On the Effective Number of Linear Regions in Shallow Univariate ReLU Networks: Convergence Guarantees and Implicit Bias

Itay Safran*
Princeton University

Gal Vardi*
TTI Chicago and Hebrew University

Jason D. Lee
Princeton University

Abstract

We study the dynamics and implicit bias of gradient flow (GF) on univariate ReLU neural networks with a single hidden layer in a binary classification setting. We show that when the labels are determined by the sign of a target network with r neurons, with high probability over the initialization of the network and the sampling of the dataset, GF converges in direction (suitably defined) to a network achieving perfect training accuracy and having at most $\mathcal{O}(r)$ linear regions, implying a generalization bound. Unlike many other results in the literature, under an additional assumption on the distribution of the data, our result holds even for mild over-parameterization, where the width is $\tilde{\mathcal{O}}(r)$ and independent of the sample size.

1 Introduction

Over-parameterized neural networks are known to be easier to train compared to their smaller counterparts, despite the resulting increase in the problem's dimensionality and the required computational resources [Daniely, 2017, Allen-Zhu et al., 2018, Safran and Shamir, 2018, Du et al., 2018, 2019, Ji et al., 2019, Zou et al., 2020, Li et al., 2020, Safran et al., 2021, Zhou et al., 2021]. However what is perhaps more surprising, is that in stark contrast to our classic understanding of generalization in machine learning models, this does not seem to degrade the generalization capabilities of the learned model in spite of the significant increase in its capacity. It is widely believed that what plays a key role in explaining this phenomenon is what is commonly referred to in the literature as *implicit bias/regularization* [Neyshabur et al., 2014, Zhang et al., 2021], where the optimization algorithm used inadvertently exhibits a bias towards empirical minimizers with a certain property that might induce better generalization. For example, such properties may include having a small norm or quasi-norm of the weights (e.g. Neyshabur [2017], Neyshabur et al. [2017], Lyu and Li [2019], Woodworth et al. [2020], Ji and Telgarsky [2020]) or a low rank solution (e.g. Razin and Cohen [2020]).

In this work, we study the dynamics and the implicit bias of GF on univariate ReLU neural networks, with the underlying assumption that the labels are determined by the sign of a target network of width r (which thus changes sign between -1 and 1 at most r times). Our analysis reveals that under the assumption of an i.i.d., normally-distributed initialization of the weights and biases of the network, over-parameterization (i.e. the use of width strictly larger than r) is necessary for attaining a small population loss. Moreover, if we also assume that the width scales at least linearly (up to logarithmic terms, excluding dependence on the confidence parameter) with the length of the shortest interval on which the labels do not change sign, then with high probability over the initialization of the network and the sampling of the data, over-parameterization is sufficient for driving the empirical loss to be

*Equal contribution

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

small enough so that all data instances are classified correctly. Thereafter, by analyzing the implicit bias of GF as the time t tends to infinity, we show that we converge in direction (see Section 2 for a formal definition) to a network having at most $\mathcal{O}(r)$ linear regions. Since the minimal number of neurons required to express a network with an arbitrarily small loss in general is r, this demonstrates that the implicit bias of GF in our setting is such that optimization converges to a solution which effectively has the optimal number of neurons up to a constant factor, which also provides a clear geometric interpretation with an intuitive generalization bound. This is in contrast to norm-based results where the characterization of the implicit bias in function space is less clear. Overall, our analysis provides an end-to-end result on the learnability of univariate ReLU neural networks with respect to GF in a binary classification setting.

The remainder of this paper is structured as follows: After specifying our contributions in more detail below, we turn to discuss related work. In Section 2, we present our notations and assumptions used throughout the paper before we present our main theorem. In Section 3 we formally present our optimization result. In Section 4 we turn to analyze the implicit bias of GF in our setting. Finally, in Section 5 we show that the implicit bias leads to a generalization bound.

Our contributions

- We prove that when training a sufficiently wide network which is initialized appropriately (Assumption 2.1) and under a suitable assumption on the distribution of the data (Assumption 2.2), then with high probability there exists some time t_0 where GF attains training error at most $\frac{1}{2n}$ on a size-n sample (Theorem 3.1). Our width requirement depends on the length of the shortest interval where the classificiation does not change sign, which in certain cases requires only mild over-parameterization with far fewer parameters in the model compared to observations in the dataset (see Remark 2.1).
- We show that if GF achieves training error smaller than $\frac{1}{n}$ at some time t_0, then it converges to zero loss and converges in direction (see Section 2 for a formal definition) to a network with at most $\mathcal{O}(r)$ linear regions (Theorem 4.2). Since this result provides a simple and intuitive geometric interpretation to the implicit bias of GF in our setting, it readily translates to a generalization bound using a standard argument (Corollary 5.1). We point out that unlike our optimization guarantee, this result holds regardless of the initialization of the network.
- Finally, we combine our optimization and generalization results to derive our main theorem (Theorem 2.3), which establishes an end-to-end learnability result for GF in our setting. This indicates that a wide model facilitates optimization, yet at the same time implicit bias prevents us from overfitting, even if the architecture we train is far wider than what is necessary. The result holds in the *rich regime*, and thus provides a guarantee which goes beyond the analysis achieved using NTK-based results (see Remark 2.3).
- As an additional contribution, we show (under our assumption on the initialization) that width at least $1.3r$ is necessary for GF to attain population loss less than an absolute constant over a particular target function (Theorem 3.2). Along with our previous results, this shows that over-parameterization is not just sufficient, but also necessary for successful learning.

We now turn to discuss some of the related work in the literature which is most relevant to ours in more detail.

Related work

Implicit bias in neural networks. The literature on the implicit bias in neural networks has rapidly expanded in recent years, and cannot be reasonably surveyed here (see Vardi [2022] for a survey). In what follows, we discuss only results which apply to depth-2 ReLU networks.

By Lyu and Li [2019], Ji and Telgarsky [2020] homogeneous neural networks (and specifically depth-2 ReLU networks) trained with exponentially-tailed classification losses converge in direction to a KKT point of the maximum-margin problem. Our analysis of the implicit bias relies on this result. We note that the aforementioned KKT point may not be a global optimum (see a discussion in Section 4). For depth-2 ReLU networks trained with the square loss there are no known guarantees on the implicit bias (cf. Vardi and Shamir [2021], Timor et al. [2022]).

Several works in recent years studied the implication of minimizing the ℓ_2 norm of the weights on the function space in depth-2 univariate ReLU networks. In Savarese et al. [2019] and Ergen and Pilanci [2021a] it is shown that a minimal-norm fit for a sample is given by the linear spline interpolation (i.e., a "connect-the-dots" function). In such linear spline interpolation the number of linear regions is small. The former work considered only regression, while the latter considered both regression and classification. Note that margin-maximization is equivalent to norm-minimization with margin at least 1. Thus, our result can also be viewed as an analysis of the implication of the bias towards norm-minimization on the learned function. We emphasize three important differences between the results from Savarese et al. [2019], Ergen and Pilanci [2021a] and ours:

1. As we already mentioned, the result of Lyu and Li [2019], Ji and Telgarsky [2020] implies a certain bias towards margin maximization, but it does not guarantee convergence to a global optimum (or even to a local optimum) of the maximum-margin problem. The only guarantee is that GF converges to a KKT point. Our result relies only on convergence to such a KKT point, and (unlike Savarese et al. [2019] and Ergen and Pilanci [2021a]) it does not assume convergence to a global optimum. As a result, we are able to obtain provable generalization bounds for GF.
2. In Savarese et al. [2019] and Ergen and Pilanci [2021a] it is shown that the linear spline interpolation minimizes the weights' norms. However, they also show that it is not a unique minimizer. Thus, in addition to the linear spline interpolation there are also other networks that fit the training set and minimize the norms. Therefore, even under the assumption that the weights' norms are minimized, their results do not guarantee convergence to a function with a small number of linear regions (as in our result).
3. Savarese et al. [2019] and Ergen and Pilanci [2021a] consider norm-minimization of the weights without the bias terms, while the implicit bias towards margin-maximization due to Lyu and Li [2019], Ji and Telgarsky [2020] (which we rely on) is w.r.t. all the parameters, including the bias terms. Hence, the implicit bias in depth-2 ReLU networks with exponentially-tailed losses does not minimize the norms in the sense considered in Savarese et al. [2019], Ergen and Pilanci [2021a].

We note that Ergen and Pilanci [2021b] showed that linear spline interpolators minimize the norms also in deep univariate networks. The result from Savarese et al. [2019] was extended to multivariate functions in Ongie et al. [2019]. Parhi and Nowak [2020] studied the relation between norm minimization and spline interpolation for a broader family of activation functions. Hanin [2021] gave a geometric characterization of all depth-2 univariate ReLU networks with a single linear unit, that minimize the ℓ_2 norm of the weights (excluding bias terms) and interpolate a given dataset (in a regression setting). Blanc et al. [2020] studied the relation between the implicit bias of SGD in depth-2 univariate ReLU networks (in a regression setting) and the number of convexity changes of the learned network. Maennel et al. [2018] showed that for a given training dataset there are only finitely many functions that GF with small initialization may converge to in depth-2 ReLU networks, independent of the network size. Chizat and Bach [2020] studied the dynamics of GF on infinite width depth-2 networks with exponentially-tailed losses and showed bias towards margin maximization w.r.t. a certain function norm known as the variation norm. Phuong and Lampert [2020] studied the implicit bias in depth-2 ReLU networks trained on orthogonally separable data.

Convergence of gradient methods under extreme over-parameterization. In recent years, many theoretical works have focused on providing convergence guarantees for training depth-2 neural networks with non-linear activations. Andoni et al. [2014] provide a convergence guarantee for learning polynomials of degree r in d-dimensional space in a regression setting using networks of width roughly d^{2r}. Since their architecture excludes bias terms which can be simulated by incrementing the input dimension by 1, their result in fact requires width 2^{2r} in our univariate setting, whereas our width requirement is typically much milder. Following the success of the NTK [Jacot et al., 2018], a spate of papers provided convergence guarantees when training using GD (e.g. [Allen-Zhu et al., 2018, Du et al., 2018, 2019, Ji et al., 2019, Zou et al., 2020]). The main difference that sets our work apart is that our width requirement is given in terms of the complexity of the teacher network, irrespective of the sample size n, whereas these works require that the width scales polynomially with n, which could be significantly larger. Moreover, as mentioned earlier, such results operate in the lazy regime where the features that are learned are dictated mainly by the initialization rather than the training process, whereas our analysis enters the rich regime once the

loss becomes sufficiently small, and provides a result that goes beyond NTK-based analyses. On the flip side, our analysis only holds for binary classification in the one-dimensional setting. Similarly to us, Soltanolkotabi et al. [2018] provide convergence guarantees by establishing that the objective function satisfies the PL-condition (see Polyak [1963]), however unlike our optimization guarantee and similarly to previously discussed works, their result requires that the network has more trainable parameters than data instances. Chizat et al. [2019] establish that by scaling a model appropriately, we can effectively interpolate between the lazy and the rich regime. While we use this observation in our optimization result to drive the loss to become sufficiently small in the first stage of optimization, our implicit bias result nevertheless operates in the rich regime regardless of this scaling.

Teacher-student setting and mild over-parameterization. In this paper, we assume that the labels of the data are determined by the sign of a teacher network of width r. Such a similar teacher-student setting but for a regression problem allowed the study of mild over-parameterization in quite a few recent works. Safran and Shamir [2018], Arjevani and Field [2020, 2021] show the existence of spurious (non-global) local minima in the loss landscape in this setting. Other works provide certain recovery guarantees; assuming absolute value activations, Li et al. [2020] provide a global convergence guarantee to loss at most $o(1/r)$, and Safran et al. [2021], Zhou et al. [2021] provide local convergence guarantees for ReLU or absolute value activations. These works require width at least poly(r) for convergence, whereas in our setting we show that width $\tilde{\mathcal{O}}(r)$ suffices in certain cases, and that over-parameterization is also necessary for successful optimization under our assumptions. While our results might superficially seem to provide stronger guarantees, we stress that the seemingly stronger bounds we derive are made possible in part due to the different assumptions made which include a univariate domain with biases compared to a multivariate domain with no biases nor output layer weights, thus highlighting the difference between the two architectures. In light of this, we argue that the bounds in these results are not directly comparable to ours.

2 Preliminaries and main result

Notations. We use bold-face letters to denote vectors, e.g., $\mathbf{x} = (x_1, \ldots, x_d)$. For $\mathbf{x} \in \mathbb{R}^d$ we denote by $\|\mathbf{x}\|$ the Euclidean norm. We denote by $\mathbb{1}[\cdot]$ the indicator function, for example $\mathbb{1}[t \geq 5]$ equals 1 if $t \geq 5$ and 0 otherwise. We denote $\mathrm{sign}(z) = 1$ if $z > 0$ and -1 otherwise. For an integer $d \geq 1$ we denote $[d] = \{1, \ldots, d\}$. We use standard asymptotic notation $\mathcal{O}(\cdot)$ to hide constant factors. A function $f : D \to \mathbb{R}$ which is twice continuously differentiable in a domain $D \subseteq \mathbb{R}^d$ is said to satisfy the *PL-condition* if there exists $\lambda > 0$ such that $\frac{1}{2}\|\nabla f(\mathbf{x})\|^2 \geq \lambda(f(\mathbf{x}) - f^*)$, where $f^* := \inf_{\mathbf{x}} f(\mathbf{x})$.

Neural networks. The ReLU activation function is defined by $\sigma(z) = \max\{0, z\}$. In this work we consider depth-2 ReLU neural networks with input dimension 1. Formally, a depth-2 network $\mathcal{N}_{\boldsymbol{\theta}}$ of width k is parameterized by $\boldsymbol{\theta} = [\mathbf{w}, \mathbf{b}, \mathbf{v}]$ where $\mathbf{w}, \mathbf{b}, \mathbf{v} \in \mathbb{R}^k$, and for every input $x \in \mathbb{R}$ we have

$$\mathcal{N}_{\boldsymbol{\theta}}(x) = \sum_{j \in [k]} v_j \sigma(w_j \cdot x + b_j) \ . \tag{1}$$

We sometimes view $\boldsymbol{\theta}$ as the vector obtained by concatenating the vectors $\mathbf{w}, \mathbf{b}, \mathbf{v}$. Thus, $\|\boldsymbol{\theta}\|$ denotes the ℓ_2 norm of the vector $\boldsymbol{\theta}$. We denote

$$\Phi(\boldsymbol{\theta}; x) := \mathcal{N}_{\boldsymbol{\theta}}(x).$$

Given a network $\mathcal{N}_{\boldsymbol{\theta}}(x)$ as above, we refer to the set of its non-differentiable points (w.r.t. the variable x) as its breakpoints.

Gradient flow (GF) and implicit bias. Let $S = \{(x_i, y_i)\}_{i=1}^n \subseteq \mathbb{R} \times \{-1, 1\}$ be a binary classification training dataset. Let $\Phi(\boldsymbol{\theta}; \cdot) : \mathbb{R} \to \mathbb{R}$ be a neural network parameterized by $\boldsymbol{\theta}$. For a loss function $\ell : \mathbb{R} \to \mathbb{R}$ the *empirical loss* of $\Phi(\boldsymbol{\theta}; \cdot)$ on the dataset S is

$$\mathcal{L}(\boldsymbol{\theta}) := \frac{1}{n} \sum_{i=1}^{n} \ell(y_i \Phi(\boldsymbol{\theta}; x_i)) \ . \tag{2}$$

We focus on the exponential loss $\ell(q) = e^{-q}$ and the logistic loss $\ell(q) = \log(1 + e^{-q})$.

We consider GF on the objective given in Eq. (2). This setting captures the behavior of GD 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 $\boldsymbol{\theta}(t)$ are given by the differential equation $\frac{d\boldsymbol{\theta}(t)}{dt} \in -\partial^\circ \mathcal{L}(\boldsymbol{\theta}(t))$. Here, ∂° denotes the *Clarke subdifferential*, which is a generalization of the derivative for non-differentiable functions (see Appendix A for a formal definition). We say that a trajectory $\boldsymbol{\theta}(t)$ *converges in direction* to $\boldsymbol{\theta}^*$ if $\lim_{t\to\infty} \frac{\boldsymbol{\theta}(t)}{\|\boldsymbol{\theta}(t)\|} = \frac{\boldsymbol{\theta}^*}{\|\boldsymbol{\theta}^*\|}$.

2.1 Assumptions

Our main result holds under the following assumption on the initialization of the weights of the network.

Assumption 2.1 (Network initialization)**.**

- *The weights and biases w_i, b_i of each neuron $i \in [k]$ in the hidden layer are i.i.d. and satisfy*
$$w_i, b_i \sim \mathcal{N}(0, \sigma_h^2).$$
- *The weights v_i, $i \in [k]$ of the output neuron are i.i.d. and satisfy*
$$v_i \sim \mathcal{N}(0, \sigma_o^2).$$

We point out that our particular choice of normally distributed weights is not essential, and that our results will also hold for example under the assumption of uniformly distributed weights, but with a slightly different proof and constants in the resulting bounds. To facilitate our analysis, we make the following assumptions on the distribution of the data and its corresponding labels.

Assumption 2.2 (Data distribution)**.** *There exist a natural $r \geq 1$ and real $R \geq 1$ and $C, \rho > 0$ such that following hold:*

- *There exists a depth-2 ReLU network $\mathcal{N}^*$ of width r such that the examples (x, y) of the data satisfy $x \sim \mathcal{D}$ and $y = \operatorname{sign}(\mathcal{N}^*(x))$.*
- *The density of $\mathcal{D}$ denoted by μ satisfies $\mu(x) = 0$ for all $x \notin [-R, R]$.*
- $\sup_x \mu(x) \leq C$.
- *$\rho > 0$ is the length of the shortest interval $I \subseteq [-R, R]$ such that $\operatorname{sign}(\mathcal{N}^*(x))$ is the same for all $x \in I$, and for all $I \subset I'$, there exist $x, x' \in I'$ such that $\operatorname{sign}(\mathcal{N}^*(x)) \neq \operatorname{sign}(\mathcal{N}^*(x'))$.*

We remark that the above assumption is mostly mild. The main non-trivial requirement is that the distribution is compactly-supported, however the last two assumptions always hold for some $C, \rho > 0$ if μ is continuous on $\mathbb{R}$ for example, and any target function that changes sign at most r times can be expressed by a network of width r.[2]

2.2 Main result

Having stated our assumptions, we now turn to present our main theorem in this paper.

Theorem 2.3. *Under Assumptions 2.1 and 2.2, given any $\varepsilon, \delta \in (0, 1)$, suppose that the following hold*

$$n \geq C_0 \cdot \frac{r \log(1/\varepsilon) + \log(2/\delta)}{\varepsilon}, \qquad k \geq 6144 \cdot \frac{R^4 \log\left(\frac{48r}{\delta}\right)}{\rho},$$
$$\sigma_h \geq 8400 \cdot \frac{n^2 C R^{3.5} \sqrt{rk}}{\delta \rho}, \qquad \sigma_o \leq \frac{1}{4kR\sigma_h \log\left(\frac{12k}{\delta}\right)},$$

where $C_0 > 0$ is a universal constant. Then with probability at least $1 - \delta$ over the randomness in the initialization of the network and the sampling of a size-n dataset, GF converges to zero loss,

[2] We also remark that the assumption $R \geq 1$ is only for simplicity of presentation, since our results also hold for any $R < 1$ with the same bounds we get when plugging $R = 1$.

and converges in direction to $\boldsymbol{\theta}^*$ such that the network $\mathcal{N}_{\boldsymbol{\theta}^*}$ has at most $32r + 67$ linear regions and satisfies

$$\mathbb{P}_{x\sim\mathcal{D}}\left[\text{sign}(\mathcal{N}_{\boldsymbol{\theta}^*}(x)) \neq \text{sign}(\mathcal{N}^*(x))\right] \leq \varepsilon.$$

Our theorem is a result of breaking the proof into two different stages and combining them using a simple union bound. Specifically, at the first stage we use Theorem 3.1, which establishes that the empirical loss drops below $\frac{1}{n}$ with high probability; and in the second stage we use Corollary 5.1 to argue that the implicit bias takes effect once the empirical loss is sufficiently small which results in a generalization bound. The theorem suggests two interesting implications: (i) We may train an arbitrarily wide network without risk of overfitting since the implicit bias of GF dictates that we converge to a model with low capacity;[3] and (ii), we further gain a sample complexity bound which is independent of ρ and R. We also note that only the direction of $\boldsymbol{\theta}$ affects the classification and the number of linear regions in $\mathcal{N}_{\boldsymbol{\theta}}$, and that the scale of $\boldsymbol{\theta}$ is not important here. Namely, for every $\boldsymbol{\theta}$, x and every $\alpha > 0$ we have $\mathcal{N}_{\alpha\boldsymbol{\theta}}(x) = \alpha^2\mathcal{N}_{\boldsymbol{\theta}}(x)$, and thus the scale of $\boldsymbol{\theta}$ affects only the scale of the outputs of $\mathcal{N}_{\boldsymbol{\theta}}$ and not the classification nor the partition to linear regions and therefore nor does it affect the generalization properties of $\boldsymbol{\theta}$.

