
Gradient Descent on Two-layer Nets: Margin Maximization and Simplicity Bias

Kaifeng Lyu^{*†}
Princeton University
klyu@cs.princeton.edu

Zhiyuan Li^{*}
Princeton University
zhiyuanli@cs.princeton.edu

Runzhe Wang^{*†}
Princeton University
runzhew@princeton.edu

Sanjeev Arora
Princeton University
arora@cs.princeton.edu

Abstract

The generalization mystery of overparametrized deep nets has motivated efforts to understand how gradient descent (GD) converges to low-loss solutions that generalize well. Real-life neural networks are initialized from small random values and trained with cross-entropy loss for classification (unlike the "lazy" or "NTK" regime of training where analysis was more successful), and a recent sequence of results (Lyu and Li, 2020; Chizat and Bach, 2020; Ji and Telgarsky, 2020a) provide theoretical evidence that GD may converge to the "max-margin" solution with zero loss, which presumably generalizes well. However, the global optimality of margin is proved only in some settings where neural nets are infinitely or exponentially wide. The current paper is able to establish this global optimality for two-layer Leaky ReLU nets trained with gradient flow on linearly separable and symmetric data, regardless of the width. The analysis also gives some theoretical justification for recent empirical findings (Kalimeris et al., 2019) on the so-called simplicity bias of GD towards linear or other "simple" classes of solutions, especially early in training. On the pessimistic side, the paper suggests that such results are fragile. A simple data manipulation can make gradient flow converge to a linear classifier with suboptimal margin.

1 Introduction

One major mystery in deep learning is why deep neural networks generalize despite overparameterization (Zhang et al., 2017). To tackle this issue, many recent works turn to study the *implicit bias* of gradient descent (GD) — what kind of theoretical characterization can we give for the low-loss solution found by GD?

The seminal works by Soudry et al. (2018a,b) revealed an interesting connection between GD and margin maximization: for linear logistic regression on linearly separable data, there can be multiple linear classifiers that perfectly fit the data, but GD with any initialization always converges to the max-margin (hard-margin SVM) solution, even when there is no explicit regularization. Thus the solution found by GD has the same margin-based generalization bounds as hard-margin SVM. Subsequent works on linear models have extended this theoretical understanding of GD to SGD (Nacson et al., 2019b), other gradient-based methods (Gunasekar et al., 2018a), other loss functions with certain poly-exponential tails (Nacson et al., 2019a), linearly non-separable data (Ji and Telgarsky, 2018, 2019b), deep linear nets (Ji and Telgarsky, 2019a; Gunasekar et al., 2018b).

^{*}Equal contribution

[†]Most of the work is done when Kaifeng Lyu and Runzhe Wang were at Tsinghua University.

Given the above results, a natural question to ask is whether GD has the same implicit bias towards max-margin solutions for machine learning models in general. [Lyu and Li \(2020\)](#) studied the relationship between GD and margin maximization on *deep homogeneous neural network*, i.e., neural network whose output function is (positively) homogeneous with respect to its parameters. For homogeneous neural networks, only the direction of parameter matters for classification tasks. For logistic and exponential loss, [Lyu and Li \(2020\)](#) assumed that GD decreases the loss to a small value and achieves full training accuracy at some time point, and then provided an analysis for the training dynamics after this time point (Theorem 3.1), which we refer to as *late phase analysis*. It is shown that GD decreases the loss to 0 in the end and converges to a direction satisfying the Karush-Kuhn-Tucker (KKT) conditions of a constrained optimization problem (P) on margin maximization.

However, given the non-convex nature of neural networks, KKT conditions do not imply global optimality for margins. Several attempts are made to prove the global optimality specifically for two-layer nets. [Chizat and Bach \(2020\)](#) provided a mean-field analysis for infinitely wide two-layer Squared ReLU nets showing that gradient flow converges to the solution with global max margin, which also corresponds to the max-margin classifier in some non-Hilbertian space of functions. [Ji and Telgarsky \(2020a\)](#) extended the proof to finite-width neural nets, but the width needs to be exponential in the input dimension (due to the use of a covering condition). Both works build upon late phase analyses. Under a restrictive assumption that the data is orthogonally separable, i.e., any data point $\mathbf{x}_i$ can serve as a perfect linear separator, [Phuong and Lampert \(2021\)](#) analyzed the full trajectory of gradient flow on two-layer ReLU nets with small initialization, and established the convergence to a piecewise linear classifier that maximizes the margin, irrespective of network width.

In this paper, we study the implicit bias of gradient flow on two-layer neural nets with Leaky ReLU activation ([Maas et al., 2013](#)) and logistic loss. To avoid the *lazy* or *Neural Tangent Kernel (NTK)* regime where the weights are initialized to large random values and do not change much during training ([Jacot et al., 2018](#); [Chizat et al., 2019](#); [Du et al., 2019b,a](#); [Allen-Zhu et al., 2018, 2019](#); [Zou et al., 2018](#); [Arora et al., 2019b](#)), we use small initialization to encourage the model to learn features actively, which is closer to real-life neural network training.

When analyzing convergence behavior of training on neural networks, one can simplify the problem and gain insights by assuming that the data distribution has a simple structure. Many works particularly study the case where the labels are generated by an unknown teacher network that is much smaller/simpler than the (student) neural network to be trained. Following [Brutzkus et al. \(2018\)](#); [Sarussi et al. \(2021\)](#) and many other works, we consider the case where the dataset is linearly separable, namely the labels are generated by a linear teacher, and study the training dynamics of two-layer Leaky ReLU nets on such dataset.

1.1 Our Contribution

Among all the classifiers that can be represented by the two-layer Leaky ReLU nets, we show **any global-max-margin classifier is exactly linear** under one more data assumption: the dataset is *symmetric*, i.e., if $\mathbf{x}$ is in the training set, then so is $-\mathbf{x}$. Note that such symmetry can be ensured by simple data augmentation.

Still, little is known about what kind of classifiers neural network trained by GD learns. Though [Lyu and Li \(2020\)](#) showed that gradient flow converges to a classifier along KKT-margin direction, we note that this result is not sufficient to guarantee the global optimality since such classifier can have nonlinear decision boundaries. See Figure 1 (left) for an example.

In this paper, we provide a multi-phase analysis for the full trajectory of gradient flow, in contrast with previous late phase analyses which only analyzes the trajectory after achieving 100% training accuracy. We show that **gradient flow with small initialization converges to a global-max-margin linear classifier** (Theorem 4.2). The proof leverages power iteration to show that neuron weights align in two directions in an early phase of training, inspired by [Li et al. \(2021\)](#). We further show the alignment at any constant training time by associating the dynamics of wide neural net with that of two-neuron neural net, and finally, extend the alignment to the infinite time limit by applying Kurdyka-Łojasiewicz (KL) inequality in a similar way as [Ji and Telgarsky \(2020a\)](#). The alignment at convergence implies that the convergent classifier is linear.

The above results also justify a recent line of works studying the so-called *simplicity bias*: GD first learns linear functions in the early phase of training, and the complexity of the solution increases

as training goes on (Kalimeris et al., 2019; Hu et al., 2020; Shah et al., 2020). Indeed, our result establishes a form of *extreme simplicity bias* of GD: *if the dataset can be fitted by a linear classifier, then GD learns a linear classifier not only in the beginning but also at convergence.*

On the pessimistic side, this paper suggests that such global margin maximization result could be fragile. Even for linearly separable data, global-max-margin classifiers may be nonlinear without the symmetry assumption. In particular, we show that for any linearly separable dataset, **gradient flow can be led to converge to a linear classifier with suboptimal margin by adding only 3 extra data points** (Theorem 6.2). See Figure 1 (right) for an example.

2 Related Works

Generalization Aspect of Margin Maximization. Margin often appears in the generalization bounds for neural networks (Bartlett et al., 2017; Neyshabur et al., 2018), and larger margin leads to smaller bounds. Jiang et al. (2020) conducted an empirical study for the causal relationships between complexity measures and generalization errors, and showed positive results for normalized margin, which is defined by the output margin divided by the product (or powers of the sum) of Frobenius norms of weight matrices from each layer. On the pessimistic side, negative results are also shown if Frobenius norm is replaced by spectral norm. In this paper, we do use the normalized margin with Frobenius norm (see Section 3).

Learning on Linearly Separable Data. Some works studied the training dynamics of (nonlinear) neural networks on linearly separable data (labels are generated by a linear teacher). Brutzkus et al. (2018) showed that SGD on two-layer Leaky ReLU nets with hinge loss fits the training set in finite steps and generalizes well. Frei et al. (2021) studied online SGD (taking a fresh sample from the population in each step) on the two-layer Leaky ReLU nets with logistic loss. For any data distribution, they proved that there exists a time step in the early phase such that the net has a test error competitive with that of the best linear classifier over the distribution, and hence generalizes well on linearly separable data. Both two papers reveal that the weight vectors in the first layer have positive correlations with the weight of the linear teacher, but their analyses do not imply that the learned classifier is linear. In the NTK regime, Ji and Telgarsky (2020b); Chen et al. (2021) showed that GD on shallow/deep neural nets learns a kernel predictor with good generalization on linearly separable data, and it suffices to have width polylogarithmic in the number of training samples. Still, they do not imply that the learned classifier is linear. Pellegrini and Biroli (2020) provided a mean-field analysis for two-layer ReLU net showing that training with hinge loss and infinite data leads to a linear classifier, but their analysis requires the data distribution to be spherically symmetric (i.e., the probability density only depends on the distance to origin), which is a more restrictive assumption than ours. Sarussi et al. (2021) provided a late phase analysis for gradient flow on two-layer Leaky ReLU nets with logistic loss, which establishes the convergence to linear classifier based on an assumption called *Neural Agreement Regime* (NAR): starting from some time point, for any training sample, the outputs of all the neurons have the same sign. However, it is unclear why this can happen a priori. Comparing with our work, we analyze the full trajectory of gradient flow and establish the convergence to linear classifier without assuming NAR. Phuong and Lampert (2021) analyzed the full trajectory for gradient flow on orthogonally separable data, but every KKT-margin direction attains the global max margin (see Appendix H) in their setting, which it is not necessarily true in general. In our setting, KKT-margin direction with suboptimal margin does exist.

Simplicity Bias. Kalimeris et al. (2019) empirically observed that neural networks in the early phase of training are learning linear classifiers, and provided evidence that SGD learns functions of increasing complexity. Hu et al. (2020) justified this view by proving that the learning dynamics of two-layer neural nets and simple linear classifiers are close to each other in the early phase, for dataset drawn from a data distribution where input coordinates are independent after some linear transformation. The aforementioned work by Frei et al. (2021) can be seen as another theoretical justification for online SGD on arbitrary data distribution. Shah et al. (2020) pointed out that extreme simplicity bias can lead to suboptimal generalization and negative effects on adversarial robustness.

Small Initialization. Several theoretical works studying neural network training with small initialization can be connected to simplicity bias. Maennel et al. (2018) uncovered a weight quantization effect in training two-layer nets with small initialization: gradient flow biases the weight vectors to a certain number of directions determined by the input data (independent of neural network width). It

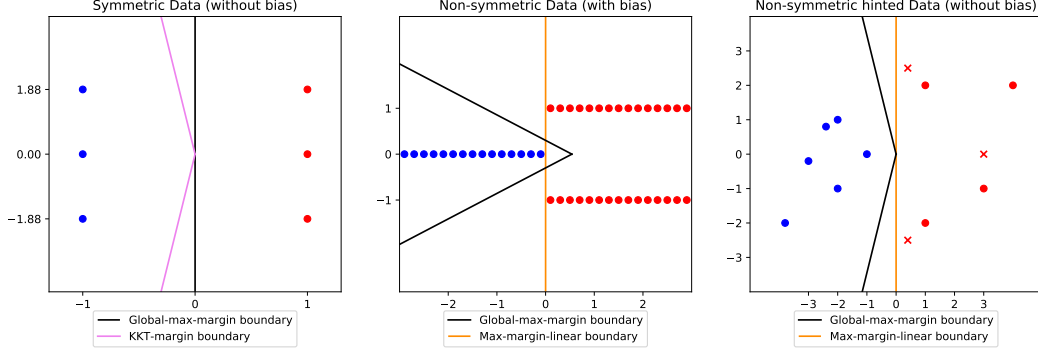


Figure 1: Two-layer Leaky ReLU nets ($\alpha_{\text{leaky}} = 1/2$) with KKT margin and global max margin on linearly separable data. See Appendix I.1 for detailed discussions. **Left:** Theorem 4.3 is not vacuous: a symmetric dataset can have KKT directions with suboptimal margin, but our theory shows that gradient flow from small initialization goes to global max margin. **Middle:** The linear classifier (orange) is along a KKT-margin direction with a much smaller margin comparing to the (nonlinear) global-max-margin classifier (black), but our theory suggests that gradient flow converges to the linear classifier. **Right:** Adding three extra data points (marked as “x”; see Definition 6.1) to a linearly separable dataset makes the linear classifier (orange) has suboptimal margin but causes the neural net to be biased to it.

is hence argued that gradient flow has a bias towards “simple” functions, but their proof is not entirely rigorous and no clear definition of simplicity is given. This weight quantization effect has also been studied under the names of weight clustering (Brutzkus and Globerson, 2019), condensation (Luo et al., 2021; Xu et al., 2021). Williams et al. (2019) studied univariate regression and showed that two-layer ReLU nets with small initialization tend to learn linear splines. For the matrix factorization problem, which can be related to training neural networks with linear or quadratic activations, we can measure the complexity of the learned solution by rank. A line of works showed that gradient descent learns solutions with gradually increasing rank (Li et al., 2018; Arora et al., 2019a; Gidel et al., 2019; Gissin et al., 2020; Li et al., 2021). Such results have been generalized to tensor factorization where the complexity measure is replaced by tensor rank (Razin et al., 2021). Beyond small initialization of our interest and large initialization in the lazy or NTK regime, Woodworth et al. (2020); Moroshko et al. (2020); Mehta et al. (2021) studied feature learning when the initialization scale transitions from small to large scale.

3 Preliminaries

We denote the set $\{1, \dots, n\}$ by $[n]$ and the unit sphere $\{\mathbf{x} \in \mathbb{R}^d : \|\mathbf{x}\|_2 = 1\}$ by $\mathbb{S}^{d-1}$. We call a function $h : \mathbb{R}^D \rightarrow \mathbb{R}$ L -homogeneous if $h(c\boldsymbol{\theta}) = c^L h(\boldsymbol{\theta})$ for all $\boldsymbol{\theta} \in \mathbb{R}^D$ and $c > 0$. For $S \subseteq \mathbb{R}^D$, $\text{conv}(S)$ denotes the convex hull of S . For locally Lipschitz function $f : \mathbb{R}^D \rightarrow \mathbb{R}$, we define Clarke’s subdifferential (Clarke, 1975; Clarke et al., 2008; Davis et al., 2020) to be $\partial^\circ f(\boldsymbol{\theta}) := \text{conv} \{ \lim_{n \rightarrow \infty} \nabla f(\boldsymbol{\theta}_n) : f \text{ differentiable at } \boldsymbol{\theta}_n, \lim_{n \rightarrow \infty} \boldsymbol{\theta}_n = \boldsymbol{\theta} \}$ (see also Appendix B.1).

3.1 Logistic Loss Minimization and Margin Maximization

For a neural net, we use $f_{\boldsymbol{\theta}}(\mathbf{x}) \in \mathbb{R}$ to denote the output logit on input $\mathbf{x} \in \mathbb{R}^d$ when the parameter is $\boldsymbol{\theta} \in \mathbb{R}^D$. We say that the neural net is L -homogeneous if $f_{\boldsymbol{\theta}}(\mathbf{x})$ is L -homogeneous with respect to $\boldsymbol{\theta}$, i.e., $f_{c\boldsymbol{\theta}}(\mathbf{x}) = c^L f_{\boldsymbol{\theta}}(\mathbf{x})$ for all $\boldsymbol{\theta} \in \mathbb{R}^D$ and $c > 0$. VGG-like CNNs can be made homogeneous if we remove all the bias terms except those in the first layer (Lyu and Li, 2020).

Throughout this paper, we restrict our attention to L -homogeneous neural nets with $f_{\boldsymbol{\theta}}(\mathbf{x})$ definable with respect to $\boldsymbol{\theta}$ in an o-minimal structure for all $\mathbf{x}$. (See Coste 2000 for reference for o-minimal structures.) This is a technical condition needed by Theorem 3.1, and it is a mild regularity condition as almost all modern neural networks satisfy this condition, including the two-layer Leaky ReLU networks studied in this paper.

For a dataset $\mathcal{S} = \{(\mathbf{x}_1, y_1), \dots, (\mathbf{x}_n, y_n)\}$, we define $q_i(\boldsymbol{\theta}) := y_i f_{\boldsymbol{\theta}}(\mathbf{x}_i)$ to be the output margin on the data point $(\mathbf{x}_i, y_i)$, and $q_{\min}(\boldsymbol{\theta}) := \min_{i \in [n]} q_i(\boldsymbol{\theta})$ to be the output margin on the dataset $\mathcal{S}$ (or

margin for short). It is easy to see that $q_1(\theta), \dots, q_n(\theta)$ are L -homogeneous functions, and so is $q_{\min}(\theta)$. We define the *normalized margin* $\gamma(\theta) := q_{\min}\left(\frac{\theta}{\|\theta\|_2}\right) = \frac{q_{\min}(\theta)}{\|\theta\|_2^L}$ to be the output margin (on the dataset) for the normalized parameter $\frac{\theta}{\|\theta\|_2}$.

We refer the problem of finding θ that maximizes $\gamma(\theta)$ as *margin maximization*. Note that once we have found an optimal solution $\theta^* \in \mathbb{R}^D$, $c\theta^*$ is also optimal for all $c > 0$. We can put the norm constraint on θ to eliminate this freedom on rescaling:

$$\max_{\theta \in \mathbb{S}^{D-1}} \gamma(\theta). \quad (\text{M})$$

Alternatively, we can also constrain the margin to have $q_{\min} \geq 1$ and minimize the norm:

$$\min \frac{1}{2} \|\theta\|_2^2 \quad \text{s.t.} \quad q_i(\theta) \geq 1, \quad \forall i \in [n]. \quad (\text{P})$$

One can easily show that θ^* is a global maximizer of (M) if and only if $\frac{\theta^*}{(q_{\min}(\theta^*))^{1/L}}$ is a global minimizer of (P). For convenience, we make the following convention: if $\frac{\theta}{\|\theta\|_2}$ is a local/global maximizer of (M), then we say θ is along a *local-max-margin direction/global-max-margin direction*; if $\frac{\theta}{(q_{\min}(\theta))^{1/L}}$ satisfies the KKT conditions of (P), then we say θ is along a *KKT-margin direction*.

Gradient flow with logistic loss is defined by the following differential inclusion,

$$\frac{d\theta}{dt} \in -\partial^\circ \mathcal{L}(\theta), \quad \text{with } \mathcal{L}(\theta) := \frac{1}{n} \sum_{i=1}^n \ell(q_i(\theta)), \quad (1)$$

where $\ell(q) := \ln(1 + e^{-q})$ is the logistic loss. [Lyu and Li \(2020\)](#); [Ji and Telgarsky \(2020a\)](#) showed that $\theta(t)/\|\theta(t)\|_2$ always converges to a KKT-margin direction. We restate the results below.

Theorem 3.1 ([Lyu and Li 2020](#); [Ji and Telgarsky 2020a](#)). *For homogeneous neural networks, if $\mathcal{L}(\theta(0)) < \frac{\ln 2}{n}$, then $\mathcal{L}(\theta(t)) \rightarrow 0$, $\|\theta(t)\|_2 \rightarrow +\infty$, and $\frac{\theta(t)}{\|\theta(t)\|_2}$ converges to a KKT-margin direction as $t \rightarrow +\infty$.*

3.2 Two-Layer Leaky ReLU Networks on Linearly Separable Data

Let $\phi(x) = \max\{x, \alpha_{\text{leaky}} x\}$ be Leaky ReLU, where $\alpha_{\text{leaky}} \in (0, 1)$. Throughout the following sections, we consider a two-layer neural net defined as below,

$$f_\theta(x) = \sum_{k=1}^m a_k \phi(w_k^\top x).$$

where $w_1, \dots, w_m \in \mathbb{R}^d$ are the weights in the first layer, $a_1, \dots, a_m \in \mathbb{R}$ are the weights in the second layer, and $\theta = (w_1, \dots, w_m, a_1, \dots, a_m) \in \mathbb{R}^D$ is the concatenation of all trainable parameters, where $D = md + m$. We can verify that $f_\theta(x)$ is 2-homogeneous with respect to θ .

Let $\mathcal{S} := \{(x_1, y_1), \dots, (x_n, y_n)\}$ be the training set. For simplicity, we assume that $\|x_i\|_2 \leq 1$. We focus on linearly separable data, thus we assume that $\mathcal{S}$ is linearly separable throughout the paper.

Assumption 3.2 (Linear Separable). There exists a $w \in \mathbb{R}^d$ such that $y_i \langle w, x_i \rangle \geq 1$ for all $i \in [n]$.

Definition 3.3 (Max-margin Linear Separator). For the linearly separable dataset $\mathcal{S}$, we say that $w^* \in \mathbb{S}^{d-1}$ is the max-margin linear separator if w^* maximizes $\min_{i \in [n]} y_i \langle w, x_i \rangle$ over $w \in \mathbb{S}^{d-1}$.

4 Training on Linearly Separable and Symmetric Data

In this section, we study the implicit bias of gradient flow assuming the training data is linearly separable and *symmetric*. We say a dataset is symmetric if whenever x is present in the training set, the input $-x$ is also present. By linear separability, x and $-x$ must have different labels because $\langle w^*, x \rangle = -\langle w^*, -x \rangle$, where w^* is the max-margin linear separator. The formal statement for this assumption is given below.

Assumption 4.1 (Symmetric). n is even and $x_i = -x_{i+n/2}, y_i = 1, y_{i+n/2} = -1$ for $1 \leq i \leq n/2$.

This symmetry can be ensured via data augmentation. Given a dataset, if it is known that the ground-truth labels are produced by an unknown linear classifier, then one can augment each data point $(\mathbf{x}, y)$ by flipping the sign, i.e., replace it with two data points $(\mathbf{x}, y)$, $(-\mathbf{x}, -y)$ (and thus the dataset size is doubled).

Our results show that gradient flow directionally converges to a global-max-margin direction for two-layer Leaky ReLU networks, when the dataset is linearly separable and symmetric. To achieve such result, the key insight is that any global-max-margin direction represents a linear classifier, which we will see in Section 4.1. Then we will present our main convergence results in Section 4.2.

4.1 Global-Max-Margin Classifiers are Linear

Theorem 4.2 below characterizes the global-max-margin direction in our case by showing that margin maximization and simplicity bias coincide with each other: a network that representing the *max-margin linear classifier* (i.e., $f_{\theta}(\mathbf{x}) = c \langle \mathbf{w}^*, \mathbf{x} \rangle$ for some $c > 0$) can simultaneously achieve the goals of being simple and maximizing the margin.

Theorem 4.2. *Under Assumptions 3.2 and 4.1, for the two-layer Leaky ReLU network with width $m \geq 2$, any global-max-margin direction $\theta^* \in \mathbb{S}^{D-1}$, f_{θ^*} represents a linear classifier. Moreover, we have $f_{\theta^*}(\mathbf{x}) = \frac{1+\alpha_{\text{leaky}}}{4} \langle \mathbf{w}^*, \mathbf{x} \rangle$ for all $\mathbf{x} \in \mathbb{R}^d$, where $\mathbf{w}^*$ is the max-margin linear separator.*

The result of Theorem 4.2 is based on the observation that replacing each neuron $(a_k, \mathbf{w}_k)$ in a network with two neurons of opposing parameters $(a_k, \mathbf{w}_k)$ and $(-a_k, -\mathbf{w}_k)$ does not decrease the normalized margin on the symmetric dataset, while making the classifier linear in function space. Thus if any direction attains the global max margin, we can construct a new global-max-margin direction which corresponds to a linear classifier. We can show that every weight vector $\mathbf{w}_k$ of this linear classifier must be in the direction of $\mathbf{w}^*$ or $-\mathbf{w}^*$. Then the original classifier must also be linear in the same direction.

4.2 Convergence to Global-Max-Margin Directions

Though Theorem 3.1 guarantees that gradient flow directionally converges to a KKT-margin direction if the loss is optimized successfully, we note that KKT-margin directions can be non-linear and have complicated decision boundaries. See Figure 1 (left) for an example. Therefore, to establish the convergence to linear classifiers, Theorem 3.1 is not enough and we need a new analysis for the trajectory of gradient flow.

We use initialization $\mathbf{w}_k \stackrel{\text{i.i.d.}}{\sim} \mathcal{N}(\mathbf{0}, \sigma_{\text{init}}^2 \mathbf{I})$, $a_k \stackrel{\text{i.i.d.}}{\sim} \mathcal{N}(0, c_{\text{ainit}}^2 \sigma_{\text{init}}^2)$, where c_{ainit} is a fixed constant throughout this paper and σ_{init} controls the initialization scale. We call this distribution as $\theta_0 \sim \mathcal{D}_{\text{init}}(\sigma_{\text{init}})$. An alternative way to generate this distribution is to first draw $\bar{\theta}_0 \sim \mathcal{D}_{\text{init}}(1)$, and then set $\theta_0 = \sigma_{\text{init}} \bar{\theta}_0$. With small initialization, we can establish the following convergence result.

Theorem 4.3. *Under Assumptions 3.2 and 4.1 and certain regularity conditions (see Assumptions 4.5 and 4.6 below), consider gradient flow on a Leaky ReLU network with width $m \geq 2$ and initialization $\theta_0 = \sigma_{\text{init}} \bar{\theta}_0$ where $\bar{\theta}_0 \sim \mathcal{D}_{\text{init}}(1)$. With probability $1 - 2^{-(m-1)}$ over the random draw of $\bar{\theta}_0$, if the initialization scale is sufficiently small, then gradient flow directionally converges and $f^\infty(\mathbf{x}) := \lim_{t \rightarrow +\infty} f_{\theta(t)/\|\theta(t)\|_2}(\mathbf{x})$ represents the max-margin linear classifier. That is,*

$$\Pr_{\bar{\theta}_0 \sim \mathcal{D}_{\text{init}}(1)} [\exists \sigma_{\text{init}}^{\max} > 0 \text{ s.t. } \forall \sigma_{\text{init}} < \sigma_{\text{init}}^{\max}, \forall \mathbf{x} \in \mathbb{R}^d, f^\infty(\mathbf{x}) = C \langle \mathbf{w}^*, \mathbf{x} \rangle] \geq 1 - 2^{-(m-1)},$$

where $C := \frac{1+\alpha_{\text{leaky}}}{4}$ is a scaling factor.

Combining Theorem 4.2 and Theorem 4.3, we can conclude that gradient flow achieves the global max margin in our case.

Corollary 4.4. *In the settings of Theorem 4.3, gradient flow on linearly separable and symmetric data directionally converges to the global-max-margin direction with probability $1 - 2^{-(m-1)}$.*

4.3 Additional Notations and Assumptions

Let $\boldsymbol{\mu} := \frac{1}{n} \sum_{i=1}^n y_i \mathbf{x}_i$, which is non-zero since $\langle \boldsymbol{\mu}, \mathbf{w}_* \rangle = \frac{1}{n} \sum_{i \in [n]} y_i \mathbf{w}_*^\top \mathbf{x}_i \geq 1$. Let $\bar{\boldsymbol{\mu}} := \frac{\boldsymbol{\mu}}{\|\boldsymbol{\mu}\|_2}$. We use $\varphi(\theta_0, t) \in \mathbb{R}^d$ to the value of θ at time t for $\theta(0) = \theta_0$.

We make the following technical assumption, which holds if we are allowed to add a slight perturbation to the training set.

Assumption 4.5. For all $i \in [n]$, $\langle \boldsymbol{\mu}, \mathbf{x}_i \rangle \neq 0$.

Another technical issue we face is that the gradient flow may not be unique due to non-smoothness. It is possible that $\varphi(\boldsymbol{\theta}_0, t)$ is not well-defined as the solution of (1) may not be unique. See Appendix I.2 for more discussions. In this case, we assign $\varphi(\boldsymbol{\theta}_0, \cdot)$ to be an arbitrary gradient flow trajectory starting from $\boldsymbol{\theta}_0$. In the case where $\varphi(\boldsymbol{\theta}_0, t)$ has only one possible value for all $t \geq 0$, we say that $\boldsymbol{\theta}_0$ is a *non-branching starting point*. We assume the following technical assumption.

Assumption 4.6. For any $m \geq 2$, there exist $r, \epsilon > 0$ such that $\boldsymbol{\theta}$ is a non-branching starting point if its neurons can be partitioned into two groups: in the first group, $a_k = \|\mathbf{w}_k\|_2 \in (0, r)$ and all $\mathbf{w}_k$ point to the same direction $\mathbf{w}^+ \in \mathbb{S}^{d-1}$ with $\|\mathbf{w}^+ - \bar{\boldsymbol{\mu}}\|_2 \leq \epsilon$; in the second group, $-a_k = \|\mathbf{w}_k\|_2 \in (0, r)$ and all $\mathbf{w}_k$ point to the same direction $\mathbf{w}^- \in \mathbb{S}^{d-1}$ with $\|\mathbf{w}^- + \bar{\boldsymbol{\mu}}\|_2 \leq \epsilon$.

5 Proof Sketch for the Symmetric Case

In this section, we provide a proof sketch for Theorem 4.3. Our proof uses a multi-phase analysis, which divides the training process into 3 phases, from small initialization to the final convergence. We will now elaborate the analyses for them one by one.

5.1 Phase I: Dynamics Near Zero

Gradient flow starts with small initialization. In Phase I, we analyze the dynamics when gradient flow does not go far away from zero. Inspired by Li et al. (2021), we relate such dynamics to power iterations and show that every weight vector $\mathbf{w}_k$ in the first layer moves towards the directions of either $\bar{\boldsymbol{\mu}}$ or $-\bar{\boldsymbol{\mu}}$. To see this, the first step is to note that $f_{\boldsymbol{\theta}}(\mathbf{x}_i) \approx 0$ when $\boldsymbol{\theta}$ is close to 0. Applying Taylor expansion on $\ell(y_i f_{\boldsymbol{\theta}}(\mathbf{x}_i))$,

$$\mathcal{L}(\boldsymbol{\theta}) = \frac{1}{n} \sum_{i \in [n]} \ell(y_i f_{\boldsymbol{\theta}}(\mathbf{x}_i)) \approx \frac{1}{n} \sum_{i \in [n]} (\ell(0) + \ell'(0) y_i f_{\boldsymbol{\theta}}(\mathbf{x}_i)). \quad (2)$$

Expanding $f_{\boldsymbol{\theta}}(\mathbf{x}_i)$ and reorganizing the terms, we have

$$\begin{aligned} \mathcal{L}(\boldsymbol{\theta}) &\approx \frac{1}{n} \sum_{i \in [n]} \ell(0) + \frac{1}{n} \sum_{i \in [n]} \ell'(0) \sum_{k \in [m]} y_i a_k \phi(\mathbf{w}_k^\top \mathbf{x}_i) = \ell(0) + \frac{\ell'(0)}{n} \sum_{k \in [m]} \sum_{i \in [n]} y_i a_k \phi(\mathbf{w}_k^\top \mathbf{x}_i) \\ &= \ell(0) - \sum_{k \in [m]} a_k G(\mathbf{w}_k), \end{aligned}$$

where G -function (Maennel et al., 2018) is defined below:

$$G(\mathbf{w}) := \frac{-\ell'(0)}{n} \sum_{i \in [n]} y_i \phi(\mathbf{w}^\top \mathbf{x}_i) = \frac{1}{2n} \sum_{i \in [n]} y_i \phi(\mathbf{w}^\top \mathbf{x}_i).$$

This means gradient flow optimizes each $-a_k G(\mathbf{w}_k)$ separately near origin.

$$\frac{d\mathbf{w}_k}{dt} \approx a_k \partial^\circ G(\mathbf{w}_k), \quad \frac{da_k}{dt} \approx G(\mathbf{w}_k). \quad (3)$$

In the case where Assumption 4.1 holds, we can pair each $\mathbf{x}_i$ with $-\mathbf{x}_i$ and use the identity $\phi(z) - \phi(-z) = \max\{z, \alpha_{\text{leaky}} z\} - \max\{-z, -\alpha_{\text{leaky}} z\} = (1 + \alpha_{\text{leaky}})z$ to show that $G(\mathbf{w})$ is linear:

$$G(\mathbf{w}) = \frac{1}{2n} \sum_{i \in [n/2]} (\phi(\mathbf{w}^\top \mathbf{x}_i) - \phi(-\mathbf{w}^\top \mathbf{x}_i)) = \frac{1}{2n} \sum_{i \in [n/2]} (1 + \alpha_{\text{leaky}}) \mathbf{w}^\top \mathbf{x}_i = \langle \mathbf{w}, \tilde{\boldsymbol{\mu}} \rangle,$$

where $\tilde{\boldsymbol{\mu}} := \frac{1+\alpha_{\text{leaky}}}{2} \boldsymbol{\mu} = \frac{1+\alpha_{\text{leaky}}}{2n} \sum_{i \in [n]} y_i \mathbf{x}_i$. Substituting this formula for G into (3) reveals that the dynamics of two-layer neural nets near zero has a close relationship to power iteration (or matrix exponentiation) of a matrix $\mathbf{M}_{\tilde{\boldsymbol{\mu}}} \in \mathbb{R}^{(d+1) \times (d+1)}$ that only depends on data.

$$\frac{d}{dt} \begin{bmatrix} \mathbf{w}_k \\ a_k \end{bmatrix} \approx \mathbf{M}_{\tilde{\boldsymbol{\mu}}} \begin{bmatrix} \mathbf{w}_k \\ a_k \end{bmatrix}, \quad \text{where} \quad \mathbf{M}_{\tilde{\boldsymbol{\mu}}} := \begin{bmatrix} \mathbf{0} & \tilde{\boldsymbol{\mu}} \\ \tilde{\boldsymbol{\mu}}^\top & 0 \end{bmatrix}.$$

Simple linear algebra shows that $\lambda_0 := \|\tilde{\mu}\|_2, \frac{1}{\sqrt{2}}(\tilde{\mu}, 1) \in \mathbb{R}^{d+1}$ are the unique top eigenvalue and eigenvector of $M_{\tilde{\mu}}$, which suggests that $(\mathbf{w}_k(t), a_k(t)) \in \mathbb{R}^{d+1}$ aligns to this top eigenvector direction if the approximation (3) holds for a sufficiently long time. With small initialization, this can indeed be true and we obtain the following lemma.

Definition 5.1 (M-norm). For parameter vector $\theta = (\mathbf{w}_1, \dots, \mathbf{w}_m, a_1, \dots, a_m)$, we define the M-norm to be $\|\theta\|_M = \max_{k \in [m]} \{\max\{\|\mathbf{w}_k\|_2, |a_k|\}\}$.

Lemma 5.2. Let $r > 0$ be a small value. With probability 1 over the random draw of $\bar{\theta}_0 = (\bar{\mathbf{w}}_1, \dots, \bar{\mathbf{w}}_m, \bar{a}_1, \dots, \bar{a}_m) \sim \mathcal{D}_{\text{init}}(1)$, if we take $\sigma_{\text{init}} \leq \frac{r^3}{\sqrt{m}\|\bar{\theta}_0\|_M}$, then any neuron $(\mathbf{w}_k, a_k)$ at time $T_1(r) := \frac{1}{\lambda_0} \ln \frac{r}{\sqrt{m}\sigma_{\text{init}}\|\bar{\theta}_0\|_M}$ can be decomposed into

$$\mathbf{w}_k(T_1(r)) = r\bar{b}_k\bar{\mu} + \Delta\mathbf{w}_k, \quad a_k(T_1(r)) = r\bar{b}_k + \Delta a_k,$$

where $\bar{b}_k := \frac{\langle \bar{\mathbf{w}}_k, \bar{\mu} \rangle + \bar{a}_k}{2\sqrt{m}\|\bar{\theta}_0\|_M}$ and the error term $\Delta\theta := (\Delta\mathbf{w}_1, \dots, \Delta\mathbf{w}_m, \Delta a_1, \dots, \Delta a_m)$ is bounded by $\|\Delta\theta\|_M \leq \frac{Cr^3}{\sqrt{m}}$ for some universal constant C .

5.2 Phase II: Near-Two-Neuron Dynamics

By Lemma 5.2, we know that at time $T_1(r)$ we have $\mathbf{w}_k(T_1(r)) \approx r\bar{b}_k\bar{\mu}$ and $a_k(T_1(r)) \approx r\bar{b}_k$, where $\bar{b} \in \mathbb{R}^d$ is some fixed vector. This motivates us to couple the training dynamics of $\theta(t) = (\mathbf{w}_1(t), \dots, \mathbf{w}_m(t), a_1(t), \dots, a_m(t))$ after the time $T_1(r)$ with another gradient flow starting from the point $(r\bar{b}_1\bar{\mu}, \dots, r\bar{b}_m\bar{\mu}, r\bar{b}_1, \dots, r\bar{b}_m)$. Interestingly, the latter dynamic can be seen as a dynamic of two neurons “embedded” into the m -neuron neural net, and we will show that $\theta(t)$ is close to this “embedded” two-neuron dynamic for a long time. Now we first introduce our idea of embedding a two-neuron network into an m -neuron network.

Embedding. For any $\mathbf{b} \in \mathbb{R}^m$, we say that $\mathbf{b}$ is a *good embedding vector* if it has at least one positive entry and one negative entry, and all the entries are non-zero. For a good embedding vector $\mathbf{b}$, we use $b_+ := \sqrt{\sum_{j \in [m]} \mathbb{1}_{[b_j > 0]} b_j^2}$ and $b_- := -\sqrt{\sum_{j \in [m]} \mathbb{1}_{[b_j < 0]} b_j^2}$ to denote the root-sum-squared of the positive entries and the negative root-sum-squared of the negative entries. For parameter $\hat{\theta} := (\hat{\mathbf{w}}_1, \hat{\mathbf{w}}_2, \hat{a}_1, \hat{a}_2)$ of a two-neuron neural net with $\hat{a}_1 > 0$ and $\hat{a}_2 < 0$, we define the *embedding* from two-neuron into m -neuron neural nets as $\pi_{\mathbf{b}}(\hat{\mathbf{w}}_1, \hat{\mathbf{w}}_2, \hat{a}_1, \hat{a}_2) = (\mathbf{w}_1, \dots, \mathbf{w}_m, a_1, \dots, a_m)$, where

$$a_k = \begin{cases} \frac{b_k}{b_+} \hat{a}_1, & \text{if } b_k > 0 \\ \frac{b_k}{b_-} \hat{a}_2, & \text{if } b_k < 0 \end{cases}, \quad \mathbf{w}_k = \begin{cases} \frac{b_k}{b_+} \hat{\mathbf{w}}_1, & \text{if } b_k > 0 \\ \frac{b_k}{b_-} \hat{\mathbf{w}}_2, & \text{if } b_k < 0 \end{cases}.$$

It is easy to check that $f_{\hat{\theta}}(\mathbf{x}) = f_{\pi_{\mathbf{b}}(\hat{\theta})}(\mathbf{x})$ by the homogeneity of the activation ($\phi(cz) = c\phi(z)$ for $c > 0$):

$$\begin{aligned} f_{\pi_{\mathbf{b}}(\hat{\theta})}(\mathbf{x}) &= \sum_{b_k > 0} a_k \phi(\mathbf{w}_k^\top \mathbf{x}) + \sum_{b_k < 0} a_k \phi(\mathbf{w}_k^\top \mathbf{x}) \\ &= \sum_{b_k > 0} \frac{b_k^2}{b_+^2} \hat{a}_1 \phi(\hat{\mathbf{w}}_1^\top \mathbf{x}) + \sum_{b_k < 0} \frac{b_k^2}{b_-^2} \hat{a}_2 \phi(\hat{\mathbf{w}}_2^\top \mathbf{x}) = \hat{a}_1 \phi(\hat{\mathbf{w}}_1^\top \mathbf{x}) + \hat{a}_2 \phi(\hat{\mathbf{w}}_2^\top \mathbf{x}) = f_{\hat{\theta}}(\mathbf{x}). \end{aligned}$$

Moreover, by taking the chain rule, we can obtain the following lemma showing that the trajectories starting from $\hat{\theta}$ and $\pi_{\mathbf{b}}(\hat{\theta})$ are essentially the same.

Lemma 5.3. Given $\hat{\theta} := (\hat{\mathbf{w}}_1, \hat{\mathbf{w}}_2, \hat{a}_1, \hat{a}_2)$ with $\hat{a}_1 > 0$ and $\hat{a}_2 < 0$, if both $\hat{\theta}$ and $\pi_{\mathbf{b}}(\hat{\theta})$ are non-branching starting points, then $\varphi(\pi_{\mathbf{b}}(\hat{\theta}), t) = \pi_{\mathbf{b}}(\varphi(\hat{\theta}, t))$ for all $t \geq 0$.

Approximate Embedding. Back to our analysis for Phase II, $\bar{b}$ is a good embedding vector with high probability (see lemma below). Let $\hat{\theta} := (\bar{b}_+, \bar{b}_+\bar{\mu}, \bar{b}_-, \bar{b}_-\bar{\mu})$. By Lemma 5.2, $\pi_{\bar{b}}(r\hat{\theta}) = (r\bar{b}_1\bar{\mu}, \dots, r\bar{b}_m\bar{\mu}, r\bar{b}_1, \dots, r\bar{b}_m) \approx \theta(T_1(r))$, which means $r\hat{\theta} \rightarrow \theta(T_1(r))$ is approximately an embedding. Suppose that the approximation happens to be exact, namely $\pi_{\bar{b}}(r\hat{\theta}) = \theta(T_1(r))$, then $\theta(T_1(r) + t) = \pi_{\bar{b}}(\varphi(r\hat{\theta}, t))$ by Lemma 5.3. Inspired by this, we consider the case where $\sigma_{\text{init}} \rightarrow 0, r \rightarrow 0$ so that the approximate embedding is infinitely close to the exact one, and prove the following lemma. We shift the training time by $T_2(r)$ to avoid trivial limits (such as 0).

Lemma 5.4. *Follow the notations in Lemma 5.2 and take $\sigma_{\text{init}} \leq \frac{r^3}{\sqrt{m}\|\bar{\theta}_0\|_M}$. Let $T_2(r) := \frac{1}{\lambda_0} \ln \frac{1}{r}$, then $T_{12} := T_1(r) + T_2(r) = \frac{1}{\lambda_0} \ln \frac{1}{\sqrt{m}\sigma_{\text{init}}\|\bar{\theta}_0\|_M}$ regardless the choice of r . For width $m \geq 2$, with probability $1 - 2^{-(m-1)}$ over the random draw of $\bar{\theta}_0 \sim \mathcal{D}_{\text{init}}(1)$, the vector $\bar{\mathbf{b}} \in \mathbb{R}^m$ is a good embedding vector, and for the two-neuron dynamics starting with rescaled initialization in the direction of $\hat{\boldsymbol{\theta}} := (\bar{\mathbf{b}}_+, \bar{\mathbf{b}}_+ \bar{\boldsymbol{\mu}}, \bar{\mathbf{b}}_-, \bar{\mathbf{b}}_- \bar{\boldsymbol{\mu}})$, the following limit exists for all t ,*

$$\tilde{\boldsymbol{\theta}}(t) := \lim_{r \rightarrow 0} \varphi(r\hat{\boldsymbol{\theta}}, T_2(r) + t) \neq \mathbf{0}, \quad (4)$$

and moreover, for the m -neuron dynamics of $\boldsymbol{\theta}(t)$, the following holds for all t ,

$$\lim_{\sigma_{\text{init}} \rightarrow 0} \boldsymbol{\theta}(T_{12} + t) = \pi_{\bar{\mathbf{b}}}(\tilde{\boldsymbol{\theta}}(t)). \quad (5)$$

5.3 Phase III: Dynamics near Global-Max-Margin Direction

With some efforts, we have the following characterization for the two-neuron dynamics.

Theorem 5.5. *For $m = 2$, if initially $a_1 = \|\mathbf{w}_1\|_2$, $a_2 = -\|\mathbf{w}_2\|_2$, $\langle \mathbf{w}_1, \mathbf{w}^* \rangle > 0$ and $\langle \mathbf{w}_2, \mathbf{w}^* \rangle < 0$, then $\boldsymbol{\theta}(t)$ directionally converges to the following global-max-margin direction,*

$$\lim_{t \rightarrow +\infty} \frac{\boldsymbol{\theta}(t)}{\|\boldsymbol{\theta}(t)\|_2} = \frac{1}{4}(\mathbf{w}^*, -\mathbf{w}^*, 1, -1),$$

where $\mathbf{w}^*$ is the max-margin linear separator.

It is not hard to verify that $\tilde{\boldsymbol{\theta}}(t)$ satisfies the conditions required by Theorem 5.5. Given this result, a first attempt to establish the convergence of $\boldsymbol{\theta}(t)$ to global-max-margin direction is to take $t \rightarrow +\infty$ on both sides of (5). However, this only proves that $\boldsymbol{\theta}(T_{12} + t)$ directionally converges to the global-max-margin direction if we take the limit $\sigma_{\text{init}} \rightarrow 0$ first then take $t \rightarrow +\infty$, while we are interested in the convergent solution when $t \rightarrow +\infty$ first then $\sigma_{\text{init}} \rightarrow 0$ (i.e., solution gradient flow converges to with infinite training time, if it starts from sufficiently small initialization). These two double limits are not equivalent because the order of limits cannot be exchanged without extra conditions.

To overcome this issue, we follow a similar proof strategy as Ji and Telgarsky (2020a) to prove local convergence near a local-max-margin direction, as formally stated below. Theorem 5.6 holds for L -homogeneous neural networks in general and we believe is of independent interest.

Theorem 5.6. *Consider any L -homogeneous neural networks with logistic loss. Given a local-max-margin direction $\bar{\boldsymbol{\theta}}^* \in \mathbb{S}^{D-1}$ and any $\delta > 0$, there exists $\epsilon_0 > 0$ and $\rho_0 \geq 1$ such that for any $\boldsymbol{\theta}_0$ with norm $\|\boldsymbol{\theta}_0\|_2 \geq \rho_0$ and direction $\left\| \frac{\boldsymbol{\theta}_0}{\|\boldsymbol{\theta}_0\|_2} - \bar{\boldsymbol{\theta}}^* \right\|_2 \leq \epsilon_0$, gradient flow starting with $\boldsymbol{\theta}_0$ directionally converges to some direction $\bar{\boldsymbol{\theta}}$ with the same normalized margin γ as $\bar{\boldsymbol{\theta}}^*$, and $\|\bar{\boldsymbol{\theta}} - \bar{\boldsymbol{\theta}}^*\|_2 \leq \delta$.*

Using Theorem 5.6, we can finish the proof for Theorem 4.3 as follows. First we note that the two-neuron global-max-margin direction $\frac{1}{4}(\mathbf{w}^*, -\mathbf{w}^*, 1, -1)$ after embedding is a global-max-margin direction for m -neurons, and we can prove that any direction with distance no more than a small constant $\delta > 0$ is still a global-max-margin direction. Then we can take t to be large enough so that $\pi_{\bar{\mathbf{b}}}(\tilde{\boldsymbol{\theta}}(t))$ satisfies the conditions in Theorem 5.6. According to (5), we can also make the conditions hold for $\boldsymbol{\theta}(T_{12} + t)$ by taking σ_{init} and r to be sufficiently small. Finally, applying Theorem 5.6 finishes the proof.

6 Non-symmetric Data Complicates the Picture

Now we turn to study the case without assuming symmetry and the question is whether the implicit bias to global-max-margin solution still holds. Unfortunately, it turns out the convergence to global-max-margin classifier is very fragile — for any linearly separable dataset, we can add 3 extra data points so that every linear classifier has suboptimal margin but still gradient flow with small initialization converges to a linear classifier.³ See Definition 6.1 for the construction and Figure 1 (right) for an example.

³Here linear classifier refers to a classifier whose decision boundary is linear.

Unlike the symmetric case, we use balanced Gaussian initialization instead of purely random Gaussian initialization: $\mathbf{w}_k \sim \mathcal{N}(\mathbf{0}, \sigma_{\text{init}}^2 \mathbf{I})$, $a_k = s_k \|\mathbf{w}_k\|_2$, where $s_k \sim \text{unif}\{\pm 1\}$. We call this distribution as $\theta_0 \sim \tilde{\mathcal{D}}_{\text{init}}(\sigma_{\text{init}})$. This adaptation can greatly simplify our analysis since it ensures that $a_k(t) = s_k \|\mathbf{w}_k(t)\|_2$ for all $t \geq 0$ (Corollary B.18). Similar as the symmetric case, an alternative way to generate this distribution is to first draw $\bar{\theta}_0 \sim \tilde{\mathcal{D}}_{\text{init}}(1)$, and then set $\theta_0 = \sigma_{\text{init}} \bar{\theta}_0$.

Definition 6.1 ($(H, K, \epsilon, \mathbf{w}_\perp)$ -Hinted Dataset). Given a linearly separable dataset $\mathcal{S}$ with max-margin linear separator $\mathbf{w}^*$, for constants $H, K, \epsilon > 0$ and unit vector $\mathbf{w}_\perp \in \mathbb{S}^{d-1}$ perpendicular to $\mathbf{w}^*$, we define the $(H, K, \epsilon, \mathbf{w}_\perp)$ -hinted dataset $\mathcal{S}'$ by the dataset containing all the data points in $\mathcal{S}$ and the following 3 data points (numbered by 1, 2, 3) that can serve as hints to the max-margin linear separator $\mathbf{w}^*$:

$$(\mathbf{x}_1, y_1) = (H\mathbf{w}^*, 1), \quad (\mathbf{x}_2, y_2) = (\epsilon\mathbf{w}^* + K\mathbf{w}_\perp, 1), \quad (\mathbf{x}_3, y_3) = (\epsilon\mathbf{w}^* - K\mathbf{w}_\perp, 1).$$

Theorem 6.2. *Given a linearly separable dataset $\mathcal{S}$ and a unit vector $\mathbf{w}_\perp \in \mathbb{S}^{d-1}$ perpendicular to the max-margin linear separator $\mathbf{w}^*$, for any sufficiently large $H > 0, K > 0$ and sufficiently small $\epsilon > 0$, the following statement holds for the $(H, K, \epsilon, \mathbf{w}_\perp)$ -Hinted Dataset $\mathcal{S}'$. Under a regularity assumption for gradient flow (see Assumption A.6), consider gradient flow on a Leaky ReLU network with width $m \geq 1$ and initialization $\theta_0 = \sigma_{\text{init}} \bar{\theta}_0$ where $\bar{\theta}_0 \sim \tilde{\mathcal{D}}_{\text{init}}(1)$. With probability $1 - 2^{-m}$ over the draw of θ_0 , if the initialization scale is sufficiently small, then gradient flow directionally converges and $f^\infty(\mathbf{x}) := \lim_{t \rightarrow +\infty} f_{\theta(t)/\|\theta(t)\|_2}(\mathbf{x})$ represents the one-Leaky-ReLU classifier $\frac{1}{2}\phi(\langle \mathbf{w}^*, \mathbf{x} \rangle)$ with linear decision boundary. That is,*

$$\Pr_{\bar{\theta}_0 \sim \tilde{\mathcal{D}}_{\text{init}}(1)} \left[\exists \sigma_{\text{init}}^{\max} > 0 \text{ s.t. } \forall \sigma_{\text{init}} < \sigma_{\text{init}}^{\max}, \forall \mathbf{x} \in \mathbb{R}^d, f^\infty(\mathbf{x}) = \frac{1}{2}\phi(\langle \mathbf{w}^*, \mathbf{x} \rangle) \right] \geq 1 - \delta.$$

Moreover, the convergent classifier only attains a suboptimal margin.

Theorem 6.2 is actually a simple corollary general theorem under data assumptions that hold for a broader class of linearly separable data. From a high-level perspective, we only require two assumptions: (1). There is a direction such that data points have large inner products with this direction on average; (2). The support vectors for the max-margin linear separator $\mathbf{w}^*$ have nearly the same labels. The first hint data point is for the first condition and the second and third data point is for the second condition. We defer formal statements of the assumptions and theorems to Appendix A.

