Mildly Overparameterized ReLU Networks on Orthogonal Data: Incremental Learning and Implicit Bias

James Town
University of Warwick
james.p.town@warwick.ac.uk

Etienne Boursier
INRIA & LMO, Université Paris-Saclay
etienne.boursier@inria.fr

Ben Lewis
University of Warwick
ben.lewis@warwick.ac.uk

Matthias Englert
University of Warwick
m.englert@warwick.ac.uk

Ranko Lazić
University of Warwick
r.s.lazic@warwick.ac.uk

Abstract

The successful training of neural networks hinges on the use of first order optimization methods, yet the theoretical characterization of these methods remains incomplete. This is especially true in settings with mild overparameterization. In this work, we study the gradient flow dynamics of two-layer ReLU networks from small initialization with orthogonal training data. We prove the flow converges to a saddle-to-saddle jump process as the initialization scale tends to zero, revealing an incremental learning phenomenon in which a new neuron activates at each saddle. This analysis recovers the known result of Dana et al. [2025] that the network interpolates the training data with high probability as soon as $m \gtrsim \log(n)$, where m is the network width and n is the number of training samples. This incremental process characterization also allows us to derive a novel implicit bias result: the learned interpolator has a squared ℓ_2-norm scaling as $\sqrt{n}$, which is within a constant factor of the minimal ℓ_2-norm interpolator. More broadly, our work provides the first rigorous proof of an incremental learning process for ReLU networks, whilst suggesting mildly overparameterized networks can converge to interpolating solutions whose complexity is of the same order as that of the optimal interpolator.

1 Introduction

Despite the remarkable empirical successes of deep learning [Lipman et al., 2023, Zagoruyko and Komodakis, 2016, Cheng et al., 2016], a precise theoretical understanding of the optimization dynamics of neural networks trained by first-order methods remains elusive. The *Neural Tangent Kernel* (NTK) framework offered an initial answer, establishing convergence to global minima for overparameterized networks [Jacot et al., 2018, Du et al., 2019, Arora et al., 2019b]. However, NTK operates under large initialization assumptions that cause network features to remain nearly unchanged during training; a regime aptly termed *lazy* by Chizat et al. [2019], and now understood to be unrepresentative of practice [Fort et al., 2020].

In the practically relevant *small initialization* regime, training induces significant changes in network features, a phenomenon known as *feature learning*. Theoretical analysis of this regime is considerably

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

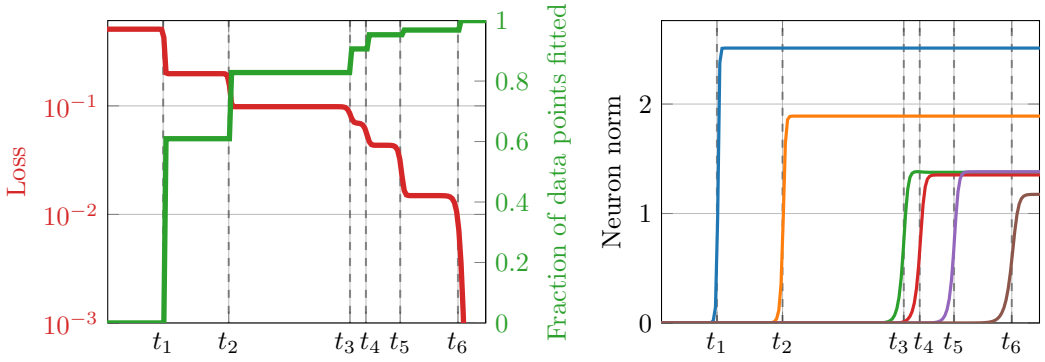


Figure 1: We train a two-layer ReLU network of six neurons on 64 orthonormal data points in 64 dimensions. The initialization scale is small (experimental details in Appendix A.1). Left: During training, the loss (red) shows long periods of little change with sudden drops in between. Those drops correspond to an additional set of data points being fitted (green). The drops in the loss happen at the times independently predicted by Algorithm 1, which are indicated by the dashed lines. Right: The norms of the hidden layer of the six neurons during the same training.

more challenging, as the resulting dynamics are highly non-convex. General convergence results exist for linear activation architectures [Arora et al., 2018, Woodworth et al., 2020] and, in the infinite-width limit, for nonlinear networks [Chizat and Bach, 2018, Wojtowytsch, 2020]. These results, however, require either smooth activations or infinite data [Holzmüller and Steinwart, 2022, Boursier and Flammarion, 2025a], and thus do not apply to the canonical case of two-layer ReLU networks.

Beyond convergence, a central motivation for studying the small initialization regime is its *implicit bias*: gradient flow is known to favor sparse or low-complexity solutions that generalize well. For classification, two-layer networks trained to interpolation have been shown to converge to min-norm max-margin estimators [Lyu and Li, 2020, Chizat and Bach, 2020, Ji and Telgarsky, 2020], which coincide with sparse, narrow networks [Safran et al., 2022, Boursier and Flammarion, 2023]. Regression is however arguably more representative of practical training dynamics, as characterizing the implicit bias requires describing the entire trajectory rather than only its asymptotic endpoint after interpolation is achieved. In the regression setting, diagonal linear networks favor sparse estimators [Woodworth et al., 2020], while fully connected linear networks favor low-rank solutions [Saxe et al., 2014, Arora et al., 2019a]. For two-layer ReLU networks, implicit bias in regression remains largely open, and constitutes the focus of this work.

Progress has been made by restricting to structured data settings. Among these, orthogonal data has emerged as a natural and tractable setting: Boursier et al. [2022] established convergence of gradient flow to a minimal ℓ_2-norm interpolator, but under an initialization condition requiring exponential overparameterization in the number of training samples. Dana et al. [2025] later showed that mild overparameterization suffices to guarantee convergence to an interpolating solution. However, their analysis leaves the norm of the reached interpolator uncharacterized, and understanding this implicit bias in the mildly overparameterized regime remained open until the present work.

Contributions. We prove a precise characterization of the implicit bias of gradient flow when training two-layer ReLU networks on orthogonal data in the mildly overparameterized regime, via a complete analysis of the training trajectory. Interestingly, our analysis reveals an *incremental learning* phenomenon, illustrated in Figure 1: as initialization scale vanishes, the loss alternates between long idle plateaus and abrupt decreases, each corresponding to the activation of a new neuron. While incremental learning has been characterized for linear networks, our work provides the first such characterization for networks with nonlinear activations. An informal statement of our main result follows, where m refers to the width of the network and n the number of training samples.

Theorem 1 (Informal)**.** *In the regression setting with orthogonal training data, as initialization scale vanishes, and with high probability over the network initialization, the gradient flow trajectory over a two-layer ReLU network with* $\log(n) \lesssim m \ll \exp\left(n\log(2)/2\right)$ *converges to a piecewise-constant process: the iterates move successively from one saddle of the training loss to another, with each transition corresponding to the creation of a new non-zero neuron. The sequence of visited saddles and transition times can be computed via Algorithm 1 without any actual training of the network.*

Building on this algorithmic description of the limit dynamics, we derive an upper bound on the parameter ℓ_2-norm of the solution found by gradient flow, showing it is within a constant factor of the minimal ℓ_2-norm interpolator.

1.1 Related work

Incremental Learning. Arpit et al. [2017], Nakkiran et al. [2019] empirically demonstrated that standard optimization of deep learning models follows a *simplicity bias*, whereby learned representations grow increasingly complex over the course of training. Separately, several works have shown that this complexity does not accumulate continuously, but rather incrementally: new *data patterns* are absorbed one at a time [Rahaman et al., 2019, Charton, 2024], giving rise to a characteristic training curve marked by long stagnation plateaus and sharp drops in the loss. The resulting training dynamics can thus be understood as a succession of jumps between saddle points of the loss landscape — a phenomenon sometimes referred to as *saddle-to-saddle dynamics* [Jacot et al., 2021].

Theoretically characterizing this incremental behavior is challenging because it is specific to the small initialization regime, where networks operate in the rich regime and the resulting dynamics are highly non-convex. For two-layer networks, Kunin et al. [2025] offered high-level intuitions for this regime: during each plateau, a dormant neuron is activated in a manner shaped by the local geometry near the saddle point, after which the activated neurons undergo rapid updates as training transitions to the next plateau. However, these intuitions fall short of a complete and rigorous characterization, which has so far been confined to specific linear activation models: for diagonal linear networks, Berthier [2023], Pesme and Flammarion [2023] precisely characterized the incremental learning dynamics, showing that the estimator's support grows one coordinate at a time; for linear networks and matrix sensing, Gidel et al. [2019], Jin et al. [2023] established, under favorable data or initialization conditions, that the learned matrix undergoes incremental rank increases throughout training.

Our work advances this line of research by providing the first theoretical characterization of incremental learning under gradient flow for networks with non-linear activations, specifically two-layer ReLU networks trained on orthogonal data. In a related vein, Abbe et al. [2023] also studied incremental learning in two-layer networks with non-linear activations; however, their analysis is tailored to a specific activation function and optimization algorithm that depart from standard practice.

Training two-layer ReLU networks on orthogonal data. Since a theoretical analysis of two-layer ReLU network training remains out of reach for general data distributions, many works have focused on specific toy settings, such as XOR data [Glasgow, 2024], positively correlated data [Wang and Ma, 2023, Chistikov et al., 2023, Min et al., 2024], or linearly separable data [Phuong and Lampert, 2021, Lyu et al., 2021]. Among these, we focus on the orthogonal data setting, first studied by Boursier et al. [2022], who showed that under small initialization, the parameters converge to a minimal ℓ_2-norm interpolator. However, their analysis requires an initialization condition that is only satisfied in a heavily overparameterized regime, where the number of neurons m is exponential in the number of training points n. Dana et al. [2025] later relaxed this assumption, showing that mild overparameterization $m \gtrsim \log(n)$ suffices to guarantee convergence to a global minimum of the training loss.

Yet the question of implicit bias in the mildly overparameterized regime remained open, as Dana et al. [2025] do not characterize the nature of the resulting estimator. Specifically, one may ask: *does the network still converge to the minimal ℓ_2-norm interpolator in the mildly overparameterized regime?* Boursier et al. [2026] answered this negatively, showing that the minimal ℓ_2-norm interpolator is not reached with high probability as soon as $m \lesssim 2^n$. Taken together, these results leave a gap precisely in the mildly overparameterized regime $\log(n) \lesssim m \lesssim 2^n$, which our work fills by characterizing the ℓ_2-norm of the interpolator obtained in this setting.

More parameters lead to simpler models. The generalization of overparameterized models remains an active area of research in deep learning theory. A phenomenon of particular interest is *double descent* [Belkin et al., 2019, Nakkiran et al., 2020], whereby overparameterization beyond the interpolation threshold leads to smaller test error. A likely driver of this behavior is that, as the number of parameters grows, the function represented at convergence becomes simpler [Wilson, 2025], exhibiting for example smaller norm [Belkin et al., 2019], fewer effective parameters [Maddox et al., 2020], or lower Kolmogorov complexity [Goldblum et al., 2024]. Numerous theoretical analyses of this phenomenon have been proposed in recent years, but remain confined to linear regression on random features [Liao et al., 2020, Belkin et al., 2020, Hastie et al., 2022, Mei and Montanari, 2022, Bach, 2024]. Our work establishes such a tendency for two-layer ReLU networks

trained on orthogonal data: mildly overparameterized networks converge to interpolators with slightly larger parameter ℓ_2-norm than those found in the heavily overparameterized regime.

2 Setting

We consider a two-layer neural network with the ReLU activation, whose output is given by

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

where m is the network width, $\sigma(z) \coloneqq \max\{0, z\}$ is the ReLU function, and $\theta = (a, W)$ are the network parameters, which consist of output weights $a \in \mathbb{R}^m$ and hidden neurons $W^\top = (w_j)_{j=1}^m \in \mathbb{R}^{d\times m}$. We consider a training dataset $(x_i, y_i)_{i=1}^n$, which consists of inputs $x_i \in \mathbb{R}^d$ and outputs $y_i \in \mathbb{R}$, satisfying the following assumption.

Assumption 1. *The training inputs are such that $x_i^\top x_{i'} = \mathbb{1}\,(i = i')$ for all $i, i' \in [n]$.*

Although restrictive, the orthogonality assumption is a useful test bed for analyzing training dynamics, which can even sometimes be relaxed to general random high dimensional data [Dana et al., 2025]. On the other hand, the norm constraint on the inputs implied by Assumption 1 is solely used for the sake of simplicity and could easily be relaxed. Training seeks to minimize the mean squared error loss over the training data,

$$\mathcal{L}(\theta) \coloneqq \frac{1}{2n} \sum_{i=1}^n (h_\theta(x_i) - y_i)^2.$$

For analytical purposes, we consider optimizing this loss function by gradient flow, which approximates (stochastic) gradient descent as the learning rate vanishes to 0:

$$\dot\theta(t) \in -\partial\mathcal{L}(\theta(t)) \text{ for almost all } t \geq 0. \tag{1}$$

Here, ∂ denotes the Clarke [1975] subdifferential, which allows to define the gradient flow beyond differentiable functions. We direct the reader to Appendix B for further details on the Clarke subdifferential. For an appropriate class of functions, the Clarke subdifferential is known to admit a chain rule [Bolte and Pauwels, 2020, 2021]. Applying this to Equation (1), we have the subdifferential of w_j as $\big(\partial\mathcal{L}(\theta)\big)_{w_j} \subseteq -\frac{1}{n}\sum_v (h(x_v) - y_v)\partial\sigma\big(w_j^\top x_v\big) a_j x_v$. Taking the inner product with x_i and using Assumption 1 implies that

$$\dot w_j^\top(t)\, x_i \in \frac{1}{n}(y_i - h_{\theta(t)}(x_i))\, a_j(t)\, \partial\sigma(w_j^\top(t)\, x_i).$$

Hence, for all $i \in [n]$, all $j \in [m]$, and almost all $t \geq 0$, if a hidden neuron component $w_j^\top(t)\, x_i$ is negative then it must be constant for all $t' \geq t$. Empirically, the same is true with "negative" replaced by "non-positive",[1] so Assumption 2 below restricts our attention to realistic training trajectories, which have that property.

Assumption 2. *For all $i \in [n]$ and all $j \in [m]$, if $w_j^\top(t)\, x_i = 0$ for some $t \geq 0$, then $w_j^\top(t')\, x_i = 0$ for all $t' > t$.*

Note that whilst it is guaranteed that a solution of Equation (1) exists locally, it may not be unique. With Assumption 2, along every solution of the subdifferential inclusion Equation (1), the left-hand side $\dot\theta(t)$ exists — possibly except at the finitely many time points that are the minima of non-empty sets $\{t \geq 0 \mid w_j^\top(t)\, x_i = 0\}$ — and is equal to the element of the right-hand side $-\partial L(\theta(t))$ that is obtained by using only 0 as the subderivative of the ReLU function σ at 0. Hence for simplicity of presentation, we fix the latter choice unless specified otherwise, meaning in the remainder of the paper we work with $\mathbb{1}\,(z > 0)$ instead of $\partial\sigma(z)$ and we may use the gradient notation to mean the subgradient obtained with that choice. Boursier et al. [2022] also proved that any other constant definition of the subgradient fails to yield global solutions of Equation (1) in presence of orthogonal data. Finally, we note that Bertoin et al. [2021] suggest that this fixed choice tends to be most efficient in practice. We consider balanced initialization of the parameters, which is specified by an initialization scale α.

Assumption 3. *For all $j \in [m]$ we have $a_j(0) = \alpha\, s_j$ and $w_j(0) = \alpha\, u_j$, where $s_j \overset{\text{iid}}{\sim} \mathcal{U}(\{\pm 1\})$ and $u_j \overset{\text{iid}}{\sim} \mathcal{U}(\mathbb{S}^{d-1})$.*

[1] E.g., in PyTorch, the derivative of the ReLU non-linearity at 0 is computed as 0.

Whilst the assumption of strictly balanced initialization is partly to simplify the analysis, the work of Azulay et al. [2021] suggests that balanced initializations encourage bias towards the rich regime. Azulay et al. [2021] also note that the commonly used Xavier initialization [Glorot and Bengio, 2010] is approximately balanced. The balanced initialization simplifies the analysis of the training dynamics, as this weight balancedness will be preserved along training, thanks to the following result.

Proposition 1 (Du et al. 2018, Theorem 2.1)**.** *For all $j \in [m]$ and all $t \geq 0$, we have $|a_j(t)| = \|w_j(t)\|$ and $\operatorname{sgn}(a_j(t)) = s_j$.*

Proposition 1 also implies that the output layer signs remain constant during training. This allows us to treat separately the data associated with positive or negative labels, as well as the neurons with positive or negative output signs. To this end, we denote for $s \in \{\pm 1\}$, the set $I_s \coloneqq \{i \in [n] \mid \operatorname{sgn}(y_i) = s\}$ of labels with sign s.

Neuron dynamics. Denote the normalized hidden-layer weights by $\overline{w}_j := w_j/\|w_j\|$ (in general we use bars to denote normalized vectors and $\|\cdot\|$ to denote the ℓ_2-norm) and define the dynamical vector

$$\mathcal{D}_j = -\frac{1}{n} \sum_{i | w_j^\top x_i > 0} (h_\theta(x_i) - y_i) x_i.$$

Equation (1), Proposition 1, and Assumption 2 imply the radial and tangential dynamics of neurons are given for almost all $t \geq 0$ by

$$\dot{a}_j = \mathcal{D}_j^\top w_j \quad \text{and} \quad \dot{\overline{w}}_j = s_j \left(\mathrm{I}_d - \overline{w}_j \overline{w}_j^\top\right) \mathcal{D}_j.$$

Acceleration of time. Our goal is to study the limiting dynamics of the parameters' trajectory as the initialization scale α goes to 0. When $\alpha \to 0$, escaping saddle points of the trajectory — starting with the initial one where all parameters are zero — takes an arbitrarily long time. Similarly to Berthier [2023], Pesme and Flammarion [2023], obtaining non-trivial limit dynamics requires a proper rescaling of time, which is given by the time acceleration $t \mapsto \log(1/\alpha)t$. We denote the accelerated parameters by $\widetilde{\theta}^\alpha(t) = \theta(\log(1/\alpha)t)$. In the following, we omit the superscript as the acceleration by a factor of $\log(1/\alpha)$ is implicit in the tilde notation. The flow defining this rescaled process then satisfies $\dot{\widetilde{\theta}}(t) \in -\log(1/\alpha)\partial\mathcal{L}(\widetilde{\theta}(t))$ for almost all $t \geq 0$.

3 Limit process

In this section, we provide intuitions for deriving the limit process of the training trajectory as $\alpha \to 0$. We here only provide high level insights, and proving the exact convergence results claimed here requires much more detailed and intricate analyses that are given in the complete proof of Theorem 2.

We build the saddles visited by the limit process and the corresponding jump times recursively. Assume that $\widetilde{\theta}(t) \overset{\alpha \to 0}{\longrightarrow} \theta^\circ(t)$ for almost every $t \in [0, t_k]$, where t_k is a jump time, and that the limit process has reached the saddle point $\theta^{(k)}$ after the k-th jump, i.e., $\theta^\circ(t_k) = \theta^{(k)}$. The limit process $\theta^\circ(t)$ will then be constant, equal to the saddle $\theta^{(k)}$ on an interval $[t_k, t_{k+1})$, and will jump to a new saddle $\theta^{(k+1)}$ at a jump time t_{k+1} to be determined. The initial saddle point is given by $\theta^{(0)} = \mathbf{0}$. We will show recursively that the limit process jumps from one saddle point to another by modifying a single neuron (w_j°, a_j°), which is zero before jump time t_{k+1}, and will reach its final, non-zero value after the jump. Also, all the training points along which the neuron j is active will be fitted after the jump. During that jump, the values of all other neurons remain unchanged. Consequently, the saddle point $\theta^{(k)}$ visited by the limit process includes two types of neurons:

- unfitted neurons, which are equal to zero, i.e., $(w_j^\circ, a_j^\circ) = \mathbf{0}$;
- fitted neurons, which interpolate the training points along which they are active: $a_j \sigma\left(w_j^{\circ\top} x_i\right) = y_i$. These neurons will not move further during the remainder of training.

Describing the limit process then calls for describing how and when it jumps from a saddle point $\theta^{(k)}$ to the next one $\theta^{(k+1)}$. For that and following the heuristics developed by Maennel et al. [2018], Kunin et al. [2025], we will decompose the tangential and radial dynamics of the parameters, using

the reparameterization $(\widetilde{\overline{w}}_j, \widetilde{\ell}_j)$ such that $\widetilde{w}_j = \alpha^{-\widetilde{\ell}_j}\widetilde{\overline{w}}_j$. This reparameterization captures the right quantities: $\widetilde{\ell}_j$ governs the scaling of a neuron's norm, while its normalized counterpart $\widetilde{\overline{w}}_j$ tracks the tangential movement of the neuron, which is key to understanding the dynamics in the vicinity of each saddle. The dynamics satisfied by these parameters can then be written as follows:

$$\frac{\mathrm{d}}{\mathrm{d}t}\widetilde{\overline{w}}_j = \log(1/\alpha) s_j \left(\mathrm{I}_d - \widetilde{\overline{w}}_j \widetilde{\overline{w}}_j^\top\right) \mathcal{D}_j; \qquad \frac{\mathrm{d}}{\mathrm{d}t}\widetilde{\ell}_j = s_j \widetilde{\overline{w}}_j^\top \mathcal{D}_j.$$

Around a saddle point $\theta^{(k)}$, note that $\mathcal{D}_j \approx 0$ for already fitted neurons, which thus do not move anymore. On the other hand, for unfitted neurons

$$\mathcal{D}_j \approx D_j^{(k)} = \frac{1}{n} \sum_{i \in S_U^{(k)}} \mathbb{1}\left(w_j(0)^\top x_i > 0\right) \mathbb{1}\left(\operatorname{sgn} y_i = s_j\right) y_i x_i,$$

where $S_U^{(k)}$ is the set of training points that are not yet fitted by the network (a rigorous definition of all quantities is given in Algorithm 1 below). For unfitted neurons, the tangential movement is then approximately Riemannian gradient ascent/descent on the sphere with objective $\widetilde{\overline{w}}_j^\top D_j^{(k)}$, similarly to observations made by Boursier et al. [2022], Min et al. [2024], Kumar and Haupt [2024]. Because of its $\log(1/\alpha)$ scaling, this tangential movement thus occurs instantly as $\alpha \to 0$, such that for $t \in (t_k, t_{k+1})$

$$\widetilde{\overline{w}}_j(t) \xrightarrow{\alpha \to 0} s_j \overline{D}_j^{(k)}.$$

Since we have $\widetilde{\overline{w}}_j(t) \approx s_j \overline{D}_j^{(k)}$, the time derivative of the norm exponent then satisfies on (t_k, t_{k+1}):

$$\frac{\mathrm{d}}{\mathrm{d}t}\widetilde{\ell}_j \approx \left\| D_j^{(k)} \right\|_2.$$

Consequently, $\widetilde{\ell}_j$ converges to a piecewise affine process ℓ_j° that starts at -1, owing to our initialization choice, and is defined recursively by:

$$\ell_j^\circ(0) = -1; \qquad \ell_j^\circ(t) = \ell_j^\circ(t_k) + (t - t_k)\left\| D_j^{(k)} \right\|_2 \quad \text{for } t \in [t_k, t_{k+1}].$$

Moreover, the next jump happens as soon as one unfitted neuron becomes non-zero (in the limit $\alpha \to 0$). With our reparameterization, this occurs as soon as $\ell_j^\circ(t) = 0$ for some unfitted neuron, for which furthermore $\ell_j^\circ(t) = 0$ for the remainder of training. The jump time t_{k+1} can then be deduced as the first time for which the exponent ℓ_j° of some unfitted neuron becomes zero, and the corresponding neuron is the one to be updated at jump time t_{k+1}. In other words, we have the following:

- $j_\star^{(k)} = \arg\max_j \ell_j^\circ(t_k) / \left\| D_j^{(k)} \right\|_2$ is the unfitted neuron that will grow in norm;
- $t_{k+1} = t_k - \ell_{j_\star^{(k)}}^\circ(t_k) / \left\| D_{j_\star^{(k)}}^{(k)} \right\|_2$ is the jump time at which $w_{j_\star^{(k)}}^\circ$ becomes non-zero.

At jump time t_{k+1}, the process then leaves the saddle $\theta^{(k)}$, and due to the time acceleration $\log(1/\alpha)$, it instantaneously reaches a new critical point of the training loss. Moreover, during this instantaneous jump, only the neuron $(w_{j_\star^{(k)}}^\circ, a_{j_\star^{(k)}}^\circ)$ of the limit process undergoes a change: it jumps from $\mathbf{0}$ to the value minimizing the new training loss over this single neuron, which is given by

$$w_{j_\star^{(k)}}^\circ(t) = s_{j_\star^{(k)}} \frac{n D_{j_\star^{(k)}}^{(k)}}{\sqrt{\left\| n D_{j_\star^{(k)}}^{(k)} \right\|_2}}; \quad a_{j_\star^{(k)}}^\circ(t) = s_{j_\star^{(k)}} \sqrt{\left\| n D_{j_\star^{(k)}}^{(k)} \right\|_2} \qquad \text{for } t \geq t_{k+1}.$$

Thanks to these high level insights, one can define formally the limit process $(\theta^\circ(t))_{t \geq 0}$ in Algorithm 1 below. Importantly, the `while` loop in Algorithm 1 eventually stops after a finite number of iterations p, such that $S_U^{(p)} \cap S_j = \emptyset$ for all $j \in N_U^{(p)}$. In the following, Assumption 4 will guarantee that we actually stop when $S_U^{(p)} = \emptyset$, which corresponds to the training loss being zero.

Algorithm 1 Construction of the limit process.

Input: data $(x_i, y_i)_{i\in[n]}$ and initial weights $(w_j(0), a_j(0))_{j\in[m]}$.

▷ *Initialization*

$p \leftarrow 0$; $t_0 = 0$; $N_U^{(0)} = [m]$; $S_U^{(0)} = [n]$

$S_j = \left\{i \in [n] \mid w_j(0)^\top x_i > 0 \text{ and } \operatorname{sgn}(y_i) = \operatorname{sgn}(a_j(0))\right\}$; and $\ell_j^\circ(0) = -1$ for any $j \in [m]$

▷ *Recursive definition of the jumps*

while $\exists j \in N_U^{(p)}, S_U^{(p)} \cap S_j \neq \emptyset$, **do**

for any $j \in [m]$, $\quad D_j^{(p)} = \begin{cases} \frac{1}{n}\sum_{i\in S_U^{(p)}\cap S_j} y_i x_i & \text{if } j \in N_U^{(p)} \\ D_j^{(p-1)} & \text{otherwise} \end{cases}$

Let $j_\star^{(p)} \in \arg\max_{j\in N_U^{(p)}} \ell_j^\circ(t_p) / \left\|D_j^{(p)}\right\|_2$ *# neuron to activate*

$t_{p+1} = t_p - \ell_{j_\star^{(p)}}^\circ(t_p) / \left\|D_{j_\star^{(p)}}^{(p)}\right\|_2$ *# jump time*

$\ell_j^\circ(t_{p+1}) = \min\left(0, \ell_j^\circ(t_p) + (t_{p+1} - t_p)\left\|D_j^{(p)}\right\|_2\right)$ for any $j \in [m]$ *# norm exponent*

$N_U^{(p+1)} = N_U^{(p)} \setminus \{j_\star^{(p)}\}$ and $S_U^{(p+1)} = S_U^{(p)} \setminus S_{j_\star^{(p)}}$ *# unfitted neurons and data*

$p \leftarrow p + 1$;

end while

▷ *Definition of limit process*

Let $t_{p+1} = +\infty$ by convention

For any $k \in [p]$ and $t \in [t_k, t_{k+1})$, define $\theta^\circ(t) = (w_j^\circ(t), a_j^\circ(t))_{j\in[m]}$ as

$$\text{for any } j \in N_U^{(k)}, \qquad (w_j^\circ(t), a_j^\circ(t)) = \mathbf{0};$$

$$\text{for any } j \notin N_U^{(k)}, \qquad w_j^\circ(t) = s_j \frac{nD_j^{(k)}}{\sqrt{\left\|nD_j^{(k)}\right\|_2}} \quad \text{and} \quad a_j^\circ(t) = s_j\sqrt{\left\|nD_j^{(k)}\right\|_2}.$$

Output: limit process $(\theta^\circ(t))_{t\geq 0}$ and jump times $(t_0, \ldots, t_{p+1})$.

4 Saddle-to-saddle dynamics

In this section, we present our main result. We first require some assumptions to ensure that the network interpolates at convergence, and avoid degenerate cases in the definition of the limit process.

Assumption 4. *Let $A \in \{0,1\}^{n\times m}$ be the matrix such that $A_{i,j} = \mathbb{1}\left(w_j^\top(0)\, x_i > 0 \text{ and } i \in I_{s_j}\right)$.*

(i) Each row of the mask matrix A is non-zero, each column is non-zero, and its columns are pairwise distinct.

(ii) The training outputs $(y_i)_{i=1}^n$ are such that they are non-zero, and in each iteration defining the limit process, the $\arg\max$ *in Algorithm 1 Line 8 is unique.*

We believe our main result could be extended to the case that columns are not pairwise distinct, at the expense of a more refined analysis, by merging neurons into groups with matching initial activation pattern in a similar manner to Boursier et al. [2022]: neurons which share activation patterns align and grow in the same directions throughout training, and the sum of their squared norms is equal to the squared norm achieved by a single neuron with this activation pattern. In this way, neurons sharing activation patterns form rank-1 sub-networks, so Algorithm 1 may be applied considering an aggregated collection of neurons with matching initial activation pattern as one neuron.

The requirement that each row be non-zero ensures loss minimization. The requirement that each column be non-zero ensures that for each neuron, there exists a direction in which it can grow at initialization to reduce the loss, otherwise such neurons remain small throughout training and can be ignored. Proposition 2 below states that Assumption 4 holds with high probability in the mild overparameterization regime $\log(n) \lesssim m \ll \exp(n\log(2)/2)$.

Proposition 2. *Let $n_+ = |I_+|$, $n_- = |I_-|$ and consider the initialization scheme of Assumption 3.*

(i) Assumption 4 (i) holds with probability at least $1 - n\left(\frac{3}{4}\right)^m - \frac{m(m+3)}{2}\left(\frac{1}{2}\right)^{\min(n_+,n_-)+1}$.

(ii) When Assumption 4 (i) holds, the set of tuples $(y_i)_{i=1}^n$ excluded by Assumption 4 (ii) has Lebesgue measure zero in $\mathbb{R}^n$.

The following result rigorously formulates the sense in which the gradient flow defined by Equation (1) converges, as the initialization scale tends to zero, to the limit process θ°.

Theorem 2. *Under Assumptions 1, 2, 3 and 4, as the initialization scale α tends to 0, the accelerated flow $\widetilde{\theta}^\alpha$ converges to θ°, uniformly on every closed subset of $\mathbb{R}_{\geq 0} \setminus \{t_1, t_2, \dots, t_p\}$, where θ° and $(t_1, \dots, t_p)$ are respectively the limit process and jump times defined in Algorithm 1.*

Theorem 2 shows that as the initialization scale decreases, the trajectory of the parameters approximates a jump process between different saddles of the loss function. As each saddle increases the rank of the network by one from the previous saddle, this result shows an example of incremental feature learning. This provides the first complete theoretical characterization of a saddle-to-saddle process in a neural network with realistic assumptions on the architecture and overparameterization. Figure 1 illustrates this convergence towards the incremental limit process θ° and additional experiments are provided in Appendix A

Instead of Assumption 4, the analysis of Boursier et al. [2022] relies on the assumption that there exist at least two neurons j_+, j_- such that $A_{i,j_+} = 1$ for all $i \in I_+$ and $A_{i,j_-} = 1$ for all $i \in I_-$. In this case, Algorithm 1 reduces to just two jumps: neuron j_+ grows to fit all positive labels, and neuron j_- grows to fit all negative labels. Moreover, as explained in Appendix D.1, these two jumps are *non-interfering*: the neurons fitting positive and negative labels evolve independently throughout training. Under mild overparameterization, however, the process is considerably more intricate: fitting positive (resp. negative) labels requires multiple neurons and hence multiple jumps, which may interact with one another.