Lastly, we remark that our initialization scheme used in Theorem 2.3 is somewhat unorthodox, in the sense that we require the hidden layer to have a rather large variance which scales polynomially with the sample size. While in practice it is more common that the weights in the hidden layer have a smaller variance (e.g. Glorot and Bengio [2010], He et al. [2015]), our scaling prevents the breakpoints of the neurons in the trained network to move too much before we are able to decrease the training error sufficiently, and the impact of the magnitude by which we scale the hidden layer upon initialization on the dynamics of GF was studied in a similar univariate regression setting [Williams et al., 2019, Sahs et al., 2020]. We now conclude the discussion of our main result with the following remarks on the setting studied in our paper.

Remark 2.1 (Mild vs. extreme over-parameterization)**.** *In this paper, we make a distinction between what we call the* mild over-parameterization *regime, where the required width of the network k scales with r but not with the sample size n; and the* extreme over-parameterization *regime, where the width k' of the network exceeds n. Under this distinction, for sufficiently large n, we will always have that $k \ll n < k'$, and thus $k \ll k'$.*

Remark 2.2 (GF vs. GD)**.** *It is important to stress that our results hold for GF which ignores computational considerations and does not necessarily imply the convergence of GD. For this reason, it is of utmost importance to generalize our results to hold for GD rather than just GF. That being said, there is some recent evidence suggesting that at least in certain cases, positive results on GF may be translated to GD [Elkabetz and Cohen, 2021]. Moreover, at the very least, Theorem 3.1 can indeed be generalized to hold for GD (see discussion after the theorem statement). In any case, for the sake of coherence we focus in this paper on GF, and we leave generalizations for GD as an important future work direction.*

Remark 2.3 (Rich vs. lazy regime)**.** *Our optimization analysis operates in the lazy regime where the hidden layer does not move much and most of the learning is performed in the output neuron. However, once our analysis goes into the second stage where the implicit bias takes effect, we essentially move into the rich (aka the feature-learning) regime, where redundant features (i.e. excess neurons) are being effectively discarded at the limit $t \to \infty$. In light of this, as was discussed in the related work section, our result provides a guarantee which goes beyond the analysis achieved using NTK-based results.*

3 Over-parameterization leads to small empirical loss

In this section, we analyze the dynamics of GF on the objective defined in Eq. (2). Our main contribution is to establish that sufficient over-parameterization guarantees that GF leads to a point with empirical loss which is sufficiently small. Formally, we present the following theorem.

[3] Note that in our univariate setting we can deduce r by analyzing the data and possibly estimate the minimal required width for attaining small training error using our derived bounds, however in more general settings (e.g. the multivariate case) r may not be easily deduced from the dataset, which might prompt us to train the widest network possible given available computational resources.

Theorem 3.1. *Under Assumptions 2.1 and 2.2, given any $\delta \in (0,1)$, suppose that the following hold*

$$k \geq 6144 \cdot \frac{R^4 \log\left(\frac{24r}{\delta}\right)}{\rho}, \quad \sigma_h \geq 4200 \cdot \frac{n^2 C R^{3.5}\sqrt{rk}}{\delta\rho} \quad \textit{and} \quad \sigma_o \leq \frac{1}{4kR\sigma_h \log\left(\frac{6k}{\delta}\right)}. \tag{3}$$

Then with probability at least $1-\delta$ over the randomness in the initialization of the network and the sampling of a size-n dataset, there exists time t_0 such that GF initialized from $\boldsymbol{\theta}(0)$ reaches a point $\boldsymbol{\theta}(t_0)$ satisfying

$$\mathcal{L}(\boldsymbol{\theta}(t_0)) \leq \frac{1}{2n}.$$

The above theorem essentially requires that we use width which is proportional to $1/\rho$ up to logarithmic factors to facilitate optimization. If $\rho = \Omega(1/r)$, then this requires that we over-parameterize by a multiplicative constant up to logarithmic factors. We remark that our result can also be adapted to hold for GD rather than GF with a polynomial number of iterations.[4] In any case, as discussed in Remark 2.2, we stress that our focus here is to show that GF attains sufficiently small loss so that our implicit bias analysis takes effect, and we leave generalizations for GD and milder width requirements for future work.

The proof of the above theorem, which appears in Appendix C, relies on over-parameterizing sufficiently to the extent of having at least three breakpoints on each constant segment where the data does not change classification, and four additional neurons that are active on all the data instances. The key observation is that under such over-parameterization, we can identify a direction in weight space which moves the current network configuration in a manner which strictly decreases the objective value. This allows us to establish that the objective function satisfies the PL-condition locally in a neighborhood around our initialization. Finally, by bounding the length of the trajectory of GF, we show that the objective value decreases to $\frac{1}{2n}$ before we can escape the neighborhood in which the PL-condition is satisfied.

Interestingly, Theorem 3.1 already implies a generalization bound if the sample size is sufficiently larger than the degrees of freedom in the student network. Nevertheless, such an approach alone is not capable of obtaining a sample complexity which is independent of ρ and R (since the width of the student network and thus also its capacity scale with these parameters, implying a generalization bound that explicitly depends on them). Moreover, understanding the implicit bias of GF is of independent interest, even if we ignore the improvement it provides to the sample complexity.

It is natural to explore what is the minimal amount of over-parameterization required for attaining a small loss in our setting. A modest requirement is that we are able to make the generalization error arbitrarily small given a sufficiently large sample. It is thus interesting to present the following theorem, which establishes that under Assumption 2.1, GF is not capable of attaining loss below an absolute constant unless $k \geq \lfloor 1.3r \rfloor$, regardless of the sample size.

Theorem 3.2. *Define the population loss of a network with weights $\boldsymbol{\theta}$ w.r.t. a distribution $\mathcal{D}$ and teacher network $\mathcal{N}^*(\cdot)$ as*

$$\mathcal{L}_{\mathcal{D}}(\boldsymbol{\theta}) \coloneqq \mathop{\mathbb{E}}_{x\sim\mathcal{D}}\left[\ell(\mathcal{N}_{\boldsymbol{\theta}}(x)\cdot \operatorname{sign}(\mathcal{N}^*(x)))\right].$$

Let $\alpha \geq 1$ and suppose that an architecture as in Eq. (1), having width $k = \lfloor \alpha r \rfloor$ is initialized according to Assumption 2.1. Then there exists a distribution $\mathcal{D}$ and $\mathcal{N}^*(\cdot)$ of width r such that for any time $t \geq 0$ and any $n \geq 1$, for a size-n sample drawn from $\mathcal{D}$ and labeled by $\operatorname{sign}(\mathcal{N}^*(\cdot))$, with probability at least 0.25 over the initialization of the network, GF trained on Eq. (2) attains population loss at least

$$\mathcal{L}_{\mathcal{D}}(\boldsymbol{\theta}(t)) \geq \frac{1}{4}\left(1 - 0.75\alpha\right).$$

Thus, under the theorem's assumptions, over-parameterization by a constant factor is required. We refer the reader to Appendix D for the construction of $\mathcal{D}$ and $\mathcal{N}^*(\cdot)$, some further discussion, and the full proof of this lower bound.

[4] To show this, one would need to bound the length of the trajectory of GD for objectives that satisfy the PL-condition locally. See Appendix C for further detail.

4 The implicit bias of GF

In this section, we show that GF converges to networks where the number of linear regions is minimal up to a constant factor. We first give some required background and discuss an important result on the implicit bias which applies to depth-2 ReLU networks, and then state our result.

4.1 Required background

The following theorem gives an important characterization of the implicit bias in depth-2 ReLU networks:

Theorem 4.1 (Lyu and Li [2019], Ji and Telgarsky [2020])**.** *Let $\Phi(\boldsymbol{\theta};\cdot)$ be a depth-2 ReLU neural network parameterized by $\boldsymbol{\theta}$. Consider minimizing either the exponential or the logistic loss over a binary classification dataset $\{(x_i,y_i)\}_{i=1}^n$ using GF. Assume that there exists time t_0 such that*

$$\mathcal{L}(\boldsymbol{\theta}(t_0)) < \frac{1}{n},$$

namely, $y_i\Phi(\boldsymbol{\theta}(t_0);x_i) > 0$ for every x_i. Then, GF converges in direction to a first order stationary point (KKT point) of the following maximum margin problem in parameter space:

$$\min_{\boldsymbol{\theta}} \frac{1}{2}\|\boldsymbol{\theta}\|^2 \quad \textit{s.t.} \quad \forall i\in[n] \;\; y_i\Phi(\boldsymbol{\theta};x_i)\geq 1\,. \tag{4}$$

Moreover, $\mathcal{L}(\boldsymbol{\theta}(t))\to 0$ and $\|\boldsymbol{\theta}(t)\|\to\infty$ as $t\to\infty$.

We note that the above theorem holds for the more general case of homogeneous neural networks (in parameter space), but for this work it suffices to consider depth-2 networks, which are indeed homogeneous. Note that in ReLU networks Problem (4) is non-smooth. Hence, the KKT conditions are defined using the Clarke subdifferential. See Appendix A for more details of the KKT conditions. Theorem 4.1 characterized the implicit bias of GF with the exponential and the logistic losses for depth-2 ReLU networks. Namely, even though there are many possible directions $\frac{\boldsymbol{\theta}}{\|\boldsymbol{\theta}\|}$ that classify the dataset correctly, GF converges only to directions that are KKT points of Problem (4). We note that such a KKT point is not necessarily a global/local optimum (cf. Vardi et al. [2021]). Thus, under the theorem's assumptions, GF *may not* converge to an optimum of Problem (4), but it is guaranteed to converge to a KKT point. This is demonstrated in the following example for the case of depth-2 univariate networks, which is our focus.

Example 1. *Let $\mathcal{N}_{\boldsymbol{\theta}}$ be a depth-2 univariate network of width 2, namely,*

$$\mathcal{N}_{\boldsymbol{\theta}}(x) = v_1\sigma(w_1x+b_1)+v_2\sigma(w_2x+b_2).$$

Let $S=\{(x_1,y_1),(x_2,y_2)\}$ be a size-2 dataset such that $x_1=4$, $x_2=-4$ and $y_1=y_2=1$. Suppose that we train $\mathcal{N}_{\boldsymbol{\theta}}$ on the dataset S using GF with the exponential or the logistic loss, and that the initialization $\boldsymbol{\theta}(0)$ is such that $b_1=v_1=1$ and $w_1=w_2=b_2=v_2=0$. Note that

$$\mathcal{L}(\boldsymbol{\theta}(0)) = \frac{1}{2}\cdot 2\ell(1) < \frac{1}{2}$$

(for both the exponential and the logistic loss). Hence, by Theorem 4.1 GF converges to zero loss, and converges in direction to a KKT point $\boldsymbol{\theta}^*$ of Problem (4). By observing the gradient $\nabla_{\boldsymbol{\theta}}\mathcal{L}(\boldsymbol{\theta})$ it is not hard to show that for every time t we have $w_2(t)=b_2(t)=v_2(t)=0$, namely, the second neuron remains inactive, and we have $\boldsymbol{\theta}^*=\boldsymbol{\theta}(0)$ (see Appendix B for details). However, $\boldsymbol{\theta}^*$ is not a global optimum of Problem (4). Indeed, consider $\boldsymbol{\theta}'$ such that $w_1'=v_1'=v_2'=\frac{1}{2}$, $w_2'=-\frac{1}{2}$, and $b_1'=b_2'=0$. Then, $y_i\mathcal{N}_{\boldsymbol{\theta}'}(x_i)=1$ for all $i\in\{1,2\}$, and we have $\|\boldsymbol{\theta}'\|^2=1<2=\|\boldsymbol{\theta}^*\|^2$.

The above example implies that although GF has a certain bias towards margin maximization (as shown in Theorem 4.1), it may not maximize the margin. Hence, we cannot obtain margin-based generalization bounds based on the bias towards margin maximization.

4.2 Characterization of the implicit bias

We now state our main result on the implicit bias:

Theorem 4.2. *Let $S = \{(x_i, y_i)\}_{i=1}^n \subseteq \mathbb{R} \times \{-1, 1\}$ be a dataset such that $x_1 < \ldots < x_n$, and for all $i \in [n]$ we have $y_i \mathcal{N}^*(x_i) > 0$, where $\mathcal{N}^* : \mathbb{R} \to \mathbb{R}$ is a depth-2 ReLU network of width r. Note that $|\{i \in [n-1] : y_i \neq y_{i+1}\}| \leq r$. Consider GF on a depth-2 neural network $\mathcal{N}_{\boldsymbol{\theta}}$ w.r.t. the dataset S. Assume that there exists time t_0 such that*

$$\mathcal{L}(\boldsymbol{\theta}(t_0)) < \frac{1}{n}.$$

Then, GF converges to zero loss, and converges in direction to a KKT point $\boldsymbol{\theta}^*$ of Problem (4), such that the network $\mathcal{N}_{\boldsymbol{\theta}^*}$ has at most $32r + 67$ linear regions.

Since the labels in the dataset S may switch sign r times, then a network that correctly classifies S must contain at least r linear regions. Hence, the theorem implies that GF minimizes the number of linear regions up to a constant factor. We remark that the constants 32 and 67 in the above result can be improved, but we preferred here a simpler proof over a tighter bound.

The formal proof of the theorem is given in Appendix E. Below we discuss the high-level approach. By Theorem 4.1, if there exists time t_0 such that $\mathcal{L}(\boldsymbol{\theta}(t_0)) < \frac{1}{n}$ then GF converges to zero loss, and converges in direction to a KKT point of Problem (4). We denote

$$\mathcal{N}_{\boldsymbol{\theta}}(x) = \sum_{j \in [k]} v_j \sigma(w_j x + b_j).$$

Assume that $\mathcal{N}_{\boldsymbol{\theta}}$ satisfies the KKT conditions of Problem (4). Thus, there are $\lambda_1, \ldots, \lambda_n \geq 0$ such that for every $j \in [k]$ we have

$$w_j = \sum_{i \in [n]} \lambda_i \frac{\partial}{\partial w_j} \left(y_i \mathcal{N}_{\boldsymbol{\theta}}(x_i)\right) = \sum_{i \in [n]} \lambda_i y_i v_j \sigma'_{i,j} x_i \ , \tag{5}$$

where $\sigma'_{i,j}$ is a subgradient of σ at $w_j \cdot x_i + b_j$, and $\lambda_i = 0$ if $y_i \mathcal{N}_{\boldsymbol{\theta}}(x_i) \neq 1$. Likewise, we have

$$b_j = \sum_{i \in [n]} \lambda_i \frac{\partial}{\partial b_j} \left(y_i \mathcal{N}_{\boldsymbol{\theta}}(x_i)\right) = \sum_{i \in [n]} \lambda_i y_i v_j \sigma'_{i,j} \ . \tag{6}$$

In the proof we show using a careful analysis of Eqs. (5) and (6) that in an interval $[x_p, x_q]$ where the labels do not switch sign (i.e., $y_p = y_{p+1} = \ldots = y_q$) the network $\mathcal{N}_{\boldsymbol{\theta}}$ has a constant number of kinks. Since the labels switch sign at most r times then we are able to conclude that $\mathcal{N}_{\boldsymbol{\theta}}$ has $\mathcal{O}(r)$ kinks as required.

5 Implicit bias leads to a generalization bound

Consider a depth-2 teacher ReLU network $\mathcal{N}^* : \mathbb{R} \to \mathbb{R}$ of width r. By Theorem 4.2, if GF reaches a sufficiently small loss at some time t_0, it converges in direction to a network $\mathcal{N}'$ with $\mathcal{O}(r)$ linear regions that classifies the training dataset correctly. Thus, even if the width k of the learned network is extremely large, the result in Theorem 4.2 guarantees that GF converges to a network with a small number of linear regions. Consider the function $f' : \mathbb{R} \to \{-1, 1\}$ defined by $f'(x) = \text{sign}(\mathcal{N}'(x))$. Since $\mathcal{N}'$ has $\mathcal{O}(r)$ linear regions then f' can be expressed by a polynomial threshold function of degree $\mathcal{O}(r)$. That is, we have $f'(x) = \text{sign}(p'(x))$ where p' is a polynomial of degree $\mathcal{O}(r)$. Since $\mathcal{N}'$ attains 100% classification accuracy on the size-n training dataset, then we can view GF as empirical risk minimization (ERM) w.r.t. the 0-1 loss over a class of degree-$\mathcal{O}(r)$ polynomial threshold functions in the realizable setting. The VC-dimension of this class is $\mathcal{O}(r)$, and thus we have the following generalization bound w.r.t. the 0-1 loss (see, e.g., Theorem 6.8 in Shalev-Shwartz and Ben-David [2014]).

Corollary 5.1. *There exists a universal constant $C_0 > 0$ such that the following holds. Let $\mathcal{N}^* : \mathbb{R} \to \mathbb{R}$ be a depth-2 ReLU network of width r. Let $\varepsilon, \delta \in (0, 1)$ and let*

$$n \geq C_0 \cdot \frac{r \log(1/\varepsilon) + \log(1/\delta)}{\varepsilon} \ .$$

Let S be a size-n binary classification dataset drawn from a distribution $\mathcal{D}$ and labeled according to $\mathcal{N}^*$. Consider GF on a depth-2 neural network $\mathcal{N}_{\boldsymbol{\theta}}$ w.r.t. the dataset S, and suppose that there exists time t_0 such that

$$\mathcal{L}(\boldsymbol{\theta}(t_0)) < \frac{1}{n}.$$

Then, GF converges in direction to $\boldsymbol{\theta}^*$ such that the network $\mathcal{N}_{\boldsymbol{\theta}^*}$ has at most $32r + 67$ linear regions, and with probability at least $1 - \delta$ over the sampling of S we have

$$\mathbb{P}_{x\sim\mathcal{D}}\left[\text{sign}(\mathcal{N}_{\boldsymbol{\theta}^*}(x)) \neq \text{sign}(\mathcal{N}^*(x))\right] \leq \varepsilon \ .$$

Acknowledgments and Disclosure of Funding

Work done while GV was at the Weizmann Institute of Science. We thank Noam Razin and Gilad Yahudai for pointing out several relevant papers to discuss in the related work section.

References

Z. Allen-Zhu, Y. Li, and Z. Song. A convergence theory for deep learning via over-parameterization. *arXiv preprint arXiv:1811.03962*, 2018.

A. Andoni, R. Panigrahy, G. Valiant, and L. Zhang. Learning polynomials with neural networks. In *Proceedings of the 31st International Conference on Machine Learning*, pages 1908–1916, 2014.

Y. Arjevani and M. Field. Analytic characterization of the hessian in shallow relu models: A tale of symmetry. *Advances in Neural Information Processing Systems*, 33:5441–5452, 2020.

Y. Arjevani and M. Field. Analytic study of families of spurious minima in two-layer relu neural networks: A tale of symmetry ii. *Advances in Neural Information Processing Systems*, 34, 2021.

G. Blanc, N. Gupta, G. Valiant, and P. Valiant. Implicit regularization for deep neural networks driven by an ornstein-uhlenbeck like process. In *Conference on learning theory*, pages 483–513. PMLR, 2020.

L. Chizat and F. 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.

L. Chizat, E. Oyallon, and F. Bach. On lazy training in differentiable programming. *Advances in Neural Information Processing Systems*, 32, 2019.

F. H. Clarke, Y. S. Ledyaev, R. J. Stern, and P. R. Wolenski. *Nonsmooth analysis and control theory*, volume 178. Springer Science & Business Media, 2008.

A. Daniely. Sgd learns the conjugate kernel class of the network. In *Advances in Neural Information Processing Systems*, pages 2422–2430, 2017.

DLMF. *NIST Digital Library of Mathematical Functions*. http://dlmf.nist.gov/, Release 1.1.3 of 2021-09-15. URL http://dlmf.nist.gov/. F. W. J. Olver, A. B. Olde Daalhuis, D. W. Lozier, B. I. Schneider, R. F. Boisvert, C. W. Clark, B. R. Miller, B. V. Saunders, H. S. Cohl, and M. A. McClain, eds.

S. Du, J. Lee, H. Li, L. Wang, and X. Zhai. Gradient descent finds global minima of deep neural networks. In *International conference on machine learning*, pages 1675–1685. PMLR, 2019.

S. S. Du, X. Zhai, B. Poczos, and A. Singh. Gradient descent provably optimizes over-parameterized neural networks. *arXiv preprint arXiv:1810.02054*, 2018.

J. Dutta, K. Deb, R. Tulshyan, and R. Arora. Approximate kkt points and a proximity measure for termination. *Journal of Global Optimization*, 56(4):1463–1499, 2013.

O. Elkabetz and N. Cohen. Continuous vs. discrete optimization of deep neural networks. *Advances in Neural Information Processing Systems*, 34, 2021.