7 Conclusions and Future Works

We study the implicit bias of gradient flow in training two-layer Leaky ReLU networks on linearly separable datasets. When the dataset is symmetric, we show any global-max-margin classifier is exactly linear and gradient flow converges to a global-max-margin direction. On the pessimistic side, we show such margin maximization result is fragile — for any linearly separable dataset, we can lead gradient flow to converge to a linear classifier with suboptimal margin by adding only 3 extra data points. A critical assumption for our convergence analysis is the linear separability of data. We left it as a future work to study simplicity bias and global margin maximization without assuming linear separability.

Acknowledgments and Disclosure of Funding

The authors acknowledge support from NSF, ONR, Simons Foundation, DARPA and SRC. ZL is also supported by Microsoft Research PhD Fellowship.

References

- Zeyuan Allen-Zhu, Yuanzhi Li, and Yingyu Liang. Learning and generalization in overparameterized neural networks, going beyond two layers, 2018.
- Zeyuan Allen-Zhu, Yuanzhi Li, and Zhao Song. A convergence theory for deep learning via over-parameterization. In Kamalika Chaudhuri and Ruslan Salakhutdinov, editors, *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pages 242–252, Long Beach, California, USA, 09–15 Jun 2019. PMLR.

- Sanjeev Arora, Nadav Cohen, Wei Hu, and Yuping Luo. Implicit regularization in deep matrix factorization. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems 32*, pages 7411–7422. Curran Associates, Inc., 2019a.
- Sanjeev Arora, Simon Du, Wei Hu, Zhiyuan Li, and Ruosong Wang. Fine-grained analysis of optimization and generalization for overparameterized two-layer neural networks. In Kamalika Chaudhuri and Ruslan Salakhutdinov, editors, *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pages 322–332. PMLR, 09–15 Jun 2019b.
- Peter L Bartlett, Dylan J Foster, and Matus J Telgarsky. Spectrally-normalized margin bounds for neural networks. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems 30*, pages 6240–6249. Curran Associates, Inc., 2017.
- Jérôme Bolte, Aris Daniilidis, Olivier Ley, and Laurent Mazet. Characterizations of Łojasiewicz inequalities: subgradient flows, talweg, convexity. *Transactions of the American Mathematical Society*, 362(6):3319–3363, 2010.
- Alon Brutzkus and Amir Globerson. Why do larger models generalize better? A theoretical perspective via the XOR problem. In Kamalika Chaudhuri and Ruslan Salakhutdinov, editors, *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pages 822–830. PMLR, 09–15 Jun 2019.
- Alon Brutzkus, Amir Globerson, Eran Malach, and Shai Shalev-Shwartz. SGD learns over-parameterized networks that provably generalize on linearly separable data. In *International Conference on Learning Representations*, 2018.
- Zixiang Chen, Yuan Cao, Difan Zou, and Quanquan Gu. How much over-parameterization is sufficient to learn deep ReLU networks? In *International Conference on Learning Representations*, 2021.
- Lénaïc Chizat and Francis Bach. Implicit bias of gradient descent for wide two-layer neural networks trained with the logistic loss. In Jacob Abernethy and Shivani Agarwal, editors, *Proceedings of Thirty Third Conference on Learning Theory*, volume 125 of *Proceedings of Machine Learning Research*, pages 1305–1338. PMLR, 09–12 Jul 2020.
- Lénaïc Chizat, Edouard Oyallon, and Francis Bach. On lazy training in differentiable programming. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems 32*, pages 2937–2947. Curran Associates, Inc., 2019.
- Francis H. Clarke, Yuri S. Ledyaev, Ronald J. Stern, and Peter R. Wolenski. *Nonsmooth analysis and control theory*, volume 178. Springer Science & Business Media, 2008.
- Frank H. Clarke. Generalized gradients and applications. *Transactions of the American Mathematical Society*, 205:247–262, 1975.
- Michel Coste. *An introduction to ϕ -minimal geometry*. Istituti editoriali e poligrafici internazionali Pisa, 2000.
- Damek Davis, Dmitriy Drusvyatskiy, Sham Kakade, and Jason D. Lee. Stochastic subgradient method converges on tame functions. *Foundations of Computational Mathematics*, 20(1):119–154, Feb 2020.
- Simon Du, Jason Lee, Haochuan Li, Liwei Wang, and Xiyu Zhai. Gradient descent finds global minima of deep neural networks. In Kamalika Chaudhuri and Ruslan Salakhutdinov, editors, *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pages 1675–1685, Long Beach, California, USA, 09–15 Jun 2019a. PMLR.

- Simon S. Du, Wei Hu, and Jason D. Lee. Algorithmic regularization in learning deep homogeneous models: Layers are automatically balanced. In S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett, editors, *Advances in Neural Information Processing Systems 31*, pages 382–393. Curran Associates, Inc., 2018.
- Simon S. Du, Xiyu Zhai, Barnabas Poczos, and Aarti Singh. Gradient descent provably optimizes over-parameterized neural networks. In *International Conference on Learning Representations*, 2019b.
- Joydeep Dutta, Kalyanmoy Deb, Rupesh Tulshyan, and Ramnik Arora. Approximate KKT points and a proximity measure for termination. *Journal of Global Optimization*, 56(4):1463–1499, 2013.
- Spencer Frei, Yuan Cao, and Quanquan Gu. Provable generalization of sgd-trained neural networks of any width in the presence of adversarial label noise. In Marina Meila and Tong Zhang, editors, *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 3427–3438. PMLR, 18–24 Jul 2021.
- Gauthier Gidel, Francis Bach, and Simon Lacoste-Julien. Implicit regularization of discrete gradient dynamics in linear neural networks. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d’Alché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems 32*, pages 3196–3206. Curran Associates, Inc., 2019.
- Daniel Gissin, Shai Shalev-Shwartz, and Amit Daniely. The implicit bias of depth: How incremental learning drives generalization. In *International Conference on Learning Representations*, 2020.
- Suriya Gunasekar, Jason Lee, Daniel Soudry, and Nathan Srebro. Characterizing implicit bias in terms of optimization geometry. In Jennifer Dy and Andreas Krause, editors, *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pages 1832–1841, Stockholmsmässan, Stockholm Sweden, 10–15 Jul 2018a. PMLR.
- Suriya Gunasekar, Jason D Lee, Daniel Soudry, and Nati Srebro. Implicit bias of gradient descent on linear convolutional networks. In S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett, editors, *Advances in Neural Information Processing Systems 31*, pages 9482–9491. Curran Associates, Inc., 2018b.
- Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Delving deep into rectifiers: Surpassing human-level performance on ImageNet classification. In *The IEEE International Conference on Computer Vision (ICCV)*, December 2015.
- Wei Hu, Lechao Xiao, Ben Adlam, and Jeffrey Pennington. The surprising simplicity of the early-time learning dynamics of neural networks. In H. Larochelle, M. Ranzato, R. Hadsell, M. F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 17116–17128. Curran Associates, Inc., 2020.
- Arthur Jacot, Franck Gabriel, and Clement Hongler. Neural tangent kernel: Convergence and generalization in neural networks. In S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett, editors, *Advances in Neural Information Processing Systems 31*, pages 8571–8580. Curran Associates, Inc., 2018.
- Ziwei Ji and Matus Telgarsky. Risk and parameter convergence of logistic regression. *arXiv preprint arXiv:1803.07300*, 2018.
- Ziwei Ji and Matus Telgarsky. Gradient descent aligns the layers of deep linear networks. In *International Conference on Learning Representations*, 2019a.
- Ziwei Ji and Matus Telgarsky. The implicit bias of gradient descent on nonseparable data. In Alina Beygelzimer and Daniel Hsu, editors, *Proceedings of the Thirty-Second Conference on Learning Theory*, volume 99 of *Proceedings of Machine Learning Research*, pages 1772–1798, Phoenix, USA, 25–28 Jun 2019b. PMLR.

- Ziwei Ji and Matus Telgarsky. Directional convergence and alignment in deep learning. In H. Larochelle, M. Ranzato, R. Hadsell, M. F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 17176–17186. Curran Associates, Inc., 2020a.
- Ziwei Ji and Matus Telgarsky. Polylogarithmic width suffices for gradient descent to achieve arbitrarily small test error with shallow relu networks. In *International Conference on Learning Representations*, 2020b.
- Yiding Jiang, Behnam Neyshabur, Hossein Mobahi, Dilip Krishnan, and Samy Bengio. Fantastic generalization measures and where to find them. In *International Conference on Learning Representations*, 2020.
- Dimitris Kalimeris, Gal Kaplun, Preetum Nakkiran, Benjamin Edelman, Tristan Yang, Boaz Barak, and Haofeng Zhang. SGD on neural networks learns functions of increasing complexity. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019.
- Yuanzhi Li, Tengyu Ma, and Hongyang Zhang. Algorithmic regularization in over-parameterized matrix sensing and neural networks with quadratic activations. In Sébastien Bubeck, Vianney Perchet, and Philippe Rigollet, editors, *Proceedings of the 31st Conference On Learning Theory*, volume 75 of *Proceedings of Machine Learning Research*, pages 2–47. PMLR, 06–09 Jul 2018.
- Zhiyuan Li, Yuping Luo, and Kaifeng Lyu. Towards resolving the implicit bias of gradient descent for matrix factorization: Greedy low-rank learning. In *International Conference on Learning Representations*, 2021.
- Tao Luo, Zhi-Qin John Xu, Zheng Ma, and Yaoyu Zhang. Phase diagram for two-layer relu neural networks at infinite-width limit. *Journal of Machine Learning Research*, 22(71):1–47, 2021.
- Kaifeng Lyu and Jian Li. Gradient descent maximizes the margin of homogeneous neural networks. In *International Conference on Learning Representations*, 2020.
- Andrew L. Maas, Awni Y. Hannun, and Andrew Y. Ng. Rectifier nonlinearities improve neural network acoustic models. In *ICML Workshop on Deep Learning for Audio, Speech and Language Processing*, 2013.
- Hartmut Maennel, Olivier Bousquet, and Sylvain Gelly. Gradient descent quantizes relu network features. *arXiv preprint arXiv:1803.08367*, 2018.
- Harsh Mehta, Ashok Cutkosky, and Behnam Neyshabur. Extreme memorization via scale of initialization. In *International Conference on Learning Representations*, 2021.
- Edward Moroshko, Blake E Woodworth, Suriya Gunasekar, Jason D Lee, Nati Srebro, and Daniel Soudry. Implicit bias in deep linear classification: Initialization scale vs training accuracy. In H. Larochelle, M. Ranzato, R. Hadsell, M. F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 22182–22193. Curran Associates, Inc., 2020.
- Mor Shpigel Nacson, Jason Lee, Suriya Gunasekar, Pedro Henrique Pamplona Savarese, Nathan Srebro, and Daniel Soudry. Convergence of gradient descent on separable data. In Kamalika Chaudhuri and Masashi Sugiyama, editors, *Proceedings of Machine Learning Research*, volume 89 of *Proceedings of Machine Learning Research*, pages 3420–3428. PMLR, 16–18 Apr 2019a.
- Mor Shpigel Nacson, Nathan Srebro, and Daniel Soudry. Stochastic gradient descent on separable data: Exact convergence with a fixed learning rate. In Kamalika Chaudhuri and Masashi Sugiyama, editors, *Proceedings of Machine Learning Research*, volume 89 of *Proceedings of Machine Learning Research*, pages 3051–3059. PMLR, 16–18 Apr 2019b.
- Behnam Neyshabur, Srinadh Bhojanapalli, and Nathan Srebro. A PAC-bayesian approach to spectrally-normalized margin bounds for neural networks. In *International Conference on Learning Representations*, 2018.

- Franco Pellegrini and Giulio Biroli. An analytic theory of shallow networks dynamics for hinge loss classification. In H. Larochelle, M. Ranzato, R. Hadsell, M. F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 5356–5367. Curran Associates, Inc., 2020.
- Mary Phuong and Christoph H Lampert. The inductive bias of ReLU networks on orthogonally separable data. In *International Conference on Learning Representations*, 2021.
- Noam Razin, Asaf Maman, and Nadav Cohen. Implicit regularization in tensor factorization. In Marina Meila and Tong Zhang, editors, *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 8913–8924. PMLR, 18–24 Jul 2021.
- Roei Sarussi, Alon Brutzkus, and Amir Globerson. Towards understanding learning in neural networks with linear teachers. In Marina Meila and Tong Zhang, editors, *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 9313–9322. PMLR, 18–24 Jul 2021.
- Harshay Shah, Kaustav Tamuly, Aditi Raghunathan, Prateek Jain, and Praneeth Netrapalli. The pitfalls of simplicity bias in neural networks. In H. Larochelle, M. Ranzato, R. Hadsell, M. F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 9573–9585. Curran Associates, Inc., 2020.
- Daniel Soudry, Elad Hoffer, Mor Shpigel Nacson, Suriya Gunasekar, and Nathan Srebro. The implicit bias of gradient descent on separable data. *Journal of Machine Learning Research*, 19(70):1–57, 2018a.
- Daniel Soudry, Elad Hoffer, and Nathan Srebro. The implicit bias of gradient descent on separable data. In *International Conference on Learning Representations*, 2018b.
- Francis Williams, Matthew Trager, Daniele Panozzo, Claudio Silva, Denis Zorin, and Joan Bruna. Gradient dynamics of shallow univariate relu networks. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d’Alché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019.
- Blake Woodworth, Suriya Gunasekar, Jason D. Lee, Edward Moroshko, Pedro Savarese, Itay Golan, Daniel Soudry, and Nathan Srebro. Kernel and rich regimes in overparametrized models. In Jacob Abernethy and Shivani Agarwal, editors, *Proceedings of Thirty Third Conference on Learning Theory*, volume 125 of *Proceedings of Machine Learning Research*, pages 3635–3673. PMLR, 09–12 Jul 2020.
- Zhi-Qin John Xu, Hanxu Zhou, Tao Luo, and Yaoyu Zhang. Towards understanding the condensation of two-layer neural networks at initial training. *arXiv preprint arXiv:2105.11686*, 2021.
- Chiyuan Zhang, Samy Bengio, Moritz Hardt, Benjamin Recht, and Oriol Vinyals. Understanding deep learning requires rethinking generalization. In *International Conference on Learning Representations*, 2017.
- Difan Zou, Yuan Cao, Dongruo Zhou, and Quanquan Gu. Stochastic gradient descent optimizes over-parameterized deep relu networks. *arXiv preprint arXiv:1811.08888*, 2018.

A Theorem Statements for the Non-symmetric Case

A.1 Assumptions and Main Theorems

For every $\mathbf{x}_i$, define $\mathbf{x}_i^+ := \mathbf{x}_i$ if $y_i = 1$ and $\mathbf{x}_i^+ := \alpha_{\text{leaky}} \mathbf{x}_i$ if $y_i = -1$. Similarly, we define $\mathbf{x}_i^- := \alpha_{\text{leaky}} \mathbf{x}_i$ if $y_i = 1$ and $\mathbf{x}_i^- := \mathbf{x}_i$ if $y_i = -1$. Then we define $\boldsymbol{\mu}^+$ to be the mean vector of $y_i \mathbf{x}_i^+$, and $\boldsymbol{\mu}^-$ to be the mean vector of $y_i \mathbf{x}_i^-$, that is,

$$\boldsymbol{\mu}^+ := \frac{1}{n} \sum_{i \in [n]} y_i \mathbf{x}_i^+, \quad \boldsymbol{\mu}^- := \frac{1}{n} \sum_{i \in [n]} y_i \mathbf{x}_i^-. \quad (6)$$

Theorem 6.2 is indeed a simple corollary of Theorem A.7 below which holds for a broader class of datasets. Now we illustrate the assumptions one by one.

We first make the following assumption saying that there is a principal direction $\mathbf{w}^\diamond \in \mathbb{S}^{d-1}$ such that data points on average have much larger inner products with $\mathbf{w}^\diamond$ than any other direction perpendicular to $\mathbf{w}^\diamond$. This ensures at small initialization, the moving direction of each neurons lies in a small cone around the the direction of $\pm \mathbf{w}^\diamond$, and thus will converge to that cone eventually. The opening angle of this small cone is $2 \arcsin \frac{\gamma^\diamond}{\max_{i \in [n]} \|\mathbf{x}_i\|_2}$, which ensures the sign pattern inside the cone $\{\langle \mathbf{w}, \mathbf{x}_i \rangle\}_{i=1}^n$ is unique and indeed equal to $\{y_i\}_{i=1}^n$, and thus all neurons converge to two directions, $\boldsymbol{\mu}^+$ and $\boldsymbol{\mu}^-$ (defined in (6)).

Assumption A.1 (Existence of Principal Direction). There exists a unit-norm vector $\mathbf{w}^\diamond$ such that $\gamma^\diamond := \min_{i \in [n]} y_i \langle \mathbf{w}^\diamond, \mathbf{x}_i \rangle > 0$ and

$$\frac{\frac{1}{n} \sum_{i \in [n]} \|\mathbf{P}^\diamond \mathbf{x}_i\|_2}{\alpha_{\text{leaky}} \langle \boldsymbol{\mu}, \mathbf{w}^\diamond \rangle} < \frac{\gamma^\diamond}{\max_{i \in [n]} \|\mathbf{P}^\diamond \mathbf{x}_i\|_2},$$

where $\mathbf{P}^\diamond := \mathbf{I} - \mathbf{w}^\diamond \mathbf{w}^{\diamond \top}$ is the projection matrix onto the space perpendicular to $\mathbf{w}^\diamond$, and $\boldsymbol{\mu} := \frac{1}{n} \sum_{i \in [n]} y_i \mathbf{x}_i$ is the mean vector of $y_i \mathbf{x}_i$.

Indeed, our main theorem is based on a weaker assumption than Assumption A.1, which is Assumption A.2 below, but the geometric meaning of Assumption A.2 is not as clear as Assumption A.1. We will show in Lemma G.1 that Assumption A.1 implies Assumption A.2.

Assumption A.2. For all $i \in [n]$, we have

$$\langle \boldsymbol{\mu}, y_i \mathbf{x}_i \rangle > \frac{1 - \alpha_{\text{leaky}}}{n \cdot \alpha_{\text{leaky}}} \sum_{j \in [n]} \max\{-\langle y_i \mathbf{x}_i, y_j \mathbf{x}_j \rangle, 0\}.$$

In general, the norms $\|\boldsymbol{\mu}^+\|_2$ and $\|\boldsymbol{\mu}^-\|_2$ should not be equal: for any given dataset $\mathcal{S}$, we can make $\|\boldsymbol{\mu}^+\|_2 \neq \|\boldsymbol{\mu}^-\|_2$ by adding arbitrarily small perturbations to the data points. This motivates us to assume that $\|\boldsymbol{\mu}^+\|_2 \neq \|\boldsymbol{\mu}^-\|_2$. Without loss of generality, we can assume that $\|\boldsymbol{\mu}^+\|_2 > \|\boldsymbol{\mu}^-\|_2$ for convenience (Assumption A.3). When the reverse is true, i.e., $\|\boldsymbol{\mu}^+\|_2 < \|\boldsymbol{\mu}^-\|_2$, we can change the direction of the inequality by flipping all the labels in the dataset so that our theorems can apply. We include the theorem statements for this reversed case in Appendix A.3.

Assumption A.3. The norm of $\boldsymbol{\mu}^+$ is strictly larger than $\boldsymbol{\mu}^-$, i.e., $\|\boldsymbol{\mu}^+\|_2 > \|\boldsymbol{\mu}^-\|_2$.

Now we define $\mathbf{w}^+$ to be the max-margin linear separator of the dataset consisting of $(\mathbf{x}_i^+, y_i)$, where $i \in [n]$, and define γ^+ to be this max margin. That is,

$$\mathbf{w}^+ := \arg \max_{\mathbf{w} \in \mathbb{S}^{d-1}} \left\{ \min_{i \in [n]} y_i \langle \mathbf{w}, \mathbf{x}_i^+ \rangle \right\}, \quad \gamma^+ := \max_{\mathbf{w} \in \mathbb{S}^{d-1}} \left\{ \min_{i \in [n]} y_i \langle \mathbf{w}, \mathbf{x}_i^+ \rangle \right\}.$$

The reason that we care about $\mathbf{w}^+$ and γ^+ is because that it can be related to margin maximization on one-neuron Leaky ReLU nets. The following lemma is easy to prove.

Lemma A.4. For $m = 1$, if $\boldsymbol{\theta} = (\mathbf{w}_1, a_1) \in \mathbb{S}^{D-1}$ is a KKT-margin direction and $a_1 \geq 0$, then $\boldsymbol{\theta} = (\frac{1}{\sqrt{2}} \mathbf{w}^+, \frac{1}{\sqrt{2}})$, and it attains the global max margin $\frac{1}{2} \gamma^+$.

The third assumption we made is that this margin cannot be obtained when all a_i are negative, regardless of the width. This assumption holds when all the support vectors $\mathbf{x}_i^+$ have positive labels, i.e., $y_i = 1$. Conceptually, this assumption is about whether nearly all the support vectors have positive labels (or negative labels in the reversed case where $\|\boldsymbol{\mu}^+\|_2 < \|\boldsymbol{\mu}^-\|_2$).

Assumption A.5. For any $m \geq 1$ and any $\boldsymbol{\theta} = (\mathbf{w}_1, \dots, \mathbf{w}_m, a_1, \dots, a_m) \in \mathbb{R}^D$, if $a_k \leq 0$ for all $k \in [m]$, then the normalized margin $\gamma(\boldsymbol{\theta})$ on the dataset $\{(\mathbf{x}_i, y_i) : i \in [n], y_i \langle \mathbf{w}^+, \mathbf{x}_i^+ \rangle = \gamma^+\}$ is less than $\frac{1}{2}\gamma^+$.

Similar to Assumption 4.6 in the symmetric case, we need Assumption A.6 on non-branching starting point due to the technical difficulty for the potential non-uniqueness of gradient flow trajectory.

Assumption A.6. For any $m \geq 1$, there exist $r, \epsilon > 0$ such that $\boldsymbol{\theta}$ is a non-branching starting point if $a_k = \|\mathbf{w}_k\|_2 \in (0, r)$ holds for all $k \in [m]$, and all $\mathbf{w}_k$ point to the same direction $\mathbf{v} \in \mathbb{S}^{d-1}$ with $\left\| \mathbf{v} - \frac{\boldsymbol{\mu}^+}{\|\boldsymbol{\mu}^+\|_2} \right\|_2 \leq \epsilon$.

Now we are ready to state our theorem, and we defer the proofs to Appendix G.

Theorem A.7. Under Assumptions 3.2, A.2, A.3, A.5 and A.6, consider gradient flow on a Leaky ReLU network with width $m \geq 1$ and initialization $\boldsymbol{\theta}_0 = \sigma_{\text{init}} \bar{\boldsymbol{\theta}}_0$ where $\bar{\boldsymbol{\theta}}_0 \sim \tilde{\mathcal{D}}_{\text{init}}(1)$. With probability $1 - 2^{-m}$ over the draw of $\bar{\boldsymbol{\theta}}_0$, if the initialization scale is sufficiently small, then gradient flow directionally converges and $f^\infty(\mathbf{x}) := \lim_{t \rightarrow +\infty} f_{\boldsymbol{\theta}(t)/\|\boldsymbol{\theta}(t)\|_2}(\mathbf{x})$ represents the one-Leaky-ReLU classifier $\frac{1}{2}\phi(\langle \mathbf{w}^+, \mathbf{x} \rangle)$ with linear decision boundary. That is,

$$\Pr_{\bar{\boldsymbol{\theta}}_0 \sim \mathcal{D}_{\text{init}}(1)} \left[\exists \sigma_{\text{init}}^{\max} > 0 \text{ s.t. } \forall \sigma_{\text{init}} < \sigma_{\text{init}}^{\max}, \forall \mathbf{x} \in \mathbb{R}^d, f^\infty(\mathbf{x}) = \frac{1}{2}\phi(\langle \mathbf{w}^+, \mathbf{x} \rangle) \right] \geq 1 - 2^{-m}.$$

A.2 Applying Theorem A.7 to prove Theorem 6.2

We give a proof of Theorem 6.2 here given the result of Theorem A.7.

Proof. With a $(H, K, \epsilon, \mathbf{w}_\perp)$ -Hinted Dataset (Definition 6.1) with proper H, K, ϵ , we only need to show that Assumptions A.2, A.3 and A.5 hold for Theorem 6.2. Specifically, we choose the parameters such that

- $K > 0$;
- $\epsilon < \alpha_{\text{leaky}} \min_{i>3} y_i \langle \mathbf{w}^*, \mathbf{x}_i \rangle$;
- $H > \max\{\epsilon, H_0, n \|\boldsymbol{\mu}^-\|_2 + \|\sum_{j>1} y_j \mathbf{x}_j^+\|_2\}$, where $H_0 = \frac{\max_{i \in [n]} \|\mathbf{P}^* \mathbf{x}_i\|_2 \sum_{i \in [n]} \|\mathbf{P}^* \mathbf{x}_i\|_2}{\alpha_{\text{leaky}} \min_{i>1} \langle y_i \mathbf{x}_i, \mathbf{w}^* \rangle} - \sum_{i>1} \langle \mathbf{w}^*, y_i \mathbf{x}_i \rangle$ and $\mathbf{P}^* = \mathbf{I} - \mathbf{w}^* \mathbf{w}^{*\top}$ is the projection matrix onto the orthogonal space of $\mathbf{w}^*$.

Notice that H_0 is independent of H as the data point $\mathbf{x}_1$ has projection $\|\mathbf{P}^* \mathbf{x}_1\|_2 = 0$. For Assumption A.1, $\mathbf{w}^\diamond = \mathbf{w}^*$ is a valid principal direction in this case, as

$$\max_{i \in [n]} \|\mathbf{P}^\diamond \mathbf{x}_i\|_2 \frac{\frac{1}{n} \sum_{i \in [n]} \|\mathbf{P}^\diamond \mathbf{x}_i\|_2}{\alpha_{\text{leaky}} \gamma^\diamond} = \frac{1}{n} (H_0 + \sum_{i>1} \langle \mathbf{w}^*, y_i \mathbf{x}_i \rangle) < \langle \boldsymbol{\mu}, \mathbf{w}^\diamond \rangle.$$

Then Assumption A.2 follows from Assumption A.1 by Lemma G.1. Since $H > n \|\boldsymbol{\mu}^-\|_2 + \|\sum_{j>1} y_j \mathbf{x}_j^+\|_2$,

$$\|\boldsymbol{\mu}^+\|_2 \geq \frac{1}{n} H - \left\| \frac{1}{n} \sum_{j>1} y_j \mathbf{x}_j^+ \right\|_2 > \|\boldsymbol{\mu}^-\|_2,$$

and thus Assumption A.3 holds. Furthermore, with $\epsilon < \alpha_{\text{leaky}} \min_{i>3} y_i \langle \mathbf{w}^*, \mathbf{x}_i \rangle$ and $H > \epsilon$, $(\mathbf{x}_2, y_2) = (\epsilon \mathbf{w}^* + K \mathbf{w}_\perp, 1)$ and $(\mathbf{x}_3, y_3) = (\epsilon \mathbf{w}^* - K \mathbf{w}_\perp, 1)$ are the only support vectors for the linear margin problem on $\{(\mathbf{x}_i, y_i)\}$ and that on $\{(\mathbf{x}_i^+, y_i)\}$ as well. Then $\mathbf{w}^+ = \mathbf{w}^*$ and $\gamma^+ = \epsilon$. For a neuron with $a_k < 0$, the total output margin on the hints $(\mathbf{x}_2, y_2)$ and $(\mathbf{x}_3, y_3)$ is $a_k \phi(\mathbf{w}_k^\top \mathbf{x}_2) + a_k \phi(\mathbf{w}_k^\top \mathbf{x}_3) \leq 2\alpha_{\text{leaky}} \epsilon |a_k| \|\mathbf{w}_k\|_2 \leq \alpha_{\text{leaky}} \epsilon (a_k^2 + \|\mathbf{w}_k\|_2^2)$. Thus the normalized margin for multiple such neurons is at most $\frac{\alpha_{\text{leaky}} \epsilon}{2} < \frac{\epsilon}{2}$, so Assumption A.5 will also be true. $\square$

A.3 Results in the Reversed Case

In a reversed case where $\|\mu^+\|_2 < \|\mu^-\|_2$, we can apply Theorem A.7 by flipping the labels in the dataset. Below we state the assumptions and the theorem in the reversed case.

Assumption A.8. $\|\mu^+\|_2 < \|\mu^-\|_2$.

Now similarly we define w^- and γ^- .

$$w^- := \arg \max_{w \in \mathbb{S}^{d-1}} \left\{ \min_{i \in [n]} y_i \langle w, x_i^- \rangle \right\}, \quad \gamma^- := \max_{w \in \mathbb{S}^{d-1}} \left\{ \min_{i \in [n]} y_i \langle w, x_i^- \rangle \right\}.$$

Assumption A.9. For any $m \geq 1$ and any $\theta = (w_1, \dots, w_m, a_1, \dots, a_m) \in \mathbb{R}^D$, if $a_k \leq 0$ for all $k \in [m]$, then the normalized margin $\gamma(\theta)$ on the dataset $\{(x_i, y_i) : i \in [n], y_i \langle w^-, x_i^- \rangle = \gamma^-\}$ is less than $\frac{1}{2}\gamma^-$.

Theorem A.10. Under Assumptions 3.2, A.2, A.6, A.8 and A.9, consider gradient flow on a Leaky ReLU network with width $m \geq 1$ and initialization $\theta_0 = \sigma_{\text{init}} \bar{\theta}_0$ where $\bar{\theta}_0 \sim \tilde{\mathcal{D}}_{\text{init}}(1)$. With probability $1 - 2^{-m}$ over the draw of θ_0 , there is an sufficiently small initialization scale, such that gradient flow directionally converges and $f^\infty(x) := \lim_{t \rightarrow +\infty} f_{\theta(t)/\|\theta(t)\|_2}(x)$ represents the one-Leaky-ReLU classifier $-\frac{1}{2}\phi(-\langle w^-, x \rangle)$ with linear decision boundary. That is,

$$\Pr_{\bar{\theta}_0 \sim \tilde{\mathcal{D}}_{\text{init}}(1)} \left[\exists \sigma_{\text{init}}^{\max} > 0 \text{ s.t. } \forall \sigma_{\text{init}} < \sigma_{\text{init}}^{\max}, \forall x \in \mathbb{R}^d, f^\infty(x) = -\frac{1}{2}\phi(-\langle w^-, x \rangle) \right] \geq 1 - 2^{-m}.$$

B Additional Preliminaries and Lemmas

In this section, we will introduce additional notations and give some preliminary results for the dynamics of the two-layer Leaky ReLU network. The only assumption we will use for the results in the section is that the input norm is bounded $\max_{i \in [n]} \|x_i\|_2 \leq 1$ and we do not assume other properties of the dataset (such as symmetry) except we assume it explicitly.

B.1 Additional Notations

For notational convenience for calculation with subgradients, we generalize the following notations for vectors to vector sets. More specifically, we define

- $\forall A, B \subseteq \mathbb{R}^d, A + B := \{x + y : x \in A, y \in B\}$ and $A - B := A + (-B)$;
- $\forall A \subseteq \mathbb{R}^d, \lambda \in \mathbb{R}, \lambda A := \{\lambda x : x \in A\}$;
- Let $\|\cdot\|$ be any norm on $\mathbb{R}^d, \forall A \subseteq \mathbb{R}^d, \|A\| := \{\|x\| : x \in A\} \subseteq \mathbb{R}$;
- $\forall A \subseteq \mathbb{R}^d$ and $y \in \mathbb{R}^d, \langle y, A \rangle \equiv \langle A, y \rangle := \{\langle x, y \rangle : x \in A\}$;
- We use $\text{dist}(x, y) := \|x - y\|_2$ to denote the L^2 -distance between $x \in \mathbb{R}^d$ and $y \in \mathbb{R}^d$, $\text{dist}(A, y) := \inf_{x \in A} \|x - y\|_2$ to denote the minimum L^2 -distance between any $x \in A$ and $y \in \mathbb{R}^d$, and $\text{dist}(A, B) := \inf_{x \in A, y \in B} \|x - y\|_2$ to denote the minimum L^2 -distance between any $x \in A$ and any $y \in B$.

By Rademacher theorem, any real-valued locally Lipschitz function on $\mathbb{R}^D$ is differentiable almost everywhere (a.e.) in the sense of Lebesgue measure. For a locally Lipschitz function $\mathcal{L} : \mathbb{R}^D \rightarrow \mathbb{R}$, we use $\nabla \mathcal{L}(\theta) \in \mathbb{R}^D$ to denote the usual gradient (if $\mathcal{L}$ is differentiable at θ) and $\partial^\circ \mathcal{L}(\theta) \subseteq \mathbb{R}^D$ to denote Clarke's subdifferential. The definition of Clarke's subdifferential is given by (7): for any sequence of differentiable points converging to θ , we collect convergent gradients from such sequences and take the convex hull as the Clarke's subdifferential at θ .

$$\partial^\circ \mathcal{L}(\theta) := \text{conv} \left\{ \lim_{n \rightarrow \infty} \nabla \mathcal{L}(\theta_n) : \mathcal{L} \text{ differentiable at } \theta_n, \lim_{n \rightarrow \infty} \theta_n = \theta \right\}. \quad (7)$$

For any full measure set $\Omega \subseteq \mathbb{R}^D$ that does not contain any non-differentiable points, (7) also has the following equivalent form:

$$\partial^\circ \mathcal{L}(\theta) = \text{conv} \left\{ \lim_{n \rightarrow \infty} \nabla \mathcal{L}(\theta_n) : \theta_n \in \Omega \text{ for all } n \text{ and } \lim_{n \rightarrow \infty} \theta_n = \theta \right\}. \quad (8)$$

The Clarke's subdifferential $\partial^\circ \mathcal{L}(\boldsymbol{\theta})$ is convex compact if $\mathcal{L}$ is locally Lipschitz, and it is upper-semicontinuous with respect to $\boldsymbol{\theta}$ (or equivalently it has closed graph) if $\mathcal{L}$ is definable. We use $\bar{\partial}^\circ \mathcal{L}(\boldsymbol{\theta}) \in \mathbb{R}^D$ to denote the min-norm gradient vector in the Clarke's subdifferential at $\boldsymbol{\theta}$, i.e., $\bar{\partial}^\circ \mathcal{L}(\boldsymbol{\theta}) := \arg \min_{\mathbf{g} \in \partial^\circ \mathcal{L}(\boldsymbol{\theta})} \|\mathbf{g}\|_2$. If $\mathcal{L}$ is continuously differentiable at $\boldsymbol{\theta}$, then $\partial^\circ \mathcal{L}(\boldsymbol{\theta}) = \{\nabla \mathcal{L}(\boldsymbol{\theta})\}$ and $\bar{\partial}^\circ \mathcal{L}(\boldsymbol{\theta}) = \nabla \mathcal{L}(\boldsymbol{\theta})$.

If $\boldsymbol{\theta}$ can be written as $\boldsymbol{\theta} = (\boldsymbol{\theta}_1, \boldsymbol{\theta}_2) \in \mathbb{R}^{D_1} \times \mathbb{R}^{D_2}$, then we use $\frac{\partial \mathcal{L}(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}_1} \in \mathbb{R}^{D_1}$ to denote the usual partial derivatives (partial gradient) and $\frac{\partial^\circ \mathcal{L}(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}_1} \subseteq \mathbb{R}^{D_1}$ to denote the partial subderivatives (partial subgradient) in the sense of Clarke.

Furthermore, we use the following notations to denote the radial and spherical components of $\bar{\partial}^\circ \mathcal{L}(\boldsymbol{\theta})$ (which will be used in analyzing Phase III):

$$\bar{\partial}_r^\circ \mathcal{L}(\boldsymbol{\theta}) := \frac{\boldsymbol{\theta} \boldsymbol{\theta}^\top}{\|\boldsymbol{\theta}\|_2^2} \bar{\partial}^\circ \mathcal{L}(\boldsymbol{\theta}), \quad \bar{\partial}_\perp^\circ \mathcal{L}(\boldsymbol{\theta}) := \left(\mathbf{I} - \frac{\boldsymbol{\theta} \boldsymbol{\theta}^\top}{\|\boldsymbol{\theta}\|_2^2} \right) \bar{\partial}^\circ \mathcal{L}(\boldsymbol{\theta}).$$

For univariate function $f : \mathbb{R} \rightarrow \mathbb{R}$, we use $f'(z) \in \mathbb{R}$ to denote the usual derivative (if f is differentiable at z) and $f^\circ(z) \subseteq \mathbb{R}$ to denote the Clarke's subdifferential.

The logistic loss is defined by $\ell(q) = \ln(1 + e^{-q})$, which satisfies $\ell(0) = \ln 2$, $\ell'(0) = -1/2$, $|\ell'(q)| \leq 1$, $|\ell''(q)| \leq 1$. Given a dataset $\mathcal{S} = \{(\mathbf{x}_1, y_1), \dots, (\mathbf{x}_n, y_n)\}$, we consider gradient flow on two-layer Leaky ReLU network with output function $f_\theta(\mathbf{x}_i)$ and logistic loss $\mathcal{L}(\boldsymbol{\theta}) := \frac{1}{n} \sum_{i \in [n]} \ell(q_i(\boldsymbol{\theta}))$, where $q_i(\boldsymbol{\theta}) := y_i f_\theta(\mathbf{x}_i)$. Following Davis et al. (2020); Lyu and Li (2020), we say that a function $\mathbf{z}(t) \in \mathbb{R}^D$ on an interval I is an *arc* if $\mathbf{z}$ is absolutely continuous on any compact subinterval of I . An arc $\boldsymbol{\theta}(t)$ is a trajectory of gradient flow on $\mathcal{L}$ if $\boldsymbol{\theta}(t)$ satisfies the following gradient inclusion for a.e. $t \geq 0$:

$$\frac{d\boldsymbol{\theta}(t)}{dt} \in -\partial^\circ \mathcal{L}(\boldsymbol{\theta}(t)).$$

Let $\Omega_{\mathcal{S}}$ be the set of parameter vectors $\boldsymbol{\theta} = (\mathbf{w}_1, \dots, \mathbf{w}_m, a_1, \dots, a_m)$ so that $\langle \mathbf{w}_k, \mathbf{x}_i \rangle \neq 0$ for all $i \in [n], k \in [m]$, i.e., no activation function has zero input. For any $\boldsymbol{\theta} \in \Omega_{\mathcal{S}}$, $f_\theta(\mathbf{x}_i)$ and $\mathcal{L}(\boldsymbol{\theta})$ are continuously differentiable at $\boldsymbol{\theta}$, and the gradients are given by

$$\frac{\partial f_\theta(\mathbf{x})}{\partial \mathbf{w}_k} = a_k \phi'(\mathbf{w}_k^\top \mathbf{x}_i) \mathbf{x}_i, \quad \frac{\partial f_\theta(\mathbf{x})}{\partial a_k} = \phi(\mathbf{w}_k^\top \mathbf{x}_i). \quad (9)$$

$$\frac{\partial \mathcal{L}(\boldsymbol{\theta})}{\partial \mathbf{w}_k} = \frac{1}{n} \sum_{i \in [n]} \ell'(q_i(\boldsymbol{\theta})) y_i a_k \phi'(\mathbf{w}_k^\top \mathbf{x}_i) \mathbf{x}_i, \quad \frac{\partial \mathcal{L}(\boldsymbol{\theta})}{\partial a_k} = \frac{1}{n} \sum_{i \in [n]} \ell'(q_i(\boldsymbol{\theta})) y_i \phi(\mathbf{w}_k^\top \mathbf{x}_i). \quad (10)$$

Then the Clarke's subdifferential for any $\boldsymbol{\theta}$ can be computed from (8) with $\Omega = \Omega_{\mathcal{S}}$ if needed.

Recall that G -function (Section 5.1) is defined by

$$G(\mathbf{w}) := \frac{-\ell'(0)}{n} \sum_{i \in [n]} y_i \phi(\mathbf{w}^\top \mathbf{x}_i) = \frac{1}{2n} \sum_{i \in [n]} y_i \phi(\mathbf{w}^\top \mathbf{x}_i).$$

Define $\tilde{\mathcal{L}}(\boldsymbol{\theta})$ to the linear approximation of $\mathcal{L}(\boldsymbol{\theta})$:

$$\tilde{\mathcal{L}}(\boldsymbol{\theta}) := \ell(0) - \sum_{k \in [m]} a_k G(\mathbf{w}_k).$$

For every $\boldsymbol{\theta}_0 \in \mathbb{R}^D$, we define $\varphi(\boldsymbol{\theta}_0, t)$ to be the value of $\boldsymbol{\theta}(t)$ for gradient flow on $\mathcal{L}(\boldsymbol{\theta})$ starting with $\boldsymbol{\theta}(0) = \boldsymbol{\theta}_0$. For every $\tilde{\boldsymbol{\theta}}_0 \in \mathbb{R}^D$, we define $\tilde{\varphi}(\tilde{\boldsymbol{\theta}}_0, t)$ to be the value of $\tilde{\boldsymbol{\theta}}(t)$ for gradient flow on $\tilde{\mathcal{L}}(\tilde{\boldsymbol{\theta}})$ starting with $\tilde{\boldsymbol{\theta}}(0) = \tilde{\boldsymbol{\theta}}_0$. In the case where the gradient flow trajectory may not be unique, we assign $\varphi(\boldsymbol{\theta}_0, \cdot)$ (or $\tilde{\varphi}(\tilde{\boldsymbol{\theta}}_0, \cdot)$) by an arbitrary trajectory of gradient flow on $\mathcal{L}$ (or $\tilde{\mathcal{L}}$) starting from $\boldsymbol{\theta}_0$ (or $\tilde{\boldsymbol{\theta}}_0$).

B.2 Grönwall's Inequality

We frequently use Grönwall's inequality in our analysis.

Lemma B.1 (Grönwall's Inequality). *Let α, β, u be real-valued functions defined on $[a, b]$. Suppose that β, u are continuous and $\min\{\alpha, 0\}$ is integrable on every compact subinterval of $[a, b]$. If $\beta \geq 0$ and u satisfies the following inequality for all $t \in [a, b]$:*

$$u(t) \leq \alpha(t) + \int_a^t \beta(\tau)u(\tau)d\tau,$$

then for all $t \in [a, b]$,

$$u(t) \leq \alpha(t) + \int_a^t \alpha(\tau)\beta(\tau) \exp\left(\int_\tau^t \beta(\tau')d\tau'\right) d\tau. \quad (11)$$

Furthermore, if α is non-decreasing, then for all $t \in [a, b]$,

$$u(t) \leq \alpha(t) \exp\left(\int_a^t \beta(\tau)d\tau\right). \quad (12)$$

B.3 Homogeneous Functions

For $L \geq 0$, we say that a function $f : \mathbb{R}^d \rightarrow \mathbb{R}$ is (positively) L -homogeneous if $f(c\theta) = c^L f(\theta)$ for all $c > 0$ and $\theta \in \mathbb{R}^d$. The proof for the following two theorems can be found in [Lyu and Li \(2020, Theorem B.2\)](#) and [Ji and Telgarsky \(2020a, Lemma C.1\)](#) respectively.

Theorem B.2. *For locally Lipschitz and L -homogeneous function $f : \mathbb{R}^d \rightarrow \mathbb{R}$, we have*

$$\partial^\circ f(c\theta) = c^{L-1} \partial^\circ f(\theta).$$

for all $\theta \in \mathbb{R}^d$.

Theorem B.3 (Euler's homogeneous function theorem). *For locally Lipschitz and L -homogeneous function $f : \mathbb{R}^d \rightarrow \mathbb{R}$, we have*

$$\forall g \in \partial^\circ f(\theta) : \quad \langle g, \theta \rangle = Lf(\theta),$$

for all $\theta \in \mathbb{R}^d$.

For the maximizer of a homogeneous function on $\mathbb{S}^{d-1}$, we have the following useful lemma.

Lemma B.4. *For locally Lipschitz and L -homogeneous function $f : \mathbb{R}^d \rightarrow \mathbb{R}$, if $\theta \in \mathbb{S}^{d-1}$ is a local/global maximizer of $f(\theta)$ on $\mathbb{S}^{d-1}$ and f is differentiable at θ , then $\nabla f(\theta) = Lf(\theta)\theta$.*

Proof. Since θ is a local/global maximizer of $f(\theta)$ on $\mathbb{S}^{d-1}$ and f is differentiable at θ , $\nabla f(\theta)$ is parallel to θ , i.e., $\nabla f(\theta) = c\theta$ for some $c \in \mathbb{R}$. By [Theorem B.3](#) we know that $\langle \nabla f(\theta), \theta \rangle = Lf(\theta)$. So $c = Lf(\theta)$. $\square$

The following is a direct corollary of [Lemma B.4](#).

Lemma B.5. *If $w \in \mathbb{S}^{d-1}$ attains the maximum of $|G(w)|$ on $\mathbb{S}^{d-1}$ and $G(w)$ is differentiable at w , then $\nabla G(w) = G(w)w$.*

Proof. Note that $G(w)$ is 1-homogeneous. If w attains the maximum of $|G(w)|$ on $\mathbb{S}^{d-1}$, then w is either a maximizer of $G(w)$ or $-G(w)$. Applying [Lemma B.4](#) gives $\nabla G(w) = G(w)w$. $\square$

B.4 Karush-Kuhn-Tucker Conditions for Margin Maximization

Definition B.6 (Feasible Point and KKT Point, [Dutta et al. 2013; Lyu and Li 2020](#)). Let $f, g_1, \dots, g_n : \mathbb{R}^D \rightarrow \mathbb{R}$ be locally Lipschitz functions. Consider the following constrained optimization problem for $\theta \in \mathbb{R}^D$:

$$\begin{aligned} \min \quad & f(\theta) \\ \text{s.t.} \quad & g_i(\theta) \leq 0, \quad \forall i \in [n]. \end{aligned}$$

We say that θ is a *feasible point* if $g_i(\theta) \leq 0$ for all $i \in [n]$. A feasible point θ is a *KKT point* if it satisfies Karush-Kuhn-Tucker Conditions: there exist $\lambda_1, \dots, \lambda_n \geq 0$ such that

1. $\mathbf{0} \in \partial^\circ f(\boldsymbol{\theta}) + \sum_{i \in [n]} \lambda_i \partial^\circ g_i(\boldsymbol{\theta})$;
2. $\forall i \in [n] : \lambda_i g_i(\boldsymbol{\theta}) = 0$.

Recall that we say that a parameter vector $\boldsymbol{\theta} \in \mathbb{R}^D$ of a L -homogeneous network is along a KKT-margin direction if $\frac{\boldsymbol{\theta}}{(q_{\min}(\boldsymbol{\theta}))^{1/L}}$ is a KKT point of (P), where $f(\boldsymbol{\theta}) = \frac{1}{2} \|\boldsymbol{\theta}\|_2^2$ and $g_i(\boldsymbol{\theta}) = 1 - q_i(\boldsymbol{\theta})$. Alternatively, we can use the following equivalent definition.

Definition B.7 (KKT-margin Direction for Homogeneous Network, [Lyu and Li 2020](#)). For a parameter vector $\boldsymbol{\theta} \in \mathbb{R}^D$ of a homogeneous network, we say $\boldsymbol{\theta}$ is along a KKT-margin direction if $q_i(\boldsymbol{\theta}) > 0$ for all $i \in [n]$ and there exist $\lambda_1, \dots, \lambda_n \geq 0$ such that

1. $\boldsymbol{\theta} \in \sum_{i \in [n]} \lambda_i \partial^\circ q_i(\boldsymbol{\theta})$;
2. For all $i \in [n]$, if $q_i(\boldsymbol{\theta}) \neq q_{\min}(\boldsymbol{\theta})$ then $\lambda_i = 0$.

For two-layer Leaky ReLU network, $q_i(\boldsymbol{\theta}) := y_i \sum_{k \in [m]} a_k \phi(\mathbf{w}_k^\top \mathbf{x}_i)$. Then the KKT-margin direction is defined as follows.

Definition B.8 (KKT-margin Direction for Two-layer Leaky ReLU Network). For a parameter vector $\boldsymbol{\theta} = (\mathbf{w}_1, \dots, \mathbf{w}_m, a_1, \dots, a_m) \in \mathbb{R}^D$ of a two-layer Leaky ReLU network, we say $\boldsymbol{\theta}$ is along a KKT-margin direction if $q_i(\boldsymbol{\theta}) > 0$ for all $i \in [n]$ and there exist $\lambda_1, \dots, \lambda_n \geq 0$ such that

1. For all $k \in [m]$, $\mathbf{w}_k \in \sum_{i \in [n]} \lambda_i y_i a_k \phi^\circ(\mathbf{w}_k^\top \mathbf{x}_i) \mathbf{x}_i$;
2. For all $k \in [m]$, $a_k = \sum_{i \in [n]} \lambda_i y_i \phi(\mathbf{w}_k^\top \mathbf{x}_i)$;
3. For all $i \in [n]$, if $q_i(\boldsymbol{\theta}) \neq q_{\min}(\boldsymbol{\theta})$ then $\lambda_i = 0$.

For $\boldsymbol{\theta}$ along a KKT-margin direction of two-layer Leaky ReLU network, Lemma B.9 below shows that $|a_k| = \|\mathbf{w}_k\|_2$ for all $k \in [m]$.

Lemma B.9. *If $\boldsymbol{\theta} = (\mathbf{w}_1, \dots, \mathbf{w}_m, a_1, \dots, a_m) \in \mathbb{R}^D$ is along a KKT-margin direction of a two-layer Leaky ReLU network, then $|a_k| = \|\mathbf{w}_k\|_2$ for all $k \in [m]$.*

Proof. By Definition B.8 and Theorem B.3, we have

$$\begin{aligned} \|\mathbf{w}_k\|_2^2 &\in \left\langle \mathbf{w}_k, \sum_{i \in [n]} \lambda_i y_i a_k \phi^\circ(\mathbf{w}_k^\top \mathbf{x}_i) \mathbf{x}_i \right\rangle = \left\{ \sum_{i \in [n]} \lambda_i y_i a_k \phi(\mathbf{w}_k^\top \mathbf{x}_i) \right\}, \\ |a_k|^2 &= a_k \cdot \sum_{i \in [n]} \lambda_i y_i \phi(\mathbf{w}_k^\top \mathbf{x}_i) = \sum_{i \in [n]} \lambda_i y_i a_k \phi(\mathbf{w}_k^\top \mathbf{x}_i). \end{aligned}$$

Therefore $\|\mathbf{w}_k\|_2^2 = |a_k|^2$. □

B.5 Lemmas for Perturbation Bounds

Recall that $\|\boldsymbol{\theta}\|_M$ is defined in Definition 5.1.