Our proof therefore requires a substantially finer analysis of the dynamics than that of Boursier et al. [2022]. The two main additional difficulties are the following. First, the norm growth of a given neuron exhibits memory across multiple saddles, which is controlled by the term ℓ_j°. Second, when a new neuron is activated, it may still be positively correlated with already fitted samples x_i, and one must ensure that its activation does not cause the loss on these samples to increase significantly.

Sketch of Proof. The result is proven by an inductive argument on the number of saddles traversed by the flow. The schematic for the proof is depicted in Figure 2 and consists of splitting the dynamics into phases around the jump from saddle k to saddle $k+1$, where ε and ε' are parameters of the proof which may be selected arbitrarily small, giving the desired convergence.

- Slow growth phase: For $\kappa \in (t_k, t_{k+1})$, τ_1 is the first time after κ that $\left\|\widetilde{w}_{j_\star^{(k)}}\right\|_2 = \alpha^\varepsilon$. We find that for $t \in [\kappa, \tau_1]$ all neurons remain aligned in the directions $D_j^{(k)}$. Hence, $\widetilde{\ell}_j$ increases towards 0 for all $j \in N_U^{(k)}$, but $\widetilde{\theta}$ does not change significantly and remains near saddle k until any $\kappa' \in (\kappa, \tau_1)$.
- Rapid fitting phase: τ_3 is the first time that $\left\|\widetilde{w}_{j_\star^{(k)}}\right\|_2^2 = n\left\|D^{(k)}_{j_\star^{(k)}}\right\|_2 - \alpha^{\varepsilon'/4}$ for a small enough ε'. We show $\tau_3 - \tau_1 = \mathcal{O}(\varepsilon, \varepsilon')$ by arguing $\widetilde{w}_{j_\star^{(k)}}$ continues to grow at an exponential rate $\left\|D^{(k)}_{j_\star^{(k)}}\right\|_2$ as a result of avoiding realignment in $\mathrm{span}(x_i \mid i \in S_F^{(k)})$. To show this we employ two techniques: first, we split the phase in two, defining τ_2 as the first time $\left\|\widetilde{w}_{j_\star^{(k)}}\right\|_2 = \eta$ for some convenient η, as having a stronger bound on neuron $j_\star^{(k)}$ helps control the loss on fitted data; second, we estimate how much neuron $j_\star^{(k)}$ realigns in $\mathrm{span}(x_i \mid i \in S_F^{(k)})$ in terms of the loss on fitted data, and using this control we can show $w_{j_\star^{(k)}}$ remains aligned in direction $D^{(k)}_{j_\star^{(k)}}$ on $[\tau_1, \tau_3]$.
- Realignment and slow growth: The unfitted neurons $j \in N_U^{(k+1)}$ respond to the change in gradients due to a new set of data being fitted by realigning rapidly in direction $D_j^{(k+1)}$ by some time τ_4, after which time this alignment remains stable until any time $\kappa'' \in (\tau_4, t_{k+2})$.

Our proofs do not provide a precise rate of convergence, however it depends on the compact set in question, with a slower rate expected for sets which are nearer to jump times. The rate of convergence derived through our analysis also becomes slower with each successive saddle, and since the number of saddles traversed is directly related to the size of the training dataset one expects a slower rate of convergence for larger training datasets.

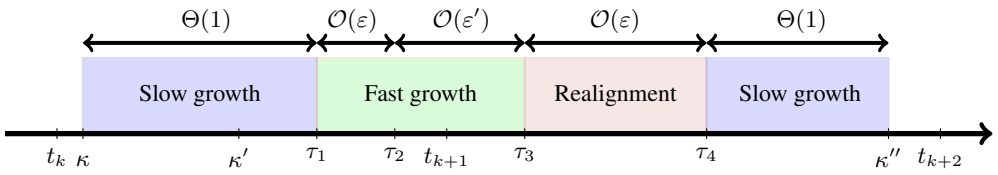


Figure 2: Timeline of the training dynamics about the jump from saddle k to saddle $k+1$. The quantities ε and ε' are parameters of the proof of Theorem 2, τ_1 is the beginning of the fast growth of neuron $j_\star^{(k)}$, τ_3 is the time by which neuron $j_\star^{(k)}$ has grown to its final norm, and τ_4 is the time by which neurons in the set $N_U^{(k+1)}$ have realigned. The quantities κ, κ', and κ'' can be viewed as fixed for the proof of convergence.

5 Implicit bias

Boursier et al. [2022] proved that gradient flow converges to a minimum ℓ_2-norm interpolator $\theta_{\rm opt}$ as the initialization scale tends to zero, which is equivalent to the sum of two neurons, one fitting all positive labels and the other fitting all negative labels. However, this norm optimality relies on a specific initialization condition, that assumes that at least one neuron is active along all positively labeled data and another neuron is active along all negatively labeled data. Under our random initialization, such a condition only holds w.h.p. in a heavily overparameterized regime with $m \gtrsim 2^n$. Boursier et al. [2026] then proved in a later work that with high probability in the mild overparameterization regime ($m \lesssim 2^n$), gradient flow does not converge towards a minimal ℓ_2-norm interpolator as the initialization scale goes to 0. We yet prove in this section that with mild overparameterization, Algorithm 1 produces with high probability an interpolator that is only larger in ℓ_2-norm than the optimal one by a constant multiplicative factor.

We denote in the following by $\theta_{\rm pred}$ the limiting parameters $\theta^\circ(t)$ for any $t \geq t_p$—the definition is independent of the choice of $t > t_p$, thanks to Theorem 2—and by $\theta_{\rm opt}$ any minimum ℓ_2-norm interpolator. The squared ℓ_2-norms of these parameters are given by

$$\frac{1}{2}\|\theta_{\rm opt}\|_2^2 = \sqrt{\sum_{i\,|\,y_i>0} y_i^2} + \sqrt{\sum_{i\,|\,y_i<0} y_i^2} \qquad \text{and} \qquad \frac{1}{2}\|\theta_{\rm pred}\|_2^2 = \sum_{j\in[m]} n \left\| D_j^{(p)} \right\|_2 . \tag{2}$$

where the last jump of the limit process p and $D_j^{(p)}$ are all defined in Algorithm 1. Theorem 3 below upper bounds the norm of the learned interpolator in the mild overparameterization regime as the initialization scale goes to 0.

Theorem 3. *Let $n_+ = |I_+|$, $n_- = |I_-|$, and $\bar{n} = \min(n_+, n_-)$. Under Assumptions 1 and 3, if $\log(m) = \mathrm{o}(\sqrt{\bar{n}}/\log(\bar{n}))$, then the final output of Algorithm 1 satisfies*

$$\frac{1}{2}\|\theta_{\rm pred}\|_2^2 \leq 4\left(\sqrt{n_+} + \sqrt{n_-}\right) \max_{i\in[n]} |y_i|,$$

with probability at least $1 - \exp(-\Omega(\sqrt{\bar{n}}))$ over the network initialization.

Theorem 3 implies that, in the presence of uniformly bounded labels, the achieved interpolator in the mild overparameterization regime as $\alpha \to 0$ scales as $\sqrt{n}$. Although this interpolator is not minimal in norm, it recovers the $\sqrt{n}$ scaling of $\|\theta_{\rm opt}\|_2^2$, thanks to Equation (2). Also note that this type of bound could be easily extended to random labels, replacing the $\max_{i\in[n]} |y_i|$ term by $\max_{i\in[n]} \sqrt{\mathbb{E}[y_i^2]}$ for, e.g., Gaussian labels. Theorem 3 does not require any lower bound on m as Algorithm 1 is still defined in that case. However, m must be of order at least $\log(n)$ to ensure that the achieved estimator interpolates the data, which is guaranteed by Item (i) Assumption 4.

The proof of Theorem 3 reflects the incremental nature of Algorithm 1. Each newly added positive (resp. negative) neuron approximately fits half of the remaining positive (resp. negative) labels. Consequently, after about $\log(n)$ increments, nearly all data points are fitted. The proof relies on a union bound over all subsets of the first $\log(n)$ neurons used to construct the estimator. This approach does not depend on the specific order induced by Algorithm 1. As a result, the bound in Theorem 3

does not capture how the norm of the interpolator decreases as the number of neurons m grows. Establishing such a result remains open, as it would require a more refined analysis that accounts for dependencies between the successive increments in Algorithm 1.

6 Conclusion

In this work, we have provided a complete analysis of the gradient flow dynamics of mildly over-parameterized two-layer ReLU networks from small initialization in the special case of orthogonal data. In doing so, we have extended existing convergence results and shown an incremental learning phenomenon.

There are several avenues for future research. It would be valuable to extend the proof of saddle-to-saddle dynamics to more general data distributions. Possibilities could include weakly interacting data [Dana et al., 2025] or clustered data with orthogonal centers [Min and Vidal, 2025]. Such work could shed light on incremental learning in more realistic settings.

In Theorem 3 on the implicit bias, there is no reference to the width except in the restriction $\log(m) = \mathrm{o}(\sqrt{\bar{n}}/\log(\bar{n}))$. Future work could seek to precisely quantify the dependence of the learned norm on the network width as one approaches the regime $m = \exp(n)$. Doing so would require a more intricate analysis of the dependence of each stage of Algorithm 1 on those that preceded it, which could be of independent interest in analyzing similar sequential algorithms.

Acknowledgments and Disclosure of Funding

The authors are thankful to Nikolaos Zygouras for insightful discussions. The authors acknowledge the Engineering and Physical Sciences Research Council (EPSRC) and DIMAP research centre at the University of Warwick for partial support. J. Town is supported by the EPSRC through the Mathematics of Systems II Centre for Doctoral Training at the University of Warwick (reference EP/S022244/1). B. Lewis is supported by the EPSRC studentship 2927289. The authors acknowledge the use of the Batch Compute System in the Department of Computer Science at the University of Warwick, and associated support services, in the completion of this work.

References

Emmanuel Abbe, Enric Boix Adserà, and Theodor Misiakiewicz. SGD learning on neural networks: leap complexity and saddle-to-saddle dynamics. In *COLT*, pages 2552–2623, 2023. 3

Sanjeev Arora, Nadav Cohen, and Elad Hazan. On the Optimization of Deep Networks: Implicit Acceleration by Overparameterization. In *ICML*, pages 244–253, 2018. 2

Sanjeev Arora, Nadav Cohen, Wei Hu, and Yuping Luo. Implicit Regularization in Deep Matrix Factorization. In *NeurIPS*, pages 7411–7422, 2019a. 2

Sanjeev Arora, Simon S. Du, Wei Hu, Zhiyuan Li, and Ruosong Wang. Fine-Grained Analysis of Optimization and Generalization for Overparameterized Two-Layer Neural Networks. In *ICML*, pages 322–332, 2019b. 1

Devansh Arpit, Stanislaw Jastrzebski, Nicolas Ballas, David Krueger, Emmanuel Bengio, Maxinder S. Kanwal, Tegan Maharaj, Asja Fischer, Aaron C. Courville, Yoshua Bengio, and Simon Lacoste-Julien. A Closer Look at Memorization in Deep Networks. In *ICML*, pages 233–242, 2017. 3

Jean-Pierre Aubin and Arrigo Cellina. *Differential Inclusions: Set-Valued Maps and Viability Theory*. Springer, 1984. 18

Shahar Azulay, Edward Moroshko, Mor Shpigel Nacson, Blake E. Woodworth, Nathan Srebro, Amir Globerson, and Daniel Soudry. On the Implicit Bias of Initialization Shape: Beyond Infinitesimal Mirror Descent. In *ICML*, pages 468–477, 2021. 5

Francis Bach. High-dimensional analysis of double descent for linear regression with random projections. *SIMODS*, 6(1):26–50, 2024. 3

Mikhail Belkin, Daniel Hsu, Siyuan Ma, and Soumik Mandal. Reconciling modern machine-learning practice and the classical bias–variance trade-off. *Proc. Natl. Acad. Sci.*, 116(32):15849–15854, 2019. 3

Mikhail Belkin, Daniel Hsu, and Ji Xu. Two Models of Double Descent for Weak Features. *SIAM J. Math. Data Sci.*, 2(4):1167–1180, 2020. 3

Raphaël Berthier. Incremental Learning in Diagonal Linear Networks. *J. Mach. Learn. Res.*, 24(171):1–26, 2023. 3, 5

David Bertoin, Jerome Bolte, Sébastien Gerchinovitz, and Edouard Pauwels. Numerical influence of ReLU'(0) on backpropagation. In *NeurIPS*, 2021. 4

Jérôme Bolte and Edouard Pauwels. A mathematical model for automatic differentiation in machine learning. In *NeurIPS*, 2020. 4, 51

Jérôme Bolte and Edouard Pauwels. Conservative set valued fields, automatic differentiation, stochastic gradient methods and deep learning. *Math. Program.*, 188(1):19–51, 2021. 4, 51

Etienne Boursier and Nicolas Flammarion. Penalising the biases in norm regularisation enforces sparsity. In *NeurIPS*, 2023. Also in CoRR abs/2303.01353, which contains a correction to Theorem 4. 2

Etienne Boursier and Nicolas Flammarion. Early Alignment in Two-Layer Networks Training is a Two-Edged Sword. *J. Mach. Learn. Res.*, 26:183:1–183:75, 2025a. 2

Etienne Boursier and Nicolas Flammarion. Simplicity bias and optimization threshold in two-layer ReLU networks. In *ICML*, 2025b. 25

Etienne Boursier, Loucas Pillaud-Vivien, and Nicolas Flammarion. Gradient flow dynamics of shallow ReLU networks for square loss and orthogonal inputs. In *NeurIPS*, 2022. 2, 3, 4, 6, 7, 8, 9

Etienne Boursier, Matthew Bowditch, Matthias Englert, and Ranko Lazić. Favorability of Loss Landscape with Regularization Requires Both Large Overparametrization and Initialization. *Trans. Mach. Learn. Res.*, 2026. 3, 9

François Charton. Learning the greatest common divisor: explaining transformer predictions. In *ICLR*, 2024. 3

Sourav Chatterjee. Convergence of gradient descent for deep neural networks. *CoRR*, abs/2203.16462, 2022. 51, 52

Heng-Tze Cheng, Levent Koc, Jeremiah Harmsen, Tal Shaked, Tushar Chandra, Hrishi Aradhye, Glen Anderson, Greg Corrado, Wei Chai, Mustafa Ispir, Rohan Anil, Zakaria Haque, Lichan Hong, Vihan Jain, Xiaobing Liu, and Hemal Shah. Wide & Deep Learning for Recommender Systems. In *DLRS*, pages 7—-10, 2016. 1

Dmitry Chistikov, Matthias Englert, and Ranko Lazić. Learning a Neuron by a Shallow ReLU Network: Dynamics and Implicit Bias for Correlated Inputs. In *NeurIPS*, 2023. 3

Lénaïc Chizat and Francis R. Bach. On the Global Convergence of Gradient Descent for Over-parameterized Models using Optimal Transport. In *NeurIPS*, pages 3040–3050, 2018. 2

Lénaïc Chizat and Francis R. Bach. Implicit Bias of Gradient Descent for Wide Two-layer Neural Networks Trained with the Logistic Loss. In *COLT*, pages 1305–1338, 2020. 2

Lénaïc Chizat, Edouard Oyallon, and Francis R. Bach. On Lazy Training in Differentiable Programming. In *NeurIPS*, pages 2933–2943, 2019. 1

Frank H. Clarke. Generalized gradients and applications. *Trans. Amer. Math. Soc.*, 205:247–262, 1975. 4, 17

Léo Dana, Francis R. Bach, and Loucas Pillaud-Vivien. Convergence of Shallow ReLU Networks on Weakly Interacting Data. In *NeurIPS*, 2025. 1, 2, 3, 4, 10

Simon S. Du, Wei Hu, and Jason D. Lee. Algorithmic Regularization in Learning Deep Homogeneous Models: Layers are Automatically Balanced. In *NeurIPS*, pages 382–393, 2018. 5

Simon S. Du, Xiyu Zhai, Barnabás Póczos, and Aarti Singh. Gradient Descent Provably Optimizes Over-parameterized Neural Networks. In *ICLR*, 2019. 1

Stanislav Fort, Gintare Karolina Dziugaite, Mansheej Paul, Sepideh Kharaghani, Daniel M. Roy, and Surya Ganguli. Deep learning versus kernel learning: an empirical study of loss landscape geometry and the time evolution of the Neural Tangent Kernel. In *NeurIPS*, 2020. 1

Gauthier Gidel, Francis Bach, and Simon Lacoste-Julien. Implicit regularization of discrete gradient dynamics in linear neural networks. In *NeurIPS*, 2019. 3

Margalit Glasgow. SGD Finds then Tunes Features in Two-Layer Neural Networks with near-Optimal Sample Complexity: A Case Study in the XOR problem. In *ICLR*, 2024. 3

Xavier Glorot and Yoshua Bengio. Understanding the difficulty of training deep feedforward neural networks. In *AISTATS*, pages 249–256, 2010. 5

Micah Goldblum, Marc Anton Finzi, Keefer Rowan, and Andrew Gordon Wilson. Position: The No Free Lunch Theorem, Kolmogorov Complexity, and the Role of Inductive Biases in Machine Learning. In *ICML*, pages 15788–15808, 2024. 3

Trevor Hastie, Andrea Montanari, Saharon Rosset, and Ryan J Tibshirani. Surprises in high-dimensional ridgeless least squares interpolation. *Ann. Stat.*, 50(2):949–986, 2022. 3

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Delving Deep into Rectifiers: Surpassing Human-Level Performance on ImageNet Classification. In *ICCV*, pages 1026–1034, 2015. 18

David Holzmüller and Ingo Steinwart. Training Two-Layer ReLU Networks with Gradient Descent is Inconsistent. *J. Mach. Learn. Res.*, 23(181):1–82, 2022. 2

Arthur Jacot, Clément Hongler, and Franck Gabriel. Neural Tangent Kernel: Convergence and Generalization in Neural Networks. In *NeurIPS*, pages 8580–8589, 2018. 1

Arthur Jacot, François Ged, Berfin Şimşek, Clément Hongler, and Franck Gabriel. Saddle-to-Saddle Dynamics in Deep Linear Networks: Small Initialization Training, Symmetry, and Sparsity. *CoRR*, abs/2106.15933, 2021. 3

Ziwei Ji and Matus Telgarsky. Directional convergence and alignment in deep learning. In *NeurIPS*, 2020. 2

Jikai Jin, Zhiyuan Li, Kaifeng Lyu, Simon S. Du, and Jason D. Lee. Understanding Incremental Learning of Gradient Descent: A Fine-grained Analysis of Matrix Sensing. In *ICML*, pages 15200–15238, 2023. 3

Akshay Kumar and Jarvis Haupt. Directional Convergence Near Small Initializations and Saddles in Two-Homogeneous Neural Networks. *Trans. Mach. Learn. Res.*, 2024. 6

Daniel Kunin, Giovanni Luca Marchetti, Feng Chen, Dhruva Karkada, James B Simon, Michael R DeWeese, Surya Ganguli, and Nina Miolane. Alternating Gradient Flows: A Theory of Feature Learning in Two-layer Neural Networks. In *NeurIPS*, 2025. 3, 5

Zhenyu Liao, Romain Couillet, and Michael W. Mahoney. A random matrix analysis of random Fourier features: beyond the Gaussian kernel, a precise phase transition, and the corresponding double descent. In *NeurIPS*, 2020. 3

Yaron Lipman, Ricky T. Q. Chen, Heli Ben-Hamu, Maximilian Nickel, and Matthew Le. Flow Matching for Generative Modeling. In *ICLR*, 2023. 1

Kaifeng Lyu and Jian Li. Gradient Descent Maximizes the Margin of Homogeneous Neural Networks. In *ICLR*, 2020. 2, 18

Kaifeng Lyu, Zhiyuan Li, Runzhe Wang, and Sanjeev Arora. Gradient Descent on Two-layer Nets: Margin Maximization and Simplicity Bias. In *NeurIPS*, pages 12978–12991, 2021. 3

Wesley J. Maddox, Gregory W. Benton, and Andrew Gordon Wilson. Rethinking Parameter Counting in Deep Models: Effective Dimensionality Revisited. *CoRR*, abs/2003.02139, 2020. 3

Hartmut Maennel, Olivier Bousquet, and Sylvain Gelly. Gradient Descent Quantizes ReLU Network Features. *CoRR*, abs/1803.08367, 2018. 5

Song Mei and Andrea Montanari. The generalization error of random features regression: Precise asymptotics and the double descent curve. *Commun. Pure Appl. Math.*, 75(4):667–766, 2022. 3

Hancheng Min and René Vidal. Gradient Flow Provably Learns Robust Classifiers for Orthonormal GMMs. In *ICML*, 2025. 10

Hancheng Min, Enrique Mallada, and René Vidal. Early Neuron Alignment in Two-layer ReLU Networks with Small Initialization. In *ICLR*, 2024. 3, 6

Preetum Nakkiran, Gal Kaplun, Dimitris Kalimeris, Tristan Yang, Benjamin L. Edelman, Fred Zhang, and Boaz Barak. SGD on Neural Networks Learns Functions of Increasing Complexity. In *NeurIPS*, pages 3491–3501, 2019. 3

Preetum Nakkiran, Gal Kaplun, Yamini Bansal, Tristan Yang, Boaz Barak, and Ilya Sutskever. Deep Double Descent: Where Bigger Models and More Data Hurt. In *ICLR*, 2020. 3

Scott Pesme and Nicolas Flammarion. Saddle-to-Saddle Dynamics in Diagonal Linear Networks. In *NeurIPS*, 2023. 3, 5

Mary Phuong and Christoph H. Lampert. The inductive bias of ReLU networks on orthogonally separable data. In *ICLR*, 2021. 3

Nasim Rahaman, Aristide Baratin, Devansh Arpit, Felix Draxler, Min Lin, Fred A. Hamprecht, Yoshua Bengio, and Aaron C. Courville. On the Spectral Bias of Neural Networks. In *ICML*, pages 5301–5310, 2019. 3

Itay Safran, Gal Vardi, and Jason D. Lee. On the Effective Number of Linear Regions in Shallow Univariate ReLU Networks: Convergence Guarantees and Implicit Bias. In *NeurIPS*, 2022. 2

Andrew M. Saxe, James L. McClelland, and Surya Ganguli. Exact solutions to the nonlinear dynamics of learning in deep linear neural networks. In *ICLR*, 2014. 2

Roman Vershynin. *High-Dimensional Probability: An Introduction with Applications in Data Science*. Cambridge University Press, 2nd edition, 2026. 55

Mingze Wang and Chao Ma. Understanding Multi-phase Optimization Dynamics and Rich Nonlinear Behaviors of ReLU Networks. In *NeurIPS*, 2023. 3

Andrew Gordon Wilson. Position: Deep Learning is Not So Mysterious or Different. In *ICML*, pages 82326–82346, 2025. 3

Stephan Wojtowytsch. On the Convergence of Gradient Descent Training for Two-layer ReLU-networks in the Mean Field Regime. *CoRR*, abs/2005.13530, 2020. 2

Blake E. 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 *COLT*, pages 3635–3673, 2020. 2

Sergey Zagoruyko and Nikos Komodakis. Wide Residual Networks. In *BMVC*, 2016. 1

Appendix

Table of contents

A Experiments **14**
A.1 Experimental details 14
A.2 Additional experiments 14

B Technical background **17**
B.1 Clarke subdifferential 17

C Recap of notations **18**

D Proof of Theorem 2 **20**
D.1 Basis case and autonomous systems 20
D.2 Phase 1: Slow growth 25
D.3 Phase 2a: Growth to constant order 31
D.4 Phase 2b: Fitting of data 39
D.5 Phase 3: Realignment and slow growth 44
D.6 Proof of Theorem 2 50

E Proof of Theorem 3 **54**

F Auxiliary results **58**
F.1 Proof of Proposition 2 58
F.2 Other auxiliary results 60

G NeurIPS Paper Checklist **62**

A Experiments

A.1 Experimental details

Figure 1 was generated by training a two-layer ReLU network with six neurons on 64 orthonormal data points in 64 dimension. The labels are chosen as the absolute values of standard Gaussian random variables. As explained in Appendix D.1, without the sign restriction on the labels, we get two independent processes: neurons with negative output weights will turn and stay in the subspace orthogonal to the space spanned by the data points with positive labels, and vice versa. The sign restriction on the labels therefore helps to more clearly illustrate one of these processes in the plots, as opposed to a loss curve that is the combination of interleaving these two independent processes.

The network weights are initialized such that the hidden layer neurons have random directions and ℓ_2 norm e^{-500} and the output weights are e^{-500}. To avoid numerical issues, we maintain a scaled version of the weights such that they are closer to 1 (i.e., initially scaled by e^{500}) and separately track the logarithm of the scale of the weights (initially -500). During training, the stored weights are kept near 1 while the logarithm of the scale is adjusted. The loss is the mean squared error and we train the network with full batch gradient descent and a learning rate of 0.01 until the loss is less than 10^{-20}. The plots in Figure 1 start at epoch $\approx 0.9 \cdot t_1$, For the purposes of the illustration, we consider a data point fitted when the output of the network for that point is at least 50% of the corresponding label. Due to the exponential growth of neuron norms, once the output of a network for a data point reaches such a significant proportion of the label, the output of the network at the data point very rapidly closes in on the label value. Therefore, the exact choice of the threshold (here 50%) is somewhat arbitrary but also immaterial in the sense that it does not greatly influence the general picture.

A.2 Additional experiments

For the same training as in Figure 1, Figure 3 shows how the ℓ_2 norms of the hidden layer neurons change over time, but with a logarithmic y-axis. This reveals that the norms grow at an exponential

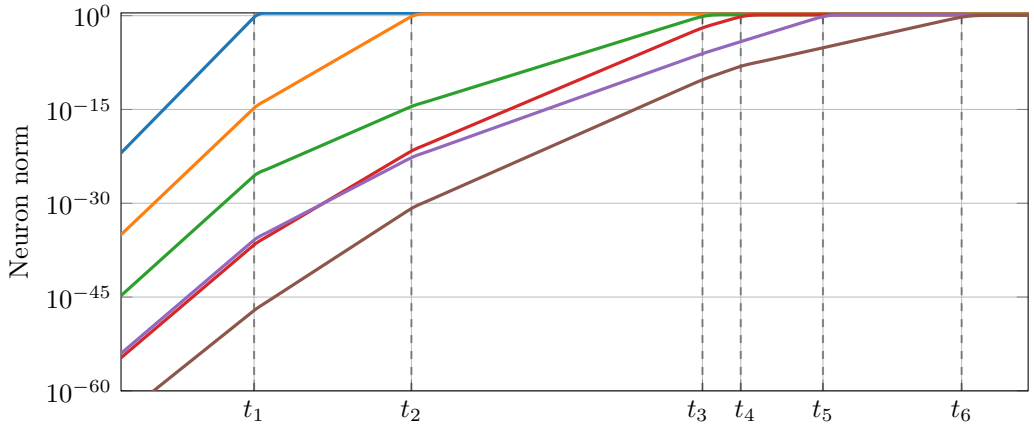


Figure 3: The same as the right plot of Figure 1 but with the y-axis on a logarithmic scale. The neuron norms increase exponentially. When new data points are fitted, the rate of this exponential increase reduces for the remaining neurons that were active on at least one of those newly fitted data points.

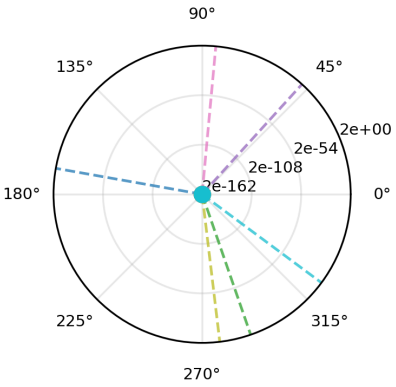
rate. At the jump times, when a neuron fits the data on which it is active, that neuron stops growing. Other neurons continue to grow. However, if they were active on some of the same data points as the newly fitted neuron, they are now growing at a slower exponential rate than before the jump time. Therefore, the logarithms of the neuron norms are piecewise linear curves with decreasing slopes. The slopes of these linear segments as well as the jump times are predicted by Algorithm 1. Specifically, between jump times t_p and t_{p+1}, the slope for neuron j is proportional to $\left\|D_j^{(p)}\right\|_2$.

Still for the same training, Figure 4 shows the norm (on a logarithmic scale) and direction (projected onto a random two-dimensional plane) of the six neurons at different points in time. Initially, the norms are very small. The neurons orient themselves in the directions predicted by Algorithm 1 and their norm starts to increase exponentially. Just before time t_1, one neuron has a significant norm while the remaining neurons have norms that are still much smaller. Just after time t_1, the neuron directions predicted by Algorithm 1 change and the neurons that did not fit any data yet turn quickly towards their new predicted directions. Time t_1 approximately corresponds to epoch 507,000. Plot (c) shows the process at epoch 506,000, plot (d) shows it at epoch 508,000, plot (e) shows it at epoch 510,000, and plot (f) shows it at epoch 538,000.

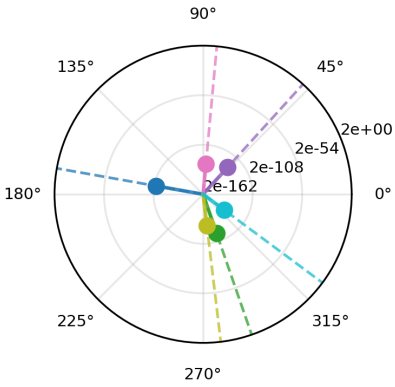
Norm minimization. Theorem 3 gives an upper bound on the squared ℓ_2 norm of the parameters of the achieved interpolator in the mild overparameterization regime. For uniformly bounded labels, this bound is of the order $\sqrt{n}$ and therefore of the same order as the minimum norm interpolator. Figure 5 experimentally shows the growth of the squared norm of the parameters of the achieved interpolator versus that of the optimal interpolator, depending on the number n of data points.

Specifically, we train a two-layer ReLU network with full batch gradient descent and a learning rate of 0.5. For different n, a network of width $m = \lceil \ln(10000 \cdot n)/\ln(4/3) \rceil$ is initialized such that the hidden neurons have random directions and norm $10^{-30}/\sqrt{m}$ and the output neurons have a random sign and absolute value $10^{-30}/\sqrt{m}$. The choice of m ensures that there are a sufficient number of neurons to ensure interpolation in a typical training run. The factor $1/\sqrt{m}$ in the initialization scale ensures that the initial squared norm of the network parameters remains the same regardless of the width of the network. The network is trained on data points corresponding to the $n \times n$ identity matrix and labels are drawn from a standard Gaussian distribution. The identity matrix is chosen for computational and memory efficiency, but we note that the training dynamics are rotation invariant and therefore the behavior for the identity matrix is the same as for any other orthonormal basis.

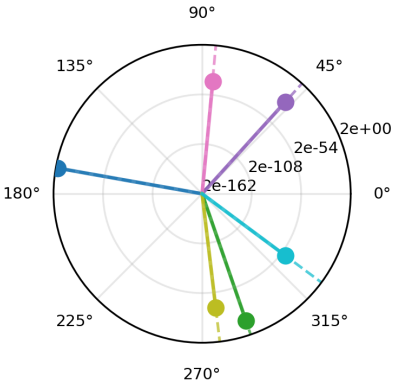
Training is run for one million epochs but stopped earlier if the loss reaches 10^{-10}. We record the squared norm of the parameters of the final network. This is repeated three times with new random initial weights. In Figure 5, we plot the mean values of the squared norms as well the optimum squared norm $\|\theta_{\text{opt}}\|_2^2$ (see Equation (2)). We also indicate the min/max deviation over the three runs, but this is barely visible due to relatively low variance of the resulting squared norms.