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

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

X. Glorot and Y. Bengio. Understanding the difficulty of training deep feedforward neural networks. In *Proceedings of the thirteenth international conference on artificial intelligence and statistics*, pages 249–256. JMLR Workshop and Conference Proceedings, 2010.

C. Gupta, S. Balakrishnan, and A. Ramdas. Path length bounds for gradient descent and flow. *Journal of Machine Learning Research*, 22(68):1–63, 2021.

B. Hanin. Ridgeless interpolation with shallow relu networks in $1d$ is nearest neighbor curvature extrapolation and provably generalizes on lipschitz functions. *arXiv preprint arXiv:2109.12960*, 2021.

K. He, X. Zhang, S. Ren, and J. Sun. Delving deep into rectifiers: Surpassing human-level performance on imagenet classification. In *Proceedings of the IEEE international conference on computer vision*, pages 1026–1034, 2015.

A. Jacot, F. Gabriel, and C. Hongler. Neural tangent kernel: Convergence and generalization in neural networks. In *Advances in neural information processing systems*, pages 8571–8580, 2018.

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

Z. Ji, M. Telgarsky, and R. Xian. Neural tangent kernels, transportation mappings, and universal approximation. *arXiv preprint arXiv:1910.06956*, 2019.

Y. Li, T. Ma, and H. R. Zhang. Learning over-parametrized two-layer neural networks beyond ntk. In *Conference on learning theory*, pages 2613–2682. PMLR, 2020.

K. Lyu and J. Li. Gradient descent maximizes the margin of homogeneous neural networks. *arXiv preprint arXiv:1906.05890*, 2019.

H. Maennel, O. Bousquet, and S. Gelly. Gradient descent quantizes relu network features. *arXiv preprint arXiv:1803.08367*, 2018.

B. Neyshabur. Implicit regularization in deep learning. *arXiv preprint arXiv:1709.01953*, 2017.

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

B. Neyshabur, S. Bhojanapalli, D. McAllester, and N. Srebro. Exploring generalization in deep learning. In *Advances in Neural Information Processing Systems*, pages 5947–5956, 2017.

G. Ongie, R. Willett, D. Soudry, and N. Srebro. A function space view of bounded norm infinite width relu nets: The multivariate case. *arXiv preprint arXiv:1910.01635*, 2019.

D. B. Owen. Tables for computing bivariate normal probabilities. *The Annals of Mathematical Statistics*, 27(4):1075–1090, 1956.

R. Parhi and R. D. Nowak. The role of neural network activation functions. *IEEE Signal Processing Letters*, 27:1779–1783, 2020.

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

B. T. Polyak. Gradient methods for minimizing functionals. *Zhurnal Vychislitel'noi Matematiki i Matematicheskoi Fiziki*, 3(4):643–653, 1963.

N. Razin and N. Cohen. Implicit regularization in deep learning may not be explainable by norms. *Advances in neural information processing systems*, 33:21174–21187, 2020.

I. Safran and O. Shamir. Spurious local minima are common in two-layer relu neural networks. In *International conference on machine learning*, pages 4433–4441. PMLR, 2018.

I. M. Safran, G. Yehudai, and O. Shamir. The effects of mild over-parameterization on the optimization landscape of shallow relu neural networks. In *Conference on Learning Theory*, pages 3889–3934. PMLR, 2021.

J. Sahs, R. Pyle, A. Damaraju, J. O. Caro, O. Tavaslioglu, A. Lu, and A. Patel. Shallow univariate relu networks as splines: initialization, loss surface, hessian, & gradient flow dynamics. *arXiv preprint arXiv:2008.01772*, 2020.

P. Savarese, I. Evron, D. Soudry, and N. Srebro. How do infinite width bounded norm networks look in function space? In *Conference on Learning Theory*, pages 2667–2690. PMLR, 2019.

S. Shalev-Shwartz and S. Ben-David. *Understanding machine learning: From theory to algorithms*. Cambridge university press, 2014.

M. Soltanolkotabi, A. Javanmard, and J. D. Lee. Theoretical insights into the optimization landscape of over-parameterized shallow neural networks. *IEEE Transactions on Information Theory*, 65(2): 742–769, 2018.

N. Timor, G. Vardi, and O. Shamir. Implicit regularization towards rank minimization in relu networks. *arXiv preprint arXiv:2201.12760*, 2022.

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

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

G. Vardi, O. Shamir, and N. Srebro. On margin maximization in linear and relu networks. *arXiv preprint arXiv:2110.02732*, 2021.

F. Williams, M. Trager, D. Panozzo, C. Silva, D. Zorin, and J. Bruna. Gradient dynamics of shallow univariate relu networks. In *Advances in Neural Information Processing Systems*, pages 8378–8387, 2019.

B. Woodworth, S. Gunasekar, J. D. Lee, E. Moroshko, P. Savarese, I. Golan, D. Soudry, and N. Srebro. Kernel and rich regimes in overparametrized models. In *Conference on Learning Theory*, pages 3635–3673. PMLR, 2020.

C. Zhang, S. Bengio, M. Hardt, B. Recht, and O. Vinyals. Understanding deep learning (still) requires rethinking generalization. *Communications of the ACM*, 64(3):107–115, 2021.

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

D. Zou, Y. Cao, D. Zhou, and Q. Gu. Gradient descent optimizes over-parameterized deep relu networks. *Machine learning*, 109(3):467–492, 2020.

Checklist

1. For all authors...
   (a) Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope? [Yes]
   (b) Did you describe the limitations of your work? [Yes]
   (c) Did you discuss any potential negative societal impacts of your work? [No] The paper is theoretical in its nature and holds no foreseeable societal impacts as far as we can discern.
   (d) Have you read the ethics review guidelines and ensured that your paper conforms to them? [Yes]
2. If you are including theoretical results...
   (a) Did you state the full set of assumptions of all theoretical results? [Yes]
   (b) Did you include complete proofs of all theoretical results? [Yes]
3. If you ran experiments...
   (a) Did you include the code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL)? [N/A]

(b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? [N/A]

(c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? [N/A]

(d) Did you include the total amount of compute and the type of resources used (e.g., type of GPUs, internal cluster, or cloud provider)? [N/A]

4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets...

   (a) If your work uses existing assets, did you cite the creators? [N/A]

   (b) Did you mention the license of the assets? [N/A]

   (c) Did you include any new assets either in the supplemental material or as a URL? [N/A]

   (d) Did you discuss whether and how consent was obtained from people whose data you're using/curating? [N/A]

   (e) Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? [N/A]

5. If you used crowdsourcing or conducted research with human subjects...

   (a) Did you include the full text of instructions given to participants and screenshots, if applicable? [N/A]

   (b) Did you describe any potential participant risks, with links to Institutional Review Board (IRB) approvals, if applicable? [N/A]

   (c) Did you include the estimated hourly wage paid to participants and the total amount spent on participant compensation? [N/A]

A Preliminaries on the Clarke subdifferential and the KKT conditions

Below we define the Clarke subdifferential, and review the definition of the KKT conditions for non-smooth optimization problems (cf. Lyu and Li [2019], Dutta et al. [2013]).

Let $f : \mathbb{R}^d \to \mathbb{R}$ be a locally Lipschitz function. The Clarke subdifferential [Clarke et al., 2008] at $\mathbf{x} \in \mathbb{R}^d$ is the convex set

$$\partial^\circ f(\mathbf{x}) := \text{conv}\left\{ \lim_{i\to\infty} \nabla f(\mathbf{x}_i) \;\middle|\; \lim_{i\to\infty} \mathbf{x}_i = \mathbf{x},\ f \text{ is differentiable at } \mathbf{x}_i \right\} .$$

If f is continuously differentiable at $\mathbf{x}$ then $\partial^\circ f(\mathbf{x}) = \{\nabla f(\mathbf{x})\}$. For the Clarke subdifferential the chain rule holds as an inclusion rather than an equation. That is, for locally Lipschitz functions $z_1, \ldots, z_n : \mathbb{R}^d \to \mathbb{R}$ and $f : \mathbb{R}^n \to \mathbb{R}$, we have

$$\partial^\circ (f \circ \mathbf{z})(\mathbf{x}) \subseteq \text{conv}\left\{ \sum_{i=1}^n \alpha_i \mathbf{h}_i : \boldsymbol{\alpha} \in \partial^\circ f(z_1(\mathbf{x}), \ldots, z_n(\mathbf{x})), \mathbf{h}_i \in \partial^\circ z_i(\mathbf{x}) \right\} .$$

Consider the following optimization problem

$$\min f(\mathbf{x}) \quad \text{s.t.} \quad \forall n \in [N] \;\; g_n(\mathbf{x}) \le 0 \,, \tag{7}$$

where $f, g_1, \ldots, g_n : \mathbb{R}^d \to \mathbb{R}$ are locally Lipschitz functions. We say that $\mathbf{x} \in \mathbb{R}^d$ is a *feasible point* of Problem (7) if $\mathbf{x}$ satisfies $g_n(\mathbf{x}) \le 0$ for all $n \in [N]$. We say that a feasible point $\mathbf{x}$ is a *KKT point* if there exists $\lambda_1, \ldots, \lambda_N \ge 0$ such that

1. $\mathbf{0} \in \partial^\circ f(\mathbf{x}) + \sum_{n\in[N]} \lambda_n \partial^\circ g_n(\mathbf{x})$;
2. For all $n \in [N]$ we have $\lambda_n g_n(\mathbf{x}) = 0$.

B Details on Example 1

We use here the notation $\Phi(\boldsymbol{\theta}; x) = \mathcal{N}_{\boldsymbol{\theta}}(x)$, and denote by $\sigma'(\cdot)$ a sub-gradient of σ, namely, $\sigma(z) = \mathbb{1}[z > 0]$ if $z \neq 0$ and $\sigma'(0) \in [0, 1]$ (the exact value in this case is not important here). For every $j \in \{1, 2\}$ we have

$$\begin{aligned}
\nabla_{w_j} \mathcal{L}(\boldsymbol{\theta}) &= \frac{1}{2} \sum_{i=1}^2 \ell'(y_i \Phi(\boldsymbol{\theta}; x_i)) \cdot y_i \nabla_{w_j} \Phi(\boldsymbol{\theta}; x_i) \\
&= \frac{1}{2} \sum_{i=1}^2 \ell'(v_1 \sigma(w_1 x_i + b_1) + v_2 \sigma(w_2 x_i + b_2)) \cdot v_j \sigma'(w_j x_i + b_j) x_i \,.
\end{aligned}$$

Likewise,

$$\begin{aligned}
\nabla_{v_j} \mathcal{L}(\boldsymbol{\theta}) &= \frac{1}{2} \sum_{i=1}^2 \ell'(y_i \Phi(\boldsymbol{\theta}; x_i)) \cdot y_i \nabla_{v_j} \Phi(\boldsymbol{\theta}; x_i) \\
&= \frac{1}{2} \sum_{i=1}^2 \ell'(v_1 \sigma(w_1 x_i + b_1) + v_2 \sigma(w_2 x_i + b_2)) \cdot \sigma(w_j x_i + b_j)
\end{aligned}$$

and

$$\begin{aligned}
\nabla_{b_j} \mathcal{L}(\boldsymbol{\theta}) &= \frac{1}{2} \sum_{i=1}^2 \ell'(y_i \Phi(\boldsymbol{\theta}; x_i)) \cdot y_i \nabla_{b_j} \Phi(\boldsymbol{\theta}; x_i) \\
&= \frac{1}{2} \sum_{i=1}^2 \ell'(v_1 \sigma(w_1 x_i + b_1) + v_2 \sigma(w_2 x_i + b_2)) \cdot v_j \sigma'(w_j x_i + b_j) \,.
\end{aligned}$$

Note that if $w_2 = b_2 = v_2 = 0$ then we have $\nabla_{w_2} \mathcal{L}(\boldsymbol{\theta}) = \nabla_{v_2} \mathcal{L}(\boldsymbol{\theta}) = \nabla_{b_2} \mathcal{L}(\boldsymbol{\theta}) = 0$. Since these parameters are initialized at zero, then they remain zero throughout the training. Moreover, Suppose that $w_1 = 0$ and $b_1 = v_1 = \alpha$ for some $\alpha > 0$, and that $w_2 = b_2 = v_2 = 0$, then we have

$$-\frac{dw_1}{dt} = \nabla_{w_1} \mathcal{L}(\boldsymbol{\theta}) = \frac{1}{2} \sum_{i=1}^2 \ell'(\alpha \sigma(\alpha)) \cdot \alpha \sigma'(\alpha) x_i = \frac{1}{2} \left(\ell'(\alpha^2) \cdot \alpha \cdot 4 + \ell'(\alpha^2) \cdot \alpha \cdot (-4) \right) = 0 \,,$$

$$-\frac{dv_1}{dt} = \nabla_{v_1}\mathcal{L}(\boldsymbol{\theta}) = \frac{1}{2}\sum_{i=1}^{2}\ell'(\alpha\sigma(\alpha))\cdot\sigma(\alpha) = \frac{1}{2}\sum_{i=1}^{2}\ell'(\alpha^2)\cdot\alpha\ ,$$

$$-\frac{db_1}{dt} = \nabla_{b_1}\mathcal{L}(\boldsymbol{\theta}) = \frac{1}{2}\sum_{i=1}^{2}\ell'(\alpha\sigma(\alpha))\cdot\alpha\sigma'(\alpha) = \frac{1}{2}\sum_{i=1}^{2}\ell'(\alpha^2)\cdot\alpha\ .$$

Hence, for every t we have $w_1(t) = 0$ and $b_1(t) = v_1(t) = \alpha(t)$ where $\alpha(t) > 0$ is monotonically increasing.

As a result, the KKT point $\boldsymbol{\theta}^*$ is such that $w_2^* = b_2^* = v_2^* = w_1^* = 0$, and $b_1^* = v_1^* = \alpha^*$ for some $\alpha^* > 0$. Since $\boldsymbol{\theta}^*$ satisfies the KKT conditions of Problem (4), then we have

$$\alpha^* = b_1^* = \sum_{i=1}^{2}\lambda_i y_i \nabla_{b_1}\Phi(\boldsymbol{\theta}^*; x_i)\ ,$$

where $\lambda_i \geq 0$ and $\lambda_i = 0$ if $y_i\Phi(\boldsymbol{\theta}^*; x_i) \neq 1$. Hence, there is i such that $y_i\Phi(\boldsymbol{\theta}^*; x_i) = 1$. Thus, we have $1 = y_i\Phi(\boldsymbol{\theta}^*; x_i) = (\alpha^*)^2$ which implies $\alpha^* = 1$. Therefore $\boldsymbol{\theta}^* = \boldsymbol{\theta}(0)$.

C Proof of Theorem 3.1

Before we prove the theorem, we first state a few definitions that are specific for this appendix. Let $S = \{(x_i, y_i)\}_{i=1}^n \subseteq [-R, R] \times \{-1, 1\}$ be a dataset such that $x_1 < \ldots < x_n$ and let $I = \{i \in [n-1] : y_i \neq y_{i+1}\}$ where we denote the elements of I using $i_1 < \ldots < i_r$, and $i_0 = -R, i_{r+1} = R$, where $r = |I|$. For all $j \in [r+1]$, define $I_j = (i_{j-1}, i_j + 1)$ which is the j-th interval where the instances in the data do not change their classification.[5] Given some function $\mathcal{L}$ and real number α, we let $L_\alpha^+(\mathcal{L}) := \{\boldsymbol{\theta} : \mathcal{L}(\boldsymbol{\theta}) \geq \alpha\}$ denote the α-superlevel set of $\mathcal{L}$.

Next, we state the following definitions, which establish sufficient conditions for our objective function to be well-behaved in the sense of having a strict direction of descent in a certain neighborhood.

Definition C.1 (Separability)**.** *Under Assumption 2.2, we say that $\boldsymbol{\theta}$ is separable from S with positive constants $\gamma, m < M, q < Q$ if for all $j \in [r+1]$, there exist three neurons with weights and biases denoted by $w_i(I_j)$ and $b_i(I_j)$ for $i \in [3]$, and breakpoints $\beta_1(I_j) < \beta_2(I_j) < \beta_3(I_j) \in I_j$, which satisfy the following items:*

1. *$m \leq |w_i(I_j)|$ and $|b_i(I_j)| \leq M$ for all $i \in [3]$.*
2. *There exist four neurons, two with breakpoints in each of the intervals $(-\infty, 0), (0, \infty)$, that are distinct from the neurons in the previous item, are active on all the data instances and whose weights w_i', b_i' satisfy $m \leq |w_i'|$ and $|b_i'| \leq M$ for all $i \in [4]$. Moreover, their breakpoints satisfy $|\beta_i| \leq Q$ for all $i \in [4]$.*
3. *$q \leq \beta_2(I_j) - \beta_1(I_j), \beta_3(I_j) - \beta_2(I_j) \leq Q$ for all $j \in [r+1]$.*
4. *$|\beta_1(I_{j+1}) - \beta_3(I_j)| \leq Q$ for all $j \in [r]$.*
5. *$x_{i_j+1} - x_{i_j} \geq \gamma$ for all $j \in [r]$.*

If the triplet of neurons satisfying the above items in an interval is not distinct, we assume w.l.o.g. that $\beta_1(\cdot)$ and $\beta_3(\cdot)$ return the left-most and right-most breakpoints satisfying the above, respectively.

The following definition is used to describe a neighborhood around the initialization point in which our separability assumption above holds.

Definition C.2 (Δ-hidden Neighborhood)**.** *Given a network $\mathcal{N}(\cdot)$, weights $\boldsymbol{\theta}$ and a constant $\Delta \geq 0$, we define the Δ-hidden Neighborhood of $\mathcal{N}$ at $\boldsymbol{\theta}$ as the set*

$$U_\Delta(\boldsymbol{\theta}) := \left\{\boldsymbol{\theta}' = [\mathbf{w}', \mathbf{b}', \mathbf{v}'] : \left\|(w_j, b_j) - (w_j', b_j')\right\|_2 \leq \Delta \quad \forall j \in [k], \quad \mathbf{v}' \in \mathbb{R}^k\right\}.$$

[5] We note that these intervals overlap and thus do contain instances that change classification with respect to the teacher network, but not with respect to the sample.

That is, the neighborhood of balls of radius Δ centered at each hidden neuron of $\boldsymbol{\theta}$ and where the output neuron weights are arbitrary.

Following the above definitions, the following auxiliary lemmas will be used in the proof of the theorem. The technical lemma below establishes that a certain binary matrix is invertible and provides a bound on the spectral norm of its inverse.

Lemma C.1. *Suppose that $A \in \{0,1\}^{d\times d}$, such that the first row of A is all-ones, and each subsequent row i is either $(1,\dots,1,0,\dots,0)$ with $i-1$ leading ones or $(0,\dots,0,1,\dots,1)$ with $i-1$ leading zeros. Then A is invertible and we have $\left\|A^{-1}\right\|_{sp} \le d$.*

Proof. The invertability of A follows from the fact that the first row of A is an all-ones vector, since we can use elementary row operations to change all subsequent rows to start with a '0' and end with a '1' if needed, resulting in an upper triangular matrix with all-ones on its main diagonal which is thus invertible. To bound $\left\|A^{-1}\right\|_{\text{sp}}$, let $I_d \in \{0,1\}^{d\times d}$ denote the identity matrix. We will use Gaussian elimination to compute the entries of A^{-1}. We first subtract the first row from all the other rows that do not have leading zeros and then multiply by the constant -1. Performing the same operation on I_d results in a matrix B whose rows are either standard unit vectors or the vector $(1,\dots,1,0,1,\dots,1)$. The resulting matrix after performing these operations on A is and upper triangular matrix with ones in all of its diagonal and above the diagonal entries. Since it is readily seen that the inverse of such a matrix is a matrix with all zero entries except for the main diagonal which is all-ones and the first diagonal above it which comprises of all -1's. Denote this matrix using B', we have that the inverse of A is given by $B \cdot B'$. The entries of A^{-1} therefore must consist of dot products of a standard unit vector and vectors $(0,\dots,0,1,-1,0,\dots,0)$, or the vector $(1,\dots,1,0,1,\dots,1)$ and vectors $(0,\dots,0,1,-1,0,\dots,0)$. In both cases the dot product is an element of $\{-1,0,1\}$, and therefore we can bound $\left\|A^{-1}\right\|_{\text{sp}}$ by the Frobenius norm of A^{-1} which is at most d. □

The following key lemma establishes that when $\boldsymbol{\theta}$ is separable from S then there exists a direction in weight space which strictly decreases our objective value.