Lemma B.10. *For $\|\mathbf{x}\|_2 \leq 1$, $|f_\theta(\mathbf{x})| \leq m \|\boldsymbol{\theta}\|_M^2$, $|f_\theta(\mathbf{x}) - f_{\tilde{\theta}}(\mathbf{x})| \leq m \|\boldsymbol{\theta} - \tilde{\boldsymbol{\theta}}\|_M \left(\|\boldsymbol{\theta}\|_M + \|\tilde{\boldsymbol{\theta}}\|_M \right)$.*

Proof. The proof is straightforward by definition of $f_\theta(\mathbf{x})$ and $\|\boldsymbol{\theta}\|_M$. For the first inequality,

$$|f_\theta(\mathbf{x})| \leq \sum_{k=1}^m |a_k \phi(\mathbf{w}_k^\top \mathbf{x})| \leq \sum_{k=1}^m |a_k| \cdot |\mathbf{w}_k^\top \mathbf{x}| \leq \sum_{k=1}^m |a_k| \cdot \|\mathbf{w}_k\|_2 \leq m \|\boldsymbol{\theta}\|_M^2.$$

For the second inequality,

$$\begin{aligned}
|f_{\boldsymbol{\theta}}(\mathbf{x}) - f_{\tilde{\boldsymbol{\theta}}}(\mathbf{x})| &\leq \sum_{k=1}^m |a_k \phi(\mathbf{w}_k^\top \mathbf{x}) - \tilde{a}_k \phi(\tilde{\mathbf{w}}_k^\top \mathbf{x})| \\
&\leq \sum_{k=1}^m |a_k \phi(\mathbf{w}_k^\top \mathbf{x}) - a_k \phi(\tilde{\mathbf{w}}_k^\top \mathbf{x})| + |a_k \phi(\tilde{\mathbf{w}}_k^\top \mathbf{x}) - \tilde{a}_k \phi(\tilde{\mathbf{w}}_k^\top \mathbf{x})| \\
&\leq \sum_{k=1}^m |a_k| \cdot \|\mathbf{w}_k - \tilde{\mathbf{w}}_k\|_2 + |a_k - \tilde{a}_k| \cdot \|\tilde{\mathbf{w}}_k\|_2 \\
&\leq m \|\boldsymbol{\theta} - \tilde{\boldsymbol{\theta}}\|_{\mathbf{M}} \left(\|\boldsymbol{\theta}\|_{\mathbf{M}} + \|\tilde{\boldsymbol{\theta}}\|_{\mathbf{M}} \right),
\end{aligned}$$

which completes the proof. $\square$

We have the following bound for the difference between $\partial^\circ \mathcal{L}(\boldsymbol{\theta})$ and $\partial^\circ \tilde{\mathcal{L}}(\boldsymbol{\theta})$.

Lemma B.11. Assume that $\|\mathbf{x}_i\|_2 \leq 1$ for all $i \in [n]$. For any $\boldsymbol{\theta} = (\mathbf{w}_1, \dots, \mathbf{w}_m, a_1, \dots, a_m) \in \mathbb{R}^D$, we have the following bounds for the partial derivatives of $\mathcal{L}(\boldsymbol{\theta}) - \tilde{\mathcal{L}}(\boldsymbol{\theta})$:

$$\left\| \frac{\partial^\circ (\mathcal{L}(\boldsymbol{\theta}) - \tilde{\mathcal{L}}(\boldsymbol{\theta}))}{\partial \mathbf{w}_k} \right\|_2 \subseteq (-\infty, m \|\boldsymbol{\theta}\|_{\mathbf{M}}^2 |a_k|], \quad \left| \frac{\partial (\mathcal{L}(\boldsymbol{\theta}) - \tilde{\mathcal{L}}(\boldsymbol{\theta}))}{\partial a_k} \right| \leq m \|\boldsymbol{\theta}\|_{\mathbf{M}}^2 \|\mathbf{w}_k\|_2.$$

for all $k \in [m]$.

Proof. We only need to prove the following bounds for gradients at any $\boldsymbol{\theta} \in \Omega_{\mathcal{S}}$, i.e., $\langle \mathbf{w}_k, \mathbf{x}_i \rangle \neq 0$ for all $i \in [n], k \in [m]$. For the general case where $\boldsymbol{\theta}$ can be non-differentiable, we can prove the same bounds for Clarke's sub-differential at every point $\boldsymbol{\theta} \in \mathbb{R}^D$ by taking limits in $\Omega_{\mathcal{S}}$ through (8).

$$\left\| \frac{\partial (\mathcal{L}(\boldsymbol{\theta}) - \tilde{\mathcal{L}}(\boldsymbol{\theta}))}{\partial \mathbf{w}_k} \right\|_2 \leq m \|\boldsymbol{\theta}\|_{\mathbf{M}}^2 |a_k|, \quad \left| \frac{\partial (\mathcal{L}(\boldsymbol{\theta}) - \tilde{\mathcal{L}}(\boldsymbol{\theta}))}{\partial a_k} \right| \leq m \|\boldsymbol{\theta}\|_{\mathbf{M}}^2 \|\mathbf{w}_k\|_2.$$

By Taylor expansion, we have

$$\ell(y_i f_{\boldsymbol{\theta}}(\mathbf{x}_i)) = \ell(0) + \ell'(0) y_i f_{\boldsymbol{\theta}}(\mathbf{x}_i) + \int_0^{y_i f_{\boldsymbol{\theta}}(\mathbf{x}_i)} \ell''(z) (y_i f_{\boldsymbol{\theta}}(\mathbf{x}_i) - z) dz.$$

Taking average over $i \in [n]$ gives

$$\begin{aligned}
\mathcal{L}(\boldsymbol{\theta}) &= \ell(0) + \frac{1}{n} \sum_{i \in [n]} \ell'(0) y_i f_{\boldsymbol{\theta}}(\mathbf{x}_i) + \frac{1}{n} \sum_{i \in [n]} \int_0^{y_i f_{\boldsymbol{\theta}}(\mathbf{x}_i)} \ell''(z) (y_i f_{\boldsymbol{\theta}}(\mathbf{x}_i) - z) dz \\
&= \tilde{\mathcal{L}}(\boldsymbol{\theta}) + \frac{1}{n} \sum_{i \in [n]} \int_0^{y_i f_{\boldsymbol{\theta}}(\mathbf{x}_i)} \ell''(z) (y_i f_{\boldsymbol{\theta}}(\mathbf{x}_i) - z) dz.
\end{aligned}$$

By Leibniz integral rule,

$$\begin{aligned}
\nabla_{\boldsymbol{\theta}} (\mathcal{L}(\boldsymbol{\theta}) - \tilde{\mathcal{L}}(\boldsymbol{\theta})) &= \nabla_{\boldsymbol{\theta}} \left(\frac{1}{n} \sum_{i \in [n]} \int_0^{y_i f_{\boldsymbol{\theta}}(\mathbf{x}_i)} \ell''(z) (y_i f_{\boldsymbol{\theta}}(\mathbf{x}_i) - z) dz \right) \\
&= -\frac{1}{n} \sum_{i \in [n]} \int_0^{y_i f_{\boldsymbol{\theta}}(\mathbf{x}_i)} \ell''(z) y_i \nabla_{\boldsymbol{\theta}} (f_{\boldsymbol{\theta}}(\mathbf{x}_i)) dz \\
&= -\frac{1}{n} \sum_{i \in [n]} \left(\int_0^{y_i f_{\boldsymbol{\theta}}(\mathbf{x}_i)} \ell''(z) dz \right) y_i \nabla_{\boldsymbol{\theta}} (f_{\boldsymbol{\theta}}(\mathbf{x}_i)).
\end{aligned}$$

Since $\ell''(z) \leq 1$, there exists $\delta_i \in [-|f_{\boldsymbol{\theta}}(\mathbf{x}_i)|, |f_{\boldsymbol{\theta}}(\mathbf{x}_i)|]$ for all $i \in [n]$ such that

$$\nabla_{\boldsymbol{\theta}} (\mathcal{L}(\boldsymbol{\theta}) - \tilde{\mathcal{L}}(\boldsymbol{\theta})) = \frac{1}{n} \sum_{i \in [n]} \delta_i \nabla_{\boldsymbol{\theta}} (f_{\boldsymbol{\theta}}(\mathbf{x}_i)). \quad (13)$$

Writing the formula with respect to $\mathbf{w}_k, a_k$, we have

$$\begin{aligned} \left\| \frac{\partial(\mathcal{L}(\boldsymbol{\theta}) - \tilde{\mathcal{L}}(\boldsymbol{\theta}))}{\partial \mathbf{w}_k} \right\|_2 &\leq \frac{1}{n} \sum_{i \in [n]} |\delta_i| \cdot \left\| \frac{\partial f_{\boldsymbol{\theta}}(\mathbf{x}_i)}{\partial \mathbf{w}_k} \right\|_2 \leq \frac{1}{n} \sum_{i \in [n]} |f_{\boldsymbol{\theta}}(\mathbf{x}_i)| \cdot \|a_k \phi'(\langle \mathbf{w}_k, \mathbf{x}_i \rangle) \mathbf{x}_i\|_2 \cdot \\ \left| \frac{\partial(\mathcal{L}(\boldsymbol{\theta}) - \tilde{\mathcal{L}}(\boldsymbol{\theta}))}{\partial a_k} \right| &\leq \frac{1}{n} \sum_{i \in [n]} |\delta_i| \cdot \left\| \frac{\partial f_{\boldsymbol{\theta}}(\mathbf{x}_i)}{\partial a_k} \right\|_2 \leq \frac{1}{n} \sum_{i \in [n]} |f_{\boldsymbol{\theta}}(\mathbf{x}_i)| \cdot |\phi(\langle \mathbf{w}_k, \mathbf{x}_i \rangle)|. \end{aligned}$$

By Lemma B.10, $|f_{\boldsymbol{\theta}}(\mathbf{x}_i)| \leq m \|\boldsymbol{\theta}\|_{\text{M}}^2$. Since Leaky ReLU is 1-Lipschitz and $\|\mathbf{x}_i\|_2 \leq 1$, we have $\|a_k \phi'(\langle \mathbf{w}_k, \mathbf{x}_i \rangle) \mathbf{x}_i\|_2 \leq |a_k|$, $|\phi(\langle \mathbf{w}_k, \mathbf{x}_i \rangle)| \leq \|\mathbf{w}_k\|_2$. Then we have

$$\begin{aligned} \left\| \frac{\partial(\mathcal{L}(\boldsymbol{\theta}) - \tilde{\mathcal{L}}(\boldsymbol{\theta}))}{\partial \mathbf{w}_k} \right\|_2 &\leq \frac{1}{n} \sum_{i \in [n]} m \|\boldsymbol{\theta}\|_{\text{M}}^2 \cdot |a_k| = m \|\boldsymbol{\theta}\|_{\text{M}}^2 \cdot |a_k|, \\ \left| \frac{\partial(\mathcal{L}(\boldsymbol{\theta}) - \tilde{\mathcal{L}}(\boldsymbol{\theta}))}{\partial a_k} \right| &\leq \frac{1}{n} \sum_{i \in [n]} m \|\boldsymbol{\theta}\|_{\text{M}}^2 \cdot \|\mathbf{w}_k\|_2 = m \|\boldsymbol{\theta}\|_{\text{M}}^2 \cdot \|\mathbf{w}_k\|_2, \end{aligned}$$

which completes the proof for $\boldsymbol{\theta} \in \Omega_{\mathcal{S}}$ and thus the same bounds hold for the general case. $\square$

Lemma B.11 is a lemma for bounding the partial subderivatives. For the full subgradient, we have the following lemma.

Lemma B.12. Assume that $\|\mathbf{x}_i\|_2 \leq 1$ for all $i \in [n]$. For any $\boldsymbol{\theta} \in \mathbb{R}^D$, we have

$$\forall \mathbf{g} \in \partial^\circ (\mathcal{L}(\boldsymbol{\theta}) - \tilde{\mathcal{L}}(\boldsymbol{\theta})) : \quad \|\mathbf{g}\|_{\text{M}} \leq m \|\boldsymbol{\theta}\|_{\text{M}}^3.$$

Proof. Note that $|a_k| \leq \|\boldsymbol{\theta}\|_{\text{M}}$ and $\|\mathbf{w}_k\|_2 \leq \|\boldsymbol{\theta}\|_{\text{M}}$. Combining this with Lemma B.11 gives $\|\partial^\circ \mathcal{L}(\boldsymbol{\theta})\|_{\text{M}} \subseteq (-\infty, m \|\boldsymbol{\theta}\|_{\text{M}}^3]$. $\square$

When $\tilde{\mathcal{L}}(\boldsymbol{\theta})$ is smooth, we have the following direct corollary.

Corollary B.13. Assume that $\|\mathbf{x}_i\|_2 \leq 1$ for all $i \in [n]$. If $\tilde{\mathcal{L}}$ is continuously differentiable at $\boldsymbol{\theta} \in \mathbb{R}^D$, then we have

$$\forall \mathbf{g} \in (\partial^\circ \mathcal{L}(\boldsymbol{\theta}) - \nabla \tilde{\mathcal{L}}(\boldsymbol{\theta})) : \quad \|\mathbf{g}\|_{\text{M}} \leq m \|\boldsymbol{\theta}\|_{\text{M}}^3.$$

Note that $\partial^\circ (\mathcal{L}(\boldsymbol{\theta}) - \tilde{\mathcal{L}}(\boldsymbol{\theta})) \neq \partial^\circ \mathcal{L}(\boldsymbol{\theta}) - \nabla \tilde{\mathcal{L}}(\boldsymbol{\theta})$ because the exact sum rule does not hold for Clarke's subdifferential when $\tilde{\mathcal{L}}(\boldsymbol{\theta})$ is not smooth. In the non-smooth case, we have the following lemma:

Lemma B.14. Assume that $\|\mathbf{x}_i\|_2 \leq 1$ for all $i \in [n]$. For any $\epsilon > 0$ and $\|\boldsymbol{\theta}\|_{\text{M}} \leq \sqrt{\frac{\epsilon}{2m}}$, we have

$$\forall k \in [m], \quad \frac{\partial^\circ \mathcal{L}(\boldsymbol{\theta})}{\partial \mathbf{w}_k} \subseteq \left\{ -\frac{a_k}{2n} \sum_{i=1}^n (1 + \epsilon_i) \alpha_i y_i \mathbf{x}_i : \alpha_i \in \phi^\circ(\mathbf{w}_k^\top \mathbf{x}_i), \epsilon_i \in [-\epsilon, \epsilon], \forall i \in [n] \right\}.$$

Proof. If $\boldsymbol{\theta} \in \Omega_{\mathcal{S}}$, by (13), there exists $\delta_i \in [-|f_{\boldsymbol{\theta}}(\mathbf{x}_i)|, |f_{\boldsymbol{\theta}}(\mathbf{x}_i)|]$ for all $i \in [n]$ such that

$$\nabla \mathcal{L}(\boldsymbol{\theta}) = \nabla \tilde{\mathcal{L}}(\boldsymbol{\theta}) + \frac{1}{n} \sum_{i \in [n]} \delta_i \nabla_{\boldsymbol{\theta}}(f_{\boldsymbol{\theta}}(\mathbf{x}_i)).$$

Writing it with respect to $\mathbf{w}_k$, we have

$$\begin{aligned} \frac{\partial \mathcal{L}}{\partial \mathbf{w}_k} &= -a_k \frac{\partial G(\mathbf{w}_k)}{\partial \mathbf{w}_k} + \frac{1}{n} \sum_{i \in [n]} \delta_i \frac{\partial f_{\boldsymbol{\theta}}(\mathbf{x}_i)}{\partial \mathbf{w}_k} \\ &= -\frac{a_k}{2n} \sum_{i \in [n]} y_i \phi'(\mathbf{w}_k^\top \mathbf{x}_i) \mathbf{x}_i + \frac{1}{n} \sum_{i \in [n]} \delta_i a_k \phi'(\mathbf{w}_k^\top \mathbf{x}_i) \mathbf{x}_i \\ &= -\frac{a_k}{2n} \sum_{i \in [n]} y_i (1 - 2y_i \delta_i) \phi'(\mathbf{w}_k^\top \mathbf{x}_i) \mathbf{x}_i. \end{aligned}$$

Regarding Clarke's subdifferential at a point $\theta \in \mathbb{R}^D$, we can take limits in a neighborhood of θ in Ω_S through (8), then

$$\frac{\partial^\circ \mathcal{L}}{\partial \mathbf{w}_k} \subseteq \left\{ -\frac{a_k}{2n} \sum_{i \in [n]} y_i (1 + \epsilon_i) \alpha_i \mathbf{x}_i : \alpha_i \in \phi^\circ(\mathbf{w}_k^\top \mathbf{x}_i), \epsilon_i \in [-2|f_\theta(\mathbf{x}_i)|, 2|f_\theta(\mathbf{x}_i)|], \forall i \in [n] \right\}.$$

We conclude the proof by noticing that $[-2|f_\theta(\mathbf{x}_i)|, 2|f_\theta(\mathbf{x}_i)|] \subseteq [-\epsilon, \epsilon]$ by Lemma B.10. $\square$

B.6 Basic Properties of Gradient Flow

The following lemma is a simple corollary from Davis et al. (2020).

Lemma B.15. *For gradient flow $\theta(t)$ on a two-layer Leaky ReLU network with logistic loss, we have*

$$\frac{d\theta(t)}{dt} = -\bar{\partial}^\circ \mathcal{L}(\theta(t)), \quad \frac{d\mathcal{L}(\theta(t))}{dt} = -\left\| \frac{d\theta(t)}{dt} \right\|_2^2$$

for a.e. $t \geq 0$.

The following lemma is from Du et al. (2018). We provide a simple proof here for completeness.

Lemma B.16. *For gradient flow $\theta(t) = (\mathbf{w}_1(t), \dots, \mathbf{w}_m(t), a_1(t), \dots, a_m(t))$ on a two-layer Leaky ReLU network with logistic loss, the following holds for all $t \geq 0$,*

$$\frac{1}{2} \frac{d\|\mathbf{w}_k\|_2^2}{dt} = \frac{1}{2} \frac{d|a_k|^2}{dt} = -\frac{1}{n} \sum_{i=1}^n \ell'(q_i(\theta)) y_i a_k \phi(\mathbf{w}_k^\top \mathbf{x}_i),$$

where $q_i(\theta) := y_i f_\theta(\mathbf{x}_i)$. Therefore, $\frac{d}{dt}(\|\mathbf{w}_k\|_2^2 - |a_k|^2) = 0$ for all $t \geq 0$.

Proof. By (9), we have the following for any $\theta \in \Omega_S$,

$$a_k \cdot \frac{\partial f_\theta(\mathbf{x})}{\partial a_k} = a_k \phi(\mathbf{w}_k^\top \mathbf{x}_i), \quad \left\langle \mathbf{w}_k, \frac{\partial f_\theta(\mathbf{x})}{\partial \mathbf{w}_k} \right\rangle = a_k \phi'(\mathbf{w}_k^\top \mathbf{x}_i) \mathbf{w}_k^\top \mathbf{x}_i.$$

By 1-homogeneity of ϕ and Theorem B.3, we have $\phi'(\mathbf{w}_k^\top \mathbf{x}_i) \mathbf{w}_k^\top \mathbf{x}_i = \phi(\mathbf{w}_k^\top \mathbf{x}_i)$, which implies that $\left\langle \mathbf{w}_k, \frac{\partial f_\theta(\mathbf{x})}{\partial \mathbf{w}_k} \right\rangle = a_k \phi(\mathbf{w}_k^\top \mathbf{x}_i)$.

For any $\theta \in \mathbb{R}^D$, we can take limits in Ω_S through (8) to show that the same equation holds in general.

$$a_k \cdot \frac{\partial^\circ f_\theta(\mathbf{x})}{\partial a_k} = \left\langle \mathbf{w}_k, \frac{\partial^\circ f_\theta(\mathbf{x})}{\partial \mathbf{w}_k} \right\rangle = \{a_k \phi(\mathbf{w}_k^\top \mathbf{x}_i)\}.$$

By chain rule, for a.e. $t \geq 0$ we have

$$\begin{aligned} \frac{1}{2} \frac{d|a_k|^2}{dt} &= \frac{da_k}{dt} \cdot a_k \in -\frac{1}{n} \sum_{i=1}^n \ell'(q_i(\theta)) y_i \frac{\partial^\circ f_\theta(\mathbf{x}_i)}{\partial a_k} \cdot a_k. \\ \frac{1}{2} \frac{d\|\mathbf{w}_k\|_2^2}{dt} &= \left\langle \frac{d\mathbf{w}_k}{dt}, \mathbf{w}_k \right\rangle \in -\frac{1}{n} \sum_{i=1}^n \ell'(q_i(\theta)) y_i \left\langle \frac{\partial^\circ f_\theta(\mathbf{x}_i)}{\partial \mathbf{w}_k}, \mathbf{w}_k \right\rangle. \end{aligned}$$

Therefore we have

$$\frac{1}{2} \frac{d|a_k|^2}{dt} = \frac{1}{2} \frac{d\|\mathbf{w}_k\|_2^2}{dt} = -\frac{1}{n} \sum_{i=1}^n \ell'(q_i(\theta)) y_i a_k \phi(\mathbf{w}_k^\top \mathbf{x}_i),$$

for a.e. $t \geq 0$. Note that $-\frac{1}{n} \sum_{i=1}^n \ell'(q_i(\theta)) y_i a_k \phi(\mathbf{w}_k^\top \mathbf{x}_i)$ is continuous in θ and thus continuous in time t . This means we can further deduce that this equation holds for all $t \geq 0$. This automatically proves that $\frac{d}{dt}(\|\mathbf{w}_k\|_2^2 - |a_k|^2) = 0$. $\square$

The following lemma shows that if a neuron has zero weights, then it stays with zero weights forever. Conversely, this also implies that the weights stay non-zero if they are initially non-zero.

Lemma B.17. *If $a_k(t_0) = 0$ and $\mathbf{w}_k(t_0) = \mathbf{0}$ at some time $t_0 \geq 0$, then $a_k(t) = 0$ and $\mathbf{w}_k(t) = \mathbf{0}$ for all $t \geq 0$.*

Proof. By Lemma B.16, we know that $\|\mathbf{w}_k\|_2 = |a_k|$ hold for all $t \geq 0$. Also, we have $\frac{1}{2} \left| \frac{d\|\mathbf{w}_k\|_2^2}{dt} \right| = \frac{1}{2} \left| \frac{d|a_k|^2}{dt} \right| \leq C \cdot |a_k| \|\mathbf{w}_k\|_2 = C \|\mathbf{w}_k\|_2^2$, where $C > 0$ is some constant. Then

$$\|\mathbf{w}_k(t)\|_2^2 \leq \|\mathbf{w}_k(t_0)\|_2^2 + \int_{t_0}^t 2C \|\mathbf{w}_k(\tau)\|_2^2 d\tau.$$

By Grönwall's inequality (12) this implies that $\|\mathbf{w}_k(t)\|_2 = 0$ for all $t \geq t_0$. Similarly,

$$\|\mathbf{w}_k(t)\|_2^2 \leq \|\mathbf{w}_k(t_0)\|_2^2 + \int_t^{t_0} 2C \|\mathbf{w}_k(\tau)\|_2^2 d\tau.$$

By Grönwall's inequality (12) again, $\|\mathbf{w}_k(t)\|_2 = 0$ for all $t \leq t_0$, which completes the proof. $\square$

A direct corollary of Lemma B.16 and Lemma B.17 is the following characterization in the case where the weights are initially balanced.

Corollary B.18. *If $|a_k| = \|\mathbf{w}_k\|_2$ initially for $t = 0$, then this equation holds for all $t \geq 0$. Moreover,*

1. *If $a_k(0) = \|\mathbf{w}_k(0)\|_2$, then $a_k(t) = \|\mathbf{w}_k(t)\|_2$ for all $t \geq 0$;*
2. *If $a_k(0) = -\|\mathbf{w}_k(0)\|_2$, then $a_k(t) = -\|\mathbf{w}_k(t)\|_2$ for all $t \geq 0$.*

B.7 A Useful Theorem for Loss Convergence

In this section we prove a useful theorem for loss convergence, which will be used later in our analysis for both symmetric and non-symmetric datasets.

Theorem B.19. *Under Assumption 3.2, for any linear separator $\mathbf{w}^*$ of the data with positive linear margin (e.g. $y_i \langle \mathbf{w}^*, \mathbf{x}_i \rangle \geq \gamma^* > 0$ for all $i \in [n]$), if initially there exists $k \in [m]$ such that*

$$\text{sgn}(a_k(0)) \langle \mathbf{w}_k(0), \mathbf{w}^* \rangle > 0, \quad \langle \mathbf{w}_k(0), \mathbf{w}^* \rangle^2 > \|\mathbf{w}_k(0)\|_2^2 - |a_k(0)|^2,$$

then $a_k(t) \neq 0$ for all $t > 0$, and $\mathcal{L}(\boldsymbol{\theta}(t)) \rightarrow 0$ and $\|\boldsymbol{\theta}(t)\|_2 \rightarrow +\infty$ as $t \rightarrow +\infty$.

Before proving Theorem B.19, we first prove a lemma on gradient lower bounds.

Lemma B.20. *For a.e. $t \geq 0$,*

$$\left\langle \text{sgn}(a_k) \frac{d\mathbf{w}_k}{dt}, \mathbf{w}^* \right\rangle \geq |a_k| \alpha_{\text{leaky}} \gamma^* \cdot \frac{1 - \exp(-n\mathcal{L})}{n}.$$

Proof. By (10), there exist $h_1^{(k)}(t), \dots, h_n^{(k)}(t) \in [\alpha_{\text{leaky}}, 1]$ such that

$$\frac{d\mathbf{w}_k}{dt} = \frac{a_k}{n} \sum_{i \in [n]} g_i(\boldsymbol{\theta}(t)) h_i^{(k)}(t) y_i \mathbf{x}_i$$

where $g_i(\boldsymbol{\theta}(t)) = -\ell'(y_i f_{\boldsymbol{\theta}(t)}(\mathbf{x}_i)) > 0$. Then we have

$$\left\langle \text{sgn}(a_k) \frac{d\mathbf{w}_k}{dt}, \mathbf{w}^* \right\rangle \geq \frac{|a_k|}{n} \sum_{i \in [n]} g_i(\boldsymbol{\theta}(t)) \alpha_{\text{leaky}} \gamma^*.$$

Note that $-\ell'(q) = \frac{1}{1+e^q} = 1 - \frac{1}{1+e^{-q}} = 1 - \exp(-\ell(q))$ for all q . So we have the following lower bound for $\sum_{i \in [n]} g_i(\boldsymbol{\theta}(t))$:

$$\begin{aligned} \sum_{i \in [n]} g_i(\boldsymbol{\theta}(t)) &= \sum_{i \in [n]} -\ell'(y_i f_{\boldsymbol{\theta}(t)}(\mathbf{x}_i)) = \sum_{i \in [n]} (1 - \exp(-\ell(y_i f_{\boldsymbol{\theta}(t)}(\mathbf{x}_i)))) \\ &\geq \max_{i \in [n]} (1 - \exp(-\ell(y_i f_{\boldsymbol{\theta}(t)}(\mathbf{x}_i)))) \\ &\geq 1 - \exp\left(-\max_{i \in [n]} \ell(y_i f_{\boldsymbol{\theta}(t)}(\mathbf{x}_i))\right) \\ &\geq 1 - \exp(-n\mathcal{L}). \end{aligned}$$

Therefore,

$$\left\langle \operatorname{sgn}(a_k) \frac{d\mathbf{w}_k}{dt}, \mathbf{w}^* \right\rangle \geq |a_k| \alpha_{\text{leaky}} \gamma^* \cdot \frac{1 - \exp(-n\mathcal{L})}{n},$$

which completes the proof. $\square$

Proof for Theorem B.19. We only need to show that there exists t_0 such that $\mathcal{L}(\boldsymbol{\theta}(t_0)) < \frac{\ln 2}{n}$, then we can apply Theorem 3.1 to show that $\mathcal{L}(\boldsymbol{\theta}(t)) \rightarrow 0$. Assume to the contrary that $\mathcal{L}(\boldsymbol{\theta}(t)) \geq \frac{\ln 2}{n}$ for all $t \geq 0$. By Lemma B.20,

$$\operatorname{sgn}(a_k) \left\langle \frac{d\mathbf{w}_k}{dt}, \mathbf{w}^* \right\rangle \geq |a_k| \cdot \frac{\alpha_{\text{leaky}} \gamma^*}{2n}.$$

Let $c := \langle \mathbf{w}_k(0), \mathbf{w}^* \rangle^2 - \|\mathbf{w}_k(0)\|_2^2 + |a_k(0)|^2 > 0$. First we show that $\operatorname{sgn}(a_k(t)) = \operatorname{sgn}(a_k(0))$ for all $t > 0$. Otherwise let $t_s := \inf\{t : \operatorname{sgn}(a_k(t)) \neq \operatorname{sgn}(a_k(0))\}$, and since $a_k(t)$ is continuous, $a_k(t_s) = 0$. We know for $t \in [0, t_s]$, $\operatorname{sgn}(a_k(0)) \frac{d}{dt} \langle \mathbf{w}_k(t), \mathbf{w}^* \rangle > 0$, and

$$\begin{aligned} |a_k(t_s)|^2 &= |a_k(0)|^2 - \|\mathbf{w}_k(0)\|_2^2 + \|\mathbf{w}_k(t_s)\|_2^2 \geq |a_k(0)|^2 - \|\mathbf{w}_k(0)\|_2^2 + \langle \mathbf{w}_k(t_s), \mathbf{w}^* \rangle^2 \\ &> |a_k(0)|^2 - \|\mathbf{w}_k(0)\|_2^2 + \langle \mathbf{w}_k(0), \mathbf{w}^* \rangle^2 = c > 0. \end{aligned}$$

This contradicts to the fact that $a_k(t_s) = 0$, and thus $\operatorname{sgn}(a_k(t))$ does not change during all time. Therefore for any $t > 0$, $a_k(t) \neq 0$. Then for all $t > 0$,

$$\begin{aligned} |a_k(t)|^2 &= |a_k(0)|^2 - \|\mathbf{w}_k(0)\|_2^2 + \|\mathbf{w}_k(t)\|_2^2 \geq |a_k(0)|^2 - \|\mathbf{w}_k(0)\|_2^2 + \langle \mathbf{w}_k(t), \mathbf{w}^* \rangle^2 \\ &> |a_k(0)|^2 - \|\mathbf{w}_k(0)\|_2^2 + \langle \mathbf{w}_k(0), \mathbf{w}^* \rangle^2 = c. \end{aligned}$$

Lemma B.15 ensures that $-\frac{d\mathcal{L}}{dt} = \left\| \frac{d\boldsymbol{\theta}}{dt} \right\|_2^2$ for a.e. $t \geq 0$. Then we have

$$-\frac{d\mathcal{L}}{dt} \geq \left\| \frac{d\mathbf{w}_k}{dt} \right\|_2^2 \geq \left\langle \frac{d\mathbf{w}_k}{dt}, \mathbf{w}^* \right\rangle^2 \geq |a_k|^2 \left(\frac{\alpha_{\text{leaky}} \gamma^*}{2n} \right)^2 \geq c^2 \cdot \left(\frac{\alpha_{\text{leaky}} \gamma^*}{2n} \right)^2.$$

Then we can conclude that

$$\mathcal{L}(\boldsymbol{\theta}(0)) - \mathcal{L}(\boldsymbol{\theta}(t)) \geq c^2 \left(\frac{\alpha_{\text{leaky}} \gamma^*}{2n} \right)^2 t.$$

Integrating on t from 0 to $+\infty$, we can see that the LHS is upper bounded by $\mathcal{L}(\boldsymbol{\theta}(0)) - \frac{\ln 2}{n}$ while the RHS is unbounded, which leads to a contradiction. Therefore, there exist time t_0 such that $\mathcal{L}(\boldsymbol{\theta}(t_0)) < \frac{\ln 2}{n}$, and thus $\mathcal{L}(\boldsymbol{\theta}(t)) \rightarrow 0$ as $t \rightarrow +\infty$. $\square$

C Proofs for Linear Maximality for the Symmetric Case

For linearly separable and symmetric data, we show that all global-max-margin directions represent linear functions in Theorem 4.2. We give a proof here.

Proof for Theorem 4.2. Let $\boldsymbol{\theta}^* = (\mathbf{w}_1, \dots, \mathbf{w}_m, a_1, \dots, a_m) \in \mathbb{S}^{D-1}$ be any global-max-margin direction with output margin $q_{\min}(\boldsymbol{\theta}^*) = \gamma(\boldsymbol{\theta}^*)$. As the dataset is symmetric,

$$\gamma(\boldsymbol{\theta}^*) = \min_{i \in [n]} \{y_i f_{\boldsymbol{\theta}^*}(\mathbf{x}_i), -y_i f_{\boldsymbol{\theta}^*}(-\mathbf{x}_i)\}.$$

Now we define $A := \sqrt{\sum_{k \in [m]} a_k^2}$ and let $\boldsymbol{\theta}' = (\mathbf{w}'_1, \dots, \mathbf{w}'_m, a'_1, \dots, a'_m)$ where

$$\mathbf{w}'_1 = \frac{1}{\sqrt{2}A} \sum_{k \in [m]} a_k \mathbf{w}_k, \quad \mathbf{w}'_2 = -\mathbf{w}'_1, \quad a'_1 = \frac{A}{\sqrt{2}}, \quad a'_2 = -a'_1,$$

and $a'_k = 0, \mathbf{w}'_k = \mathbf{0}$ for $k > 2$. We claim that $\gamma(\boldsymbol{\theta}') \geq \gamma(\boldsymbol{\theta}^*)$. First we prove that $q_i(\boldsymbol{\theta}') \geq \gamma(\boldsymbol{\theta}^*)$ by repeatedly applying $\phi(z) - \phi(-z) = (1 + \alpha_{\text{leaky}})z$.

$$\begin{aligned}
q_i(\boldsymbol{\theta}') &= y_i f_{\boldsymbol{\theta}'}(\mathbf{x}_i) = y_i (a'_1 \phi(\langle \mathbf{w}'_1, \mathbf{x}_i \rangle) + a'_2 \phi(\langle \mathbf{w}'_2, \mathbf{x}_i \rangle)) \\
&= y_i (1 + \alpha_{\text{leaky}}) a'_1 \langle \mathbf{w}'_1, \mathbf{x}_i \rangle \\
&= \frac{y_i}{2} \sum_{k \in [m]} \langle (1 + \alpha_{\text{leaky}}) a_k \mathbf{w}_k, \mathbf{x}_i \rangle \\
&= \frac{y_i}{2} \sum_{k \in [m]} (a_k \phi(\mathbf{w}_k^\top \mathbf{x}_i) - a_k \phi(-\mathbf{w}_k^\top \mathbf{x}_i)) \\
&= \frac{1}{2} (y_i f_{\boldsymbol{\theta}^*}(\mathbf{x}_i) - y_i f_{\boldsymbol{\theta}^*}(-\mathbf{x}_i)) \geq \gamma(\boldsymbol{\theta}^*).
\end{aligned}$$

Meanwhile, by the Cauchy-Schwarz inequality,

$$\|\boldsymbol{\theta}'\|_2^2 = A^2 + \left\| \sum_{k \in [m]} \frac{a_k}{A} \mathbf{w}_k \right\|_2^2 \leq A^2 + \sum_{k \in [m]} \left(\frac{a_k}{A} \right)^2 \cdot \sum_{k \in [m]} \|\mathbf{w}_k\|_2^2 = A^2 + \sum_{k \in [m]} \|\mathbf{w}_k\|_2^2 = \|\boldsymbol{\theta}^*\|_2^2.$$

Thus $\gamma(\boldsymbol{\theta}') = \frac{q_{\min}(\boldsymbol{\theta}')}{\|\boldsymbol{\theta}'\|_2} \geq \gamma(\boldsymbol{\theta}^*)$. As $\boldsymbol{\theta}^*$ is already a global-max-margin direction, equalities should hold in all the inequalities above, so

$$\min_{i \in [n]} \{y_i f_{\boldsymbol{\theta}^*}(\mathbf{x}_i) - y_i f_{\boldsymbol{\theta}^*}(-\mathbf{x}_i)\} = 2\gamma(\boldsymbol{\theta}^*), \quad \left\| \sum_{k \in [m]} \frac{a_k}{A} \mathbf{w}_k \right\|_2^2 = \sum_{k \in [m]} \left(\frac{a_k}{A} \right)^2 \cdot \sum_{k \in [m]} \|\mathbf{w}_k\|_2^2.$$

Then we know the following:

- There is $\mathbf{c} \in \mathbb{R}^d$ that $\mathbf{w}_k = a_k \mathbf{c}$ for all k ;
- There is $j \in [n]$ that $y_j f_{\boldsymbol{\theta}^*}(\mathbf{x}_j) = -y_j f_{\boldsymbol{\theta}^*}(-\mathbf{x}_j) = \gamma(\boldsymbol{\theta}^*)$.

Note that $\phi(z) + \phi(-z) = (1 - \alpha_{\text{leaky}})|z|$. Then we have

$$0 = f_{\boldsymbol{\theta}^*}(\mathbf{x}_j) + f_{\boldsymbol{\theta}^*}(-\mathbf{x}_j) = \sum_{k \in [m]} a_k (\phi(a_k \mathbf{c}^\top \mathbf{x}_j) + \phi(-a_k \mathbf{c}^\top \mathbf{x}_j)) = \sum_{k=1}^m (1 - \alpha_{\text{leaky}}) a_k |a_k \mathbf{c}^\top \mathbf{x}_j|.$$

Certainly $\mathbf{c}^\top \mathbf{x}_j \neq 0$ as otherwise the margin would be zero. Then $\sum_{k \in [m]} a_k |a_k| = 0$, which means $\sum_{k: a_k \geq 0} a_k^2 = \sum_{k: a_k < 0} a_k^2 = \frac{1}{2} A^2$, and therefore

$$\begin{aligned}
f_{\boldsymbol{\theta}^*}(\mathbf{x}) &= \sum_{k=1}^m a_k \phi(a_k \mathbf{c}^\top \mathbf{x}) = \sum_{k=1}^m a_k |a_k| \phi(\text{sgn}(a_k) \mathbf{c}^\top \mathbf{x}) \\
&= \frac{1}{2} A^2 (\phi(\mathbf{c}^\top \mathbf{x}) - \phi(-\mathbf{c}^\top \mathbf{x})) = \frac{1}{2} A^2 (1 + \alpha_{\text{leaky}}) \mathbf{c}^\top \mathbf{x}
\end{aligned}$$

is a linear function in $\mathbf{x}$.

Finally, let $\gamma_{\mathbf{w}^*} = \min_{i \in [n]} y_i \langle \mathbf{w}^*, \mathbf{x}_i \rangle$ be the maximum linear margin, where $\mathbf{w}^* \in \mathbb{S}^{d-1}$ is the max-margin linear separator. As $\|\boldsymbol{\theta}^*\|_2^2 = 1 = (1 + \|\mathbf{c}\|_2^2) A^2$,

$$\begin{aligned}
\gamma(\boldsymbol{\theta}^*) &= \frac{1}{2} A^2 (1 + \alpha_{\text{leaky}}) \min_{i \in [n]} y_i \mathbf{c}^\top \mathbf{x}_i \leq \frac{1}{2} A^2 (1 + \alpha_{\text{leaky}}) \|\mathbf{c}\|_2 \gamma_{\mathbf{w}^*} \\
&= \frac{\|\mathbf{c}\|_2}{2(1 + \|\mathbf{c}\|_2^2)} (1 + \alpha_{\text{leaky}}) \gamma_{\mathbf{w}^*} \leq \frac{1}{4} (1 + \alpha_{\text{leaky}}) \gamma_{\mathbf{w}^*}.
\end{aligned}$$

By choosing $\mathbf{c} = \mathbf{w}^*$ with $A = \frac{1}{\sqrt{2}}$, the network is able to attain the margin $\frac{1}{4} (1 + \alpha_{\text{leaky}}) \gamma_{\mathbf{w}^*}$. As $\boldsymbol{\theta}^*$ is already a global-max-margin direction, we know again that the equalities must hold. Therefore we know

- $\min_{i \in [n]} y_i \mathbf{c}^\top \mathbf{x}_i = \|\mathbf{c}\|_2 \gamma_{\mathbf{w}^*};$
- $\frac{\|\mathbf{c}\|_2}{1 + \|\mathbf{c}\|_2^2} = \frac{1}{2}.$

Then we know $\mathbf{c} = \mathbf{w}^*$ due to the uniqueness of the max-margin linear separator, and thus $A = \frac{1}{\sqrt{2}}$. Therefore the function is $f_{\theta^*}(\mathbf{x}) = \frac{1 + \alpha_{\text{leaky}}}{4} \langle \mathbf{w}^*, \mathbf{x} \rangle$. $\square$

D Proofs for Phase I

In the subsequent sections we first show the proofs for the symmetric datasets under Assumption 4.1. Additional proofs for the non-symmetric counterparts are provided in Appendix G.

As we have illustrated in Section 5.1, we have $G(\mathbf{w}) = \langle \mathbf{w}, \tilde{\boldsymbol{\mu}} \rangle$ under Assumption 4.1. Then we have

$$\tilde{\mathcal{L}}(\boldsymbol{\theta}) := \ell(0) - \sum_{k \in [m]} a_k G(\mathbf{w}_k) = \ell(0) - \sum_{k \in [m]} a_k \langle \mathbf{w}_k, \tilde{\boldsymbol{\mu}} \rangle.$$

This means the dynamics of $\tilde{\boldsymbol{\theta}}(t) = (\tilde{\mathbf{w}}_1(t), \dots, \tilde{\mathbf{w}}_m(t), \tilde{a}_1(t), \dots, \tilde{a}_m(t)) = \tilde{\varphi}(\tilde{\boldsymbol{\theta}}_0, t)$ can be described by linear ODE:

$$\frac{d\tilde{\mathbf{w}}_k}{dt} = a_k \tilde{\boldsymbol{\mu}}, \quad \frac{d\tilde{a}_k}{dt} = \langle \tilde{\mathbf{w}}_k, \tilde{\boldsymbol{\mu}} \rangle.$$

Lemma D.1. *Let $\tilde{\boldsymbol{\theta}}(t) = \tilde{\varphi}(\tilde{\boldsymbol{\theta}}_0, t)$. Then*

$$\|\tilde{\boldsymbol{\theta}}(t)\|_{\text{M}} \leq \exp(t\lambda_0) \|\tilde{\boldsymbol{\theta}}_0\|_{\text{M}}.$$

Proof. By definition and Cauchy-Schwartz inequality,

$$\left\| \frac{d\tilde{\mathbf{w}}_k}{dt} \right\|_2 \leq \|\tilde{a}_k \tilde{\boldsymbol{\mu}}\|_2 \leq \lambda_0 |\tilde{a}_k|, \quad \left| \frac{d\tilde{a}_k}{dt} \right| \leq |\tilde{\mathbf{w}}_k^\top \tilde{\boldsymbol{\mu}}| \leq \lambda_0 \|\tilde{\mathbf{w}}_k\|_2.$$

So we have $\|\tilde{\boldsymbol{\theta}}(t)\|_{\text{M}} \leq \|\tilde{\boldsymbol{\theta}}_0\|_{\text{M}} + \int_0^t \lambda_0 \|\tilde{\boldsymbol{\theta}}(\tau)\|_{\text{M}} d\tau$. Then we can finish the proof by Grönwall's inequality (11). $\square$

Lemma D.2. *For initial point $\boldsymbol{\theta}_0 \neq \mathbf{0}$, we have*

$$\|\boldsymbol{\theta}(t) - \tilde{\varphi}(\boldsymbol{\theta}_0, t)\|_{\text{M}} \leq \frac{4m\|\boldsymbol{\theta}_0\|_{\text{M}}^3}{\lambda_0} \exp(3\lambda_0 t),$$

for all $t \leq \frac{1}{\lambda_0} \ln \frac{\sqrt{\lambda_0/4}}{\sqrt{m}\|\boldsymbol{\theta}_0\|_{\text{M}}}$.

Proof. Let $\tilde{\boldsymbol{\theta}}(t) = \tilde{\varphi}(\boldsymbol{\theta}_0, t)$. By Corollary B.13, the following holds for a.e. $t \geq 0$,

$$\begin{aligned} \left\| \frac{d\boldsymbol{\theta}}{dt} - \frac{d\tilde{\boldsymbol{\theta}}}{dt} \right\|_{\text{M}} &\leq \sup \left\{ \|\boldsymbol{\delta} - \nabla \tilde{\mathcal{L}}(\boldsymbol{\theta})\|_{\text{M}} : \boldsymbol{\delta} \in \partial^\circ \mathcal{L}(\boldsymbol{\theta}) \right\} + \|\nabla \tilde{\mathcal{L}}(\boldsymbol{\theta}) - \nabla \tilde{\mathcal{L}}(\tilde{\boldsymbol{\theta}})\|_{\text{M}} \\ &\leq m\|\boldsymbol{\theta}(t)\|_{\text{M}}^3 + \lambda_0 \|\boldsymbol{\theta} - \tilde{\boldsymbol{\theta}}\|_{\text{M}}. \end{aligned}$$

Taking integral, we have

$$\|\boldsymbol{\theta}(t) - \tilde{\boldsymbol{\theta}}(t)\|_{\text{M}} \leq \int_0^t \left(m\|\boldsymbol{\theta}(\tau)\|_{\text{M}}^3 + \lambda_0 \|\boldsymbol{\theta}(\tau) - \tilde{\boldsymbol{\theta}}(\tau)\|_{\text{M}} \right) d\tau.$$

Let $t_0 := \inf\{t \geq 0 : \|\boldsymbol{\theta}(t)\|_{\text{M}} \geq 2\|\boldsymbol{\theta}_0\|_{\text{M}} \exp(\lambda_0 t)\}$. Then for all $0 \leq t \leq t_0$ (or for all $t \geq 0$ if $t_0 = +\infty$),

$$\begin{aligned} \|\boldsymbol{\theta}(t) - \tilde{\boldsymbol{\theta}}(t)\|_{\text{M}} &\leq \int_0^t \left(8m\|\boldsymbol{\theta}_0\|_{\text{M}}^3 \exp(3\lambda_0 \tau) + \lambda_0 \|\boldsymbol{\theta}(\tau) - \tilde{\boldsymbol{\theta}}(\tau)\|_{\text{M}} \right) d\tau \\ &\leq \frac{8m\|\boldsymbol{\theta}_0\|_{\text{M}}^3}{3\lambda_0} \exp(3\lambda_0 t) + \lambda_0 \int_0^t \|\boldsymbol{\theta}(\tau) - \tilde{\boldsymbol{\theta}}(\tau)\|_{\text{M}} d\tau. \end{aligned}$$

By Grönwall's inequality (12),

$$\begin{aligned}\|\boldsymbol{\theta}(t) - \tilde{\boldsymbol{\theta}}(t)\|_{\text{M}} &\leq \frac{8m\|\boldsymbol{\theta}_0\|_{\text{M}}^3}{3\lambda_0} \left(\exp(3\lambda_0 t) + \lambda_0 \int_0^t \exp(3\lambda_0 \tau) \exp(\lambda_0(t - \tau)) d\tau \right) \\ &= \frac{8m\|\boldsymbol{\theta}_0\|_{\text{M}}^3}{3\lambda_0} \left(\exp(3\lambda_0 t) + \frac{1}{2} \exp(3\lambda_0 t) \right) = \frac{4m\|\boldsymbol{\theta}_0\|_{\text{M}}^3}{\lambda_0} \exp(3\lambda_0 t),\end{aligned}$$

If $t_0 < \frac{1}{2\lambda_0} \ln \frac{\lambda_0}{4m\|\boldsymbol{\theta}_0\|_{\text{M}}^2}$, then

$$\begin{aligned}\|\boldsymbol{\theta}(t)\|_{\text{M}} &\leq \|\tilde{\boldsymbol{\theta}}(t)\|_{\text{M}} + \frac{4m\|\boldsymbol{\theta}_0\|_{\text{M}}^3}{\lambda_0} \exp(3\lambda_0 t) \\ &\leq \|\tilde{\boldsymbol{\theta}}(t)\|_{\text{M}} + \frac{4m\|\boldsymbol{\theta}_0\|_{\text{M}}^2}{\lambda_0} \exp(2\lambda_0 t_0) \cdot \|\boldsymbol{\theta}_0\|_{\text{M}} \exp(\lambda_0 t) \\ &< \|\tilde{\boldsymbol{\theta}}(t)\|_{\text{M}} + \|\boldsymbol{\theta}_0\|_{\text{M}} \exp(\lambda_0 t).\end{aligned}$$

By Lemma D.1, $\|\tilde{\boldsymbol{\theta}}(t)\|_{\text{M}} \leq \|\boldsymbol{\theta}_0\|_{\text{M}} \exp(\lambda_0 t)$. So $\|\boldsymbol{\theta}(t)\|_{\text{M}} < 2\|\boldsymbol{\theta}_0\|_{\text{M}} \exp(\lambda_0 t)$ for all $0 \leq t \leq t_0$, which contradicts to the definition of t_0 . Therefore, $t_0 \geq \frac{1}{2\lambda_0} \ln \frac{\lambda_0}{4m\|\boldsymbol{\theta}_0\|_{\text{M}}^2} = \frac{1}{\lambda_0} \ln \frac{\sqrt{\lambda_0/4}}{\sqrt{m}\|\boldsymbol{\theta}_0\|_{\text{M}}}$. $\square$

Proof for Lemma 5.2. Let $\tilde{\boldsymbol{\theta}}(t) = (\tilde{\boldsymbol{w}}_1(t), \dots, \tilde{\boldsymbol{w}}_m(t), \tilde{a}_1(t), \dots, \tilde{a}_m(t)) = \tilde{\varphi}(\boldsymbol{\theta}_0, t)$. Then

$$[\tilde{\boldsymbol{w}}_k(t), \tilde{a}_k(t)]^\top = \exp(T_1(r)\boldsymbol{M}_{\tilde{\boldsymbol{\mu}}})[\tilde{\boldsymbol{w}}_k(0), \tilde{a}_k(0)]^\top = \exp(T_1(r)\boldsymbol{M}_{\tilde{\boldsymbol{\mu}}})[\sigma_{\text{init}}\bar{\boldsymbol{w}}_k, \sigma_{\text{init}}\bar{a}_k]^\top,$$

where $\boldsymbol{M}_{\tilde{\boldsymbol{\mu}}}$ is defined in Section 5.1,

$$\boldsymbol{M}_{\tilde{\boldsymbol{\mu}}} := \begin{bmatrix} \mathbf{0} & \tilde{\boldsymbol{\mu}} \\ \tilde{\boldsymbol{\mu}}^\top & 0 \end{bmatrix}.$$

Let $\bar{\boldsymbol{\mu}}_2 := \frac{1}{\sqrt{2}}[\bar{\boldsymbol{\mu}}, 1]^\top$ be the top eigenvector of $\boldsymbol{M}_{\tilde{\boldsymbol{\mu}}}$, which is associated with eigenvalue λ_0 . All the other eigenvalues of $\boldsymbol{M}_{\tilde{\boldsymbol{\mu}}}$ are no greater than 0. Note that

$$\exp(T_1(r)\lambda_0)\bar{\boldsymbol{\mu}}_2\bar{\boldsymbol{\mu}}_2^\top \begin{bmatrix} \sigma_{\text{init}}\bar{\boldsymbol{w}}_k \\ \sigma_{\text{init}}\bar{a}_k \end{bmatrix} = \frac{r}{\sqrt{m}\|\boldsymbol{\theta}_0\|_{\text{M}}} \left(\frac{\sigma_{\text{init}}}{\sqrt{2}}(\bar{\boldsymbol{\mu}}^\top \bar{\boldsymbol{w}}_k + \bar{a}_k) \right) \bar{\boldsymbol{\mu}}_2 = \sqrt{2}r\bar{b}_k\bar{\boldsymbol{\mu}}_2 = r\bar{b}_k \begin{bmatrix} \bar{\boldsymbol{\mu}} \\ 1 \end{bmatrix}.$$

So we have

$$\begin{aligned}\left\| \begin{bmatrix} \tilde{\boldsymbol{w}}_k(T_1(r)) \\ \tilde{a}_k(T_1(r)) \end{bmatrix} - r\bar{b}_k \begin{bmatrix} \bar{\boldsymbol{\mu}} \\ 1 \end{bmatrix} \right\|_2 &= \left\| (\exp(T_1(r)\boldsymbol{M}_{\tilde{\boldsymbol{\mu}}}) - \exp(T_1(r)\lambda_0)\bar{\boldsymbol{\mu}}_2\bar{\boldsymbol{\mu}}_2^\top) \begin{bmatrix} \sigma_{\text{init}}\bar{\boldsymbol{w}}_k \\ \sigma_{\text{init}}\bar{a}_k \end{bmatrix} \right\|_2 \\ &\leq \sigma_{\text{init}} \left\| \begin{bmatrix} \bar{\boldsymbol{w}}_k \\ \bar{a}_k \end{bmatrix} \right\|_2 \leq \sqrt{2}\sigma_{\text{init}}\|\bar{\boldsymbol{\theta}}_0\|_{\text{M}}.\end{aligned}$$

With probability 1, $\bar{\boldsymbol{\theta}}_0 \neq \mathbf{0}$. For $r \leq \sqrt{\lambda_0/4}$, we have $T_1(r) = \frac{1}{\lambda_0} \ln \frac{r}{\sqrt{m}\|\boldsymbol{\theta}_0\|_{\text{M}}} \leq \frac{1}{\lambda_0} \ln \frac{\sqrt{\lambda_0/4}}{\sqrt{m}\|\boldsymbol{\theta}_0\|_{\text{M}}}$. Then by Lemma D.2,

$$\|\boldsymbol{\theta}(T_1(r)) - \tilde{\boldsymbol{\theta}}(T_1(r))\|_{\text{M}} \leq \frac{4m\|\boldsymbol{\theta}_0\|_{\text{M}}^3}{\lambda_0} \exp(3\lambda_0 T_1(r)) = \frac{4r^3}{\lambda_0\sqrt{m}}.$$

By triangle inequality, we have

$$\begin{aligned}\left\| \begin{bmatrix} \boldsymbol{w}_k(T_1(r)) \\ a_k(T_1(r)) \end{bmatrix} - r\bar{b}_k \begin{bmatrix} \bar{\boldsymbol{\mu}} \\ 1 \end{bmatrix} \right\|_{\text{M}} &\leq \left\| \begin{bmatrix} \boldsymbol{w}_k(T_1(r)) \\ a_k(T_1(r)) \end{bmatrix} - \begin{bmatrix} \tilde{\boldsymbol{w}}_k(T_1(r)) \\ \tilde{a}_k(T_1(r)) \end{bmatrix} \right\|_{\text{M}} + \left\| \begin{bmatrix} \tilde{\boldsymbol{w}}_k(T_1(r)) \\ \tilde{a}_k(T_1(r)) \end{bmatrix} - r\bar{b}_k \begin{bmatrix} \bar{\boldsymbol{\mu}} \\ 1 \end{bmatrix} \right\|_{\text{M}} \\ &\leq \frac{4r^3}{\lambda_0\sqrt{m}} + \sqrt{2}\sigma_{\text{init}}\|\bar{\boldsymbol{\theta}}_0\|_{\text{M}} \leq \frac{Cr^3}{\sqrt{m}},\end{aligned}$$

for some universal constant C , where the last step is due to our choice of $\sigma_{\text{init}} \leq \frac{r^3}{\sqrt{m}\|\bar{\boldsymbol{\theta}}_0\|_{\text{M}}}$. $\square$

E Proofs for Phase II

E.1 Proof for Exact Embedding

To prove Lemma 5.3, we start from the following lemma.

