Proceedings of Machine Learning Research vol 201:1–31, 2023 34th International Conference on Algorithmic Learning Theory

Implicit Regularization Towards Rank Minimization in ReLU Networks

Nadav Timor NADAV.TIMOR@WEIZMANN.AC.IL
Department of Computer Science and Applied Mathematics, Weizmann Institute of Science

Gal Vardi GALVARDI@TTIC.EDU
TTI Chicago and Hebrew University

Ohad Shamir OHAD.SHAMIR@WEIZMANN.AC.IL
Department of Computer Science and Applied Mathematics, Weizmann Institute of Science

Editors: Shipra Agrawal and Francesco Orabona

Abstract

We study the conjectured relationship between the implicit regularization in neural networks, trained with gradient-based methods, and rank minimization of their weight matrices. Previously, it was proved that for linear networks (of depth 2 and vector-valued outputs), gradient flow (GF) w.r.t. the square loss acts as a rank minimization heuristic. However, understanding to what extent this generalizes to nonlinear networks is an open problem. In this paper, we focus on nonlinear ReLU networks, providing several new positive and negative results. On the negative side, we prove (and demonstrate empirically) that, unlike the linear case, GF on ReLU networks may no longer tend to minimize ranks, in a rather strong sense (even approximately, for "most" datasets of size 2). On the positive side, we reveal that ReLU networks of sufficient depth are provably biased towards low-rank solutions in several reasonable settings.

Keywords: Implicit regularization, Rank minimization

1. Introduction

A central puzzle in the theory of deep learning is how neural networks generalize even when trained without any explicit regularization, and when there are far more learnable parameters than training examples. In such an underdetermined optimization problem, there are many global minima with zero training loss, and gradient descent seems to prefer solutions that generalize well (see Zhang et al. (2017)). Hence, it is believed that gradient descent induces an *implicit regularization* (a.k.a. *implicit bias*) (Neyshabur et al., 2015, 2017), and characterizing this bias has been a subject of extensive research.

Several works in recent years studied the relationship between the implicit regularization in *linear* neural networks and rank minimization. Much effort was directed at the matrix factorization problem, which corresponds to training a depth-2 linear neural network with multiple outputs w.r.t. the square loss, and is considered a well-studied test-bed for studying implicit regularization in deep learning. Gunasekar et al. (2018b) initially conjectured that the implicit regularization in matrix factorization can be characterized by the nuclear norm of the corresponding linear predictor. This conjecture was further studied in a string of works (e.g., Belabbas (2020); Arora et al. (2019); Razin and Cohen (2020)) and was formally refuted by Li et al. (2020). Razin and Cohen (2020) conjectured that the implicit regularization in matrix factorization can be explained by rank minimization, and also hypothesized that some notion of rank minimization may be key to explaining generalization in deep learning. Li et al. (2020) established evidence that the implicit regularization in matrix factorization is a heuristic for rank minimization. Razin et al. (2021) studied implicit regularization in tensor factorization (a generalization of matrix factorization). They demonstrated, both theoretically and empirically, implicit bias towards low-rank tensors. Going beyond factorization problems, Ji and Telgarsky (2018, 2020) showed that in linear networks of output dimension 1, gradient flow (GF) w.r.t. exponentially-tailed classification losses converges to networks where the weight matrix of every layer is of rank 1.

However, once we move to nonlinear neural networks (which are by far the more common in practice), things are less clear. Empirically, a series of works studying neural network compression (cf. Denton et al. (2014); Yu et al. (2017); Alvarez and Salzmann (2017); Arora et al. (2018); Tukan et al. (2020)) showed that replacing the weight matrices by low-rank approximations results in only a small drop in accuracy. This suggests that the weight matrices in practice are not too far from being low-rank. However, whether they provably behave this way remains unclear.

Our contribution. In this work we consider fully-connected nonlinear networks employing the popular ReLU activation function, and study whether GF is biased towards networks where the weight matrices have low ranks. On the negative side, we show that for small ReLU networks (i.e., of depth and width 2), there is no rank-minimization bias in a rather strong sense. On the positive side, for deeper and possibly wider overparameterized networks, we identify reasonable settings where GF is biased towards low-rank solutions. In more detail, our contributions are as follows:

- We begin by considering depth-2 width-2 ReLU networks with multiple outputs, trained with the square loss. Li et al. (2020) gave evidence that, in *linear* networks with the same architecture, the implicit bias of GF can be characterized as a heuristic for rank minimization. In contrast, we show that for ReLU networks the situation is quite different. Specifically, we present a simple negative example in which GF does not converge to a low-rank solution. It holds for datasets of size 2, $\{(\mathbf{x}_1, \mathbf{y}_1), (\mathbf{x}_2, \mathbf{y}_2)\} \subseteq \mathbb{S}^1 \times \mathbb{S}^1$, whenever the angle between $\mathbf{x}_1$ and $\mathbf{x}_2$ is in $(\pi/2, \pi)$ and $\mathbf{y}_1, \mathbf{y}_2$ are linearly independent. Thus, rank minimization does not occur even if we just consider "most" datasets of this size. Moreover, we show that with at least constant probability, the solutions that GF converges to are not even close to having low rank, under any reasonable approximate rank metric. We also demonstrate these results empirically.

- Next, for ReLU networks that are overparameterized in terms of depth and have width ≥ 2, we identify interesting settings in which GF is biased towards low ranks:

  - First, we consider ReLU networks trained w.r.t. the square loss. We show that for sufficiently deep networks, if GF converges to a network that attains zero loss and minimizes the ℓ_2 norm of the parameters, then the average ratio between the spectral and the Frobenius norms of the weight matrices is close to 1. Since the squared inverse of this ratio is the *stable rank* (which is a continuous approximation of the rank, and equals 1 if and only if the matrix has rank 1), the result implies a bias towards low ranks. Although GF in ReLU networks w.r.t. the square loss is not known to have a bias towards solutions that minimize the ℓ_2 norm, in practice it is common to use explicit ℓ_2 regularization, which encourages norm minimization. Thus, our result suggests that GF in deep networks trained with the square loss and explicit ℓ_2 regularization encourages rank minimization.

– Shifting our attention to binary classification problems, we consider ReLU networks trained with exponentially-tailed classification losses. By Lyu and Li (2019), GF in such networks is biased towards networks that maximize the ℓ_2 margin. We show that for sufficiently deep networks, maximizing the margin implies rank minimization, where the rank is measured by the ratio between the norms as in the former case.

Additional Related Work

The implicit regularization of GF trained with the square loss in matrix factorization and linear neural networks was extensively studied as a first step toward understanding implicit regularization in more complex models (see, e.g., Gunasekar et al. (2018b); Razin and Cohen (2020); Arora et al. (2019); Belabbas (2020); Eftekhari and Zygalakis (2020); Li et al. (2018); Ma et al. (2018); Woodworth et al. (2020); Gidel et al. (2019); Li et al. (2020); Yun et al. (2020); Azulay et al. (2021); Razin et al. (2021)). As we already discussed, some of these works showed bias towards low ranks. Ergen and Pilanci (2021b) also considered bias towards low ranks in deep linear and ReLU networks for one-dimensional datasets.

The implicit regularization of GF in nonlinear neural networks with the square loss was studied in several works. Vardi and Shamir (2021) and Azulay et al. (2021) studied the implicit regularization in single-neuron networks. By Vardi and Shamir (2021), in single-neuron networks and single-hidden-neuron networks with the ReLU activation, the implicit regularization cannot be expressed by any non-trivial regularization function. Namely, there is no non-trivial regularization function $\mathcal{R}(\boldsymbol{\theta})$, where $\boldsymbol{\theta}$ are the parameters of the model, such that if GF with the square loss converges to a global minimum, then it is a global minimum that minimizes $\mathcal{R}$. However, this negative result does not imply that GF is not implicitly biased towards low-rank solutions, for two reasons. First, bias toward low ranks would not have implications in the case of networks of width 1 that these authors studied, and hence it would not contradict their negative result. Second, their result rules out the existence of a non-trivial regularization function which expresses the implicit bias for all possible datasets and initializations, but it does not rule out the possibility that GF acts as a heuristic for rank minimization, in the sense that it minimizes the ranks for "most" datasets and initializations. Boursier et al. (2022) showed that if the training examples are orthogonal, then GF on two-layer ReLU networks, under certain assumptions on the initialization, converges to a zero-loss solution that minimizes the ℓ_2 norm of the parameters.

The implicit bias of neural networks in classification tasks was also widely studied in recent years (see Vardi (2022) for a survey). Below we discuss several works that apply to GF in nonlinear neural networks. Lyu and Li (2019) and Ji and Telgarsky (2020) showed that GF on homogeneous neural networks, with exponentially-tailed losses, converges in direction to a KKT point of the maximum-margin problem in the parameter space. Vardi et al. (2021) studied in which settings this KKT point is guaranteed to be a global/local optimum of the maximum-margin problem. Chizat and Bach (2020) studied the dynamics of GF on infinite-width homogeneous two-layer networks with exponentially-tailed losses and showed bias towards margin maximization w.r.t. a certain function norm known as the variation norm. Lyu et al. (2021) studied the implicit bias in two-layer leaky ReLU networks trained on linearly separable and symmetric data and showed that GF converges to a linear classifier which maximizes the ℓ_2 margin. Sarussi et al. (2021) studied GF on two-layer leaky ReLU networks, where the training data is linearly separable. They showed convergence to a linear classifier based on a certain assumption called *Neural Agreement Regime (NAR)*. Phuong and Lampert (2020) studied the implicit bias in two-layer ReLU networks trained on orthogonally-separable data (i.e., where for every pair of labeled examples $(\mathbf{x}_i, y_i), (\mathbf{x}_j, y_j)$ we have $\mathbf{x}_i^\top \mathbf{x}_j > 0$ if $y_i = y_j$ and $\mathbf{x}_i^\top \mathbf{x}_j \leq 0$ otherwise). Safran et al. (2022) proved implicit bias towards minimizing the number of linear regions in univariate two-layer ReLU networks.

Organization. We introduce necessary notations and definitions in Section 2. In Section 3, we present our negative results for depth-2 networks, provide the proofs ideas, and discuss supporting empirical evidence. In Section 4 and 5, we state our positive results for deep ReLU networks. All formal proofs are deferred to the appendix.

2. Preliminaries

Notations. We use boldface letters to denote vectors. For $\mathbf{x} \in \mathbb{R}^d$ we denote by $\|\mathbf{x}\|$ the Euclidean norm. For a matrix X we denote by $\|X\|_F$ the Frobenius norm and by $\|X\|_\sigma$ the spectral norm. We denote $\mathbb{R}_+ := \{x \in \mathbb{R} : x \geq 0\}$. For an integer $d \geq 1$ we denote $[d] := \{1, \dots, d\}$. The angle between a pair of vectors $\mathbf{u}_1, \mathbf{u}_2 \in \mathbb{R}^d$ is $\measuredangle(\mathbf{u}_1, \mathbf{u}_2) := \arccos\left(\frac{\mathbf{u}_1^\top \mathbf{u}_2}{\|\mathbf{u}_1\| \cdot \|\mathbf{u}_2\|}\right) \in [0, \pi]$. The unit d-sphere is $\mathbb{S}^d := \{\mathbf{u} \in \mathbb{R}^{d+1} \mid \|\mathbf{u}\| = 1\}$. An open d-ball that is centered at the origin is denoted by $B_d(\epsilon) := \{\mathbf{u} \in \mathbb{R}^d \mid \|\mathbf{u}\| < \epsilon\}$ for some $\epsilon \in \mathbb{R}$. The closure of a set $A \in \mathbb{R}^d$, denoted by $\operatorname{cl} A$, is the intersection of all closed sets containing A. The boundary of A is $\partial A := (\operatorname{cl} A) \cap \left(\operatorname{cl}\left(\mathbb{R}^d \setminus A\right)\right)$.

Neural networks. A *fully-connected neural network* $N_{\boldsymbol{\theta}}$ of *depth* $k \geq 2$ is a mathematical object parameterized by a collection of *weight matrices* (also known as *layers*) $\boldsymbol{\theta} := [W^{(l)}]_{l=1}^k$, such that for every layer $l \in [k]$ we have $W^{(l)} \in \mathbb{R}^{d_l \times d_{l-1}}$. We denote by $d_{\text{in}} := d_0$, $d_{\text{out}} := d_k$ the input and output dimensions of the network, respectively. The network computes a function $N_{\boldsymbol{\theta}} : \mathbb{R}^{d_{\text{in}}} \to \mathbb{R}^{d_{\text{out}}}$ recursively as follows. For an input $\mathbf{x} \in \mathbb{R}^{d_{\text{in}}}$ we set $\mathbf{h}'_0 := \mathbf{x}$, and define for every $j \in [k-1]$ the input to the j-th layer as $\mathbf{h}_j := W^{(j)} \mathbf{h}'_{j-1}$, and the output of the j-th layer as $\mathbf{h}'_j := \sigma(\mathbf{h}_j)$, where $\sigma : \mathbb{R} \to \mathbb{R}$ is an *activation* function that acts coordinate-wise on vectors. In this work we focus on the *ReLU* activation $\sigma(z) = \max\{0, z\}$. Lastly, we define $N_{\boldsymbol{\theta}}(\mathbf{x}) := W^{(k)} \mathbf{h}'_{k-1}$ (without an additional, final activation). The l-th layer has d_l *neurons*. The j-th neuron of the l-th layer is a mathematical object that computes the function $\mathbf{h} \mapsto \sigma\left((W^{(l)}\mathbf{h})_j\right)$ if $l \neq k$, and $\mathbf{h} \mapsto (W^{(l)}\mathbf{h})_j$ otherwise. A neuron is *active* for a given input to the network if the corresponding output of this neuron is strictly positive. The neurons in layers $[k-1]$ are called *hidden neurons*. The *width of the l-th layer* is d_l, i.e., the number of neurons it has. The *width of the network $N_{\boldsymbol{\theta}}$* is the maximal width of its layers, i.e., $\max_{l \in [k]} d_l$. We sometimes apply the activation function σ also on matrices, in which case it acts entry-wise. The parameters $\boldsymbol{\theta}$ of the neural network are given by a collection of matrices, but we often view $\boldsymbol{\theta}$ as the vector obtained by concatenating the matrices in the collection. Thus, $\|\boldsymbol{\theta}\|$ denotes the ℓ_2 norm of the vector $\boldsymbol{\theta}$.

We often consider depth-2 networks. For matrices $W \in \mathbb{R}^{d_1 \times d_0}$ and $V \in \mathbb{R}^{d_2 \times d_1}$ we denote by $N_{W,V}$ the depth-2 ReLU network where $W^{(1)} = W$ and $W^{(2)} = V$. We denote by $\mathbf{w}_1^\top, \dots, \mathbf{w}_{d_1}^\top$ the rows of W, namely, the *incoming weight vectors* to the neurons in the hidden layer, and by $\mathbf{v}_1, \dots, \mathbf{v}_{d_1}$ the columns of V, namely, the *outgoing weight vectors* from the neurons in the hidden layer.

Let $\mathbf{x}_1, \dots, \mathbf{x}_n$ be n inputs, and let $X \in \mathbb{R}^{d_{\text{in}} \times n}$ be a matrix whose columns are $\mathbf{x}_1, \dots, \mathbf{x}_n$. We denote by $N_{\boldsymbol{\theta}}(X) \in \mathbb{R}^{d_{\text{out}} \times n}$ the matrix whose i-th column is $N_{\boldsymbol{\theta}}(\mathbf{x}_i)$.

Optimization problem. Let $S = \{(\mathbf{x}_i, \mathbf{y}_i)\}_{i=1}^n \subseteq \mathbb{R}^{d_{\text{in}}} \times \mathbb{R}^{d_{\text{out}}}$ be a training dataset. We often represent the dataset by matrices $(X, Y) \in \mathbb{R}^{d_{\text{in}} \times n} \times \mathbb{R}^{d_{\text{out}} \times n}$. For a neural network $N_{\boldsymbol{\theta}}$ we consider empirical-loss minimization w.r.t. the square loss. Thus, the objective is given by:

$$L_{X,Y}(\boldsymbol{\theta}) := \frac{1}{2} \sum_{i=1}^{n} \|N_{\boldsymbol{\theta}}(\mathbf{x}_i) - \mathbf{y}_i\|^2 = \frac{1}{2} \|N_{\boldsymbol{\theta}}(X) - Y\|_F^2 \,. \tag{1}$$

We assume that the data is realizable, that is, $\min_{\boldsymbol{\theta}} L_{X,Y}(\boldsymbol{\theta}) = 0$. Moreover, we focus on settings where the network is *overparameterized*, in the sense that $L_{X,Y}$ has multiple (or even infinitely many) global minima.