Lemma C.2. *Under Assumption 2.2, suppose that $\boldsymbol{\theta}$ is separable from S with constants γ, m, M, q, Q, and that $\mathcal{L}(\boldsymbol{\theta}) \ge \frac{1}{2n}$. Then*

$$\frac{1}{2}\left\|\nabla\mathcal{L}(\boldsymbol{\theta})\right\|_2^2 \ge \frac{\gamma^2 q^2 m^6}{259200 n^4 Q^2 M^4}.$$

Proof. Since $\mathcal{L}(\boldsymbol{\theta}) > \frac{1}{2n}$, there must exist some $i \in [n]$ such that $\ell(y_i\Phi(\boldsymbol{\theta};x_i)) > \frac{1}{2n}$. We now consider two possible cases, depending on the location of x_i with respect to the breakpoints whose existence is guaranteed by Def. C.1, where will show the existence of a direction $\mathbf{u}$ which guarantees that $\mathcal{L}(\boldsymbol{\theta})$ is strictly decreasing.

- Suppose that $x_i \in I_\ell$ for some $\ell \in [r+1]$, such that $x_i \in (\beta_1(I_\ell), \beta_3(I_\ell))$. Assume without loss of generality that $x_i \in (\beta_2(I_\ell), \beta_3(I_\ell))$ (the proof is symmetric otherwise), and for ease of notation denote $\beta_2 \coloneqq \beta_1(I_\ell), \beta_3 \coloneqq \beta_2(I_\ell), \beta_4 \coloneqq \beta_3(I_\ell)$. Then by Item 2 in the separability assumption, there exists a breakpoints $\beta_1 < \beta_2$ such that all data instances in (β_2, β_4) have the same classification and the neuron with breakpoint at β_1 is active on all the data instances. Moreover, Item 2 also guarantees the existence of two breakpoints that are distinct from the previous ones, which we denote by β_5, β_6, where $\beta_5 < \beta_1 < 0$ and $\beta_6 > 0$ are active on all the data points. We will now show the existence of a depth-2 ReLU network which consists of six hidden neurons with weights $\mathbf{w} = (w_1,\dots,w_6)$ and biases $\mathbf{b} = (b_1,\dots,b_6)$ (corresponding to the breakpoints $\beta_1,\dots,\beta_6$ defined above) which computes the piece-wise linear function $f : [\beta_1, \beta_6) \to \mathbb{R}$ given by

$$f(x) \coloneqq \begin{cases} 0 & x \in [\beta_1, \beta_2] \\ \frac{1}{\beta_3-\beta_2}x - \frac{\beta_2}{\beta_3-\beta_2} & x \in (\beta_2, \beta_3) \\ \frac{1}{\beta_3-\beta_4}x - \frac{\beta_4}{\beta_3-\beta_4} & x \in [\beta_3, \beta_4] \\ 0 & x \in (\beta_4, \beta_6) \end{cases}.$$

  The intuition behind the approximation is that we can use the first four neurons to approximate the slopes of the function f, and the last two remaining neurons to simulate a bias term

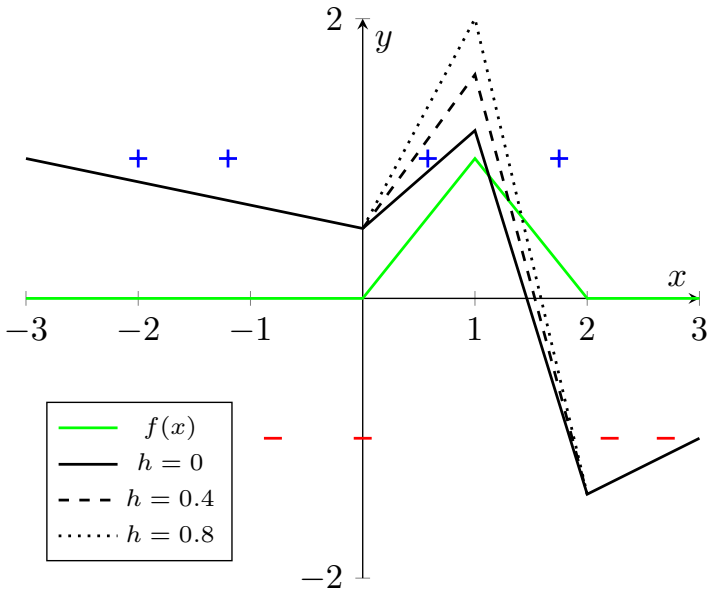

Figure 1: The plots of $f(x)$ (green) and $\Phi(\boldsymbol{\theta}; x) + h \cdot f(x)$ (black) for various values of h. Moving $\boldsymbol{\theta}$ in the direction of $\mathbf{u}$ which computes $f(\cdot)$ strictly decreases the loss over the positively-labeled instances in the interval $(0, 2)$ without affecting the rest of the dataset. Best viewed in color.

which would shift the function approximated by the network to overlap f in the relevant domain of approximation (see Figure 1 for an illustration).

More formally, define $W \coloneqq \operatorname{diag}(w_1, \ldots, w_4) \in \mathbb{R}^{4\times 4}$, define the masking matrix $A \in \mathbb{R}^{4\times 4}$ with entries $a_{j,j'} = \mathbb{1}\{j\text{-th neuron is active on the interval starting with } \beta_{j'}\}$ and let

$$\mathbf{d} = \left(0, \frac{1}{\beta_3 - \beta_2}, \frac{1}{\beta_3 - \beta_4}, 0\right).$$

Thus, to match the slopes computed by the a depth-2 ReLU network with weights $\mathbf{w}, \mathbf{b}$ to those of f, we first want the output neuron's weights $\mathbf{v} = [\mathbf{v}_1, \mathbf{v}_2] \in \mathbb{R}^4 \times \mathbb{R}^2$ to satisfy the equality $A^\top W \mathbf{v}_1 = \mathbf{d}$. To this end, we have by Lemma C.1 that A is invertible and $\left\|A^{-\top}\right\|_{\text{sp}} \leq 4$. It then follows that $\mathbf{v}_1 = W^{-1}A^{-\top}\mathbf{d}$, which entails

$$\|\mathbf{v}_1\|_2 \leq \left\|W^{-1}\right\|_{\text{sp}} \left\|A^{-\top}\right\|_{\text{sp}} \|\mathbf{d}\|_2 . \leq \frac{8\sqrt{2}}{qm}, \tag{8}$$

where we used the separability assumption, implying that $\|\mathbf{d}\|_2 \leq \sqrt{2}q^{-1}$ due to Item 3 which guarantees that $\beta_3 - \beta_2, \beta_4 - \beta_3 \geq q$, and the lower bound assumption $|w_i| \geq m$ for all $i \in [4]$ which holds by Items 1 and 2. Next, we use the two neurons with breakpoints at β_5, β_6 to shift the network by a constant so that it overlaps with f on the interval $[-R, R]$. To perform this shift, we first compute the magnitude by which we wish to shift which is given by the expression

$$b_0 \coloneqq -\sum_{j=1}^{4} v_j b_j \mathbb{1}\{w_j x + b_j > 0 \;\; \forall x \in (\beta_{j'}, \beta_{j'+1})\}.$$

Letting

$$P \coloneqq \begin{pmatrix} b_5 & b_6 \\ w_5 & w_6 \end{pmatrix},$$

we have that the neurons with breakpoints at β_5, β_6 compute a function which equals b_0 on the interval $[\beta_5, \beta_6]$ when the equality $P \cdot \mathbf{v}_2 = (b_0, 0)^\top$ is satisfied. We will now compute P^{-1} and show that it is well-defined. The inverse of a 2×2 matrix is given by

$$P^{-1} = \frac{1}{b_5 w_6 - b_6 w_5} \begin{pmatrix} w_6 & -b_6 \\ -w_5 & b_5 \end{pmatrix} = \frac{1}{\beta_6 - \beta_5} \begin{pmatrix} \frac{1}{w_5} & -\frac{b_6}{w_5 w_6} \\ -\frac{1}{w_6} & \frac{b_5}{w_5 w_6}. \end{pmatrix}$$

Using the above, we can upper bound the spectral norm of P^{-1} by upper bounding $1/(\beta_6 - \beta_5)$ with $\frac{1}{2R} \leq \frac{1}{2}$ since β_5, β_6 are outside the interval $[-R, R]$ at opposite sides and $R \geq 1$ by Assumption 2.2, and by upper bounding the spectral norm of the matrix with its Frobenius norm by using Item 2, to obtain

$$\left\|P^{-1}\right\|_{\text{sp}} \leq \frac{M}{m^2}.$$

Similarly, we derive an upper bound on $|b_0|$ using Cauchy-Schwartz and Items 1 and 2 to obtain

$$|b_0| \leq \|\mathbf{v}_1\|_2 \|\mathbf{b}\|_2 \leq 2M \|\mathbf{v}_1\|_2 .$$

With the above, we can bound the norm of $\mathbf{v}_2$ as follows

$$\|\mathbf{v}_2\|_2 = \left\|P^{-1} \cdot (b_0, 0)^\top\right\|_{\text{sp}} \leq \frac{2M^2}{m^2} \|\mathbf{v}_1\|_2 . \tag{9}$$

We now define $\mathbf{u}$ as the all-zero vector, except for the six output neuron entries corresponding to the neurons with breakpoints $\beta_1, \ldots, \beta_6$, where the coordinates of $\mathbf{u}$ take the values $y_i v_1, \ldots, y_i v_6$. Note that this entails

$$\|\mathbf{u}\|_2 = \sqrt{\|\mathbf{v}_1\|_2^2 + \|\mathbf{v}_2\|_2^2} \leq \|\mathbf{v}_1\|_2 \sqrt{1 + \frac{4M^4}{m^4}} \leq \frac{8}{qm}\sqrt{2 + \frac{8M^4}{m^4}} \leq 8\sqrt{10}\frac{M^2}{qm^3}, \tag{10}$$

where we have used Eq. (8) and the fact that $1 < M/m$. Next, we have for all $j \in [n]$ and $h > 0$ that

$$y_j \Phi\left(\boldsymbol{\theta} + \frac{h}{\|\mathbf{u}\|}\mathbf{u}; x_j\right) = y_j \left(\Phi(\boldsymbol{\theta}; x_j) + \frac{h}{\|\mathbf{u}\|} y_i f(x_j)\right).$$

Observe that all data points satisfy $x_j \in [\beta_5, \beta_6]$, and are therefore unaffected by the value $\Phi(\cdot, x)$ attains for x's outside of this interval. Additionally, $f(x) = 0$ for all $x \in [\beta_1, \beta_2] \cup (\beta_4, \infty)$, which also keeps $\Phi(\cdot, x)$ unaffected by moving in the direction of $\mathbf{u}$. Moreover, since the sign of data instances in $x_j \in (\beta_2, \beta_4]$ is always y_i, we have that

$$y_j \Phi\left(\boldsymbol{\theta} + \frac{h}{\|\mathbf{u}\|}\mathbf{u}; x_j\right) = y_j \Phi(\boldsymbol{\theta}; x_j) + \frac{h}{\|\mathbf{u}\|} f(x_j) \geq y_j \Phi(\boldsymbol{\theta}; x_j).$$

Lastly, for x_i we have that $x_i \in [\beta_3, \beta_4]$, and that x_i is at distance at least γ from the boundary by Item 5 in our separability assumption. By Item 3, this implies that $f(x_i)$ is at least $\frac{\gamma}{\beta_4 - \beta_3} \geq \frac{\gamma}{Q}$, which with the above equation and our bound from Eq. (10) yields

$$y_i \Phi\left(\boldsymbol{\theta} + \frac{h}{\|\mathbf{u}\|}\mathbf{u}; x_i\right) \geq y_i \Phi(\boldsymbol{\theta}; x_i) + h\frac{\gamma}{Q\|\mathbf{u}\|} \geq y_i \Phi(\boldsymbol{\theta}; x_i) + h\frac{\gamma q m^3}{8\sqrt{10} Q M^2}. \tag{11}$$

- Suppose that $x_i \in I_\ell$ for some $\ell \in \{1, \ldots, r+1\}$, such that $x_i \notin (\beta_1(I_\ell), \beta_3(I_\ell))$. Assume without loss of generality that $x_i \leq \beta_4 := \beta_1(I_\ell)$ (the proof is symmetric otherwise). Then by our separability assumption, there exist $\beta_2 := \beta_1(I_{\ell-1}), \beta_3 := \beta_3(I_{\ell-1}), \beta_5 := \beta_3(I_\ell)$ and β_1 whose neuron is active on all the data points where $\beta_1 \leq \beta_2$. Note that by the definition of $\beta_i(\cdot)$, it must hold that β_4 is the smallest element in the interval I_ℓ which satisfies Items 1 and 3 (since otherwise we would have that $x_i \in (\beta_3, \beta_5) \subseteq I_\ell$, which is handled in the previous case), and therefore $\beta_3 \notin I_\ell$, implying that $x_i \in (\beta_3, \beta_4)$. Moreover, we note that we may assume that $\ell > 1$, since otherwise we can take the smallest two breakpoints that are active on all the data along with β_4 which reduces us to the previous case. We thus denote the largest data instance in $I_{\ell-1}$ as x_{i-1} which implies $x_{i-1} < x_i$. Lastly, Item 2 also guarantees the existence of two breakpoints that are distinct from the previous ones, which we denote by β_6, β_7, where $\beta_6 < \beta_1 < 0$ and $\beta_7 > 0$ are active on all the data points.

  Following a similar approach as in the previous case, we will now show the existence of a depth-2 ReLU network which consists of seven hidden neurons with weights $\mathbf{w} = (w_1, \ldots, w_7)$ and biases $\mathbf{b} = (b_1, \ldots, b_7)$ and computes the piece-wise linear function

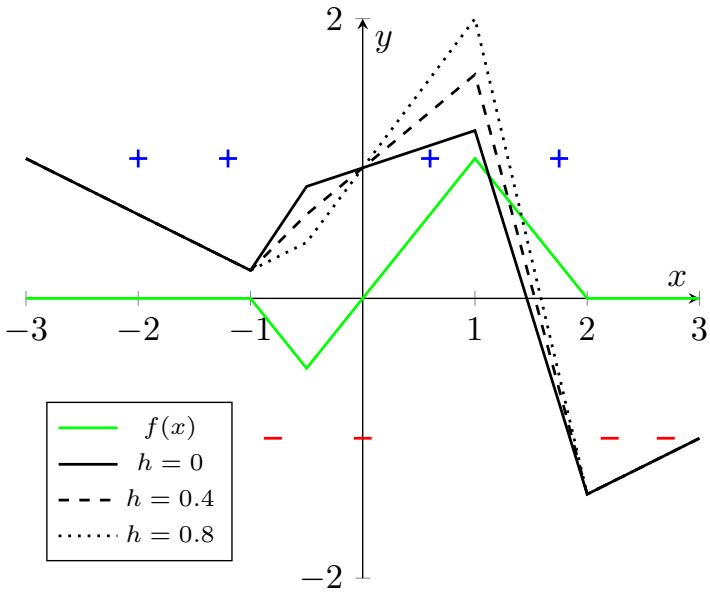

Figure 2: The plots of $f(x)$ (green) and $\Phi(\boldsymbol{\theta}; x) + h \cdot f(x)$ (black) for various values of h. Moving $\boldsymbol{\theta}$ in the direction of $\mathbf{u}$ which computes $f(\cdot)$ strictly decreases the loss over the positively-labeled instances in the interval $(0, 2)$ without degrading the loss over the rest of the dataset. Unlike the previous simpler case, since $\boldsymbol{\theta}$ has no breakpoints between the negative instance at $x = 0$ and the positive instance at $x = 0.6$, we use a function $f(x)$ which pivots around $x = 0$ to prevent the prediction over the negatively-labeled instances in the interval $[-1, 0]$ from increasing. Best viewed in color.

$f : [\beta_1, \beta_7) \to \mathbb{R}$ given by

$$f(x) := \begin{cases} 0 & x \in [\beta_1, \beta_2] \\ \frac{1}{\beta_2 - \beta_3} x - \frac{\beta_2}{\beta_2 - \beta_3} & x \in (\beta_2, \beta_3) \\ \frac{1}{x_{i-1} - \beta_3} x - \frac{x_{i-1}}{x_{i-1} - \beta_3} & x \in [\beta_3, \beta_4] \\ \frac{\beta_4 - x_{i-1}}{x_{i-1} - \beta_3} \left(\frac{1}{\beta_4 - \beta_5} x - \frac{\beta_5}{\beta_4 - \beta_5} \right) & x \in (\beta_4, \beta_5) \\ 0 & x \in [\beta_5, \beta_7) \end{cases},$$

where the first five neurons are used to compute the slopes of f and the remaining last two neurons are used to simulate a bias term to shift the network to accord with f in its domain (see Figure 2 for an illustration).

Define $W := \operatorname{diag}(w_1, \ldots, w_5) \in \mathbb{R}^{5 \times 5}$, define the masking matrix $A \in \mathbb{R}^{5 \times 5}$ with entries $a_{j,j'} = \mathbb{1}\{j\text{-th neuron is active on the interval starting with } \beta_{j'}\}$ and let

$$\mathbf{d} = \left(0, \frac{1}{\beta_2 - \beta_3}, \frac{1}{x_{i-1} - \beta_3}, \frac{\beta_4 - x_{i-1}}{x_{i-1} - \beta_3} \cdot \frac{1}{\beta_4 - \beta_5}, 0 \right).$$

Thus, to match the slopes computed by the a depth-2 ReLU network with weights $\mathbf{w}, \mathbf{b}$ to those of f, we first want the output neuron's weights $\mathbf{v} = [\mathbf{v}_1, \mathbf{v}_2] \in \mathbb{R}^5 \times \mathbb{R}^2$ to satisfy the equality $A^\top W \mathbf{v}_1 = \mathbf{d}$. To this end, we have by Lemma C.1 that A is invertible and $\left\| A^{-\top} \right\|_{\text{sp}} \leq 5$. It then follows that $\mathbf{v}_1 = W^{-1} A^{-\top} \mathbf{d}$, which entails

$$\|\mathbf{v}_1\|_2 \leq \left\| W^{-1} \right\|_{\text{sp}} \left\| A^{-\top} \right\|_{\text{sp}} \|\mathbf{d}\|_2 \leq 5 \left\| W^{-1} \right\|_{\text{sp}} \|\mathbf{d}\|_2 .$$

To upper bound the above, we first bound $\|\mathbf{d}\|_2$. By our separability assumption we have

$$\|\mathbf{d}\|_2 \leq \sqrt{\frac{1}{q^2} + \frac{1}{(x_{i-1}-\beta_3)^2} + \frac{4Q^2}{q^2(x_{i-1}-\beta_3)^2}}$$
$$= \sqrt{\frac{q^2 + (x_{i-1}-\beta_3)^2 + 4Q^2}{q^2(x_{i-1}-\beta_3)^2}} \leq \frac{1}{q(x_{i-1}-\beta_3)}\sqrt{8Q^2+q^2}$$

where we used Item 3 to upper bound the denominators, and Item 4 which entails $\beta_4 - x_{i-1}, x_{i-1} - \beta_3 \leq \beta_4 - \beta_3 \leq Q$. By Item 1, we have $|w_i| \geq m$ for all $i \in [5]$, and thus we obtain

$$\|\mathbf{v}_1\|_2 \leq \frac{5}{mq(x_{i-1}-\beta_3)}\sqrt{40Q^2+5q^2} \leq \frac{15\sqrt{5}Q}{mq(x_{i-1}-\beta_3)}. \tag{12}$$

Now, similarly to the previous case, we wish to shift the function computed by the depth-2 ReLU network by a constant. This is done in the exact same manner as in the previous case, where we shift it by a magnitude given by

$$b_0 = -\sum_{j=1}^{5} v_j b_j \mathbb{1}\left\{w_j x + b_j > 0 \;\; \forall x \in (\beta_{j'}, \beta_{j'+1})\right\}.$$

Bounding $|b_0|$ using its above definition, Cauchy-Schwartz and Item 1 in our separability assumption, we obtain

$$|b_0| \leq \|\mathbf{v}_1\|_2 \|\mathbf{b}\|_2 \leq \sqrt{5}M \|\mathbf{v}_1\|_2.$$

From the above and Eq. (9), we can shift the network by the desired magnitude using a vector $\mathbf{v}_2 = (v_6, v_7)$ satisfying

$$\|\mathbf{v}_2\|_2 \leq \frac{\sqrt{5}M^2}{m^2}\|\mathbf{v}_1\|_2.$$

We now define $\mathbf{u}$ as the all-zero vector, except for the output neuron entries corresponding to the neurons with breakpoints $\beta_1, \ldots, \beta_7$, where the coordinates of $\mathbf{u}$ take the values $y_i v_1, \ldots, y_i v_7$. Note that this entails

$$\|\mathbf{u}\|_2 = \sqrt{\|\mathbf{v}_1\|_2^2 + \|\mathbf{v}_2\|_2^2} \leq \|\mathbf{v}_1\|_2 \sqrt{1 + \frac{5M^4}{m^4}} \leq \frac{15\sqrt{30}QM^2}{m^3 q(x_{i-1}-\beta_3)} \leq \frac{90QM^2}{m^3 q(x_{i-1}-\beta_3)}, \tag{13}$$

where we used Eq. (12) and the fact that $1 < M/m$. We therefore have for all $j \in [n]$ and $h > 0$ that

$$y_j \Phi\left(\boldsymbol{\theta} + \frac{h}{\|\mathbf{u}\|}\mathbf{u}; x_j\right) = y_j(\Phi(\boldsymbol{\theta}; x_j) + \frac{h}{\|\mathbf{u}\|} y_i f(x_j)).$$

Observe that all data points satisfy $x_j \in [\beta_6, \beta_7]$, and are therefore unaffected by the value $\Phi(\cdot, x)$ attains for x's outside of this interval. Additionally, $f(x) = 0$ for all $x \in [\beta_1, \beta_2] \cup [\beta_5, \infty)$, which also keeps $\Phi(\cdot, x)$ unaffected by moving in the direction of $\mathbf{u}$. In the interval (β_2, x_{i-1}), the sign of $\Phi(\cdot, x)$ is $-y_i$, and in the interval (x_{i-1}, β_5) its sign changes to y_i. For this reason, similarly to the previous case, we have that

$$y_j \Phi\left(\boldsymbol{\theta} + \frac{h}{\|\mathbf{u}\|}\mathbf{u}; x_j\right) = y_j \Phi(\boldsymbol{\theta}; x_j) + \frac{h}{\|\mathbf{u}\|} f(x_j) \geq y_j \Phi(\boldsymbol{\theta}; x_j), \tag{14}$$

for all $x_j \in (\beta_2, \beta_5)$. Lastly, for x_i we have that $x_i \in [\beta_3, \beta_4]$, and that x_i is at distance at least γ from x_{i-1} by Item 5 in our separability assumption. This implies that

$$f(x_i) = \frac{x_i}{x_{i-1}-\beta_3} - \frac{x_{i-1}}{x_{i-1}-\beta_3} \geq \frac{\gamma}{x_{i-1}-\beta_3},$$

and therefore $hf(x_i) \geq \frac{\gamma}{x_{i-1}-\beta_3} h$, which with Eqs. (13,14) implies that

$$y_i \Phi\left(\boldsymbol{\theta} + \frac{h}{\|\mathbf{u}\|}\mathbf{u}; x_i\right) \geq y_i \Phi(\boldsymbol{\theta}; x_i) + h\frac{\gamma}{\|\mathbf{u}\|(x_{i-1}-\beta_3)} \geq y_i \Phi(\boldsymbol{\theta}; x_i) + h\frac{\gamma q m^3}{90QM^2}, \tag{15}$$

which is a weaker lower bound than the one derived in Eq. (11), and thus always holds if Eq. (11) is satisfied.

