EARLY NEURON ALIGNMENT IN TWO-LAYER ReLU NETWORKS WITH SMALL INITIALIZATION

Anonymous authors
Paper under double-blind review

ABSTRACT

This paper studies the problem of training a two-layer ReLU network for binary classification using gradient flow with small initialization. We consider a training dataset with well-separated input vectors: Any pair of input data with the same label are positively correlated, and any pair with different labels are negatively correlated. Our analysis shows that, during the early phase of training, neurons in the first layer try to align with either the positive data or the negative data, depending on its corresponding weight on the second layer. A careful analysis of the neurons’ directional dynamics allows us to provide an $O\left(\log n / \sqrt{\mu}\right)$ upper bound on the time it takes for all neurons to achieve good alignment with the input data, where $n$ is the number of data points and $\mu$ measures how well the data are separated. After the early alignment phase, the loss converges to zero at a $O\left(\frac{1}{t}\right)$ rate, and the weight matrix on the first layer is approximately low-rank. Numerical experiments on the MNIST dataset illustrate our theoretical findings.

1 INTRODUCTION

Neural networks have shown excellent empirical performance in many application domains such as vision (Krizhevsky et al., 2012), speech (Hinton et al., 2012) and video games (Silver et al., 2016). Despite being highly overparametrized, networks trained by gradient descent with random initialization and without explicit regularization enjoy good generalization performance. One possible explanation for this phenomenon is the implicit bias or regularization induced by first-order algorithms under certain initialization assumptions. For instance, first-order methods applied to (deep) matrix factorization models may produce solutions that have low nuclear norm (Gunasekar et al., 2017) and/or low rank (Arora et al., 2019), and similar phenomena have been observed for deep tensor factorization (Razin et al., 2022). Moreover, prior work such as (Saxe et al., 2014; Stöger & Soltanolkotabi, 2021) has found that deep linear networks sequentially learn the dominant singular values of the input-output correlation matrix.

It is widely known that these sparsity-inducing biases can often be achieved by small initialization. This has motivated a series of works that theoretically analyze the training dynamics of first-order methods for neural networks with small initialization. For linear networks, the implicit bias of small initialization has been studied in the context of linear regression (Saxe et al., 2014; Gidel et al., 2019; Min et al., 2021; Varre et al., 2023) and matrix factorization (Gunasekar et al., 2017; Arora et al., 2019; Li et al., 2018; 2021; Stöger & Soltanolkotabi, 2021; Yaras et al., 2023; Soltanolkotabi et al., 2023). Recently, the effect of small initialization has been studied for two-layer ReLU networks (Maennel et al., 2018; Lyu et al., 2021; Phuong & Lampert, 2021; Boursier et al., 2022). For example, Maennel et al. (2018) observes that during the early stage of training, neurons in the first layer converge to one out of finitely many directions determined by the dataset. Based on this observation, Phuong & Lampert (2021) shows that in the case of well-separated data, where any pair of input data with the same label are positively correlated and any pair with different labels are negatively correlated, there are only two directions the neurons tend to converge to: the positive data center and the negative one. Moreover, Phuong & Lampert (2021) shows that if such directional convergence holds, then the loss converges, and the resulting first-layer weight matrix is low-rank. However, directional convergence is assumed in their analysis; there is no explicit characterization of how long it takes to achieve directional convergence and how the time to convergence depends on the initialization scale.
Paper contributions: In this paper, we provide a complete analysis of the dynamics of gradient flow for the problem of training a two-layer ReLU network on well-separated data under the assumption of small initialization. Specifically, we show that if the initialization is sufficiently small, during the early phase of training the neurons in the first layer try to align with either the positive data or the negative data, depending on its corresponding weight on the second layer. Moreover, through a careful analysis of the neuron’s directional dynamics we show that the time it takes for all neurons to achieve good alignment with the input data is upper bounded by $O(\frac{\mu}{\sigma^2})$, where $\mu$ is the number of data points and $\mu$ measures how well the data are separated. We also show that after the early alignment phase the loss converges to zero at a $O(\frac{1}{t})$ rate and that the weight matrix on the first layer is approximately low-rank.

Notation: We denote the Euclidean norm of a vector $x$ by $\|x\|$, the inner product between vectors $x$ and $y$ by $(x, y) = x^\top y$, and the cosine of the angle between them as $\cos(x, y) = \langle \frac{x}{\|x\|}, \frac{y}{\|y\|} \rangle$. For an $n \times m$ matrix $A$, we let $A^\top$ denote its transpose. We also let $\|A\|_2$ and $\|A\|_F$ denote the spectral norm and Frobenius norm of $A$, respectively. For a scalar-valued or matrix-valued function of time, $F(t)$, we let $\dot{F} = \frac{d}{dt}F(t) = \frac{d}{dt}F(t)$ denote its time derivative. Additionally, we define $\mathbb{I}_A$ to be the indicator for a statement $A$: $\mathbb{I}_A = 1$ if $A$ is true and $\mathbb{I}_A = 0$ otherwise. We also let $I$ denote the identity matrix, and $\mathcal{N}(\mu, \sigma^2)$ denote the normal distribution with mean $\mu$ and variance $\sigma^2$.

2 Preliminaries

In this section, we first discuss the problem setting. We then present some key ingredients for analyzing the training dynamics of ReLU networks under small initialization, and discuss some of the weaknesses/issues from prior work.

2.1 Problem setting

We are interested in a binary classification problem with dataset $[x_1, \ldots, x_n] \in \mathbb{R}^{D \times n}$ (input data) and $[y_1, \ldots, y_n]^\top \in \{-1, +1\}^n$ (labels). For the classifier, $f : \mathbb{R}^D \rightarrow \mathbb{R}$, we consider a two-layer ReLU network:

$$f(x; W, v) = v^\top \sigma(W^\top x) = \sum_{j=1}^{h} v_j \sigma(w_j^\top x),$$

parametrized by network weights $W := [w_1, \ldots, w_h] \in \mathbb{R}^{D \times h}, v := [v_1, \ldots, v_h]^\top \in \mathbb{R}^{h \times 1}$, where $\sigma(\cdot) = \max\{\cdot, 0\}$ is the ReLU activation function. We aim to find the network weights that minimize the training loss $\mathcal{L}(W, v) = \sum_{i=1}^{n} \ell(y_i, f(x_i; W, v))$, where $\ell : \mathbb{R} \times \mathbb{R} \rightarrow \mathbb{R}_{\geq 0}$ is either the exponential loss $\ell(y, \hat{y}) = \exp(-y\hat{y})$ or the logistic loss $\ell(y, \hat{y}) = \log(1 + \exp(-y\hat{y}))$. The network is trained via the gradient flow (GF) dynamics

$$\dot{W} \in \partial_W \mathcal{L}(W, v), \quad \dot{v} \in \partial_v \mathcal{L}(W, v),$$

where $\partial_W \mathcal{L}$, $\partial_v \mathcal{L}$ are Clarke sub-differentials of $\mathcal{L}$. Therefore, (2) is a differential inclusion (Bolte et al., 2010). For simplicity of the presentation, instead of directly working on this differential inclusion, our theoretical results will be stated for the Caratheodory solution (Reed, 1971) of (2) when the ReLU subgradient is fixed as $\sigma'(x) = \mathbb{I}_{x \geq 0}$. In Appendix E we show that under the data assumption of our interest (to be introduced later), Caratheodory solution(s) $\{W(t), v(t)\}$ exist globally for all $t \in [0, \infty)$, which we call solution(s) of (2) throughout this paper.

To initialize the weights, we consider the following initialization scheme. First, we start from a weight matrix $W_0 \in \mathbb{R}^{D \times h}$, and then and then initialize the weights as

$$W(0) = \epsilon W_0, \quad v_j(0) \in \{||w_j(0)||, -||w_j(0)||\}, \forall j \in [h].$$

That is, the weight matrix $W_0$ determines the initial shape of the first-layer weights $W(0)$ and we use $\epsilon$ to control the initialization scale and we are interested in the regime where $\epsilon$ is sufficiently small. For the second layer weights $v(0)$, each $v_j(0)$ has magnitude $||w_j(0)||$ and we only need to decide its sign. Our results in later sections are stated for a deterministic choice of $\epsilon$, $W_0$, and $v(0)$, then we comment on the case where $W_0$ is chosen randomly via some distribution.

The resulting weights in (3) are always "balanced", i.e., $v_j^2(0) - ||w_j(0)||^2 = 0, \forall j \in [h]$, because $v_j(0)$ can only take two values: either $||w_j(0)||$ or $-||w_j(0)||$. More importantly, under GF (2), this
balancedness is preserved (Du et al., 2018): \( v_j^2(t) - \|w_j(t)\|^2 = 0, \forall t \geq 0, \forall j \in [h]. \) In addition, it is shown in Boursier et al. (2022) that \( \text{sign}(v_j(t)) = \text{sign}(v_j(0)), \forall t \geq 0, \forall j \in [h], \) and the dynamical behaviors of neurons will be divided into two types, depending on \( \text{sign}(v_j(0)) \).

**Remark 1.** For our theoretical results, the balancedness condition is assumed for technical purposes: it simplifies the dynamics of GF and thus the analysis. It is a common assumption for many existing works on both linear (Arora et al., 2018b) and nonlinear (Phuong & Lampert, 2021; Boursier et al., 2022) neural networks. For the experiments in Section 4, we use a standard Gaussian initialization with small variance, which is not balanced.

**Remark 2.** Without loss of generality, we consider the case where all columns of \( W_0 \) are nonzero, i.e., \( \|w_j(0)\| > 0, \forall j \in [h] \). We make this assumption because whenever \( w_j(0) = 0 \), we also have \( v_j(0) = 0 \) from the balancedness, which together would imply \( v_j \equiv 0, w_j \equiv 0 \) under gradient flow. As a result, \( w_j \) and \( v_j \) would remain zero and thus they could be ignored in the convergence analysis.

**Remark 3.** Our main results will depend on both \( \max_j \|w_j(0)\| \) and \( \min_j \|w_j(0)\| \), as shown in our proofs in Appendices C and D. Therefore, whenever we speak of small initialization, we will say that \( \epsilon \) is small without worrying about the scale of \( W_0 \), which is already considered in our results.

## 2.2 Neural Alignment with Small Initialization: An Overview

Prior work argues that the gradient flow dynamics (2) under small initialization (3), i.e., when \( \epsilon \) is sufficiently small, can be roughly described as “align then fit” (Maennel et al., 2018; Boursier et al., 2022) : During the early phase of training, every neuron \( w_j, j \in [h] \) keeps a small norm \( \|w_j\|^2 \ll 1 \) while changing their directions \( \frac{w_j}{\|w_j\|} \) significantly in order to locally maximize a “signed coverage” (Maennel et al., 2018) of itself w.r.t. the training data. After the alignment phase, part of (potentially all) the neurons grow their norms in order to fit the training data, and the loss decreases significantly. The analysis for the fitting phase generally depends on the resulting neuron directions at the end of the alignment phase (Phuong & Lampert, 2021; Boursier et al., 2022). However, prior analysis of the alignment phase either is based on a vanishing initialization argument that can not be directly translated into the case finite but small initialization (Maennel et al., 2018) or assumes some stringent assumption on the data (Boursier et al., 2022). In this section, we provide a brief overview of the existing analysis for neural alignment and then point out several weaknesses in prior work.

**Prior analysis of the alignment phase:** Since during the alignment phase all neurons have small norm, prior work mainly focuses on the directional dynamics, i.e., \( \frac{d}{dt} \frac{w_j}{\|w_j\|} \), of the neurons. The analysis relies on the following approximation of the dynamics of every neuron \( w_j, j \in [h] \):

\[
\frac{d}{dt} \frac{w_j}{\|w_j\|} \approx \text{sign}(v_j(0)) P_{w_j} x_a(w_j), \tag{4}
\]

where \( P_w = I - \frac{ww^T}{\|w\|^2} \) is the projection onto the subspace orthogonal to \( w \) and

\[
x_a(w) := \sum_{i: \langle x_i, w \rangle > 0} y_i x_i \tag{5}
\]

denotes the signed combination of the data points activated by \( w \). First of all, (4) implies that the dynamics \( \frac{w_j}{\|w_j\|} \) are approximately decoupled, and thus one can study each \( \frac{w_j}{\|w_j\|} \) separately.

Moreover, as illustrated in Figure 1, if \( \text{sign}(v_j(0)) > 0 \), the flow (4) pushes \( w_j \) towards \( x_a(w_j) \), since \( w_j \) is attracted by its currently activated positive data and repelled by its currently activated negative data. Intuitively, during the alignment phase, a neuron \( w_j \) with \( \text{sign}(v_j(0)) > 0 \) would try to find a direction where it can activate as much positive data and as less negative data as possible. If \( \text{sign}(v_j(0)) < 0 \), the opposite holds.

Indeed, Maennel et al. (2018) claims that the neuron \( w_j \) would be aligned with some “extreme vectors,” defined as vector \( w \in S^{L-1} \) that locally maximizes \( \sum_{i \in [n]} y_i \sigma(\langle x_i, w \rangle) \) (similarly, \( w_j \) with \( \text{sign}(v_j(0)) < 0 \) would be aligned with the local minimizer), and there are only finitely many
such vectors; thus the neurons are expected to converge to one of these extreme vectors in direction. The analysis is done by taking the limit $\epsilon \to 0$, under which the approximation in (4) is exact.

Weakness in prior analyses: Although [Maennel et al., 2018] provides great insights into the dynamical behavior of the neurons in the alignment phase, the validity of the aforementioned approximation for finite but small $\epsilon$ remains in question. First, one needs to make sure that the error $\left\| \frac{d}{d\epsilon} w_{ij} x_i \left(\epsilon(0)\right) P_{w_j} x_a \left(\epsilon\right) \right\|$ is sufficiently small when $\epsilon$ is finite in order to justify (4) as a good approximation. Second, the error bound needs to hold for the entire alignment phase. [Maennel et al., 2018] assumes $\epsilon \to 0$; hence there is no formal error bound. In addition, prior analyses on small initialization [Stöger & Soltanolkotabi, 2021; Boursier et al., 2022] suggest the alignment phase only holds for $\Theta(\log \frac{1}{\epsilon})$ time. Thus, the claim in [Maennel et al., 2018] would only hold if good alignment is achieved before the alignment phase ends. However, [Maennel et al., 2018] provides no upper bound on the time it takes to achieve good alignment. Therefore, without a finite $\epsilon$ analysis, [Maennel et al., 2018] fails to fully explain the training dynamics under small initialization. Understanding the alignment phase with finite $\epsilon$ requires additional quantitative analysis. To the best of our knowledge, this has only been studied under a stringent assumption that all data points are orthogonal to each other (Boursier et al., 2022).

Goal of this paper: In this paper, we want to address some of the aforementioned issues by developing a formal analysis for the early alignment phase with a finite but small initialization scale $\epsilon$. We first discuss our main theorem that shows that a directional convergence can be achieved within bounded time under data assumptions that are less restrictive and have more practical relevance. Then, we discuss the error bound for justifying (4) in the proof sketch of the main theorem.

3 Convergence of Two-layer ReLU Networks with Small Initialization

In this section, we present our main results, which require the following assumption on the training data (we will compare our assumption with those in prior work after the main theorem):

**Assumption 1.** Any pair of input data with the same label are positively correlated, and any pair of inputs with different labels are negatively correlated, i.e., $\min_{i,j} \langle x_i, y_i, x_j, y_j \rangle =: \mu > 0$.

Given a training dataset, we define $S_+ := \{ z \in \mathbb{R}^D : 1_{(x,z)>0} = 1_{y_i, z_i} \forall i \}$ to be the cone in $\mathbb{R}^n$ such that whenever neuron $w \in S_+$, $w$ is activated exclusively by every $x_i$ with a positive label (see Figure 2). Similarly, for $x_i$ with negative labels, we define $S_- := \{ z \in \mathbb{R}^D : 1_{(x,z)>0} = 1_{y_i, z_i} \forall i \}$. Finally, we define $S_{\text{dead}} := \{ z \in \mathbb{R}^D : \langle z, x_i \rangle \leq 0, \forall i \}$ to be the cone such that whenever $w \in S_{\text{dead}}$, no data activates $w$. Given Assumption 1, it can be shown (see Appendix A) that $S_+$ ($S_-$) is a non-empty, convex cone, containing all positive data $x_i, i \in I_+$ (negative data $x_i, i \in I_-$). $S_{\text{dead}}$ is a convex cone as well, but not necessarily non-empty. We illustrate these cones in Figure 2 given some training data (red solid arrow denotes positive data and blue denotes negative ones).

Moreover, given some initialization from $\mathcal{I}$, we define $I_+ := \{ i \in [n] : y_i \geq 0 \}$ to be the set of indices of positive data, and $I_- := \{ i \in [n] : y_i < 0 \}$ for negative data. We also define $V_+ := \{ j \in [h] : \text{sign}(v_j(t)) > 0 \}$ to be the set of indices of neurons with positive second-layer entry and $V_- := \{ j \in [h] : \text{sign}(v_j(t)) < 0 \}$ for neurons with negative second-layer entry. Note that, as discussed in Section 2.2, $\text{sign}(v_j(t))$ does not change under balanced initialization, thus $V_+, V_-\text{ are time invariant.}$ Further, as we discussed in Section 2.2 about the early alignment phase, we expect that every neuron in $V_+$ will drift toward the region where positive data concentrate and thus eventually reach $S_+$ or $S_{\text{dead}}$, as visualized in Figure 2($x_+, x_-$ shown in the figure are defined in Assumption 2). Similarly, all neurons in $V_-$ would chase after negative data and thus reach $S_-$ or $S_{\text{dead}}$. Our theorem precisely characterizes this behavior.
3.1 Main results

In this work, we study the properties of solutions of the GF dynamics (2) that satisfy a certain notion of regularity, as defined below.

**Definition 1.** A solution \( \{W(t), v(t)\} \) to (2) is regular if it satisfies that \( w_j(t_0) \in S_{\text{dead}} \) for some \( j \in [h] \) and some \( t_0 \geq 0 \) implies \( w_j(t) \in S_{\text{dead}} \) for all \( t \geq t_0 \).

In rare cases, solutions to (2) could be non-unique and there are potentially “irregular solutions” (please refer to Appendix E.4 for details) that allows some neuron to regain activation even after becoming completely deactivated in \( S_{\text{dead}} \). We deem such irregular solutions of little practical relevance since when implementing gradient descent algorithm in practice, neurons in \( S_{\text{dead}} \) would receive zero update and thus stay in \( S_{\text{dead}} \). Therefore, our main theorem concerns the regular solutions to (2) (the existence of such solutions is shown in Appendix E.2).

Before we present our main theorem, we need the following assumption on the initialization, for technical reasons, essentially asking the neuron \( w_j(0), j \in V_+ \) (or \( w_j(0), j \in V_- \), resp.) to not be completely aligned with \( x_+ \) (or \( x_- \), resp.).

**Assumption 2.** The initialization from (3) satisfies that \( \max_{j \in V_+} \left( \frac{w_j(0)}{\|w_j(0)\|}, \frac{x_+}{\|x_+\|} \right) < 1 \), and \( \max_{j \in V_-} \left( \frac{w_j(0)}{\|w_j(0)\|}, \frac{x_-}{\|x_-\|} \right) < 1 \), where \( x_+ = \sum_{i \in I_+} x_i \) and \( x_- = \sum_{i \in I_-} x_i \).