Gradient flow (GF). We consider *gradient flow* on the objective given in Eq. (1). This setting captures the behavior of gradient descent with an infinitesimally small step size. Let $\boldsymbol{\theta}(t)$ be the trajectory of GF. Starting from an initial point $\boldsymbol{\theta}(0)$, the dynamics of $\boldsymbol{\theta}(t)$ is given by the differential equation $\frac{d\boldsymbol{\theta}(t)}{dt} = -\nabla L_{X,Y}(\boldsymbol{\theta}(t))$. Note that the ReLU function is not differentiable at 0. Practical implementations of gradient methods define the derivative $\sigma'(0)$ to be some constant in $[0, 1]$. In this work, we assume for convenience that $\sigma'(0) = 0$. We say that GF *converges* if $\lim_{t\to\infty} \boldsymbol{\theta}(t)$ exists. In this case, we denote $\boldsymbol{\theta}(\infty) := \lim_{t\to\infty} \boldsymbol{\theta}(t)$.

3. Gradient flow does not even approximately minimize ranks

This section considers rank minimization in depth-2 networks $N_{W,V}$ trained with the square loss. We show that even for the simple case of size-2 datasets, GF does not converge to a minimum-rank solution even approximately, under mild assumptions.

In what follows, we consider ReLU networks with vector-valued outputs. For linear networks with the same architecture, it was shown that GF could be viewed as a heuristic for rank minimization (cf. Li et al. (2020); Razin and Cohen (2020)). Specifically, let

$$(X, Y) \in \mathbb{R}^{2\times 2} \times \mathbb{R}^{2\times 2}$$

be a training dataset, and let $W, V \in \mathbb{R}^{2\times 2}$ be weight matrices such that

$$N_{W,V}(X) = V\sigma(WX)$$

is a zero-loss solution. Note that if $\text{rank}(Y) = 2$, then we must have $\text{rank}(V) = 2$: Indeed, by the definition of $N_{W,V}$, we necessarily have

$$\text{rank}(Y) = \text{rank}(N_{W,V}(X)) \leq \text{rank}(V) \,.$$

Therefore, to understand rank minimization in this simple setting, we consider the rank of W in a zero-loss solution. Trivially, $\text{rank}(W) \leq 2$, so W can be considered low-rank only if $\text{rank}(W) \leq 1$.

To make the setting non-trivial, we need to show that such low-rank zero-loss solutions exist at all. The following theorem shows that this is true for almost all size-2 datasets:

Theorem 1 *Given any labeled dataset $(X, Y) \in \mathbb{R}^{2\times 2} \times \mathbb{R}^{2\times 2}$ of two inputs $\mathbf{x}_1, \mathbf{x}_2 \in \mathbb{R}^2$ with a strictly positive angle between them, i.e., $\measuredangle(\mathbf{x}_1, \mathbf{x}_2) > 0$, there exists a zero-loss solution $N_{W,V}$ with $W, V \in \mathbb{R}^{2\times 2}$, such that $\text{rank}(W) = 1$.*

The theorem follows by constructing a network where the weight vectors of the neurons in the first layer have opposite directions (and hence the weight matrix is of rank 1), such that each neuron is active for exactly one input. Then, it is possible to show that for an appropriate choice of the weights in the second layer, the network achieves zero loss. See Appendix A for the formal proof.

Theorem 1 implies that zero-loss solutions of rank 1 exist. However, we now show that GF does not converge to such solutions. We prove this result under the following assumptions:

Assumption 1 *The two target vectors $\mathbf{y}_1, \mathbf{y}_2$ are on the unit sphere $\mathbb{S}^1$ and are linearly independent.*

Assumption 2 *The two inputs $\mathbf{x}_1, \mathbf{x}_2$ are on the unit sphere $\mathbb{S}^1$, and satisfy $\frac{\pi}{2} < \measuredangle(\mathbf{x}_1, \mathbf{x}_2) < \pi$.*

The assumptions that $\mathbf{x}_i, \mathbf{y}_i$ are of unit norm are mostly for technical convenience, and we believe that they are not essential. Then, we have:

Theorem 2 *Let $(X, Y) \in \mathbb{R}^{2\times 2} \times \mathbb{R}^{2\times 2}$ be a labeled dataset that satisfies Assumptions 1 and 2. Consider GF w.r.t. the loss function $L_{X,Y}(W, V)$ from Eq. (1). Suppose that $W, V \in \mathbb{R}^{2\times 2}$ are initialized such that*

$$\|\mathbf{w}_i(0)\| < \min\left\{\frac{1}{2}, \frac{\sqrt{3}}{2}\cos\frac{\measuredangle(\mathbf{x}_1, \mathbf{x}_2)}{2}\right\}$$

and $\|\mathbf{v}_i(0)\| < \frac{1}{2}$ for all $i \in \{1, 2\}$. If GF converges to a zero-loss solution $N_{W(\infty),V(\infty)}$, then $\mathrm{rank}(W(\infty)) = 2$.

By the above theorem, GF does not minimize the rank even in a very simple setting where the dataset contains two inputs with an angle larger than $\pi/2$ (as long as the initialization point is sufficiently close to 0). In particular, if the dataset is drawn from the uniform distribution on the sphere, then this condition holds with probability $1/2$. We note that we focus on networks of width 2 in order to show our negative result already for a simple setting, but our proof technique can also be extended to wider networks (trained with a size-2 dataset). Moreover, we note that assuming a sufficiently small initialization scale is common in the literature on implicit bias in matrix factorization and linear neural networks (e.g., Li et al. (2020); Razin and Cohen (2020); Gunasekar et al. (2018b); Arora et al. (2019)).

While Theorem 2 shows that GF does not minimize the rank, it does not rule out the possibility that it converges to a solution which is close to a low-rank solution. There are many ways to define such closeness, such as the ratio of the Frobenius and spectral norms, the Frobenius distance from a low-rank solution, or the exponential of the entropy of the singular values (cf. Rudelson and Vershynin (2007); Sanyal et al. (2019); Razin and Cohen (2020); Roy and Vetterli (2007)). However, for 2×2 matrices they all boil down to either having the two rows of the matrix being nearly aligned or having at least one of them very small (at least compared to the other). In the following theorem, we show that under the assumptions stated above, for any fixed dataset, with at least constant probability, GF converges to a zero-loss solution, where the two row vectors are bounded away from 0, the ratio of their norms are bounded, and the angle between them is bounded away from 0 and from π (all by explicit constants that depend just on the dataset and are large in general). Thus, with at least constant probability, GF does not minimize any reasonable approximate notion of rank.

Theorem 3 *Let $(X, Y) \in \mathbb{R}^{2\times 2} \times \mathbb{R}^{2\times 2}$ be a labeled dataset that satisfies Assumptions 1 and 2. Consider GF w.r.t. the loss function $L_{X,Y}(W, V)$ from Eq. (1). Suppose that $W, V \in \mathbb{R}^{2\times 2}$ are initialized such that for all $i \in \{1, 2\}$ we have $\mathbf{v}_i(0) = \mathbf{0}$, and $\mathbf{w}_i(0)$ is drawn from a spherically symmetric distribution with*

$$\|\mathbf{w}_i(0)\| \leq \frac{\sqrt{3}}{2} \min\left\{\sin\left(\frac{\pi - \measuredangle(\mathbf{x}_1, \mathbf{x}_2)}{4}\right), \sin\left(\measuredangle(\mathbf{x}_1, \mathbf{x}_2) - \frac{\pi}{2}\right)\right\} .$$

Let E be the event that GF converges to a zero-loss solution $N_{W(\infty),V(\infty)}$ such that

(i) $\measuredangle\left(\mathbf{w}_1(\infty), \mathbf{w}_2(\infty)\right) \in \left[\frac{\pi}{2} - \left(\measuredangle(\mathbf{x}_1, \mathbf{x}_2) - \frac{\pi}{2}\right), \frac{3\pi}{4} + \frac{\measuredangle(\mathbf{x}_1,\mathbf{x}_2) - \pi/2}{2}\right]$,

(ii) $\|\mathbf{w}_i(\infty)\| \in \left(\frac{\sqrt{3}}{2}, \sqrt{\frac{1}{4} + \frac{4}{3(\sin\measuredangle(\mathbf{x}_1,\mathbf{x}_2))^2}}\right)$ *for all* $i \in \{1, 2\}$.

Then, $\Pr[E] \geq 2 \cdot \left(\frac{\measuredangle(\mathbf{x}_1,\mathbf{x}_2)}{2\pi}\right)^2$.

We note that in Theorem 3, the weights in the second layer are initialized to zero, while in Theorem 2, the assumption on the initialization is weaker. This difference is for technical convenience. We believe that Theorem 3 should also hold under weaker assumptions on the initialization.

3.1. Proof ideas

We define the following regions (see Figure 1):

$$\begin{aligned}
\mathcal{D} &= \{\mathbf{w} \in \mathbb{R}^2 \mid \forall i \in \{1, 2\}, \sigma(\mathbf{w}^\top \mathbf{x}_i) \leq 0\}, \\
\mathcal{S} &= \{\mathbf{w} \in \mathbb{R}^2 \mid \forall i \in \{1, 2\}, \sigma(\mathbf{w}^\top \mathbf{x}_i) > 0\}, \\
\mathcal{S}_1 &= \{\mathbf{w} \in \mathbb{R}^2 \mid \sigma(\mathbf{w}^\top \mathbf{x}_1) > 0, \sigma(\mathbf{w}^\top \mathbf{x}_2) \leq 0\}, \\
\mathcal{S}_2 &= \{\mathbf{w} \in \mathbb{R}^2 \mid \sigma(\mathbf{w}^\top \mathbf{x}_2) > 0, \sigma(\mathbf{w}^\top \mathbf{x}_1) \leq 0\}.
\end{aligned}$$

Intuitively, $\mathcal{D}$ defines the "dead" region, where neurons output 0 on both inputs, $\mathbf{x}_1$ and $\mathbf{x}_2$. $\mathcal{S}$ is the "active" region, in which neurons output a strictly positive value given both $\mathbf{x}_1, \mathbf{x}_2$. $\mathcal{S}_1$ and $\mathcal{S}_2$ are the "partially active" regions, where neurons output a strictly positive value on one input and 0 on the other.

3.1.1. PROOF IDEA OF THEOREM 2.

Assume towards contradiction that GF converges to some zero-loss network $N_{W(\infty),V(\infty)}$ with $\mathrm{rank}(W(\infty)) < 2$. Since $N_{W(\infty),V(\infty)}$ attains zero loss, then

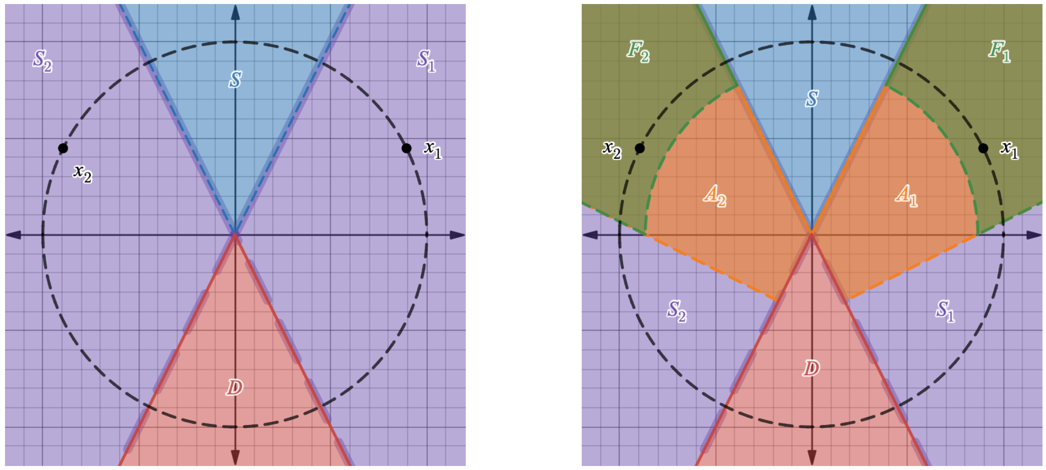
$$Y = V(\infty)\sigma\left(W(\infty)X\right) .$$

Hence

$$2 = \mathrm{rank}(Y) = \mathrm{rank}\left(V(\infty)\sigma\left(W(\infty)X\right)\right) \leq \mathrm{rank}\left(\sigma\left(W(\infty)X\right)\right) . \tag{2}$$

Therefore, the weight vectors $\mathbf{w}_1(\infty)$ and $\mathbf{w}_2(\infty)$ are not in the region $\mathcal{D}$. Indeed, if $\mathbf{w}_1(\infty)$ or $\mathbf{w}_2(\infty)$ are in $\mathcal{D}$, then at least one of the rows of $\sigma(W(\infty)X)$ is zero, in contradiction to Eq. (2). In particular, it implies that $\mathbf{w}_1(\infty)$ and $\mathbf{w}_2(\infty)$ are non-zero. Since by our assumption we have $\mathrm{rank}(W(\infty)) < 2$, then we conclude that $\mathrm{rank}(W(\infty)) = 1$. We denote $\mathbf{w}_2(\infty) = \alpha\mathbf{w}_1(\infty)$ where $\alpha \neq 0$. Note that if $\alpha > 0$, then

$$\sigma(\mathbf{w}_2(\infty)^\top \mathbf{x}_j) = \alpha\sigma(\mathbf{w}_1(\infty)^\top \mathbf{x}_j)$$



(a) Regions $\mathcal{D}$, $\mathcal{S}_1$, $\mathcal{S}_2$, $\mathcal{S}$.

(b) Regions $\mathcal{F}_i$ (in green) and $\mathcal{A}_i$ (the union of the green and orange regions).

Figure 1: Regions of interest. Dashed lines mark open boundaries.

for all $j \in \{1, 2\}$, in contradiction to Eq. (2). Thus, $\alpha < 0$. Since we also have $\mathbf{w}_1(\infty), \mathbf{w}_2(\infty) \notin \mathcal{D}$, then one of these weight vectors is in $\mathcal{S}_1 \setminus \partial\mathcal{S}_1$ and the other is in $\mathcal{S}_2 \setminus \partial\mathcal{S}_2$ (as can be seen from Figure 1). Assume w.l.o.g. that $\mathbf{w}_1(\infty) \in \mathcal{S}_1 \setminus \partial\mathcal{S}_1$ and $\mathbf{w}_2(\infty) \in \mathcal{S}_2 \setminus \partial\mathcal{S}_2$.

By observing the gradients of $L_{X,Y}$ w.r.t. $\mathbf{w}_i$ for $i \in \{1, 2\}$, the following facts follow. First, if $\mathbf{w}_i(t) \in \mathcal{D}$ at some time t, then $\frac{d}{dt}\mathbf{w}_i(t) = \mathbf{0}$, hence $\mathbf{w}_i$ remains at $\mathcal{D}$ indefinitely, in contradiction to $\mathbf{w}_i(\infty) \in \mathcal{S}_i \setminus \partial\mathcal{S}_i$. Thus, the trajectory $\mathbf{w}_i(t)$ does not visit $\mathcal{D}$. Second, if $\mathbf{w}_i(t) \in \mathcal{S}_i$ at time t, then $\frac{d}{dt}\mathbf{w}_i(t) \in \mathrm{span}\{\mathbf{x}_i\}$. Since $\mathbf{w}_i(\infty) \in \mathcal{S}_i \setminus \partial\mathcal{S}_i$, we can consider the last time t' that $\mathbf{w}_i$ enters $\mathcal{S}_i$, which can be either at the initialization (i.e., $t' = 0$) or when moving from $\mathcal{S}$ (i.e., $t' > 0$). For all time $t \geq t'$ we have $\frac{d}{dt}\mathbf{w}_i(t) \in \mathrm{span}\{\mathbf{x}_i\}$. It allows us to conclude that $\mathbf{w}_i(\infty)$ must be in a region $\mathcal{A}_i$ which is illustrated in Figure 1 (by the union of the orange and green regions).

Furthermore, we show that $\|\mathbf{w}_i(\infty)\|$ cannot be too small, namely, obtaining a lower bound on $\|\mathbf{w}_i(\infty)\|$. First, a theorem from Du et al. (2018) implies that $\|\mathbf{w}_i(t)\|^2 - \|\mathbf{v}_i(t)\|^2$ remains constant throughout the training. Since at the initialization both $\|\mathbf{w}_i(0)\|$ and $\|\mathbf{v}_i(0)\|$ are small, the consequence is that $\|\mathbf{v}_i(\infty)\|$ is small if $\|\mathbf{w}_i(\infty)\|$ is small.

Also, since $N_{W(\infty),V(\infty)}$ attains zero loss and $\mathbf{w}_i(\infty) \in \mathcal{S}_i$ for all $i \in \{1, 2\}$, then we have

$$\mathbf{y}_i = \mathbf{v}_i(\infty)(\mathbf{w}_i(\infty)^\top \mathbf{x}_i) \,.$$

Namely, only the i-th hidden neuron contributes to the output of $N_{W(\infty),V(\infty)}$ for the input $\mathbf{x}_i$. Since $\|\mathbf{y}_i\| = \|\mathbf{x}_i\| = 1$, it is impossible that both $\|\mathbf{w}_i(\infty)\|$ and $\|\mathbf{v}_i(\infty)\|$ are small. Hence, we are able to obtain a lower bound on $\|\mathbf{w}_i(\infty)\|$, which implies that $\mathbf{w}_i(\infty)$ is in a region $\mathcal{F}_i$ which is illustrated in Figure 1.