We now turn to lower bound the norm of the gradient by analyzing the directional derivative of $\mathcal{L}(\cdot)$ in the direction of the vector $\mathbf{u}$ defined by the above two cases. To this end, denote $\bar{\mathbf{u}} = \frac{\mathbf{u}}{\|\mathbf{u}\|}$ and compute

$$\begin{aligned}
\|\nabla\mathcal{L}(\boldsymbol{\theta})\|_2 \geq |\langle\nabla\mathcal{L}(\boldsymbol{\theta}), \bar{\mathbf{u}}\rangle| &= \left|\lim_{h\to 0}\frac{1}{hn}\sum_{j=1}^{n}\left(\ell(y_j\Phi(\boldsymbol{\theta}+h\bar{\mathbf{u}};x_j)) - \ell(y_j\Phi(\boldsymbol{\theta};x_j))\right)\right| \\
&= \lim_{h\to 0}\frac{1}{hn}\sum_{j=1}^{n}\left(\ell(y_j\Phi(\boldsymbol{\theta};x_j)) - \ell(y_j\Phi(\boldsymbol{\theta}+h\bar{\mathbf{u}};x_j))\right) \\
&\geq \lim_{h\to 0}\frac{1}{hn}\left(\ell(y_i\Phi(\boldsymbol{\theta};x_i)) - \ell(y_i\Phi(\boldsymbol{\theta}+h\bar{\mathbf{u}};x_i))\right) \\
&\geq \lim_{h\to 0}\frac{1}{hn}\left(\ell(y_i\Phi(\boldsymbol{\theta};x_i)) - \ell\left(y_i\Phi(\boldsymbol{\theta};x_i) + h\frac{\gamma q m^3}{90QM^2}\right)\right) \\
&= -\frac{\gamma q m^3}{90nQM^2}\ell'(y_i\Phi(\boldsymbol{\theta};x_i)). \qquad (16)
\end{aligned}$$

In the above, the first inequality is by Cauchy-Schwartz; the second equality is due to Eq. (14), which guarantees that each summand is non-positive; the second inequality is another application of Eq. (14) which guarantees that we're omitting only non-negative terms; and the last inequality is due to Eq. (15). We now lower bound the above expression depending on whether $\ell(\cdot)$ is the exponential or the logistic loss.

First assume that $\ell(\cdot)$ is the exponential loss. Then we have that $-\ell'(x) = \ell(x)$ for all $x \in \mathbb{R}$, and therefore we can directly use the inequality $\ell(y_i\Phi(\boldsymbol{\theta};x_i)) > 1/2n$.

In the case where $\ell(\cdot)$ is the logistic loss, we have that

$$-\ell'(x) = \frac{1}{1+\exp(x)} = 1 - \frac{1}{1+\exp(-x)}.$$

By the inequality $\ell(y_i\Phi(\boldsymbol{\theta};x_i)) > 1/2n$, we have $1+\exp(-y_i\Phi(\boldsymbol{\theta};x_i)) > \exp(1/2n)$, implying that

$$-\ell'(y_i\Phi(\boldsymbol{\theta};x_i)) > 1 - \exp\left(-\frac{1}{2n}\right) \geq \frac{1}{4n},$$

where we used the inequality $\exp(-x) \leq 1 - 0.5x$ which holds for all $x \in [0,1]$. Combining both loss cases and Eq. (16), we arrived at

$$\|\nabla\mathcal{L}(\boldsymbol{\theta})\|_2 \geq \frac{\gamma q m^3}{360n^2QM^2}.$$

Squaring the above and dividing by 2, the lemma follows. □

The following proposition establishes the separability (Def. C.1) of a neighborhood in weight space around our initialization point (Def. C.2) from the dataset S.

Proposition C.1 (Bounded Gradient with High Probability)**.** *Under Assumptions 2.1 and 2.2, given any $\delta \in (0,1)$, suppose that the following hold*

$$k \geq 6144 \cdot \frac{R^4 \log\left(\frac{24r}{\delta}\right)}{\rho} \quad \textit{and} \quad \Delta = \frac{\delta\rho\sigma_h}{24nkCR^3}. \qquad (17)$$

Then with probability at least $1-\delta$, for all $\boldsymbol{\theta} \in U_\Delta(\boldsymbol{\theta}(0)) \cap L^+_{1/2n}(\mathcal{L})$, we have that

$$\frac{1}{2}\|\nabla\mathcal{L}(\boldsymbol{\theta})\|_2^2 \geq 3\cdot 10^{-11}\frac{\delta^2\rho^2}{n^6r^2C^2R^8}\sigma_h^2.$$

Proof. To prove the proposition, we will show that for all $\boldsymbol{\theta} \in U_\Delta(\boldsymbol{\theta}(0)) \cap L^+_{1/2n}(\mathcal{L})$, $\boldsymbol{\theta}$ is separable from S with high probability. By Lemma C.2, this would imply the proposition. We will begin with proving items 1-4 jointly, and then we will show item 5 separately.

1-4. Under Assumption 2.1, suppose we are given some $a \in [0, R]$ and $\boldsymbol{\xi} := (\xi_1, \xi_2) \in \{-1, 1\}^2$, and sufficiently small $\varepsilon > 0$ such that

$$\varepsilon \leq \frac{512R^4 \log\left(\frac{24r}{\delta}\right)}{k} \leq \frac{\rho}{6} \leq \frac{R}{6}, \tag{18}$$

where the first inequality is by assumption on ε, the second inequality is by the lower bound on k in Eq. (17), and the last inequality follows from $\rho \leq \frac{2R}{r+1} \leq R$ since ρ must be smaller than the average length of an interval and since we assume $r \geq 1$. We now consider the event denoted by $E_{a,\varepsilon,\boldsymbol{\xi}}$ where the weights w_i, b_i of the i-th neuron satisfy

$$w_i \in \left(\xi_1 \frac{\sigma_{\mathrm{h}}}{R}, \xi_1 \frac{2\sigma_{\mathrm{h}}}{R}\right) \quad \text{and} \quad -\frac{b_i}{w_i} \in (\xi_2 a, \xi_2(a+\varepsilon)), \tag{19}$$

Since such an event is symmetric about 0, it is unaffected by the signs of $\boldsymbol{\xi}$. We can therefore assume without loss of generality that both intervals in Eq. (19) are contained in the positive real line and omit $\boldsymbol{\xi}$ from our notation. Under this assumption, the probability of $E_{a,\varepsilon}$ can be given in terms of Owen's T function which is defined by

$$T(h, a) := \frac{1}{2\pi} \int_0^a \frac{\exp\left(-\frac{1}{2}h^2(1+x^2)\right)}{1+x^2} dx$$

(see Owen [1956]), yielding

$$\mathbb{P}[E_{a,\varepsilon}] = T\left(\frac{1}{R}, a+\varepsilon\right) - T\left(\frac{1}{R}, a\right) - \left(T\left(\frac{2}{R}, a+\varepsilon\right) - T\left(\frac{2}{R}, a\right)\right).$$

Using the definition of $T(\cdot, \cdot)$, the above can be simplified to

$$\begin{aligned}
\mathbb{P}[E_{a,\varepsilon}] &= \frac{1}{2\pi} \int_a^{a+\varepsilon} \frac{\exp\left(-\frac{1}{2R^2}(1+x^2)\right)\left(1 - \exp\left(-\frac{3}{2R^2}(1+x^2)\right)\right)}{1+x^2} dx \\
&\geq \frac{1}{2\pi}\left(1 - \exp\left(-\frac{3}{2R^2}\right)\right) \int_a^{a+\varepsilon} \frac{\exp\left(-\frac{1}{2R^2}(1+x^2)\right)}{1+x^2} dx \\
&\geq \frac{3}{8\pi R^2} \int_a^{a+\varepsilon} \frac{\exp\left(-\frac{1}{2R^2}(1+x^2)\right)}{1+x^2} dx \\
&\geq \frac{3}{8\pi R^2} \int_a^{a+\varepsilon} \frac{\exp\left(-\frac{1}{2R^2}(1+(R+\varepsilon)^2)\right)}{1+(R+\varepsilon)^2} dx \geq \frac{\varepsilon}{512R^4}.
\end{aligned} \tag{20}$$

In the above, the second inequality follows from the inequality $1 - \exp(-x) \geq 0.5x$ which holds for all $x \in [0, 1.5]$ and from the fact that $1 \leq R$; the third inequality follows from the fact that the integrand is a monotonically decreasing function and $|a| \leq R$; and the last inequality follows from $\varepsilon \leq R$ which is implied by Eq. (18) and allows us to lower bound the numerator of the integrand by $\exp(-2.5)$ and upper bound the denominator by $5R^2$, and the fact that $3/(40\pi \exp(2.5)) \geq 1/512$.

Next, given some interval $I_j := (x_{i_j}, x_{i_{j+1}+1})$, $j \in [r+1]$, where the classification does not change signs on the data, we consider the three sub-intervals given by

$$\begin{aligned}
I_{j_1} &:= (x_{i_j}, x_{i_j} + \varepsilon), \\
I_{j_2} &:= \left(\frac{x_{i_j} + x_{i_{j+1}+1}}{2} - \frac{\varepsilon}{2}, \frac{x_{i_j} + x_{i_{j+1}+1}}{2} + \frac{\varepsilon}{2}\right), \\
I_{j_3} &:= \left(x_{i_{j+1}+1} - \varepsilon, x_{i_{j+1}+1}\right).
\end{aligned}$$

We remark that due to Eq. (18), the above sub-intervals are all disjoint and the distance between the intervals is positive. We now wish to show that Items 1 and 3 hold. We have from Eq. (20) that the probability that a given sub-interval of length ε contains no breakpoint is at most

$$(1 - \mathbb{P}[E_{a,\varepsilon}])^k \leq \left(1 - \frac{\log\left(\frac{24r}{\delta}\right)}{k}\right)^k \leq \exp\left(-\log\left(\frac{24r}{\delta}\right)\right) = \frac{\delta}{24r},$$

where we used the inequality $(1-x/y)^y \leq \exp(-x)$ which holds for all $x, y > 0$. There are exactly $3 \cdot (r+1) \leq 6r$ sub-intervals, therefore by a union bound we have that Items 1 and 3 hold for some positive q, Q, m, M with probability at least $1 - \frac{\delta}{4}$.

Next, we show Item 2. By Eq. (20), we have

$$\mathbb{P}\left[E_{R,R/6}\right] \geq \frac{1}{3072R^3}.$$

Thus, the probability of initializing a neuron with breakpoint in $(R, \frac{7}{6}R)$ which is active on all the data points is at least $\frac{1}{6144R^3}$, since there's an independent 0.5 probability that it has the correct orientation. This entails that the probability of initializing at most one neuron which is active on all the data points and has a breakpoint in $(R, \frac{7}{6}R)$ is upper bounded by

$$\begin{aligned}\left(1-\mathbb{P}\left[E_{R,R/6}\right]\right)^k + k\left(1-\mathbb{P}\left[E_{R,R/6}\right]\right)^{k-1}\mathbb{P}\left[E_{R,R/6}\right] &\leq 2\left(1-\frac{\mathbb{P}\left[E_{R,R/6}\right]}{2}\right)^k \\ &\leq 2\left(1-\frac{1}{6144R^3}\right)^k \leq 2\exp\left(-\frac{k}{6144R^3}\right) \\ &\leq 2\exp\left(-\frac{R\log(24r/\delta)}{\rho}\right) \leq \frac{2\delta}{24r} \leq \frac{\delta}{12}.\end{aligned}$$

In the above, the first inequality follows from the inequality $(1-x)^k + k(1-x)^{k-1}x \leq 2(1-x/2)^k$ which holds for any natural k and all $x \in [0,1]$,[6] the third inequality follows from $(1-1/x)^x \leq \exp(-1)$ for all $x > 0$, the fourth inequality follows from our lower bound on k in Eq. (17), and the penultimate inequality holds due to Eq. (18) which entails $\rho \leq R$. Therefore, by the above and a union bound on the symmetric event where two neurons are initialized in $(-\frac{7}{6}R, -R)$, we have that Item 2 holds for some m, M with probability at least $1 - \frac{\delta}{6} \geq 1 - \frac{\delta}{4}$.

We will now derive explicit bounds on the constants m, M, q, Q, and in addition we will show that Item 4 holds. Applying a union bound on the two previous cases, we have that Items 1-3 hold with probability at least $1 - \delta/2$. In such a case, we get an explicit lower bound on q as follows

$$q \geq \frac{|I_j| - 3\varepsilon}{2} \geq \frac{\rho - 3\varepsilon}{2} \geq \frac{\rho}{4},$$

where in the second inequality we used the fact that $|I_j| \geq \rho$ for all $j \in [r+1]$ which holds by the definition of ρ and in the last inequality we used Eq. (18). To bound Q in Item 4, we first argue that under the realization of $E_{R,R/6}$, the four neurons that are active on all the data points have a breakpoint with absolute value at most $\frac{7}{6}R \leq 2R$. To upper bound Q in Item 3, observe that under the realization of the previous events we have that

$$\beta_{i+1}(I_j) - \beta_i(I_j) \leq \frac{|I_j| + \varepsilon}{2} \leq R + \frac{\varepsilon - \rho}{2} \leq R,$$

for all $j \in [r+1]$ and $i \in [2]$, which follows from $|I_j| \leq 2R - \rho$ since $r \geq 1$ and $I_j \subseteq [-R, R]$ (i.e. there exists at least one interval other than I_j which has length at least ρ), and from the inequality $\varepsilon \leq \rho/6$ which holds by Eq. (18). To upper bound Q in Item 4, we bound the term $|\beta_1(I_{j+1}) - \beta_3(I_j)|$. Observe that under the realization of the above events we have that $\beta_1(I_{j+1}) \in (x, x+\varepsilon)$ and $\beta_3(I_j) \in (x'-\varepsilon, x')$ where $x < x'$ are the largest and smallest data instances in I_j, I_{j+1}, respectively. We therefore have

$$|\beta_1(I_{j+1}) - \beta_3(I_j)| \leq \max\{|x-x'|, |x-x'+2\varepsilon|\} \leq |x-x'| + 2\varepsilon \leq \frac{7}{3}R,$$

where the second inequality follows from the triangle inequality and the last inequality follows from the fact that $x_i \in [-R, R]$ for all $i \in [n]$ and from Eq. (18). Turning to bound

[6] To show this inequality holds, consider k i.i.d. random variables $X_j \sim U([0,1])$. Then the left-hand side equals $\mathbb{P}[|\{x_j : x_j \in [0,x]\}| \leq 1]$. The occurrence of the complement of this event is implied if $x_i \in [0, x/2]$ and $x_{i'} \in [x/2, x]$ hold for some $i \neq i'$, therefore to upper bound the left-hand side it suffices to upper bound the complement of the event where $x_i \in [0, x/2]$ and $x_{i'} \in [x/2, x]$ hold for some $i \neq i'$. This in turn follows from applying a union bound on $\mathbb{P}[|\{x_j : x_j \in [0, x/2]\}| = 0]$ and $\mathbb{P}[|\{x_j : x_j \in [x/2, x]\}| = 0]$.

w_i, b_i, we have by Eqs. (18,19) that when $E_{a,\varepsilon}$ or $E_{R,R/6}$ occur then the i-th neuron satisfies $|w_i| \geq \sigma/R$ and

$$|b_i| \leq (R + \frac{1}{6}R)w_i \leq \frac{7}{6}R \cdot \frac{2\sigma_{\mathrm{h}}}{R} = \frac{7}{3}\sigma_{\mathrm{h}},$$

concluding the derivation of Items 1-4.

5. It will suffice to lower bound the probability of the event denoted by A where $x_i \notin (\alpha_j - \gamma/2, \alpha_j + \gamma/2)$ for all $i \in [n]$ and $j \in [r]$, where α_j is the j-th sign change of the ground truth function labelling y_i. The set $\cup_{j=1}^{r}(\alpha_j - \gamma/2, \alpha_j + \gamma/2)$ has Lebesgue measure of at most γr, and since by Assumption 2.2 we have that $\mu(x) \leq C$ for all $x \in [-R, R]$, we lower bound the probability of the event by the expression

$$\mathbb{P}[A] \geq (1 - \gamma r C)^n.$$

Plugging $\gamma = \frac{\delta}{4nrC}$ in the above which entails $\gamma r C \leq 1$ and using Bernoulli's inequality we have

$$\mathbb{P}[A] \geq 1 - \frac{\delta}{4}.$$

To conclude the derivation so far, using another union bound, we have shown that with probability at least $1 - 0.75\delta$, $\boldsymbol{\theta}(0)$ is separable from S with constants

$$\gamma_0 = \frac{\delta}{4nrC}, \quad q_0 = \frac{\rho}{4}, \quad Q_0 = \frac{7}{3}R, \quad m_0 = \frac{\sigma_{\mathrm{h}}}{R}, \quad M_0 = \frac{7}{3}\sigma_{\mathrm{h}}. \tag{21}$$

We will now show that the separability also holds in a Δ-hidden neighborhood of $\boldsymbol{\theta}(0)$ for an appropriately chosen $\Delta > 0$. To this end, we first establish that

$$\mathbb{P}\left[\min_{i\in[n], j\in[k]} |\beta_j - x_i| > \frac{\delta}{8nkC}\right] \geq 1 - \frac{\delta}{4}.$$

Suppose we have n data instances in $[-R, R]$, then a cover of radius $\frac{\delta\pi}{8kn}$ over these data points has a (one-dimensional Lebesgue) measure of at most $\frac{\delta\pi}{4k}$. Thus, the probability that a breakpoint will not be initialized within distance less than $\frac{\delta\pi}{8kn}$ from any point is at least $(1 - \frac{\delta}{4k})$. This is true since Assumption 2.1 implies that the distribution of a breakpoint is a standard Cauchy distribution with density at most $1/\pi$. We thus have that

$$\mathbb{P}\left[\min_{i\in[n], j\in[k]} |\beta_j - x_i| > \frac{\delta}{8nkC}\right] \geq \left(1 - \frac{\delta}{4k}\right)^k \geq 1 - \frac{\delta}{4}, \tag{22}$$

where the last inequality follows from Bernoulli's inequality. A final union bound now implies that the above bound holds with the previous implications with probability at least $1 - \delta$. Define

$$\Delta := \frac{\delta\rho\sigma_{\mathrm{h}}}{24nkCR^3},$$

we will now show that this implies the uniform separability of any $\boldsymbol{\theta} \in U_\Delta(\boldsymbol{\theta}(0))$ from S, by proving Items 1-4 jointly and Item 5 separately.

