2\varepsilon}\right)\left(1 - \sum_k \langle \mathsf{w}_j^0, x_k\rangle^2\left(\mathbb{1}_{y_k<0}\mathbb{1}_{\langle \mathsf{w}_j^0, x_k\rangle>0} + \mathbb{1}_{\langle \mathsf{w}_j^0, x_k\rangle<0}\right)\right) \\ &< a^2,\end{aligned}$$

where the two last inequalities are valid for λ small enough. Hence $t \mapsto f(t)$ is decreasing and $f'(t) \le -a^2 + f(0)^2$, that is if we call $b := a^2 - f(0)^2 > 0$, we have $f(t) \le -bt + f(0)$, and for $\tau = f(0)/b$, we have $f(\tau) = 0$. And as $\tau < t_{-1}$, we have $f(t_{-1}) = 0$. This concludes the proof of condition (iii) and of the Lemma. □

We end this subsection by a lemma that shows that until $t_{\pm 1}$, the neurons w_j^t with $j \in S_{\pm 1}$, could not have collapsed to 0.

Lemma 9. *Define $\underline{c} = \min \|w_j^0\|/\lambda$, then, if $\lambda \le \lambda^*$,*

(i) for all $t \le t_{+1}$, $\forall j \in S_{+,1}$, $\|w_j^t\| > \underline{c}\lambda^{1+\varepsilon}/2$,

(ii) for all $t \le t_{-1}$, $\forall j \in S_{-,1}$, $\|w_j^t\| > \underline{c}\lambda^{1+r\varepsilon}/2$.

Proof. Let us begin with the first point. As stated in the proof of Lemma 7, $y_k - h_{\theta^t}(x_k)$ and y_k have the same sign for any $t \le t_{+1}$. As a consequence, for $j \in S_{+,1}$, it yields

$$\begin{aligned}\frac{\mathrm{d}\|w_j^t\|}{\mathrm{d}t} &\ge \langle D_{-,j}^t, w_j^t\rangle \ge -\|D_{-,j}^t\|\|w_j^t\| \ge -\left(\|D_-\| + \|\frac{1}{n}\sum_{k,y_k<0} h_{\theta^t}(x_k)x_k\|\right)\|w_j^t\| \\ &\ge -\left(\|D_+\| + 4mc^2\lambda^{2(1-\varepsilon)}\right)\|w_j^t\|.\end{aligned}$$

Then, by Grönwall's lemma, this gives for any $t \leq t_{+1}$, $\|w_j^t\| \geq \|w_j^0\| e^{-\left(\|D_+\| + 4mc^2\lambda^{2(1-\varepsilon)}\right)t}$. This shows that, for $t \leq t_{+1}$,

$$\|w_j^t\| \geq \underline{c}\lambda^{1+\varepsilon} e^{\frac{4\varepsilon m c^2 \lambda^{2(1-\varepsilon)}}{\|D_+\|}\ln(\lambda)} > \underline{c}\lambda^{1+\varepsilon}/2.$$

This concludes the first point of the lemma. The second point is very similar to the first one. Indeed, in this case also, for $t \leq t_{-1}$, y_k rules the sign of the residual so that for $j \in S_{-,1}$,

$$\begin{aligned}\frac{\mathrm{d}\|w_j^t\|}{\mathrm{d}t} &\geq \langle D_{+,j}^t, w_j^t\rangle \geq -\|D_{+,j}^t\|\|w_j^t\| \geq -\left(\|D_+\| + \|\frac{1}{n}\sum_{k,y_k>0} h_{\theta^t}(x_k)x_k\|\right)\|w_j^t\| \\ &\geq -\left(\|D_+\| + 4mc^2\lambda^{2(1-r\varepsilon)}\right)\|w_j^t\|.\end{aligned}$$

Then, by Grönwall's lemma, this gives for any $t \leq t_{-1}$, $\|w_j^t\| \geq \|w_j^0\| e^{-\left(\|D_+\| + 4mc^2\lambda^{2(1-r\varepsilon)}\right)t}$. This shows that, as $t \leq t_{-1}$,

$$\|w_j^t\| \geq \underline{c}\lambda^{1+r\varepsilon} e^{\frac{4\varepsilon r m c^2 \lambda^{2(1-r\varepsilon)}}{\|D_+\|}\ln(\lambda)} > \underline{c}\lambda^{1+r\varepsilon}/2.$$

This concludes the proof. □

B.4 Phase 2: proof of Lemma 3

During the second phase, the norm of the neurons in $S_{+,1}$ (which are aligned with D_+) grows until perfectly fitting all the positive labels of the training points. Meanwhile, all the other neurons do not move significantly. We define the ending time of the second phase as

$$\begin{aligned}t_2 = \inf\Bigg\{ t \geq t_{+1} \mid \text{either } \exists j \notin S_{+,1}, \|w_j^t\| \geq 2c\lambda^\varepsilon \\ \text{or } \sum_{j\in S_{+,1}} \|w_j^t\|^2 \geq n\|D_+\| - \lambda^{\frac{\varepsilon}{5}} \\ \text{or } \exists j \in S_{+,1}, \langle \mathsf{w}_j^t, D_+\rangle \leq \|D_+\| - \lambda^{\frac{\varepsilon}{2}} \Bigg\}.\end{aligned} \tag{17}$$

The following lemma corresponds to Lemma 3 of the main text.

Lemma 10. *If $\lambda \leq \lambda_*$, then we have the following inequalities*

(i) $t_2 \leq -\frac{1+3\varepsilon}{\|D_+\|}\ln(\lambda)$,

(ii) $\forall j \in [\![m]\!] \setminus S_{+,1}, \|w_j^{t_2}\| < 2c\lambda^\varepsilon$,

(iii) $\forall j \in S_{+,1}, \langle \mathsf{w}_j^{t_2}, D_+\rangle > \|D_+\| - \lambda^{\frac{\varepsilon}{2}}$.

The first point of the lemma states that the second phase lasts a time of order $-\ln(\lambda)$. The other two points state that the first and third conditions in the definition of t_2 do not hold at the end of the phase. Thus, the second condition in Equation (17) holds at t_2, meaning that the norm of the neurons in $S_{+,1}$ have grown enough to fit the positive labels:

$$\sum_{j\in S_{+,1}} \|w_j^{t_2}\|^2 = n\|D_+\| - \lambda^{\frac{\varepsilon}{5}}$$

A direct consequence of this[2] is that at the end of the second phase, the training loss on the positive labels is of order $\lambda^{\frac{2\varepsilon}{5}}$ (at least).

Proof. Preliminaries: define for this proof $h(t) = \|D_+\| - \min_{j\in S_{+,1}}\langle \mathsf{w}_j^t, D_+\rangle$. By definition of the second phase, we have $h(t) \leq \lambda^{\frac{\varepsilon}{2}}$ for any $t \in [t_{+1}, t_2]$.

[2]This comes from the decomposition given by Equation (19) in the proof.

We first show that during the whole second phase, D_+^t is almost colinear with D_+. For any $j \in S_{+,1}$ and $t \in [t_{+1}, t_2]$, define $d_j^t = \mathsf{w}_j^t - \frac{D_+}{\|D_+\|}$. Since w_j^t and $\frac{D_+}{\|D_+\|}$ are both of norm 1, we have $\|d_j^t\|^2 = -2\langle d_j^t, \frac{D_+}{\|D_+\|}\rangle$. By definition of $h(t)$,

$$0 \geq \langle d_j^t, D_+\rangle \geq -h(t).$$

So finally, we have for any $j \in S_{+,1}$ and $t \in [t_{+1}, t_2]$ the decomposition

$$\mathsf{w}_j^t = \frac{D_+}{\|D_+\|} + d_j^t \qquad \text{with} \qquad \|d_j^t\|^2 \leq 2\frac{h(t)}{\|D_+\|}. \tag{18}$$

Now let any k such that $y_k > 0$. Using Equation (18) and the fact that $\|w_j^t\| \leq 2c\lambda^\varepsilon$ for any $j \notin S_{+,1}$, we have for any $t \in [t_{+1}, t_2]$:

$$\begin{aligned}
h_{\theta^t}(x_k) &= \sum_{j\in S_{+,1}} \|w_j^t\|\langle w_j^t, x_k\rangle + \sum_{j\notin S_{+,1}} \mathsf{s}_j\|w_j^t\|\langle w_j^t, x_k\rangle \\
&= \sum_{j\in S_{+,1}} \frac{\|w_j^t\|^2}{\|D_+\|}\langle D_+, x_k\rangle + \underbrace{\sum_{j\in S_{+,1}} \|w_j^t\|^2\langle d_j^t, x_k\rangle + \sum_{j\notin S_{+,1}} \mathsf{s}_j\|w_j^t\|\langle w_j^t, x_k\rangle}_{} \\
&= \sum_{j\in S_{+,1}} \frac{\|w_j^t\|^2}{\|D_+\|}\langle D_+, x_k\rangle + \qquad\qquad\qquad\qquad h_k(t),
\end{aligned}$$

where $|h_k(t)| \leq \sum_{j\in S_{+,1}} \|w_j^t\|^2|\langle d_j^t, x_k\rangle| + 4mc^2\lambda^{2\varepsilon}$. It follows that

$$\begin{aligned}
D_+^t &= -\frac{1}{n}\sum_{k, y_k>0} h_{\theta^t}(x_k)x_k + D_+ \\
&= (1 - \frac{\sum_{j\in S_{+,1}} \|w_j^t\|^2}{n\|D_+\|})D_+ - \frac{1}{n}\sum_{k,y_k>0} h_k(t)x_k.
\end{aligned}$$

Roughly, this means that as long as $1 - \frac{\sum_{j\in S_{+,1}} \|w_j^t\|^2}{n\|D_+\|}$ is large enough, D_+^t is almost colinear with D_+. Precisely, we have for any $t \in [t_{+1}, t_2]$ the following decomposition

$$D_+^t = \left(1 - \frac{\sum_{j\in S_{+,1}} \|w_j^t\|^2}{n\|D_+\|} + h_+(t)\right) D_+ + h_\perp(t), \tag{19}$$

where $\langle h_\perp(t), D_+\rangle = 0$ and $|h_+(t)|\|D_+\| \vee \|h_\perp(t)\| \leq \frac{1}{n}\sum_{j\in S_{+,1}} \|w_j^t\|^2\sqrt{2\frac{h(t)}{\|D_+\|}} + 4mc^2\lambda^{2\varepsilon}$.