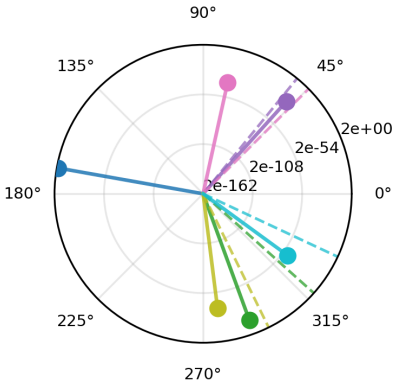
(a) Near the start



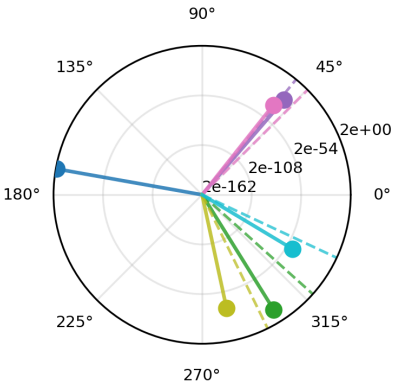
(b) Around time $t_1/2$



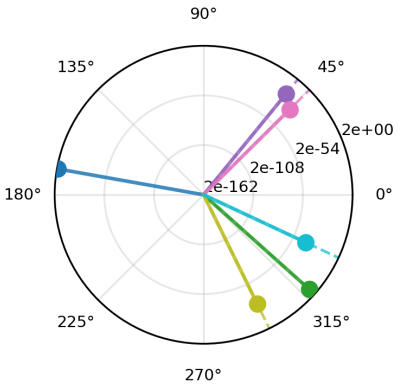
(c) Just before time t_1



(d) Just after time t_1



(e) A short while after t_1



(f) Around time $(t_1 + t_2)/2$

Figure 4: For the same training as in Figure 1, these polar plots show the norm of each neuron (on a logarithmic scale) as well as the direction of the neuron projected onto a random two-dimensional plane, at different points in time. The projections of the neuron directions from the limit process $\theta^\circ(t)$ predicted by Algorithm 1 are indicated by dashed lines.

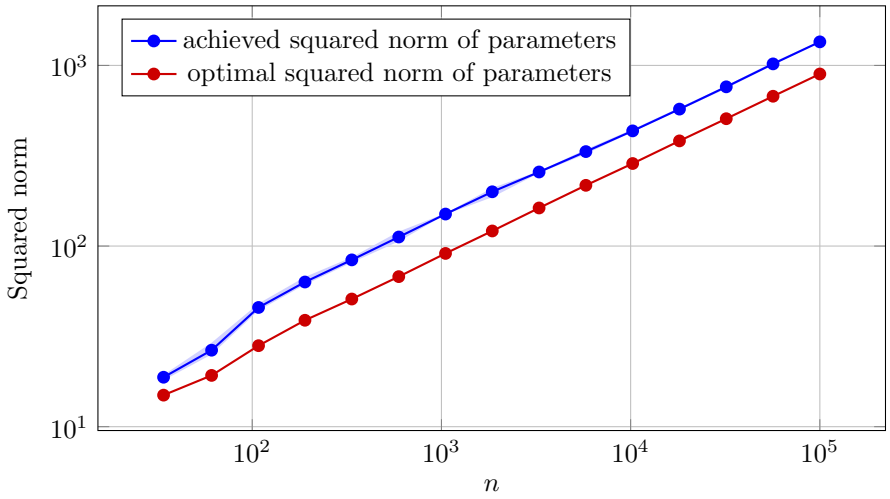


Figure 5: We train a two-layer ReLU network with increasing data dimensionality and number of data points n, and plot the square of the norm of the parameters of the final network as well as that of the optimal interpolator (see Equation (2)). The hidden layer size m increases logarithmically in n, and we use a balanced initialization of scale $10^{-30}/\sqrt{m}$. Each value plotted is the mean over three runs, with each run using fixed orthogonal inputs and different random labels which are standard Gaussians. We use full batch gradient descent with a learning rate of 0.5 for one million epochs or until the loss reaches 10^{-10}. The area between the minimum and maximum values is shaded, however they are relatively concentrated.

In the plot, both axes are logarithmic. The straight lines of slope about $1/2$ therefore indicate that both squared norms grow proportional to $\sqrt{n}$ and that the achieved squared norm is only by a constant factor larger than the optimal one.

Larger overparameterization. Our next experiment investigates what happens when we increase the overparameterization for two different initialization schemes. The first initialization scheme is the same as before used to produce Figure 5 (i.e., balanced small initialization with hidden layer neuron norms equal to $10^{-30}/\sqrt{m}$). The second scheme is a uniform He initialization which in our network draws hidden layer weights uniformly at random from $[-1/\sqrt{d}, 1/\sqrt{d}]$ and output layer weights uniformly at random from $[-1/\sqrt{m}, 1/\sqrt{m}]$. This is the default initialization for linear layers in PyTorch. The experimental details are the same as before, except that n is now fixed at 32 while m is varied and the learning rate is 1 when using the first initialization scheme and 0.001 when using the second initialization scheme.

Each point in the plot of Figure 6 is the mean of three runs with different random initial weights. We also indicate the min/max deviation over the three runs, but this is barely visible due to relatively low variance of the resulting squared norms.

We observe that with small initialization, the squared ℓ_2 norm of the trained network approaches the theoretical optimum as the width of the network increases. This is in contrast to the uniform He initialization, for which the square ℓ_2-norm of the trained network keeps increasing.

Compute. The experiment for Figures 1, 3 and 4 takes only about one minute to generate on a standard desktop CPU. Generating Figure 5 takes about 3 hours, and generating Figure 6 takes about 10 minutes, both on a consumer GPU.

B Technical background

B.1 Clarke subdifferential

We introduce the Clarke subdifferential [Clarke, 1975] and a basic result on the existence of gradient flows using this generalized derivative.

Definition 1. *The Clarke subdifferential of a locally Lipschitz function $f : \mathcal{X} \to \mathbb{R}$ at the point $x \in \mathcal{X}$ is defined as the convex set*

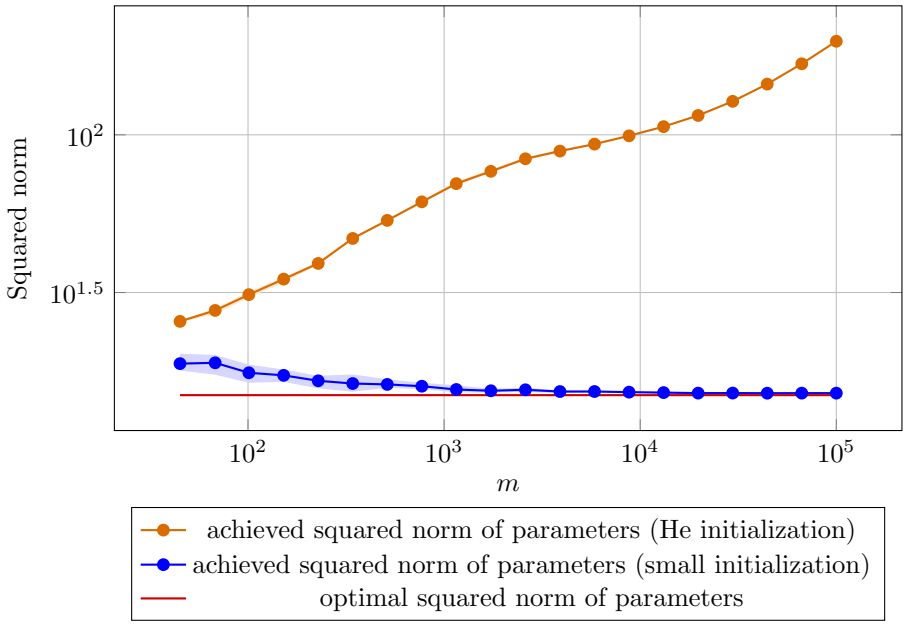


Figure 6: We train two-layer ReLU networks of varying widths m on 32 orthonormal data points in 32 dimensions and labels drawn from the standard Gaussian distribution. We compare two initialization schemes: balanced small initialization where hidden neurons have norm $10^{-30}/\sqrt{m}$ and uniform He initialization [He et al., 2015], which is the default used in PyTorch to initialize linear layers. We train using full batch gradient descent for one million epochs or until the loss reaches 10^{-10}. Each value plotted is the mean over three runs with new random initial weights. The area between the minimum and maximum values is shaded, however they are relatively concentrated.

$$\partial f(x) = \operatorname{conv}\left\{ \lim_{k\to\infty} \nabla f(x_k) \;\middle|\; x_k \to x,\ f \text{ is differentiable at } x_k \right\}.$$

One may see Lyu and Li [2020, Section 3] for further details. We present the following result guaranteeing the local existence of solutions.

Theorem 4. *[Aubin and Cellina, 1984, Theorem 3, Chapter 2] Let $\mathcal{X}$ be a Hilbert space and $Q \subset \mathbb{R} \times \mathcal{X}$ be an open subset containing $(0, x_0)$. Let F be an upper semicontinuous map from Q into the non-empty closed convex subsets of $\mathcal{X}$. We assume that $(t, x) \mapsto m(F(t, x))$ is locally compact, where for $A \subset \mathcal{X}$ we define $m(A) := \inf\{\|v\| \mid v \in A\}$. Then there exist $T > 0$ and an absolutely continuous function $x(\cdot)$ defined on $[0, T]$, a solution to the differential inclusion*

$$x'(t) \in F(t, x(t)); \qquad x(0) = x_0.$$

C Recap of notations

For ease of reference, Table 1 collects notation used throughout the paper, including some symbols that appear only in the appendix.

Table 1: Definitions and descriptions of named quantities, collected here from the text. Hyperlinks in the Symbol column take the reader to the first point in the paper where a given symbol is introduced.

Symbol	Definition	Description
I_s	$\{i \in [n] \mid \operatorname{sgn}(y_i) = s\}$	The set of data i with sign s, for $s \in \{\pm 1\}$.
$A_{i,j}$	$\mathbb{1}\left(w_j^\top(0)\, x_i > 0 \text{ and } i \in I_{s_j}\right)$	Indicator of data i on which, at initialization, neuron j is active with correct second-layer weight sign.
J_s	$\{j \in [m] \mid \operatorname{sgn}(a_j(0)) = s\}$	The set of neurons j with sign s, for $s \in \{\pm 1\}$.
t_k	$t_0 = 0$ $t_{k+1} = \min_{j \in N_U^{(k)}} \left\{ t_k - \frac{\ell_j^\circ(t_k)}{\left\|D_j^{(k)}\right\|} \right\}$	The jump times.
$\ell_j^\circ(t)$	$\ell_j^\circ(0) = -1$ $\ell_j^\circ(t) = \min\left(\ell_j^\circ(t_k) + (t - t_k)\|D_j^{(k)}\|, 0\right)$	The exponents of the neurons' norms, where $t \in [t_k, t_{k+1}]$.
$j_\star^{(k)}$	$\arg\min_{j \in N_U^{(k)}} \left\{ t_k - \frac{\ell_j^\circ(t_k)}{\left\|D_j^{(k)}\right\|} \right\}$	The neuron which fits when moving from saddle k to saddle $k+1$.
$N_U^{(k)}$	$N_U^{(0)} = [m]$ $N_U^{(k+1)} = N_U^{(k)} \setminus \{j^\star\}$	The set of unfitted neurons at saddle k.
$N_F^{(k)}$	$[m] \setminus N_U^{(k)}$	The set of fitted neurons at saddle k.
$S_j^{(k)}$	$S_j^{(0)} = S_j = \{i \in [n] \mid A_{i,j} = 1\}$ $S_j^{(k+1)} = \begin{cases} S_j^{(k)} \setminus S_{j_\star^{(k)}}^{(k)} & \text{if } j \neq j_\star^{(k)} \\ S_j^{(k)} & \text{if } j = j_\star^{(k)} \end{cases}$	The set of unfitted data on which neuron j is active with correct second-layer weight at saddle k, and which has not yet been fitted by any other neuron.
$S_U^{(k)}$	$S_U^{(0)} = [n]$ $S_U^{(k+1)} = S_U^{(k)} \setminus S_{j^\star}^{(k)}$	The set of unfitted data at saddle k.
$S_F^{(k)}$	$[n] \setminus S_U^{(k)}$	The set of fitted data at saddle k.
$D_j^{(k)}$	$D_j^{(k)} = \frac{1}{n} \sum_{i \in S_j^{(k)}} y_i x_i$	The label-weighted sum of data in the set $S_j^{(k)}$, which is the direction in which neuron j aligns and grows at saddle k.
$E^{(k)}$	$E^{(k)} = -\frac{1}{n} \sum_{i \in S_F^{(k)}} (h_{\widetilde{\theta}}(x_i) - y_i) x_i$	A dynamical vector encoding the loss on data which has already been fitted, with $n \left\|E^{(k)}\right\|^2$ directly corresponding to these terms of the loss.
$\mathcal{D}_j$	$\mathcal{D}_j = -\frac{1}{n} \sum_{\substack{i \in [n] \mid \\ w_j^\top x_i > 0}} (h_{\widetilde{\theta}}(x_i) - y_i) x_i$	A dynamical vector encoding the loss on all data on which neuron j is active, with $n\|\mathcal{D}_j\|^2$ directly corresponding to these terms of the loss.

D Proof of Theorem 2

The proof of convergence to the limit process defined in Algorithm 1 is due to a non-asymptotic analysis of the dynamics of all neurons around the jump from saddle k to saddle $k+1$. We use this to construct an inductive argument for convergence, with the number of saddles traversed by the flow serving as the parameter of the induction. Specifically, we will assume the following inductive hypothesis on convergence at saddle k:

Hypothesis 1 (for $k \in \{0, 1, \dots, p\}$)**.** *Given any $\kappa \in (t_k, t_{k+1})$, there exists $\delta > 0$ such that for all $\xi > 0$, there exists $\alpha^* > 0$ such that for all $\alpha \in (0, \alpha^*)$:*

1. $\forall j \in N_U^{(k)},\ \left|\ell_j^\circ(\kappa) + \log_\alpha(\|\widetilde{w}_j(\kappa)\|)\right| < \xi.$
2. $\forall j \in N_F^{(k)},\ \left|n\left\|D_j^{(k)}\right\| - \|\widetilde{w}_j(\kappa)\|^2\right| < \alpha^\delta.$
3. $\forall i \in S_F^{(k)},\ |h_{\widetilde{\theta}(\kappa)}(x_i) - y_i| < \alpha^\delta.$
4. $\forall j \in [m],\ s_j \widetilde{\overline{w}}_j^\top(\kappa)\overline{D}_j^{(k)} > 1 - \alpha^\delta.$

In the above, and throughout the proof, we use $\|\cdot\|$ to denote the ℓ_2-norm. The reader should note that δ may degenerate as κ approaches either endpoint t_k or t_{k+1}. This does not present a problem, since the proof is of convergence on compact subsets of $\mathbb{R}$ which exclude $(t_k)_{k=1}^p$. Therefore, κ may be viewed as a quantity which is fixed at the beginning of the proof depending on the compact set in question. The following lemma states that on compact intervals not containing a jump time, we have the desired uniform convergence to the limit process under Hypothesis 1.

Lemma 1. *Suppose that Hypothesis 1 holds for some $k \in \{0, 1, \dots, p-1\}$. Then given any $\kappa, \kappa' > 0$ such that $t_k < \kappa < \kappa' < t_{k+1}$, $\widetilde{\theta}^\alpha$ converges uniformly to θ° on $[\kappa, \kappa']$.*

The following lemma recovers Hypothesis 1 at saddle $k+1$.

Lemma 2. *If Hypothesis 1 holds for some $k \in \{0, 1, \dots, p-1\}$, then it holds for $k+1$.*

Proofs of each of these lemmas, as well as of Theorem 2, are provided in Appendix D.6. Final convergence in time is proven in Lemma 9.

D.1 Basis case and autonomous systems

We provide here an analysis of the early time dynamics, specifically a proof of the alignment of neurons towards the directions $D_j^{(0)}$ by an arbitrary positive time, where

$$D_j^{(0)} := \frac{1}{n}\sum_{i\in[n]} A_{i,j} y_i x_i,$$

and we recall that

$$A_{i,j} = \mathbb{1}\left(w_j^\top(0)\, x_i > 0 \text{ and } i \in I_{s_j}\right).$$

This will amount to a proof of a statement similar to the case $k = 0$ for Hypothesis 1. At this point, it will also be helpful to introduce the vector

$$D_j^{(k)} := \frac{1}{n}\sum_{i\in S_j^{(k)}} y_i x_i.$$

In the above display, $S_j^{(k)}$ is recursively defined by

$$S_j^{(0)} = \{i \in [n] \mid A_{i,j} = 1\}$$
$$S_j^{(k+1)} = \begin{cases} S_j^{(k)} \setminus S_{j_\star^{(k)}}^{(k)} & \text{if } j \neq j_\star^{(k)} \\ S_j^{(k)} & \text{if } j = j_\star^{(k)} \end{cases}$$

and denotes the set of data on which neuron j is active at saddle k and on which no neuron in the set

$$N_F^{(k)} := [m] \setminus N_U^{(k)}$$

is active. Note that S_j (as introduced in Algorithm 1) is the same set as $S_j^{(0)}$. In a similar way to how we define the set of neurons which have fitted their data, we can also define the set of fitted data itself by

$$S_F^{(k)} := [n] \setminus S_U^{(k)}.$$

One may see Algorithm 1 for the construction of the sets $N_U^{(k)}$ and $S_U^{(k)}$. Our arguments will use the equations

$$\dot{\widetilde{a}}_j = \log(1/\alpha)\widetilde{w}_j^\top \mathcal{D}_j \qquad\qquad \dot{\widetilde{w}}_j = \log(1/\alpha)\widetilde{a}_j \mathcal{D}_j,$$

where we have

$$\mathcal{D}_j = -\frac{1}{n} \sum_{i \in [n] \mid \widetilde{w}_j^\top x_i > 0} (h_{\widetilde{\theta}}(x_i) - y_i) x_i.$$

As noted in Section 2, these equations combined with Proposition 1 and Assumption 2 imply that

$$\dot{\widetilde{\overline{w}}}_j(t) = s_j \log(1/\alpha) \left(\mathcal{D}_j(t) - \widetilde{\overline{w}}_j(t) \widetilde{\overline{w}}_j^\top(t) \mathcal{D}_j(t) \right).$$

Using these equations, we will ensure that by an arbitrary time $\kappa > 0$, all neurons have aligned towards the directions $D_j^{(0)}$. Furthermore, we will verify by introducing an auxiliary parameter of the proof, $\widehat{\kappa}$, that $-\widetilde{\ell}_j(\kappa) \approx -\ell_j^\circ(\kappa)$ with arbitrary precision. The sketch of the proof is to ensure that by any time $\widehat{\kappa} > 0$, the realignment has occurred and then remains stable until time κ. Therefore, the precision of the exponents of the unfitted neurons is controlled by the arbitrary parameter $\widehat{\kappa}$. To state the lemma below, we define for $s \in \{\pm 1\}$ the set of neurons with output sign given by s:

$$J_s := \{j \in [m] \mid \operatorname{sgn}(a_j(0)) = s\}.$$

Lemma 3. *For any $\kappa \in \left(0, \frac{1}{4}\sqrt{\frac{n}{2\mathcal{L}(0)}}\right)$ there exists $\delta > 0$ such that for any $\xi > 0$ there exists $\alpha^*(\kappa, \xi) > 0$ such that for all $\alpha \in (0, \alpha^*)$ and both signs $s \in \{\pm 1\}$ the following are true:*

1. $\forall j \in J_s,\ s_j \widetilde{\overline{w}}_j^\top(\kappa) \overline{D}_j^{(0)} > 1 - \alpha^\delta.$
2. $\forall j \in J_s,\ \left|\ell_j^\circ(\kappa) + \log_\alpha(\|\widetilde{w}_j(\kappa)\|)\right| < \xi\,.$
3. $\forall i \in I_{-s}, \forall j \in J_s,\ \widetilde{w}_j^\top(\kappa) x_i \leq 0.$

The monotonicity of the gradient flow means there exists a maximum rate of growth it is possible for any neuron to achieve. Therefore, the upper bound on κ appears as a technicality to ensure that even if a neuron were to have grown at this maximum rate from initialization, it could not have exceeded the threshold $\alpha^{1/2}$ which we introduce for the purpose of the proof. This kind of argument will be

used several times throughout subsequent analyses (see Proposition 3 and its application in the proof of Lemmas 4 to 7).

Item 3 Lemma 3 is a particularly important point, because it ensures that neurons deactivate within an arbitrarily short length of time on the data for which $A_{i,j} = 0$. Combining this with Assumption 2 will allow us to see that the network splits into two sub-networks, consisting only of neurons with positive (resp. negative) second-layer weight and which are active on data with only positive (resp. negative) labels. This allows us to consider w.l.o.g. the case that $s_j > 0$ for all $j \in [m]$ and $\mathrm{sgn}(y_i) > 0$ for all $i \in [n]$. This is explained more precisely after the proof.

We stress now that the statement of Lemma 3 does not exactly correspond to the statement of Hypothesis 1, since the interval $\left(0, \frac{1}{4}\sqrt{\frac{n}{2\mathcal{L}(0)}}\right)$ does not correspond to the interval $(0, t_1)$. However, when uniform convergence to θ° is proven, we will apply Lemma 1, and this statement is sufficient to obtain uniform convergence on arbitrary closed intervals contained in $(0, t_1)$. Concretely, if we were to consider a compact set partially or entirely contained in $\left(\frac{1}{4}\sqrt{\frac{n}{2\mathcal{L}(0)}}, t_1\right)$, we could obtain convergence by applying Lemma 1 on some interval whose lower endpoint is contained in $\left(0, \frac{1}{4}\sqrt{\frac{n}{2\mathcal{L}(0)}}\right)$ and whose upper endpoint is greater than the maximum of the compact set.

Proof. Define $\tau := \inf\{t \geq 0 \mid \exists\, j \in [m] \text{ s.t. } \|\widetilde{w}_j\| \geq \alpha^{1/2}\}$. As a preliminary, we check that $\kappa < \tau$. On $[0, \tau]$,

$$\begin{aligned} \frac{1}{\log(1/\alpha)} \frac{\mathrm{d}}{\mathrm{d}t} \|\widetilde{w}_j\| &= -s_j \frac{\|\widetilde{w}_j\|}{n} \sum_{i \in [n] \,|\, \widetilde{w}_j^\top x_i > 0} (h_{\widetilde{\theta}}(x_i) - y_i) \widetilde{w}_j^\top x_i \\ &\leq \sqrt{\frac{2\mathcal{L}(0)}{n}} \|\widetilde{w}_j\|. \end{aligned} \tag{3}$$

This follows from the Cauchy-Schwarz inequality combined with the fact that $\mathcal{L}$ is monotone decreasing, by definition of the gradient flow. Grönwall's inequality gives that on this interval,

$$\begin{aligned} \|\widetilde{w}_j(t)\| &\leq \|\widetilde{w}_j(0)\| \alpha^{-\sqrt{\frac{2\mathcal{L}(0)}{n}} t} \\ &= \alpha^{1 - \sqrt{\frac{2\mathcal{L}(0)}{n}} t}. \end{aligned}$$

Since $\kappa < \frac{1}{4}\sqrt{\frac{n}{2\mathcal{L}(0)}}$, we indeed have $\kappa < \tau$. We may reverse the inequality and consequently obtain for $t \in [0, \tau]$ the lower bound

$$\|\widetilde{w}_j(t)\| > \alpha^{1 + \sqrt{\frac{2\mathcal{L}(0)}{n}} t}.$$

We conclude that for $t \in [0, \kappa]$

$$\alpha^{1 + \sqrt{\frac{2\mathcal{L}(0)}{n}} t} < \|\widetilde{w}_j(t)\| < \alpha^{1 - \sqrt{\frac{2\mathcal{L}(0)}{n}} t}. \tag{4}$$

Observe at this point that the signs are symmetric, and so we prove w.l.o.g. the statements for $s_j = 1$.

Proof of Item 1. For the claimed alignment, we compute that

$$\begin{aligned}\frac{1}{\log(1/\alpha)}\frac{\mathrm{d}}{\mathrm{d}t}\widetilde{\overline{w}}_j^\top D_j^{(0)} &\geq \left\|D_j^{(0)}\right\|^2 - \left(\widetilde{\overline{w}}_j^\top D_j^{(0)}\right)^2 \\ &\quad - \frac{1}{n}\sum_{i\,|\,\widetilde{w}_j^\top x_i>0} h_{\widetilde{\theta}}(x_i)\left(x_i^\top D_j^{(0)} - x_i^\top \widetilde{\overline{w}}_j \widetilde{\overline{w}}_j^\top D_j^{(0)}\right) \\ &> \left\|D_j^{(0)}\right\|^2 - \left(\widetilde{\overline{w}}_j^\top D_j^{(0)}\right)^2 - 2m\left\|D_j^{(0)}\right\|\alpha,\end{aligned}$$

where the first inequality is from discarding negatively labeled data on which neuron j is active. This is because such data add a positive contribution for α small enough that $|h_{\widetilde{\theta}}(x_i)| < |y_i|$. We then use the fact that $f'(t) = \log(1/\alpha)(a^2 - f(t)^2)$ is solved by $a\tanh(a\log(1/\alpha)t + t_0)$, where t_0 is an integration constant. Since the inner product is positive at initialization by definition, t_0 is positive. By the fact that $\tanh$ is monotone increasing, we obtain a lower bound by discarding this integration constant. This leaves

$$\begin{aligned}\widetilde{\overline{w}}_j^\top D_j^{(0)} &> \left(\left\|D_j^{(0)}\right\|^2 - 2m\alpha\right)^{1/2}\tanh\left(\log(1/\alpha)\left(\left\|D_j^{(0)}\right\|^2 - 2m\alpha\right)^{1/2} t\right) \\ &\geq \left(\left\|D_j^{(0)}\right\| - \sqrt{2m\alpha}\right)\left(1 - \alpha^{2\left(\left\|D_j^{(0)}\right\| - \sqrt{2m\alpha}\right)t}\right),\end{aligned}$$

where above we applied the inequalities $\sqrt{1-x^2} \geq 1 - x$ and $\tanh(x) \geq 1 - 2\exp(-2x)$. Note that there exists α sufficiently small that for any $j \in [m]$, $2\left(\left\|D_j^{(0)}\right\| - \sqrt{2m\alpha}\right) > \left\|D_j^{(0)}\right\|$, so for α below this threshold at time $t \in [0, \kappa]$ one has

$$\begin{aligned}\widetilde{\overline{w}}_j^\top(t) D_j^{(0)} &> \left(\left\|D_j^{(0)}\right\| - \sqrt{2m\alpha}\right)\left(1 - \alpha^{\left\|D_j^{(0)}\right\| t}\right) \\ &> \left\|D_j^{(0)}\right\|\left(1 - \alpha^{\left\|D_j^{(0)}\right\| t} - \alpha^{1/4}\right),\end{aligned} \tag{5}$$

where the second inequality is valid for α sufficiently small. This directly implies that for α small enough and any $\kappa \in \left(0, \frac{1}{4}\sqrt{\frac{n}{2\mathcal{L}(0)}}\right)$, there exists $\delta > 0$ such that the alignment of any neuron j with direction $D_j^{(0)}$ is bounded by below by $1 - \alpha^\delta$ at time κ, as required. Note here that δ depends on κ, and α^* on the data, width, and directions $D_j^{(0)}$. □

Proof of Item 2. We select some arbitrary $\widehat{\kappa} \in (0, \kappa)$. Since $\widehat{\kappa} < \tau$, we may insert it into Equation (4) to obtain

$$\alpha^{1+\sqrt{\frac{2\mathcal{L}(0)}{n}}\widehat{\kappa}} < \|\widetilde{w}_j(\widehat{\kappa})\| < \alpha^{1-\sqrt{\frac{2\mathcal{L}(0)}{n}}\widehat{\kappa}}.$$

We express this deviation in terms of the quantity $\ell_j^\circ(\widehat{\kappa})$. We note that $-\ell_j^\circ(\widehat{\kappa}) = 1 - \left\|D_j^{(0)}\right\|\widehat{\kappa}$ for all $j \in [m]$. Therefore,

$$\alpha^{-\ell_j^\circ(\widehat{\kappa}) + \left(\left\|D_j^{(0)}\right\| + \sqrt{\frac{2\mathcal{L}(0)}{n}}\right)\widehat{\kappa}} < \|\widetilde{w}_j(\widehat{\kappa})\| < \alpha^{-\ell_j^\circ(\widehat{\kappa}) + \left(\left\|D_j^{(0)}\right\| - \sqrt{\frac{2\mathcal{L}(0)}{n}}\right)\widehat{\kappa}}.$$

Then, Equation (5) gives that for $\widehat{\kappa}$ sufficiently small, α may be chosen sufficiently small that for any time $t \in [\widehat{\kappa}, \kappa]$

$$\left\|D_j^{(0)}\right\|\left(1 - \alpha^{\left\|D_j^{(0)}\right\|\widehat{\kappa}/2}\right) < \widetilde{\overline{w}}_j^\top(t) D_j^{(0)} \leq \left\|D_j^{(0)}\right\|. \tag{6}$$

Note that by definition of τ, $|h_{\widetilde{\theta}}(x_i)| < m\alpha$ for all $i \in [n]$. Putting this back into Equation (3), we obtain as an upper bound that

$$\begin{aligned}\|\widetilde{w}_j(\kappa)\| &< \|\widetilde{w}_j(\widehat{\kappa})\|\alpha^{-\left(\left\|D_j^{(0)}\right\|+m\alpha\right)(\kappa-\widehat{\kappa})} \\ &< \alpha^{-\ell_j^\circ(\kappa)+\left(\left\|D_j^{(0)}\right\|-\sqrt{\frac{2\mathcal{L}(0)}{n}}\right)\widehat{\kappa}-m\alpha(\kappa-\widehat{\kappa})} \\ &< \alpha^{-\ell_j^\circ(\kappa)+\left(\left\|D_j^{(0)}\right\|-2\sqrt{\frac{2\mathcal{L}(0)}{n}}\right)\widehat{\kappa}}.\end{aligned}$$

For the lower bound, we get

$$\begin{aligned}\|\widetilde{w}_j(\kappa)\| &> \|\widetilde{w}_j(\widehat{\kappa})\|\exp_\alpha\left(-\left(\left\|D_j^{(0)}\right\|\left(1-\alpha^{\left\|D_j^{(0)}\right\|\widehat{\kappa}/2}\right)-m\alpha\right)(\kappa-\widehat{\kappa})\right) \\ &> \exp_\alpha\left(-\ell_j^\circ(\kappa)+\left(\left\|D_j^{(0)}\right\|+\sqrt{\frac{2\mathcal{L}(0)}{n}}\right)\widehat{\kappa}+\left(\left\|D_j^{(0)}\right\|\alpha^{\left\|D_j^{(0)}\right\|\widehat{\kappa}/2}+m\alpha\right)(\kappa-\widehat{\kappa})\right) \\ &> \alpha^{-\ell_j^\circ(\kappa)+\left(\left\|D_j^{(0)}\right\|+2\sqrt{\frac{2\mathcal{L}(0)}{n}}\right)\widehat{\kappa}},\end{aligned}$$

where the final inequality is valid for α sufficiently small. Since $\widehat{\kappa}$ is arbitrary, we conclude that given any $\xi > 0$, we may select α sufficiently small that the deviation of the neurons' norms from those predicted by the Algorithm 1 does not exceed ξ. □

Proof of Item 3. We compute that for some $y_i < 0$ such that $\widetilde{\overline{w}}_j^\top(0)x_i > 0$, and for $t \in [0, \kappa]$,

$$\begin{aligned}\frac{1}{\log(1/\alpha)}\frac{\mathrm{d}}{\mathrm{d}t}\widetilde{\overline{w}}_j^\top x_i &= \mathcal{D}_j^\top x_i - \mathcal{D}_j^\top\widetilde{\overline{w}}_j\widetilde{\overline{w}}_j^\top x_i \\ &= \frac{1}{n}(y_i - h_{\widetilde{\theta}}(x_i)) + \frac{1}{n}\sum_{v\,|\,\widetilde{\overline{w}}_j^\top x_v>0}(h_{\widetilde{\theta}}(x_v)-y_v)\widetilde{\overline{w}}_j^\top x_v\widetilde{\overline{w}}_j^\top x_i \\ &\le \frac{1}{n}y_i - \frac{1}{n}\sum_{\substack{v\,|\,y_v<0,\\ \widetilde{\overline{w}}_j^\top x_v>0}} y_v\widetilde{\overline{w}}_j^\top x_v\widetilde{\overline{w}}_j^\top x_i + 2m\alpha.\end{aligned}$$

This is obtained by plugging in the bound on the network output. This bound also implies that for α sufficiently small, $\mathbb{1}(y_v > 0)(y_v - h_\theta(x_v)) > 0$, which allowed us to discard the sum over positively labeled data. Since neurons are aligned by time $\widehat{\kappa}$, we conclude using Equation (6) and Proposition 8 that

$$\frac{\mathrm{d}}{\mathrm{d}t}\widetilde{\overline{w}}_j^\top x_i \le \log(1/\alpha)\left(\frac{y_i}{n} + \sqrt{2}\alpha^{\widehat{\kappa}\left\|D_j^{(0)}\right\|/4} + 2m\alpha\right),$$

where this holds for α sufficiently small and any $t \in [\widehat{\kappa}, \kappa]$. Therefore, by time $2\widehat{\kappa}$ and for α sufficiently small, $\widetilde{\overline{w}}_j^\top x_i = 0$ for all $i \notin I_{s_j}$. □

This completes the proof of Lemma 3. ■

From Lemma 3 and Assumption 2, we conclude that given any $\kappa > 0$, there exists α^* sufficiently small that $\theta_+ := (a_j, w_j)_{j\in J_+}$ satisfies

$$\dot{\theta}_+(t) \in -\partial\mathcal{L}_+(\theta_+(t)) \text{ for almost all } t \ge \kappa,$$

where

$$\mathcal{L}_+(\theta_+) \coloneqq \frac{1}{2n}\sum_{i\in I_+}(h_{\theta_+}(x_i)-y_i)^2.$$

By the symmetry of the signs, an analogous statement is true of $\theta_- := (a_i, w_j)_{j\in J_-}$. Therefore, on the interval $[\kappa,\infty)$ neurons in the set J_+ evolve independently from those in the set J_- according to the autonomous systems just described above. Consequently, we can treat the systems θ_+ and θ_- independently from now on, which we do by assuming w.l.o.g. $s_j = 1$ for all $j \in [m]$ and $\operatorname{sgn} y_i = 1$ for all $i \in [n]$. We note that with this assumption, $A_{i,j} = \mathbb{1}\left(\widetilde{w}_j^\top(0)x_i > 0\right)$.

