When Does Gradient Descent with Logistic Loss Find Interpolating Two-Layer Networks?

Niladri S. Chatterji
University of California, Berkeley
chatterji@berkeley.edu

Philip M. Long
Google
plong@google.com

Peter L. Bartlett
University of California, Berkeley & Google
peter@berkeley.edu

July 2, 2021

Abstract

We study the training of finite-width two-layer smoothed ReLU networks for binary classification using the logistic loss. We show that gradient descent drives the training loss to zero if the initial loss is small enough. When the data satisfies certain cluster and separation conditions and the network is wide enough, we show that one step of gradient descent reduces the loss sufficiently that the first result applies.

1 Introduction

The success of deep learning has led to a lot of recent interest in understanding the properties of "interpolating" neural network models, that achieve (near-)zero training loss [Zha+17a; Bel+19]. One aspect of understanding these models is to theoretically characterize how first-order gradient methods (with appropriate random initialization) seem to reliably find interpolating solutions to non-convex optimization problems.

In this paper, we show that, under two sets of conditions, training fixed-width two-layer networks with gradient descent drives the logistic loss to zero. The networks have smooth "Huberized" ReLUs [Tat+20, see equation (1) and Figure 1] and the output weights are not trained.

The first result only requires the assumption that the initial loss is small, but does not require any assumption about either the width of the network or the number of samples. It guarantees that if the initial loss is small then gradient descent drives the logistic loss to zero.

For our second result we assume that the inputs come from four clusters, two per class, and that the clusters corresponding to the opposite labels are appropriately separated. Under these assumptions, we show that random Gaussian initialization along with a single step of gradient descent is enough to guarantee that the loss reduces sufficiently that the first result applies.

A few proof ideas that facilitate our results are as follows: under our first set of assumptions, when the loss is small, we show that the negative gradient aligns well with the parameter vector. This yields a lower bound on the norm of the gradient in terms of the loss and the norm of the

current weights. This implies that, if the weights are not too large, the loss is reduced rapidly at the beginning of the gradient descent step. Exploiting the Huberization of the ReLUs, we also show that the loss is a smooth function of the weights, so that the loss continues to decrease rapidly throughout the step, as long as the step-size is not too big. Crucially, we show that the loss is decreased significantly compared with the size of the change to the weights. This implies, in particular, that the norm of the weights does not increase by too much, so that progress can continue.

The preceding analysis requires a small loss to "get going". Our second result provides one example when this provably happens. A two-layer network may be viewed as a weighted vote over predictions made by the hidden units. Units only vote on examples that fall in halfspaces where their activation functions are non-zero. When the network is randomly initialized, we can think of each hidden unit as "capturing" roughly half of the examples—each example is turn captured by roughly half of the hidden units. Some capturing events are helpful, and some are harmful. At initialization, these are roughly equal. Using the properties of the Gaussian initialization (including concentration and anti-concentration) we show that each example (x_s, y_s) is captured by many nodes whose first updates contribute to improving its loss. For this to happen, the updates for this example must not be offset by updates for other examples. This happens with sufficient probability at each individual node that the cumulative effect of these "good" nodes overwhelms the effects of potentially confounding nodes, which tend to cancel one another. Consequently, with $2p$ hidden nodes, the loss after one iteration is at most $\exp(-\Omega(p^{1/2-\beta}))$ for $\beta > 0$. By comparison, under similar, but weaker, clustering assumptions, Li and Liang [LL18] used a neural tangent kernel (NTK) analysis to show that the loss is $1/\mathsf{poly}(p)$ after $\mathsf{poly}(p)$ steps. Our proof uses more structure of the problem than the NTK proof, for example, that (loosely speaking) the reduction in the loss is exponential in the number of hidden units improved.

We work with smooth Huberized ReLUs to facilitate theoretical analysis. We analyze networks with Huberized ReLUs instead of the increasingly popular Swish [RZL18], which is also a smooth approximation to the ReLU, to facilitate a simple analysis. We describe some preliminary experiments with artificial data supporting our theoretical analysis, and suggesting that networks with Huberized ReLUs behave similarly to networks with standard ReLUs.

Related results, under weaker assumptions, have been obtained for the quadratic loss [Du+18; Du+19; ALS19; OS20], using the NTK [JGH18; COB19]. The logistic loss is qualitatively different; among other things, driving the logistic loss to zero requires the weights to go to infinity, far from their initial values, so that a Taylor approximation around the initial values cannot be applied. The NTK framework has also been applied to analyze training with the logistic loss. A typical result [LL18; ALS19; Zou+20] is that after $\mathsf{poly}(1/\varepsilon)$ updates, a network of size/width $\mathsf{poly}(1/\varepsilon)$ achieves ε loss. Thus to guarantee loss very close to zero, these analyses require larger and larger networks. The reason for this appears to be that a key part of these analyses is to show that a wider network can achieve a certain fixed loss by traveling a shorter distance in parameter space. Since it seems that, to drive the logistic loss to zero with a fixed-width network, the parameters must travel an unbounded distance, the NTK approach cannot be applied to obtain the results of this paper.

In a recent paper, Lyu and Li [LL20] studied the margin maximization of ReLU networks for the logistic loss. Lyu and Li [LL20] also proved the convergence of gradient descent to zero, but that result requires positive homogeneity and smoothness, which rules out the ReLU and similar nonlinearities like the Huberized ReLU studied here. Their results do apply in

the case that the ReLU is raised to a power strictly greater than two. Lyu and Li used both assumptions of positive homogeneity and smoothness to prove the results in their paper that are most closely related to this paper, so that a substantially different analysis was needed here. (See, for example, the proof of Lemma E.7 of their paper.) As far as we know, the analysis of the alignment between the negative gradient and the weights originated in their paper: in this paper, we establish such alignment under weaker conditions.

Building on this work by Lyu and Li [LL20], Ji and Telgarsky [JT20] studied finite-width deep ReLU neural networks and showed that starting from a small loss, gradient flow coupled with logistic loss leads to convergence of the directions of the parameter vectors. They also demonstrate alignment between the parameter vector directions and the negative gradient. However, they do not prove that the training loss converges to zero.

The remainder of the paper is organized as follows. In Section 2 we introduce notation, definitions, assumptions, and present both of our main results. We provide a proof of Theorem 1 in Section 3 and we prove Theorem 2 in Section 4. Section 5 is devoted to some numerical simulations. Section 6 points to other related work and we conclude with a discussion in Section 7.

2 Preliminaries and Main Results

This section includes notational conventions, a description of the setting, and the statements of the main results.

2.1 Notation

Given a vector v, let $\|v\|$ denote its Euclidean norm. Given a matrix M, let $\|M\|$ denote its Frobenius norm and $\|M\|_{op}$ denote its operator norm. For any $k \in \mathbb{N}$, we denote the set $\{1, \dots, k\}$ by $[k]$. For a number d of inputs, we denote the set of unit-length vectors in $\mathbb{R}^d$ by $\mathbb{S}^{d-1}$. Given an event A, we let $\mathbf{1}_A$ denote the indicator of this event. The symbol $\wedge$ is used to denote the logical "AND" operation. At multiple points in the proof we will use the standard "big Oh notation" [see, e.g., Cor+09] to denote how certain quantities scale with the number of hidden units ($2p$), while viewing all other problem parameters that are not specifically set as a function of p as constants. We will use $C_1, C_2, \dots$ to denote absolute constants whose values are fixed throughout the paper, and $c', c_1, \dots$ to denote "local" constants, which may take different values in different contexts.

2.2 The Setting

We will analyze gradient descent applied to minimize the training loss of a two-layer network.

Let d be the number of inputs, and $2p$ be the number of hidden nodes. We consider the case that the weights connected to the output nodes are fixed: p of them take the value 1, and the other p take the value -1.

We work with Huberized ReLUs that are defined as follows:

$$\phi(z) := \begin{cases} 0 & \text{if } z < 0, \\ \frac{z^2}{2h} & \text{if } z \in [0, h], \\ z - h/2 & \text{otherwise.} \end{cases} \tag{1}$$

See Figure 1. We set the value of the bandwidth parameter $h = 1/p$ throughout the paper.

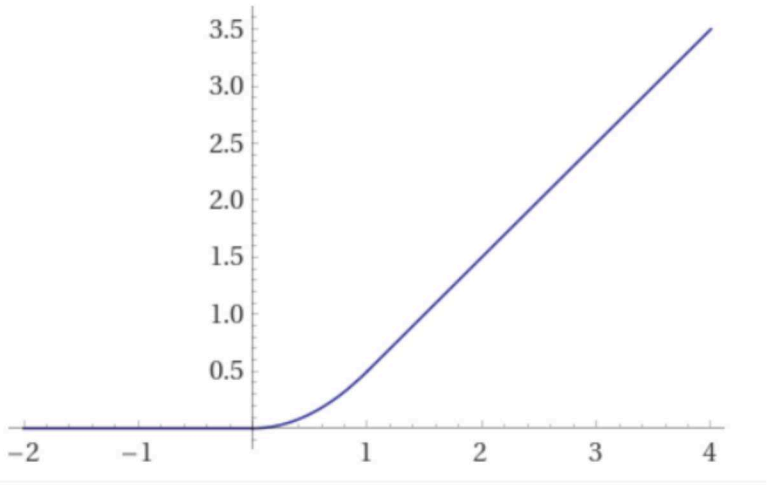

Figure 1: A plot of the Huberized ReLU in the case $h = 1$.

For $i \in \{1, \ldots, 2p\}$, let $v_i \in \mathbb{R}^d$ be vector of weights from the inputs to the ith hidden node, and let $u_1, \ldots, u_{2p} \in \mathbb{R}$ be the weights connecting the hidden nodes to the output node. Set $u_1 = \ldots = u_p = 1$ and $u_{p+1} = \ldots = u_{2p} = -1$. Let b_i be the bias for the ith hidden node. Let $\theta = (v_1, b_1, \ldots, v_{2p}, b_{2p})$ consist of all of the trainable parameters in the network. Let f_θ denote the function computed by the network, which maps x to

$$f_\theta(x) = \sum_{i=1}^{2p} u_i \phi(v_i \cdot x + b_i).$$

Consider a training set $(x_1, y_1), \ldots, (x_n, y_n) \in \mathbb{S}^{d-1} \times \{-1, 1\}$. Define the training loss L by

$$L(\theta) := \frac{1}{n} \sum_{s=1}^{n} \log(1 + \exp(-y_s f_\theta(x_s))),$$

and refer to loss on example s by

$$L(\theta; x_s, y_s) := \log(1 + \exp(-y_s f_\theta(x_s))).$$

The gradient of the loss evaluated at θ is

$$\nabla_\theta L(\theta) = \frac{1}{n} \sum_{s=1}^{n} \frac{-y_s \nabla_\theta f_\theta(x_s)}{1 + \exp(y_s f_\theta(x_s))}.$$

Note that, since ϕ'' is not defined at 0 and h, the Hessian of L is not defined everywhere. We use the following weak derivative of ϕ' to define a weak Hessian ($\nabla^2_\theta L(\theta)$):

$$\gamma(z) := \begin{cases} \frac{1}{h} & \text{if } z \in [0, h], \\ 0 & \text{otherwise.} \end{cases} \tag{2}$$

2.3 A General Bound

We first analyze the iterates $\theta^{(1)}, \theta^{(2)}, \ldots$ defined by

$$\theta^{(t+1)} := \theta^{(t)} - \alpha_t \nabla_\theta L|_{\theta=\theta^{(t)}},$$

in terms of properties of $\theta^{(1)}$.

Theorem 1. *There is an absolute constant $C_1 > 0$ such that, for all $n \geq 2$, for all $p \geq 1$, for any initial parameters $\theta^{(1)}$ and dataset $(x_1, y_1), \ldots, (x_n, y_n) \in \mathbb{S}^{d-1} \times \{-1, 1\}$, for any positive $Q_1 \leq \widetilde{Q}_1$ and positive $Q_2 \leq \widetilde{Q}_2(Q_1)$ (where $\widetilde{Q}_1$ and $\widetilde{Q}_2$ are defined in eqs. (6)-(7)) the following holds for all $t \geq 1$. If $h = 1/p$ and each step-size $\alpha_t = Q_1 \log^2(1/L(\theta^{(t)}))$, and if $L(\theta^{(1)}) \leq \frac{1}{n^{1+C_1}}$ then, for all $t \geq 1$,*

$$L(\theta^{(t)}) \leq \frac{L(\theta^{(1)})}{Q_2 \cdot (t-1) + 1}.$$

The proof of this theorem is presented in Section 3 below.

We reiterate that this theorem makes no assumption about the number of hidden nodes (p) and makes a mild assumption on the number of samples required ($n \geq 2$). The only other constraint is that the initial loss needs to be smaller than $1/n^{1+C_1}$, for some universal constant $C_1 > 0$. Our choice of a time-varying step-size α_t that grows with $\log^2(1/L(\theta^{(t)}))$ leads to an upper bound on the loss that scales with $1/t$.

2.4 Clustered Data with Random Initialization

We next consider gradient descent after random initialization by showing that, after one iteration, $\theta^{(1)}$ has the favorable properties needed to apply Theorem 1. We assume that all trainable parameters are initialized by being chosen independently at random from $\mathcal{N}(0, \sigma^2)$. Let $\theta^{(0)}$ be the initial value of the parameters and α_0 be the original step-size (which will be chosen as a function of p).

This analysis depends on cluster and separation conditions. We shall use k and ℓ to index over the clusters (ranging from 1 to 4), and s and q to index over individual samples (ranging from 1 to n). We assume that the training data can be divided into four clusters $\mathcal{K}_1, \ldots, \mathcal{K}_4$. All examples (x_s, y_s) in clusters $\mathcal{K}_1$ and $\mathcal{K}_2$ have $y_s = 1$ and all examples (x_s, y_s) in clusters $\mathcal{K}_3$ and $\mathcal{K}_4$ have $y_s = -1$. For some cluster index k, let y_k be the label shared by all examples in cluster $\mathcal{K}_k$. With some abuse of notation we will often use $s \in \mathcal{K}_k$ to denote that the example (x_s, y_s) belongs to the cluster $\mathcal{K}_k$.

We make the following assumptions about the clustered training data.

- For $\varepsilon > 0$, for each cluster $\mathcal{K}_k$, we assume $n_k := |\mathcal{K}_k|$ satisfies $(1/4 - \varepsilon)\, n \leq n_k \leq (1/4 + \varepsilon)\, n$.
- Assume that $\|x_s\| = 1$ for all $s \in [n]$.
- For a radius $r > 0$, each cluster $\mathcal{K}_k$ has a center μ_k with $\|\mu_k\| = 1$, such that for all $s \in \mathcal{K}_k$, $\|x_s - \mu_k\| \leq r$.
- For a separation parameter $\Delta > 0$, we assume that for all k, ℓ with $y_k \neq y_\ell$, $\mu_k \cdot \mu_\ell \leq \Delta$.

Under these assumptions we demonstrate that with high probability random initialization followed by one step of gradient descent leads to a network whose training loss is at most $\exp\left(-\Omega(p^{1/2-\beta})\right)$ for $\beta > 0$. Theorem 1 then implies that gradient descent in the subsequent steps leads to a solution with training loss approaching zero.

Theorem 2. *For any $\beta \in (0, 1/2)$, there are absolute constants $C_2, \ldots, C_6 > 0$ such that, under the assumptions described above, the following holds for all $\delta < 1/2$ and $n \geq 4$. If $\Delta < C_2$, $r < C_2$, $\varepsilon < C_2$, $\alpha_0 = \frac{1}{p^{1/2+\beta}}$, $\sigma = \frac{1}{p^{1/2+\beta/2}}$, $h = \frac{1}{p}$, and $p \geq \log^{C_3}(nd/\delta)$ both of the following hold:*

(a) With probability $1-\delta$,

$$L(\theta^{(1)}) \leq \exp(-C_4 p^{(1/2-\beta)});$$

(b) if, for all $t \geq 1$, $\alpha_t = \frac{C_5 \log^2(1/L(\theta^{(t)}))}{p}$, *then with probability* $1-\delta$, *for all* $t > 1$,

$$L(\theta^{(t)}) \leq \frac{C_6 d}{p^{1-3\beta} t}.$$

This theorem is proved in Section 4. It shows that if the data satisfies the cluster and separation conditions then the loss after a single step of gradient descent decreases by an amount that is exponential in $p^{1/2-\beta}$ with high probability. This result only requires the width p to be poly-logarithmic in the number of samples, input dimension and $1/\delta$.

3 Proof of Theorem 1

In this section, we prove Theorem 1. Our proof is by induction. As mentioned in the introduction, the key lemma is a lower bound on the norm of the gradient. Our lower bound (Lemma 10) is in terms of the loss, and also the norm of the weights. Roughly speaking, for it to provide leverage, we need that the loss is small relative to the size of the weights, or, in other words, that the model doesn't excessively "waste weight". The bound of Theorem 1 accounts for the amount of such wasted weight at initialization, so we do not need a wasted-weight assumption. On the other hand, we need a condition on the wasted weight in our inductive hypothesis—we need to prove that training does not increase the amount of wasted weight too much.

Lemma 10 also requires an upper bound on the step size – another part of the inductive hypothesis ensures that this requirement is met throughout training.

Before the proof, we lay some groundwork. First, to simplify expressions, we reduce to the case that the biases are zero. Then we establish some lemmas that will be used in the inductive step, about the progress in an iteration, smoothness, etc. Finally, we applied these tools in the inductive proof.

3.1 Reduction to the Zero-Bias Case

We first note that, applying a standard reduction, without loss of generality, we may assume

- $b_1, \ldots, b_{2p}$ are fixed to 0, and not trained, and
- for all s, q, $x_s \cdot x_q \geq 0$.

The idea is to adopt the view that the inputs have an additional component that acts as a placeholder for the bias term, which allows us to view the bias term as another component of v_i. The details are in Appendix A. We will make the above assumptions from now on. Since the bias terms are fixed at zero, for a matrix V whose rows are the weights of the hidden units, we will refer to the resulting loss as $L(V)$, f_θ as f_V, and so on. Let $V^{(t)}$ be the tth iterate.

3.2 Additional Definitions

Definition 3. *For all iterates* t, *define* $L_{ts} := L(V^{(t)}; x_s, y_s)$ *and let* $L_t := \frac{1}{n}\sum_{s=1}^{n} L_{ts}$. *Additionally for all* t, *define* $\nabla L_t := \nabla_V L|_{V=V^{(t)}}$. *We will also use* $\nabla_{V^{(t)}} L$ *to refer to the gradient* $\nabla_V L|_{V=V^{(t)}}$.