Lemma E.1. *Given $\hat{\theta}_0 := (\hat{w}_1, \hat{w}_2, \hat{a}_1, \hat{a}_2)$ with $\hat{a}_1 > 0$ and $\hat{a}_2 < 0$, then $\theta(t) = \pi_b(\varphi(\hat{\theta}_0, t))$ is a gradient flow trajectory on $\mathcal{L}(\theta)$ starting from $\theta(0) = \pi_b(\hat{\theta}_0)$.*

First we notice the following fact.

Lemma E.2. *For any $\hat{\theta}$ and $g \in \partial^\circ \mathcal{L}(\hat{\theta})$, $\pi_b(g) \in \partial^\circ \mathcal{L}(\pi_b(\hat{\theta}))$.*

Below we use $\pi_b(S) = \{\pi_b(s) : s \in S\}$ to denote the embedding of a parameter set.

Proof. For every $\hat{\theta} = (\hat{w}_1, \hat{w}_2, \hat{a}_1, \hat{a}_2) \in \Omega_S$ (i.e., no activation function has zero input), let $\theta = \pi_b(\hat{\theta}) = (w_1, \dots, w_m, a_1, \dots, a_m)$, and clearly $\theta \in \Omega_S$. Then $\partial^\circ \mathcal{L}(\hat{\theta}) = \{\nabla \mathcal{L}(\hat{\theta})\}$ and $\partial^\circ \mathcal{L}(\theta) = \{\nabla \mathcal{L}(\theta)\}$ are the usual differentials. In this case, we can apply the chain rule as

$$\begin{aligned}\nabla \mathcal{L}(\hat{\theta}) &= \frac{1}{n} \sum_{i \in [n]} y_i \ell'(y_i f_{\hat{\theta}}(x_i)) \frac{\partial f_{\hat{\theta}}(x_i)}{\partial \hat{\theta}}, \\ \nabla \mathcal{L}(\theta) &= \frac{1}{n} \sum_{i \in [n]} y_i \ell'(y_i f_{\theta}(x_i)) \frac{\partial f_{\theta}(x_i)}{\partial \theta}.\end{aligned}$$

Notice that the embedding preserves the function value,

$$\begin{aligned}f_{\theta}(x_i) &= \sum_{j=1}^m a_j \phi(w_j^\top x_i) = \sum_{j: b_j > 0} \frac{b_j^2}{b_+^2} \hat{a}_1 \phi(\hat{w}_1^\top x_i) + \sum_{j: b_j < 0} \frac{b_j^2}{b_-^2} \hat{a}_2 \phi(\hat{w}_2^\top x_i) \\ &= \hat{a}_1 \phi(\hat{w}_1^\top x_i) + \hat{a}_2 \phi(\hat{w}_2^\top x_i) = f_{\hat{\theta}}(x_i);\end{aligned}$$

and the also preserves the gradient

$$\begin{aligned}\frac{\partial f_{\theta}(x_i)}{\partial w_k} &= a_k \phi'(w_k^\top x_i) x_i = \begin{cases} \frac{b_k}{b_+} \hat{a}_1 \phi'(\hat{w}_1^\top x_i) x_i & \text{if } b_k > 0 \\ \frac{b_k}{b_-} \hat{a}_2 \phi'(\hat{w}_2^\top x_i) x_i & \text{if } b_k < 0 \end{cases}, \\ \frac{\partial f_{\theta}(x_i)}{\partial a_k} &= \phi(w_k^\top x_i) = \begin{cases} \frac{b_k}{b_+} \phi(\hat{w}_1^\top x_i) & \text{if } b_k > 0 \\ \frac{b_k}{b_-} \phi(\hat{w}_2^\top x_i) & \text{if } b_k < 0 \end{cases},\end{aligned}$$

so $\frac{\partial f_{\theta}(x_i)}{\partial \theta} = \pi_b \left(\frac{\partial f_{\hat{\theta}}(x_i)}{\partial \hat{\theta}} \right)$. Then from the chain rule above we can see $\nabla \mathcal{L}(\theta) = \pi_b(\nabla \mathcal{L}(\hat{\theta}))$, and we proved the lemma in this case.

In the general case, by the definition of Clarke's subdifferential,

$$\partial^\circ \mathcal{L}(\theta) := \text{conv} \left\{ \lim_{n \rightarrow \infty} \nabla \mathcal{L}(\theta_n) : \mathcal{L} \text{ differentiable at } \theta_n, \lim_{n \rightarrow \infty} \theta_n = \theta \right\}.$$

For any $\hat{\theta}_n \rightarrow \hat{\theta}$ with $\hat{\theta}_n \in \Omega_S$, $\pi_b(\hat{\theta}_n) \rightarrow \pi_b(\hat{\theta})$, and

$$\lim_{n \rightarrow \infty} \nabla \mathcal{L}(\pi_b(\hat{\theta}_n)) = \lim_{n \rightarrow \infty} \pi_b(\nabla \mathcal{L}(\hat{\theta}_n)) = \pi_b \left(\lim_{n \rightarrow \infty} \nabla \mathcal{L}(\hat{\theta}_n) \right).$$

Taking the convex hull, it follows that $\pi_b(\partial^\circ \mathcal{L}(\hat{\theta})) \subseteq \partial^\circ \mathcal{L}(\pi_b(\hat{\theta}))$, and we finished the proof. $\square$

Proof for Lemma E.1. For notations we write $\hat{\theta}(t) := \varphi(\hat{\theta}_0, t)$ and $\theta(t) = \pi_b(\hat{\theta}(t))$. Then $\frac{d}{dt} \hat{\theta}(t) \in -\partial^\circ \mathcal{L}(\hat{\theta}(t))$ for a.e. t . At these t , $\frac{d}{dt} \theta(t) = \pi_b \left(\frac{d}{dt} \hat{\theta}(t) \right) \in \pi_b(-\partial^\circ \mathcal{L}(\hat{\theta}(t)))$. From Lemma E.2 we know $\pi_b(\partial^\circ \mathcal{L}(\hat{\theta}(t))) \subseteq \partial^\circ \mathcal{L}(\theta(t))$. Then $\frac{d}{dt} \theta(t) \in -\partial^\circ \mathcal{L}(\theta(t))$ for a.e. t , and therefore $\theta(t)$ is indeed a gradient flow trajectory. $\square$

Proof for Lemma 5.3. By Lemma E.1, $\pi_b(\varphi(\hat{\theta}_0, t))$ is indeed a gradient flow trajectory. Then, as $\pi_b(\varphi(\hat{\theta}_0, 0)) = \pi_b(\hat{\theta}_0)$, as well as the fact that $\hat{\theta}_0$ and $\pi_b(\hat{\theta}_0)$ are non-branching starting points, the gradient flow trajectory is unique and therefore $\pi_b(\varphi(\hat{\theta}_0, t)) = \varphi(\pi_b(\hat{\theta}_0), t)$ for all $t \geq 0$. $\square$

E.2 A General Theorem for Limiting Trajectory Near Zero

Before analyzing Phase II, we first give a characterization for gradient flow on Leaky ReLU networks with logistic loss, starting near $r\hat{\theta}$, where $\hat{\theta}$ is a well-aligned parameter vector defined below. We only assume that the inputs are bounded $\|x_i\|_2 \leq 1$ and $\lambda := \max\{|G(\mathbf{w})| : \mathbf{w} \in \mathbb{S}^{d-1}\} > 0$. Theorems in the section will be used again in the non-symmetric case.

Definition E.3 (Well-aligned Parameter Vector). We say that $\hat{\theta} := (\hat{\mathbf{w}}_1, \dots, \hat{\mathbf{w}}_m, \hat{a}_1, \dots, \hat{a}_m)$ is a *well-aligned parameter vector* if it satisfies the following for some $1 \leq p \leq m$:

1. For $1 \leq k \leq p$, $\frac{\hat{\mathbf{w}}_k}{\|\hat{\mathbf{w}}_k\|_2}$ attains the maximum value of $|G(\mathbf{w})|$ on $\mathbb{S}^{d-1}$, i.e., $\left|G\left(\frac{\hat{\mathbf{w}}_k}{\|\hat{\mathbf{w}}_k\|_2}\right)\right| = \lambda$;
2. For $1 \leq k \leq p$, $\hat{a}_k = \text{sgn}(G(\hat{\mathbf{w}}_k))\|\hat{\mathbf{w}}_k\|_2$;
3. For $1 \leq k \leq p$, $\langle \hat{\mathbf{w}}_k, \mathbf{x}_i \rangle \neq 0$ for all $i \in [n]$;
4. For $p+1 \leq k \leq m$, $\hat{\mathbf{w}}_k = \mathbf{0}$, $\hat{a}_k = 0$.

Our analysis for Phase I shows that weight vectors approximately align to either of $\bar{\mu}$ or $-\bar{\mu}$, and both of them are maximizers of $|G(\mathbf{w})|$. Therefore, gradient flow goes near a well-aligned parameter vector (with $p = m$) at the end of Phase I.

The following is the main theorem of this subsection.

Theorem E.4. Let $\hat{\theta}$ be a well-aligned parameter vector. Let $\hat{\Delta} := \min_{k \in [p], i \in [n]} \frac{|\langle \hat{\mathbf{w}}_k, \mathbf{x}_i \rangle|}{\|\hat{\theta}\|_{\text{M}}} > 0$. Define $T_2(r) := \frac{1}{\lambda} \ln \frac{1}{r}$ and let t_0 be the following time constant

$$t_0 := \frac{1}{2\lambda} \ln \frac{\lambda \hat{\Delta}}{16m \|\hat{\theta}\|_{\text{M}}^2}. \quad (14)$$

Then for all $t \in (-\infty, t_0]$, the following is true:

1. $\lim_{r \rightarrow 0} \varphi(r\hat{\theta}, T_2(r) + t)$ exists. This limit is independent of the choice of φ when the gradient flow may not be unique.
2. $\lim_{r \rightarrow 0} \varphi(r\hat{\theta}, T_2(r) + t)$ lies near $e^{\lambda t} \hat{\theta}$:

$$\left\| \lim_{r \rightarrow 0} \varphi(r\hat{\theta}, T_2(r) + t) - e^{\lambda t} \hat{\theta} \right\|_{\text{M}} \leq \frac{4m \|\hat{\theta}\|_{\text{M}}^3}{\lambda} e^{3\lambda t}.$$

3. Let $\theta_1, \theta_2, \dots$ be a series of parameters converging to $\mathbf{0}$, $r_1, r_2, \dots$ be a series of positive real numbers converging to 0. If $\|\theta_s - r_s \hat{\theta}\|_2 \leq C r_s^{1+\kappa}$ for some $C > 0, \kappa > 0$, then

$$\lim_{s \rightarrow \infty} \varphi(\theta_s, T_2(r_s) + t) = \lim_{r \rightarrow 0} \varphi(r\hat{\theta}, T_2(r) + t).$$

Now we prove Theorem E.4. Throughout this subsection, we fix a well-aligned parameter vector $\hat{\theta} := (\hat{\mathbf{w}}_1, \dots, \hat{\mathbf{w}}_m, \hat{a}_1, \dots, \hat{a}_m)$ with constant $p \in [m]$. We also use t_0 and $T_2(r)$ to denote the same constant t_0 defined by (14) and the same function $T_2(r) := \frac{1}{\lambda} \ln \frac{1}{r}$ as in Theorem E.4.

For any parameter $\theta = (\mathbf{w}_1, \dots, \mathbf{w}_m, a_1, \dots, a_m)$, we use $\|\theta\|_{\text{P}}$, $\|\theta\|_{\text{R}}$ to denote the following semi-norms respectively,

$$\|\theta\|_{\text{P}} := \max_{k \in [p]} \{\max\{\|\mathbf{w}_k\|_2, |a_k|\}\}, \quad \|\theta\|_{\text{R}} := \max_{p < k \leq m} \{\max\{\|\mathbf{w}_k\|_2, |a_k|\}\}.$$

The M-norm can be expressed in terms of P-norm and R-norm: $\|\theta\|_{\text{M}} = \max\{\|\theta\|_{\text{P}}, \|\theta\|_{\text{R}}\}$. Also note that Condition 4 in Definition E.3 is now equivalent to $\|\hat{\theta}\|_{\text{R}} = 0$.

For $k \in [p]$, define $\widehat{\mathcal{W}}_k := \{\mathbf{w} \in \mathbb{R}^d : \langle \hat{\mathbf{w}}_k, \mathbf{x}_i \rangle \cdot \langle \mathbf{w}, \mathbf{x}_i \rangle > 0, \forall i \in [n]\}$ to be the set of weights that share the same activation pattern as $\hat{\mathbf{w}}_k$.

Lemma E.5. If $r > 0$ is small enough and the initial point θ_0 of gradient flow satisfies $\|\theta_0 - r\hat{\theta}\|_{\text{M}} \leq C r^{1+\kappa}$ for some $C > 0, \kappa > 0$, then for any $-T_2(r) \leq t \leq t_0$, the following four properties hold:

1. For all $k \in [p]$, $\mathbf{w}_k(T_2(r) + t) \in \widehat{\mathcal{W}}_k$;
2. $\|\varphi(\boldsymbol{\theta}_0, T_2(r) + t)\|_{\mathbf{M}} \leq 2e^{\lambda t} \|\hat{\boldsymbol{\theta}}\|_{\mathbf{M}}$;
3. $\|\varphi(\boldsymbol{\theta}_0, T_2(r) + t) - e^{\lambda t} \hat{\boldsymbol{\theta}}\|_{\mathbf{P}} \leq Cr^\kappa e^{\lambda t} + \frac{4m\|\hat{\boldsymbol{\theta}}\|_{\mathbf{M}}^3}{\lambda} e^{3\lambda t}$;
4. $\|\varphi(\boldsymbol{\theta}_0, T_2(r) + t) - e^{\lambda t} \hat{\boldsymbol{\theta}}\|_{\mathbf{R}} \leq 2Cr^\kappa e^{\lambda t}$.

Proof. Let $\boldsymbol{\theta}(t) := \varphi(\boldsymbol{\theta}_0, t)$ be gradient flow on $\mathcal{L}$ starting from $\boldsymbol{\theta}_0$, and $\tilde{\boldsymbol{\theta}}(t) := re^{\lambda t} \hat{\boldsymbol{\theta}}$. Let t_1, t_2 be the following time constants and define $t_{\max} := \min\{t_0, t_1, t_2\}$:

$$\begin{aligned} t_1 &:= \inf\{t \geq 0 : \exists k \in [p], \mathbf{w}_k(t) \notin \widehat{\mathcal{W}}_k\} - T_2(r), \\ t_2 &:= \inf\{t \geq 0 : \|\boldsymbol{\theta}(t)\|_{\mathbf{M}} \geq 2re^{\lambda t} \|\hat{\boldsymbol{\theta}}\|_{\mathbf{M}}\} - T_2(r). \end{aligned}$$

We also define $r_{\max} := \left(\frac{\|\hat{\boldsymbol{\theta}}\|_{\mathbf{M}} \hat{\Delta}}{8C}\right)^{1/\kappa}$. We only consider the dynamics for $r \leq r_{\max}$, $t < T_2(r) + t_{\max}$. Our goal is to show that

$$\|\boldsymbol{\theta}(t) - \tilde{\boldsymbol{\theta}}(t)\|_{\mathbf{P}} \leq Cr^{1+\kappa} e^{\lambda t} + \frac{4m\|\hat{\boldsymbol{\theta}}\|_{\mathbf{M}}^3}{\lambda} r^3 e^{3\lambda t}, \quad \|\boldsymbol{\theta}(t) - \tilde{\boldsymbol{\theta}}(t)\|_{\mathbf{R}} \leq 2Cr^{1+\kappa} e^{\lambda t}$$

within the time interval $[0, T_2(r) + t_{\max}]$ (and thus it also holds for $[0, T_2(r) + t_{\max}]$ by continuity), and to show that t_0 is actually equal to $t_{\max}$, i.e., t_0 is the minimum among t_0, t_1, t_2 . It is easy to see that proving these suffice to deduce the original lemma statement, given the translation of time $e^{\lambda T_2(r)} = \frac{1}{r}$.

For $k \in [m]$, by Lemma B.11 we have

$$\left\| \frac{d\mathbf{w}_k}{dt} - a_k \partial^\circ G(\mathbf{w}_k) \right\|_2 \subseteq (-\infty, m\|\boldsymbol{\theta}\|_{\mathbf{M}}^2 |a_k|], \quad \left| \frac{da_k}{dt} - G(\mathbf{w}_k) \right| \leq m\|\boldsymbol{\theta}\|_{\mathbf{M}}^2 \|\mathbf{w}_k\|_2. \quad (15)$$

For $\tilde{\boldsymbol{\theta}}(t)$, a simple calculus shows that for all $t \geq 0$,

$$\forall k \in [p] : \quad \frac{d\tilde{\mathbf{w}}_k}{dt} = \lambda \tilde{a}_k \frac{\hat{\mathbf{w}}_k}{\|\hat{\mathbf{w}}_k\|_2}, \quad \frac{d\tilde{a}_k}{dt} = \lambda \left\langle \frac{\hat{\mathbf{w}}_k}{\|\hat{\mathbf{w}}_k\|_2}, \tilde{\mathbf{w}}_k \right\rangle. \quad (16)$$

$$\forall p < k \leq m : \quad |\tilde{a}_k| = \|\tilde{\mathbf{w}}_k\|_2 = 0. \quad (17)$$

Bounding $\|\boldsymbol{\theta}(t) - \tilde{\boldsymbol{\theta}}(t)\|_{\mathbf{P}}$. For $k \in [p]$, $\partial^\circ G(\mathbf{w}_k) = \{\nabla G(\mathbf{w}_k)\} = \{\nabla G(\hat{\mathbf{w}}_k)\}$. Also note that $\nabla G(\hat{\mathbf{w}}_k) = \nabla G(\frac{\hat{\mathbf{w}}_k}{\|\hat{\mathbf{w}}_k\|_2}) = \lambda \frac{\hat{\mathbf{w}}_k}{\|\hat{\mathbf{w}}_k\|_2}$ by Lemma B.5. Then $a_k \partial^\circ G(\mathbf{w}_k) = \{\lambda a_k \frac{\hat{\mathbf{w}}_k}{\|\hat{\mathbf{w}}_k\|_2}\}$ and $G(\mathbf{w}_k) = \lambda \left\langle \frac{\hat{\mathbf{w}}_k}{\|\hat{\mathbf{w}}_k\|_2}, \mathbf{w}_k \right\rangle$. Combining these with (15) gives

$$\max \left\{ \left\| \frac{d\mathbf{w}_k}{dt} - \lambda a_k \frac{\hat{\mathbf{w}}_k}{\|\hat{\mathbf{w}}_k\|_2} \right\|_2, \left| \frac{da_k}{dt} - \lambda \left\langle \frac{\hat{\mathbf{w}}_k}{\|\hat{\mathbf{w}}_k\|_2}, \mathbf{w}_k \right\rangle \right| \right\} \leq m\|\boldsymbol{\theta}\|_{\mathbf{M}}^3. \quad (18)$$

Then by (16) we have

$$\begin{aligned} \left\| \frac{d\boldsymbol{\theta}}{dt} - \frac{d\tilde{\boldsymbol{\theta}}}{dt} \right\|_{\mathbf{P}} &\leq m\|\boldsymbol{\theta}\|_{\mathbf{M}}^3 + \max_{k \in [p]} \left\{ \left\| \lambda(a_k - \tilde{a}_k) \frac{\hat{\mathbf{w}}_k}{\|\hat{\mathbf{w}}_k\|_2} \right\|_2, \left| \lambda \left\langle \frac{\hat{\mathbf{w}}_k}{\|\hat{\mathbf{w}}_k\|_2}, \mathbf{w}_k - \tilde{\mathbf{w}}_k \right\rangle \right| \right\} \\ &\leq m\|\boldsymbol{\theta}\|_{\mathbf{M}}^3 + \lambda \|\boldsymbol{\theta} - \tilde{\boldsymbol{\theta}}\|_{\mathbf{P}}. \end{aligned}$$

Taking the integral gives $\|\boldsymbol{\theta}(t) - \tilde{\boldsymbol{\theta}}(t)\|_{\mathbf{P}} \leq \|\boldsymbol{\theta}(0) - \tilde{\boldsymbol{\theta}}(0)\|_{\mathbf{P}} + \int_0^t (m\|\boldsymbol{\theta}(\tau)\|_{\mathbf{M}}^3 + \lambda \|\boldsymbol{\theta}(\tau) - \tilde{\boldsymbol{\theta}}(\tau)\|_{\mathbf{P}}) d\tau$. Note that $t_{\max} \leq t_2$. Then

$$\begin{aligned} \|\boldsymbol{\theta}(t) - \tilde{\boldsymbol{\theta}}(t)\|_{\mathbf{P}} &\leq \|\boldsymbol{\theta}(0) - \tilde{\boldsymbol{\theta}}(0)\|_{\mathbf{P}} + \int_0^t \left(8mr^3 e^{3\lambda t} \|\hat{\boldsymbol{\theta}}\|_{\mathbf{M}}^3 + \lambda \|\boldsymbol{\theta}(\tau) - \tilde{\boldsymbol{\theta}}(\tau)\|_{\mathbf{P}} \right) d\tau \\ &\leq Cr^{1+\kappa} + \frac{8}{3\lambda} mr^3 e^{3\lambda t} \|\hat{\boldsymbol{\theta}}\|_{\mathbf{M}}^3 + \lambda \int_0^t \|\boldsymbol{\theta}(\tau) - \tilde{\boldsymbol{\theta}}(\tau)\|_{\mathbf{P}} d\tau. \end{aligned}$$

By Grönwall's inequality (11), we have

$$\begin{aligned}
\|\theta(t) - \tilde{\theta}(t)\|_P &\leq Cr^{1+\kappa} + \frac{8}{3\lambda}mr^3e^{3\lambda t}\|\hat{\theta}\|_M^3 + \int_0^t \left(Cr^{1+\kappa} + \frac{8}{3\lambda}mr^3e^{3\lambda\tau}\|\hat{\theta}\|_M^3 \right) \lambda e^{\lambda(t-\tau)} d\tau \\
&\leq Cr^{1+\kappa} + Cr^{1+\kappa}(e^{\lambda t} - 1) + \frac{8}{3\lambda}mr^3e^{3\lambda t}\|\hat{\theta}\|_M^3 + \frac{8}{3\lambda}mr^3 \cdot \frac{e^{\lambda t}}{2}(e^{2\lambda t} - 1)\|\hat{\theta}\|_M^3 \\
&\leq Cr^{1+\kappa}e^{\lambda t} + \frac{8}{3\lambda}mr^3e^{3\lambda t}(1 + 1/2)\|\hat{\theta}\|_M^3.
\end{aligned}$$

Therefore we can conclude that

$$\|\theta(t) - \tilde{\theta}(t)\|_P \leq Cr^{1+\kappa}e^{\lambda t} + \frac{4m\|\hat{\theta}\|_M^3}{\lambda}r^3e^{3\lambda t}. \quad (19)$$

Bounding $\|\theta(t) - \tilde{\theta}(t)\|_R$. For $p < k \leq m$, we can combine Theorem B.3 and (15) to give the following bound for the norm growth:

$$\frac{1}{2} \frac{d\|\mathbf{w}_k\|_2^2}{dt} = \frac{1}{2} \frac{d|a_k|^2}{dt} \leq a_k G(\mathbf{w}_k) + |a_k| \cdot m\|\theta\|_M^2 \|\mathbf{w}_k\|_2.$$

This implies

$$\frac{1}{2} \frac{d|a_k|^2}{dt} = \frac{1}{2} \frac{d\|\mathbf{w}_k\|_2^2}{dt} \leq \|\theta\|_R^2 (\lambda + m\|\theta\|_M^2). \quad (20)$$

Taking the integral gives $\|\theta(t)\|_R^2 \leq \|\theta(0)\|_R^2 + \int_0^t 2\|\theta(\tau)\|_R^2 (\lambda + m\|\theta(\tau)\|_M^2) d\tau$. Note that $t_{\max} \leq t_2$ and $\|\theta(0)\|_R \leq Cr^{1+\kappa}$. Then

$$\|\theta(t)\|_R^2 \leq C^2r^{2(1+\kappa)} + \int_0^t 2\|\theta(\tau)\|_R^2 (\lambda + 4mr^2e^{2\lambda\tau}\|\hat{\theta}\|_M^2) d\tau$$

By Grönwall's inequality (12), we have

$$\begin{aligned}
\|\theta(t)\|_R^2 &\leq C^2r^{2(1+\kappa)} \exp \left(\int_0^t 2(\lambda + 4mr^2e^{2\lambda\tau}\|\hat{\theta}\|_M^2) d\tau \right) \\
&\leq C^2r^{2(1+\kappa)} \exp \left(2\lambda t + \frac{4m\|\hat{\theta}\|_M^2}{\lambda}r^2e^{2\lambda t} \right).
\end{aligned}$$

Taking the square root gives

$$\|\theta(t)\|_R \leq Cr^{1+\kappa} \exp \left(\lambda t + \frac{2m\|\hat{\theta}\|_M^2}{\lambda}r^2e^{2\lambda t} \right).$$

For $t \leq T(r) + t_{\max} \leq T(r) + t_0$, we can use the definition (14) of t_0 to deduce that $\frac{2m\|\hat{\theta}\|_M^2}{\lambda}r^2e^{2\lambda t} \leq \frac{2m\|\hat{\theta}\|_M^2}{\lambda}r^2e^{2\lambda t_0} = \hat{\Delta}/8 \leq 1/8$. Therefore, we have

$$\|\theta(t) - \tilde{\theta}(t)\|_R = \|\theta(t)\|_R \leq Cr^{1+\kappa}e^{\lambda t+1/8} < Cr^{1+\kappa}e^{\lambda t+\ln 2} = 2Cr^{1+\kappa}e^{\lambda t}. \quad (21)$$

Bounding $t_{\max}$. To prove the lemma, now we only need to show that $t_{\max} = t_0$. Combining (19) and (21), we have for $t \leq T_2(r) + t_{\max}$,

$$\|\theta(t) - \tilde{\theta}(t)\|_M \leq 2Cr^{1+\kappa}e^{\lambda t} + \frac{4m\|\hat{\theta}\|_M^3}{\lambda}r^3e^{3\lambda t}.$$

Since $r \leq r_{\max}$, $2Cr^\kappa \leq \frac{1}{4}\|\hat{\theta}\|_M\hat{\Delta}$. By definition (14) of t_0 , $\frac{4m\|\hat{\theta}\|_M^3}{\lambda}r^2e^{2\lambda t} \leq \frac{4m\|\hat{\theta}\|_M^3}{\lambda}r^2e^{2\lambda t_0} \leq \frac{1}{4}\|\hat{\theta}\|_M\hat{\Delta}$. Then we have $2Cr^\kappa + \frac{4m\|\hat{\theta}\|_M^3}{\lambda}r^2e^{2\lambda t} \leq \frac{1}{2}\|\hat{\theta}\|_M\hat{\Delta}$ and thus

$$\|\theta(t) - \tilde{\theta}(t)\|_M \leq re^{\lambda t} \left(2Cr^\kappa + \frac{4m\|\hat{\theta}\|_M^3}{\lambda}r^2e^{2\lambda t} \right) \leq \frac{1}{2}re^{\lambda t}\|\hat{\theta}\|_M\hat{\Delta}. \quad (22)$$

For all time $0 \leq t < T_2(r) + t_{\max}$, we can use (22) to deduce

$$\begin{aligned} \operatorname{sgn}(\langle \hat{\mathbf{w}}_k, \mathbf{w}_i \rangle) \langle \mathbf{w}(t), \mathbf{x}_i \rangle &\geq \operatorname{sgn}(\langle \hat{\mathbf{w}}_k, \mathbf{x}_i \rangle) \langle re^{\lambda t} \hat{\mathbf{w}}_k, \mathbf{x}_i \rangle - \frac{1}{2} re^{\lambda t} \|\hat{\boldsymbol{\theta}}\|_{\mathbf{M}} \hat{\Delta} \\ &= re^{\lambda t} \left(|\langle \hat{\mathbf{w}}_k, \mathbf{x}_i \rangle| - \frac{1}{2} \|\hat{\boldsymbol{\theta}}\|_{\mathbf{M}} \hat{\Delta} \right) \\ &\geq re^{\lambda t} \|\hat{\boldsymbol{\theta}}\|_{\mathbf{M}} \hat{\Delta} / 2 > 0, \end{aligned}$$

which implies $t_1 > t_{\max}$.

For norm growth, we can again use (22) to deduce

$$\begin{aligned} \|\boldsymbol{\theta}(t)\|_{\mathbf{M}} &\leq \|\tilde{\boldsymbol{\theta}}(t)\|_{\mathbf{M}} + \frac{1}{2} re^{\lambda t} \|\hat{\boldsymbol{\theta}}\|_{\mathbf{M}} \hat{\Delta} = re^{\lambda t} \left(\|\hat{\boldsymbol{\theta}}\|_{\mathbf{M}} + \frac{1}{2} \|\hat{\boldsymbol{\theta}}\|_{\mathbf{M}} \hat{\Delta} \right) \\ &\leq \frac{3}{2} re^{\lambda t} \|\hat{\boldsymbol{\theta}}\|_{\mathbf{M}} < 2re^{\lambda t} \|\hat{\boldsymbol{\theta}}\|_{\mathbf{M}}, \end{aligned}$$

which implies $t_2 > t_{\max}$.

Now we have $t_1 > t_{\max}, t_2 > t_{\max}$. Recall that $t_{\max} := \min\{t_0, t_1, t_2\}$ by definition. Then $t_{\max} = t_0$ must hold, which completes the proof. $\square$

Lemma E.6. *If $r > 0$ is small enough and the initial point $\boldsymbol{\theta}_0$ of gradient flow satisfies $\|\boldsymbol{\theta}_0 - r\hat{\boldsymbol{\theta}}\|_{\mathbf{M}} \leq Cr^{1+\kappa}$ for some $C > 0, \kappa > 0$, then for all $t \in [-T_2(r), t_0]$,*

$$\|\varphi(\boldsymbol{\theta}_0, T_2(r) + t) - \varphi(r\hat{\boldsymbol{\theta}}, T_2(r) + t)\|_{\mathbf{M}} \leq 4Cr^{\kappa} e^{\lambda t}.$$

Proof. Let $\boldsymbol{\theta}(t) := \varphi(\boldsymbol{\theta}_0, t)$ and $\tilde{\boldsymbol{\theta}}(t) := \varphi(r\hat{\boldsymbol{\theta}}, t)$ be gradient flows starting from $\boldsymbol{\theta}_0$ and $r\hat{\boldsymbol{\theta}}$. For notation simplicity, let $h_{ki} = y_i \phi'(\hat{\mathbf{w}}_k^\top \mathbf{x}_i)$. Let $g_i := -\ell'(y_i f_{\boldsymbol{\theta}}(\mathbf{x}_i))$, $\tilde{g}_i := -\ell'(y_i f_{\tilde{\boldsymbol{\theta}}}(\mathbf{x}_i))$.

By Lemma E.5, we can make r to be small enough so that the four properties hold for both $\boldsymbol{\theta}(T_2(r) + t)$ and $\tilde{\boldsymbol{\theta}}(T_2(r) + t)$ when $t \leq t_0$.

Bounding the Difference for $1 \leq k \leq p$. For all $t \leq t_0$ and $k \in [p]$, we know that $\phi'(\mathbf{w}_k^\top \mathbf{x}_i) = \phi'(\tilde{\mathbf{w}}_k^\top \mathbf{x}_i) = h_{ki}$, and thus for $\mathbf{w}_k, \tilde{\mathbf{w}}_k$ we have

$$\begin{aligned} \left\| \frac{d\mathbf{w}_k}{dt} - \frac{d\tilde{\mathbf{w}}_k}{dt} \right\|_2 &= \left\| \frac{a_k}{n} \sum_{i=1}^n g_i h_{ki} \mathbf{x}_i - \frac{\tilde{a}_k}{n} \sum_{i=1}^n \tilde{g}_i h_{ki} \mathbf{x}_i \right\|_2 \\ &\leq |a_k - \tilde{a}_k| \cdot \underbrace{\left\| \frac{1}{n} \sum_{i=1}^n \tilde{g}_i h_{ki} \mathbf{x}_i \right\|_2}_{\Lambda(t)} + |a_k| \cdot \underbrace{\left\| \frac{1}{n} \sum_{i=1}^n (g_i - \tilde{g}_i) h_{ki} \mathbf{x}_i \right\|_2}_{\Delta(t)} \\ &=: \Lambda(t) \cdot |a_k - \tilde{a}_k| + |a_k| \cdot \Delta(t). \end{aligned}$$

and for $a_k, \tilde{a}_k$ we have

$$\begin{aligned} \left\| \frac{da_k}{dt} - \frac{d\tilde{a}_k}{dt} \right\|_2 &= \left\| \frac{1}{n} \sum_{i=1}^n g_i \phi(\mathbf{w}_k^\top \mathbf{x}_i) - \frac{1}{n} \sum_{i=1}^n \tilde{g}_i \phi(\tilde{\mathbf{w}}_k^\top \mathbf{x}_i) \right\|_2 \\ &= \left\| \frac{1}{n} \sum_{i=1}^n g_i h_{ki} \mathbf{w}_k^\top \mathbf{x}_i - \frac{1}{n} \sum_{i=1}^n \tilde{g}_i h_{ki} \tilde{\mathbf{w}}_k^\top \mathbf{x}_i \right\|_2 \\ &= \|\mathbf{w}_k - \tilde{\mathbf{w}}_k\|_2 \cdot \left\| \frac{1}{n} \sum_{i=1}^n \tilde{g}_i h_{ki} \mathbf{x}_i \right\|_2 + \|\mathbf{w}_k\|_2 \cdot \left\| \frac{1}{n} \sum_{i=1}^n (g_i - \tilde{g}_i) h_{ki} \mathbf{x}_i \right\|_2 \\ &= \Lambda(t) \cdot \|\mathbf{w}_k - \tilde{\mathbf{w}}_k\|_2 + \|\mathbf{w}_k\|_2 \cdot \Delta(t). \end{aligned}$$

Therefore, $\left\| \frac{d\boldsymbol{\theta}}{dt} - \frac{d\tilde{\boldsymbol{\theta}}}{dt} \right\|_{\mathbf{P}} \leq \Lambda(t) \cdot \|\boldsymbol{\theta} - \tilde{\boldsymbol{\theta}}\|_{\mathbf{M}} + \|\boldsymbol{\theta}\|_{\mathbf{M}} \cdot \Delta(t)$. Now we turn to bound $\Lambda(t)$ and $\Delta(t)$.

By Lipschitzness of ℓ' and Lemma B.10, we have

$$|-\ell'(0) - \tilde{g}_i| \leq m \|\tilde{\boldsymbol{\theta}}\|_{\mathbf{M}}^2, \quad |g_i - \tilde{g}_i| \leq m \|\boldsymbol{\theta} - \tilde{\boldsymbol{\theta}}\|_{\mathbf{M}} \left(\|\boldsymbol{\theta}\|_{\mathbf{M}} + \|\tilde{\boldsymbol{\theta}}\|_{\mathbf{M}} \right).$$

For $\Lambda(t)$, by triangle inequality and Lemma B.5 we have

$$\Lambda(t) \leq \left\| \frac{-\ell'(0)}{n} \sum_{i=1}^n h_{ki} \mathbf{x}_i \right\|_2 + m \|\tilde{\boldsymbol{\theta}}\|_{\mathbf{M}}^2 = \|\nabla G(\hat{\mathbf{w}}_k)\|_2 + m \|\tilde{\boldsymbol{\theta}}\|_{\mathbf{M}}^2 = \lambda + m \|\tilde{\boldsymbol{\theta}}\|_{\mathbf{M}}^2,$$

For $\Delta(t)$, we use triangle inequality again to give the following bound:

$$\Delta(t) \leq \frac{1}{n} \sum_{i=1}^n |g_i - \tilde{g}_i| \leq m \|\boldsymbol{\theta} - \tilde{\boldsymbol{\theta}}\|_{\mathbf{M}} \left(\|\boldsymbol{\theta}\|_{\mathbf{M}} + \|\tilde{\boldsymbol{\theta}}\|_{\mathbf{M}} \right).$$

Therefore, we can conclude that

$$\begin{aligned} \left\| \frac{d\boldsymbol{\theta}}{dt} - \frac{d\tilde{\boldsymbol{\theta}}}{dt} \right\|_{\mathbf{P}} &\leq (\lambda + m \|\tilde{\boldsymbol{\theta}}\|_{\mathbf{M}}^2) \cdot \|\boldsymbol{\theta} - \tilde{\boldsymbol{\theta}}\|_{\mathbf{M}} + \|\boldsymbol{\theta}\|_{\mathbf{M}} \cdot m \|\boldsymbol{\theta} - \tilde{\boldsymbol{\theta}}\|_{\mathbf{M}} \left(\|\boldsymbol{\theta}\|_{\mathbf{M}} + \|\tilde{\boldsymbol{\theta}}\|_{\mathbf{M}} \right) \\ &\leq \left(\lambda + 3m \max\{\|\boldsymbol{\theta}\|_{\mathbf{M}}, \|\tilde{\boldsymbol{\theta}}\|_{\mathbf{M}}\}^2 \right) \|\boldsymbol{\theta} - \tilde{\boldsymbol{\theta}}\|_{\mathbf{M}} \\ &\leq \left(\lambda + 12mr^2 e^{2\lambda t} \|\hat{\boldsymbol{\theta}}\|_{\mathbf{M}}^2 \right) \|\boldsymbol{\theta} - \tilde{\boldsymbol{\theta}}\|_{\mathbf{M}}, \end{aligned}$$

where the last inequality uses the 2nd property in Lemma E.5. Note that $\|\boldsymbol{\theta}_0 - r\hat{\boldsymbol{\theta}}\|_{\mathbf{P}} \leq Cr^{1+\kappa}$. So we can write it into the integral form:

$$\|\boldsymbol{\theta}(t) - \tilde{\boldsymbol{\theta}}(t)\|_{\mathbf{P}} \leq Cr^{1+\kappa} + \int_0^t \left(\lambda + 12mr^2 e^{2\lambda\tau} \|\hat{\boldsymbol{\theta}}\|_{\mathbf{M}}^2 \right) \|\boldsymbol{\theta}(\tau) - \tilde{\boldsymbol{\theta}}(\tau)\|_{\mathbf{M}} d\tau. \quad (23)$$

Bounding the Difference for $p < k \leq m$. By Lemma B.17, $\|\tilde{\boldsymbol{\theta}}(t)\|_{\mathbf{R}} = 0$ for all $t \geq 0$, so $\|\boldsymbol{\theta} - \tilde{\boldsymbol{\theta}}\|_{\mathbf{R}} = \|\boldsymbol{\theta}\|_{\mathbf{R}}$. By the 4th property in Lemma E.5, we then have

$$\|\boldsymbol{\theta}(t) - \tilde{\boldsymbol{\theta}}(t)\|_{\mathbf{R}} = \|\boldsymbol{\theta}(t)\|_{\mathbf{R}} = \|\boldsymbol{\theta}(t) - re^{\lambda t} \hat{\boldsymbol{\theta}}\|_{\mathbf{R}} \leq 2Cr^{1+\kappa} e^{\lambda t}.$$

So we can verify that $\|\boldsymbol{\theta}(t) - \tilde{\boldsymbol{\theta}}(t)\|_{\mathbf{R}}$ satisfies the following inequality:

$$\|\boldsymbol{\theta}(t) - \tilde{\boldsymbol{\theta}}(t)\|_{\mathbf{R}} \leq 2Cr^{1+\kappa} + \int_0^t \lambda \|\boldsymbol{\theta}(\tau) - \tilde{\boldsymbol{\theta}}(\tau)\|_{\mathbf{R}} d\tau. \quad (24)$$

Bounding the Difference for All. Combining Lemma E.5 and Lemma E.5, we have the following inequality for $\|\boldsymbol{\theta}(t) - \tilde{\boldsymbol{\theta}}(t)\|_{\mathbf{M}}$:

$$\|\boldsymbol{\theta}(t) - \tilde{\boldsymbol{\theta}}(t)\|_{\mathbf{M}} \leq 2Cr^{1+\kappa} + \int_0^t \left(\lambda + 12mr^2 e^{2\lambda\tau} \|\hat{\boldsymbol{\theta}}\|_{\mathbf{M}}^2 \right) \|\boldsymbol{\theta}(\tau) - \tilde{\boldsymbol{\theta}}(\tau)\|_{\mathbf{M}} d\tau.$$

By Grönwall's inequality (12),

$$\begin{aligned} \|\boldsymbol{\theta}(t) - \tilde{\boldsymbol{\theta}}(t)\|_{\mathbf{M}} &\leq 2Cr^{1+\kappa} \exp \left(\int_0^t \left(\lambda + 12mr^2 e^{2\lambda\tau} \|\hat{\boldsymbol{\theta}}\|_{\mathbf{M}}^2 \right) d\tau \right) \\ &\leq 2Cr^{1+\kappa} \exp \left(\lambda t + \frac{6m \|\hat{\boldsymbol{\theta}}\|_{\mathbf{M}}^2}{\lambda} r^2 e^{2\lambda t} \right). \end{aligned}$$

By definition (14) of t_0 , we have $\frac{6m \|\hat{\boldsymbol{\theta}}\|_{\mathbf{M}}^2}{\lambda} r^2 e^{2\lambda t} \leq \frac{6m \|\hat{\boldsymbol{\theta}}\|_{\mathbf{M}}^2}{\lambda} r^2 e^{2\lambda t_0} = \frac{3\hat{\Delta}}{8} \leq 3/8 < \ln 2$. Therefore we have the following bound for $\|\boldsymbol{\theta}(t) - \tilde{\boldsymbol{\theta}}(t)\|_{\mathbf{M}}$:

$$\|\boldsymbol{\theta}(t) - \tilde{\boldsymbol{\theta}}(t)\|_{\mathbf{M}} \leq 2Cr^{1+\kappa} e^{\lambda t + \ln 2} = 4Cr^{1+\kappa} e^{\lambda t}.$$

At time $T_2(r) + t \in [0, T_2(r) + t_0]$, this bound can be rewritten as

$$\|\boldsymbol{\theta}(T_2(r) + t) - \tilde{\boldsymbol{\theta}}(T_2(r) + t)\|_{\mathbf{M}} \leq 4Cr^{\kappa} e^{\lambda t},$$

which completes the proof. $\square$

Proof for Theorem E.4. First we show that $\lim_{r \rightarrow 0} \varphi(r\hat{\boldsymbol{\theta}}, T_2(r) + t)$ exists. We consider the case of $r \leq r_{\max}$, where $r_{\max}$ is chosen to be small enough so that the properties in Lemma E.5 hold. For any $r' < r$, by Lemma E.5 we have

$$\left\| \varphi \left(r'\hat{\boldsymbol{\theta}}, T_2(r') + \frac{1}{\lambda} \ln r \right) - r\hat{\boldsymbol{\theta}} \right\|_{\text{M}} \leq \frac{4m\|\hat{\boldsymbol{\theta}}\|_{\text{M}}^3}{\lambda} r^3 \leq C' r^{1+\kappa'},$$

where $C' = \frac{4m\|\hat{\boldsymbol{\theta}}\|_{\text{M}}^3}{\lambda}$, $\kappa' = 2$. Applying Lemma E.6, we then have

$$\left\| \varphi \left(\varphi \left(r'\hat{\boldsymbol{\theta}}, T_2(r') + \frac{1}{\lambda} \ln r \right), T_2(r) + t \right) - \varphi \left(r\hat{\boldsymbol{\theta}}, T_2(r) + t \right) \right\|_{\text{M}} \leq 4C' r^{\kappa'} e^{\lambda t}.$$

Note that $T_2(r') + \frac{1}{\lambda} \ln r + T_2(r) + t = T_2(r') + t$. So this proves

$$\|\varphi(r'\hat{\boldsymbol{\theta}}, T_2(r') + t) - \varphi(r\hat{\boldsymbol{\theta}}, T_2(r) + t)\|_{\text{M}} \leq 4C' r^{\kappa'} e^{\lambda t}.$$

For any fixed $t \leq t_0$, the RHS converges to 0 as $r \rightarrow 0$, which implies Cauchy convergence of the limit $\lim_{r \rightarrow 0} \varphi(r\hat{\boldsymbol{\theta}}, T_2(r) + t)$ and thus the limit exists. By the 1st property in Lemma E.5, we know that there is no activation pattern switch in the time interval $t \in [0, T_2(r) + t_0]$ if r is small enough. This means $\mathcal{L}$ is locally smooth near the trajectory of $\varphi(r\hat{\boldsymbol{\theta}}, T_2(r) + t)$ and thus the trajectory is unique. Therefore, the limit $\lim_{r \rightarrow 0} \varphi(r\hat{\boldsymbol{\theta}}, T_2(r) + t)$ is uniquely defined.

By Lemma E.5,

$$\|\varphi(r\hat{\boldsymbol{\theta}}, T_2(r) + t) - e^{\lambda t} \hat{\boldsymbol{\theta}}\|_{\text{M}} \leq \frac{4m\|\hat{\boldsymbol{\theta}}\|_{\text{M}}^3}{\lambda} e^{3\lambda t}.$$

Taking $r \rightarrow 0$ on both sides gives the range of the limit $\lim_{r \rightarrow 0} \varphi(r\hat{\boldsymbol{\theta}}, T_2(r) + t)$:

$$\left\| \lim_{r \rightarrow 0} \varphi(r\hat{\boldsymbol{\theta}}, T_2(r) + t) - e^{\lambda t} \hat{\boldsymbol{\theta}} \right\|_{\text{M}} \leq \frac{4m\|\hat{\boldsymbol{\theta}}\|_{\text{M}}^3}{\lambda} e^{3\lambda t}.$$

For $s \rightarrow \infty$, by Lemma E.6, we have

$$\lim_{s \rightarrow \infty} \left\| \varphi \left(\hat{\boldsymbol{\theta}}_s, T_2(r_s) + t \right) - \varphi \left(r_s \hat{\boldsymbol{\theta}}, T_2(r_s) + t \right) \right\|_{\text{M}} = 0.$$

So $\lim_{s \rightarrow \infty} \varphi(\hat{\boldsymbol{\theta}}_s, T_2(r_s) + t) = \lim_{r \rightarrow 0} \varphi(r\hat{\boldsymbol{\theta}}, T_2(r) + t)$ is proved. $\square$

E.3 Proof for Approximate Embedding

To analyze Phase II, we need to deal with approximate embedding instead of the exact one. For this, we further divide Phase II into Phase II.1 and II.2 and analyze them in order. At the end of this subsection we will prove Lemma 5.4.

E.3.1 Proofs for Phase II.1

Given the discussions in the previous sections, we are ready to present proofs for the phase II dynamics (Lemma 5.4) here.

We subdivide the dynamics of Phase II into Phase II.1 and Phase II.2. At the end of Phase I, we show that the parameters grow to norm $O(r)$ in time $T_1(r)$. In Phase II.1, we extend the dynamic to time $T_1(r) + T_2(r)$ so that the parameters grow into constant norms (irrelevant to r and σ_{init}). Then, when the initialization scale becomes sufficiently small, at the end of Phase II.1 the parameters become sufficiently close to the embedded parameters from two neurons at constant norms, so the subsequent dynamics is a good approximate embedding until the norm of the parameters grow sufficiently large to ensure directional convergence in Phase III. Here we show the results in Phase II.1.

Lemma E.7. For $m \geq 2$, with probability $1 - 2^{-(m-1)}$ over the random draw of $\bar{\boldsymbol{\theta}}_0 \sim \mathcal{D}_{\text{init}}(1)$, the vector $\bar{\mathbf{b}} \in \mathbb{R}^m$ with entries $\bar{b}_k := \frac{\langle \bar{\mathbf{w}}_k, \bar{\boldsymbol{\mu}} \rangle + \bar{a}_k}{2\sqrt{m}\|\bar{\boldsymbol{\theta}}_0\|_{\text{M}}}$ defined as in Lemma 5.2 is a good embedding vector.