First point: denote $u_+(t) = \sum_{j\in S_{+,1}} \|w_j^t\|^2$. We then have by balancedness and Equation (3)

$$\frac{1}{2}\frac{\mathrm{d}u_+(t)}{\mathrm{d}t} = \sum_{j\in S_{+,1}} \|w_j^t\|\langle D_j^{\theta^t}, w_j^t\rangle.$$

Note that $\langle D_+^t, x_k\rangle \geq 0$ for any $t \in [t_{+1}, t_2]$. We thus have $\frac{\mathrm{d}}{\mathrm{d}t}\langle w_j^t, x_k\rangle \leq 0$ for $j \in S_{+,1}$ and k such that $y_k < 0$ during this phase. Thanks to the last point of Lemma 7, $\langle D_j^{\theta^t}, w_j^t\rangle = \langle D_+^t, w_j^t\rangle$ for any $j \in S_{+,1}$ and $t \in [t_{+1}, t_2]$. Using Equations (18) and (19), this yields

$$\begin{aligned}
\frac{1}{2}\frac{\mathrm{d}u_+(t)}{\mathrm{d}t} &= \sum_{j\in S_{+,1}} \|w_j^t\|^2\langle D_+^t, \mathsf{w}_j^t\rangle \\
&\geq \sum_{j\in S_{+,1}} \|w_j^t\|^2\langle\left(1 - \frac{u_+(t)}{n\|D_+\|} + h_+(t)\right) D_+ + h_\perp(t), \mathsf{w}_j^t\rangle \\
&\geq \sum_{j\in S_{+,1}} \|w_j^t\|^2 \left(1 - \frac{u_+(t)}{n\|D_+\|} + h_+(t)\right)\left(\|D_+\| - \lambda^{\frac{\varepsilon}{2}}\right) - \sum_{j\in S_{+,1}} \|w_j^t\|^2\|h_\perp(t)\|\|d_j^t\|.
\end{aligned}$$

So we have the following growth comparison

$$\frac{u_+'(t)}{2} \geq \left(\|D_+\| - \lambda^{\frac{\varepsilon}{2}}\right)\left(1 - \frac{u_+(t)}{n\|D_+\|} + h_+(t)\right)u_+(t) - \|h_\perp(t)\| \max_{j\in S_{+,1}} \|d_j^t\| u_+(t).$$

By definition of the second phase, $u_+(t) \leq n\|D_+\|$ for any $t \in [t_{+1}, t_2]$. We chose λ small enough, so that $\sqrt{2\|D_+\|}\lambda^{\frac{\varepsilon}{4}} \geq 4mc^2\lambda^{2\varepsilon}$. This implies that $\|h_\perp(t)\| \vee |h_+(t)|\|D_+\| \leq 2\sqrt{2\|D_+\|}\lambda^{\frac{\varepsilon}{4}}$. So we finally have the following growth comparison during the second phase

$$u_+'(t) \geq 2\left(\|D_+\| - \lambda^{\frac{\varepsilon}{2}}\right)\left(1 - \left(2\sqrt{\frac{2}{\|D_+\|}}\right)\lambda^{\frac{\varepsilon}{4}} - \frac{u_+(t)}{n\|D_+\|}\right)u_+(t) - 8\lambda^{\frac{\varepsilon}{2}}u_+(t)$$

$$u_+'(t) \geq 2\left(\|D_+\| - \lambda^{\frac{\varepsilon}{2}}\right)\left(1 - \left(2\sqrt{\frac{2}{\|D_+\|}}\right)\lambda^{\frac{\varepsilon}{4}} - \frac{4}{\|D_+\| - \lambda^{\frac{\varepsilon}{2}}}\lambda^{\frac{\varepsilon}{2}} - \frac{u_+(t)}{n\|D_+\|}\right)u_+(t). \quad (20)$$

Solution of the ODE $f'(t) = af(t) - bf(t)^2$ with $f(0) \in (0, \frac{a}{b})$ are of the form $f(t) = \frac{a}{b}\frac{e^{a(t-\tau)}}{1+e^{a(t-\tau)}}$. Note in the following

$$\begin{cases} a(\lambda) = 2\left(\|D_+\| - \lambda^{\frac{\varepsilon}{2}}\right)\left(1 - \left(2\sqrt{\frac{2}{\|D_+\|}}\right)\lambda^{\frac{\varepsilon}{4}} - \frac{4}{\|D_+\| - \lambda^{\frac{\varepsilon}{2}}}\lambda^{\frac{\varepsilon}{2}}\right), \\ b(\lambda) = 2\frac{\|D_+\| - \lambda^{\frac{\varepsilon}{2}}}{n\|D_+\|}. \end{cases}$$

By Grönwall comparison, for any $t \in [t_{+1}, t_2]$,

$$u_+(t) \geq \frac{a(\lambda)}{b(\lambda)}\frac{e^{a(\lambda)(t-\tau)}}{1 + e^{a(\lambda)(t-\tau)}} \quad \text{where} \quad u_+(t_{+1}) = \frac{a(\lambda)}{b(\lambda)}\frac{e^{a(\lambda)(t_{+1}-\tau)}}{1 + e^{a(\lambda)(t_{+1}-\tau)}}. \quad (21)$$

Thanks to Lemma 9, $u_+(t_{+1}) \geq \frac{\underline{c}^2}{4}\lambda^{2(1+\varepsilon)}$. This implies that

$$\tau \leq t_{+1} - \frac{2(1+\varepsilon)}{a(\lambda)}\ln(\lambda) - \frac{1}{a(\lambda)}\ln\left(\frac{\underline{c}^2 b(\lambda)}{4a(\lambda)}\right),$$

and so

$$u_+(t_2) \geq \frac{a(\lambda)}{b(\lambda)} - \frac{a(\lambda)}{b(\lambda)}\exp\left(-a(\lambda)(t_2 - t_{+1}) - 2(1+\varepsilon)\ln(\lambda) - \ln\left(\frac{\underline{c}^2 b(\lambda)}{a(\lambda)}\right)\right).$$

In particular, if $t_2 - t_{+1} > -\frac{1+2\varepsilon}{\|D_+\|}\ln(\lambda)$, then

$$u_+(t_2) > \frac{a(\lambda)}{b(\lambda)} - \frac{a(\lambda)^2}{\underline{c}^2 b(\lambda)^2}\lambda^{\frac{a(\lambda)}{\|D_+\|}(1+2\varepsilon)-2(1+\varepsilon)}.$$

Note that $\frac{a(\lambda)}{b(\lambda)} = n\|D_+\| - \mathcal{O}\left(\lambda^{\frac{\varepsilon}{4}}\right)$ and $\lim_{\lambda\to 0}\frac{a(\lambda)}{\|D_+\|}(1+2\varepsilon)-2(1+\varepsilon) = 2\varepsilon$. As a consequence, for λ small enough, we have $u_+(t_2) > n\|D_+\| - \lambda^{\frac{\varepsilon}{5}}$ if $t_2 - t_{+1} > -\frac{1+2\varepsilon}{\|D_+\|}\ln(\lambda)$. This would break the second condition in the definition of the second phase Equation (17), so that $t_2 - t_{+1} \leq -\frac{1+2\varepsilon}{\|D_+\|}\ln(\lambda)$ and thanks to Lemma 7, $t_2 \leq -\frac{1+3\varepsilon}{\|D_+\|}\ln(\lambda)$.

Second point: let $j \in [\![m]\!] \setminus S_{+,1}$. Similarly to the proof of Lemma 8, we can show if $\mathsf{s}_j = -1$ during the second phase that

$$\frac{\mathrm{d}\|w_j^t\|}{\mathrm{d}t} \leq \left(\|D_-\| + 4mc^2\lambda^{2\varepsilon}\right)\|w_j^t\|.$$

If $\mathsf{s}_j = 1$ instead, there is some k_j such that $y_{k_j} > 0$ and $\langle w_j^t, x_{k_j}\rangle < 0$ thanks to Lemma 6 and the continuous initialisation. In that case we have the following inequalities

$$\begin{aligned} \frac{\mathrm{d}\|w_j^t\|}{\mathrm{d}t} &\leq -\frac{1}{n}\sum_{k|y_k>0}(h_{\theta^t}(x_k) - y_k)\langle w_j^t, x_k\rangle_+ \\ &\leq \frac{1}{n}\sum_{k|y_k>0} y_k\langle w_j^t, x_k\rangle_+ + 4mc^2\lambda^{2\varepsilon}\|w_j^t\| \\ &\leq \left(\frac{1}{n}\sqrt{\sum_{k\neq k_j|y_k>0} y_k^2} + 4mc^2\lambda^{2\varepsilon}\right)\|w_j^t\| \\ &\leq \left((1-\alpha)\|D_+\| + 4mc^2\lambda^{2\varepsilon}\right)\|w_j^t\|, \end{aligned}$$

where we recall $\alpha = \frac{\min_{k|y_k>0} y_k^2}{2\|D_+\|^2} > 0$. The previous inequalities are also valid during the first phase. In any case, for any $j \notin S_{+,1}$ and $t \le t_2$:

$$\frac{\mathrm{d}\|w_j^t\|}{\mathrm{d}t} \le \left(\max\left((1-\alpha)\|D_+\|, \|D_-\|\right) + 4mc^2\lambda^{2\varepsilon}\right)\|w_j^t\|.$$

Note that we chose ε small enough in Appendix B.1, so that $(1+3\varepsilon)\frac{\max((1-\alpha)\|D_+\|,\|D_-\|)}{\|D_+\|} \le 1-\varepsilon$. Since $t_2 \le -\frac{1+3\varepsilon}{\|D_+\|}\ln(\lambda)$, Grönwall inequality yields for any $j \notin S_{+,1}$ and λ small enough

$$\begin{aligned}\|w_j^{t_2}\| &\le \|w_j^0\| e^{-(1-\varepsilon)\ln(\lambda)} e^{-(1+3\varepsilon)\frac{4mc^2\lambda^{2\varepsilon}}{\|D_+\|}\ln(\lambda)} \\ &< 2c\lambda^{\varepsilon}.\end{aligned}$$

Third point: let $j \in S_{+,1}$. Recall that we have for any $t \in [t_{+1}, t_2]$

$$\begin{aligned}\frac{\mathrm{d}\langle \mathsf{w}_j^t, D_+\rangle}{\mathrm{d}t} &= \langle D_j^{\theta_t}, D_+\rangle - \langle \mathsf{w}_j^t, D_j^{\theta_t}\rangle\langle \mathsf{w}_j^t, D_+\rangle \\ &= \langle D_+^t, D_+\rangle - \langle \mathsf{w}_j^t, D_+^t\rangle\langle \mathsf{w}_j^t, D_+\rangle.\end{aligned}$$

Thanks to Equations (18) and (19)

$$\frac{\mathrm{d}\langle \mathsf{w}_j^t, D_+\rangle}{\mathrm{d}t} \ge \left(1 - \frac{u_+(t)}{n\|D_+\|} + h_+(t)\right)\left(\|D_+\|^2 - \langle \mathsf{w}_j^t, D_+\rangle^2\right) - \|D_+\|\|d_j^t\|\|h_\perp(t)\|. \tag{22}$$

Let us now define

$$t_u = \inf\left\{t \ge t_{+1} \mid u_+(t) \ge \frac{n\|D_+\|}{7} \text{ or } h(t) \ge 5\|D_+\|\lambda^{\varepsilon}\right\}.$$

For any $t \in [t_{+1}, t_u]$, we have for λ small enough

$$\frac{\mathrm{d}\langle \mathsf{w}_j^t, D_+\rangle}{\mathrm{d}t} \ge \frac{5}{7}\left(\|D_+\|^2 - \langle \mathsf{w}_j^t, D_+\rangle^2\right) - \frac{20}{7}\|D_+\|^2\lambda^{\varepsilon}.$$

The positive solution of the ODE $f'(t) = a^2 - b^2 f(t)^2$ is either increasing if $f(0) \le \frac{a}{b}$ or remains larger than $\frac{a}{b}$. The following inequality thus implies by Grönwall comparison for any $t \in [t_{+1}, t_u]$ and λ small enough

$$\begin{aligned}\langle \mathsf{w}_j^t, D_+\rangle &\ge \min\left(\|D_+\| - 4\|D_+\|\lambda^{\varepsilon}, \langle \mathsf{w}_j^{t_{+1}}, D_+\rangle\right) \\ &\ge \|D_+\| - 4\|D_+\|\lambda^{\varepsilon}.\end{aligned}$$