We are now ready to present our main result (given Assumption 1 and Assumption 2):

**Theorem 1.** Given some initialization from (3), if \( \epsilon = O\left(\frac{1}{\sqrt{n}} \exp\left(-\frac{n}{\sqrt{p}} \log n\right)\right) \), then for any regular solution to the gradient flow dynamics (2), we have

1. (Directional convergence in early alignment phase) \( \exists t_1 = O\left(\frac{\log n}{\sqrt{p}}\right) \), such that

   a. \( \forall j \in V_+ \), either \( w_j(t_1) \in S_+ \) or \( w_j(t_1) \in S_{\text{dead}} \). Moreover, if \( \max_{i \in I_+} \langle w_j(0), x_i \rangle > 0 \), then \( w_j(t_1) \in S_+ \).

   b. \( \forall j \in V_- \), either \( w_j(t_1) \in S_- \) or \( w_j(t_1) \in S_{\text{dead}} \). Moreover, if \( \max_{i \in I_-} \langle w_j(0), x_i \rangle > 0 \), then \( w_j(t_1) \in S_- \).

2. (Final convergence and low-rank bias) \( \forall t \geq t_1 \) and \( \forall j \in [h] \), neuron \( w_j(t) \) stays within \( S_+ \) (\( S_- \), or \( S_{\text{dead}} \)) if \( w_j(t_1) \in S_+ \) (\( S_- \), or \( S_{\text{dead}} \) resp.). Moreover, if both \( S_+ \) and \( S_- \) contains at least one neuron at time \( t_1 \), then

   a. \( \exists \alpha > 0 \) and \( \exists t_2 \) with \( t_1 \leq t_2 = \Theta\left(\frac{1}{\sqrt{n}} \log \frac{1}{\sqrt{p}}\right) \), such that \( \mathcal{L}(t) \leq \frac{L(t_2)}{\mathcal{L}(t_2)\alpha(t_2-t_2)^2} + 1 \), \( \forall t \geq t_2 \).

   b. As \( t \to \infty \), \( \mathcal{W}(t) \to \infty \) and \( \|W(t)\|_{2,p} \leq 2\|W(t)\|_{2,p} + O(\epsilon) \). Thus, the stable rank of \( W(t) \) satisfies \( \limsup_{t \to \infty} \|W(t)\|_{2,p} / \|W(t)\|_{2,p} \leq 2 \).

We provide a proof sketch that highlights the technical novelty of our results in Section 3.3. Our \( O(\cdot) \) notations hide additional constants that depend on the data and initialization, for which we refer readers to the complete proof of Theorem 1 in Appendix C and D. We make the following remarks:

**Early neuron alignment:** The first part of the Theorem 1 describes the configuration of all neurons at the end of the alignment phase. Every neuron in \( V_+ \) reaches either \( S_+ \) or \( S_{\text{dead}} \) by time \( t_1 \), and stays there for the remainder of training. Obviously, we care about those neurons reaching \( S_+ \) as any neuron in \( S_{\text{dead}} \) does not contribute to the final convergence at all. Luckily, Theorem 1 suggests that any neuron in \( V_+ \) that starts with some activation on the positive data, i.e., it is initialized in the union of halfspaces \( \bigcup_{i \in I_+} \{ w : \langle w, x_i \rangle > 0 \} \), will eventually reach \( S_+ \). A similar discussion holds for neurons in \( V_- \). We argue that randomly initializing \( W_0 \) ensures that with high probability, there will be at least a pair of neurons reaching \( S_+ \) and \( S_- \) by time \( t_1 \) (please see the next remark). Lastly, we note that it is possible that \( S_{\text{dead}} = \emptyset \), in which case every neuron reaches either \( S_+ \) or \( S_- \).

**Merits of random initialization:** Our theorem is stated for a deterministic initialization (3) given an initial shape \( W_0 \). In practice, one would use random initialization to find a \( W_0 \), for example, \( [W_0]_{ij} \overset{i.i.d.}{\sim} \mathcal{N}(0, 1/D) \). First, our Theorem 1 applies to this Gaussian initialization: Assumption 2 is satisfied with probability one because the events \( \left\{ \frac{w_j(0)}{\|w_j(0)\|}, \frac{x_+}{\|x_+\|} = 1 \right\} \) and \( \left\{ \frac{w_j(0)}{\|w_j(0)\|}, \frac{x_-}{\|x_-\|} = 1 \right\} \)
have probability zero. Moreover, any neuron in $V_+$ has at least probability $1/2$ of being initialized within the union of halfspaces $\cup_{i \in I} \{ w : \langle w, x_i \rangle > 0 \}$, which ensures that this neuron reaches $S_+$. Thus when there are $m$ neurons in $V_+$, the probability that $S_+$ has at least one neuron at time $t_1$ is lower bounded by $1 - 2^{-m}$ (same argument holds for $S_-$). Therefore, with only very mild overparametrization on the network width $h$, one can make sure that with high probability there is at least one neuron in both $S_+$ and $S_-$, leading to final convergence.

**Importance of a quantitative bound on $t_1$:** The analysis for neural alignment relies on the approximation in (4), which, through our analysis (see Lemma 1), is shown to only hold before $T = \Theta(\frac{1}{n} \log \frac{1}{\sqrt{\mu}})$, thus if one proves, through the approximation in (4), that good alignment is achieved within $t_1$ time, then the initialization scale $\epsilon$ must be chosen to be $\mathcal{O}\left(\frac{1}{\sqrt{\mu}} \exp(-nt_1)\right)$ so that $t_1 \leq T$, i.e. the proved alignment should finish before the approximation (4) fails. Therefore, without an explicit bound on $t_1$, one does not know a priori how small $\epsilon$ should be. Our analysis shows that under (4), directional convergence, i.e., all neurons reaching either $S_+, S_-, S_{\text{dead}}$, is achieved within $t_1 = \mathcal{O}(\frac{\log n}{\sqrt{\mu}})$ time. This bound, in return, determines the bound for initialization scale $\epsilon$.

Moreover, our bound quantitatively reveals the non-trivial dependency on the "data separation" $\mu$ for such directional convergence to occur. Indeed, through a numerical illustration in Appendix A.2 we show that the dependence on the data separability $\mu > 0$ is crucial in determining the scale of the initialization: As $\mu$ approaches zero, the time needed for the desired alignment increases, necessitate the use of a smaller $\epsilon$.

**Refined alignment within $S_+, S_-:$** Once a neuron in $V_+$ reaches $S_+$, it never leaves $S_+$. Moreover, it always gets attracted by $x_+$. Therefore, every neuron gets well aligned with $x_+$, i.e., $\cos(\langle w, x_+ \rangle) \simeq 1, \forall w \in S_+$. A similar argument shows neurons in $S_-$ get attracted by $x_-$. We opt not to formally state it in Theorem 1 as the result would be similar to that in (Boursier et al., 2022), and alignment with $x_+, x_-$ is not necessary to guarantee convergence. Instead, we show this refined alignment through our numerical experiment in Section 4.

**Final convergence and low-rank bias:** We present the final convergence results mostly for the completeness of the analysis. GF after $t_1$ can be viewed as fitting positive data $x_i, i \in I_+$, with a subnetwork consisting of neurons in $S_+$, and fitting negative data with neurons in $S_-$. By the fact that all neurons in $S_+$ activate all $x_i, i \in I_+$, the resulting subnetwork is linear, and so is the subnetwork for fitting $x_i, i \in I_-$. The convergence analysis reduces to establishing $\mathcal{O}(1/t)$ convergence for two linear networks (Arora et al., 2018a; Min et al., 2021; Yun et al., 2020). The non-trivial and novel part is to show that right after the alignment phase ends, one can expect a substantial decrease in the loss (starting from time $t_2 = \Theta(\frac{1}{n} \log \frac{1}{\sqrt{\mu}})$). An alternative way of proving convergence is by observing that at $t_1$, all data has been correctly classified (w.r.t. sign of $f$), which is sufficient for showing $\mathcal{O}(\frac{1}{\log T})$ convergence (Lyu & Li, 2019; Ji & Telgarsky, 2020) of the loss, but this asymptotic rate does not suggest a time after which the loss start to decrease significantly. As for the stable rank, our result follows the analysis in Le & Jegelka (2022), but in a simpler form since ours is for linear networks. Although convergence is established partially by existing results, we note that these analyses are all possible because we have quantitatively bound $t_1$ in the alignment phase.

3.2 **Comparison with prior work**

Our results provide a complete (from alignment to convergence), non-asymptotic (finite $\epsilon$), quantitative (bounds on $t_1$, $t_2$) analysis for the GF in (2) under small initialization. Similar neural alignment has been studied in prior work for **orthogonally separable** data (same as ours) and for **orthogonal** data, and we shall discuss them separately.

**Alignment under orthogonally separable data:** Phuong & Lampert (2021) assumes that there exists a time $t_1$ such that at $t_1$, the neurons are in either $S_+, S_-, S_{\text{dead}}$ and their main contribution is the analysis of the implicit bias for the later stage of the training. They justify their assumption by the analysis in Maennel et al. (2018), which does not necessarily apply to the case of finite $\epsilon$, as we discussed in Section 2.2. Later Wang & Pilanci (2022) shows $t_1$ exists, provided that the initialization scale $\epsilon$ is sufficiently small, but still with no explicit analysis showing how $t_1$ depends on the data separability $\mu$ and the size of the training data $n$. Moreover, there is no quantification on how small $\epsilon$ should be. In our work, all the results are non-asymptotic and quantitative: we show that good alignment is achieved within $t_1 = \mathcal{O}(\frac{\log n}{\sqrt{\mu}})$ time and provide an explicit upper bound on
Moreover, our results highlight the dependence on the separability $\mu > 0$, (Further illustrated in Appendix A.2) which is not studied in Phuong & Lampert (2021), Wang & Pilanci (2022).

**Alignment under orthogonal data:** In Boursier et al. (2022), the neuron alignment is carefully analyzed for the case all data points are orthogonal to each other, i.e., $\langle x_i, x_j \rangle = 0, \forall i \neq j \in [n]$. We point out that neuron behavior is different under orthogonal data (illustrated in Appendix A.3): only the positive (negative) neurons initially activate all the positive (negative) data will end up in $S_+$ ($S_-$). In our case, all positive (negative) neurons will arrive at $S_+$ ($S_-$), unless they become a dead neuron. Moreover, due to such distinction, the analysis is different: Boursier et al. (2022) restrict their results to positive (negative) neurons $w_j$ that initially activate all the positive (negative) data, and there is no need for analyzing neuron activation. However, since our analysis is on all positive neurons, regardless of their initial activation pattern, it utilizes novel techniques to track the evolution of the activation pattern (see Section 3.3).

### 3.3 Proof sketch for the alignment phase

In this section, we sketch the proof for our Theorem 1. First of all, it can be shown that $S_+, S_{\text{dead}}$ are trapping regions for all $w_j(t), j \in V_+$, that is, whenever $w_j(t)$ gets inside $S_+$ (or $S_{\text{dead}}$), it never leaves $S_+$ (or $S_{\text{dead}}$). Similarly, $S_-, S_{\text{dead}}$ are trapping regions for all $w_j(t), j \in V_-$. The alignment phase analysis concerns how long it takes for all neurons to reach one of the trapping regions, followed by the final convergence analysis on fitting data with +1 label by neurons in $S_+$ and fitting data with −1 label by those in $S_-$. We have discussed the final convergence analysis in the remark "Final convergence and low-rank bias", thus we focus on the proof sketch for the early alignment phase here, which is considered as our main technical contribution.

**Approximating $\frac{d}{dt} \|w_j\|^2$:** Our analysis for the neural alignment is rooted in the following Lemma:

**Lemma 1.** Given some initialization from (3), if $\epsilon = \mathcal{O}(\frac{1}{\sqrt{n}})$, then there exists $T = \Theta(\frac{1}{\epsilon^2})$ such that any solution to the gradient flow dynamics (2) satisfies that $\forall t \leq T$,

$$\max_j \left| \frac{d}{dt} \|w_j(t)\|^2 - \text{sign}(v_j(0))\mathcal{P}_{w_j(t)} x_{\alpha}(w_j(t)) \right| = \mathcal{O}\left(\epsilon n \sqrt{\mu} \right).$$

(6)

This Lemma shows that the error between $\frac{d}{dt} \|w_j(t)\|^2$ and $\text{sign}(v_j(0))\mathcal{P}_{w_j(t)} x_{\alpha}(w_j(t))$ can be arbitrarily small with some appropriate choice of $\epsilon$ (to be determined later). This allows one to analyze the true directional dynamics $\frac{d}{dt} \|w_j(t)\|^2$ using some property of $\mathcal{P}_{w_j(t)} x_{\alpha}(w_j(t))$, which leads to a $t_1 = \mathcal{O}(\frac{n}{\mu})$ upper bound on the time it takes for the neuron direction to converge to the sets $S_+$, $S_-$, or $S_{\text{dead}}$. Moreover, it also suggests $\epsilon$ can be made sufficiently small so that the error bound holds until the directional convergence is achieved, i.e. $T \geq t_1$. We will first illustrate the analysis for directional convergence, then close the proof sketch with the choice of a sufficiently small $\epsilon$.

**Activation pattern evolution:** Given a sufficiently small $\epsilon$, one can show that under Assumption [1] for every neuron $w_j$ that is not in $S_{\text{dead}}$ we have:

$$\frac{d}{dt} \left( \frac{w_j}{\|w_j\|} \cdot \frac{x_i y_i}{\|x_i\|} \right)_{(w_j, x_i) = 0} > 0, \forall i \in [n], j \in V_+.$$  

(7)

$$\frac{d}{dt} \left( \frac{w_j}{\|w_j\|} \cdot \frac{x_i y_i}{\|x_i\|} \right)_{(w_j, x_i) = 0} < 0, \forall i \in [n], j \in V_-.$$  

(8)

This is because if a neuron satisfies $(x_i, w_j) = 0$ for some $i$, and is not in $S_{\text{dead}}$, GF moves $w_j$ towards $x_{\alpha}(w_j) = \sum_i: (x_i, w_j) > 0 x_i y_i$. Interestingly, Assumption [1] implies $(x_i y_i, x_{\alpha}(w_j)) > 0, \forall i \in [n]$, which makes $\frac{d}{dt} \|w_j\|^2 \approx \text{sign}(v_j(0))\mathcal{P}_{w_j} x_{\alpha}(w_j)$ point inward (or outward) the halfspace $(x_i y_i, w_j) > 0$, if sign$(v_j(0)) > 0$ (or sign$(v_j(0)) < 0$, respectively). See Figure 3 for illustration.

As a consequence, a neuron can only change its activation pattern in a particular manner: a neuron in $V_+$, whenever it is activated by some $x_i$ with $y_i = +1$, never loses the activation on $x_i$ thereafter, because (7) implies that GF pushes $\frac{w_j}{\|w_j\|}$ towards $x_i$ at the boundary $\langle w_j, x_i \rangle = 0$. Moreover, (7) also shows that a neuron in $V_+$ will never regain activation on a $x_i$ with $y_i = -1$ once it loses the
We use a toy example in Appendix A.1 to clearly visualize the neuron alignment during training (due to space constraints). In the main body of this paper, we validate our theorem using a binary classification task for two MNIST digits. Such training data do not satisfy Assumption 1 since every data vector is a grayscale image with non-negative entries, making the inner product between any pair of data non-negative, regardless of their labels. However, we can preprocess the training data by centering: \( x = x - \bar{x} \), where \( \bar{x} = \frac{1}{n} \sum_{i=1}^{n} x_i \). The preprocessed data, then, approximately satisfies our assumption (see the left-most plot in Figure 5): a pair of data points is very likely to have a positive correlation if they have the same label and to have a negative correlation if they have a small negative correlation.

**Numerical Experiments**

We use a toy example in Appendix A.1 to clearly visualize the neuron alignment during training (due to space constraints). In the main body of this paper, we validate our theorem using a binary classification task for two MNIST digits. Such training data do not satisfy Assumption 1 since every data vector is a grayscale image with non-negative entries, making the inner product between any pair of data non-negative, regardless of their labels. However, we can preprocess the training data by centering: \( x = x - \bar{x} \), where \( \bar{x} = \frac{1}{n} \sum_{i=1}^{n} x_i \). The preprocessed data, then, approximately satisfies our assumption (see the left-most plot in Figure 5): a pair of data points is very likely to have a positive correlation if they have the same label and to have a negative correlation if they have a small negative correlation.
We build a two-layer ReLU network with $\alpha$ careful analysis of the neurons’ directional dynamics allows us to provide an upper bound on the $x$ preprocessed data. For the remaining section, we use $x_i, i \in [n]$, to denote the preprocessed (centered) data and use $\bar{x}$ to denote the mean of the original data.

We build a two-layer ReLU network with $h = 50$ neurons and initialize all entries of the weights as $[W]_{ij} \sim \mathcal{N}(0, \alpha^2)$, $v_j \sim \mathcal{N}(0, \alpha^2)$, $\forall i \in [n], j \in [h]$ with $\alpha = 10^{-6}$. Then we run gradient descent on both $W$ and $v$ with step size $\eta = 2 \times 10^{-5}$. Notice that here the weights are not initialized to be balanced as in [3]. The numerical results are shown in Figure 5.

**Alignment phase:** Without balancedness, one no longer has $\text{sign}(v_j(t)) = \text{sign}(v_j(0))$. With a little abuse of notation, we denote $V_+(t) = \{ j \in [h] : \text{sign}(v_j(t)) > 0 \}$ and $V_-(t) = \{ j \in [h] : \text{sign}(v_j(t)) < 0 \}$, and we expect that at the end of the alignment phase, neurons in $V_+$ are aligned with $x_+ = \sum_{i \in I_+} x_i$, and neurons in $V_-$ with $x_- = \sum_{i \in I_-} x_i$. The second plot in Figure 5 shows such an alignment between neurons and $x_+, x_-$. In the top part, the red solid line shows $\cos(\bar{w}_+, x_+)$ during training, where $\bar{w}_+ = \sum_{j \in V_+} w_j / |V_+|$, and the shaded region defines the range between $\min_{j \in V_+} \cos(w_j, x_+)$ and $\max_{j \in V_+} \cos(w_j, x_+)$. Similarly, in the bottom part, the green solid line shows $\cos(\bar{w}_-, x_-)$ during training, where $\bar{w}_- = \sum_{j \in V_-} w_j / |V_-|$, and the shaded region delineates the range between $\min_{j \in V_-} \cos(w_j, x_-)$ and $\max_{j \in V_-} \cos(w_j, x_-)$. Initially, every neuron is approximately orthogonal to $x_+, x_-$ due to random initialization. Then all neurons in $V_+$ ($V_-$) start to move towards $x_+$ ($x_-$) and achieve good alignment after $\sim 2000$ iterations. When the loss starts to decrease, the alignment drops. We conjecture that because Assumption [1] is not exactly satisfied, neurons in $V_+$ have to fit some negative data, for which $x_+$ is not the best direction.

**Final convergence:** After $\sim 3000$ iterations, the norm $\|W\|_2^2$ starts to grow and the loss decreases, as shown in the third plot in Figure 5. Moreover, the stable rank $\|W\|_F^2 / \|W\|_2^2$ decreases below 2. For this experiment, we almost have $\cos(x_+, x_-) \approx -1$, thus the neurons in $V_+$ (aligned with $x_+$) and those in $V_-$ (aligned with $x_-)$ are almost co-linear. Therefore, the stable rank $\|W\|_F^2 / \|W\|_2^2$ is almost 1, as seen from the plot. Finally, at iteration 15000, we visualize the mean neuron $\bar{w}_+ = \sum_{j \in V_+} w_j / |V_+|$, $\bar{w}_- = \sum_{j \in V_-} w_j / |V_-|$ as grayscale images, and compare them with $\bar{x}_+ = x_+/|I_+|$, $\bar{x}_- = x_-/|I_-|$, showing good alignment.