Proof. Note that $\bar{\mathbf{b}}$ is a good embedding vector iff $\bar{\mathbf{b}}' = 2\sqrt{m}\|\bar{\boldsymbol{\theta}}_0\|_{\text{M}}\bar{\mathbf{b}}$ is a good embedding vector. Recall that $\bar{\mathbf{w}}_k \stackrel{\text{i.i.d.}}{\sim} \mathcal{N}(\mathbf{0}, \mathbf{I})$, $\bar{a}_k \stackrel{\text{i.i.d.}}{\sim} \mathcal{N}(0, c_{\text{a init}}^2)$. By the property of Gaussian variables,

$$\bar{b}'_k = \langle \bar{\mathbf{w}}_k, \bar{\boldsymbol{\mu}} \rangle + \bar{a}_k \sim \mathcal{N}(0, 1 + c_{\text{a init}}^2).$$

Thus $\bar{\mathbf{b}}' \sim \mathcal{N}(\mathbf{0}, (1 + c_{\text{ainit}}^2)\mathbf{I})$. Since it is a continuous probability distribution, $\bar{\mathbf{b}}' \neq \mathbf{0}$ with probability 1. By symmetry and independence, we know that $\Pr[\forall k \in [m] : \bar{b}'_k > 0] = 2^{-m}$ and $\Pr[\forall k \in [m] : \bar{b}'_k < 0] = 2^{-m}$. So $\bar{\mathbf{b}}'$ is a good embedding vector with probability $1 - 2^{-m} - 2^{-m} = 1 - 2^{-(m-1)}$, and so is $\bar{\mathbf{b}}$. $\square$

Lemma E.8. Let $T_2(r) := \frac{1}{\lambda_0} \ln \frac{1}{r}$, then $T_{12} := T_1(r) + T_2(r) = \frac{1}{\lambda_0} \ln \frac{1}{\sqrt{m}\sigma_{\text{init}}\|\bar{\boldsymbol{\theta}}_0\|_{\text{M}}}$ regardless the choice of r . For random draw of $\bar{\boldsymbol{\theta}}_0 \sim \mathcal{D}_{\text{init}}(1)$, if $\bar{\mathbf{b}} \in \mathbb{R}^m$ defined as in Lemma 5.2 is a good embedding vector, then there exists $t_0 \in \mathbb{R}$ such that the following holds:

1. For the two-neuron dynamics starting with rescaled initialization in the direction of $\hat{\boldsymbol{\theta}} := (\bar{b}_+, \bar{b}_+ \bar{\boldsymbol{\mu}}, \bar{b}_-, \bar{b}_- \bar{\boldsymbol{\mu}})$, for all $t \in (-\infty, t_0]$, the limit $\tilde{\boldsymbol{\theta}}(t) := \lim_{r \rightarrow 0} \varphi(r\hat{\boldsymbol{\theta}}, T_2(r) + t)$ exists and lies near $e^{\lambda_0 t} \hat{\boldsymbol{\theta}}$:

$$\left\| \tilde{\boldsymbol{\theta}}(t) - e^{\lambda_0 t} \hat{\boldsymbol{\theta}} \right\|_{\text{M}} \leq \frac{4m\|\hat{\boldsymbol{\theta}}\|_{\text{M}}^3}{\lambda_0} e^{3\lambda_0 t} = O(e^{3\lambda_0 t}).$$

2. For the m -neuron dynamics $\boldsymbol{\theta}(t)$, the following holds for all $t \in (-\infty, t_0]$,

$$\lim_{\sigma_{\text{init}} \rightarrow 0} \boldsymbol{\theta}(T_{12} + t) = \pi_{\bar{\mathbf{b}}}(\tilde{\boldsymbol{\theta}}(t)).$$

Proof. Under Assumptions 4.1 and 4.5, the maximum value of $|G(\mathbf{w})|$ on $\mathbb{S}^{d-1}$ is λ_0 and is attained at $\pm \bar{\boldsymbol{\mu}}$. Given a good embedding vector $\bar{\mathbf{b}}$, both $\hat{\boldsymbol{\theta}}$ and $\hat{\boldsymbol{\theta}}_{\pi} := \pi_{\bar{\mathbf{b}}}(\hat{\boldsymbol{\theta}})$ are well-aligned parameter vectors (Definition E.3) for width-2 and width- m Leaky ReLU nets respectively. By Theorem E.4, there exists $t_0 \in \mathbb{R}$ such that the following two limits exist for all $t \in (-\infty, t_0]$:

$$\tilde{\boldsymbol{\theta}}(t) := \lim_{r \rightarrow 0} \varphi(r\hat{\boldsymbol{\theta}}, T_2(r) + t), \quad \tilde{\boldsymbol{\theta}}_{\pi}(t) := \lim_{r \rightarrow 0} \varphi(r\hat{\boldsymbol{\theta}}_{\pi}, T_2(r) + t).$$

Note that by Lemma E.1, we have $\pi_{\bar{\mathbf{b}}}(\varphi(r\hat{\boldsymbol{\theta}}, T_2(r) + t))$ is a trajectory of gradient flow starting from $r\hat{\boldsymbol{\theta}}_{\pi}$. The uniqueness of $\tilde{\boldsymbol{\theta}}_{\pi}(t)$ (for all possible choices of φ) shows that

$$\pi_{\bar{\mathbf{b}}}(\tilde{\boldsymbol{\theta}}(t)) = \lim_{r \rightarrow 0} \pi_{\bar{\mathbf{b}}}(\varphi(r\hat{\boldsymbol{\theta}}, T_2(r) + t)) = \tilde{\boldsymbol{\theta}}_{\pi}(t).$$

By Lemma 5.2, for σ_{init} small enough, if we choose r so that $\sigma_{\text{init}} = \frac{r^3}{\sqrt{m}\|\bar{\boldsymbol{\theta}}_0\|_{\text{M}}}$, then for some universal constant C we have

$$\|\boldsymbol{\theta}(T_1(r)) - r\hat{\boldsymbol{\theta}}_{\pi}\|_{\text{M}} \leq \frac{Cr^3}{\sqrt{m}}.$$

Applying Theorem E.4 proves the following for all $t \in (-\infty, t_0]$:

$$\lim_{\sigma_{\text{init}} \rightarrow 0} \varphi(\boldsymbol{\theta}(T_1(r)), T_2(r) + t) = \tilde{\boldsymbol{\theta}}_{\pi}(t).$$

Therefore $\lim_{\sigma_{\text{init}} \rightarrow 0} \boldsymbol{\theta}(T_{12} + t) = \pi_{\bar{\mathbf{b}}}(\tilde{\boldsymbol{\theta}}(t))$. $\square$

E.4 Proofs for Phase II.2

Next, at the end of Phase II.1, $\boldsymbol{\theta}(T_{12} + t_0)$ has a constant norm. Then we show the trajectory convergence with respect to the initialization scale in Phase II.2.

Lemma E.9. If $\pi_{\bar{\mathbf{b}}}(\tilde{\boldsymbol{\theta}}(t_0))$ is non-branching and $\lim_{\sigma_{\text{init}} \rightarrow 0} \boldsymbol{\theta}(T_{12} + t_0) = \pi_{\bar{\mathbf{b}}}(\tilde{\boldsymbol{\theta}}(t_0))$ for some constant t_0 , then for all $t > t_0$, $\lim_{\sigma_{\text{init}} \rightarrow 0} \boldsymbol{\theta}(T_{12} + t) = \pi_{\bar{\mathbf{b}}}(\tilde{\boldsymbol{\theta}}(t))$.

We first start with a simple lemma on gradient upper bounds, and then show that the trajectory of gradient flow is Lipschitz with time.

Lemma E.10. For every $\boldsymbol{\theta} \in \mathbb{R}^D$, $\|\mathbf{g}\|_2 \leq \|\boldsymbol{\theta}\|_2$ for all $\mathbf{g} \in \partial^{\circ} \mathcal{L}(\boldsymbol{\theta})$.

Proof. Note that $|\ell'(q)| \leq 1$, $|\phi'(z)| \leq 1$, $\|\mathbf{x}_i\|_2 \leq 1$, $|y_i| \leq 1$. For every $\boldsymbol{\theta} \in \Omega_{\mathcal{S}}$ (i.e., no activation function has zero input), by the chain rule (10), we have

$$\begin{aligned}\left\| \frac{\partial \mathcal{L}(\boldsymbol{\theta})}{\partial \mathbf{w}_k} \right\|_2 &= \left\| \frac{1}{n} \sum_{i \in [n]} \ell'(q_i(\boldsymbol{\theta})) y_i a_k \phi'(\mathbf{w}_k^\top \mathbf{x}_i) \mathbf{x}_i \right\|_2 \leq \frac{1}{n} \sum_{i \in [n]} |a_k| = |a_k|, \\ \left\| \frac{\partial \mathcal{L}(\boldsymbol{\theta})}{\partial a_k} \right\|_2 &= \left\| \frac{1}{n} \sum_{i \in [n]} \ell'(q_i(\boldsymbol{\theta})) y_i \phi(\mathbf{w}_k^\top \mathbf{x}_i) \right\|_2 \leq \frac{1}{n} \sum_{i \in [n]} |\mathbf{w}_k^\top \mathbf{x}_i| \leq \|\mathbf{w}_k\|_2.\end{aligned}$$

So $\|\nabla \mathcal{L}(\boldsymbol{\theta})\|_2 \leq \|\boldsymbol{\theta}\|_2$. We can finish the proof for any $\boldsymbol{\theta} \in \mathbb{R}^D$ by taking limits in $\Omega_{\mathcal{S}}$. $\square$

Lemma E.11. *The gradient flow trajectory $\boldsymbol{\theta}(T_{12} + t)$, in the interval $t \in [t_0, t_s]$ for any $t_s > t_0$, is Lipschitz in t with Lipschitz constant $\|\boldsymbol{\theta}(T_{12} + t_0)\|_2 e^{(t_s - t_0)}$.*

Proof. By Lemma E.10, $\left\| \frac{d}{dt} \boldsymbol{\theta}(T_{12} + t) \right\|_2 \leq \|\boldsymbol{\theta}(T_{12} + t)\|_2$. So $\frac{d\|\boldsymbol{\theta}(T_{12} + t)\|_2}{dt} \leq \|\boldsymbol{\theta}(T_{12} + t)\|_2$. By Grönwall's inequality (12), $\|\boldsymbol{\theta}(T_{12} + t)\|_2 \leq \|\boldsymbol{\theta}(T_{12} + t_0)\|_2 e^{t - t_0}$. Then $\left\| \frac{d}{dt} \boldsymbol{\theta}(T_{12} + t) \right\|_2 \leq \|\boldsymbol{\theta}(T_{12} + t)\|_2 \leq \|\boldsymbol{\theta}(T_{12} + t_0)\|_2 e^{t_s - t_0}$. $\square$

Now we are ready to prove Lemma E.9.

Proof of Lemma E.9. When $\sigma_{\text{init}} \rightarrow 0$, as $\boldsymbol{\theta}(T_{12} + t_0) \rightarrow \pi_{\tilde{\mathbf{b}}}(\tilde{\boldsymbol{\theta}}(t_0))$, we can select a countable sequence $(\sigma_{\text{init}})_i \rightarrow 0$ and trajectories $\boldsymbol{\theta}_i(T_{12} + t)$ with initialization scale $(\sigma_{\text{init}})_i$. We show that for every $t \geq t_0$, there must be $\boldsymbol{\theta}_i(T_{12} + t) = \varphi(\boldsymbol{\theta}_i(T_{12} + t_0), t - t_0) \rightarrow \pi_{\tilde{\mathbf{b}}}(\tilde{\boldsymbol{\theta}}(t)) = \varphi(\pi_{\tilde{\mathbf{b}}}(\tilde{\boldsymbol{\theta}}(t_0)), t - t_0)$.

If this does not hold for some $t = T$, then there must be a limit point $\mathbf{q}_T$ of points in $\{\varphi(\boldsymbol{\theta}_i(T_{12} + t_0), T - t_0)\}_{i=1}^\infty$ such that $\mathbf{q}_T \neq \varphi(\pi_{\tilde{\mathbf{b}}}(\tilde{\boldsymbol{\theta}}(t_0)), T - t_0)$ and a converging subsequence in $\{\varphi(\boldsymbol{\theta}_i(T_{12} + t_0), T - t_0)\}_{i=1}^\infty$ to $\mathbf{q}_T$. Thus WLOG we assume that the sequence is chosen so that

$$\lim_{i \rightarrow \infty} \varphi(\boldsymbol{\theta}_i(T_{12} + t_0), T - t_0) = \mathbf{q}_T \neq \varphi(\pi_{\tilde{\mathbf{b}}}(\tilde{\boldsymbol{\theta}}(t_0)), T - t_0).$$

We then show that there is a trajectory of the gradient flow that starts from $\pi_{\tilde{\mathbf{b}}}(\tilde{\boldsymbol{\theta}}(t_0))$ and reaches $\mathbf{q}_T$ at time $T - t_0$, thereby causing a contradiction to our assumption that $\pi_{\tilde{\mathbf{b}}}(\tilde{\boldsymbol{\theta}}(t_0))$ is non-branching.

For any pair of L_0 -Lipschitz continuous functions $\mathbf{f}, \mathbf{g} : [t_0, T] \rightarrow \mathbb{R}^D$, define the L^∞ -distance to be $\|\mathbf{f} - \mathbf{g}\|_\infty := \sup_{t \in [t_0, T]} \|\mathbf{f}(t) - \mathbf{g}(t)\|_2$. Note that the space of L_0 -Lipschitz functions with bounded function values is compact under L^∞ -distance, and therefore any sequence of functions in this space has a converging subsequence whose limit is also L_0 -Lipschitz.

Let $C := \sup_i \{\|\boldsymbol{\theta}_i(T_{12} + t_0)\|_2\}$, then as $\{\boldsymbol{\theta}_i(T_{12} + t_0)\}$ is converging to $\pi_{\tilde{\mathbf{b}}}(\tilde{\boldsymbol{\theta}}(t_0)) \neq \infty$, $C < \infty$. By Lemma E.11 we know each trajectory $\boldsymbol{\theta}_i(T_{12} + t)$ is $(Ce^{T - t_0})$ -Lipschitz for $t \in [t_0, T]$. This means we can find a subsequence $1 \leq i_1 < i_2 < i_3 < \dots$ that the trajectory $\{\boldsymbol{\theta}_{i_j}(T_{12} + t)\}$ L^∞ -converges on $[t_0, T]$ as $j \rightarrow \infty$. Then the pointwise limit $\mathbf{q}(t) := \lim_{j \rightarrow \infty} \boldsymbol{\theta}_{i_j}(T_{12} + t)$ exists for every $t \in [t_0, T]$. $\mathbf{q}(t_0) = \pi_{\tilde{\mathbf{b}}}(\tilde{\boldsymbol{\theta}}(t_0))$, $\mathbf{q}(T) = \mathbf{q}_T$.

Finally we show that $\mathbf{q}(t)$ is indeed a valid gradient flow trajectory. Notice that $\mathbf{q}(t)$ is $(Ce^{T - t_0})$ -Lipschitz, then by Rademacher theorem for $\mathbf{q}(t)$ is differentiable for a.e. $t \in [t_0, T]$. We are left to show $\mathbf{q}'(t) \in \partial^\circ \mathcal{L}(\mathbf{q}(t))$ whenever $\mathbf{q}$ is differentiable at t .

For any $\epsilon > 0$ that $[t, t + \epsilon] \subseteq [t_0, T]$, we investigate the behaviour of $\mathbf{q}(t)$ in the ϵ -neighborhood of t . Let Ω_j be the set of $\tau \in [t_0, T]$ so that $\frac{d}{d\tau} \boldsymbol{\theta}_{i_k}(T_{12} + \tau) \in -\partial^\circ \mathcal{L}(\boldsymbol{\theta}_{i_k}(T_{12} + \tau))$. By definition of differential inclusion, Ω_j has full measure in $[t_0, T]$. Define $B_{j,\epsilon}$ be the following closed convex hull:

$$B_{j,\epsilon} = \overline{\text{conv}} \left\{ \frac{d}{d\tau} \boldsymbol{\theta}_{i_k}(T_{12} + \tau) : k \geq j, \tau \in [t, t + \epsilon] \cap \Omega_j \right\}.$$

It is easy to see that $B_{j,\epsilon}$ is monotonic with respect to j . Then we know that for any j ,

$$\frac{\boldsymbol{\theta}_{i_j}(T_{12} + t + \epsilon) - \boldsymbol{\theta}_{i_j}(T_{12} + t)}{\epsilon} = \frac{1}{\epsilon} \int_t^{t+\epsilon} \frac{d}{d\tau} \boldsymbol{\theta}_{i_j}(T_{12} + \tau) d\tau \in B_{j,\epsilon},$$

Then taking the limits $j \rightarrow \infty$, as all $B_{j,\epsilon}$ are closed, we know $\frac{\mathbf{q}(t+\epsilon)-\mathbf{q}(t)}{\epsilon} \in \lim_{j \rightarrow \infty} B_{j,\epsilon}$.

Now let $C_{j,\epsilon}, C_\epsilon$ be the following closed convex hull of subgradients:

$$C_{j,\epsilon} = \overline{\text{conv}} \left(\bigcup_{\substack{k \geq j \\ \tau \in [t, t+\epsilon]}} \partial^\circ \mathcal{L}(\theta_{i_k}(T_{12} + \tau)) \right), \quad C_\epsilon = \overline{\text{conv}} \left(\bigcup_{\tau \in [t, t+\epsilon]} \partial^\circ \mathcal{L}(\mathbf{q}(T_{12} + \tau)) \right).$$

Then we know $B_{j,\epsilon} \subseteq -C_{j,\epsilon}$ for all $j \geq 1$ and $\epsilon > 0$. Notice that $C_{j,\epsilon}$ and C_ϵ are also monotonic with respect to j and ϵ respectively so we can take the respective limit. As for $\tau \in [t, t+\epsilon]$, $\lim_{j \rightarrow \infty} \theta_{i_j}(T_{12} + \tau) = \mathbf{q}(T_{12} + \tau)$, by the upper-semicontinuity of $\partial^\circ \mathcal{L}$, $\lim_{j \rightarrow \infty} C_{j,\epsilon} \subseteq C_\epsilon$. Then $\frac{\mathbf{q}(t+\epsilon)-\mathbf{q}(t)}{\epsilon} \in \lim_{j \rightarrow \infty} B_{j,\epsilon} \subseteq \lim_{j \rightarrow \infty} C_{j,\epsilon} \subseteq C_\epsilon$.

When $t \in [t_0, T)$ and $\mathbf{q}(t)$ is differential at t , we can take the limit $\epsilon \rightarrow 0$, and by the upper-semicontinuity of $\partial^\circ \mathcal{L}$ again, we have

$$\mathbf{q}'(t) \in \lim_{\epsilon \rightarrow 0} C_\epsilon \subseteq \overline{\text{conv}}(-\partial^\circ \mathcal{L}(\mathbf{q}(T_{12} + t))) = -\partial^\circ \mathcal{L}(\mathbf{q}(T_{12} + t))$$

as $\partial^\circ \mathcal{L}(\mathbf{q}(T_{12} + t))$ is closed convex for any t . Therefore $\mathbf{q}(t)$ is indeed a gradient flow trajectory. $\square$

Proof for Lemma 5.4. We can prove Lemma 5.4 by combining Lemmas E.7 to E.9 together. For $-\infty < t \leq t_0$, by Lemma E.8, $\|\tilde{\theta}(t) - e^{\lambda_0 t} \hat{\theta}\|_{\mathbf{M}} \leq \frac{4m\|\hat{\theta}\|_{\mathbf{M}}^3}{\lambda_0} e^{3\lambda_0 t}$. With $\hat{\theta} := (\bar{b}_+, \bar{b}_+ \bar{\mu}, \bar{b}_-, \bar{b}_- \bar{\mu})$, by choosing a threshold $t_s < t_0$ small enough, we can have for any $t \leq t_s$,

- $\tilde{a}_1(t) \geq e^{\lambda_0 t} \bar{b}_+ - \frac{4m\|\hat{\theta}\|_{\mathbf{M}}^3}{\lambda_0} e^{3\lambda_0 t} > 0;$
- $\tilde{a}_2(t) \leq e^{\lambda_0 t} \bar{b}_- + \frac{4m\|\hat{\theta}\|_{\mathbf{M}}^3}{\lambda_0} e^{3\lambda_0 t} < 0;$
- $\langle \tilde{w}_1(t), \mathbf{w}^* \rangle \geq e^{\lambda_0 t} \bar{b}_+ \langle \bar{\mu}, \mathbf{w}^* \rangle - \frac{4m\|\hat{\theta}\|_{\mathbf{M}}^3}{\lambda_0} e^{3\lambda_0 t} > \frac{8m\|\hat{\theta}\|_{\mathbf{M}}^3}{\lambda_0} e^{3\lambda_0 t} > 0;$
- $\langle \tilde{w}_2(t), \mathbf{w}^* \rangle \leq e^{\lambda_0 t} \bar{b}_- \langle \bar{\mu}, \mathbf{w}^* \rangle + \frac{4m\|\hat{\theta}\|_{\mathbf{M}}^3}{\lambda_0} e^{3\lambda_0 t} < -\frac{8m\|\hat{\theta}\|_{\mathbf{M}}^3}{\lambda_0} e^{3\lambda_0 t} < 0.$

Then $\tilde{\theta}(t) \neq 0$ for all $t \leq t_s$. For $t > t_s$, we know $\tilde{\theta}(t) \neq 0$ by applying Theorem B.19. Finally by Lemmas E.8 and E.9 we know $\lim_{\sigma_{\text{init}} \rightarrow 0} \theta(T_{12} + t) = \pi_{\tilde{\theta}}(\tilde{\theta}(t))$ for all t . $\square$

F Proofs for Phase III

F.1 Two Neuron Case: Margin Maximization

In this subsection we prove Theorem 5.5 for the symmetric datasets. By Theorem B.19 and Theorem 3.1, we know that gradient flow must converge in a KKT-margin direction of width-2 two-layer Leaky ReLU network (Definition B.8). Thus we first give some characterizations for KKT-margin directions by proving Lemma F.1 and Lemma F.2.

Lemma F.1. *Given $\mathbf{u}_1, \mathbf{u}_2 \in \mathbb{R}^d$, if $y_i(\phi(\langle \mathbf{u}_1, \mathbf{x}_i \rangle) - \phi(-\langle \mathbf{u}_2, \mathbf{x}_i \rangle)) \geq 1$ for all $i \in [n]$, then*

$$y_i(h_i^{(1)} \langle \mathbf{u}_2, \mathbf{x}_i \rangle + h_i^{(2)} \langle \mathbf{u}_1, \mathbf{x}_i \rangle) \geq 1,$$

for all Clarke's sub-differentials $h_i^{(1)} \in \phi^\circ(\langle \mathbf{u}_1, \mathbf{x}_i \rangle), h_i^{(2)} \in \phi^\circ(-\langle \mathbf{u}_2, \mathbf{x}_i \rangle)$.

Proof. We prove by cases for any fixed $i \in [n]$. By Assumption 4.1 we have

$$y_i(\phi(\langle \mathbf{u}_1, \mathbf{x}_i \rangle) - \phi(-\langle \mathbf{u}_2, \mathbf{x}_i \rangle)) \geq 1, \quad -y_i(\phi(-\langle \mathbf{u}_1, \mathbf{x}_i \rangle) - \phi(\langle \mathbf{u}_2, \mathbf{x}_i \rangle)) \geq 1.$$

Case 1. Suppose that $\langle \mathbf{u}_1, \mathbf{x}_i \rangle \neq 0$, $\langle \mathbf{u}_2, \mathbf{x}_i \rangle \neq 0$. Then we have $h_i^{(1)}, h_i^{(2)} \in \{\alpha_{\text{leaky}}, 1\}$. If $h_i^{(1)} = h_i^{(2)}$, then $h_i^{(1)} \langle \mathbf{u}_2, \mathbf{x}_i \rangle = -\phi(-\langle \mathbf{u}_2, \mathbf{x}_i \rangle)$, $h_i^{(2)} \langle \mathbf{u}_1, \mathbf{x}_i \rangle = \phi(\langle \mathbf{u}_1, \mathbf{x}_i \rangle)$, and thus we have

$$\begin{aligned} y_i(h_i^{(1)} \langle \mathbf{u}_2, \mathbf{x}_i \rangle + h_i^{(2)} \langle \mathbf{u}_1, \mathbf{x}_i \rangle) &= y_i(-\phi(-\langle \mathbf{u}_2, \mathbf{x}_i \rangle) + \phi(\langle \mathbf{u}_1, \mathbf{x}_i \rangle)) \\ &= y_i(\phi(\langle \mathbf{u}_1, \mathbf{x}_i \rangle) - \phi(-\langle \mathbf{u}_2, \mathbf{x}_i \rangle)) \geq 1. \end{aligned}$$

Otherwise, $h_i^{(1)} \neq h_i^{(2)}$, then we have $h_i^{(1)} \langle \mathbf{u}_2, \mathbf{x}_i \rangle = \phi(\langle \mathbf{u}_2, \mathbf{x}_i \rangle)$, $h_i^{(2)} \langle \mathbf{u}_1, \mathbf{x}_i \rangle = -\phi(-\langle \mathbf{u}_1, \mathbf{x}_i \rangle)$, and thus

$$\begin{aligned} y_i(h_i^{(1)} \langle \mathbf{u}_2, \mathbf{x}_i \rangle + h_i^{(2)} \langle \mathbf{u}_1, \mathbf{x}_i \rangle) &= y_i(\phi(\langle \mathbf{u}_2, \mathbf{x}_i \rangle) - \phi(-\langle \mathbf{u}_1, \mathbf{x}_i \rangle)) \\ &= -y_i(\phi(-\langle \mathbf{u}_1, \mathbf{x}_i \rangle) - \phi(\langle \mathbf{u}_2, \mathbf{x}_i \rangle)) \geq 1. \end{aligned}$$

Case 2. Suppose that $\langle \mathbf{u}_1, \mathbf{x}_i \rangle = 0$ or $\langle \mathbf{u}_2, \mathbf{x}_i \rangle = 0$. WLOG we assume that $\langle \mathbf{u}_1, \mathbf{x}_i \rangle = 0$ (the case of $\langle \mathbf{u}_2, \mathbf{x}_i \rangle = 0$ can be proved similarly). Then we have

$$-y_i \phi(-\langle \mathbf{u}_2, \mathbf{x}_i \rangle) \geq 1, \quad y_i \phi(\langle \mathbf{u}_2, \mathbf{x}_i \rangle) \geq 1.$$

If $\langle \mathbf{u}_2, \mathbf{x}_i \rangle = 0$, then the feasibility cannot be satisfied. So we must have $\langle \mathbf{u}_2, \mathbf{x}_i \rangle \neq 0$ and $h_i^{(2)} \in \{\alpha_{\text{leaky}}, 1\}$. This implies that $y_i \langle \mathbf{u}_2, \mathbf{x}_i \rangle \geq \frac{1}{\alpha_{\text{leaky}}}$.

Since $\langle \mathbf{u}_1, \mathbf{x}_i \rangle = 0$, we have $h_i^{(1)} \in [\alpha_{\text{leaky}}, 1]$. Therefore,

$$y_i(h_i^{(1)} \langle \mathbf{u}_2, \mathbf{x}_i \rangle + h_i^{(2)} \langle \mathbf{u}_1, \mathbf{x}_i \rangle) = y_i h_i^{(1)} \langle \mathbf{u}_2, \mathbf{x}_i \rangle \geq y_i \alpha_{\text{leaky}} \langle \mathbf{u}_2, \mathbf{x}_i \rangle \geq 1,$$

which completes the proof. $\square$

Lemma F.2. If $(\mathbf{w}_1, \mathbf{w}_2, a_1, a_2)$ is along a KKT-margin direction of width-2 two-layer Leaky ReLU network and $a_1 > 0, a_2 < 0$, then $\mathbf{w}_1 = -\mathbf{w}_2$, $a_1 = -a_2 = \|\mathbf{w}_1\|_2$.

Proof. WLOG we assume that $q_{\min}(\boldsymbol{\theta}) = 1$. By Definition B.8 and Lemma B.9, there exist $\lambda_1, \dots, \lambda_n \geq 0$ and $h_1^{(1)}, \dots, h_n^{(1)} \in \mathbb{R}, h_1^{(2)}, \dots, h_n^{(2)} \in \mathbb{R}$ such that $h_i^{(1)} \in \phi^\circ(\langle \mathbf{w}_1, \mathbf{x}_i \rangle)$, $h_i^{(2)} \in \phi^\circ(\langle \mathbf{w}_2, \mathbf{x}_i \rangle)$, and the following conditions hold:

1. $\mathbf{w}_1 = a_1 \sum_{i \in [n]} \lambda_i y_i h_i^{(1)} \mathbf{x}_i$, $\mathbf{w}_2 = a_2 \sum_{i \in [n]} \lambda_i y_i h_i^{(2)} \mathbf{x}_i$;
2. $a_1 = \|\mathbf{w}_1\|_2$, $a_2 = -\|\mathbf{w}_2\|_2$;
3. For all $i \in [n]$, if $q_i(\boldsymbol{\theta}) \neq 1$ then $\lambda_i = 0$.

Let $\mathbf{u}_1 = a_1 \mathbf{w}_1$ and $\mathbf{u}_2 = -a_2 \mathbf{w}_2$. Let $\bar{\mathbf{u}}_1 := \frac{\mathbf{u}_1}{\|\mathbf{u}_1\|_2}$, $\bar{\mathbf{u}}_2 := -\frac{\mathbf{u}_2}{\|\mathbf{u}_2\|_2}$. Then the following conditions hold for all $i \in [n]$:

$$\bar{\mathbf{u}}_1 - \sum_{i=1}^n \lambda_i h_i^{(1)} y_i \mathbf{x}_i = 0, \tag{25}$$

$$\bar{\mathbf{u}}_2 - \sum_{i=1}^n \lambda_i h_i^{(2)} y_i \mathbf{x}_i = 0, \tag{26}$$

$$\lambda_i (1 - y_i (\phi(\langle \mathbf{u}_1, \mathbf{x}_i \rangle) - \phi(-\langle \mathbf{u}_2, \mathbf{x}_i \rangle))) = 0. \tag{27}$$

By homogeneity, $h_i^{(1)} \cdot \langle \mathbf{u}_1, \mathbf{x}_i \rangle = \phi(\langle \mathbf{u}_1, \mathbf{x}_i \rangle)$, $h_i^{(2)} \cdot \langle \mathbf{u}_2, \mathbf{x}_i \rangle = -\phi(-\langle \mathbf{u}_2, \mathbf{x}_i \rangle)$. Left-multiplying $(\mathbf{u}_1)^\top$ or $(\mathbf{u}_2)^\top$ on both sides of (25), we have

$$\|\mathbf{u}_1\|_2 - \sum_{i=1}^n \lambda_i y_i \phi(\langle \mathbf{u}_1, \mathbf{x}_i \rangle) = 0, \tag{28}$$

$$\langle \bar{\mathbf{u}}_1, \bar{\mathbf{u}}_2 \rangle \|\mathbf{u}_2\|_2 - \sum_{i=1}^n \lambda_i h_i^{(1)} y_i \langle \mathbf{u}_2, \mathbf{x}_i \rangle = 0. \tag{29}$$

Similarly, we have

$$\|\mathbf{u}_2\|_2 + \sum_{i=1}^n \lambda_i y_i \phi(-\langle \mathbf{u}_2, \mathbf{x}_i \rangle) = 0, \quad (30)$$

$$\langle \bar{\mathbf{u}}_1, \bar{\mathbf{u}}_2 \rangle \|\mathbf{u}_1\|_2 - \sum_{i=1}^n \lambda_i h_i^{(2)} y_i \langle \mathbf{u}_1, \mathbf{x}_i \rangle = 0. \quad (31)$$

Combining (28) and (30), we have

$$\|\mathbf{u}_1\|_2 + \|\mathbf{u}_2\|_2 = \sum_{i=1}^n \lambda_i y_i \phi(\langle \mathbf{u}_1, \mathbf{x}_i \rangle) - \sum_{i=1}^n \lambda_i y_i \phi(-\langle \mathbf{u}_2, \mathbf{x}_i \rangle) \quad (32)$$

$$= \sum_{i=1}^n \lambda_i y_i (\phi(\langle \mathbf{u}_1, \mathbf{x}_i \rangle) - \phi(-\langle \mathbf{u}_2, \mathbf{x}_i \rangle)) \quad (33)$$

$$= \sum_{i=1}^n \lambda_i, \quad (34)$$

where the last equality is due to (27).

Combining (29) and (31), we have

$$\langle \bar{\mathbf{u}}_1, \bar{\mathbf{u}}_2 \rangle (\|\mathbf{u}_1\|_2 + \|\mathbf{u}_2\|_2) = \sum_{i=1}^n \lambda_i h_i^{(1)} y_i \langle \mathbf{u}_2, \mathbf{x}_i \rangle + \sum_{i=1}^n \lambda_i h_i^{(2)} y_i \langle \mathbf{u}_1, \mathbf{x}_i \rangle \quad (35)$$

$$= \sum_{i=1}^n \lambda_i y_i (h_i^{(1)} \langle \mathbf{u}_2, \mathbf{x}_i \rangle + h_i^{(2)} \langle \mathbf{u}_1, \mathbf{x}_i \rangle) \quad (36)$$

$$\geq \sum_{i=1}^n \lambda_i, \quad (37)$$

where the last inequality is due to Lemma F.1. Since we have deduced that $\|\mathbf{u}_1\|_2 + \|\mathbf{u}_2\|_2 = \sum_{i=1}^n \lambda_i$, we further have

$$\langle \bar{\mathbf{u}}_1, \bar{\mathbf{u}}_2 \rangle (\|\mathbf{u}_1\|_2 + \|\mathbf{u}_2\|_2) \geq \|\mathbf{u}_1\|_2 + \|\mathbf{u}_2\|_2.$$

Combining this with $\langle \bar{\mathbf{u}}_1, \bar{\mathbf{u}}_2 \rangle \leq \|\bar{\mathbf{u}}_1\|_2 \|\bar{\mathbf{u}}_2\|_2 \leq 1$, we have $1 \leq \langle \bar{\mathbf{u}}_1, \bar{\mathbf{u}}_2 \rangle \leq 1$. So all the inequalities become equalities, and thus $\bar{\mathbf{u}}_1 = \bar{\mathbf{u}}_2$. (36) also equals to (37), so

$$y_i (h_i^{(1)} \langle \mathbf{u}_2, \mathbf{x}_i \rangle + h_i^{(2)} \langle \mathbf{u}_1, \mathbf{x}_i \rangle) = 1, \quad (38)$$

whenever $\lambda_i \neq 0$.

By (27), we have $y_i (h_i^{(1)} \langle \mathbf{u}_1, \mathbf{x}_i \rangle + h_i^{(2)} \langle \mathbf{u}_2, \mathbf{x}_i \rangle) = 1$ whenever $\lambda_i \neq 0$. Combining this with (38), we have

$$y_i (h_i^{(1)} - h_i^{(2)}) \langle \mathbf{u}_1, \mathbf{x}_i \rangle = y_i (h_i^{(1)} - h_i^{(2)}) \langle \mathbf{u}_2, \mathbf{x}_i \rangle.$$

Then we prove that $\langle \mathbf{u}_1, \mathbf{x}_i \rangle = \langle \mathbf{u}_2, \mathbf{x}_i \rangle$ by discussing two cases:

1. If $\langle \mathbf{u}_1, \mathbf{x}_i \rangle = 0$, then $\langle \mathbf{u}_2, \mathbf{x}_i \rangle = 0$ since $\bar{\mathbf{u}}_1 = \bar{\mathbf{u}}_2$;
2. Otherwise, we have $(h_i^{(1)}, h_i^{(2)}) = (1, \alpha_{\text{leaky}})$ or $(\alpha_{\text{leaky}}, 1)$ by symmetry, so $h_i^{(1)} \neq h_i^{(2)}$ and thus $\langle \mathbf{u}_1, \mathbf{x}_i \rangle = \langle \mathbf{u}_2, \mathbf{x}_i \rangle$.

This means $\mathbf{u}_1$ and $\mathbf{u}_2$ have the same projection onto the linear space spanned by $\{\mathbf{x}_i : \lambda_i \neq 0\}$. By (25) and (26), $\mathbf{u}_1$ and $\mathbf{u}_2$ are in the span of $\{\mathbf{x}_i : i \in [n], \lambda_i \neq 0\}$. Therefore, $\mathbf{u}_1 = \mathbf{u}_2$ and we can easily deduce that $\mathbf{w}_1 = -\mathbf{w}_2$, $a_1 = -a_2 = \|\mathbf{w}_1\|_2$. $\square$

Lemma F.3. *If $\boldsymbol{\theta} = (\mathbf{w}_1, \mathbf{w}_2, a_1, a_2)$ is along a KKT-margin direction of width-2 two-layer Leaky ReLU network and $\|\boldsymbol{\theta}\|_2 = 1$, $a_1 \geq 0$ and $a_2 \leq 0$, then one of the following three cases is true:*

1. $\boldsymbol{\theta} = \frac{1}{2}(\mathbf{w}^*, -\mathbf{w}^*, 1, -1)$;

2. $\boldsymbol{\theta} = \frac{1}{\sqrt{2}}(\mathbf{w}^*, \mathbf{0}, 1, 0)$;
3. $\boldsymbol{\theta} = \frac{1}{\sqrt{2}}(\mathbf{0}, -\mathbf{w}^*, 0, -1)$.

Proof. Suppose $a_1 > 0$ and $a_2 < 0$, then by Lemma F.2, we know $\mathbf{w}_1 = -\mathbf{w}_2$, $a_1 = -a_2 = \|\mathbf{w}_1\|_2$. Since $q_i(\boldsymbol{\theta}) > 0, \forall i$, we know $\langle \mathbf{w}_1, \mathbf{x}_i \rangle \neq 0, \forall i$, which implies $q_i(\boldsymbol{\theta})$ is differentiable at $\boldsymbol{\theta}$. Let $\boldsymbol{\theta}' = (\mathbf{w}_1, a_1)$ and $[\boldsymbol{\theta}'; -\boldsymbol{\theta}'] = (\mathbf{w}_1, -\mathbf{w}_1, a_1, -a_1)$, we know $\boldsymbol{\theta}'$ is along the KKT direction of the following optimization problem:

$$\begin{aligned} \min \quad & f([\boldsymbol{\theta}'; -\boldsymbol{\theta}']) \\ \text{s.t.} \quad & g_i([\boldsymbol{\theta}'; -\boldsymbol{\theta}']) \leq 0, \quad \forall i \in [n], \end{aligned}$$

where $f([\boldsymbol{\theta}'; -\boldsymbol{\theta}']) = \|\boldsymbol{\theta}'; -\boldsymbol{\theta}'\|_2^2 = 2\|\mathbf{w}_1\|_2^2 + 2a_1^2$, and $g_i([\boldsymbol{\theta}'; -\boldsymbol{\theta}']) = 1 - q_i([\boldsymbol{\theta}'; -\boldsymbol{\theta}']) = y_i a_i (\phi(\langle \mathbf{w}_1, \mathbf{x}_i \rangle) - \phi(-\langle \mathbf{w}_1, \mathbf{x}_i \rangle)) = a_1(1 + \alpha_{\text{leaky}}) \langle \mathbf{w}_1, y_i \mathbf{x}_i \rangle$. With a standard analysis, we know $\mathbf{w}_1$ be in the direction of the max-margin classifier of the original problem, $\mathbf{w}^*$.

Next we discuss the case where $a_2 = 0$ ($a_1 = 0$ follows the same analysis). When $a_2 = 0$, since $q_i(\boldsymbol{\theta}) > 0$ for all i , we know $a_1 y_i \langle \mathbf{x}_i, \mathbf{x}_1 \rangle > 0$ for all i . Thus $q_i(\boldsymbol{\theta}) > q_{i+\frac{n}{2}}(\boldsymbol{\theta}) = \alpha_{\text{leaky}} q_i(\boldsymbol{\theta})$, which means only the second half constraints might be active. This reduces the optimization problem to a standard linear-max-margin problem, and $\mathbf{w}_1$ will be aligned with $\mathbf{w}^*$. $\square$

Proof for Theorem 5.5. By Theorem B.19 and Theorem 3.1, we know $\lim_{t \rightarrow +\infty} \frac{\boldsymbol{\theta}(t)}{\|\boldsymbol{\theta}(t)\|_2}$ must be along a KKT-margin direction. By Lemma F.3, we know that there are only 3 KKT-margin directions:

$$\frac{1}{2}(\mathbf{w}^*, -\mathbf{w}^*, 1, -1), \quad \frac{1}{\sqrt{2}}(\mathbf{w}^*, \mathbf{0}, 1, 0), \quad \frac{1}{\sqrt{2}}(\mathbf{0}, -\mathbf{w}^*, 0, -1).$$

Thus it suffices to show $\lim_{t \rightarrow +\infty} \frac{\boldsymbol{\theta}(t)}{\|\boldsymbol{\theta}(t)\|_2} \neq \frac{1}{\sqrt{2}}(\mathbf{w}^*, \mathbf{0}, 1, 0)$. ($\lim_{t \rightarrow +\infty} \frac{\boldsymbol{\theta}(t)}{\|\boldsymbol{\theta}(t)\|_2} \neq \frac{1}{\sqrt{2}}(\mathbf{w}^*, \mathbf{0}, 1, 0)$ would hold for the same reason.)

For convenience, we define $i' := i + n/2$ if $1 \leq i \leq n/2$ and $i' := i - n/2$ if $n/2 < i \leq n$. By Assumption 4.1 we know that $\mathbf{x}_{i'} = -\mathbf{x}_i$ and $y_{i'} = -y_i$.

We first define the angle between $\mathbf{w}^*$ and $\mathbf{w}_1(t)$ as $\beta_1(t) := \arccos \frac{\langle \mathbf{w}^*, \mathbf{w}_1(t) \rangle}{\|\mathbf{w}_1(t)\|_2}$ and angle between $-\mathbf{w}^*$ and $\mathbf{w}_2(t)$ as $\beta_2(t) := \arccos \frac{\langle -\mathbf{w}^*, \mathbf{w}_2(t) \rangle}{\|\mathbf{w}_2(t)\|_2}$. Since $\langle \mathbf{w}^*, \mathbf{w}_1(0) \rangle > 0$ and $\langle -\mathbf{w}^*, \mathbf{w}_2(0) \rangle > 0$, by Lemma B.20 we know that $\beta_1(t), \beta_2(t) \in [0, \pi/2)$ for all $t \geq 0$.

We also define $\epsilon := \min_{i \in [n]} \left\{ \arcsin \frac{\langle y_i \mathbf{x}_i, \mathbf{w}^* \rangle}{\|\mathbf{x}_i\|_2} \right\}$, which can be understood as the angle between $\mathbf{x}_i$ and the decision boundary determined by the linear separator $\mathbf{w}^*$.

Below we will prove by contradiction. Suppose $\lim_{t \rightarrow +\infty} \frac{\boldsymbol{\theta}(t)}{\|\boldsymbol{\theta}(t)\|_2} = \frac{1}{\sqrt{2}}(\mathbf{w}^*, \mathbf{0}, 1, 0) =: \bar{\boldsymbol{\theta}}_\infty$ holds. Then $\beta_1(t) \rightarrow 0$ and $\frac{\|\mathbf{w}_2(t)\|_2}{\|\mathbf{w}_1(t)\|_2} \rightarrow 0$ as $t \rightarrow +\infty$. Thus there must exist $T_1 > 0$ such that $\beta_1(t) \leq \epsilon/2$.

Note that $f_{\bar{\boldsymbol{\theta}}_\infty}(\mathbf{x}_i) = \frac{1}{2}\phi(\langle \mathbf{x}_i, \mathbf{w}^* \rangle)$ for all $i \in [n]$. By symmetry, for $i \in [n/2]$ we have

$$q_i(\bar{\boldsymbol{\theta}}_\infty) - q_{i'}(\bar{\boldsymbol{\theta}}_\infty) = f_{\bar{\boldsymbol{\theta}}_\infty}(\mathbf{x}_i) - f_{\bar{\boldsymbol{\theta}}_\infty}(-\mathbf{x}_i) = \frac{1}{2}\langle \mathbf{x}_i, \mathbf{w}^* \rangle - \frac{\gamma^*}{2}\langle \mathbf{x}_i, \mathbf{w}^* \rangle = \frac{1-\gamma^*}{2}\langle \mathbf{x}_i, \mathbf{w}^* \rangle > 0.$$

By Theorem B.19, we also know $\|\boldsymbol{\theta}(t)\|_2 \rightarrow \infty$, so $q_i(\boldsymbol{\theta}) - q_{i'}(\boldsymbol{\theta}) = \|\boldsymbol{\theta}(t)\|_2(q_i(\bar{\boldsymbol{\theta}}_\infty) - q_{i'}(\bar{\boldsymbol{\theta}}_\infty)) \rightarrow +\infty$ for all $i \in [n/2]$. Let $g_i(\boldsymbol{\theta}) := -\ell'(q_i(\boldsymbol{\theta}))$. Then

$$\frac{g_i(\boldsymbol{\theta}(t))}{g_{i'}(\boldsymbol{\theta}(t))} \sim \frac{\exp(-g_i(\boldsymbol{\theta}(t)))}{\exp(-g_{i'}(\boldsymbol{\theta}(t)))} = e^{-(g_i(\boldsymbol{\theta}(t)) - g_{i'}(\boldsymbol{\theta}(t)))} \rightarrow 0.$$

Thus there must exist $T_2 > 0$ such that $\frac{g_i(\boldsymbol{\theta}(t))}{g_{i'}(\boldsymbol{\theta}(t))} \leq \max \left\{ \frac{\cos \epsilon - \alpha_{\text{leaky}}}{1 - \alpha_{\text{leaky}} \cos \epsilon}, 1 \right\}$ for all $i \in [n/2]$ and $t \geq T_2$.

We will use these to show that $\frac{\langle \mathbf{w}_2(t), -\mathbf{w}^* \rangle}{\langle \mathbf{w}_1(t), \mathbf{w}^* \rangle}$ is non-decreasing for $t \geq T := \max\{T_1, T_2\}$, which further implies $\frac{\|\mathbf{w}_2(t)\|_2}{\|\mathbf{w}_1(t)\|_2}$ is lower bounded by some constant. Thus it contradicts with the assumption of convergence.

By Corollary B.18, we know that $a_1(t) = \|\mathbf{w}_1(t)\|_2$ and $a_2(t) = -\|\mathbf{w}_2(t)\|_2$ for all $t \geq 0$. Then for all $i \in [n]$, we have

$$f_{\theta}(\mathbf{x}_i) = a_1\phi(\mathbf{w}_1^\top \mathbf{x}_i) + a_2\phi(\mathbf{w}_2^\top \mathbf{x}_i) = \|\mathbf{w}_1\|_2\phi(\mathbf{w}_1^\top \mathbf{x}_i) - \|\mathbf{w}_2\|_2\phi(\mathbf{w}_2^\top \mathbf{x}_i).$$

By (10), if $\mathbf{w}_1(t), \mathbf{w}_2(t) \in \Omega_{\mathcal{S}}$ then we have

$$\frac{d\mathbf{w}_1}{dt} = \frac{\|\mathbf{w}_1\|_2}{n} \sum_{i \in [n]} g_i(\theta)\phi'(\mathbf{w}_1^\top \mathbf{x}_i)y_i\mathbf{x}_i, \quad -\frac{d\mathbf{w}_2}{dt} = \frac{\|\mathbf{w}_2\|_2}{n} \sum_{i \in [n]} g_i(\theta)\phi'(\mathbf{w}_2^\top \mathbf{x}_i)y_i\mathbf{x}_i.$$

By symmetry, we can rewrite them as

$$\frac{d\mathbf{w}_1(t)}{dt} = \frac{\|\mathbf{w}_1(t)\|_2}{n} \sum_{i \in [n/2]} \sigma_i^{(1)}(t)\mathbf{x}_i, \quad -\frac{d\mathbf{w}_2(t)}{dt} = \frac{\|\mathbf{w}_2(t)\|_2}{n} \sum_{i \in [n/2]} \sigma_i^{(2)}(t)\mathbf{x}_i. \quad (39)$$

where $\sigma_i^{(k)}(t) := g_i(\theta(t))\phi'(\mathbf{w}_k^\top(t)\mathbf{x}_i) + g_{i'}(\theta(t))\phi'(-\mathbf{w}_k^\top(t)\mathbf{x}_i)$. Note that this only holds for $\mathbf{w}_k(t) \in \Omega_{\mathcal{S}}$. By taking limits through (8), we know that for a.e. $t \geq 0$, there exists $\sigma_i^{(k)}(t)$ such that (39) holds and

$$\sigma_i^{(k)}(t) \in \begin{cases} \{g_i(\theta) + \alpha_{\text{leaky}}g_{i'}(\theta)\} & \text{if } \mathbf{w}_k^\top \mathbf{x}_i > 0; \\ \{\alpha_{\text{leaky}}g_i(\theta) + g_{i'}(\theta)\} & \text{if } \mathbf{w}_k^\top \mathbf{x}_i < 0; \\ \{\lambda g_i(\theta) + (1 + \alpha_{\text{leaky}} - \lambda)g_{i'}(\theta) : \alpha_{\text{leaky}} \leq \lambda \leq 1\} & \text{if } \mathbf{w}_k^\top \mathbf{x}_i = 0. \end{cases} \quad (40)$$

By chain rule, for a.e. $t \geq 0$ we have:

$$\begin{aligned} \frac{d}{dt} \ln \frac{\langle \mathbf{w}_2, -\mathbf{w}^* \rangle}{\langle \mathbf{w}_1, \mathbf{w}^* \rangle} &= \frac{\langle \frac{d\mathbf{w}_2}{dt}, -\mathbf{w}^* \rangle}{\langle \mathbf{w}_2, -\mathbf{w}^* \rangle} - \frac{\langle \frac{d\mathbf{w}_1}{dt}, \mathbf{w}^* \rangle}{\langle \mathbf{w}_1, \mathbf{w}^* \rangle} \\ &= \frac{\|\mathbf{w}_2\|_2}{\langle \mathbf{w}_2, -\mathbf{w}^* \rangle} \cdot \frac{1}{n} \sum_{i \in [n/2]} \sigma_i^{(2)} \langle \mathbf{x}_i, \mathbf{w}^* \rangle - \frac{\|\mathbf{w}_1\|_2}{\langle \mathbf{w}_1, \mathbf{w}^* \rangle} \cdot \frac{1}{n} \sum_{i \in [n/2]} \sigma_i^{(1)} \langle \mathbf{x}_i, \mathbf{w}^* \rangle \\ &= \frac{1}{n} \sum_{i \in [n/2]} \left(\frac{\sigma_i^{(2)}}{\cos \beta_2} - \frac{\sigma_i^{(1)}}{\cos \beta_1} \right) \langle \mathbf{x}_i, \mathbf{w}^* \rangle. \end{aligned}$$

Now we are ready to prove $\frac{d}{dt} \ln \frac{\langle \mathbf{w}_2, -\mathbf{w}^* \rangle}{\langle \mathbf{w}_1, \mathbf{w}^* \rangle} \geq 0$ for $t \geq T$. For this, we only need to show that

$$\frac{\sigma_i^{(2)}}{\cos \beta_2} \geq \frac{\sigma_i^{(1)}}{\cos \beta_1} \text{ in two cases.}$$

Case 1. When $\beta_1 < \beta_2$, it suffices to show $\sigma_i^{(1)} \leq \sigma_i^{(2)}$. By our choice of T_1 , we have $\beta_1 \leq \epsilon/2$, which implies $\mathbf{w}_1^\top \mathbf{x}_i > 0$ and $\sigma_i^{(1)} = g_i(\theta) + \alpha_{\text{leaky}}g_{i'}(\theta)$ for all $i \in [n/2]$. Note that $g_i(\theta) \leq g_{i'}(\theta)$ according to our choice of T_2 . Then for any $\lambda \in [\alpha_{\text{leaky}}, 1]$ we have

$$\sigma_i^{(1)} = g_i(\theta) + \alpha_{\text{leaky}}g_{i'}(\theta) \leq \lambda g_i(\theta) + (1 + \alpha_{\text{leaky}} - \lambda)g_{i'}(\theta).$$

By (40), we therefore have $\sigma_i^{(1)} \leq \sigma_i^{(2)}$.