Similar autonomous systems arguments have been used in Boursier and Flammarion [2025b, Appendix F.3], but are even simpler here due to the orthogonal data assumption.

D.2 Phase 1: Slow growth

In this stage, we verify that up to a time arbitrarily close to the next jump time, t_{k+1}, all neurons in the set $N_U^{(k)}$ are still small and aligned in the directions $D_j^{(k)}$, and all neurons in the set $N_F^{(k)}$ have not changed significantly. For $\varepsilon > 0$, define the time

$$\tau_1 := \inf\Big\{ t \ge \kappa \,\Big|\, \exists\, j \in [m] \text{ s.t. } \widetilde{\overline{w}}_j^\top \overline{D}_j^{(k)} \le 1-\alpha^{\varepsilon/2} \tag{7}$$

$$\text{or } \exists\, i \in S_F^{(k)} \text{ s.t. } |h_{\widetilde{\theta}}(x_i)-y_i| \ge \alpha^{\varepsilon} \tag{8}$$

$$\text{or } \exists\, j \in N_U^{(k)} \text{ s.t. } \|\widetilde{w}_j\| \ge \alpha^{\varepsilon}\Big\} \bigwedge t_{k+1}. \tag{9}$$

At this point, we make some useful observations on differentiability. Note that in the positive autonomous system, we have the inequalities $a_j\sigma\left(w_j^\top x_i\right) \le h(x_i) \le |y_i| + \sqrt{2n\mathcal{L}(0)}$. Hence, $\left(w_j^\top x_i\right)^2 \le |y_i| + \sqrt{2n\mathcal{L}(0)}$, implying boundedness of the parameters. Since the Clarke subdifferential of $\mathcal{L}$ is locally bounded and solutions of the differential inclusion are absolutely continuous, we conclude that the trajectories defined by Equation (1) are Lipschitz continuous on $[0,\infty)$. We may therefore appeal to Rademacher's theorem to see that compositions of Lipshitz functions with the parameters are differentiable almost everywhere for α fixed. It will also be useful to note that the arguments proving Lemma 3 give the lower bound $\|\widetilde{w}_j(t)\| \ge \alpha^{1+\sqrt{\frac{2\mathcal{L}(0)}{n}}t}$, so $t \mapsto \widetilde{\overline{w}}_j$ is Lipschitz on compact time intervals. In subsequent arguments, we may appeal briefly to Rademacher's theorem for differentiability, and it is implicit that this argument has been used.

In the result below and throughout the account of subsequent phases, we use the notation $j^\star$ in place of $j_\star^{(k)}$ for the neuron which grows to constant order when passing from saddle k to saddle $k+1$ to improve readability. Assuming Hypothesis 1, we will prove in the lemma below that on the interval $[\kappa,\tau_1]$, the alignment of all neurons remains stable and the loss on fitted data does not increase significantly. We will also show that provided ε is selected sufficiently small we have $\|\widetilde{w}_{j^\star}(\tau_1)\| = \alpha^{\varepsilon}$, meaning neuron $j^\star$ is the largest of all of the unfitted neurons at time τ_1. We will also prove that as ε decreases, τ_1 approaches t_{k+1}.

Lemma 4. *Suppose that Hypothesis 1 holds. Then for all $\kappa \in (t_k, t_{k+1})$, there exists ε^* such that for all $\varepsilon \in (0,\varepsilon^*)$ there exists $\alpha^* > 0$ such that for all $\alpha \in (0,\alpha^*)$ the following are true:*

1. $\forall i \in S_F^{(k)}, \forall t \in [\kappa,\tau_1],\ |h_{\widetilde{\theta}}(x_i)-y_i| < \alpha^{\varepsilon}$.
2. $\forall j \in [m], \forall t \in [\kappa,\tau_1],\ \widetilde{\overline{w}}_j^\top \overline{D}_j^{(k)} > 1-\alpha^{\varepsilon/2}$.
3. $\forall j \in N_F^{(k)}, \forall t \in [\kappa,\tau_1],\ \sqrt{n\left\|D_j^{(k)}\right\|} - \alpha^{\varepsilon/4} < \|\widetilde{w}_j\| < \sqrt{n\left\|D_j^{(k)}\right\|} + \alpha^{\varepsilon/4}$.
4. $t_{k+1} - \frac{3\varepsilon}{2\left\|D_{j^\star}^{(k)}\right\|} < \tau_1 < t_{k+1} - \frac{\varepsilon}{2\left\|D_{j^\star}^{(k)}\right\|}$.

5. $\|\widetilde{w}_{j^\star}(\tau_1)\| = \alpha^{\varepsilon}$.

6. $\forall j \in N_U^{(k)}$, $\forall t \in [\kappa, \tau_1]$, $\left|\ell_j^\circ(t) + \log_\alpha(\|\widetilde{w}_j\|)\right| < \frac{\varepsilon}{2}$.

Proof. Proof of Item 1 Lemma 4. Let δ be such that the conditions of Hypothesis 1 hold. In order to check that $|h_{\widetilde{\theta}}(x_i) - y_i| < \alpha^{\varepsilon}$, we will bound the growth in norm of the quantity

$$E^{(k)} = -\frac{1}{n} \sum_{i \in S_F^{(k)}} \left(h_{\widetilde{\theta}}(x_i) - y_i\right) x_i.$$

To begin this section, it will be useful to note now that

$$\begin{aligned} \|\mathcal{D}_j\| &\leq \frac{1}{n} \sqrt{\left(\sum_{i \in [n]} (h_{\widetilde{\theta}}(x_i) - y_i)^2\right)} \\ &\leq \sqrt{\frac{2\mathcal{L}(0)}{n}}, \end{aligned} \tag{10}$$

where we have applied the Cauchy-Schwarz inequality and the monotonicity of the gradient flow. We will use this inequality repeatedly. Using Rademacher's theorem for differentiability, we compute that

$$\begin{aligned} \frac{1}{2\log(1/\alpha)} \frac{\mathrm{d}}{\mathrm{d}t} \left\|E^{(k)}\right\|^2 &= \frac{1}{\log(1/\alpha)} E^{(k)\top} \frac{\mathrm{d}}{\mathrm{d}t} E^{(k)} \\ &= -\frac{1}{n} \sum_{i \in S_F^{(k)}} \sum_{j \in [m]} A_{i,j} \|\widetilde{w}_j\|^2 \left(\left(x_i^\top E^{(k)}\right)^2 + \mathcal{D}_j^\top \widetilde{\overline{w}}_j \widetilde{\overline{w}}_j^\top x_i x_i^\top E^{(k)}\right) \\ &\leq -\frac{1}{n} \sum_{i \in S_F^{(k)}} \sum_{j \in [m]} A_{i,j} \|\widetilde{w}_j\|^2 \mathcal{D}_j^\top \widetilde{\overline{w}}_j \widetilde{\overline{w}}_j^\top x_i x_i^\top E^{(k)} \\ &\leq -\frac{1}{n} \sum_{i \in S_F^{(k)}} \sum_{j \in N_U^{(k)}} A_{i,j} \|\widetilde{w}_j\|^2 \mathcal{D}_j^\top \widetilde{\overline{w}}_j \widetilde{\overline{w}}_j^\top x_i x_i^\top E^{(k)}, \end{aligned} \tag{11}$$

where the final inequality follows from the fact that for any $j \in N_F^{(k)}$, the sum may be written as

$$-\frac{\|\widetilde{w}_j\|^2}{n^3} \left(\sum_{v \in [n]} A_{v,j} (h_{\widetilde{\theta}}(x_v) - y_v) \widetilde{\overline{w}}_j^\top x_v\right) \left(\sum_{i \in S_F^{(k)}} A_{i,j} (h_{\widetilde{\theta}}(x_i) - y_i) \widetilde{\overline{w}}_j^\top x_i\right) \leq 0.$$

The above inequality follows from noting that, since we are considering some $j \in N_F^{(k)}$, $\{v \in [n] \mid A_{v,j} = 1\} = \{i \in S_F^{(k)} \mid A_{i,j} = 1\}$ (such neurons are by definition active only on fitted data) so this is a square. We have therefore only to obtain a bound on the contribution of those $j \in N_U^{(k)}$. Note that for any $j \in [m]$, the fact that $\widetilde{\overline{w}}_j$ is a unit vector combined with (10) implies

$$|\widetilde{\overline{w}}_j^\top \mathcal{D}_j| \leq \|\mathcal{D}_j\| \leq \sqrt{\frac{2\mathcal{L}(0)}{n}}.$$

Also, we have for any $j \in N_U^{(k)}$ by (7) and Proposition 8 that $\widetilde{\overline{w}}_j^\top x_i < \sqrt{2}\alpha^{\varepsilon/4}$ on $[\kappa, \tau_1]$. We combine these observations with (9) to obtain

$$\begin{aligned}\frac{1}{\log(1/\alpha)}E^{(k)\top}\frac{\mathrm{d}}{\mathrm{d}t}E^{(k)} &\leq \frac{1}{n}\sum_{i\in S_F^{(k)}} m\sqrt{\frac{4\mathcal{L}(0)}{n}}\alpha^{9\varepsilon/4}\left|x_i^\top E^{(k)}\right| \\ &\leq \frac{2m}{n}\sqrt{\mathcal{L}(0)}\alpha^{9\varepsilon/4}\left\|E^{(k)}\right\|,\end{aligned}$$

where we have used the Cauchy-Schwarz inequality in the final line. This affords us the bound

$$\begin{aligned}\frac{\mathrm{d}}{\mathrm{d}t}\left\|E^{(k)}\right\| &\leq \frac{2m}{n}\sqrt{\mathcal{L}(0)}\log(1/\alpha)\alpha^{9\varepsilon/4} \\ &< \alpha^{2\varepsilon},\end{aligned}$$

for α sufficiently small. Since $|h_{\widetilde{\theta}^\kappa}(x_i) - y_i| \leq \alpha^\delta$ by Item 3 Hypothesis 1, we have $\left\|E^{(k)}(\kappa)\right\| < \alpha^\delta/\sqrt{n}$. Integrating and using this initial condition, we have

$$\begin{aligned}\left\|E^{(k)}\right\| &< \left\|E^{(k)}(\kappa)\right\| + (t-\kappa)\alpha^{2\varepsilon} \\ &< \frac{\alpha^\delta}{\sqrt{n}} + (t_{k+1}-\kappa)\alpha^{2\varepsilon} \\ &< \alpha^{3\varepsilon/2},\end{aligned}$$

for $\varepsilon < \frac{\delta}{2}$ and α sufficiently small. This implies that

$$\begin{aligned}\max_{i\in S_F^{(k)}}|h_{\widetilde{\theta}}(x_i) - y_i| &< n\alpha^{3\varepsilon/2} \\ &< \alpha^\varepsilon,\end{aligned}$$

for all $t \in [\kappa, \tau_1]$, concluding the proof of this item. □

Proof of Item 2 Lemma 4. We verify now the stability of alignment of the neurons. We split this into the case $j \in N_F^{(k)}$ and the case $j \in N_U^{(k)}$. For the former, first note that for $j \in N_F^{(k)}$, $\{i \in S_U^{(k)} \mid A_{i,j} = 1\} = \emptyset$ by definition. Consequently,

$$\left\|\mathcal{D}_j\right\| < \frac{1}{\sqrt{n}}\alpha^\varepsilon.$$

Inserting this into the differential equation governing tangential dynamics, we get

$$\begin{aligned}\frac{1}{\log(1/\alpha)}\frac{\mathrm{d}}{\mathrm{d}t}\widetilde{\overline{w}}_j^\top D_j^{(k)} &= \mathcal{D}_j^\top D_j^{(k)} - \mathcal{D}_j^\top\widetilde{\overline{w}}_j\widetilde{\overline{w}}_j^\top D_j^{(k)} \\ &\geq -2\left\|D_j^{(k)}\right\|\left\|\mathcal{D}_j\right\| \\ &> -\frac{2\left\|D_j^{(k)}\right\|}{\sqrt{n}}\alpha^\varepsilon, \qquad (12)\end{aligned}$$

We simply integrate, using Item 4 Hypothesis 1 for the alignment value at time κ, to obtain

$$\begin{aligned}\widetilde{\overline{w}}_j^\top D_j^{(k)} &> \left\|D_j^{(k)}\right\|\left(1-\alpha^\delta\right) - \frac{2\left\|D_j^{(k)}\right\|}{\sqrt{n}}(t_{k+1}-\kappa)\log(1/\alpha)\alpha^\varepsilon \\ &> \left\|D_j^{(k)}\right\|\left(1-\alpha^{\varepsilon/2}\right),\end{aligned}$$

where the final inequality is again valid for ε and α sufficiently small. For the case $j \in N_U^{(k)}$, we first use Equation (9) to bound the network output on data $i \in S_U^{(k)}$:

$$\begin{aligned} h_{\widetilde{\theta}}(x_i) &= \sum_{j \in [m]} \|\widetilde{w}_j\|^2 \left(\widetilde{\overline{w}}_j^\top x_i\right)_+ \\ &< m\alpha^{2\varepsilon}. \end{aligned} \tag{13}$$

Using this, we estimate the alignment of an arbitrary neuron $j \in N_U^{(k)}$ with direction $D_j^{(k)}$ as follows:

$$\begin{aligned} \frac{1}{\log(1/\alpha)} \frac{\mathrm{d}}{\mathrm{d}t} \widetilde{\overline{w}}_j^\top D_j^{(k)} &= \left\|D_j^{(k)}\right\|^2 - \left(\widetilde{\overline{w}}_j^\top D_j^{(k)}\right)^2 \\ &\quad - \frac{1}{n} \sum_{i \in S_j^{(k)}} h_{\widetilde{\theta}}(x_i) \left(x_i^\top D_j^{(k)} - x_i^\top \widetilde{\overline{w}}_j \widetilde{\overline{w}}_j^\top D_j^{(k)}\right) \\ &\quad - \frac{1}{n} \sum_{i \in S_F^{(k)}} A_{i,j}(h_{\widetilde{\theta}}(x_i) - y_i) \left(x_i^\top D_j^{(k)} - x_i^\top \widetilde{\overline{w}}_j \widetilde{\overline{w}}_j^\top D_j^{(k)}\right) \\ &> \left\|D_j^{(k)}\right\|^2 - \left(\widetilde{\overline{w}}_j^\top D_j^{(k)}\right)^2 - 2\left\|D_j^{(k)}\right\| m\alpha^{9\varepsilon/4} - 2\left\|D_j^{(k)}\right\| \alpha^{5\varepsilon/4} \\ &> \left\|D_j^{(k)}\right\|^2 - \left(\widetilde{\overline{w}}_j^\top D_j^{(k)}\right)^2 - \alpha^{6\varepsilon/5}. \end{aligned} \tag{14}$$

The first inequality is an application of Corollary 1 to both sums (given (7)) as well as using (13) for the sum over $i \in S_j^{(k)}$ and (8) for the sum over $i \in S_F^{(k)}$. The second inequality follows for α sufficiently small. The above implies that the solutions of the differential equation $f'(t) = b(a^2 - f(t)^2)$, where $a^2 = \left\|D_j^{(k)}\right\|^2 - \alpha^{6\varepsilon/5}$ and $b = \log(1/\alpha)$, furnish us with a lower bound on the alignment. Solutions of this differential equation approach a limiting value of a monotonically. Since $a > \left\|D_j^{(k)}\right\| - \alpha^{3\varepsilon/5}$ from the inequality $\sqrt{1-x^2} > 1 - x$, and Item 4 Hypothesis 1 allows us to lower bound the alignment value at time κ, we have that for all $t \in [\kappa, \tau_1]$,

$$\widetilde{\overline{w}}_j^\top D_j^{(k)} > \left\|D_j^{(k)}\right\| \min\left(1 - \alpha^\delta, 1 - \frac{\alpha^{3\varepsilon/5}}{\left\|D_j^{(k)}\right\|}\right) > \left\|D_j^{(k)}\right\| \left(1 - \alpha^{\varepsilon/2}\right),$$

for ε and α sufficiently small. □

Proof of Item 3 Lemma 4. We check that all the neurons in the set $N_F^{(k)}$ cannot change in norm significantly in the first phase. Recall that for $j \in N_F^{(k)}$, $\{i \in S_U^{(k)} \mid A_{i,j} = 1\} = \emptyset$. Therefore, the differential equation governing radial dynamics admits the bound

$$\begin{aligned} \frac{1}{\log(1/\alpha)} \frac{\mathrm{d}}{\mathrm{d}t} \|\widetilde{w}_j\| &= \mathcal{D}_j^\top \widetilde{w}_j \\ &\geq -\left\|E^{(k)}\right\| \|\widetilde{w}_j\|. \end{aligned}$$

Applying Grönwall's inequality combined with the initial conditions from Item 2 Hypothesis 1, we get that for any $t \in [\kappa, \tau_1]$ and $j \in N_F^{(k)}$,

$$\begin{aligned}
\|\widetilde{w}_j\| &\geq \|\widetilde{w}_j(\kappa)\| \exp\left(-\log(1/\alpha)\int_\kappa^{\tau_1} \left\|E^{(k)}\right\| ds\right) \\
&> \|\widetilde{w}_j(\kappa)\| \exp\left(-\log(1/\alpha)\frac{\alpha^\varepsilon}{\sqrt{n}}(t_{k+1}-\kappa)\right) \\
&> \|\widetilde{w}_j(\kappa)\| \exp\left(-\alpha^{\varepsilon/2}\right) \\
&\geq \|\widetilde{w}_j(\kappa)\| \left(1-\alpha^{\varepsilon/2}\right) \\
&> \sqrt{n\left\|D_j^{(k)}\right\| - \alpha^\delta}\left(1-\alpha^{\varepsilon/2}\right) \\
&> \sqrt{n\left\|D_j^{(k)}\right\|} - \alpha^{\varepsilon/4}. \qquad (15)
\end{aligned}$$

The above inequalities are valid for ε and α sufficiently small. An analogous argument yields the claimed upper bound

$$\|\widetilde{w}_j(t)\| < \sqrt{n\left\|D_j^{(k)}\right\|} + \alpha^{\varepsilon/4}.$$

□

Proof of Item 6 Lemma 4. To estimate the growth of any neuron $j \in N_U^{(k)}$, we compute that

$$\begin{aligned}
\frac{1}{\log(1/\alpha)}\frac{\mathrm{d}}{\mathrm{d}t}\|\widetilde{w}_j\| &= -\frac{\|\widetilde{w}_j\|}{n}\sum_{i\in[n]} A_{i,j}(h_{\widetilde{\theta}}(x_i)-y_i)\widetilde{\overline{w}}_j^\top x_i \\
&> \left(\left\|D_j^{(k)}\right\|\left(1-\alpha^{\varepsilon/2}\right) - m\alpha^{2\varepsilon}\right)\|\widetilde{w}_j\| - \frac{\|\widetilde{w}_j\|}{n}\sum_{i\notin S_j^{(k)}} A_{i,j}(h_{\widetilde{\theta}}(x_i)-y_i)\widetilde{\overline{w}}_j^\top x_i \\
&> \left(\left\|D_j^{(k)}\right\|\left(1-\alpha^{\varepsilon/2}\right) - m\alpha^{2\varepsilon}\right)\|\widetilde{w}_j\| - \sqrt{2}\alpha^{5\varepsilon/4}\|\widetilde{w}_j\|. \qquad (16)
\end{aligned}$$

The first inequality is from the assumption on alignment, (7), combined with the assumed bound on the norms of the unfitted neurons, (9). The second is the combination of this assumed alignment with (8) and (8). The above implies that for α sufficiently small,

$$\frac{\mathrm{d}}{\mathrm{d}t}\|\widetilde{w}_j\| > \log(1/\alpha)\left(\left\|D_j^{(k)}\right\| - 2\alpha^{\varepsilon/2}\right)\|\widetilde{w}_j\|.$$

Grönwall's inequality then gives that for $t \in [\kappa, \tau_1]$,

$$\begin{aligned}
\|\widetilde{w}_j\| &> \|\widetilde{w}_j(\kappa)\| \exp\left(\log(1/\alpha)\left(\left\|D_j^{(k)}\right\| - 2\alpha^{\varepsilon/2}\right)(t-\kappa)\right) \\
&= \|\widetilde{w}_j(\kappa)\|\alpha^{-\left\|D_j^{(k)}\right\|(t-\kappa)+2\alpha^{\varepsilon/2}(t-\kappa)} \\
&\overset{(1)}{>} \alpha^{-\ell_j^\circ(\kappa)-\left\|D_j^{(k)}\right\|(t-\kappa)+\varepsilon/3+2\alpha^{\varepsilon/2}(t_{k+1}-\kappa)} \\
&\overset{(2)}{>} \alpha^{-\ell_j^\circ(\kappa)-\left\|D_j^{(k)}\right\|(t-\kappa)+\varepsilon/2} \\
&= \alpha^{-\ell_j^\circ(t)+\varepsilon/2}.
\end{aligned}$$

In the above, (1) is from an application of Item 1 Hypothesis 1 with the choice $\xi = \varepsilon/3$ and (2) is valid for α sufficiently small. For the upper bound, we compute in a similar manner that for any $j \in N_U^{(k)}$,

$$\frac{1}{\log(1/\alpha)}\frac{\mathrm{d}}{\mathrm{d}t}\|\widetilde{w}_j\| = -\frac{\|w_j\|}{n}\sum_{i\in[n]} A_{i,j}(h_{\widetilde{\theta}}(x_i) - y_i)\widetilde{\overline{w}}_j^{\top} x_i$$
$$< \left(\left\|D_j^{(k)}\right\| + \sqrt{2}\alpha^{5\varepsilon/4}\right)\|\widetilde{w}_j\|.$$

Again by Grönwall's inequality combined with the initial conditions from Hypothesis 1 with $\xi = \varepsilon/3$, we have the bound

$$\begin{aligned}\|\widetilde{w}_j\| &< \alpha^{-\ell_j^\circ(\kappa) - \left\|D_j^{(k)}\right\|(t-\kappa) - \varepsilon/3 - \sqrt{2}\alpha^{5\varepsilon/4}(t_{k+1}-\kappa)} \\ &< \alpha^{-\ell_j^\circ(t)-\varepsilon/2}. \end{aligned} \tag{17}$$

□

Proof of Items 4 and 5 Lemma 4. Note that for $t, t' \in [t_k, t_{k+1}]$, $-\ell_j^\circ(t) - \left\|D_j^{(k)}\right\|(t'-t) = -\ell_j^\circ(t')$. Using this fact and that $\|\widetilde{w}_j\| < \alpha^\varepsilon$ on $[\kappa, \tau_1)$, we obtain that for all $j \in N_U^{(k)}$,

$$\begin{aligned}\varepsilon &< -\ell_j^\circ(\tau_1) + \frac{\varepsilon}{2} \\ &= -\ell_j^\circ(t_{k+1}) - \left\|D_j^{(k)}\right\|(\tau_1 - t_{k+1}) + \frac{\varepsilon}{2}.\end{aligned}$$

We have $-\ell_{j^\star}^\circ(t_{k+1}) = 0$, and hence the upper bound τ_1 of

$$\tau_1 < t_{k+1} - \frac{\varepsilon}{2\left\|D_{j^\star}^{(k)}\right\|}. \tag{18}$$

To verify that neuron $j^\star$ is the largest at time τ_1, note the upper bound on the neurons' norms provided by (17) is monotone increasing in t. From this observation followed by inserting the upper bound on τ_1 derived in Equation (18), we get

$$\begin{aligned}\|\widetilde{w}_j(\tau_1)\| &< \exp_\alpha\left(-\ell_j^\circ(\tau_1) - \frac{\varepsilon}{2}\right) \\ &< \exp_\alpha\left(-\ell_j^\circ(t_{k+1}) + \frac{1}{2}\left(\frac{\left\|D_j^{(k)}\right\|}{\left\|D_{j^\star}^{(k)}\right\|} - 1\right)\varepsilon\right).\end{aligned} \tag{19}$$

Note that $-\ell_j^\circ(t_{k+1}) > 0$ for $j \in N_U^{(k)} \setminus \{j^\star\}$. This implies that for such j there exists ε sufficiently small that

$$-\ell_j^\circ(t_{k+1}) + \frac{1}{2}\left(\frac{\left\|D_j^{(k)}\right\|}{\left\|D_{j^\star}^{(k)}\right\|} - 1\right)\varepsilon > \varepsilon.$$

Hence, $\|\widetilde{w}_j(\tau_1)\| < \alpha^\varepsilon$ for $j \in N_U^{(k)} \setminus \{j^\star\}$. This proves that no neuron other than $j^\star$ could have $\|\widetilde{w}_j(\tau_1)\| = \alpha^\varepsilon$ for ε sufficiently small, and since no other condition breaks we have $\|\widetilde{w}_{j^\star}(\tau_1)\| = \alpha^\varepsilon$. We obtain a lower bound on τ_1 by considering that

$$\varepsilon > -\left\|D_{j^\star}^{(k)}\right\|(\tau_1 - t_{k+1}) - \frac{\varepsilon}{2}$$

imposes the condition

$$\tau_1 > t_{k+1} - \frac{3\varepsilon}{2\left\|D_{j^\star}^{(k)}\right\|}. \tag{20}$$

Note that by selecting ε sufficiently small depending on κ, we can indeed ensure that $\tau_1 > \kappa$. □

This completes the proof of Lemma 4. ■

D.3 Phase 2a: Growth to constant order

In this section, we deal with the rapid growth of neuron $j^\star$, the account of which is split into two parts. To begin, it will be important to obtain a uniform bound on the norms of neurons in the set $j \in N_U^{(k)} \setminus \{j^\star\}$, in much the same way that in the proof of Lemma 4 we had by definition of τ_1 that $\|\widetilde{w}_j\| < \alpha^\varepsilon$ on $[\kappa, \tau_1]$ or in the proof of Lemma 3 we had $\|\widetilde{w}_j\| < \alpha^{1/2}$. The idea of the proof is to verify that, given the upper bound on neurons in the set $N_U^{(k)} \setminus \{j^\star\}$ at time τ_1, if these neurons were to subsequently grow at the fastest possible rate, they would still not have exceeded some threshold by a certain time after τ_1.

Proposition 3. *For all $\kappa \in (t_k, t_{k+1})$, there exists $\varepsilon^* > 0$ such that for all $\varepsilon \in (0, \varepsilon^*)$ there exists $\alpha^* > 0$ such that for all $\alpha \in (0, \alpha^*)$ the following holds:*

$$\forall j \in N_U^{(k)} \setminus \{j^\star\}, \forall t \in \left[\tau_1, \tau_1 + 2\sqrt{\frac{n}{2\mathcal{L}(0)}}c_U\right], \|\widetilde{w}_j\| < \alpha^{c_U}, \tag{21}$$

where $c_U := -\frac{1}{4}\max_{j' \in N_U^{(k)} \setminus \{j^\star\}} \ell_{j'}^\circ(t_{k+1})$.

Proof. Recall that for any $j \in [m]$,

$$\frac{1}{\log(1/\alpha)}\frac{\mathrm{d}}{\mathrm{d}t}\|\widetilde{w}_j\| \leq \sqrt{\frac{2\mathcal{L}(0)}{n}}\|\widetilde{w}_j\|.$$

Recall also the upper bound on neurons in the set $N_U^{(k)} \setminus \{j^\star\}$ at time τ_1 provided by (19). Using Grönwall's inequality we obtain from these observations that on the considered interval,

$$\begin{aligned}
\|\widetilde{w}_j\| &\leq \|\widetilde{w}_j(\tau_1)\| \exp\left(\sqrt{\frac{2\mathcal{L}(0)}{n}}\log(1/\alpha)(t-\tau_1)\right) \\
&< \exp_\alpha\left(-\ell_j^\circ(\tau_1) - \frac{\varepsilon}{2} - \sqrt{\frac{2\mathcal{L}(0)}{n}}(t-\tau_1)\right) \\
&< \exp_\alpha\left(-\ell_j^\circ(t_{k+1}) - \frac{\varepsilon}{2} - \sqrt{\frac{2\mathcal{L}(0)}{n}}(t-\tau_1)\right) \\
&\leq \exp_\alpha\left(-\ell_j^\circ(t_{k+1}) - \frac{\varepsilon}{2} + \frac{1}{2}\max_{j' \in N_U^{(k)} \setminus \{j^\star\}} \ell_{j'}(t_{k+1})\right) \\
&\leq \exp_\alpha\left(-\frac{1}{2}\max_{j' \in N_U^{(k)} \setminus \{j^\star\}} \ell_j^\circ(t_{k+1}) - \frac{\varepsilon}{2}\right) \\
&< \exp_\alpha\left(-\frac{1}{4}\max_{j' \in N_U^{(k)} \setminus \{j^\star\}} \ell_j^\circ(t_{k+1})\right).
\end{aligned}$$

The final inequality is valid for ε sufficiently small, proving the claim. ■

We will throughout the account of several subsequent results assume that ε is selected small enough that $\varepsilon < c_U$, meaning that, informally, contributions of order c_U can be thought of as negligible relative to those of order ε. We define the quantity

$$\eta := \frac{1}{2}\min\left(\frac{\sqrt{n}}{60\sqrt{2}}\frac{\left\|D_{j^\star}^{(k)}\right\|}{\sqrt[4]{2\mathcal{L}(0)}}, \sqrt{n\left\|D_{j^\star}^{(k)}\right\|}\right).$$

This is a constant threshold which we will use to break the account of the growth of neuron $j^\star$ in two. We do so by defining

$$\tau_2 := \inf\Big\{t \geq \tau_1 \mid \widetilde{\overline{w}}_{j^\star}^\top \overline{D}_{j^\star}^{(k)} \leq 1 - \alpha^{\varepsilon/12} \tag{22}$$

$$\text{or } \exists\, i \in S_F^{(k)} \text{ s.t. } |h_{\widetilde{\theta}}(x_i) - y_i| \geq \alpha^{\varepsilon/6} \tag{23}$$

$$\text{or } \|\widetilde{w}_{j^\star}\| \geq \eta\Big\} \bigwedge \left(\tau_1 + 2\sqrt{\frac{n}{2\mathcal{L}(0)}}c_U\right). \tag{24}$$

We will prove in the lemma below that $\|\widetilde{w}_{j^\star}(\tau_2)\| = \eta$, meaning that neuron $j^\star$ has grown to constant order. In order to do so, we will verify that the loss on fitted data does not increase significantly during this phase. As a consequence, the alignment of neurons in the set $N_F^{(k)} \cup \{j^\star\}$ remains stable.

The use of the constant threshold η is that there is a tradeoff between proving the growth conditions and the stability of the loss on fitted data. Specifically, one must ensure that the growth phase lasts long enough that neuron $j^\star$ can grow to constant order. On the other hand, this increases the time interval on which one must control $\left\|E^{(k)}\right\|$ and increases the coefficients of terms in (11). Introducing this constant threshold, the specific choice of which simply emerges as a consequence of the way we bound $\left\|E^{(k)}\right\|$, allows us to balance these two competing aims across both parts of the growth phase.