**Comparison with other training schemes:** For two-layer ReLU networks, there is another line of work [Brutzkus et al., 2018; Wang et al., 2019] that studies GD/SGD only on the first-layer weights $W$ and keeping the second-layer $v$ fixed throughout training. In Appendix A.5 we compare our training schemes to those in [Brutzkus et al., 2018; Wang et al., 2019], and show that while both schemes achieve small training loss, the aforementioned two-phase training (alignment then final convergence) does no happen if only the first-layer in trained.

**5 Conclusion**

This paper studies the problem of training a binary classifier via gradient flow on two-layer ReLU networks under small initialization. We consider a training dataset with well-separated input vectors. A careful analysis of the neurons’ directional dynamics allows us to provide an upper bound on the time it takes for all neurons to achieve good alignment with the input data. Numerical experiment on classifying two digits from the MNIST dataset correlates with our theoretical findings.
REFERENCES


Mary Phuong and Christoph H Lampert. The inductive bias of relu networks on orthogonally separable data. In International Conference on Learning Representations, 2021.


A ADDITIONAL EXPERIMENTS

A.1 ILLUSTRATIVE EXAMPLE

We illustrate our theorem using a toy example: we train a two-layer ReLU network with \( h = 50 \) neurons under a toy dataset in \( \mathbb{R}^2 \) (See Figure 6) that satisfies our Assumption 1 and initialize all entries of the weights as \( [W]_{ij} \overset{i.i.d.}{\sim} \mathcal{N}(0, \alpha^2), v_j \overset{i.i.d.}{\sim} \mathcal{N}(0, \alpha^2), \forall i \in [n], j \in [h] \) with \( \alpha = 10^{-6} \). Then we run gradient descent on both \( W \) and \( v \) with step size \( \eta = 2 \times 10^{-3} \). Our theorem well predicts the dynamics of neurons at the early stage of the training: aside from neurons that ended up in \( S_{\text{dead}} \), neurons in \( V_+ \) reach \( S_+ \) and achieve good alignment with \( x_+ \), and neurons in \( V_- \) are well aligned with \( x_- \) in \( S_- \). Note that after alignment, the loss experiences two sharp decreases before it gets close to zero, which is studied and explained in Boursier et al. (2022).

![Figure 6: Illustration of gradient descent on two-layer ReLU network with small initialization. The marker represents either a data point or a neuron. Solid lines represent the directions of neurons. (a) at initialization, all neurons have small norm and are pointing in different directions; (b) around the end of the alignment phase, all neurons are in \( S_+, S_-, \) or \( S_{\text{dead}} \). Moreover, neurons in \( S_+ (S_-) \) are well aligned with \( x_+ (x_-); \) (c) With good alignment, neurons in \( S_-, S_+ \) start to grow in norm and the loss decreases. When the loss is close to zero, the resulting network has its first-layer weight approximately low-rank.](image)

A.2 EFFECT OF DATA SEPARABILITY \( \mu \)

This section investigates the effect of data separability \( \mu \) on the time required to achieve the desired alignment as in Theorem 1 through a simple example. We consider a similar setting as in A.1, and explore the cases when data separability \( \mu \ll 1 \). We expect that as separability \( \mu \) decreases, the time for neurons to achieve the desired alignment as in Theorem 1 increases, necessitating a smaller initialization scale. For simplicity, we consider a dataset with only two positive data \((x_1, y_1 = +1), (x_2, y_2 = +1)\).

In Figure 7, we first set \( \mu = \langle x_1, x_2 \rangle = \sin(0.1) \), and the neuron alignment is consistent with Theorem 1: positive neurons (that are not dead) eventually enter \( S_+ \), activating both data points, and then final convergence follows.

![Figure 7: Neural alignment under two data points with small positive correlation: \( x_1 = [1, 0], y_1 = +1, x_2 = [\sin(0.1), \cos(0.1)], y_2 = +1 \). The experimental setting is exactly the same as the illustrative example in Appendix A.1 (initialization scale \( \alpha = 10^{-6} \)). The marker represents either a data point or a neuron. Solid lines represent the directions of neurons. In the alignment phase, positive neurons are aligned with \( x_+ \), and then grow their norm for final convergence.](image)
However, in Figure 8, as we decrease the separability $\mu$ to $\sin(0.001)$ (other settings remain unchanged), the neural alignment becomes slower: 1) at iteration 7000, there are still neurons (that are not dead) outside $S_+$, namely those aligned with either $x_1$ or $x_2$, while in our previous setting ($\mu = \sin(0.1)$), all neurons (that are not dead) have reached $S_+$; 2) In this particular instance of the experiment, we also see one neuron remains outside $S_+$ at the late stage of the training (at iteration 21000). This clearly shows that as data separability $\mu$ decreases, the time needed for all neurons (that are not dead) to reach $S_+$ increases, and if the initialization scale is not small enough for the alignment phase to hold for a long time, there will be neurons remains outside $S_+$.

Figure 8: Neural alignment under two data points with tiny positive correlation: $x_1 = [1, 0], y_1 = +1$, $x_2 = [\sin(0.001), \cos(0.001)], y_2 = +1$. The experimental setting is exactly the same as the illustrative example in Appendix A.1 (initialization scale $\alpha = 10^{-6}$). The marker represents either a data point or a neuron. Solid lines represent the directions of neurons. In the alignment phase, positive neurons are aligned with $x_+$, but the alignment is slower.

A.3 Neuron Dynamics under Orthogonal Data

We have seen in the last section how a small $\mu$ affects the neuron dynamics. The orthogonal data assumption studied in Boursier et al. (2022) is precisely the extreme case of $\mu \to 0$, where the neuron behavior changes substantially. We follow exactly the same setting in Appendix A.2 and consider the case of $\mu = 0$.

In Figure 9, we see that $S_+$ is no longer the region that contains all (non-dead) positive neurons at the end of the alignment phase. Depending on where each neuron is initialized, it could end up being in $S_+$, aligned with $x_1$, or aligned with $x_2$. Moreover, for final convergence, only the neurons ended up in $S_+$ grow their norms and fit the data, whose number is clearly less than that in the case of $\mu > 0$.

Figure 9: Neural alignment under two orthogonal data points: $x_1 = [1, 0], y_1 = +1, x_2 = [0, 1], y_2 = +1$. The experimental setting is exactly the same as the illustrative example in Appendix A.1. The marker represents either a data point or a neuron. Solid lines represent the directions of neurons. In the alignment phase, positive neurons are aligned with one of these directions: $x_1, x_2, x_+, x_-$, and only those aligned with $x_+$ grow their norm for final convergence.

This difference in neurons’ dynamical behavior makes the analysis in Boursier et al. (2022) different than ours: First, Boursier et al. (2022) only studies the dynamics of the positive (negative) neurons that initially activate all positive (negative) data, which will end up in $S_+$ ($S_-$) and fit the data, and the analysis does not evolve the changes in their activation pattern. In our case, any positive (negative) neurons could potentially end up in $S_+$ ($S_-$), and in particular, it will if it initially activates at least
one positive (negative) data, thus it becomes necessary to track the evolution of the activation pattern of all these neurons (novelty in our analysis). Moreover, consider the case that neurons are being randomly initialized, Boursier et al. (2022) requires the set of positive (negative) neurons that initially activate all positive (negative) data being non-empty, which needs the number of neurons \( h \) to scale exponentially in number training data \( n \) (extremely overparametrized). In our case, we only require \( h = \Theta(1) \) (See Merits of overparametrization after Theorem 1), a mild overparamerization.

In summary, while Boursier et al. (2022) also provides quantitative analysis on neural alignment under small initialization, it is done under the assumption that all data are orthogonal to each other, leading to a different neuron dynamical behavior than ours. Due to such differences, their analysis cannot be directly applied to the case of orthogonally separable data (ours), for which we develop novel analyses on the evolution of neuron activation patterns (See proof sketch in Section 3.3).

A.4 ADDITIONAL EXPERIMENTS ON MNIST DATASET

We use exactly the same experimental setting as in the main paper and only use a different pair of digits. The results are as follows:

Figure 10: Binary classification on MNIST Digits 3 and 8.

Figure 11: Binary classification on MNIST Digits 1 and 7.

A.5 DISCUSSION ON THE TWO-PHASE CONVERGENCE

With the same two-digit MNIST dataset in Section 4, we further discuss the two-phase convergence under small initialization. We use a two-layer ReLU network with \( h = 50 \) neurons and initialize all entries of the weights as \( [W]_{ij} \overset{i.i.d.}{\sim} \mathcal{N}(0, \alpha^2) \), \( v_j \overset{i.i.d.}{\sim} \mathcal{N}(0, \alpha^2) \), \( \forall i \in [n], j \in [h] \) with \( \alpha = 10^{-6} \). Then we run stochastic gradient descent (SGD) with batch size 2000 on both \( W \) and \( v \) with step size \( \eta = 2 \times 10^{-3} \). For comparison, we also consider the training schemes studied in Brutzkus et al. (2018), Wang et al. (2019), where only the first-layer weight \( W \) is trained starting from a small initialization \( [W]_{ij} \overset{i.i.d.}{\sim} \mathcal{N}(0, \alpha^2) \), and \( v_j \) are chosen to be either +1 or −1 with equal probability, then fixed throughout training.

We consider the changes in neuron norms and directions separately. In particular, these quantities are defined as

\[
\sum_i \frac{d}{dt} ||w_j||^2 |_{w_j = -\nabla w_j \mathcal{L}} = \sum_i 2 \left( -\nabla w_j \mathcal{L}, w_j \right) \quad \text{(changes in neuron norms)}
\]
Figure 12: Two-phase training under small initialization (SGD with a batch size of 2000, step size of $2 \times 10^{-3}$). At the early phase of the training, only neuron directions are changing while neurons’ norms do not grow. In the second stage, neurons start to grow their norms and loss starts to decrease. See Appendix A.5 for the precise definitions of “changes in neuron norms” and “changes in neuron directions.”

Training both layers: In Figure 12 we show the changes in neuron norms and directions over the training trajectory when we run stochastic gradient descent (SGD) on both first- and second-layer weights. The two-phase (alignment phase then final convergence) is clearly shown by comparing the relative scale of changes in neuron norms and directions in different phases of the training.

Training only the first layer: In Figure 13 we show the changes in neuron norms and directions over the training trajectory when we run stochastic gradient descent (SGD) on ONLY the first-layer weights (Brutzkus et al., 2018; Wang et al., 2019). The plot indicates that two-phase training does not happen in this case.
B PROOF OF LEMMA 1: NEURON DYNAMICS UNDER SMALL INITIALIZATION

The following property of \( \ell \) (exponential loss \( \ell(y, \hat{y}) = \exp(-y\hat{y}) \) or logistic loss \( \ell(y, \hat{y}) = 2\log(1+\exp(-y\hat{y})) \)) will be used throughout the Appendix for proofs of several results:

**Lemma 2.** For \( \ell \), we have

\[
| - \nabla_y \ell(y, \hat{y}) - y | \leq 2|\hat{y}|, \quad \forall y \in \{-1, 1\}, \quad \forall |\hat{y}| \leq 1.
\]  

**Proof.** **Exponential loss:** when \( \ell(y, \hat{y}) = \exp(-y\hat{y}) \):

\[
| - \nabla_y \ell(y, \hat{y}) - y | = |y\exp(-y\hat{y}) - y | \\
\leq |y||\exp(-y\hat{y}) - 1| \\
\leq |\exp(-y\hat{y}) - 1| \leq 2|\hat{y}|,
\]

where the last inequality is due to the fact that \( 2x \geq \max\{1 - \exp(-x), \exp(x) - 1\}, \forall x \in [0, 1] \).

**Logistic loss:** when \( \ell(y, \hat{y}) = 2\log(1+\exp(-y\hat{y})) \):

\[
| - \nabla_y \ell(y, \hat{y}) - y | = 2y \left| \frac{\exp(-y\hat{y})}{1+\exp(-y\hat{y})} - y \right| \\
= |y\exp(-y\hat{y}) - y | \\
\leq |y||\exp(-y\hat{y}) - 1| \\
\leq |\exp(-y\hat{y}) - 1| \leq 2|\hat{y}|,
\]

\[ \square \]

**Remark 4.** More generally, our subsequent results regarding neuron dynamics under small initialization hold for any loss function that satisfies the condition stated in Lemma 2, which includes the \( l_2 \) loss \( \ell(y, \hat{y}) = \frac{1}{2}(y - \hat{y})^2 \) studied in Boursier et al. [2022].

B.1 FORMAL STATEMENT

Our results for neuron direction dynamics during the early phase of the training will be stated for networks with any \( \alpha \)-leaky ReLU activation \( \sigma(x) = \max\{x, \alpha x\} \) with \( \alpha \in [0, 1] \). In particular, it is the ReLU activation when \( \alpha = 0 \), which is the activation function we considered in the main paper, and it is the linear activation when \( \alpha = 1 \).

Denote: \( X_{\text{max}} = \max_i ||x_i||, W_{\text{max}} = \max_j ||W_{0, j}|| \). The formal statement of Lemma 1 is as follow:

**Lemma 1.** Let the activation function be an \( \alpha \)-leaky ReLU activation \( \sigma(x) = \max\{x, \alpha x\} \). Given some initialization from \( \mathcal{D} \), for any \( \epsilon \leq \frac{1}{4nX_{\text{max}}^2W_{\text{max}}^2} \), then any solution to the gradient flow dynamics (2) satisfies that \( \forall t \leq T = \frac{1}{4nX_{\text{max}}^2W_{\text{max}}^2} \log \frac{1}{\sqrt{\epsilon}} \),

\[
\max_j \left\| \frac{d}{dt} \mathbf{w}_j(t) - \text{sign}(v_j(0))P_{w_j(t)}x_a(w_j(t)) \right\| \leq 4\epsilon n \sqrt{n}X_{\text{max}}^2 W_{\text{max}}^2,
\]

where

\[
x_a(w_j) = \sum_{i=1}^{n} x_i y_i \sigma'((x_i, w_j)) = \sum_{i:(x_i, w_j) > 0} x_i y_i + \alpha \sum_{i:(x_i, w_j) \leq 0} x_i y_i.
\]

With Lemma 1 and set \( \alpha = 0 \), we obtain the results stated in the main paper. Lemma 1 is a direct result of the following two lemmas.

**Lemma 3.** Let the activation function be an \( \alpha \)-leaky ReLU activation \( \sigma(x) = \max\{x, \alpha x\} \). Given some initialization in \( \mathcal{D} \), then for any \( \epsilon \leq \frac{1}{4nX_{\text{max}}^2W_{\text{max}}^2} \), any solution to the gradient flow dynamics (2) satisfies

\[
\max_j ||w_j(t)||^2 \leq 2\epsilon W_{\text{max}}^2, \quad \max_i |f(x_i, W(t), v(t))| \leq 2\epsilon \sqrt{n}X_{\text{max}} W_{\text{max}}^2,
\]

\( \forall t \leq \frac{1}{4nX_{\text{max}}^2} \log \frac{1}{\sqrt{\epsilon}} \).
Lemma 4. Let the activation function be an $\alpha$-leaky ReLU activation $\sigma(x) = \max\{x, \alpha x\}$. Consider any solution to the gradient flow dynamic (2) starting from initialization (3). Whenever $\max_{i\in[n]} |f(x_i; W, v)| \leq 1$, we have, $\forall t \in [n]$,  
\[
\left\| \frac{d}{dt} w_j \right\| - \text{sign}(v_j(0)) \left( I - \frac{w_j w_j^\top}{\|w_j\|^2} \right) x_a(w_j) \right\| \leq 2nX_{\max} \max_{i} |f(x_i; W, v)|, \tag{11}
\]
where 
\[
x_a(w_j) = \sum_{i=1}^{n} x_i y_i \sigma'((x_i, w_j)) = \sum_{i: (x_i, w_j) > 0} x_i y_i + \alpha \sum_{i: (x_i, w_j) \leq 0} x_i y_i.
\]

Remark 5. By stating our approximation results for neuron directional dynamics with any $\alpha$-leaky ReLU activation function, we highlight that even for some networks with other activation functions than ReLU, there is a similar notion of neuron alignment at the early stage of the training, and the analytical tools used in this paper can be applied to them. However, we note that our main results (Theorem 7) will not directly apply as the neuron directional dynamics have changed as we consider an activation function different than ReLU (see the general definition of $x_a(w_j)$), and additional efforts are required to establish the directional convergence for general leaky-ReLU functions.

B.2 PROOF OF LEMMA 3: BOUNDS ON NEURON NORMS

Proof of Lemma 3 Under gradient flow, we have
\[
\frac{d}{dt} w_j = -\sum_{i=1}^{n} \mathbb{1}_{(x_i, w_j) > 0} \nabla \ell(y_i, f(x_i; W, v)) x_i v_j. \tag{12}
\]
Balanced initialization enforces $v_j = \text{sign}(v_j(0))\|w_j\|$, hence
\[
\frac{d}{dt} w_j = -\sum_{i=1}^{n} \sigma'((x_i, w_j)) \nabla \ell(y_i, f(x_i; W, v)) x_i \text{sign}(v_j(0))\|w_j\|. \tag{13}
\]
Let $T := \inf\{t : \max_{i} |f(x_i; W(t), v(t))| > 2e\sqrt{h}X_{\max} W_{\max}^2 \}$, then $\forall t \leq T, j \in [h]$, we have
\[
\frac{d}{dt} \|w_j\|^2 = \left\langle w_j, \frac{d}{dt} w_j \right\rangle \\
= -2 \sum_{i=1}^{n} \sigma'((x_i, w_j)) \nabla \ell(y_i, f(x_i; W, v)) \langle x_i, w_j \rangle \text{sign}(v_j(0))\|w_j\| \\
\leq 2 \sum_{i=1}^{n} |\nabla \ell(y_i, f(x_i; W, v))| |\langle x_i, w_j \rangle| \|w_j\| \\
\leq 2 \sum_{i=1}^{n} (|y_i| + 2|f(x_i; W, v)|) |\langle x_i, w_j \rangle| \|w_j\| \tag{by Lemma 2} \\
\leq 2 \sum_{i=1}^{n} (1 + 4e\sqrt{h}X_{\max} W_{\max}^2) |\langle x_i, w_j \rangle| \|w_j\| \\
\leq 2 \sum_{i=1}^{n} (1 + 4e\sqrt{h}X_{\max} W_{\max}^2) \|x_i\| \|w_j\|^2 \tag{Since $t \leq T$} \\
\leq 2n(X_{\max} + 4e\sqrt{h}X_{\max} W_{\max}^2) \|w_j\|^2. \tag{14}
\]
Let $\tau_j := \inf\{t : \|w_j(t)\|^2 > \frac{2eW_{\max}^2}{\sqrt{h}} \}$, and let $j^* := \arg\min_{j} \tau_j$, then $\tau_{j^*} = \min_{j} \tau_j \leq T$ due to the fact that
\[
|f(x_i; W, v)| = \left| \sum_{j \in [h]} \sigma'((x_i, w_j))^j \langle w_j, x_i \rangle \right| \leq \sum_{j \in [h]} \|w_j\|^2 \|x_i\| \leq hX_{\max} \max_{j \in [h]} \|w_j\|^2,
\]
which implies \(|f(x_i; W(t), v(t))| > 2\epsilon \sqrt{h} X_{\text{max}} W^2_{\text{max}} \Rightarrow \exists j, s.t. \|w_j(t)\|^2 > 2\epsilon^2 W^2_{\text{max}}\).