We thus have $h(t_u) < 5\|D_+\|\lambda^{\varepsilon}$. So $u_+(t_u) = \frac{n\|D_+\|}{7}$. Let us now bound $t_2 - t_u$. Similarly to Equation (21), we actually have for any $t \in [t_u, t_2]$

$$u_+(t) \ge \frac{a(\lambda)}{b(\lambda)}\frac{e^{a(\lambda)(t-\tau_u)}}{1+e^{a(\lambda)(t-\tau_u)}} \quad \text{where} \quad u_+(t_u) = \frac{a(\lambda)}{b(\lambda)}\frac{e^{a(\lambda)(t_u-\tau_u)}}{1+e^{a(\lambda)(t_u-\tau_u)}}.$$

We showed $u_+(t_u) = \frac{n\|D_+\|}{7}$ and so $\tau_u \le t_u - \frac{1}{a(\lambda)}\ln(\frac{b(\lambda)}{a(\lambda)}\frac{n\|D_+\|}{7})$, i.e. for any $t \in [t_u, t_2]$:

$$u_+(t) \ge \frac{a(\lambda)}{b(\lambda)} - \frac{7a(\lambda)^2}{b(\lambda)^2 n\|D_+\|}e^{-a(\lambda)(t-t_u)}.$$

Similarly to the proof of the first point, we can then show that for λ small enough, $t_2 - t_u \le -\frac{\varepsilon}{5\|D_+\|}\ln(\lambda)$. Now note that for any $t \in [t_u, t_2]$, Equation (22) yields:

$$\frac{\mathrm{d}\langle \mathsf{w}_j^t, D_+\rangle}{\mathrm{d}t} \ge -2\|D_+\|h(t) - 4\sqrt{2}\|D_+\|mc^2\lambda^{2\varepsilon}.$$

This directly leads to

$$h'(t) \le 2\|D_+\|h(t) + 4\sqrt{2}\|D_+\|mc^2\lambda^{2\varepsilon}.$$

By Grönwall's comparison, we thus have for λ small enough

$$\begin{aligned}h(t_2) &\le \left(h(t_u) + 2\sqrt{2}mc^2\lambda^{2\varepsilon}\right)e^{2\|D_+\|(t_2-t_u)} \\ &\le 6\|D_+\|\lambda^{\varepsilon}e^{-\frac{2\varepsilon}{5}\ln(\lambda)} < \lambda^{\frac{\varepsilon}{2}}.\end{aligned}$$

□

B.5 Phase 3: proof of Lemma 4

Similarly to the second phase with the positive labels, the third phase aims at fitting the negative labels. During this third phase, the norm of the neurons in $S_{-,1}$ (which are aligned with $-D_-$) grows until perfectly fitting all the negative labels of the training points; while all the other neurons do not change significantly. We define the ending time of the second phase as

$$\begin{aligned} t_3 = \inf \Bigg\{ t \geq t_{-1} \mid \quad & \exists j \notin S_{+,1} \cup S_{-,1}, \|w_j^t\| \geq 3c\lambda^\varepsilon \\ & \text{or } \sum_{j \in S_{-,1}} \|w_j^t\|^2 \geq n\|D_-\| - \lambda^{\frac{\varepsilon}{29}} \\ & \text{or } \exists j \in S_{-,1}, \langle \mathsf{w}_j^t, -D_- \rangle \leq \|D_-\| - \lambda^{\frac{\varepsilon}{14}} \\ & \text{or } \left(t \geq t_2 \text{ and } \|D_+^t\| \geq \lambda^{\frac{\varepsilon}{14}}\right) \Bigg\}. \end{aligned} \tag{23}$$

The following lemma corresponds to Lemma 4 of the main text.

Lemma 11. *If $\lambda \leq \lambda_*$,then the following inequalities hold*

(i) $t_3 \leq -\frac{1+3r\varepsilon}{\|D_-\|}\ln(\lambda)$,
(ii) $\forall j \notin S_{+,1} \cup S_{-,1}, \|w_j^{t_3}\| < 3c\lambda^\varepsilon$,
(iii) $\forall j \in S_{-,1}, \langle \mathsf{w}_j^{t_3}, -D_- \rangle > \|D_-\| - \lambda^{\frac{\varepsilon}{14}}$,
(iv) $\|D_+^{t_3}\| < \lambda^{\frac{\varepsilon}{14}}$.

The first point of the lemma states that the third phase also lasts a time of order $-\ln(\lambda)$. It actually ends after the second one ($t_3 > t_2$), since the neurons in $S_{-,1}$ do not grow in norm during the second one. The last point states that the positive labels remain fitted during the third phase. As a consequence, this also means that the neurons of $S_{+,1}$ do not change a lot after t_2. The other points imply that the second condition in Equation (23) holds at t_3, meaning that the norm of the neurons in $S_{-,1}$ have grown enough to fit the negative labels.

Proof. Preliminaries: the three first points of this proof share similarities with the proof of Lemma 10. For conciseness and clarity, parts of the proof are shortened, as they follow the same lines as the proof of Lemma 10. Similarly to the proof of the second phase, we define $h(t) = \|D_-\| - \min_{j \in S_{-,1}} \langle w_j^t, -D_- \rangle$ and we have for any $j \in S_{-,1}$ the decomposition

$$\mathsf{w}_j^t = -\frac{D_-}{\|D_-\|} + d_j^t \quad \text{with} \quad \|d_j^t\|^2 \leq \frac{2h(t)}{\|D_-\|}. \tag{24}$$

We have for any k such that $y_k < 0$

$$\begin{aligned} h_{\theta^t}(x_k) &\geq \sum_{j \in S_{-,1}} -\|w_j^t\| \langle w_j^t, x_k \rangle + \sum_{\substack{j \notin S_{-,1} \cup S_{+,1} \\ \langle w_j^t, x_k \rangle > 0}} \mathsf{s}_j \|w_j^t\| \langle w_j^t, x_k \rangle \\ &\geq \sum_{j \in S_{-,1}} \frac{\|w_j^t\|^2}{\|D_-\|} \langle D_-, x_k \rangle + h_k(t) = \sum_{j \in S_{-,1}} \frac{\|w_j^t\|^2}{n\|D_-\|} y_k + h_k(t), \end{aligned} \tag{25}$$

where

$$|h_k(t)| \leq \sum_{j \in S_{-,1}} \|w_j^t\|^2 |\langle d_j^t, x_k \rangle| + 9mc^2\lambda^{2\varepsilon}.$$

Since we chose λ small enough, this implies that $h_{\theta^t}(x_k) \geq y_k$ for k such that $y_k < 0$ and $t \in [t_{-1}, t_3]$. Moreover, thanks to Lemma 7, $\langle w_j^{t_{+1}}, x_k \rangle = 0$ for any such k and $j \in S_{+,1}$. Because of this, we have for any $[t_{-1}, t_3]$

$$\langle w_j^t, x_k \rangle \leq 0 \qquad \text{for all } j \in S_{+,1} \text{ and } k \text{ such that } y_k < 0.$$

Equation (25) is thus an equality on $[t_{-1}, t_3]$. Similarly to the second phase for D_+^t, we can now decompose D_-^t as

$$\begin{aligned} D_-^t &= D_- - \frac{\sum_{j\in S_{-,1}} \|w_j^t\|^2}{n\|D_-\|} \sum_{k,y_k<0} \langle D_-, x_k\rangle x_k - \frac{1}{n}\sum_{k,y_k<0} h_k(t)x_k \\ &= \left(1 - \frac{\sum_{j\in S_{-,1}} \|w_j^t\|^2}{n\|D_-\|}\right) D_- - \frac{1}{n}\sum_{k,y_k<0} h_k(t)x_k. \end{aligned}$$

And then for any $t \in [t_{-1}, t_3]$:

$$D_-^t = \left(1 - \frac{\sum_{j\in S_{-,1}} \|w_j^t\|^2}{n\|D_-\|} + h_-(t)\right) D_- + h_\perp(t), \tag{26}$$

where $\langle h_\perp(t), D_-\rangle = 0$ and $|h_-(t)|\|D_-\| \vee \|h_\perp(t)\| \leq \frac{1}{n}\sum_{j\in S_{-,1}} \|w_j^t\|^2 \sqrt{\frac{2h(t)}{\|D_-\|}} + 9mc^2\lambda^{2\varepsilon}$.

Using the previous decompositions, we also have the following inequalities for any $j \in S_{-,1}$

$$\begin{aligned} \langle D_j^{\theta^t}, \mathsf{w}_j^t\rangle &= \langle D_-^t, \mathsf{w}_j^t\rangle + \sum_{\substack{k,y_k>0 \\ \langle \mathsf{w}_j^t, x_k\rangle > 0}} \langle D_+^t, x_k\rangle \langle d_j^t, x_k\rangle \\ &= \langle D_-^t, \mathsf{w}_j^t\rangle + g_j(t), \end{aligned}$$

where $|g_j(t)| \leq \|d_j^t\|\|D_+^t\|$.

First point: with the previous decompositions, we can now prove the first point of Lemma 11. Define $u_-(t) = \sum_{j\in S_{-,1}} \|w_j^t\|^2$. Based on the previous inequalities, we have for any $t \in [t_{-1}, t_3]$

$$\begin{aligned} \frac{1}{2}\frac{\mathrm{d}u_-(t)}{\mathrm{d}t} &= \sum_{j\in S_{-,1}} \|w_j^t\|\langle -D_j^{\theta^t}, w_j^t\rangle \\ &= \sum_{j\in S_{-,1}} \|w_j^t\|^2\langle -D_-^t, \mathsf{w}_j^t\rangle - \|w_j^t\|^2 g_j(t) \\ &\geq u_-(t)\left(1 - \frac{u_-(t)}{n\|D_-\|} + h_-(t)\right)\left(\|D_-\| - \lambda^{\frac{\varepsilon}{14}}\right) - u_-(t)\sqrt{\frac{2h(t)}{\|D_-\|}}\left(\|h_\perp(t)\| + \|D_+^t\|\right). \end{aligned}$$

From there, we can show similarly to the proof of the second phase that $t_3 \leq -\frac{1+3r\varepsilon}{\|D_-\|}\ln(\lambda)$.

Second point: first consider $j \notin S_{+,1} \cup S_{-,1}$ such that $\mathsf{s}_j = 1$. Thanks to Lemma 10, we already have $\|w_j^t\| < 2c\lambda^\varepsilon$ for any $t \leq t_2$. For $t \in [t_2, t_3]$, we then have

$$\frac{\mathrm{d}\|w_j^t\|}{\mathrm{d}t} \leq \|D_+^t\|\|w_j^t\|.$$

Grönwall inequality then gives for λ small enough: $\|w_j^{t_3}\| \leq 2c\lambda^\varepsilon e^{-\lambda^{\frac{\varepsilon}{14}}\frac{1+3r\varepsilon}{\|D_-\|}\ln(\lambda)} < 3c\lambda^\varepsilon$.