The final stopping condition given by the time $\tau_1 + 2\sqrt{\frac{n}{2\mathcal{L}(0)}}c_U$ is required to ensure, using Proposition 3, that we have a uniform bound on all of the neurons in the set $N_U^{(k)} \setminus \{j^\star\}$ on the interval $[\tau_1, \tau_2]$. That we apply this kind of bound is reflective of the fact we do not consider at all the precise alignment or growth rates of the neurons in the set $N_U^{(k)} \setminus \{j^\star\}$ during the growth phases; we only ensure that the duration of these phases is sufficiently short that these neurons do not change significantly.

Lemma 5. *Suppose the Hypothesis 1 holds. Then for all $\kappa \in (t_k, t_{k+1})$, there exists $\varepsilon^* > 0$ such that for all $\varepsilon \in (0, \varepsilon^*)$ there exists $\alpha^* > 0$ such that for all $\alpha \in (0, \alpha^*)$ the following are true:*

1. $\tau_2 - \tau_1 < \frac{3\varepsilon}{\left\|D_{j^\star}^{(k)}\right\|}$.
2. $\forall j \in N_F^{(k)}, \forall t \in [\tau_1, \tau_2],\ \sqrt{n\left\|D_j^{(k)}\right\|} - \alpha^{\varepsilon/8} < \|\widetilde{w}_j\| < \sqrt{n\left\|D_j^{(k)}\right\|} + \alpha^{\varepsilon/8}$.
3. $\forall t \in [\tau_1, \tau_2],\ \max_{i \in S_F^{(k)}} \left(\widetilde{\overline{w}}_{j^\star}^\top x_i\right)_+ < \alpha^{\varepsilon/6}$.
4. $\forall t \in [\tau_1, \tau_2],\ \widetilde{\overline{w}}_{j^\star}^\top \overline{D}_{j^\star}^{(k)} > 1 - \alpha^{\varepsilon/12}$.
5. $\forall i \in S_F^{(k)}, \forall t \in [\tau_1, \tau_2],\ |h_{\widetilde{\theta}}(x_i) - y_i| < \alpha^{\varepsilon/6}$.
6. $\|\widetilde{w}_{j^\star}(\tau_2)\| = \eta$.

We note at this point that although Item 4 Lemma 5 would seem to imply a statement on realignment with fitted data, we are required in the proof to obtain Item 3 Lemma 5 first in order to prove Item 4 Lemma 5. This is an interesting result of the fact that a bootstrapping argument, in which one would assume some level of alignment of neuron $j^\star$ in direction $D_{j^\star}^{(k)}$ and then insert this into (11), leads to

too weak a control on $\left\|E^{(k)}\right\|$ to be able to use this to get the assumed alignment back at the end (and vice versa). This circularity was not an issue in the arguments leading to (14), because in this setting we had a the growth conditions $\|\widetilde{w}_j\| \leq \alpha^{\varepsilon}$ for all $j \in N_U^{(k)}$ on the interval $[\kappa, \tau_1]$. The reader may also note that our statement includes no control on the alignment of neurons in the set $N_F^{(k)}$. This is because we can simply control the loss on fitted data and use this when needed to estimate the alignment of such neurons across several phases (as will ultimately be done in Lemma 8).

Proof. We consider the items in the statement of Lemma 5.

Proof of Item 1. We bound the phase duration above by examining the growth of neuron $j^\star$. We have that on $[\tau_1, \tau_2]$,

$$
\begin{aligned}
&\frac{1}{2\log(1/\alpha)}\frac{\mathrm{d}}{\mathrm{d}t}\|\widetilde{w}_{j^\star}\|^2 \\
&\quad = \|\widetilde{w}_{j^\star}\|^2 \left(-\frac{1}{n}\sum_{i\in S_F^{(k)}} A_{i,j^\star}(h_{\widetilde{\theta}}(x_i) - y_i)\widetilde{w}_{j^\star}^\top x_i - \frac{1}{n}\sum_{i\in S_{j^\star}^{(k)}} (h_{\widetilde{\theta}}(x_i) - y_i)\widetilde{w}_{j^\star}^\top x_i \right) \\
&\quad > \|\widetilde{w}_{j^\star}\|^2 \left(\|D_{j^\star}^{(k)}\|(1-\alpha^{\varepsilon/12}) - \alpha^{\varepsilon/6} - \frac{1}{n}\sum_{i\in S_{j^\star}^{(k)}} h_{\widetilde{\theta}}(x_i)\widetilde{w}_{j^\star}^\top x_i \right) \qquad (25) \\
&\quad = \|\widetilde{w}_{j^\star}\|^2 \Bigg(\|D_{j^\star}^{(k)}\|(1-\alpha^{\varepsilon/12}) - \alpha^{\varepsilon/6} \\
&\qquad\qquad - \frac{1}{n}\sum_{i\in S_{j^\star}^{(k)}} \left(\|\widetilde{w}_{j^\star}\|^2 \widetilde{w}_{j^\star}^\top x_i + \sum_{j\neq j^\star} A_{i,j}\|\widetilde{w}_j\|^2 \widetilde{w}_j^\top x_i \right) \widetilde{w}_{j^\star}^\top x_i \Bigg) \\
&\quad \geq \|\widetilde{w}_{j^\star}\|^2 \left(\|D_{j^\star}^{(k)}\|(1-\alpha^{\varepsilon/12}) - \alpha^{\varepsilon/6} - \frac{\|\widetilde{w}_{j^\star}\|^2}{n} - m\alpha^{2c_U} \right) \qquad (26) \\
&\quad > \|\widetilde{w}_{j^\star}\|^2 \left(\|D_{j^\star}^{(k)}\| - \alpha^{\varepsilon/13} \right) - \frac{\|\widetilde{w}_{j^\star}\|^4}{n}. \qquad (27)
\end{aligned}
$$

In the above, (25) is from the definition of τ_2 (where we have not used the alignment in the estimate of the sum over the fitted data as it is not necessary here) and (26) is from the fact that for any $i \in S_{j^\star}^{(k)}$, $\{j \in N_F^{(k)} \mid A_{i,j} = 1\} = \emptyset$. A function satisfying the differential inequality $f'(t) > 2\log(1/\alpha)af(t) - 2\log(1/\alpha)bf(t)^2$ has a lower bound

$$
f(t) > \frac{a}{b}\frac{\exp\left(2\log(1/\alpha)a(t-C)\right)}{1+\exp\left(2\log(1/\alpha)a(t-C)\right)} = \frac{a}{b}\frac{1}{\alpha^{2a(t-C)}+1},
$$

where C is an integration constant. In our case, $f = \|\widetilde{w}_{j^\star}\|^2$, $a = \left\|D_{j^\star}^{(k)}\right\| - \alpha^{\varepsilon/13}$, and $b = \frac{1}{n}$. We will obtain an upper bound on this constant C, since this will in turn provide a lower bound on $f(t)$ and hence an upper bound on τ_2. By Item 5 Lemma 4, $\|\widetilde{w}_{j^\star}(\tau_1)\|^2 = \alpha^{2\varepsilon}$. Simple algebra then gives

$$
C < \tau_1 - \frac{\log\left(\frac{a}{b}\alpha^{-2\varepsilon} - 1\right)}{2a\log(\alpha)}.
$$

This implies that for $t \in [\tau_1, \tau_2]$,

$$f(t) > \frac{a}{b}\frac{1}{1+\exp_\alpha\left(2a(t-\tau_1)+\frac{\log\left(\frac{a}{b}\alpha^{-2\varepsilon}-1\right)}{\log(\alpha)}\right)}.$$

Rearranging

$$\frac{a}{b}\frac{1}{1+\exp_\alpha\left(2a(t-\tau_1)+\frac{\log\left(\frac{a}{b}\alpha^{-2\varepsilon}-1\right)}{\log(\alpha)}\right)} < \eta^2$$

gives

$$\begin{aligned} t-\tau_1 &< \frac{1}{2a\log(\alpha)}\left(\log\left(\frac{a}{b\eta^2}-1\right)-\log\left(\frac{a}{b}\alpha^{-2\varepsilon}-1\right)\right) \\ &< \frac{1}{2a\log(\alpha)}\left(\log\left(\frac{a}{b\eta^2}-1\right)-\log\left(\frac{a}{b}\alpha^{-2\varepsilon}\right)\right) \\ &< \frac{3\varepsilon}{\left\|D_{j^\star}^{(k)}\right\|}, \end{aligned}$$

valid provided α is taken sufficiently small owing to the fact we take $\eta < \sqrt{n\left\|D_{j^\star}^{(k)}\right\|}$ fixed. □

Proof of Items 3 and 5. We derive an upper bound on the alignment of neuron $j^\star$ with data in the set $S_F^{(k)}$ in terms of $\left\|E^{(k)}\right\|$. We have for any $i \in S_F^{(k)}$

$$\begin{aligned} \frac{1}{\log(1/\alpha)}\frac{\mathrm{d}}{\mathrm{d}t}\widetilde{w}_{j^\star}^\top x_i = &-\frac{1}{n}A_{i,j^\star}(h_{\widetilde{\theta}}(x_i)-y_i) \\ &+\frac{1}{n}\sum_{v\in[n]}A_{v,j^\star}(h_{\widetilde{\theta}}(x_v)-y_v)\widetilde{w}_{j^\star}^\top x_v \widetilde{w}_{j^\star}^\top x_i. \end{aligned} \tag{28}$$

For the first term, if $i \in S_F^{(k)}$, then

$$-\frac{1}{n}(h_{\widetilde{\theta}}(x_i)-y_i) \le \left\|E^{(k)}\right\|. \tag{29}$$

For the sum, we obtain separate estimates over unfitted and fitted data. Recall that by construction, $A_{v,j} = 0$ for all $j \in N_F^{(k)}$ and $v \in S_U^{(k)}$. Therefore, considering that $S_{j^\star}^{(k)} \subseteq S_U^{(k)}$, the only contribution to the network output on such data is from neurons in the set $N_U^{(k)}$. This leads to the inequalities

$$\begin{aligned} h_{\widetilde{\theta}}(x_v)-y_v &= \|\widetilde{w}_{j^\star}\|^2\widetilde{w}_{j^\star}^\top x_v + \sum_{j\in N_U^{(k)}\setminus\{j^\star\}}\|\widetilde{w}_j\|^2\widetilde{w}_j^\top x_v - y_v \\ &\overset{(1)}{<} \|\widetilde{w}_{j^\star}\|^2\widetilde{w}_{j^\star}^\top x_v + m\alpha^{2c_U} - y_v \\ &\overset{(2)}{<} \eta^2\left(\frac{y_v}{n\left\|D_{j^\star}^{(k)}\right\|}+\sqrt{2}\alpha^{\varepsilon/24}\right)+m\alpha^{2c_U}-y_v \\ &\overset{(3)}{<} y_v\left(\frac{\eta^2}{n\left\|D_{j^\star}^{(k)}\right\|}-1\right)+\alpha^{\varepsilon/25}, \end{aligned}$$

where (1) is from Proposition 3, (2) is an application of Proposition 8 to (22), and (3) is valid for α sufficiently small. Using the same application of Proposition 8 to (22), we obtain

$$\begin{aligned}
&\frac{1}{n}\sum_{v\in S_{j^\star}^{(k)}}(h_{\widetilde{\theta}}(x_v)-y_v)\widetilde{\overline{w}}_{j^\star}^\top x_v \\
&\quad< \frac{1}{n}\sum_{v\in S_{j^\star}^{(k)}}\left(y_v\left(\frac{\eta^2}{n\left\|D_{j^\star}^{(k)}\right\|}-1\right)+\alpha^{\varepsilon/25}\right)\left(\frac{y_v}{n\left\|D_{j^\star}^{(k)}\right\|}-\sqrt{2}\alpha^{\varepsilon/24}\right) \\
&\quad< \sum_{v\in S_{j^\star}^{(k)}}\frac{y_v^2}{n^2\left\|D_{j^\star}^{(k)}\right\|}\left(\frac{\eta^2}{n\left\|D_{j^\star}^{(k)}\right\|}-1\right)+\alpha^{\varepsilon/26} \\
&\quad= \left\|D_{j^\star}^{(k)}\right\|\left(\frac{\eta^2}{n\left\|D_{j^\star}^{(k)}\right\|}-1\right)+\alpha^{\varepsilon/26},
\end{aligned} \tag{30}$$

where the second inequality is valid for α sufficiently small. For $v\in S_F^{(k)}$, we have the obvious inequality $\widetilde{\overline{w}}_{j^\star}^\top x_v \le \max_{v\in S_F^{(k)}}\widetilde{\overline{w}}_{j^\star}^\top x_v$. Using the Cauchy-Schwarz inequality gives

$$\frac{1}{n}\sum_{v\in S_F^{(k)}}A_{v,j^\star}(h_{\widetilde{\theta}}(x_v)-y_v)\widetilde{\overline{w}}_{j^\star}^\top x_v \le \sqrt{n}\left\|E^{(k)}\right\|\max_{v\in S_F^{(k)}}\widetilde{\overline{w}}_{j^\star}^\top x_v. \tag{31}$$

Inserting (29), (30), and (31) into (28), we obtain that for $[\tau_1,\tau_2]$ and $i\in S_F^{(k)}$,

$$\begin{aligned}
\frac{1}{\log(1/\alpha)}\frac{\mathrm{d}}{\mathrm{d}t}\widetilde{\overline{w}}_{j^\star}^\top x_i \le \left\|E^{(k)}\right\| &+ \left(\left\|D_{j^\star}^{(k)}\right\|\left(\frac{\eta^2}{n\left\|D_{j^\star}^{(k)}\right\|}-1\right)+\alpha^{\varepsilon/26}\right)\widetilde{\overline{w}}_{j^\star}^\top x_i \\
&+\sqrt{n}\left\|E^{(k)}\right\|\max_{v\in S_F^{(k)}}\widetilde{\overline{w}}_{j^\star}^\top x_v\widetilde{\overline{w}}_{j^\star}^\top x_i.
\end{aligned}$$

Since the above inequality holds for any $i\in S_F^{(k)}$, we have that, when differentiable (as is the case almost everywhere by Rademacher's theorem), the continuous function $f(t):=\max_{i\in S_F^{(k)}}\left(\widetilde{\overline{w}}_{j^\star}^\top x_i\right)_+$ obeys the differential inequality

$$\begin{aligned}
\frac{1}{\log(1/\alpha)}f'(t) &\le \left\|E^{(k)}\right\|\left(1+\sqrt{n}f(t)^2\right)-\left(\left\|D_{j^\star}^{(k)}\right\|\left(1-\frac{\eta^2}{n\left\|D_{j^\star}^{(k)}\right\|}\right)-\alpha^{\varepsilon/26}\right)f(t) \\
&< \left\|E^{(k)}\right\|\left(1+\sqrt{n}f(t)^2\right),
\end{aligned} \tag{32}$$

where the second inequality is valid for α sufficiently small. We use (22) and Proposition 8 to see that $f(t)^2<2\alpha^{\varepsilon/12}$ on $[\tau_1,\tau_2]$. Plugging this back in and integrating yields

$$f(t)<\sqrt{2}\alpha^{\varepsilon/4}+\left(1+2\sqrt{n}\alpha^{\varepsilon/12}\right)\log(1/\alpha)\int_{\tau_1}^t\left\|E^{(k)}\right\|ds. \tag{33}$$

We use this bound to derive a differential equation governing the growth of the residual. We must again estimate (11), which we recall states

$$-\frac{1}{2\log(1/\alpha)}\frac{\mathrm{d}}{\mathrm{d}t}\left\|E^{(k)}\right\|^2 \leq -\frac{1}{n}\sum_{i\in S_F^{(k)}}\sum_{j\in N_U^{(k)}} A_{i,j}\|\widetilde{w}_j\|^2\mathcal{D}_j^\top\widetilde{\widetilde{w}}_j\widetilde{\widetilde{w}}_j^\top x_i x_i^\top E^{(k)}.$$

We consider separately the cases $j \in N_U^{(k)} \setminus j^\star$ and $j = j^\star$. Beginning with the former, Proposition 3 implies $\|\widetilde{w}_j\|^2 < \alpha^{2c_U}$. We simply reuse (10) combined with the Cauchy-Schwarz inequality to obtain

$$-\frac{1}{n}\sum_{i\in S_F^{(k)}}\sum_{j\in N_U^{(k)}\setminus\{j^\star\}} A_{i,j}\|\widetilde{w}_j\|^2\mathcal{D}_j^\top\widetilde{\widetilde{w}}_j\widetilde{\widetilde{w}}_j^\top x_i x_i^\top E^{(k)} \leq \frac{m}{n}\sqrt{2\mathcal{L}(0)}\alpha^{2c_U}\left\|E^{(k)}\right\|. \tag{34}$$

Note that we have simply estimated $|\widetilde{\widetilde{w}}_j^\top x_i| \leq 1$, since we have imposed no assumption on the alignment of neurons $j \in N_U^{(k)}$ during this phase. For the contribution of the term corresponding to $j = j^\star$, we have

$$-\frac{1}{n}\sum_{i\in S_F^{(k)}} A_{i,j^\star}\|\widetilde{w}_{j^\star}\|^2\mathcal{D}_{j^\star}^\top\widetilde{\widetilde{w}}_{j^\star}\widetilde{\widetilde{w}}_{j^\star}^\top x_i x_i^\top E^{(k)} < \frac{1}{n}\sqrt{2\mathcal{L}(0)}\eta^2 f(t)\left\|E^{(k)}\right\|,$$

where we have applied the bound (24) on the norm of neuron $j^\star$ in the definition of τ_2 and (10) combined with the Cauchy-Schwarz inequality. Substituting in (33), we obtain

$$\begin{aligned}-\frac{1}{n}\sum_{i\in S_F^{(k)}} & A_{i,j^\star}\|\widetilde{w}_{j^\star}\|^2\mathcal{D}_{j^\star}^\top\widetilde{\widetilde{w}}_{j^\star}\widetilde{\widetilde{w}}_{j^\star}^\top x_i x_i^\top E^{(k)} \\ &\leq \frac{1}{n}\sqrt{2\mathcal{L}(0)}\eta^2\left(\sqrt{2}\alpha^{\varepsilon/4} + \left(1+2\sqrt{n}\alpha^{\varepsilon/12}\right)\log(1/\alpha)\int_{\tau_1}^t\left\|E^{(k)}\right\|ds\right)\left\|E^{(k)}\right\|\end{aligned} \tag{35}$$

Inserting (34) and (35) into (11), we have

$$\begin{aligned}\frac{\mathrm{d}}{\mathrm{d}t}\left\|E^{(k)}\right\| &\leq \frac{1}{n}\sqrt{2\mathcal{L}(0)}\eta^2\log(1/\alpha)\left(\sqrt{2}\alpha^{\varepsilon/4} + \left(1+2\sqrt{n}\alpha^{\varepsilon/12}\right)\log(1/\alpha)\int_{\tau_1}^t\left\|E^{(k)}\right\|ds\right) \\ &\qquad + \frac{m}{n}\sqrt{2\mathcal{L}(0)}\log(1/\alpha)\alpha^{2c_U} \\ &< \frac{2}{n}\sqrt{2\mathcal{L}(0)}\eta^2\log(1/\alpha)\alpha^{\varepsilon/4} + \frac{2}{n}\sqrt{2\mathcal{L}(0)}\eta^2\log(1/\alpha)^2\int_{\tau_1}^t\left\|E^{(k)}\right\|ds,\end{aligned}$$

where the second inequality is valid for α sufficiently small by increasing the coefficient of the $\alpha^{\varepsilon/4}$ term (recall here that $c_U > \varepsilon$) and by using the estimate $a = 1 + 2\sqrt{n}\alpha^{\varepsilon/12} < 2$ for α sufficiently small. Differentiability almost everywhere is again due to Rademacher's theorem. This is

a differential inequality of the form $F'' < pF + q$, where $F = \int_{\tau_1}^t \left\|E^{(k)}\right\| ds$. Applying Lemma 10,

$$\begin{aligned}
F'(t) &< \left(\frac{q}{\sqrt{p}} + F'(\tau_1)\right) \exp\left(\sqrt{p}(t-\tau_1)\right) \\
&= \left(\sqrt{\frac{2}{n}}\sqrt[4]{2\mathcal{L}(0)}\eta\alpha^{\varepsilon/4} + \alpha^{\varepsilon}\right) \exp_\alpha\left(-\sqrt{\frac{2}{n}}\sqrt[4]{2\mathcal{L}(0)}\eta(t-\tau_1)\right) \\
&\overset{(1)}{<} 2\sqrt{\frac{2}{n}}\sqrt[4]{2\mathcal{L}(0)}\eta\alpha^{\varepsilon/4} \exp_\alpha\left(-\sqrt{\frac{2}{n}}\sqrt[4]{2\mathcal{L}(0)}\eta(t-\tau_1)\right) \\
&\overset{(2)}{<} \frac{4}{\sqrt{n}}\sqrt[4]{2\mathcal{L}(0)}\eta\alpha^{\varepsilon/4} \exp_\alpha\left(-\sqrt{\frac{2}{n}}\sqrt[4]{2\mathcal{L}(0)}\eta\frac{3\varepsilon}{\left\|D_{j^\star}^{(k)}\right\|}\right) \\
&\overset{(3)}{<} \alpha^{\varepsilon/5}.
\end{aligned} \tag{36}$$

In these workings, (1) is valid for α sufficiently small, (2) follows from inserting the upper bound on $\tau_2 - \tau_1$ derived in Item 1 Lemma 5, and (3) follows from the fact $\eta < \frac{\sqrt{n}}{60\sqrt{2}} \frac{\left\|D_{j^\star}^{(k)}\right\|}{\sqrt[4]{2\mathcal{L}(0)}}$.

We are now able to plug (36) back into (33) to estimate

$$\begin{aligned}
\max_{i \in S_F^{(k)}} A_{i,j^\star} \widetilde{\overline{w}}_{j^\star}^\top x_i &< \sqrt{2}\alpha^{\varepsilon/4} + a\log(1/\alpha)\alpha^{\varepsilon/5}(t-\tau_1) \\
&\leq \sqrt{2}\alpha^{\varepsilon/4} + a\log(1/\alpha)\alpha^{\varepsilon/5}(\tau_2-\tau_1) \\
&< \alpha^{11\varepsilon/60},
\end{aligned} \tag{37}$$

where the final inequality is valid for α sufficiently small and gives the claim. Item 3 Lemma 5 then follows for α sufficiently small. □

Proof of Item 4. Next, we verify the stability of alignment of neuron $j^\star$. We have

$$\begin{aligned}
\frac{1}{\log(1/\alpha)}\frac{\mathrm{d}}{\mathrm{d}t}\widetilde{\overline{w}}_{j^\star}^\top D_{j^\star}^{(k)} &= \left\|D_{j^\star}^{(k)}\right\|^2 - \left(\widetilde{\overline{w}}_{j^\star}^\top D_{j^\star}^{(k)}\right)^2 \\
&\quad - \frac{1}{n}\sum_{i \in S_{j^\star}^{(k)}} h_{\widetilde{\theta}}(x_i)\left(x_i^\top D_{j^\star}^{(k)} - x_i^\top \widetilde{\overline{w}}_{j^\star}\widetilde{\overline{w}}_{j^\star}^\top D_{j^\star}^{(k)}\right) \\
&\quad - \frac{1}{n}\sum_{i \in S_F^{(k)}} (h_{\widetilde{\theta}}(x_i) - y_i)\left(x_i^\top D_{j^\star}^{(k)} - x_i^\top \widetilde{\overline{w}}_{j^\star}\widetilde{\overline{w}}_{j^\star}^\top D_{j^\star}^{(k)}\right) \\
&> \left\|D_{j^\star}^{(k)}\right\|^2 - \left(\widetilde{\overline{w}}_{j^\star}^\top D_{j^\star}^{(k)}\right)^2 \\
&\quad - \frac{1}{n}\sum_{i \in S_{j^\star}^{(k)}} h_{\widetilde{\theta}}(x_i)\left(x_i^\top D_{j^\star}^{(k)} - x_i^\top \widetilde{\overline{w}}_{j^\star}\widetilde{\overline{w}}_{j^\star}^\top D_{j^\star}^{(k)}\right) \\
&\quad - 2\left\|D_{j^\star}^{(k)}\right\|\alpha^{5\varepsilon/24},
\end{aligned} \tag{38}$$

This bound comes from (23) combined with an application of Corollary 1 to (22). We recall that for $i \in S_{j^\star}^{(k)}$, $h_{\widetilde{\theta}}(x_i) \leq \|\widetilde{w}_{j^\star}\|^2 \widetilde{\overline{w}}_{j^\star}^\top x_i + m\alpha^{2c_U}$. The growth to constant order of the norm of neuron $j^\star$ means that the previously used bootstrapping argument, where Corollary 1 was combined with a uniform growth bound, is no longer possible. Instead, from (37), we have that if $i \in S_F^{(k)}$, then on the interval $[\tau_1, \tau_2]$, $\widetilde{\overline{w}}_{j^\star}^\top x_i \leq \alpha^{11\varepsilon/60}$. We apply this to see

$$
\begin{aligned}
&\frac{1}{n}\sum_{i\in S_{j^\star}^{(k)}} h_{\widetilde{\theta}}(x_i)\left(x_i^\top D_{j^\star}^{(k)} - x_i^\top \widetilde{\overline{w}}_{j^\star}\widetilde{\overline{w}}_{j^\star}^\top D_{j^\star}^{(k)}\right)\\
&= \frac{1}{n}\sum_{i\in S_{j^\star}^{(k)}}\left(\|\widetilde{w}_{j^\star}\|^2\widetilde{\overline{w}}_{j^\star}^\top x_i + \sum_{j\neq j^\star} A_{i,j}\|\widetilde{w}_j\|^2\widetilde{\overline{w}}_j^\top x_i\right)\left(x_i^\top D_{j^\star}^{(k)} - x_i^\top \widetilde{\overline{w}}_{j^\star}\widetilde{\overline{w}}_{j^\star}^\top D_{j^\star}^{(k)}\right)\\
&\leq \frac{1}{n}\|\widetilde{w}_{j^\star}\|^2\widetilde{\overline{w}}_{j^\star}^\top D_{j^\star}^{(k)}\left(1-\sum_{i\in S_{j^\star}^{(k)}}\left(\widetilde{\overline{w}}_{j^\star}^\top x_i\right)^2\right) + 2m\left\|D_{j^\star}^{(k)}\right\|\alpha^{2c_U}\\
&= \frac{1}{n}\|\widetilde{w}_{j^\star}\|^2\widetilde{\overline{w}}_{j^\star}^\top D_{j^\star}^{(k)}\left(\sum_{i\notin S_{j^\star}^{(k)}}\left(\widetilde{\overline{w}}_{j^\star}^\top x_i\right)^2\right) + 2m\left\|D_{j^\star}^{(k)}\right\|\alpha^{2c_U}\\
&\leq \eta^2\left\|D_{j^\star}^{(k)}\right\|\alpha^{11\varepsilon/30} + 2m\left\|D_{j^\star}^{(k)}\right\|\alpha^{2c_U}. \qquad (39)
\end{aligned}
$$

Putting (39) into (38), we have for $t\in[\tau_1,\tau_2]$ and for α sufficiently small that

$$
\begin{aligned}
\frac{1}{\log(1/\alpha)}\frac{\mathrm{d}}{\mathrm{d}t}\widetilde{\overline{w}}_{j^\star}^\top D_{j^\star}^{(k)} &> \left\|D_{j^\star}^{(k)}\right\|^2 - \left(\widetilde{\overline{w}}_{j^\star}^\top D_{j^\star}^{(k)}\right)^2\\
&\quad - \eta^2\left\|D_{j^\star}^{(k)}\right\|\alpha^{11\varepsilon/30} - 2m\left\|D_{j^\star}^{(k)}\right\|\alpha^{2c_U} - 2\left\|D_{j^\star}^{(k)}\right\|\alpha^{5\varepsilon/24}\\
&> \left\|D_{j^\star}^{(k)}\right\|^2 - \left(\widetilde{\overline{w}}_{j^\star}^\top D_{j^\star}^{(k)}\right)^2 - \alpha^{9\varepsilon/48}.
\end{aligned}
$$

We can now use arguments similar to those employed in the proof of Item 2 Lemma 4. From the initial conditions given by Item 2 Lemma 4 and the inequality $\sqrt{1-x^2} > 1-x$, we obtain that on $[\tau_1,\tau_2]$ the alignment is lower bounded a function converging monotonically to $\left\|D_{j^\star}^{(k)}\right\|\min\left(1-\alpha^{\varepsilon/2}, 1-\alpha^{9\varepsilon/96}/\left\|D_{j^\star}^{(k)}\right\|\right)$. It then follows that for α sufficiently small,

$$
\widetilde{\overline{w}}_{j^\star}^\top D_{j^\star}^{(k)} > \left\|D_{j^\star}^{(k)}\right\|(1-\alpha^{\varepsilon/12}).
$$

□

Proof of Item 2. We verify that the norms of neurons in the set $N_F^{(k)}$ did not change significantly on the interval $[\tau_1,\tau_2]$. Item 3 Lemma 4 provides upper and lower bounds on the norms of neurons in the set $N_F^{(k)}$ at time τ_1. In light of these initial conditions and the bound on the loss on data in the set $S_F^{(k)}$ given in (23), identical workings to those yielding (15) give

$$
\sqrt{n\left\|D_j^{(k)}\right\|} - \alpha^{\varepsilon/8} < \|\widetilde{w}_j\| < \sqrt{n\left\|D_j^{(k)}\right\|} + \alpha^{\varepsilon/8}
$$

for all $t\in[\tau_1,\tau_2]$ for α sufficiently small.

□

Proof. Proof of Item 6 We have concluded that no condition except for that on the norm of neuron $j^\star$ could have broken at time τ_2, therefore $\|\widetilde{w}_{j^\star}(\tau_2)\| = \eta$, as claimed.

□

This completes the proof of Lemma 5. ■

Remark. We observe that in passing to (32), the inequality is far from sharp. We could have proceeded from the previous line by noting we have the differential inequality $f'(t) < a\left\|E^{(k)}\right\|\log(1/\alpha) - b\log(1/\alpha)f(t)$, where $a = \left(1+2\sqrt{n}\alpha^{\varepsilon/12}\right)$ and $b = \left\|D_{j^\star}^{(k)}\right\|\left(1-\frac{\eta^2}{n\left\|D_{j^\star}^{(k)}\right\|}\right) + \alpha^{\varepsilon/26}$. Multiplying both sides by $\exp(b\log(1/\alpha)t)$, we get $(\exp(b\log(1/\alpha)t)f(t))' < a\log(1/\alpha)\left\|E^{(k)}\right\|\exp(b\log(1/\alpha)t)$. Integration leads to

$$\begin{aligned} f(t) &< \alpha^{b(t-\tau_1)}f(\tau_1) + a\log(1/\alpha)\alpha^{bt}\int_{\tau_1}^{t}\alpha^{-bs}\left\|E^{(k)}\right\|ds \\ &< \sqrt{2}\alpha^{\varepsilon/4} + a\log(1/\alpha)\alpha^{bt}\int_{\tau_1}^{t}\alpha^{-bs}\left\|E^{(k)}\right\|ds. \end{aligned}$$

It seems possible that carefully estimating this integral expression may allow the growth phases to be combined into a single phase.

D.4 Phase 2b: Fitting of data

Define for $\varepsilon' > 0$ the time

$$\tau_3 := \inf\Big\{t \geq \tau_2 \;\Big|\; \widetilde{\overline{w}}_{j^\star}^\top \overline{D}_{j^\star}^{(k)} \leq 1-\alpha^{\varepsilon'} \tag{40}$$

$$\text{or } \exists\, i \in S_F^{(k)} \text{ s.t. } |h_{\widetilde{\theta}}(x_i) - y_i| \geq \alpha^{\varepsilon'} \tag{41}$$

$$\text{or } \|\widetilde{w}_{j^\star}\|^2 \geq n\left\|D_{j^\star}^{(k)}\right\| - \alpha^{\varepsilon'/4}\Big\} \bigwedge \left(\tau_1 + 2\sqrt{\frac{n}{2\mathcal{L}(0)}}c_U\right). \tag{42}$$

The lemma below provides an account of the phase where neuron $j^\star$ fits the data on which is it active. We prove that the alignment of neuron $j^\star$ remains stable, the loss on fitted data does not increase significantly, and by the end of the phase neuron $j^\star$ has grown very close to its final norm as predicted by Algorithm 1. The proof for this phase is almost identical to that for Phase 2a.