Finally, we show that since $\mathbf{w}_1(\infty) \in \mathcal{F}_1$ and $\mathbf{w}_2(\infty) \in \mathcal{F}_2$, then the angle between $\mathbf{w}_1(\infty)$ and $\mathbf{w}_2(\infty)$ is smaller than π, in contradiction to $\mathbf{w}_2(\infty) = \alpha\mathbf{w}_1(\infty)$.

3.1.2. PROOF IDEA OF THEOREM 3.

We show that if the initialization is such that $\mathbf{w}_1(0) \in \mathcal{S}_1 \setminus \partial\mathcal{S}_1$ and $\mathbf{w}_2(0) \in \mathcal{S}_2 \setminus \partial\mathcal{S}_2$ (or, equivalently, that $\mathbf{w}_1(0) \in \mathcal{S}_2 \setminus \partial\mathcal{S}_2$ and $\mathbf{w}_2(0) \in \mathcal{S}_1 \setminus \partial\mathcal{S}_1$), then GF converges to a zero-loss net-

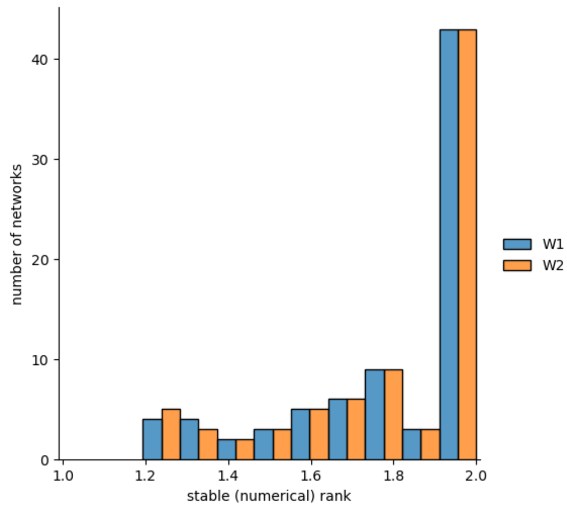


Figure 2: A histogram of the stable (numerical) ranks at convergence. In all runs, we converge to networks with stable ranks which seem bounded away from 1. Namely, gradient descent does not even approximately minimize the ranks.

work, and $\|\mathbf{w}_1(\infty)\|, \|\mathbf{w}_2(\infty)\|, \measuredangle(\mathbf{w}_1(\infty), \mathbf{w}_2(\infty))$ are in the required intervals. Since by simple geometric arguments we can show that the initialization satisfies this requirement with probability at least $2 \cdot \left(\frac{\measuredangle(\mathbf{x}_1,\mathbf{x}_2)}{2\pi}\right)^2$, the theorem follows. Indeed, suppose that $\mathbf{w}_1(0) \in \mathcal{S}_1 \setminus \partial\mathcal{S}_1$ and $\mathbf{w}_2(0) \in \mathcal{S}_2 \setminus \partial\mathcal{S}_2$. We argue that GF converges to a zero-loss network and $\|\mathbf{w}_1(\infty)\|, \|\mathbf{w}_2(\infty)\|, \measuredangle(\mathbf{w}_1(\infty), \mathbf{w}_2(\infty))$ are in the required intervals, as follows. By analyzing the dynamics of GF for such an initialization, we show that for all t and i we have $\frac{d}{dt}\mathbf{w}_i(t) = C_i(t)\mathbf{x}_i$ for some $C_i(t) \geq 0$. Thus, $\mathbf{w}_i(t)$ moves only in the direction of $\mathbf{x}_i$, and $\mathbf{w}_i(t) \in \mathcal{S}_i \setminus \partial\mathcal{S}_i$ for all t. Moreover, we are able to prove that these properties of the trajectories $\mathbf{w}_1(t)$ and $\mathbf{w}_2(t)$ imply that GF converges to a zero-loss network $N_{W(\infty),V(\infty)}$. Then, by similar arguments to the proof of Theorem 2, we have $\mathbf{w}_i(\infty) \in \mathcal{F}_i$ for all $i \in \{1, 2\}$, where $\mathcal{F}_i$ are the regions from Figure 1, and it allows us to obtain the required bounds on $\|\mathbf{w}_1(\infty)\|, \|\mathbf{w}_2(\infty)\|$, and $\measuredangle(\mathbf{w}_1(\infty), \mathbf{w}_2(\infty))$.

3.2. An empirical result

Our theorems imply that GF will not converge close to low-rank solutions, with some positive probability over the initialization. We now present a simple experiment that corroborates this and suggests that, furthermore, this holds with high probability.

We trained ReLU networks in the same setup as in the previous subsections (w.r.t. two 2×2 weight matrices $W^{(1)}, W^{(2)}$) on the two data points $\{(\mathbf{x}_i, \mathbf{y}_i)\}_{i=1}^2$ where $\mathbf{y}_1, \mathbf{y}_2$ are the standard basis vectors in $\mathbb{R}^2$, and $\mathbf{x}_1, \mathbf{x}_2$ are $(1, 0.99)$ and $(-1, 0.99)$ normalized to have unit norm. At initialization, every row of $W^{(1)}$ and every column of $W^{(2)}$ is sampled uniformly at random from the sphere of radius 10^{-4} around the origin. To simulate GF, we performed $3 \cdot 10^6$ epochs of full-batch gradient descent of step size 10^{-4}, w.r.t. the square loss. Of 288 repeats of this experiment, 79 converged to negligible loss (defined as $< 10^{-4}$). In Figure 2, we plot a histogram of the *stable (numerical) ranks* of the resulting weight matrices, i.e. the ratio $\|W^{(\ell)}\|_F^2 / \|W^{(\ell)}\|_\sigma^2$

of layer $\ell \in [2]$. The figure clearly suggests that whenever convergence to zero loss occurs, the solutions are all of rank 2, and none are even close to being low-rank (in terms of the stable rank).

4. Rank minimization in deep networks with small ℓ_2 norm

When training neural networks with gradient descent, it is common to use explicit ℓ_2 regularization on the parameters. In this case, gradient descent is biased towards solutions that minimize the ℓ_2 norm of the parameters. Below we show that, for deep overparameterized ReLU networks, such bias towards small ℓ_2 norm implies small ratios between the Frobenius and the spectral norms in the weight matrices. As discussed in the previous section, the ratio between these norms acts as a continuous surrogate for the exact rank. Formally, we have:

Theorem 4 *Let $\{(\mathbf{x}_i, y_i)\}_{i=1}^n \subseteq \mathbb{R}^{d_{in}} \times \mathbb{R}_+$ be a dataset, and assume that there is $i \in [n]$ with $\|\mathbf{x}_i\| \leq 1$ and $y_i \geq 1$. Assume that there is a fully-connected neural network N of width $m \geq 2$ and depth $k \geq 2$, such that for all $i \in [n]$ we have $N(\mathbf{x}_i) = y_i$, and the weight matrices $W_1, \ldots, W_k$ of N satisfy $\|W_i\|_F \leq B$ for some $B > 0$. Let $N_{\boldsymbol{\theta}}$ be a fully-connected neural network of width $m' \geq m$ and depth $k' > k$ parameterized by $\boldsymbol{\theta}$. Let $\boldsymbol{\theta}^* = \left[W_1^*, \ldots, W_{k'}^*\right]$ be a global optimum of the following problem:*

$$\min_{\boldsymbol{\theta}} \|\boldsymbol{\theta}\| \quad s.t. \quad \forall i \in [n],\ N_{\boldsymbol{\theta}}(\mathbf{x}_i) = y_i \ . \tag{3}$$

Then,

$$\frac{1}{k'} \sum_{i=1}^{k'} \frac{\|W_i^*\|_\sigma}{\|W_i^*\|_F} \geq \left(\frac{1}{B}\right)^{\frac{k}{k'}} \ . \tag{4}$$

Equivalently, we have the following upper bound on the harmonic mean of the ratios $\frac{\|W_i^*\|_F}{\|W_i^*\|_\sigma}$:

$$\frac{k'}{\sum_{i=1}^{k'} \left(\frac{\|W_i^*\|_F}{\|W_i^*\|_\sigma}\right)^{-1}} \leq B^{\frac{k}{k'}} \ . \tag{5}$$

By the above theorem, if k' is much larger than k, then the average ratio between the spectral and the Frobenius norms (Eq. (4)) is at least roughly 1. Likewise, the harmonic mean of the ratio between the Frobenius and the spectral norms (Eq. (5)), namely, the square root of the stable rank, is at most roughly 1. Noting that both these ratios equal 1 if and only if the matrix is of rank 1, we see that there is a bias towards low-rank solutions as the depth k' of the trained network increases. Since the dependence on k' in the theorem is exponential, increasing k' has a significant effect. Note that the result does not depend on the width of the networks. Namely, even if the width m' is large, the average ratio remains close to 1. The theorem implies that, even though many depth-k' networks can interpolate the data, some with large ranks and some with low ranks, under the theorem's conditions, there is a bias towards networks with low ranks.

Theorem 4 considers the implications of minimizing the ℓ_2 norm of the parameters on the ranks. We emphasize that even when training with explicit ℓ_2 regularization, gradient descent may not converge to such a minimizer. However, since explicit ℓ_2 regularization encourages norm minimization, it is natural to ask what the implications of norm minimization are. Indeed, several prior works considered the implications of minimizing the ℓ_2-norm of the parameters for different architectures of neural networks. For example, Savarese et al. (2019); Ergen and Pilanci (2021a) studied univariate two-layer neural networks and Ongie et al. (2019) extended the analysis to the multivariate case, Gunasekar et al. (2018a) studied fully-connected, diagonal and convolutional linear networks, and Jagadeesan et al. (2022); Pilanci and Ergen (2020); Dai et al. (2021) studied larger families of convolutional linear networks.

The intuition for the proof of Theorem 4 can be roughly described as follows. Since for the input $\mathbf{x}_i$ with $\|\mathbf{x}_i\| \leq 1$ the output of the network is of size at least 1, then the average spectral norm of the layers is at least 1. On the other hand, minimizing the ℓ_2 norm of the parameters induces bias towards small Frobenius norms of the weight layers, and in deep networks, these norms can be only slightly larger than 1. Hence, the average ratio between the spectral and the Frobenius norms is not much smaller than 1. The formal proof is given in the appendix.

5. Rank minimization in deep networks with exponentially-tailed losses

In this section, we turn to consider GF in classification tasks with exponentially-tailed losses, namely, the exponential loss or the logistic loss.

Let us first formally define the setting. We consider neural networks of output dimension 1, i.e., $d_{\text{out}} = 1$. Let

$$S = \{(\mathbf{x}_i, y_i)\}_{i=1}^n \subseteq \mathbb{R}^d \times \{-1, 1\}$$

be a binary classification training dataset. Let $X \in \mathbb{R}^{d_{\text{in}} \times n}$ and $\mathbf{y} \in \mathbb{R}^n$ be the data matrix and labels that correspond to S. Let $N_{\boldsymbol{\theta}}$ be a neural network parameterized by $\boldsymbol{\theta}$. For a loss function $\ell : \mathbb{R} \to \mathbb{R}$, the empirical loss of $N_{\boldsymbol{\theta}}$ on the dataset S is

$$L_{X,\mathbf{y}}(\boldsymbol{\theta}) := \sum_{i=1}^{n} \ell\left(y_i N_{\boldsymbol{\theta}}(\mathbf{x}_i)\right) \ . \tag{6}$$

We focus on the exponential loss $\ell(q) = e^{-q}$ and the logistic loss $\ell(q) = \log(1 + e^{-q})$. We say that the dataset is *correctly classified* by the network $N_{\boldsymbol{\theta}}$ if for all $i \in [n]$ we have $y_i N_{\boldsymbol{\theta}}(\mathbf{x}_i) > 0$. We consider GF on the objective given in Eq. (6). We say that a network $N_{\boldsymbol{\theta}}$ is *homogeneous* if there exists $M > 0$ such that for every $\alpha > 0$ and $\boldsymbol{\theta}, \mathbf{x}$ we have $N_{\alpha\boldsymbol{\theta}}(\mathbf{x}) = \alpha^M N_{\boldsymbol{\theta}}(\mathbf{x})$. Note that fully-connected ReLU networks are homogeneous. We say that a trajectory $\boldsymbol{\theta}(t)$ of GF *converges in direction* to $\tilde{\boldsymbol{\theta}}$ if $\lim_{t\to\infty} \frac{\boldsymbol{\theta}(t)}{\|\boldsymbol{\theta}(t)\|} = \frac{\tilde{\boldsymbol{\theta}}}{\|\tilde{\boldsymbol{\theta}}\|}$ The following well-known result shows an important property of the implicit bias in homogeneous neural networks trained with the logistic or the exponential loss:

Lemma 1 (Paraphrased from Lyu and Li (2019) and Ji and Telgarsky (2020)) *Let $N_{\boldsymbol{\theta}}$ be a homogeneous ReLU neural network. Consider minimizing either the exponential or the logistic loss over a binary classification dataset using GF. Suppose that there exists time t_0 such that the empirical loss is smaller than* 1. *Then, GF converges to zero loss, and converges in direction to a first order stationary point (KKT point) of the following maximum margin problem in parameter space:*

$$\min_{\boldsymbol{\theta}} \frac{1}{2}\|\boldsymbol{\theta}\|^2 \quad \textit{s.t.} \quad \forall i \in [n],\ y_i N_{\boldsymbol{\theta}}(\mathbf{x}_i) \geq 1 \ . \tag{7}$$

The lemma guarantees that GF converges in direction to a KKT point of the maximum margin problem in parameter space (assuming that it reaches a small loss). However, this KKT point may not be a global optimum (see Vardi et al. (2021)). Intuitively, the lemma implies a certain bias towards margin maximization in parameter space, although it does not guarantee convergence to a maximum-margin solution. As we discussed in Section 4 (for the case of norm minimization), it is natural to ask what the implications of margin maximization are. The theorem below shows that, in deep overparameterized ReLU networks, if GF converges in direction to an optimal solution of Problem (7) (from the lemma), then the ratios between the Frobenius and the spectral norms in the weight matrices tend to be small. Formally, we have the following:

Theorem 5 *Let $\{(\mathbf{x}_i, y_i)\}_{i=1}^n \subseteq \mathbb{R}^{d_{in}} \times \{-1, 1\}$ be a binary classification dataset, and assume that there is $i \in [n]$ with $\|\mathbf{x}_i\| \leq 1$. Assume that there is a fully-connected neural network N of width $m \geq 2$ and depth $k \geq 2$, such that for all $i \in [n]$ we have $y_i N(\mathbf{x}_i) \geq 1$, and the weight matrices $W_1, \ldots, W_k$ of N satisfy $\|W_i\|_F \leq B$ for some $B > 0$. Let $N_{\boldsymbol{\theta}}$ be a fully-connected neural network of width $m' \geq m$ and depth $k' > k$ parameterized by $\boldsymbol{\theta}$. Let $\boldsymbol{\theta}^* = \left[W_1^*, \ldots, W_{k'}^*\right]$ be a global optimum of Problem (7). Namely, $\boldsymbol{\theta}^*$ parameterizes a minimum-norm fully-connected network of width m' and depth k' that labels the dataset correctly with margin 1. Then, we have*

$$\frac{1}{k'} \sum_{i=1}^{k'} \frac{\|W_i^*\|_\sigma}{\|W_i^*\|_F} \geq \frac{1}{\sqrt{2}} \cdot \left(\frac{\sqrt{2}}{B}\right)^{\frac{k}{k'}} \cdot \sqrt{\frac{k'}{k'+1}} \,. \tag{8}$$

Equivalently, we have the following upper bound on the harmonic mean of the ratios $\frac{\|W_i^*\|_F}{\|W_i^*\|_\sigma}$:

$$\frac{k'}{\sum_{i=1}^{k'} \left(\frac{\|W_i^*\|_F}{\|W_i^*\|_\sigma}\right)^{-1}} \leq \sqrt{2} \cdot \left(\frac{B}{\sqrt{2}}\right)^{\frac{k}{k'}} \cdot \sqrt{\frac{k'+1}{k'}} \,. \tag{9}$$

By the above theorem, if k' is much larger than k, then the average ratio between the spectral and the Frobenius norms (Eq. (8)) is at least roughly $1/\sqrt{2}$. Likewise, the harmonic mean of the ratio between the Frobenius and the spectral norms (Eq. (9)), i.e., the square root of the stable rank, is at most roughly $\sqrt{2}$. Since the dependence on k' in the theorem is exponential, increasing k' has a significant effect. Note that the result does not depend on the width of the networks. Namely, it holds even if the width m' is very large. Similarly to the case of Theorem 4, we note that the theorem implies that even though the data can be realized by many depth-k' networks, some with large ranks and some with low ranks, under the theorem's conditions, there is a bias towards networks with low ranks. The proof of the theorem follows similar ideas to the proof of Theorem 4, and is given in the appendix.

The combination of the above result with Lemma 1 suggests that, in overparameterized deep fully-connected networks, GF might tend to converge in direction to neural networks with low ranks. Note that we consider the exponential and the logistic losses, and hence if the loss tends to zero as $t \to \infty$, then we have $\|\boldsymbol{\theta}(t)\| \to \infty$. To conclude, in our case, the parameters tend to have an infinite norm and to converge in direction to a low-rank solution. Moreover, note that the ratio between the spectral and the Frobenius norms is invariant to scaling, and hence it suggests that after a sufficiently long time, GF tends to reach a network with low ranks.