Definition 4. *For any weight matrix V, define*

$$g_s(V) := \frac{1}{1 + \exp\left(y_s f_V(x_s)\right)}.$$

We often will use g_s as shorthand for $g_s(V)$ when V can be determined from context. Further, for all $t \in \{0, 1, \ldots\}$, define $g_{ts} := g_s(V^{(t)})$.

Informally, $g_s(V)$ is the size of the contribution of example s to the gradient.

3.3 Technical Tools

In this subsection we assemble several technical tools required to prove Theorem 1. The proofs that are omitted in this subsection are presented in Appendix B.

We start with the following lemma, which is a slight variant of a standard inequality, and provides a bound on the loss after a step of gradient descent when the loss function is locally smooth. It is proved in Appendix B.1.

Lemma 5. *For $\alpha_t > 0$, let $V^{(t+1)} = V^{(t)} - \alpha_t \nabla L_t$. If, for all convex combinations W of $V^{(t)}$ and $V^{(t+1)}$, we have $\|\nabla^2_W L\|_{op} \le M$, then if $\alpha_t \le \frac{1}{3M}$, we have*

$$L_{t+1} \le L_t - \frac{5\alpha_t \|\nabla L_t\|^2}{6}.$$

To apply Lemma 5 we need to show that the loss function L is smooth near L_t; the following lemma is a start. It is proved in Appendix B.2.

Lemma 6. *If $h = 1/p$, for any weight matrix $V \in \mathbb{R}^{2p \times (d+1)}$, $\|\nabla^2_V L\|_{op} \le 5pL(V)$.*

Next, we show that L changes slowly in general, and especially slowly when it is small. The proof is in Appendix B.3.

Lemma 7. *For any weight matrix $V \in \mathbb{R}^{2p \times (d+1)}$, $\|\nabla_V L\| \le \sqrt{2p} \min\{L(V), 1\}$.*

The following lemma applies Lemma 5 (along with Lemma 6) to show that if the step-size at step t is small enough then the loss decreases by an amount that is proportional to the squared norm of the gradient. Its proof is in Appendix B.4.

Lemma 8. *If $\alpha_t L_t \le \frac{1}{30p}$, then $L_{t+1} \le L_t - \frac{5\alpha_t \|\nabla L_t\|^2}{6}$.*

We need the following technical lemma which is proved in Appendix B.5.

Lemma 9. *If $\psi : (0, M] \to \mathbb{R}$ is a continuous, concave function such that $\lim_{\eta \to 0^+} \psi(\eta)$ exists. Then the infimum of $\sum_{i=1}^n \psi(z_i)$ subject to $z_1, \ldots, z_n > 0$ and $\sum_{i=1}^n z_i = M$ is $\psi(M) + (n-1) \lim_{\eta \to 0^+} \psi(\eta)$.*

The next lemma establishes a lower bound on the norm of the gradient of the loss in the later iterations.

Lemma 10. *For all large enough C_1, for any $t \ge 1$, if $L_t \le 1/n^{1+C_1}$, then*

$$\|\nabla L_t\| \ge \frac{5L_t \log(1/L_t)}{6\|V^{(t)}\|}. \tag{3}$$

Proof Since

$$\|\nabla L_t\| = \sup_{a:\|a\|=1} (\nabla L_t \cdot a) \geq (\nabla L_t) \cdot \left(\frac{-V^{(t)}}{\|V^{(t)}\|} \right),$$

we seek a lower bound on $-\nabla L_t \cdot \frac{V^{(t)}}{\|V^{(t)}\|}$. We have

$$\begin{aligned} -\nabla L_t \cdot \frac{V^{(t)}}{\|V^{(t)}\|} &= \frac{1}{\|V^{(t)}\|} \sum_{i\in[2p]} \frac{u_i}{n} \sum_{s=1}^{n} g_{ts} y_s \phi'(v_i^{(t)} \cdot x_s) v_i^{(t)} \cdot x_s \\ &= \frac{1}{n\|V^{(t)}\|} \sum_{s=1}^{n} g_{ts} y_s \left[\sum_{i\in[2p]} u_i \phi'(v_i^{(t)} \cdot x_s)(v_i^{(t)} \cdot x_s) \right]. \end{aligned}$$

Note that, by definition of the Huberized ReLU for any $z \in \mathbb{R}$, $\phi(z) \leq \phi'(z)z \leq \phi(z) + h/2$, and therefore,

$$\begin{aligned} -\nabla L_t \cdot \frac{V^{(t)}}{\|V^{(t)}\|} &= \frac{1}{n\|V^{(t)}\|} \sum_{s=1}^{n} g_{ts} \left[y_s \sum_{i\in[2p]} u_i \phi(v_i^{(t)} \cdot x_s) \right] \\ &\quad + \frac{1}{n\|V^{(t)}\|} \sum_{s=1}^{n} g_{ts} \left[\sum_{i\in[2p]} y_s u_i \left(\phi'(v_i^{(t)} \cdot x_s)(v_i^{(t)} \cdot x_s) - \phi(v_i^{(t)} \cdot x_s) \right) \right] \\ &\geq \frac{1}{n\|V^{(t)}\|} \sum_{s=1}^{n} g_{ts} \left[y_s \sum_{i\in[2p]} u_i \phi(v_i^{(t)} \cdot x_s) \right] - \frac{1}{n\|V^{(t)}\|} \sum_{s=1}^{n} g_{ts} \left(\frac{h}{2} \sum_{i\in[2p]} |y_s u_i| \right) \\ &\overset{(i)}{=} \frac{1}{n\|V^{(t)}\|} \sum_{s=1}^{n} g_{ts} \left[y_s \sum_{i\in[2p]} u_i \phi(v_i^{(t)} \cdot x_s) \right] - \frac{hp}{n\|V^{(t)}\|} \sum_{s=1}^{n} g_{ts} \\ &\overset{(ii)}{\geq} \frac{1}{n\|V^{(t)}\|} \sum_{s=1}^{n} g_{ts} \left[y_s \sum_{i\in[2p]} u_i \phi(v_i^{(t)} \cdot x_s) \right] - \frac{L_t}{\|V^{(t)}\|} \\ &= \frac{1}{n\|V^{(t)}\|} \sum_{s=1}^{n} \frac{y_s f_{V^{(t)}}(x_s)}{1 + \exp\left(y_s f_{V^{(t)}}(x_s)\right)} - \frac{L_t}{\|V^{(t)}\|}, \end{aligned} \tag{4}$$

where (i) follows as $|y_s u_i| = 1$ for all $i \in [2p]$ and the inequality in (ii) follows since $g_{ts} \leq L_{ts}$ for all samples by Lemma 20 and because $h = 1/p$.

For every sample s, $L_{ts} = \log\left(1 + \exp\left(-y_s f_{V^{(t)}}(x_s)\right)\right)$ which implies

$$y_s f_{V^{(t)}}(x_s) = \log\left(\frac{1}{\exp(L_{ts}) - 1} \right) \quad \text{and} \quad \frac{1}{1 + \exp\left(y_s f_{V^{(t)}}(x_s)\right)} = 1 - \exp(-L_{ts}).$$

Plugging this into inequality (4) we derive,

$$-\nabla L_t \cdot \frac{V^{(t)}}{\|V^{(t)}\|} \geq \frac{1}{n\|V^{(t)}\|} \sum_{s=1}^{n} (1 - \exp(-L_{ts})) \log\left(\frac{1}{\exp(L_{ts}) - 1} \right) - \frac{L_t}{\|V^{(t)}\|}.$$

Observe that the function $(1-\exp(-z))\log\left(\frac{1}{\exp(z)-1}\right)$ is continuous and concave with $\lim_{z\to 0^+}(1-\exp(-z))\log\left(\frac{1}{\exp(z)-1}\right) = 0$. Also recall that $\sum_s L_{ts} = L_t n$. Therefore by Lemma 9,

$$-\nabla L_t \cdot \frac{V^{(t)}}{\|V^{(t)}\|} \geq \frac{1}{\|V^{(t)}\|}\left[\frac{1-\exp(-L_t n)}{n}\log\left(\frac{1}{\exp(L_t n)-1}\right) - L_t\right]. \tag{5}$$

We know that for any $z \in [0,1]$

$$\exp(z) \leq 1+2z \quad \text{and} \quad \exp(-z) \leq 1 - z + z^2.$$

Since $L_t \leq \frac{1}{n^{1+C_1}}$ and $n \geq 2$ for large enough C_1 these bounds on the exponential function combined with inequality (5) yields

$$\begin{aligned}
-\nabla L_t \cdot \frac{V^{(t)}}{\|V^{(t)}\|} &\geq \frac{1}{\|V^{(t)}\|}\left[(L_t - nL_t^2)\log\left(\frac{1}{2nL_t}\right) - L_t\right] \\
&= \frac{1}{\|V^{(t)}\|}\left[L_t\log\left(\frac{1}{L_t}\right) + nL_t^2\log(2n) - L_t(1+\log(2n)) - nL_t^2\log\left(\frac{1}{L_t}\right)\right] \\
&\geq \frac{1}{\|V^{(t)}\|}\left[L_t\log\left(\frac{1}{L_t}\right) - L_t(1+\log(2n)) - nL_t^2\log\left(\frac{1}{L_t}\right)\right] \\
&= \frac{L_t\log(1/L_t)}{\|V^{(t)}\|}\left[1 - \frac{1+\log(2)+\log(n)}{\log(1/L_t)} - nL_t\right].
\end{aligned}$$

Recalling again that $L_t \leq \frac{1}{n^{1+C_1}}$ and $n \geq 2$,

$$\begin{aligned}
-\nabla L_t \cdot \frac{V^{(t)}}{\|V^{(t)}\|} &\geq \frac{L_t\log(1/L_t)}{\|V^{(t)}\|}\left[1 - \frac{1+\log(2)+\log(n)}{(1+C_1)\log(n)} - \frac{1}{n^{C_1}}\right] \\
&\geq \frac{L_t\log(1/L_t)}{\|V^{(t)}\|}\left[1 - \frac{1+2\log(2)}{(1+C_1)\log(2)} - \frac{1}{2^{C_1}}\right] \\
&\geq \frac{5L_t\log(1/L_t)}{6\|V^{(t)}\|},
\end{aligned}$$

where the final inequality holds for a large enough value of C_1. ■

We are now ready to prove our theorem.

3.4 The Proof

As mentioned above, the proof of Theorem 1 is by induction. Given the initial weight matrix $V^{(1)}$ and p, the values $\widetilde{Q}_1$ and $\widetilde{Q}_2(Q_1)$ can be chosen as stated below:

$$\widetilde{Q}_1 = \min\left\{\frac{1}{30pL_1\log^2(1/L_1)}, \frac{108\|V^{(1)}\|^2}{125L_1\log^4(1/L_1)}, \frac{e^2}{120p}\right\} \quad \text{and} \tag{6}$$

$$\widetilde{Q}_2(Q_1) = \frac{125Q_1L_1\log^4(1/L_1)}{216\|V^{(1)}\|^2}. \tag{7}$$

The proof goes through for any positive $Q_1 \leq \widetilde{Q}_1$ and any positive $Q_2 \leq \widetilde{Q}_2(Q_1)$. Recall that the sequence of step-sizes is given by $\alpha_t = Q_1\log^2(1/L_t)$. We will use the following multi-part inductive hypothesis:

(I1) $L_t \leq \frac{L_1}{Q_2 \cdot (t-1)+1}$;

(I2) $\alpha_t L_t \leq \frac{1}{30p}$;

(I3) $\frac{\log^2(1/L_t)}{\|V^{(t)}\|} \geq \frac{\log^2(1/L_1)}{\|V^{(1)}\|}$.

The base case is trivially true for the first and the third part of the inductive hypothesis. It is true for the second part since $Q_1 \leq \frac{1}{30pL_1 \log^2(1/L_1)}$.

Now let us assume that the inductive hypothesis holds for a step $t \geq 1$ and prove that it holds for the next step $t+1$. We start with Part I1.

Lemma 11. *If the inductive hypothesis holds at step t then,*

$$L_{t+1} \leq \frac{L_1}{Q_2 t + 1}.$$

Proof Since $\alpha_t L_t < 1/(30p)$ by applying Lemma 8

$$L_{t+1} \leq L_t - \frac{5\alpha_t}{6}\|\nabla L_t\|^2.$$

By the lower bound on the norm of the gradient established in Lemma 10 since $L_t \leq L_1 \leq 1/n^{1+C_1}$ we have

$$\begin{aligned} L_{t+1} \leq L_t - \frac{125\alpha_t L_t^2 \log^2(1/L_t)}{216\|V^{(t)}\|^2} &= L_t - \frac{125 Q_1 L_t^2 \log^4(1/L_t)}{216\|V^{(t)}\|^2} \\ &\leq L_t \left(1 - \frac{125 Q_1 L_t \log^4(1/L_1)}{216\|V^{(1)}\|^2}\right), \end{aligned} \tag{8}$$

where the final inequality makes use of the third part of the inductive hypothesis. For any $z \geq 0$, the quadratic function

$$z - z^2 \frac{125 Q_1 \log^4(1/L_1)}{216\|V^{(1)}\|^2}$$

is a monotonically increasing function in the interval

$$\left[0, \frac{108\|V^{(1)}\|^2}{125 Q_1 \log^4(1/L_1)}\right].$$

Thus, because $L_t \leq \frac{L_1}{Q_2(t-1)+1}$, if $\frac{L_1}{Q_2(t-1)+1} \leq \frac{108\|V^{(1)}\|^2}{125 Q_1 \log^4(1/L_1)}$, the RHS of (8) is bounded above by its value when $L_t = \frac{L_1}{Q_2(t-1)+1}$. But this is easy to check: by our choice of the constant Q_1 we have,

$$\begin{aligned} Q_1 \leq \widetilde{Q}_1 &\leq \frac{108\|V^{(1)}\|^2}{125 L_1 \log^4(1/L_1)} \\ \Rightarrow L_1 &\leq \frac{108\|V^{(1)}\|^2}{125 Q_1 \log^4(1/L_1)} \\ \Rightarrow \frac{L_1}{Q_2(t-1)+1} &\leq \frac{108\|V^{(1)}\|^2}{125 Q_1 \log^4(1/L_1)}. \end{aligned}$$

Bounding the RHS of inequality (8) by using the worst case that $L_t = \frac{L_1}{Q_2(t-1)+1}$, we get

$$\begin{aligned}
L_{t+1} &\leq \frac{L_1}{Q_2(t-1)+1}\left(1 - \frac{L_1}{Q_2(t-1)+1}\frac{125Q_1\log^4(1/L_1)}{216\|V^{(1)}\|^2}\right) \\
&= \frac{L_1}{Q_2t+1}\left(\frac{Q_2t+1}{Q_2(t-1)+1}\right)\left(1 - \frac{Q_2}{Q_2(t-1)+1}\frac{125Q_1L_1\log^4(1/L_1)}{216Q_2\|V^{(1)}\|^2}\right) \\
&= \frac{L_1}{Q_2t+1}\left(1 + \frac{Q_2}{Q_2(t-1)+1}\right)\left(1 - \frac{Q_2}{Q_2(t-1)+1}\frac{125Q_1L_1\log^4(1/L_1)}{216Q_2\|V^{(1)}\|^2}\right) \\
&\leq \frac{L_1}{Q_2t+1}\left(1 - \left(\frac{Q_2}{Q_2(t-1)+1}\right)^2\right) \\
&\qquad\qquad \left(\text{since } Q_2 \leq \tfrac{125Q_1L_1\log^4(1/L_1)}{216\|V^{(1)}\|^2}\right) \\
&\leq \frac{L_1}{Q_2t+1}.
\end{aligned}$$

This establishes the desired upper bound on the loss at step $t+1$. ■

In the next lemma we ensure that the second part of the inductive hypothesis holds.

Lemma 12. *Under the setting of Theorem 1 if the induction hypothesis holds at step t then,*

$$\alpha_{t+1}L_{t+1} \leq \frac{1}{30p}.$$

Proof We know by the previous lemma that if the induction hypothesis holds at step t, then $L_{t+1} \leq L_t \leq 1$. The function $z\log^2(1/z)$ is no more than $4/e^2$ for $z \in (0,1]$. Since $Q_1 \leq e^2/(120p)$ we have

$$\alpha_{t+1}L_{t+1} = Q_1L_{t+1}\log^2(1/L_{t+1}) \leq \frac{1}{30p}.$$

■

Finally, we shall establish that the third part of the inductive hypothesis holds.

Lemma 13. *Under the setting of Theorem 1 if the induction hypothesis holds at step t then,*

$$\frac{\log^2\left(\frac{1}{L_{t+1}}\right)}{\|V^{(t+1)}\|} \geq \frac{\log^2\left(\frac{1}{L_1}\right)}{\|V^{(1)}\|}.$$

Proof We know from Lemma 8 that $L_{t+1} \leq L_t\left(1-5\alpha_t\|\nabla L_t\|^2/(6L_t)\right)$, and by the triangle inequality $\|V^{(t+1)}\| \leq \|V^{(t)}\| + \alpha_t\|\nabla L_t\|$, hence

$$\begin{aligned}
\frac{\log^2\left(\frac{1}{L_{t+1}}\right)}{\|V^{(t+1)}\|} &\geq \frac{\log^2\left(\frac{1}{L_t\left(1-\frac{5\alpha_t}{6L_t}\|\nabla L_t\|^2\right)}\right)}{\|V^{(t)}\| + \alpha_t\|\nabla L_t\|} \\
&= \frac{\left(\log\left(\frac{1}{L_t}\right) - \log\left(1-\frac{5\alpha_t}{6L_t}\|\nabla L_t\|^2\right)\right)^2}{\|V^{(t)}\| + \alpha_t\|\nabla L_t\|} \\
&= \frac{\log^2\left(\frac{1}{L_t}\right) - 2\log\left(\frac{1}{L_t}\right)\log\left(1-\frac{5\alpha_t}{6L_t}\|\nabla L_t\|^2\right) + \log^2\left(1-\frac{5\alpha_t}{6L_t}\|\nabla L_t\|^2\right)}{\|V^{(t)}\| + \alpha_t\|\nabla L_t\|} \\
&\overset{(i)}{\geq} \frac{\log^2\left(\frac{1}{L_t}\right)\left(1 - \frac{2\log\left(1-\frac{5\alpha_t}{6L_t}\|\nabla L_t\|^2\right)}{\log\left(\frac{1}{L_t}\right)}\right)}{\|V^{(t)}\|\left(1+\frac{\alpha_t\|\nabla L_t\|}{\|V^{(t)}\|}\right)} \\
&\overset{(ii)}{\geq} \frac{\log^2\left(\frac{1}{L_t}\right)}{\|V^{(t)}\|}\left\{\frac{1+\frac{5\alpha_t\|\nabla L_t\|^2}{3L_t\log\left(\frac{1}{L_t}\right)}}{1+\frac{\alpha_t\|\nabla L_t\|}{\|V^{(t)}\|}}\right\}
\end{aligned} \tag{9}$$

where in (i) the lower bound follows as we are dropping a positive lower-order term, and (ii) follows since $\log(1-z) \leq -z$ for all $z < 1$ and

$$\begin{aligned}
\frac{5\alpha_t}{6L_t}\|\nabla L_t\|^2 &\leq \frac{10p\alpha_t L_t}{6} \qquad \text{(by Lemma 7)} \\
&< 1
\end{aligned}$$

by the inductive hypothesis.