Then for \(t \leq \tau_j\), we have
\[
\frac{d}{dt}\|w_j\|^2 \leq 2n(X_{\text{max}} + 4\epsilon \sqrt{h} X^2_{\text{max}} W^2_{\text{max}})\|w_j\|^2.
\] (15)

By Grönwall’s inequality, we have \(\forall t \leq \tau_j\)
\[
\|w_j(t)\|^2 \leq \exp\left(2n(X_{\text{max}} + 4\epsilon \sqrt{h} X^2_{\text{max}} W^2_{\text{max}})t\right)\|w_j(0)\|^2,
\]
\[
= \exp\left(2n(X_{\text{max}} + 4\epsilon \sqrt{h} X^2_{\text{max}} W^2_{\text{max}})t\right) \epsilon^2\|W_0\|_2^2\|
\]
\[
\leq \exp\left(2n(X_{\text{max}} + 4\epsilon \sqrt{h} X^2_{\text{max}} W^2_{\text{max}})t\right) \epsilon^2 W^2_{\text{max}}.
\]

Suppose \(\tau_j < \frac{1}{4nX_{\text{max}}} \log \left(\frac{1}{\epsilon^2\sqrt{h}}\right)\), then by the continuity of \(\|w_j(t)\|^2\), we have
\[
\frac{2\epsilon W^2_{\text{max}}}{\sqrt{h}} \leq \|w_j(t)\|^2 \leq \exp\left(\frac{1}{4nX_{\text{max}}} \log \left(\frac{1}{\sqrt{h}\epsilon^2}\right)\right) \epsilon^2 W^2_{\text{max}}
\]
\[
\leq \exp\left(\left(1 + 4\epsilon \sqrt{h} X^2_{\text{max}} W^2_{\text{max}}\right) \log \left(\frac{1}{\sqrt{h}\epsilon^2}\right)\right) \epsilon^2 W^2_{\text{max}}
\]
\[
\leq \exp\left(\log \left(\frac{1}{\sqrt{h}\epsilon^2}\right)\right) \epsilon^2 W^2_{\text{max}} = \frac{\epsilon^2 W^2_{\text{max}}}{\sqrt{h}}
\]

which leads to a contradiction \(2\epsilon \leq \epsilon\). Therefore, one must have \(T \geq \tau_j \geq \frac{1}{4nX_{\text{max}}} \log \left(\frac{1}{\sqrt{h}\epsilon^2}\right)\).

This finishes the proof. \(\Box\)

### B.3 Proof of Lemma \[\text{Directional Dynamics of Neurons}\]

**Proof of Lemma** As we showed in the proof for Lemma \[\text{under balanced initialization},\]
\[
\frac{d}{dt} w_j = -\sum_{i=1}^{n} 1_{(x_i, w_j) > 0} \nabla g\ell(y_i, f(x_i; W, v)) x_i \text{sign}(v_j(0)) \|w_j\|.
\] (16)

Then for any \(i \in [n],\)
\[
\frac{d}{dt} \frac{w_j}{\|w_j\|} = -\text{sign}(v_j(0)) \sum_{i=1}^{n} 1_{(x_i, w_j) > 0} \nabla g\ell(y_i, f(x_i; W, v)) \left(x_i - \frac{\langle x_i, w_j \rangle}{\|w_j\|^2} w_j\right)
\]
\[
= -\text{sign}(v_j(0)) \left( I - \frac{w_j w_j^T}{\|w_j\|^2} \right) \left(\sum_{i=1}^{n} \sigma'(\langle x_i, w_j \rangle) \nabla g\ell(y_i, f(x_i; W, v)) x_i \right).
\]

Therefore, whenever \(\max_i |f(x_i; W, v)| \leq 1,\)
\[
\frac{d}{dt} \frac{w_j}{\|w_j\|} - \text{sign}(v_j(0)) \left( I - \frac{w_j w_j^T}{\|w_j\|^2} \right) x_a(w_j)
\]
\[
= -\text{sign}(v_j(0)) \left(\sum_{i=1}^{n} \sigma'(\langle x_i, w_j \rangle) \left(\nabla g\ell(y_i, f(x_i; W, v)) + y_i\right) x_i\right)
\]
\[
\leq \sum_{i=1}^{n} |\nabla g\ell(y_i, f(x_i; W, v)) + y_i| \cdot \|x_i\|
\]
\[
\leq \sum_{i=1}^{n} 2 |f(x_i; W, v)| \cdot \|x_i\| \leq 2n M_x \max_i |f(x_i; W, v)|.
\] (17)
C PROOF FOR THEOREM 1: EARLY ALIGNMENT PHASE

We break the proof of Theorem 1 into two parts: In Appendix C we prove the first part regarding directional convergence. Then in Appendix D we prove the remaining statement on final convergence and low-rank bias.

C.1 AUXILIARY LEMMAS

The first several Lemmas concern mostly some conic geometry given the data assumption:

Consider the following conic hull

\[ K = CH(\{x_i y_i, i \in [n]\}) = \left\{ \sum_{i=1}^{n} a_i x_i y_i : a_i \geq 0, i \in [n] \right\}. \]  (18)

It is clear that \( x_i y_i \in K, \forall i \), and \( x_a(w) \in K, \forall w \). The following lemma shows any pair of vectors in \( K \) is \( \mu \)-coherent.

**Lemma 5.** \( \cos(z_1, z_2) \geq \mu, \forall 0 \neq z_1, z_2 \in K. \)

**Proof.** Since \( z_1, z_2 \in K \), we let \( z_1 = \sum_{i=1}^{n} x_i y_i a_{i1}, \) and \( z_2 = \sum_{j=1}^{n} x_j y_j a_{2j} \), where \( a_{1i}, a_{2j} \geq 0 \) but not all of them.

\[
\cos(z_1, z_2) = \frac{1}{\|z_1\| \|z_2\|} \langle z_1, z_2 \rangle = \frac{1}{\|z_1\| \|z_2\|} \sum_{i,j \in [n]} a_{1i} a_{2j} \langle x_i y_i, x_j y_j \rangle
\]

\[
= \sum_{i,j \in [n]} \|x_i\| \|x_j\| \|a_{1i} a_{2j}\| \geq \mu,
\]

where the last inequality is due to

\[
\|z_1\| \|z_2\| \leq \left( \sum_{i=1}^{n} \|x_i\| a_{i1} \right) \left( \sum_{j=1}^{n} \|x_j\| a_{2j} \right) = \sum_{i,j \in [n]} \|x_i\| \|x_j\| a_{1i} a_{2j}.
\]

The following lemma is some basic results regarding \( S_+ \) and \( S_- \): \( S_+ \) and \( S_- \) are convex cones (excluding the origin).

**Lemma 6.** \( S_+ \) and \( S_- \) are convex cones (excluding the origin).

**Proof.** Since \( 1_{(x_i, z_i)} = 1_{(x_i, a_z)}, \forall i \in [n], a > 0, S_+, S_- \) are cones. Moreover, \( \langle x_i, z_1 \rangle > 0 \) and \( \langle x_i, z_2 \rangle > 0 \) implies \( \langle x_i, a_1 z_1 + a_2 z_2 \rangle > 0, \forall a_1, a_2 > 0 \), thus \( S_+, S_- \) are convex cones.

Now we consider the complete metric space \( S^{D-1} \) (w.r.t. \( \arccos(\langle \cdot, \cdot \rangle) \)) and we are interested in its subsets \( K \cap S^{D-1}, S_+ \cap S^{D-1}, \) and \( S_- \cap S^{D-1} \). First, we have (we use \( \text{Int}(S) \) to denote the interior of \( S \))

**Lemma 7.** \( K \cap S^{D-1} \subset \text{Int}(S_+ \cap S^{D-1}), \) and \( -K \cap S^{D-1} \subset \text{Int}(S_- \cap S^{D-1}) \)

**Proof.** Consider any \( x_c = \sum_{j=1}^{n} a_j x_j y_j \in K \cap S^{D-1} \). For any \( x_i, y_i, i \in [n] \), we have

\[
\langle x_c, x_i \rangle = \sum_{i,j} a_j \|x_j\| \left( \frac{x_j y_j}{\|x_j\|} \cdot \frac{x_i y_i}{\|x_i\|} \right) \geq \mu y_i \|x_i\| \sum_{i,j} a_j \|x_j\| \geq \mu X^{\min} > 0, \quad y_i > 0
\]

Depending on the sign of \( y_i \), we have either

\[
\langle x_c, x_i \rangle = \sum_{i,j} a_j \|x_j\| \left( \frac{x_j y_j}{\|x_j\|} \cdot \frac{x_i y_i}{\|x_i\|} \right) \geq \mu y_i \|x_i\| \sum_{i,j} a_j \|x_j\| \geq \mu X^{\min} > 0, \quad (y_i = +1)
\]

21
or
\[
\langle x_c, x_i \rangle = \sum_{i=j}^{n} a_j \| x_j \| \left( \frac{x_j y_j}{\| x_j \|} \right) \| x_i \| \leq \mu \sum_{i=j}^{n} a_j \| x_j \| \leq -\mu X_{\min} < 0, \ (y_i = -1)
\]

where we use the fact that \( 1 = \| x_c \| = \| \sum_{j=1}^{n} a_j x_j y_j \| \leq \sum_{j=1}^{n} a_j \| x_j \| \). This already tells us \( x_c \in S_+ \cap S^{D-1} \).

Since \( f_i(z) = \langle z, x_i \rangle \) is a continuous function of \( z \in S^{D-1} \). There exists an open ball \( B(x_c, \delta_i) \) centered at \( x_c \) with some radius \( \delta_i > 0 \), such that \( \forall z \in B(x_c, \delta_i) \), one have \( |f_i(z) - f_i(x_c)| \leq \frac{\mu X_{\min}}{\delta_i} \), which implies
\[
\langle z, x_i \rangle \begin{cases} 
\geq \mu X_{\min}/2 > 0, & y_i > 0 \\
\leq -\mu X_{\min}/2 < 0, & y_i < 0.
\end{cases}
\]

Hence \( \cap_{i=1}^{n} B \left( \frac{x_c}{\| x_c \|}, \delta_i \right) \in S_+ \cap S^{D-1} \). Therefore, \( x_c \in \text{Int}(S_+ \cap S^{D-1}) \). This suffices to show \( K \cap S^{D-1} \subset \text{Int}(S_+ \cap S^{D-1}) \). The other statement \( -K \cap S^{D-1} \subset \text{Int}(S_- \cap S^{D-1}) \) is proved similarly.

The following two lemmas are some direct results of Lemma 7

**Lemma 8.** \( \exists \zeta > 0 \) such that
\[
S_{x_+}^j \subset S_+, \quad S_{x_-}^j \subset S_-,
\]
where \( S_x^j := \{ z \in \mathbb{R}^D : \cos(z, x) \geq \sqrt{1 - \zeta} \} \).

**Proof.** By Lemma 7 \( \frac{x_+}{\| x_+ \|} \in K \subset \text{Int}(S_+) \). Since \( S^{D-1} \) is a complete metric space (w.r.t \( \arccos(\langle \cdot, \cdot \rangle) \)), there exists an open ball centered at \( \frac{x_+}{\| x_+ \|} \) of some radius \( \arccos(\sqrt{1 - \zeta}) \) that is a subset of \( S_+ \), from which one can show \( S_{x_+}^j \subset S_+ \). The other statement \( S_{x_-}^j \subset S_- \) simply comes from the fact that \( x_+ = -x_- \) and \( \text{Int}(S_+) = -\text{Int}(S_-) \).

**Lemma 9.** \( \exists \xi > 0 \) such that
\[
\sup_{x_1 \in K \cap S^{D-1}, x_2 \in (S_+ \cap S^{D-1})^c \cap (S_- \cap S^{D-1})^c} |\cos(x_1, x_2)| \leq \sqrt{1 - \xi}.
\]

\((S^c \) here is defined to be \( S^{D-1} - S \), the set complement w.r.t. complete space \( S^{D-1} \))

**Proof.** Notice that
\[
\sup_{x_1 \in K \cap S^{D-1}, x_2 \in (\text{Int}(S_+ \cap S^{D-1}))^c} \langle x_1, x_2 \rangle = \inf_{x_1 \in K \cap S^{D-1}, x_2 \in \text{Int}(S_+ \cap S^{D-1})^c} \arccos \langle x_1, x_2 \rangle.
\]

Since \( S^{D-1} \) is a complete metric space (w.r.t \( \arccos(\langle \cdot, \cdot \rangle) \)) and \( K \cap S^{D-1} \) and \( x_2 \in \text{Int}(S_+ \cap S^{D-1})^c \) are two of its compact subsets. Suppose
\[
\inf_{x_1 \in K \cap S^{D-1}, x_2 \in \text{Int}(S_+ \cap S^{D-1})^c} \arccos \langle x_1, x_2 \rangle = 0,
\]
then \( \exists x_1 \in K \cap S^{D-1}, x_2 \in \text{Int}(S_+ \cap S^{D-1})^c \) such that \( \arccos \langle x_1, x_2 \rangle = 0 \), i.e., \( x_1 = x_2 \), which contradicts the fact that \( K \cap S^{D-1} \subset \text{Int}(S_+ \cap S^{D-1}) \) (Lemma 7). Therefore, we have the infimum strictly larger than zero, then
\[
\sup_{x_1 \in K \cap S^{D-1}, x_2 \in (S_+ \cap S^{D-1})^c} \langle x_1, x_2 \rangle \leq \sup_{x_1 \in K \cap S^{D-1}, x_2 \in \text{Int}(S_+ \cap S^{D-1})^c} \langle x_1, x_2 \rangle < 1. \tag{21}
\]

Similarly, one can show that
\[
\sup_{x_1 \in -K \cap S^{D-1}, x_2 \in (S_- \cap S^{D-1})^c} \langle x_1, x_2 \rangle < 1. \tag{22}
\]

Finally, find \( \xi < 1 \) such that
\[
\max \left\{ \sup_{x_1 \in K \cap S^{D-1}, x_2 \in (S_+ \cap S^{D-1})^c} \langle x_1, x_2 \rangle, \sup_{x_1 \in -K \cap S^{D-1}, x_2 \in (S_- \cap S^{D-1})^c} \langle x_1, x_2 \rangle \right\} = \sqrt{1 - \xi},
\]
then for any \( x_1 \in K \cap S^{D-1} \) and \( x_2 \in (S_+ \cap S^{D-1}) \cap (S_- \cap S^{D-1})^c \), we have
\[
-\sqrt{1-\xi} \leq \langle x_1, x_2 \rangle \leq \sqrt{1-\xi},
\]
which is the desired result.

The remaining two lemmas are technical but extensively used in the main proof.

**Lemma 10.** Consider any solution to the gradient flow dynamic (2) starting from initialization (3). Let \( x_r \in S^{n-1} \) be some reference direction, we define
\[
\psi_{rj} = \left\langle x_r, \frac{w_j}{\|w_j\|} \right\rangle, \quad \psi_{ra} = \left\langle x_r, \frac{x_a(w_j)}{\|x_a(w_j)\|} \right\rangle, \quad \psi_{aj} = \left\langle \frac{w_j}{\|w_j\|}, \frac{x_a(w_j)}{\|x_a(w_j)\|} \right\rangle,
\]
where \( x_a(w_j) = \sum_{i: (x_i, w_j)>0} y_i x_i \).

Whenever \( \max_i |f(x_i; W, v)| \leq 1 \), we have
\[
\left| \frac{d}{dt} \psi_{rj} - \text{sign}(v_j(0)) (\psi_{ra} - \psi_{rj}\psi_{aj}) \|x_a(w_j)\| \right| \leq 2nX_{\text{max}} \max_i |f(x_i; W, v)|. \tag{24}
\]

**Proof.** A simple application of Lemma 5 together with Cauchy-Schwartz:
\[
\left| \frac{d}{dt} \psi_{rj} - \text{sign}(v_j(0)) (\psi_{ra} - \psi_{rj}\psi_{aj}) \|x_a(w_j)\| \right| = \left| x_r^\top \left( \frac{d}{dt} \frac{w_j}{\|w_j\|} - \text{sign}(v_j(0)) \left( I - \frac{w_j w_j^\top}{\|w_j\|^2} \right) \left( \sum_{i: (x_i, w_j)>0} y_i x_i \right) \right) \right| \leq 2nX_{\text{max}} \max_i |f(x_i; W, v)|. \tag{25}
\]

**Lemma 11.**
\[
\|x_a(w)\| \geq \sqrt{n_a(w)}X_{\text{min}},
\]
where \( n_a(w) = |\{i \in [n] : \langle x_i, w \rangle > 0\}|. \tag{26}
\]

**Proof.** Let \( I_a(w) \) denote \( \{i \in [n] : \langle x_i, w \rangle > 0\} \), then
\[
\|x_a(w)\| = \left\| \sum_{i: \langle x_i, w \rangle > 0} x_i y_i \right\| = \left\| \sum_{i \in I_a(w)} \|x_i\|^2 y_i^2 + \sum_{i, j \in I_a(w), i < j} \|x_i\| \|x_j\| y_i y_j \mu \right\|
\[
\geq \sqrt{\sum_{i \in I_a(w)} \|x_i\|^2 y_i^2 + \sum_{i, j \in I_a(w), i < j} \|x_i\| \|x_j\| y_i y_j \mu}
\[
\geq \sqrt{n_a(w)X_{\text{min}}^2 + \mu n_a(w) (n_a(w) - 1) X_{\text{min}}^2}
\[
\geq \sqrt{n_a(w)(1 + \mu (n_a(w) - 1)) X_{\text{min}}}
\[
\geq \sqrt{\sqrt{\mu n_a(w)}X_{\text{min}}}. \tag{27}
\]

C.2 Proof for Early Alignment Phase

**Proof of Theorem 1.** First Part. Given some initialization in (3), by Assumption 2 \( \exists \zeta_2 > 0 \), such that
\[
\max_{j \in V_+} \cos(w_j(0), x_-) < \sqrt{1 - \zeta_2}, \quad \max_{j \in V_-} \cos(w_j(0), x_+) < \sqrt{1 - \zeta_2}. \tag{26}
\]

We define \( \zeta := \max\{\zeta_1, \zeta_2\} \), where \( \zeta_1 \) is from Lemma 8. In addition, by Lemma 9 \( \exists \xi > 0 \), such that
\[
\sup_{x_1 \in K \cap S^{D-1}, x_2 \in S_- \cap S_+^c \cap S^{D-1}} |\cos(x_1, x_2)| \leq \sqrt{1 - \xi}. \tag{27}
\]
We pick an initialization scale $\epsilon$ that satisfies:

$$
\epsilon \leq \min \left\{ \frac{\min\{\mu, \zeta, \xi\} \sqrt{\beta X_{\min}}}{4\sqrt{\ln X_{\max}} W_{\max}^2}, \frac{1}{\sqrt{h}} \exp\left(-\frac{64nX_{\max}}{\min\{\zeta, \xi\} \sqrt{\beta X_{\min}} \log n}\right) \right\} \leq \frac{1}{4\sqrt{\ln X_{\max}} W_{\max}^2}.
$$

(28)

By Lemma 3, $\forall t < T = \frac{1}{4nX_{\max}} \log \frac{1}{\epsilon}$, we have

$$
\max_i |f(x_i; W, v)| \leq \frac{\min\{\mu, \zeta, \xi\} \sqrt{\beta X_{\min}}}{4nX_{\max}},
$$

(29)

which is the key to analyzing the alignment phase. For the sake of simplicity, we only discuss the analysis of neurons in $V_+$ here, the proof for neurons in $V_-$ is almost identical.