Let now $j \notin S_{+,1} \cup S_{-,1}$ such that $\mathsf{s}_j = -1$. By definition, there is some k_j such that $y_{k_j} < 0$ and $\langle w_j^t, x_{k_j}\rangle < 0$. Similarly to the proof of Lemma 10, we then have for any $t \in [0, t_3]$:

$$\begin{aligned} \frac{\mathrm{d}\|w_j^t\|}{\mathrm{d}t} &= -\langle w_j^t, D_j^{\theta^t}\rangle \\ &\leq \left((1-\beta)\|D_-\| + \lambda^{\frac{\varepsilon}{15}}\right)\|w_j^t\|, \end{aligned}$$

where we recall $\beta = \frac{\min_{k,y_k<0} y_k^2}{2\|D_-\|^2} > 0$. Note that we chose ε small enough, so that $(1+3r\varepsilon)(1-\beta) \leq 1-\varepsilon$. Grönwall inequality then yields for λ small enough $\|w_j^{t_3}\| < 3c\lambda^\varepsilon$.

Third point: let $j \in S_{-,1}$. Recall that for any $t \in [t_{-1}, t_3]$, $\langle \mathsf{w}_j^t, D_j^{\theta^t} \rangle = \langle \mathsf{w}_j^t, D_-^t \rangle + g_j(t)$. This yields for any $t \in [t_{-1}, t_3]$

$$\begin{aligned}\frac{\mathrm{d}\langle \mathsf{w}_j^t, -D_- \rangle}{\mathrm{d}t} &= \langle D_-^t, D_- \rangle - \langle \mathsf{w}_j^t, D_-^t \rangle \langle \mathsf{w}_j^t, D_- \rangle - g_j(t) \langle \mathsf{w}_j^t, D_- \rangle \\ &\geq \left(1 - \frac{u_-(t)}{n\|D_-\|} + h_-(t)\right) \left(\|D_-\|^2 - \langle \mathsf{w}_j^t, -D_- \rangle^2\right) - \sqrt{2\|D_-\| h(t)} \|h_\perp(t)\| - \|D_-\| |g_j(t)|.\end{aligned}$$

Let us now define

$$t_u = \inf \left\{ t \geq t_{-1} \mid u_-(t) \geq \frac{n\|D_-\|}{7} \text{ or } h(t) \geq 5\|D_-\| \lambda^{\frac{\varepsilon}{8}} \right\}.$$

Similarly to the third phase, we can show for λ small enough the following sequence of properties

- $h(t_u) < 5\|D_-\| \lambda^{\frac{\varepsilon}{8}}$
- $u_-(t_u) = \frac{n\|D_-\|}{7}$
- $t_3 - t_u \leq -\frac{\varepsilon}{57\|D_-\|} \ln(\lambda)$
- $\langle \mathsf{w}_j^{t_3}, -D_- \rangle > \|D_-\| - \lambda^{\frac{\varepsilon}{14}}$.

Fourth point: define for this point

$$\tau = \inf \left\{ t \geq t_{-1} \mid u_-(t) \geq n\lambda^{\frac{\varepsilon}{5}} \right\} \wedge t_3.$$

Thanks to Lemma 10, $\tau \geq t_2$. For any k such that $y_k > 0$, we have

$$\frac{\mathrm{d}h_{\theta^t}(x_k)}{\mathrm{d}t} = \sum_{j, \langle w_j^t, x_k \rangle > 0} \|w_j^t\|^2 \langle D_j^{\theta^t}, x_k \rangle + \langle w_j^t, D_j^{\theta^t} \rangle \langle w_j^t, x_k \rangle.$$

And so

$$\begin{aligned}\frac{\mathrm{d}D_+^t}{\mathrm{d}t} &= -\frac{1}{n} \sum_{k, y_k > 0} \frac{\mathrm{d}h_{\theta^t}(x_k)}{\mathrm{d}t} x_k \\ &= -\frac{1}{n} \sum_{k, y_k > 0} \sum_{j, \langle w_j^t, x_k \rangle > 0} \|w_j^t\|^2 \langle D_+^t, x_k \rangle x_k + \langle w_j^t, D_j^{\theta^t} \rangle \langle w_j^t, x_k \rangle x_k.\end{aligned}$$

Recall the for any $j \in S_{+,1}$, $\langle w_j^t, x_k \rangle = 0$ for any k such that $y_k < 0$. From there, we have $\langle w_j^t, D_j^{\theta^t} \rangle = \langle w_j^t, D_+^t \rangle$. This leads to the following inequalities

$$\begin{aligned}\frac{1}{2} \frac{\mathrm{d}\|D_+^t\|^2}{\mathrm{d}t} &= -\frac{1}{n} \sum_{k, y_k > 0} \sum_{j, \langle w_j^t, x_k \rangle > 0} \|w_j^t\|^2 \langle D_+^t, x_k \rangle^2 + \langle w_j^t, D_j^{\theta^t} \rangle \langle w_j^t, x_k \rangle \langle D_+^t, x_k \rangle \qquad (27) \\ &\leq -\frac{\sum_{j \in S_{+,1}} \|w_j^t\|^2}{n} \sum_{k, y_k > 0} \langle D_+^t, x_k \rangle^2 - \frac{1}{n} \sum_{j \in S_{+,1}} \langle w_j^t, D_+^t \rangle \sum_{k, y_k > 0} \langle w_j^t, x_k \rangle \langle D_+^t, x_k \rangle \\ &\quad - \frac{1}{n} \sum_{k | y_k > 0} \sum_{\substack{j \notin S_{+,1} \\ \langle w_j^t, x_k \rangle > 0}} \langle w_j^t, D_j^{\theta^t} \rangle \langle w_j^t, x_k \rangle \langle D_+^t, x_k \rangle \\ &\leq -\frac{\sum_{j \in S_{+,1}} \|w_j^t\|^2}{n} \sum_{k, y_k > 0} \langle D_+^t, x_k \rangle^2 - \frac{1}{n} \sum_{j \in S_{+,1}} \langle w_j^t, D_+^t \rangle^2 \\ &\quad - \frac{1}{n} \sum_{k | y_k > 0} \sum_{\substack{j \notin S_{+,1} \\ \langle w_j^t, x_k \rangle > 0}} \langle w_j^t, D_j^{\theta^t} \rangle \langle w_j^t, x_k \rangle \langle D_+^t, x_k \rangle.\end{aligned}$$

Given the bounds on $\|D_+^t\|$ and t_3, a simple Grönwall argument implies that $\sum_{j\in S_{+,1}} \|w_j^t\|^2$ did not change significantly between t_2 and t_3. In particular for $t \in [t_2, \tau]$ and λ small enough:

$$\frac{\mathrm{d}\|D_+^t\|}{\mathrm{d}t} \leq -\left(\|D_+\| - \lambda^{\frac{\varepsilon}{15}}\right)\|D_+^t\| + \left(\|D_-\| + 2\lambda^{\frac{\varepsilon}{29}}\right)\lambda^{\frac{\varepsilon}{5}}.$$

Since $\|D_+^{t_2}\| \leq 2\|D_+\|\lambda^{\frac{\varepsilon}{5}}$ thanks to Lemma 10, the above inequality implies by Grönwall comparison that $\|D_+^t\| \leq (1 + 2\|D_+\|)\,\lambda^{\frac{\varepsilon}{5}}$ for any $t \in [t_2, \tau]$ and λ small enough.

For any $j \in S_{-,1}$, define

$$\alpha_j(t) = \sqrt{\sum_{\substack{k|y_k>0 \\ \langle \mathsf{w}_j^t, x_k\rangle > 0}} \langle \mathsf{w}_j^t, x_k\rangle^2}.$$

Since $\frac{\mathrm{d}\langle \mathsf{w}_j^t, x_k\rangle}{\mathrm{d}t} = \mathsf{s}_j\left(\langle D_+^t, x_k\rangle - \langle D_j^{\theta^t}, \mathsf{w}_j^t\rangle\langle \mathsf{w}_j^t, x_k\rangle\right)\mathbb{1}_{\langle \mathsf{w}_j^t, x_k\rangle>0}$, we have

$$\begin{aligned}\frac{1}{2}\frac{\mathrm{d}\alpha_j(t)^2}{\mathrm{d}t} &= -\sum_{\substack{k|y_k>0 \\ \langle \mathsf{w}_j^t, x_k\rangle > 0}} \langle D_+^t, x_k\rangle\langle \mathsf{w}_j^t, x_k\rangle - \langle D_j^{\theta^t}, \mathsf{w}_j^t\rangle\langle \mathsf{w}_j^t, x_k\rangle^2 \\ &\leq \|D_+^t\|\alpha_j(t) + \langle D_j^{\theta^t}, \mathsf{w}_j^t\rangle\alpha_j(t)^2.\end{aligned}$$

Note that for λ small enough $\langle D_j^{\theta^t}, \mathsf{w}_j^t\rangle \leq -\frac{\|D_-\|}{2}$ for any $j \in S_{-,1}$ and $t \in [t_{-1}, \tau]$. We thus have for any $t \in [t_2, \tau]$

$$\frac{\mathrm{d}\alpha_j(t)}{\mathrm{d}t} \leq (1 + 2\|D_+\|)\,\lambda^{\frac{\varepsilon}{5}} - \frac{\|D_-\|}{2}\alpha_j(t).$$

Moreover, recall that $\langle D_+^t, x_k\rangle > 0$ before t_2. Thanks to Lemma 8, this actually implies that $\alpha_j^{t_2} = 0$ and by Grönwall comparison: $\alpha_j^\tau \leq \left(4r + \frac{2}{\|D_-\|}\right)\lambda^{\frac{\varepsilon}{5}}$.

Now note that $\langle D_j^{\theta^t}, \mathsf{w}_j^t\rangle \leq 0$ for any $t \in [\tau, t_3]$, which leads for any $t \in [\tau, t_3]$ to

$$\alpha_j(t) \leq \left(4r + \frac{2}{\|D_-\|}\right)\lambda^{\frac{\varepsilon}{5}} + \int_\tau^t \|D_+^s\|\mathrm{d}s.$$

Equation (27) then becomes for any $t \in [\tau, t_3]$ and λ small enough

$$\begin{aligned}\frac{\mathrm{d}\|D_+^t\|}{\mathrm{d}t} &\leq -\left(\|D_+\| - \lambda^{\frac{\varepsilon}{15}}\right)\|D_+^t\| + \left(\|D_-\| + \lambda^{\frac{\varepsilon}{29}}\right)^2 \max_{j\in S_{-,1}} \alpha_j(t) + 9mc^2\lambda^{2\varepsilon} \\ &\leq -\left(\|D_+\| - \lambda^{\frac{\varepsilon}{15}}\right)\|D_+^t\| + \left(\|D_-\| + \lambda^{\frac{\varepsilon}{29}}\right)^2 \int_\tau^t \|D_+^s\|\mathrm{d}s + 5\|D_-\|\,(1 + \|D_+\|)\,\lambda^{\frac{\varepsilon}{5}}.\end{aligned}$$

Thanks to Lemma 16, this implies for any $t \in [\tau, t_3]$ and λ small enough

$$\|D_+^t\| \leq \left(2\|D_+\| + 6\|D_-\| + \frac{6}{r}\right)\lambda^{\frac{\varepsilon}{5}}\left(1 + e^{\frac{\left(\|D_-\| + \lambda^{\frac{\varepsilon}{29}}\right)^2}{\left(\|D_+\| - \lambda^{\frac{\varepsilon}{15}}\right)}(t-\tau)}\right).$$

Similarly to the first point, we can also show that $t_3 - \tau \leq -\frac{\varepsilon}{8\|D_-\|}\ln(\lambda)$, which finally yields that $\|D_+^t\| \leq \lambda^{\frac{\varepsilon}{14}}$ on $[t_2, t_3]$ for λ small enough. □

B.6 Final phase: proof of Theorem 1

To prove Theorem 1, we need to first prove some auxiliary lemmas. Lemma 12 shows that at time t_3 the neural network is in the vicinity of some identifiable (i.e. independent of λ) interpolator. Lemmas 13 to 15 allow to apply Chatterjee [2022] convergence result when a local PL is satisfied. Finally, we restate the main theorem in Theorem 2 for the sake of clearness and prove it.