We want the term in curly brackets in inequality (9) to be at least 1, that is,

$$\begin{aligned}
1 + \frac{5\alpha_t\|\nabla L_t\|^2}{3L_t\log\left(\frac{1}{L_t}\right)} &\geq 1 + \frac{\alpha_t\|\nabla L_t\|}{\|V^{(t)}\|} \\
\Leftarrow \|\nabla L_t\| &\geq \frac{3L_t\log\left(\frac{1}{L_t}\right)}{5\|V^{(t)}\|},
\end{aligned}$$

which follows from Lemma 10 which ensures that $\|\nabla L_t\| \geq 5L_t\log(1/L_t)/(6\|V^{(t)}\|)$ (since $5/6 \geq 3/5$). Thus we can infer that

$$\frac{\log^2\left(\frac{1}{L_{t+1}}\right)}{\|V^{(t+1)}\|} \geq \frac{\log^2\left(\frac{1}{L_t}\right)}{\|V^{(t)}\|} \geq \frac{\log^2\left(\frac{1}{L_1}\right)}{\|V^{(1)}\|}.$$

This proves that the ratio is lower bounded at step $t+1$ by its initial value and establishes our claim. ■

Combining the results of Lemmas 11, 12 and 13 completes the proof of theorem.

4 Proof of Theorem 2

The proof of Theorem 2 has two parts. First, we analyze the first step and show that the loss decreases by a factor that is exponentially large in $p^{1/2-\beta}$. After this, we complete the proof by invoking Theorem 1.

4.1 The Effect of the Reduction on the Clusters

We reduce to the case that the bias terms are fixed at zero in the context of Theorem 2. In this case, we can assume the following without loss of generality:

- $b_1, \ldots, b_{2p}$ are fixed to 0, and not trained,
- for all s, $\|x_s\| = 1$,
- for all $s, q \in [n]$, $x_s \cdot x_q \geq 0$,
- for all $k, \ell \in [4]$ for $y_k \neq y_\ell$, $\mu_k \cdot \mu_\ell \leq (1+\Delta)/2$, and
- for all s, k, $\ s \in \mathcal{K}_k$, $\|x_s - \mu_k\| \leq r/\sqrt{2}$.

The details are in Appendix C.

4.2 Analysis of the Initial Step

Our analysis of the first step will make reference to the set of hidden units that "capture" an example by a sufficient margin, further dividing them into helpful and harmful units.

Definition 14. *Define*

$$I_{+s} := \left\{ i \in [2p] : (u_i = y_s) \wedge \left(v_i^{(0)} \cdot x_s \geq h + 4\alpha_0 \right) \right\} \text{ and}$$
$$I_{-s} := \left\{ i \in [2p] : (u_i = -y_s) \wedge \left(v_i^{(0)} \cdot x_s \geq h + 4\alpha_0 \right) \right\}.$$

Next, we prove that the random initialization satisfies a number of properties with high probability. (We will later show that they are sufficient for convergence.) The proof is in Appendix D. (Recall that $C_2, \ldots, C_4$ are specified in the statement of Theorem 2. We also remind the reader that sufficiently large C_3 means that p is sufficiently large.)

Lemma 15. *There exists a real-valued function χ such that, for all $(x_1, y_1), \ldots, (x_n, y_n) \in \mathbb{S}^{d-1} \times \{-1, 1\}$, for all small enough C_2 and all large enough C_3, with probability $1 - \delta$ over the draw of $V^{(0)}$, all of the following hold.*

1. *For all $s \in [n]$,*

$$\sum_{i \in I_{+s}} v_i^{(0)} \cdot x_s \geq p\chi(h, \alpha_0, \sigma) - 1, \text{ and}$$
$$\sum_{i \in I_{-s}} v_i^{(0)} \cdot x_s \leq p\chi(h, \alpha_0, \sigma) + 1.$$

2. *For all samples* $s \in [n]$,

$$(1/2 - o(1))p \le |I_{+s}| \le (1/2 + o(1))p, \text{ and}$$
$$(1/2 - o(1))p \le |I_{-s}| \le (1/2 + o(1))p.$$

3. *For all samples* $s \in [n]$

$$\frac{1}{2} - o(1) \le g_{0s} \le \frac{1}{2} + o(1).$$

4. *For all clusters* $k \in [4]$,

$$|\{i \in [2p] : \forall\ s \in \mathcal{K}_k,\ i \in I_{+s})\}| \ge \left(\frac{1-\sqrt{r}}{2} - o(1)\right) p, \text{ and}$$
$$|\{i \in [2p] : \forall\ s \in \mathcal{K}_k,\ i \in I_{-s})\}| \ge \left(\frac{1-\sqrt{r}}{2} - o(1)\right) p.$$

5. *For all pairs* $s, q \in [n]$ *such that* $y_s \neq y_q$,

$$\left|\left\{i \in [2p] : (i \in I_{+s}) \wedge \left(v_i^{(0)} \cdot x_q \ge 0\right)\right\}\right| \le \left(\frac{1}{3} + \frac{\Delta}{4} + r + o(1)\right) p, \text{ and}$$
$$\left|\left\{i \in [2p] : (i \in I_{-s}) \wedge \left(v_i^{(0)} \cdot x_q \ge 0\right)\right\}\right| \le \left(\frac{1}{3} + \frac{\Delta}{4} + r + o(1)\right) p.$$

6. *For all samples* $s \in [n]$,

$$\left|\left\{i \in [2p] : \left(-\alpha_0\left(\frac{1}{2} + 2(\Delta + r)\right) \le v_i^{(0)} \cdot x_s \le h + 4\alpha_0\right) \wedge (u_i \neq y_s)\right\}\right| \le \left(\frac{\sqrt{2}}{\sigma\sqrt{\pi}}\left(h + 5\alpha_0(2 + \Delta + r)\right)\right) p.$$

7. *The norm of the weight matrix after one iteration satisfies* $\frac{3}{5}\sqrt{\frac{d}{p^\beta}} \le \|V^{(1)}\| \le 3\sqrt{\frac{d}{p^\beta}}$.

Definition 16. *If the random initialization satisfies all of the conditions of Lemma 15, let us refer to the entire ensuing training process as* a good run.

Armed with Lemma 15, it suffices to show that the loss bounds of Theorem 2 hold on a good run. For the rest of the proof, let us assume that we are analyzing a good run.

Lemma 17. *For all small enough* $C_2 > 0$, *all large enough* C_3 *and all small enough* $C_4 > 0$, *the loss after the initial step of gradient descent is bounded above as follows:*

$$L_1 \le \exp\left(-C_4 p^{(1/2-\beta)}\right).$$

Proof Let us examine the loss of each example after one step. Consider an example $s \in [n]$. Without loss of generality let us assume that $y_s = 1$ and that it belongs to cluster $\mathcal{K}_k$:

$$L_{1s} = \log\left(1 + \exp\left(-\sum_{i=1}^{2p} u_i \phi(v_i^{(1)} \cdot x_s)\right)\right).$$

Since s is fixed and $y_s = 1$, we simplify the notation for I_{+s} and I_{-s} and define their complements, dividing the hidden nodes into four groups:

1. I_+ where $v_i^{(0)} \cdot x_s \geq h + 4\alpha_0$ and $u_i = 1$;
2. I_- where $v_i^{(0)} \cdot x_s \geq h + 4\alpha_0$ and $u_i = -1$;
3. $\widetilde{I}_+$ where $v_i^{(0)} \cdot x_s < h + 4\alpha_0$ and $u_i = 1$;
4. $\widetilde{I}_-$ where $v_i^{(0)} \cdot x_s < h + 4\alpha_0$ and $u_i = -1$.

We have

$$L_{1s} = \log\left(1 + \exp\left(-\sum_{i\in I_+}\phi(v_i^{(1)}\cdot x_s) + \sum_{i\in I_-}\phi(v_i^{(1)}\cdot x_s) - \sum_{i\in \widetilde{I}_+}\phi(v_i^{(1)}\cdot x_s) + \sum_{i\in \widetilde{I}_-}\phi(v_i^{(1)}\cdot x_s)\right)\right). \tag{10}$$

By definition of the gradient descent update we have, for each node i,

$$v_i^{(1)}\cdot x_s = v_i^{(0)}\cdot x_s + \frac{\alpha_0 u_i}{n}\sum_{q: v_i^{(0)}\cdot x_q \geq 0} y_q g_{0q}\phi'(v_i^{(0)}\cdot x_q)(x_q \cdot x_s).$$

Note that the groups I_+ and I_- are defined such that even after one step of gradient descent, for any node $i \in I_+ \cup I_-$

$$\phi(v_i^{(1)}\cdot x_s) = v_i^{(1)}\cdot x - h/2. \tag{11}$$

That is, $v_i^{(1)}\cdot x_s$ continues to lie in the linear region of ϕ after the first step. To see this, notice that for all q,

$$g_{0q}, \phi'(v_i^{(0)}\cdot x_q), x_q \cdot x_s \in [0,1],$$

and hence $|v_i^{(1)}\cdot x_s - v_i^{(0)}\cdot x_s| \leq \alpha_0$.

Our proof will proceed using four steps. Each step analyzes the contribution of nodes in a particular group. We give the outline here, deferring the proof of some parts to lemmas that follow.

Steps 1 and 2: In Lemma 18 we will show that, for an absolute constant c,

$$\sum_{i\in I_+}\phi(v_i^{(1)}\cdot x_s) \geq p\chi(h,\alpha_0,\sigma) + \frac{\alpha_0 p}{48}\left(1 - c(\Delta + \sqrt{r} + \varepsilon) - o(1)\right) - 2 \tag{12}$$

and

$$\sum_{i\in I_-}\phi(v_i^{(1)}\cdot x_s) \leq p\chi(h,\alpha_0,\sigma) - \frac{\alpha_0 p}{48}\left(1 - c(\Delta + \sqrt{r} + \varepsilon) - o(1)\right) + 2. \tag{13}$$

Step 3: Since the Huberized ReLU is non-negative a simple bound on the contribution of nodes in $\widetilde{I}_+$ is $\sum_{i\in\widetilde{I}_+}\phi(v_i^{(1)}\cdot x_k) \geq 0$.

Step 4: Finally in Lemma 19 we will show that the contribution of the nodes in $\widetilde{I}_-$ is bounded above by

$$\sum_{i\in\widetilde{I}_-}\phi(v_i^{(1)}\cdot x_s) \leq \frac{\sqrt{2}}{\sigma\sqrt{\pi}}\left(h + 5\alpha_0(2+\Delta+r)\right)^2 p. \tag{14}$$

Combining the bounds in inequalities (12), (13) and (14) with the decomposition of the loss in (10) we infer,

$$\begin{aligned} L_{1s} &\leq \log\left(1+\exp\left(-\frac{\alpha_0 p}{24}\left(1-c(\Delta+\sqrt{r}+\varepsilon)-o(1)\right)+\frac{\sqrt{2}}{\sigma\sqrt{\pi}}\left(h+5\alpha_0(2+\Delta+r)\right)^2 p\right)\right) \\ &\leq \log\left(1+\exp\left(-\frac{\alpha_0 p}{24}\left[1-c(\Delta+\sqrt{r}+\varepsilon)-o(1)\right]\right)\right), \end{aligned}$$

since $h=o(\alpha_0)$, and $\alpha_0=o(\sigma)$. Now since $\Delta, r, \varepsilon < C_2$, where C_2 is a small enough constant, $\alpha_0 = 1/p^{1/2+\beta}$ and because p is bigger than a suitably large constant we have,

$$L_{1s} \leq \log\left(1+\exp\left(-C_4\alpha_0 p\right)\right) = \log\left(1+\exp\left(-C_4 p^{1/2-\beta}\right)\right) \leq \exp\left(-C_4 p^{1/2-\beta}\right).$$

Recall that the sample s was chosen without loss of generality above. Therefore, by averaging over the n samples we have

$$L_1 = \frac{1}{n}\sum_{s=1}^{n} L_{1s} \leq \exp\left(-C_4 p^{1/2-\beta}\right)$$

establishing our claim. ■

Next, as promised in the proof of Lemma 17, we bound the contribution due to the nodes in I_+ and I_- after one step.

Lemma 18. *Borrowing all notation from the proof of Lemma 17 above, for all small enough C_2 and large enough C_3, there is an absolute constant c such that, on a good run*

$$\begin{aligned} \sum_{i\in I_+}\phi(v_i^{(1)}\cdot x_s) &\geq p\chi(h,\alpha_0,\sigma)+\frac{\alpha_0 p}{48}\left(1-c(\Delta+\sqrt{r}+\varepsilon)-o(1)\right)-2, \text{ and} \\ \sum_{i\in I_-}\phi(v_i^{(1)}\cdot x_s) &\leq p\chi(h,\alpha_0,\sigma)-\frac{\alpha_0 p}{48}\left(1-c(\Delta+\sqrt{r}+\varepsilon)-o(1)\right)+2. \end{aligned}$$

Proof We begin by analyzing the contribution of nodes in group I_+.

$$\begin{aligned} &\sum_{i\in I_+}\phi(v_i^{(1)}\cdot x_s) \\ &= \sum_{i\in I_+}\phi\left(v_i^{(0)}\cdot x_s+\frac{\alpha_0}{n}\sum_{q:v_i^{(0)}\cdot x_q\geq 0} y_q g_{0q}\phi'(v_i^{(0)}\cdot x_q)(x_q\cdot x_s)\right) \\ &= \sum_{i\in I_+}\left(v_i^{(0)}\cdot x_s+\frac{\alpha_0}{n}\sum_{q:v_i^{(0)}\cdot x_q\geq 0} y_q g_{0q}\phi'(v_i^{(0)}\cdot x_q)(x_q\cdot x_s)\right)-\frac{h|I_+|}{2} \\ &\qquad \text{(since the } i\in I_+ \text{ satisfy (11))} \\ &= \sum_{i\in I_+}\left(v_i^{(0)}\cdot x_s+\frac{\alpha_0 g_{0s}}{n}+\frac{\alpha_0}{n}\sum_{q\neq s:v_i^{(0)}\cdot x_q\geq 0} y_q g_{0q}\phi'(v_i^{(0)}\cdot x_q)(x_q\cdot x_s)\right)-\frac{h|I_+|}{2}. \end{aligned}$$

Since we are analyzing a good run, Parts 1 and 2 of Lemma 15 imply that $\sum_{i\in I_+} v_i^{(0)} \cdot x_k \geq p\chi(h,\alpha_0,\sigma) - 1$ and that $h|I_+| \leq 1/2 + o(1)$, therefore, for p larger than a constant,

$$\begin{aligned}
&\sum_{i\in I_+} \phi(v_i^{(1)} \cdot x_s) \\
&\geq p\chi(h,\alpha_0,\sigma) + \frac{\alpha_0}{n} \sum_{i\in I_+} \left(g_{0s} + \sum_{q\neq s: v_i^{(0)}\cdot x_q \geq 0} y_q g_{0q} \phi'(v_i^{(0)} \cdot x_q)(x_q \cdot x_s) \right) - 2 \\
&\geq p\chi(h,\alpha_0,\sigma) + \frac{\alpha_0}{n} \left(\sum_{i\in I_+} g_{0s} + \sum_{i\in I_+} \sum_{q\in\mathcal{K}_k - \{s\}, v_i^{(0)}\cdot x_q \geq h} g_{0q}(x_q \cdot x_s) \right. \\
&\qquad \left. - \sum_{i\in I_+} \sum_{q: v_i^{(0)}\cdot x_q \geq 0, y_q = -1} g_{0q} \phi'(v_i^{(0)} \cdot x_q)(x_q \cdot x_s) \right) - 2 \qquad (15)
\end{aligned}$$

where the previous inequality above follows in part by recalling that $s \in \mathcal{K}_k$ where $y_k = 1$, and noting that, since $x_s \cdot x_q \geq 0$ for all pairs, we can ignore contributions that have $y_q = 1$. Evolving this further

$$\begin{aligned}
\sum_{i\in I_+} \phi(v_i^{(1)} \cdot x_s) &\overset{(i)}{\geq} p\chi(h,\alpha_0,\sigma) + \frac{\alpha_0}{n} \left(\sum_{i\in I_+} g_{0s} + \sum_{i\in I_+} \sum_{q: q\in\mathcal{K}_k - \{s\}, v_i^{(0)}\cdot x_q \geq h} g_{0q}(1-2r) \right. \\
&\qquad \left. - \sum_{i\in I_+} \sum_{q: v_i^{(0)}\cdot x_q \geq 0, y_q = -1} g_{0q} \phi'(v_i^{(0)} \cdot x_q)(x_q \cdot x_s) \right) - 2 \\
&\overset{(ii)}{\geq} p\chi(h,\alpha_0,\sigma) + \frac{\alpha_0}{n} \left(\underbrace{\sum_{i\in I_+} g_{0s} + \sum_{i\in I_+} \sum_{q: q\in\mathcal{K}_k - \{s\}, v_i^{(0)}\cdot x_q \geq h} g_{0q}(1-2r)}_{=:\Xi_1} \right. \\
&\qquad \left. \underbrace{- \sum_{i\in I_+} \sum_{q: v_i^{(0)}\cdot x_q \geq 0, y_q = -1} g_{0q}(x_q \cdot x_s)}_{=:\Xi_2} \right) - 2, \qquad (16)
\end{aligned}$$

where (i) follows since, when s and q are in the same cluster, $x_s \cdot x_q \geq 1 - 2r$ (which is proved in Lemma 21 below) and (ii) follows since ϕ is 1-Lipschitz. Next we provide a lower bound on

the term Ξ_1

$$\begin{aligned}
\Xi_1 &= \sum_{i\in I_+} g_{0s} + \sum_{i\in I_+} \sum_{q\in\mathcal{K}_k-\{s\}, v_i^{(0)}\cdot x_q\geq h} g_{0q}(1-2r) \\
&= g_{0s}|I_+| + \sum_{i\in I_+} \sum_{q\in\mathcal{K}_k-\{s\}, v_i^{(0)}\cdot x_q\geq h} g_{0q}(1-2r) \\
&\overset{(i)}{\geq} \left(\frac{1}{2}-o(1)\right)|I_+| \\
&\qquad + \left(\frac{1}{2}-o(1)\right)(1-2r)\sum_{q:q\in\mathcal{K}_k-\{s\}} \left|\left\{i\in[2p] : i\in I_+ \text{ and } v_i^{(0)}\cdot x_q\geq h\right\}\right| \\
&\overset{(ii)}{\geq} \left(\frac{1}{2}-o(1)\right)\left[\left(\frac{1}{2}-o(1)\right)p + (1-2r)(|\mathcal{K}_k|-1)\left(\frac{1-\sqrt{r}}{2}-o(1)\right)p\right] \\
&\geq \left(\frac{1}{2}-o(1)\right)|\mathcal{K}_k|(1-2r)\left(\frac{1-\sqrt{r}}{2}-o(1)\right)p \\
&\overset{(iii)}{\geq} \left(\frac{1}{2}-o(1)\right)\left(\frac{1}{4}-\varepsilon\right)(1-2r)\left(\frac{1-\sqrt{r}}{2}-o(1)\right)np \\
&\overset{(iv)}{\geq} \frac{1}{16}\left(1-c_1(\sqrt{r}+\varepsilon)-o(1)\right)np, \qquad (17)
\end{aligned}$$

for an absolute positive constant c_1, where (i) follows since, by Part 3 of Lemma 15, on a good run, $g_{0s}\geq 1/2-o(1)$ for all samples, (ii) follows by using Parts 2 and 4 of Lemma 15, (iii) is by the assumption that $|\mathcal{K}_k|\geq(1/4-\varepsilon)n$ and the simplification in (iv) follows since both $r,\varepsilon<C_2$ for a small enough constant C_2.