Acknowledgments

This research is supported in part by European Research Council (ERC) grant 754705.

References

Jose M Alvarez and Mathieu Salzmann. Compression-aware training of deep networks. *Advances in neural information processing systems*, pages 856–867, 2017.

Sanjeev Arora, Rong Ge, Behnam Neyshabur, and Yi Zhang. Stronger generalization bounds for deep nets via a compression approach. In *International Conference on Machine Learning*, pages 254–263. PMLR, 2018.

Sanjeev Arora, Nadav Cohen, Wei Hu, and Yuping Luo. Implicit regularization in deep matrix factorization. In *Advances in Neural Information Processing Systems*, pages 7413–7424, 2019.

Shahar Azulay, Edward Moroshko, Mor Shpigel Nacson, Blake Woodworth, Nathan Srebro, Amir Globerson, and Daniel Soudry. On the implicit bias of initialization shape: Beyond infinitesimal mirror descent. In *International Conference on Machine Learning*, pages 468–477, 2021.

Mohamed Ali Belabbas. On implicit regularization: Morse functions and applications to matrix factorization. *arXiv preprint arXiv:2001.04264*, 2020.

Etienne Boursier, Loucas Pillaud-Vivien, and Nicolas Flammarion. Gradient flow dynamics of shallow relu networks for square loss and orthogonal inputs. *arXiv preprint arXiv:2206.00939*, 2022.

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

Zhen Dai, Mina Karzand, and Nathan Srebro. Representation costs of linear neural networks: Analysis and design. *Advances in Neural Information Processing Systems*, 34:26884–26896, 2021.

Emily L Denton, Wojciech Zaremba, Joan Bruna, Yann LeCun, and Rob Fergus. Exploiting linear structure within convolutional networks for efficient evaluation. In *Advances in neural information processing systems*, pages 1269–1277, 2014.

Simon S Du, Wei Hu, and Jason D Lee. Algorithmic regularization in learning deep homogeneous models: Layers are automatically balanced. *Advances in Neural Information Processing Systems*, pages 384–395, 2018.

Armin Eftekhari and Konstantinos Zygalakis. Implicit regularization in matrix sensing: A geometric view leads to stronger results. *arXiv preprint arXiv:2008.12091*, 2020.

Tolga Ergen and Mert Pilanci. Convex geometry and duality of over-parameterized neural networks. *Journal of machine learning research*, 2021a.

Tolga Ergen and Mert Pilanci. Revealing the structure of deep neural networks via convex duality. In *International Conference on Machine Learning*, pages 3004–3014. PMLR, 2021b.

Gauthier Gidel, Francis Bach, and Simon Lacoste-Julien. Implicit regularization of discrete gradient dynamics in linear neural networks. In *Advances in Neural Information Processing Systems*, pages 3202–3211, 2019.

Suriya Gunasekar, Jason D Lee, Daniel Soudry, and Nati Srebro. Implicit bias of gradient descent on linear convolutional networks. In *Advances in Neural Information Processing Systems*, pages 9461–9471, 2018a.

Suriya Gunasekar, Blake Woodworth, Srinadh Bhojanapalli, Behnam Neyshabur, and Nathan Srebro. Implicit regularization in matrix factorization. In *2018 Information Theory and Applications Workshop (ITA)*, pages 1–10. IEEE, 2018b.

Meena Jagadeesan, Ilya Razenshteyn, and Suriya Gunasekar. Inductive bias of multi-channel linear convolutional networks with bounded weight norm. In *Conference on Learning Theory*, pages 2276–2325. PMLR, 2022.

Ziwei Ji and Matus Telgarsky. Gradient descent aligns the layers of deep linear networks. In *International Conference on Learning Representations*, 2018.

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

Yuanzhi Li, Tengyu Ma, and Hongyang Zhang. Algorithmic regularization in over-parameterized matrix sensing and neural networks with quadratic activations. In *Conference On Learning Theory*, pages 2–47. PMLR, 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*, 2020.

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

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

Cong Ma, Kaizheng Wang, Yuejie Chi, and Yuxin Chen. Implicit regularization in nonconvex statistical estimation: Gradient descent converges linearly for phase retrieval and matrix completion. In *International Conference on Machine Learning*, pages 3345–3354. PMLR, 2018.

Behnam Neyshabur, Ryota Tomioka, and Nathan Srebro. In search of the real inductive bias: On the role of implicit regularization in deep learning. In *ICLR (Workshop)*, 2015.

Behnam Neyshabur, Srinadh Bhojanapalli, David McAllester, and Nati Srebro. Exploring generalization in deep learning. In *Advances in Neural Information Processing Systems*, pages 5947–5956, 2017.

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

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

Mert Pilanci and Tolga Ergen. Neural networks are convex regularizers: Exact polynomial-time convex optimization formulations for two-layer networks. In *International Conference on Machine Learning*, pages 7695–7705. PMLR, 2020.

Noam Razin and Nadav Cohen. Implicit regularization in deep learning may not be explainable by norms. *Advances in Neural Information Processing Systems*, 2020.

Noam Razin, Asaf Maman, and Nadav Cohen. Implicit regularization in tensor factorization. *arXiv preprint arXiv:2102.09972*, 2021.

Olivier Roy and Martin Vetterli. The effective rank: A measure of effective dimensionality. In *2007 15th European Signal Processing Conference*, pages 606–610. IEEE, 2007.

Mark Rudelson and Roman Vershynin. Sampling from large matrices: An approach through geometric functional analysis. *Journal of the ACM (JACM)*, 54(4):21–es, 2007.

Itay Safran, Gal Vardi, and Jason D Lee. On the effective number of linear regions in shallow univariate relu networks: Convergence guarantees and implicit bias. *arXiv preprint arXiv:2205.09072*, 2022.

Amartya Sanyal, Philip H Torr, and Puneet K Dokania. Stable rank normalization for improved generalization in neural networks and gans. In *International Conference on Learning Representations*, 2019.

Roei Sarussi, Alon Brutzkus, and Amir Globerson. Towards understanding learning in neural networks with linear teachers. In *International Conference on Machine Learning*, pages 9313–9322. PMLR, 2021.

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

Murad Tukan, Alaa Maalouf, Matan Weksler, and Dan Feldman. Compressed deep networks: Goodbye svd, hello robust low-rank approximation. *arXiv preprint arXiv:2009.05647*, 2020.

Gal Vardi. On the implicit bias in deep-learning algorithms. *arXiv preprint arXiv:2208.12591*, 2022.

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

Gal Vardi, Ohad Shamir, and Nathan Srebro. On margin maximization in linear and relu networks. *arXiv preprint arXiv:2110.02732*, 2021.

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

Xiyu Yu, Tongliang Liu, Xinchao Wang, and Dacheng Tao. On compressing deep models by low rank and sparse decomposition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 7370–7379, 2017.

Chulhee Yun, Shankar Krishnan, and Hossein Mobahi. A unifying view on implicit bias in training linear neural networks. In *International Conference on Learning Representations*, 2020.

Chiyuan Zhang, Samy Bengio, Moritz Hardt, Benjamin Recht, and Oriol Vinyals. Understanding deep learning requires rethinking generalization. In *International Conference on Learning Representations*, 2017.

Appendix A. Proof of Theorem 1

Consider a matrix $W \in \mathbb{R}^{2\times 2}$ whose rows $\mathbf{w}_1^\top, \mathbf{w}_2^\top$ satisfy

$$\begin{aligned}\mathbf{w}_1 &= \frac{\mathbf{x}_1}{\|\mathbf{x}_1\|} - \frac{\mathbf{x}_2}{\|\mathbf{x}_2\|},\\ \mathbf{w}_2 &= -\mathbf{w}_1.\end{aligned}$$

The matrix W has rank 1. To complete the proof, we need to show that we can choose a matrix $V \in \mathbb{R}^{2\times 2}$ such that $N_{W,V}$ attains zero loss. According to Lemma 2 below, it is enough to show that $\mathbf{w}_1^\top \mathbf{x}_1 > 0$ and $\mathbf{w}_1^\top \mathbf{x}_2 < 0$. Since the angle between the inputs is strictly positive, namely, $\measuredangle(\mathbf{x}_1, \mathbf{x}_2) > 0$, it holds that $\frac{\mathbf{x}_1^\top \mathbf{x}_2}{\|\mathbf{x}_1\|\cdot\|\mathbf{x}_2\|} < 1$. Thus,

$$\|\mathbf{x}_1\| \cdot \|\mathbf{x}_2\| - \mathbf{x}_1^\top \mathbf{x}_2 > 0\ .$$

Then,

$$\mathbf{w}_1^\top \mathbf{x}_1 = \frac{\mathbf{x}_1^\top \mathbf{x}_1}{\|\mathbf{x}_1\|} - \frac{\mathbf{x}_1^\top \mathbf{x}_2}{\|\mathbf{x}_2\|} = \|\mathbf{x}_1\| - \frac{\mathbf{x}_1^\top \mathbf{x}_2}{\|\mathbf{x}_2\|} = \frac{\|\mathbf{x}_1\| \cdot \|\mathbf{x}_2\| - \mathbf{x}_1^\top \mathbf{x}_2}{\|\mathbf{x}_2\|} > 0,$$

while

$$\mathbf{w}_1^\top \mathbf{x}_2 = \frac{\mathbf{x}_1^\top \mathbf{x}_2}{\|\mathbf{x}_1\|} - \frac{\mathbf{x}_2^\top \mathbf{x}_2}{\|\mathbf{x}_2\|} = \frac{\mathbf{x}_1^\top \mathbf{x}_2}{\|\mathbf{x}_1\|} - \|\mathbf{x}_2\| = \frac{\mathbf{x}_1^\top \mathbf{x}_2 - \|\mathbf{x}_1\| \cdot \|\mathbf{x}_2\|}{\|\mathbf{x}_1\|} < 0.$$

■

Lemma 2 *Let $(X, Y) \in \mathbb{R}^{d_{in}\times n} \times \mathbb{R}^{d_{out}\times n}$ be a labeled dataset. Let $W \in \mathbb{R}^{d_{hidden}\times d_{in}}$. Suppose that for every data point $\mathbf{x}_j$ there is at least one row $\mathbf{w}_i^\top$ in W such that $\mathbf{w}_i^\top \mathbf{x}_j > 0$, and $\mathbf{w}_i^\top \mathbf{x}_\ell \leq 0$ for all $\ell \neq j$. Then, there exists V such that $N_{W,V}(X) = Y$.*

Proof Consider the matrix $\sigma(WX)$ of size $d_{\text{hidden}} \times n$, where σ acts entrywise. Note that our assumption on W implies that $\text{rank}\left(\sigma(WX)\right) = n$. Thus, the $d_{\text{hidden}} \times d_{\text{hidden}}$ matrix $Z := \begin{bmatrix}\sigma(WX)^\dagger \\ 0\end{bmatrix}$ satisfies $Z\sigma(WX) = \begin{bmatrix}I_n \\ 0\end{bmatrix}$, where $A^\dagger$ denotes the Moore-Penrose inverse of a matrix A, and I_n is the $n \times n$ identity matrix. Hence, the matrix $M := \begin{bmatrix}Y & 0\end{bmatrix}$ of dimensions $d_{\text{out}} \times d_{\text{hidden}}$ yields $MZ\sigma(WX) = Y$. By setting $V := MZ$, the network $N_{W,V}$ achieves zero loss. Namely, $N_{W,V}(X) = Y$. ■

Appendix B. Proof of Theorem 2

Definition 1 *We define the following regions of interest:*

$$\mathcal{D} := \{\mathbf{w} \in \mathbb{R}^2 \mid \forall i \in \{1,2\}, \sigma(\mathbf{w}^\top \mathbf{x}_i) \leq 0\},$$
$$\mathcal{S} := \{\mathbf{w} \in \mathbb{R}^2 \mid \forall i \in \{1,2\}, \sigma(\mathbf{w}^\top \mathbf{x}_i) > 0\}.$$

Also, for $j \in \{1,2\}$ we define

$$\mathcal{S}_j := \{\mathbf{w} \in \mathbb{R}^2 \mid \sigma(\mathbf{w}^\top \mathbf{x}_j) > 0, \sigma(\mathbf{w}^\top \mathbf{x}_{3-j}) \leq 0\}\ .$$

The regions in the above definition appear in Figure 1. Note that each of the regions of Definition 1, denoted as $\mathcal{P}$, is disjoint from the others and satisfies $c \cdot p \in \mathcal{P}$ for all $p \in \mathcal{P}$ and $c \in \mathbb{R}$ where $c > 0$. Assumption 2 induces the following geometry: Each of the four regions $\mathcal{D}$, $\mathcal{S}_1$, $\mathcal{S}_2$ and $\mathcal{S}$ is nonempty, and any straight line on the plane that goes through the origin intersects exactly two regions: Either (i) the $\mathcal{S}$ and $\mathcal{D}$ regions, or (ii) one of the $\mathcal{S}_i$ regions and the $\mathcal{D}$ region, or (iii) the $\mathcal{S}_1 \setminus \partial\mathcal{S}_1$ and $\mathcal{S}_2 \setminus \partial\mathcal{S}_2$ regions.

Assume, for the sake of contradiction, that GF converges to some zero-loss network $N_{W(\infty),V(\infty)}$ with $\text{rank}(W(\infty)) < 2$. On the one hand, in Lemma 3 we show that the weight vectors $\mathbf{w}_1(\infty)$ and $\mathbf{w}_2(\infty)$ are non-zero, and satisfy $\mathbf{w}_2(\infty) = \alpha\mathbf{w}_1(\infty)$ with $\alpha < 0$. It implies that the straight line that connects $\mathbf{w}_1(\infty)$ and $\mathbf{w}_2(\infty)$, denoted as $\mathbf{w}_1\mathbf{w}_2$, goes through the origin. On the other hand, in Lemma 4 we show that $\mathbf{w}_i(\infty) \notin \mathcal{D}$ for every $i \in \{1,2\}$. In other words, $\mathbf{w}_1\mathbf{w}_2$ cannot intersect the $\mathcal{D} \setminus \{\mathbf{0}\}$ region. Thus, one neuron must lie in $\mathcal{S}_1 \setminus \partial\mathcal{S}_1$ and the other neuron in $\mathcal{S}_2 \setminus \partial\mathcal{S}_2$. W.l.o.g., let $\mathbf{w}_i(\infty) \in \mathcal{S}_i \setminus \partial\mathcal{S}_i$ for all $i \in \{1,2\}$. Therefore, by Lemma 6, it holds that $\measuredangle\big(\mathbf{w}_1(\infty), \mathbf{w}_2(\infty)\big) \in \Big[\pi - \measuredangle(\mathbf{x}_1, \mathbf{x}_2), \measuredangle(\mathbf{x}_1, \mathbf{x}_2) + 2\arcsin\frac{2\max_{i\in[2]}\|\mathbf{w}_i(0)\|}{\sqrt{3}}\Big)$. To complete the proof by contradiction, it remains to show that $\measuredangle\big(\mathbf{w}_1(\infty), \mathbf{w}_2(\infty)\big) < \pi$ so that $\mathbf{w}_2(\infty) \neq \alpha\mathbf{w}_1(\infty)$. Recall that we initialize the network such that $\|\mathbf{w}_i(0)\| < \frac{\sqrt{3}}{2}\cos\left(\frac{\measuredangle(\mathbf{x}_1,\mathbf{x}_2)}{2}\right) = \frac{\sqrt{3}}{2}\sin\left(\frac{\pi}{2} - \frac{\measuredangle(\mathbf{x}_1,\mathbf{x}_2)}{2}\right)$. Hence, $\measuredangle\big(\mathbf{w}_1(\infty), \mathbf{w}_2(\infty)\big) < \pi$, as required.

■

Lemma 3 *Let $(X, Y) \in \mathbb{R}^{2\times 2} \times \mathbb{R}^{2\times 2}$ be a labeled dataset that satisfies Assumption 1. Consider a zero-loss ReLU network $N_{W,V}$ where $W, V \in \mathbb{R}^{2\times 2}$ and $\text{rank}(W) < 2$. Then, the weight vectors $\mathbf{w}_1$ and $\mathbf{w}_2$ are non-zero, and satisfy $\mathbf{w}_2 = \alpha\mathbf{w}_1$ with $\alpha < 0$.*

Proof First, by Lemma 4 we have $\mathbf{w}_1 \neq \mathbf{0}$ and $\mathbf{w}_2 \neq \mathbf{0}$. Thus, $\text{rank}(W) > 0$. Since by our assumption we also have $\text{rank}(W) < 2$ then we must have $\text{rank}(W) = 1$. Hence, we can denote $\mathbf{w}_2 = \alpha \mathbf{w}_1$ for some $\alpha \in \mathbb{R}$ with $\alpha \neq 0$.