Case 2. If $\beta_1 \geq \beta_2$, then by our choice of T_1 we have $\epsilon/2 \geq \beta_1 \geq \beta_2$. Then for all $i \in [n/2]$, $\mathbf{w}_2^\top \mathbf{x}_i \leq 0$. So we have

$$\frac{\sigma_i^{(1)}}{\sigma_i^{(2)}} = \frac{g_i(\theta) + \alpha_{\text{leaky}}g_{i'}(\theta)}{\alpha_{\text{leaky}}g_i(\theta) + g_{i'}(\theta)} = \frac{\frac{g_i(\theta)}{g_{i'}(\theta)} + \alpha_{\text{leaky}}}{\alpha_{\text{leaky}}\frac{g_i(\theta)}{g_{i'}(\theta)} + 1} \leq \cos \epsilon \leq \cos \beta_1(t) \leq \frac{\cos \beta_1(t)}{\cos \beta_2(t)}.$$

$$\text{Thus } \frac{\sigma_i^{(2)}}{\cos \beta_2} \geq \frac{\sigma_i^{(1)}}{\cos \beta_1}.$$

Now we have shown that $\frac{\langle \mathbf{w}_2(t), -\mathbf{w}^* \rangle}{\langle \mathbf{w}_1(t), \mathbf{w}^* \rangle} \geq \frac{\langle \mathbf{w}_2(T), -\mathbf{w}^* \rangle}{\langle \mathbf{w}_1(T), \mathbf{w}^* \rangle} =: r_0$, where r_0 is a constant (ratio at time T). So for $t \geq T$,

$$\frac{\|\mathbf{w}_2(t)\|_2}{\|\mathbf{w}_1(t)\|_2} = \frac{\langle \mathbf{w}_2(t), -\mathbf{w}^* \rangle \cos \beta_1(t)}{\langle \mathbf{w}_1(t), \mathbf{w}^* \rangle \cos \beta_2(t)} \geq \frac{\langle \mathbf{w}_2(t), -\mathbf{w}^* \rangle \cos \epsilon}{\langle \mathbf{w}_1(t), \mathbf{w}^* \rangle} \geq r_0 \cos \epsilon, \quad (41)$$

is lower bounded, which contradicts with $\lim_{t \rightarrow +\infty} \frac{\theta(t)}{\|\theta(t)\|_2} = \frac{1}{\sqrt{2}}(\mathbf{w}^*, \mathbf{0}, 1, 0)$. $\square$

F.2 Directional Convergence of L -homogeneous Neural Nets

In this section we consider general L -homogeneous neural nets with logistic loss following the settings introduced in Section 3.1. We define $\alpha(\boldsymbol{\theta})$ and $\tilde{\gamma}(\boldsymbol{\theta})$ to be smoothed margin and its normalized version following [Lyu and Li \(2020\)](#).

$$\alpha(\boldsymbol{\theta}) = \ell^{-1}(n\mathcal{L}(\boldsymbol{\theta})), \quad \tilde{\gamma}(\boldsymbol{\theta}) = \frac{\alpha(\boldsymbol{\theta})}{\|\boldsymbol{\theta}\|_2^L}.$$

Define $\zeta(t) := \int_0^t \left\| \frac{d}{d\tau} \frac{\boldsymbol{\theta}(\tau)}{\|\boldsymbol{\theta}(\tau)\|_2} \right\|_2 d\tau$ to be the length of the trajectory swept by $\boldsymbol{\theta}/\|\boldsymbol{\theta}\|_2$ from time 0 to t . Define $\beta(t)$ to be the cosine of the angle between $\boldsymbol{\theta}(t)$ and $\frac{d\boldsymbol{\theta}(t)}{dt}$.

$$\beta(t) := \frac{\left\langle \frac{d\boldsymbol{\theta}(t)}{dt}, \boldsymbol{\theta}(t) \right\rangle}{\left\| \frac{d\boldsymbol{\theta}(t)}{dt} \right\|_2 \cdot \|\boldsymbol{\theta}(t)\|_2}, \quad \text{for a.e. } t \geq 0.$$

F.2.1 Lemmas from Previous Works

We leverage the following two lemmas from [Ji and Telgarsky \(2020a\)](#) on desingularizing function. Formally, we say that $\Psi : [0, \nu)$ is a desingularizing function if Ψ is continuous on $[0, \nu)$ with $\Psi(0) = 0$ and continuously differentiable on $(0, \nu)$ with $\Psi' > 0$.

Lemma F.4 (Lemma 3.6, [Ji and Telgarsky 2020a](#)). *Given a locally Lipschitz definable function f with an open domain $D \subseteq \{\boldsymbol{\theta} : \|\boldsymbol{\theta}\|_2 > 1\}$, for any $c, \eta > 0$, there exists $\nu > 0$ and a definable desingularizing function Ψ on $[0, \nu)$ such that*

$$\Psi'(f(\boldsymbol{\theta})) \cdot \|\boldsymbol{\theta}\|_2 \|\bar{\partial}^\circ f(\boldsymbol{\theta})\|_2 \geq 1,$$

whenever $f(\boldsymbol{\theta}) \in (0, \nu)$ and $\|\bar{\partial}_\perp f(\boldsymbol{\theta})\|_2 \geq c\|\boldsymbol{\theta}\|_2^\eta \|\bar{\partial}_r f(\boldsymbol{\theta})\|_2$.

Lemma F.5 (Corollary of Lemma 3.7, [Ji and Telgarsky 2020a](#)). *Given a locally Lipschitz definable function f with an open domain $D \subseteq \{\boldsymbol{\theta} : \|\boldsymbol{\theta}\|_2 > 1\}$, for any $\lambda > 0$, there exists $\nu > 0$ and a definable desingularizing function Ψ on $[0, \nu)$ such that*

$$\Psi'(f(\boldsymbol{\theta})) \cdot \|\boldsymbol{\theta}\|_2^{1+\lambda} \|\bar{\partial}^\circ f(\boldsymbol{\theta})\|_2 \geq 1,$$

whenever $f(\boldsymbol{\theta}) \in (0, \nu)$.

For $\tilde{\gamma}(\boldsymbol{\theta})$, we have the following decomposition lemma from [Ji and Telgarsky \(2020a\)](#).

Lemma F.6 (Lemma 3.4, [Ji and Telgarsky 2020a](#)). *If $\mathcal{L}(\boldsymbol{\theta}(t)) < \ell(0)/n$ at time $t = t_0$, it holds for a.e. $t \geq t_0$ that*

$$\frac{d\tilde{\gamma}(\boldsymbol{\theta}(t))}{dt} = \|\bar{\partial}_r^\circ \tilde{\gamma}(\boldsymbol{\theta}(t))\|_2 \|\bar{\partial}_r^\circ \mathcal{L}(\boldsymbol{\theta}(t))\|_2 + \|\bar{\partial}_\perp^\circ \tilde{\gamma}(\boldsymbol{\theta}(t))\|_2 \|\bar{\partial}_\perp^\circ \mathcal{L}(\boldsymbol{\theta}(t))\|_2.$$

For $a \in \mathbb{R} \cup \{+\infty, -\infty\}$, we say that v is an asymptotic Clarke critical value of a locally Lipschitz function $f : \mathbb{R}^D \rightarrow \mathbb{R}$ if there exists a sequence of $(\boldsymbol{\theta}_j, \mathbf{g}_j)$, where $\boldsymbol{\theta}_j \in \mathbb{R}^D$ and $\mathbf{g}_j \in \partial^\circ f(\boldsymbol{\theta}_j)$, such that $\lim_{j \rightarrow +\infty} f(\boldsymbol{\theta}_j) = v$ and $\lim_{j \rightarrow +\infty} (1 + \|\boldsymbol{\theta}_j\|_2) \|\mathbf{g}_j\|_2 = 0$.

Lemma F.7 (Corollary of Lemma B.10, [Ji and Telgarsky 2020a](#)). *$\tilde{\gamma}(\boldsymbol{\theta})$ only has finitely many asymptotic Clarke critical values.*

For $\beta(\boldsymbol{\theta})$, we have the following lemma from [Lyu and Li \(2020\)](#).

Lemma F.8 (Lemma C.12, [Lyu and Li 2020](#)). *If $\mathcal{L}(\boldsymbol{\theta}(t)) < \ell(0)/n$ at time $t = t_0$, then there exists a sequence $t_1, t_2, \dots$ such that $t_j \rightarrow +\infty$ and $\beta(t_j) \rightarrow 1$ as $j \rightarrow +\infty$.*

F.2.2 Characterizing Margin Maximization with Asymptotic Clarke Critical Value

Before proving Theorem 5.6, we first prove the following theorem that characterizes margin maximization using asymptotic Clarke critical value.

Theorem F.9. *For homogeneous nets, if $\mathcal{L}(\boldsymbol{\theta}(0)) < \ell(0)/n$, then $\frac{\boldsymbol{\theta}(t)}{\|\boldsymbol{\theta}(t)\|_2}$ converges to some direction $\bar{\boldsymbol{\theta}}$ and $\gamma(\bar{\boldsymbol{\theta}})$ is an asymptotic Clarke critical value of $\tilde{\gamma}$.*

Proof. Note that Theorem 3.1 already implies that $\frac{\theta(t)}{\|\theta(t)\|_2}$ converges to some direction $\bar{\theta}$. We only need to show that $\gamma(\bar{\theta})$ is an asymptotic Clarke critical value of $\tilde{\gamma}$.

By Lemma F.8 and definition of β , there exists a sequence of $(\theta_j, \mathbf{h}_j)$, where $\mathbf{h}_j \in -\partial^\circ \mathcal{L}(\theta_j)$, such that $\tilde{\gamma}(\theta_j) \rightarrow \gamma(\bar{\theta})$, $\|\theta_j\|_2 \rightarrow +\infty$, $\frac{\theta_j}{\|\theta_j\|_2} \rightarrow \bar{\theta}$, $\frac{\langle \mathbf{h}_j, \theta_j \rangle}{\|\mathbf{h}_j\|_2 \|\theta_j\|_2} \rightarrow 1$ as $j \rightarrow +\infty$. By chain rule, we know that

$$\begin{aligned} \partial^\circ \tilde{\gamma}(\theta) &= \frac{\partial^\circ \alpha(\theta)}{\|\theta\|_2^L} + \frac{L\alpha(\theta)}{\|\theta\|_2^{L+2}} \theta = \frac{n(\ell^{-1})'(n\mathcal{L}(\theta))\partial^\circ \mathcal{L}(\theta)}{\|\theta\|_2^L} + \frac{L\alpha(\theta)}{\|\theta\|_2^{L+2}} \theta \\ &= \frac{n\partial^\circ \mathcal{L}(\theta)}{\|\theta\|_2^L \ell'(\alpha(\theta))} + \frac{L\alpha(\theta)}{\|\theta\|_2^{L+2}} \theta. \end{aligned}$$

This means $\mathbf{g}_j := \frac{L\alpha(\theta_j)}{\|\theta_j\|_2^{L+2}} \theta_j - \frac{n\mathbf{h}_j}{\|\theta_j\|_2^L \ell'(\alpha(\theta_j))} \in \partial^\circ \tilde{\gamma}(\theta)$. By definition of asymptotic Clarke critical value, it suffices to show that $\|\theta_j\|_2 \cdot \|\mathbf{g}_j\|_2 \rightarrow 0$ as $j \rightarrow +\infty$.

By Lemma C.5 in (Ji and Telgarsky, 2020a), $\left| \frac{n\langle -\mathbf{h}_j, \theta_j \rangle}{L \cdot \ell'(\alpha(\theta_j))} - \alpha(\theta) \right| \leq 2 \ln n + 1$. So

$$\lim_{j \rightarrow +\infty} \left| \frac{n\langle -\mathbf{h}_j, \theta_j \rangle}{L \cdot \|\theta_j\|_2^L \ell'(\alpha(\theta_j))} - \tilde{\gamma}(\theta_j) \right| = \lim_{j \rightarrow +\infty} \frac{1}{\|\theta_j\|_2^L} \left| \frac{n\langle -\mathbf{h}_j, \theta_j \rangle}{L \cdot \ell'(\alpha(\theta_j))} - \alpha(\theta_j) \right| = 0,$$

which implies that $\lim_{j \rightarrow +\infty} \frac{n\langle -\mathbf{h}_j, \theta_j \rangle}{L \cdot \|\theta_j\|_2^L \ell'(\alpha(\theta_j))} = \lim_{j \rightarrow +\infty} \tilde{\gamma}(\theta_j) = \gamma(\bar{\theta})$. Now for the radial component of $\mathbf{g}_j$ we have

$$\|\theta_j\|_2 \cdot \left\langle \mathbf{g}_j, \frac{\theta_j}{\|\theta_j\|_2} \right\rangle = \frac{n\langle \mathbf{h}_j, \theta_j \rangle}{\|\theta_j\|_2^L \ell'(\alpha(\theta_j))} + \frac{L\alpha(\theta_j)}{\|\theta_j\|_2^L} \rightarrow -L\gamma(\bar{\theta}) + L\gamma(\bar{\theta}) = 0.$$

And for the tangential component we have

$$\begin{aligned} \|\theta_j\|_2 \cdot \left\| \left(\mathbf{I} - \frac{\theta_j \theta_j^\top}{\|\theta_j\|_2^2} \right) \mathbf{g}_j \right\|_2 &= \frac{n}{\|\theta_j\|_2^{L-1} \ell'(\alpha(\theta_j))} \left\| \left(\mathbf{I} - \frac{\theta_j \theta_j^\top}{\|\theta_j\|_2^2} \right) \mathbf{h}_j \right\|_2 \\ &= \frac{n\|\mathbf{h}_j\|_2}{\|\theta_j\|_2^{L-1} \ell'(\alpha(\theta_j))} \sqrt{1 - \frac{\langle \theta_j, \mathbf{h}_j \rangle^2}{\|\theta_j\|_2^2 \|\mathbf{h}_j\|_2^2}} \\ &= \frac{n\langle -\mathbf{h}_j, \theta_j \rangle}{\|\theta_j\|_2^L \ell'(\alpha(\theta_j))} \frac{\|\mathbf{h}_j\|_2 \|\theta_j\|_2}{\langle -\mathbf{h}_j, \theta_j \rangle} \sqrt{1 - \frac{\langle \theta_j, \mathbf{h}_j \rangle^2}{\|\theta_j\|_2^2 \|\mathbf{h}_j\|_2^2}} \\ &\rightarrow L\gamma(\bar{\theta}) \cdot 1 \cdot 0 = 0. \end{aligned}$$

Combining these proves that $\|\theta_j\|_2 \cdot \|\mathbf{g}_j\|_2 \rightarrow 0$. $\square$

F.2.3 Proof for Theorem 5.6

Given Lemmas F.4 and F.5 from Ji and Telgarsky (2020a), we have the following inequality around any direction.

Lemma F.10. *Given any parameter direction $\bar{\theta}^* \in \mathbb{S}^{D-1}$, for any $\kappa \in (L/2, L)$, there exists $\nu > 0$ and a definable desingularizing function Ψ on $[0, \nu)$ such that the following holds.*

1. For any θ , if $\gamma(\bar{\theta}^*) - \tilde{\gamma}(\theta) \in (0, \nu)$ and

$$\|\bar{\partial}^\circ \tilde{\gamma}(\theta)\|_2 \geq \frac{\gamma(\bar{\theta}^*)}{4 \ln n + 2} \|\theta\|_2^{L-\kappa} \|\bar{\partial}^\circ \tilde{\gamma}(\theta)\|_2, \quad (42)$$

then

$$\Psi'(\gamma(\bar{\theta}^*) - \tilde{\gamma}(\theta)) \cdot \|\theta\|_2 \|\bar{\partial}^\circ \tilde{\gamma}(\theta)\|_2 \geq 1. \quad (43)$$

2. For any θ , if $\gamma(\bar{\theta}^*) - \tilde{\gamma}(\theta) \in (0, \nu)$,

$$\Psi'(\gamma(\bar{\theta}^*) - \tilde{\gamma}(\theta)) \cdot \|\theta\|_2^{2\kappa-L+1} \|\bar{\partial}^\circ \tilde{\gamma}(\theta)\|_2 \geq 1. \quad (44)$$

Proof. Applying Lemma F.4 with $f(\boldsymbol{\theta}) = \gamma(\bar{\boldsymbol{\theta}}^*) - \tilde{\gamma}(\boldsymbol{\theta})$, $c = \frac{\gamma(\bar{\boldsymbol{\theta}}^*)}{4 \ln n + 2}$, $\eta = L - \kappa$, we know that there exists $\nu_1 > 0$ and a definable desingularizing function Ψ_1 on $[0, \nu_1)$ such that Item 1 holds for Ψ_1 , i.e.,

$$\Psi_1'(\gamma^* - \tilde{\gamma}(\boldsymbol{\theta})) \cdot \|\boldsymbol{\theta}\|_2 \|\bar{\partial}^\circ \tilde{\gamma}(\boldsymbol{\theta})\|_2 \geq 1,$$

whenever (42) holds.

Applying Lemma F.5 with $f(\boldsymbol{\theta}) = \gamma(\bar{\boldsymbol{\theta}}^*) - \tilde{\gamma}(\boldsymbol{\theta})$, $\lambda = 2\kappa - L$, we know that there exists $\nu_2 > 0$ and a definable desingularizing function Ψ_2 on $[0, \nu_2)$ such that Item 2 holds for Ψ_2 , i.e.,

$$\Psi_2'(\gamma(\bar{\boldsymbol{\theta}}^*) - \tilde{\gamma}(\boldsymbol{\theta})) \cdot \|\boldsymbol{\theta}\|_2^{2\kappa - L + 1} \|\bar{\partial}^\circ \tilde{\gamma}(\boldsymbol{\theta})\|_2 \geq 1.$$

Since $\Psi_1'(x) - \Psi_2'(x)$ is definable, there exists a sufficiently small constant $\nu > 0$ such that either $\Psi_1'(x) - \Psi_2'(x) \geq 0$ holds for all $x \in [0, \nu)$, or $\Psi_1'(x) - \Psi_2'(x) \leq 0$ holds for all $x \in [0, \nu)$. This means either $\Psi_1'(x) \geq \Psi_2'(x)$ for all $x \in [0, \nu)$ or $\Psi_2'(x) \geq \Psi_1'(x)$ for all $x \in [0, \nu)$. Let $\Psi(x) = \Psi_1(x)$ in the former case and $\Psi(x) = \Psi_2(x)$ in the latter case. Then $\Psi'(x) \geq \Psi_1'(x)$ and $\Psi'(x) \geq \Psi_2'(x)$, and thus both Items 1 and 2 hold. $\square$

Now we prove the following lemma, which will directly lead to Theorem 5.6. The core idea of the proof is essentially the same as that for Lemma 3.3 in Ji and Telgarsky (2020a). The key difference here is that the desingularizing function Ψ in their lemma has dependence on the initial point, while our lemma does not have such dependence.

Lemma F.11. *Consider any L -homogeneous neural networks with definable output $f_\theta(x_i)$ and logistic loss. Given a local-max-margin direction $\bar{\boldsymbol{\theta}}^* \in \mathbb{S}^{D-1}$, there is a desingularizing function on $[0, \nu)$ and two constants $\epsilon_0 > 0$, $\rho_0 \geq 1$ such that for any $\boldsymbol{\theta}_0$ with norm $\|\boldsymbol{\theta}_0\|_2 \geq \rho_0$ and direction $\left\| \frac{\boldsymbol{\theta}_0}{\|\boldsymbol{\theta}_0\|_2} - \bar{\boldsymbol{\theta}}^* \right\|_2 \leq \epsilon_0$, the gradient flow $\boldsymbol{\theta}(t)$ starting with $\boldsymbol{\theta}_0$ satisfies*

$$\frac{d\zeta(t)}{dt} \leq -c \frac{d\Psi(\gamma(\bar{\boldsymbol{\theta}}^*) - \tilde{\gamma}(\boldsymbol{\theta}(t)))}{dt}, \quad \text{for a.e. } t \in [0, T),$$

where $T := \inf\{t \geq 0 : \tilde{\gamma}(\boldsymbol{\theta}(t)) \geq \gamma(\bar{\boldsymbol{\theta}}^*)\} \in \mathbb{R} \cup \{+\infty\}$.

Proof. Fix an arbitrary $\kappa \in (L/2, L)$. Let Ψ be the desingularizing function on $[0, \nu)$ obtained from Lemma F.10. WLOG, we can make $\nu < \gamma(\bar{\boldsymbol{\theta}}^*)/2$.

Let $\tilde{\gamma}_{\inf}(\rho, \epsilon)$ be the following lower bound for the initial smoothed margin $\tilde{\gamma}(\boldsymbol{\theta}_0)$:

$$\tilde{\gamma}_{\inf}(\rho, \epsilon) := \inf \left\{ \tilde{\gamma}(\boldsymbol{\theta}) : \|\boldsymbol{\theta}\|_2 \geq \rho, \left\| \frac{\boldsymbol{\theta}}{\|\boldsymbol{\theta}\|_2} - \bar{\boldsymbol{\theta}}^* \right\|_2 \leq \epsilon \right\}. \quad (45)$$

We set ρ_0 to be sufficiently large and ϵ_0 to be sufficiently small so that $\tilde{\gamma}_{\inf}(\rho_0, \epsilon_0) > \frac{1}{2}\gamma(\bar{\boldsymbol{\theta}}^*)$ and $\rho_0^{L-\kappa} \geq (4 \ln n + 2)/\gamma(\bar{\boldsymbol{\theta}}^*)$. By chain rule, it suffices to prove

$$\frac{d\tilde{\gamma}(\boldsymbol{\theta}(t))}{dt} \geq \frac{1}{c\Psi'(\gamma(\bar{\boldsymbol{\theta}}^*) - \tilde{\gamma}(\boldsymbol{\theta}(t)))} \frac{d\zeta(t)}{dt}, \quad \text{for a.e. } t \in [0, T), \quad (46)$$

where $c = \max \left\{ 2, \frac{\gamma(\bar{\boldsymbol{\theta}}^*)}{2 \ln n + 1} \right\}$.

We consider two cases, where assume (42) is true in Case 1 and (42) is not true in Case 2. According to our choice of ρ_0 and the monotonicity of $\|\boldsymbol{\theta}(t)\|_2$, we have $\|\boldsymbol{\theta}(t)\|_2^{L-\kappa} \geq \rho_0^{L-\kappa} \geq \frac{4 \ln n + 2}{\gamma(\bar{\boldsymbol{\theta}}^*)}$, and thus $\frac{\gamma(\bar{\boldsymbol{\theta}}^*)}{4 \ln n + 2} \|\boldsymbol{\theta}(t)\|_2^{L-\kappa} \geq 1$. This means

$$\|\bar{\partial}_r^\circ \tilde{\gamma}(\boldsymbol{\theta}(t))\|_2 + \|\bar{\partial}_\perp^\circ \tilde{\gamma}(\boldsymbol{\theta}(t))\|_2 \leq 2 \|\bar{\partial}_r^\circ \tilde{\gamma}(\boldsymbol{\theta}(t))\|_2 \quad (47)$$

in Case 1, and

$$\|\bar{\partial}_r^\circ \tilde{\gamma}(\boldsymbol{\theta}(t))\|_2 + \|\bar{\partial}_\perp^\circ \tilde{\gamma}(\boldsymbol{\theta}(t))\|_2 < \frac{\gamma(\bar{\boldsymbol{\theta}}^*)}{2 \ln n + 1} \|\boldsymbol{\theta}(t)\|_2^{L-\kappa} \|\bar{\partial}_r^\circ \tilde{\gamma}(\boldsymbol{\theta}(t))\|_2 \quad (48)$$

in Case 2.

Case 1. For any $t \geq 0$, if $\frac{d\boldsymbol{\theta}(t)}{dt} = -\bar{\partial}^\circ \mathcal{L}(\boldsymbol{\theta}(t))$ and (42) hold for $\boldsymbol{\theta}(t)$. By Lemma F.6, we have the following lower bound for $\frac{d\tilde{\gamma}(\boldsymbol{\theta}(t))}{dt}$.

$$\frac{d\tilde{\gamma}(\boldsymbol{\theta}(t))}{dt} \geq \|\bar{\partial}_\perp^\circ \tilde{\gamma}(\boldsymbol{\theta}(t))\|_2 \|\bar{\partial}_\perp^\circ \mathcal{L}(\boldsymbol{\theta}(t))\|_2. \quad (49)$$

By triangle inequality and (47),

$$\|\bar{\partial}^\circ \tilde{\gamma}(\boldsymbol{\theta}(t))\|_2 \leq \|\bar{\partial}_r^\circ \tilde{\gamma}(\boldsymbol{\theta}(t))\|_2 + \|\bar{\partial}_\perp^\circ \tilde{\gamma}(\boldsymbol{\theta}(t))\|_2 \leq 2 \|\bar{\partial}_\perp^\circ \tilde{\gamma}(\boldsymbol{\theta}(t))\|_2.$$

So $\|\bar{\partial}_\perp^\circ \tilde{\gamma}(\boldsymbol{\theta}(t))\|_2 \geq \frac{1}{2} \|\bar{\partial}^\circ \tilde{\gamma}(\boldsymbol{\theta}(t))\|_2$. Combining this with (49) and noting that $\frac{d\zeta(t)}{dt} = \frac{1}{\|\boldsymbol{\theta}(t)\|_2} \|\bar{\partial}_\perp^\circ \mathcal{L}(\boldsymbol{\theta}(t))\|_2$, we have

$$\frac{d\tilde{\gamma}(\boldsymbol{\theta}(t))}{dt} \geq \frac{1}{2} \|\bar{\partial}^\circ \tilde{\gamma}(\boldsymbol{\theta}(t))\|_2 \cdot \left(\|\boldsymbol{\theta}(t)\|_2 \cdot \frac{d\zeta(t)}{dt} \right).$$

Applying (43) gives

$$\frac{d\tilde{\gamma}(\boldsymbol{\theta}(t))}{dt} \geq \frac{1}{2\Psi'(\gamma(\bar{\boldsymbol{\theta}}^*) - \tilde{\gamma}(\boldsymbol{\theta}(t)))} \frac{d\zeta(t)}{dt} \geq \frac{1}{c\Psi'(\gamma(\bar{\boldsymbol{\theta}}^*) - \tilde{\gamma}(\boldsymbol{\theta}(t)))} \frac{d\zeta(t)}{dt}.$$

Case 2. For any $t \geq 0$, if $\frac{d\boldsymbol{\theta}(t)}{dt} = -\bar{\partial}^\circ \mathcal{L}(\boldsymbol{\theta}(t))$ and (42) does not hold for $\boldsymbol{\theta}(t)$, i.e.,

$$\|\bar{\partial}_\perp^\circ \tilde{\gamma}(\boldsymbol{\theta}(t))\|_2 < \frac{\gamma(\bar{\boldsymbol{\theta}}^*)}{4 \ln n + 2} \|\boldsymbol{\theta}(t)\|_2^{L-\kappa} \|\bar{\partial}_r^\circ \tilde{\gamma}(\boldsymbol{\theta}(t))\|_2, \quad (50)$$

By Lemma F.6, we have the following lower bound for $\frac{d\tilde{\gamma}(\boldsymbol{\theta}(t))}{dt}$.

$$\frac{d\tilde{\gamma}(\boldsymbol{\theta}(t))}{dt} \geq \|\bar{\partial}_r^\circ \tilde{\gamma}(\boldsymbol{\theta}(t))\|_2 \|\bar{\partial}_r^\circ \mathcal{L}(\boldsymbol{\theta}(t))\|_2. \quad (51)$$

We lower bound $\|\bar{\partial}_r^\circ \tilde{\gamma}(\boldsymbol{\theta}(t))\|_2$ and $\|\bar{\partial}_r^\circ \mathcal{L}(\boldsymbol{\theta}(t))\|_2$ respectively in order to apply KL inequality (44).

Bounding $\|\bar{\partial}_r^\circ \tilde{\gamma}(\boldsymbol{\theta}(t))\|_2$ in Case 2. By triangle inequality and (48),

$$\begin{aligned} \|\bar{\partial}^\circ \tilde{\gamma}(\boldsymbol{\theta}(t))\|_2 &\leq \|\bar{\partial}_r^\circ \tilde{\gamma}(\boldsymbol{\theta}(t))\|_2 + \|\bar{\partial}_\perp^\circ \tilde{\gamma}(\boldsymbol{\theta}(t))\|_2 \\ &< \frac{\gamma(\bar{\boldsymbol{\theta}}^*)}{2 \ln n + 1} \|\boldsymbol{\theta}(t)\|_2^{L-\kappa} \|\bar{\partial}_r^\circ \tilde{\gamma}(\boldsymbol{\theta}(t))\|_2, \end{aligned}$$

which can be restated as

$$\|\bar{\partial}_r^\circ \tilde{\gamma}(\boldsymbol{\theta}(t))\|_2 \geq \frac{2 \ln n + 1}{\gamma(\bar{\boldsymbol{\theta}}^*)} \|\boldsymbol{\theta}(t)\|_2^{\kappa-L} \|\bar{\partial}^\circ \tilde{\gamma}(\boldsymbol{\theta}(t))\|_2. \quad (52)$$

Bounding $\|\bar{\partial}_r^\circ \mathcal{L}(\boldsymbol{\theta}(t))\|_2$ in Case 2. By Lemma C.3 in Ji and Telgarsky (2020a),

$$\|\bar{\partial}_r^\circ \tilde{\gamma}(\boldsymbol{\theta}(t))\|_2 \leq \frac{L \cdot (2 \ln n + 1)}{\|\boldsymbol{\theta}(t)\|_2^{L+1}}.$$

Combining this with (50),

$$\|\bar{\partial}_\perp^\circ \tilde{\gamma}(\boldsymbol{\theta}(t))\|_2 < \frac{\gamma(\bar{\boldsymbol{\theta}}^*)}{4 \ln n + 2} \|\boldsymbol{\theta}(t)\|_2^{L-\kappa} \cdot \frac{L \cdot (2 \ln n + 1)}{\|\boldsymbol{\theta}(t)\|_2^{L+1}} = \frac{\gamma(\bar{\boldsymbol{\theta}}^*)}{2} L \|\boldsymbol{\theta}(t)\|_2^{-(1+\kappa)},$$

which can be rewritten as

$$\frac{L\gamma(\bar{\boldsymbol{\theta}}^*)}{2} > \|\bar{\partial}_\perp^\circ \tilde{\gamma}(\boldsymbol{\theta}(t))\|_2 \|\boldsymbol{\theta}(t)\|_2^{1+\kappa}. \quad (53)$$

By the chain rule and Lemma C.5 in Ji and Telgarsky (2020a),

$$\|\bar{\partial}_r^\circ \alpha(\boldsymbol{\theta}(t))\|_2 = \frac{L \cdot \langle \boldsymbol{\theta}(t), \bar{\partial}^\circ \alpha(\boldsymbol{\theta}(t)) \rangle}{\|\boldsymbol{\theta}(t)\|_2} \geq \frac{L\alpha(\boldsymbol{\theta}(t))}{\|\boldsymbol{\theta}(t)\|_2} = L\tilde{\gamma}(\boldsymbol{\theta}(t))\|\boldsymbol{\theta}(t)\|_2^{L-1}. \quad (54)$$

By the monotonicity of $\tilde{\gamma}(\boldsymbol{\theta}(t))$ during training, $\tilde{\gamma}(\boldsymbol{\theta}(t)) \geq \tilde{\gamma}(\boldsymbol{\theta}(0))$. Also note that

$$\tilde{\gamma}(\boldsymbol{\theta}(0)) \geq \tilde{\gamma}_0 > \frac{1}{2}\gamma(\bar{\boldsymbol{\theta}}^*),$$

where the first inequality is by definition of $\tilde{\gamma}_0$, and the second inequality is due to our choice of ρ_0, ϵ_0 . So we can replace $\tilde{\gamma}(\boldsymbol{\theta}(t))$ with $\frac{1}{2}\gamma(\bar{\boldsymbol{\theta}}^*)$ in the RHS of (54) and obtain

$$\|\bar{\partial}_r^\circ \alpha(\boldsymbol{\theta}(t))\|_2 \geq \frac{L\gamma(\bar{\boldsymbol{\theta}}^*)}{2} \|\boldsymbol{\theta}(t)\|_2^{L-1}.$$

Combining this with (53) and noting that $\bar{\partial}_\perp^\circ \alpha(\boldsymbol{\theta}(t)) = \|\boldsymbol{\theta}(t)\|_2^L \bar{\partial}_\perp^\circ \tilde{\gamma}(\boldsymbol{\theta}(t))$, we have

$$\|\bar{\partial}_r^\circ \alpha(\boldsymbol{\theta}(t))\|_2 \geq \|\boldsymbol{\theta}(t)\|_2^{L+\kappa} \|\bar{\partial}_\perp^\circ \tilde{\gamma}(\boldsymbol{\theta}(t))\|_2 \geq \|\boldsymbol{\theta}(t)\|_2^\kappa \|\bar{\partial}_\perp^\circ \alpha(\boldsymbol{\theta}(t))\|_2. \quad (55)$$

Recall that $\alpha(\boldsymbol{\theta}) = \ell^{-1}(\mathcal{L}(\boldsymbol{\theta}))$. By chain rule, $\bar{\partial}^\circ \alpha(\boldsymbol{\theta})$ is equal to the subgradient $\bar{\partial}^\circ \mathcal{L}(\boldsymbol{\theta})$ rescaled by some factor. Thus (55) implies

$$\|\bar{\partial}_r^\circ \mathcal{L}(\boldsymbol{\theta}(t))\|_2 \geq \|\boldsymbol{\theta}(t)\|_2^\kappa \|\bar{\partial}_\perp^\circ \mathcal{L}(\boldsymbol{\theta}(t))\|_2. \quad (56)$$

Applying (44) for Case 2. Putting (51), (52) and (56) together gives

$$\begin{aligned} \frac{d\tilde{\gamma}(\boldsymbol{\theta}(t))}{dt} &\geq \left(\frac{2\ln n + 1}{\gamma(\bar{\boldsymbol{\theta}}^*)} \|\boldsymbol{\theta}(t)\|_2^{\kappa-L} \|\bar{\partial}^\circ \tilde{\gamma}(\boldsymbol{\theta}(t))\|_2 \right) \cdot (\|\boldsymbol{\theta}(t)\|_2^\kappa \|\bar{\partial}_\perp^\circ \mathcal{L}(\boldsymbol{\theta}(t))\|_2) \\ &\geq \frac{2\ln n + 1}{\gamma(\bar{\boldsymbol{\theta}}^*)} \|\boldsymbol{\theta}(t)\|_2^{2\kappa-L} \|\bar{\partial}^\circ \tilde{\gamma}(\boldsymbol{\theta}(t))\|_2 \cdot \|\bar{\partial}_\perp^\circ \mathcal{L}(\boldsymbol{\theta}(t))\|_2 \\ &= \frac{2\ln n + 1}{\gamma(\bar{\boldsymbol{\theta}}^*)} \|\boldsymbol{\theta}(t)\|_2^{2\kappa-L+1} \|\bar{\partial}^\circ \tilde{\gamma}(\boldsymbol{\theta}(t))\|_2 \cdot \frac{d\zeta(t)}{dt}, \end{aligned}$$

where the last equality is due to $\frac{d\zeta(t)}{dt} = \frac{1}{\|\boldsymbol{\theta}(t)\|_2} \|\bar{\partial}_\perp^\circ \mathcal{L}(\boldsymbol{\theta}(t))\|_2$. Applying (44) gives

$$\frac{d\tilde{\gamma}(\boldsymbol{\theta}(t))}{dt} \geq \frac{2\ln n + 1}{\gamma(\bar{\boldsymbol{\theta}}^*)} \cdot \frac{1}{\Psi'(\gamma(\bar{\boldsymbol{\theta}}^*) - \tilde{\gamma}(\boldsymbol{\theta}(t)))} \frac{d\zeta(t)}{dt} \geq \frac{1}{c\Psi'(\gamma(\bar{\boldsymbol{\theta}}^*) - \tilde{\gamma}(\boldsymbol{\theta}(t)))} \frac{d\zeta(t)}{dt}.$$

Final Proof Step. For a.e. $t \geq 0$, $\boldsymbol{\theta}(t)$ lies in either Case 1 or Case 2, so (46) holds, and we can rewrite it as

$$c\Psi'(\gamma(\bar{\boldsymbol{\theta}}^*) - \tilde{\gamma}(\boldsymbol{\theta}(t))) \frac{d\tilde{\gamma}(\boldsymbol{\theta}(t))}{dt} \geq \frac{d\zeta(t)}{dt}, \quad \text{for a.e. } t \in [0, T].$$

By chain rule, the LHS is equal to $\frac{d}{dt} (c\Psi(\gamma(\bar{\boldsymbol{\theta}}^*) - \tilde{\gamma}(\boldsymbol{\theta}(t))))$, which completes the proof. $\square$

Proof for Theorem 5.6. By Lemma F.11, we can choose ϵ_0, ρ_0 such that

$$\frac{d\zeta(t)}{dt} \leq -c \frac{d\Psi(\gamma(\bar{\boldsymbol{\theta}}^*) - \tilde{\gamma}(\boldsymbol{\theta}(t)))}{dt}, \quad \text{for a.e. } t \in [0, T],$$

where $T := \inf\{t \geq 0 : \tilde{\gamma}(\boldsymbol{\theta}(t)) \geq \gamma(\bar{\boldsymbol{\theta}}^*)\} \in \mathbb{R} \cup \{+\infty\}$. Then for all $t \in (0, T)$,

$$\zeta(t) \leq c\Psi(\gamma(\bar{\boldsymbol{\theta}}^*) - \tilde{\gamma}(\boldsymbol{\theta}_0)) \leq \delta(\epsilon_0, \rho_0) := c\Psi(\gamma(\bar{\boldsymbol{\theta}}^*) - \tilde{\gamma}_{\inf}(\rho_0, \epsilon_0)), \quad (57)$$

where $\tilde{\gamma}_{\inf}$ is defined in (45). We can choose ϵ_0 small enough and ρ_0 large enough so that $\delta(\epsilon_0, \rho_0) > 0$ is as small as we want.

If $T = +\infty$, then (57) implies that $\frac{\boldsymbol{\theta}(t)}{\|\boldsymbol{\theta}(t)\|_2}$ converges to some $\bar{\boldsymbol{\theta}}$ as $t \rightarrow +\infty$, and $\|\bar{\boldsymbol{\theta}} - \bar{\boldsymbol{\theta}}^*\|_2 \leq \delta$ if $\delta(\epsilon_0, \rho_0) \leq \delta$.

If T is finite, then by triangle inequality we have $\left\| \frac{\boldsymbol{\theta}(T)}{\|\boldsymbol{\theta}(T)\|_2} - \bar{\boldsymbol{\theta}}^* \right\|_2 \leq \epsilon_0 + \delta(\epsilon_0, \rho_0)$. Since $\bar{\boldsymbol{\theta}}^*$ is a local-max-margin direction, when ϵ_0 and $\delta(\epsilon_0, \rho_0)$ are sufficiently small, $\tilde{\gamma}(\boldsymbol{\theta}) \leq \gamma(\boldsymbol{\theta}) \leq \gamma(\bar{\boldsymbol{\theta}}^*)$ holds for any $\boldsymbol{\theta}$ satisfying $\left\| \frac{\boldsymbol{\theta}}{\|\boldsymbol{\theta}\|_2} - \bar{\boldsymbol{\theta}}^* \right\|_2 \leq 2(\epsilon_0 + \delta(\epsilon_0, \rho_0))$. The definition of T then implies that $\tilde{\gamma}(\boldsymbol{\theta}(T)) = \gamma(\boldsymbol{\theta}(T)) = \gamma(\bar{\boldsymbol{\theta}}^*)$. By Lemma B.1 from Lyu and Li (2020), $\tilde{\gamma}(\boldsymbol{\theta}(t))$ is non-decreasing over time, and if it stops increasing at some value, then the time derivative of $\frac{\boldsymbol{\theta}(t)}{\|\boldsymbol{\theta}(t)\|_2}$ must be zero.

Thus we have $\tilde{\gamma}(\boldsymbol{\theta}(t)) = \gamma(\bar{\boldsymbol{\theta}}^*)$ and $\frac{d}{dt} \frac{\boldsymbol{\theta}(t)}{\|\boldsymbol{\theta}(t)\|_2} = 0$ for all $t \geq T$, which implies that $\frac{\boldsymbol{\theta}(t)}{\|\boldsymbol{\theta}(t)\|_2}$ converges to $\bar{\boldsymbol{\theta}} := \frac{\boldsymbol{\theta}(T)}{\|\boldsymbol{\theta}(T)\|_2}$ as $t \rightarrow +\infty$. This again proves that $\|\bar{\boldsymbol{\theta}} - \bar{\boldsymbol{\theta}}^*\|_2 \leq \delta$.

Now we only need to show that $\gamma(\bar{\boldsymbol{\theta}}) = \gamma(\bar{\boldsymbol{\theta}}^*)$. In the case where T is finite, we have $\gamma(\bar{\boldsymbol{\theta}}) = \gamma\left(\frac{\boldsymbol{\theta}(T)}{\|\boldsymbol{\theta}(T)\|_2}\right) = \gamma(\bar{\boldsymbol{\theta}}^*)$. In the case where $T = +\infty$, $\gamma(\bar{\boldsymbol{\theta}})$ is a asymptotic Clarke critical value of $\tilde{\gamma}$ by Theorem F.9. Since there are only finitely many asymptotic Clarke critical values (Lemma F.7), we can make $\delta(\epsilon_0, \rho_0)$ to be small enough so that the only asymptotic Clarke critical value that can be achieved near $\bar{\boldsymbol{\theta}}^*$ is $\gamma(\bar{\boldsymbol{\theta}}^*)$ itself. $\square$

F.3 Proof for Theorem 4.3

Proof. By Lemma 5.4, $\lim_{\sigma_{\text{init}} \rightarrow 0} \boldsymbol{\theta}(T_{12} + t) = \pi_{\bar{\mathbf{b}}}(\tilde{\boldsymbol{\theta}}(t))$. Using Lemma E.8 and noting that $\langle \boldsymbol{\mu}, \mathbf{w}^* \rangle = \frac{1}{n} \sum_{i \in [n]} \langle y_i \mathbf{x}_i, \mathbf{w}^* \rangle > 0$, we know that there exists $t \leq t_0$ such that $\tilde{\boldsymbol{\theta}}(t) = (\tilde{\mathbf{w}}_1(t), \tilde{\mathbf{w}}_2(t), \tilde{a}_1(t), \tilde{a}_2(t))$ satisfies $\tilde{a}_1(t) = \|\tilde{\mathbf{w}}_1(t)\|_2$, $\tilde{a}_2(t) = -\|\tilde{\mathbf{w}}_2(t)\|_2$, $\langle \tilde{\mathbf{w}}_1(t), \mathbf{w}^* \rangle > 0$ and $\langle \tilde{\mathbf{w}}_2(t), \mathbf{w}^* \rangle < 0$. Then by Theorem 5.5,

$$\lim_{t \rightarrow +\infty} \frac{\tilde{\boldsymbol{\theta}}(t)}{\|\tilde{\boldsymbol{\theta}}(t)\|_2} = \frac{1}{2}(\mathbf{w}^*, -\mathbf{w}^*, 1, -1) =: \tilde{\boldsymbol{\theta}}_\infty,$$

which also implies that

$$\begin{aligned} \lim_{t \rightarrow +\infty} \lim_{\sigma_{\text{init}} \rightarrow 0} \frac{\boldsymbol{\theta}(T_{12} + t)}{\|\boldsymbol{\theta}(T_{12} + t)\|_2} &= \lim_{t \rightarrow +\infty} \frac{\pi_{\bar{\mathbf{b}}}(\tilde{\boldsymbol{\theta}}(t))}{\|\pi_{\bar{\mathbf{b}}}(\tilde{\boldsymbol{\theta}}(t))\|_2} = \pi_{\bar{\mathbf{b}}} \left(\lim_{t \rightarrow +\infty} \frac{\tilde{\boldsymbol{\theta}}(t)}{\|\tilde{\boldsymbol{\theta}}(t)\|_2} \right) = \pi_{\bar{\mathbf{b}}}(\tilde{\boldsymbol{\theta}}_\infty). \\ \lim_{t \rightarrow +\infty} \lim_{\sigma_{\text{init}} \rightarrow 0} \|\boldsymbol{\theta}(T_{12} + t)\|_2 &= \lim_{t \rightarrow +\infty} \|\pi_{\bar{\mathbf{b}}}(\tilde{\boldsymbol{\theta}}(t))\|_2 = +\infty. \end{aligned}$$

This means that for any $\epsilon > 0$ and $\rho > 0$, we can choose a time $t_1 \in \mathbb{R}$ such that $\left\| \frac{\boldsymbol{\theta}(T_{12} + t_1)}{\|\boldsymbol{\theta}(T_{12} + t_1)\|_2} - \pi_{\bar{\mathbf{b}}}(\tilde{\boldsymbol{\theta}}_\infty) \right\|_2 \leq \epsilon$ and $\|\boldsymbol{\theta}(T_{12} + t_1)\|_2 \geq \rho$ for any σ_{init} small enough. By Theorem 4.2, $\pi_{\bar{\mathbf{b}}}(\tilde{\boldsymbol{\theta}}_\infty)$ is a global-max-margin direction. Then Theorem 5.6 shows that there exists $\sigma_{\text{init}}^{\max}$ such that for all $\sigma_{\text{init}} < \sigma_{\text{init}}^{\max}$, $\frac{\boldsymbol{\theta}(t)}{\|\boldsymbol{\theta}(t)\|_2} \rightarrow \bar{\boldsymbol{\theta}}$, where $\gamma(\bar{\boldsymbol{\theta}}) = \gamma(\pi_{\bar{\mathbf{b}}}(\tilde{\boldsymbol{\theta}}_\infty))$ and $\|\bar{\boldsymbol{\theta}} - \pi_{\bar{\mathbf{b}}}(\tilde{\boldsymbol{\theta}}_\infty)\|_2 \leq \delta$. Therefore, $\bar{\boldsymbol{\theta}}$ is a global-max-margin direction and $f^\infty(\mathbf{x}) = \frac{1 + \alpha_{\text{leaky}}}{4} \langle \mathbf{w}^*, \mathbf{x} \rangle$ by Theorem 4.2. $\square$

G Trajectory-based Analysis for Non-symmetric Case

The proofs for the non-symmetric case follow similar manners from phase I to phase III. The high-level idea is to show the following in the 3 phases:

1. In Phase I, every weight vector $\mathbf{w}_k$ in the first layer moves towards the direction of either $\boldsymbol{\mu}^+$ or $-\boldsymbol{\mu}^-$. At the end of Phase I the weight vectors towards $-\boldsymbol{\mu}^-$ have much smaller norms than those towards $\boldsymbol{\mu}^+$, thereby becoming negligible.
2. In Phase II, we show that the dynamics of $\boldsymbol{\theta}(t)$ is close to a one-neuron dynamic (after embedding) for a long time.
3. In Phase III, we show that the one-neuron classifier converges to the max-margin solution among one-neuron neural nets (while the embedded classifier may have suboptimal margin among m -neuron neural nets), and the gradient flow $\boldsymbol{\theta}(t)$ on the m -neuron neural net gets stuck at a KKT-direction near this embedded classifier.

G.1 Additional Notations

In this section we highlight the additional notations that allow us to adapt the results from previous sections. For $\delta \geq 0$, define $\mathcal{C}^\delta$ to be the convex cone containing all the unit weight vectors that have δ margin over the dataset $\{(\mathbf{x}_i, y_i)\}_{i \in [n]}$.

$$\begin{aligned} \mathcal{C}^\delta &:= \{ \lambda \mathbf{w} : \langle \mathbf{w}, y_i \mathbf{x}_i \rangle \geq \delta, \mathbf{w} \in \mathbb{S}^{d-1}, \lambda > 0, \forall i \in [n] \}, \\ \mathcal{C} &:= \mathcal{C}^0 := \{ \mathbf{w} \in \mathbb{R}^d : \mathbf{w} \neq \mathbf{0}, \langle \mathbf{w}, y_i \mathbf{x}_i \rangle \geq 0, \forall i \in [n] \}. \end{aligned}$$

For $0 < \epsilon < 1$, we define

$$\mathcal{H}^\epsilon := \left\{ \frac{1}{2n} \sum_{i \in [n]} (1 + \epsilon_i) \alpha_i y_i \mathbf{x}_i : \alpha_i \in [\alpha_{\text{leaky}}, 1], \epsilon_i \in [-\epsilon, \epsilon], \forall i \in [n] \right\},$$

$$\mathcal{H} := \mathcal{H}^0 := \left\{ \frac{1}{2n} \sum_{i \in [n]} \alpha_i y_i \mathbf{x}_i : \alpha_i \in [\alpha_{\text{leaky}}, 1], \forall i \in [n] \right\}.$$

By Lemma B.14 we know $-\frac{\partial^\circ \mathcal{L}(\boldsymbol{\theta})}{\partial \mathbf{w}_k} \subseteq a_k \mathcal{H}^\epsilon$ if $\|\boldsymbol{\theta}\|_{\text{M}} \leq \sqrt{\frac{\epsilon}{2m}}$. Further we define

$$\mathcal{K}^\epsilon := \bigcup_{\lambda > 0} \lambda \mathcal{H}^\epsilon \quad \text{and} \quad \mathcal{K} := \mathcal{K}^0 = \bigcup_{\lambda > 0} \lambda \mathcal{H}^0$$

Then $-\frac{\partial^\circ \mathcal{L}(\boldsymbol{\theta})}{\partial \mathbf{w}_k} \subseteq \text{sgn}(a_k) \mathcal{K}^\epsilon$. For a set S , we will use $\mathring{S}$ to denote the interior of S .

Recall in Appendix A, for every $\mathbf{x}_i$, we define $\mathbf{x}_i^+ := \mathbf{x}_i$ if $y_i = 1$ and $\mathbf{x}_i^+ := \alpha_{\text{leaky}} \mathbf{x}_i$ if $y_i = -1$. Similarly, we define $\mathbf{x}_i^- := \alpha_{\text{leaky}} \mathbf{x}_i$ if $y_i = 1$ and $\mathbf{x}_i^- := \mathbf{x}_i$ if $y_i = -1$. Then we define $\boldsymbol{\mu}^+$ to be the mean vector of $y_i \mathbf{x}_i^+$, and $\boldsymbol{\mu}^-$ to be the mean vector of $y_i \mathbf{x}_i^-$, that is,

$$\boldsymbol{\mu}^+ := \frac{1}{n} \sum_{i \in [n]} y_i \mathbf{x}_i^+, \quad \boldsymbol{\mu}^- := \frac{1}{n} \sum_{i \in [n]} y_i \mathbf{x}_i^-.$$

We use $\bar{\boldsymbol{\mu}}^+ := \frac{\boldsymbol{\mu}^+}{\|\boldsymbol{\mu}^+\|_2}$, $\bar{\boldsymbol{\mu}}^- := \frac{\boldsymbol{\mu}^-}{\|\boldsymbol{\mu}^-\|_2}$ to denote $\boldsymbol{\mu}^+$, $\boldsymbol{\mu}^-$ after normalization. Similar to $\mathcal{K}^\epsilon$, we define $\mathcal{M}_+^\epsilon$ and $\mathcal{M}_-^\epsilon$ as the perturbed versions of $\boldsymbol{\mu}^+$ and $\boldsymbol{\mu}^-$ in the sense that $\mathcal{M}_+ := \{\lambda \boldsymbol{\mu}^+ : \lambda > 0\}$ and $\mathcal{M}_- := \{\lambda \boldsymbol{\mu}^- : \lambda > 0\}$.

$$\mathcal{M}_+^\epsilon = \left\{ \frac{\lambda}{2n} \sum_{i \in [n]} (1 + \epsilon_i) y_i \mathbf{x}_i^+ : \epsilon_i \in [-\epsilon, \epsilon], \lambda > 0 \right\},$$

$$\mathcal{M}_-^\epsilon = \left\{ \frac{\lambda}{2n} \sum_{i \in [n]} (1 + \epsilon_i) y_i \mathbf{x}_i^- : \epsilon_i \in [-\epsilon, \epsilon], \lambda > 0 \right\}.$$

G.2 More about Our Assumptions

The following lemma shows that Assumption A.1 is a weaker assumption than Assumption A.2.