Now we upper bound Ξ_2 to get,

$$\begin{aligned}
\Xi_2 &= \sum_{i\in I_+} \sum_{q:v_i^{(0)}\cdot x_q\geq 0, y_q=-1} g_{0q}(x_q\cdot x_s) \\
&\overset{(i)}{\leq} \left(\frac{1}{2}+o(1)\right)\sum_{i\in I_+} \sum_{q:v_i^{(0)}\cdot x_q\geq 0, y_q=-1} x_q\cdot x_s \\
&\overset{(ii)}{\leq} \left(\frac{1}{2}+o(1)\right)\sum_{i\in I_+} \sum_{q:v_i^{(0)}\cdot x_q\geq 0, y_q=-1} \left(\frac{1+\Delta}{2}+2r\right) \\
&= \left(\frac{1}{2}+o(1)\right)\left(\frac{1+\Delta}{2}+2r\right)\sum_{q:y_q=-1} \left|\left\{i\in[2p] : i\in I_+ \text{ and } v_i^{(0)}\cdot x_q\geq 0\right\}\right| \\
&\overset{(iii)}{\leq} \left(\frac{1}{2}+o(1)\right)\left(\frac{1+\Delta+4r}{2}\right)(|\mathcal{K}_3|+|\mathcal{K}_4|)\left(\frac{1}{3}+\frac{\Delta}{4}+r+o(1)\right)p \\
&\overset{(iv)}{\leq} \left(\frac{1}{2}+o(1)\right)\left(\frac{1+\Delta+4r}{2}\right)\left(\frac{1}{2}+2\varepsilon\right)\left(\frac{1}{3}+\frac{\Delta}{4}+r+o(1)\right)np \\
&\overset{(v)}{\leq} \frac{(1+c_2(\Delta+r+\varepsilon)+o(1))\,np}{24}, \qquad (18)
\end{aligned}$$

for an absolute positive constant c_2, where (i) follows as, by Part 3 of Lemma 15, on a good run, for all samples $g_{0q}\leq 1/2+o(1)$, (ii) follows from the fact that, for q and s from opposite

classes, $x_q \cdot x_s \leq \frac{1+\Delta}{2} + 2r$ (which is proved in Lemma 21 below), (iii) is obtained by invoking Part 5 of Lemma 15, (iv) is by the assumption that all clusters have at most $(1/4+\varepsilon)n$ examples and the simplification in (v) follows since $\Delta, r, \varepsilon < C_2$ where C_2 is a small enough constant.

Combining the conclusion of inequality (16) with the bounds in (17) and (18) completes the proof of the first part of the lemma:

$$\sum_{i \in I_+} \phi(v_i^{(1)} \cdot x_s) \geq p\chi(h, \alpha_0, \sigma) + \frac{\alpha_0 p}{48}\left(1 - c_3(\Delta + \sqrt{r} + \varepsilon) - o(1)\right) - 2.$$

Now we move on to analyzing the contribution of the group I_-.

$$\begin{aligned}
\sum_{i \in I_-} \phi(v_i^{(1)} \cdot x_s) &\leq \sum_{i \in I_-} \left(v_i^{(1)} \cdot x_s\right) \\
&= \sum_{i \in I_-} \left(v_i^{(0)} \cdot x_s - \frac{\alpha_0}{n} \sum_{q : v_i^{(0)} \cdot x_q \geq 0} y_q g_{0q} \phi'(v_i^{(0)} \cdot x_q)(x_q \cdot x_s) \right) \\
&\overset{(i)}{\leq} \sum_{i \in I_-} v_i^{(0)} \cdot x_s - \frac{\alpha_0}{n} \left(\sum_{i \in I_-} g_{0s} + \sum_{i \in I_-} \sum_{q \in \mathcal{K}_k - \{s\}, v_i^{(0)} \cdot x_q \geq h} g_{0q}(x_q \cdot x_s) \right) \\
&\qquad + \frac{\alpha_0}{n} \sum_{i \in I_-} \sum_{q : v_i^{(0)} \cdot x_q \geq 0, y_q = -1} g_{0q} \phi'(v_i^{(0)} \cdot x_q)(x_q \cdot x_q) \\
&\overset{(ii)}{\leq} p\chi(h, \alpha_0, p) - \frac{\alpha_0}{n} \left(\sum_{i \in I_-} g_{0s} + \sum_{i \in I_-} \sum_{q \in \mathcal{K}_k - \{s\}, v_i^{(0)} \cdot x_q \geq h} g_{0q}(x_q \cdot x_s) \right. \\
&\qquad \left. + \sum_{i \in I_-} \sum_{q : v_i^{(0)} \cdot x_q \geq 0, y_q = -1} g_{0q} \phi'(v_i^{(0)} \cdot x_q)(x_q \cdot x_q) \right) + 1,
\end{aligned}$$

where (i) follows by noting that $x_s \cdot x_q \geq 0$ for all pairs, therefore we can ignore contributions that have $y_q = 1$, and (ii) is by Part 1 of Lemma 15. Now by using an argument that is identical to that in first part of the proof that bounded the contribution of I_+ above starting from inequality (15) we conclude

$$\sum_{i \in I_-} \phi(v_i^{(1)} \cdot x_s) \leq p\chi(h, \alpha_0, \sigma) - \frac{\alpha_0 p}{48}\left(1 - c_3(\Delta + \sqrt{r} + \varepsilon) - o(1)\right) + 2.$$

This establishes our bound on the contribution of the nodes in I_-. ■

In the following lemma we bound the contribution of the nodes in $\widetilde{I}_-$ defined in the proof of Lemma 17.

Lemma 19. *Borrowing all notation from the proof of Lemma 17 above, on a good run,*

$$\sum_{i \in \widetilde{I}_-} \phi(v_i^{(1)} \cdot x_s) \leq \frac{\sqrt{2}}{\sigma\sqrt{\pi}} \left(h + 5\alpha_0(2 + \Delta + r)\right)^2 p.$$

Proof Recalling that $v_i^{(1)}$ is obtained by taking a gradient step

$$
\begin{aligned}
&\sum_{i\in\widetilde{I}_-}\phi(v_i^{(1)}\cdot x_s)\\
&=\sum_{i\in\widetilde{I}_-}\phi\left(v_i^{(0)}\cdot x_s-\frac{\alpha_0}{n}\sum_{q:v_i^{(0)}\cdot x_q\geq 0}y_q g_{0q}\phi'(v_i^{(0)}\cdot x_q)(x_q\cdot x_s)\right)\\
&\overset{(i)}{\leq}\sum_{i\in\widetilde{I}_-}\phi\left(v_i^{(0)}\cdot x_s+\frac{\alpha_0}{n}\sum_{q:v_i^{(0)}\cdot x_q\geq 0,y_q=-1}g_{0q}\phi'(v_i^{(0)}\cdot x_q)(x_q\cdot x_s)\right)\\
&\overset{(ii)}{\leq}\sum_{i\in\widetilde{I}_-}\phi\left(v_i^{(0)}\cdot x_s+\frac{\alpha_0}{n}\sum_{q:v_i^{(0)}\cdot x_q\geq 0,y_q=-1}(x_q\cdot x_s)\right)\\
&\overset{(iii)}{\leq}\sum_{i\in\widetilde{I}_-}\phi\left(v_i^{(0)}\cdot x_s+\frac{\alpha_0}{n}\left(n\left(\frac{1}{2}+\frac{\Delta}{2}+2r\right)\right)\right)\\
&\leq\sum_{i\in\widetilde{I}_-}\phi\left(v_i^{(0)}\cdot x_s+\alpha_0\left(\frac{1}{2}+2\left(\Delta+r\right)\right)\right)\\
&\overset{(iv)}{\leq}\left|\left\{i\in[p+1,\ldots,2p]:-\alpha_0\left(\frac{1}{2}+2(\Delta+r)\right)\leq v_i^{(0)}\cdot x_s\leq h+4\alpha_0\right\}\right|\\
&\qquad\qquad\times\left(h+4\alpha_0+\alpha_0\left(\frac{1}{2}+2(\Delta+r)\right)\right)\\
&\overset{(v)}{\leq}\left(\frac{\sqrt{2}}{\sigma\sqrt{\pi}}\left(h+5\alpha_0(2+\Delta+r)\right)\right)p\times\left(h+4\alpha_0+\alpha_0\left(\frac{1}{2}+2(\Delta+r)\right)\right)\\
&\leq\frac{\sqrt{2}}{\sigma\sqrt{\pi}}\left(h+5\alpha_0(2+\Delta+r)\right)^2p,
\end{aligned}
$$

where (i) follows by discarding the contribution of the examples with the same label $y_q=1$, (ii) is because $g_{0\ell}$ and ϕ' are non-negative and bounded by 1, (iii) follows by the bound $x_q\cdot x_s\leq(1+\Delta+4r)/2$ established in Lemma 21 below. Inequality (iv) follows from the facts that $\phi(z)=0$ for all $z<0$ and $v_i^{(0)}\cdot x_s\leq h+4\alpha_0$ for all $i\in\widetilde{I}_+$, and finally (v) follows from Part 6 of Lemma 15. This establishes the claim. ■

4.3 Proof of Theorem 2

Having analyzed the first step we are now ready to prove Theorem 2.

Part (a) of the theorem follows by invoking Lemma 17 that shows that after the first step $L_1\leq\exp\left(-C_4p^{(1/2-\beta)}\right)$ with probability at least $1-\delta$.

Part (b) of the theorem shall follow by invoking Theorem 1. Since $p\geq\log^{C_3}(nd/\delta)$ for a large enough constant C_3 we know that $L_1\leq 1/n^{1+C_1}$ as required by Theorem 1. Also note that Part 7 of Lemma 15 ensures that on a good run, $\frac{3}{5}\sqrt{\frac{d}{p^\beta}}\leq\|V^{(1)}\|\leq 3\sqrt{\frac{d}{p^\beta}}$.

Set the value of $Q_1 = \frac{e^2}{120p}$. (This sets the step-size $\alpha_t = Q_1 \log^2(1/L_t) = \frac{e^2 \log^2(1/L_t)}{120p}$.) To invoke Theorem 1 we need to ensure that $Q_1 \leq \widetilde{Q}_1$ (see its definition in equation (6)), but this is easy to check since

$$\begin{aligned}
\widetilde{Q}_1 &= \min\left\{\frac{1}{30pL_1\log^2(1/L_1)}, \frac{108\|V^{(1)}\|^2}{125L_1\log^4(1/L_1)}, \frac{e^2}{120p}\right\} \\
&\geq \min\left\{\frac{\exp(C_4 p^{(1/2-\beta)})}{30C_4^2 p^{2-2\beta}}, \frac{972d\exp(C_4 p^{(1/2-\beta)})}{3125C_4^4 p^{2-3\beta}}, \frac{e^2}{120p}\right\} \\
&\qquad \text{(since } L_1 \leq \exp\left(-C_4 p^{(1/2-\beta)}\right) \text{ and } \|V^{(1)}\| \geq \tfrac{3}{5}\sqrt{\tfrac{d}{p^\beta}}\text{)} \\
&= \frac{e^2}{120p},
\end{aligned}$$

where the final equality holds since $p \geq \log^{C_3} d$. Next we set $Q_2 = \widetilde{Q}_2(Q_1)$ (recall its definition from equation (7) above):

$$Q_2 = \widetilde{Q}_2(Q_1) = \frac{125Q_1L_1\log^4(1/L_1)}{216\|V^{(1)}\|^2}.$$

With these valid choices of Q_1 and Q_2 we now invoke Theorem 1 to get that, for all $t > 1$

$$\begin{aligned}
L_t &\leq \frac{L_1}{Q_2 \cdot (t-1) + 1} \\
&\leq \frac{L_1}{\frac{125Q_1L_1\log^4(1/L_1)p^\beta(t-1)}{1944d} + 1} \qquad \text{(since } \|V^{(1)}\| \leq 3\sqrt{\tfrac{d}{p^\beta}}\text{)} \\
&= \frac{L_1}{\frac{c_1L_1\log^4(1/L_1)(t-1)}{dp^{1-\beta}} + 1} \\
&\leq \frac{L_1}{\max\left\{\frac{c_1L_1\log^4(1/L_1)(t-1)}{dp^{1-\beta}}, 1\right\}} \\
&= \min\left\{\frac{dp^{1-\beta}}{c_1\log^4(1/L_1)(t-1)}, L_1\right\} \\
&\leq \min\left\{\frac{dp^{1-\beta}}{c_1C_4^4p^{2-4\beta}(t-1)}, L_1\right\} \\
&= \min\left\{\frac{d}{c_2p^{1-3\beta}(t-1)}, L_1\right\}.
\end{aligned}$$

Combining this with Part (a), together with the assumption that $p \geq \log^{C_3} d$, proves Part (b).

5 Simulations

In this section, we experimentally verify the convergence results of Theorem 1. We performed 100 rounds of batch gradient descent to minimize the softmax loss on random training data. The training data was for a two-class classification problem. There were 128 random examples drawn from a distribution in which each of two equally likely classes was distributed as a

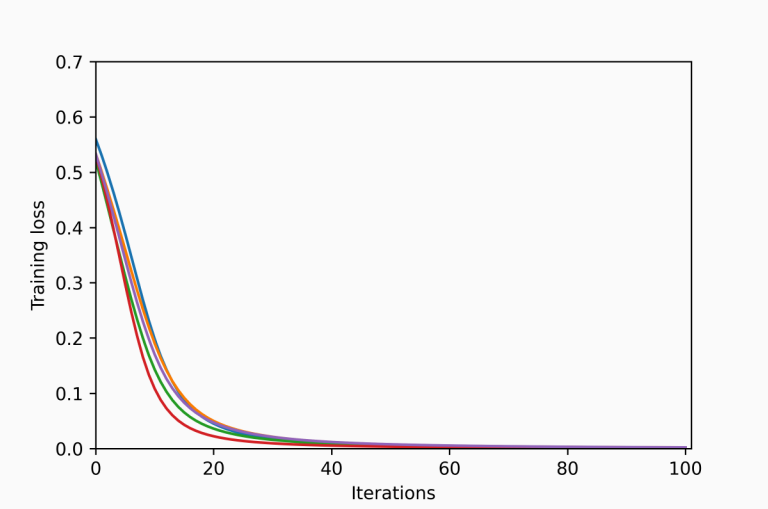

Figure 2. Training loss as a function of the number of gradient descent steps for data distributed as mixtures of Gaussians, where the means have an XOR structure. Details are in Section 5.

mixture of Gaussians whose centers had an XOR structure: the positive examples came from an equal mixture of

$$\mathcal{N}\left(\left(\sqrt{\frac{1}{2}}, -\sqrt{\frac{1}{2}}\right), \frac{I}{100}\right) \text{ and } \mathcal{N}\left(\left(-\sqrt{\frac{1}{2}}, \sqrt{\frac{1}{2}}\right), \frac{I}{100}\right),$$

and the negative examples came from an equal mixture of

$$\mathcal{N}\left(\left(\sqrt{\frac{1}{2}}, \sqrt{\frac{1}{2}}\right), \frac{I}{100}\right) \text{ and } \mathcal{N}\left(\left(-\sqrt{\frac{1}{2}}, -\sqrt{\frac{1}{2}}\right), \frac{I}{100}\right).$$

The number p of hidden units per class was 100. The activation functions were Huberized ReLUs with $h = 1/p$. The weights were initialized using $\mathcal{N}(0, (4p)^{-5/4})$ and the initial step size was $(4p)^{-3/4}$. (These correspond to the choice $\beta = 1/4$ in Theorem 2.) For the other updates, the step-size on iteration t was $\log^2(1/L_t)/p$. The process of randomly generating data, randomly initializing a network, and running gradient descent was repeated 5 times, and the curves of training error as a function of update number are plotted in Figure 2. The decrease in the loss with the number of iterations is roughly in line with our upper bounds.

We performed a similar collection of simulations, except with a different, more challenging, data distribution, which we call the "shoulders" distribution. The means of the mixture components of the positive examples were $(1, 0)$ and $(0, 1)$ while the means of mixture components of the negative examples remain at the same place. The positive centers start to crowd the negative center $(1/\sqrt{2}, 1/\sqrt{2})$ making it more difficult to pick out examples from the negative center. Plots for this data distribution, which also scale roughly like our upper bounds, are shown in Figure 3.

Next, we performed ten training runs as described above for the shoulders data, except that, for five of them, the Huberized ReLU was replaced by a standard ReLU. The results are in Figure 4. While there is evidence that training with the non-smooth objective arising from the standard ReLU leads to a limited extent of "overshooting", the shapes of the loss curves agree on a coarser scale.