Lemma 12. *For any $\lambda \leq \lambda^*$, there exists $\theta^* \in \operatorname{argmin}_{L(\theta)=0}\|\theta\|^2$ independent of λ such that*

$$\|\theta^{t_3} - \theta^*\| \leq \lambda^{\frac{\varepsilon}{31}},$$

where t_3 is defined in Equation (23).

Lemmas 10 and 11 directly imply Lemma 12 for some θ^*_λ that depends on λ. Extra work is required to prove that this minimal norm interpolator does not actually depend on λ.

Proof. Lemmas 10 and 11 imply the following properties

(i) for any $j \in S_{+,1}$, $\|\mathsf{w}_j^{t_3} - \frac{D_+}{\|D_+\|}\| \leq \lambda^{\frac{\varepsilon}{15}}$,

(ii) $\left|\sum_{j\in S_{+,1}} \|w_j^t\|^2 - n\|D_+\|\right| \leq \lambda^{\frac{\varepsilon}{15}}$,

(iii) for any $j \in S_{-,1}$, $\|\mathsf{w}_j^{t_3} + \frac{D_-}{\|D_-\|}\| \leq \sqrt{\frac{2}{\|D_-\|}}\lambda^{\frac{\varepsilon}{28}}$,

(iv) $\left|\sum_{j\in S_{-,1}} \|w_j^t\|^2 - n\|D_-\|\right| \leq \lambda^{\frac{\varepsilon}{29}}$,

(v) for any $j \notin S_{+,1} \cup S_{-,1}$, $\|w_j^t\| \leq 3c\lambda^{\varepsilon}$.

The balancedness property along these 5 properties guarantee there exists some $\theta^*_\lambda \in \operatorname{argmin}_{L(\theta)=0}\|\theta\|^2$ such that $\|\theta^{t_3} - \theta^*_\lambda\| \leq \lambda^{\frac{\varepsilon}{30}}$. To show that this θ^*_λ actually does not depend on λ, it remains to show that the norm of any individual neuron in $S_{+,1} \cup S_{-,1}$ is close to some constant (independent from λ).

Consider any pair $j, j' \in S_{+,1}$. We recall Equation (9):

$$\frac{\mathrm{d}\rho_j^t}{\mathrm{d}t} = \langle D_j^{\theta^t}, \mathsf{w}_j^t\rangle \quad \text{and} \quad \frac{\mathrm{d}\mathsf{w}_j^t}{\mathrm{d}t} = D_j^{\theta^t} - \langle D_j^{\theta^t}, \mathsf{w}_j^t\rangle \mathsf{w}_j^t.$$

We note in the following $\widetilde{\rho}_j^t, \widetilde{\mathsf{w}}_j^t$ any solutions of the ODEs

$$\begin{aligned} \frac{\mathrm{d}\widetilde{\rho}_j^t}{\mathrm{d}t} &= \langle \widetilde{D}_j^t, \widetilde{\mathsf{w}}_j^t\rangle \quad \text{with} \quad \widetilde{\rho}_j^0 = \ln(\|w_j^0\|/\lambda) \\ \frac{\mathrm{d}\widetilde{\mathsf{w}}_j^t}{\mathrm{d}t} &= \widetilde{D}_j^t - \langle \widetilde{D}_j^t, \widetilde{\mathsf{w}}_j^t\rangle \widetilde{\mathsf{w}}_j^t \quad \text{with} \quad \widetilde{\mathsf{w}}_j^0 = \mathsf{w}_j^0, \end{aligned} \tag{28}$$

where $\widetilde{D}_j^t = \frac{1}{n}\sum_k y_k x_k \mathbb{1}_{\langle \widetilde{\mathsf{w}}_j^t, x_k\rangle > 0}$. Remark that the process $(\widetilde{\mathsf{w}}, \widetilde{\rho})$ does not depend on λ as $\|w_j^0\| = \lambda g_j$ with the g_j's being standard Gaussian vectors. We first have

$$\begin{aligned} \frac{1}{2}\frac{\mathrm{d}\|\mathsf{w}_j^t - \widetilde{\mathsf{w}}_j^t\|^2}{\mathrm{d}t} &= \sum_k \left(\mathbb{1}_{\langle \widetilde{\mathsf{w}}_j^t, x_k\rangle > 0} - \mathbb{1}_{\langle \mathsf{w}_j^t, x_k\rangle > 0}\right) \frac{y_k}{n}\langle \widetilde{\mathsf{w}}_j^t - \mathsf{w}_j^t, x_k\rangle + \sum_k \frac{y_k + h_{\theta^t}(x_k)}{n}\langle \mathsf{w}_j^t, x_k\rangle_+ \langle \widetilde{\mathsf{w}}_j^t - \mathsf{w}_j^t, \mathsf{w}_j^t\rangle \\ &\quad - \sum_k \frac{y_k}{n}\langle \widetilde{\mathsf{w}}_j^t, x_k\rangle_+ \langle \widetilde{\mathsf{w}}_j^t - \mathsf{w}_j^t, \widetilde{\mathsf{w}}_j^t\rangle + \sum_k \mathbb{1}_{\langle \mathsf{w}_j^t, x_k\rangle > 0}\frac{h_{\theta^t}(x_k)}{n}\langle \widetilde{\mathsf{w}}_j^t - \mathsf{w}_j^t, x_k\rangle. \end{aligned}$$

For $y_k > 0$, both $\mathbb{1}_{\langle \widetilde{\mathsf{w}}_j^t, x_k\rangle > 0}$ and $\mathbb{1}_{\langle \mathsf{w}_j^t, x_k\rangle > 0}$ remain positive until t_{+1}. As a consequence, the first sum is non-positive, which leads to

$$\frac{1}{2}\frac{\mathrm{d}\|\mathsf{w}_j^t - \widetilde{\mathsf{w}}_j^t\|^2}{\mathrm{d}t} \leq \frac{2}{n}\sqrt{\sum_k y_k^2}\|\mathsf{w}_j^t - \widetilde{\mathsf{w}}_j^t\|^2 + \frac{2}{n}\sqrt{\sum_k h_{\theta^t}(x_k)^2}\|\mathsf{w}_j^t - \widetilde{\mathsf{w}}_j^t\|,$$

and then

$$\frac{\mathrm{d}\|\mathsf{w}_j^t - \widetilde{\mathsf{w}}_j^t\|}{\mathrm{d}t} \leq \frac{2}{n}\sqrt{\sum_k y_k^2}\|\mathsf{w}_j^t - \widetilde{\mathsf{w}}_j^t\| + \frac{2}{n}\sqrt{\sum_k h_{\theta^t}(x_k)^2}.$$

Since $|h_{\theta^t}(x_k)| \leq 4mc^2\lambda^{2(1-\varepsilon)}$ during the first phase, Grönwall lemma implies that

$$\forall t \in [0, t_{+1}], \|\mathsf{w}_j^t - \widetilde{\mathsf{w}}_j^t\| \leq \frac{4mc^2}{\sqrt{n\left(\|D_+^2 + D_-\|^2\right)}}\lambda^{2(1-\varepsilon)}\left(e^{2\sqrt{\|D_+^2+D_-\|^2}t} - 1\right).$$

We thus have $\|\mathsf{w}_j^{t_{+1}} - \widetilde{\mathsf{w}}_j^{t_{+1}}\| \leq \frac{4mc^2}{\sqrt{n\left(\|D_+^2+D_-\|^2\right)}}\lambda^{2-2(1+\sqrt{2})\varepsilon}$. From there, note that we also have

$$\begin{aligned}\frac{\mathrm{d}\widetilde{\rho}_j^t - \rho_j^t}{\mathrm{d}t} &= \sum_k \frac{y_k}{n}\left(\langle \widetilde{\mathsf{w}}_j^t, x_k\rangle_+ - \langle \mathsf{w}_j^t, x_k\rangle_+\right) - \sum_k \frac{h_{\theta^t}(x_k)}{n}\langle \mathsf{w}_j^t, x_k\rangle_+ \\ &\leq \sqrt{\|D_+\|^2 + \|D_-\|^2}\|\widetilde{\mathsf{w}}_j^t - \mathsf{w}_j^t\| + \frac{1}{n}\sqrt{\sum_k h_{\theta^t}(x_k)^2}.\end{aligned}$$

As a consequence:

$$\left|\widetilde{\rho}_j^{t_{+1}} - \rho_j^{t_{+1}} - \ln(\lambda)\right| \leq -\frac{8mc^2\varepsilon}{\|D_+\|}\ln(\lambda)\lambda^{2-2(1+\sqrt{2})\varepsilon}.$$

Moreover, for any $t \in [t_{+1}, t_3]$, we have $\frac{\mathrm{d}\rho_j^t - \rho_{j'}^t}{\mathrm{d}t} \leq -2\sqrt{\frac{2}{\|D_+\|}}\|D_+^t\|\frac{3\varepsilon}{\|D_+\|}\ln(\lambda)\lambda^{\frac{\varepsilon}{14}}$, and so for λ small enough:

$$|\rho_j^{t_3} - \rho_{j'}^{t_3}| \leq |\widetilde{\rho}_j^{t_{+1}} - \widetilde{\rho}_{j'}^{t_{+1}}| + 2\lambda^{\frac{\varepsilon}{15}}. \tag{29}$$

The quantity $|\widetilde{\rho}_j^{t_{+1}} - \widetilde{\rho}_{j'}^{t_{+1}}|$ only depends on λ because of the t_{+1} term. The $\widetilde{\rho}$ are indeed independent of λ.