Lemma 6. *Suppose that Hypothesis 1 holds. Then for all $\kappa \in (t_k, t_{k+1})$, there exists $\varepsilon^* > 0$ such that for all $\varepsilon \in (0, \varepsilon^*)$, there exists $\varepsilon'^* > 0$ such that for all $\varepsilon' \in (0, \varepsilon'^*)$ there exists α^* such that for $\alpha \in (0, \alpha^*)$ the following are true:*

1. $\tau_3 - \tau_2 < \frac{\varepsilon'}{\left\|D_{j^\star}^{(k)}\right\|}$.
2. $\forall j \in N_F^{(k)}, \forall t \in [\tau_2, \tau_3],\ \sqrt{n\left\|D_j^{(k)}\right\|} - \alpha^{\varepsilon'/2} < \|\widetilde{w}_j\| < \sqrt{n\left\|D_j^{(k)}\right\|} + \alpha^{\varepsilon'/2}$.
3. $\forall i \in S_F^{(k)}, \forall t \in [\tau_2, \tau_3],\ |h_{\widetilde{\theta}}(x_i) - y_i| < \alpha^{\varepsilon'}$.
4. $\forall i \in S_F^{(k)}, \forall t \in [\tau_2, \tau_3],\ \max_{i \in S_F^{(k)}}\left(\widetilde{\overline{w}}_{j^\star}^\top x_i\right)_+ < \alpha^{5\varepsilon'/4}$.
5. $\forall t \in [\tau_2, \tau_3],\ \widetilde{\overline{w}}_{j^\star}^\top \overline{D}_{j^\star}^{(k)} > 1-\alpha^{\varepsilon'}$.
6. $\|\widetilde{w}_{j^\star}(\tau_3)\|^2 = n\left\|D_{j^\star}^{(k)}\right\| - \alpha^{\varepsilon'/4}$.

Proof. We work out a useful preliminary bound to begin this section. We upper bound the network output on $i \in S_{j^\star}^{(k)}$:

$$
\begin{aligned}
h_{\widetilde{\theta}}(x_i) &\overset{(1)}{<} \widetilde{\overline{w}}_{j^\star}^\top x_i \|\widetilde{w}_{j^\star}\|^2 + m\alpha^{2c_U} \\
&\overset{(2)}{<} \left(\frac{y_i}{n \left\| D_{j^\star}^{(k)} \right\|} + \sqrt{2}\alpha^{\varepsilon'/2} \right) \left(n \left\| D_{j^\star}^{(k)} \right\| - \alpha^{\varepsilon'/4} \right) + m\alpha^{2c_U} \\
&\overset{(3)}{<} y_i - \alpha^{\varepsilon'/3}.
\end{aligned} \tag{43}
$$

Here, (1) is from Proposition 3, (2) is from combining (42) with an application of Proposition 8 to (40), and (3) is valid for α sufficiently small.

Proof of Item 1. We estimate the growth of neuron $j^\star$. Similar workings to those which allowed us to obtain (27) in the proof of Item 1 Lemma 5 allow us to conclude that for α sufficiently small, on $[\tau_2, \tau_3]$ we have

$$
\begin{aligned}
\frac{1}{2\log(1/\alpha)} \frac{\mathrm{d}}{\mathrm{d}t} \|\widetilde{w}_{j^\star}\|^2 &> \|\widetilde{w}_{j^\star}\|^2 \left(\left\| D_{j^\star}^{(k)} \right\| (1 - \alpha^{\varepsilon'}) - \alpha^{\varepsilon'} - \frac{\|\widetilde{w}_{j^\star}\|^2}{n} - m\alpha^{2c_U} \right) \\
&> \|\widetilde{w}_{j^\star}\|^2 \left(\left\| D_{j^\star}^{(k)} \right\| - \alpha^{\varepsilon'/2} \right) - \frac{\|\widetilde{w}_{j^\star}\|^4}{n}.
\end{aligned}
$$

As before, solutions of differential inequality $f'(t) = 2\log(1/\alpha)af(t) - 2\log(1/\alpha)bf(t)^2$ for $t \in [\tau_2, \tau_3]$ have the lower bound

$$
f(t) > \frac{a}{b} \frac{1}{\alpha^{2a(t-C)} + 1}.
$$

From Lemma 5, we have $\|\widetilde{w}_{j^\star}(\tau_2)\|^2 = \eta^2$, and determining the integration constant yields

$$
f(t) > \frac{a}{b} \frac{1}{1 + \exp_\alpha \left(2a(t - \tau_2) + \frac{\log\left(\frac{a}{b}\eta^{-2} - 1\right)}{\log(\alpha)} \right)}.
$$

Therefore,

$$
\begin{aligned}
f\left(\tau_2 + \frac{\varepsilon'}{\left\| D_{j^\star}^{(k)} \right\|} \right) &\overset{(1)}{>} \frac{a}{b} \frac{1}{1 + \exp_\alpha \left(\frac{2\varepsilon'}{\left\| D_{j^\star}^{(k)} \right\|} \left(\left\| D_{j^\star}^{(k)} \right\| - \alpha^{\varepsilon'/2} \right) + \frac{\log\left(\frac{a}{b}\eta^{-2} - 1\right)}{\log(\alpha)} \right)} \\
&\overset{(1)}{>} \frac{a}{b} \frac{1}{1 + \alpha^{\varepsilon'}} \\
&> \frac{a}{b}(1 - 2\alpha^{\varepsilon'}) \\
&\overset{(2)}{>} n \left\| D_{j^\star}^{(k)} \right\| - \alpha^{\varepsilon'/4},
\end{aligned} \tag{44}
$$

which contradicts the definition of τ_3. Inequality (1) comes from noting $\frac{a}{b}\eta^{-2} - 1$ is bounded away from 0 as α decreases owing to the fact η is fixed and inequality (2) comes from estimating $\frac{a}{b}$. Note that each of the inequalities above is only valid for α sufficiently small, depending on ε and the data. Necessarily, we have $\tau_3 < \tau_2 + \frac{\varepsilon'}{\left\| D_{j^\star}^{(k)} \right\|}$. □

Proof of Items 3 and 4. In this section it will again be important to derive a bound on the realignment of neuron $j^\star$ with fitted data in terms of the residual on fitted data. We recall that for $i \in S_F^{(k)}$,

$$\begin{aligned}\frac{1}{\log(1/\alpha)}\frac{\mathrm{d}}{\mathrm{d}t}\widetilde{\overline{w}}_{j^\star}^\top x_i &= -\frac{1}{n}(h_{\widetilde{\theta}}(x_i)-y_i)\mathbb{1}\left(\widetilde{w}_{j^\star}^\top x_i>0\right)\\ &\quad+\frac{1}{n}\sum_{v\in[n]}A_{v,j^\star}(h_{\widetilde{\theta}}(x_v)-y_v)\widetilde{\overline{w}}_{j^\star}^\top x_v\widetilde{\overline{w}}_{j^\star}^\top x_i\\ &\le \left\|E^{(k)}\right\|\left(1+\sqrt{n}\max_{v\in S_F^{(k)}}\widetilde{\overline{w}}_{j^\star}^\top x_v\widetilde{\overline{w}}_{j^\star}^\top x_i)\right),\end{aligned}$$

where we have discarded the sum over $v\in S_{j^\star}^{(k)}$ since (43) gives that this adds a negative contribution. Since this holds for all $i\in S_F^{(k)}$ and for all $t\in[\tau_2,\tau_3]$, we have that $f(t):=\max_{i\in S_F^{(k)}}\left(\widetilde{\overline{w}}_{j^\star}^\top x_i\right)_+$ obeys the inequality

$$\frac{1}{\log(1/\alpha)}f'(t)\le\left\|E^{(k)}\right\|\left(1+2\sqrt{n}\alpha^{\varepsilon'}\right),$$

where again we have estimated $f(t)^2$ by plugging in the assumed bound on alignment. Differentiability almost everywhere follows from Rademacher's theorem. Integration gives

$$f(t)<\sqrt{2}\alpha^{\varepsilon/24}+\left(1+2\sqrt{n}\alpha^{\varepsilon'}\right)\log(1/\alpha)\int_{\tau_2}^t\left\|E^{(k)}\right\|ds. \tag{45}$$

We proceed to estimate (11), which we recall for convenience is

$$\frac{1}{2\log(1/\alpha)}\frac{\mathrm{d}}{\mathrm{d}t}\left\|E^{(k)}\right\|^2\le-\frac{1}{n}\sum_{i\in S_F^{(k)}}\sum_{j\in N_U^{(k)}}A_{i,j}\|\widetilde{w}_j\|^2\mathcal{D}_j^\top\widetilde{\overline{w}}_j\widetilde{\overline{w}}_j^\top x_ix_i^\top E^{(k)}.$$

We can again discard the sum over $j\in N_F^{(k)}$, leaving only the cases $j=j^\star$ and $j\in N_U^{(k)}\setminus\{j^\star\}$. For the latter, we apply Proposition 3 in a similar manner to the previous phase to get

$$-\frac{1}{n}\sum_{i\in S_F^{(k)}}\sum_{j\in N_U^{(k)}\setminus\{j^\star\}}A_{i,j}\|\widetilde{w}_j\|^2\mathcal{D}_j^\top\widetilde{\overline{w}}_j\widetilde{\overline{w}}_j^\top x_ix_i^\top E^{(k)}\le\frac{m}{n}\sqrt{2\mathcal{L}(0)}\alpha^{2c_U}\left\|E^{(k)}\right\|. \tag{46}$$

For the case $j=j^\star$, we obtain the estimate

$$\begin{aligned}&-\frac{1}{n}\sum_{i\in S_F^{(k)}}A_{i,j^\star}\|\widetilde{w}_{j^\star}\|^2\mathcal{D}_{j^\star}^\top\widetilde{\overline{w}}_{j^\star}\widetilde{\overline{w}}_{j^\star}^\top x_ix_i^\top E^{(k)}\\ &\qquad\le\left(n\left\|D_{j^\star}^{(k)}\right\|-\alpha^{\varepsilon'/4}\right)\sqrt{\frac{2\mathcal{L}(0)}{n}}\frac{1}{n}\sum_{i\in S_F^{(k)}}A_{i,j^\star}\widetilde{\overline{w}}_{j^\star}^\top x_ix_i^\top E^{(k)}\\ &\qquad\le\left\|D_{j^\star}^{(k)}\right\|\sqrt{2\mathcal{L}(0)}f(t)\left\|E^{(k)}\right\|.\end{aligned} \tag{47}$$

where in the final line we apply the Cauchy-Schwarz inequality. Inserting (47) and (46) into (11), we get

$$\begin{aligned}\frac{\mathrm{d}}{\mathrm{d}t}\left\|E^{(k)}\right\|&\le\left\|D_{j^\star}^{(k)}\right\|\sqrt{2\mathcal{L}(0)}\log(1/\alpha)\left(\sqrt{2}\alpha^{\varepsilon/24}+\left(1+2\sqrt{n}\alpha^{\varepsilon'}\right)\log(1/\alpha)\int_{\tau_2}^t\left\|E^{(k)}\right\|ds\right)\\ &\qquad+\frac{m}{n}\sqrt{2\mathcal{L}(0)}\log(1/\alpha)\alpha^{2c_U}\\ &<2\left\|D_{j^\star}^{(k)}\right\|\sqrt{2\mathcal{L}(0)}\log(1/\alpha)\alpha^{\varepsilon/24}+2\left\|D_{j^\star}^{(k)}\right\|\sqrt{2\mathcal{L}(0)}\log(1/\alpha)^2\int_{\tau_2}^t\left\|E^{(k)}\right\|ds,\end{aligned}$$

for α sufficiently small. Again, we have a differential inequality of the form $F'' < pF + q$, for which we employ Lemma 10:

$$\begin{aligned}
F'(t) &< \left(\frac{q}{\sqrt{p}} + F'(\tau_2)\right) \exp\left(\sqrt{p}(t-\tau_2)\right) \\
&< \left(\left(2\left\|D_{j^\star}^{(k)}\right\|\sqrt{2\mathcal{L}(0)}\right)^{1/2}\alpha^{\varepsilon/24} + \alpha^{\varepsilon/6}\right)\exp_\alpha\left(-\left(2\left\|D_{j^\star}^{(k)}\right\|\sqrt{2\mathcal{L}(0)}\right)^{1/2}(t-\tau_2)\right) \\
&< \alpha^{\varepsilon/25}\exp_\alpha\left(-\left(2\left\|D_{j^\star}^{(k)}\right\|\sqrt{2\mathcal{L}(0)}\right)^{1/2}\frac{\varepsilon'}{\left\|D_{j^\star}^{(k)}\right\|}\right) \\
&\overset{(2)}{<} \alpha^{3\varepsilon'/2},
\end{aligned} \tag{48}$$

provided ε' is selected sufficiently small depending on ε. We may now plug (48) back into (45) to get that for $t \in [\tau_2, \tau_3]$,

$$\begin{aligned}
\max_{i\in S_F^{(k)}} A_{i,j^\star}\widetilde{\overline{w}}_{j^\star}^\top x_i &< \sqrt{2}\alpha^{\varepsilon/24} + \left(1+2\sqrt{n}\alpha^{\varepsilon'}\right)\log(1/\alpha)\int_{\tau_2}^t \left\|E^{(k)}\right\| ds \\
&< \sqrt{2}\alpha^{\varepsilon/24} + \left(1+2\sqrt{n}\alpha^{\varepsilon'}\right)\log(1/\alpha)\alpha^{3\varepsilon'/2}\frac{\varepsilon'}{\left\|D_{j^\star}^{(k)}\right\|} \\
&< \alpha^{5\varepsilon'/4},
\end{aligned} \tag{49}$$

where the final inequality is valid for ε' sufficiently small depending on ε and for α sufficiently small. □

Proof of Item 5. To estimate the alignment of neuron $j^\star$, we have from (49) that if $i \in S_F^{(k)}$ then $\widetilde{\overline{w}}_{j^\star}^\top x_i < \alpha^{5\varepsilon'/4}$. Combined with (48), we have

$$\begin{aligned}
\frac{1}{\log(1/\alpha)}\frac{\mathrm{d}}{\mathrm{d}t}\widetilde{\overline{w}}_{j^\star}^\top D_{j^\star}^{(k)} &= \left\|D_{j^\star}^{(k)}\right\|^2 - \left(\widetilde{\overline{w}}_{j^\star}^\top D_{j^\star}^{(k)}\right)^2 \\
&\quad - \frac{1}{n}\sum_{i\in S_{j^\star}^{(k)}} h_{\widetilde{\theta}}(x_i)\left(x_i^\top D_{j^\star}^{(k)} - x_i^\top \widetilde{\overline{w}}_{j^\star}\widetilde{\overline{w}}_{j^\star}^\top D_{j^\star}^{(k)}\right) \\
&\quad - \frac{1}{n}\sum_{i\in S_F^{(k)}} (h_{\widetilde{\theta}}(x_i) - y_i)\left(x_i^\top D_{j^\star}^{(k)} - x_i^\top \widetilde{\overline{w}}_{j^\star}\widetilde{\overline{w}}_{j^\star}^\top D_{j^\star}^{(k)}\right) \\
&> \left\|D_{j^\star}^{(k)}\right\|^2 - \left(\widetilde{\overline{w}}_{j^\star}^\top D_{j^\star}^{(k)}\right)^2 \\
&\quad - \frac{1}{n}\sum_{i\in S_{j^\star}^{(k)}} h_{\widetilde{\theta}}(x_i)\left(x_i^\top D_{j^\star}^{(k)} - x_i^\top \widetilde{\overline{w}}_{j^\star}\widetilde{\overline{w}}_{j^\star}^\top D_{j^\star}^{(k)}\right) \\
&\quad - 2\left\|D_{j^\star}^{(k)}\right\|\alpha^{11\varepsilon'/4}.
\end{aligned} \tag{50}$$

As in the previous section, we apply the bound on the norms of all neurons $j \in N_U^{(k)} \setminus \{j^\star\}$ to get

$$
\begin{aligned}
&\frac{1}{n}\sum_{i\in S_{j^\star}^{(k)}} h_{\widetilde{\theta}}(x_i)\left(x_i^\top D_{j^\star}^{(k)} - x_i^\top \widetilde{\overline{w}}_{j^\star}\widetilde{\overline{w}}_{j^\star}^\top D_{j^\star}^{(k)}\right)\\
&= \frac{1}{n}\sum_{i\in S_{j^\star}^{(k)}}\left(\|\widetilde{w}_{j^\star}\|^2\widetilde{\overline{w}}_{j^\star}^\top x_i + \sum_{j\neq j^\star} A_{i,j}\|\widetilde{w}_j\|^2\widetilde{\overline{w}}_j^\top x_i\right)\left(x_i^\top D_{j^\star}^{(k)} - x_i^\top \widetilde{\overline{w}}_{j^\star}\widetilde{\overline{w}}_{j^\star}^\top D_{j^\star}^{(k)}\right)\\
&< \frac{1}{n}\|\widetilde{w}_{j^\star}\|^2\widetilde{\overline{w}}_{j^\star}^\top D_{j^\star}^{(k)}\left(1 - \sum_{i\in S_{j^\star}^{(k)}}\left(\widetilde{\overline{w}}_{j^\star}^\top x_i\right)^2\right) + 2m\left\|D_{j^\star}^{(k)}\right\|\alpha^{2c_U}\\
&= \frac{1}{n}\|\widetilde{w}_{j^\star}\|^2\widetilde{\overline{w}}_{j^\star}^\top D_{j^\star}^{(k)}\left(\sum_{i\notin S_{j^\star}^{(k)}}\left(\widetilde{\overline{w}}_{j^\star}^\top x_i\right)^2\right) + 2m\left\|D_{j^\star}^{(k)}\right\|\alpha^{2c_U}\\
&< n\left\|D_{j^\star}^{(k)}\right\|^2\alpha^{5\varepsilon'/2} + 2m\left\|D_{j^\star}^{(k)}\right\|\alpha^{2c_U},
\end{aligned}
\tag{51}
$$

where in the final inequality we used (49) and (42). Putting (51) into (50), we have for $t \in [\tau_2, \tau_3]$ and for α sufficiently small that

$$
\frac{1}{\log(1/\alpha)}\frac{\mathrm{d}}{\mathrm{d}t}\widetilde{\overline{w}}_{j^\star}^\top D_{j^\star}^{(k)} > \left\|D_{j^\star}^{(k)}\right\|^2 - \left(\widetilde{\overline{w}}_{j^\star}^\top D_{j^\star}^{(k)}\right)^2 - \alpha^{2\varepsilon'}.
$$

From the initial conditions provided by Item 4 combined with the inequality $\sqrt{1-x^2} > 1-x$, we obtain for $t \in [\tau_2, \tau_3]$

$$
\widetilde{\overline{w}}_{j^\star}^\top D_{j^\star}^{(k)} > \left\|D_{j^\star}^{(k)}\right\|\left(1-\alpha^{\varepsilon'}\right),
$$

which implies the required alignment for α sufficiently small.

□

Proof. Proof of Item 2 We verify that the norms of neurons in the set $N_F^{(k)}$ did not change significantly on the interval $[\tau_1, \tau_2]$. Item 2 Lemma 5 provides upper and lower bounds on the norms of neurons in the set $N_F^{(k)}$ at time τ_2. In light of these initial conditions and the bound on the loss on data in the set $S_F^{(k)}$ given in (41), identical workings to those yielding (15) give

$$
\sqrt{n\left\|D_j^{(k)}\right\|} - \alpha^{\varepsilon'/2} < \|\widetilde{w}_j\| < \sqrt{n\left\|D_j^{(k)}\right\|} + \alpha^{\varepsilon'/2}
$$

for all $t \in [\tau_2, \tau_3]$ for α sufficiently small.

□

Proof of Item 6. For this point, we simply note that (44) in the proof of Item 1 Lemma 6 gives that if no other condition in the definition of τ_3 breaks then it must be that $\|\widetilde{w}_{j^\star}(\tau_3)\|^2 = n\left\|D_{j^\star}^{(k)}\right\| - \alpha^{\varepsilon'/4}$. □

This completes the proof of Lemma 6. ■

D.5 Phase 3: Realignment and slow growth

Define the time

$$\tau_4 := \inf\left\{t \geq \tau_3 \,\middle|\, \exists\, i \in S_F^{(k+1)} \text{ s.t. } |h_{\widetilde{\theta}}(x_i) - y_i| \geq \alpha^{\varepsilon'/6}\right\} \bigwedge \left(\tau_3 + \frac{3\varepsilon}{2n\left\|D_{j^\star}^{(k)}\right\|}\right) \bigwedge \left(\tau_1 + 2\sqrt{\frac{n}{2\mathcal{L}(0)}}c_U\right). \tag{52}$$

This phase is that during which neurons in the set $N_U^{(k)} \setminus \{j^\star\} = N_U^{(k+1)}$ realign in directions $D_j^{(k+1)}$ to respond to the fact that neuron $j^\star$ has grown to reduce the loss on the data on which it is active. We will find that $\tau_4 = \tau_3 + \frac{3\varepsilon}{2\left\|D_{j^\star}^{(k)}\right\|}$, and combining this with the bounds on the duration of phases 2a and 2b, we will see that the norms of the neurons in the set $N_U^{(k+1)}$ did not change significantly on the interval $[\tau_1, \tau_4]$. This choice is motivated by the lower bound on τ_1 afforded to us by (20), as we can then guarantee that $\tau_4 > t_{k+1}$. The reader may note that in the lemma below, we do not track the alignment of neuron $j^\star$, the reason being that it is now a member of the set $S_F^{(k+1)}$. As noted before Lemma 5, we can avoid laboriously tracking the alignment of all of these neurons throughout each phase, because this is not necessary to ensure the loss on fitted data remains small. Once this has been done, we will be able to estimate the alignment across several phases at once, using the alignment at time τ_3 as the initial condition for neuron $j^\star$ and the alignment at time τ_1 as the initial condition for neurons in the set $N_F^{(k)}$.

Lemma 7. *Suppose that Hypothesis 1 holds. Then for all $\kappa \in (t_k, t_{k+1})$, there exists $\varepsilon^* > 0$ such that for all $\varepsilon \in (0, \varepsilon^*)$ there exists ε'^* such that for all $\varepsilon' \in (0, \varepsilon'^*)$ then there exists $\alpha^* > 0$ such that for all $\alpha \in (0, \alpha^*)$ the following are true:*

1. $t_{k+1} < \tau_4 < t_{k+1} + \frac{4\varepsilon}{\left\|D_{j^\star}^{(k)}\right\|} + \frac{\varepsilon'}{\left\|D_{j^\star}^{(k)}\right\|}$.
2. $\forall i \in S_F^{(k+1)},\ \forall t \in [\tau_3, \tau_4],\ |h_{\widetilde{\theta}}(x_i) - y_i| < \alpha^{\varepsilon'/6}$.
3. $\forall j \in N_F^{(k+1)},\ \forall t \in [\tau_3, \tau_4],\ \sqrt{n\left\|D_j^{(k)}\right\|} - \alpha^{\varepsilon'/8} < \|\widetilde{w}_j\| < \sqrt{n\left\|D_j^{(k)}\right\|} + \alpha^{\varepsilon'/8}$
4. $\forall j \in N_U^{(k+1)},\ \widetilde{\overline{w}}_j^\top(\tau_4)\overline{D}_j^{(k+1)} > 1 - \alpha^{\varepsilon'/15}$.
5. $\forall j \in N_U^{(k+1)},\ \left|\ell_j^\circ(\tau_4) + \log_\alpha(\|\widetilde{w}_j(\tau_4)\|)\right| < \zeta_j(\varepsilon, \varepsilon')$, *where we define the quantity* $\zeta_j(\varepsilon, \varepsilon') := \frac{\varepsilon}{2} + \left(\left\|D_j^{(k)}\right\| + \sqrt{\frac{2\mathcal{L}(0)}{n}}\right)\left(\frac{9\varepsilon}{2\left\|D_{j^\star}^{(k)}\right\|} + \frac{\varepsilon'}{\left\|D_{j^\star}^{(k)}\right\|}\right)$.

Proof. The reader should note it will be convenient to derive the bounds in Item 1 Lemma 7 as consequences of analyses for the other points, as it is not possible to derive the upper and lower bounds at the same time.

Proof of Item 2. For an upper bound on τ_4, combining the upper bound on τ_1 provided by Item 4 Lemma 4, the upper bound on τ_2 provided by Item 1 Lemma 5, and the upper bound on τ_3 provided by Item 1 Lemma 6, we conclude that

$$\begin{aligned}\tau_4 &< \tau_1 + \frac{3\varepsilon}{\left\|D_{j^\star}^{(k)}\right\|} + \frac{\varepsilon'}{\left\|D_{j^\star}^{(k)}\right\|} + \frac{3\varepsilon}{2\left\|D_{j^\star}^{(k)}\right\|} \\ &< t_{k+1} + \frac{4\varepsilon}{\left\|D_{j^\star}^{(k)}\right\|} + \frac{\varepsilon'}{\left\|D_{j^\star}^{(k)}\right\|}.\end{aligned} \tag{53}$$

As ε and ε' are arbitrary, we may select these quantities sufficiently small that $\tau_3 + \frac{3\varepsilon}{2\left\|D_{j^\star}^{(k)}\right\|} < \tau_1 + 2\sqrt{\frac{n}{2\mathcal{L}(0)}}c_U$. In this case, we may apply Proposition 3 to bound the norms of the neurons in the set $N_U^{(k)} \setminus \{j^\star\}$. Therefore, we have on $[\tau_3, \tau_4]$ the bound $\|\widetilde{w}_j\| < \alpha^{c_U}$ for all $j \in N_U^{(k)} \setminus \{j^\star\}$. To determine the loss on data fitted by neuron $j^\star$, this observation combined with Item 5 and Item 6 Lemma 6 implies that for $i \in S_{j^\star}^{(k)}$, $|h_{\widetilde{\theta}}(x_i) - y_i| \lesssim \alpha^{\varepsilon'/4}$. Furthermore, by definition of τ_3, $\left\|E^{(k)}(\tau_3)\right\| < \alpha^{\varepsilon'}$. From these facts we conclude that for α sufficiently small,

$$\left\|E^{(k+1)}(\tau_3)\right\| < \alpha^{\varepsilon'/5}.$$

We now use arguments similar to those employed in the proof of Item 1 Lemma 4 to control the residual on fitted data. The same reasoning used to obtain (11) leads us to conclude

$$\frac{1}{2\log(1/\alpha)}\frac{\mathrm{d}}{\mathrm{d}t}\left\|E^{(k+1)}\right\|^2 \leq -\frac{1}{n}\sum_{i\in S_F^{(k+1)}}\sum_{j\in N_U^{(k+1)}} A_{i,j}\|\widetilde{w}_j\|^2 \mathcal{D}_j^\top \widetilde{\overline{w}}_j \widetilde{\overline{w}}_j^\top x_i x_i^\top E^{(k+1)}.$$

We use Proposition 3 to obtain the estimate

$$\begin{aligned}\frac{\mathrm{d}}{\mathrm{d}t}\left\|E^{(k+1)}\right\| &\leq \frac{m}{n}\sqrt{2\mathcal{L}(0)}\log(1/\alpha)\alpha^{2c_U} \\ &< \alpha^{\varepsilon'},\end{aligned}$$

where the final inequality is valid for α sufficiently small. Integration yields

$$\left\|E^{(k+1)}\right\| < \alpha^{\varepsilon'/5} + (\tau_4 - \tau_3)\alpha^{\varepsilon'}.$$

Hence,

$$\begin{aligned}\max_{i\in S_F^{(k+1)}} |h_{\widetilde{\theta}^t}(x_i) - y_i| &< n\left(\alpha^{\varepsilon'/5} + \varepsilon'\alpha^{\varepsilon'}\right) \\ &< \alpha^{\varepsilon'/6},\end{aligned}$$

as claimed. □

Proof. Proof of Item 3 We verify that the norms of neurons in the set $N_F^{(k)}$ did not change significantly on the interval $[\tau_3, \tau_4]$, which will lead us to conclude in combination with Proposition 3 that the condition on neurons' norms did not fail on this interval. Item 2 Lemma 6 and Item 6 provide upper and lower bounds on the norms of neurons in the set $N_F^{(k+1)}$ at time τ_3. In light of these initial conditions and the bound on the loss on data in the set $S_F^{(k+1)}$ given in (52), identical workings to those yielding (15) give

$$\sqrt{n\left\|D_j^{(k)}\right\|} - \alpha^{\varepsilon'/8} < \|\widetilde{w}_j\| < \sqrt{n\left\|D_j^{(k)}\right\|} + \alpha^{\varepsilon'/8}$$

for all $t \in [\tau_3, \tau_4]$ for α sufficiently small. Since none of the other conditions in the definition of τ_4 break, as a result of our analysis so far, we conclude that $\tau_4 = \tau_3 + \frac{3\varepsilon}{2\left\|D_{j^\star}^{(k)}\right\|}$. This leads to the easy lower bound

$$\begin{aligned}\tau_4 &> \tau_1 + \frac{3\varepsilon}{2\left\|D_{j^\star}^{(k)}\right\|}\\ &> t_{k+1},\end{aligned}$$

where the first inequality is by definition of τ_2, τ_3, and τ_4 and the second is from the lower bound on τ_1 provided by Item 4 Lemma 4

□

Proof of Item 4. We verify that by time $\tau_3 + \frac{3\varepsilon}{2\left\|D_{j^\star}^{(k)}\right\|}$, the neurons in the set $N_U^{(k+1)}$ have realigned. We first check that in the interval $[\tau_1, \tau_3]$, no such neurons deactivated on the data in the set $S_j^{(k+1)}$. We recall that $\frac{\mathrm{d}}{\mathrm{d}t}\widetilde{w}_j^\top x_i = -\frac{1}{n}\mathbb{1}\left(w_j^\top x_i > 0\right)\log(1/\alpha)(h_{\theta^t}(x_i) - y_i)\left\|\widetilde{w}_j\right\|$. By (21), on $[\tau_1, \tau_3]$ we had for every $i \in S_U^{(k+1)}$ that $h_{\theta^t}(x_i) - y_i \leq 0$ for α sufficiently small. From the alignment at time τ_1 given by Item 2 Lemma 4 we conclude that no deactivation occurred. Similarly to the proof of Lemma 3, we estimate

$$\begin{aligned}\frac{1}{\log(1/\alpha)}\frac{\mathrm{d}}{\mathrm{d}t}\widetilde{\overline{w}}_j^\top D_j^{(k+1)} &= \left\|D_j^{(k+1)}\right\|^2 - \left(\widetilde{\overline{w}}_j^\top D_j^{(k+1)}\right)^2\\ &\quad - \frac{1}{n}\sum_{i\in S_j^{(k+1)}} h_{\widetilde{\theta}}(x_i)\left(x_i^\top D_j^{(k+1)} - x_i^\top \widetilde{\overline{w}}_j\widetilde{\overline{w}}_j^\top D_j^{(k+1)}\right)\\ &\quad - \frac{1}{n}\sum_{i\in S_F^{(k+1)}} (h_{\widetilde{\theta}}(x_i) - y_i)\left(x_i^\top D_j^{(k+1)} - x_i^\top \widetilde{\overline{w}}_j\widetilde{\overline{w}}_j^\top D_j^{(k+1)}\right)\\ &> \left\|D_j^{(k+1)}\right\|^2 - \left(\widetilde{\overline{w}}_j^\top D_j^{(k+1)}\right)^2\\ &\quad - 2m\left\|D_j^{(k+1)}\right\|\alpha^{2c_U} - 2\left\|D_j^{(k+1)}\right\|\alpha^{\varepsilon'/6}\\ &> \left\|D_j^{(k+1)}\right\|^2 - \left(\widetilde{\overline{w}}_j^\top D_j^{(k+1)}\right)^2 - \alpha^{\varepsilon'/7},\end{aligned}$$

where the final inequality is valid for α sufficiently small. Similar workings to those which led to (5) give

$$\begin{aligned}&\widetilde{\overline{w}}_j^\top\left(\tau_3 + \frac{3\varepsilon}{2\left\|D_{j^\star}^{(k)}\right\|}\right)D_j^{(k+1)}\\ &\quad > \left(\left\|D_j^{(k+1)}\right\| - \alpha^{\varepsilon'/14}\right)\left(1 - \exp_\alpha\left(\left(\left\|D_j^{(k+1)}\right\| - \alpha^{\varepsilon'/14}\right)\frac{3\varepsilon}{2\left\|D_{j^\star}^{(k)}\right\|}\right)\right)\\ &\quad > \left\|D_j^{(k+1)}\right\| - \alpha^{\varepsilon'/15},\end{aligned}$$

where the final inequality is valid for ε' selected sufficiently small dependent on ε (and α sufficiently small in turn). □

Proof of Item 5. At this point, we are in a position to estimate the deviation of the norms of the neurons $j \in N_U^{(k+1)}$ from that predicted by Algorithm 1. A simple Grönwall argument gives that

$$\|\widetilde{w}_j(\tau_4)\| \leq \|\widetilde{w}_j(\tau_1)\|\alpha^{-(\tau_4-\tau_1)\sqrt{\frac{2\mathcal{L}(0)}{n}}}$$
$$< \exp_\alpha\left(-\ell_j^\circ(\tau_1) - \frac{\varepsilon}{2} - \sqrt{\frac{2\mathcal{L}(0)}{n}}\left(\frac{9\varepsilon}{2\left\|D_{j^\star}^{(k)}\right\|} + \frac{\varepsilon'}{\left\|D_{j^\star}^{(k)}\right\|}\right)\right), \tag{54}$$

where above we have combined (19) with the upper bound on $\tau_4 - \tau_1$ from (53). Note that $|\ell_j^\circ(t) - \ell_j^\circ(t')| \leq \left\|D_j^{(k)}\right\|(t-t')$ for any $t, t' \in (t_k, \infty)$. Using (53) again, we get

$$|-\ell_j^\circ(\tau_1) + \ell_j^\circ(\tau_4)| < \left\|D_j^{(k)}\right\||\tau_1 - \tau_4|$$
$$< \left\|D_j^{(k)}\right\|\left(\frac{9\varepsilon}{2\left\|D_{j^\star}^{(k)}\right\|} + \frac{\varepsilon'}{\left\|D_{j^\star}^{(k)}\right\|}\right). \tag{55}$$

Putting (55) back into (54), we conclude that for any $j \in N_U^{(k+1)}$,

$$\|\widetilde{w}_j(\tau_4)\| < \exp_\alpha\left(-\ell_j^\circ(\tau_4) - \frac{\varepsilon}{2} - \left(\left\|D_j^{(k)}\right\| + \sqrt{\frac{2\mathcal{L}(0)}{n}}\right)\left(\frac{9\varepsilon}{2\left\|D_{j^\star}^{(k)}\right\|} + \frac{\varepsilon'}{\left\|D_{j^\star}^{(k)}\right\|}\right)\right). \tag{56}$$

For the converse inequality, we argue in an identical fashion to obtain

$$\|\widetilde{w}_j(\tau_4)\| > \exp_\alpha\left(-\ell_j^\circ(\tau_4) + \frac{\varepsilon}{2} + \left(\left\|D_j^{(k)}\right\| + \sqrt{\frac{2\mathcal{L}(0)}{n}}\right)\left(\frac{9\varepsilon}{2\left\|D_{j^\star}^{(k)}\right\|} + \frac{\varepsilon'}{\left\|D_{j^\star}^{(k)}\right\|}\right)\right). \tag{57}$$

□

This completes the proof of Lemma 7. ■

We can now provide a result from which Lemma 2 follows as a corollary. We will verify that the alignment, norm, and residual conditions propagate forwards after the jump having applied Lemma 7. We will also verify that the norms of neurons in the set $N_U^{(k+1)}$ consequently grow at a rate close to that predicted by Algorithm 1 after the jump. Define for any $\kappa'' \in (t_{k+1}, t_{k+2})$ the time

$$\tau_5 := \inf\Big\{t \geq \tau_4 \;\Big|\; \exists\, j \in [m] \text{ s.t. } \widetilde{\overline{w}}_j^\top \overline{D}_j^{(k+1)} \leq 1 - \alpha^{\varepsilon'/32}$$
$$\text{or } \exists\, i \in S_F^{(k+1)} \text{ s.t. } |h_{\widetilde{\theta}}(x_i) - y_i| \geq \alpha^{\varepsilon'/8} \tag{58}$$
$$\text{or } \exists\, j \in N_U^{(k+1)} \text{ s.t. } \|\widetilde{w}_j\| \geq \alpha^{\varepsilon'}\Big\} \bigwedge \kappa'' \tag{59}$$

This phase is very similar to that described in Appendix D.2, in that it is also a period of slow growth during which the network parameters do not change significantly. Indeed, the structure of the proof is almost identical to that of Lemma 4.