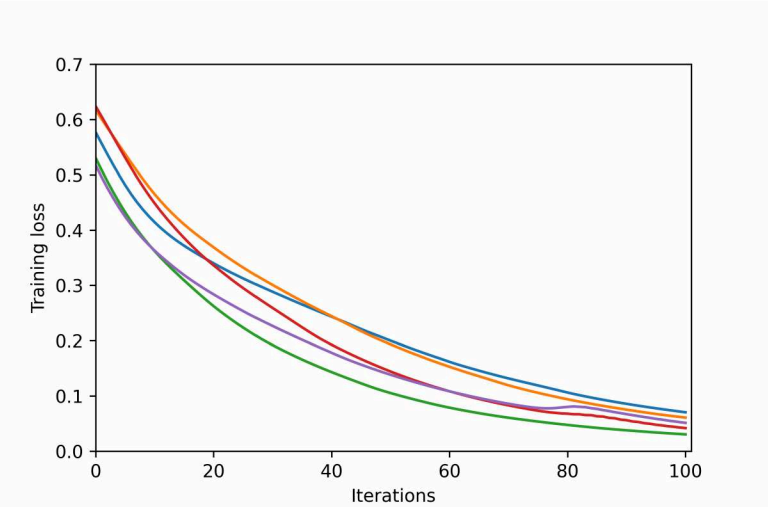

Figure 3. Training loss as a function of the number of gradient descent steps for the "shoulders" distribution. Details are in Section 5.

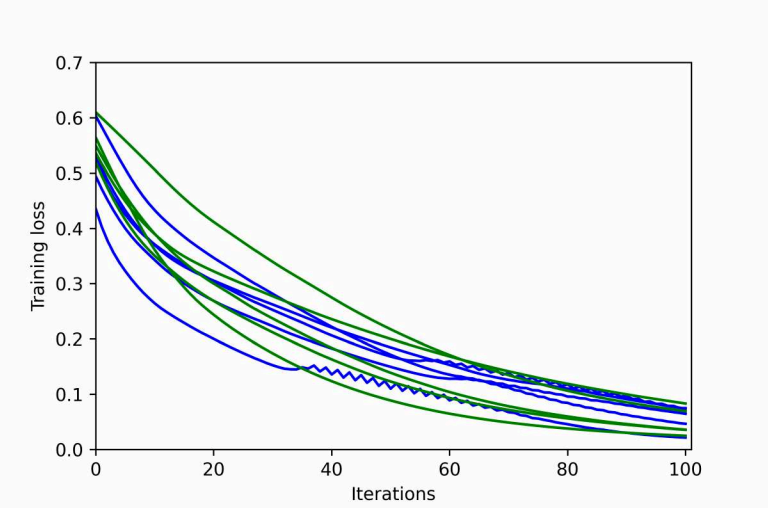

Figure 4. Training loss as a function of the number of gradient descent steps with the ReLU (blue) and the Huberized ReLU (green) on the "shoulders" data.

6 Additional Related Work

Chizat, Oyallon, and Bach [COB19] analyzed gradient flow for a general class of smoothly parameterized models, showing that scaling up the initialization, while scaling down the loss, ensures that a first-order Taylor approximation around the initial solution remains accurate until convergence.

Chizat and Bach [CB20], building on [CB18; MMM19], show that infinitely wide two-layer squared ReLU networks trained with gradient flow on the logistic loss leads to a max-margin classifier in a particular non-Hilbertian space of functions. (See also the videos in a talk about this work [Chi20].) Brutzkus et al. [Bru+18] show that finite-width two-layer leaky ReLU networks can be trained up to zero-loss using stochastic gradient descent with the hinge loss, when the underlying data is linearly separable.

The papers [BG19; Wei+19; JT19b] identify when it is possible to efficiently learn XOR-type data using neural networks with stochastic gradient descent on the logistic loss.

Chen et al. [Che+20] analyzed regularized training with gradient flow on infinitely wide networks. When training is regularized, the weights also may travel far from their initial values.

Our study is motivated in part by the line of work that has emerged which emphasizes the need to understand the behavior of interpolating (zero training loss/error) classifiers and regressors [see, e.g., Zha+17a; Bel+19, among others]. A number of recent papers have analyzed the properties of interpolating methods in linear regression [Has+19; Bar+20; Mut+20b; TB20; BL20], linear classification [Mon+19; CL21; LS20; Mut+20a; HMX20], kernel regression [LR20; MM19; LRZ20] and simplicial nearest neighbor methods [BHM18].

Also related are the papers that study the implicit bias of gradient methods [NTS15; Sou+18; JT19c; Gun+18a; Gun+18b; LMZ18; Aro+19a; JT19a].

A number of recent papers also theoretically study the optimization of neural networks including [And+14; LY17; Zho+17; Zha+17b; GLM18; PSZ18; Du+18; SS18; Zha+19; Aro+19b; Dan20; DM20; BN20].

In particular, the proof of Daniely and Malach [DM20] demonstrated that the first iteration of gradient descent learned useful features for the parity-learning problem studied there.

7 Discussion

We demonstrated that gradient descent drives the logistic loss of finite-width two-layer Huberized ReLU networks to zero if the initial loss is small enough. This result makes no assumptions about the width or the number of samples. We also showed that when the data is structured, and the data satisfies certain cluster and separation conditions, random initialization followed by gradient descent drives the loss to zero.

After a preliminary version of this paper was posted on arXiv [CLB20], related results were obtained [CLB21] for deep networks with smoothed approximations to the ReLU, under conditions that include Swish. This analysis included adapting the NTK techniques to these activation functions. This provides a broader set of circumstances under which Theorem 1 of this paper can be applied.

Another interesting way forward would be to examine whether the loss can be shown to decrease super-polynomially with the width when there are more than two clusters per label or if the number of samples per cluster is imbalanced.

It would be interesting to see if the corresponding results hold for ReLU activations, which, despite the success of Swish, remain popular.

Now that we have established conditions under which gradient descent can drive the training error to zero, future work could study the implicit bias of this limit and potentially use this to study the generalization behavior of the final interpolating solution. One step towards this could be establish a more precise directional alignment result to show that gradient descent maximizes the margin of Huberized ReLU networks for logistic loss [as LL20; JT20, did for ReLU networks trained using gradient flow].

Theorems 1 and 2 use a concrete choice of a learning rate schedule (at least, up to a constant factor). We believe that our techniques can be extended to apply to a wider variety of learning-rate schedules, with corresponding changes to the convergence rate.

In our paper, we assumed that the features are all unit-length vectors to simplify the proofs. We believe that the results of Theorem 1 can be easily extended to the case where the features have arbitrary bounded lengths. We also expect that the results of Theorem 2 can be extended to the case where the examples in the four clusters are drawn from sub-Gaussian distributions with suitably small variances.

Acknowledgements

We thank Zeshun Zong for alerting us to a mistake in an earlier version of this paper.

We gratefully acknowledge the support of the NSF through grants DMS-2031883 and DMS-2023505 and the Simons Foundation through award 814639.

A Reduction to the Case of No Bias

Denoting the components of x_s by $x_{s1}, \ldots, x_{sd}$, define $\tilde{x}_s = (x_{s1}/\sqrt{2}, \ldots, x_{sd}/\sqrt{2}, 1/\sqrt{2})$. We consider the process of training a model using $(\tilde{x}_1, y_1), \ldots, (\tilde{x}_n, y_n)$.

Consider

$$(\theta^{(1)}, \widetilde{V}^{(1)}), (\theta^{(2)}, \widetilde{V}^{(2)}), \ldots$$

defined as follows. First, $\theta^{(1)}, \theta^{(2)}, \ldots$ are generated as described in Section 2. Each row $\tilde{v}_i^{(1)}$ of $\widetilde{V}^{(1)} \in \mathbb{R}^{2p \times (d+1)}$ is $\sqrt{2}(v_{i1}^{(1)}, \ldots, v_{id}^{(1)}, b_i^{(1)})$.

Define $\widetilde{L}$ to be, informally, L, but without the bias terms, and applied to

$$(\tilde{x}_1, y_1), \ldots, (\tilde{x}_n, y_n).$$

That is

$$\widetilde{L}(\widetilde{V}) := \frac{1}{n} \sum_{s=1}^{n} \log \left(1 + \exp \left(-y_s \sum_{i=1}^{2p} u_i \phi(\tilde{v}_i \cdot \tilde{x}_s) \right) \right).$$

Then, for $\tilde{\alpha}_1 = 2\alpha_1, \tilde{\alpha}_2 = 2\alpha_2, \ldots > 0$, we define $\widetilde{V}^{(2)}, \widetilde{V}^{(3)}, \ldots$ to be the iterates of gradient descent applied to $\widetilde{L}$, except replacing $\alpha_1, \alpha_2, \ldots$ by $\tilde{\alpha}_1, \tilde{\alpha}_2, \ldots$.

We claim that, for all t,

- for all i, $\tilde{v}_i^{(t)} = \sqrt{2}(v_{i1}^{(t)}, \ldots, v_{id}^{(t)}, b_i^{(t)})$
- for all i and all s, $\tilde{v}_i^{(t)} \cdot \tilde{x}_s = v_i^{(t)} \cdot x_s + b_i^{(t)}$, so that $\widetilde{L}(\widetilde{V}^{(t)}) = L(\theta^{(t)})$.

The first condition is easily seen to imply the second. Further, the first condition holds at $t = 1$ by construction. What remains is to prove that the inductive hypothesis for iteration t implies the first condition at iteration $t + 1$. If f_V is the function computed by the network with weights V and no biases, we have

$$\begin{aligned}
\tilde{v}_i^{(t+1)} &= \tilde{v}_i^{(t)} + \tilde{\alpha}_t \sum_{s=1}^{n} \frac{1}{1 + \exp(y_s f_{\tilde{V}}^{(t)}(\tilde{x}_s))} \phi'(\tilde{v}_i^{(t)} \cdot \tilde{x}_s) y_s u_i \tilde{x}_s \\
&= \sqrt{2}(v_{i1}^{(t)}, \ldots, v_{id}^{(t)}, b_i^{(t)}) + \tilde{\alpha}_t \sum_{s=1}^{n} \frac{1}{1 + \exp(y_s f_{\theta}^{(t)}(x_s))} \phi'(v_i^{(t)} \cdot x_s + b_i^{(t)}) y_s u_i \tilde{x}_s \\
&\qquad\qquad \text{(by the inductive hypothesis)} \\
&= \sqrt{2}(v_{i1}^{(t+1)}, \ldots, v_{id}^{(t+1)}, b_i^{(t+1)}),
\end{aligned}$$

because $\tilde{x}_s = (x_{s1}/\sqrt{2}, \ldots, x_{sd}/\sqrt{2}, 1/\sqrt{2})$ and $\tilde{\alpha}_t = 2\alpha_t$, completing the induction.

Finally, note that

$$\tilde{x}_s \cdot \tilde{x}_q = \frac{x_s \cdot x_q + 1}{2} \geq 0,$$

since x_s and x_q are unit length.

B Omitted Proofs from Section 3.3

In this section we provide proofs of Lemmas 5-8.

B.1 Proof of Lemma 5

Lemma 5. *For $\alpha_t > 0$, let $V^{(t+1)} = V^{(t)} - \alpha_t \nabla L_t$. If, for all convex combinations W of $V^{(t)}$ and $V^{(t+1)}$, we have $\|\nabla_W^2 L\|_{op} \leq M$, then if $\alpha_t \leq \frac{1}{3M}$, we have*

$$L_{t+1} \leq L_t - \frac{5\alpha_t \|\nabla L_t\|^2}{6}.$$

Proof For any $W \in [V^{(t)}, V^{t+1}]$ we have that

$$\nabla_W L(W) - \nabla L_t = \int_0^1 \left(\nabla_{\bar{W}}^2 L|_{\bar{W} = sV^{(t)} + (1-s)W} \right) (W - V^{(t)}) \, \mathrm{d}s,$$

where, as stated above, the weak Hessian is defined using the weak derivative γ of ϕ'. Thus,

$$\|\nabla_W L(W) - \nabla L_t\| \leq \left[\sup_{s \in [0,1]} \left\| \nabla_{\bar{W}}^2 L|_{\bar{W} = sV^{(t)} + (1-s)W} \right\|_{op} \right] \|W - V^{(t)}\| \leq M \|W - V^{(t)}\|.$$

This shows that along the line segment joining $V^{(t)}$ to $V^{(t+1)}$ the function is M-smooth. Therefore, by using a standard argument [see, e.g, Bub15, Lemma 3.4] we get that

$$\begin{aligned}
L_{t+1} &\leq L_t + \nabla L_t \cdot (V^{(t+1)} - V^{(t)}) + \frac{M}{2} \|V^{(t+1)} - V^{(t)}\|^2 \\
&= L_t - \alpha_t \|\nabla L_t\|^2 + \frac{\alpha_t^2 M}{2} \|\nabla L_t\|^2 \\
&= L_t - \alpha_t \left(1 - \frac{\alpha_t M}{2}\right) \|\nabla L_t\|^2 \\
&\leq L_t - \frac{5\alpha_t \|\nabla L_t\|^2}{6}.
\end{aligned}$$

This completes the proof. ■

B.2 Proof of Lemma 6

Lemma 6. *If $h = 1/p$, for any weight matrix $V \in \mathbb{R}^{2p\times(d+1)}$, $\|\nabla_V^2 L\|_{op} \le 5pL(V)$.*

Proof We know that the gradient of the loss with respect to v_i is

$$\nabla_{v_i} L = \frac{-u_i}{n}\sum_{s=1}^{n}\frac{\phi'(v_i\cdot x_s)y_s x_s}{1+\exp\left(y_s f_V(x_s)\right)}.$$

The weak Hessian $\nabla^2 L$ is a block matrix with $4p^2$ blocks, where the $(i,j)^{th}$ block is $\nabla_{v_i}\nabla_{v_j} L$.

First, if $i \neq j$

$$\nabla_{v_i}\nabla_{v_j} L = \frac{u_i u_j}{n}\sum_{s=1}^{n}\frac{\phi'(v_i\cdot x_s)\phi'(v_j\cdot x_s)\exp(y_s f_V(x_s))}{\left(1+\exp\left(y_s f_V(x_s)\right)\right)^2}x_s x_s^\top. \tag{19}$$

If $i = j$,

$$\nabla_{v_i}^2 L = \frac{1}{n}\sum_{s=1}^{n}\left[\frac{-u_i\gamma(v_i\cdot x_s)y_s}{1+\exp\left(y_s f_V(x_s)\right)} + \frac{\phi'(v_i\cdot x_s)^2\exp(y_s f_V(x_s))}{\left(1+\exp\left(y_s f_V(x_s)\right)\right)^2}\right]x_s x_s^\top. \tag{20}$$

By definition of the operator norm,

$$\|\nabla_V^2 L\|_{op} = \sup_{a:\|a\|=1}\left\|\left(\nabla_V^2 L\right)a\right\|. \tag{21}$$

Let a be a unit length member of $\mathbb{R}^{2p(d+1)}$ and let us decompose a into segments $a_1,\ldots,a_{2p}$ of $(d+1)$ components each, so that a is the concatenation of $a_1,\ldots,a_{2p}\in\mathbb{R}^{d+1}$. Note that $\sum_{i=1}^{2p}\|a_i\|^2 = 1$.

The squared norm of $(\nabla^2 L)a$ is

$$\begin{aligned}\left\|(\nabla^2 L)a\right\|^2 &= \sum_{i\in[2p]}\left\|\sum_{j\in[2p]}\left(\nabla_{v_i}\nabla_{v_j}L\right)a_j\right\|^2\\ &= \sum_{i,j,k\in[2p]} a_k^\top\left[\left(\nabla_{v_i}\nabla_{v_k}L\right)\left(\nabla_{v_i}\nabla_{v_j}L\right)\right]a_j\\ &\le \sum_{i,j,k\in[2p]}\|a_j\|\|a_k\|\|\nabla_{v_i}\nabla_{v_k}L\|_{op}\|\nabla_{v_i}\nabla_{v_j}L\|_{op}.\end{aligned} \tag{22}$$

By definition of the Huberized ReLU (in equation (1)) and its weak Hessian (in equation (2)) we know that, for any $z\in\mathbb{R}$, $|\phi'(z)|<1$ and $|\gamma(z)|<1/h$. Further, by Lemma 20, we know that for all s

$$\frac{\exp(y_s f_V(x_s))}{\left(1+\exp\left(y_s f_V(x_s)\right)\right)^2} \le \frac{1}{1+\exp\left(y_s f_V(x_s)\right)} \le L(V;x_s,y_s) = \log(1+\exp(-y_s f_V(x_s))).$$

Also recall that for all $s \in [n]$, $\|x_s\| = 1$ and for all $i \in [2p]$, $|u_i| = 1$. Applying these to equation (19), when $i \neq j$ we get that

$$\|(\nabla_{v_i}\nabla_{v_j}L)\|_{op} \leq L, \tag{23}$$

and, using equation (20), when $i = j$ yields the bound

$$\|\nabla^2_{v_i}L\|_{op} \leq L\left(1 + \frac{1}{h}\right) \leq 2L/h. \tag{24}$$

Returning to inequality (22),

$$\begin{aligned}
\left\|(\nabla^2 L)a\right\|^2 &\leq \sum_{i,j,k\in[2p]} \|a_j\|\|a_k\|\|\nabla_{v_i}\nabla_{v_k}L\|_{op}\|\nabla_{v_i}\nabla_{v_j}L\|_{op} \\
&= \sum_{i,j,k\in[2p]:i=j=k} \|a_j\|^2\|\nabla^2_{v_i}L\|^2_{op} \\
&\quad + \sum_{i,j,k\in[2p]:(i\neq j)\wedge(i\neq k)} \|a_j\|\|a_k\|\|\nabla_{v_i}\nabla_{v_k}L\|_{op}\|\nabla_{v_i}\nabla_{v_j}L\|_{op} \\
&\quad + \sum_{i,j,k\in[2p]:i=j\neq k} \|a_j\|\|a_k\|\|\nabla_{v_i}\nabla_{v_k}L\|_{op}\|\nabla^2_{v_i}L\|_{op} \\
&\quad + \sum_{i,j,k\in[2p]:i=k\neq j} \|a_j\|\|a_k\|\|\nabla_{v_i}\nabla_{v_j}L\|_{op}\|\nabla^2_{v_i}L\|_{op}.
\end{aligned} \tag{25}$$

Recall that $h = 1/p$, therefore, by inequality (24), the first term in the inequality above can be bounded by

$$\sum_{i,j,k\in[2p]:i=j=k} \|a_j\|^2\|\nabla^2_{v_i}L\|^2_{op} \leq (2L/h)^2\sum_j \|a_j\|^2 = 4L^2p^2.$$

Using inequality (23), the second term in the RHS of inequality (25) is

$$\begin{aligned}
&\sum_{i,j,k\in[2p]:(i\neq j)\wedge(i\neq k)} \|a_j\|\|a_k\|\|\nabla_{v_i}\nabla_{v_k}L\|_{op}\|\nabla_{v_i}\nabla_{v_j}L\|_{op} \\
&\qquad \leq L^2 \sum_{i,j,k\in[2p]:(i\neq j)\wedge(i\neq k)} \|a_j\|\|a_k\| \\
&\qquad \leq L^2 \sum_{i,j,k\in[2p]:(i\neq j)\wedge(i\neq k)} \frac{\|a_j\|^2 + \|a_k\|^2}{2} \\
&\qquad = \frac{(2p-1)L^2}{2}\sum_{i\in[2p]}\left(\sum_{j\in[2p]:i\neq j}\|a_j\|^2 + \sum_{k\in[2p]:i\neq k}\|a_k\|^2\right) \\
&\qquad = \frac{(2p-1)L^2}{2}\sum_{i\in[2p]} 2(1-\|a_i\|^2) \\
&\qquad = (2p-1)^2L^2 \leq 4p^2L^2.
\end{aligned}$$

Finally, the last two terms in inequality (25) can each be bounded by

$$\begin{aligned}
\sum_{i,j,k\in[2p]:i=j\neq k} \|a_j\|\|a_k\|\|\nabla_{v_i}\nabla_{v_k}L\|_{op}\|\nabla^2_{v_i}L\|_{op} &\leq (2L/h)\cdot L \sum_{j,k\in[2p]:j\neq k}\|a_j\|\|a_k\| \\
&= 2pL^2 \sum_{j,k\in[2p]:j\neq k}\|a_j\|\|a_k\| \\
&\leq 2pL^2 \sum_{j,k\in[2p]:j\neq k}\frac{\|a_j\|^2+\|a_k\|^2}{2} \\
&= pL^2\sum_{j\in[2p]}\left(\|a_j\|^2+\sum_{k\in[2p]:k\neq j}\|a_k\|^2\right) \\
&= pL^2\sum_{i\in[2p]}\left(\|a_j\|^2+(1-\|a_j\|^2)\right) \\
&= 2p^2L^2.
\end{aligned}$$

The bounds on these four terms combined with inequality (25) tells us that

$$\left\|(\nabla^2 L)a\right\|^2 \leq 12L^2p^2.$$

Taking square roots along with the definition of the operator norm in equation (21) completes the proof. ■

B.3 Proof of Lemma 7

Lemma 7. *For any weight matrix* $V\in\mathbb{R}^{2p\times(d+1)}$, $\|\nabla_V L\|\leq\sqrt{2p}\min\{L(V),1\}$.