Now, we prove that $\alpha < 0$. Assume for the sake of contradiction that $\alpha > 0$. Then, we have $\sigma(\mathbf{w}_2^\top \mathbf{x}_j) = \alpha\sigma(\mathbf{w}_1^\top \mathbf{x}_j)$ for all $j \in [2]$. Thus, $\text{rank}(\sigma(WX)) \leq 1$. Therefore, $\text{rank}(V\sigma(WX)) \leq \min\{\text{rank}(V), \text{rank}(\sigma(WX))\} \leq 1$. Since by Assumption 1 we have $\text{rank}(Y) = 2$, then we conclude that $Y \neq V\sigma(WX)$, in contradiction to the zero-loss assumption. Therefore, $\alpha < 0$, as required. ∎

Lemma 4 *Let $(X, Y) \in \mathbb{R}^{2\times 2} \times \mathbb{R}^{2\times 2}$ be a labeled dataset that satisfies Assumption 1. Consider a zero-loss ReLU network $N_{W,V}$ where $W, V \in \mathbb{R}^{2\times 2}$. Then, we have $\mathbf{w}_i \notin \mathcal{D}$ for all $i \in \{1, 2\}$.*

Proof Assume that there is $i \in [2]$ such that $\mathbf{w}_i \in \mathcal{D}$. Hence, $\sigma(\mathbf{w}_i^\top \mathbf{x}_j) = 0$ for all $j \in [2]$. Thus, $\text{rank}(\sigma(WX)) \leq 1$. Therefore, $\text{rank}(V\sigma(WX)) \leq \min\{\text{rank}(V), \text{rank}(\sigma(WX))\} \leq 1$. Since by Assumption 1 we have $\text{rank}(Y) = 2$, then we conclude that $Y \neq V\sigma(WX)$, in contradiction to the zero-loss assumption. ∎

Lemma 5 *Let $(X, Y) \in \mathbb{R}^{2\times 2} \times \mathbb{R}^{2\times 2}$ be a labeled dataset that satisfies Assumption 1. Consider GF w.r.t. the loss function $L_{X,Y}(W, V)$ for $W, V \in \mathbb{R}^{2\times 2}$, and assume that it converges to a network $N_{W(\infty),V(\infty)}$. Suppose that there exist $i \in [2]$ and time $t \geq 0$ such that $\mathbf{w}_i(t) \in \mathcal{D}$. Then, we have $N_{W(\infty),V(\infty)}(X) \neq Y$.*

Proof Note that if $\mathbf{w}_i(t) \in \mathcal{D}$ then the gradient of $L_{X,Y}$ w.r.t. $\mathbf{w}_i$ is zero. Hence $\mathbf{w}_i$ remains constant for all $t' \geq t$. Therefore, $\mathbf{w}_i(\infty) \in \mathcal{D}$. The claim now follows from Lemma 4. ∎

Lemma 6 *Let (X, Y) be a labeled dataset that satisfies Assumptions 1 and 2. Consider GF w.r.t. the loss function $L_{X,Y}(W, V)$. Suppose that $W, V \in \mathbb{R}^{2\times 2}$ are initialized such that for all $i \in [2]$ we have $\|\mathbf{w}_i(0)\|, \|\mathbf{v}_i(0)\| < \frac{1}{2}$. If GF converges to a zero-loss network $N_{W(\infty),V(\infty)}$ such that $\mathbf{w}_i(\infty) \in \mathcal{S}_i \setminus \partial\mathcal{S}_i$ for all $i \in [2]$, then*

$$\|\mathbf{w}_i(\infty)\| \in \left(\frac{\sqrt{3}}{2}, \sqrt{\frac{1}{4} + \frac{4}{3\left(\cos\max\left\{\arcsin\frac{2\|\mathbf{w}_i(0)\|}{\sqrt{3}}, \measuredangle(\mathbf{x}_1, \mathbf{x}_2) - \frac{\pi}{2}\right\}\right)^2}}\right)$$

and

$$\measuredangle(\mathbf{w}_1(\infty), \mathbf{w}_2(\infty)) \in \left[\pi - \measuredangle(\mathbf{x}_1, \mathbf{x}_2), \measuredangle(\mathbf{x}_1, \mathbf{x}_2) + 2\arcsin\frac{2\max_{i\in[2]}\|\mathbf{w}_i(0)\|}{\sqrt{3}}\right).$$

Proof First, in Lemma 7 we show that GF induces a dynamic on each neuron $\mathbf{w}_i$ that lies in a $\mathcal{S}_i$ region, such that the neuron can only move in the direction of $\mathbf{x}_i$. Formally, for every $\mathbf{w}_i \in \mathcal{S}_i$ we have $\frac{d}{dt}\mathbf{w}_i(t) = c_t^{(i)}\mathbf{x}_i$, where $c_t^{(i)} \in \mathbb{R}$. We denote by $t_0^{(i)}$ the last time that $\mathbf{w}_i$ enters $\mathcal{S}_i$. That is,

$$t_0^{(i)} := \inf\{t \mid \mathbf{w}_i(t') \in \mathcal{S}_i \text{ for all } t' \geq t\} .$$

Thus,

$$\mathbf{w}_i(\infty) = \mathbf{w}_i(t_0^{(i)}) + C^{(i)}\mathbf{x}_i, \tag{10}$$

for some constant $C^{(i)} \in \mathbb{R}$. Note that since $\mathbf{w}_i(\infty) \in \mathcal{S}_i \setminus \partial\mathcal{S}_i$, then there exists some $t' \geq 0$ with $\mathbf{w}_i(t') \in \mathcal{S}_i$. We will further delimit the location of $\mathbf{w}_i(t_0^{(i)})$. There are only two cases for $t_0^{(i)}$:

Case $t_0^{(i)} = 0$: If the last time that $\mathbf{w}_i$ enters $\mathcal{S}_i$ is at initialization, then we have $t_0^{(i)} = 0$. Our assumptions on the initialization imply that:

$$\text{For } t_0^{(i)} = 0, \quad \mathbf{w}_i(t_0^{(i)}) \in \mathcal{T}_i := \mathcal{S}_i \cap (\operatorname{cl} B_2(\|\mathbf{w}_i(0)\|)) \ .$$

Note that by Lemma 5 it is not possible that $\mathbf{w}_i(0) \in \mathcal{D}$, and hence we cannot have $\mathbf{w}_i(0) \in \partial\mathcal{S}_i \cap \mathcal{D}$.

Otherwise (i.e., $t_0^{(i)} > 0$): In that case, $t_0^{(i)}$ is when the neuron moves from some other region to $\mathcal{S}_i$. The other region can only be $\mathcal{S}$ or $\mathcal{D}$, due to the geometry that Assumption 2 imposes. Since Lemma 5 implies that at any time no neuron is in $\mathcal{D}$, then the previous region is necessarily $\mathcal{S}$. Hence, we have:

$$\text{For } t_0^{(i)} > 0, \quad \mathbf{w}_i(t_0^{(i)}) \in \mathcal{U}_i := \partial(\mathcal{S}_i) \setminus \mathcal{D} \ .$$

In any case, we conclude that:

$$\mathbf{w}_i(t_0^{(i)}) \in \mathcal{E}_i := \mathcal{T}_i \cup \mathcal{U}_i \ .$$

Therefore, the region of all neurons that are reachable under the aforementioned dynamics of GF is

$$\mathbf{w}_i(\infty) \in \mathcal{A}_i := \{\mathbf{w} + \lambda\mathbf{x}_i \mid \mathbf{w} \in \mathcal{E}_i, \lambda \geq 0\}.$$

We can assume that $\lambda \geq 0$ in the above definition, because every $\bar{\mathbf{a}} \in \{\mathbf{w} + \lambda\mathbf{x}_i \mid \mathbf{w} \in \mathcal{E}_i, \lambda < 0\} \setminus \mathcal{A}_i$ satisfies $\bar{\mathbf{a}} \notin \mathcal{S}_i$.

We denote $\epsilon_0^{(i)} := \|\mathbf{w}_i(0)\|^2 - \|\mathbf{v}_i(0)\|^2$. By Lemma 9 we have $\epsilon_0^{(i)} = \|\mathbf{w}_i(t)\|^2 - \|\mathbf{v}_i(t)\|^2$ for any time $t \geq 0$, and hence $\epsilon_0^{(i)} = \|\mathbf{w}_i(\infty)\|^2 - \|\mathbf{v}_i(\infty)\|^2$. By Lemma 8 we obtain $\|\mathbf{w}_i(\infty)\| \geq \sqrt{1 - |\epsilon_0^{(i)}|}$ for every $i \in [2]$. We define a new region of interest: The set of all feasible neurons at the convergence of GF, i.e., neurons that are reachable and satisfy the minimal norm requirement. Formally,

$$\mathbf{w}_i(\infty) \in \mathcal{F}_i := \left\{\mathbf{w} \in \mathcal{A}_i \;\middle|\; \|\mathbf{w}\| \geq \sqrt{1 - |\epsilon_0^{(i)}|}\right\} = \mathcal{A}_i \setminus B_2\left(\sqrt{1 - |\epsilon_0^{(i)}|}\right) \ .$$

The regions $\mathcal{A}_i$ and $\mathcal{F}_i$ are illustrated in Figure 1. Recall that all neurons are initialized such that $\|\mathbf{w}_i(0)\|, \|\mathbf{v}_i(0)\| < \frac{1}{2}$ for all $i \in [2]$. Thus, we have $\left|\epsilon_0^{(i)}\right| < (\frac{1}{2})^2 = \frac{1}{4}$ for all $i \in [2]$. Hence,

$$\|\mathbf{w}_i(\infty)\| > \frac{\sqrt{3}}{2} \ , \tag{11}$$

as required.

We now consider the angle between $\mathbf{w}_1(\infty)$ and $\mathbf{w}_2(\infty)$. On the one hand, the minimal angle between the neurons is achieved when $\mathbf{w}_1(\infty)$ and $\mathbf{w}_2(\infty)$ lie on the "non-dead boundaries" of $\mathcal{S}_1, \mathcal{S}_2$. That is,

$$\measuredangle\big(\mathbf{w}_1(\infty), \mathbf{w}_2(\infty)\big) \geq \measuredangle(\mathbf{b}_1, \mathbf{b}_2) = \pi - \measuredangle(\mathbf{x}_1, \mathbf{x}_2) \ , \tag{12}$$

where $\mathbf{b}_i \in \partial(\mathcal{S}_i) \setminus \mathcal{D}$. On the other hand, the angle between the neurons is maximized when

$$\measuredangle\big(\mathbf{w}_1(\infty), \mathbf{w}_2(\infty)\big) = \measuredangle(\mathbf{x}_1, \mathbf{x}_2) + \sum_{i=1}^{2} \measuredangle\big(\mathbf{w}_i(\infty), \mathbf{x}_i\big) \ .$$

Note that in the above expression the angle $\measuredangle\big(\mathbf{w}_i(\infty), \mathbf{x}_i\big)$ corresponds to the case where $\mathbf{w}_i(\infty)$ is in the direction w.r.t. $\mathbf{x}_i$ which is closer to $\mathcal{D}$ and farther from $\mathcal{S}$. Due to Eq. (10) and the definition of $\mathcal{F}_i$, the appropriate angle $\measuredangle\big(\mathbf{w}_i(\infty), \mathbf{x}_i\big)$ in the above expression can be upper bounded by $\arcsin \frac{\|\mathbf{w}_i(0)\|}{\|\mathbf{w}_i(\infty)\|}$. It corresponds to the case where $\mathbf{w}_i$ is initialized in $\mathcal{S}_i$ such that $\measuredangle(\mathbf{w}_i(0), \mathbf{x}_i)$ is close to $\pi/2$, and $\mathbf{w}_i$ follows the trajectory from Eq. (10). Using Eq. (11) we have $\arcsin \frac{\|\mathbf{w}_i(0)\|}{\|\mathbf{w}_i(\infty)\|} < \arcsin \frac{2\|\mathbf{w}_i(0)\|}{\sqrt{3}}$. Hence, we get

$$\measuredangle\big(\mathbf{w}_1(\infty), \mathbf{w}_2(\infty)\big) < \measuredangle(\mathbf{x}_1, \mathbf{x}_2) + 2 \arcsin \frac{2 \max_{i \in [2]} \|\mathbf{w}_i(0)\|}{\sqrt{3}} \ .$$

Combining the above with Eq. (12) we obtain

$$\measuredangle\big(\mathbf{w}_1(\infty), \mathbf{w}_2(\infty)\big) \in \left[\pi - \measuredangle(\mathbf{x}_1, \mathbf{x}_2),\ \measuredangle(\mathbf{x}_1, \mathbf{x}_2) + 2 \arcsin \frac{2 \max_{i \in [2]} \|\mathbf{w}_i(0)\|}{\sqrt{3}}\right) \ .$$

Finally, we obtain an upper bound for $\|\mathbf{w}_i(\infty)\|$. We have

$$\mathbf{w}_i(\infty)^\top \mathbf{x}_i = \|\mathbf{w}_i(\infty)\| \cdot \|\mathbf{x}_i\| \cos \measuredangle\big(\mathbf{w}_i(\infty), \mathbf{x}_i\big) > \frac{\sqrt{3}}{2} \cos \measuredangle\big(\mathbf{w}_i(\infty), \mathbf{x}_i\big)$$

for all $i \in [2]$. Note that $\measuredangle\big(\mathbf{w}_i(\infty), \mathbf{x}_i\big)$ corresponds either to the case where $\mathbf{w}_i(\infty)$ is in the direction w.r.t. $\mathbf{x}_i$ which is closer to $\mathcal{D}$ and farther from $\mathcal{S}$, or closer to $\mathcal{S}$ and farther from $\mathcal{D}$. For the former case, we saw that $\measuredangle\big(\mathbf{w}_i(\infty), \mathbf{x}_i\big) < \arcsin \frac{2\|\mathbf{w}_i(0)\|}{\sqrt{3}}$. In the latter case, $\measuredangle\big(\mathbf{w}_i(\infty), \mathbf{x}_i\big) = \measuredangle\big(\mathbf{x}_1, \mathbf{x}_2\big) - \measuredangle\big(\mathbf{w}_i(\infty), \mathbf{x}_{3-i}\big) \leq \measuredangle\big(\mathbf{x}_1, \mathbf{x}_2\big) - \frac{\pi}{2}$. Therefore,

$$\mathbf{w}_i(\infty)^\top \mathbf{x}_i > \frac{\sqrt{3}}{2} \cos \max \left\{ \arcsin \frac{2\|\mathbf{w}_i(0)\|}{\sqrt{3}}, \measuredangle\big(\mathbf{x}_1, \mathbf{x}_2\big) - \frac{\pi}{2} \right\} .$$

Since the network has zero-loss, i.e., it interpolates the entire dataset, then we have that

$$\mathbf{v}_i(\infty) = \frac{1}{\mathbf{w}_i(\infty)^\top \mathbf{x}_i} \mathbf{y}_i \ .$$

Hence,

$$\|\mathbf{v}_i(\infty)\| = \frac{1}{\mathbf{w}_i(\infty)^\top \mathbf{x}_i} \|\mathbf{y}_i\| < \frac{2}{\sqrt{3} \cos \max \left\{ \arcsin \frac{2\|\mathbf{w}_i(0)\|}{\sqrt{3}}, \measuredangle\big(\mathbf{x}_1, \mathbf{x}_2\big) - \frac{\pi}{2} \right\}} \ .$$

By Lemma 9, we have $\|\mathbf{w}_i(\infty)\|^2 - \|\mathbf{v}_i(\infty)\|^2 = \|\mathbf{w}_i(0)\|^2 - \|\mathbf{v}_i(0)\|^2 < \frac{1}{4}$. Therefore,

$$\|\mathbf{w}_i(\infty)\|^2 < \frac{1}{4} + \frac{4}{3 \left(\cos \max \left\{ \arcsin \frac{2\|\mathbf{w}_i(0)\|}{\sqrt{3}}, \measuredangle\big(\mathbf{x}_1, \mathbf{x}_2\big) - \frac{\pi}{2} \right\} \right)^2}$$

as required. ■

Lemma 7 *Let $(X, Y) \in \mathbb{R}^{2\times 2} \times \mathbb{R}^{2\times 2}$ be a labeled dataset that satisfies Assumption 2. Consider GF on a ReLU network $N_{W,V}$ with $W, V \in \mathbb{R}^{2\times 2}$, w.r.t. $L_{X,Y}(W, V)$. Assume that at time t we have $\mathbf{w}_i(t) \in \mathcal{S}_i$ for some $i \in [2]$. Then, there exists $c_t^{(i)} \in \mathbb{R}$ such that $\frac{d}{dt}\mathbf{w}_i(t) = c_t^{(i)}\mathbf{x}_i$.*

Proof We have

$$\frac{d}{dt}\mathbf{w}_i(t) = -\frac{\partial}{\partial \mathbf{w}_i} L_{X,Y}(W(t), V(t)) .$$

The derivative of the $L_{X,Y}$ w.r.t. the matrix W is

$$\frac{\partial}{\partial W} L_{X,Y}(W, V) = \left(\sigma'(WX) \odot \left(V^\top \left(V\sigma(WX) - Y\right)\right)\right) X^\top .$$