**Activation pattern evolution:** Pick any $w_j$ in $V_+$ and pick $x_r = x_i y_i$ for some $i \in [n]$, and consider the case when $\langle w_j, x_i \rangle = 0$. From Lemma 10, we have

$$
\left| \frac{d}{dt} \psi_{rj} - (\psi_{ra} - \psi_{rj}\psi_{aj}) \|x_a(w_j)\| \right| \leq 2nX_{\max} \max_i |f(x_i; W, v)|.
$$

Then whenever $w_j \notin S_{\text{dead}}$, we have

$$
\left| \frac{d}{dt} \psi_{rj}|_{(w_j, x_i) = 0} \psi_{ra}\|x_a(w_j)\| - 2nX_{\max} \max_i |f(x_i; W, v)| \right|
\geq \mu\|x_a(w_j)\| - 2nX_{\max} \max_i |f(x_i; W, v)|
\geq \mu^{3/2} X_{\min} - 2nX_{\max} \max_i |f(x_i; W, v)|
\geq \mu^{3/2} X_{\min}/2 > 0.
$$

(29)

This is precisely \(29\) in Section 3.3.

**Bound on activation transitions and duration:** Next we show that if at time $t_0 < T$, $w_j(t_0) \notin S_+ \cup S_{\text{dead}}$, and the activation pattern of $w_j$ is $\mathbb{1}_{\{x_r(w_j(t_0)) > 0\}}$, then $\mathbb{1}_{\{x_r(w_j(\tau)) > 0\}} = \mathbb{1}_{\{x_r(w_j(t)) > 0\}}$, i.e. $w_j$ fixes its activation during $[t_0, t_0 + \Delta t]$, then we have $x_a(w_j(t_0 + \tau)) = x_a(w_j(t_0))$, $\forall 0 \leq \tau \leq \Delta t$. Let us pick $x_r = x_a(w_j(t_0))$, then Lemma 10 leads to

$$
\left| \frac{d}{dt} \cos(w_j, x_a(w_j)) - (1 - \cos^2(w_j, x_a(w_j))) \|x_a(w_j)\| \right| \leq 2nX_{\max} \max_i |f(x_i; W, v)|.
$$

Since $x_a(w_j)$ is fixed, we have $\forall t \in [t_0, t_0 + \Delta t]$,

$$
\left| \frac{d}{dt} \cos(w_j, x_a(w_j(t_0))) - (1 - \cos^2(w_j, x_a(w_j(t_0)))) \|x_a(w_j(t_0))\| \right| \leq 2nX_{\max} \max_i |f(x_i; W, v)|,
$$

$$
\left| \frac{d}{dt} \cos(w_j, x_a(w_j(t_0))) \geq (1 - \cos^2(w_j, x_a(w_j(t_0)))) \|x_a(w_j(t_0))\|ight|
\geq -2nX_{\max} \max_i |f(x_i; W, v)|
\geq \xi \|x_a(w_j(t_0))\| - 2nX_{\max} \max_i |f(x_i; W, v)|
\geq \xi \sqrt{n a(w_j(t_0))} X_{\min} - 2nX_{\max} \max_i |f(x_i; W, v)|
\geq \xi \sqrt{n a(w_j(t_0))} X_{\min}/2.
$$

(29)

(27)
which implies that, by the Fundamental Theorem of Calculus,
\[
\begin{align*}
\cos(w_j(t_0 + \Delta t), x_a(w_j(t_0))] &= \cos(w_j(t_0), x_a(w_j(t_0))] + \int_0^{\Delta t} \frac{d}{dt} \cos(w_j(t_0 + \tau), x_a(w_j(t_0))] d\tau \\
&\geq \cos(w_j(t_0), x_a(w_j(t_0))] + \Delta t \cdot \min\{\xi, \zeta\} \sqrt{\mu_{w_a}(w_j(t_0))] X_{\text{min}}/2 \\
&= \cos(w_j(t_0), x_a(w_j(t_0))] + 2 \geq 1,
\end{align*}
\]
which leads to \(\cos(w_j(t_0 + \Delta t), x_a(w_j(t_0))] = 1\). This would imply \(w_j(t_0 + \Delta t) \in S_+\) because \(x_a(w_j(t_0))] \in S_+\), which contradicts our original assumption that \(w_j\) fixes the activation pattern. Therefore, \(\exists \tau_0 \leq \Delta t\) such that \(\vec{x}(w_j(t_0 + \tau_0)) \neq \vec{x}(w_j(t_0)) \geq 0\), due to the restriction on how \(w_j\) can change its activation pattern, it cannot return to its previous activation pattern, then one must have \(\vec{x}(x, w_j(t_0 + \Delta t)) \neq \vec{x}(x, w_j(t_0)) \geq 0\).

\(\blacksquare\) The other case is when \(w_j(t_0) \in S_-\). For this case, we need first show that \(w_j(t_0 + \tau) \in S_{-\zeta}, \forall 0 \leq \tau \leq \Delta t\), or more generally, \(S_r^\zeta\) does not contain any \(w_j \in V_r\) during \([0, T]\). To see this, let us pick \(x_r = x_-\), then Lemma 10 suggests that
\[
\left| \frac{d}{dt} \psi_{r} - \overline{\psi_{r}} \right| \|x_a(w_j))] \leq 2nX_{\text{max}} \max_i |f(x_i; W, v)|.
\]
Consider the case when \(\cos(w_j, x_-) = \sqrt{1-\zeta}\), i.e. \(w_j\) is at the boundary of \(S_{-\zeta}\). We know that in this case, \(w_j \in S_{-\zeta}^\zeta \subseteq S_-\), thus \(x_a(w_j) = -x_-\), and
\[
\left| \frac{d}{dt} \cos(w_j, x_-) \right|_{\cos(w_j, x_-) = \sqrt{1-\zeta}} + \left(1 - \cos^2(w_j, x_-) \right) \|x_-\| \leq 2nX_{\text{max}} \max_i |f(x_i; W, v)|,
\]
which is
\[
\left| \frac{d}{dt} \cos(w_j, x_-) \right|_{\cos(w_j, x_-) = \sqrt{1-\zeta}} + \zeta \|x_-\| \leq 2nX_{\text{max}} \max_i |f(x_i; W, v)|
\]
\[
\Rightarrow \left| \frac{d}{dt} \cos(w_j, x_-) \right|_{\cos(w_j, x_-) = \sqrt{1-\zeta}} \leq -\zeta \|x_-\| + 2nX_{\text{max}} \max_i |f(x_i; W, v)|
\]
\[
\leq -\zeta \sqrt{\mu X_{\text{min}}} + 2nX_{\text{max}} \max_i |f(x_i; W, v)| \quad \text{(by Lemma 11)}
\]
\[
\leq -\zeta \sqrt{\mu X_{\text{min}}/2} < 0. \quad \text{(by 29)}
\]
Therefore, during \([0, T]\), neuron \(w_j \in V_+\) cannot enter \(S^\zeta_{-\zeta}\) if at initialization, \(w_j(0) \notin S^\zeta_{-\zeta}\), which is guaranteed by 26.

With the argument above, we know that \(w_j(t_0 + \tau) \notin S^\zeta_{-\zeta}, \forall 0 \leq \tau \leq \Delta t\). Again we suppose that \(w_j(t) \in S_- - S^\zeta_{-\zeta}, \forall t \in [t_0, t_0 + \Delta t]\), i.e., \(w_j\) fixes its activation during \([t_0, t_0 + \Delta t]\). Let us pick \(x_r = x_-\), then Lemma 10 suggests that
\[
\left| \frac{d}{dt} \cos(w_j, x_-) + \left(1 - \cos^2(w_j, x_-) \right) \|x_-\| \right| \leq 2nX_{\text{max}} \max_i |f(x_i; W, v)|,
\]
which leads to \(\forall t \in [t_0, t_0 + \Delta t],\)
\[
\frac{d}{dt} \cos(w_j, x_-) \leq -\left(1 - \cos^2(w_j, x_-) \right) \|x_-\| + 2nX_{\text{max}} \max_i |f(x_i; W, v)|
\]
\[
\leq -\zeta \|x_-\| + 2nX_{\text{max}} \max_i |f(x_i; W, v)| \quad \text{(w_j \notin S^\zeta_{-\zeta})}
\]
\[
\leq -\zeta \sqrt{\mu a(w_j(t_0))] X_{\text{min}} + 2nX_{\text{max}} \max_i |f(x_i; W, v)| \quad \text{(by Lemma 11)}
\]
\[
\leq -\zeta \sqrt{\mu a(w_j(t_0))] X_{\text{min}}/2. \quad \text{(by 29)}
\]
\[
\leq - \min\{\xi, \zeta\} \sqrt{\mu a(w_j(t_0))] X_{\text{min}}/2,
\]
Similarly, by FTC, we have

$$\cos(w_j(t_0 + \Delta t), x_-) \leq -1.$$  

This would imply $w_j(t_0 + \Delta t) \in \mathcal{S}_\Delta$ because $-x_- = x_a(w_j(t_0)) \in \mathcal{S}_\Delta$, which contradicts our original assumption that $w_j$ fixes its activation pattern. Therefore, one must have $1(x, w_j(t_0+\Delta t)) \neq 1(x, w_j(t_0)) > 0$.

In summary, we have shown that, during $[0, T]$, a neuron in $\mathcal{V}_+$ cannot keep a fixed activation pattern for a time longer than $\Delta t = \frac{4}{\min\{\zeta, \xi\} \sqrt{\mu X_{\min} n_a}}$, where $n_a$ is the number of data points that activate $w_j$ under the fixed activation pattern.

**Bound on total travel time until directional convergence** As we have discussed in Section 3.3 and also formally proved here, during alignment phase $[0, T]$, a neuron in $\mathcal{V}_+$ must change its activation pattern within $\Delta t = \frac{4}{\min\{\zeta, \xi\} \sqrt{\mu X_{\min} n_a}}$ time unless it is in either $\mathcal{S}_+$ or $\mathcal{S}_{\text{dead}}$. And the new activation it is transitioning into must contain no new activation on negative data points and must keep all existing activation on positive data points, together it shows that a neuron must reach either $\mathcal{S}_+$ or $\mathcal{S}_{\text{dead}}$ within a fixed amount of time, which is the remaining thing we need to formally show here.

For simplicity of the argument, we first assume $T = \infty$, i.e., the alignment phase lasts indefinitely, and we show that a neuron in $\mathcal{V}_+$ must reach $\mathcal{S}_+$ or $\mathcal{S}_{\text{dead}}$ before $t_1 = \frac{16 \log n}{\min\{\zeta, \xi\} \sqrt{\mu X_{\min}}}$ Lastly, such directional convergence can be achieved if $t_1 \leq T$, which is guaranteed by our choice of $\epsilon$ in Section 3.3.

- For a neuron in $\mathcal{V}_+$ that reaches $\mathcal{S}_{\text{dead}}$, the analysis is easy: It must start with no activation on positive data and then lose activation on negative data one by one until losing all of its activation. Therefore, it must reach $\mathcal{S}_{\text{dead}}$ before

  \[
  \sum_{k=1}^{n_a(w_j(0))} \frac{4}{\min\{\zeta, \xi\} \sqrt{\mu X_{\min} k}} \leq \frac{4}{\min\{\zeta, \xi\} \sqrt{\mu X_{\min}}} \left( \sum_{k=1}^{n} \frac{1}{k} \right) \leq \frac{16 \log n}{\min\{\zeta, \xi\} \sqrt{\mu X_{\min}}} = t_1.
  \]

- For a neuron in $\mathcal{V}_+$ that reaches $\mathcal{S}_+$, there is no difference conceptually, but it can switch its activation pattern in many ways before reaching $\mathcal{S}_+$, so it is not straightforward to see its travel time until $\mathcal{S}_+$ is upper bounded by $t_1$.

To formally show the upper bound on the travel time, we need some definition of a path that keeps a record of the activation patterns of a neuron $w_j(t)$ before it reaches $\mathcal{S}_+$.

Let $n_+ = |\mathcal{I}_+|$, $n_- = |\mathcal{I}_-|$ be the number of positive, negative data respectively, then we call $\mathcal{P}(k^{(0)}, k^{(1)}, \ldots, k^{(L)})$ a path of length $L$, if

1. $0 \leq l < L$, we have $k^{(l)} = (k^{(l)}_+, k^{(l)}_-) \in \mathbb{N} \times \mathbb{N}$ with $0 \leq k^{(l)}_+ \leq n_+$, $0 \leq k^{(l)}_- \leq n_-;
2. For $k^{(l_1)}_+, k^{(l_2)}_-$ with $l_1 < l_2$, we have either $k^{(l_1)}_+ > k^{(l_2)}_+$ or $k^{(l_1)}_- < k^{(l_2)}_-;
3. k^{(L)} = (n_+, 0);
4. $k^{(l)} \neq (0, 0), \forall 0 \leq l \leq L$.

Given all our analysis on how a neuron $w_j(t)$ can switch its activation pattern in previous parts, we know that for any $w_j(t)$ that reaches $\mathcal{S}_+$, there is an associated $\mathcal{P}(k^{(0)}, k^{(1)}, \ldots, k^{(L)})$ that keeps an ordered record of encountered values of

$$\{(i \in \mathcal{I}_+ : \langle x_i, w_j(t) \rangle > 0\}, \{(i \in \mathcal{I}_- : \langle x_i, w_j(t) \rangle > 0\}|,$$

before $w_j$ reaches $\mathcal{S}_+$. That is, a neuron $w_j$ starts with some activation pattern that activates $k_+(0)$ positive data and $k_-(0)$ negative data, then switch its activation pattern (by either losing negative data or gaining positive data) to one that activates $k_+(1)$ positive data and $k_-(1)$ negative data. By keep doing so, it reaches $\mathcal{S}_+$ that activates $k_+(L) = n_+$ positive data and $k_-(L) = 0$ negative data. Please see Figure 14 for an illustration of a path.

Given a path $\mathcal{P}(k^{(0)}, k^{(1)}, \ldots, k^{(L)})$ of neuron $w_j$, we define the travel time of this path as

$$T(\mathcal{P}(k^{(0)}, k^{(1)}, \ldots, k^{(L)})) = \sum_{l=0}^{L-1} \frac{4}{\min\{\zeta, \xi\} \sqrt{\mu X_{\min} (k^{(l)}_+ + k^{(l)}_-)}},$$

26
Under review as a conference paper at ICLR 2024

Figure 14: Illustration of a path of length-10. Each dot on the grid represents one $k^{(l)}$.

Figure 15: Illustration of a path and the maximal path

which is exactly the traveling time from $k^{(0)}$ to $k^{(L)}$ if one spends $\frac{4}{\min\{\zeta, \xi\} \sqrt{\mu X_{\min}(k^{(l)} + k^{(l+1)})}}$ on the edge between $k^{(l)}$ and $k^{(l+1)}$.

Our analysis shows that if $w_j$ reaches $S_+$, then

$$\inf\{t : w_j(t) \in S_+\} \leq T(P_{(k^{(0)}, k^{(1)}, \ldots, k^{(L)})}).$$

Now we define the maximal path $P_{\text{max}}$ as a path that has the maximum length $n = n_+ + n_-$, which is uniquely determined by the following trajectory of $k^{(l)}$

$$(0, n_-), (0, n_- - 1), (0, n_- - 2), \ldots, (0, 1), (1, 1), (1, 0), \ldots, (n_+ - 1, 0), (n_+, 0).$$

Please see Figure 15 for an illustration.

The traveling time for $P_{\text{max}}$ is

$$T(P_{\text{max}}) = \frac{4}{\min\{\zeta, \xi\} \sqrt{\mu X_{\min}}} \left( \sum_{k=1}^{n} \frac{1}{k} + \frac{n_+ - 1}{2} \right) \leq \frac{16 \log n}{\min\{\zeta, \xi\} \sqrt{\mu X_{\min}}} = t_1.$$

The proof is complete by the fact that any path satisfies

$$T(P_{(k^{(0)}, k^{(1)}, \ldots, k^{(L)})}) \leq T(P_{\text{max}}).$$

This is because there is a one-to-one correspondence between the edges $(k^{(l)}, k^{(l+1)})$ in $P_{(k^{(0)}, k^{(1)}, \ldots, k^{(L)})}$ and a subset of edges in $P_{\text{max}}$, and the travel time from of edge $(k^{(l)}, k^{(l+1)})$ is shorter than the corresponding edge in $P_{\text{max}}$. Formally stating such correspondence is tedious and a visual illustration in Figure 16 is more effective (Putting all correspondence makes a clustered plot thus we split them into two figures):

Therefore, if $w_j$ reaches $S_+$, then it reaches $S_+$ within $t_1$:

$$\inf\{t : w_j(t) \in S_+\} \leq T(P_{(k^{(0)}, k^{(1)}, \ldots, k^{(L)})}) \leq T(P_{\text{max}}) \leq t_1.$$

So far we have shown when the alignment phase lasts long enough, i.e., $T$ large enough, the directional convergence is achieved by $t_1$. We simply pick $\epsilon$ such that

$$T = \frac{1}{4nX_{\max}} \log \frac{1}{\sqrt{n\epsilon}} \geq t_1 = \frac{16 \log n}{\min\{\zeta, \xi\} \sqrt{\mu X_{\min}}},$$

and (28) suffices.
Figure 16: Correspondence between edges in $\mathcal{P}_{(k(0), k(1), \ldots, k(L))}$ and $\mathcal{P}_{\max}$. (Part 1)

Figure 17: Correspondence between edges in $\mathcal{P}_{(k(0), k(1), \ldots, k(L))}$ and $\mathcal{P}_{\max}$. (Part 2)

D Proof for Theorem 1: Final Convergence

Since we have proved the first part of Theorem 1 in Section C, we will use it as a fact, then prove the remaining part of Theorem 1.

D.1 Auxiliary Lemmas

First, we show that $S_+, S_-, S_{\text{dead}}$ are trapping regions.

**Lemma 12.** Consider any solution to the gradient flow dynamic (2), we have the following:

- If at some time $t_1 \geq 0$, we have $w_j(t_1) \in S_{\text{dead}}$, then $w_j(t_1 + \tau) \in S_{\text{dead}}$, $\forall \tau \geq 0$;
- If at some time $t_1 \geq 0$, we have $w_j(t_1) \in S_+$ for some $j \in \mathcal{V}_+$, then $w_j(t_1 + \tau) \in S_+$, $\forall \tau \geq 0$;
- If at some time $t_1 \geq 0$, we have $w_j(t_1) \in S_-$ for some $j \in \mathcal{V}_-$, then $w_j(t_1 + \tau) \in S_-$, $\forall \tau \geq 0$;

**Proof.** The first statement is simple, if $w_j \in S_{\text{dead}}$, then one have $\dot{w}_j \equiv 0$, thus $w_j$ remains in $S_{\text{dead}}$.