1-4. First, by Assumption 2.2, we have $1 = \int_{-R}^{R} \mu(x)dx \leq 2RC$ which with Eq. (18) implies that $\Delta \leq \frac{\delta\sigma_{\mathrm{h}}}{24nkR} \leq \frac{\sigma_{\mathrm{h}}}{24R}$. Since the weight and bias of each neuron in $U_\Delta(\boldsymbol{\theta}(0))$ change by at most Δ, we have

$$m \geq \frac{\sigma_{\mathrm{h}}}{R} - \frac{\sigma_{\mathrm{h}}}{24R} = \frac{23\sigma_{\mathrm{h}}}{24R} \quad \text{and} \quad M \leq \frac{7}{3}\sigma_{\mathrm{h}} + \frac{\sigma_{\mathrm{h}}}{24R} \leq \frac{57}{24}\sigma_{\mathrm{h}}.$$

To bound q and Q, we will first show that under our assumptions the breakpoints cannot move much. To this end, we show that for each neuron, the function $f(w, b) := -\frac{b}{w}$ is Lipschitz on $U_\Delta(\boldsymbol{\theta}(0))$. We have

$$\nabla f(w, b) = \left(\frac{b}{w^2}, -\frac{1}{w}\right),$$

and therefore for any neuron $(w, b) \in \boldsymbol{\theta}$ such that $\boldsymbol{\theta} \in U_\Delta(\boldsymbol{\theta}(0))$ we get

$$\|\nabla f(w,b)\| = \sqrt{\frac{b^2}{w^4} + \frac{1}{w^2}} \leq \sqrt{\frac{M^2}{m^4} + \frac{1}{m^2}} \leq \frac{24R}{23\sigma_{\mathrm{h}}}\sqrt{\frac{57^2}{23^2}R^2 + 1} < \frac{3R^2}{\sigma_{\mathrm{h}}},$$

where the last inequality follows from $1 \leq R^2$. This implies that

$$\left| -\frac{b}{w} + \frac{b_0}{w_0} \right| \leq \|\nabla f(w,b)\| \cdot \|(w,b) - (w_0, b_0)\| < \frac{3R^2}{\sigma_{\mathrm{h}}}\Delta \leq \frac{\delta\rho}{8nkCR}.$$

That is, we have that the breakpoint of each neuron moves a distance strictly less than $\frac{\delta\rho}{8nkCR} \leq \frac{\delta}{8nkC}$, which along with Eq. (22) guarantees that $\mathcal{L}(\cdot)$ is differentiable on $U_\Delta(\boldsymbol{\theta}(0))$ since no ReLU crosses a data instance. Since Q is the upper bound on the difference between two breakpoints where each moves by at most $\frac{\delta}{8nkC}$, this also yields a bound on Q as follows

$$Q \leq Q_0 + 2\frac{\delta}{8nkC} \leq \frac{7}{3}R + \frac{1}{8}R \leq 2.5R,$$

where we used the upper bound on Q_0 from Eq. (21), Eq. (17) which implies $k \geq 4$ (since $\rho \leq R$ by Eq. (18)), and $1/C \leq 2R$. Likewise, to lower bound q, compute

$$q \geq q_0 - 2\frac{\delta\rho}{8nkCR} \geq \frac{\rho}{4} - \frac{\rho}{20R} \geq \frac{\rho}{5},$$

where again we used Eq. (21), Eq. (17) which implies $k \geq 10$, and $1/C \leq 2R$.

5. Since γ depends on S and not on $\boldsymbol{\theta}$, it remains unchanged and we have $\gamma = \gamma_0$.

We can now use the assumption $\boldsymbol{\theta} \in L^+_{1/2n}(\mathcal{L})$ and Lemma C.2 to conclude

$$\frac{1}{2}\|\nabla\mathcal{L}(\boldsymbol{\theta})\|_2^2 \geq \frac{\gamma^2 q^2 m^6}{259200 n^4 Q^2 M^4} \geq \frac{1}{259200n^4} \cdot \frac{\delta^2}{4^2 n^2 r^2 C^2} \cdot \frac{4\rho^2}{25^2 R^2} \cdot \frac{23^6}{57^4 24^2 R^6} \cdot \sigma_{\mathrm{h}}^2.$$

Simplifying the above, the proposition follows.

□

Having established the required machinery for proving Theorem 3.1, we now turn to do so.

Proof of Theorem 3.1. We begin with bounding the loss upon initialization with high probability. First, consider $3k$ i.i.d. random variables $X_j \sim \mathcal{N}(0,1)$. We have that

$$\mathbb{P}\left[\max_{j\in[3k]} |X_j| \leq x\right] = \left(\mathrm{erf}\left(\frac{x}{\sqrt{2}}\right)\right)^{3k} \geq \left(1 - \exp\left(-0.5x^2\right)\right)^{3k} \geq 1 - 3k\exp\left(-0.5x^2\right),$$

where the first inequality follows from $1 - \mathrm{erf}(x) < \exp(-x^2)$ for all $x \geq 0$ (see Eq. (7.8.3) in DLMF) and the second inequality follows from Bernoulli's inequality since $\exp(-0.5x^2) < 1$. Plugging $x = \sqrt{2\log(6k/\delta)}$ in the above, we have

$$\mathbb{P}\left[\max_{j\in[3k]} |X_j| \leq \sqrt{2\log(6k/\delta)}\right] \geq 1 - \frac{3k\delta}{6k} = 1 - \frac{\delta}{2}.$$

Thus, with probability at least $1 - \frac{\delta}{2}$, we have that all the weights of $\boldsymbol{\theta}(0)$ are at most $\sqrt{2\log(6k/\delta)}$ standard deviations away from zero. With this bound, we can derive for all $x \in [-R, R]$

$$\mathcal{N}_{\boldsymbol{\theta}(0)}(x) \leq \sum_{j\in[k]} |v_j|\sigma(|w_j| \cdot |x| + |b_j|) \leq 4kR\sigma_{\mathrm{h}}\sigma_{\mathrm{o}}\log\left(\frac{6k}{\delta}\right),$$

which for both the exponential and logistic losses implies

$$\mathcal{L}(\boldsymbol{\theta}(0)) \leq \exp\left(4kR\sigma_{\mathrm{h}}\sigma_{\mathrm{o}}\log\left(\frac{6k}{\delta}\right)\right) \leq e, \tag{23}$$

where the last inequality is by our assumption $\sigma_{\mathrm{o}} \leq \frac{1}{4kR\sigma_{\mathrm{h}}\log\left(\frac{6k}{\delta}\right)}$. Letting

$$\lambda := 10^{-11}\frac{\delta^2\rho^2}{n^6r^2C^2R^8}$$

and observing that our lower bound assumption on σ_{h} in Eq. (3) implies it's at least 0.5 since $\rho \leq R$ and $C \geq 1/2R$, we can invoke Proposition C.1 with confidence $\frac{\delta}{2}$ to obtain

$$\begin{aligned}\frac{1}{2}\left\|\nabla\mathcal{L}(\boldsymbol{\theta}(t))\right\|_2^2 &\geq 3\cdot 10^{-11}\frac{\delta^2\rho^2}{n^6r^2C^2R^8}\sigma_{\mathrm{h}}^2 \\ &\geq 3\cdot 10^{-11}\frac{\delta^2\rho^2}{n^6r^2C^2R^8}\sigma_{\mathrm{h}}^2\cdot\frac{\mathcal{L}(\boldsymbol{\theta}(t))}{\mathcal{L}(\boldsymbol{\theta}(0))} \\ &\geq \lambda\sigma_{\mathrm{h}}^2\cdot\mathcal{L}(\boldsymbol{\theta}(t)),\end{aligned} \tag{24}$$

where the second inequality holds since $\frac{\mathcal{L}(\boldsymbol{\theta}(t))}{\mathcal{L}(\boldsymbol{\theta}(0))} \leq 1$ because the flow is non-increasing, and the last inequality holds due to Eq. (23) which implies $\frac{1}{\mathcal{L}(\boldsymbol{\theta}(0))} \geq \exp(-1) \geq \frac{1}{3}$. By a union bound, the above holds with probability at least $1-\delta$.

Denote $D := U_\Delta(\boldsymbol{\theta}(0)) \cap L^+_{\frac{1}{2n}}(\mathcal{L})$ where Δ is defined in Eq. (17), and define $t' \in [0,\infty)$ to be the smallest time such that $\boldsymbol{\theta}(t')$ is on the boundary of $U_\Delta(\boldsymbol{\theta}(0))$ (where $t' = \infty$ if there exists no such time). We will now show that the flow attains loss at most $\frac{1}{2n}$ in time $t_0 := \frac{\log(2n\mathcal{L}(\boldsymbol{\theta}(0)))}{2\lambda\sigma_{\mathrm{h}}^2}$, by analyzing several different cases.

- Suppose that $t' > t_0$.
  - If $\{\boldsymbol{\theta}(t) : t \in [0,t_0]\} \subseteq D$, then by the PL-condition shown in Eq. (24) we have for all $t \in [0,t_0]$ that GF enjoys a convergence rate of

    $$\mathcal{L}(\boldsymbol{\theta}(t)) \leq \exp\left(-2\lambda\sigma_{\mathrm{h}}^2 t\right)\cdot\mathcal{L}(\boldsymbol{\theta}(0)).$$

    Plugging $t = t_0$ in the above and simplifying, we have $\mathcal{L}(\boldsymbol{\theta}(t_0)) \leq \frac{1}{2n}$.
  - If $\{\boldsymbol{\theta}(t) : t \in [0,t_0]\} \not\subseteq D$, then there exists a time $t'' \leq t_0$ such that $\boldsymbol{\theta}(t'') \notin D$. Since $t'' \leq t_0 < t'$, it must hold that $t'' \notin L^+_{\frac{1}{2n}}(\mathcal{L})$, and therefore $\mathcal{L}(\boldsymbol{\theta}(t'')) < \frac{1}{2n}$ which implies $\mathcal{L}(\boldsymbol{\theta}(t_0)) < \frac{1}{2n}$ since the flow is non-increasing.
- Suppose that $t' \leq t_0$. Assume by contradiction that $\{\boldsymbol{\theta}(t) : t \in [0,t']\} \subseteq D$. We will now show that the length of the trajectory of GF cannot have been long enough to reach the boundary of D, which will result in a contradiction. To this end, we use a similar technique as in Gupta et al. [2021, Thm. 9]. Define the potential function $\varepsilon(t) = \sqrt{\mathcal{L}(\boldsymbol{\theta}(t))}$. Taking the derivative of $\varepsilon(t)$ with respect to t and using the chain rule we have

  $$\dot{\varepsilon}(t) = \frac{\frac{d\mathcal{L}(\boldsymbol{\theta}(t))}{dt}}{2\sqrt{\mathcal{L}(\boldsymbol{\theta}(t))}} = -\frac{\left\|\nabla\mathcal{L}(\boldsymbol{\theta}(t))\right\|_2^2}{2\sqrt{\mathcal{L}(\boldsymbol{\theta}(t))}} \leq -\sqrt{\frac{\lambda}{2}}\sigma_{\mathrm{h}}\cdot\left\|\nabla\mathcal{L}(\boldsymbol{\theta}(t))\right\|_2,$$

  where the inequality follows from Eq. (24). We can now bound the length of the trajectory up until time t' by using the fundamental theorem of calculus and obtain

  $$\begin{aligned}\int_0^{t'}\left\|\nabla\mathcal{L}(\boldsymbol{\theta}(t))\right\|_2 dt &\leq -\frac{1}{\sigma_{\mathrm{h}}}\sqrt{\frac{2}{\lambda}}\int_0^{t'}\dot{\varepsilon}(t)dt \leq -\frac{1}{\sigma_{\mathrm{h}}}\sqrt{\frac{2}{\lambda}}\left[\sqrt{\mathcal{L}(\boldsymbol{\theta}(t))}\right]_0^{t'} \\ &\leq \frac{1}{\sigma_{\mathrm{h}}}\sqrt{\frac{2\mathcal{L}(\boldsymbol{\theta}(0))}{\lambda}} \leq \frac{1}{\sigma_{\mathrm{h}}}\sqrt{\frac{2e}{\lambda}} < \Delta,\end{aligned}$$

  where in the second line, the first inequality uses the fact that $\mathcal{L}(\cdot) > 0$, the second inequality follows from Eq. (23), and the last inequality follows from our bound on σ_{h} assumed in Eq. (3) and the definition of Δ in Eq. (17). In contrast, since $\boldsymbol{\theta}(t')$ is on the boundary of $U_\Delta(\boldsymbol{\theta}(0))$, this implies that there exists some neuron with weight and bias $w(t), b(t)$ at time

$t \geq 0$ such that $\|(w(t'), b(t')) - (w(0), b(0))\|_2 = \Delta$. From this and the path length upper bound we have

$$\Delta \leq \|\boldsymbol{\theta}(t') - \boldsymbol{\theta}(0)\|_2 \leq \int_0^{t'} \|\nabla \mathcal{L}(\boldsymbol{\theta}(t))\|_2 \, dt < \Delta,$$

which is a contradiction. We therefore must have that $\{\boldsymbol{\theta}(t) : t \in [0, t']\} \not\subseteq D$. When this holds, there exists a time $t'' \leq t' \leq t_0$ such that $\boldsymbol{\theta}(t'') \notin D$. Since $t'' \leq t'$, it must hold that $t'' \notin L^+_{\frac{1}{2n}}(\mathcal{L})$, and therefore $\mathcal{L}(\boldsymbol{\theta}(t'')) < \frac{1}{2n}$ which implies $\mathcal{L}(\boldsymbol{\theta}(t_0)) < \frac{1}{2n}$ since the flow is non-increasing.

□

D Over-parameterization is necessary

In this appendix, we further discuss and formally prove Theorem 3.2, which establishes that in general under Assumption 2.1, an over-parameterization of magnitude at least $1.3r$ is necessary for achieving population loss below a constant. Our analysis is based on the following specific construction, where the labels are determined by a function f_r parameterized by a natural number r for all $x \in [-1, 1]$, expressible by the sign of a teacher network of width r and defined as

$$f_r(x) \coloneqq \text{sign}(\sin(0.5\pi(r+1)(x+1))). \tag{25}$$

That is, f_r changes value r times between -1 and 1 on the interval $[-1, 1]$, and is constant along intervals of length $\frac{2}{r+1}$. We now define the distribution $\mathcal{D}$ over the inputs of the dataset used in our lower bound and its corresponding labelling rule. We have

$$x \sim U[-1, 1] \quad \text{and} \quad y = f_r(x). \tag{26}$$

Recall the statement of Theorem 3.2, we have for example that if $\alpha = 1.3$, then the width of the network being trained is no more than $1.3r$ and GF attains loss at least $\frac{1}{160}$ in this case. While a lower bound of r neurons for the construction specified in Eq. (26) is trivially implied by function approximation considerations, our lower bound merely improves upon this quantity by a constant multiplicative factor. Nevertheless, it is interesting to compare our lower bound to other similar settings in the literature, since it is typically difficult to derive lower bounds that require strictly more than r neurons. For example, in a teacher-student setting where networks of the form $\mathbf{x} \mapsto \sum_{i=1}^r \sigma\left(\mathbf{w}_i^\top \mathbf{x}\right)$ are considered, it is known that there are spurious (non-global) minima already when $r \geq 6$ [Safran and Shamir, 2018, Arjevani and Field, 2020, 2021], and that empirically we are more likely to get stuck in those minima the larger r is [Safran and Shamir, 2018], but in spite of this ample empirical evidence, there is no proof that optimization will fail for any natural number $r \geq 6$. In contrast, in our univariate setting, it is possible to show a non-trivial lower bound since we utilize bias terms. This highlights the difference between settings that omit and include biases, which impacts the associated optimization problem in a non-trivial manner.

The proof of our lower bound, which appears below in Appendix D.1, relies on the observation that under Assumption 2.1, the breakpoints of the trained network upon initialization follow a standard Cauchy distribution. In such a case, neurons with a breakpoint outside the support of the data and with the wrong orientation will remain dormant throughout the optimization process, which requires initializing at least a fraction more of the minimal number of neurons required so that sufficiently many will be optimized and could improve the approximation of the target function. While one can circumvent this issue by scaling the breakpoints to the support of the data, this would require (i) an initialization scheme which is different than Assumption 2.1, which is used in our upper bounds; and (ii) this may even prove detrimental to optimization, as our positive result requires neurons that are active on all the data instances. We stress that our lower bound given here applies to training over a sample of any size, since it relies on approximation arguments. Additionally, we remark that by scaling the distribution $\mathcal{D}$ to be supported on a smaller interval we can increase the required magnitude of over-parameterization up to a factor of $\alpha = 2$, however due to the common practice of scaling the data to have unit norm, we assume it is supported on $[-1, 1]$. We also remark that a common initialization scheme is to set the bias terms to zero [He et al., 2015]. This results in breakpoints that are initialized at the origin and circumvents the issue of dormant neurons upon

initialization, however the main motivation for using such an initialization scheme is to achieve numerical stability and avoid exploding gradients when training very deep networks, which is not an issue for the shallow architecture we consider here. In any case, we stress that the goal of our lower bounds is to exemplify that over-parameterization is necessary in a setting complementary to our upper bound in Theorem 3.1, and we leave the derivation of stronger lower bounds under more general initialization schemes as a tantalizing future work direction.

D.1 Proof of Theorem 3.2

To prove the theorem, we would need the following auxiliary lemmas. The first lemma below establishes that if we approximate the function f_r which is defined in Eq. (25) by a function which does not change its sign over an interval of length larger than $\frac{2}{r+1}$, then this results in a strictly positive loss which is roughly proportional to the length of the approximation interval.

Lemma D.1. *Let f_r be as defined in Eq. (25), let $\beta_1, \beta_2 \in [-1, 1]$, $\mathcal{N} : [\beta_1, \beta_2] \to \mathbb{R}$ such that its sign is fixed on $[\beta_1, \beta_2]$, and let ℓ be either the exponential or logistic loss. Then*

$$\int_{\beta_1}^{\beta_2} \ell(\mathcal{N}(x) \cdot f_r(x))dx \geq \frac{1}{2}\ell(0)\left(\beta_2 - \beta_1 - \frac{2}{r+1}\right).$$

Proof. If $\beta_2 - \beta_1 \leq \frac{2}{r+1}$, then the right-hand side is non-positive and the lemma follows since $\ell(\cdot) > 0$. If $\beta_2 - \beta_1 > \frac{2}{r+1}$, then the (one-dimensional Lebesgue) measure of the set $A \coloneqq \{x \in [\beta_1, \beta_2] : \mathcal{N}(x) \cdot f_r(x) \leq 0\}$ is at least

$$\frac{1}{2}\left(\beta_2 - \beta_1 - \frac{2}{r+1}\right),$$

since the measure of the complementary set $\{x \in [\beta_1, \beta_2] : \mathcal{N}(x) \cdot f_r(x) > 0\}$ is at most $\frac{1}{2}\left(\beta_2 - \beta_1 + \frac{2}{r+1}\right)$, where the upper bound is attained when $\mathcal{N}(x) \cdot f_r(x) > 0$ for all

$$x \in \left[\beta_1, \beta_1 + \frac{2}{r+1}\right] \cup \left[\beta_2 - \frac{2}{r+1}, \beta_2\right].$$

We can therefore lower bound the integral in the lemma by

$$\int_{\beta_1}^{\beta_2} \ell(\mathcal{N}(x) \cdot f_r(x))dx \geq \int_{x \in A} \ell(\mathcal{N}(x) \cdot f_r(x))dx \geq \int_{x \in A} \ell(0)dx \geq \frac{1}{2}\ell(0)\left(\beta_2 - \beta_1 - \frac{2}{r+1}\right).$$

□

The following lemma shows that approximating f_r using a ReLU network with just r' neurons results in loss proportional to $1 - r'/r$.

Lemma D.2. *Suppose that f_r as defined in Eq. (25). Then for any ReLU network $\mathcal{N}_{\boldsymbol{\theta}}$ of width at most r', we have*

$$\mathcal{L}_{\mathcal{D}}(\boldsymbol{\theta}) \geq \frac{1}{4}\left(1 - \frac{r'}{r}\right).$$

Proof. Denote by $\beta_1, \ldots, \beta_{r'}$ the set of points where $\mathcal{N}$ changes sign in $(-1, 1)$. Note that this set is always of size at most r', and we may assume without loss of generality that it is of size exactly r' (since otherwise we can prove a stronger claim, where the lemma holds for some $r'' < r'$). Further define the boundaries $\beta_0 \coloneqq -1$ and $\beta_{r'+1} \coloneqq 1$. We compute

$$\begin{aligned}
\mathcal{L}_{\mathcal{D}}(\boldsymbol{\theta}) &= \int_{-1}^{1} \frac{1}{2}\ell(f_r(x) \cdot \mathcal{N}_{\boldsymbol{\theta}}(x))dx = \frac{1}{2}\sum_{i=0}^{r'} \int_{\beta_i}^{\beta_{i+1}} \ell(f_r(x) \cdot \mathcal{N}_{\boldsymbol{\theta}}(x))dx \\
&\geq \frac{1}{4}\ell(0)\sum_{i=0}^{r'}\left(\beta_{i+1} - \beta_i - \frac{2}{r+1}\right) = \frac{1}{4}\left(2 - 2\frac{r'+1}{r+1}\right) \\
&= \frac{1}{2} \cdot \frac{r - r'}{r+1} \geq \frac{1}{4}\left(1 - \frac{r'}{r}\right),
\end{aligned}$$

where the first inequality uses Lemma D.1, the equality that follows is due to the sum telescoping and since $\ell(0) = 1$ for both the exponential and logistic losses, and the final inequality is due to $r \geq 1$. $\square$

With the above auxiliary lemmas, we can now turn to the proof of the theorem.