Here, $\odot$ denotes the Hadamard product (i.e., the entrywise product). Note that $\frac{\partial L_{X,Y}\left(N_{W,V}\right)}{\partial W}$ is a matrix whose (i, j)-th entry is $\frac{\partial L_{X,Y}(W,V)}{\partial W_{i,j}}$. We denote the i-th row of $\sigma'(WX)$ by $\sigma'(WX)_i$. We have

$$\sigma'\left(WX\right)_i = \sigma'\left(\begin{bmatrix}\mathbf{w}_i^\top \mathbf{x}_1 & \mathbf{w}_i^\top \mathbf{x}_2\end{bmatrix}\right) = \begin{bmatrix}\sigma'(\mathbf{w}_i^\top \mathbf{x}_1) & \sigma'(\mathbf{w}_i^\top \mathbf{x}_2)\end{bmatrix} .$$

If $\mathbf{w}_i \in \mathcal{S}_i$ then the j-th entry of the aforementioned row vector is

$$\sigma'(WX)_{ij} = \mathbb{1}\{i = j\} .$$

Thus, there exists a constant $\alpha^{(i)} \in \mathbb{R}$ such that

$$\left(\sigma'(WX) \odot \left(V^\top \left(V\sigma(WX) - Y\right)\right)\right)_{ij} = \mathbb{1}\{i = j\}\alpha^{(i)} .$$

Since the derivative of the loss w.r.t. the i-th neuron $\mathbf{w}_i$ is the i-th row of $\frac{\partial}{\partial W} L_{X,Y}(W, V)$, we conclude that

$$\frac{\partial}{\partial \mathbf{w}_i} L_{X,Y}(W, V) = \alpha^{(i)}\mathbf{x}_i .$$

By setting $c_t^{(i)} = -\alpha^{(i)}$, the proof is done. ■

Lemma 8 *Let $(X, Y) \in \mathbb{R}^{2\times 2} \times \mathbb{R}^{2\times 2}$ be a labeled dataset that satisfies Assumptions 1 and 2. Let $N_{W,V}$ be a zero-loss network with $W, V \in \mathbb{R}^{2\times 2}$, such that $\mathbf{w}_i \in \mathcal{S}_i$ for all $i \in [2]$. Let $\epsilon^{(i)} := \|\mathbf{w}_i\|^2 - \|\mathbf{v}_i\|^2$. Then $\|\mathbf{w}_i\|, \|\mathbf{v}_i\| \geq \sqrt{1 - |\epsilon^{(i)}|}$ for all $i \in [2]$.*

Proof Since the network has zero loss, for all $i \in [2]$ we have

$$\mathbf{y}_i = V\sigma(W\mathbf{x}_i) = \begin{bmatrix}\sum_{k=1}^2 V_{1,k}\sigma(\mathbf{w}_k^\top \mathbf{x}_i) \\ \\ \sum_{k=1}^2 V_{2,k}\sigma(\mathbf{w}_k^\top \mathbf{x}_i)\end{bmatrix} .$$

Since $\mathbf{w}_i \in \mathcal{S}_i$ for every $i \in [2]$, we have $\sigma(\mathbf{w}_k^\top \mathbf{x}_i) = \begin{cases} \mathbf{w}_i^\top \mathbf{x}_i & \text{if } k = i \\ 0 & \text{otherwise} \end{cases}$. Hence, the above expression is equal to

$$\begin{bmatrix} V_{1,i} \cdot \mathbf{w}_i^\top \mathbf{x}_i \\ \\ V_{2,i} \cdot \mathbf{w}_i^\top \mathbf{x}_i \end{bmatrix} = \mathbf{v}_i(\mathbf{w}_i^\top \mathbf{x}_i) \ .$$

Therefore,

$$1 = \|\mathbf{y}_i\| = \|\mathbf{v}_i(\mathbf{w}_i^\top \mathbf{x}_i)\| = \|(\mathbf{v}_i \mathbf{w}_i^\top)\mathbf{x}_i\| \leq \|\mathbf{v}_i \mathbf{w}_i^\top\|_F \cdot \|\mathbf{x}_i\| = \|\mathbf{v}_i \mathbf{w}_i^\top\|_F = \|\mathbf{v}_i\| \cdot \|\mathbf{w}_i\| \ .$$

Now, there are two cases:
Case $\|\mathbf{w}_i\| \leq \|\mathbf{v}_i\|$: We have that $\|\mathbf{v}_i\|^2 \geq 1$. Then,

$$\|\mathbf{w}_i\| = \sqrt{\|\mathbf{v}_i\|^2 + \epsilon^{(i)}} \geq \sqrt{1 + \epsilon^{(i)}} = \sqrt{1 - |\epsilon^{(i)}|} \ .$$

Otherwise: Similarly, we have $\|\mathbf{w}_i\|^2 \geq 1$. Then,

$$\|\mathbf{v}_i\| = \sqrt{\|\mathbf{w}_i\|^2 - \epsilon^{(i)}} \geq \sqrt{1 - \epsilon^{(i)}} = \sqrt{1 - |\epsilon^{(i)}|} \ .$$

In any case, we conclude that

$$\|\mathbf{w}_i\|, \|\mathbf{v}_i\| \geq \sqrt{1 - |\epsilon^{(i)}|} \ .$$

■

Lemma 9 (Du et al. (2018)) *Let $N_{\boldsymbol{\theta}}$ be a fully-connected depth-k ReLU network, where $k > 1$. Denote $\boldsymbol{\theta} = [W^{(1)}, \ldots, W^{(k)}]$. Consider minimizing any differentiable loss function (e.g., the square loss) over a dataset using GF. Then, for every $l \in [k-1]$ at all time t we have*

$$\frac{d}{dt}\left(\left\|W^{(l)}(t)\right\|_F^2 - \left\|W^{(l+1)}(t)\right\|_F^2\right) = 0 \ .$$

Moreover, for every $l \in [k-1]$ and $i \in [d_l]$ at all time t we have

$$\frac{d}{dt}\left(\left\|W^{(l)}[i,:](t)\right\|^2 - \left\|W^{(l+1)}[:,i](t)\right\|^2\right) = 0 \ ,$$

where $W^{(l)}[i,:]$ is the vector of incoming weights to the i-th neuron in the l-th hidden layer (i.e., the i-th row of $W^{(l)}$), and $W^{(l+1)}[:,i]$ is the vector of outgoing weights from this neuron (i.e., the i-th column of $W^{(l+1)}$).

Appendix C. Proof of Theorem 3

Consider the partition of $\mathbb{R}^2$ into regions as described in Definition 1. If $\mathbf{w}_i \in \mathcal{S}_i \setminus \partial\mathcal{S}_i$ for all $i \in [2]$, then the gradient of $L_{X,Y}(W,V)$ is given by:

$$\frac{\partial}{\partial \mathbf{v}_i} L_{X,Y} = \left(\mathbf{w}_i^\top \mathbf{x}_i\right) \phi_i \,,$$
$$\frac{\partial}{\partial \mathbf{w}_i} L_{X,Y} = \mathbf{v}_i^\top \phi_i \mathbf{x}_i \,, \tag{13}$$

for all $i \in [2]$, where $\phi_i := (\mathbf{w}_i^\top \mathbf{x}_i)\mathbf{v}_i - \mathbf{y}_i = N_{W,V}(\mathbf{x}_i) - \mathbf{y}_i$. We denote the parameters of the network by $\boldsymbol{\theta} = [W, V]$. Moreover, when $\mathbf{w}_i \in \mathcal{S}_i \setminus \partial\mathcal{S}_i$ for all $i \in [2]$ we denote $L^i_{X,Y}(\boldsymbol{\theta}) = \frac{1}{2}\|\phi_i\|^2$. Then, we have $L_{X,Y}(\boldsymbol{\theta}) = \sum_{i=1}^2 L^i_{X,Y}(\boldsymbol{\theta})$.

Lemma 10 *Let $t_1 > 0$ and suppose that for all $t \in [0, t_1]$ and $i \in [2]$ we have $\mathbf{w}_i(t) \in \mathcal{S}_i \setminus \partial\mathcal{S}_i$, and that $\mathbf{v}_i(0) = \mathbf{0}$. Then, we have $L^i_{X,Y}(\boldsymbol{\theta}(t_1)) < L^i_{X,Y}(\boldsymbol{\theta}(0))$. Moreover, for every time t where $\mathbf{w}_i(t) \in \mathcal{S}_i \setminus \partial\mathcal{S}_i$ for all $i \in [2]$ we have $\frac{d}{dt} L^i_{X,Y}(\boldsymbol{\theta}(t)) \leq 0$.*

Proof For time t such that $\mathbf{w}_i(t) \in \mathcal{S}_i \setminus \partial\mathcal{S}_i$ for all $i \in [2]$ we denote $F_i(t) := L^i_{X,Y}(\boldsymbol{\theta}(t)) = \frac{1}{2}\|\phi_i(t)\|^2$. Let $\boldsymbol{\theta}_i := [\mathbf{w}_i, \mathbf{v}_i]$. We have

$$\begin{aligned}
\frac{d}{dt} F_i(t) &= \left(\nabla_{\boldsymbol{\theta}} L^i_{X,Y}(\boldsymbol{\theta}(t))\right)^\top \frac{d\boldsymbol{\theta}(t)}{dt} \\
&= \left(\nabla_{\boldsymbol{\theta}_i} L^i_{X,Y}(\boldsymbol{\theta}(t))\right)^\top \frac{d\boldsymbol{\theta}_i(t)}{dt} \\
&= -\left(\nabla_{\boldsymbol{\theta}_i} L^i_{X,Y}(\boldsymbol{\theta}(t))\right)^\top \nabla_{\boldsymbol{\theta}_i} L_{X,Y}(\boldsymbol{\theta}(t)) \\
&= -\left(\nabla_{\boldsymbol{\theta}_i} L^i_{X,Y}(\boldsymbol{\theta}(t))\right)^\top \nabla_{\boldsymbol{\theta}_i} L^i_{X,Y}(\boldsymbol{\theta}(t)) \\
&= -\left\|\nabla_{\boldsymbol{\theta}_i} L^i_{X,Y}(\boldsymbol{\theta}(t))\right\|^2 \,,
\end{aligned} \tag{14}$$

where we used the fact that $L^i_{X,Y}(\boldsymbol{\theta})$ depends only on $\boldsymbol{\theta}_i$. Therefore, $\frac{d}{dt} L^i_{X,Y}(\boldsymbol{\theta}(t)) \leq 0$.

Note that $\left\|\nabla_{\boldsymbol{\theta}_i} L^i_{X,Y}(\boldsymbol{\theta}(t))\right\|^2$ is continuous as a function of t, and at time 0 we have

$$\begin{aligned}
\left\|\nabla_{\boldsymbol{\theta}_i} L^i_{X,Y}(\boldsymbol{\theta}(0))\right\| &= \|\nabla_{\boldsymbol{\theta}_i} L_{X,Y}(\boldsymbol{\theta}(0))\| \geq \left\|\frac{\partial}{\partial \mathbf{v}_i} L_{X,Y}(\boldsymbol{\theta}(0))\right\| = \left\|(\mathbf{w}_i^\top(0)\mathbf{x}_i)\phi_i(0)\right\| \\
&= \left\|(\mathbf{w}_i^\top(0)\mathbf{x}_i)\left((\mathbf{w}_i^\top(0)\mathbf{x}_i)\mathbf{v}_i(0) - \mathbf{y}_i\right)\right\| = \left\|(\mathbf{w}_i^\top(0)\mathbf{x}_i)\left(-\mathbf{y}_i\right)\right\| > 0 \,,
\end{aligned}$$

where the last inequality is since $\mathbf{w}_i^\top(0)\mathbf{x}_i > 0$ and $\mathbf{y}_i \neq \mathbf{0}$. Combining the above with Eq. (14), we conclude that there is some small enough $t_0 \in (0, t_1)$ such that for all $t \in [0, t_0]$ we have $\frac{d}{dt} F_i(t) < 0$. Moreover, Eq. (14) implies that for all $t \in [t_0, t_1]$ we have $\frac{d}{dt} F_i(t) \leq 0$. Hence, $F_i(t_1) \leq F_i(t_0) < F_i(0)$. ■

Lemma 11 *Suppose that we initialize $\boldsymbol{\theta}(0)$ such that $\mathbf{w}_i(0) \in \mathcal{S}_i \setminus \partial\mathcal{S}_i$ and $\mathbf{v}_i(0) = \mathbf{0}$ for all $i \in [2]$. For every sufficiently small $t' > 0$ we have for every $t \in [0, t']$ and $i \in [2]$ that $\mathbf{w}_i(t) \in \mathcal{S}_i \setminus \partial S_i$, and at time t' we have $\mathbf{v}_i^\top(t')\phi_i(t') < 0$ and $\mathbf{v}_i(t') \in \operatorname{span}\{\mathbf{y}_i\}$. Moreover, $L^i_{X,Y}(\boldsymbol{\theta}(t')) < L^i_{X,Y}(\boldsymbol{\theta}(0))$.*

Proof Since $\mathbf{w}_i(0)$ is in the open set $\mathcal{S}_i \setminus \partial\mathcal{S}_i$ for all $i \in [2]$, then for every small enough $t > 0$ we have $\mathbf{w}_i(t) \in \mathcal{S}_i \setminus \partial\mathcal{S}_i$. Also, by Lemma 10, for every small enough $t > 0$ we have $L^i_{X,Y}(\boldsymbol{\theta}(t)) < L^i_{X,Y}(\boldsymbol{\theta}(0))$. Let $\tilde{t}$ be such that the two conditions above hold for all $t \in \left(0, \tilde{t}\,\right]$.

Let $g_i(t) := \mathbf{v}_i^\top(t)\phi_i(t)$. We show that for every small enough $0 < t' < \tilde{t}$ we have $g_i(t') < 0$. First, note that $g_i(0) = 0$ since $\mathbf{v}_i(0) = \mathbf{0}$. Moreover, $g_i(t)$ is continuously differentiable, and satisfies

$$\frac{d}{dt}g_i(t) = \left(\frac{d}{dt}\mathbf{v}_i^\top(t)\right)\phi_i(t) + \mathbf{v}_i^\top(t)\left(\frac{d}{dt}\phi_i(t)\right) = -\left(\mathbf{w}_i^\top(t)\mathbf{x}_i\right)\phi_i^\top(t)\phi_i(t) + \mathbf{v}_i^\top(t)\left(\frac{d}{dt}\phi_i(t)\right) .$$

Therefore,

$$\frac{d}{dt}g_i(0) = -\left(\mathbf{w}_i^\top(0)\mathbf{x}_i\right)\|\phi_i(0)\|^2 + \mathbf{v}_i^\top(0)\left(\frac{d}{dt}\phi_i(0)\right) = -\left(\mathbf{w}_i^\top(0)\mathbf{x}_i\right)\|\phi_i(0)\|^2 .$$

Since $\mathbf{w}_i(0) \in \mathcal{S}_i \setminus \partial\mathcal{S}_i$ then $\mathbf{w}_i^\top(0)\mathbf{x}_i > 0$ and hence we obtain $\frac{d}{dt}g_i(0) < 0$.

Overall, the function g_i is continuously differentiable with $g_i(0) = 0$ and $\frac{d}{dt}g_i(0) < 0$ and therefore we have $g_i(t') < 0$ for every small enough $t' > 0$.

It remains to show that $\mathbf{v}_i(t') \in \mathrm{span}\{\mathbf{y}_i\}$. Since for every $t \in [0, t']$ we have $\mathbf{w}_i(t) \in \mathcal{S}_i \setminus \partial\mathcal{S}_i$, then for every $t \in [0, t']$ we have

$$\frac{d}{dt}\mathbf{v}_i(t) = -(\mathbf{w}_i^\top(t)\mathbf{x}_i)\phi_i(t) = -(\mathbf{w}_i^\top(t)\mathbf{x}_i)\left((\mathbf{w}_i^\top(t)\mathbf{x}_i)\mathbf{v}_i(t) - \mathbf{y}_i\right) \in \mathrm{span}\{\mathbf{v}_i(t), \mathbf{y}_i\} .$$

Since the above holds for all $t \in [0, t']$ and $\mathbf{v}_i(0) = \mathbf{0}$, then for all $t \in [0, t']$ we have $\mathbf{v}_i(t) \in \mathrm{span}\{\mathbf{y}_i\}$. Thus, $\mathbf{v}_i$ remains on the line $\mathrm{span}\{\mathbf{y}_i\}$. ■

Lemma 12 *Suppose that we initialize $\boldsymbol{\theta}(0)$ such that $\mathbf{w}_i(0) \in \mathcal{S}_i \setminus \partial\mathcal{S}_i$ and $\mathbf{v}_i(0) = \mathbf{0}$ for all $i \in [2]$. Let $t' > 0$ as in Lemma 11, and denote $\mathbf{w}'_i := \mathbf{w}_i(t')$ for $i \in [2]$. Let*

$$\begin{aligned} G := \big\{\boldsymbol{\theta} \ :\ & \text{for all } i \in [2] \text{ we have} \\ & \mathbf{w}_i = \mathbf{w}'_i + c_i\mathbf{x}_i \text{ for } c_i \geq 0 \,, \\ & \mathbf{v}_i \in \mathrm{span}\{\mathbf{y}_i\} \,, \\ & L^i_{X,Y}(W,V) \leq L^i_{X,Y}(W(t'),V(t')) < L^i_{X,Y}(W(0),V(0)) \,, \\ & \mathbf{v}_i^\top\phi_i \leq 0\big\} \,. \end{aligned}$$

Then, for all $t \geq t'$ we have $\boldsymbol{\theta}(t) \in G$.