For the second statement, we have, since $j \in \mathcal{V}_+$,

$$
\frac{d}{dt} w_j = - \sum_{i=1}^{n} 1_{(x_i, w_j) > 0} \nabla y_i f(x_i; W, v) x_i \|w_j\|.
$$

When $\ell$ is the exponential loss, by the Fundamental Theorem of Calculus, one writes, $\forall \tau \geq 0$,

$$
w_j(t_1 + \tau) = w_j(t_1) + \int_0^\tau \frac{d}{dt} w_j d\tau
= w_j(t_1) + \int_0^\tau - \sum_{i=1}^{n} 1_{(x_i, w_j) > 0} \nabla y_i f(x_i; W, v) x_i \|w_j\| d\tau
= w_j(t_1) + \int_0^\tau \sum_{i=1}^{n} 1_{(x_i, w_j) > 0} y_i \exp(-y_i f(x_i; W, v)) x_i \|w_j\| d\tau
= w_j(t_1) + \sum_{i \in \mathcal{I}_+} \left( \int_0^\tau \exp(-y_i f(x_i; W, v)) \|w_j\| d\tau \right) x_i.
$$

Here $w_j(t_1) \in S_+$ by our assumption, $\tilde{x}_+ \in K \subseteq S_+$ because $\tilde{x}_+$ is a conical combination of $x_i, i \in \mathcal{I}_+$. Since $S_+$ is a convex cone, we have $w_j(t_1 + \tau) \in S_+$ as well.
When $\ell$ is the logistic loss, we have, similarly,
\[
    w_j(t_1 + \tau) = w_j(t_1) + \int_0^\tau \sum_{i=1}^n \mathbb{1}_{(x_i, w_j) > 0} y_i \frac{2 \exp(-y_i f(x_i; W, v))}{1 + \exp(-y_i f(x_i; W, v))} x_i \|w_j\| d\tau
\]
\[
    = w_j(t_1) + \sum_{i \in \mathcal{I}_+} \left( \int_0^\tau \frac{2 \exp(-y_i f(x_i; W, v))}{1 + \exp(-y_i f(x_i; W, v))} \|w_j\| d\tau \right) x_i \in \mathcal{S}_+.
\]

The proof of the third statement is almost identical (we only show the case of exponential loss here): when $j \in \mathcal{V}_-$, we have
\[
    \frac{d}{dt} w_j = \sum_{i=1}^n \mathbb{1}_{(x_i, w_j) > 0} \nabla y \ell(y_i, f(x_i; W, v)) x_i \|w_j\|
\]
and
\[
    w_j(t_1 + \tau) = w_j(t_1) + \sum_{i \in \mathcal{I}_-} \left( \int_0^\tau \exp(-y_i f(x_i; W, v)) \|w_j\| d\tau \right) x_i.
\]

Again, here $w_j(t_1) \in \mathcal{S}_-$ by our assumption. $\hat{x}_- \in -K \subseteq \mathcal{S}_-$ because $\hat{x}_-$ is a conical combination of $x_i, i \in \mathcal{I}_-$. Since $\mathcal{S}_-$ is a convex cone, we have $w_j(t_1 + \tau) \in \mathcal{S}_+$ as well. \qed

Then the following Lemma provides a lower bound on neuron norms upon $t_1$.

**Lemma 13.** Consider any solution to the gradient flow dynamic (2) starting from initialization (3). Let $t_1$ be the time when directional convergence is achieved, as defined in Theorem 7 and we define $\tilde{\mathcal{V}}_+ : \{j : w_j(t_1) \in \mathcal{S}_+ \}$ and $\tilde{\mathcal{V}}_- : \{j : w_j(t_1) \in \mathcal{S}_- \}$. If both $\tilde{\mathcal{V}}_+$ and $\tilde{\mathcal{V}}_-$ are non-empty, we have
\[
    \sum_{j \in \tilde{\mathcal{V}}_+} \|w_j(t_1)\|^2 \geq \exp(-4n X_{\max} t_1) \sum_{j \in \tilde{\mathcal{V}}_+} \|w_j(0)\|^2,
\]
\[
    \sum_{j \in \tilde{\mathcal{V}}_-} \|w_j(t_1)\|^2 \geq \exp(-4n X_{\max} t_1) \sum_{j \in \tilde{\mathcal{V}}_-} \|w_j(0)\|^2.
\]

**Proof.** We have shown that
\[
    \frac{d}{dt} \|w_j\|^2 = -2 \sum_{i=1}^n \mathbb{1}_{(x_i, w_j) > 0} \nabla y \ell(y_i, f(x_i; W, v)) \langle x_i, w_j \rangle \text{sign}(w_j(0)) \|w_j\|.
\]
Then before $t_1$, we have $\forall j \in [h]$
\[
    \frac{d}{dt} \|w_j\|^2 = -2 \sum_{i=1}^n \mathbb{1}_{(x_i, w_j) > 0} \nabla y \ell(y_i, f(x_i; W, v)) \langle x_i, w_j \rangle \text{sign}(w_j(0)) \|w_j\|
\]
\[
    \geq -2 \sum_{i=1}^n (\|y_i\| + 2 \max_i |f(x_i; W, v)|) \|x_i\| \|w_j\|^2
\]
\[
    \geq -4 \sum_{i=1}^n \|x_i\| \|w_j\|^2 \geq -4n X_{\max} \|w_j\|^2,
\]
where the second last inequality is because $\max_i |f(x_i; W, v)| \leq \frac{1}{2}$ before $t_1$. Summing over $j \in \tilde{\mathcal{V}}_+$, we have
\[
    \frac{d}{dt} \sum_{j \in \tilde{\mathcal{V}}_+} \|w_j\|^2 \geq -4n X_{\max} \sum_{j \in \tilde{\mathcal{V}}_+} \|w_j\|^2.
\]
Therefore, we have the following bound:
\[
    \sum_{j \in \tilde{\mathcal{V}}_+} \|w_j(t_1)\|^2 \geq \exp(-4n X_{\max} t_1) \sum_{j \in \tilde{\mathcal{V}}_+} \|w_j(0)\|^2.
\]
\qed
Moreover, after $t_1$, the neuron norms are non-decreasing, as suggested by

**Lemma 14.** Consider any solution to the gradient flow dynamic \( (2) \) starting from initialization \( (3) \). Let \( t_1 \) be the time when directional convergence is achieved, as defined in Theorem\( 4 \) and we define \( \tilde{V}_+ : \{ j : w_j(t_1) \in S_+ \} \) and \( \tilde{V}_- : \{ j : w_j(t_1) \in S_- \} \). If both \( \tilde{V}_+ \) and \( \tilde{V}_- \) are non-empty, we have \( \forall \tau \geq 0 \) and \( t_2 \geq t_1 \),

\[
\sum_{j \in \tilde{V}_+} \| w_j(t_2 + \tau) \|^2 \geq \sum_{j \in \tilde{V}_+} \| w_j(t_2) \|, \quad \sum_{j \in \tilde{V}_-} \| w_j(t_2 + \tau) \|^2 \geq \sum_{j \in \tilde{V}_-} \| w_j(t_2) \| \quad (30)
\]

**Proof.** It suffices to show that after \( t_1 \), the following derivatives:

\[
\frac{d}{dt} \sum_{j \in \tilde{V}_+} \| w_j(t) \|^2, \quad \frac{d}{dt} \sum_{j \in \tilde{V}_-} \| w_j(t) \|^2,
\]

are non-negative.

For \( j \in \tilde{V}_+ \), \( w_j \) stays in \( S_+ \) by Lemma\( 12 \) and we have

\[
\frac{d}{dt} \| w_j \|^2 = -2 \sum_{i \in I_+} \nabla_y \ell(y_i, f(x_i; W, v)) \langle x_i, w_j \rangle \| w_j \|.
\]

\[
= \begin{cases} 
2 \sum_{i \in I_+} y_i \exp(-y_i f(x_i; W, v)) \langle x_i, w_j \rangle \| w_j \| \quad (\ell \text{ is exponential}) \\
2 \sum_{i \in I_+} y_i \frac{2 \exp(-y_i f(x_i; W, v))}{1 + \exp(-y_i f(x_i; W, v))} \langle x_i, w_j \rangle \| w_j \| \quad (\ell \text{ is logistic})
\end{cases}
\]

\[
\geq 0.
\]

Summing over \( j \in \tilde{V}_+ \), we have \( \frac{d}{dt} \sum_{j \in \tilde{V}_+} \| w_j(t) \|^2 \geq 0 \). Similarly one has \( \frac{d}{dt} \sum_{j \in \tilde{V}_-} \| w_j(t) \|^2 \geq 0 \).

Finally, the following lemma is used for deriving the final convergence.

**Lemma 15.** Consider the following loss function

\[
\mathcal{L}_{lin}(W, v) = \sum_{i=1}^n \ell \left( y_i, v^T W^T x_i \right),
\]

if \( \{ x_i, y_i \}, i \in [n] \) are linearly separable, i.e., \( \exists \gamma > 0 \) and \( z \in \mathbb{S}^{D-1} \) such that \( y_i \langle z, x_i \rangle \geq \gamma, \forall i \in [n] \), then under the gradient flow on \( \mathcal{L}_{lin}(W, v) \), whenever \( y_i v^T W^T x_i \geq 0, \forall i \), we have

\[
\dot{\mathcal{L}}_{lin} \leq -\frac{1}{4} \| v \|^2 \mathcal{L}^2 \gamma^2. \quad (31)
\]

**Proof.** For \( \ell \) being exponential loss, we have:

\[
\dot{\mathcal{L}} = -\| \nabla_W \mathcal{L} \|_F^2 - \| \nabla_v \mathcal{L} \|_2^2 \leq -\| \nabla_W \mathcal{L} \|_F^2.
\]

\[
= -\| \sum_{i=1}^n y_i \ell(y_i, v^T W^T x_i) x_i v^T \|_F^2
\]

\[
= -\| v \|^2 \left( \sum_{i=1}^n y_i \ell(y_i, v^T W^T x_i) x_i \right)^2
\]

\[
\leq -\| v \|^2 \left( z, \sum_{i=1}^n y_i \ell(y_i, v^T W^T x_i) x_i \right)^2
\]

\[
\leq -\| v \|^2 \sum_{i=1}^n \ell(y_i, v^T W^T x_i) \gamma^2
\]

\[
\leq -\| v \|^2 L^2 \gamma^2 \leq -\frac{1}{4} \| v \|^2 \mathcal{L}^2 \gamma^2.
\]
For $\ell$ being logistic loss, we have:

$$
\hat{\mathcal{L}} = -\|\nabla_W \mathcal{L}\|_F^2 - \|\nabla_v \mathcal{L}\|_F^2 \\
= -\left\| \sum_{i=1}^n y_i \frac{2 \exp(-y_i v^T W x_i)}{1 + \exp(-y_i v^T W x_i)} x_i v^T \right\|_F^2
$$

We define $\mathcal{L} = \hat{\mathcal{L}}$ where \(\mathcal{L}\) implies, by Lemma 14, \(\forall \mathcal{L}\) where the last inequality uses the fact that \(2^{\frac{1}{1+z}} \geq \log(1 + z)\) when \(z \in [0, 1]\).

D.2 PROOF OF FINAL CONVERGENCE

Proof of Theorem 2 Second Part. By Lemma 12 we know that after $t_1$, neurons in $\mathcal{S}_+$ ($\mathcal{S}_-$) stays in $\mathcal{S}_+$ ($\mathcal{S}_-$). Thus the loss can be decomposed as

$$
\mathcal{L} = \sum_{i \in \mathcal{I}_+} \ell \left( y_i, \sum_{j \in \mathcal{V}_+} v_j \langle w_j, x_i \rangle \right) + \sum_{i \in \mathcal{I}_-} \ell \left( y_i, \sum_{j \in \mathcal{V}_-} v_j \langle w_j, x_i \rangle \right),
$$

where $\mathcal{V}_+ : \{ j : w_j(t_1) \in \mathcal{S}_+ \}$ and $\mathcal{V}_- : \{ j : w_j(t_1) \in \mathcal{S}_- \}$. Therefore, the training after $t_1$ is decoupled into 1) using neurons in $\mathcal{V}_+$ to fit positive data in $\mathcal{I}_+$ and 2) using neurons in $\mathcal{V}_-$ to fit positive data in $\mathcal{I}_-$. We define $f_+(x; W, v) = \sum_{j \in \mathcal{V}_+} v_j \langle w_j, x \rangle$ and let $t_2^+ = \inf \{ t : \max_{i \in \mathcal{I}_+} |f_+(x_i; W, v)| > \frac{1}{4} \}$. Similarly, we also define $f_-(x; W, v) = \sum_{j \in \mathcal{V}_-} v_j \langle w_j, x \rangle$ and let $t_2^- = \inf \{ t : \max_{i \in \mathcal{I}_-} |f_-(x_i; W, v)| > \frac{1}{4} \}$. Then $t_1 \leq \min\{ t_2^+, t_2^- \}$, by Lemma 3.

$O(1/t)$ convergence after $t_2$: We first show that when both $t_2^+, t_2^-$ are finite, then it implies $O(1/t)$ convergence on the loss. Then we show that they are indeed finite and $t_2 := \max\{ t_2^+, t_2^- \} = O(\frac{1}{n \log \frac{1}{\gamma}})$.

At $t_2 = \max\{ t_2^+, t_2^- \}$, by definition, $\exists i_+ \in \mathcal{I}_+$ such that

$$
\frac{1}{4} \leq f_+(x_{i_+}; W, v) \leq \sum_{j \in \mathcal{V}_+} v_j \langle w_j, x_{i_+} \rangle \leq \sum_{j \in \mathcal{V}_+} \| w_j \|^2 \| x_{i_+} \|,
$$

which implies, by Lemma 14 $\forall t \geq t_2$

$$
\sum_{j \in \mathcal{V}_+} \| w_j(t) \|^2 \geq \sum_{j \in \mathcal{V}_+} \| w_j(t_2) \|^2 \geq \frac{1}{4 \| x_{i_+} \|} \geq \frac{1}{4 X_{\text{max}}}.
$$
Similarly, we have ∀t ≥ t₂,
\[ \sum_{j \in V_+} \|w_j(t)\|^2 \geq \frac{1}{4X_{\max}}. \] (35)

Under the gradient flow dynamics (2), we apply Lemma 15 to the decomposed loss (32)
\[ 4\dot{\mathcal{L}} \leq - \left( \sum_{j \in V_+} \|w_j\|^2 \right) \cdot \mathcal{L}_+^2 \cdot (\mu_{X_{\min}})^2 - \left( \sum_{j \in V_+} \|w_j\|^2 \right) \cdot \mathcal{L}_-^2 \cdot (\mu_{X_{\min}})^2. \]

Here, we can pick the same γ = μ_{X_{\min}} for both 𝖀_+ and 𝖀_- because \{x_i, y_i\}, i ∈ I_+ is linearly separable with \( z = \frac{w_j}{\|z\|}; \langle z, x_iy_i \rangle \geq \mu \|x_i\| \geq \mu_{X_{\min}} \) by Assumption 1. And similarly, \{x_i, y_i\}, i ∈ I_- is linearly separable with \( \langle z, x_iy_i \rangle \geq \mu \|x_i\| \geq \mu_{X_{\min}} \). Replace \( v_t^* \) by \( \|w_j\|^2 \) from balancedness, together with (33), (35), we have
\[ 4\dot{\mathcal{L}} \leq - \left( \sum_{j \in V_+} \|w_j\|^2 \right) \cdot \mathcal{L}_+^2 \cdot (\mu_{X_{\min}})^2 - \left( \sum_{j \in V_+} \|w_j\|^2 \right) \cdot \mathcal{L}_-^2 \cdot (\mu_{X_{\min}})^2 \]
\[ \leq - \frac{(\mu_{X_{\min}})^2}{4X_{\max}} (\mathcal{L}_+^2 + \mathcal{L}_-^2) \leq - \frac{(\mu_{X_{\min}})^2}{8X_{\max}} (\mathcal{L}_+ + \mathcal{L}_-)^2 \]
which is
\[ \frac{1}{\mathcal{L}^2} \dot{\mathcal{L}} \leq \frac{(\mu_{X_{\min}})^2}{32X_{\max}}. \]
Integrating both side from \( t_2 \) to any \( t \geq t_2 \), we have
\[ \frac{1}{\mathcal{L}} \bigg| \bigg|_{t_2}^{t} \leq - \frac{(\mu_{X_{\min}})^2}{32X_{\max}} (t - t_2), \]
which leads to
\[ \mathcal{L}(t) \leq \frac{\mathcal{L}(t_2)}{\mathcal{L}(t_2) \alpha(t - t_2) + 1} , \text{ where } \alpha = \frac{(\mu_{X_{\min}})^2}{32X_{\max}}. \]

**Showing** \( t_2 = \mathcal{O}(\frac{1}{n} \log \frac{1}{\epsilon}) \): The remaining thing is to show \( t_2 \) is \( \mathcal{O}(\frac{1}{n} \log \frac{1}{\epsilon}) \).

Since after \( t_1 \), the gradient dynamics are fully decoupled into two gradient flow dynamics (on 𝖀_+ and on 𝖀_-), it suffices to show \( t_2 = \mathcal{O}(\frac{1}{n} \log \frac{1}{\epsilon}) \) and \( t_2 = \mathcal{O}(\frac{1}{n} \log \frac{1}{\epsilon}) \) separately, then combine them to show \( t_2 = \max\{t_2^1, t_2^2\} = \mathcal{O}(\frac{1}{n} \log \frac{1}{\epsilon}) \). The proof is almost identical for 𝖀_+ and 𝖀_-; thus we only prove \( t_2^1 = \mathcal{O}(\frac{1}{n} \log \frac{1}{\epsilon}) \) here.

Suppose
\[ t_2 \geq t_1 + \frac{6}{\sqrt{\|x_{\min}\|X_{\min}}} + \frac{4}{\sqrt{\|x_{\min}\|X_{\min}}} \left( \log \frac{2}{\epsilon^2 \sqrt{\|x_{\min}\|W_{\min}^2}} + \log X_{\max} t_1 \right) , \] (36)
where \( n_+ = |I_+| \). It takes two steps to show a contradiction: First, we show that for some \( t_a \geq 0 \), a refined alignment \( \cos(w_j(t_1 + t_a), x_+) \geq \frac{1}{4}, \forall j \in \tilde{V}_+ \) is achieved, and such refined alignment is maintained until at least \( t_a^2 \geq \frac{1}{4}, \forall j \in \tilde{V}_+ \) for all \( t_1 + t_a \leq t \leq t_a^2 \). Then, keeping this refined alignment leads to a contradiction.

- For \( j \in \tilde{V}_+ \), we have
\[ \frac{d}{dt} \frac{w_j}{\|w_j\|} = \left( I - \frac{w_j w_j^\top}{\|w_j\|^2} \right) \left( \sum_{i \in I_+} -\nabla_{\hat{y}_i} \ell(y_i, f_+(x_i; W, w)) x_i \right) := \hat{x}_a. \]

Then
\[ \frac{d}{dt} \cos(x_+, w_j) = (\cos(x_+, \hat{x}_a) - \cos(x_+, w_j) \cos(\hat{x}_a, w_j)) \|\hat{x}_a\| \]
\[ \geq (\cos(x_+, \hat{x}_a) - \cos(x_+, w_j) \|\hat{x}_a\|. \]
We can show that $\cos(x_+, \hat{x}_a) \geq \frac{1}{3}$ and $\|\hat{x}_a\| \geq \sqrt{\mu n + X_{\min}}/2$ when $t_1 \leq t \leq t_2^+$ (we defer the proof to the end as it breaks the flow), thus within $[t_1, t_2^+]$, we have

$$\frac{d}{dt} \cos(x_+, w_j) \geq \left(\frac{1}{3} - \cos(x_+, w_j)\right) \sqrt{\mu n + X_{\min}}/2.$$  
(37)

We use (37) in two ways: First, since

$$\cos(x_+, w_j) \geq \frac{1}{4}$$

is a trapping region for $w_j$ during $[t_1, t_2^+]$. Define $t_a := \inf\{t \geq t_1 : \min_{j \in \hat{V}_+} \cos(x_+, w_j(t)) \geq \frac{1}{4}\}$, then clearly, if $t_a \leq t_2^+$, then $\cos(w_j(t), x_+) \geq \frac{1}{4}, \forall j \in \hat{V}_+$ for all $t_1 + t_a \leq t \leq t_2^+$.