Proof Recall the definition of $g_s=(1+\exp(y_s f_V(x_s)))^{-1}$. By using the expression for the gradient of the loss

$$\begin{aligned}
\|\nabla_V L\|^2 &= \sum_{i=1}^{2p}\|\nabla_{v_i}L(V)\|^2 \\
&= \frac{1}{n^2}\sum_{i=1}^{2p}\left\|\sum_{s=1}^{n} g_s\phi'(v_i\cdot x_s)y_s x_s\right\|^2 \\
&= \frac{1}{n^2}\sum_{i=1}^{2p}\sum_{s=1}^{n}\sum_{q=1}^{n} g_s g_q\phi'(v_i\cdot x_s)\phi'(v_i\cdot x_q)y_s y_q x_s\cdot x_q.
\end{aligned}$$

By definition we know that $|\phi'(v_i\cdot x_s)|<1$ for all $s\in[n]$ and $|y_s y_q x_s\cdot x_q|\leq 1$ for any pair $s,q\in[n]$. Therefore,

$$\|\nabla_V L\|^2 \leq \frac{1}{n^2}\sum_{i=1}^{2p}\sum_{s=1}^{n}\sum_{q=1}^{n} g_s g_q = \frac{2p}{n^2}\sum_{s=1}^{n}\sum_{q=1}^{n} g_s g_q.$$

Since $g_k, g_\ell\leq 1$, this implies $\|\nabla_V L\|^2\leq 2p$.

To get the stronger bound when $L(V)$ is small, by Lemma 20, Part 1 we know that $g_s g_q \leq L_s L_q$. Thus,

$$\|\nabla_V L\|^2 \leq \frac{2p}{n^2}\left(\sum_{s=1}^{n}\sum_{q=1}^{n} L_s L_q\right) = 2p\left(\frac{1}{n}\sum_{s=1}^{n} L_s\right)^2 = 2pL(V)^2$$

completing the proof. ■

B.4 Proof of Lemma 8

Lemma 8. *If* $\alpha_t L_t \leq \frac{1}{30p}$, *then* $L_{t+1} \leq L_t - \frac{5\alpha_t\|\nabla L_t\|^2}{6}$.

Proof In order to apply Lemma 5, we would like to bound $\|\nabla^2_W L\|_{op}$ for all convex combinations W of $V^{(t)}$ and $V^{(t+1)}$. For $N = \left\lceil \frac{\sqrt{2p}\|V^{(t+1)}-V^{(t)}\|}{L_t} \right\rceil$, we will prove the following by induction:

> For all $s \in \{0,\ldots,N\}$, for all $\eta \in [0, s/N]$, for $W = \eta V^{(t+1)} + (1-\eta)V^{(t)}$, $\|\nabla^2_W L\|_{op} \leq 10pL_t$.

The base case, where $s = 0$ follows directly from Lemma 6. Now, assume that the inductive hypothesis holds from some s, and, for $\eta \in (s/N, (s+1)/N]$, consider $W = \eta V^{(t+1)} + (1-\eta)V^{(t)}$. Let $\widetilde{W} = (s/N)V^{(t+1)} + (1-s/N)V^{(t)}$. Applying Lemma 5 along with the inductive hypothesis, $L(\widetilde{W}) \leq L_t$. Applying Lemma 7,

$$\begin{aligned} L(W) &\leq L(\widetilde{W}) + (\sqrt{2p})\|W - \widetilde{W}\| \\ &\leq L_t + \frac{\sqrt{2p}\|V^{(t+1)} - V^{(t)}\|}{N} \\ &\leq 2L_t. \end{aligned}$$

Applying Lemma 6, this implies $\|\nabla^2_W L\|_{op} \leq 10pL(V^{(t)})$, completing the proof of the inductive step.

So, now we know that, for all convex combinations W of $V^{(t)}$ and $V^{(t+1)}$, $\|\nabla^2_W L\|_{op} \leq 10pL(V^{(t)})$. Applying Lemma 5, we have

$$L_{t+1} \leq L_t - \frac{5\alpha_t\|\nabla L_t\|^2}{6},$$

which is the desired result. ■

B.5 Proof of Lemma 9

Lemma 9. *If* $\psi : (0, M] \to \mathbb{R}$ *is a continuous, concave function such that* $\lim_{\eta\to 0^+}\psi(\eta)$ *exists. Then the infimum of* $\sum_{i=1}^n \psi(z_i)$ *subject to* $z_1,\ldots,z_n > 0$ *and* $\sum_{i=1}^n z_i = M$ *is* $\psi(M) + (n-1)\lim_{\eta\to 0^+}\psi(\eta)$.

Proof If $n = 1$ the lemma is trivial. Consider the case $n > 1$. Consider an arbitrary feasible point $z_1, \ldots, z_n$ with $z_1, \ldots, z_n > 0$ and $\sum_{i=1}^n z_i = M$. Assume without loss of generality that $z_1 \geq z_2 \geq \ldots \geq z_n$. For an arbitrarily small $\eta > 0$, we claim that the point $z_1 + z_2 - \eta, \eta, z_3, \ldots, z_n$ is at least as good. Since ψ is concave

$$\psi(z_1) \geq \frac{z_2 - \eta}{z_1 + z_2 - 2\eta}\psi(\eta) + \frac{z_1 - \eta}{z_1 + z_2 - 2\eta}\psi(z_1 + z_2 - \eta), \quad \text{and}$$
$$\psi(z_2) \geq \frac{z_1 - \eta}{z_1 + z_2 - 2\eta}\psi(\eta) + \frac{z_2 - \eta}{z_1 + z_2 - 2\eta}\psi(z_1 + z_2 - \eta).$$

So by adding these two inequalities we infer

$$\psi(z_1 + z_2 - \eta) + \psi(\eta) + \sum_{i=3}^n \psi(z_i) \leq \psi(z_1) + \psi(z_2) + \sum_{i=3}^n \psi(z_i).$$

Repeating this for the other $(n-2)$ components of the solution, we find that

$$\psi(M - (n-1)\eta) + (n-1)\psi(\eta) \leq \sum_{i=1}^n \psi(z_i).$$

Since ψ is a continuous function by taking the limit $\eta \to 0^+$ we get that,

$$\psi(M) + (n-1)\lim_{\eta \to 0^+} \psi(\eta) \leq \sum_{i=1}^n \psi(z_i).$$

Given that $z_1, \ldots, z_n$ was an arbitrary feasible point, the previous inequality establishes our claim. ■

C Reduction to the Case of No Bias with Random Initialization

We once again consider the process of training a model using $(\tilde{x}_1, y_1), \ldots, (\tilde{x}_n, y_n)$, where $\tilde{x}_s$ is defined as in Appendix A.

Let $\widetilde{\sigma} = \sqrt{2}\sigma$. A sample from $\mathcal{N}(0, \widetilde{\sigma}^2)$ can be generated by sampling from $\mathcal{N}(0, \sigma^2)$, and scaling the result up by a factor of $\sqrt{2}$.

For some $\sigma > 0$, and $\alpha_0, \alpha_1, \alpha_2, \ldots > 0$, $h \geq 0$, consider the joint distribution on

$$(\theta^{(0)}, \widetilde{V}^{(0)}), (\theta^{(1)}, \widetilde{V}^{(1)}), \ldots$$

defined as follows. First, $\theta^{(0)}, \theta^{(1)}, \ldots$ are generated as described in Section 2.4. Each row $\tilde{v}_i^{(0)}$ of $\widetilde{V}^{(0)} \in \mathbb{R}^{2p \times (d+1)}$ is $\sqrt{2}(v_{i1}^{(0)}, \ldots, v_{id}^{(0)}, b_i^{(0)})$ (so that they are mutually independent draws from $\mathcal{N}(0, 2\sigma^2)$).

Define $\widetilde{L}$ as in Appendix A: informally, L, but without the bias terms.

Then, for $\tilde{\alpha}_0 = 2\alpha_0, \tilde{\alpha}_1 = 2\alpha_1, \ldots > 0$, we define $\widetilde{V}^{(1)}, \widetilde{V}^{(2)}, \ldots$ to be the iterates of gradient descent applied to $\widetilde{L}$, except replacing $\alpha_0, \alpha_1, \alpha_2, \ldots$ by $\tilde{\alpha}_0, \tilde{\alpha}_1, \tilde{\alpha}_2, \ldots$.

Arguing as in Appendix A, except starting with round 0, we can see that, for all t,

- for all i, $\tilde{v}_i^{(t)} = \sqrt{2}(v_{i1}^{(t)}, \ldots, v_{id}^{(t)}, b_i^{(t)})$

- for all i and all s, $\tilde{v}_i^{(t)} \cdot \tilde{x}_s = v_i^{(t)} \cdot x_s + b_i^{(t)}$, so that $\widetilde{L}(\widetilde{V}^{(t)}) = L(\theta^{(t)})$.

For each cluster k, define $\tilde{\mu}_k$ by $\tilde{\mu}_k = \left(\frac{\mu_{k1}}{\sqrt{2}}, \ldots, \frac{\mu_{kd}}{\sqrt{2}}, \frac{1}{\sqrt{2}}\right)$. Note that $\|\tilde{x}_s - \tilde{\mu}_k\| = \frac{\|x_s - \mu_k\|}{\sqrt{2}}$ and, for all clusters k and ℓ

$$\tilde{\mu}_k \cdot \tilde{\mu}_\ell = \frac{\mu_k \cdot \mu_\ell + 1}{2}.$$

D Proof of Lemma 15

We begin by restating the lemma here.

Lemma 15. *There exists a real-valued function χ such that, for all $(x_1, y_1), \ldots, (x_n, y_n) \in \mathbb{S}^{d-1} \times \{-1, 1\}$, for all small enough C_2 and all large enough C_3, with probability $1 - \delta$ over the draw of $V^{(0)}$, all of the following hold.*

1. *For all $s \in [n]$,*

$$\sum_{i \in I_{+s}} v_i^{(0)} \cdot x_s \geq p\chi(h, \alpha_0, \sigma) - 1, \text{ and}$$
$$\sum_{i \in I_{-s}} v_i^{(0)} \cdot x_s \leq p\chi(h, \alpha_0, \sigma) + 1.$$

2. *For all samples $s \in [n]$,*

$$(1/2 - o(1))p \leq |I_{+s}| \leq (1/2 + o(1))p, \text{ and}$$
$$(1/2 - o(1))p \leq |I_{-s}| \leq (1/2 + o(1))p.$$

3. *For all samples $s \in [n]$*

$$\frac{1}{2} - o(1) \leq g_{0s} \leq \frac{1}{2} + o(1).$$

4. *For all clusters $k \in [4]$,*

$$|\{i \in [2p] : \forall\, s \in \mathcal{K}_k,\ i \in I_{+s})\}| \geq \left(\frac{1 - \sqrt{r}}{2} - o(1)\right) p, \text{ and}$$
$$|\{i \in [2p] : \forall\, s \in \mathcal{K}_k,\ i \in I_{-s})\}| \geq \left(\frac{1 - \sqrt{r}}{2} - o(1)\right) p.$$

5. *For all pairs $s, q \in [n]$ such that $y_s \neq y_q$,*

$$\left|\left\{i \in [2p] : (i \in I_{+s}) \wedge \left(v_i^{(0)} \cdot x_q \geq 0\right)\right\}\right| \leq \left(\frac{1}{3} + \frac{\Delta}{4} + r + o(1)\right) p, \text{ and}$$
$$\left|\left\{i \in [2p] : (i \in I_{-s}) \wedge \left(v_i^{(0)} \cdot x_q \geq 0\right)\right\}\right| \leq \left(\frac{1}{3} + \frac{\Delta}{4} + r + o(1)\right) p.$$

6. *For all samples $s \in [n]$,*

$$\left|\left\{i \in [2p] : \left(-\alpha_0\left(\frac{1}{2} + 2(\Delta + r)\right) \le v_i^{(0)} \cdot x_s \le h + 4\alpha_0\right) \wedge (u_i \neq y_s)\right\}\right| \le \left(\frac{\sqrt{2}}{\sigma\sqrt{\pi}}\left(h + 5\alpha_0(2 + \Delta + r)\right)\right) p.$$

7. *The norm of the weight matrix after one iteration satisfies $\frac{3}{5}\sqrt{\frac{d}{p^\beta}} \le \|V^{(1)}\| \le 3\sqrt{\frac{d}{p^\beta}}$.*

The different parts of the lemma are proved one at a time in the subsections below. The lemma holds by taking a union bound over all the different parts. Throughout the proof of this lemma we fix the samples $(x_1, y_1), \ldots, (x_n, y_n) \in \mathbb{S}^{d-1} \times \{-1, 1\}$. Conditioned on their value, for all $i \in [2p]$ and for all $s \in [n]$, the random variables $v_i^{(0)} \cdot x_s \sim \mathcal{N}(0, \sigma^2)$.

D.1 Proof of Part 1

Consider a fixed sample s. Without loss of generality, suppose that $y_s = 1$. We want to demonstrate a high probability lower bound on

$$\sum_{i \in I_{+s}} v_i^{(0)} \cdot x_s = \sum_{i \in [p]} (v_i^{(0)} \cdot x_s) \mathbf{1}_{i \in I_{+s}}.$$

Now the expected value of this sum,

$$\mathbb{E}\left[\sum_{i \in [p]} (v_i^{(0)} \cdot x_s) \mathbf{1}_{i \in I_{+s}}\right] = p\mathbb{E}\left[\left(v_1^{(0)} \cdot x_s\right) \mathbf{1}_{1 \in I_{+s}}\right].$$

Choose the function χ in the statement of the result to be

$$\chi(h, \alpha_0, \sigma) := \mathbb{E}\left[\left(v_1^{(0)} \cdot x_s\right) \mathbf{1}_{1 \in I_{+s}}\right].$$

By applying Hoeffding's inequality [see Ver18, Theorem 2.6.2] (since $v_i^{(0)} \cdot x_s \sim \mathcal{N}(0, \sigma^2)$, the truncated random variable $(v_i^{(0)} \cdot x_s)\mathbf{1}_{i \in I_{+s}}$ is also $c_1\sigma$-sub-Gaussian for an appropriate positive constant c_1),

$$\mathbb{P}\left[\sum_{i \in I_{+s}} v_i^{(0)} \cdot x_s \le \mathbb{E}\left[\sum_{i \in I_{+s}} (v_i^{(0)} \cdot x_s)\right] - \eta p\right] \le \exp\left(-c_2\eta^2 p/\sigma^2\right) = \exp\left(-c_2\eta^2 p^{2+\beta}\right)$$

since $\sigma = 1/p^{1/2+\beta/2}$. Setting $\eta = 1/p$ we get

$$\mathbb{P}\left[\sum_{i \in I_{+s}} v_i^{(0)} \cdot x_s \le \mathbb{E}\left[\sum_{i \in I_{+s}} (v_i^{(0)} \cdot x_s)\right] - 1\right] \le \exp\left(-c_2 p^\beta\right).$$

Since $p \geq \log^{C_3}(nd/\delta)$, for any $C_3 \geq c_3/\beta$, for a large enough constant c_3 we can establish that

$$\mathbb{P}\left[\sum_{i\in I_{+s}} v_i^{(0)} \cdot x_s \leq \mathbb{E}\left[\sum_{i\in I_{+s}} (v_i^{(0)} \cdot x_s)\right] - 1\right] \leq \frac{\delta}{20n}.$$

Finally, a union bound over all n samples completes the proof for I_{+s}:

$$\mathbb{P}\left[\exists s \in [n] : \sum_{i\in I_{+s}} v_i^{(0)} \cdot x_s \leq \mathbb{E}\left[\sum_{i\in I_{+s}} (v_i^{(0)} \cdot x_s)\right] - 1\right] \leq \frac{\delta}{20}.$$

An identical argument holds for the sum: $\sum_{i\in I_{-s}} v_i^{(0)} \cdot x_s$ which completes the proof of this part of the lemma.