Lemma G.1. *Assumption A.1 implies Assumption A.2.*

Proof. Let $\mathbf{w}^\diamond$ be the principal direction defined in Assumption A.1. We can decompose $\boldsymbol{\mu} = \boldsymbol{\mu}_\perp + \boldsymbol{\mu}_\parallel$, where $\boldsymbol{\mu}_\parallel$ is the along the direction of $\mathbf{w}^\diamond$ and $\boldsymbol{\mu}_\perp$ is orthogonal to $\mathbf{w}^\diamond$. Assumption A.1 implies that for all $i, j \in [n]$,

$$\begin{aligned} -\langle y_i \mathbf{x}_i, y_j \mathbf{x}_j \rangle &= -\langle y_i \mathbf{x}_i, \mathbf{w}^\diamond \rangle \cdot \langle y_j \mathbf{x}_j, \mathbf{w}^\diamond \rangle - \langle \mathbf{P}^\diamond(y_i \mathbf{x}_i), \mathbf{P}^\diamond(y_j \mathbf{x}_j) \rangle \\ &\leq \|\mathbf{P}^\diamond \mathbf{x}_i\|_2 \|\mathbf{P}^\diamond \mathbf{x}_j\|_2. \end{aligned}$$

Then for all $i \in [n]$, we have

$$\begin{aligned} \frac{1}{n} \sum_{j \in [n]} \max\{-\langle y_i \mathbf{x}_i, y_j \mathbf{x}_j \rangle, 0\} &\leq \frac{1}{n} \sum_{j \in [n]} \|\mathbf{P}^\diamond \mathbf{x}_i\|_2 \|\mathbf{P}^\diamond \mathbf{x}_j\|_2 \\ &\leq \|\mathbf{P}^\diamond \mathbf{x}_i\|_2 \cdot \alpha_{\text{leaky}} \langle \boldsymbol{\mu}, \mathbf{w}^\diamond \rangle \frac{\gamma^\diamond}{\max_{j \in [n]} \|\mathbf{P}^\diamond \mathbf{x}_j\|_2} \\ &\leq \alpha_{\text{leaky}} \langle \boldsymbol{\mu}, \mathbf{w}^\diamond \rangle \gamma^\diamond. \end{aligned}$$

On the other hand, recall that $\gamma^\diamond := \min_{i \in [n]} y_i \langle \mathbf{w}^\diamond, \mathbf{x}_i \rangle$, then we have

$$\begin{aligned}
\langle \boldsymbol{\mu}, y_i \mathbf{x}_i \rangle &= \langle \boldsymbol{\mu}, \mathbf{w}^\diamond \rangle \langle y_i \mathbf{x}_i, \mathbf{w}^\diamond \rangle + \langle \mathbf{P}^\diamond \boldsymbol{\mu}, \mathbf{P}^\diamond (y_i \mathbf{x}_i) \rangle \\
&\geq \langle \boldsymbol{\mu}, \mathbf{w}^\diamond \rangle \langle y_i \mathbf{x}_i, \mathbf{w}^\diamond \rangle - \frac{1}{n} \sum_{j \in [n]} \|\mathbf{P}^\diamond \mathbf{x}_j\|_2 \|\mathbf{P}^\diamond \mathbf{x}_i\|_2 \\
&\geq \langle \boldsymbol{\mu}, \mathbf{w}^\diamond \rangle \gamma^\diamond - \|\mathbf{P}^\diamond \mathbf{x}_i\|_2 \cdot \alpha_{\text{leaky}} \langle \boldsymbol{\mu}, \mathbf{w}^\diamond \rangle \frac{\gamma^\diamond}{\max_{j \in [n]} \|\mathbf{P}^\diamond \mathbf{x}_j\|_2} \\
&\geq (1 - \alpha_{\text{leaky}}) \langle \boldsymbol{\mu}, \mathbf{w}^\diamond \rangle \gamma^\diamond.
\end{aligned}$$

Combining these proves that $\langle \boldsymbol{\mu}, y_i \mathbf{x}_i \rangle \geq \frac{1 - \alpha_{\text{leaky}}}{n \cdot \alpha_{\text{leaky}}} \sum_{j \in [n]} \max\{-\langle y_i \mathbf{x}_i, y_j \mathbf{x}_j \rangle, 0\}$. $\square$

Lemma G.2 gives the main property we will use from Assumption A.2, i.e. $\mathcal{K} \subseteq \mathring{\mathcal{C}}$.

Lemma G.2. *For linearly separable dataset $\{(\mathbf{x}_i, y_i)\}_{i \in [n]}$ and $\alpha_{\text{leaky}} \in (0, 1]$, Assumption A.2 is equivalent to $\mathcal{K} \subseteq \mathring{\mathcal{C}}$.*

Proof. By definition, we know

$$\begin{aligned}
\mathcal{C} &= \{\mathbf{w} \in \mathbb{R}^d : \mathbf{w} \neq \mathbf{0}, \langle \mathbf{w}, y_i \mathbf{x}_i \rangle \geq 0, \forall i \in [n]\}, \\
\mathring{\mathcal{C}} &= \{\mathbf{w} \in \mathbb{R}^d : \langle \mathbf{w}, y_i \mathbf{x}_i \rangle > 0, \forall i \in [n]\}.
\end{aligned}$$

and $\mathcal{K} = \left\{ \lambda \sum_{i \in [n]} \alpha_i y_i \mathbf{x}_i : \alpha_i \in [\alpha_{\text{leaky}}, 1], \lambda > 0 \right\}$. For any $i \in [n]$, we have

$$\begin{aligned}
&\langle \boldsymbol{\mu}, y_i \mathbf{x}_i \rangle > \frac{1 - \alpha_{\text{leaky}}}{n \cdot \alpha_{\text{leaky}}} \sum_{j \in [n]} \max\{-\langle y_i \mathbf{x}_i, y_j \mathbf{x}_j \rangle, 0\} \\
&\iff \frac{1}{n} \sum_{j \in [n]} \left(\frac{1 - \alpha_{\text{leaky}}}{\alpha_{\text{leaky}}} \min\{\langle y_i \mathbf{x}_i, y_j \mathbf{x}_j \rangle, 0\} + \langle y_i \mathbf{x}_i, y_j \mathbf{x}_j \rangle \right) > 0 \\
&\iff \sum_{j \in [n]} \underbrace{((1 - \alpha_{\text{leaky}}) \min\{\langle y_i \mathbf{x}_i, y_j \mathbf{x}_j \rangle, 0\} + \alpha_{\text{leaky}} \langle y_i \mathbf{x}_i, y_j \mathbf{x}_j \rangle)}_{\Delta_{ij}} > 0
\end{aligned}$$

Note that $\Delta_{ij} = \min_{\alpha_j \in [\alpha_{\text{leaky}}, 1]} \langle y_i \mathbf{x}_i, \alpha_j y_j \mathbf{x}_j \rangle$. So

$$\sum_{j \in [n]} \Delta_{ij} = \sum_{j \in [n]} \min_{\alpha_j \in [\alpha_{\text{leaky}}, 1]} \langle y_i \mathbf{x}_i, \alpha_j y_j \mathbf{x}_j \rangle = \min_{\boldsymbol{\alpha} \in [\alpha_{\text{leaky}}, 1]^n} \left\langle y_i \mathbf{x}_i, \sum_{j \in [n]} \alpha_j y_j \mathbf{x}_j \right\rangle.$$

Therefore we have the following equivalence:

$$\text{Assumption A.2} \iff \forall i \in [n] : \min_{\boldsymbol{\alpha} \in [\alpha_{\text{leaky}}, 1]^n} \left\langle y_i \mathbf{x}_i, \sum_{j \in [n]} \alpha_j y_j \mathbf{x}_j \right\rangle > 0 \quad (58)$$

$$\iff \forall \mathbf{w} \in \mathcal{K}, \forall i \in [n] : \langle y_i \mathbf{x}_i, \mathbf{w} \rangle > 0 \quad (59)$$

$$\iff \mathcal{K} \subseteq \mathring{\mathcal{C}}. \quad (60)$$

which completes the proof. $\square$

Lemma G.2 shows that every direction in $\mathcal{K}$ has non-zero margin. Below we let the δ be the minimum of the margin of unit-norm linear separators in $\mathcal{K}$:

$$\delta := \min_{\mathbf{w} \in \mathcal{K} \cap \mathbb{S}^{d-1}} \min_{i \in [n]} \langle y_i \mathbf{x}_i, \mathbf{w} \rangle.$$

By (58) we have $\delta > 0$, and thus $\mathcal{K} \subseteq \mathcal{C}^\delta$.

G.3 Phase I

The overall result we will prove for phase I in the non-symmetric case is Lemma G.5. Compared to the symmetric case, even G function is not linear anymore. Recall G is defined as below:

$$G(\mathbf{w}) := \frac{-\ell'(0)}{n} \sum_{i \in [n]} y_i \phi(\mathbf{w}^\top \mathbf{x}_i) = \frac{1}{2n} \sum_{i \in [n]} y_i \phi(\mathbf{w}^\top \mathbf{x}_i).$$

It holds that $\forall \mathbf{w} \in \mathbb{R}^d$, $\partial^\circ G(\mathbf{w}) \subseteq \mathcal{K}$. Moreover, we have $\mathbf{w} \in \mathring{\mathcal{C}} \implies \partial^\circ G(\mathbf{w}) = \{\boldsymbol{\mu}^+\}$ and $\mathbf{w} \in -\mathring{\mathcal{C}} \implies \partial^\circ G(\mathbf{w}) = \{\boldsymbol{\mu}^-\}$. Thanks to Assumption A.2, we can show each neuron $\mathbf{w}_k(t)$ will eventually converge to areas with fixed sign pattern $\pm \mathcal{C}^{\delta/3}$ and thus G will become linear. Lemma G.3 states this idea more formally. Its proof is a simplication to the realistic case, Lemma G.4, and thus omitted. We will not use Lemma G.3 in the future.

Lemma G.3. *For any dataset $\{(\mathbf{x}_i, y_i)\}_{i \in [n]}$ satisfying Assumption A.2, suppose $\mathbf{w}(0) \neq \lambda \boldsymbol{\mu}^-$, $\forall \lambda \geq 0$, and it holds that*

$$\frac{d\mathbf{w}}{dt} \in \|\mathbf{w}\|_2 \cdot \partial^\circ G(\mathbf{w}), \quad (61)$$

then there exists $T_0 > 0$, such that $\mathbf{w}(T_0) \in \mathcal{C}^{\delta/2}$.

However, in the realistic setting, each $\mathbf{w}_k$ is not following gradient flow of G exactly — there are tiny correlations between different $\mathbf{w}_k$. And we will control those correlations by setting initialization very small. This yields Lemma G.4.

Lemma G.4. *Under Assumption A.2, if $\bar{\boldsymbol{\theta}} = (\bar{\mathbf{w}}_1, \dots, \bar{\mathbf{w}}_m, \bar{a}_1, \dots, \bar{a}_m)$ satisfies the following three conditions:*

1. *For all $k \in [m]$, $|\bar{a}_k| = \|\bar{\mathbf{w}}_k\|_2 \neq 0$;*
2. *If $\bar{a}_k > 0$, then $\bar{\mathbf{w}}_k \neq \lambda \boldsymbol{\mu}^-$ for any $\lambda > 0$;*
3. *If $\bar{a}_k < 0$, then $\bar{\mathbf{w}}_k \neq -\lambda \boldsymbol{\mu}^+$ for any $\lambda > 0$;*

then there exist $T_0, \sigma_{\text{init}}^{\max} > 0$, such that for any $\sigma_{\text{init}} < \sigma_{\text{init}}^{\max}$, the gradient flow $\boldsymbol{\theta}(t) = (\mathbf{w}_1(t), \dots, \mathbf{w}_m(t), a_1(t), \dots, a_m(t)) = \varphi(\sigma_{\text{init}} \bar{\boldsymbol{\theta}}, t)$ satisfies the following at time T_0 ,

$$\mathbf{w}_k(T_0) \in \begin{cases} \mathcal{C}^{\delta/3}, & \text{if } \bar{a}_k > 0, \\ -\mathcal{C}^{\delta/3}, & \text{if } \bar{a}_k < 0. \end{cases} \quad (62)$$

Moreover, there are constants $A, B > 0$ such that $A\sigma_{\text{init}} \leq \|\mathbf{w}_k(T_0)\|_2 \leq B\sigma_{\text{init}}$.

It is easy to see that the three conditions in Lemma G.4 hold with probability 1 over the random draw of $\bar{\boldsymbol{\theta}}_0 \sim \mathcal{D}_{\text{init}}(1)$. Then after time T_0 , all the neurons $\mathbf{w}_k$ are either in $\mathcal{C}^{\delta/3}$ or $-\mathcal{C}^{\delta/3}$, and will not leave it until $T_{\sigma_{\text{init}}}^\epsilon$, which implies the sign patterns $\text{sgn}(\langle \mathbf{x}_i, \mathbf{w}_k(t) \rangle) = s_k y_i$ is fixed for $t \in [T_0, T_{\sigma_{\text{init}}}^\epsilon]$. Thus similar to the symmetric case, $\boldsymbol{\theta}(t)$ evolves approximately under power iteration and yields the following lemma.

Lemma G.5. *Suppose that Assumptions A.2 and A.3 hold. Let $T_1(\sigma_{\text{init}}, r) := \frac{1}{\lambda_0^+} \ln \frac{r}{\sqrt{m}\sigma_{\text{init}}}$. With probability 1 over the random draw of $\bar{\boldsymbol{\theta}}_0 = (\bar{\mathbf{w}}_1, \dots, \bar{\mathbf{w}}_m, \bar{a}_1, \dots, \bar{a}_m) \sim \mathcal{D}_{\text{init}}(1)$, the prerequisites of Lemma G.4 are satisfied. In this case, there exists a vector $\bar{\mathbf{b}}(\sigma_{\text{init}}) \in \mathbb{R}^m$ for any $\sigma_{\text{init}} > 0$ such that the following statements hold:*

1. *There exist constants $C_1 > 0, C_2 > 0, T_0 \geq 0, r_{\max} > 0$ such that for $r \in (0, r_{\max}), \sigma_{\text{init}} \in (0, C_1 r^3)$, any neuron $(\mathbf{w}_k, a_k)$ at time $T_0 + T_1(\sigma_{\text{init}}, r)$ can be decomposed into*

$$\begin{aligned} \text{If } \bar{a}_k > 0: \quad & \mathbf{w}_k(T_0 + T_1(\sigma_{\text{init}}, r)) = r \bar{b}_k(\sigma_{\text{init}}) \bar{\boldsymbol{\mu}}^+ + \Delta \mathbf{w}_k, \\ & a_k(T_0 + T_1(\sigma_{\text{init}}, r)) = r \bar{b}_k(\sigma_{\text{init}}) + \Delta a_k, \\ \text{If } \bar{a}_k < 0: \quad & \mathbf{w}_k(T_0 + T_1(\sigma_{\text{init}}, r)) = r^{1-\kappa} \bar{b}_k(\sigma_{\text{init}}) \bar{\boldsymbol{\mu}}^- + \Delta \mathbf{w}_k, \\ & a_k(T_0 + T_1(\sigma_{\text{init}}, r)) = r^{1-\kappa} \bar{b}_k(\sigma_{\text{init}}) + \Delta a_k, \end{aligned}$$

where the error term $\Delta \boldsymbol{\theta} := (\Delta \mathbf{w}_1, \dots, \Delta \mathbf{w}_m, \Delta a_1, \dots, \Delta a_m)$ is upper bounded by $\|\Delta \boldsymbol{\theta}\|_{\text{M}} \leq C_2 r^3$ and κ is the gap $1 - \frac{\|\boldsymbol{\mu}^-\|_2}{\|\boldsymbol{\mu}^+\|_2} > 0$.

2. There exist constants $\bar{A}, \bar{B} > 0$ such that $|\bar{b}_k(\sigma_{\text{init}})| \in [\bar{A}, \bar{B}]$ whenever $\bar{a}_k > 0$ and $|\bar{b}_k(\sigma_{\text{init}})| \in [\sigma_{\text{init}}^\kappa \bar{A}, \sigma_{\text{init}}^\kappa \bar{B}]$ whenever $\bar{a}_k < 0$.

As $\sigma_{\text{init}} \rightarrow 0$, $|\bar{b}_k(\sigma_{\text{init}})| \rightarrow 0$ for neurons with $\bar{a}_k < 0$, while $|\bar{b}_k(\sigma_{\text{init}})| \in [\bar{A}, \bar{B}]$ remains for neurons with $\bar{a}_k > 0$. This means when the initialization scale is small, only the neurons with $\bar{a}_k > 0$ remain effective and the others become negligible. Those effective neurons move their weight vectors towards the direction of $\bar{\mu}^+$, until the error term $\Delta\theta$ becomes large.

G.3.1 Proof of Lemma G.4

Proof of Lemma G.4. Let $s_k := \text{sgn}(\bar{a}_k)$. By Corollary B.18, $a_k(t) = s_k \|\mathbf{w}(t)\|_2$ for all $t \geq 0$. Define $T_{\sigma_{\text{init}}}^\epsilon := \inf \{t \geq 0 : \|\theta(t)\|_{\text{M}} \geq \sqrt{\frac{\epsilon}{m}}\}$. By Lemma B.14, we have $\forall t \leq T_{\sigma_{\text{init}}}^\epsilon$, $-\frac{\partial^\circ \mathcal{L}(\theta(t))}{\partial \mathbf{w}_k} \subseteq a_k(t) \mathcal{H}^\epsilon \subseteq s_k \mathcal{K}^\epsilon$. Since $\mathcal{K} \subseteq \mathcal{C}^\delta$, there exists $\epsilon_1 > 0$, such that for all $\epsilon < \epsilon_1$, $\mathcal{K}^\epsilon \subseteq \mathcal{C}^{2\delta/3}$. The high-level idea of the proof is that suppose $-\frac{\partial^\circ \mathcal{L}(\theta(t))}{\partial \mathbf{w}_k} \subseteq s_k \mathcal{C}^{2\delta/3}$ holds for sufficiently long time T_0 , $\mathbf{w}_k(t)$ will eventually end up in a cone $s_k \mathcal{C}^{\delta/3}$ slightly wider than $s_k \mathcal{C}^{2\delta/3}$, as long as the total distance traveled is sufficiently long. On the other hand, we can make $\sigma_{\text{init}}^{\text{max}}$ sufficiently small, such that $T_{\sigma_{\text{init}}}^\epsilon \geq T_0$ for all $\sigma_{\text{init}} < \sigma_{\text{init}}^{\text{max}}$.

By Lemma B.16 and Lipschitzness of ℓ ,

$$\left| \frac{1}{2} \frac{d \|\mathbf{w}_k\|_2^2}{dt} \right| = \left| \frac{1}{n} \sum_{i=1}^n \ell'(q_i(\theta)) y_i a_k \phi(\mathbf{w}_k^\top \mathbf{x}_i) \right| \leq \frac{1}{n} \sum_{i=1}^n |a_k| \cdot \|\mathbf{w}_k\|_2 \leq \|\mathbf{w}_k\|_2^2.$$

Then we have

$$\forall t \leq T_{\sigma_{\text{init}}}^\epsilon, \quad \|\mathbf{w}_k(t)\|_2 \in [\|\mathbf{w}_k(0)\|_2 e^{-t}, \|\mathbf{w}_k(0)\|_2 e^t]. \quad (63)$$

Thus for any $T_0 \geq 0$, if $\sigma_{\text{init}} \leq e^{-T_0} \sqrt{\frac{\epsilon}{m}} / \|\theta\|_{\text{M}}$, we have $T_0 \leq T_{\sigma_{\text{init}}}^\epsilon$.

In order to lower bound the total travel distance for each $\mathbf{w}_k(t)$, it turns out that it suffices to lower bound the $\inf_{t \in [0, T_{\sigma_{\text{init}}}^\epsilon]} \|\mathbf{w}_k(t)\|_2$ by $\bar{D} \sigma_{\text{init}}$, where $\bar{D} > 0$ is some constant. We will first show that we can guarantee the existence of such constant $\bar{D}$ by picking sufficiently small ϵ . Then we will formally prove the original claim of Lemma G.4.

Existence of $\bar{D}$. By definitions of $\mathcal{M}_+$ and $\mathcal{M}_-$, it holds that $\forall k \in [m]$,

$$\bar{\mathbf{w}}_k \notin \begin{cases} \mathcal{M}_+, & \text{if } \bar{a}_k < 0; \\ -\mathcal{M}_-, & \text{if } \bar{a}_k > 0. \end{cases}$$

In other words

$$\bar{d} := \min \left\{ \min_{k: \bar{a}_k < 0} \text{dist}(\bar{\mathbf{w}}_k - \mathcal{M}_+, \mathbf{0}), \min_{k: \bar{a}_k > 0} \text{dist}(\bar{\mathbf{w}}_k + \mathcal{M}_-, \mathbf{0}) \right\} > 0.$$

By the continuity of the distance function, there exists $\epsilon_2 > 0$ such that $\forall \epsilon \in (0, \epsilon_2)$, it holds that

$$\min \left\{ \min_{k: \bar{a}_k < 0} \text{dist}(\bar{\mathbf{w}}_k - \mathcal{M}_+^\epsilon, \mathbf{0}), \min_{k: \bar{a}_k > 0} \text{dist}(\bar{\mathbf{w}}_k + \mathcal{M}_-^\epsilon, \mathbf{0}) \right\} \geq \frac{\bar{d}}{2}.$$

Now we take $\epsilon = \min\{\epsilon_1, \epsilon_2\}$. We will first show the existence of such $\bar{D}$ for $k \in [m]$ with $\bar{a}_k > 0$. And the same argument holds for k with negative $\bar{a}_k$. Let $t_k := \sup\{t \leq T_{\sigma_{\text{init}}}^\epsilon : \mathbf{w}_k(t) \in -\mathcal{C}^{\delta/3}\}$. We note that $\mathbf{w}_k(t) \in -\mathcal{C}^{\delta/3}$ for all $t \leq t_k$. Otherwise, $\mathbf{w}_k(t') \notin -\mathcal{C}^{\delta/3}$ for some $t' < t_k$. On the one hand, we have $\mathbf{w}_k(t_k) \in \mathbf{w}_k(t') + \mathcal{K}^\epsilon \subseteq \mathbf{w}_k(t') + \mathcal{C}^{2\delta/3}$; on the other hand, we also know that $\mathbf{w}_k(t_k) \in -\mathcal{C}^{\delta/3}$ by continuity of the trajectory of $\mathbf{w}_k(t)$. This implies $-\mathcal{C}^{\delta/3} \cap (\mathcal{C}^{2\delta/3} + \mathbf{w}_k(t')) \neq \emptyset$, and thus $\mathbf{w}_k(t') \in -\mathcal{C}^{\delta/3} - \mathcal{C}^{2\delta/3} \subseteq -\mathcal{C}^{\delta/3} - \mathcal{C}^{\delta/3} \subseteq -\mathcal{C}^{\delta/3}$. Contradiction.

Now we have $\mathbf{w}_k(t) \in -\mathcal{C}^{\delta/3}$ for all $t \leq t_k$, and this implies that $\langle \mathbf{w}_k(t), \mathbf{x}_i \rangle < 0$ for all $i \in [n]$. Then $\frac{d\mathbf{w}_k(t)}{dt} = -\nabla_{\mathbf{w}_k} \mathcal{L}(\theta(t)) \in \mathcal{M}_-^\epsilon$. Therefore we have $\inf_{t \in [0, t_k]} \|\mathbf{w}_k(t)\|_2 \geq \text{dist}(\mathbf{w}_k(0) + \mathcal{M}_-^\epsilon, \mathbf{0}) = \sigma_{\text{init}} \text{dist}(\bar{\mathbf{w}}_k + \mathcal{M}_-^\epsilon, \mathbf{0}) \geq \frac{\bar{d} \sigma_{\text{init}}}{2}$.

Below we show the norm lower bound for any t such that $t \in [t_k, T_{\sigma_{\text{init}}}^\epsilon]$. Let $\bar{d}'$ be the minimum distance between any point in $-\mathcal{C}^{2\delta/3}$ and any point on unit sphere but not in $-\mathcal{C}^{\delta/2}$, that is,

$$\bar{d}' := \text{dist}(\mathbb{S}^{d-1} \setminus (-\mathcal{C}^{\delta/2}), -\mathcal{C}^{2\delta/3})$$

We claim that $\bar{d}' > 0$. Otherwise there is a sequence of $\{\mathbf{w}_j\}$ with unit norm and $\mathbf{w}_j \notin -\mathcal{C}^{\delta/2}$ satisfying that $\lim_{n \rightarrow \infty} \text{dist}(\mathbf{w}_k, -\mathcal{C}^{2\delta/3}) = 0$. Let $\bar{\mathbf{w}}$ be a limit point, then $\bar{\mathbf{w}} \in -\mathcal{C}^{2\delta/3}$ since $-\mathcal{C}^{2\delta/3}$ is closed. Since $-\mathcal{C}^{2\delta/3} \subseteq -\mathcal{C}^{\delta/2}$, we further have $\bar{\mathbf{w}} \in -\mathcal{C}^{\delta/2}$, which contradicts with the definition of limit point.

By the continuity of $\mathbf{w}_k(t)$, we know $\mathbf{w}_k(t_k) \notin -\mathcal{C}^{\delta/2}$. Thus for any $t \in [t_k, T_{\sigma_{\text{init}}}^\epsilon]$, we have $\mathbf{w}_k(t) \in \mathbf{w}_k(t_k) + \mathcal{C}^{2\delta/3}$ and $\inf_{t \in [t_k, T_{\sigma_{\text{init}}}^\epsilon]} \|\mathbf{w}_k(t)\|_2 \geq \text{dist}(\mathbf{0}, \mathbf{w}_k(t_k) + \mathcal{C}^{2\delta/3}) = \text{dist}(-\mathcal{C}^{2\delta/3}, \mathbf{w}_k(t_k)) = \|\mathbf{w}_k(t_k)\|_2 \text{dist}(-\mathcal{C}^{2\delta/3}, \frac{\mathbf{w}_k(t_k)}{\|\mathbf{w}_k(t_k)\|_2}) \geq \|\mathbf{w}_k(t_k)\|_2 \cdot \bar{d}' \geq \frac{\bar{d}' \sigma_{\text{init}}}{2}$. We can apply the same argument for those k with $\bar{a}_k < 0$, and finally we can conclude that $\|\mathbf{w}_k(t)\|_2 \geq \bar{D} \sigma_{\text{init}}$ for all $t \in [0, T_{\sigma_{\text{init}}}^\epsilon]$ and $k \in [m]$, where $\bar{D} := \max\{1, \bar{d}'\}$.

Convergence to $\mathcal{C}^{\delta/3}$. For $c \geq 0$ and $i \in [n]$ define $\Gamma_i^c(\mathbf{w}) := \langle \mathbf{w}, y_i \mathbf{x}_i \rangle - c \|\mathbf{w}\|_2$. For all $k \in [m]$ and $t \leq T_{\sigma_{\text{init}}}^\epsilon$, it holds that

$$\begin{aligned} \frac{d\Gamma_i^{\delta/3}(s_k \mathbf{w}_k)}{dt} &= \left\langle \frac{d\mathbf{w}_k}{dt}, s_k y_i \mathbf{x}_i \right\rangle - (\delta/3) \left\langle \frac{d\mathbf{w}_k}{dt}, \frac{\mathbf{w}_k}{\|\mathbf{w}_k\|_2} \right\rangle \\ &\geq (2\delta/3) \left\| \frac{d\mathbf{w}_k}{dt} \right\|_2 - (\delta/3) \left\| \frac{d\mathbf{w}_k}{dt} \right\|_2 = (\delta/3) \left\| \frac{d\mathbf{w}_k}{dt} \right\|_2, \end{aligned}$$

where the inequality is because $\frac{d\mathbf{w}_k}{dt} \subseteq a_k \mathcal{H}^\epsilon \subseteq a_k \mathcal{C}^{2\delta/3}$ and $\langle \frac{d\mathbf{w}_k}{dt}, \frac{\mathbf{w}_k}{\|\mathbf{w}_k\|_2} \rangle \leq \left\| \frac{d\mathbf{w}_k}{dt} \right\|_2$.

Let $h_{\min} := \inf_{\mathbf{w} \in \mathcal{H}^\epsilon} \|\mathbf{w}\|_2 = \min_{\mathbf{w} \in \mathcal{H}^\epsilon} \|\mathbf{w}\|_2 > 0$. Note that $|a_k(t)| = \|\mathbf{w}_k(t)\|_2 \geq \bar{D} \sigma_{\text{init}}$. Using $\frac{d\mathbf{w}_k}{dt} \subseteq a_k \mathcal{H}^\epsilon$ again we have

$$\frac{d\Gamma_i^{\delta/3}(s_k \mathbf{w}_k(t))}{dt} \geq \frac{\delta h_{\min} \bar{D} \sigma_{\text{init}}}{3}.$$

Thus if we pick

$$T_0 := \max \left\{ \frac{3}{\delta h_{\min} \bar{D}} \max_{i \in [n], k \in [m]} \{-\Gamma_i^{\delta/3}(s_k \bar{\mathbf{w}}_k)\}, 0 \right\}$$

and set $\sigma_{\text{init}}^{\max} \leq e^{-T_0} \frac{\sqrt{\frac{\epsilon}{m}}}{\|\bar{\boldsymbol{\theta}}\|_M}$ then it holds that $T_0 \leq T_{\sigma_{\text{init}}}^\epsilon$ for all $\sigma_{\text{init}} \leq \sigma_{\text{init}}^{\max}$ and that

$$\begin{aligned} \Gamma_i^{\delta/3}(s_k \mathbf{w}_k(T_0)) &\geq \frac{\delta h_{\min} \bar{D} \sigma_{\text{init}} T_0}{3} + \Gamma_i^{\delta/3}(s_k \mathbf{w}_k(0)) \\ &\geq \sigma_{\text{init}} \left(\frac{\delta h_{\min} \bar{D} T_0}{3} + \Gamma_i^{\delta/3}(s_k \bar{\mathbf{w}}_k(0)) \right) \\ &\geq 0, \end{aligned}$$

which implies (62).

Finally, by (63), it suffices to pick $A = e^{-T_0} \min_{k \in [m]} \|\bar{\mathbf{w}}_k\|_2$ and $B = e^{T_0} \max_{k \in [m]} \|\bar{\mathbf{w}}_k\|_2$. $\square$

G.3.2 Proof of Lemma G.5

Note that $G(\mathbf{w}) = \langle \mathbf{w}, \boldsymbol{\mu}^+ \rangle$ for $\mathbf{w} \in \mathcal{C}^{\delta/3}$ and $G(\mathbf{w}) = \langle \mathbf{w}, \boldsymbol{\mu}^- \rangle$ for $\mathbf{w} \in -\mathcal{C}^{\delta/3}$. Similar to the first-phase analysis to the symmetric case, we use $\tilde{\varphi}(\bar{\boldsymbol{\theta}}_0, t)$ to denote the trajectory of gradient flow on $\tilde{\mathcal{L}}$:

$$\tilde{\mathcal{L}}(\boldsymbol{\theta}) := \ell(0) + \sum_{k \in [m]} a_k G(\mathbf{w}_k).$$

Throughout this subsection, we will set T_0 and ϵ as defined in the proof of Lemma G.4, and therefore by Lemma G.4, we know there is $\sigma_{\text{init}}^{\max} > 0$, s.t. $a_k(T_0) \mathbf{w}_k(T_0) \in \mathcal{C}^{\delta/3}$ for all $\sigma_{\text{init}} \leq \sigma_{\text{init}}^{\max}$. This

means the dynamics of $\tilde{\boldsymbol{\theta}}(t) = (\tilde{\mathbf{w}}_1(t), \dots, \tilde{\mathbf{w}}_m(t), \tilde{a}_1(t), \dots, \tilde{a}_m(t)) = \tilde{\varphi}(\boldsymbol{\theta}(T_0), t - T_0)$ can be described by linear ODE for $T_0 \leq t \leq T_{\sigma_{\text{init}}}^\epsilon$.

$$\begin{aligned} \text{If } \bar{a}_k > 0: \quad & \frac{d\tilde{\mathbf{w}}_k}{dt} = \tilde{a}_k \boldsymbol{\mu}^+, \quad \frac{d\tilde{a}_k}{dt} = \langle \tilde{\mathbf{w}}_k, \boldsymbol{\mu}^+ \rangle; \\ \text{If } \bar{a}_k < 0: \quad & \frac{d\tilde{\mathbf{w}}_k}{dt} = \tilde{a}_k \boldsymbol{\mu}^-, \quad \frac{d\tilde{a}_k}{dt} = \langle \tilde{\mathbf{w}}_k, \boldsymbol{\mu}^- \rangle. \end{aligned}$$

Let $M_+ := \begin{bmatrix} \mathbf{0} & \boldsymbol{\mu}^+ \\ (\boldsymbol{\mu}^+)^\top & 0 \end{bmatrix}$ and $M_- := \begin{bmatrix} \mathbf{0} & \boldsymbol{\mu}^- \\ (\boldsymbol{\mu}^-)^\top & 0 \end{bmatrix}$. The largest eigenvalues for M_+ and M_- are $\lambda_0^+ := \|\boldsymbol{\mu}^+\|_2$ and $\lambda_0^- := \|\boldsymbol{\mu}^-\|_2$ respectively. Then the above linear ODE can be solved as

$$\text{If } \bar{a}_k > 0: \quad \begin{bmatrix} \tilde{\mathbf{w}}_k(T_0 + t) \\ \tilde{a}_k(T_0 + t) \end{bmatrix} = \exp(tM_+) \begin{bmatrix} \mathbf{w}_k(T_0) \\ a_k(T_0) \end{bmatrix}; \quad (64)$$

$$\text{If } \bar{a}_k < 0: \quad \begin{bmatrix} \tilde{\mathbf{w}}_k(T_0 + t) \\ \tilde{a}_k(T_0 + t) \end{bmatrix} = \exp(tM_-) \begin{bmatrix} \mathbf{w}_k(T_0) \\ a_k(T_0) \end{bmatrix}. \quad (65)$$

Lemma G.6. *Let $\tilde{\boldsymbol{\theta}}(t) = \tilde{\varphi}(\boldsymbol{\theta}(T_0), t - T_0)$. Then for all $T_0 \leq t \leq T_{\sigma_{\text{init}}}^\epsilon$, it holds that*

$$\|\tilde{\boldsymbol{\theta}}(t)\|_{\mathbf{M}} \leq \exp((t - T_0)\lambda_0^+) \|\tilde{\boldsymbol{\theta}}(T_0)\|_{\mathbf{M}}.$$

Proof. By Assumption A.3, we have $\lambda_0^+ > \lambda_0^-$. By definition and Cauchy-Schwartz inequality,

$$\left\| \frac{d\tilde{\mathbf{w}}_k}{dt} \right\|_2 \leq \lambda_0^+ |\tilde{a}_k|, \quad \left| \frac{d\tilde{a}_k}{dt} \right| \leq \lambda_0^+ \|\tilde{\mathbf{w}}_k\|_2.$$

So we have $\|\tilde{\boldsymbol{\theta}}(t)\|_{\mathbf{M}} \leq \|\boldsymbol{\theta}(T_0)\|_{\mathbf{M}} + \int_{T_0}^{T_{\sigma_{\text{init}}}^\epsilon} \lambda_0^+ \|\tilde{\boldsymbol{\theta}}(\tau)\|_{\mathbf{M}} d\tau$. Then we can finish the proof by Grönwall's inequality (12). $\square$

Lemma G.7. *For $\boldsymbol{\theta}(T_0)$ with $|a_k(T_0)| = \|\mathbf{w}_k(T_0)\|_2$ and $a_k(T_0)\mathbf{w}_k(T_0) \in \mathcal{C}^{\delta/3}$, we have*

$$\|\boldsymbol{\theta}(t) - \tilde{\varphi}(\boldsymbol{\theta}(T_0), t - T_0)\|_{\mathbf{M}} \leq \frac{4m\|\boldsymbol{\theta}(T_0)\|_{\mathbf{M}}^3}{\lambda_0^+} \exp(3\lambda_0(t - T_0)),$$

for all $T_0 \leq t \leq \frac{1}{\lambda_0^+} \ln \frac{\sqrt{\min\{\epsilon, \lambda_0^+\}}}{\sqrt{4m}\|\boldsymbol{\theta}(T_0)\|_{\mathbf{M}}}$.

Proof. Let $\tilde{\boldsymbol{\theta}}(t) = \tilde{\varphi}(\boldsymbol{\theta}(T_0), t - T_0)$. Let

$$t_0 := \min\{T_{\sigma_{\text{init}}}^\epsilon, \inf\{t \geq T_0 : \|\boldsymbol{\theta}(t)\|_{\mathbf{M}} \geq 2\|\boldsymbol{\theta}(T_0)\|_{\mathbf{M}} \exp(\lambda_0^+(t - T_0))\}\}.$$

and it holds that $\forall T_0 \leq t \leq t_0$, all neurons of $\tilde{\boldsymbol{\theta}}(t)$, $\boldsymbol{\theta}(t)$ are either in $\mathcal{C}^{\delta/3}$ or $-\mathcal{C}^{\delta/3}$, thus $\tilde{\boldsymbol{\theta}}(t)$, $\boldsymbol{\theta}(t)$ are in the same differentiable region of $\tilde{\mathcal{L}}$. By Corollary B.13, the following holds for a.e. $t \geq 0$,

$$\begin{aligned} \left\| \frac{d\boldsymbol{\theta}}{dt} - \frac{d\tilde{\boldsymbol{\theta}}}{dt} \right\|_{\mathbf{M}} &\leq \sup \left\{ \|\boldsymbol{\delta} - \nabla \tilde{\mathcal{L}}(\boldsymbol{\theta})\|_{\mathbf{M}} : \boldsymbol{\delta} \in \partial^\circ \mathcal{L}(\boldsymbol{\theta}) \right\} + \|\nabla \tilde{\mathcal{L}}(\boldsymbol{\theta}) - \nabla \tilde{\mathcal{L}}(\tilde{\boldsymbol{\theta}})\|_{\mathbf{M}} \\ &\leq m\|\boldsymbol{\theta}(t)\|_{\mathbf{M}}^3 + \lambda_0^+ \|\boldsymbol{\theta} - \tilde{\boldsymbol{\theta}}\|_{\mathbf{M}}. \end{aligned}$$

Then we can argue as the proof for Lemma D.2 to show that

$$\|\boldsymbol{\theta}(t) - \tilde{\boldsymbol{\theta}}(t)\|_{\mathbf{M}} \leq \frac{4m\|\boldsymbol{\theta}(T_0)\|_{\mathbf{M}}^3}{\lambda_0} \exp(3\lambda_0^+(t - T_0))$$

for all $t \in [T_0, t_0]$. If $t_0 < T_0 + \frac{1}{2\lambda_0^+} \ln \frac{\min\{\lambda_0^+, \epsilon\}}{4m\|\boldsymbol{\theta}(T_0)\|_{\mathbf{M}}^2}$, then for all $T_0 \leq t \leq t_0$, we have

$$\|\boldsymbol{\theta}(t)\|_{\mathbf{M}} \leq \|\boldsymbol{\theta}(T_0)\|_{\mathbf{M}} \sqrt{\frac{\min\{\lambda_0^+, \epsilon\}}{4m\|\boldsymbol{\theta}(T_0)\|_{\mathbf{M}}^2}} < \sqrt{\frac{\epsilon}{m}},$$

which implies that $t_0 < T_{\sigma_{\text{init}}}^\epsilon$ by definition of $T_{\sigma_{\text{init}}}^\epsilon$. Moreover,

$$\begin{aligned}\|\boldsymbol{\theta}(t)\|_{\text{M}} &\leq \|\tilde{\boldsymbol{\theta}}(t)\|_{\text{M}} + \frac{4m\|\boldsymbol{\theta}(T_0)\|_{\text{M}}^3}{\lambda_0^+} \exp(3\lambda_0^+(t - T_0)) \\ &\leq \|\tilde{\boldsymbol{\theta}}(t)\|_{\text{M}} + \frac{4m\|\boldsymbol{\theta}(T_0)\|_{\text{M}}^2}{\lambda_0^+} \exp(2\lambda_0^+(t_0 - T_0)) \cdot \|\boldsymbol{\theta}(T_0)\|_{\text{M}} \exp(\lambda_0^+(t - T_0)) \\ &< \|\tilde{\boldsymbol{\theta}}(t)\|_{\text{M}} + \|\boldsymbol{\theta}(T_0)\|_{\text{M}} \exp(\lambda_0^+(t - T_0)).\end{aligned}$$

By Lemma G.6, $\|\tilde{\boldsymbol{\theta}}(t)\|_{\text{M}} \leq \|\boldsymbol{\theta}(T_0)\|_{\text{M}} \exp(\lambda_0^+(t - T_0))$. So $\|\boldsymbol{\theta}(t)\|_{\text{M}} < 2\|\boldsymbol{\theta}(T_0)\|_{\text{M}} \exp(\lambda_0^+(t - T_0))$ for all $T_0 \leq t \leq t_0$, which contradicts to the definition of t_0 . Therefore, $t_0 \geq \frac{1}{2\lambda_0^+} \ln \frac{\min\{\epsilon, \lambda_0^+\}}{4m\|\boldsymbol{\theta}(T_0)\|_{\text{M}}^2} = \frac{1}{\lambda_0^+} \ln \frac{\sqrt{\min\{\epsilon, \lambda_0^+\}}}{\sqrt{4m}\|\boldsymbol{\theta}(T_0)\|_{\text{M}}}$. $\square$

Proof for Lemma G.5. Let $r_{\text{max}} := \frac{\sqrt{\min\{\lambda_0^+, \epsilon\}}}{2}$ and $C_1 := \sigma_{\text{init}} r_{\text{max}}^{-3}$. We only need to prove the statements for all $\sigma_{\text{init}} < \sigma_{\text{init}}^{\text{max}} = C_1 r_{\text{max}}^3$.

We fix a pair of $\sigma_{\text{init}} < \sigma_{\text{init}}^{\text{max}}$ and $r < r_{\text{max}}$ satisfying $\sigma_{\text{init}} < C_1 r^3$. For convenience, we use $\bar{\mathbf{b}}, T_1$ to denote $\bar{\mathbf{b}}(\sigma_{\text{init}}), T_1(\sigma_{\text{init}}, r)$ for short.

Let $\boldsymbol{\theta}(t) = \varphi(\sigma_{\text{init}} \bar{\boldsymbol{\theta}}, t)$. It is easy to see that the prerequisites of Lemma G.4 are satisfied with probability 1. Below we only focus on the case where the prerequisites of Lemma G.4 are satisfied. Let $T_0, \sigma_{\text{init}}^{\text{max}}, A, B$ be the constants from Lemma G.4. Let $\tilde{\boldsymbol{\theta}}(t) = \tilde{\varphi}(\boldsymbol{\theta}(T_0), t - T_0)$.

For $\bar{a}_k > 0$, we define

$$\bar{b}_k := \frac{\langle \mathbf{w}_k(T_0), \bar{\boldsymbol{\mu}}^+ \rangle + a_k(T_0)}{2\sqrt{m}\sigma_{\text{init}}},$$

and for $\bar{a}_k < 0$, we define

$$\bar{b}_k := \frac{\langle \mathbf{w}_k(T_0), \bar{\boldsymbol{\mu}}^- \rangle + a_k(T_0)}{2(\sqrt{m}\sigma_{\text{init}})^{1-\kappa}}.$$

$r \leq r_{\text{max}}$ and $\sigma_{\text{init}} < \sigma_{\text{init}}^{\text{max}}$.

Proof for Item 1. By Lemma G.7, we have

$$\|\boldsymbol{\theta}(T_0 + T_1) - \tilde{\boldsymbol{\theta}}(T_0 + T_1)\|_{\text{M}} \leq \frac{4m\|\boldsymbol{\theta}(T_0)\|_{\text{M}}^3}{\lambda_0^+} \exp(3\lambda_0^+ T_1) = \frac{4\|\boldsymbol{\theta}(T_0)\|_{\text{M}}^3}{\lambda_0^+ \sqrt{m}\sigma_{\text{init}}^3} r^3. \quad (66)$$

Now we turn to characterize $\tilde{\boldsymbol{\theta}}(T_0 + T_1)$. Note that $\bar{\boldsymbol{\mu}}_2^+ := \frac{1}{\sqrt{2}}[\bar{\boldsymbol{\mu}}^+, 1]^\top$ and $\bar{\boldsymbol{\mu}}_2^- := \frac{1}{\sqrt{2}}[\bar{\boldsymbol{\mu}}^-, 1]^\top$ are the top eigenvectors of $\mathbf{M}_+$ and $\mathbf{M}_-$ respectively. Let $\kappa := 1 - \frac{\|\boldsymbol{\mu}^-\|_2}{\|\boldsymbol{\mu}^+\|_2}$. Recall that $T_1 := \frac{1}{\lambda_0^+} \ln \frac{r}{\sqrt{m}\sigma_{\text{init}}}$. Then for $\bar{a}_k > 0$, we have

$$\exp(T_1 \lambda_0^+) \bar{\boldsymbol{\mu}}_2^+ (\bar{\boldsymbol{\mu}}_2^+)^{\top} \begin{bmatrix} \mathbf{w}_k(T_0) \\ a_k(T_0) \end{bmatrix} = \left(\frac{r}{\sqrt{m}\sigma_{\text{init}}} \right) \bar{\boldsymbol{\mu}}_2^+ (\bar{\boldsymbol{\mu}}_2^+)^{\top} \begin{bmatrix} \mathbf{w}_k(T_0) \\ a_k(T_0) \end{bmatrix} = r \bar{b}_k \begin{bmatrix} \bar{\boldsymbol{\mu}}^+ \\ 1 \end{bmatrix},$$

where the last equality is by definition of $\bar{b}_k$. Similarly for $\bar{a}_k < 0$, we have

$$\exp(T_1 \lambda_0^-) \bar{\boldsymbol{\mu}}_2^- (\bar{\boldsymbol{\mu}}_2^-)^{\top} \begin{bmatrix} \mathbf{w}_k(T_0) \\ a_k(T_0) \end{bmatrix} = \left(\frac{r}{\sqrt{m}\sigma_{\text{init}}} \right)^{1-\kappa} \bar{\boldsymbol{\mu}}_2^- (\bar{\boldsymbol{\mu}}_2^-)^{\top} \begin{bmatrix} \mathbf{w}_k(T_0) \\ a_k(T_0) \end{bmatrix} = r^{1-\kappa} \bar{b}_k \begin{bmatrix} \bar{\boldsymbol{\mu}}^- \\ 1 \end{bmatrix}.$$

Combining these with (64) and (65), then for $\bar{a}_k > 0$ we have

$$\begin{aligned}\left\| \begin{bmatrix} \tilde{\mathbf{w}}_k(T_0 + t) \\ \tilde{a}_k(T_0 + t) \end{bmatrix} - r \bar{b}_k \begin{bmatrix} \bar{\boldsymbol{\mu}}^+ \\ 1 \end{bmatrix} \right\|_2 &\leq \left\| (\exp(T_1 \mathbf{M}_+) - \exp(T_1 \lambda_0^+) \bar{\boldsymbol{\mu}}_2^+ (\bar{\boldsymbol{\mu}}_2^+)^{\top}) \begin{bmatrix} \mathbf{w}_k(T_0) \\ a_k(T_0) \end{bmatrix} \right\|_2 \\ &\leq \left\| \begin{bmatrix} \mathbf{w}_k(T_0) \\ a_k(T_0) \end{bmatrix} \right\|_2 \\ &\leq \sqrt{2} \|\boldsymbol{\theta}(T_0)\|_{\text{M}}.\end{aligned}$$

and for $\bar{a}_k < 0$ we have

$$\begin{aligned} \left\| \begin{bmatrix} \tilde{\mathbf{w}}_k(T_0 + t) \\ \tilde{a}_k(T_0 + t) \end{bmatrix} - r^{1-\kappa} \bar{b}_k \begin{bmatrix} \bar{\boldsymbol{\mu}}^- \\ 1 \end{bmatrix} \right\|_2 &\leq \left\| (\exp(T_1 \mathbf{M}_-) - \exp(T_1 \lambda_0^-) \bar{\boldsymbol{\mu}}_2^- (\bar{\boldsymbol{\mu}}_2^-)^\top) \begin{bmatrix} \mathbf{w}_k(T_0) \\ a_k(T_0) \end{bmatrix} \right\|_2 \\ &\leq \left\| \begin{bmatrix} \mathbf{w}_k(T_0) \\ a_k(T_0) \end{bmatrix} \right\|_2 \\ &\leq \sqrt{2} \|\boldsymbol{\theta}(T_0)\|_{\text{M}}. \end{aligned}$$

Then by definition of $\Delta\boldsymbol{\theta}$ and (66), we have

$$\|\Delta\boldsymbol{\theta}\|_{\text{M}} \leq \frac{4\|\boldsymbol{\theta}(T_0)\|_{\text{M}}^3}{\lambda_0^+ \sqrt{m} \sigma_{\text{init}}^3} r^3 + \sqrt{2} \|\boldsymbol{\theta}(T_0)\|_{\text{M}}.$$

Applying the upper bound $\|\mathbf{w}_k(T_0)\|_2 \leq B\sigma_{\text{init}}$ from Lemma G.4, we then have

$$\|\Delta\boldsymbol{\theta}\|_{\text{M}} \leq \frac{4B^3 \sigma_{\text{init}}^3}{\lambda_0^+ \sqrt{m} \sigma_{\text{init}}^3} r^3 + \sqrt{2} B \sigma_{\text{init}} = \frac{4B^3}{\lambda_0^+ \sqrt{m}} r^3 + \sqrt{2} B \sigma_{\text{init}},$$

Finally, recalling that $\sigma_{\text{init}} \leq C_1 r^3$, we can conclude that $\|\Delta\boldsymbol{\theta}\|_{\text{M}} \leq C_2 r^3$, where $C_2 := \frac{4B^3}{\lambda_0^+ \sqrt{m}} + \sqrt{2} B C_1$.

Item 2. Now it only remains to lower and upper bound $|\bar{b}_k|$. By Lemma G.4, $a_k(T_0) \mathbf{w}_k(T_0) \in \mathcal{C}^{\delta/3}$. Then $\text{sgn}(a_k(T_0)) \langle \mathbf{w}_k(T_0), \bar{\boldsymbol{\mu}}^+ \rangle \geq 0$ and thus

$$\begin{aligned} \text{if } \bar{a}_k > 0 : \quad &\langle \mathbf{w}_k(T_0), \bar{\boldsymbol{\mu}}^+ \rangle + a_k(T_0) \in [\|\mathbf{w}_k(T_0)\|_2, 2 \cdot \|\mathbf{w}_k(T_0)\|_2] \subseteq [A\sigma_{\text{init}}, 2B\sigma_{\text{init}}]; \\ \text{if } \bar{a}_k < 0 : \quad &\langle \mathbf{w}_k(T_0), \bar{\boldsymbol{\mu}}^- \rangle + a_k(T_0) \in [-2 \cdot \|\mathbf{w}_k(T_0)\|_2, -\|\mathbf{w}_k(T_0)\|_2] \subseteq [-2B\sigma_{\text{init}}, -A\sigma_{\text{init}}]. \end{aligned}$$

Then for every $\bar{b}_k$,

$$\begin{aligned} \text{if } \bar{a}_k > 0 : \quad &|\bar{b}_k| = \frac{|\langle \mathbf{w}_k(T_0), \bar{\boldsymbol{\mu}}^+ \rangle + a_k(T_0)|}{2\sqrt{m} \sigma_{\text{init}}} \in \left[\frac{A}{2\sqrt{m}}, \frac{B}{\sqrt{m}} \right]; \\ \text{if } \bar{a}_k < 0 : \quad &|\bar{b}_k| = \frac{|\langle \mathbf{w}_k(T_0), \bar{\boldsymbol{\mu}}^- \rangle + a_k(T_0)|}{2(\sqrt{m} \sigma_{\text{init}})^{1-\kappa}} \in \left[\frac{\sigma_{\text{init}}^\kappa A}{2m^{(1-\kappa)/2}}, \frac{\sigma_{\text{init}}^\kappa B}{m^{(1-\kappa)/2}} \right]. \end{aligned}$$

Letting $\bar{A} := \frac{A}{2m^{(1-\kappa)/2}}$ and $\bar{B} := \frac{B}{m^{(1-\kappa)/2}}$ completes the proof. $\square$

G.4 Phase II

As shown in our analysis for Phase I, if the initialization scale is small, the weight vectors of neurons with $\bar{a}_k > 0$ move towards the direction of $\bar{\boldsymbol{\mu}}^+$, and all the other neurons are negligible. Now we show that the dynamic of $\boldsymbol{\theta}(t)$ is close to that of a one-neuron dynamic in a similar manner as we do for the symmetric case.

First we slightly extend the definition of embedding. For $\hat{\boldsymbol{\theta}} = (\hat{\mathbf{w}}_1, \hat{\mathbf{w}}_2, \hat{a}_1, \hat{a}_2)$ and an embedding vector $\mathbf{b} \in \mathbb{R}^m$, we say that $\mathbf{b}$ is compatible with $\hat{\boldsymbol{\theta}}$ if the following holds:

1. If $b_+ = 0$, then $\|\hat{\mathbf{w}}_1\|_2 = |\hat{a}_1| = 0$;
2. If $b_- = 0$, then $\|\hat{\mathbf{w}}_2\|_2 = |\hat{a}_2| = 0$.