Similarly to the proof of Lemma 7, we can show that after some time t_j, $\widetilde{D}_j^t = D_+$ for all $t \geq t_j$. From there, Equation (28) imply for any $t \geq t_j$

$$\begin{aligned}\frac{\mathrm{d}\widetilde{\rho}_j^t}{\mathrm{d}t} &= \langle D_+, \widetilde{\mathsf{w}}_j^t\rangle \\ \frac{\mathrm{d}\langle \widetilde{\mathsf{w}}_j^t, D_+\rangle}{\mathrm{d}t} &= \|D_+\|^2 - \langle D_+, \widetilde{\mathsf{w}}_j^t\rangle^2.\end{aligned}$$

Solving these ODEs, we thus have for some constants τ_j, c_j and any $t \geq t_j$:

$$\begin{aligned}\langle \widetilde{\mathsf{w}}_j^t, D_+\rangle &= \|D_+\|\tanh\left(\|D_+\|(t-\tau_j)\right) \\ \widetilde{\rho}_j^t &= \ln\left(\cosh\left(\|D_+\|(t-\tau_j)\right)\right) + c_j.\end{aligned}$$

For λ small enough, $t_{+1} \geq t_j \vee t_{j'}$ and then:

$$\begin{aligned}\widetilde{\rho}_j^{t_{+1}} - \widetilde{\rho}_{j'}^{t_{+1}} &= c_j - c_{j'} + \|D_+\|\left(\tau_{j'} - \tau_j\right) + \ln\left(\frac{1 + e^{2\tau_j}\lambda^{2\varepsilon}}{1 + e^{2\tau_{j'}}\lambda^{2\varepsilon}}\right) \\ &= d_{j,j'} + h_{j,j'}(\lambda),\end{aligned}$$

where $d_{j,j'} = c_j - c_{j'} + \|D_+\|\left(\tau_{j'} - \tau_j\right)$ and $|h_{j,j'}(\lambda)| \leq e^{2(\tau_j \vee \tau_{j'})}\lambda^{2\varepsilon}$. Using Equation (29), this leads for λ small enough to

$$\begin{aligned}\|w_j^{t_3}\| &= \|w_{j'}^{t_3}\|e^{\rho_j^t - \rho_{j'}^t} \\ &= \|w_{j'}^{t_3}\|e^{d_{j,j'}}\left(1 + g_{j,j'}(\lambda)\right),\end{aligned}$$

where $|g_{j,j'}(\lambda)| \leq 4\lambda^{\frac{\varepsilon}{15}}$. As the sum of the norms of the w_j is known, this actually fixes the norm of each individual neuron. Precisely we have for $|f(\lambda)| \leq \lambda^{\frac{\varepsilon}{15}}$

$$\begin{aligned}n\|D_+\| - f(\lambda) &= \sum_{i \in S_{+,1}} \|w_i^{t_3}\|^2 \\ &= \|w_j^{t_3}\|^2 \sum_{i \in S_{+,1}} e^{2d_{i,j}}\left(1 + 2g_{i,j} + g_{i,j}^2\right).\end{aligned}$$

And so we finally have for any $j \in S_{+,1}$, $\|w_j^{t_3}\| = \sqrt{\frac{n\|D_+\|}{\sum_{i \in S_{+,1}} e^{2d_{i,j}}}} + \mathcal{O}\left(\lambda^{\frac{\varepsilon}{15}}\right)$. Similarly, we can show that for any $j \in S_{-,1}$, $\|w_j^{t_3}\| = \sqrt{\frac{n\|D_-\|}{\sum_{i \in S_{-,1}} e^{2d_{i,j}}}} + \mathcal{O}\left(\lambda^{\frac{\varepsilon}{30}}\right)$. We can now define θ^* as follows:

- $w_j^* = \sqrt{\frac{n\|D_+\|}{\sum_{i\in S_{+,1}} e^{2d_{i,j}}}}\frac{D_+}{\|D_+\|}$ and $a_j^* = \|w_j^*\|$ if $j \in S_{+,1}$
- $w_j^* = -\sqrt{\frac{n\|D_-\|}{\sum_{i\in S_{-,1}} e^{2d_{i,j}}}}\frac{D_-}{\|D_-\|}$ and $a_j^* = -\|w_j^*\|$ if $j \in S_{-,1}$
- $w_j^* = 0$ and $a_j^* = 0$ if $j \notin S_{+,1} \cup S_{-,1}$.

By construction, θ^* does not depend on λ and $\|\theta^{t_3} - \theta^*\| \leq \lambda^{\frac{\varepsilon}{31}}$ for λ small enough. Moreover, thanks to Proposition 1, $\theta^* \in \operatorname{argmin}_{L(\theta)=0}\|\theta\|^2$. □

For this final phase, we know that at time t_3, given Lemma 11, the neural net is arrived at a point θ^{t_3} satisfying the following: for $\varepsilon' = \frac{\varepsilon}{30}$,

(i) For all $t > 0$, $\theta^t \in \Theta := \{\theta = (a_j, w_j)_{j\leq m}, \text{ such that } \forall j \in [\![m]\!],\ |a_j|^2 = \|w_j\|^2\}$.

(ii) For all $j \in [\![m]\!] \setminus (S_{+,1} \cup S_{-,1})$, $\|w_j^{t_3}\| \leq \lambda^{\varepsilon'}$.

(iii) We have $\left|\frac{1}{n}\sum_{j\in S_{+,1}} \|w_j^{t_3}\|^2 - \|D_+\|\right| \leq \lambda^{\epsilon'}$ and $\left|\frac{1}{n}\sum_{j\in S_{-,1}} \|w_j^{t_3}\|^2 - \|D_-\|\right| \leq \lambda^{\epsilon'}$.

(iv) For all $j \in S_{+,1}$, $\left|\langle \mathsf{w}_j^{t_3}, D_+\rangle - \|D_+\|\right| \leq \lambda^{\epsilon'}$ and $j \in S_{-,1}$, $\left|\langle \mathsf{w}_j^{t_3}, D_-\rangle - \|D_-\|\right| \leq \lambda^{\epsilon'}$.

Let us fist show an auxiliary Lemma that states that when these four conditions are satisfied, then the loss is almost zero.

Lemma 13. *For θ such that conditions (i), (ii), (iii) and (iv) are satisfied then, for λ small enough,*

$$L(\theta) \leq \lambda^{\varepsilon'/2}. \tag{30}$$

Proof. Assume that θ satisfies conditions (i), (ii), (iii) and (iv). Then, for k such that $y_k > 0$,

$$\begin{aligned}
|h_\theta(x_k) - y_k| &= \left|\sum_{j|\langle w_j, x_k\rangle > 0} \mathsf{s}_j \|w_j\| \langle w_j, x_k\rangle - y_k\right| \\
&\leq \left|\sum_{j\in S_{+,1}} \mathsf{s}_j \|w_j\| \langle w_j, x_k\rangle - y_k\right| + \left|\sum_{j\notin S_{+,1}\cup S_{1-}} \mathsf{s}_j \|w_j\| \langle w_j, x_k\rangle\right| \\
&\leq \left|\sum_{j\in S_{+,1}} \|w_j\|^2 \langle \mathsf{w}_j, x_k\rangle - y_k\right| + \sum_{j\notin S_{+,1}\cup S_{-,1}} \|w_j\|^2.
\end{aligned}$$

Using the (ii) property, on the one hand, the second term is upper bounded by $m\lambda^{2\varepsilon'}$, and, on the other hand, as $\|\mathsf{w}_j - \frac{D_+}{\|D_+\|}\|^2 \leq 2\frac{\lambda^{\varepsilon'}}{\|D_+\|}$, we have by adding and subtracting $\frac{D_+}{\|D_+\|}$ in the inner product of the first term

$$\begin{aligned}
|h_\theta(x_k) - y_k| &\leq \left|\sum_{j\in S_{+,1}} \frac{\|w_j\|^2}{\|D_+\|}\langle D_+, x_k\rangle - y_k\right| + \sum_{j\in S_{+,1}} \|w_j\|^2 \sqrt{2\frac{\lambda^{\varepsilon'}}{\|D_+\|}} + m\lambda^{2\varepsilon'} \\
&\leq \left|\frac{1}{n}\sum_{j\in S_{+,1}} \frac{\|w_j\|^2}{\|D_+\|} - 1\right| y_k + \sum_{j\in S_{+,1}} \|w_j\|^2 \sqrt{2\frac{\lambda^{\varepsilon'}}{\|D_+\|}} + m\lambda^{2\varepsilon'} \\
&\leq \frac{\lambda^{\varepsilon'}}{\|D_+\|} y_k + 2n\sqrt{\|D_+\|}\lambda^{\varepsilon'/2} + \frac{n\lambda^{3\varepsilon'/2}}{\sqrt{\|D_+\|}} + m\lambda^{2\varepsilon'} \\
&\leq 2\lambda^{\varepsilon'/4},
\end{aligned}$$

for λ small enough. And the same goes similarly for $y_k < 0$. Hence,

$$L(\theta) = \frac{1}{2n}\sum_k \left(h_\theta(x_k) - y_k\right)^2 \leq \lambda^{\varepsilon'/2}.$$

This concludes the proof of the lemma. □

We show a local estimate of the PL inequality when balancedness, i.e. (i), is assumed.

Lemma 14. *For all $\theta \in \Theta$, we have*

$$\|\nabla L(\theta)\|^2 \geq 2L(\theta) \cdot \min\left\{\frac{1}{n}\sum_{j\in S_{-,1}}\|w_j\|^2, \frac{1}{n}\sum_{j\in S_{+,1}}\|w_j\|^2\right\}. \tag{31}$$

Proof. Indeed, thanks to the balancedness property we have the following calculation

$$\begin{aligned}
\|\nabla L(\theta)\|^2 &= \sum_{j=1}^m \langle D_j, w_j\rangle^2 + \sum_{j=1}^m \|D_j\|^2\|w_j\|^2 \\
&\geq \sum_{j\in S_{+,1}}^m \|D_j\|^2\|w_j\|^2 + \sum_{j\in S_{-,1}}^m \|D_j\|^2\|w_j\|^2 \\
&\geq n\min_{j\in S_{+,1}}\|D_j\|^2 \cdot \frac{1}{n}\sum_{j\in S_{+,1}}^m \|w_j\|^2 + n\min_{j\in S_{-,1}}\|D_j\|^2 \cdot \frac{1}{n}\sum_{j\in S_{-,1}}^m \|w_j\|^2 \\
&\geq \left(n\min_{j\in S_{+,1}}\|D_j\|^2 + n\min_{j\in S_{-,1}}\|D_j\|^2\right) \cdot \min\left\{\frac{1}{n}\sum_{j\in S_{-,1}}\|w_j\|^2, \frac{1}{n}\sum_{j\in S_{+,1}}\|w_j\|^2\right\}.
\end{aligned}$$

Furthermore for all $j \in S_{+,1}$,

$$n\|D_j\|^2 = \frac{1}{n}\sum_{k|\langle w_j, x_k\rangle > 0}(h_\theta(x_k) - y_k)^2 \geq \frac{1}{n}\sum_{k|y_k>0}(h_\theta(x_k) - y_k)^2,$$

where the last inequality is implied by the definition of the set $S_{+,1}$. As the same goes for $S_{-,1}$ replacing the sum over the positive (y_k)'s by the negative ones, we have that

$$\begin{aligned}
\left(n\min_{j\in S_{+,1}}\|D_j\|^2 + n\min_{j\in S_{-,1}}\|D_j\|^2\right) &\geq \frac{1}{n}\sum_{k|y_k>0}(h_\theta(x_k) - y_k)^2 + \frac{1}{n}\sum_{k|y_k<0}(h_\theta(x_k) - y_k)^2 \\
&= 2L(\theta).
\end{aligned}$$

This concludes the proof of the claimed inequality. □

Second we show that on a neighbourhood of θ^{t_3} intersected with Θ, the local PL constant can be lower bounded, where we recall Θ is the set of balanced parameters.