Moreover, for all $t_2 \geq t_1 \geq t'$ and all $i \in [2]$ we have

$$\mathbf{w}_i^\top(t_2)\mathbf{x}_i \geq \mathbf{w}_i^\top(t_1)\mathbf{x}_i > 0 .$$

Proof By Lemma 11 we have $\boldsymbol{\theta}(t') \in G$. Let $t \geq t'$ and suppose that $\boldsymbol{\theta}(t) \in G$. Note that for all $i \in [2]$ we have $\mathbf{w}_i(t) = \mathbf{w}'_i + c_i(t)\mathbf{x}_i$ for some $c_i(t) \geq 0$. Since $\mathbf{w}'_i \in \mathcal{S}_i \setminus \partial\mathcal{S}_i$ then we also have $\mathbf{w}_i(t) \in \mathcal{S}_i \setminus \partial\mathcal{S}_i$. Hence,

$$\frac{d}{dt}\mathbf{w}_i(t) = -\frac{\partial}{\partial\mathbf{w}_i}L_{X,Y}(\boldsymbol{\theta}(t)) = -\mathbf{v}_i^\top(t)\phi_i(t)\mathbf{x}_i . \tag{15}$$

Since by the definition of G we have $\mathbf{v}_i^\top(t)\phi_i(t) \leq 0$ then the above can be written as $c_i'(t)\mathbf{x}_i$ for some $c_i'(t) \geq 0$. Moreover,

$$\frac{d}{dt}\mathbf{v}_i(t) = -\frac{\partial}{\partial \mathbf{v}_i} L_{X,Y}(\boldsymbol{\theta}(t)) = -(\mathbf{w}_i^\top(t)\mathbf{x}_i)\phi_i(t) = -(\mathbf{w}_i^\top(t)\mathbf{x}_i)\left((\mathbf{w}_i^\top(t)\mathbf{x}_i)\mathbf{v}_i(t) - \mathbf{y}_i\right) .$$

Since by the definition of G we have $\mathbf{v}_i(t) \in \mathrm{span}\{\mathbf{y}_i\}$, then the above is also in $\mathrm{span}\{\mathbf{y}_i\}$.

Moreover, by Lemma 10 we have $\frac{d}{dt} L^i_{X,Y}(\boldsymbol{\theta}(t)) \leq 0$.

The above observations imply that as long as $\mathbf{v}_i^\top(t)\phi_i(t) \leq 0$ the parameters $\mathbf{w}_i(t)$ and $\mathbf{v}_i(t)$ satisfy the conditions in G. We now show that if $\mathbf{v}_i^\top(t)\phi_i(t) = 0$ then $\frac{d}{dt}\mathbf{w}_i(t) = \frac{d}{dt}\mathbf{v}_i(t) = \mathbf{0}$, and hence GF will get stuck at $\mathbf{w}_i(t), \mathbf{v}_i(t)$. Thus, GF cannot reach $\mathbf{w}_i, \mathbf{v}_i$ with $\mathbf{v}_i^\top \phi_i > 0$.

Suppose that $\mathbf{v}_i^\top(t)\phi_i(t) = 0$, $\mathbf{v}_i(t) \in \mathrm{span}\{\mathbf{y}_i\}$, and $L^i_{X,Y}(\boldsymbol{\theta}(t)) \leq L^i_{X,Y}(\boldsymbol{\theta}(t')) < L^i_{X,Y}(\boldsymbol{\theta}(0))$. Note that $\mathbf{v}_i(t) \neq \mathbf{0}$, since otherwise we have

$$\begin{aligned} L^i_{X,Y}(\boldsymbol{\theta}(t)) &= \frac{1}{2}\|\phi_i(t)\|^2 = \frac{1}{2}\left\|(\mathbf{w}_i^\top(t)\mathbf{x}_i)\mathbf{0} - \mathbf{y}_i\right\|^2 = \frac{1}{2}\left\|(\mathbf{w}_i^\top(0)\mathbf{x}_i)\mathbf{0} - \mathbf{y}_i\right\|^2 \\ &= \frac{1}{2}\left\|(\mathbf{w}_i^\top(0)\mathbf{x}_i)\mathbf{v}_i(0) - \mathbf{y}_i\right\|^2 = L^i_{X,Y}(\boldsymbol{\theta}(0)) , \end{aligned}$$

in contradiction to our assumption. Now, since $\mathbf{v}_i(t) \in \mathrm{span}\{\mathbf{y}_i\}$, then $\phi_i(t) = (\mathbf{w}_i^\top(t)\mathbf{x}_i)\mathbf{v}_i(t) - \mathbf{y}_i \in \mathrm{span}\{\mathbf{y}_i\}$. Thus, both $\mathbf{v}_i(t)$ and $\phi_i(t)$ are in $\mathrm{span}\{\mathbf{y}_i\}$, and we have $\mathbf{v}_i(t) \neq \mathbf{0}$ and $\mathbf{v}_i^\top(t)\phi_i(t) = 0$. Therefore, $\phi_i(t) = \mathbf{0}$. By Eq. (13) it implies that $\frac{d}{dt}\mathbf{w}_i(t) = \frac{d}{dt}\mathbf{v}_i(t) = \mathbf{0}$.

Thus, $\boldsymbol{\theta}(t) \in G$ for all $t \geq t'$. It remains to show that for all $t_2 \geq t_1 \geq t'$ and all $i \in [2]$ we have $\mathbf{w}_i^\top(t_2)\mathbf{x}_i \geq \mathbf{w}_i^\top(t_1)\mathbf{x}_i$. By Eq. (15) and since $\mathbf{v}_i^\top(t)\phi_i(t) \leq 0$ for all $t \geq t'$, we can write $\mathbf{w}_i(t_1) = \mathbf{w}_i' + \gamma_1\mathbf{x}_i$ and $\mathbf{w}_i(t_2) = \mathbf{w}_i' + \gamma_2\mathbf{x}_i$ where $\gamma_2 \geq \gamma_1 \geq 0$. Therefore

$$\mathbf{w}_i^\top(t_2)\mathbf{x}_i = \mathbf{w}_i'^\top\mathbf{x}_i + \gamma_2\|\mathbf{x}_i\|^2 \geq \mathbf{w}_i'^\top\mathbf{x}_i + \gamma_1\|\mathbf{x}_i\|^2 = \mathbf{w}_i^\top(t_1)\mathbf{x}_i > 0 .$$

■

Lemma 13 *Suppose that we initialize $\boldsymbol{\theta}(0)$ such that $\mathbf{w}_i(0) \in \mathcal{S}_i \backslash \partial\mathcal{S}_i$ and $\mathbf{v}_i(0) = \mathbf{0}$ for all $i \in [2]$. Then, GF converges (i.e., $W(\infty)$ and $V(\infty)$ exist) and $L_{X,Y}(W(\infty), V(\infty)) = 0$. Moreover $\mathbf{w}_i(\infty) \in \mathcal{S}_i \setminus \partial\mathcal{S}_i$ for all $i \in [2]$*

Proof By Lemma 12, there is $t' > 0$ such that for all $i \in [2]$ and $t \geq t'$ we have $\mathbf{w}_i(t) = \mathbf{w}_i(t') + c_i(t)\mathbf{x}_i$ for $c_i(t) \geq 0$. Hence, $\mathbf{w}_i(t) \in \mathcal{S}_i \setminus \partial\mathcal{S}_i$ for all $t \geq t'$. We have $\frac{d}{dt}L_{X,Y}(\boldsymbol{\theta}(t)) = (\nabla L_{X,Y}(\boldsymbol{\theta}(t)))^\top \frac{d}{dt}\boldsymbol{\theta}(t) = -\|\nabla L_{X,Y}(\boldsymbol{\theta}(t))\|^2$. Hence, for $T \geq t'$ we have

$$L_{X,Y}(\boldsymbol{\theta}(T)) = L_{X,Y}(\boldsymbol{\theta}(t')) + \int_{t=t'}^{T} \frac{d}{dt}L_{X,Y}(\boldsymbol{\theta}(t))dt = L_{X,Y}(\boldsymbol{\theta}(t')) - \int_{t=t'}^{T} \|\nabla L_{X,Y}(\boldsymbol{\theta}(t))\|^2 dt .$$

Therefore,

$$\int_{t=t'}^{T} \|\nabla L_{X,Y}(\boldsymbol{\theta}(t))\|^2 dt = L_{X,Y}(\boldsymbol{\theta}(t')) - L_{X,Y}(\boldsymbol{\theta}(T)) \leq L_{X,Y}(\boldsymbol{\theta}(t')) .$$

Since it holds for every $T \geq t'$, then we have

$$\int_{t=t'}^{\infty} \|\nabla L_{X,Y}(\boldsymbol{\theta}(t))\|^2 dt \leq L_{X,Y}(\boldsymbol{\theta}(t')) < \infty . \tag{16}$$

Moreover, since $\mathbf{w}_i(t) \in \mathcal{S}_i \setminus \partial\mathcal{S}_i$ for all $i \in [2]$ and $t \geq t'$, then by Eq. (13) we have

$$L_{X,Y}(\boldsymbol{\theta}(t)) = \frac{1}{2}\sum_{i=1}^{2} \|\phi_i(t)\|^2 = \frac{1}{2}\sum_{i=1}^{2} \frac{1}{\left(\mathbf{w}_i^\top(t)\mathbf{x}_i\right)^2} \left\| \frac{\partial}{\partial \mathbf{v}_i} L_{X,Y}(\boldsymbol{\theta}(t)) \right\|^2$$
$$\leq \left(\frac{1}{2}\sum_{i=1}^{2} \frac{1}{\left(\mathbf{w}_i^\top(t)\mathbf{x}_i\right)^2} \right) \|\nabla L_{X,Y}(\boldsymbol{\theta}(t))\|^2 .$$

By Lemma 12 we have $\left(\mathbf{w}_i^\top(t)\mathbf{x}_i\right)^2 \geq \left(\mathbf{w}_i^\top(t')\mathbf{x}_i\right)^2$. Therefore

$$L_{X,Y}(\boldsymbol{\theta}(t)) \leq \left(\frac{1}{2}\sum_{i=1}^{2} \frac{1}{\left(\mathbf{w}_i^\top(t')\mathbf{x}_i\right)^2} \right) \|\nabla L_{X,Y}(\boldsymbol{\theta}(t))\|^2 .$$

Letting $K := \frac{1}{2}\sum_{i=1}^{2} \frac{1}{\left(\mathbf{w}_i^\top(t')\mathbf{x}_i\right)^2}$ and combining the above with Eq. (16), we get

$$\frac{1}{K}\int_{t=t'}^{\infty} L_{X,Y}(\boldsymbol{\theta}(t))dt < \infty .$$

Since $L_{X,Y}(\boldsymbol{\theta}(t))$ is non-negative, and since by Lemma 10 it is monotonically non-increasing as a function of t, then we conclude that $\lim_{t\to\infty} L_{X,Y}(\boldsymbol{\theta}(t)) = 0$.

It remains to show that $\boldsymbol{\theta}(\infty)$ exists, namely, that GF converges. Since $\mathbf{w}_i(t) \in \mathcal{S}_i \setminus \partial\mathcal{S}_i$ for all $t \geq t'$ and $\lim_{t\to\infty} L_{X,Y}(\boldsymbol{\theta}(t)) = 0$, then $\lim_{t\to\infty} L^i_{X,Y}(\boldsymbol{\theta}(t)) = 0$ for all $i \in [2]$. That is, $(\mathbf{w}_i^\top(t)\mathbf{x}_i)\mathbf{v}_i(t) \to \mathbf{y}_i$ as $t \to \infty$. By Lemma 12 we can write $\mathbf{w}_i(t) = \mathbf{w}'_i + a_i(t)\mathbf{x}_i$ and $\mathbf{v}_i(t) = b_i(t)\mathbf{y}_i$, for some $a_i(t), b_i(t)$ with $a_i(t) \geq 0$ for all t. Since $\mathbf{w}_i^\top(t)\mathbf{x}_i > 0$ and $(\mathbf{w}_i^\top(t)\mathbf{x}_i)\mathbf{v}_i(t) \to \mathbf{y}_i$ then we also have $b_i(t) > 0$ for large enough t.

By Lemma 9, $\|\mathbf{v}_i(t)\|^2 - \|\mathbf{w}_i(t)\|^2$ remains constant throughout the training. Hence, we can write

$$b_i(t) = \|b_i(t)\mathbf{y}_i\| = \|\mathbf{v}_i(t)\| = \sqrt{C + \|\mathbf{w}_i(t)\|^2} = \sqrt{C + \|\mathbf{w}'_i + a_i(t)\mathbf{x}_i\|^2} ,$$

for some constant C. Therefore,

$$(\mathbf{w}_i^\top(t)\mathbf{x}_i)\mathbf{v}_i(t) = \left(\mathbf{w}_i'^\top\mathbf{x}_i + a_i(t)\|\mathbf{x}_i\|^2\right)\sqrt{C + \|\mathbf{w}'_i + a_i(t)\mathbf{x}_i\|^2} \cdot \mathbf{y}_i .$$

Since $(\mathbf{w}_i^\top(t)\mathbf{x}_i)\mathbf{v}_i(t) \to \mathbf{y}_i$, then we conclude that for

$$g_i(a) := \left(\mathbf{w}_i'^\top\mathbf{x}_i + a\|\mathbf{x}_i\|^2\right)\sqrt{C + \|\mathbf{w}'_i + a\mathbf{x}_i\|^2}$$

we have $\lim_{t\to\infty} g_i(a_i(t)) = 1$. The function $g_i(a)$ on $[0,\infty)$ is continuous and strictly increasing, and $\lim_{a\to\infty} g(a) = \infty$. Also, $g(0) \leq 1$ since otherwise we cannot have $\lim_{t\to\infty} g_i(a_i(t)) = 1$. Thus, there is exactly one point $a'_i \geq 0$ such that $g(a'_i) = 1$, and we have $\lim_{t\to\infty} a_i(t) = a'_i$. Hence, $\mathbf{w}_i(\infty)$ and $\mathbf{v}_i(\infty)$ exist. Moreover, $\mathbf{w}_i(\infty) = \mathbf{w}'_i + a'_i\mathbf{x}_i \in \mathcal{S}_i \setminus \partial\mathcal{S}_i$. ■

Proof [Proof of Theorem 3] We denote

$$\mathcal{W} = \left\{ W :\ \|\mathbf{w}_i\| \in \left(\frac{\sqrt{3}}{2}, \sqrt{\frac{1}{4} + \frac{4}{3\left(\cos\max\left\{\arcsin\frac{2\|\mathbf{w}_i(0)\|}{\sqrt{3}}, \measuredangle(\mathbf{x}_1,\mathbf{x}_2) - \frac{\pi}{2}\right\}\right)^2}} \right) \ \forall i \in [2], \right.$$
$$\left. \text{and } \measuredangle\left(\mathbf{w}_1,\mathbf{w}_2\right) \in \left[\pi - \measuredangle(\mathbf{x}_1,\mathbf{x}_2), \measuredangle(\mathbf{x}_1,\mathbf{x}_2) + 2\arcsin\frac{2\max_{i\in[2]}\|\mathbf{w}_i(0)\|}{\sqrt{3}}\right] \right\} .$$

By Lemma 13 if we initialize $\mathbf{v}_i(0) = \mathbf{0}$ and $\mathbf{w}_i(0) \in \mathcal{S}_i \setminus \partial\mathcal{S}_i$ for all $i \in [2]$, then GF converges and we have $L_{X,Y}(\boldsymbol{\theta}(\infty)) = 0$ and $\mathbf{w}_i(\infty) \in \mathcal{S}_i \setminus \partial\mathcal{S}_i$ for all $i \in [2]$. Also, by our assumption we have

$$\|\mathbf{w}_i(0)\| \leq \frac{\sqrt{3}}{2} \sin\left(\frac{\pi - \measuredangle(\mathbf{x}_1, \mathbf{x}_2)}{4}\right) < \frac{\sqrt{3}}{2} \sin\left(\frac{\pi}{8}\right) < \frac{1}{2} \,.$$

Therefore, by Lemma 6, $W(\infty) \in \mathcal{W}$. From the same arguments, $W(\infty) \in \mathcal{W}$ also if the initialization of $\mathbf{w}_i$ is such that $\mathbf{w}_i(0) \in \mathcal{S}_{3-i} \setminus \partial\mathcal{S}_{3-i}$ for all $i \in [2]$. Hence,

$$\begin{aligned}
&\Pr\left[W(\infty) \in \mathcal{W} \text{ and } L_{X,Y}(\boldsymbol{\theta}(\infty)) = 0\right] \\
&\quad\geq \Pr\left[\mathbf{w}_i(0) \in \mathcal{S}_i \setminus \partial\mathcal{S}_i \; \forall i \in [2] \text{ or } \mathbf{w}_i(0) \in \mathcal{S}_{3-i} \setminus \partial\mathcal{S}_{3-i} \; \forall i \in [2]\right] \\
&\quad= 2 \cdot \Pr\left[\mathbf{w}_i(0) \in \mathcal{S}_i \setminus \partial\mathcal{S}_i \; \forall i \in [2]\right] \\
&\quad= 2 \cdot \frac{\alpha(\mathcal{S}_1)}{2\pi} \cdot \frac{\alpha(\mathcal{S}_2)}{2\pi} \,,
\end{aligned} \tag{17}$$

where $\alpha(\mathcal{S}_i)$ is the angle that corresponds to the region $\mathcal{S}_i$. Formally, the angle of a region $\mathcal{S}_i$ is defined by $\alpha(\mathcal{S}_i) = \measuredangle(\mathbf{a}_1, \mathbf{a}_2)$ where $\mathbf{a}_1, \mathbf{a}_2 \in \partial\mathcal{S}_i$ are linearly independent.