D.2 Proof of Part 2

Consider a sample s. Without loss of generality, suppose that $y_s = 1$. Recall the definition of the set

$$I_{+s} = \{i \in [2p] : v_i^{(0)} \cdot x_s \geq h + 4\alpha_0 \text{ and } u_i = y_s = 1\}.$$

Note that the variable $v_i^{(0)} \cdot x_s$ has a Gaussian distribution with zero-mean and variance σ^2. Also, recall that $u_i = 1$ for all $i \in [p]$. Therefore, for each $i \in [p]$,

$$\begin{aligned}
\zeta := \mathbb{P}\left[v_i^{(0)} \cdot x_s \geq h + 4\alpha_0\right] &= \frac{1}{2} - \mathbb{P}\left[v_i^{(0)} \cdot x_s \in [0, h + 4\alpha_0]\right] \\
&\overset{(i)}{=} \frac{1}{2} - O\left(\frac{h + \alpha_0}{\sigma}\right) \\
&= \frac{1}{2} - O\left(\frac{1/p + 1/p^{1/2+\beta}}{1/p^{1/2+\beta/2}}\right) \\
&= \frac{1}{2} - o(1),
\end{aligned}$$

where (i) follows by an upper bound of $1/(\sigma\sqrt{2\pi})$ on the density of a Gaussian random variable with variance σ^2. A Hoeffding bound implies that, for any $\eta > 0$

$$\mathbb{P}\left[\left|\, |I_{+s}| - \zeta p \right| \geq \eta p\right] \leq 2\exp\left(-c'\eta^2 p\right).$$

Thus by a union bound over all samples

$$\mathbb{P}\left[\exists s \in [n] : \left||I_{+s}| - \zeta p\right| \geq \eta p\right] \leq 2n\exp\left(-c'\eta^2 p\right).$$

Setting $\eta = 1/p^{1/4}$ and recalling that $\zeta = 1/2 - o(1)$ and $p \geq \log^{C_3}(nd/\delta)$ completes the argument for the sets I_{+s}. An identical argument goes through for the second claim that establishes a bound on the size of the sets I_{-s}.

D.3 Proof of Part 3

By definition $g_{0s} = \left(1 + \exp\left(y_s \sum_{i=1}^{2p} u_i \phi(v_i^0 \cdot x_s)\right)\right)^{-1}$. Recall that $v_i^{(0)}$ is drawn from a zero-mean Gaussian with variance $\sigma^2 I$. Therefore, for each i, $v_i^{(0)} \cdot x_s$ is a zero-mean Gaussian with variance σ^2 (since $\|x_s\| = 1$). For ease of notation let us define $\xi_i := v_i^{(0)} \cdot x_s$. The sigmoid function $1/(1+\exp(t))$ is 1-Lipschitz. Therefore,

$$\left| \frac{1}{1 + \exp\left(y_s \sum_{i=1}^{2p} u_i \phi(\xi_i)\right)} - \frac{1}{1 + \exp\left(y_s \mathbb{E}\left[\sum_{i=1}^{2p} u_i \phi(\xi_i)\right]\right)} \right| \leq \left| \sum_{i=1}^{2p} u_i \phi(\xi_i) - \mathbb{E}\left[\sum_{i=1}^{2p} u_i \phi(\xi_i)\right] \right|.$$

Additionally, by its definition the Huberized ReLU ϕ is also 1-Lipschitz. Therefore for any pair $z_1, z_2 \in \mathbb{R}^{2p}$

$$\left| \sum_{i=1}^{2p} u_i \left(\phi(z_{1i}) - \phi(z_{2i})\right) \right| \leq \sum_{i=1}^{2p} |\phi(z_{1i}) - \phi(z_{2i})| \leq \sum_{i=1}^{2p} |z_{1i} - z_{2i}| = \|z_1 - z_2\|_1 \leq \sqrt{2p}\|z_1 - z_2\|.$$

Hence, the function $y_s \sum_{i=1}^{2p} u_i \phi(\xi_i)$ is $\sqrt{2p}$-Lipschitz with respect to its argument

$$(\xi_1, \ldots, \xi_{2p}).$$

By the Borell-Tsirelson-Ibragimov-Sudakov inequality for the concentration of Lipschitz functions of Gaussian random variables [see Wai19, Theorem 2.4],

$$\mathbb{P}\left[\left|\sum_{i=1}^{2p} u_i \phi(\xi_i) - \mathbb{E}\left[\sum_{i=1}^{2p} u_i \phi(\xi_i)\right]\right| \geq \eta\right] \leq 2\exp\left(-\frac{\eta^2}{4p\sigma^2}\right).$$

Recall that $\sigma = 1/p^{1/2+\beta/2}$, thus,

$$\mathbb{P}\left[\left|\sum_{i=1}^{2p} u_i \phi(\xi_i)) - \mathbb{E}\left[\sum_{i=1}^{2p} u_i \phi(\xi_i))\right]\right| \geq \eta\right] \leq 2\exp\left(-c_1 p^{\beta} \eta^2\right).$$

By choosing $\eta = 1/p^{\beta/4}$,

$$\mathbb{P}\left[\left|\sum_{i=1}^{2p} u_i \phi(\xi_i)) - \mathbb{E}\left[\sum_{i=1}^{2p} u_i \phi(\xi_i))\right]\right| \geq \frac{1}{p^{\beta/4}}\right] \leq 2\exp\left(-c_2 p^{\beta/2}\right).$$

This tells us that with probability at least $1 - 2\exp(-c_2 p^{\beta/2})$,

$$\frac{1}{1 + \exp\left(y_s \mathbb{E}\left[\sum_{i=1}^{2p} u_i \phi(\xi_i))\right]\right)} - \frac{1}{p^{\beta/4}} \leq \frac{1}{1 + \exp\left(y_s \sum_{i=1}^{2p} u_i \phi(\xi_i))\right)} \leq \frac{1}{1 + \exp\left(y_s \mathbb{E}\left[\sum_{i=1}^{2p} u_i \phi(\xi_i))\right]\right)} + \frac{1}{p^{\beta/4}}. \tag{26}$$

Next, we calculate the value of $\mathbb{E}\left[\sum_{i=1}^{2p} u_i\phi(\xi_i))\right]$. Note that all the random variables $\{\xi_i\}_{i\in[2p]}$ are identically distributed. Recall that, $u_i = 1$ if $i \in \{1,\dots,p\}$ and $u_i = -1$ if $i \in \{p+1,\dots,2p\}$, thus

$$\mathbb{E}\left[\sum_{i=1}^{2p} u_i\phi(\xi_i)\right] = \mathbb{E}\left[\sum_{i=1}^{p}\phi(\xi_i)\right] - \mathbb{E}\left[\sum_{i=p+1}^{2p}\phi(\xi_i)\right] = p\mathbb{E}\left[\phi(\xi_1)\right] - p\mathbb{E}[\phi(\xi_1)] = 0.$$

Thus by inequality (26) we know that with probability at least $1 - 2\exp(-c_2 p^{\beta/2})$

$$\frac{1}{2} - o(1) \le \frac{1}{1+\exp(y_s\sum_{i=1}^{2p} u_i\phi(v_i^{(0)}\cdot x_s))} \le \frac{1}{2} + o(1).$$

A union bound over all n samples completes the proof, since $p \ge \log^{C_3}(nd/\delta)$ for a large enough constant C_3.

D.4 Proof of Part 4

We will first prove the first claim of this part of the lemma. Without loss of generality consider the cluster $\mathcal{K}_1$ (recall that for all examples $s \in \mathcal{K}_1$, $y_s = 1$). For any pair $s, q \in \mathcal{K}_1$

$$\begin{aligned}
&\mathbb{P}\left[v_i^{(0)}\cdot x_s \ge h + 4\alpha_0 \text{ and } v_i^{(0)}\cdot x_q \ge h + 4\alpha_0\right] \\
&\ge \mathbb{P}\left[v_i^{(0)}\cdot x_s \ge 0 \text{ and } v_i^{(0)}\cdot x_q \ge 0\right] - \mathbb{P}[v_i^{(0)}\cdot x_s \in [0, h+4\alpha_0]] - \mathbb{P}[v_i^{(0)}\cdot x_q \in [0, h+4\alpha_0]] \\
&\overset{(i)}{\ge} \mathbb{P}\left[v_i^{(0)}\cdot x_s \ge 0 \text{ and } v_i^{(0)}\cdot x_q \ge 0\right] - O\left(\frac{h+\alpha_0}{\sigma}\right) \\
&= \mathbb{P}\left[v_i^{(0)}\cdot x_s \ge 0 \text{ and } v_i^{(0)}\cdot x_q \ge 0\right] - O\left(\frac{1/p + 1/p^{1/2+\beta}}{1/p^{1/2+\beta/2}}\right) \\
&\overset{(ii)}{=} \frac{1 - \arccos(x_s\cdot x_q)/\pi}{2} - o(1) \\
&\overset{(iii)}{\ge} \frac{1 - \arccos(1-2r)/\pi}{2} - o(1) \\
&\ge \frac{1-\sqrt{r}}{2} - o(1)
\end{aligned}$$

where (i) follows by an upper bound of $1/(\sqrt{2\pi}\sigma)$ on the density of a Gaussian random variable, (ii) follows by noting that the conditional probability of $v_i^{(0)}\cdot x_q \ge 0$ conditioned on the event that $v_i^{(0)}\cdot x_s \ge 0$ is $1 - \frac{\arccos(x_s\cdot x_q)}{\pi}$, and (iii) follows since $x_s\cdot x_q \ge 1 - 2r$ by Lemma 21.

Define $\zeta := \mathbb{P}\left[v_i^{(0)}\cdot x_s \ge h + 4\alpha_0 \text{ and } v_i^{(0)}\cdot x_q \ge h + 4\alpha_0\right]$. A Hoeffding bound implies that, for any $\eta > 0$

$$\mathbb{P}\left[\left|\left\{i\in[p] : \left(v_i^{(0)}\cdot x_s \ge h + 4\alpha_0\right) \wedge \left(v_i^{(0)}\cdot x_q \ge h + 4\alpha_0\right)\right\}\right| \le (\zeta-\eta)p\right] \le \exp\left(-c_1\eta^2 p\right).$$

Recall the definition of the set $I_{+s} = \left\{i \in [2p] : (u_i = y_s = 1) \wedge \left(v_i^{(0)}\cdot x_s \ge h + 4\alpha_0\right)\right\}$. Therefore, a union bound over all sample pairs $s, q \in \mathcal{K}_1$ implies that

$$\mathbb{P}\left[\left|\{i\in[p] : \forall\, s\in\mathcal{K}_1, i \in I_{+s}\}\right| \le (\zeta-\eta)p\right] \le n^2\exp\left(-c_1\eta^2 p\right).$$

Finally, by taking a union bound over all 4 clusters we get that

$$\mathbb{P}\left[\exists k \in [4] : |\{i \in [2p] : \forall\, s \in \mathcal{K}_k, i \in I_{+s}\}| \leq (\zeta - \eta)p\right] \leq 4n^2 \exp\left(-c_1 \eta^2 p\right).$$

Choosing $\eta = 1/p^{1/4}$, recalling that $\zeta \geq (1 - \sqrt{r} - o(1))/2$ and $p \geq \log^{C_3}(nd/\delta)$ for a large enough constant C_3 completes the proof of the first claim.

The second claim of this part of the lemma follows by an identical argument.

D.5 Proof of Part 5

Without loss of generality, consider a node $i \in [p]$ with $u_i = 1$ and a fixed pair $s, q \in [n]$ such that $y_s = 1$ and $y_q = -1$. Since each of $v_i^{(0)} \cdot x_s$ and $v_i^{(0)} \cdot x_q$ are distributed as $\mathcal{N}(0, \sigma^2)$, we have

$$\begin{aligned}
\mathbb{P}[v_i^{(0)} \cdot x_s \geq h + 4\alpha_0 \text{ and } v_i^{(0)} \cdot x_q \geq 0] &\leq \mathbb{P}[v_i^{(0)} \cdot x_s \geq 0 \text{ and } v_i^{(0)} \cdot x_q \geq 0] \\
&\overset{(i)}{=} \frac{1 - \frac{\arccos(x_s \cdot x_q)}{\pi}}{2} \\
&\overset{(ii)}{\leq} \frac{1 - \frac{\arccos((1+\Delta)/2 + 2r)}{\pi}}{2} \\
&\leq \frac{1}{3} + \frac{\Delta}{4} + r,
\end{aligned}$$

if the bound C_2 on r and Δ is small enough, where (i) follows by noting that the conditional probability of $v_i^{(0)} \cdot x_q \geq 0$ conditioned on the event that $v_i^{(0)} \cdot x_s \geq 0$ is $1 - \frac{\arccos(x_s \cdot x_q)}{\pi}$, while (ii) follows since by Lemma 21, $x_s \cdot x_q \leq (1 + \Delta)/2 + 2r$ for samples where $y_s \neq y_q$.

Now, define $\zeta := \mathbb{P}[v_i^0 \cdot x_s \geq h + 4\alpha_0 \text{ and } v_i^{(0)} \cdot x_q \geq 0]$; a Hoeffding bound implies that, for any $\eta > 0$

$$\mathbb{P}[|\{i \in [p] : v_i^{(0)} \cdot x_s \geq h + 4\alpha_0 \text{ and } v_i^{(0)} \cdot x_q \geq 0]\}| \geq (\zeta + \eta)p] \leq \exp(-c'\eta^2 p).$$

Choosing $\eta = 1/p^{1/4}$ and recalling that $\zeta \leq \frac{1}{3} + \frac{\Delta}{4} + r$ along with a union bound over the pairs of samples completes the proof of the first claim. An identical argument works to establish the second claim.

D.6 Proof of Part 6

Without loss of generality, consider a node $i \in [p+1, \ldots, 2p]$ with $u_i = -1$ and fix a sample s with $y_s = 1$. Since each $v_i^0 \cdot x_s$ is distributed as $\mathcal{N}(0, \sigma^2)$, we have,

$$\zeta := \mathbb{P}\left[-\alpha_0(1/2 + 2(\Delta + r)) \leq v_i^{(0)} \cdot x_s \leq h + 4\alpha_0\right] \leq \frac{1}{\sigma\sqrt{2\pi}}\left(h + 5\alpha_0(1 + \Delta + r)\right), \quad (27)$$

where the bound on the probability above follows by an upper bound of $1/(\sqrt{2\pi}\sigma)$ on the density of a Gaussian random variable. A Hoeffding bound implies that, for any $\eta > 0$

$$\begin{aligned}
&\mathbb{P}\left[|\{i \in \{p+1, \ldots, 2p\} : -\alpha_0(1/2 + 2(\Delta + r)) \leq v_i^{(0)} \cdot x_s \leq h + 4\alpha_0\}| \geq (\zeta + \eta)p\right] \\
&\leq \exp\left(-c'\eta^2 p\right).
\end{aligned}$$

By choosing $\eta = \frac{5\alpha_0}{\sigma\sqrt{2\pi}} = \frac{5}{\sqrt{2\pi}p^{\beta/2}}$, recalling the upper bound on ζ established in (27) and a union bound over all the n samples completes the proof since $p \geq \log^{C_3}(nd/\delta)$ for a large enough constant C_3.

D.7 Proof of Part 7

We know that each $v_i^{(0)} \sim \mathcal{N}(0, \sigma^2 I_{(d+1)\times(d+1)})$. Thus by a concentration inequality for the lower tail of a χ^2-random variable with $2(d+1)p$ degrees of freedom [see LM00, Lemma 1] we have that, for any $\eta > 0$

$$\mathbb{P}\left[\frac{\|V^{(0)}\|}{\sqrt{2(d+1)p}} \leq \sigma\sqrt{1-\eta}\right] \leq \exp(-(d+1)p\eta^2/2).$$

Recall that $\sigma = 1/p^{1/2+\beta/2}$, thus by setting $\eta = 1/32$ we get that

$$\mathbb{P}\left[\|V^{(0)}\| \leq \frac{6\sqrt{d+1}}{5p^{\beta/2}}\right] \leq \exp(-c_1(d+1)p).$$

Since $p \geq \log^{C_3}(nd/\delta)$ for a large enough value of C_3, this ensures that

$$\|V^{(0)}\| \geq 6\sqrt{d+1}/(5p^{\beta/2})$$

with probability at least $1 - \delta/c_2$. By the reverse triangle inequality,

$$\|V^{(1)}\| \geq \|V^{(0)}\| - \alpha_0\|\nabla L_0\| \overset{(i)}{\geq} \|V^{(0)}\| - \alpha_0\sqrt{2p} \geq \frac{6\sqrt{d+1}}{5p^{\beta/2}} - \frac{\sqrt{2}}{p^{\beta}} \overset{(ii)}{\geq} \frac{3}{5}\sqrt{\frac{d}{p^{\beta}}},$$

where (i) follows by the bound on the norm of gradient established in Lemma 7 and (ii) follows since $d \geq 2$ under our clustering assumptions. Hence

$$\mathbb{P}\left[\|V^{(1)}\| \geq \frac{3}{5}\sqrt{\frac{d}{p^{\beta}}}\right] \geq 1 - \delta/c_2, \tag{28}$$

which establishes the desired lower bound on the norm of $V^{(1)}$. To establish the upper bound we will use the Borell-TIS inequality for Lipschitz functions of Gaussian random variables [see Wai19, Example 2.28]. By this inequality we have that, for any $\eta > 0$

$$\mathbb{P}\left[\frac{\|V^{(0)}\|}{\sqrt{2(d+1)p}} \geq \sigma(1+\eta)\right] \leq \exp(-(d+1)p\eta^2).$$

Once again because $\sigma = 1/p^{1/2+\beta/2}$, by setting $\eta = 1/32$ we get that

$$\mathbb{P}\left[\|V^{(0)}\| \geq \frac{3\sqrt{d+1}}{2p^{\beta/2}}\right] \leq \exp(-c_1(d+1)p).$$

Since $p \geq \log^{C_3}(nd/\delta)$ for a large enough value of C_3, this ensures that

$$\|V^{(0)}\| \leq 3\sqrt{d+1}/(2p^{\beta/2}) \leq 5\sqrt{d}/2p^{\beta/2}$$

with probability at least $1 - \delta/c_2$. By the triangle inequality,

$$\|V^{(1)}\| \leq \|V^{(0)}\| + \alpha_0\|\nabla L_0\| \overset{(i)}{\leq} \|V^{(0)}\| + \alpha_0\sqrt{2p} \leq \frac{5\sqrt{d}}{2p^{\beta/2}} + \frac{\sqrt{2}}{p^{\beta}} \leq 3\sqrt{\frac{d}{p^{\beta}}},$$

where (i) follows by the bound on the norm of gradient established in Lemma 7. Hence

$$\mathbb{P}\left[\|V^{(1)}\| \leq 3\sqrt{\frac{d}{p^\beta}}\right] \geq 1 - \delta/c_2.$$

Combining this with inequality (28) above we get that

$$\mathbb{P}\left[\frac{3}{5}\sqrt{\frac{d}{p^\beta}} \leq \|V^{(1)}\| \leq 3\sqrt{\frac{d}{p^\beta}}\right] \geq 1 - 2\delta/c_2$$

completing our proof.