Lemma 15. *For λ small enough, we have the following lower bound on the PL constant*

$$\inf_{\theta\in B(\theta^{t_3}, \lambda^{\frac{\varepsilon'}{8}})\cap\Theta} \frac{\|\nabla L(\theta)\|^2}{L(\theta)} \geq \min\{\|D_+\|, \|D_-\|\}. \tag{32}$$

Proof. Indeed, fix any $r > 0$ and take $\theta \in B(\theta^{t_3}, r) \cap \Theta$. Let us denote by $(w_j^{t_3})_j$ the components of the hidden layer of θ^{t_3}. We have

$$\|w_j\|^2 = \|w_j - w_j^{t_3} + w_j^{t_3}\|^2 \geq \|w_j^{t_3}\|^2 - 2\|w_j - w_j^{t_3}\|\|w_j^{t_3}\| \geq \|w_j^{t_3}\|^2 - 2r\|w_j^{t_3}\|.$$

Furthermore,

$$\frac{1}{n}\sum_{j\in S_{+,1}}\|w_j^{t_3}\| \leq \frac{\sqrt{m}}{n}\sqrt{\sum_{j\in S_{+,1}}\|w_j^{t_3}\|^2} \leq \sqrt{\frac{m}{n}}\sqrt{\sum_{j\in S_{+,1}}\|w_j^{t_3}\|^2} \leq \sqrt{\frac{m}{n}}\left(\sqrt{\|D_+\|} + \lambda^{\varepsilon'/2}\right).$$

Hence,

$$\begin{aligned}
\frac{1}{n}\sum_{j\in S_{+,1}}\|w_j\|^2 &\geq \frac{1}{n}\sum_{j\in S_{+,1}}\|w_j^{t_3}\|^2 - 2r\frac{1}{n}\sum_{j\in S_{+,1}}\|w_j^{t_3}\| \\
&\geq \|D_+\| - \lambda^{\varepsilon'} - 2r\sqrt{\frac{m}{n}}\left(\sqrt{\|D_+\|} + \lambda^{\varepsilon'/2}\right),
\end{aligned}$$

and for $r = \lambda^{\varepsilon'/8}$ and λ small enough such that $\lambda^{\varepsilon'} + 2\lambda^{\varepsilon'/8}\sqrt{\frac{m}{n}}\left(\sqrt{\|D_+\|} + \lambda^{\varepsilon'/2}\right) \leq \|D_+\|/2$, we have

$$\frac{1}{n}\sum_{j\in S_{+,1}} \|w_j\|^2 \geq \|D_+\|/2,$$

and the exact same inequality stands for the sum over $j \in S_{-,1}$ with alignment vector D_-. □

Thanks to the lower bound given by Lemma 15, we can now conclude that the gradient flow will not go out $B(\theta^{t_3},\ \lambda^{\frac{\varepsilon'}{8}})$ and will converge exponentially fast to some θ_λ^∞. Then we can take the limit $\lambda \to 0$ to characterise in this low initialisation regime the limit of the gradient flow.

Theorem 2. *For λ small enough,*

- *The gradient flow $(\theta^t)_{t>0}$ converges to some θ_λ^∞ of zero training loss , i.e $L(\theta_\lambda^\infty) = 0$.*
- *There exists θ^* such that we have the following limit:*

$$\lim_{\lambda\to 0}\lim_{t\to\infty} \theta^t = \theta^* \in \operatorname*{argmin}_{L(\theta)=0} \|\theta\|^2, \tag{33}$$

and in its last phase, the gradient flow $(\theta^t)_{t\geq t_3}$ stays in $B(\theta^*,\ \lambda^{\frac{\varepsilon}{240}}) \cap \Theta$ for which the convergence is exponential.

Proof. From Lemma 15, Equation (32), as $\varepsilon' = \varepsilon/30$ we have

$$\inf_{\theta\in B(\theta^{t_3},\ \lambda^{\frac{\varepsilon}{240}})\cap\Theta} \frac{\|\nabla L(\theta)\|^2}{L(\theta)} \geq \min\left\{\|D_+\|, \|D_-\|\right\},$$

and for λ small enough, $L(\theta^{t_3})/\lambda^{\varepsilon/120} \leq \lambda^{\varepsilon/60}/\lambda^{\varepsilon/120} = \lambda^{\varepsilon/120} \leq \min\{\|D_+\|, \|D_-\|\}$. Hence for $r = \lambda^{\varepsilon/240}$,

$$\inf_{\theta\in B(\theta^{t_3},\, r)\cap\Theta} \frac{\|\nabla L(\theta)\|^2}{L(\theta)} \geq \frac{L(\theta^{t_3})}{r^2}.$$

Then, Theorem 2.1 of Chatterjee [2022] applies (at least a benign modification of it restricting the flow to Θ) and this shows that the gradient flow $(\theta^t)_{t\geq t_3}$ stays in $B(\theta^{t_3},\ \lambda^{\frac{\varepsilon}{240}}) \cap \Theta$, and converges towards some θ_λ^∞ of zero loss at exponential speed. Furthermore, from Lemma 12, there exists θ^*, independent of λ, and belonging to $\operatorname{argmin}_{L(\theta)=0}\|\theta\|^2$, such that $\theta^{t_3} \in B(\theta^*,\ \lambda^{\frac{\varepsilon}{31}}) \cap \Theta$. Hence, $(\theta^t)_{t\geq t_3} \in B(\theta^*,\ 2\lambda^{\frac{\varepsilon}{240}}) \cap \Theta$ and finally: $\lim_{\lambda\to 0}\lim_{t\to\infty}\theta^t = \lim_{\lambda\to 0}\theta_\lambda^\infty = \theta^*$. □

B.7 Auxiliary Lemmas

This section stats the auxiliary Lemma 16 used in the proof of Lemma 11.

Lemma 16. *Suppose a non-negative function f verifies for non-negative constants a, b and c*

$$\forall t \in \mathbb{R}_+, \qquad f'(t) \leq -af(t) + b\int_0^t f(s)\mathrm{d}s + \frac{c}{a},$$

then f is bounded as

$$\forall t \in \mathbb{R}_+, \qquad f(t) \leq (f(0)+c)\left(1 + e^{\frac{b}{a}t}\right).$$

Proof. For $g(t) = -af(t) + b\int_0^t f(s)\mathrm{d}s + c$, we have

$$g'(t) = -af(t) + bf(t) \geq -af'(t) \geq -ag(t).$$

By Grönwall lemma, we then have $g(t) \geq -af(0)$ on $\mathbb{R}_+$. It then follows

$$\begin{aligned} f(t) &= -\frac{g(t)}{a} + \frac{b}{a}\int_0^t f(s)\mathrm{d}s + \frac{c}{a} \\ &\leq f(0) + \frac{c}{a} + \frac{b}{a}\int_0^t f(s)\mathrm{d}s. \end{aligned}$$

For $F(t) = \int_0^t f(s)\mathrm{d}s$, Grönwall comparison yields

$$\begin{aligned} F(t) &\leq \frac{a}{b} e^{\frac{b}{a}t} \left(1 - e^{-\frac{b}{a}t}\right) \left(f(0) + \frac{c}{a}\right) \\ &\leq \left(f(0) + \frac{c}{a}\right) \frac{a}{b} e^{\frac{b}{a}t}. \end{aligned}$$

The previous inequality $f(t) \leq f(0) + \frac{c}{a} + \frac{b}{a} F(t)$ yields the lemma. □

C On global solutions of the minimum norm problem

In this section, we study the following optimisation problem:

$$\theta^* \in \underset{L(\theta)=0}{\operatorname{argmin}} \|\theta\|_2^2. \tag{34}$$

An important remark is that the optimisation problem in Equation (34) is not convex because of the constraint set. Hence, the KKT conditions stated below are only necessary conditions: there exist real numbers $(\lambda_k)_{k \in [\![n]\!]}$ such that

$$\theta^* = \sum_{k=1}^{n} \lambda_k \nabla_\theta h_{\theta^*}(x_k) \quad \text{and for all } k \in [\![n]\!], \quad h_{\theta^*}(x_k) = y_k. \tag{35}$$

In terms of (a^*, W^*), it implies

$$a_j^* = \sum_{\langle w_j^*, x_k \rangle > 0} \lambda_k \langle x_k, w_j^* \rangle \tag{36}$$

$$w_j^* = \sum_{\langle w_j^*, x_k \rangle > 0} \lambda_k x_k \, a_j^*. \tag{37}$$

In the orthonormal case, it is possible to solve explicitly the non-convex optimisation problem defined in (34). Recall the definition of balanced networks: $\Theta = \{(a, W) \text{ such that } \forall j \in [\![m]\!], |a_j| = \|w_j\|\}$. For $\theta \in \Theta$, this means that there exists $(\mathsf{s}_j)_{j \in [\![m]\!]} \in \{-1, 1\}^m$ such that $a_j = \mathsf{s}_j \|w_j\|$. Let us call $\mathsf{S}_0^\theta = \{j \in [\![m]\!] \,|\, w_j = 0\}$, $\mathsf{S}_+^\theta = (\mathsf{S}_0^\theta)^c \cap \{j \in [\![m]\!] \,|\, \mathsf{s}_j = +1\}$ and $\mathsf{S}_-^\theta = (\mathsf{S}_0^\theta)^c \cap \{j \in [\![m]\!] \,|\, \mathsf{s}_j = -1\}$, where $(\mathsf{S}_0^\theta)^c := [\![m]\!] \setminus \mathsf{S}_0^\theta$. Note that the family S_0^θ, S_+^θ and S_-^θ form a partition of $[\![m]\!]$.

We have the following proposition.

Proposition 1. *All KKT points of the problem* (34) *that are balanced, i.e.* $\theta \in \Theta$*, are in fact global minimisers with objective value* $2n(\|D_+\| + \|D_-\|)$*. More precisely, we have the following description of the global minimisers of Equation* (34)*:*

$$\begin{aligned} \underset{L(\theta)=0}{\operatorname{argmin}} \|\theta\|_2^2 = \{\theta \in \Theta \textit{ such that } \forall j \in \mathsf{S}_+^\theta, \langle w_j, \quad D_+ \rangle = \|w_j\| \|D_+\| \textit{ and } \sum_{j \in \mathsf{S}_+} \|w_j\|^2 = n\|D_+\| \\ \forall j \in \mathsf{S}_-^\theta, \langle w_j, -D_- \rangle = \|w_j\| \|D_-\| \textit{ and } \sum_{j \in \mathsf{S}_-} \|w_j\|^2 = n\|D_-\|\}. \end{aligned}$$

Proof. First the constraint set is non-empty if $m \geq 2$: one can consider the hidden weights defined as $W = (\sqrt{\frac{n}{\|D_+\|}} D_+, -\sqrt{\frac{n}{\|D_-\|}} D_-, 0, \ldots, 0)$ and the outputs $a = (\sqrt{n\|D_+\|}, -\sqrt{n\|D_-\|}, 0, \ldots, 0)$ so that for all $k \in [\![m]\!]$, $h_{(a,W)}(x_k) = \langle a, \sigma(W x_k) \rangle = y_k$. Hence the minimum is attained in the closed ball centred in the origin and of radius $\|(a, W)\|$. Moreover, the set $C := \{\theta | \forall k \leq n, \; h_\theta(x_k) = y_k\} = \cap_{k \leq n} C_k$, where each C_k is a closed subset as it is the pre-image of y_k by the continuous function: $\theta \mapsto h_\theta(x_k)$. Hence the minimum is to be found in the intersection of a compact and a close subset, that is, a compact set overall. By continuity of the norm to be minimised over this compact set, there exists a global minimum.