When $\mathbf{b}$ is compatible with $\hat{\boldsymbol{\theta}}$, we define the (exact) embedding from two-neuron into m -neuron neural nets as $\pi_{\mathbf{b}}(\hat{\boldsymbol{\theta}}) := (\mathbf{w}_1, \dots, \mathbf{w}_m, a_1, \dots, a_m)$, where

$$a_k = \begin{cases} \frac{b_k}{b_+} \hat{a}_1, & \text{if } b_k > 0 \\ \frac{b_k}{b_-} \hat{a}_2, & \text{if } b_k < 0, \\ \mathbf{0}, & \text{if } b_k = 0 \end{cases}, \quad \mathbf{w}_k = \begin{cases} \frac{b_k}{b_+} \hat{\mathbf{w}}_1, & \text{if } b_k > 0 \\ \frac{b_k}{b_-} \hat{\mathbf{w}}_2, & \text{if } b_k < 0. \\ \mathbf{0}, & \text{if } b_k = 0 \end{cases}.$$

One can easily show that Lemma 5.3 continue to hold when $\mathbf{b}$ is compatible with $\hat{\boldsymbol{\theta}}$.

Lemma G.8. Let $\bar{\mathbf{b}}(\sigma_{\text{init}})$ be the same vector as in the statement of Lemma G.5. Let $T_{12}(\sigma_{\text{init}}) := T_0 + \frac{1}{\lambda_0^+} \ln \frac{1}{\sqrt{m}\sigma_{\text{init}}}$ and $T_2(r) := \frac{1}{\lambda_0^+} \ln \frac{1}{r}$. For width $m \geq 1$, the following statements hold with probability $1 - 2^{-m}$ over the random draw of $\bar{\boldsymbol{\theta}}_0 = (\bar{\mathbf{w}}_1, \dots, \bar{\mathbf{w}}_m, \bar{a}_1, \dots, \bar{a}_m) \sim \mathcal{D}_{\text{init}}(1)$. Let $\sigma_1, \sigma_2, \dots$ be any sequence of initialization scales so that σ_j converges to 0 as $j \rightarrow +\infty$ and the limit $\hat{\mathbf{b}} := \lim_{j \rightarrow +\infty} \bar{\mathbf{b}}(\sigma_j)$ exists.

1. $\hat{b}_+ > 0$ and $\hat{b}_- = 0$;
2. For the two-neuron dynamics starting with rescaled initialization in the direction of $\hat{\boldsymbol{\theta}} := (\hat{b}_+ \bar{\boldsymbol{\mu}}^+, \mathbf{0}, \hat{b}_+, 0)$, the following limit exists for all $t \geq 0$,

$$\tilde{\boldsymbol{\theta}}(t) := \lim_{r \rightarrow 0} \varphi\left(r\hat{\boldsymbol{\theta}}, T_2(r) + t\right) \neq \mathbf{0}; \quad (67)$$

3. For the m -neuron dynamics of $\boldsymbol{\theta}_j(t)$ with initialization scale $\sigma_{\text{init}} = \sigma_j$, the following holds for all $t \geq 0$,

$$\lim_{j \rightarrow \infty} \boldsymbol{\theta}_j(T_{12}(\sigma_j) + t) = \pi_{\hat{\mathbf{b}}}(\tilde{\boldsymbol{\theta}}(t)). \quad (68)$$

Proof. The proof is similar to Lemma 5.4 for the symmetric case. Apply Theorem E.4 and then the lemma is straightforward. $\square$

G.5 Phase III

In Phase III, we show that the dynamic of $\boldsymbol{\theta}(t)$ converges to the same classifier as the one-neuron dynamic.

Let $\mathcal{S}^+ := \arg \min_{i \in [n]} \{y_i \langle \mathbf{w}^+, \mathbf{x}_i^+ \rangle\} \subseteq [n]$. Let $\Delta^{h-1} = \{\mathbf{p} \in \mathbb{R}^h : \sum_{i \in [h]} p_i = 1, p_i \geq 1\}$ be the probability simplex. Let $\Lambda^+ := \{\boldsymbol{\lambda} \in \Delta^{n-1} : \lambda_i = 0, \forall i \notin \mathcal{S}^+\}$.

The theorem below characterizes the solution found by the one-neuron dynamic.

Theorem G.9. Under Assumption 3.2, for $m = 1$, if initially $a_1 = \|\mathbf{w}_1\|_2$, $\langle \mathbf{w}_1, \mathbf{w}^* \rangle > 0$, then $\boldsymbol{\theta}(t)$ directionally converges to the following global-max-margin direction,

$$\lim_{t \rightarrow +\infty} \frac{\boldsymbol{\theta}(t)}{\|\boldsymbol{\theta}(t)\|_2} = \frac{1}{\sqrt{2}}(\mathbf{w}^+, 1).$$

Proof. By Theorem B.19, $\mathcal{L}(\boldsymbol{\theta}(t)) \rightarrow 0$. Then by Theorem 3.1, $\frac{\boldsymbol{\theta}(t)}{\|\boldsymbol{\theta}(t)\|_2}$ converges along a KKT-margin direction. Combining this with Lemma B.17, we know that this direction must have the form $\frac{1}{\sqrt{2}}(\bar{\mathbf{w}}, 1)$ for some $\bar{\mathbf{w}} \in \mathbb{S}^{d-1}$.

By Definition B.8, $y_i \cdot \frac{1}{2} \phi(\langle \bar{\mathbf{w}}, \mathbf{x}_i \rangle) > 0$ and $\bar{\mathbf{w}}$ can be expressed by a convex combination of $y_i \phi'(\langle \bar{\mathbf{w}}, \mathbf{x}_i \rangle) \mathbf{x}_i$ among $i \in \arg \min \{\frac{1}{2} \phi(\langle \bar{\mathbf{w}}, \mathbf{x}_i \rangle)\}$. Equivalently, we know that $y_i \langle \bar{\mathbf{w}}, \mathbf{x}_i^+ \rangle$ and $\bar{\mathbf{w}}$ can be expressed by a convex combination of $y_i \mathbf{x}_i^+$ among $i \in \mathcal{S}^+$. Then the only possibility is $\bar{\mathbf{w}} = \mathbf{w}^+$. $\square$

Now we turn to analyze the trajectory of $\boldsymbol{\theta}(t)$ on m -neuron neural net. First we prove the following lemma, then we prove Theorem G.11 for local-max-margin directions.

Lemma G.10. Let $\Theta_- := \{\boldsymbol{\theta} = (\mathbf{w}_1, \dots, \mathbf{w}_m, a_1, \dots, a_m) : m \geq 1, a_k \leq 0\}$. Then we have the following characterization for the global maximum of the normalized margin on the dataset $\{(\mathbf{x}_i, y_i) : i \in \mathcal{S}^+\}$:

$$\sup_{\boldsymbol{\theta} \in \Theta_-} \left\{ \frac{\min_{i \in \mathcal{S}^+} q_i(\boldsymbol{\theta})}{\|\boldsymbol{\theta}\|_2^2} \right\} = \inf_{\boldsymbol{\lambda} \in \Lambda^+} \sup_{\mathbf{u} \in \mathbb{S}^{d-1}} \left\{ -\frac{1}{2} \sum_{i \in [n]} \lambda_i y_i \phi(\langle \mathbf{u}, \mathbf{x}_i \rangle) \right\}$$

Proof. The proof is inspired by Chizat and Bach (2020, Proposition 12). By Lemma B.9, the maximum normalized margin is attained when $|a_k| = \|\mathbf{w}_k\|_2$ for all $k \in [m]$. Note that we can

rewrite each neuron output $a_k \phi(\langle \mathbf{w}_k, \mathbf{x}_i \rangle)$ as $-a_k^2 \phi(\langle \mathbf{w}_k / \|\mathbf{w}_k\|_2, \mathbf{x}_i \rangle)$ for any such solution, and it is easy to see $\sum_{k \in [m]} a_k^2 = \frac{1}{2} \|\boldsymbol{\theta}\|_2^2$. Let $\mathcal{V}$ be the set of probability distributions supported on finitely many points of $\mathbb{S}^{d-1}$. Then

$$\begin{aligned} \sup_{\boldsymbol{\theta} \in \Theta_-} \left\{ \frac{\min_{i \in \mathcal{S}^+} q_i(\boldsymbol{\theta})}{\|\boldsymbol{\theta}\|_2^2} \right\} &= \sup_{\substack{m \geq 1, \mathbf{p} \in \Delta^{m-1} \\ \mathbf{u}_1, \dots, \mathbf{u}_m \in \mathbb{S}^{d-1}}} \min_{i \in \mathcal{S}^+} \left\{ -\frac{1}{2} y_i \sum_{k \in [m]} p_k \phi(\langle \mathbf{u}_k, \mathbf{x}_i \rangle) \right\} \\ &= \sup_{\nu \in \mathcal{V}} \min_{i \in \mathcal{S}^+} \mathbb{E}_{\mathbf{u} \sim \nu} \left[-\frac{1}{2} y_i \phi(\langle \mathbf{u}, \mathbf{x}_i \rangle) \right]. \end{aligned}$$

By minimax theorem, we can swap the order between sup and min in the following way:

$$\begin{aligned} \sup_{\nu \in \mathcal{V}} \min_{i \in \mathcal{S}^+} \mathbb{E}_{\mathbf{u} \sim \nu} \left[-\frac{1}{2} y_i \phi(\langle \mathbf{u}, \mathbf{x}_i \rangle) \right] &= \sup_{\nu \in \mathcal{V}} \inf_{\boldsymbol{\lambda} \in \Lambda^+} \mathbb{E}_{\mathbf{u} \sim \nu} \left[-\frac{1}{2} \sum_{i \in \mathcal{S}^+} \lambda_i y_i \phi(\langle \mathbf{u}, \mathbf{x}_i \rangle) \right] \\ &= \inf_{\boldsymbol{\lambda} \in \Lambda^+} \sup_{\nu \in \mathcal{V}} \mathbb{E}_{\mathbf{u} \sim \nu} \left[-\frac{1}{2} \sum_{i \in \mathcal{S}^+} \lambda_i y_i \phi(\langle \mathbf{u}, \mathbf{x}_i \rangle) \right] \\ &= \inf_{\boldsymbol{\lambda} \in \Lambda^+} \sup_{\mathbf{u} \in \mathbb{S}^{d-1}} \left\{ -\frac{1}{2} \sum_{i \in \mathcal{S}^+} \lambda_i y_i \phi(\langle \mathbf{u}, \mathbf{x}_i \rangle) \right\}, \end{aligned}$$

which proves the claim. $\square$

Theorem G.11. Let $\hat{\boldsymbol{\theta}} := (\frac{1}{\sqrt{2}} \mathbf{w}^+, \frac{1}{\sqrt{2}}, \mathbf{0}, 0)$ and P be a non-empty subset of $[m]$. Let $\bar{\mathcal{Q}}$ be the following subset of $\mathbb{S}^{D-1}$:

$$\bar{\mathcal{Q}} := \{\boldsymbol{\theta} = (\mathbf{w}_1, \dots, \mathbf{w}_m, a_1, \dots, a_m) \in \mathbb{S}^{D-1} : a_k \geq 0 \text{ for all } k \in P \text{ and } a_k \leq 0 \text{ otherwise}\}.$$

For any embedding vect $\mathbf{b}$ be an embedding vector satisfying the following:

- $\mathbf{b}$ is compatible with $\hat{\boldsymbol{\theta}}$;
- $b_k > 0$ for all $k \in P$;
- $b_k = 0$ for all $k \notin P$;

the following statements are true under Assumption A.5,

1. $\pi_{\mathbf{b}}(\hat{\boldsymbol{\theta}})$ is a local maximizer of $\gamma(\boldsymbol{\theta})$ among $\boldsymbol{\theta} \in \bar{\mathcal{Q}}$;
2. If $\boldsymbol{\theta} \in \bar{\mathcal{Q}}$ has the same normalized margin as $\pi_{\mathbf{b}}(\hat{\boldsymbol{\theta}})$ and $\boldsymbol{\theta}$ is sufficiently close to $\pi_{\mathbf{b}}(\hat{\boldsymbol{\theta}})$, then $f_{\boldsymbol{\theta}}(\mathbf{x}) = f_{\hat{\boldsymbol{\theta}}}(\mathbf{x})$ for all $\mathbf{x} \in \mathbb{R}^d$.

Proof. It is easy to see that $\pi_{\mathbf{b}}(\hat{\boldsymbol{\theta}})$ is a KKT-margin direction with $\gamma(\pi_{\mathbf{b}}(\hat{\boldsymbol{\theta}})) = \gamma(\hat{\boldsymbol{\theta}}) = \frac{1}{2} \gamma^+$. Also, $\arg \min_{i \in [n]} \{q_i(\pi_{\mathbf{b}}(\hat{\boldsymbol{\theta}}))\} = \arg \min_{i \in [n]} \{q_i(\hat{\boldsymbol{\theta}})\} = \mathcal{S}^+$. Let $\epsilon > 0$ be a small constant such that the following holds whenever $\|\boldsymbol{\theta} - \pi_{\mathbf{b}}(\hat{\boldsymbol{\theta}})\|_{\text{M}} < \epsilon$:

1. $\text{sgn}(\langle \mathbf{w}_k, \mathbf{x}_i \rangle) = \text{sgn}(\langle \mathbf{w}^+, \mathbf{x}_i \rangle)$ for all $i \in [n]$ and for all $k \in P$;
2. $\arg \min_{i \in [n]} \{q_i(\boldsymbol{\theta})\} \subseteq \mathcal{S}^+$.

Let $\boldsymbol{\theta} \in \bar{\mathcal{Q}}$ be any parameter satisfying $\|\boldsymbol{\theta} - \pi_{\mathbf{b}}(\hat{\boldsymbol{\theta}})\|_{\text{M}} < \epsilon$. We can decompose $\boldsymbol{\theta}$ into $\boldsymbol{\theta}^+ + \boldsymbol{\theta}^-$, where $\boldsymbol{\theta}^+ = (\mathbf{w}_1^+, \dots, \mathbf{w}_m^+, a_1^+, \dots, a_m^+)$, $\boldsymbol{\theta}^- = (\mathbf{w}_1^-, \dots, \mathbf{w}_m^-, a_1^-, \dots, a_m^-)$, and

$$\mathbf{w}_k^+ = \mathbb{1}_{[k \in P]} \mathbf{w}_k, a_k^+ = \mathbb{1}_{[k \in P]} a_k, \quad \mathbf{w}_k^- = \mathbb{1}_{[k \notin P]} \mathbf{w}_k, a_k^- = \mathbb{1}_{[k \notin P]} a_k.$$

Let $r_+ = \|\boldsymbol{\theta}^+\|_2$ and $r_- = \|\boldsymbol{\theta}^-\|_2$. Define $\bar{\boldsymbol{\theta}}^+$ and $\bar{\boldsymbol{\theta}}^-$ to be two unit-norm parameters so that $\boldsymbol{\theta}^+ = r_+ \bar{\boldsymbol{\theta}}^+$, $\boldsymbol{\theta}^- = r_- \bar{\boldsymbol{\theta}}^-$. Then we have

$$\gamma(\boldsymbol{\theta}) = \min_{i \in \mathcal{S}^+} \{q_i(\boldsymbol{\theta})\} = \min_{i \in \mathcal{S}^+} \{q_i(\boldsymbol{\theta}^+) + q_i(\boldsymbol{\theta}^-)\} = \min_{i \in \mathcal{S}^+} \{r_+^2 q_i(\bar{\boldsymbol{\theta}}^+) + r_-^2 q_i(\bar{\boldsymbol{\theta}}^-)\}$$

Note that $r_+^2 + r_-^2 = 1$. By minimax theorem (similar to Lemma G.10),

$$\min_{i \in S^+} \{r_+^2 q_i(\bar{\theta}^+) + r_-^2 q_i(\bar{\theta}^-)\} \leq \min_{\lambda \in \Lambda^+} \max \left\{ \sum_{i \in S^+} \lambda_i q_i(\bar{\theta}^+), \sum_{i \in S^+} \lambda_i q_i(\bar{\theta}^-) \right\}.$$

By definition of $\mathbf{w}^+$ and KKT conditions, we can find $\lambda^* \in \Lambda^+$ so that $\sum_{i \in S^+} \lambda_i^* \mathbf{x}_i^+ = \gamma^+ \mathbf{w}^+$. Letting $\lambda = \lambda^*$ for the above inequality, we can obtain

$$\gamma(\theta) \leq \max \left\{ \sum_{i \in S^+} \lambda_i^* q_i(\bar{\theta}^+), \sum_{i \in S^+} \lambda_i^* q_i(\bar{\theta}^-) \right\}.$$

We only need to prove that both $\sum_{i \in S^+} \lambda_i^* q_i(\bar{\theta}^+)$ and $\sum_{i \in S^+} \lambda_i^* q_i(\bar{\theta}^-)$ are no more than $\frac{1}{2} \gamma^+$. Note that combining Assumption A.5 and Lemma G.10 directly implies that $\sum_{i \in S^+} \lambda_i^* q_i(\bar{\theta}^-) < \frac{1}{2} \gamma^+$. Now we focus on $\sum_{i \in S^+} \lambda_i^* q_i(\bar{\theta}^+)$.

According to our choice of ϵ , we have $a_k \phi(\langle \mathbf{w}_k, \mathbf{x}_i^+ \rangle) = \langle a_k \mathbf{w}_k, \mathbf{x}_i^+ \rangle$. For $\sum_{i \in S^+} \lambda_i^* q_i(\bar{\theta}^+)$, we have

$$\begin{aligned} \sum_{i \in S^+} \lambda_i^* q_i(\bar{\theta}^+) &= \sum_{i \in S^+} \lambda_i^* y_i \sum_{k \in P} \langle a_k \mathbf{w}_k, \mathbf{x}_i^+ \rangle = \sum_{k \in P} a_k \left\langle \mathbf{w}_k, \sum_{i \in S^+} \lambda_i^* y_i \mathbf{x}_i^+ \right\rangle \\ &= \sum_{k \in P} a_k \langle \mathbf{w}_k, \gamma^+ \mathbf{w}^+ \rangle. \end{aligned}$$

By Cauchy-Schwartz inequality,

$$\sum_{k \in P} a_k \langle \mathbf{w}_k, \gamma^+ \mathbf{w}^+ \rangle \leq \sqrt{\sum_{k \in P} a_k^2} \cdot \sqrt{\sum_{k \in P} \langle \mathbf{w}_k, \gamma^+ \mathbf{w}^+ \rangle^2} \leq \frac{1}{\sqrt{2}} \cdot \frac{1}{\sqrt{2}} \gamma^+ = \frac{1}{2} \gamma^+.$$

This proves that $\sum_{i \in S^+} \lambda_i^* q_i(\bar{\theta}^+) \leq \frac{1}{2} \gamma^+$, and thus $\gamma(\theta) \leq \frac{1}{2} \gamma^+ = \gamma(\hat{\theta})$. Therefore Item 1 is true.

For Item 2, we only need to note that the equality in $\gamma(\theta) \leq \frac{1}{2} \gamma^+$ only holds if $r_- = 0$ and $\mathbf{w}_k = a_k \mathbf{w}^+$ for all $k \in P$, so f_θ represents the same function as $f_{\hat{\theta}}$. $\square$

For proving Theorem A.7, we only need to show this:

Theorem G.12. *For any sequence of $\sigma_1, \sigma_2, \dots$ converging to 0, there is a subsequence $\sigma_{p_1}, \sigma_{p_2}, \dots$ and a constant $\sigma_{\text{init}}^{\max}$ such that Theorem A.7 holds for $\sigma_{\text{init}} = \sigma_{p_i}$ as long as $\sigma_{p_i} < \sigma_{\text{init}}^{\max}$.*

Proof for Theorem A.7. Assume to the contrary that Theorem A.7 does not hold. Then there exists $\mathbf{x} \in \mathbb{R}^d$ and a sequence of initialization scales $\sigma_1, \sigma_2, \dots$ converging to 0 such that $f^\infty(\mathbf{x}) \neq \frac{1}{2} \phi(\langle \mathbf{w}^+, \mathbf{x} \rangle)$ for any σ_j . However, by Theorem G.12, we can find a subsequence $\sigma_{p_1}, \sigma_{p_2}, \dots$ and a constant $\sigma_{\text{init}}^{\max}$ such that $\frac{1}{2} \phi(\langle \mathbf{w}^+, \mathbf{x} \rangle)$ holds for σ_{p_i} as long as $\sigma_{p_i} < \sigma_{\text{init}}^{\max}$, contradiction. $\square$

Proof for Theorem G.12. With probability 1 over the random draw of $\bar{\theta}_0 \sim \mathcal{D}_{\text{init}}(1)$, by Lemma G.5, the prerequisites of Lemma G.4 hold and we can find a subsequence of initialization scales $\sigma_{p_1}, \sigma_{p_2}, \dots$ so that the limit $\hat{\mathbf{b}} := \lim_{j \rightarrow +\infty} \bar{\mathbf{b}}(\sigma_{p_j})$ exists.

Let $\theta_j(t) = \varphi(\sigma_{p_j} \bar{\theta}_0, t)$. By Lemma G.8, with probability $1 - 2^{-m}$, $\lim_{j \rightarrow \infty} \theta_j(T_{12}(\sigma_{p_j}) + t) = \pi_{\hat{\mathbf{b}}}(\bar{\theta}(t))$. By Theorem G.9, $\lim_{t \rightarrow +\infty} \frac{\bar{\theta}(t)}{\|\bar{\theta}(t)\|_2} = \frac{1}{\sqrt{2}}(\mathbf{w}^+, \mathbf{0}, 1, 0) =: \tilde{\theta}_\infty$. Then we can argue in a similar way as Theorem 4.3 to show that for any $\epsilon > 0$ and $\rho > 0$, we can choose a time $t_1 \in \mathbb{R}$ such that $\left\| \frac{\theta_j(T_{12}(\sigma_{p_j}) + t_1)}{\|\theta_j(T_{12}(\sigma_{p_j}) + t_1)\|_2} - \pi_{\hat{\mathbf{b}}}(\tilde{\theta}_\infty) \right\|_2$ and $\|\theta_j(T_{12}(\sigma_{p_j}) + t_1)\|_2 \geq \rho$ for σ_{p_j} small enough.

By Corollary B.18, the trajectory of gradient flow starting with $\sigma_{p_j} \bar{\theta}_0$ lies in the set $\mathcal{Q} := \{\theta : a_k \bar{a}_k \geq 0 \text{ for all } k \in [m]\}$ for all $j \geq 1$, that is, every a_k has the same sign as its initial value during training. By a variant of Theorem 5.6, there exists $\sigma_{\text{init}}^{\max}$ such that for all $\sigma_{p_j} < \sigma_{\text{init}}^{\max}$, $\frac{\theta(t)}{\|\theta(t)\|_2} \rightarrow \bar{\theta} \in \mathcal{Q}$, where $\gamma(\bar{\theta}) = \gamma(\pi_{\hat{\mathbf{b}}}(\tilde{\theta}_\infty))$ and $\|\bar{\theta} - \pi_{\hat{\mathbf{b}}}(\tilde{\theta}_\infty)\|_2 \leq \delta$. Applying Theorem G.11 proves that $f^\infty(\mathbf{x}) = \frac{1}{2} \phi(\langle \mathbf{w}^+, \mathbf{x} \rangle)$ for $\sigma_{p_j} < \sigma_{\text{init}}^{\max}$. $\square$

H Proofs for the Orthogonally Separable Case

In this section, we revisit the orthogonally separable setting considered by [Phuong and Lampert \(2021\)](#). Surprisingly, in this setting, all KKT points which contains at least one positive neuron and negative neuron are indeed global-max-margin directions and unique in function space. This means it is possible to prove the global optimality of margin in [Phuong and Lampert \(2021\)](#)'s setting even without a trajectory-based analysis.

Definition H.1 (Orthogonally Separable Data, [Phuong and Lampert 2021](#)). A binary classification dataset $\{(\mathbf{x}_1, y_1), \dots, (\mathbf{x}_n, y_n)\}$ is called *orthogonally separable* if for all $i, j \in [n]$, if $\mathbf{x}_i^\top \mathbf{x}_j > 0$ whenever $y_i = y_j$ and $\mathbf{x}_i^\top \mathbf{x}_j \leq 0$ whenever $y_i = -y_j$.

Let $\boldsymbol{\theta} = (\mathbf{w}_1, \dots, \mathbf{w}_m, a_1, \dots, a_m) \in \mathbb{R}^D$ and $f_{\boldsymbol{\theta}}(\mathbf{x}) := \sum_{i=1}^m a_i \phi(\langle \mathbf{x}, \mathbf{w}_i \rangle)$ where ϕ is ReLU, i.e., $\phi(x) = \max\{0, x\}$. The following theorem shows that for orthogonally separable data, all KKT-margin directions are global-max-margin directions.

Theorem H.2. Suppose the dataset is orthogonally separable, for all KKT-margin directions $\boldsymbol{\theta} \in \mathbb{S}^D$, their corresponding functions $f_{\boldsymbol{\theta}}$ are the same and thus they are all global-max-margin directions.

The Theorem H.2 is a simple corollary of the following lemma Lemma H.3.

Lemma H.3. If $\boldsymbol{\theta}$ satisfies the KKT conditions of (P), then for $a_k \neq 0$, $|a_k| = \|\mathbf{w}_k\|_2$ and $(\sum_{j: a_j a_k > 0} a_j^2) \frac{\mathbf{w}_k}{a_k}$ is the global minimizer of the following optimization problem (Q):

$$\min_{\mathbf{w}} \quad \frac{1}{2} \|\mathbf{w}\|_2^2 \quad \text{s.t.} \quad \langle \mathbf{w}, \mathbf{x}_i \rangle \geq 1, \quad \text{for all } i \in [n] \text{ with } y_i = \text{sgn}(a_k). \quad (\text{Q})$$

In other words, all the non-zero a_k , $\mathbf{w}_k$ can be split into 2 groups according to the sign of a_k , where in each group, $\frac{\mathbf{w}_k}{a_k}$ is the same.

Proof of Theorem H.2. By Lemma H.3, we know for any $\boldsymbol{\theta}$ satisfying the KKT condition of (P),

$$\mathbf{w}_k = \left(\sum_{j: a_j a_k > 0} a_j^2 \right)^{-1} a_k \mathbf{w}^{\text{sgn}(a_k)}, \quad (69)$$

where $\mathbf{w}^{\text{sgn}(a_k)}$ ($\mathbf{w}^+$ or $\mathbf{w}^-$) are the unique global minimzer of the constrained convex optimization of (Q).

Thus $\|\boldsymbol{\theta}\|_2^2 = \sum_{i \in [m]} (|a_i|^2 + \|\mathbf{x}_i\|_2^2) = 2 \sum_{i \in [m]} |a_i|^2 = \|\mathbf{w}^-\|_2 + \|\mathbf{w}^+\|_2$ is the same for all $\boldsymbol{\theta}$ satisfying the condition in the theorem statement. Here the last equality uses (69) and $|a_k| = \|\mathbf{w}_k\|_2$.

Next we check the uniqueness of $f_{\boldsymbol{\theta}}$. For any $\mathbf{x}$, we have

$$\begin{aligned} f_{\boldsymbol{\theta}}(\mathbf{x}) &= \sum_{k \in [m]} a_k \phi(\langle \mathbf{x}, \mathbf{w}_k \rangle) = \phi \left(\left\langle \mathbf{x}, \sum_{k: a_k > 0} a_k \mathbf{w}_k \right\rangle \right) + \phi \left(\left\langle \mathbf{x}, \sum_{k: a_k < 0} a_k \mathbf{w}_k \right\rangle \right) \\ &= \phi(\langle \mathbf{x}, \mathbf{w}^+ \rangle) + \phi(\langle \mathbf{x}, \mathbf{w}^- \rangle), \end{aligned}$$

which completes the proof. $\square$

Proof of Lemma H.3. By KKT conditions (Definition B.8), there exist $\lambda_1, \dots, \lambda_n \geq 0$, such that for each $k \in [m]$, there are $h_1^{(k)}, \dots, h_n^{(k)} \in \mathbb{R}$ such that for all $i \in [n]$, $h_i^{(k)} \in \phi^\circ(\langle \mathbf{w}_k, \mathbf{x}_i \rangle)$, and the following conditions hold:

$$\mathbf{w}_k = a_k \sum_{i \in [n]} \lambda_i h_i^{(k)} y_i \mathbf{x}_i, \quad a_k = \sum_{i \in [n]} \lambda_i y_i \phi(\mathbf{w}_k^\top \mathbf{x}_i),$$

and $\lambda_i = 0$ whenever $y_i f_{\boldsymbol{\theta}}(\mathbf{x}_i) > 1$. By Lemma B.9, $\|\mathbf{w}_k\|_2 = |a_k|$.

We claim that for all $i \in [n]$ so that $\lambda_i h_i^{(k)} > 0$, it holds that $y_i = \text{sgn}(a_k)$ and $\langle \mathbf{w}_k, \mathbf{x}_i \rangle > 0$. Let $i \in [n]$ be any index so that $\lambda_i h_i^{(k)} > 0$. Then $h_i^{(k)} > 0$. By KKT conditions,

$$\langle \mathbf{w}_k, \mathbf{x}_i \rangle = \left\langle a_k \sum_{j \in [n]} \lambda_j h_j^{(k)} y_j \mathbf{x}_j, \mathbf{x}_i \right\rangle = a_k y_i \sum_{j \in [n]} \lambda_j h_j^{(k)} \langle y_j \mathbf{x}_j, y_i \mathbf{x}_i \rangle. \quad (70)$$

Since $\phi(x) = \max\{x, 0\}$, it holds that $\langle \mathbf{w}_k, \mathbf{x}_i \rangle \geq 0$; otherwise $h_i^{(k)} \in \phi^\circ(\langle \mathbf{w}_k, \mathbf{x}_i \rangle) = \{0\}$, which contradicts to $h_i^{(k)} > 0$. Then (70) implies that the product of $a_k y_i$ and $\sum_{j \in [n]} \lambda_j h_j^{(k)} \langle y_j \mathbf{x}_j, y_i \mathbf{x}_i \rangle$ is non-negative. By orthogonal separability, $\langle y_j \mathbf{x}_j, y_i \mathbf{x}_i \rangle \geq 0$ and thus $\sum_{j \in [n]} \lambda_j h_j^{(k)} \langle y_j \mathbf{x}_j, y_i \mathbf{x}_i \rangle \geq \lambda_i h_i^{(k)} \|y_i \mathbf{x}_i\|_2^2 > 0$. Then we can conclude that $a_k y_i \geq 0$ and thus $y_i = \text{sgn}(a_k)$. Since $y_i = \text{sgn}(a_k)$ and $a_k \neq 0$, indeed we have $a_k y_i > 0$. Now using (70) again, we obtain $\langle \mathbf{w}_k, \mathbf{x}_i \rangle \geq a_k y_i \cdot \lambda_i h_i^{(k)} \|y_i \mathbf{x}_i\|_2^2 > 0$ if $\lambda_i h_i^{(k)} > 0$.

Furthermore, for any $a_k \neq 0$, since $\|\mathbf{w}_k\|_2 = |a_k| > 0$, there is at least one index $j_* \in [n]$ such that $\lambda_{j_*} h_{j_*}^{(k)} > 0$ (otherwise $\mathbf{w}_k = \mathbf{0}$ by KKT conditions). For all $i \in [n]$, again by (70), it holds that

$$\text{sgn}(a_k) y_i \langle \mathbf{w}_k, \mathbf{x}_i \rangle = |a_k| \sum_{j \in [n]} \lambda_j h_j^{(k)} \langle y_j \mathbf{x}_j, y_i \mathbf{x}_i \rangle \geq |a_k| \lambda_{j_*} h_{j_*}^{(k)} \langle y_{j_*} \mathbf{x}_{j_*}, y_i \mathbf{x}_i \rangle > 0,$$

where the last inequality is from the assumption of orthogonally separability. This further implies $h_i^{(k)} = \mathbb{1}_{[y_i = \text{sgn}(a_k)]}$ and thus $\mathbf{w}_k = a_k \sum_{i=1}^n \mathbb{1}_{[y_i = \text{sgn}(a_k)]} \lambda_i y_i \mathbf{x}_i$ for all $k \in [m]$.

Therefore we can split the neurons with non-zero a_k into two parts: $K^+ = \{k \in [m] : a_k > 0\}$, $K^- = \{k \in [m] : a_k < 0\}$. Every $k \in K^+$ satisfies the following:

$$a_k = \|\mathbf{w}_k\|_2, \quad (71)$$

$$\mathbf{w}_k = a_k \sum_{i=1}^n \mathbb{1}_{[y_i=1]} \lambda_i \mathbf{x}_i. \quad (72)$$

This implies $\forall k \in K^+$, $\frac{\mathbf{w}_k}{a_k} = \frac{\mathbf{w}_k}{\|\mathbf{w}_k\|_2} = \sum_{i=1}^n \mathbb{1}_{[y_i=1]} \lambda_i \mathbf{x}_i$. Define $\bar{\mathbf{w}} := \sum_{k \in K^+} a_k \mathbf{w}_k$, then

$$\bar{\mathbf{w}} = \left(\sum_{k \in K^+} a_k^2 \right) \sum_{i=1}^n \mathbb{1}_{[y_i=1]} \lambda_i \mathbf{x}_i.$$

Recall that $\lambda_i = 0$ whenever $y_i f_\theta(\mathbf{x}_i) > 1$. When $y_i = 1$, $f_\theta(\mathbf{x}_i)$ can be rewritten as

$$f_\theta(\mathbf{x}_i) = \sum_{k \in [m]} a_k \mathbb{1}_{[\text{sgn}(a_k)=1]} \langle \mathbf{w}_k, \mathbf{x}_i \rangle = \langle \mathbf{x}_i, \bar{\mathbf{w}} \rangle.$$

So we can verify that $\bar{\mathbf{w}}$ satisfies the KKT conditions of the following constrained convex optimization problem:

$$\min_{\mathbf{w}} \|\mathbf{w}\|_2^2 \quad (73)$$

$$\text{s.t. } \langle \mathbf{w}, \mathbf{x}_i \rangle \geq 1, \quad \text{for all } i \in [n] \text{ with } y_i = 1. \quad (74)$$

By convexity, $\bar{\mathbf{w}}$ is the unique minimizer of the above problem. The negative part K^- can be analyzed in the same way. $\square$

I Additional Discussions

I.1 Illustrations for Figure 1

In this section we further illustrate the relationship between KKT-margin and max-margin directions, as the examples have showed in Figure 1.

I.1.1 Left: Symmetric Data

Example. For some symmetric data, there are KKT-margin directions with non-linear decision boundary (and thus by Theorem 4.2 are not global-max-margin directions).

Let λ_i be the dual variable for (x_i, y_i) , then the KKT conditions (Definition B.8 and Lemma B.9) ask

1. for all $k \in [m]$, $\mathbf{w}_k \in \sum_{i \in [n]} \lambda_i y_i a_k \phi^\circ(\mathbf{w}_k^\top \mathbf{x}_i) \mathbf{x}_i$;
2. for all $k \in [m]$, $|a_k| = \|\mathbf{w}_k\|_2$;

3. for all $i \in [n]$, if $q_i(\theta) \neq q_{\min}(\theta)$ then $\lambda_i = 0$ (recall that $q_i(\theta) = y_i f_\theta(\mathbf{x}_i)$).

For $\alpha_{\text{leaky}} = 0$, the example is simpler. Consider the following case: the data points are $\mathbf{x}_1 = (1, -1)$, $\mathbf{x}_2 = (1, 0)$, $\mathbf{x}_3 = (1, 1)$ with label 1 and the symmetric counterpart $\mathbf{x}_4 = (-1, 1)$, $\mathbf{x}_5 = (-1, 0)$, $\mathbf{x}_6 = (-1, -1)$ with label -1 . As we have proved, the global-max-margin solution is a linear function and in this case $\mathbf{w}^* = (1, 0)$. On the other hand, for hidden neurons $m \geq 3$, one KKT-margin direction is as follows:

$$\begin{cases} a_1 = 2^{-1/4} & \mathbf{w}_1 = \frac{1}{2^{3/4}}(1, 1) \\ a_2 = 2^{-1/4} & \mathbf{w}_2 = \frac{1}{2^{3/4}}(1, -1) \\ a_3 = -1 & \mathbf{w}_3 = (-1, 0) \\ a_k = 0 & \mathbf{w}_k = \mathbf{0} \end{cases} \quad \text{for all } k > 3.$$

In this case, all the data points $\mathbf{x}_i$ share the same output margin $q_i(\theta)$, so they are all support vectors. A possible choice of dual variables is $\lambda = (\frac{1}{\sqrt{2}}, 0, \frac{1}{\sqrt{2}}, 0, 1, 0)$. It is easy to verify that this KKT-margin direction does not have linear decision boundary and is thus not global-max-margin.

For $\alpha_{\text{leaky}} > 0$, we can adapt the above case to construct a KKT point. Let β be a solution to the equation

$$(2 \sin^2 \beta + \cos \beta) \alpha_{\text{leaky}}^2 - (1 + \cos \beta) \alpha_{\text{leaky}} + \cos 2\beta = 0.$$

Let the data be $\mathbf{x}_1 = (1, \cot \beta)$, $\mathbf{x}_2 = (1, 0)$, $\mathbf{x}_3 = (1, -\cot \beta)$ with label 1 and the corresponding opposites $\mathbf{x}_4 = (-1, -\cot \beta)$, $\mathbf{x}_5 = (-1, 0)$, $\mathbf{x}_6 = (-1, \cot \beta)$ with label -1 . Then we can have a KKT point with $\lambda = (\frac{\sin^2 \beta}{\cos \beta(1-\alpha_{\text{leaky}})}, 0, \frac{\sin^2 \beta}{\cos \beta(1-\alpha_{\text{leaky}})}, 0, 1 - \frac{2\alpha_{\text{leaky}} \sin^2 \beta}{(1-\alpha_{\text{leaky}}) \cos \beta}, 0)$ and

$$\begin{cases} a_1 = (2(1 + \alpha_{\text{leaky}}) \cos \beta)^{-1/2} & \mathbf{w}_1 = a_1 \sin \beta (\cot \beta, 1) \\ a_2 = a_1 & \mathbf{w}_2 = a_2 \sin \beta (\cot \beta, -1) \\ a_3 = -(1 + \alpha_{\text{leaky}})^{-1/2} & \mathbf{w}_3 = -a_3 (-1, 0) \\ a_k = 0 & \mathbf{w}_k = \mathbf{0} \end{cases} \quad \text{for all } k > 3.$$

When $\cos 2\beta = 0$ we already have solution $\alpha_{\text{leaky}} = 0$, and it is easy to verify that for any $\alpha_{\text{leaky}} \in [0, 1)$ there is a solution β that satisfy the KKT conditions. Thus in the leaky ReLU case we are considering in the previous chapters, there are also KKT-margin directions that have non-linear decision boundaries and therefore have sub-optimal margin.

I.1.2 Middle and Right: Non-symmetric Data

In Figure 1 we further show two examples of non-symmetric data that gradient flow from small initialization converges to a linear-boundary classifier that has a suboptimal margin.

The idea of the middle plot dataset comes from [Shah et al. \(2020\)](#). In the middle subplot, we exhibit a data example that is linear separable in the first dimension x but not linear separable in the second dimension y . The data is distributed on $(A_\epsilon, 1)$ and $(A_\epsilon, -1)$ with label 1 and on $(-A_\epsilon, 0)$ with label -1 (here $A_\epsilon = [c, \infty)$ is an interval in one dimension). We add identical entries c to all the data in the third dimension z so in the $x - y$ plane with $z = c$ the two-layer ReLU network can represent decision patterns with bias.

To apply Theorem A.7 on this dataset, we need to make c smaller than ϵ and $\epsilon/\epsilon' \ll \alpha_{\text{leaky}}$, so that the points at $(\epsilon, 1)$ and $(\epsilon, -1)$ becomes the support vectors for the one-neuron function. Also we can control the principal direction by taking more data points from the positive class, and then gradient flow will converge to the one-neuron max-margin solution as predicted by Theorem A.7. This solution cannot be global max-margin when $\epsilon \ll 1$, as a two-neuron network can express a function where these two support vectors have much larger distances to the decision boundary (and possess larger output margins).

In the right plot, we add three hints to a linear separable dataset so that gradient flow converges to the solution with a linear decision boundary and suboptimal margin. The result follows from Theorem 6.2.

I.1.3 Experimental Results

We run gradient descent with small learning rate and 0.001 times the He initialization ([He et al., 2015](#)) on the two-layer LeakyReLU network for the examples in Figure 1. The contours of the neural net

outputs are displayed in Figure 2. In the three settings the neural nets actually converge to linear classifiers.

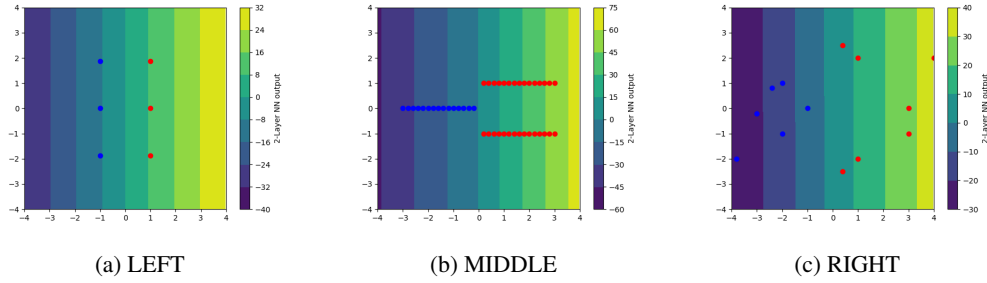


Figure 2: Two-layer Leaky ReLU neural nets converge to functions with linear decision boundary for the examples in Figure 1. The output contours are displayed in colors, and lighter colors mean higher outputs.

I.2 On the Non-branching Starting Point Assumptions

In the proofs of the main theorems we make assumptions regarding the starting point of gradient flow trajectories being non-branching (Assumption 4.6 for the symmetric case and Assumption A.6 for the non-symmetric case). The assumptions address a technical difficulty due to the potential non-uniqueness of gradient flow trajectories on general non-smooth loss functions. The motivations for these assumptions are explained below.

I.2.1 The non-uniqueness of gradient flow trajectories

Gradient flow trajectories are unique on smooth loss functions by the classic theory of ordinary differential equations. In this case, for trajectory defined by $\frac{d\theta}{dt} = -\nabla\mathcal{L}(\theta)$, at any point θ_0 , if both $\nabla\mathcal{L}(\theta_0)$ and $\nabla^2\mathcal{L}(\theta_0)$ are continuous, then the trajectory is unique as long as it exists.

For the non-smooth case with differential inclusion $\frac{d\theta}{dt} \in -\partial^\circ\mathcal{L}(\theta)$, when $\mathcal{L}$ is continuous and convex, the Clarke subdifferentials agree with the subdifferentials for convex functions, and gradient flow trajectory is also unique (for instance see Bolte et al. 2010). However, on loss functions that are non-smooth and non-convex, gradient flow may not be unique and the trajectory may branch at non-differentiable points (see Figure 3). When a non-differentiable point is atop a “ridge”, a gradient flow reaching it may go down different slopes next. Then any starting points wherefrom gradient flow can reach such on-the-ridge points are not non-branching starting points as the trajectory is not unique. For instance, with $\mathcal{L}(\theta) = -|\langle\theta, w\rangle|$, then the trajectory with $\theta(t) = 0$ for $t < t_s$ and $\theta(t) = \pm(t - t_s)w$ for $t \geq t_s$ is a valid gradient flow trajectory for any $t_s \geq 0$. On the other hand, when the point is either at the bottom of a “valley” or at a “refraction edge”, the trajectory would not split. Figure 3 sketches in red the possible gradient flow trajectories in different circumstances.

In the case of two-layer Leaky ReLU network dynamics, there are settings where Assumption 4.6 or Assumption A.6 holds. When data points are orthogonally separable (Definition H.1), all starting points are non-branching. In this case, the output of each Leaky ReLU neuron will change monotonically. By the chain rule, for any neuron $k \in [m]$, on any data sample $i \in [n]$,

$$\left\langle \frac{d\mathbf{w}_k}{dt}, x_i \right\rangle \in -\frac{a_k}{n} \sum_{j \in [n]} \ell'(q_j(\theta)) y_j \phi^\circ(\langle \mathbf{w}_k, \mathbf{x}_i \rangle) \langle x_i, x_j \rangle.$$

Then as $y_i y_j \langle x_i, x_j \rangle \geq 0$ by the orthogonally separability, the sign of RHS is controlled by $\text{sgn}(a_k y_i)$. With Theorem B.19, we know each a_k does not change its sign along the gradient flow trajectory, and therefore $\langle \mathbf{w}_k, x_i \rangle$ changes monotonically. Then following the arguments of the classic theory of ordinary differential equation, by applying Grönwall’s inequality to both intervals $\{t : \langle \mathbf{w}_k(t), x_i \rangle > 0\}$ and $\{t : \langle \mathbf{w}_k(t), x_i \rangle \leq 0\}$ we know the trajectory is unique. In this setting all the non-differentiable landscapes resemble the “refraction edges”.

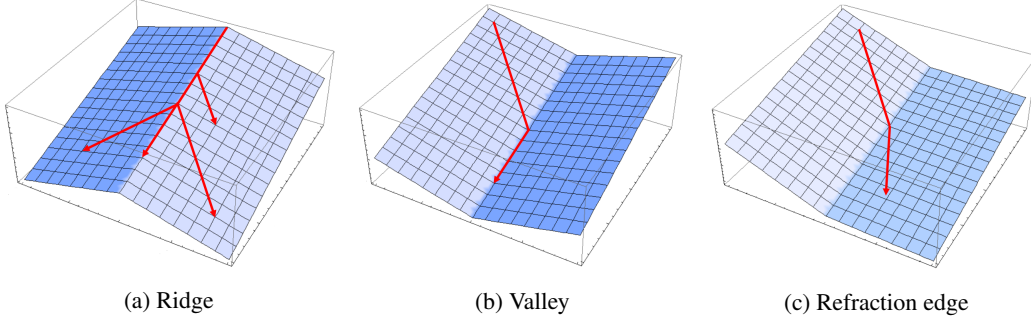


Figure 3: Gradient flow trajectories behave differently in different landscapes. The trajectory may be non-unique only after arriving at a point on the "ridge".

In the general cases, it is a future research direction to find other analyses that can replace the non-branching starting point assumptions, and doing so may deepen our understanding in the trajectory behaviors in non-smooth settings.

J Additional Experiments

We conducted several additional experiments on synthetic datasets. The goal is to show that 2-layer Leaky ReLU networks actually converges to the max-margin linear classifiers in different settings with moderately small initialization. The results are summarized in Table 1 and Figure 4.

Dataset size	SVM test error	2-Layer neural net test error
10	30.2%	30.5 %
20	19.7%	18.9%
30	17.6%	15.6%
40	8.0%	7.1%
50	6.4%	5.9%
60	6.3%	5.1%
70	7.6%	6.5%
80	3.9%	3.1%
90	6.1%	5.2%
100	2.9%	2.9%

Table 1: Test errors for SVM max-margin linear classifiers and 2-Layer ReLU neural networks are nearly the same across different data size.

Data. $n = 10, 20, \dots, 100$ data points are randomly sampled from the standard gaussian distribution $\mathcal{N}(0, \mathbf{I})$ in the space of dimension $d = 50$, and are classified with a linear classifier through zero. Then the points are translated mildly away from the classifier to make a small nonzero margin that assists learning.

Model and Training. We used the two-layer leaky ReLU network with hidden layer width $m = 100$ and with bias terms. In our setting the bias term is equivalent to adding an extra dimension of value 0.1 to all the data points. We trained our model with the gradient descent method from 0.001 times the He initialization (He et al., 2015) and initial learning rate 0.01. The learning rate is raised after interpolation to boost margin increase.

We compare the neural network output with the max-margin linear classifier produced by the support vector machine (SVM) on hinge loss. In Table 1, the test errors are calculated from 10000 test points from the same distribution. In Figure 4, we draw the decision boundaries for both the SVM max-margin linear classifier and the neural network restricted to a plane passing 0. The results show that the neural network classifier converges to the max-margin linear classifier in our setting.

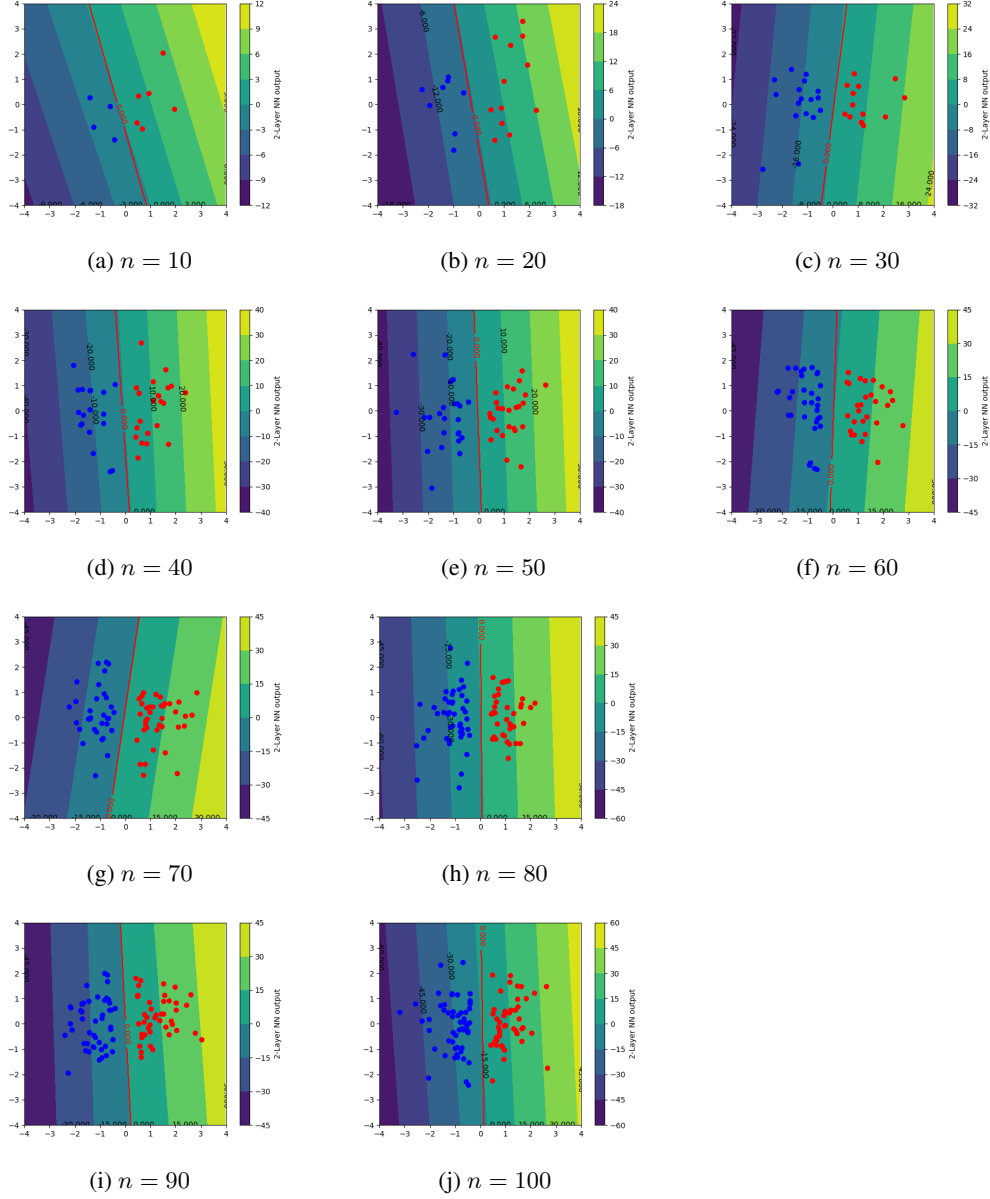


Figure 4: Two-layer Leaky ReLU neural net converges in direction to the SVM max-margin linear classifier. **Red and Blue Dots:** two classes of data points. **Red Lines:** the decision boundaries of the SVM max-margin linear classifiers. **Background:** the contours of two-layer leaky-ReLU neural network outputs. Lighter colors mean higher outputs. The underlying true separator is the vertical line through zero.