Now we use (37) again to show that $t_a \leq t_1 + \frac{6}{\sqrt{\mu n + X_{\min}}}$: Suppose that $t_a \geq t_1 + \frac{6}{\sqrt{\mu n + X_{\min}}}$, then $\exists j^*$ such that $\cos(x_+, w_{j^*}(t)) \leq \frac{1}{4}$, $\forall t \in [t_1 + t_a, t_1 + \frac{6}{\sqrt{\mu n + X_{\min}}}]$, and we have

$$\frac{d}{dt} \cos(x_+, w_{j^*}) \geq \left(\frac{1}{3} - \cos(x_+, w_{j^*})\right) \sqrt{\mu n + X_{\min}}/2 \geq \frac{\sqrt{\mu n + X_{\min}}}{24}.$$  
(38)

This shows

$$\cos(x_+, w_{j^*}(t_1 + 1)) \geq \cos(x_+, w_{j^*}(t_1)) + \frac{1}{4} \geq \frac{1}{4},$$

which contradicts that $\cos(x_+, w_{j^*}(t)) < \frac{1}{4}$. Hence we know $t_a \leq t_1 + \frac{6}{\sqrt{\mu n + X_{\min}}}$.

In summary, we have $\cos(w_j(t), x_+) \geq \frac{1}{4}, \forall j \in \hat{V}_+$ for all $t_1 + \frac{6}{\sqrt{\mu n + X_{\min}}} \leq t \leq t_2^+$.

- Now check the dynamics of $\sum_{j \in \hat{V}_+} \|w_j(t)\|^2$ during $t_1 + \frac{6}{\sqrt{\mu n + X_{\min}}} \leq t \leq t_2^+$. For simplicity, we denote $t_1 + \frac{6}{\sqrt{\mu n + X_{\min}}} := t_1^\prime$.

For $j \in \hat{V}_+$, we have, for $t_1^\prime \leq t \leq t_2^+$,

$$\frac{d}{dt} \|w_j\|^2 = 2 \sum_{i \in I_+} \nabla_i \ell(y_i, f(x_i; W, v)) \langle x_i, w_j \rangle \|w_j\|$$

$$\geq \sum_{i \in I_+} \langle x_i, w_j \rangle \|w_j\|$$

(by (40))

$$= \langle x_+, w_j \rangle \|w_j\|$$

$$= \|x_+\| \|w_j\|^2 \cos(x_+, w_j)$$

$$\geq \frac{1}{4} \|x_+\| \|w_j\|^2$$

(Since $t \geq t_1^\prime$)

$$\geq \frac{\sqrt{\mu n + X_{\min}}}{4} \|w_j\|^2,$$

(by Lemma 11)

which leads to (summing over $j \in \hat{V}_+$)

$$\frac{d}{dt} \sum_{j \in \hat{V}_+} \|w_j\|^2 \geq \frac{\sqrt{\mu n + X_{\min}}}{4} \sum_{j \in \hat{V}_+} \|w_j\|^2.$$

33
By Gronwall’s inequality, we have

\[
\sum_{j \in \tilde{V}_+} \|w_j(t_2^+)^2 \geq \exp \left( \frac{\sqrt{\mu_n} n_{\max}}{4} (t_2^* - t'_1) \right) \sum_{j \in \tilde{V}_+} \|w_j(t'_1)\|^2 \geq \exp \left( \frac{\sqrt{\mu_n} n_{\max}}{4} (t_2^* - t'_1) \right) \sum_{j \in \tilde{V}_+} \|w_j(t_1)\|^2 \geq \exp \left( \frac{\sqrt{\mu_n} n_{\max}}{4} (t_2^* - t'_1) \right) \exp (-4n X_{\max} t_1) \sum_{j \in \tilde{V}_+} \|w_j(0)\|^2 \geq \frac{2}{\sqrt{n} X_{\min}}.
\]

(By Lemma 14)

However, at \(t_2^*\), we have

\[
\frac{1}{4} \geq \frac{1}{n_+} \sum_{i \in \tilde{I}_+} f_+(x_i; W, v) = \frac{1}{n_+} \sum_{i \in \tilde{I}_+} \sum_{j \in \tilde{V}_+} v_j \langle w_j, x_i \rangle = \frac{1}{n_+} \sum_{j \in \tilde{V}_+} \|w_j\|^2 \cos(w_j, x_+) \|x_+\| \geq \frac{1}{4n} \sum_{j \in \tilde{V}_+} \|w_j\|^2 \|x_+\| \geq \frac{1}{4} \sum_{j \in \tilde{V}_+} \|w_j\|^2 \sqrt{n} X_{\min} \text{, (by Lemma 11)}
\]

which suggests \(\sum_{j \in \tilde{V}_+} \|w_j\|^2 \leq \frac{1}{\sqrt{n} X_{\min}}\). A contradiction.

Therefore, we must have

\[
t_2^* \leq t_1 + \frac{6}{\sqrt{\mu_n} n_{\max}} + \frac{4}{\sqrt{\mu_n} n_{\max}} \left( \log \frac{2}{\epsilon^2 \sqrt{n} X_{\min} W_{\min}^2} + 4n X_{\max} t_1 \right) \text{. (39)}
\]

Since the dominant term here is \(\frac{4}{\sqrt{\mu_n} n_{\max}} \log \frac{2}{\epsilon^2 \sqrt{n} X_{\min} W_{\min}^2}\), we have \(t_2^* = O(\frac{1}{n} \log \frac{1}{\epsilon})\). A similar analysis shows \(t_2 = O(\frac{1}{n} \log \frac{1}{\epsilon})\).

Complete the missing pieces We have two claims remaining to be proved. The first is \(\cos(x_+, \tilde{x}_a) \geq \frac{1}{2}\) when \(t_1 \leq t \leq t_2^*\). Since \(x_+ = \sum_{i \in \tilde{I}_+} x_i\) and \(\tilde{x}_a = \sum_{i \in \tilde{I}_+} -\nabla g(y_i, f_+(x_i; W, v)) x_i\). We simply use the fact that before \(t_2^*\), we have, by Lemma 2

\[
\frac{1}{2} \leq -\nabla g(y_i, f_+(x_i; W, v)) = \leq \frac{3}{2} \text{. (40)}
\]
to show the following
\[
\cos(x_+, \tilde{x}_a) = \frac{\langle x_+, \tilde{x}_a \rangle}{\| x_+ \| \| \tilde{x}_a \|} = \frac{\sum_{i,j \in \mathcal{I}_+} (-\nabla \phi(y_i, f_+(x_i; W, v))) \langle x_i, x_j \rangle}{\sqrt{\sum_{i,j \in \mathcal{I}_+} (x_i, x_j) \sqrt{\sum_{i,j \in \mathcal{I}_+} (-\nabla \phi(y_i, f_+(x_i; W, v)))^2 (x_i, x_j)}} \geq \frac{\frac{1}{2} \sum_{i,j \in \mathcal{I}_+} (x_i, x_j)}{\sqrt{\sum_{i,j \in \mathcal{I}_+} (x_i, x_j) \sqrt{\sum_{i,j \in \mathcal{I}_+} (\frac{3}{2})^2 (x_i, x_j)}} \geq \frac{1}{\sqrt{3}},
\]

since all \( \langle x_i, x_j \rangle, i, j \in \mathcal{I}_+ \) are non-negative.

The second claim is \( \| \tilde{x}_a \| \geq \sqrt{\mu n + X_{\text{min}}}/2 \) is due to that
\[
\| \tilde{x}_a \| = \sqrt{\sum_{i,j \in \mathcal{I}_+} (-\nabla \phi(y_i, f_+(x_i; W, v)))^2 (x_i, x_j)} \geq \frac{1}{2} \sqrt{\sum_{i,j \in \mathcal{I}_+} (x_i, x_j)} = \frac{\| x_+ \|}{2} \geq \frac{\sqrt{\mu n + X_{\text{min}}}}{2},
\]

where the last inequality is from Lemma [11].

\[ \square \]

D.3 Proof of Low-rank Bias

So far we have proved the directional convergence at the early alignment phase and final \( O(1/t) \) convergence of the loss in the later stage. The only thing that remains to be shown is the low-rank bias. The proof is quite straightforward but we need some additional notations.

As we proved above, after \( t_1 \), neurons in \( \mathcal{S}_+ (\mathcal{S}_-) \) stays in \( \mathcal{S}_+ (\mathcal{S}_-) \). Thus the loss can be decomposed as
\[
\mathcal{L} = \sum_{i \in \mathcal{I}_+} \ell \left( y_i, \sum_{j \in \mathcal{V}_+} v_j \langle w_j, x_i \rangle \right) + \sum_{i \in \mathcal{I}_-} \ell \left( y_i, \sum_{j \in \mathcal{V}_-} v_j \langle w_j, x_i \rangle \right),
\]

where \( \mathcal{V}_+ : \{ j : w_j(t_1) \in \mathcal{S}_+ \} \) and \( \mathcal{V}_- : \{ j : w_j(t_1) \in \mathcal{S}_- \} \). Therefore, the training after \( t_1 \) is decoupled into 1) using neurons in \( \hat{\mathcal{V}}_+ \) to fit positive data in \( \mathcal{I}_+ \) and 2) using neurons in \( \hat{\mathcal{V}}_- \) to fit positive data in \( \mathcal{I}_- \). We use
\[
W_+ = [W]_{\mathcal{V}_+}, \quad W_- = [W]_{\mathcal{V}_-}
\]
to denote submatrices of \( W \) by picking only columns in \( \hat{\mathcal{V}}_+ \) and \( \hat{\mathcal{V}}_- \), respectively. Similarly, we define
\[
v_+ = [v]_{\hat{\mathcal{V}}_+}, \quad v_- = [v]_{\hat{\mathcal{V}}_-}
\]
for the second layer weight \( v \). Lastly, we also define
\[
W_{\text{dead}} = [W]_{\mathcal{V}_- \setminus \mathcal{V}_+}, v_{\text{dead}} = [v]_{\hat{\mathcal{V}}_{\text{dead}}},
\]
where \( \hat{\mathcal{V}}_{\text{dead}} := \{ j : w_j(t_1) \in \mathcal{S}_{\text{dead}} \} \). Given these notations, after \( t_1 \) the loss is decomposed as
\[
\mathcal{L} = \sum_{i \in \mathcal{I}_+} \ell \left( y_i, x_i^T W_+ v_+ \right) + \sum_{i \in \mathcal{I}_-} \ell \left( y_i, x_i^T W_- v_- \right),
\]

and the GF on \( \mathcal{L} \) is equivalent to GF on \( \mathcal{L}_+ \) and \( \mathcal{L}_- \) separately. It suffices to study one of them. For GF on \( \mathcal{L}_+ \), we have the following important invariance [Arora et al., 2018b]
\[
W_+^T(t) W_+(t) - v_+(t) v_+^T(t) = W_+^T(t_1) W_+(t_1) - v_+(t_1) v_+^T(t_1),
\]

35
from which one has
\[
\|W^+_t(t)W_+(t) - v_+(t)v^+_t(t)\|_2 = \|W^+_t(t_1)W_+(t_1) - v_+(t_1)v^+_t(t_1)\|_2 \\
\leq \|W^+_t(t_1)W_+(t_1)\|_2 - \|v_+(t_1)v^+_t(t_1)\|_2 \\
\leq \text{tr}(W^+_t(t_1)W_+(t_1)) + \|v_+(t_1)\|^2 \\
= 2 \sum_{j \in \mathcal{V}_+} \|w_j(t_1)\|^2 \leq \frac{4\epsilon W_{\text{max}}^2}{\sqrt{\lambda}} |\hat{\mathcal{V}}_+|,
\]
where the last inequality is by Lemma 3. Then one can immediately get
\[
\|v_+(t)v^+_t(t)\|_2 - \|W^+_t(t)W_+(t)\|_2 \leq \|W^+_t(t)W_+(t) - v_+(t)v^+_t(t)\|_2 \leq \frac{4\epsilon W_{\text{max}}^2}{\sqrt{\lambda}} |\hat{\mathcal{V}}_+|,
\]
which is precisely
\[
\|W_+(t)\|_F^2 \leq \|W_+(t)\|_2^2 + \frac{4\epsilon W_{\text{max}}^2}{\sqrt{\lambda}} |\hat{\mathcal{V}}_+| . \tag{41}
\]
Similarly, we have
\[
\|W_-(t)\|_F^2 \leq \|W_-(t)\|_2^2 + \frac{4\epsilon W_{\text{max}}^2}{\sqrt{\lambda}} |\hat{\mathcal{V}}_-| . \tag{42}
\]
Lastly, one has
\[
\|W_{\text{dead}}\|_F^2 = \sum_{j \in \mathcal{V}_{\text{dead}}} \|w_j(t_1)\|^2 \leq \frac{4\epsilon W_{\text{max}}^2}{\sqrt{\lambda}} |\hat{\mathcal{V}}_{\text{dead}}| \tag{43}
\]
Adding (41), (42), (43) together, we have
\[
\|W(t)\|_F^2 = \|W_+(t)\|_F^2 + \|W_-(t)\|_F^2 + \|W_{\text{dead}}\|_F^2 \\
\leq \|W_+(t)\|_2^2 + \|W_-(t)\|_2^2 + \frac{4\sqrt{\lambda} \epsilon W_{\text{max}}^2}{\sqrt{\lambda}} \leq 2\|W(t)\|_2^2 + 4\sqrt{\lambda} \epsilon W_{\text{max}}^2 .
\]
Finally, since we have shown $\mathcal{L} \to 0$ as $t \to \infty$, then $\forall i \in [n]$, we have $\ell(y_i, f(x_i; W, v)) \to 0$. This implies
\[
f(x_i; W, v) = -\frac{1}{y_i} \log \ell(y_i, f(x_i; W, v)) \to \infty .
\]
Because we have shown that
\[
f(x_i; W, v) \leq \sum_{j \in [h]} \|w_j\|^2 |x_i| \leq \|W\|_F^2 X_{\text{max}},
\]
f($x_i; W, v$) $\to \infty$ enforces $\|W\|_F^2 \to \infty$ as $t \to \infty$, thus $\|W\|_2^2 \to \infty$ as well. This gets us
\[
\lim_{t \to \infty} \|W\|_2^2 = 2 .
\]
E  EXISTENCE OF CARATHEODORY SOLUTION UNDER FIXED SUBGRADIENT

$\sigma'(x) = 1_{x>0}$

In this Appendix, we first introduce the notion of solution we are interested in for the GF (2): Caratheodory solutions that satisfy (2) for almost all time $t$. Next, in Appendix E.2, we show that if we fix the ReLU subgradient as $\sigma'(x) = 1_{x>0}$, then global Caratheodory solutions exists for (2) under Assumption[1]. Finally, we use simple examples to illustrate two points: 1) Caratheodory solutions cease to exist when ReLU subgradient at zero is chosen to be a fixed non-zero value, highlighting the importance of choosing the right subgradient for analysis; 2) Caratheodory solutions are potentially non-unique, the neurons’ dynamical behavior could become somewhat irregular if certain solutions are not excluded, justifying the introduction of regular solutions (Definition 1).

E.1  CARATHEODORY SOLUTIONS

Given an differential equation

$$\dot{\theta} = F(\theta), \theta(0) = \theta_0,$$ (44)

with $F$ potentially be discontinuous, $\theta(t)$ is said to be a Caratheodory solution of (44) if it satisfies the following integral equation

$$\theta(t) = \theta_0 + \int_0^t F(\theta(\tau))d\tau,$$ (45)

for all $t \in [0, a)$, where $a \in \mathbb{R}_{>0} \cup \infty$. In this section, we are interested in global Caratheodory solutions: $\theta(t)$ that satisfies (45) for all time $t \geq 0$.

E.2  PROOF OF EXISTENCE OF REGULAR CARATHEODORY SOLUTIONS UNDER ASSUMPTION[1]

In this section, we show the existence of global regular (Definition[1]) Caratheodory solutions to $\dot{\theta} = F(\theta), \theta(0) = \theta_0$, where $\theta := \{W, v\}$ and $F := \nabla_{W,v}\mathcal{L}$ defined from a fixed choice of ReLU subgradient $\sigma'(x) = 1_{x>0}$, under Assumption[1]. For the sake of a clear presentation, we first discuss the case of $S_{\text{dead}} = \emptyset$, where all solutions are regular, then discuss the modifications one needs to make when $S_{\text{dead}} \neq \emptyset$.

Existence of Caratheodory solutions when $S_{\text{dead}} = \emptyset$: First of all, notice that $\nabla_{W,v}\mathcal{L}$ is continuous almost everywhere except for a zero measure set $\mathcal{A} = \{W, v: \exists i \in [n], j \in [h] s.t. \langle x_i, w_j \rangle = 0\}$, since discontinuity only happens when one has to evaluate $\sigma'(\langle x_i, w_j \rangle)$ at $\langle x_i, w_j \rangle = 0$ for some $i, j$. Being a finite union of hyperplanes, $\mathcal{A}$ has zero measure.

For points outside $\mathcal{A}$, the existence of a local solution is guaranteed by the generalized Caratheodory existence theorem in Persson (1975). We refer readers to Appendix E.5 for the construction of such a local solution. The local solution can be extended to a global solution, as long as it does not encounter any point in $\mathcal{A}$ (the set where the flow is discontinuous). Whenever a point in $\mathcal{A}$ is reached, one requires extra certificates to extend the solution beyond that point. Simply speaking, the existence of a local solution around a point in $\mathcal{A}$ requires that the flow around this point does not push trajectories towards $\mathcal{A}$ from both sides of the zero measure set, causing an infinite number of crossings of $\mathcal{A}$, called Zeno behavior [van der Schaft & Schumacher, 2000; Maennel et al., 2018]. See Figure 18 and 19 for an illustration. In Appendix E.5, we formally show that if there is no Zeno behavior, then a solution can be extended until reaching discontinuity in $\mathcal{A}$, and gets extended by leaving $\mathcal{A}$ immediately.[1]

One sufficient condition for avoiding Zeno behavior is to show: For each hyperplane $\mathcal{A}_{ij} := \{\langle x_i, w_j \rangle = 0\}$, all points in a neighborhood around this hyperplane $\mathcal{A}_{ij}$ must satisfy that the inner products between the normal vector of $\mathcal{A}_{ij}$ and the flow $F$ have the same sign. Formally speaking, we need that there exists $\delta > 0$, such that for all pair of $\theta_k, \theta_l \in \{\theta = (W, v): 0 < |\langle x_i, w_j \rangle| < \delta\}$, we have $\langle N_{\mathcal{A}_{ij}}, F(\theta_k) \rangle \langle N_{\mathcal{A}_{ij}}, F(\theta_l) \rangle > 0$, here $N_{\mathcal{A}_{ij}}$ should be a fixed choice of the normal vector of hyperplane $\mathcal{A}_{ij}$.

[1] Strictly speaking, Appendix E.5 is part of the proof but discussing the technical part right now disrupts the presentation.
Figure 18: Non-existence of Caratheodory solution around points of discontinuity (Does not happen under Assumption 1). Since flow $F(\theta)$ is continuous, the solution $\theta(t)$ can be extended until it reaches the points of discontinuity in $A$, after which the solution is forced to stay within $A$ (often referred as Zeno behavior: the solution is crossing $A$ infinitely many times), and Caratheodory solution ceases to exist.

Figure 19: Existence of Caratheodory solution around points of discontinuity (Guaranteed by Assumption 1). When the solution $\theta(t)$ reaches $A$, it immediately leaves $A$ since the flow on the opposite side is flowing outward. This is a valid Caratheodory solution.