E Auxiliary Lemmas

In this section we list a couple of lemmas that are useful in various proofs above.

Lemma 20. *For any $x \in \mathbb{R}^d$ and $y \in \{-1, 1\}$ and any weight matrix V we have the following*

1. $$\frac{1}{1 + \exp\left(y f_V(x)\right)} \leq \log(1 + \exp(-y f_V(x))) = L(V; x, y).$$

2. $$\frac{\exp\left(y f_V(x)\right)}{\left(1 + \exp\left(y f_V(x)\right)\right)^2} \leq \frac{1}{1 + \exp\left(y f_V(x)\right)} \leq L(V; x, y).$$

Proof Part 1 follows since for any $z \in \mathbb{R}$, we have the inequality $(1 + \exp(z))^{-1} \leq \log(1 + \exp(-z))$.

Part 2 follows since for any $z \in \mathbb{R}^d$, we have the inequality $\exp(z)/\left(1 + \exp(z)\right)^2 \leq (1 + \exp(z))^{-1}$. ■

Lemma 21. *Given an $r < 1$ suppose that for any $k \in [4]$ all samples $s \in \mathcal{K}_k$ satisfy the bound $\|x_s - \mu_k\| \leq r/\sqrt{2}$ and for all $k \in [4]$, $\|\mu_k\| = 1$.*

1. *Then for any pair of clusters $\mathcal{K}_k, \mathcal{K}_\ell$ such that $y_k \neq y_\ell$, and $\mu_k \cdot \mu_\ell \leq (1 + \Delta)/2$ we have, for all $s \in \mathcal{K}_k$ and $q \in \mathcal{K}_\ell$*

$$x_s \cdot x_q \leq \frac{1 + \Delta}{2} + 2r.$$

2. *Given a cluster $\mathcal{K}_k$, if $s, s' \in \mathcal{K}_k$ then,*

$$x_s \cdot x_{s'} \geq 1 - 2r.$$

Proof *Proof of Part 1:* By evaluating the inner product and applying the Cauchy-Schwarz inequality

$$\begin{aligned} x_s \cdot x_q &= (x_s - \mu_k + \mu_k) \cdot (x_q - \mu_\ell + \mu_\ell) \\ &= \mu_k \cdot \mu_\ell + (x_s - \mu_k) \cdot \mu_\ell + \mu_k \cdot (x_q - \mu_\ell) + (x_s - \mu_k) \cdot (x_q - \mu_\ell) \\ &\leq \frac{1 + \Delta}{2} + \sqrt{2}r + \frac{r^2}{2} \\ &\leq \frac{1 + \Delta}{2} + 2r. \end{aligned}$$

Proof of Part 2: Recall that $\|\mu_k\| = 1$. Thus, given two samples $s, s' \in \mathcal{K}_k$,

$$\begin{aligned} x_s \cdot x'_s &= (x_s - \mu_k + \mu_k) \cdot (x_{s'} - \mu_k + \mu_k) \\ &= \mu_k \cdot \mu_k + (x_s - \mu_k) \cdot \mu_k + \mu_k \cdot (x_{s'} - \mu_k) + (x_s - \mu_k) \cdot (x_{s'} - \mu_k) \\ &\geq 1 - \sqrt{2}r - \frac{r^2}{2} \\ &\geq 1 - 2r \end{aligned}$$

as claimed. ■

References

[ALS19] Zeyuan Allen-Zhu, Yuanzhi Li, and Zhao Song. "A convergence theory for deep learning via over-parameterization". In: *International Conference on Machine Learning*. 2019, pp. 242–252 (Cited on page 2).

[And+14] Alexandr Andoni, Rina Panigrahy, Gregory Valiant, and Li Zhang. "Learning polynomials with neural networks". In: *International Conference on Machine Learning*. 2014, pp. 1908–1916 (Cited on page 24).

[Aro+19a] Sanjeev Arora, Nadav Cohen, Wei Hu, and Yuping Luo. "Implicit regularization in deep matrix factorization". In: *Advances in Neural Information Processing Systems*. 2019, pp. 7413–7424 (Cited on page 24).

[Aro+19b] Sanjeev Arora, Simon S Du, Wei Hu, Zhiyuan Li, and Ruosong Wang. "Fine-grained analysis of optimization and generalization for overparameterized two-layer neural networks". In: *International Conference on Machine Learning*. 2019, pp. 322–332 (Cited on page 24).

[BL20] Peter L Bartlett and Philip M Long. "Failures of model-dependent generalization bounds for least-norm interpolation". In: *arXiv preprint arXiv:2010.08479* (2020) (Cited on page 24).

[Bar+20] Peter L Bartlett, Philip M Long, Gábor Lugosi, and Alexander Tsigler. "Benign overfitting in linear regression". In: *Proceedings of the National Academy of Sciences* 117.48 (2020), pp. 30063–30070 (Cited on page 24).

[Bel+19] Mikhail Belkin, Daniel J Hsu, Siyuan Ma, and Soumik Mandal. "Reconciling modern machine-learning practice and the classical bias–variance trade-off". In: *Proceedings of the National Academy of Sciences* 116.32 (2019), pp. 15849–15854 (Cited on pages 1, 24).

[BHM18] Mikhail Belkin, Daniel J Hsu, and Partha Mitra. "Overfitting or perfect fitting? Risk bounds for classification and regression rules that interpolate". In: *Advances in Neural Information Processing Systems*. 2018, pp. 2300–2311 (Cited on page 24).

[BN20] Guy Bresler and Dheeraj Nagaraj. "A corrective view of neural networks: Representation, memorization and learning". In: *Conference on Learning Theory*. 2020, pp. 848–901 (Cited on page 24).

[BG19] Alon Brutzkus and Amir Globerson. "Why do larger models generalize better? A theoretical perspective via the XOR problem". In: *International Conference on Machine Learning*. 2019, pp. 822–830 (Cited on page 24).

[Bru+18] Alon Brutzkus, Amir Globerson, Eran Malach, and Shai Shalev-Shwartz. "SGD learns over-parameterized networks that provably generalize on linearly separable data". In: *International Conference on Learning Representations*. 2018 (Cited on page 24).

[Bub15] Sébastien Bubeck. "Convex optimization: algorithms and complexity". In: *Foundations and Trends® in Machine Learning* 8.3-4 (2015), pp. 231–357 (Cited on page 26).

[CL21] Niladri S Chatterji and Philip M Long. "Finite-sample analysis of interpolating linear classifiers in the overparameterized regime". In: *Journal of Machine Learning Research* 22.129 (2021), pp. 1–30 (Cited on page 24).

[CLB20] Niladri S Chatterji, Philip M Long, and Peter L Bartlett. "When does gradient descent with logistic loss find interpolating two-layer networks?" In: *arXiv preprint arXiv:2012.02409* (2020) (Cited on page 24).

[CLB21] Niladri S Chatterji, Philip M Long, and Peter L Bartlett. "When does gradient descent with logistic loss interpolate using deep networks with smoothed ReLU activations?" In: *Conference on Learning Theory*. 2021 (Cited on page 24).

[Che+20] Zixiang Chen, Yuan Cao, Quanquan Gu, and Tong Zhang. "A generalized neural tangent kernel analysis for two-layer neural networks". In: *Advances in Neural Information Processing Systems*. 2020 (Cited on page 24).

[Chi20] Lénaïc Chizat. *Analysis of gradient descent on wide two-layer ReLU neural networks*. Talk at MSRI. 2020. URL: https://www.msri.org/workshops/928/schedules/28397 (Cited on page 24).

[CB18] Lénaïc Chizat and Francis Bach. "On the global convergence of gradient descent for over-parameterized models using optimal transport". In: *Advances in Neural Information Processing Systems*. 2018, pp. 3036–3046 (Cited on page 24).

[CB20] Lénaïc Chizat and Francis Bach. "Implicit bias of gradient descent for wide two-layer neural networks trained with the logistic loss". In: *Conference on Learning Theory*. 2020 (Cited on page 24).

[COB19] Lénaïc Chizat, Edouard Oyallon, and Francis Bach. "On lazy training in differentiable programming". In: *Advances in Neural Information Processing Systems*. 2019, pp. 2937–2947 (Cited on pages 2, 24).

[Cor+09] Thomas H Cormen, Charles E Leiserson, Ronald L Rivest, and Clifford Stein. *Introduction to algorithms*. MIT Press, 2009 (Cited on page 3).

[Dan20] Amit Daniely. "Neural networks learning and memorization with (almost) no over-parameterization". In: *Advances in Neural Information Processing Systems*. 2020, pp. 9007–9016 (Cited on page 24).

[DM20] Amit Daniely and Eran Malach. "Learning parities with neural networks". In: *Advances in Neural Information Processing Systems* 33 (2020), pp. 20356–20365 (Cited on page 24).

[Du+19] Simon S Du, Jason D Lee, Haochuan Li, Liwei Wang, and Xiyu Zhai. "Gradient descent finds global minima of deep neural networks". In: *International Conference on Machine Learning*. 2019, pp. 1675–1685 (Cited on page 2).

[Du+18] Simon S Du, Xiyu Zhai, Barnabas Poczos, and Aarti Singh. "Gradient descent provably optimizes over-parameterized neural networks". In: *International Conference on Learning Representations*. 2018 (Cited on pages 2, 24).

[GLM18] Rong Ge, Jason D Lee, and Tengyu Ma. "Learning one-hidden-layer neural networks with landscape design". In: *International Conference on Learning Representations* (2018) (Cited on page 24).

[Gun+18a] Suriya Gunasekar, Jason D Lee, Daniel Soudry, and Nathan Srebro. "Characterizing implicit bias in terms of optimization geometry". In: *International Conference on Machine Learning*. 2018, pp. 1832–1841 (Cited on page 24).

[Gun+18b] Suriya Gunasekar, Jason D Lee, Daniel Soudry, and Nathan Srebro. "Implicit bias of gradient descent on linear convolutional networks". In: *Advances in Neural Information Processing Systems*. 2018, pp. 9461–9471 (Cited on page 24).

[Has+19] Trevor Hastie, Andrea Montanari, Saharon Rosset, and Ryan J Tibshirani. "Surprises in high-dimensional ridgeless least squares interpolation". In: *arXiv preprint arXiv:1903.08560* (2019) (Cited on page 24).

[HMX20] Daniel J Hsu, Vidya Muthukumar, and Ji Xu. "On the proliferation of support vectors in high dimensions". In: *arXiv preprint arXiv:2009.10670* (2020) (Cited on page 24).

[JGH18] Arthur Jacot, Franck Gabriel, and Clément Hongler. "Neural tangent kernel: Convergence and generalization in neural networks". In: *Advances in Neural Information Processing Systems*. 2018, pp. 8571–8580 (Cited on page 2).

[JT19a] Ziwei Ji and Matus Telgarsky. "Gradient descent aligns the layers of deep linear networks". In: *International Conference on Learning Representations*. 2019 (Cited on page 24).

[JT19b] Ziwei Ji and Matus Telgarsky. "Polylogarithmic width suffices for gradient descent to achieve arbitrarily small test error with shallow ReLU networks". In: *International Conference on Learning Representations*. 2019 (Cited on page 24).

[JT19c] Ziwei Ji and Matus Telgarsky. "The implicit bias of gradient descent on nonseparable data". In: *Conference on Learning Theory*. 2019, pp. 1772–1798 (Cited on page 24).

[JT20] Ziwei Ji and Matus Telgarsky. "Directional convergence and alignment in deep learning". In: *Advances in Neural Information Processing Systems*. 2020 (Cited on pages 3, 25).

[LM00] Beatrice Laurent and Pascal Massart. "Adaptive estimation of a quadratic functional by model selection". In: *The Annals of Statistics* (2000), pp. 1302–1338 (Cited on page 38).

[LL18] Yuanzhi Li and Yingyu Liang. "Learning overparameterized neural networks via stochastic gradient descent on structured data". In: *Advances in Neural Information Processing Systems*. 2018, pp. 8157–8166 (Cited on page 2).

[LMZ18] Yuanzhi Li, Tengyu Ma, and Hongyang Zhang. "Algorithmic regularization in over-parameterized matrix sensing and neural networks with quadratic activations". In: *Conference On Learning Theory*. 2018, pp. 2–47 (Cited on page 24).

[LY17] Yuanzhi Li and Yang Yuan. "Convergence analysis of two-layer neural networks with ReLU activation". In: *Advances in Neural Information Processing Systems*. 2017, pp. 597–607 (Cited on page 24).

[LR20] Tengyuan Liang and Alexander Rakhlin. "Just interpolate: Kernel "ridgeless" regression can generalize". In: *The Annals of Statistics* 48.3 (2020), pp. 1329–1347 (Cited on page 24).

[LRZ20] Tengyuan Liang, Alexander Rakhlin, and Xiyu Zhai. "On the multiple descent of minimum-norm interpolants and restricted lower isometry of kernels". In: *Conference on Learning Theory*. 2020, pp. 2683–2711 (Cited on page 24).

[LS20] Tengyuan Liang and Pragya Sur. "A precise high-dimensional asymptotic theory for boosting and min-ℓ_1-norm interpolated classifiers". In: *arXiv preprint arXiv:2002.01586* (2020) (Cited on page 24).

[LL20] Kaifeng Lyu and Jian Li. "Gradient descent maximizes the margin of homogeneous neural networks". In: *International Conference on Learning Representations*. 2020 (Cited on pages 2, 3, 25).

[MMM19] Song Mei, Theodor Misiakiewicz, and Andrea Montanari. "Mean-field theory of two-layers neural networks: dimension-free bounds and kernel limit". In: *Conference on Learning Theory*. 2019, pp. 2388–2464 (Cited on page 24).

[MM19] Song Mei and Andrea Montanari. "The generalization error of random features regression: Precise asymptotics and double descent curve". In: *arXiv preprint arXiv:1908.05355* (2019) (Cited on page 24).

[Mon+19] Andrea Montanari, Feng Ruan, Youngtak Sohn, and Jun Yan. "The generalization error of max-margin linear classifiers: High-dimensional asymptotics in the overparametrized regime". In: *arXiv preprint arXiv:1911.01544* (2019) (Cited on page 24).

[Mut+20a] Vidya Muthukumar, Adhyyan Narang, Vignesh Subramanian, Mikhail Belkin, Daniel J Hsu, and Anant Sahai. "Classification vs regression in overparameterized regimes: Does the loss function matter?" In: *arXiv preprint arXiv:2005.08054* (2020) (Cited on page 24).

[Mut+20b] Vidya Muthukumar, Kailas Vodrahalli, Vignesh Subramanian, and Anant Sahai. "Harmless interpolation of noisy data in regression". In: *IEEE Journal on Selected Areas in Information Theory* (2020) (Cited on page 24).

[NTS15] Behnam Neyshabur, Ryota Tomioka, and Nathan Srebro. "In search of the real inductive bias: On the role of implicit regularization in deep learning." In: *International Conference on Learning Representations (Workshop)*. 2015 (Cited on page 24).

[OS20] Samet Oymak and Mahdi Soltanolkotabi. "Towards moderate overparameterization: global convergence guarantees for training shallow neural networks". In: *IEEE Journal on Selected Areas in Information Theory* (2020) (Cited on page 2).

[PSZ18] Rina Panigrahy, Sushant Sachdeva, and Qiuyi Zhang. "Convergence results for neural networks via electrodynamics". In: *Innovations in Theoretical Computer Science*. 2018 (Cited on page 24).

[RZL18] Prajit Ramachandran, Barret Zoph, and Quoc V Le. "Searching for activation functions". In: *International Conference on Learning Representations (Workshop)*. 2018 (Cited on page 2).

[SS18] Itay Safran and Ohad Shamir. "Spurious local minima are common in two-layer ReLU neural networks". In: *International Conference on Machine Learning*. 2018, pp. 4433–4441 (Cited on page 24).

[Sou+18] Daniel Soudry, Elad Hoffer, Mor Shpigel Nacson, Suriya Gunasekar, and Nathan Srebro. "The implicit bias of gradient descent on separable data". In: *Journal of Machine Learning Research* 19.1 (2018), pp. 2822–2878 (Cited on page 24).

[Tat+20] Norman Tatro, Pin-Yu Chen, Payel Das, Igor Melnyk, Prasanna Sattigeri, and Rongjie Lai. "Optimizing mode connectivity via neuron alignment". In: *Advances in Neural Information Processing Systems*. 2020 (Cited on page 1).

[TB20] Alexander Tsigler and Peter L Bartlett. "Benign overfitting in ridge regression". In: *arXiv preprint arXiv:2009.14286* (2020) (Cited on page 24).

[Ver18] Roman Vershynin. *High-dimensional probability: An introduction with applications in data science*. Vol. 47. Cambridge university press, 2018 (Cited on page 33).

[Wai19] Martin J Wainwright. *High-dimensional statistics: A non-asymptotic viewpoint*. Vol. 48. Cambridge University Press, 2019 (Cited on pages 35, 38).

[Wei+19] Colin Wei, Jason D Lee, Qiang Liu, and Tengyu Ma. "Regularization matters: Generalization and optimization of neural nets vs. their induced kernel". In: *Advances in Neural Information Processing Systems*. 2019, pp. 9712–9724 (Cited on page 24).

[Zha+17a] Chiyuan Zhang, Samy Bengio, Moritz Hardt, Benjamin Recht, and Oriol Vinyals. "Understanding deep learning requires rethinking generalization". In: *International Conference on Learning Representations*. 2017 (Cited on pages 1, 24).

[Zha+19] Xiao Zhang, Yaodong Yu, Lingxiao Wang, and Quanquan Gu. "Learning one-hidden-layer ReLU networks via gradient descent". In: *International Conference on Artificial Intelligence and Statistics*. 2019, pp. 1524–1534 (Cited on page 24).

[Zha+17b] Yuchen Zhang, Jason D Lee, Martin Wainwright, and Michael Jordan. "On the learnability of fully-connected neural networks". In: *International Conference on Artificial Intelligence and Statistics*. 2017, pp. 83–91 (Cited on page 24).

[Zho+17] Kai Zhong, Zhao Song, Prateek Jain, Peter L Bartlett, and Inderjit S. Dhillon. "Recovery Guarantees for One-hidden-layer Neural Networks". In: *International Conference on Machine Learning*. 2017, pp. 4140–4149 (Cited on page 24).

[Zou+20] Difan Zou, Yuan Cao, Dongruo Zhou, and Quanquan Gu. "Gradient descent optimizes over-parameterized deep ReLU networks". In: *Machine Learning* 109.3 (2020), pp. 467–492 (Cited on page 2).