Proof of Theorem 3.2. By Assumption 2.1, the breakpoints of $\mathcal{N}$ at initialization follow a standard Cauchy distribution. Since $\mathcal{D}$ is supported on $[-1, 1]$, we have with probability exactly 0.5 that the breakpoint of a given neuron falls outside of $[-1, 1]$. Moreover, with an independent probability of 0.5, the orientation of the neuron is such that it is off on all the data instances. I.e., such a neuron remains dormant throughout the optimization process of GF with probability 0.25. It can be verified that for any integer $k \geq 1$, at least $\lceil 0.25k \rceil$ neurons will be dormant upon initialization with probability at least 0.25.[7] Thus, with probability at least 0.25, we have that out of αr neurons, there are at most $r' = \lfloor 0.75\alpha r \rfloor$ neurons that are effectively being trained with breakpoints in $[-1, 1]$. By Lemma D.2, this results in a lower bound on the population loss of

$$\mathcal{L}_{\mathcal{D}}(\mathcal{N}_{\boldsymbol{\theta}(t)}) \geq \frac{1}{4}\left(1 - \frac{\lfloor 0.75\alpha r \rfloor}{r}\right) \geq \frac{1}{4}\left(1 - 0.75\alpha\right),$$

for any time $t \geq 0$. $\square$

E Proof of Theorem 4.2

By Theorem 4.1, if there exists time t_0 such that $\mathcal{L}(\boldsymbol{\theta}(t_0)) < \frac{1}{n}$ then GF converges to zero loss, and converges in direction to a KKT point of Problem (4). We denote $\mathcal{N}_{\boldsymbol{\theta}}(x) = \sum_{j\in[k]} v_j \sigma(w_j x + b_j)$. Thus, $\mathcal{N}_{\boldsymbol{\theta}}$ is a network of width k, where the weights in the first layer are $w_1, \ldots, w_k$, the bias terms are $b_1, \ldots, b_k$, and the weights in the second layer are $v_1, \ldots, v_k$. We denote $J := \{j \in [k] : v_j \neq 0\}$, $J^+ := \{j \in J : v_j > 0\}$, and $J^- := \{j \in J : v_j < 0\}$. Since neurons with output weight $v_j = 0$ do not affect the function that the network computes, then in this proof we ignore them. We also denote $I := [n]$, and $I' = \{i \in I : y_i \mathcal{N}_{\boldsymbol{\theta}}(x_i) = 1\}$. Thus, I' are the indices of the examples where $\mathcal{N}_{\boldsymbol{\theta}}$ attains margin of exactly 1.

Assume that $\mathcal{N}_{\boldsymbol{\theta}}$ satisfies the KKT conditions of Problem (4). Thus, there are $\lambda_1, \ldots, \lambda_n$ such that for every $j \in J$ we have

$$w_j = \sum_{i\in I} \lambda_i \frac{\partial}{\partial w_j}\left(y_i \mathcal{N}_{\boldsymbol{\theta}}(x_i)\right) = \sum_{i\in I} \lambda_i y_i v_j \sigma'_{i,j} x_i \,, \tag{27}$$

where $\sigma'_{i,j}$ is a subgradient of σ at $w_j \cdot x_i + b_j$, i.e., if $w_j \cdot x_i + b_j \neq 0$ then $\sigma'_{i,j} = \mathbb{1}[w_j \cdot x_i + b_j > 0]$, and otherwise $\sigma'_{i,j}$ is some value in $[0, 1]$ (we emphasize that in this case $\sigma'_{i,j}$ may be any value in $[0, 1]$ and in this proof we do not have any further assumptions on it). Also we have $\lambda_i \geq 0$ for all $i \in I$, and $\lambda_i = 0$ if $i \notin I'$. Likewise, we have

$$b_j = \sum_{i\in I} \lambda_i \frac{\partial}{\partial b_j}\left(y_i \mathcal{N}_{\boldsymbol{\theta}}(x_i)\right) = \sum_{i\in I} \lambda_i y_i v_j \sigma'_{i,j} \,. \tag{28}$$

We say that $\mathcal{N}_{\boldsymbol{\theta}}$ has an *activation point* at x if there is $j \in [J]$ with $w_j \neq 0$ such that $w_j \cdot x + b_j = 0$. In this case we say that *the activation point x corresponds to the neuron j*. Note that if $w_j = 0$ then the neuron computes a constant function and thus it does not affect the number of linear regions in $\mathcal{N}_{\boldsymbol{\theta}}$.

Lemma E.1. *We denote $I' = \{i_1, \ldots, i_q\}$ where $1 \leq i_1 < \ldots < i_q \leq [n]$. For every $\ell \in [q-1]$ the network $\mathcal{N}_{\boldsymbol{\theta}}$ has at most two activation points in the open interval $(x_{i_\ell}, x_{i_{\ell+1}})$. Moreover, $\mathcal{N}_{\boldsymbol{\theta}}$ has at most one activation point in $(-\infty, i_1)$ and at most one activation points in (i_q, ∞).*

[7] Essentially, this holds true since the median of a binomially-distributed random variable $B(n, k)$ is $\lceil nk \rceil$ or $\lfloor nk \rfloor$, and since deviating from the median by at most 1 never increases the probability of the tail to more than 0.75.

Proof. Let $x \in (x_{i_\ell}, x_{i_{\ell+1}})$ be an activation point, and let $j \in J$ such that $w_j \neq 0$ and $w_j \cdot x + b_j = 0$. Suppose first that $w_j > 0$. Since $w_j \cdot x + b_j = 0$ then for every $x' > x$ we have $w_j \cdot x' + b_j > 0$, and for every $x' < x$ we have $w_j \cdot x' + b_j < 0$. By Eq. (27) we have

$$w_j = \sum_{i \in I} \lambda_i y_i v_j \sigma'_{i,j} x_i = \sum_{i \in I'} \lambda_i y_i v_j \sigma'_{i,j} x_i \ .$$

Since $x \neq x_i$ for all $i \in I'$ then $w_j \cdot x_i + b_j \neq 0$, and we have $\sigma'_{i,j} = \mathbb{1}(w_j \cdot x_i + b_j > 0) = \mathbb{1}[x_i > x]$. Therefore, the above displayed equation equals

$$\sum_{i \in I'} \lambda_i y_i v_j \mathbb{1}[x_i > x] x_i = \sum_{i \in I', \, i \geq i_{\ell+1}} \lambda_i y_i v_j x_i \ . \tag{29}$$

Likewise, by Eq. (28) we have

$$b_j = \sum_{i \in I} \lambda_i y_i v_j \sigma'_{i,j} = \sum_{i \in I'} \lambda_i y_i v_j \sigma'_{i,j} = \sum_{i \in I'} \lambda_i y_i v_j \mathbb{1}[x_i > x] = \sum_{i \in I', \, i \geq i_{\ell+1}} \lambda_i y_i v_j \ . \tag{30}$$

By Eq. (29) and Eq. (30), the activation point x satisfies

$$x = \frac{-b_j}{w_j} = \frac{-\sum_{i \in I', \, i \geq i_{\ell+1}} \lambda_i y_i v_j}{\sum_{i \in I', \, i \geq i_{\ell+1}} \lambda_i y_i v_j x_i} = \frac{-\sum_{i \in I', \, i \geq i_{\ell+1}} \lambda_i y_i}{\sum_{i \in I', \, i \geq i_{\ell+1}} \lambda_i y_i x_i} \ .$$

Therefore if x and x' are two activation points in $(x_{i_\ell}, x_{i_{\ell+1}})$ that correspond to $w_j > 0$ and $w_{j'} > 0$ respectively, then $x = x'$. Thus, there is at most one activation point $x \in (x_{i_\ell}, x_{i_{\ell+1}})$ that corresponds to some $w_j > 0$.

Moreover, by similar arguments there is at most one activation point $x \in (x_{i_\ell}, x_{i_{\ell+1}})$ that corresponds to $w_j < 0$. Overall, in the interval $(x_{i_\ell}, x_{i_{\ell+1}})$ there are at most two activation points.

If $x \in (-\infty, i_1)$ then from similar argument we get that there is at most one activation point that corresponds to a neuron j with $w_j > 0$. Also, an activation point in $(-\infty, i_1)$ that corresponds to a neuron with $w_j < 0$ does not exist, since such neuron is not active for any input x_i with $i \in I'$, and hence by Eq. (27) we must have $w_j = 0$. The proof of the claim for the interval (i_q, ∞) is similar. □

Let $1 \leq a < b \leq n$ be indices such that for every $a \leq i < i' \leq b$ we have $y_i = y_{i'}$. Thus the labels do not switch signs for the inputs $x_a, x_{a+1}, \ldots, x_b$. Intuitively, the proof follows by showing that in the interval $[x_a, x_b]$ the network $\mathcal{N}_{\boldsymbol{\theta}}$ has a constant number of linear regions, and then concluding that the overall number of linear regions in $\mathcal{N}_{\boldsymbol{\theta}}$ must be $\mathcal{O}(r)$. We first consider the case where $x_b > x_a \geq 0$ and for all $a \leq i \leq b$ we have $y_i = 1$. In the following lemmas we analyze the activation points in this case and obtain a bound on the number of linear regions. Then, we will extend this result also to the cases where $y_i = -1$ and where $x_a < x_b \leq 0$. For a given activation point x we say that the derivative of the network increases (respectively, decreases) in x if for every sufficiently small $\varepsilon > 0$ the derivative of the network at $x - \varepsilon$ is smaller (respectively, larger) than the derivative at $x + \varepsilon$.

Lemma E.2. *Suppose that $x_b > x_a \geq 0$ and for all $a \leq i \leq b$ we have $y_i = 1$. In the interval $[x_a, x_b]$ the network $\mathcal{N}_{\boldsymbol{\theta}}$ has at most two activation points where the derivative decreases.*

Proof. If the derivative decreases at an activation point $x \geq 0$, then there is at least one neuron $j \in J$ where $w_j \cdot x + b_j = 0$ and the derivative (of the function computed by this neuron) decreases in x. There are two types of such neurons: either (a) $w_j > 0$, $b_j \leq 0$ and $v_j < 0$; or (b) $w_j < 0$, $b_j \geq 0$ and $v_j < 0$.

We now show that in the interval $[x_a, x_b]$ there is at most one activation point that corresponds to a neuron of type (a) and at most one activation point that corresponds to a neuron of type (b). We note that an activation point might correspond to multiple neurons, namely, to a set $Q \subseteq J$ of neurons of size larger than 1. However, we show that if x, x' are activation points in $[x_a, x_b]$ that correspond to sets Q_x and $Q_{x'}$ of neurons (respectively) and both sets $Q_x, Q_{x'}$ contain neurons of type (a) then $x = x'$. Likewise, if both sets $Q_x, Q_{x'}$ contain neurons of type (b) then we also have $x = x'$.

Suppose towards contradiction that $x \in [x_a, x_b]$ is an activation point that corresponds to a neuron j of type (a), and $x' \in [x_a, x_b]$ is an activation point with $x' > x$ that corresponds to a neuron j'

of type (a). Since both neurons j, j' are of type (a), then we have $w_j \cdot z + b_j > 0$ iff $z > x$ and $w_{j'} \cdot z + b_{j'} > 0$ iff $z > x'$. By Eq. (27) we have

$$\frac{1}{v_{j'}} \cdot w_{j'} = \frac{1}{v_{j'}} \left(\sum_{i \in I} \lambda_i y_i v_{j'} \sigma'_{i,j'} x_i \right) = \sum_{i \in I} \lambda_i y_i \sigma'_{i,j'} x_i \leq \sum_{i \in I} \lambda_i y_i \mathbb{1}[x_i \geq x'] x_i \ ,$$

where the last inequality is since $\sigma'_{i,j'} = \mathbb{1}[x_i \geq x']$ if $x_i \neq x'$, and when $x_i = x'$ we have $\sigma'_{i,j'} \leq \mathbb{1}[x_i \geq x']$ (and recall that for $a \leq i \leq b$ we have $y_i = 1$, $x_i \geq 0$ and $\lambda_i \geq 0$). The above RHS equals

$$\begin{aligned}
\sum_{i \in I} \lambda_i y_i \mathbb{1}[x_i > x] x_i - \sum_{i \in I} \lambda_i y_i \mathbb{1}[x < x_i < x'] x_i &\leq \sum_{i \in I} \lambda_i y_i \mathbb{1}[x_i > x] x_i \\
&\leq \sum_{i \in I} \lambda_i y_i \sigma'_{i,j} x_i \\
&= \frac{1}{v_j} \left(\sum_{i \in I} \lambda_i y_i v_j \sigma'_{i,j} x_i \right) \\
&= \frac{1}{v_j} \cdot w_j \ .
\end{aligned}$$

Since $v_j < 0$ we conclude that

$$w_j \leq \frac{v_j}{v_{j'}} \cdot w_{j'} \ . \tag{31}$$

Likewise, by Eq. (28) we have

$$\begin{aligned}
\frac{1}{v_{j'}} \cdot b_{j'} = \frac{1}{v_{j'}} \left(\sum_{i \in I} \lambda_i y_i v_{j'} \sigma'_{i,j'} \right) &= \sum_{i \in I} \lambda_i y_i \sigma'_{i,j'} \leq \sum_{i \in I} \lambda_i y_i \mathbb{1}[x_i \geq x'] \\
&= \sum_{i \in I} \lambda_i y_i \mathbb{1}[x_i > x] - \sum_{i \in I} \lambda_i y_i \mathbb{1}[x < x_i < x'] \\
&\leq \sum_{i \in I} \lambda_i y_i \sigma'_{i,j} = \frac{1}{v_j} \left(\sum_{i \in I} \lambda_i y_i v_j \sigma'_{i,j} \right) = \frac{1}{v_j} \cdot b_j \ .
\end{aligned}$$

Hence, we conclude that

$$b_j \leq \frac{v_j}{v_{j'}} \cdot b_{j'} \ . \tag{32}$$

Since $0 \leq x < x'$, then by using Eq. (31) and (32) we have

$$0 < w_j \cdot x' + b_j \leq \frac{v_j}{v_{j'}} \cdot w_{j'} \cdot x' + \frac{v_j}{v_{j'}} \cdot b_{j'} = \frac{v_j}{v_{j'}} \left(w_{j'} \cdot x' + b_{j'} \right) = 0 \ .$$

Thus, we reached a contradiction.

Next, suppose that $x \in [x_a, x_b]$ is an activation point that corresponds to a neuron j of type (b), and $x' \in [x_a, x_b]$ is an activation point with $x' > x$ that corresponds to a neuron j' of type (b). We will reach a contradiction using similar arguments to the case of type (a) neurons, with some required modifications.

Since both neurons j, j' are of type (b), then we have $w_j \cdot z + b_j > 0$ iff $z < x$ and $w_{j'} \cdot z + b_{j'} > 0$ iff $z < x'$. By Eq. (27) we have

$$\frac{1}{v_{j'}} \cdot w_{j'} = \frac{1}{v_{j'}} \left(\sum_{i \in I} \lambda_i y_i v_{j'} \sigma'_{i,j'} x_i \right) = \sum_{i \in I} \lambda_i y_i \sigma'_{i,j'} x_i \geq \sum_{i \in I} \lambda_i y_i \mathbb{1}[x_i < x'] x_i \ ,$$

where the last inequality is since $\sigma'_{i,j'} = \mathbb{1}[x_i < x']$ if $x_i \neq x'$, and when $x_i = x'$ we have $\sigma'_{i,j'} \geq \mathbb{1}[x_i < x']$. The above RHS equals

$$\begin{aligned}\sum_{i\in I}\lambda_i y_i \mathbb{1}[x_i \leq x]x_i + \sum_{i\in I}\lambda_i y_i \mathbb{1}[x < x_i < x']x_i &\geq \sum_{i\in I}\lambda_i y_i \mathbb{1}[x_i \leq x]x_i \\ &\geq \sum_{i\in I}\lambda_i y_i \sigma'_{i,j} x_i \\ &= \frac{1}{v_j}\left(\sum_{i\in I}\lambda_i y_i v_j \sigma'_{i,j} x_i\right) \\ &= \frac{1}{v_j}\cdot w_j \ .\end{aligned}$$

Since $v_j < 0$ we conclude that

$$w_j \geq \frac{v_j}{v_{j'}}\cdot w_{j'} \ . \tag{33}$$

Likewise, by Eq. (28) we have

$$\begin{aligned}\frac{1}{v_{j'}}\cdot b_{j'} &= \frac{1}{v_{j'}}\left(\sum_{i\in I}\lambda_i y_i v_{j'}\sigma'_{i,j'}\right) = \sum_{i\in I}\lambda_i y_i \sigma'_{i,j'} \geq \sum_{i\in I}\lambda_i y_i \mathbb{1}[x_i < x'] \\ &= \sum_{i\in I}\lambda_i y_i \mathbb{1}[x_i \leq x] + \sum_{i\in I}\lambda_i y_i \mathbb{1}[x < x_i < x'] \\ &\geq \sum_{i\in I}\lambda_i y_i \sigma'_{i,j} = \frac{1}{v_j}\left(\sum_{i\in I}\lambda_i y_i v_j \sigma'_{i,j}\right) = \frac{1}{v_j}\cdot b_j \ .\end{aligned}$$

Hence, we conclude that

$$b_j \geq \frac{v_j}{v_{j'}}\cdot b_{j'} \ . \tag{34}$$

Since $0 \leq x < x'$, and by using Eq. (33) and (34), we have

$$0 > w_j \cdot x' + b_j \geq \frac{v_j}{v_{j'}}\cdot w_{j'}\cdot x' + \frac{v_j}{v_{j'}}\cdot b_{j'} = \frac{v_j}{v_{j'}}\left(w_{j'}\cdot x' + b_{j'}\right) = 0 \ .$$

Thus, we reached a contradiction. □

We denote $\mathcal{I}_{a,b} := \{a, a+1, \ldots, b\} \subseteq I$ and $\mathcal{I}'_{a,b} := \{i \in \mathcal{I}_{a,b} : y_i \mathcal{N}_{\boldsymbol{\theta}}(x_i) = 1\} = \mathcal{I}_{a,b} \cap I'$. Thus, $\mathcal{I}'_{a,b}$ are the indices of the examples in the interval $[x_a, x_b]$ where $\mathcal{N}_{\boldsymbol{\theta}}$ attains margin of exactly 1. We denote $\mathcal{I}'_{a,b} = \{i_1, \ldots, i_m\}$, where $a \leq i_1 < \ldots < i_m \leq b$.