This inner product $\langle N_{A_i}, F(\theta_k) \rangle$ between the normal vector and the flow is exactly computed as $\langle x, \nabla w, \mathcal{L} \rangle$. Under Assumption 1, we have a much stronger result than what is required in the last paragraph: we can show that on the entire parameter space, we have (shown in Appendix E.5)

$$y_i \text{sign}(v_j) \langle x, \nabla w, \mathcal{L} \rangle > 0,$$

As such, since $v_j(t)$ does not change sign, Assumption 1 prevents Zeno behavior and ensures the existence of local solution around points in $A$.

In summary, from any initialization, the Caratheodory solution can be extended until the trajectory encounters points of discontinuity in $A$, then the existence of a local solution is guaranteed by ensuring that the flow forces the solution to leave $A$ immediately. Moreover, (46) ensures that $A$ can only be crossed a finite number of times (every hyperplane can only be crossed once), after which no discontinuity is encountered and the solution can be extended to $t = \infty$. Therefore a global Caratheodory solution always exists.

**Existence of Caratheodory solutions when $S_{\text{dead}} \neq \emptyset$:** Notice that when $S_{\text{dead}} \neq \emptyset$, $A$ contains boundary of $S_{\text{dead}}$. If the solution gets extended to $A$ where one neuron lands on the boundary of $S_{\text{dead}}$, then this neuron stays at the boundary of $S_{\text{dead}}$, i.e. the solution stays at $A$. Therefore, the previous argument about existence does not apply.

However, one only needs very a minor modification: If at time $t_0$, the solution enters $A$ by having one neuron (say $w_j(t)$) land on the boundary of $S_{\text{dead}}$, set $w_j(t) \equiv w_j(t_0)$ and $v_j(t) \equiv v_j(t_0)$ for $t \geq t_0$, then exclude $\{w_j, v_j\}$ from the parameter space and continue constructing and extending local solutions for other parameters via the previous argument. This shows the existence of the Caratheodory solution under non-empty $S_{\text{dead}}$, and by our construction, the solution is regular.

**E.3 Non-existence of Caratheodory solution under other fixed subgradient**

Consider the following simple example: The training data consists of a single data point $x = [1, 0]^T$, $y = -1$, and the network consists of a single neuron $\{w, v\}$ initialized at $\{w(0) = [0, 1]^T, v(0) = 1\}$. See Figure 20 for an illustration.

When the ReLU subgradient is chosen to be $\sigma'(x) = 1_{x > 0}$, the Caratheodory solution $\{w(t) \equiv [0, 1], v(t) \equiv 1\}$ exists, i.e. the neuron stays at the boundary of $S_{\text{dead}} := \{w : \langle x, w \rangle \leq 0\}$.

If the ReLU subgradient is chosen to be $\sigma'(0) = a > 0$, then the Caratheodory solution ceases to exist: the neuron cannot stay at the boundary $\langle x, w \rangle = 0$ of $S_{\text{dead}}$, because the non-zero $\sigma'(0)$ pushes
We consider the case when the ReLU subgradient $\sigma'(0) = a > 0$. For any $t > 0$, $w(t)$ cannot stay at $[0, 1]^\top$, because the subgradient $\sigma'(0) = a > 0$ is positive, leading to non-zero $\dot{w}$. At the same time, it cannot enter the interior of $S_{\text{dead}}$, because there is no flow in the interior. It towards the interior of $S_{\text{dead}}$. However, the neuron cannot enter the interior of $S_{\text{dead}}$ because the flow is all zero within the interior of $S_{\text{dead}}$.

To see this formally, suppose $\{w(t) = w(0), v(t) = v(0)\}$ for $t \in [0, t_0]$ (neuron stay at $\{w(0), v(0)\}$), then by definition of Caratheodory solution, we have

$$\int_0^{t_0} \nabla_{w,v} L(w(0), v(0)) dt = t_0 \nabla_{w,v} L(w(0), v(0)) = 0,$$

suggesting $\nabla_{w,v} L(w(0), v(0)) = 0$, which is not true when $\sigma'(0) > 0$, thus a contradiction. Now suppose $w(t_0) \in \text{Int}(S_{\text{dead}})$ for some $t_0$, then it must be that $w(t) \in \text{Int}(S_{\text{dead}}), \forall 0 < t \leq t_0$, otherwise it leads to the same contradiction as in previous paragraph. By definition of Caratheodory solution, we have

$$\int_0^{t_0} \nabla_{w,v} L(w(t), v(t)) dt = w(t_0) - w(0).$$

The left-hand side is zero because $w(t) \in \text{Int}(S_{\text{dead}}) \Rightarrow \nabla_{w,v} L(w(t), v(t)) = 0, \forall 0 < t \leq t_0$. The right-hand side is non-zero because $w(t_0) \in \text{Int}(S_{\text{dead}})$, thus a contradiction. Similarly, $w(t)$ cannot enter $S_{\text{dead}}^\text{int}$. Therefore, the Caratheodory solution $\{w(t), v(t)\}$ does not exist for any $t > 0$.

### E.4 Non-uniqueness of Caratheodory solutions

Consider the following simple example: The training data consists of a single data point $x = [1, 0]^\top$, $y = 1$, and the network consists of a single neuron $(w, v)$ initialized at $w(0) = [0, 1]^\top, v(0) = 1$. See Figure 21 for an illustration.

We consider the case when the ReLU subgradient is chosen to be $\sigma'(x) = \mathbb{1}_{x > 0}$. There exists one Caratheodory solution $w(t) \equiv w(0), v(t) \equiv v(0)$, i.e. the neuron stays at the boundary of $S_{\text{dead}} := \{w : \langle x, w \rangle \leq 0\}$. However, consider $\tilde{w}(t), \tilde{v}(t)$ being the solution to the following ode (the one that neuron follows once enters the positive orthant):

$$\dot{\tilde{w}} = y \exp(-y \tilde{v}(x, \tilde{w})) \tilde{v}, \dot{\tilde{v}} = y \exp(-y \tilde{v}(x, \tilde{w})) \langle x, \tilde{w} \rangle, \tilde{w}(0) = w(0), \tilde{v}(0) = v(0). \quad (47)$$

Then for any $t_0 \geq 0$,

$$w(t) = \mathbb{1}_{t < t_0} w(0) + \mathbb{1}_{t \geq t_0} \tilde{w}(t - t_0), v(t) = \mathbb{1}_{t < t_0} v(0) + \mathbb{1}_{t \geq t_0} \tilde{v}(t - t_0)$$

is a Caratheodory solution. This example shows that the Caratheodory solution could be non-unique. This is somewhat troublesome for our analysis, one would like that all neurons in $S_{\text{dead}}$ stay within $S_{\text{dead}}$, but Caratheodory solutions do not have this property, and in fact, as long as the neuron is on
the boundary of $S_{\text{dead}}$, and the flow outside $S_{\text{dead}}$ is pointing away from the boundary, the neuron can leave $S_{\text{dead}}$ at any time and it does not violate the definition of a Caratheodory solution. Therefore, for our main theorem, we added an additional regularity condition (Definition 1) on the solution, forcing neurons to stay within $S_{\text{dead}}$.

**Remark 6.** This issue of having irregular solutions is not specific to our choice of the notion of solutions. Even if one considers more generally the Filippov solution Filippov (1971) of the differential inclusion in (2), the same issue of non-uniqueness persists and needs attention when analyzing neuron dynamics.

**Remark 7.** Although irregular solutions are not desired for analyzing neuron behaviors, as we see in this example, they are rare cases under very specific initialization of the neurons and thus can be avoided by randomly initializing the weights.

### E.5 Constructing Global Caratheodory Solution

In this section, we formally show that if there is no Zeno behavior, then a solution can be extended until reaching discontinuity in $\mathcal{A}$, and gets extended by leaving $\mathcal{A}$ immediately, leading to a construction of global Caratheodory solution. The only ingredient that is needed is the existence theorem in Persson (1975) Theorem 2.3, showing that if $F(\theta)$ is continuous and $\forall \theta < 1$

$$\|F(\theta)\|_F \leq M(1 + \|\theta\|_F),$$

for some $M > 0$, then global solution of $\dot{\theta} = F(\theta)$ exists. Obviously, this result cannot be applied directly for two reasons: a) it requires continuity of the flow; b) it requires linear growth of $\|F(\theta)\|_F$ w.r.t. $\|\theta\|_F$. The key idea is constructing a local solution by restricting the flow to a neighborhood of initial conditions where a) and b) are satisfied, and then extending this solution to a global one.

As we discussed in Appendix E.2, we can assume $S_{\text{dead}} = \emptyset$ without loss of generality. Moreover, it suffices to show that starting from an initialization $\theta(0) = \{W(0), v(0)\}$ outside $\mathcal{A}$ we can construct either: 1) a global solution without encountering any point in $\mathcal{A}$; or 2) a local solution that lands on $\mathcal{A}$ at some $t_0$ then leave $\mathcal{A}$ immediately. Because if 2) happens, we take the end of this local solution as a new initial condition and repeat this argument. Importantly, 2) cannot happen infinitely many times because we have shown in Appendix E.2 that $\mathcal{A}$ can only be crossed finitely many times, thus 1) must happen, resulting in a global solution.

**Construct local solution from initial condition:** Now given an initial condition $\theta(0) = \{W(0), v(0)\}$, define the following two sets (Notation-wise, we drop the dependency on $\{W(0), v(0)\}$ for simplicity):

$$\Theta_0 := \{\theta = (W, v) : L(W, v) \leq L(W(0), v(0)), \text{sign}(v_j) = \text{sign}(v_j(0)), \forall j \in [h]\},$$

$$\Theta_1 := \{\theta = (W, v) : \forall i \in [n], j \in [h], \langle x_i, w_j \rangle > 0 \},$$

$\Theta_1$ is the positive invariant set of $\{W(0), v(0)\}$; all solutions from $\{W(0), v(0)\}$ never leaves $\Theta_1$, so it suffices to study the flow within $\Theta_1$ for the existence of solutions. Moreover, $\Theta_0$ is the intersection of a closed set $\{v : \text{sign}(v_j) = \text{sign}(v_j(0))\}$ and the pre-image of a continuous function $L$ on the range $[0, L(W(0), v(0))]$ thus closed. $\Theta_2$ is the largest connected set that contains $\{W(0), v(0)\}$ without point of discontinuity.

Consider the following set

$$\tilde{\Theta}_1 := \Theta_0 \cap \text{cl}(\Theta_1).$$

Then $\tilde{\Theta}_1$ is closed. Consider a new flow $\tilde{F}^1$ on $\tilde{\Theta}_1$, such that $\tilde{F}^1 = F = \nabla_{W,v} L$ for all $\theta \in \text{Int}(\tilde{\Theta}_1)$, and $\tilde{F}^1(\theta) = \lim_{k \to \infty} F(\theta_k)$ for all $\theta \in \tilde{\Theta}_1 \setminus \text{Int}(\tilde{\Theta}_1)$, where $\theta_k \in \text{Int}(\tilde{\Theta}_1), k = 1, 2, \cdots$ is a convergent sequence to $\theta$.

$\tilde{F}^1_{\tilde{\Theta}_1}$ is continuous by construction, and we can show that (at the end of this section)

$$\|\tilde{F}^1(\theta)\|_F \leq C\|\theta\|_F, \forall \theta \in \tilde{\Theta}_1.$$  

(50)

By a generalized version of the Tietze extension theorem Ercan (1997), there exists continuous $\tilde{F}^1$ on the entire parameter space, such that

$$\tilde{F}^1(\theta) = \tilde{F}^1_1(\theta), \forall \theta \in \tilde{\Theta}_1.$$  

(51)

---

Under review as a conference paper at ICLR 2024
We construct $\tilde{F}_1(\theta) F \leq C\|\theta\|_F, \forall \theta$. 

Because now $F = F^cl = \hat{F}_1$ on $\text{Int}(\hat{\Theta}_1)$, any solution $\hat{\theta}(t)$ of $\hat{\theta}_1 = \hat{F}_1(\hat{\theta}_1), \hat{\theta}(0) = \theta(0)$ (existence guaranteed by Persson (1975)) gives a local solution of $\hat{\theta} = F(\theta), \theta(0) = \theta(0)$, for $t \leq t_0$, where $t_0 := \inf\{ t : \hat{\theta}_1(t) \notin \Theta_1 \}$.

If $t_0 = \infty$, one has a global solution and the construction is finished. If $t_0 < \infty$, it must be that $\hat{\theta}_1(t_0) \in A$ (since $\hat{\theta}_1$ must leave $\Theta$ via the boundary of $\Theta_2$). Now we need to construct a solution that leaves $A$ immediately.

**Construct local solution that leaves $A$:** As we discussed, $A$ is a union of hyperplanes. For simplicity, let us assume $\hat{\theta}(t_0)$ is not at the intersection of two hyperplanes (the treatment is similar but tedious, we will make remarks in the end).

Now $\hat{\theta}(t_0)$ lands on a single hyperplane, let it be $\{ \theta : \langle x_i^*, w_j^* \rangle = 0 \}$, we define

$$\Theta_2 := \{ \theta = (W, v) : \forall i \neq i^*, \forall j \neq j^*, \langle x_i, w_j \rangle \langle x_i, w_j(0) \rangle > 0, \langle x_i^*, w_j^* \rangle \langle x_i^*, w_j^*(0) \rangle < 0 \},$$

and we let

$$\hat{\Theta}_2 := \Theta_2 \cap \text{cl}(\Theta_2).$$

It is clear that, from the definition of $\Theta_2$, any solution we construct that leaves $A$ immediately after $t_0$ must enter $\text{Int}(\hat{\Theta}_2)$. To construct the solution, we just need to repeat the first part, but now for $\Theta_2$: We construct $F^2_2$ that is continuous on $\Theta_2$ and agrees with $F$ on the interior, then extends $F^2_2$ to $\tilde{F}$ on the entire parameter space. Consider the solution $\hat{\theta}_2(t)$ of

$$\hat{\theta}_2 = F(\hat{\theta}_2), \hat{\theta}_2(0) = \hat{\theta}_1(t_0),$$

which gives a local solution of

$$\hat{\theta} = F(\theta), \theta(0) = \hat{\theta}_1(t_0).$$

Because we have shown that Zeno behavior does not happen, $\hat{\theta}_2(t)$ leaves $A$ immediately and enters $\text{Int}(\hat{\Theta}_2)$. We just pick any $\tau_0 > 0$ such that $\hat{\theta}_2(\tau_0) \in \text{Int}(\hat{\Theta}_2)$ then

$$\hat{\theta}(t) = 1_{t \leq \tau_0} \hat{\theta}_1(t) + 1_{\tau_0 < t \leq \tau_0 + \tau_0} \hat{\theta}_2(t - t_0),$$

is a Caratheodory solution to $\hat{\theta} = F(\theta), \theta(0) = \theta(0)$ for $t \leq t_0 + \tau_0$. This is exactly what we intended to show.

**Remark 8.** When $\hat{\theta}(t_0)$ lands at the intersection of two (or more) hyperplanes, the only difference is that now there could be more regions to escape to. But under Assumption (7), (66) suggests that the solution must cross all hyperplanes after $t_0$, leaving one unique region similar to $\Theta$. Then one constructs the local solution following previous procedures.

**Complete the missing pieces** To complete the proof, there are two statements ((46) and (50)) left to be shown.

To show (46), we start from the derivative

$$\nabla_{w_j} L = -\sum_{k=1}^n 1_{\langle x_k, w_j \rangle > 0} \nabla_{y_k} \ell(y_k, f(x_k; W, v)) x_k \text{sign}(v_j(0)) \|w_j\|, \|v_j\|, \|w_j\|,$$

and we have

$$y_i \text{sign}(v_j) \langle x_i, \nabla_{w_j} L \rangle = \sum_{k=1}^n 1_{\langle x_k, w_j \rangle > 0} \exp(-y_k f(x_k; W, v)) \langle y_i x_k, y_k x_k \rangle \|w_j\|$$

$$\geq \sum_{k=1}^n 1_{\langle x_k, w_j \rangle > 0} \exp(-y_k f(x_k; W, v)) \mu \|x_k\| \|x_i\| \|w_j\| > 0,$$
since there is at least one summand \( (S_{\text{dead}} = \emptyset) \), the summation is always positive.

To show (50), we first consider \( \theta \in \text{Int}(\tilde{\Theta}_1) \), and we have

\[
\sum_{j=1}^{h} \| \nabla_{w_j} \mathcal{L} \|^2 = \sum_{j=1}^{h} \left\| \sum_{k=1}^{n} \mathbb{1}_{\{x_k, w_j\} > 0} y_k \exp(-y_k f(x_k; W, v)) x_k v_j \right\|^2 \\
\leq \sum_{j=1}^{h} \left( \sum_{k=1}^{n} \exp(-y_k f(x_k; W, v)) \| x_k \| \| v_j \| \right)^2 \\
\leq \sum_{j=1}^{h} \| v_j \|^2 \cdot (X_{\max} \sum_{k=1}^{n} \exp(-y_k f(x_k; W, v)))^2 \\
= \sum_{j=1}^{h} \| v_j \|^2 \cdot (X_{\max} \mathcal{L}(W, v))^2 \\
\leq \sum_{j=1}^{h} \| v_j \|^2 \cdot (X_{\max} \mathcal{L}(W(0), v(0)))^2 = X_{\max}^2 \mathcal{L}^2(W(0), v(0)) \| v \|^2,
\]

similarly, we also have

\[
\sum_{j=1}^{h} \| \nabla_{v_j} \mathcal{L} \|^2 = \sum_{j=1}^{h} \left\| \sum_{k=1}^{n} \mathbb{1}_{\{x_k, w_j\} > 0} y_k \exp(-y_k f(x_k; W, v)) \langle x_k, w_j \rangle \right\|^2 \\
\leq \sum_{j=1}^{h} \left( \sum_{k=1}^{n} \exp(-y_k f(x_k; W, v)) \| x_k \| \| w_j \| \right)^2 \\
\leq \sum_{j=1}^{h} \| w_j \|^2 \cdot (X_{\max} \sum_{k=1}^{n} \exp(-y_k f(x_k; W, v)))^2 \\
= \sum_{j=1}^{h} \| w_j \|^2 \cdot (X_{\max} \mathcal{L}(W, v))^2 \\
\leq \sum_{j=1}^{h} \| w_j \|^2 \cdot (X_{\max} \mathcal{L}(W(0), v(0)))^2 = X_{\max}^2 \mathcal{L}^2(W(0), v(0)) \| W \|^2_F.
\]

Therefore, we have \( \forall \theta \in \text{Int}(\tilde{\Theta}_1) \)

\[
\| F^1_{\theta} \|_F^2 = \| F(\theta) \|_F^2 = \sum_{j=1}^{h} (\| \nabla_{w_j} \mathcal{L} \|^2 + \| \nabla_{v_j} \mathcal{L} \|^2) \\
\leq X_{\max}^2 \mathcal{L}^2(W(0), v(0)) \| W \|^2_F + \| v \|^2 = X_{\max}^2 \mathcal{L}^2(W(0), v(0)) \| \theta \|^2_F,
\]

which gives (50) with \( C = X_{\max} \mathcal{L}(W(0), v(0)) \).

Then for \( \theta \in \hat{\Theta} \setminus \text{Int}(\tilde{\Theta}_1) \), \( \| F^1_{\theta} \| = \lim_{k \to \infty} \| F(\theta_k) \| \leq C \lim_{k \to \infty} \| \theta_k \| = C \| \theta \| \), given some Cauchy sequence \( \theta_k \in \text{Int}(\tilde{\Theta}_1), k = 1, 2, \cdots \) convergent to \( \theta \). This finishes proving (50).

---

3We show it for exponential loss, the case of logistic loss is similar