Lemma 8. *For all κ and κ'' such that $\kappa \in (t_k, t_{k+1})$ and $\kappa'' \in (t_{k+1}, t_{k+2})$, there exists ε^* such that for all $\varepsilon \in (0, \varepsilon^*)$ there exists ε'^* such that for all $\varepsilon' \in (0, \varepsilon'^*)$ there exists α^* such that for all $\alpha \in (0, \alpha^*)$ the following are true:*

1. $\tau_5 = \kappa''$.
2. $\forall i \in S_F^{(k+1)}$, $\forall t \in [\tau_4, \tau_5]$, $|h_{\widetilde{\theta}}(x_i) - y_i| < \alpha^{\varepsilon'/8}$.

3. $\forall j \in [m]$, $\forall t \in [\tau_4, \tau_5]$, $\widetilde{\overline{w}}_j^\top \overline{D}_j^{(k+1)} > 1 - \alpha^{\varepsilon'/32}$

4. $\forall j \in N_F^{(k+1)}, \forall t \in [\tau_4, \tau_5]$, $\sqrt{n\left\|D_j^{(k)}\right\|} - \alpha^{\varepsilon'/10} < \|\widetilde{w}_j\| < \sqrt{n\left\|D_j^{(k)}\right\|} + \alpha^{\varepsilon'/10}$.

5. $\forall j \in N_U^{(k+1)}$, $\forall t \in [\tau_4, \tau_5]$, $\left|\ell_j^\circ(t) + \log_\alpha(\|\widetilde{w}_j\|)\right| < \zeta_j(\varepsilon, \varepsilon') + \varepsilon'$, *where* $\zeta_j(\varepsilon, \varepsilon')$ *is as defined in Lemma 7.*

Proof. Given any $\kappa'' \in (t_{k+1}, t_{k+2})$, Item 1 Lemma 7 guarantees that we may select ε and ε' sufficiently small that $\tau_4 < \kappa''$. Assume that this has been done and that α has been selected sufficiently small that all of the conditions of Lemma 7 hold. It remains to check that all of these conditions remain stable for $t \in [\tau_4, \kappa'']$.

Proof of Item 2. In a similar manner to workings for previous phases, we have only to estimate

$$\frac{1}{2\log(1/\alpha)} \frac{\mathrm{d}}{\mathrm{d}t} \left\|E^{(k+1)}\right\|^2 \leq -\frac{1}{n} \sum_{i \in S_F^{(k+1)}} \sum_{j \in N_U^{(k+1)}} A_{i,j} \|\widetilde{w}_j\|^2 \mathcal{D}_j^\top \widetilde{\overline{w}}_j \widetilde{\overline{w}}_j^\top x_i x_i^\top E^{(k+1)}.$$

From the uniform bound on $j \in N_U^{(k+1)}$ given by (59), we have for $i \in S_U^{(k+1)}$ that on $[\tau_4, \kappa'']$, $|h_{\widetilde{\theta}}(x_i)| < m\alpha^{2\varepsilon'}$. Hence,

$$\frac{\mathrm{d}}{\mathrm{d}t} \left\|E^{(k+1)}\right\| \leq \frac{m}{n} \sqrt{2\mathcal{L}(0)} \log(1/\alpha) \alpha^{2\varepsilon'}.$$

Integration implies that

$$\left\|E^{(k+1)}\right\| < \alpha^{\varepsilon'/6} + \frac{m}{n} \sqrt{2\mathcal{L}(0)} \log(1/\alpha) \alpha^{2\varepsilon'} (\kappa'' - \tau_4).$$

Hence,

$$\begin{aligned} \max_{i \in S_F^{(k+1)}} |h_{\widetilde{\theta}}(x_i) - y_i| &< n \left(\alpha^{\varepsilon'/6} + \frac{m}{n} \sqrt{2\mathcal{L}(0)} \log(1/\alpha) \alpha^{2\varepsilon'} (\kappa'' - \tau_4) \right) \\ &< \alpha^{\varepsilon'/8} \end{aligned}$$

for α sufficiently small, as required. □

Proof of Item 3. We again split this argument into the cases $j = j^\star$, $j \in N_F^{(k+1)} \setminus \{j^\star\}$, and $j \in N_U^{(k+1)}$. For the case $j \in N_F^{(k+1)} \setminus \{j^\star\}$, note that Item 2 Lemma 4 provides an initial condition for the alignment of these neurons in the directions $D_j^{(k+1)}$ (since for such neurons, $D_j^{(k+1)} = D_j^{(k)}$ by construction). We can reuse arguments similar to those which led to (12) to obtain

$$\frac{1}{\log(1/\alpha)} \frac{\mathrm{d}}{\mathrm{d}t} \widetilde{\overline{w}}_j^\top D_j^{(k)} > -\frac{2\left\|D_j^{(k+1)}\right\|}{\sqrt{n}} \alpha^{\varepsilon'/8},$$

given that $\|E^{(k)}\| < \alpha^{\varepsilon'/8}$ for $t \in [\tau_1, \tau_5]$. Integrating over this time interval, we get

$$\begin{aligned}\widetilde{\overline{w}}_j^\top D_j^{(k)} &> \widetilde{\overline{w}}_j^\top(\tau_1) D_j^{(k)} - \frac{2\left\|D_j^{(k+1)}\right\|}{\sqrt{n}} \log(1/\alpha)\alpha^{\varepsilon'/8}(\tau_5 - \tau_1) \\ &> \left\|D_j^{(k+1)}\right\| \left(1 - \alpha^{\varepsilon/2}\right) - \frac{2\left\|D_j^{(k+1)}\right\|}{\sqrt{n}} \log(1/\alpha)\alpha^{\varepsilon'/8}(\tau_5 - \tau_1) \\ &> \left\|D_j^{(k)}\right\| \left(1 - \alpha^{\varepsilon'/9}\right),\end{aligned}$$

where the final inequality is valid for α sufficiently small. For the case $j = j^\star$, note that Item 5 Lemma 6 provides an initial condition for the alignment of neuron $j^\star$ in the direction $D_{j^\star}^{(k)}$ (where again we have $D_{j^\star}^{(k+1)} = D_{j^\star}^{(k)}$ by construction). Identical reasoning to the case $j \in N_F^{(k+1)} \setminus \{j^\star\}$, except integrating from τ_3 to τ_5 with initial conditions $\widetilde{\overline{w}}_j^\top(\tau_3) D_j^{(k+1)} > \left\|D_j^{(k+1)}\right\| \left(1 - \alpha^{\varepsilon'}\right)$, gives the desired lower bound for α sufficiently small. For the case $j \in N_U^{(k+1)}$, we have

$$\begin{aligned}\frac{1}{\log(1/\alpha)} \frac{\mathrm{d}}{\mathrm{d}t} \widetilde{\overline{w}}_j^\top D_j^{(k+1)} &= \left\|D_j^{(k+1)}\right\|^2 - \left(\widetilde{\overline{w}}_j^\top D_j^{(k+1)}\right)^2 \\ &\quad - \frac{1}{n} \sum_{i \in S_j^{(k+1)}} h_{\widetilde{\theta}}(x_i) \left(x_i^\top D_j^{(k+1)} - x_i^\top \widetilde{\overline{w}}_j \widetilde{\overline{w}}_j^\top D_j^{(k+1)}\right) \\ &\quad - \frac{1}{n} \sum_{i \in S_F^{(k+1)}} A_{i,j}(h_{\widetilde{\theta}}(x_i) - y_i) \left(x_i^\top D_j^{(k+1)} - x_i^\top \widetilde{\overline{w}}_j \widetilde{\overline{w}}_j^\top D_j^{(k+1)}\right) \\ &> \left\|D_j^{(k+1)}\right\|^2 - \left(\widetilde{\overline{w}}_j^\top D_j^{(k+1)}\right)^2 \\ &\quad - 2\left\|D_j^{(k+1)}\right\| m\alpha^{129\varepsilon'/64} - 2\left\|D_j^{(k+1)}\right\| \alpha^{9\varepsilon'/64} \\ &> \left\|D_j^{(k+1)}\right\|^2 - \left(\widetilde{\overline{w}}_j^\top D_j^{(k+1)}\right)^2 - \alpha^{\varepsilon'/8}.\end{aligned}$$

By similar arguments to previous phases, we have convergence of alignment monotonically for $t \in [\tau_4, \tau_5]$ to a value lower bounded by $\left\|D_j^{(k+1)}\right\| \left(1 - \alpha^{\varepsilon'/16}\right)$. Combining this with the initial conditions from Item 4 Lemma 7, we conclude that

$$\widetilde{\overline{w}}_j^\top D_j^{(k+1)} > \left\|D_j^{(k+1)}\right\| \left(1 - \alpha^{\varepsilon'/32}\right)$$

for all $t \in [\tau_4, \tau_5]$ provided α is selected sufficiently small. □

Proof of Item 4. We verify that the norms of neurons in the set $N_F^{(k)}$ did not change significantly on the interval $[\tau_4, \tau_5]$. Item 3 Lemma 7 provides upper and lower bounds on the norms of neurons in the set $N_F^{(k+1)}$ at time τ_4. In light of these initial conditions and the bound on the loss on data in the set $S_F^{(k+1)}$ given in (58), identical workings to those yielding (15) give

$$\sqrt{n\left\|D_j^{(k)}\right\|} - \alpha^{\varepsilon'/10} < \|\widetilde{w}_j\| < \sqrt{n\left\|D_j^{(k)}\right\|} + \alpha^{\varepsilon'/10}$$

for all $t \in [\tau_4, \tau_5]$ for α sufficiently small.

□

Proof of Items 1 and 5. To estimate the growth of any neuron $j \in N_U^{(k+1)}$, we follow similar arguments to those which yielded (16) in the proof of Lemma 4 to obtain

$$\frac{1}{\log(1/\alpha)}\frac{\mathrm{d}}{\mathrm{d}t}\|\widetilde{w}_j\| > \left(\left\|D_j^{(k+1)}\right\| - \alpha^{\varepsilon'/32} - m\alpha^{2\varepsilon'}\right)\|\widetilde{w}_j\| - \sqrt{2}\alpha^{9\varepsilon'/64}\|\widetilde{w}_j\|$$
$$> \left(\left\|D_j^{(k+1)}\right\| - \alpha^{\varepsilon'/33}\right)\|\widetilde{w}_j\|.$$

Grönwall's inequality yields for α sufficiently small the inequalities

$$\|\widetilde{w}_j(t)\| > \|\widetilde{w}_j(\tau_4)\|\alpha^{-\left\|D_j^{(k+1)}\right\|(t-\tau_4)+\varepsilon'}$$
$$> \alpha^{-\ell_j^\circ(t)+\zeta_j(\varepsilon,\varepsilon')+\varepsilon'},$$

where in the second inequality we inserted (57). Analogous arguments yield the claimed upper bound

$$\|\widetilde{w}_j(t)\| < \alpha^{-\ell_j^\circ(t)-\zeta_j(\varepsilon,\varepsilon')-\varepsilon'},$$

where we have inserted (56). By construction, for every $j \in N_U^{(k+1)}$, $-\ell_j^\circ(\kappa'') > 0$. Therefore, we may select ε and ε' sufficiently small that for all such j, $\|\widetilde{w}_j(\kappa'')\| < \alpha^{\varepsilon'}$. For this choice of ε and ε', we conclude that necessarily $\tau_5 = \kappa''$, since none of the other conditions in the definition of τ_5 break. □

This completes the proof of Lemma 8. ■

D.6 Proof of Theorem 2

In this section, we prove Lemma 1 and Lemma 2 using the multiphase analysis of the previous subsections. Once these two lemmas are proven, it will be possible to provide a short proof of Theorem 2. The proof of Lemma 1 is a fairly direct application of Lemma 4, with the case of the final saddle accounted for by Lemma 9.

Proof of Lemma 1. From Item 4 Lemma 4, we conclude there exists ε sufficiently small such that there exists α sufficiently small that $\tau_1 > \kappa'$. For such ε and α, we have by definition of τ_1 that for all $j \in N_U^{(k)}$ and for all $t \in [\kappa, \kappa']$, $\|\widetilde{w}_j\| < \alpha^\varepsilon$. As a consequence, neurons in the set $N_U^{(k)}$ converge uniformly to $(\mathbf{0})_{j\in N_U^{(k)}}$. For the neurons in the set $N_F^{(k)}$, the combination of Item 2 Lemma 4 and Item 3 Lemma 4 imply

$$\left\|\widetilde{a}_j\widetilde{w}_j - nD_j^{(k)}\right\|^2 = \left\|\widetilde{w}_j\right\|^4 + n^2\left\|D_j^{(k)}\right\|^2 - 2n\widetilde{a}_j\widetilde{w}_j^\top D_j^{(k)}$$
$$< \left(\sqrt{n\left\|D_j^{(k)}\right\|} + \alpha^{\varepsilon/4}\right)^4 + n^2\left\|D_j^{(k)}\right\|^2$$
$$- 2n\left\|D_j^{(k)}\right\|\left(\sqrt{n\left\|D_j^{(k)}\right\|} - \alpha^{\varepsilon/4}\right)^2\left(1-\alpha^{\varepsilon/2}\right)$$
$$< \alpha^{\varepsilon/5},$$

for α sufficiently small. Therefore, $(\widetilde{a}_j\widetilde{w}_j)_{j\in N_F^{(k)}}$ converges uniformly to $\left(nD_j^{(k)}\right)_{j\in N_F^{(k)}}$ on $[\kappa, \kappa']$. ■

Next, we prove Lemma 2, which recovers Hypothesis 1 at saddle $k+1$ and allows us to construct the argument for uniform convergence of the whole process. To prove this lemma, we require that given any $\kappa'' \in (t_{k+1}, t_{k+2})$ (where we will take the convention that $t_{p+1} = \infty$), there exists $\delta'' > 0$ such that for all $\xi'' > 0$, there exists $\alpha^* > 0$ such that for all $\alpha \in (0, \alpha^*)$:

1. $\forall j \in N_U^{(k+1)},\ \left|\ell_j^\circ(\kappa'') + \log_\alpha(\|\widetilde{w}_j(\kappa'')\|)\right| < \xi''$.
2. $\forall j \in N_F^{(k+1)},\ \left|n\left\|D_j^{(k+1)}\right\| - \|\widetilde{w}_j(\kappa'')\|^2\right| < \alpha^{\delta''}$
3. $\forall i \in S_F^{(k+1)},\ |h_{\widetilde{\theta}(\kappa'')}(x_i) - y_i| < \alpha^{\delta''}$.
4. $\forall j \in [m],\ \widetilde{\overline{w}}_j^\top(\kappa'')\overline{D}_j^{(k+1)} > 1 - \alpha^{\delta''}$.

The proof consists of two applications of Lemma 8: one choice of $\varepsilon, \varepsilon'$ in order to get the existence of δ'' such that the alignment, norm, and loss conditions all hold at time κ''; the other choice of $\varepsilon, \varepsilon'$ in order to get the arbitrary precision of the norms of the unfitted neurons.

Proof of Lemma 2. Lemma 8 immediately guarantees that for ε_0 sufficiently small, for ε_0' sufficiently small, and for α sufficiently small, Item 4, Item 2, and Item 3 immediately above all hold with $\delta'' = \varepsilon_0'/32$. For Item 1 immediately above, Item 5 Lemma 8 states that for ε_1 sufficiently small one can choose ε_1' sufficiently small that for α sufficiently small, $\alpha^{-\ell_j^\circ(\kappa'') + \zeta_j(\varepsilon_1, \varepsilon_1') + \varepsilon_1'} < \|\widetilde{w}_j(\kappa'')\| < \alpha^{-\ell_j^\circ(\kappa'') - \zeta_j(\varepsilon_1, \varepsilon_1') - \varepsilon_1'}$. One then chooses ε_1 followed by ε_1' such that $|\zeta_j(\varepsilon_1, \varepsilon_1') + \varepsilon_1'| < \xi''$ for all $j \in N_U^{(k+1)}$. Simply selecting the smaller of the thresholds α_0^* and α_1^* gives the claim. ■

We prove a PL inequality which allows proof of final convergence using arguments lifted from Chatterjee [2022]. Specifically, we will prove uniform convergence of the flow to θ_{pred} on $[T, \infty)$ as $\alpha \to 0$ for any $T > t_p$. The arguments we present will also show that the parameters themselves converge at an exponential rate to a global minimizer of the loss, allowing the following exchange of limits: $\lim_{\alpha\to 0}\lim_{t\to\infty}\widetilde{\theta}(t) = \theta_{\text{pred}} = \lim_{t\to\infty}\lim_{\alpha\to 0}\widetilde{\theta}(t)$. It will not be possible in forthcoming arguments to bound those neurons in the set $N_U^{(p)}$ away from activation boundaries, so we will in fact work with any valid element of the Clarke subdifferential defining Equation (1).

Lemma 9. *Assume that the conditions of Hypothesis 1 hold with $k = p$ at some time $T > t_p$, with a corresponding $\delta > 0$. Define*

$$\tau := \inf\{t \geq T \mid \widetilde{\theta}(t) \notin B(\widetilde{\theta}(T), \alpha^{\delta/4})\}.$$

Then for α sufficiently small, $\tau = \infty$ and $\mathcal{L}(\widetilde{\theta}(t)) \leq \mathcal{L}(\widetilde{\theta}(T))\,\alpha^{\lambda(t-T)}$ for all $t \geq T$, where $\lambda := \min_{j\in N_F^{(p)}}\left\|D_j^{(p)}\right\|$. Furthermore, we have that $\widetilde{\theta}^\alpha$ converges uniformly to θ_{pred} on $[T, \infty)$ as $\alpha \to 0$.

Remark. We remark that in non-rescaled time, this rate is $\exp\left(-\lambda(t - T\log(1/\alpha))\right)$ and so is actually independent of α.

Proof. Because the Clarke subdifferential of the loss is a conservative field [Bolte and Pauwels, 2020, 2021], we have almost everywhere that

$$\begin{aligned}\frac{\mathrm{d}}{\mathrm{d}t}\mathcal{L}(\widetilde{\theta}(t)) &= -\log(1/\alpha)\left\|\widetilde{G}\right\|^2 \\ &\leq -\log(1/\alpha)\inf_{G'\in\partial\mathcal{L}(\widetilde{\theta}(t))}\|G'\|^2, \end{aligned} \tag{60}$$

where above $\widetilde{G}(t) \coloneqq -\dot{\widetilde{\theta}}(t)/\log(1/\alpha)$, which lies in $\partial\mathcal{L}(\widetilde{\theta}(t))$ for almost every t. We thus proceed to lower bound $\left\|\widetilde{G}\right\|^2$. We find by computation that

$$\left\|\widetilde{G}\right\|^2 \geq \frac{1}{n^2}\sum_{j\in[m]}\sum_{i\in[n]}(h_{\widetilde{\theta}}(x_i) - y_i)^2\|\widetilde{w}_j\|^2\inf_{\mathrm{d}'\in\partial\sigma\left(\widetilde{w}_j^\top x_i\right)}(\mathrm{d}')^2.$$

From an application of Lemma 2 with $k = p - 1$, there exists $\delta > 0$ such that for α sufficiently small for all $j \in N_F^{(p)}$ the following are true: $\left| n \left\| D_j^{(p)} \right\| - \|\widetilde{w}_j(T)\|^2 \right| < \alpha^\delta$; $\widetilde{\overline{w}}_j^\top \overline{D}_j^{(p)} \geq 1 - \alpha^\delta$; and $\mathcal{L}(\widetilde{\theta}(T)) \leq \alpha^{2\delta}$. It is clear from the definition of τ that for α sufficiently small, we have that both

$$\|\widetilde{w}_j\|^2 > \frac{n}{2} \left\| D_j^{(p)} \right\|$$

and $\widetilde{w}_j^\top x_i > 0$ for all $i \in S_j^{(p)}$. For any $j \in N_F^{(p)}$ and $t \in [T, \tau)$, we thus obtain

$$\begin{aligned}
\|\widetilde{w}_j\|^2 &= \|\widetilde{w}_j - \widetilde{w}_j(T) + \widetilde{w}_j(T)\|^2 \\
&\geq \|\widetilde{w}_j(T)\|^2 - 2\|\widetilde{w}_j - \widetilde{w}_j(T)\| \, \|\widetilde{w}_j(T)\| \\
&\geq \|\widetilde{w}_j(T)\|^2 - 2\alpha^{\delta/4} \|\widetilde{w}_j(T)\| \\
&> \frac{n}{2} \left\| D_j^{(p)} \right\|.
\end{aligned}$$

Noting that $S_F^{(p)} = [n]$, we have along the trajectory that whenever $\widetilde{\theta}(t) \in B(\widetilde{\theta}(T), \alpha^{\delta/4})$,

$$\inf_{G' \in \partial \mathcal{L}(\widetilde{\theta})} \|G'\|^2 \geq \min_{j \in N_F^{(p)}} \left\| D_j^{(p)} \right\| \mathcal{L}(\widetilde{\theta}) = \lambda \mathcal{L}(\widetilde{\theta}).$$

Substituting this into (60), we see by Grönwall's inequality that for any $t \in (T, \tau)$,

$$\mathcal{L}(\widetilde{\theta}(t)) \leq \mathcal{L}(\widetilde{\theta}(T)) \alpha^{\lambda(t-T)}. \tag{61}$$

We shall use this fact to conclude that $\tau = \infty$. Suppose $\mathcal{L}(\widetilde{\theta}(t_0)) = 0$ for some $t_0 < \tau$. The loss is nonincreasing, so it stays 0, and then every element of $\partial \mathcal{L}$ is 0. Hence $\dot{\widetilde{\theta}} = 0$ almost everywhere on $[t_0, \infty)$, so $\widetilde{\theta}$ is constant and $\tau = \infty$ directly. Otherwise $\mathcal{L} > 0$ on $[T, \tau)$, and we suppose for contradiction that $\tau < \infty$ and define $g(t) := \sqrt{\mathcal{L}(\widetilde{\theta}(t))}$. The supposition $\mathcal{L} > 0$ on $[T, \tau)$ implies that g is bounded away from 0 on any compact interval $[a, b] \subset [T, \tau)$. On such an interval, g is an absolutely continuous function (as it is a composition of such functions) and hence differentiable almost everywhere:

$$g'(t) = \frac{1}{2g(t)} \frac{\mathrm{d}}{\mathrm{d}t} \mathcal{L}(\widetilde{\theta}(t)) = -\log(1/\alpha) \frac{\left\| \widetilde{G}(t) \right\|^2}{2g(t)}.$$

From here, the proof of convergence follows the lines of Chatterjee [2022]. We have

$$\begin{aligned}
\left\|\widetilde{\theta}(b) - \widetilde{\theta}(a)\right\| &\leq \int_a^b \left\|\dot{\widetilde{\theta}}(s)\right\| \mathrm{d}s \\
&= \log(1/\alpha) \int_a^b \left\|\widetilde{G}(s)\right\| \mathrm{d}s \\
&= -2 \int_a^b \frac{g(s)g'(s)}{\left\|\widetilde{G}(s)\right\|} \mathrm{d}s \\
&\leq -\frac{2}{\sqrt{\lambda}} \int_a^b g'(s) \mathrm{d}s \\
&= \frac{2}{\sqrt{\lambda}} \left(\sqrt{\mathcal{L}(\widetilde{\theta}(a))} - \sqrt{\mathcal{L}(\widetilde{\theta}(b))} \right) \\
&\leq 2\sqrt{\frac{\mathcal{L}(\widetilde{\theta}(a))}{\lambda}}.
\end{aligned}$$

As g is continuous on $[T, \tau]$ we may take $a \to T$ and $b \to \tau$ to conclude that

$$\begin{aligned}
\left\|\widetilde{\theta}(\tau) - \widetilde{\theta}(T)\right\| &\leq 2\sqrt{\frac{\mathcal{L}(\widetilde{\theta}(T))}{\lambda}} \\
&\leq \frac{2}{\sqrt{\lambda}} \alpha^{\delta} \\
&< \alpha^{\delta/4},
\end{aligned}$$

where the penultimate inequality follows from the bound supplied by Hypothesis 1 and the final inequality from taking α sufficiently small. This contradiction implies that $\tau = \infty$. Convergence of the loss to zero is immediate from noting that (61) is valid for all $t > T$, supplying the inequality

$$\mathcal{L}(\widetilde{\theta}(t)) \leq \alpha^{2\delta + \lambda(t-T)} \underset{t \to \infty}{\to} 0,$$

from which we see that the rate is indeed exponential. To prove that the parameters themselves converge, we use similar reasoning to the above to see that for s, t such that $T \leq s < t$ we have

$$\left\|\widetilde{\theta}(t) - \widetilde{\theta}(s)\right\| \leq 2\sqrt{\frac{\mathcal{L}(\widetilde{\theta}(T))}{\lambda}} \alpha^{\lambda(s-T)/2}.$$

Since this bound, independent of t, approaches 0 as $s \to \infty$, we conclude that the flow satisfies the Cauchy property and therefore converges. At this point we can conclude the following uniform convergence on $[T, \infty)$:

$$\begin{aligned}
\sup_{t \geq T} \left\|\widetilde{\theta}(t) - \theta_{\mathrm{pred}}\right\| &\leq \sup_{t \geq T} \left\|\widetilde{\theta}(T) - \theta_{\mathrm{pred}}\right\| + \sup_{t \geq T} \left\|\widetilde{\theta}(t) - \widetilde{\theta}(T)\right\| \\
&\underset{\alpha \to 0}{\longrightarrow} 0.
\end{aligned}$$

■

Finally, we are able to provide a proof of Theorem 2. This proof consists of iteratively concluding uniform convergence on arbitrary closed intervals contained in (t_k, t_{k+1}) using Hypothesis 1, and then using Lemma 2 to apply the same reasoning to the next interval.

Proof of Theorem 2. Let $\mathcal{S} \subset \mathbb{R}_{\geq 0} \setminus \{t_1, \dots, t_p\}$ be a closed set. Consider the subset $\overline{\mathcal{S}} := \mathcal{S} \cap \{t \mid \frac{1}{4}\sqrt{\frac{n}{2\mathcal{L}(0)}} \leq t \leq T\}$ for some $T > t_p$. The Heine-Borel theorem implies that $\overline{\mathcal{S}}$ is compact, so there exists a cover of $\overline{\mathcal{S}}$ by disjoint closed intervals $([\kappa_k, \kappa_k'])_{k=0}^p$ whose endpoints satisfy $\kappa_k, \kappa_k' \in (t_k, t_{k+1})$. By Lemma 3, we have that Hypothesis 1 holds for any $\kappa_0 \in \left(0, \frac{1}{4}\sqrt{\frac{n}{2\mathcal{L}(0)}}\right)$. From Lemma 1, we obtain that for such κ_0 and for any $\kappa_0' \in (\kappa_0, t_1)$, $\widetilde{\theta}$ converges uniformly to $\theta^\circ = \mathbf{0}$ on $[\kappa_0, \kappa_0']$. Note that uniform convergence on $[0, \kappa_0)$ is a consequence of the arguments proving Lemma 3. Furthermore, we have from Lemma 2 that for any $\kappa_1 \in (t_1, t_2)$, Hypothesis 1 also holds. Iterating this argument, we can use Lemma 1 to conclude that there exists $\alpha_k^* > 0$ such that $\widetilde{\theta}$ converges uniformly to θ° on $[\kappa_k, \kappa_k']$ with rate α^{δ_k} for some $\delta_k > 0$, followed by Lemma 2 to conclude that Hypothesis 1 holds at each $\kappa_{k+1} \in (t_{k+1}, t_{k+2})$. From this reasoning, we see that Hypothesis 1 holds at time T, after which Lemma 9 provides uniform convergence on $\mathcal{S} \setminus \overline{\mathcal{S}}$. Since the number of saddles is finite, we may select $\alpha \in (0, \min_{k \in \{0,\dots,p\}} \alpha_k^*)$ so that each application of Lemma 1 and Lemma 2 holds. This gives the claimed uniform convergence, the rate of which is α^{δ^*} where $\delta^* = \min_{k \in \{0,\dots,p\}} \delta_k$. ■

E Proof of Theorem 3

We consider in this section the setting where we have only positive labels and condition all the probabilities on the fact that $a_j > 0$ for all $j \in [m]$. The argument generalizes to the case of mixed signs, by treating positive and negative labels separately.