Lemma E.3. *Suppose that $x_b > x_a \geq 0$ and for all $a \leq i \leq b$ we have $y_i = 1$. There are at most 2 indices $\ell \in [m-1]$ such that $\mathcal{N}_{\boldsymbol{\theta}}(x) > 1$ for some $x \in [x_{i_\ell}, x_{i_{\ell+1}}]$.*

Proof. Assume that $\mathcal{N}_{\boldsymbol{\theta}}(x) > 1$ for $x \in [x_{i_\ell}, x_{i_{\ell+1}}]$. Since $\mathcal{N}_{\boldsymbol{\theta}}(x_{i_\ell}) = \mathcal{N}_{\boldsymbol{\theta}}(x_{i_{\ell+1}}) = 1$, then we have $x \in (x_{i_\ell}, x_{i_{\ell+1}})$. Now, since $\mathcal{N}_{\boldsymbol{\theta}}(x_{i_\ell}) = \mathcal{N}_{\boldsymbol{\theta}}(x_{i_{\ell+1}}) = 1$ and $\mathcal{N}_{\boldsymbol{\theta}}(x) > 1$ for some $x \in (x_{i_\ell}, x_{i_{\ell+1}})$, then there must be an activation point in $(x_{i_\ell}, x_{i_{\ell+1}})$ where the derivative decreases. Since by Lemma E.2 there are at most two such activation points in $[x_a, x_b]$ then the lemma follows. □

Lemma E.4. *Suppose that $x_b > x_a \geq 0$ and for all $a \leq i \leq b$ we have $y_i = 1$. There are at most 5 indices $\ell \in [m-1]$ such that $\mathcal{N}_{\boldsymbol{\theta}}(x) < 1$ for some $x \in [x_{i_\ell}, x_{i_{\ell+1}}]$.*

Proof. Assume that $\mathcal{N}_{\boldsymbol{\theta}}(x) < 1$ for $x \in [x_{i_\ell}, x_{i_{\ell+1}}]$. Since $\mathcal{N}_{\boldsymbol{\theta}}(x_{i_\ell}) = \mathcal{N}_{\boldsymbol{\theta}}(x_{i_{\ell+1}}) = 1$, then $x \in (x_{i_\ell}, x_{i_{\ell+1}})$. If $\ell \neq m-1$ then we have $\mathcal{N}_{\boldsymbol{\theta}}(x) < 1$ and $\mathcal{N}_{\boldsymbol{\theta}}(x_{i_{\ell+1}}) = \mathcal{N}_{\boldsymbol{\theta}}(x_{i_{\ell+2}}) = 1$. Hence, the interval $(x, x_{i_{\ell+1}})$ contains a point with positive derivative, and the interval $(x_{i_{\ell+1}}, x_{i_{\ell+2}})$ contains a point with non-positive derivative. Thus, there must be an activation point in $(x, x_{i_{\ell+2}})$ where the derivative decreases. Therefore, there is an activation point with decreasing derivative either in the interval $(x_{i_\ell}, x_{i_{\ell+1}}]$ or in the interval $[x_{i_{\ell+1}}, x_{i_{\ell+2}})$ (and possibly in both). Thus, an interval

$[x_{i_\ell}, x_{i_{\ell+1}}]$ for $\ell \neq m-1$ might contain some x with $\mathcal{N}_{\boldsymbol{\theta}}(x) < 1$ only if there is an activation point with decreasing derivative in $(x_{i_\ell}, x_{i_{\ell+1}}]$ or $[x_{i_{\ell+1}}, x_{i_{\ell+2}})$. Since by Lemma E.2 there are at most two such activation points in $[x_a, x_b]$, then there are at most 4 intervals $[x_{i_\ell}, x_{i_{\ell+1}}]$ with $\ell \neq m-1$ that contain some x with $\mathcal{N}_{\boldsymbol{\theta}}(x) < 1$. The interval $[x_{i_{m-1}}, x_{i_m}]$ might also contain such x. Overall, there are at most 5 indices $\ell \in [m-1]$ such that $\mathcal{N}_{\boldsymbol{\theta}}(x) < 1$ for some $x \in [x_{i_\ell}, x_{i_{\ell+1}}]$. □

Lemma E.5. *Suppose that $x_b > x_a \geq 0$ and for all $a \leq i \leq b$ we have $y_i = 1$. There are at most 30 boundaries between linear regions in $[x_{i_1}, x_{i_m}]$.*

Proof. By Lemmas E.3 and E.4 there are at most 7 indices $\ell \in [m-1]$ such that $\mathcal{N}_{\boldsymbol{\theta}}(x) \neq 1$ for some $x \in [x_{i_\ell}, x_{i_{\ell+1}}]$. We denote the set of these indices by R. Let $x \in (x_{i_1}, x_{i_m})$ be a boundary between two linear regions. Note that if $x \in (x_{i_\ell}, x_{i_{\ell+1}})$ for some $\ell \in [m-1]$ then $\ell \in R$. Also, if $x = x_{i_\ell}$ for some $2 \leq \ell \leq m-1$ then either $\ell \in R$ or $\ell - 1 \in R$. In any case, we have $x \in [x_{i_\ell}, x_{i_{\ell+1}}]$ for some $\ell \in R$. Note that each boundary between linear regions is also an activation point. Therefore, the number of boundaries between linear regions in (x_{i_1}, x_{i_m}) is at most the number of activation points in the intervals $[x_{i_\ell}, x_{i_{\ell+1}}]$ with $\ell \in R$. By Lemma E.1 each interval $[x_{i_\ell}, x_{i_{\ell+1}}]$ contains at most 4 activation points: two points in $(x_{i_\ell}, x_{i_{\ell+1}})$ and two in $\{x_{i_\ell}, x_{i_{\ell+1}}\}$. Overall, there are at most $|R| \cdot 4 \leq 28$ boundaries between linear regions in (x_{i_1}, x_{i_m}). Thus, there are at most 30 boundaries between linear regions in $[x_{i_1}, x_{i_m}]$. □

In the above lemmas we considered the case where $x_b > x_a \geq 0$ and for all $a \leq i \leq b$ we have $y_i = 1$, and proved that $\mathcal{I}'_{a,b}$ is such that there are at most 30 boundaries between linear regions in $[x_{i_1}, x_{i_m}]$. In Subsection E.1 we show analogous results for the case where $x_b > x_a \geq 0$ and for all $a \leq i \leq b$ we have $y_i = -1$. Thus, if $x_b > x_a \geq 0$ and the labels do not switch sign in the interval $[x_a, x_b]$ (i.e., either all labels are 1 or all labels are -1) then there are at most 30 boundaries between linear regions in $[x_{i_1}, x_{i_m}]$. The case where $x_a < x_b \leq 0$ (and the labels do not switch sign in the interval $[x_a, x_b]$) can be handled in a similar manner. Thus, even where the inputs are negative, $\mathcal{I}'_{a,b}$ is such that there are at most 30 boundaries between linear regions in $[x_{i_1}, x_{i_m}]$. The proof for this case is similar and for conciseness we do not repeat it.

We are now ready to finish the proof of the theorem. Consider the set I' of indices where $\mathcal{N}_{\boldsymbol{\theta}}$ attains margin 1 and denote $I' = \{i_1, \ldots, i_q\}$. Note that if I' is an empty set, then by Eq. (27) and (28) all neurons have $w_j = b_j = 0$ and hence the network $\mathcal{N}_{\boldsymbol{\theta}}$ is the zero function. Let $\ell \leq \ell'$ be such that the labels of the examples in the dataset do not change sign in the interval $[x_{i_\ell}, x_{i_{\ell'}}]$, and either $0 \leq x_{i_\ell} \leq x_{i_{\ell'}}$ or $x_{i_\ell} \leq x_{i_{\ell'}} \leq 0$. Thus, the interval $[x_{i_\ell}, x_{i_{\ell'}}]$ contains at most 30 boundaries between linear regions. Also, by Lemma E.1 the interval $(x_{i_{\ell-1}}, x_{i_\ell})$ (or $(-\infty, x_{i_\ell})$ if $\ell = 1$) contains at most two boundaries between linear regions. Likewise, the interval $(x_{i_{\ell'}}, x_{i_{\ell'+1}})$ (or $(x_{i_{\ell'}}, \infty)$ if $\ell' = q$) contains at most two boundaries between linear regions. Recall that the labels in the dataset switch sign at most r times. Overall, we get that the number of boundaries between linear regions in the whole domain $\mathbb{R}$ is at most $30(r+2) + 2(r+3) = 32r + 66$. Indeed, if one of the $r+1$ intervals where $\mathcal{N}_{\boldsymbol{\theta}}$ do not switch sign contains 0 then we split it into two intervals, and thus we obtain $r+2$ intervals. Each of these intervals includes at most 30 boundaries, and outside of these intervals there are at most $2(r+3)$ boundaries. Thus, that are at most $32r + 67$ linear regions.

E.1 Lemmas for the case $y_i = -1$

Lemma E.6. *Suppose that $x_b > x_a \geq 0$ and for all $a \leq i \leq b$ we have $y_i = -1$. In the interval $[x_a, x_b]$ the network $\mathcal{N}_{\boldsymbol{\theta}}$ has at most two activation points where the derivative increases.*

Proof. If the derivative increases at an activation point $x \geq 0$, then there is a least one neuron $j \in J$ where $w_j \cdot x + b_j = 0$ and the derivative (of the function computed by this neuron) increases in x. There are two types of such neurons: either (a) $w_j > 0$, $b_j \leq 0$ and $v_j > 0$; or (b) $w_j < 0$, $b_j \geq 0$ and $v_j > 0$.

We now show that in the interval $[x_a, x_b]$ there is at most one activation point that corresponds to a neuron of type (a) and at most one activation point that corresponds to a neuron of type (b). We note that an activation point might correspond to multiple neurons, namely, to a set $Q \subseteq J$ of neurons of size larger than 1. However, we show that if x, x' are activation points in $[x_a, x_b]$ that correspond to sets Q_x and $Q_{x'}$ of neurons (respectively) and both sets $Q_x, Q_{x'}$ contain neurons of type (a) then $x = x'$. Likewise, if both sets $Q_x, Q_{x'}$ contain neurons of type (b) then we also have $x = x'$.

Suppose towards contradiction that $x \in [x_a, x_b]$ is an activation point that corresponds to a neuron j of type (a), and $x' \in [x_a, x_b]$ is an activation point with $x' > x$ that corresponds to a neuron j' of type (a). Since both neurons j, j' are of type (a), then we have $w_j \cdot z + b_j > 0$ iff $z > x$ and $w_{j'} \cdot z + b_{j'} > 0$ iff $z > x'$. By Eq. (27) we have

$$\frac{1}{v_{j'}} \cdot w_{j'} = \frac{1}{v_{j'}} \left(\sum_{i \in I} \lambda_i y_i v_{j'} \sigma'_{i,j'} x_i \right) = \sum_{i \in I} \lambda_i y_i \sigma'_{i,j'} x_i \geq \sum_{i \in I} \lambda_i y_i \mathbb{1}[x_i \geq x'] x_i \ ,$$

where the last inequality is since $\sigma'_{i,j'} = \mathbb{1}[x_i \geq x']$ if $x_i \neq x'$, and when $x_i = x'$ we have $\sigma'_{i,j'} \leq \mathbb{1}[x_i \geq x']$ (and recall that for $a \leq i \leq b$ we have $y_i = -1$, $x_i \geq 0$ and $\lambda_i \geq 0$). The above RHS equals

$$\begin{aligned}
\sum_{i \in I} \lambda_i y_i \mathbb{1}[x_i > x] x_i - \sum_{i \in I} \lambda_i y_i \mathbb{1}[x < x_i < x'] x_i &\geq \sum_{i \in I} \lambda_i y_i \mathbb{1}[x_i > x] x_i \\
&\geq \sum_{i \in I} \lambda_i y_i \sigma'_{i,j} x_i \\
&= \frac{1}{v_j} \left(\sum_{i \in I} \lambda_i y_i v_j \sigma'_{i,j} x_i \right) \\
&= \frac{1}{v_j} \cdot w_j \ .
\end{aligned}$$

We conclude that

$$w_j \leq \frac{v_j}{v_{j'}} \cdot w_{j'} \ . \tag{35}$$

Likewise, by Eq. (28) we have

$$\begin{aligned}
\frac{1}{v_{j'}} \cdot b_{j'} = \frac{1}{v_{j'}} \left(\sum_{i \in I} \lambda_i y_i v_{j'} \sigma'_{i,j'} \right) &= \sum_{i \in I} \lambda_i y_i \sigma'_{i,j'} \geq \sum_{i \in I} \lambda_i y_i \mathbb{1}[x_i \geq x'] \\
&= \sum_{i \in I} \lambda_i y_i \mathbb{1}[x_i > x] - \sum_{i \in I} \lambda_i y_i \mathbb{1}[x < x_i < x'] \\
&\geq \sum_{i \in I} \lambda_i y_i \sigma'_{i,j} = \frac{1}{v_j} \left(\sum_{i \in I} \lambda_i y_i v_j \sigma'_{i,j} \right) = \frac{1}{v_j} \cdot b_j \ .
\end{aligned}$$

Hence, we conclude that

$$b_j \leq \frac{v_j}{v_{j'}} \cdot b_{j'} \ . \tag{36}$$

Since $0 \leq x < x'$, then by using Eq. (35) and (36) we have

$$0 < w_j \cdot x' + b_j \leq \frac{v_j}{v_{j'}} \cdot w_{j'} \cdot x' + \frac{v_j}{v_{j'}} \cdot b_{j'} = \frac{v_j}{v_{j'}} \left(w_{j'} \cdot x' + b_{j'} \right) = 0 \ .$$

Thus, we reached a contradiction.

Next, suppose that $x \in [x_a, x_b]$ is an activation point that corresponds to a neuron j of type (b), and $x' \in [x_a, x_b]$ is an activation point with $x' > x$ that corresponds to a neuron j' of type (b). We will reach a contradiction using similar arguments to the case of type (a) neurons, with some required modifications.

Since both neurons j, j' are of type (b), then we have $w_j \cdot z + b_j > 0$ iff $z < x$ and $w_{j'} \cdot z + b_{j'} > 0$ iff $z < x'$. By Eq. (27) we have

$$\frac{1}{v_{j'}} \cdot w_{j'} = \frac{1}{v_{j'}} \left(\sum_{i \in I} \lambda_i y_i v_{j'} \sigma'_{i,j'} x_i \right) = \sum_{i \in I} \lambda_i y_i \sigma'_{i,j'} x_i \leq \sum_{i \in I} \lambda_i y_i \mathbb{1}[x_i < x'] x_i \ ,$$

where the last inequality is since $\sigma'_{i,j'} = \mathbb{1}[x_i < x']$ if $x_i \neq x'$, and when $x_i = x'$ we have $\sigma'_{i,j'} \geq \mathbb{1}[x_i < x']$ (and $y_i = -1$). The above RHS equals

$$\begin{aligned}\sum_{i\in I} \lambda_i y_i \mathbb{1}[x_i \leq x] x_i + \sum_{i\in I} \lambda_i y_i \mathbb{1}[x < x_i < x'] x_i &\leq \sum_{i\in I} \lambda_i y_i \mathbb{1}[x_i \leq x] x_i \\ &\leq \sum_{i\in I} \lambda_i y_i \sigma'_{i,j} x_i \\ &= \frac{1}{v_j} \left(\sum_{i\in I} \lambda_i y_i v_j \sigma'_{i,j} x_i \right) \\ &= \frac{1}{v_j} \cdot w_j \ .\end{aligned}$$

We conclude that

$$w_j \geq \frac{v_j}{v_{j'}} \cdot w_{j'} \ . \tag{37}$$

Likewise, by Eq. (28) we have

$$\begin{aligned}\frac{1}{v_{j'}} \cdot b_{j'} &= \frac{1}{v_{j'}} \left(\sum_{i\in I} \lambda_i y_i v_{j'} \sigma'_{i,j'} \right) = \sum_{i\in I} \lambda_i y_i \sigma'_{i,j'} \leq \sum_{i\in I} \lambda_i y_i \mathbb{1}[x_i < x'] \\ &= \sum_{i\in I} \lambda_i y_i \mathbb{1}[x_i \leq x] + \sum_{i\in I} \lambda_i y_i \mathbb{1}[x < x_i < x'] \\ &\leq \sum_{i\in I} \lambda_i y_i \sigma'_{i,j} = \frac{1}{v_j} \left(\sum_{i\in I} \lambda_i y_i v_j \sigma'_{i,j} \right) = \frac{1}{v_j} \cdot b_j \ .\end{aligned}$$

Hence, we conclude that

$$b_j \geq \frac{v_j}{v_{j'}} \cdot b_{j'} \ . \tag{38}$$

Since $0 \leq x < x'$, and by using Eq. (37) and (38), we have

$$0 > w_j \cdot x' + b_j \geq \frac{v_j}{v_{j'}} \cdot w_{j'} \cdot x' + \frac{v_j}{v_{j'}} \cdot b_{j'} = \frac{v_j}{v_{j'}} \left(w_{j'} \cdot x' + b_{j'} \right) = 0 \ .$$

Thus, we reached a contradiction. □

We use the notations $\mathcal{I}_{a,b} := \{a, a+1, \ldots, b\} \subseteq I$ and $\mathcal{I}'_{a,b} := \{i \in \mathcal{I}_{a,b} : y_i \mathcal{N}_{\boldsymbol{\theta}}(x_i) = 1\} = \mathcal{I} \cap I'$. Thus, $\mathcal{I}'$ are the indices of the examples in the interval $[x_a, x_b]$ where $\mathcal{N}_{\boldsymbol{\theta}}$ attains margin of exactly 1. We denote $\mathcal{I}'_{a,b} = \{i_1, \ldots, i_m\}$, where $a \leq i_1 < \ldots < i_m \leq b$.

Lemma E.7. *Suppose that $x_b > x_a \geq 0$ and for all $a \leq i \leq b$ we have $y_i = -1$. There are at most 2 indices $\ell \in [m-1]$ such that $\mathcal{N}_{\boldsymbol{\theta}}(x) < -1$ for some $x \in [x_{i_\ell}, x_{i_{\ell+1}}]$.*

Proof. Assume that $\mathcal{N}_{\boldsymbol{\theta}}(x) < -1$ for $x \in [x_{i_\ell}, x_{i_{\ell+1}}]$. Since $\mathcal{N}_{\boldsymbol{\theta}}(x_{i_\ell}) = \mathcal{N}_{\boldsymbol{\theta}}(x_{i_{\ell+1}}) = -1$, then we have $x \in (x_{i_\ell}, x_{i_{\ell+1}})$. Now, since $\mathcal{N}_{\boldsymbol{\theta}}(x_{i_\ell}) = \mathcal{N}_{\boldsymbol{\theta}}(x_{i_{\ell+1}}) = -1$ and $\mathcal{N}_{\boldsymbol{\theta}}(x) < -1$ for some $x \in (x_{i_\ell}, x_{i_{\ell+1}})$, then there must be an activation point in $(x_{i_\ell}, x_{i_{\ell+1}})$ where the derivative increases. Since by Lemma E.6 there are at most two such activation points in $[x_a, x_b]$ then the lemma follows. □

Lemma E.8. *Suppose that $x_b > x_a \geq 0$ and for all $a \leq i \leq b$ we have $y_i = -1$. There are at most 5 indices $\ell \in [m-1]$ such that $\mathcal{N}_{\boldsymbol{\theta}}(x) > -1$ for some $x \in [x_{i_\ell}, x_{i_{\ell+1}}]$.*

Proof. Assume that $\mathcal{N}_{\boldsymbol{\theta}}(x) > -1$ for $x \in [x_{i_\ell}, x_{i_{\ell+1}}]$. Since $\mathcal{N}_{\boldsymbol{\theta}}(x_{i_\ell}) = \mathcal{N}_{\boldsymbol{\theta}}(x_{i_{\ell+1}}) = -1$, then $x \in (x_{i_\ell}, x_{i_{\ell+1}})$. If $\ell \neq m-1$ then we have $\mathcal{N}_{\boldsymbol{\theta}}(x) > -1$ and $\mathcal{N}_{\boldsymbol{\theta}}(x_{i_{\ell+1}}) = \mathcal{N}_{\boldsymbol{\theta}}(x_{i_{\ell+2}}) = -1$. Hence, the interval $(x, x_{i_{\ell+1}})$ contains a point with negative derivative, and the interval $(x_{i_{\ell+1}}, x_{i_{\ell+2}})$ contains a point with non-negative derivative. Thus, there must be an activation point in $(x, x_{i_{\ell+2}})$ where the derivative increases. Therefore, there is an activation point with increasing derivative either in the interval $(x_{i_\ell}, x_{i_{\ell+1}}]$ or in the interval $[x_{i_{\ell+1}}, x_{i_{\ell+2}})$ (and possibly in both). Thus, an interval $[x_{i_\ell}, x_{i_{\ell+1}}]$ for $\ell \neq m-1$ might contain some x with $\mathcal{N}_{\boldsymbol{\theta}}(x) > -1$ only if there is an activation point

with increasing derivative in $(x_{i_\ell}, x_{i_{\ell+1}}]$ or $[x_{i_{\ell+1}}, x_{i_{\ell+2}})$. Since by Lemma E.6 there are at most two such activation points in $[x_a, x_b]$, then there are at most 4 intervals $[x_{i_\ell}, x_{i_{\ell+1}}]$ with $\ell \neq m-1$ that contain some x with $\mathcal{N}_{\boldsymbol{\theta}}(x) > -1$. The interval $[x_{i_{m-1}}, x_{i_m}]$ might also contain such x. Overall, there are at most 5 indices $\ell \in [m-1]$ such that $\mathcal{N}_{\boldsymbol{\theta}}(x) > -1$ for some $x \in [x_{i_\ell}, x_{i_{\ell+1}}]$. □

Lemma E.9. *Suppose that $x_b > x_a \geq 0$ and for all $a \leq i \leq b$ we have $y_i = -1$. There are at most 30 boundaries between linear regions in $[x_{i_1}, x_{i_m}]$.*

Proof. The proof is similar to the proof of Lemma E.5. The only difference is that here we use Lemmas E.6 and E.7 in order to conclude that there are at most 7 indices $\ell \in [m-1]$ such that $\mathcal{N}_{\boldsymbol{\theta}}(x) \neq -1$ for some $x \in [x_{i_\ell}, x_{i_{\ell+1}}]$, and denote the set of these indices by R. □