Moreover, let (a^*, W^*) be a global minimiser of the optimisation problem. Note that if there exists $j \in [\![m]\!]$ such that $|a_j^*|^2 \neq \|w_j^*\|^2$, then, as for $c = |a_j^*|/\|w_j^*\|$, $|a_j^*|^2/c + c\|w_j^*\|^2 < |a_j^*|^2 + \|w_j^*\|^2$, without changing the constraint set, we have found a strictly better minimum. Hence, for all $j \in [\![m]\!]$ we have: $|a_j^*|^2 = \|w_j^*\|^2$.

Suppose that $w_j^* \neq 0$ for all $j \in [\![m]\!]$. Otherwise the analysis can be restricted to the indices of non-zero w_j^*, without changing neither the objective nor the constraint set. Set $j \in [\![m]\!]$, we have that $w_j^* \in \mathrm{span}(x_1, \dots, x_n)$, otherwise, this would simply add weights in the objective without changing $(h_\theta^*(x_k))_{k\leq n}$. Hence, $w_j^* = \sum_k \langle w_j^*, x_k\rangle x_k$. Then, defining the Lagrange multipliers $(\lambda_k)_{k\leq n}$ as in the KKT condition stated above, we have that, for all $k \leq n$

$$\text{If } \langle w_j^*, x_k\rangle > 0, \text{ then } \lambda_k = \mathsf{s}_j \langle \frac{w_j^*}{\|w_j^*\|}, x_k\rangle,$$
$$\text{else if } \langle w_j^*, x_k\rangle \leq 0, \text{ then } \langle w_j^*, x_k\rangle = 0.$$

On the other side, for all $k \leq n$, $h_{\theta^*}(x_k) = y_k$, and thus

$$\sum_{j|\langle w_j^*, x_k\rangle > 0} \mathsf{s}_j \|w_j^*\| \langle w_j^*, x_k\rangle = y_k \quad \iff \quad \lambda_k \left(\sum_{j|\langle w_j^*, x_k\rangle > 0} \|w_j\|^2 \right) = y_k.$$

From there, we deduce that λ_k and y_k have the same sign and

$$A_k := \{j \in [\![m]\!] \,|\, \langle w_j^*, x_k\rangle > 0\} = \{j \in [\![m]\!] \,|\, \mathsf{s}_j \lambda_k > 0\} = \{j \in [\![m]\!] \,|\, \mathsf{s}_j y_k > 0\} = \begin{cases} \mathsf{S}_+^\theta & \text{if } y_k > 0 \\ \mathsf{S}_-^\theta & \text{if } y_k < 0. \end{cases}$$

Finally define $\mathsf{W}_+^* = \sum_{j\in \mathsf{S}_+^\theta} \|w_j^*\|^2$ and $\mathsf{W}_-^* = \sum_{j\in \mathsf{S}_-^\theta} \|w_j^*\|^2$, we have

$$\text{If } y_k > 0, \text{ then } \lambda_k = (\mathsf{W}_+^*)^{-1} y_k,$$
$$\text{else if } y_k < 0, \text{ then } \lambda_k = (\mathsf{W}_-^*)^{-1} y_k.$$

Finally note that

$$B_j := \{k \in [\![n]\!] \,|\, \langle w_j^*, x_k\rangle > 0\} = \{k \in [\![n]\!] \,|\, \mathsf{s}_j \lambda_k > 0\} = \begin{cases} \{k \in [\![n]\!] \,|\, y_k > 0\} & \text{if } j \in \mathsf{S}_+^\theta \\ \{k \in [\![n]\!] \,|\, y_k < 0\} & \text{if } j \in \mathsf{S}_-^\theta. \end{cases}$$

And the KKT condition (36) reads, e.g. for $j \in \mathsf{S}_+^\theta$,

$$\|w_j^*\| = \sum_{k|y_k>0} \lambda_k \langle w_j^*, x_k\rangle = \sum_{k|y_k>0} \lambda_k^2 \|w_j^*\|.$$

Hence, $\sum_{k|y_k>0} \lambda_k^2 = 1$ and a similar reasoning on S_-^θ gives that $\sum_{k|y_k<0} \lambda_k^2 = 1$. Overall this gives that $\sum_{k|y_k>0} y_k^2 = (\mathsf{W}_+^*)^2$ and $\sum_{k|y_k<0} y_k^2 = (\mathsf{W}_-^*)^2$. And as $\mathsf{S}_+^\theta, \mathsf{S}_-^\theta$ are a partition of $[\![m]\!]$,

$$\|\theta^*\|^2 = 2\mathsf{W}_+^* + 2\mathsf{W}_-^* = 2\sqrt{\sum_{k|y_k>0} y_k^2} + 2\sqrt{\sum_{k|y_k<0} y_k^2} = 2n(\|D_+\| + \|D_-\|).$$

Hence all KKT points that are balanced have the same objective value and hence are global minimisers of the objective. The description of the set of minimisers directly follows from the above proof. □

D On the existence of gradient flows

This section shows that a global solution of the ODE followed by the gradient flow only exists for the choice of subdifferential $\sigma'(0) = 0$. Precisely, the gradient flow follows the following ODE

$$\frac{\mathrm{d}\theta^t}{\mathrm{d}t} \in -\partial L(\theta^t), \tag{38}$$

where ∂f is the Clarke subdifferential of f. The subdifferential of the loss is uniquely defined up to the choice of the subdifferential of the activation function at 0. If we consistently choose a fixed value $\sigma_0 \in [0, 1]$ for the latter, we then have

$$\frac{\mathrm{d}\theta^t}{\mathrm{d}t} = \left(\left(\frac{1}{n} \sum_{k=1}^n (y_k - h_{\theta^t}(x_k)) \sigma(\langle w_j^t, x_k\rangle) \right)_j, \left(\frac{a_j^t}{n} \sum_{k=1}^n (y_k - h_{\theta^t}(x_k)) \sigma'(\langle w_j^t, x_k\rangle) x_k \right)_j \right),$$
$$\text{where} \qquad \sigma'(z) = \begin{cases} 0 \text{ if } z < 0, \\ \sigma_0 \text{ if } z = 0, \\ 1 \text{ if } z > 0. \end{cases} \tag{39}$$

Note that an empirical study of the influence of σ_0 on the dynamics has been conducted by Bertoin et al. [2021]. We have the following proposition, justifying the choice $\sigma'(0) = 0$.

Proposition 2. *The ODE* (39) *admits (at least) one solution on* $[0, \infty)$ *if and only if* $\sigma_0 = 0$.

Proof. First of all note that $\langle w_j^0, x_k \rangle \neq 0$ for all j and k. The Peano theorem then implies there exists a local solution of this ODE, i.e., there exists a time t_0 such that Equation (39) admits a continuous solution θ^t on $[0, t_0)$. Assume now that $t_0 = \infty$. As long as $\langle w_j^0, x_k \rangle \neq 0$ for all j and k, the analysis does not depend on the choice of σ_0. We can thus show similarly to the proof of the first phase that for some time τ and some j, k: $\langle w_j^\tau, x_k \rangle = 0$. Moreover, still following the lines of the first phase, we have some $\delta > 0$ such that $|h_{\theta^t}(x_k)| < \frac{|y_k|}{2}$ on $[0, \tau + \delta]$. Recall that

$$\frac{\mathrm{d}\langle w_j^t, x_k \rangle}{\mathrm{d}t} = -\frac{y_k - h_{\theta^t}(x_k)}{n} a_j^t \sigma'(\langle w_j^t, x_k \rangle).$$

As a consequence, $\langle w_j^t, x_k \rangle$ is monotone on $[0, \tau + \delta]$, and thus decreasing. In particular, $\langle w_j^t, x_k \rangle \leq 0$ on $[\tau, \tau + \delta]$. Moreover, note that $\langle w_j^t, x_k \rangle$ can not become (strictly) negative, as its derivative is 0 as soon as it becomes negative. Indeed, if $t_- \coloneqq \inf\{t \mid \langle w_j^t, x_k \rangle < 0\} < \tau + \delta$, then the scalar product has a zero derivative on $(t_-, \tau + \delta)$ and is thus constant, equal to 0 by continuity, on this interval. So we finally have $\langle w_j^t, x_k \rangle = 0$ on $[\tau, \tau + \delta]$ and θ^t is then a solution of Equation (39) (almost everywhere) only if $\sigma_0 = 0$.

Conversely, let $\sigma_0 = 0$ now. Assume we have a maximal solution of the ODE in finite time, i.e., let $t_0 \in \mathbb{R}_+$ and θ_*^t defined on $[0, t_0)$ such that θ_*^t verifies Equation (39) almost everywhere on $[0, t_0)$ and there exists no $\delta > 0$ and $\widetilde{\theta}$ such that $\widetilde{\theta}^0 = \theta_*^0$ and $\widetilde{\theta}^t$ verifies Equation (39) almost everywhere on $[0, t_0 + \delta)$.

By definition of the gradient flow, the training loss is decreasing. As a consequence, $|y_k - h_{\theta_*^t}(x_k)|$ is uniformly bounded in time. The ODE then leads for some constant $L > 0$ to $\frac{\mathrm{d}\|\theta_*^t\|}{\mathrm{d}t} \leq L\|\theta_*^t\|$. By Grönwall argument, $\|\theta_*^t\|$ is bounded on $[0, t_0)$. This implies that θ_*^t is Lipschitz with time on $[0, t_0)$. In particular, θ_*^t admits a limit in t_0. Now consider the alternative ODE

$$\begin{gathered}\frac{\mathrm{d}\theta^t}{\mathrm{d}t} = \left(\left(\frac{1}{n} \sum_{k=1}^n (y_k - h_{\theta^t}(x_k)) \sigma(\langle w_j^t, x_k \rangle) \right)_j, \left(\frac{a_j^t}{n} \sum_{k=1}^n (y_k - h_{\theta^t}(x_k)) \mathbb{1}_{\langle w_{*,j}^{t_0}, x_k \rangle > 0} x_k \right)_j \right), \\ \theta^{t_0} = \theta_*^{t_0}.\end{gathered} \tag{40}$$

We replaced $\mathbb{1}_{\langle w_j^t, x_k \rangle > 0}$ in the original ODE by $\mathbb{1}_{\langle w_{*,j}^{t_0}, x_k \rangle > 0}$, which makes it Lipschitz in θ^t. Cauchy-Lipschitz theorem then implies that Equation (40) admits a (unique) local solution $\widetilde{\theta}^t$ on $[t_0 - \delta, t_0 + \delta]$ for some $\delta > 0$. Moreover, $\widetilde{\theta}^t = 0$ on the whole interval if $\langle w_{*,j}^{t_0}, x_k \rangle = 0$. By continuity, we can thus choose $\delta > 0$ small enough so that $\mathbb{1}_{\langle w_{*,j}^{t_0}, x_k \rangle > 0} = \mathbb{1}_{\langle \widetilde{w}_j^t, x_k \rangle > 0}$ on $[t_0, t_0 + \delta]$. θ_* can then be extended by $\widetilde{\theta}$ on $[t_0, t_0 + \delta]$ and still verifies Equation (39) on $(t_0, t_0 + \delta]$, contradicting its maximality. As a consequence, there exists a solution of the ODE on $\mathbb{R}_+$. □