Let p denote the number of iterations of the `while` loop in Algorithm 1, and let $(j_\star^{(0)}, \dots, j_\star^{(p-1)})$ denote the neurons successively selected at Line 8 of Algorithm 1. For any $k \in \{0, \dots, p-1\}$, the neuron $j_\star^{(k)}$ is removed from the set of unfitted neurons at iteration k, so that

$$D^{(p)}_{j_\star^{(k)}} = D^{(k)}_{j_\star^{(k)}} = \frac{1}{n} \sum_{i \in S_U^{(k)} \cap S_{j_\star^{(k)}}} y_i x_i, \qquad \text{and thus} \qquad n \left\| D^{(p)}_{j_\star^{(k)}} \right\|_2 = \sqrt{\sum_{i \in S_U^{(k)} \cap S_{j_\star^{(k)}}} y_i^2}$$

by Assumption 1. For any neuron $j \in N_U^{(p)}$, i.e., any neuron that is never selected, we have $D_j^{(p)} = \frac{1}{n} \sum_{i \in S_U^{(p)} \cap S_j} y_i x_i = 0$, since $S_U^{(p)} \cap S_j = \emptyset$ by the termination condition of the `while` loop. Recall that Theorem 3 provides a bound on the squared ℓ_2-norm of θ_{pred}, which thus satisfies

$$\begin{aligned} \frac{1}{2}\|\theta_{\mathrm{pred}}\|^2 &= \sum_{j \in [m]} n \|D_j^{(p)}\|_2 \\ &= \sum_{k=0}^{p-1} \sqrt{\sum_{i \in S_U^{(k)} \cap S_{j_\star^{(k)}}} y_i^2}. \end{aligned} \tag{62}$$

In contrast with a proof based on a fixed ordering of the neurons, which is independent of both data and initialization, we work here directly with the ordering $(j_\star^{(0)}, \dots, j_\star^{(p-1)})$ produced by Algorithm 1. To do so, we establish a property of the initialization which holds with high probability simultaneously for all sets of neurons of at most logarithmic cardinality, and hence in particular for the sets of neurons successively selected by Algorithm 1.

Note that when initializing the neurons according to Assumption 3, and conditioning on the fact that $a_i > 0$, $(A_{i,j})_{j \in [m],\, i \in [n]}$ are independent $\mathrm{Bernoulli}\left(\frac{1}{2}\right)$ variables. For any set of neurons $K \subseteq [m]$, we define

$$P_K = \{i \in [n] \mid A_{i,j} = 0 \text{ for all } j \in K\} = [n] \setminus \bigcup_{j \in K} S_j,$$

which is the set of data points on which all the neurons in K are inactive, i.e., the data points that remain to be fitted once all the neurons in K have grown. Since $S_U^{(0)} = [n]$ and $S_U^{(k+1)} = S_U^{(k)} \setminus S_{j_\star^{(k)}}$ in Algorithm 1, we have

$$S_U^{(k)} = P_{\left\{j_\star^{(0)}, \dots, j_\star^{(k-1)}\right\}} \qquad \text{for any } k \in \{0, \dots, p\}. \tag{63}$$

For a fixed $\delta \in (0, \frac{1}{2})$, denote

$$L = \left\lfloor \frac{\log(n)}{2\log\left(\frac{1}{\frac{1}{2}+\delta}\right)} \right\rfloor,$$

and define the event

$$E = \Bigg\{ \text{for all } K \subseteq [m] \text{ with } |K| \leq L \text{ and all } j \in [m] \setminus K, \\ |P_K| \geq \sqrt{n} \implies \left| \left|P_K \setminus S_j\right| - \frac{1}{2}|P_K| \right| \leq \delta |P_K| \Bigg\}.$$

The interpretation of the event E—which has high probability, as shown below—is that, as long as at most L neurons have been fitted and at least $\sqrt{n}$ data points remain to be fitted, fitting any additional neuron almost halves the cardinality of the set of unfitted data. Note that the event E is defined in terms of the matrix A only, and does not refer to Algorithm 1. Note also that, since $\left|P_K \cap S_j\right| = |P_K| - \left|P_K \setminus S_j\right|$, the inequality in the definition of E is equivalent to

$$\left| \left|P_K \cap S_j\right| - \frac{1}{2}|P_K| \right| \leq \delta |P_K|, \tag{64}$$

so that in particular both $\left|P_K \setminus S_j\right| \leq \left(\frac{1}{2} + \delta\right)|P_K|$ and $\left|P_K \cap S_j\right| \leq \left(\frac{1}{2} + \delta\right)|P_K|$.

Proposition 4. *For any $\delta \in (0, \frac{1}{2})$, n and m,*

$$\mathbb{P}\left(\neg E\right) \leq 2(L+1)\exp\left((L+1)\log(m) - 2\delta^2\sqrt{n}\right).$$

Proof. Fix a set $K \subseteq [m]$ with $|K| \leq L$ and a neuron $j \in [m] \setminus K$. The set P_K is a function of $(A_{i,j'})_{i\in[n],\, j'\in K}$ only. Since the variables $(A_{i,j})_{i,j}$ are independent and $j \notin K$, $(A_{i,j})_{i\in[n]}$ is independent from $(A_{i,j'})_{i\in[n],\, j'\in K}$. Consequently,

$$\left|P_K \setminus S_j\right| = \sum_{i \in P_K} (1 - A_{i,j}) \;\Bigg|\; (A_{i,j'})_{\substack{i\in[n] \\ j'\in K}} \sim \mathrm{Bin}\left(|P_K|, \frac{1}{2}\right).$$

For any fixed set $P \subseteq [n]$, the event $\{P_K = P\}$ depends only on $(A_{i,j'})_{i\in[n],\, j'\in K}$, and is thus independent from $(A_{i,j})_{i\in[n]}$, since the variables $(A_{i,j})_{i,j}$ are independent and $j \notin K$. Moreover, $\left|P \setminus S_j\right| = \sum_{i\in P}(1 - A_{i,j}) \sim \mathrm{Bin}\left(|P|, \frac{1}{2}\right)$. Using Hoeffding's inequality [see, e.g., Vershynin, 2026, Theorem 2.2.1], we have for any $N \geq 1$ and $X \sim \mathrm{Bin}(N, \frac{1}{2})$ that $\mathbb{P}\left(\left|X - \frac{N}{2}\right| \geq \delta N\right) \leq 2\exp\left(-2\delta^2 N\right)$.

Since P_K takes values in the finite set of subsets of $[n]$, the events $\{P_K = P\}$, $P \subseteq [n]$, form a partition of the sample space. Moreover, on the event $\{P_K = P\}$, the event

$$\left\{ |P_K| \geq \sqrt{n} \text{ and } \left||P_K \setminus S_j| - \frac{1}{2}|P_K|\right| > \delta |P_K| \right\}$$

coincides with

$$\left\{ |P| \geq \sqrt{n} \text{ and } \left||P \setminus S_j| - \frac{1}{2}|P|\right| > \delta |P| \right\},$$

which is empty if $|P| < \sqrt{n}$ and otherwise depends only on $(A_{i,j})_{i\in[n]}$. Hence, by the law of total probability and independence, summing over all the possible realizations P of P_K with $|P| \geq \sqrt{n}$,

$$
\begin{aligned}
&\mathbb{P}\left(|P_K| \geq \sqrt{n} \text{ and } \left|\, \left|P_K \setminus S_j\right| - \frac{1}{2}|P_K| \,\right| > \delta|P_K|\right) \\
&\quad = \sum_{\substack{P\subseteq[n] \\ |P|\geq\sqrt{n}}} \mathbb{P}\left(P_K = P\right) \mathbb{P}\left(\left|\, \left|P \setminus S_j\right| - \frac{1}{2}|P| \,\right| > \delta|P|\right) \\
&\quad \leq \sum_{\substack{P\subseteq[n] \\ |P|\geq\sqrt{n}}} \mathbb{P}\left(P_K = P\right) 2\exp\left(-2\delta^2|P|\right) \\
&\quad \leq 2\exp\left(-2\delta^2\sqrt{n}\right) \sum_{P\subseteq[n]} \mathbb{P}\left(P_K = P\right) \\
&\quad = 2\exp\left(-2\delta^2\sqrt{n}\right).
\end{aligned}
$$

We then apply a union bound over all the pairs (K, j) with $K \subseteq [m]$, $|K| \leq L$ and $j \in [m] \setminus K$.[2] For any $l \in \{0, \ldots, L\}$, the number of sets $K \subseteq [m]$ with $|K| = l$ is $\binom{m}{l} \leq m^l$, and there are at most m choices for j, so that the total number of such pairs is at most $\sum_{l=0}^{L} m^{l+1} \leq (L+1)m^{L+1}$. Thus

$$
\begin{aligned}
\mathbb{P}\left(\neg E\right) &\leq (L+1)m^{L+1} \cdot 2\exp\left(-2\delta^2\sqrt{n}\right) \\
&= 2(L+1)\exp\left((L+1)\log(m) - 2\delta^2\sqrt{n}\right). \qquad \blacksquare
\end{aligned}
$$

Note that as δ approaches zero here, the probability bound degrades, but the event E gets more restrictive, as δ controls the relative distance by which the cardinality of the set of unfitted data after fitting an additional neuron can deviate from half of its previous value.

We now apply the event E repeatedly along the iterations of Algorithm 1, until the cardinality of the set of unfitted data falls below $\sqrt{n}$. To this end, define

$$
\hat{k} = \min\left(\left\{k \in \{0, \ldots, p\} \;\middle|\; \left|S_U^{(k)}\right| < \sqrt{n}\right\} \cup \{p\}\right),
$$

In natural language, $\hat{k}$ is the first iteration of Algorithm 1 after which fewer than $\sqrt{n}$ data points remain to be fitted, and $\hat{k} = p$ if there is no such iteration. By definition, $\left|S_U^{(k)}\right| \geq \sqrt{n}$ for any $k < \hat{k}$.

Proposition 5. *The event E implies that*

$$
\left|S_U^{(k)}\right| \leq n\left(\frac{1}{2} + \delta\right)^k \qquad \textit{for any } k \in \{0, \ldots, \hat{k}\}, \tag{65}
$$

$$
\textit{and} \qquad \left|S_U^{(k)} \cap S_{j_\star^{(k)}}\right| \leq n\left(\frac{1}{2} + \delta\right)^{k+1} \qquad \textit{for any } k \in \{0, \ldots, \hat{k} - 1\}. \tag{66}
$$

Proof. We assume that E holds and prove Equation (65) by induction on k, establishing Equation (66) along the way. For $k = 0$, we have $\left|S_U^{(0)}\right| = n$. Now let $k \in \{0, \ldots, \hat{k} - 1\}$ and assume that $\left|S_U^{(k)}\right| \leq n\left(\frac{1}{2} + \delta\right)^k$. Since $k < \hat{k}$, we have $\left|S_U^{(k)}\right| \geq \sqrt{n}$, so that

$$
\begin{aligned}
&\sqrt{n} \leq n\left(\frac{1}{2} + \delta\right)^k \\
\implies \qquad &k \leq \frac{\log(n)}{2\log\left(\frac{1}{\frac{1}{2}+\delta}\right)}.
\end{aligned}
$$

Since k is an integer, this gives $k \leq L$. Let $K = \left\{j_\star^{(0)}, \ldots, j_\star^{(k-1)}\right\}$ and consider neuron $j_\star^{(k)}$. The neurons $j_\star^{(0)}, \ldots, j_\star^{(k-1)}$ are pairwise distinct, as each selected neuron is removed from the set of

[2] Since the event E holds simultaneously for all such pairs, its conclusion applies in particular to the sets of neurons selected by Algorithm 1, whatever they are. It is thus never necessary to know the distribution of the variables $(A_{i,j})_{i,j}$ conditionally on the previous iterations of Algorithm 1.

unfitted neurons, so that $|K| = k \leq L$. Moreover, $j_\star^{(k)} \in N_U^{(k)} = [m] \setminus K$. By Equation (63), $S_U^{(k)} = P_K$, so that $|P_K| \geq \sqrt{n}$. The event E and Equation (64) then give

$$\left|P_K \setminus S_{j_\star^{(k)}}\right| \leq \left(\frac{1}{2} + \delta\right)|P_K| \qquad \text{and} \qquad \left|P_K \cap S_{j_\star^{(k)}}\right| \leq \left(\frac{1}{2} + \delta\right)|P_K|.$$

Since $S_U^{(k+1)} = S_U^{(k)} \setminus S_{j_\star^{(k)}} = P_K \setminus S_j$ and $S_U^{(k)} \cap S_{j_\star^{(k)}} = P_K \cap S_{j_\star^{(k)}}$, we obtain

$$\begin{aligned} \left|S_U^{(k+1)}\right| &\leq \left(\frac{1}{2} + \delta\right)\left|S_U^{(k)}\right| \leq n\left(\frac{1}{2} + \delta\right)^{k+1} \\ \text{and} \qquad \left|S_U^{(k)} \cap S_{j_\star^{(k)}}\right| &\leq \left(\frac{1}{2} + \delta\right)\left|S_U^{(k)}\right| \leq n\left(\frac{1}{2} + \delta\right)^{k+1}, \end{aligned}$$

which concludes the induction and proves Equation (66).

■

Proposition 5 demonstrates that, under the event E, the cardinality of the set $S_U^{(k)}$ decreases by roughly one half at each iteration of Algorithm 1, until it falls below $\sqrt{n}$. Note that this holds for the ordering $(j_\star^{(0)}, \ldots, j_\star^{(p-1)})$ returned by Algorithm 1, and irrespective of how ties in the $\arg\max$ at Line 8 would be broken.

In the following, we work under the event E and bound the squared ℓ_2-norm of the estimator returned by Algorithm 1 by splitting Equation (62) into two terms: that induced by the first $\hat{k}$ iterations of the algorithm, and the remaining ones.

Proposition 6. *If for all $i \in [n]$, $|y_i| \leq y_{\max}$, then*

$$E \implies \sum_{k=0}^{\hat{k}-1} \sqrt{\sum_{i \in S_U^{(k)} \cap S_{j_\star^{(k)}}} y_i^2} \leq \frac{\sqrt{\frac{1}{2} + \delta}}{1 - \sqrt{\frac{1}{2} + \delta}} y_{\max} \sqrt{n}.$$

Proof. The bound $|y_i| \leq y_{\max}$ directly implies that

$$\sum_{k=0}^{\hat{k}-1} \sqrt{\sum_{i \in S_U^{(k)} \cap S_{j_\star^{(k)}}} y_i^2} \leq y_{\max} \sum_{k=0}^{\hat{k}-1} \sqrt{\left|S_U^{(k)} \cap S_{j_\star^{(k)}}\right|}.$$

Note that E implies that for any $k \leq \hat{k} - 1$, $\left|S_U^{(k)} \cap S_{j_\star^{(k)}}\right| \leq n\left(\frac{1}{2} + \delta\right)^{k+1}$ thanks to Equation (66). So that if it holds,

$$\begin{aligned} \sum_{k=0}^{\hat{k}-1} \sqrt{\sum_{i \in S_U^{(k)} \cap S_{j_\star^{(k)}}} y_i^2} &\leq y_{\max} \sqrt{n} \sum_{k=0}^{\hat{k}-1} \left(\frac{1}{2} + \delta\right)^{(k+1)/2} \\ &\leq y_{\max} \sqrt{n} \cdot \frac{\sqrt{\frac{1}{2} + \delta}}{1 - \sqrt{\frac{1}{2} + \delta}}. \end{aligned}$$

■

Finally, we bound the squared ℓ_2-norm of the remaining increments. Note that, unlike Proposition 6, the following bound holds deterministically.

Proposition 7. *If for all $i \in [n]$, $|y_i| \leq y_{\max}$, then*

$$\sum_{k=\hat{k}}^{p-1} \sqrt{\sum_{i \in S_U^{(k)} \cap S_{j_\star^{(k)}}} y_i^2} \leq y_{\max} \sqrt{n}.$$

Proof. If $\hat{k} = p$, the sum is empty and the bound holds trivially. Otherwise, $\left|S_U^{(\hat{k})}\right| < \sqrt{n}$ by definition of $\hat{k}$. By subadditivity of the square root, we have

$$\begin{aligned}\sum_{k=\hat{k}}^{p-1} \sqrt{\sum_{i \in S_U^{(k)} \cap S_{j_\star^{(k)}}} y_i^2} &\leq \sum_{k=\hat{k}}^{p-1} \sum_{i \in S_U^{(k)} \cap S_{j_\star^{(k)}}} |y_i| \\ &= \sum_{i \in S_U^{(\hat{k})}} |y_i| \\ &\leq y_{\max} \left|S_U^{(\hat{k})}\right| \\ &< y_{\max} \sqrt{n}. \end{aligned}$$

■

Combining Equation (62) with Propositions 4, 6 and 7, and noting that $\frac{\sqrt{\frac{1}{2}+\delta}}{1-\sqrt{\frac{1}{2}+\delta}} + 1 = \frac{1}{1-\sqrt{\frac{1}{2}+\delta}}$, we obtain that for any $\delta \in (0, \frac{1}{2})$, n and m,

$$\sum_{j \in [m]} n \|D_j^{(p)}\| \leq \frac{1}{1 - \sqrt{\frac{1}{2} + \delta}} y_{\max} \sqrt{n}$$

with probability at least $1 - 2(L+1)\exp\left((L+1)\log(m) - 2\delta^2\sqrt{n}\right)$. We conclude taking $\delta = \frac{1}{16}$, for which $\sqrt{\frac{1}{2}+\delta} = \frac{3}{4}$, $2\delta^2 = \frac{1}{128}$ and $L = \left\lfloor \frac{\log(n)}{2\log(16/9)} \right\rfloor$, that for any n and m,

$$\sum_{j \in [m]} n \|D_j^{(p)}\| \leq 4 y_{\max} \sqrt{n}$$

with probability at least $1 - 2(L+1)\exp\left((L+1)\log(m) - \frac{\sqrt{n}}{128}\right)$. In particular, if $\log(m) = o(\sqrt{n}/\log(n))$, this probability is at least $1 - \exp(-\Omega(\sqrt{n}))$. This concludes the proof when considering only positive labels and neurons, and Theorem 3 follows from a symmetric argument.

F Auxiliary results

We provide in this section auxiliary results that are useful for our analysis, as well as a proof of Proposition 2.

F.1 Proof of Proposition 2

Proof of Item (i). The proof is a simple union bound. Indeed, observe that 1) $(A_{i,j})_{j \in [m]}$ are independent i.i.d. Bernoulli$\left(\frac{1}{4}\right)$ variables. Then, the probability for a row i to be zero is exactly $\left(\frac{3}{4}\right)^m$, so that the probability that any row is zero is smaller than $n\left(\frac{3}{4}\right)^m$.

2) Conditionally on a_j, $(A_{i,j})_{i \in [n]}$ are independent variables, each of them being either the deterministic variable 0 if $i \notin I_{s_j}$, or a Bernoulli$\left(\frac{1}{2}\right)$ $i \in I_{s_j}$. Since $|I_{s_j}| \geq \min(n_-, n_+)$ almost surely by definition, we then have that a column j is zero with probability at most $\left(\frac{1}{2}\right)^{\min(n_-,n_+)}$. So that the probability for any column to be zero is bounded by $m\left(\frac{1}{2}\right)^{\min(n_-,n_+)}$.

3) Note that two non-zero columns j and j' are equal if and only if both $\operatorname{sgn}(a_j) = \operatorname{sgn}(a_{j'})$ and $\mathbb{1}(w_j^\top x_i) = \mathbb{1}(w_{j'}^\top x_i)$ for all $i \in I_{s_j}$. All these variables are independent Bernoulli$\left(\frac{1}{2}\right)$, so that a union bound argument again yields that the probability for two non-zero columns to be equal is at most $\frac{m(m-1)}{2}\left(\frac{1}{2}\right)^{\min(n_-,n_+)+1}$.

A union bound on these three events and summing the different terms then yields Item (i) of Proposition 2. □

Proof of Item (ii). Similarly to the proof of Theorem 3, we here fix an arbitrary ordering $\pi = (j_1, \dots, j_m)$, and define an alternative Algorithm 1 with the fixed ordering π, i.e.,

$$S_U^0 = [n], \quad S_U^{k+1} = S_U^k \setminus S_{j_k}$$
$$S_j = \{i \in [n] \mid A_{i,j} = 1\}.$$

Similarly, we also define the increment times as $t'_0 = 0$, and for any $k \geq 0$:

$$t'_{k+1} = \max(t'_k - \frac{\ell'_{j_{k+1}}(t'_k)}{\|D^k_{j_{k+1}}\|}, t'_k)$$
$$\text{where} \quad D^k_j = \frac{1}{n} \sum_{i \in S_U^k \cap S_j} y_i x_i,$$
$$\ell'_j(0) = -1 \text{ and } \quad \ell'_j(t) = \ell'_j(t'_k) + (t - t'_k)\|D^k_j\| \text{ for any } t \in [t'_k, t'_{k+1}],$$

where $\frac{\ell'_{j_{k+1}}(t'_k)}{\|D^k_{j_{k+1}}\|} = -\infty$ when $\|D^k_{j_{k+1}}\| = 0$ by convention. Note that when π coincides with $j_\star$ on its first p elements, we have the equality $(t_0, \dots, t_p) = (t'_0, t'_p)$, as well as $\ell'_j(t) = \ell^\circ_j(t)$ on $[0, t_p]$ and $D^k_j = D^{(k)}_j$ for $k \leq p$.

From there, we can see all t'_k, $\ell'_j(t)$ and $\|D^k_j\|$ as functions of $(y_1, \dots, y_n)$. Consider two neurons $j, j^\dagger$. Since we assume Assumption 4 (i) holds, the columns of the masking matrix $A_{\cdot,j}$ and $A_{\cdot,j^\dagger}$ are both non-zero and distinct; and also consider k such that $j_l \notin \{j, j^\dagger\}$ for all $l \leq k$, i.e., both neurons j and $j^\dagger$ have not been selected by the alternative algorithm in the first k increments. From there, we want to show that the following system of inequalities

$$-\frac{\ell'_j(t'_k)}{\|D^k_j\|} = -\frac{\ell'_{j^\dagger}(t'_k)}{\|D^k_{j^\dagger}\|}, \tag{67}$$
$$\|D^k_j\| > 0 \quad \text{and} \quad \max(\ell'_{j^\dagger}(t'_k), \ell'_j(t'_k)) < 0 \tag{68}$$

only holds on a measure zero set for the values of $(y_1, \dots, y_n)$. Indeed, since the two masking columns are different, there exists some $i \in [n]$ such that, w.l.o.g., $A_{i,j} = 1$ and $A_{i,j^\dagger} = 0$. We then have two cases.

Either $i \in S_U^k$, i.e., i remains to be fitted at iteration k. In that case, note that all $(t'_0, \dots, t'_k)$, $\ell'_{j^\dagger}$ and $(D^0_{j^\dagger}, \dots, D^k_{j^\dagger})$ do not depend on the value y_i. On the other hand, $-\frac{\ell'_j(t'_k)}{\|D^k_j\|}$ decreases (strictly) with $|y_i|$, when Equation (68) holds. Indeed, note that

$$-\ell'_j(t'_k) = 1 - \sum_{l<k}(t'_{l+1} - t'_l)\|D^l_j\| \quad \text{and} \quad \|D^l_j\| = \frac{1}{n}\sqrt{\sum_{i' \in S_U^l \cap S_j} y_{i'}^2}.$$

so that $-\ell'_j(t'_k)$ decreases with $|y_i|$ and is positive, while $\|D^k_j\|$ increases with it. Consequently, $-\frac{\ell'_j(t'_k)}{\|D^k_j\|}$ is a decreasing function of $|y_i|$, so that for fixed values $(y_{i'})_{i' \neq i}$, there is at most one value of $|y_i|$ such that the system of Equations (67) and (68) holds. This implies over all labels that the system of Equations (67) and (68) can only be satisfied on a zero measure set.

The other possible case is when for all i such that $A_{i,j} \neq A_{i,j^\dagger}$, $i \notin S_U^k$. In other words, this means that $S_U^k \cap S_j = S_U^k \cap S_{j^\dagger}$. By definition of the quantities of interest, if the system of Equations (67) and (68) holds for k, the following system also holds

$$\ell'_j(t'_{\overline{k}}) = \ell'_{j^\dagger}(t'_{\overline{k}}), \tag{69}$$
$$\|D^{\overline{k}}_j\| > 0 \quad \text{and} \quad \max(\ell'_{j^\dagger}(t'_{\overline{k}}), \ell'_j(t'_{\overline{k}})) < 0, \tag{70}$$

where $\overline{k} \leq k$ is the largest integer such that $S_U^{\overline{k}} \cap S_j \neq S_U^{\overline{k}} \cap S_{j^\dagger}$—indeed, for any $\overline{k} < l \leq k$, we would have $D^l_j = D^l_{j^\dagger}$. Now note that we could have chosen i such that $A_{i,j} \neq A_{i,j^\dagger}$ and it is the last such i to be fitted by the algorithm. In that case, we necessarily have $i \in S_U^{\overline{k}}$ and we are in a case very similar to the one treated above. We can then show with a similar study—which is even

simpler here, as we only have to consider the numerator of the previous fraction—that the system of Equations (69) and (70) only holds for a zero measure set.

We thus have proven that the system of Equations (67) and (68) only holds on a zero measure set for the values of $(y_1, \dots, y_n)$, for a fixed ordering π, iteration k, and neurons $j, j^\dagger$ not yet fitted at iteration k. By doing a finite union, it remains true that this system is only satisfied on a zero measure set for all choices of ordering, iterations and neuron couples. In particular, it holds for some ordering that coincides with the one returned by Algorithm 1 on the first p iterations, i.e., $(j_\star^{(1)}, \dots, j_\star^{(p)})$. For this specific choice of ordering, we observe that the case of a non-unique minimizer of the loop in Algorithm 1 corresponds to a couple satisfying the system of Equations (67) and (68), such that both neurons have not fitted yet, which concludes the proof of Item (ii) of Proposition 2. □

F.2 Other auxiliary results

The following result estimates the normalized component of $\widetilde{w}_j$ in the direction x_i under the assumption that this neuron is aligned in direction $D_j^{(k)}$.

Proposition 8. *For any $j \in [m]$ and $k \in [p]$, if $\widetilde{\overline{w}}_j^\top \overline{D}_j^{(k)} > 1 - \alpha$, then for all $i \in [n]$,*

$$\left| \widetilde{\overline{w}}_j^\top x_i - \frac{\mathbb{1}\left(i \in S_j^{(k)}\right) y_i}{n \left\| D_j^{(k)} \right\|} \right| < \sqrt{2} \alpha^{1/2}.$$

Proof. Let $u = \sum_i u_i x_i$ and $v = \sum_i v_i x_i$ be the representations of two unit vectors in an extension of $(x_i)_{i=1}^d$ to an orthonormal basis given by such that $u^\top v > 1 - \alpha$. We have for any i_0

$$\begin{aligned}(u_{i_0} - v_{i_0})^2 &\leq \sum_i (u_i - v_i)^2 \\ &= \sum_i \left(u_i^2 + v_i^2 - 2 u_i v_i\right) \\ &< 2\alpha.\end{aligned}$$

Taking square roots yields $|u_{i_0} - v_{i_0}| < \sqrt{2}\alpha^{1/2}$. Recalling that

$$\overline{D}_j^{(k)} = \frac{1}{n \left\| D_j^{(k)} \right\|} \sum_{i \in [n]} \mathbb{1}\left(i \in S_j^{(k)}\right) y_i x_i$$

gives the claim. ■

The following result applies the preceding proposition to deduce an inequality which will prove useful in our analysis.

Corollary 1. *If $\widetilde{\overline{w}}_j^\top \overline{D}_j^{(k)} > 1 - \alpha$, then*

$$\left| x_i^\top D_j^{(k)} - (x_i^\top \widetilde{\overline{w}}_j)(\widetilde{\overline{w}}_j^\top D_j^{(k)}) \right| < 2 \left\| D_j^{(k)} \right\| \alpha^{1/2}$$

for α sufficiently small.

Proof. We simply plug in the bounds from Proposition 8 to obtain

$$\begin{aligned}-\sqrt{2} \left\| D_j^{(k)} \right\| \alpha^{1/2} &< x_i^\top D_j^{(k)} - (x_i^\top \widetilde{\overline{w}}_j)(\widetilde{\overline{w}}_j^\top D_j^{(k)}) \\ &< \frac{y_i}{n}\alpha + \sqrt{2} \left\| D_j^{(k)} \right\| \left(\alpha^{1/2} - \alpha^{3/2}\right),\end{aligned}$$

from which the claim follows. ■

Lemma 10. *Suppose that the following hold:*

1. *$F \in \mathcal{C}^1([T_1, T_2])$ and F' is Lipschitz continuous.*

2. $F \geq 0$ *and* $F(T_1) = 0$.

3. $F''(t) \leq pF(t) + q$ *for almost every* $t \in [T_1, T_2]$, *where* $p, q > 0$.

Then for every $t \in [T_1, T_2]$,

$$F'(t) \leq \left(\frac{q}{\sqrt{p}} + F'(T_1) \right) \exp(\sqrt{p}(t - T_1)).$$

Proof. Let $G = F + q/p$. Then G obeys for almost every $t \in [T_1, T_2]$ the differential inequality $G'' < pG$. We compute using this inequality that

$$\frac{\mathrm{d}}{\mathrm{d}t} \left(G' + \sqrt{p}G \right) \leq pG + \sqrt{p}G' = \sqrt{p} \left(\sqrt{p}G + G' \right).$$

Since F' is Lipschitz continuous, both F' and F are in fact absolutely continuous. Therefore, $\sqrt{p}G + G'$ is absolutely continuous. We may apply Grönwall's inequality to see that

$$\sqrt{p}G + G' \leq \exp\left(\sqrt{p}(t - T_1) \right) \left(\sqrt{p}G(T_1) + G'(T_1) \right).$$

Noting that $G \geq F$, the imposition of the condition that $F \geq 0$ gives

$$G' \leq \exp\left(\sqrt{p}(t - T_1) \right) \left(\sqrt{p}G(T_1) + G'(T_1) \right).$$

Substituting F into the above and using the initial condition $F(T_1) = 0$, we get

$$F'(t) \leq \left(\frac{q}{\sqrt{p}} + F'(T_1) \right) \exp(\sqrt{p}(t - T_1)).$$

as claimed. ■

G NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: A formal version of our main theorem is provided in the introduction, as well as a succinct mention of other results.

   Guidelines:

   - The answer [N/A] means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A [No] or [N/A] answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: The different assumptions are clearly stated and discussed. The conclusion section discusses possible ways to overcome these limitations.

   Guidelines:

   - The answer [N/A] means that the paper has no limitation while the answer [No] means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

Answer: [Yes]

Justification: Assumptions and theorems are numbered and referenced. The complete proofs are provided in the appendix, with a proof sketch in the main text.

Guidelines:

- The answer [N/A] means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: All details are given in Appendix A.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - If the paper includes experiments, a [No] answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
     (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
     (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
     (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

Answer: [Yes]

Justification: Code to reproduce results is provided in the supplementary material.

Guidelines:

- The answer [N/A] means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so [No] is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

   Answer: [Yes]

   Justification: All details are given in Appendix A.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [Yes]

   Justification: Where applicable, plots indicate min/max values in addition to mean values. However, these can be difficult to see because the deviations tend to be very small.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
   - The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
   - The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)

- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
- If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: The experiments do not require any significant compute resources. Each experiments can be run within a few minutes to a couple of hours on normal consumer hardware. This is detailed at the end of Appendix A.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: This is a theoretical work on optimization for Deep Learning, with experiments run on toy data.

   Guidelines:

   - The answer [N/A] means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer [No], they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [N/A]

    Justification: This is a foundational research paper and it is not tied to any particular application.

    Guidelines:

    - The answer [N/A] means that there is no societal impact of the work performed.
    - If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.

- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

    Answer: [N/A]

    Justification: Our experiments only consider toy data and models.

    Guidelines:

    - The answer [N/A] means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [N/A]

    Justification: The paper does not use existing assets.

    Guidelines:

    - The answer [N/A] means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
    - If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.

- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [N/A]

    Justification: The paper does not release new assets.

    Guidelines:

    - The answer [N/A] means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [N/A]

    Justification: the paper does not involve crowdsourcing nor research with human subjects.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [N/A]

    Justification: the paper does not involve crowdsourcing nor research with human subjects.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
    - We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
    - For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

Answer: [N/A]

Justification: the core method development in this research does not involve LLMs as any important, original, or non-standard components.

Guidelines:

- The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
- Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.