Let $\mathbf{s}_i \in (\partial\mathcal{S}_i) \cap (\partial\mathcal{S})$ and let $\mathbf{d}_i \in (\partial\mathcal{S}_i) \cap (\partial\mathcal{D})$. Note that $\measuredangle(\mathbf{s}_i, \mathbf{x}_i) = \measuredangle(\mathbf{x}_1, \mathbf{x}_2) - \frac{\pi}{2}$ and that $\measuredangle(\mathbf{d}_i, \mathbf{x}_i) = \frac{\pi}{2}$. Thus,

$$\alpha(\mathcal{S}_i) = \measuredangle(\mathbf{s}_i, \mathbf{d}_i) = \measuredangle(\mathbf{s}_i, \mathbf{x}_i) + \measuredangle(\mathbf{d}_i, \mathbf{x}_i) = \measuredangle(\mathbf{x}_1, \mathbf{x}_2) - \frac{\pi}{2} + \frac{\pi}{2} = \measuredangle(\mathbf{x}_1, \mathbf{x}_2) \,.$$

Combining the above with Eq. (17) we get

$$\Pr\left[W(\infty) \in \mathcal{W} \text{ and } L_{X,Y}(\boldsymbol{\theta}(\infty)) = 0\right] \geq 2 \cdot \left(\frac{\measuredangle(\mathbf{x}_1, \mathbf{x}_2)}{2\pi}\right)^2 \,.$$

Finally, since

$$\|\mathbf{w}_i(0)\| \leq \min\left\{\frac{\sqrt{3}}{2} \sin\left(\frac{\pi - \measuredangle(\mathbf{x}_1, \mathbf{x}_2)}{4}\right), \frac{\sqrt{3}}{2} \sin\left(\measuredangle(\mathbf{x}_1, \mathbf{x}_2) - \frac{\pi}{2}\right)\right\} \,,$$

then $W(\infty) \in \mathcal{W}$ implies that for all $i \in [2]$ we have

$$\begin{aligned}
\|\mathbf{w}_i(\infty)\| &\in \left(\frac{\sqrt{3}}{2}, \sqrt{\frac{1}{4} + \frac{4}{3\left(\cos\max\left\{\arcsin\frac{2 \cdot \frac{\sqrt{3}}{2} \sin\left(\measuredangle(\mathbf{x}_1, \mathbf{x}_2) - \frac{\pi}{2}\right)}{\sqrt{3}}, \measuredangle\left(\mathbf{x}_1, \mathbf{x}_2\right) - \frac{\pi}{2}\right\}\right)^2}}\right) \\
&= \left(\frac{\sqrt{3}}{2}, \sqrt{\frac{1}{4} + \frac{4}{3\left(\cos(\measuredangle\left(\mathbf{x}_1, \mathbf{x}_2\right) - \frac{\pi}{2})\right)^2}}\right) \\
&= \left(\frac{\sqrt{3}}{2}, \sqrt{\frac{1}{4} + \frac{4}{3\left(\sin\measuredangle\left(\mathbf{x}_1, \mathbf{x}_2\right)\right)^2}}\right) \,,
\end{aligned}$$

and

$$\measuredangle\left(\mathbf{w}_1(\infty), \mathbf{w}_2(\infty)\right) \in \left[\pi - \measuredangle(\mathbf{x}_1, \mathbf{x}_2), \measuredangle(\mathbf{x}_1, \mathbf{x}_2) + 2\arcsin\frac{2 \cdot \frac{\sqrt{3}}{2}\sin\left(\frac{\pi - \measuredangle(\mathbf{x}_1,\mathbf{x}_2)}{4}\right)}{\sqrt{3}}\right]$$
$$= \left[\pi - \measuredangle(\mathbf{x}_1, \mathbf{x}_2), \measuredangle(\mathbf{x}_1, \mathbf{x}_2) + \frac{\pi - \measuredangle(\mathbf{x}_1, \mathbf{x}_2)}{2}\right] .$$

■

Appendix D. Proof of Theorem 4

Let $\alpha = \left(\frac{1}{B}\right)^{\frac{k'-k}{k'}}$. Consider the following fully-connected network N' of width m and depth k'. The weight matrices of layers $i \in [k]$ in N' are $W'_i = \alpha W_i$. Note that the k-th layer in N' contains a single neuron, and that since the weights in the first k layers of N' are obtained from the weights of N by scaling with the parameter α, then for every input $\mathbf{x}_i$ in the dataset the input to the neuron in layer k in N' is $\alpha^k \cdot N(\mathbf{x}_i) = \alpha^k y_i \geq 0$. The layers $i \in \{k+1, \ldots, k'\}$ in N' are of width 1. Hence, their weight matrices are of dimension 1×1. We define these weights by $W'_i = \beta$ for $\beta := \left(\frac{1}{B}\right)^{-\frac{k}{k'}}$. Thus, for an input $\mathbf{x}_i$ we have

$$N'(\mathbf{x}_i) = \alpha^k y_i \cdot \beta^{k'-k} = y_i \left(\frac{1}{B}\right)^{\frac{k'-k}{k'} \cdot k} \left(\frac{1}{B}\right)^{-\frac{k}{k'} \cdot (k'-k)} = y_i \ .$$

Let $\boldsymbol{\theta}' = \left[W'_1, \ldots, W'_{k'}\right]$ be the parameters of N'. Let $N^* := N_{\boldsymbol{\theta}^*}$ be the network with the parameters $\boldsymbol{\theta}^*$ that achieves a global optimum of Problem (3). Since the network N' is of depth k' and width $m \leq m'$ and since the network N^* is a global optimum, then we have $\|\boldsymbol{\theta}^*\| \leq \|\boldsymbol{\theta}\|$. Therefore,

$$\begin{aligned}
\|\boldsymbol{\theta}^*\|^2 &\leq \left\|\boldsymbol{\theta}'\right\|^2 \\
&= \left(\sum_{i=1}^{k} \alpha^2 \|W_i\|_F^2\right) + (k'-k)\beta^2 \\
&\leq \alpha^2 B^2 k + \beta^2 (k'-k) \\
&= \left(\frac{1}{B^2}\right)^{\frac{k'-k}{k'}} B^2 k + \left(\frac{1}{B^2}\right)^{-\frac{k}{k'}} (k'-k) \\
&= \left(\frac{1}{B^2}\right)^{-\frac{k}{k'}} k' \ .
\end{aligned} \tag{18}$$

In the following lemma, we show that since N^* is a global optimum of Eq. (3), then its layers must be balanced:

Lemma 14 *For every $1 \leq i < j \leq k'$ we have $\|W_i^*\|_F = \left\|W_j^*\right\|_F$.*

Proof Let $1 \leq i < j \leq k'$. For $\gamma > 0$ we define a network N_γ which is obtained from N^* as follows. The network N_γ is obtained by multiplying the weight matrix W_i^* by γ, and the weight matrix W_j^* by $1/\gamma$. Note that for every input $\mathbf{x}$ we have $N_\gamma(\mathbf{x}) = N^*(\mathbf{x})$.

We have

$$\frac{d}{d\gamma}\left(\|\gamma W_i^*\|_F^2 + \left\|\frac{1}{\gamma}W_j^*\right\|_F^2\right) = 2\gamma\|W_i^*\|_F^2 - \frac{2}{\gamma^3}\|W_j^*\|_F^2 .$$

When $\gamma = 1$ the above expression equals $2\|W_i^*\|_F^2 - 2\left\|W_j^*\right\|_F^2$. Hence, if $\|W_i^*\|_F \neq \left\|W_j^*\right\|_F$ then the derivative at $\gamma = 1$ is non-zero, in contradiction to the optimality of N^*. ■

By the above lemma, there is $B^* > 0$ such that $B^* = \|W_i^*\|_F$ for all $i \in [k']$. By Eq. (18) we have

$$(B^*)^2 \cdot k' = \|\boldsymbol{\theta}^*\|^2 \leq \left(\frac{1}{B^2}\right)^{-\frac{k}{k'}} k' .$$

Hence, for every $i \in [k']$ we have

$$\|W_i^*\|_F^2 = (B^*)^2 \leq \left(\frac{1}{B^2}\right)^{-\frac{k}{k'}} . \tag{19}$$

Moreover, since there is $i \in [n]$ with $\|\mathbf{x}_i\| \leq 1$ and $y_i \geq 1$, then the network N^* satisfies

$$1 \leq y_i = N^*(\mathbf{x}_i) \leq \|\mathbf{x}_i\| \prod_{i \in [k']} \|W_i^*\|_\sigma \leq \prod_{i \in [k']} \|W_i^*\|_\sigma \leq \left(\frac{1}{k'}\sum_{i \in [k']} \|W_i^*\|_\sigma\right)^{k'} ,$$

where the last inequality follows from the AM-GM inequality. Therefore, we have

$$\frac{1}{k'}\sum_{i \in [k']} \|W_i^*\|_\sigma \geq 1 .$$

Combining the above with Eq. (19) we get

$$\frac{1}{k'}\sum_{i \in [k']} \frac{\|W_i^*\|_\sigma}{\|W_i^*\|_F} = \frac{1}{B^*} \cdot \frac{1}{k'}\sum_{i \in [k']} \|W_i^*\|_\sigma \geq \left(\frac{1}{B}\right)^{\frac{k}{k'}} .$$

■

Appendix E. Proof of Theorem 5

Let $\alpha = \left(\frac{\sqrt{2}}{B}\right)^{\frac{k'-k}{k'}}$. Consider the following fully-connected network N' of width m and depth k'. The weight matrices of layers $i \in [k-1]$ in N' are $W_i' = \alpha W_i$. Let $\mathbf{u}$ be the weight vector of the output neuron in N. The k-th layer in N' is defined by the weight matrix $W_k' = \alpha \cdot \begin{bmatrix} \mathbf{u}^\top \\ -\mathbf{u}^\top \end{bmatrix}$.

That is, the k-th layer in N' has two neurons: the first neuron corresponds to the output neuron of N, and the second neuron to its negation. Note that since the weights in N' are obtained from the weights of N by scaling with the parameter α, then for every input $\mathbf{x}$ the input to the first neuron in layer k in N' is $\alpha^k \cdot N(\mathbf{x})$, and the input to the second neuron in layer k is $-\alpha^k \cdot N(\mathbf{x})$. The layers $i \in \{k+1, \dots, k'-1\}$ in N' are defined by the weight matrices $W'_i = \beta I_2$, where $\beta := \left(\frac{\sqrt{2}}{B}\right)^{-\frac{k}{k'}}$ and I_2 is the identity matrix of dimension 2. Finally, the k'-th layer in N' is defined by the weight vector $\beta \cdot \begin{pmatrix} 1 \\ -1 \end{pmatrix}$. Note that given an input $\mathbf{x}$, the first k layers in N' compute $\begin{pmatrix} \sigma\left(\alpha^k \cdot N(\mathbf{x})\right) \\ \sigma\left(-\alpha^k \cdot N(\mathbf{x})\right) \end{pmatrix}$, then the next $k'-k-1$ layers compute $\begin{pmatrix} \beta^{k'-k-1}\sigma\left(\alpha^k \cdot N(\mathbf{x})\right) \\ \beta^{k'-k-1}\sigma\left(-\alpha^k \cdot N(\mathbf{x})\right) \end{pmatrix}$, and finally the last layer returns

$$
\begin{aligned}
\beta^{k'-k}\sigma\left(\alpha^k \cdot N(\mathbf{x})\right) - \beta^{k'-k}\sigma\left(-\alpha^k \cdot N(\mathbf{x})\right) &= \beta^{k'-k}\alpha^k \cdot N(\mathbf{x}) \\
&= \left(\frac{\sqrt{2}}{B}\right)^{-\frac{k}{k'}\cdot(k'-k)} \cdot \left(\frac{\sqrt{2}}{B}\right)^{\frac{k'-k}{k'}\cdot k} \cdot N(\mathbf{x}) \\
&= N(\mathbf{x}) \; .
\end{aligned}
$$

Thus, $N'(\mathbf{x}) = N(\mathbf{x})$.

Let $\boldsymbol{\theta}' = \left[W'_1, \dots, W'_{k'}\right]$ be the parameters of N'. Let $N^* := N_{\boldsymbol{\theta}^*}$ be the network with the parameters $\boldsymbol{\theta}^*$ that achieves a global optimum of Problem (7). Since the network N' is of depth k' and width $m \le m'$ and since the network N^* is a global optimum, then we have $\|\boldsymbol{\theta}^*\| \le \|\boldsymbol{\theta}\|$. Therefore,

$$
\begin{aligned}
\|\boldsymbol{\theta}^*\|^2 &\le \left\|\boldsymbol{\theta}'\right\|^2 \\
&= \left(\sum_{i=1}^{k-1} \alpha^2 \|W_i\|_F^2\right) + \alpha^2\left(2\|W_k\|_F^2\right) + (k'-k-1)\beta^2 \cdot 2 + \beta^2 \cdot 2 \\
&\le \alpha^2(k-1)B^2 + \alpha^2 \cdot 2B^2 + \left(2(k'-k-1)+2\right)\beta^2 \\
&= \alpha^2 B^2 (k+1) + \beta^2 \cdot 2(k'-k) \\
&= \left(\frac{2}{B^2}\right)^{\frac{k'-k}{k'}} B^2(k+1) + \left(\frac{2}{B^2}\right)^{-\frac{k}{k'}} \cdot 2(k'-k) \\
&= 2 \cdot \left(\frac{2}{B^2}\right)^{-\frac{k}{k'}} (k+1) + \left(\frac{2}{B^2}\right)^{-\frac{k}{k'}} \cdot 2(k'-k) \\
&= 2 \cdot \left(\frac{2}{B^2}\right)^{-\frac{k}{k'}} (k'+1) \; .
\end{aligned} \tag{20}
$$

The following lemma shows that since N^* is a global optimum of Eq. (7), then its layers must be balanced:

Lemma 15 *For every $1 \le i < j \le k'$ we have $\|W_i^*\|_F = \left\|W_j^*\right\|_F$.*

The proof of the lemma is similar to the proof of Lemma 14. By the lemma, there is $B^* > 0$ such that $B^* = \|W_i^*\|_F$ for all $i \in [k']$. By Eq. (20) we have

$$(B^*)^2 \cdot k' = \|\boldsymbol{\theta}^*\|^2 \leq 2 \cdot \left(\frac{2}{B^2}\right)^{-\frac{k}{k'}} (k'+1) \;.$$

Hence, for every $i \in [k']$ we have

$$\|W_i^*\|_F^2 = (B^*)^2 \leq 2 \cdot \left(\frac{2}{B^2}\right)^{-\frac{k}{k'}} \cdot \frac{k'+1}{k'} \;. \tag{21}$$

Moreover, since there is $i \in [n]$ with $\|\mathbf{x}_i\| \leq 1$ and $|y_i| = 1$, then the network N^* satisfies

$$1 \leq y_i N^*(\mathbf{x}_i) \leq |N^*(\mathbf{x}_i)| \leq \|\mathbf{x}_i\| \prod_{i \in [k']} \|W_i^*\|_\sigma \leq \prod_{i \in [k']} \|W_i^*\|_\sigma \leq \left(\frac{1}{k'} \sum_{i \in [k']} \|W_i^*\|_\sigma\right)^{k'} \;,$$

where the last inequality follows from the AM-GM inequality. Therefore, we have

$$\frac{1}{k'} \sum_{i \in [k']} \|W_i^*\|_\sigma \geq 1 \;.$$

Combining the above with Eq. (21) we get

$$\begin{aligned}
\frac{1}{k'} \sum_{i \in [k']} \frac{\|W_i^*\|_\sigma}{\|W_i^*\|_F} &= \frac{1}{B^*} \cdot \frac{1}{k'} \sum_{i \in [k']} \|W_i^*\|_\sigma \\
&\geq \left(2 \cdot \left(\frac{2}{B^2}\right)^{-\frac{k}{k'}} \cdot \frac{k'+1}{k'}\right)^{-1/2} \cdot 1 \\
&= \frac{1}{\sqrt{2}} \cdot \left(\frac{\sqrt{2}}{B}\right)^{\frac{k}{k'}} \cdot \sqrt{\frac{k'}{k'+1}} \;.
\end{aligned}$$

■